Stochastic Control

edited by
Chris Myers

Stochastic Control

Edited by Chris Myers

Contents

Preface

Uncertainty presents significant challenges in the reasoning about and controlling of complex dynamical systems. To address this challenge, numerous researchers are developing improved methods for stochastic analysis. This book presents a diverse collection of some of the latest research in this important area. In particular, this book gives an overview of some of the theoretical methods and tools for stochastic analysis, and it presents the applications of these methods to problems in systems theory, science, and economics.

The first section of the book presents theoretical methods and tools for the analysis of stochastic systems. The first two chapters by Sharma et al. present the Fokker-Planck equation and the Ito calculus. In Chapter 3, Charles presents the use of colored noise with stochastic differential equations. In Chapter 4, Lopez-Ruiz and Sanchez discuss coupled map lattices and cellular automata. In Chapter 5, Ohtsubo presents a game theoretic approach. In Chapter 6, Doria presents an approach that uses Hausdorff outer and inner measures. In Chapter 7, Warin and Vialle for analysis using distributed algorithms. Finally, in Chapter 8, Spiring explores the use of Mathematica7.

The second section of the book presents the application of stochastic methods in systems theory. In Chapter 9, Yang et al. present a learning algorithm for the parity problem. In Chapter 10, Nechval and Pugailis present an improved technique for state estimation. In Chapter 11, Serra presents a fuzzy identification method. In Chapter 12, Ferreira and Serra present an application of fuzzy methods to dynamic systems. The next three chapters by Yan et al., Perez et al., and Courmontagne explore the problem of filtering for stochastic systems. In Chapter 16, Olama et al. look at wireless fading channel models. In Chapter 17, Liang considers information flow and causality quantification. The last two chapters of this section by Zhang and Zhang and Sokolov consider control systems.

The third section of the book presents the application of stochastic methods to problems in science. In Chapter 20, Marano and Sgobba present design criteria for vibration control. In Chapter 21, Kaminski and Lauke consider reinforced elastomers. In Chapter 22, Alessandro and Francesco discuss structural design. In Chapter 23, Lam et al. apply stochastic methods to the modeling of earthquake ground motion. In Chapter 24, Rastovic addresses laser-plasma interactions. Finally, in Chapter 25, Kuwahara et al. apply new, efficient stochastic simulation methods to biological systems.

The final section of the book presents the application of stochastic methods to problems in economics. In Chapter 26, Nechval and Purgailis consider the problem of determining a products lifetime. In Chapter 27, Gontis et al. applies a stochastic model to financial markets. In Chapter 28, Mania et al. take on the problem of hedging in the market. In Chapter 29, Simovic and Simovic apply stochastic control approaches to tactical and strategic operations in the market. Finally, in Chapter 30, Darya et al. consider optimal control problems in fractional bio-economic systems.

Editor

Chris Myers
University of Utah
U.S.A.

The Fokker-Planck equation

Shambhu N. Sharma † and Hiren G. Patel‡
Department of Electrical Engineering†
National Institute of Technology, Surat, India
snsvolterra@gmail.com
Department of Electrical Engineering‡
National Institute of Technology, Surat, India
hgp@eed.svnit.ac.in

In 1984, H. Risken authored a book (H. Risken, *The Fokker-Planck Equation: Methods of Solution, Application*s, Springer-Verlag, Berlin, New York) discussing the Fokker-Planck equation for one variable, several variables, methods of solution and its applications, especially dealing with laser statistics. There has been a considerable progress on the topic as well as the topic has received greater clarity. For these reasons, it seems worthwhile again to summarize previous as well as recent developments, spread in literature, on the topic. The Fokker-Planck equation describes the evolution of conditional probability density for given initial states for a Markov process, which satisfies the Itô stochastic differential equation. The structure of the Fokker-Planck equation for the vector case is

$$\frac{\partial p(x,t|x_{t_o},t_0)}{\partial t} = -tr(\frac{\partial f(x,t)p(x,t|x_{t_0},t_0)}{\partial x}) + \frac{1}{2}tr(\frac{\partial^2 GG^T(x,t)p(x,t|x_{t_0},t_0)}{\partial x \partial x^T}),$$

where $f(x_t,t)$ is the system non-linearity, $G(x_t,t)$ is termed as the process noise coefficient, and $p(x,t|x_{t_0},t_o)$ is the conditional probability density. The Fokker-Planck equation, a prediction density evolution equation, has found its applications in developing prediction algorithms for stochastic problems arising from physics, mathematical control theory, mathematical finance, satellite mechanics, as well as wireless communications. In this chapter, the Authors try to summarize elementary proofs as well as proofs constructed from the standard theories of stochastic processes to arrive at the Fokker-Planck equation. This chapter encompasses an approximate solution method to the Fokker-Planck equation as well as a Fokker-Planck analysis of a Stochastic Duffing-van der Pol (SDvdP) system, which was recently analysed by one of the Authors.

Key words: The Duffing-van der Pol system, the Galerkin approximation, the Ornstein-Uhlenbeck process, prediction density, second-order fluctuation equations.

1. Introduction

The stochastic differential equation formalism arises from stochastic problems in diverse field, especially the cases, where stochastic problems are analysed from the dynamical systems' point of view. Stochastic differential equations have found applications in population dynamics, stochastic control, radio-astronomy, stochastic networks, helicopter rotor dynamics, satellite trajectory estimation problems, protein kinematics, neuronal activity, turbulence diffusion, stock pricing, seismology, statistical communication theory, and structural mechanics. A greater detail about stochastic differential equations' applications can be found in Kloeden and Platen (1991). Some of the standard structures of stochastic differential equations are the Itô stochastic differential equation, the Stratonovich stochastic differential equation, the stochastic differential equation involving p-differential, stochastic differential equation in Hida sense, non-Markovian stochastic differential equations as well as the Ornstein-Uhlenbeck (OU) process-driven stochastic differential equation. The Itô stochastic differential equation is the standard formalism to analyse stochastic differential systems, since non-Markovian stochastic differential equations can be re-formulated as the Itô stochastic differential equation using the extended phase space formulation, unified coloured noise approximation (Hwalisz *et al.* 1989). Stochastic differential systems can be analysed using the Fokker-Planck equation (Jazwinski 1970). The Fokker-Planck equation is a parabolic linear homogeneous differential equation of order two in partial differentiation for the transition probability density. The Fokker-Planck operator is an adjoint operator. In literature, the Fokker-Planck equation is also known as the Kolmogorov forward equation. The Kolmogorov forward equation can be proved using mild regularity conditions involving the notion of drift and diffusion coefficients (Feller 2000). The Fokker-Planck equation, definition of the conditional expectation, and integration by part formula allow to derive the evolution of the conditional moment. In the Risken's book, the stochastic differential equation involving the Langevin force was considered and subsequently, the Fokker-Planck equation was derived. The stochastic differential equation with the Langevin force can be regarded as the white noise-driven stochastic differential equation, where the input process satisfies $\langle w_t \rangle = 0, \langle w_t w_s \rangle = \delta(t - s)$. He considered the approximate solution methods to the scalar and vector Fokker-Planck equations involving change of variables, matrix continued-fraction method, numerical integration method, etc. (Risken 1984, p. 158). Further more, the laser Fokker-Planck equation was derived.

This book chapter is devoted to summarize alternative approaches to derive the Fokker-Planck equation involving elementary proofs as well as proofs derived from the Itô differential rule. In this chapter, the Fokker-Planck analysis hinges on the stochastic differential equation in the Itô sense in contrast to the Langevin sense. From the mathemacians' point of view, the Itô stochastic differential equation involves rigorous interpretation in contrast to the Langevin stochastic differential equation. On the one hand, the stochastic differential equation in Itô sense is described as $dx_t = f(x_t, t)dt + G(x_t, t)dB_t$, on the other, the Langevin stochastic differential equation assumes the structure $\dot{x}_t = f(x_t, t) + G(x_t, t)w_t$, where B_t and w_t are the Brownian and white noises respectively. The white noise can be regarded as an informal

non-existent time derivative $\dot{B}_t$ of the Brownian motion B_t. Kiyoshi Itô, a famous Japanese mathematician, considered the term $'dB_t'=\dot{B}_t dt$ and developed Itô differential rule. The results of Itô calculus were published in two seminal papers of Kiyoshi Itô in 1945. The approach of this chapter is different and more exact in contrast to the Risken's book in the sense that involving the Itô stochastic differential equation, introducing relatively greater discussion on the Kolmogorov forward and Backward equations. This chapter discusses a Fokker-Planck analysis of a stochastic Duffing-van der Pol system, an appealing case, from the dynamical systems' point of view as well.

This chapter is organised as follows: (i) section 2 discusses the evolution equation of the prediction density for the Itô stochastic differential equation. A brief discussion about approximate methods to the Fokker-Planck equation, stochastic differential equation is also given in section 2 (ii) in section 3, the stochastic Duffing-van der Pol system was analysed to demonstrate a usefulness of the Fokker-Planck equation. (iii) Section 4 is about the numerical simulation of the mean and variance evolutions of the SDvdP system. Concluding remarks are given in section (5).

2. Evolution of conditional probability density

The Fokker-Planck equation describes the evolution of conditional probability density for given initial states for the Itô stochastic differential system. The equation is also known as the prediction density evolution equation, since it can be utilized to develop prediction algorithms, especially where observations are not available at every time instant. One of the potential applications of the Fokker-Planck equation is to develop estimation algorithms for the satellite trajectory estimation. This chapter summarizes four different proofs to arrive at the Fokker-Planck equation. The first two proofs can be regarded as elementary proofs and the last two utilize the Itô differential rule. Moreover, the Fokker-Planck equation for the OU process-driven stochastic differential equation is discussed here, where the input process has non-zero, finite, relatively smaller correlation time.

The *first proof* of this chapter begins with the Chapman-Kolmogorov equation. The Chapman-Kolmogorov equation is a consequence of the theory of the Markov process. This plays a key role in proving the Kolmogorov backward equation (Feller 2000). Here, we describe briefly the Chapman-Kolmogorov equation and subsequently, the concept of the conditional probability density as well as transition probability density are introduced to derive the evolution of conditional probability density for the non-Markov process. The Fokker-Planck equation becomes a special case of the resulting equation. The conditional probability density

$$p(x_1, x_2 | x_3) = p(x_1 | x_2, x_3)....p(x_2 | x_3).$$

Consider the random variables $x_{t_1}, x_{t_2}, x_{t_3}$ at the time instants t_1, t_2, t_3, where $t_1 > t_2 > t_3$ and take values x_1, x_2, x_3. In the theory of the Markov process, the above can be re-stated as

$$p(x_1, x_2 | x_3) = p(x_1 | x_2) p(x_2 | x_3),$$

integrating over the variable x_2, we have

$$p(x_1|x_3) = \int p(x_1|x_2)p(x_2|x_3)dx_2,$$

introducing the notion of the transition probability density and time instants

$$q_{t+s}(x_1,x_3) = \int q_t(x_1,x_2)q_s(x_2,x_3)dx_2.$$

Consider the multi-dimensional probability density $p(x_1,x_2) = p(x_1|x_2)p(x_2)$ and integrating over the variable x_2, we have

$$p(x_1) = \int p(x_1|x_2)p(x_2)dx_2,$$

or

$$p(x_1) = \int q_{t_1,t_2}(x_1,x_2)p(x_2)dx_2, \qquad (1)$$

where $q_{t_1,t_2}(x_1,x_2)$ is the transition probability density and $t_1 > t_2$. The transition probability density $q_{t_1,t_2}(x_1,x_2)$ is the inverse Fourier transform of the characteristic function $Ee^{iu(x_{t_1}-x_{t_2})}$, i.e.

$$q_{t_1,t_2}(x_1,x_2) = \frac{1}{2\pi}\int e^{-iu(x_1-x_2)}Ee^{iu(x_{t_1}-x_{t_2})}du. \qquad (2)$$

Equation (1) in combination with equation (2) leads to

$$p(x_1) = \frac{1}{2\pi}\int e^{-iu(x_1-x_2)}(Ee^{iu(x_{t_1}-x_{t_2})})p(x_2)dx_2du. \qquad (3)$$

The characteristic function is the moment generating function, the characteristic function $Ee^{iu(x_{t_1}-x_{t_1})} = \sum_{0\leq n}\frac{(iu)^n}{n!}\langle(x_{t_1}-x_{t_2})^n\rangle$. After introducing the definition of the characteristic function, equation (3) can be recast as

$$p(x_1) = \frac{1}{2\pi}\int e^{-iu(x_1-x_2)}(\sum_{0\leq n}\frac{(iu)^n}{n!}\langle(x_{t_1}-x_{t_2})^n\rangle)p(x_2)dx_2du$$

$$= \sum_{0\leq n}\int \frac{1}{n!}(\frac{1}{2\pi}\int(iu)^n e^{-iu(x_1-x_2)}du)\langle(x_{t_1}-x_{t_2})^n\rangle p(x_2)dx_2.$$

The term $\dfrac{1}{2\pi}\int e^{-iu(x_1-x_2)}(iu)^n\,du = (-\dfrac{\partial}{\partial x_1})^n\delta(x_1-x_2)$ and leads to the probability density

$$p(x_1)=\sum_{0\leq n}\int\frac{1}{n!}(-\frac{\partial}{\partial x_1})^n\delta(x_1-x_2)\Big\langle(x_{t_1}-x_{t_2})^n\Big\rangle p(x_2)dx_2. \qquad (4)$$

For the short hand notation, introducing the notion of the stochastic process, taking $x_1=x_\tau, x_2=x$, where the time instants $t_1=t+\tau, t_2=t$, equation (4) can be recast as

$$p(x_\tau)=\sum_{0\leq n}\int\frac{1}{n!}(-\frac{\partial}{\partial x_\tau})^n\delta(x_\tau-x)\Big\langle(x_\tau-x)^n\Big\rangle p(x)dx$$

$$=\sum_{0\leq n}\int\frac{1}{n!}(-\frac{\partial}{\partial x_\tau})^n\delta(x_\tau-x)k_n(x)\tau\,p(x)dx,$$

where $\left\langle\dfrac{(x_\tau-x)^n}{\tau}\right\rangle=k_n(x)$ and the time interval condition $\tau\to 0$ leads to

$$\underset{\tau\to 0}{Lt}\frac{p(x_\tau)-p(x)}{\tau}=\sum_{1\leq n}\frac{1}{n!}(-\frac{\partial}{\partial x_\tau})^n k_n(x)p(x),$$

or

$$\dot{p}(x)=\sum_{1\leq n}\frac{1}{n!}(-\frac{\partial}{\partial x})^n k_n(x\,)p(x).$$

Note that the above density evolution equation is derived for the arbitrary stochastic process $X=(x_t, 0\leq t<\infty)$. Here, the arbitrary process means that there is no restriction imposed on the process while deriving the density evolution equation and can be regarded as the non-Markov process. Consider a Markov process, which satisfies the Itô stochastic differential equation, the evolution of conditional probability density retains only the first two terms $k_1(x)$ and $k_2(x)$, which is a direct consequence of the stochastic differential rule for the Itô stochastic differential equation in combination with the definition $\left\langle\dfrac{(x_\tau-x)^n}{\tau}\right\rangle=k_n(x)$. As a result of these, the evolution of conditional probability density for the scalar stochastic differential equation of the form

$$dx_t=f(x_t, t)dt+g(x_t, t)dB_t,$$

leads to the Fokker-Planck equation,

$$\dot{p}(x) = -\frac{\partial}{\partial x} f(x, t) p(x) + \frac{1}{2} \frac{\partial^2 g^2(x, t)}{\partial x^2} p(x),$$

where $k_1(x) = f(x, t)$, $k_2(x) = g^2(x, t)$. The Fokker-Planck equation can be recast as $dp(x) = Lp(x)dt$, where the vector version of the Fokker-Planck operator

$$L(.) = -\sum_i \frac{\partial}{\partial x_i} f_i(x, t)(.) + \frac{1}{2} \sum_{i,j} \frac{\partial^2 (GG^T)_{ij}(x, t)(.)}{\partial x_i \partial x_j}.$$

The Fokker-Planck operator is an adjoint operator, since $\langle \phi, L(p) \rangle = \langle L'\phi, p \rangle$, where $L'(.)$ is the Kolmogorov backward operator. This property is utilized in deriving the evolution $d\hat{\phi}(x_t)$ of the conditional moment (Jazwinski 1970). The Fokker-Planck equation is also known as the Kolmogorov Forward equation.

The *second proof* of this chapter begins with the Green function, the Kolmogorov forward and backward equations involve the notion of the drift and diffusion coefficients as well as mild regularity conditions (Feller 2000). The drift and diffusion coefficients are regarded as the system non-linearity and the 'stochastic perturbation in the variance evolution' respectively in noisy dynamical system theory. Here, we explain briefly about the formalism associated with the proof of the Kolmogorov forward and backward equations. Consider the Green's function

$$u_t(x) = \int q_t(x, y) u_0(y) dy, \tag{5}$$

where $q_t(x, y)$ is the transition probability density, $u_t(x)$ is a scalar function, x is the initial point and y is the final point. Equation (5) is modified at the time duration $t + h$ as

$$u_{t+h}(x) = \int q_{t+h}(x, y) u_0(y) dy. \tag{6}$$

The Chapman-Kolmogorov equation can be stated as

$$q_{t+h}(x, y) = \int q_t(x, \varsigma) q_h(\varsigma, y) d\varsigma. \tag{7}$$

Making the use of equations (6)-(7) and the Taylor series expansion with mild regularity conditions leads to

$$\frac{\partial u_t(x)}{\partial t} = b(x)\frac{\partial u_t(x)}{\partial x} + \frac{1}{2}a(x)\frac{\partial^2 u_t(x)}{\partial x^2}, \tag{8}$$

where $\dfrac{\partial u_t(x)}{\partial t} = \underset{h \to 0}{Lt} \dfrac{u_{t+h} - u_t}{h}$, $b(x)$ and $a(x)$ are the drift and diffusion coefficients respectively (Feller 2000), and the detailed proof of equation (8) can be found in a celebrated book authored by Feller (2000). For the vector case, the Kolmogorov backward equation can be recast as

$$\frac{\partial u_t(x)}{\partial t} = \sum b_i(x)\frac{\partial u_t(x)}{\partial x_i} + \sum \frac{1}{2}a_{ij}(x)\frac{\partial^2 u_t(x)}{\partial x_i \partial x_j},$$

where the summation is extended for $1 \le i \le n, 1 \le j \le n$. From the dynamical systems' point of view, the vector case of the Kolmogorov backward equation can be reformulated as

$$\frac{\partial u_t(x)}{\partial t} = \sum f_i(x,\,t)\frac{\partial u_t(x)}{\partial x_i} + \frac{1}{2}\sum (GG)_{ij}(x,\,t)\frac{\partial^2 u_t(x)}{\partial x_i \partial x_j},$$

where the mappings f and G are the system non-linearity and process noise coefficient matrix respectively and the Kolmogorov backward operator

$$L'(.) = \sum f_i(x,t)\frac{\partial(.)}{\partial x_i} + \sum_i \frac{1}{2}(GG)_{ij}(x,\,t)\frac{\partial^2(.)}{\partial x_i \partial x_j}.$$

Note that the Kolmogorov backward equation is a parabolic linear homogeneous differential equation of order two in partial differentiation, since the backward operator is a linear operator and the homogeneity condition holds. The Kolmogorov forward equation can be derived using the relation

$$v_s(y) = \int q_s(x,y)v_0(x)dx,$$

in combination with integration by part formula as well as mild regularity conditions (Feller 2000) lead to the expression

$$\frac{\partial v_s(y)}{\partial s} = -\frac{\partial b(y)v_s(y)}{\partial y} + \frac{1}{2}\frac{\partial^2 a(y)v_s(y)}{\partial y^2}. \tag{9}$$

The terms $b(y)$ and $a(y)$ of equation (9) have similar interpretations as the terms of equation (8). The vector version of equation (9) is

$$\frac{\partial v_s(y)}{\partial s} = -\sum_i \frac{\partial b_i(y)v_s(y)}{\partial y_i} + \frac{1}{2}\sum_{i,j}\frac{\partial^2 a_{ij}(y)v_s(y)}{\partial y_i \partial y_j},$$

and the Kolmogorov forward operator $L(.) = -\sum_i \frac{\partial b_i(y)(.)}{\partial y_i} + \frac{1}{2}\sum_{i,j}\frac{\partial^2 a_{ij}(y)(.)}{\partial y_i \partial y_j}$. For

$b_i = f_i$ and $a_{ij} = (GG^T)_{ij}$, the Kolmogorov forward operator assumes the structure of the Fokker-Planck operator and is termed as the Kolmogorov-Fokker-Planck operator. The *third proof* of the chapter explains how the Fokker-Planck equation can be derived using the definition of conditional expectation and Itô differential rule.

$$E\phi(x_{t+dt}) = E(E(\phi(x_{t+dt})|x_t = x)). \tag{10}$$

The Taylor series expansion of the scalar function $\phi(x_{t+dt}) = \sum_m \frac{(dx_t)^m}{m!}\phi^m(x_t)$ and

$$E(\phi(x_{t+dt})|x_t = x) = \sum_{0 \le m \le 2} \frac{\mu_m(x,t)}{m!}\phi^m(x), \tag{11}$$

where $\mu_m(x,t) = E((dx_t)^m|x_t = x)$ and the summing variable m takes values upto two for the Brownian motion process-driven stochastic differential equation that can be explained via the Itô differential rule. Equation (10) in conjunction with equation (11) leads to

$$E\phi(x_{t+dt}) = \sum_{0 \le m \le 2} E(\frac{\mu_m(x,t)}{m!}\phi^m(x)),$$

the definition of 'expectation' leads to the following expression:

$$\int \phi(x)p(x,t+dt)dx = \sum_{0 \le m \le 2}\int \frac{\mu_m(x,t)}{m!}\phi^m(x)p(x,t)dx, \tag{12}$$

the integration by part, applying to equation (12), leads to the Fokker-Planck equation,

$$\frac{\partial p(x,t)}{\partial t} = \sum_{1 \le m \le 2}\frac{(-1)^m}{m!}\frac{\partial^m \mu_m(x,t)p(x,t)}{\partial x^m},$$

where

$$\mu_1 = f(x,t), \ \mu_2 = g^2(x,t).$$

Finally, we derive the Fokker-Planck equation using the concept of the evolution of the conditional moment and the conditional characteristic function. Consider the state vector $x_t \in U$, $\phi : U \to R$, i.e. $\phi(x_t) \in R$, and the phase space $U \subset R^n$. The state vector x_t satisfies the Itô SDE as well. Suppose the function $\phi(x_t)$ is twice differentiable. The evolution $d\widehat{\phi}(x_t)$ of the conditional moment is the standard formalism to analyse stochastic differential systems. Further more, $d\widehat{\phi}(x_t) = E(d\,\phi(x_t)\big| x_{t_0}, t_0)$ holds. A greater detail can be found in Sharma (2008). The stochastic evolution $d\phi(x_t)$ of the scalar function $\phi(x_t)$ (Sage and Melsa 1971) can be stated as

$$d\phi(x_t) = (\sum_i \frac{\partial \phi(x_t)}{\partial x_i} f_i(x_t, t) + \frac{1}{2}\sum_i (GG^T)_{ii}(x_t, t)\frac{\partial^2 \phi(x_t)}{\partial x_i^2}$$

$$+ \sum_{i<j}(GG^T)_{ij}(x_t, t)\frac{\partial^2 \phi(x_t)}{\partial x_i \partial x_j})dt + \sum_{1 \le i \le n, 1 \le \gamma \le r}\frac{\partial \phi(x_t)}{\partial x_i}G_{i\gamma}(x_t, t)dB_\gamma, \qquad (13)$$

thus

$$d\widehat{\phi}(x_t) = (\sum_p \overbrace{f_p(x_t,t)\frac{\partial \phi(x_t)}{\partial x_p}}^{\wedge} + \frac{1}{2}\sum_p \overbrace{(GG^T)_{pp}(x_t, t)\frac{\partial^2 \phi(x_t)}{\partial x_p^2}}^{\wedge} + \sum_{p<q}\overbrace{(GG^T)_{pq}(x_t, t)\frac{\partial^2 \phi(x_t)}{\partial x_p \partial x_q}}^{\wedge})dt.$$

Note that the expected value of the last term of the right-hand side of equation (13) vanishes, i.e. $\langle G_{i\gamma}(x_t, t)dB_\gamma \rangle = 0$. Consider $\phi(x_t) = e^{S^T x_t}$, the evolution of the characteristic function becomes

$$dE(e^{S^T x_t}) = (\sum_p \overbrace{S_p f_p(x_t,t)e^{S^T x_t}}^{\wedge} + \frac{1}{2}\sum_p \overbrace{S_p^2(GG^T)_{pp}(x_t, t)\frac{\partial^2 \phi(x_t)}{\partial x_p^2}}^{\wedge} + \sum_{p<q}\overbrace{S_p S_q(GG^T)_{pq}(x_t, t)e^{S^T x_t}}^{\wedge})dt.$$

Making the use of the definition of the characteristic function as well as the integration by part formula, we arrive at the Fokker-Planck equation.
The *Kushner* equation, the filtering density evolution equation for the Itô stochastic differential equation, is a 'generalization' of the Fokker-Planck equation. The Kushner equation is a partial-integro stochastic differential equation, i.e.

$$dp = L(p)dt + (h - \widehat{h})^T \varphi_n^{-1}(dz_t - \widehat{h}dt)p, \qquad (14)$$

where $L(.)$ is the Fokker-Planck operator, $p = p(x,t|z_\tau, t_0 \le \tau \le t)$, the observation

$$z_t = \int_{t_0}^t h(x_\tau, \tau) d\tau + B_t, \quad \text{and} \quad h(x_t, t) \text{ is the measurement non-linearity. Harald J}$$

Kushner first derived the expression of the filtering density and subsequently, the filtering density evolution equation using the stochastic differential rule (Jazwinski 1970). Liptser-Shiryayev discovered an alternative proof of the filtering density evolution, equation (14), involving the following steps: (i) derive the stochastic evolution $d\widehat{\phi(x_t)}$ of the conditional moment, where $\widehat{\phi(x_t)} = E(\phi(x_t)|z_\tau, t_0 \le \tau \le t)$ (ii) subsequently, the stochastic evolution of the conditional characteristic function can be regarded as a special case of the conditional moment evolution, where $\phi(x_t) = e^{S^T x_t}$ (iii) the definition of the conditional expectation as well as integration by part formula lead to the filtering density evolution equation, see Liptser and Shiryayev (1977). RL Stratonovich developed the filtering density evolution for stochastic differential equation involving the $\frac{1}{2}$-differential as well. For this reason, the filtering density evolution equation is also termed as the Kushner-Stratonovich equation.

Consider the stochastic differential equation of the form

$$\dot{x}_t = f(x_t) + g(x_t)\xi_t, \tag{15}$$

where ξ_t is the Ornstein-Uhlenbeck process and generates the process x_t, a non-Markov process. The evolution of conditional probability density for the non-Markov process with the input process with a non-zero, finite, smaller correlation time τ_{cor}, i.e. $0 < \tau_{cor} << 1$, reduces to the Fokker-Planck equation. One of the approaches to arrive at the Fokker-Planck equation for the OU process-driven stochastic differential equation with smaller correlation time is function calculus. The function calculus approach involves the notion of the functional derivative. The evolution of conditional probability density for the output process x_t, where the input process ξ_t is a zero mean, stationary and Gaussian process, can be written (Hänggi 1995, p.85) as

$$\dot{p}_t(x) = -\frac{\partial f(x) p}{\partial x} + \frac{\partial}{\partial x} g(x) \left(\frac{\partial}{\partial x} \int_{t_o}^t C_2(t-s) \left\langle \delta(x_t - x) \frac{\delta x_t}{\delta \xi_s} \right\rangle ds \right), \tag{16}$$

where the second-order cumulant of the zero mean, stationary and Gaussian process is

$$C_2(t,s) = \text{cov}(\xi_t, \xi_s) = R_{\xi\xi}(t-s) \text{ and } \frac{\delta x_t}{\delta \xi_s} \text{ is the functional derivative of the process}$$

x_t with respect to the input process ξ_s. The integral counterpart of equation (15) is

$$x_t = x_{t_0} + \int_{t_0}^{t} f(x_\tau) + g(x_\tau)\xi_\tau d\tau.$$

The functional derivative $\dfrac{\delta x_t}{\delta \xi_s}$ depends on the time interval $s \le \tau \le t$ and can be stated as

$$\frac{\delta x_t}{\delta \xi_s} = \int_{s}^{t}(f'(x_\tau)\frac{\delta x_\tau}{\delta \xi_s} + g'(x_\tau)\frac{\delta x_\tau}{\delta \xi_s}\xi_\tau + g(x_\tau)\frac{\delta \xi_\tau}{\delta \xi_s})d\tau,$$

$$= g(x_s) + \int_{s}^{t}(f'(x_\tau)\frac{\delta x_\tau}{\delta \xi_s} + g'(x_\tau)\frac{\delta x_\tau}{\delta \xi_s}\xi_\tau)d\tau, \tag{17}$$

Making the repetitive use of the expression $\dfrac{\delta x_t}{\delta \xi_s}$ within the integral sign of equation (17), we have

$$\frac{\delta x_t}{\delta \xi_s} = g(x_s)\exp(\int_{s}^{t}\frac{\partial \dot{x}_\tau}{\partial x_\tau}d\tau) \;,$$

$$= g(x_s)\exp(\int_{s}^{t}(f'(x_\tau) + g'(x_\tau)\xi_\tau d\tau)). \tag{18}$$

More over, the time derivative of the process noise coefficient $g(x_t)$ of equation (15) can be written as

$$\dot{g}(x_t) = g'(x_t)\dot{x}_t$$
$$= g'(x_t)(f(x_t) + g(x_t)\xi_t),$$

after some calculations, the integral counterpart of the above equation can be stated as

$$g(x_s) = g(x_t)\exp(-(\int_{s}^{t}(\frac{g'(x_\tau)f(x_\tau)}{g(x_\tau)} + g'(x_\tau)\xi_\tau)d\tau)). \tag{19}$$

Equation (18) in combination with equation (19) leads to

$$\frac{\delta x_t}{\delta \xi_s} = g(x_t) \exp(\int_s^t (f'(x_\tau) - g'(x_\tau) \frac{f(x_\tau)}{g(x_\tau)}) d\tau). \tag{20}$$

Furthermore, the Taylor series expansion of the functional derivative $\dfrac{\delta x_t}{\delta \xi_s}$ in powers of

$(s-t)$ can be stated as

$$\frac{\delta x_t}{\delta \xi_s} = \frac{\delta x_t}{\delta \xi_t} + (s-t)(\frac{\partial}{\partial s}(\frac{\delta x_t}{\delta \xi_s})\big| s = t) + O((s-t)^2). \tag{21}$$

From equation (20), we have

$$\frac{\delta x_t}{\delta \xi_t} = g(x_t), \tag{22}$$

$$\frac{\partial}{\partial s}\frac{\delta x_t}{\delta \xi_s}\big| s = t = -g(x_t)(f'(x_t) - g'(x_t)\frac{f(x_t)}{g(x_t)}). \tag{23}$$

After retaining the first two terms of the right-hand side of equation (21) and equations (22)-(23) in combination with equation (21) lead to

$$\frac{\delta x_t}{\delta \xi_s} = g(x_t) + (t-s)\, g(x_t)(f'(x_t) - g'(x_t)\frac{f(x_t)}{g(x_t)})$$

$$= g(x_t)(1 + (t-s)\, (\frac{f'(x_t)g(x_t) - g'(x_t)f(x_t)}{g(x_t)})),$$

thus

$$\left\langle \delta(x_t - x)\frac{\delta x_t}{\delta \xi_s} \right\rangle = g(x)(1 + (t-s)g(x)(\frac{f(x)}{g(x)})')\, p(x). \tag{24}$$

The autocorrelation $R_{\xi\xi}(t-s)$ of the OU process satisfying the stochastic differential

equation $d\xi_t = \dfrac{-1}{\tau_{cor}}\xi_t dt + \dfrac{\sqrt{2D}}{\tau_{cor}}dB_t$ becomes

$$R_{\xi\xi}(t-s) = \frac{D}{\tau_{cor}}e^{-\frac{|t-s|}{\tau_{cor}}}. \tag{25}$$

Equations (24)-(25) in conjunction with equation (16) give

$$\dot{p}(x) = -\frac{\partial}{\partial x} fp + D \frac{\partial}{\partial x} (g \frac{\partial}{\partial x} g(1 + \tau_{cor} g(\frac{f}{g})')p).$$

The Kolmogorov-Fokker-Planck equation and the Kolmogorov backward equation are exploited to analyse the Itô stochastic differential equation by deriving the evolution of the conditional moment. The evolutions of conditional mean and variance are the special cases of the conditional moment evolution. The conditional mean and variance evolutions are infinite dimensional as well as involve higher-order moments. For these reasons, approximate mean and variance evolutions are derived and examined involving numerical experiments. Alternatively, the Carleman linearization to the exact stochastic differential equation resulting the bilinear stochastic differential equation has found applications in developing the approximate estimation procedure. The Carleman linearization transforms a finite dimensional non-linear system into a system of infinite dimensional linear systems (Kowalski and Steeb 1991).

The exact solution of the Fokker-Planck equation is possible for the simpler form of the stochastic differential equation, e.g.

$$dx_t = adB_t. \tag{26}$$

The Fokker-Planck equation for equation (26) becomes

$$\frac{\partial p(x, t | x_{t_0}, t_0)}{\partial t} = \frac{1}{2} a^2 \frac{\partial^2 p(x, t | x_{t_0}, t_0)}{\partial x^2}.$$

Consider the process $N(0, a^2 t)$ and its probability density

$$p(x, t | x_{t_0}, t_0) = \frac{1}{\sqrt{2\pi a^2 t}} e^{-\frac{x^2}{2a^2 t}}$$

satisfies equation (26). However, the closed-form solution to the Fokker-Planck equation for the non-linear stochastic differential equation is not possible, the approximate solution to the Fokker-Planck equation is derived. The Galerkin approximation to the Fokker-Planck equation received some attention in literature. The Galerkin approximation can be applied to the Kushner equation as well. More generally, the usefulness of the Galerkin approximation to the partial differential equation and the stochastic differential equation for the approximate solution can be explored. The theory of the Galerkin approximation is grounded on the orthogonal projection lemma. For a greater detail, an authoritative book, computational Galerkin methods, authored by C A J Fletcher can be consulted (Fletcher 1984).

3. A stochastic Duffing-van der Pol system

The second-order fluctuation equation describes a dynamical system in noisy environment. The second-order fluctuation equation can be regarded as

$$\ddot{x}_t = F(t, x_t, \dot{x}_t, \dot{B}_t).$$

The phase space formulation allows transforming a single equation of order n into a system of n first-order differential equations. Choose $x_t = x_1$

$$\dot{x}_1 = x_2,$$
$$\dot{x}_2 = F(t, x_1, x_2, \dot{B}_t),$$

by considering a special case of the above system of equations, we have

$$dx_1 = x_2 dt,$$
$$dx_2 = f_2(t, x_1, x_2)dt + g_2(t, x_1, x_2)dB_t,$$

in matrix-vector format

$$d\xi_t = f(t, x_1, x_2)dt + G(t, x_1, x_2)dB_t, \tag{27}$$

where

$$\xi_t = (\xi_1, \xi_2)^T = (x_1, x_2)^T, \ f(t, \xi_t) = (x_2, f_2)^T, \ G(t, \xi_t) = (0_2, g_2)^T.$$

The stochastic Duffing-van der Pol system can be formulated in the form of equation (27) (Sharma 2008), where

$$f(\xi_t, t) = \begin{pmatrix} x_2 \\ \alpha x_1 + \beta x_2 - a x_1^3 - b x_2 x_1^2 \end{pmatrix}, \ G(\xi_t, t) = \begin{pmatrix} 0 \\ \sigma_B x_1^n \end{pmatrix}, \tag{28}$$

and $G(\xi_t, t)$ is the process noise coefficient matrix. The Fokker-Planck equation can be stated as (Sage and Melsa 1971, p.100)

$$\frac{\partial p(\xi, t | \xi_{t_0}, t_0)}{\partial t} = -\sum_i \frac{\partial f_i(\xi, t) p(\xi, t | \xi_{t_0}, t_0)}{\partial \xi_i}$$

$$+\frac{1}{2}\sum_{i,j}\frac{\partial^2(GG^T)_{i,j}(\xi,t)p(\xi,t|\xi_{t_0},t_0)}{\partial\xi_i\partial\xi_j}, \quad (29)$$

where $\xi=(x_1,x_2,...,x_n)^T$. Equation (29) in combination with equation (28) leads to the Fokker-Planck equation for the stochastic system of this chapter, i.e.

$$\frac{\partial p}{\partial t}=-x_2\frac{\partial p}{\partial x_1}-\alpha x_1\frac{\partial p}{\partial x_2}-\beta x_2\frac{\partial p}{\partial x_2}-\beta p+ax_1^3\frac{\partial p}{\partial x_2}+bx_1^2 p$$

$$+bx_2x_1^2\frac{\partial p}{\partial x_2}+\frac{\sigma_B^2}{2}(2n)(2n-1)x_1^{2n-2}p.$$

Alternatively, the stochastic differential system can be analysed qualitatively involving the Itô differential rule, see equation (13) of the chapter. The energy function for the stochastic system of this chapter is

$$E(x_1,x_2)=\frac{1}{2}x_2^2-\alpha\frac{x_1^2}{2}+a\frac{x_1^4}{4}. \quad (30)$$

From equations (13), (28), and (30), we obtain

$$dE(x_1,x_2)=((-\alpha x_1+ax_1^3)x_2+x_2(\alpha x_1+\beta x_2-ax_1^3-bx_2x_1^2)+\frac{1}{2}\sigma_B^2x_1^{2n})dt$$

$$+\sigma_B x_1^n dw_t.$$

After a simple calculation, we have the following SDE:

$$dE=(\beta x_2^2-bx_2^2x_1^2+\frac{1}{2}\sigma_B^2x_1^{2n}+\frac{1}{2}\sigma_u^2)dt+\sigma_B x_1^n dw_t+\sigma_u dv_t.$$

The qualitative analysis of the stochastic problem of this chapter using the multi-dimensional Itô differential rule illustrates the contribution of diffusion parameters to the stochastic evolution of the energy function. The energy evolution equation suggests the system will exhibit either increasing oscillations or decreasing depending on the choice of the parameters $\beta, b,$ and the diffusion parameters σ_B, σ_u. The numerical experiment also confirms the qualitative analysis of this chapter, see figures (1)-(2). This chapter discusses a Fokker-Planck analysis of the SDvdP system, recently analysed and published by one of the Authors (Sharma 2008).

Making use of the Fokker-Planck equation, Kolmogorov backward equation, the evolutions of condition mean and variances (Jazwinski 1970, p. 363) can be stated as

$$d\widehat{x}_i = \overbrace{f_i(x_t, t)}^{\Lambda}dt, \tag{31}$$

$$dP_{ij} = (\overbrace{x_i f_j}^{\Lambda} - \hat{x}_i \hat{f}_j + \overbrace{f_i x_j}^{\Lambda} - \hat{f}_i \hat{x}_j + \overbrace{(GG^T)_{ij}(x_t,t)}^{\Lambda})dt, \tag{32}$$

where the state vector $x_t = (x_1, x_2, ..., x_n)^T$, the component-wise stochastic differential equation is

$$dx_i(t) = f_i(x_t, t)dt + \sum_{\phi} G_{i\phi}(x_t, t)dB_{\phi},$$

and

$$\widehat{x}_i = E(x_i(t)\big|x_{t_0}, t_0), \quad P_{ij} = E(\widetilde{x}_i \widetilde{x}_j \big| x_{t_0}, t_0),$$

$$\widetilde{x}_i = x_i - E(x_i\big|x_{t_0}, t_0), \quad \overbrace{f_i(x_t,t)}^{\Lambda} = E(f_i(x_t, t)\big|x_{t_0}, t_0),$$

and $f(x_t, t)$ is the system non-linearity and $G(x_t, t)$ is the dispersion matrix. The dispersion matrix is also known as the process noise coefficient matrix in mathematical control theory. The mean and variance evolutions using the third-order approximation can be derived involving the following: (i) first, develop the conditional moment evolution $d\widehat{\phi}(x_t)$ using the Fokker-Planck equation (ii) secondly, the mean and variance evolutions can be derived by considering $\phi(x_t)$ as x_i and $\widetilde{x}_i\widetilde{x}_j$ respectively (iii) finally, the third-order partials of the system non-linearity and diffusion coefficient are introduced into the exact evolution equations (31) and (32). Thus, we have

$$d\widehat{x}_i = (f_i(\widehat{x}_t, t) + \frac{1}{2}\sum_{p, q} P_{pq} \frac{\partial^2 f_i(\widehat{x}_t, t)}{\partial \widehat{x}_p \partial \widehat{x}_q})dt, \tag{33}$$

$$(dP_t)_{ij} = (\sum_p P_{ip} \frac{\partial f_j(\widehat{x}_t, t)}{\partial \widehat{x}_p} + \sum_p P_{jp} \frac{\partial f_i(\widehat{x}_t, t)}{\partial \widehat{x}_p} + \frac{1}{2}\sum_{p, q, r} P_{ip}P_{qr} \frac{\partial^3 f_j(\widehat{x}_t, t)}{\partial \widehat{x}_p \partial \widehat{x}_q \partial \widehat{x}_r}$$

$$+ \frac{1}{2}\sum_{p, q, r} P_{jp}P_{qr} \frac{\partial^3 f_i(\widehat{x}_t,t)}{\partial \widehat{x}_p \partial \widehat{x}_q \partial \widehat{x}_r} + (GG^T)_{ij}(\widehat{x}_t, t) + \frac{1}{2}\sum_{p, q} P_{pq} \frac{\partial^2 (GG^T)_{ij}(\widehat{x}_t, t)}{\partial \widehat{x}_p \partial \widehat{x}_q})dt. \tag{34}$$

The moment evolution equations using the second-order approximation can be found in Jazwinski (1970, p. 363) and become a   special case of the evolution equations, i.e. (33) and

(34). The evolutions of the i-th component of the mean vector and (i, j) element of the variance matrix, resulting from combining equations (28), (33), (34), are

$$d\widehat{\xi}_i = A_i(\widehat{\xi}_t, P_t, t)dt,$$

$$dP_{ij} = B_{ij}(\widehat{\xi}_t, P_t, t)dt,$$

where $A = (A_i)$, $B = (B_{ij})$ with $1 \le i \le 2, 1 \le j \le 2$ and

$$A_1 = \widehat{x}_2$$

$$A_2 = \alpha\widehat{x}_1 + \beta\widehat{x}_2 - a\widehat{x}_1^3 - b\widehat{x}_2\widehat{x}_1^2 + (-3a\widehat{x}_1 - b\widehat{x}_2)P_{11} + (-2b\widehat{x}_1)P_{12},$$

$$B_{11} = 2P_{12},$$

$$B_{12} = B_{21} = P_{11}(\alpha - 3a\widehat{x}_1^2 - 2b\widehat{x}_1\widehat{x}_2) + P_{12}(\beta - b\widehat{x}_1^2) + P_{22} - 3aP_{11}^2 - 3bP_{11}P_{12},$$

$$B_{22} = 2P_{12}(\alpha - 3a\widehat{x}_1^2 - 2b\widehat{x}_1\widehat{x}_2) + 2(\beta - b\widehat{x}_1^2)P_{22} - 6aP_{11}P_{12} - 4bP_{12}^2 - 2bP_{11}P_{22}$$

$$+ \sigma_B^2\widehat{x}_1^{2n} + \frac{1}{2}(2n)(2n-1)\sigma_B^2 P_{11}\widehat{x}_1^{2n-2}.$$

Evolution equations (33) and (34) involve the partial differential equation formalism. The mean and variance evolutions for the stochastic problem of concern here become the special cases of equations (33) and (34) as well as assume the structure of ODEs.

4. Numerical simulations

Approximate evolution equations, equations (33) and (34), are intractable theoretically, since the global properties are replaced with the local. Numerical experiments under a variety of conditions allow examining the effectiveness of the approximate estimation procedure. The following set of initial conditions and system parameters can be chosen for the numerical testing:

$\alpha = -1$, $a = 0.001$, $\beta = -0.2$, $b = 0.8$, $\sigma_B = 0.028$, $\sigma_u = 0.07$, $\widehat{x}_1(0) = 0.1$, $\widehat{x}_2(0) = 0.5$, $P_{11}(0) = 1$, $P_{12}(0) = 0$, $P_{22}(0) = 2$, $n = 3$.

Here the initial variances are chosen 'non-zero' and covariances take zero values, which illustrate uncertainties in initial conditions and the uncertainties are initially uncorrelated respectively. The order n of the state-dependent perturbation $\sigma_B x_t^n dB_t$ is three, since this choice of the order contributes to higher-order partials of the diffusion coefficient $(GG^T)(x_t, t)$ and allows to examine the efficacy of higher-order estimation algorithms.

Other choices can be made about the 'state-dependent perturbation order' provided $n \geq 1$. The diffusion parameters σ_B and σ_u are selected so that the contribution to the force from the random forcing term is smaller than the contribution from the deterministic part. Thanks to a pioneering paper of H. J. Kushner on stochastic estimation theory that the initial data can be adjusted for the convenience of the estimation procedure, however, it must be tested under a variety of conditions (Kushner 1967, p. 552). The choice of an estimation procedure is also dictated by some experimentation and guesswork. More over, the scanty numerical evidence will not suffice to adjudge the usefulness of the estimation procedure. As a result of these, numerical experiments of this chapter encompass three *different approximations.*

 In this chapter, the three different estimation procedures are the third-order, second-order, and first-order approximate evolution equations. The third-order approximate variance evolution equation involves the additional correction terms, i.e.

$$\frac{1}{2} \sum_{p,\,q,r} P_{ip} P_{qr} \frac{\partial^3 f_j(\widehat{\xi}_t, t)}{\partial \widehat{\xi}_p \partial \widehat{\xi}_q \partial \widehat{\xi}_r} \text{ and } \frac{1}{2} \sum_{p,\,q,r} P_{jp} P_{qr} \frac{\partial^3 f_i(\widehat{\xi}_t, t)}{\partial \widehat{\xi}_p \partial \widehat{\xi}_q \partial \widehat{\xi}_r} .$$

In the second-order variance evolution equation, these additional terms are not accounted for. The structure of the second-order mean evolution will be the same as the third-order, since the third-order moment vanishes with 'nearly Gaussian assumption'. The graphs of this chapter illustrate unperturbed trajectories correspond to the bilinear approximation, since the mean trajectories involving the bilinear approximation do not involve the variance term. On the other hand, the perturbed trajectories correspond to the second-order and third-order approximations, see figures (1)-(2). The qualitative analysis of the stochastic problem of concern here confirms the 'mean evolution pattern' using the third-order approximation. This chapter discusses briefly about the numerical simulation of the stochastic Duffing-van der Pol system. A greater detail about the Fokker-Planck analysis of the stochastic problem considered here can be found in a paper recently published by one of the Authors (Sharma 2008).
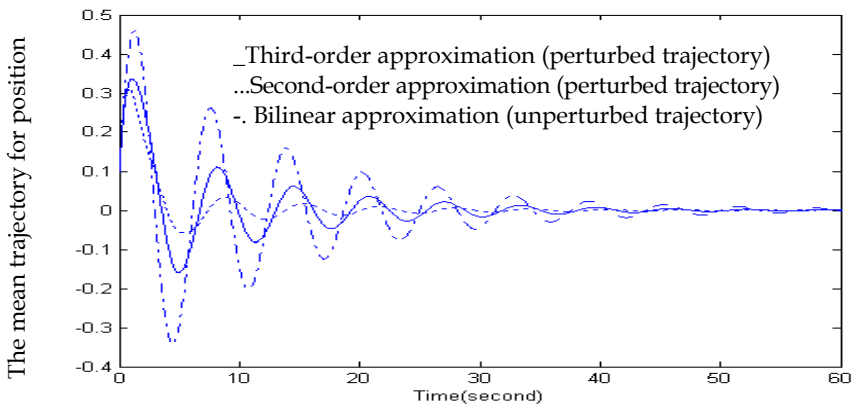
Fig. 1. A comparison between the mean trajectories for position using three approximations
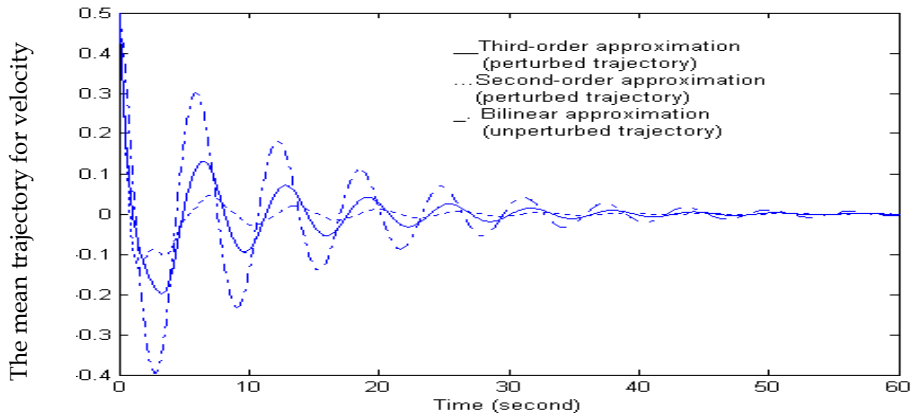
Fig. 2. A comparison between the mean trajectories for velocity using three approximations
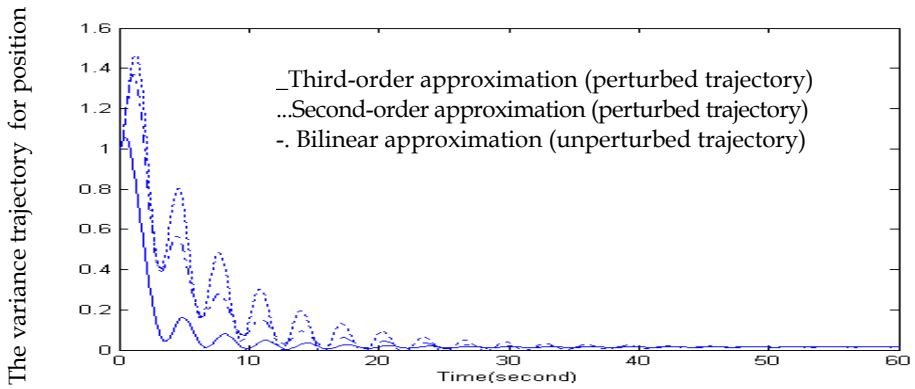
Fig. 3. A comparison between the variance trajectories for position using three approximations
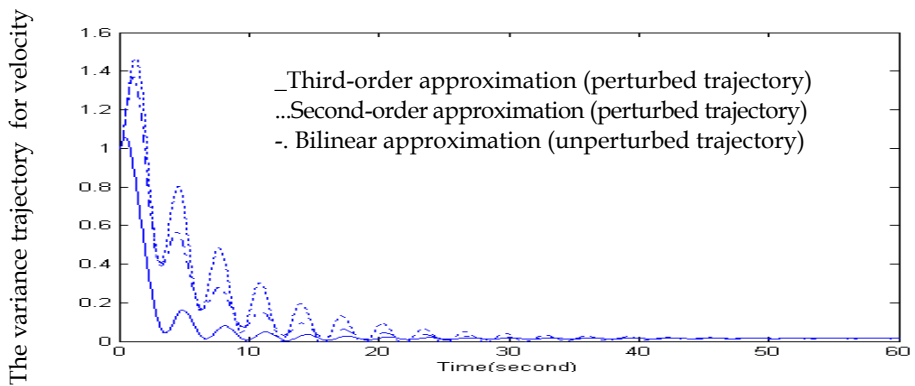
Fig. 4. A comparison between the variance trajectories for velocity using three approximations

5. Conclusion

In this chapter, the Authors have summarized four different methods to derive the Fokker-Planck equation, including two elementary proofs. The Fokker-Planck equation of the OU process-driven stochastic differential system, which received relatively less attention in literature, is also discussed. Most notably, in this chapter, the OU process with non-zero, finite and smaller correlation time was considered. This chapter discusses briefly approximate methods to the Fokker-Planck equation, stochastic differential equations as well as lists 'celebrated books' on the topic. It is believed that the Fokker-Planck analysis of the stochastic problem discussed here will be useful for analysing stochastic problems from diverse field.

Acknowledgement

6. References

Feller, W. (2000). *An Introduction to Probability Theory and its Applications,* John Wiley and Sons, New York and Chichester, vol.2.

Fletcher, C. A. J. (1984). *Computational Galerkin Methods*, Springer-Verlag, New York, Berlin.

Hänggi, P. (1995). The functional derivative and its use in the description of noisy dynamical systems, *Stochastic Processes Applied to Physics* (L. Pesquera and M. Rodriguez, eds.), World Scientific, Heyden, Philadelphia, pp. 69–95.

Hwalisz, L., Jung, P., Hänggi, P., Talkner, P. & Schimansky-Geier, L. (1989). Colored noise driven systems with inertia, *Z. Physik B*, 77, 471–483.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory,* Academic Press, New York and London.

Kloeden, P. E. and Platen, E. (1991). *The Numerical Solutions of Stochastic Differential Equations* (Applications of Mathematics), Springer, New York.

Kowalski, K. & Steeb W-H. (1991). *Non-Linear Dynamical Systems and Carleman Linearization,* World Scientific, Singapore, New Jersy.

Kushner, H. J. (1967). Approximations to optimal non-linear filter*s, IEEE Trans. Automat. Contr.* **12**(5), 546-556.

Liptser, R. S. and Shiryayev, A. N. (1977). *Statistics of Random Processes* 1, Springer, Berlin.

Risken, H. (1984). *The Fokker-Planck Equation: Methods of Solution and Applications,* Springer-Verlag, Berlin.

Sage, A. P. and Melsa, M. L. (1971). *Estimation Theory with Applications to Communications and Control*, Mc-Graw Hill, New York.

Sharma, Shambhu N. (Oct. 2008). A Kolmogorov-Fokker-Planck approach for a stochastic Duffing-van der Pol system, *Differential Equations and Dynamical Systems* (An international Journal for theory, applications and computer simulations), 16(4), 351-377. DOI: 10.1007/s12591-008-0019-x.
          http://www.springerlink.com/content/t5315t2k62154151/

The Itô calculus for a noisy dynamical system

Shambhu N. Sharma
Department of Electrical Engineering
National Institute of Technology, Surat, India
snsvolterra@gmail.com

The deterministic versions of dynamical systems have been studied extensively in literature. The notion of noisy dynamical systems is attributed to random initial conditions and small perturbations felt by dynamical systems. The stochastic differential equation formalism is utilized to describe noisy dynamical systems. The Itô calculus, a pioneering contribution of Kiyoshi Itô, is regarded as a path-breaking discovery in the branch of mathematical science in which the term $'dB_t{}'= \dot{B}_t dt$, where the Brownian motion $B = \{B_t, t_0 \leq t < \infty\}$. The Itô theory deals with multi-dimensional Itô differential rule, Itô stochastic integral and subsequently, can be exploited to analyse non-linear stochastic differential systems.

   This chapter discusses the usefulness of Itô theory to analysing a noisy dynamical system. In this chapter, we consider a system of two coupled second-order fluctuation equations, which has central importance in noisy dynamical systems. Consider the system of the coupled fluctuation equations of the form

$$\ddot{x}_1 = F_1(t, x_1, \dot{x}_1, x_2, \dot{x}_2, \dot{B}_1),$$
$$\ddot{x}_2 = F_2(t, x_1, \dot{x}_1, x_2, \dot{x}_2, \dot{B}_2),$$

where the state vector $x_t = (x_1, x_2, \dot{x}_1, \dot{x}_2)^T$ and the vector Brownian motion $B_t = (B_1, B_2)^T$. Interestingly, a suitable choice of the right-hand side terms F_1, F_2 of the above formalism describes the motion of an orbiting satellite in noisy environment, which w'd be the subject of discussion. After accomplishing the phase space formulation, the structure of the dynamical system of concern here becomes a multi-dimensional stochastic differential equation. Remarkably, in this chapter, the resulting SDE is analysed using the Itô differential rule in contrast to the Fokker-Planck approach. This chapter aims to open the topic to a broader audience as well as provides guidance for understanding the estimation-theoretic scenarios of stochastic differential systems.

Key words: Brownian motion, Itô differential rule, Fokker-Planck approach, second-order fluctuation equations, multi-dimensional stochastic differential equation

1. Introduction

The Ordinary Differential Equation (ODE) formalism is utilized to analyse dynamical systems deterministically. After accounting the effect of random initial conditions and small perturbations felt by dynamical systems gives rise to the concept of stochastic processes and subsequently, stochastic differential equations, a branch of mathematical science. As a result of these, the SDE confirms actual physical situations in contrast to the ODE. A remarkable success of stochastic differential equations can be found in different branches of sciences, i.e. stochastic control, satellite trajectory estimations, helicopter rotor, stochastic networks, mathematical finance, blood clotting dynamics, protein kinematics, population dynamics, neuronal activity. A nice exposition about the application of stochastic processes and Stochastic Differential Equations in sciences can be found in celebrated books authored by Karatzas and Shreve (1991), Kloeden and Platen (1991), Campen (2007) . The stochastic differential equation in the Itô sense is a standard form to describe dynamical systems in noisy environments. Alternatively, stochastic differential equations can be re-written involving $\dfrac{1}{2}$ differential, i.e. the Stratonovich sense, as well as p-differential, where $0 \le p \le 1$ (Pugachev and Synstin 1977). The Itô stochastic differential equation describes stochastic differential systems driven by the Brownian motion process. The Brownian motion process has greater conceptual depth and ageless beauty. The Brownian motion process is a Gauss-Markov process as well as satisfies the martingale properties, i.e. $E(x_t | F_s) = x_s, t \ge s$ and the sigma algebra $F_s = \underset{r \le s}{\cup} F_r$ (Revuz and Yor 1991, Strook and Varadhan 1979). The Central Limit Theorem (CLT) of stochastic processes confirms the usefulness of the Brownian motion for analysing randomly perturbed dynamical systems. The Brownian motion process can be utilized to generate the Ornstein-Uhlenbeck (OU) process, a colored noise (Wax 1954). This suggests that the stochastic differential system driven by the OU process can be reformulated as the Itô stochastic differential equation by introducing the notion of 'augmented state vector approach'. Moreover, the state vector, which satisfies the stochastic differential equation driven by the OU process, will be non-Markovian. On the other hand, the augmented state vector, after writing down the SDE for the OU process, becomes the Markovian. For these reasons, the Itô stochastic differential equation would be the cornerstone formalism in this chapter. The white noise can be regarded as informal non-existent time derivative $\dot{B}_t$ of the Brownian motion B_t. Kiyoshi Ito considered the term 'dB_t' resulting from the multiplication between the white noise $\dot{B}_t$ and the time differential dt.

This chapter demonstrates the usefulness of the Itô theory to analysing the motion of an orbiting satellite accounting for stochastic accelerations. Without accounting the effect of stochastic accelerations, stochastic estimation algorithms may lead to inaccurate estimation of positioning of the orbiting particle.

After introducing the phase space formulation, the stochastic problem of concern here can be regarded as a dynamical system perturbed by the Brownian motion process. In this chapter, the multi-dimensional Itô differential rule is exploited to analyse the stochastic differential system, which is the subject of discussion, in contrast to the Fokker-Planck

approach (Sharma and Parthasarathy 2007). The Fokker-Planck Equation (FPE) is a parabolic linear homogeneous differential equation of order two in partial differentiation for the transition probability density. A discussion on the Fokker-Planck equation is given in appendix 2. The chapter encompasses estimation-theoretic scenarios as well as qualitative analysis of the stochastic problem considered here.

This chapter is organized as follows: section (2) begins by writing a generalized structure of two-coupled second-order fluctuation equations. Subsequently, approximate evolutions of conditional mean vector and variance matrix are derived.   In section (3), numerical experiments were accomplished. Concluding remarks are given in section (4). Furthermore, a qualitative analysis of the stochastic problem of concern here can be found in 'appendix'1.

2. The structure of a noisy dynamical system and evolution equations

In dynamical systems' theory, second-order fluctuation equations describe dynamical systems perturbed by noise processes. Here, first we consider a system of two coupled second-order equations, which is an appealing case in dynamical systems and the theory of ordinary differential equations (Arnold 1995),

$$\ddot{x}_1 = F_1(t, x_1, \dot{x}_1, x_2, \dot{x}_2),$$
$$\ddot{x}_2 = F_2(t, x_1, \dot{x}_1, x_2, \dot{x}_2),$$

after introducing the noise processes along the components (x_1, x_2) of the coupled equations, the above can be re-written as

$$\ddot{x}_1 = F_1(t, x_1, \dot{x}_1, x_2, \dot{x}_2, \dot{B}_1), \tag{1}$$
$$\ddot{x}_2 = F_2(t, x_1, \dot{x}_1, x_2, \dot{x}_2, \dot{B}_2). \tag{2}$$

Equations (1)-(2) constitute a system of two coupled second-order fluctuation equations. After accomplishing the phase space formulation, the above system of fluctuation equations leads to a multi-dimensional stochastic differential equation. Choose

$$\dot{x}_1 = x_3,$$
$$\dot{x}_2 = x_4,$$

and

$$\dot{x}_3 = F_1(t, x_1, x_2, x_3, x_4, \dot{B}_1),$$
$$\dot{x}_4 = F_2(t, x_1, x_2, x_3, x_4, \dot{B}_2).$$

By considering a special case of the above system of equations, we have

$$dx_1 = x_3 dt,$$
$$dx_2 = x_4 dt,$$

and

$$dx_3 = f_3(t,x_1,x_2,x_3,x_4)dt + g_3(t,x_1,x_2,x_3,x_4)dB_1,$$
$$dx_3 = f_4(t,x_1,x_2,x_3,x_4)dt + g_4(t,x_1,x_2,x_3,x_4)dB_2.$$

The resulting stochastic differential equation is a direct consequence of the Itô theory, i.e. $'dB_t' = \dot{B}_t dt$. More precisely,

$$dx_t = f(t,x_1,x_2,x_3,x_4)dt + G(t,x_1,x_2,x_3,x_4)dB_t, \tag{3}$$

where

$$x_t = (x_1,x_2,x_3,x_4)^T, \ f(t,x_1,x_2,x_3,x_4) = (x_3, \ x_4, f_3, f_4)^T,$$

$$G(t,x_1,x_2,x_3,x_4) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ g_3 & 0 \\ 0 & g_4 \end{pmatrix}, \ dB_t = (dB_1, dB_2)^T.$$

Equation (3) can be regarded as the stochastic differential equation in the Itô sense. Alternatively, the above stochastic differential equation can be expressed in the Stratonovich sense. The Stratonovich stochastic differential equation can be re-written as the Itô stochastic differential equation using mean square convergence. A greater detail can be found in Jazwinski (1970), Protter (2005) and Pugachev and Synstin (1977). Here, the Itô SDE w'd be the cornerstone formalism for the stochastic problem of concern here. It is interesting to note that the motion of an orbiting particle accounting for stochastic dust particles' perturbations can be modeled in the form of stochastic differential equation, i.e. equation (3), where

$$x_t = (x_1,x_2,x_3,x_4)^T = (r, \ \phi, \ v_r, \ \omega)^T,$$

$$f(x_t,t) = (x_3,x_4,f_3,f_4)^T = (v_r, \ \omega, \ (\omega^2 r - V'(r)), \ -\frac{2v_r\omega}{r})^T,$$

$$G(x_t,t) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ \sigma_r r & 0 \\ 0 & \dfrac{\sigma_\phi}{r} \end{pmatrix}, \tag{4}$$

and r, ϕ are the radial and angular co-ordinates respectively. The radial and angular

components of the stochastic velocity are $\sigma_r r\, dB_r$ and $\dfrac{\sigma_\phi}{r} dB_\phi$ respectively. A

procedure for deriving the equation of motion of the stochastic differential system of
concern here involves the following: (i) write down the Lagrangian of the orbiting particle

$$L(r, \dot{r}, \dot{\phi}) = \frac{1}{2} m(\dot{r}^2 + r^2 \dot{\phi}^2) - mV(r).$$

This form of the Lagrangian is stated in Landau (1976), which results from the Lagrangian

$$L(r, \theta, \dot{r}, \dot{\theta}, \dot{\phi}) = \frac{1}{2} m(\dot{r}^2 + r^2 \dot{\theta}^2 + r^2 (\sin\theta)^2 \dot{\phi}^2) - mV(r) \text{ evaluated at } \theta = \frac{\pi}{2}.$$

(ii) Subsequently, the use of the Euler-Lagrange equation with additional random forces
along $(r(t),\, \phi(t))$ results stochastic two-body dynamics, a system of two coupled second-
order fluctuation equations assuming the structure of equations (1)-(2) (iii) accomplish
phase space formulation, which leads to the multi-dimentional stochastic differential
equation. For a greater detail about the motion of the orbiting particle in a stochastic dust
environment, the Royal Society paper (Sharma and Partasarathy 2007) can be consulted. A
theoretical justification explaining 'why the Brownian motion process is accurate to describe
the dust perturbation' hinges on the Central Limit Theorem of stochastic processes.
 Equation (3) in conjunction with equation (4) can be re-stated in the standard format as

$$dx_t = f(x_t, t)dt + G(x_t, t, \sigma_r, \sigma_\phi)dB_t,$$

where x_t is the state vector, $f(x_t, t)$ is the system non-linearity, $G(x_t, t, \sigma_r, \sigma_\phi)$ is the

dispersion matrix, σ_r and σ_ϕ are diffusion parameters. The on-line estimation of the

diffusion parameters σ_r and σ_ϕ can be accomplished from experiments by taking

measurements on the particle trajectory at discrete-time instants using the Maximum
Likelihood Estimate (MLE). The MLE involves the notion of the conditional probability

density $p(z_{\tau_n}, z_{\tau_{n-1}}, \ldots z_{\tau_2}, z_{\tau_1}, z_{\tau_0} | \sigma_r, \sigma_\phi)$, where z_{τ_i} denotes the observation vector

at i th   time   instant, $0 \le i \le n$.   The   estimated   parameter

vector $(\sigma_r, \sigma_\phi)^T = \max\limits_{\sigma_r, \sigma_\phi} p(z_{\tau_n}, z_{\tau_{n-1}}, \ldots z_{\tau_2}, z_{\tau_1}, z_{\tau_0} | \sigma_r, \sigma_\phi)$.   Moreover,   the

conditional probability density $p(z_{\tau_n}, z_{\tau_{n-1}}, \ldots z_{\tau_2}, z_{\tau_1}, z_{\tau_0} | \sigma_r, \sigma_\phi)$ can be regarded as

the   conditional   expectation   of   the   conditional   probability   density

$p(z_{\tau_n}, z_{\tau_{n-1}}, \ldots z_{\tau_2}, z_{\tau_1}, z_{\tau_0} | x_{\tau_n}, x_{\tau_{n-1}}, x_{\tau_{n-2}}, \ldots x_{\tau_2}, x_{\tau_1}, x_{\tau_0})$, i.e.

$$p(z_{\tau_n}, z_{\tau_{n-1}}, \ldots, z_{\tau_2}, z_{\tau_1}, z_{\tau_0} \mid \sigma_r, \sigma_\phi)$$

$$= E(p(z_{\tau_n}, z_{\tau_{n-1}}, \ldots, z_{\tau_2}, z_{\tau_1}, z_{\tau_0} \mid x_{\tau_n}, x_{\tau_{n-1}}, x_{\tau_{n-2}}, \ldots x_{\tau_2}, x_{\tau_1}, x_{\tau_0}) \mid \sigma_r, \sigma_\phi),$$

where

$$x_{\tau_{k+1}} = x_{\tau_k} + f(x_{\tau_k}, \tau_k)(\tau_{k+1} - \tau_k) + \sqrt{\tau_{k+1} - \tau_k}\ w_{\tau_{k+1}},\ w_{\tau_{k+1}} \text{ is } N(0,1).$$

After determining the diffusion parameters on the basis of the MLE, the diffusion parameters are plugged into the above diffusion equation, i.e. stochastic differential equation. As a result of this, we have

$$dx_t = f(x_t, t)dt + G(x_t, t)\, dB_t, \tag{5}$$

where $dB_t \sim N(0, Idt)$. A detailed discussion about the on-line estimation of unknown parameters of the stochastic differential system can be found in Dacunha-Castelle and Florens-Zmirou (1986). The above stochastic differential equation, equation (5), in conjunction with equation (4) can be analysed using the Fokker-Planck approach. Making the use of the FPE, we derive the evolution of the conditional moment, conditional expectation of the scalar function of an n-dimensional state vector. Note that the Fokker-Planck operator is an adjoint operator.

This chapter is intended to analyse the stochastic problem of concern here using the multi-dimensional Itô differential rule in contrast to the FPE approach. Here, we explain the Itô theory briefly and subsequently, its usefulness for analysing the noisy dynamical system. Consider the state vector $x_t = (x_1, x_2)^T \in U$ is a solution vector of the above SDE, $\phi : U \to R$, i.e. $\phi(x_t) \in R$, and the phase space $U \subset R^n$. Suppose the function $\phi(x_t)$ is twice differentiable. The stochastic evolution $d\phi(x_t)$ of the scalar function of the n-dimensional state vector using the stochastic differential rule can be stated as

$$d\phi(x_t) = \sum_i \frac{\partial \phi(x_t)}{\partial x_i} dx_i + \frac{1}{2} \sum_{i,j} dx_i dx_j\ \frac{\partial^2 \phi(x_t)}{\partial x_i \partial x_j}.$$

After plugging the i th component of stochastic differential equation, i.e. equation (5), in the above evolution, we have

$$d\phi(x_t) = (\sum_i \frac{\partial \phi(x_t)}{\partial x_i} f_i(x_t, t) + \frac{1}{2} \sum_i\ (GG^T)_{ii}(x_t, t) \frac{\partial^2 \phi(x_t)}{\partial x_i^2}$$

$$+\sum_{i<j}(GG^T)_{ij}(x_t,t)\frac{\partial^2\phi(x_t)}{\partial x_i\partial x_j})dt \quad +\sum_{1\le i\le n,1\le\gamma\le r}\frac{\partial\phi(x_t)}{\partial x_i}G_{i\gamma}(x_t,t)dB_\gamma, \qquad (6)$$

where the size of the vector Brownian motion process is r. Note that the contribution to the term $d\phi(x_t)$ coming from the second and third terms of the right-hand side of equation (6) is attributed to the property $dB_\phi dB_\gamma = \delta_{\phi\gamma}dt$. The integral counterpart of equation (6) can be written as

$$\phi(x_t)=\phi(x_{t_0})+(\sum_i\int_{t_0}^t\frac{\partial\phi(x_s)}{\partial x_i(s)}f_i(x_s,s)ds+\frac{1}{2}\sum_i\int_{t_0}^t(GG^T)_{ii}(x_s,s)\frac{\partial^2\phi(x_s)}{\partial x_i^2(s)}ds$$

$$+\sum_{i<j}\int_{t_0}^t(GG^T)_{ij}(x_s,s)\frac{\partial^2\phi(x_s)}{\partial x_i(s)\partial x_j(s)}ds)$$

$$+\sum_{1\le i\le n,1\le\gamma\le r}\int_{t_0}^t\frac{\partial\phi(x_s)}{\partial x_i}G_{i\gamma}(x_s,s)dB_\gamma(s).$$

The evolution $d\hat{\phi}(x_t)$ of the conditional moment is the standard formalism to analyse stochastic differential systems. The contribution to the term $d\hat{\phi}(x_t)$ comes from the system non-linearity and dispersion matrix, since the term $\phi(x_t)$ is a scalar function of the $n-$dimensional state vector. The state vector satisfies the Itô stochastic differential equation, see equation (5). As a result of this, the expectation and differential operators can be interchanged.

$$d\hat{\phi}(x_t) = E(d\,\phi(x_t)\big|x_{t_0},t_0).$$

The above equation in conjunction with equation (6) leads to

$$d\hat{\phi}(x_t)=(\sum_p\overbrace{f_p(x_t,t)\frac{\partial\phi(x_t)}{\partial x_p}}^{\wedge}+\frac{1}{2}\sum_p\overbrace{(GG^T)_{pp}(x_t,t)\frac{\partial^2\phi(x_t)}{\partial x_p^2}}^{\wedge}+\sum_{p<q}\overbrace{(GG^T)_{pq}(x_t,t)\frac{\partial^2\phi(x_t)}{\partial x_p\partial x_q}}^{\wedge})dt.$$

Note that the expected value of the last term of the right-hand side of equation (6) vanishes, i.e. $\left\langle G_{i\gamma}(x_t,t)dB_\gamma\right\rangle=0$. For the exact mean and variance evolutions, we consider $\phi(x_t)$ as $x_i(t)$ and $\tilde{x}_i\tilde{x}_j$ respectively, where $\tilde{x}_i=x_i-\hat{x}_i$. Thus, we have

$$d\hat{x}_i(t) = \hat{f}_i(x_t, t)dt, \tag{7}$$

$$dP_{ij} = (\overbrace{x_i f_j}^{\Lambda} - \hat{x}_i \hat{f}_j + \overbrace{f_i x_j}^{\Lambda} - \hat{f}_i \hat{x}_j + \overbrace{(GG^T)_{ij}(x_t, t)}^{\Lambda})dt, \tag{8}$$

where

$$\hat{x}_i(t) = E(x_i(t)|x_{t_0}, t_0),$$

$$P_{ij} = E((x_i - E(x_i(t)|x_{t_0}, t_0))(x_j - E(x_j|x_{t_0}, t_0))|x_{t_0}, t_0)$$

$$= E(\tilde{x}_i \tilde{x}_j|x_{t_0}, t_0).$$

The analytical and numerical solutions of the exact estimation procedure for the non-linear stochastic differential system are not possible, since its evolutions are infinite dimensional and require knowledge of higher-order moment evolutions. For these reasons, approximate evolutions, which preserve some of the qualitative characteristics of the exact evolutions, are analysed. Here, the bilinear and second-order approximations are the subject of investigation. The second-order approximate evolution equations can be derived by introducing second-order partials of the system non-linearity $f(x_t, t)$ and the diffusion coefficient $(GG^T)(x_t, t)$ into the exact mean and variance evolutions, equations (7)-(8). Thus, the mean and variance evolutions for the non-linear stochastic differential system, using the second-order approximation, are

$$d\hat{x}_i = (f_i(\hat{x}_t, t) + \frac{1}{2}\sum_{p,q} P_{pq} \frac{\partial^2 f_i(\hat{x}_t, t)}{\partial \hat{x}_p \partial \hat{x}_q})dt, \tag{9}$$

$$(dP_t)_{ij} = (\sum_p P_{ip} \frac{\partial f_j(\hat{x}_t, t)}{\partial \hat{x}_p} + \sum_p P_{jp} \frac{\partial f_i(\hat{x}_t, t)}{\partial \hat{x}_p} + (GG^T)_{ij}(\hat{x}_t, t)$$

$$+ \frac{1}{2}\sum_{p,q} P_{pq} \frac{\partial^2 (GG^T)_{ij}(\hat{x}_t, t)}{\partial \hat{x}_p \partial \hat{x}_q})dt. \tag{10}$$

Making the use of the above conditional moment evolutions for the system non-linearity and process noise coefficient matrix stated in equation (4), leads to the following mean and variance evolutions for the stochastic differential system considered here:

Mean evolutions

$$d\widehat{r} = \widehat{v}_r dt , \quad d\widehat{\phi} = \widehat{\omega}dt,$$

$$d\widehat{v}_r = (\widehat{r}\widehat{\omega}^2 - V'(\widehat{r}) - \frac{1}{2}\frac{\partial^2 V'(\widehat{r})P_{rr}}{\partial \widehat{r}^2} + \widehat{r}P_{\omega\omega} + 2\widehat{\omega}P_{r\omega})dt,$$

$$d\widehat{\omega} = (-\frac{2\widehat{v}_r\widehat{\omega}}{\widehat{r}} + \frac{2\widehat{\omega}P_{rv_r}}{\widehat{r}^2} + \frac{2\widehat{v}_r P_{r\omega}}{\widehat{r}^2} - \frac{2\ \widehat{v}_r\ \widehat{\omega}P_{rr}}{\widehat{r}^3} - \frac{2P_{v_r\omega}}{\widehat{r}})dt.$$

Variance evolutions

$$dP_{rr} = 2P_{rv_r}dt, \quad\quad dP_{r\phi} = (P_{v_r\phi} + P_{r\omega})dt,$$

$$dP_{rv_r} = ((\widehat{\omega}^2 - \frac{\partial V'(\widehat{r})}{\partial \widehat{r}})P_{rr} + 2\widehat{r}\widehat{\omega}P_{r\omega} + P_{v_r v_r})dt,$$

$$dP_{r\omega} = (\frac{2\widehat{r}\widehat{\omega}P_{rr}}{\widehat{r}^2} - \frac{2\widehat{\omega}P_{rv_r}}{\widehat{r}} - 2\frac{\widehat{v}_r P_{r\omega}}{\widehat{r}} + P_{v_r\omega})dt,$$

$$dP_{\phi\phi} = 2P_{\phi\omega}dt \quad, \quad dP_{\phi v_r} = ((\widehat{\omega}^2 - \frac{\partial V'(\widehat{r})}{\partial \widehat{r}})P_{r\phi} + 2\widehat{r}\ \widehat{\omega}P_{\phi\omega} + P_{v_r\omega})dt,$$

$$dP_{\phi\omega} = (\frac{2\widehat{v}_r\widehat{\omega}P_{r\phi}}{\widehat{r}^2} - \frac{2\widehat{\omega}P_{\phi v_r}}{\widehat{r}} - 2\frac{\widehat{v}_r P_{\phi\omega}}{\widehat{r}} + P_{\omega\omega})dt,$$

$$dP_{v_r v_r} = (2(\widehat{\omega}^2 - \frac{\partial V'(\widehat{r})}{\partial\widehat{r}})P_{rv_r} + 4\widehat{r}\widehat{\omega}P_{v_r\omega} + \sigma_r^2\widehat{r}^2 + \sigma_r^2 P_{rr})dt,$$

$$dP_{\omega v_r} = (\frac{2\widehat{v}_r\widehat{\omega}P_{rv_r}}{\widehat{r}^2} - \frac{2\widehat{\omega}P_{v_r v_r}}{\widehat{r}} - 2\frac{\widehat{v}_r P_{v_r\omega}}{\widehat{r}} + (\widehat{\omega}^2 - \frac{\partial V'(\widehat{r})}{\partial\widehat{r}})P_{r\omega} + 2\widehat{r}\widehat{\omega}P_{\omega\omega})dt$$

$$dP_{\omega\omega} = (\frac{4\widehat{v}_r\widehat{\omega}P_{r\omega}}{\widehat{r}^2} - 4\frac{\widehat{\omega}P_{v_r\omega}}{\widehat{r}} - 4\frac{\widehat{v}_r P_{\omega\omega}}{\widehat{r}} + \frac{\sigma_\phi^2}{\widehat{r}^2} + 3\sigma_\phi^2\frac{P_{rr}}{\widehat{r}^4}).$$

The second-order approximation of the function f, and the diffusion coefficient matrix

$GG^T(x_t,t)$ around the mean trajectory gives

$$f(x_t,t) \approx f(\widehat{x}_t,t) + \sum_p \widetilde{x}_p \frac{\partial f(\widehat{x}_t,t)}{\partial \widehat{x}_p} + \frac{1}{2}\sum_{p,q} \widetilde{x}_p \widetilde{x}_q \frac{\partial^2 f(\widehat{x}_t,t)}{\partial \widehat{x}_p \partial \widehat{x}_q},$$

$$GG^T(x_t,t) \approx GG^T(\widehat{x}_t,t) + \sum_p \widetilde{x}_p \frac{\partial GG^T(\widehat{x}_t,t)}{\partial \widehat{x}_p} + \frac{1}{2}\sum_{p,q} \widetilde{x}_p \widetilde{x}_q \frac{\partial^2 GG^T(\widehat{x}_t,t)}{\partial \widehat{x}_p \partial \widehat{x}_q},$$

where $\widetilde{x}_p = x_p - \widehat{x}_p$. The approximate conditional moment evolutions using bilinear approximation can be obtained by considering only the first-order partials of the system non-linearity and diffusion coefficients. In other words, the terms $\frac{1}{2}\sum_{p,q} \widetilde{x}_p \widetilde{x}_q \frac{\partial^2 f(\widehat{x}_t,t)}{\partial \widehat{x}_p \partial \widehat{x}_q}$ and $\frac{1}{2}\sum_{p,q} \widetilde{x}_p \widetilde{x}_q \frac{\partial^2 GG^T(\widehat{x}_t,t)}{\partial \widehat{x}_p \partial \widehat{x}_q}$ vanish for the bilinear approximation. As a result of this,

$$d\widehat{r} = \widehat{v}_r dt, \quad d\widehat{\phi} = \widehat{\omega} dt,$$

$$d\widehat{v}_r = (\widehat{r}\widehat{\omega}^2 - V'(\widehat{r}))dt, \quad d\widehat{\omega} = (-\frac{2\widehat{v}_r\widehat{\omega}}{\widehat{r}})dt,$$

and

$$dP_{rr} = 2P_{rv_r}dt, \quad dP_{r\phi} = (P_{v_r\phi} + P_{r\omega})dt,$$

$$dP_{rv_r} = ((\widehat{\omega}^2 - \frac{\partial V'(\widehat{r})}{\partial \widehat{r}})P_{rr} + 2\widehat{r}\widehat{\omega}P_{r\omega} + P_{v_rv_r})dt,$$

$$dP_{r\omega} = (\frac{2\widehat{r}\widehat{\omega}P_{rr}}{\widehat{r}^2} - \frac{2\widehat{\omega}P_{rv_r}}{\widehat{r}} - 2\frac{\widehat{v}_r P_{r\omega}}{\widehat{r}} + P_{v_r\omega})dt,$$

$$dP_{\phi\phi} = 2P_{\phi\omega}dt, \quad dP_{\phi v_r} = ((\widehat{\omega}^2 - \frac{\partial V'(\widehat{r})}{\partial \widehat{r}})P_{r\phi} + 2\widehat{r}\widehat{\omega}P_{\phi\omega} + P_{v_r\omega})dt,$$

$$dP_{\phi\omega} = (\frac{2\widehat{v}_r\widehat{\omega}P_{r\phi}}{\widehat{r}^2} - \frac{2\widehat{\omega}P_{\phi v_r}}{\widehat{r}} - 2\frac{\widehat{v}_r P_{\phi\omega}}{\widehat{r}} + P_{\omega\omega})dt,$$

$$dP_{v_r v_r} = (2(\widehat{\omega}^2 - \frac{\partial V^{'}(\widehat{r})}{\partial \widehat{r}})P_{rv_r} + 4\widehat{r}\widehat{\omega}P_{v_r\omega} + \sigma_r^2\widehat{r}^2)dt,$$

$$dP_{\omega v_r} = (\frac{2\widehat{v}_r\widehat{\omega}P_{rv_r}}{\widehat{r}^2} - \frac{2\widehat{\omega}P_{v_r v_r}}{\widehat{r}} - 2\frac{\widehat{v}_r P_{v_r\omega}}{\widehat{r}} + (\widehat{\omega}^2 - \frac{\partial V^{'}(\widehat{r})}{\partial \widehat{r}})P_{r\omega} + 2\widehat{r}\widehat{\omega}P_{\omega\omega})dt,$$

$$dP_{\omega\omega} = (\frac{4\widehat{v}_r\widehat{\omega}P_{r\omega}}{\widehat{r}^2} - 4\frac{\widehat{\omega}P_{v_r\omega}}{\widehat{r}} - 4\frac{\widehat{v}_r P_{\omega\omega}}{\widehat{r}} + \frac{\sigma_\phi^2}{\widehat{r}^2})dt.$$

The mean trajectory for the stochastically perturbed dynamical system using bilinear approximation does not include *variance terms* in the mean evolution. The term $\langle GG^T(x(t),t)\rangle$, the expected value of the diffusion coefficient, in the variance evolution accounts for the stochastic perturbation felt by the orbiting particle. For this reason, the bilinear approximation leads to the 'unperturbed mean trajectory', see figures (1)-(4) as well. On the other hand, the variance evolution using bilinear approximation for the dust-perturbed model includes perturbation effects, i.e. $GG^T(\widehat{x}_t,t)$. In order to account for the stochastic perturbation in the mean evolution, we utilize the second-order approximation in the mean evolution. The second-order approximation includes 'the second-order partials' of the system non-linearity $f(x_t,t)$ and variance terms in the mean trajectory, which leads to better *estimation* of the trajectory. The variance evolution dP_{v_r} of the radial velocity, using the second-order approximation, involves an additional term $\sigma_r^2 P_{rr}$ in contrast to the bilinear approximation. The variance evolution dP_ω of the angular velocity, using the second-order approximation, accounts for a correction term $3\sigma_\phi^2\frac{P_{rr}}{\widehat{r}^4}$, in contrast to the bilinear approximation as well.

 Note that the conditional moment evolutions derived in this chapter for the stochastic problem of concerns here agree with the evolutions stated in a Royal society paper (Sharma and Parthasarathy 2007). However, the approach of this chapter, multi-dimensional Itô rule, is different from the Fokker-Plank approach adopted in the Royal Society contribution.

3. Numerical experiments

The simulations of the mean and variance evolutions are accomplished using a simple, but effective finite difference method-based numerical scheme. The discrete version of the standard stochastic differential equation is

$$x_{t_{k+1}} = x_{t_k} + f(x_{t_k}, t_k)(t_{k+1} - t_k) + G(x_{t_k}, t_k)\sqrt{t_{k+1} - t_k}\, w_{t_{k+1}},$$

where $w_{t_{k+1}}$ is a standard normal variable. The dimension of the phase space of the stochastic problem of concern here is four, since the state vector $x_t = (x_1, x_2, x_3, x_4)^T = (r, \phi, v_r, \omega)^T \in U \subset R^4$. The size of the mean state vector is four and the number of entries in the variance matrix of the state is sixteen. Since the state vector is a real-valued vector stochastic process, the condition $P_{ij} = P_{ji}$ holds. The total number of distinct entries in the variance matrix w'd be ten. The initial conditions are chosen as

$$\hat{r}(0) = 1\,\text{AU}, \hat{\phi}(0) = 1\,\text{rad}, \hat{v}_r(0) = 0.01\,\text{AU/TU}, \hat{\omega}(0) = 1.1\,\text{rad/TU}, P_{xy}(0) = 0,$$

for the state variable x and y.

The initial conditions considered here are in canonical system of units. Astronomers adopt a normalized system of units, i.e. 'canonical units', for the simplification purposes. In canonical units, the physical quantities are expressed in terms of Time Unit (TU) and Astronomical Unit (AU). The diffusion parameters $\sigma_r = 0.0121(\text{TU})^{-\frac{3}{2}}$ and $\sigma_\phi = 2.2 \times 10^{-4}\, \dfrac{AU}{(TU)^{\frac{3}{2}}}$ are chosen for numerical simulations. Here we consider a set of deterministic initial conditions, which implies that the initial variance matrix w'd be zero. Note that random initial conditions lead to the non-zero initial variance matrix. The system is deterministic at $t = t_0$ and becomes stochastic at $t > t_0$ because of the stochastic perturbation. This makes the contribution to the variance evolution coming from the 'system non-linearity coupled with 'initial variance terms' will be zero at $t = t_1$. The contribution to the variance evolution at $t = t_1$ comes from the perturbation term $(GG^T)(x_t, t)$ only.

For $t > t_1$, the contribution to the variance evolution comes from the system non-linearity as well as the perturbation term. This assumption allows to study the effect of random perturbations explicitly on the dynamical system. The values of diffusion parameters are selected so that the contribution to the force coming from the random part is smaller than the force coming from the deterministic part. It has been chosen for simulational convenience only.
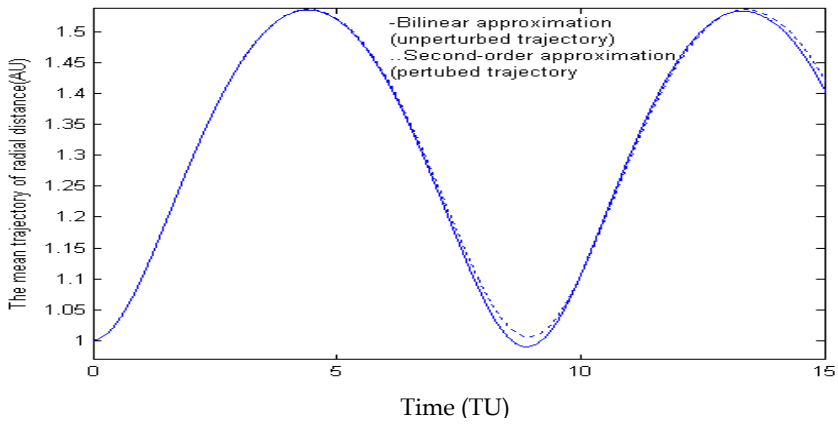
Fig. 1.
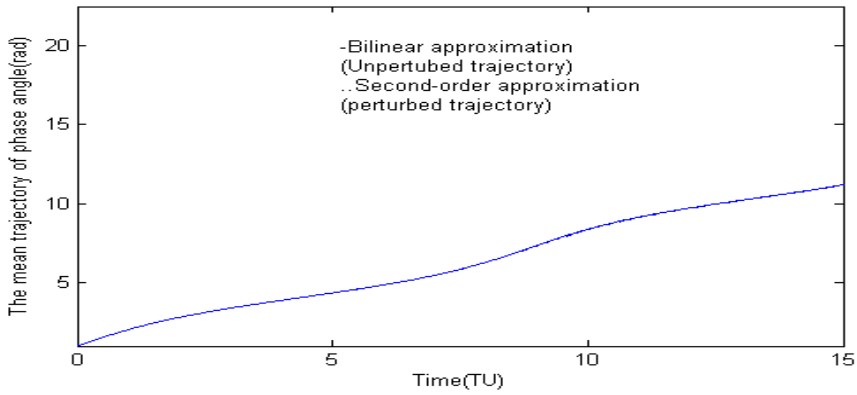
Fig. 2.
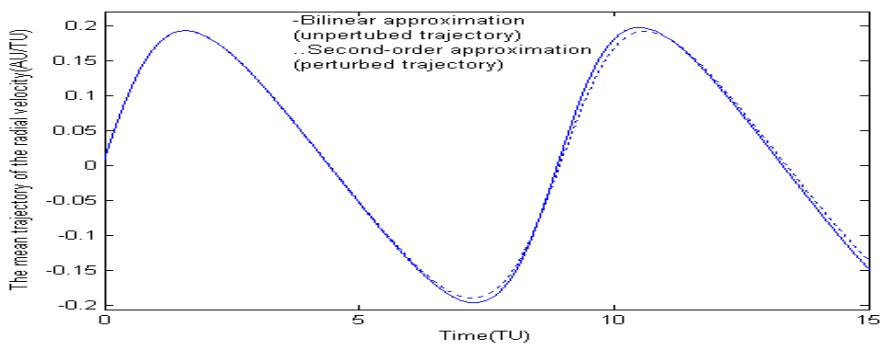
Fig. 3.
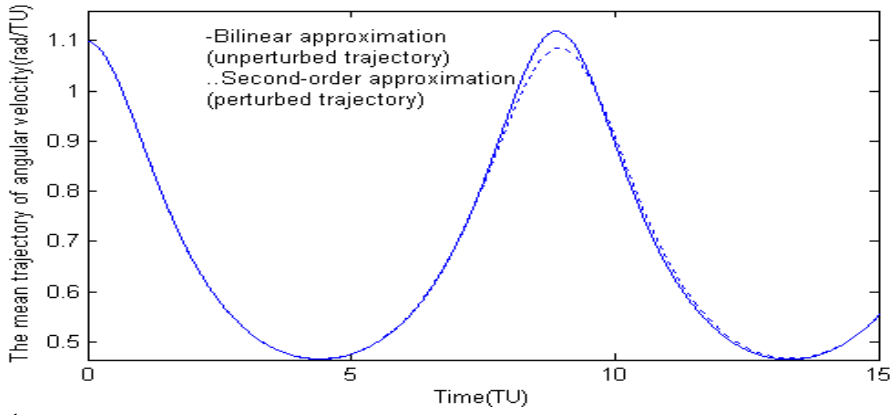
Fig. 4.
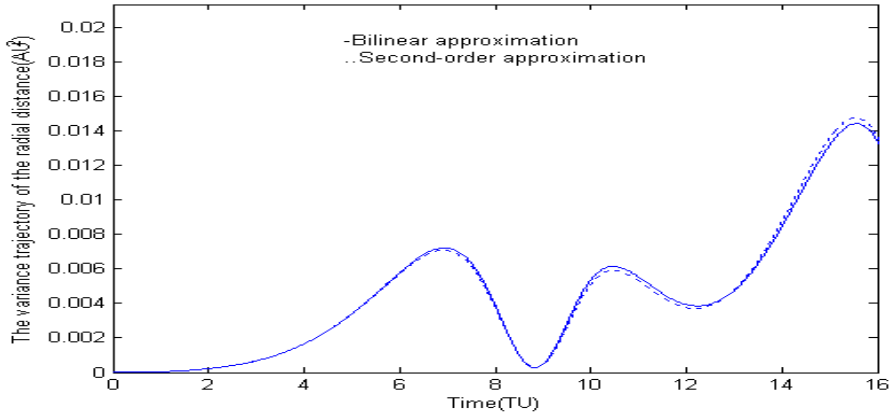
Fig. 5.
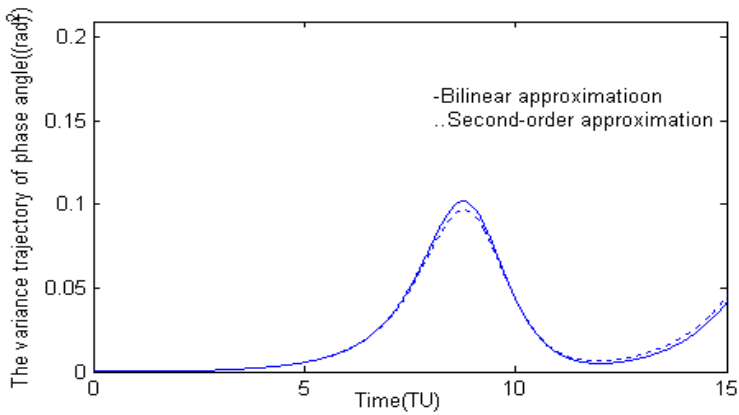
Fig. 6.
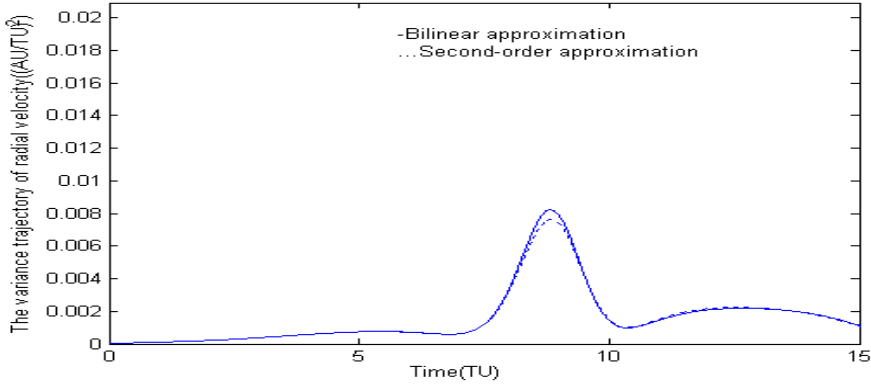
Fig. 7.
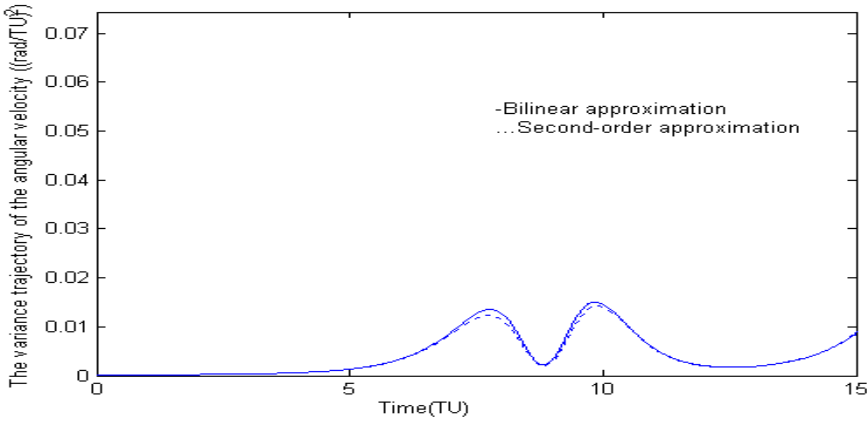
Fig. 8.

Here, we analyse the stochastic problem involving the numerical simulation of approximate conditional moment evolutions. The approximate conditional moment evolutions, i.e. conditional mean and variance evolutions, were derived in the previous section using the second-order and bilinear approximations. The variance evolutions using the second-order approximation result reduced variances of the state variables rather than the bilinear, see figures (5), (6), (7), and (8). These illustrate that the second-order approximation of the mean evolution produces less random fluctuations in the mean trajectory, which are attributed to the second-order partials of the system non-linearity $f(x_t, t)$, i.e.

$$\frac{1}{2}\sum_{p,q}\tilde{x}_p\tilde{x}_q\,\frac{\partial^2 f(\hat{x}_t,t)}{\partial\hat{x}_p\,\partial\hat{x}_q}.$$ The expectation of the partial terms leads to

$\dfrac{1}{2}\sum\limits_{p,q} P_{pq} \dfrac{\partial^2 f(\widehat{x}_t,t)}{\partial \widehat{x}_p \partial \widehat{x}_q}$. The correction term $\dfrac{1}{2}\sum\limits_{p,q} P_{pq} \dfrac{\partial^2 f(\widehat{x}_t,t)}{\partial \widehat{x}_p \partial \widehat{x}_q}$ involves the variance

term P_{pq}. The evolution of the variance term P_{ij} encompasses the contributions from the preceding variances, partials of the system non-linearity, the diffusion coefficient $(GG^T)_{ij}(\widehat{x}_t,t)$ as well as the second-order partial term $\dfrac{1}{2}\sum\limits_{p,q} P_{pq}(GG^T)_{ij}(\widehat{x}_t,t)$.

Significantly, the variance terms are also accounted for in the mean trajectory. This explains the second-order approximation leads to the perturbed mean trajectory. This section discusses very briefly about the numerical testing for the mean and variance evolutions derived in the previous section. A greater detail is given in the Author's Royal Society contribution. This chapter is intended to demonstrate the usefulness of the Itô theory for stochastic problems in dynamical systems by taking up an appealing case in satellite mechanics.

4. Conclusion

In this chapter, the Author has derived the conditional moment evolutions for the motion of an orbiting satellite in dust environment, i.e. a noisy dynamical system. The noisy dynamical system was modeled in the form of multi-dimensional stochastic differential equation. Subsequently, the Itô calculus for 'the Brownian motion process as well as the dynamical system driven by the Brownian motion' was utilized to study the stochastic problem of concern here. Furthermore, the Itô theory was utilized to analyse the resulting stochastic differential equation qualitatively. The Markovian stochastic differential system can be analysed using the Kolmogorov-Fokker-Planck Equation (KFPE) as well. The KFPE-based analysis involves the definition of conditional expectation, the adjoint property of the Fokker-Planck operator as well as integration by part formula. On the other hand, the Itô differential rule involves relatively fewer steps, i.e. Taylor series expansion, the Brownian motion differential rule. It is believed that the approach of this chapter will be useful for analysing stochastic problems arising from physics, mathematical finance, mathematical control theory, and technology.

Appendix 1

The qualitative analysis of the non-linear autonomous system can be accomplished by taking the Lie derivative of the scalar function ϕ, where $\phi : U \to R$, U is the phase space of the non-linear autonomous system and $\phi(x_t) \in R$. The function ϕ is said to be the first integral if the Lie derivative $L_v\phi$ vanishes (Arnold 1995). The problem of analysing the non-linear stochastic differential system qualitatively becomes quite difficult, since it involves multi-dimensional diffusion equation formalism. The Itô differential rule (Liptser and Shirayayev 1977, Sage and Melsa 1971) allows us to obtain the stochastic evolution of the function ϕ. Equation (6) of this chapter can be re-written as

$$\frac{d\phi(x_t)}{dt} = \sum_i \frac{\partial \phi(x_t)}{\partial x_i} f_i(x_t, t) + \frac{1}{2} \sum_i (GG^T)_{ii}(x_t, t) \frac{\partial^2 \phi(x_t)}{\partial x_i^2}$$

$$+ \sum_{i<j} (GG^T)_{ij}(x_t, t) \frac{\partial^2 \phi(x_t)}{\partial x_i \partial x_j} + \sum_{1 \le i \le n, 1 \le \gamma \le r} \frac{\partial \phi(x_t)}{\partial x_i} G_{i\gamma}(x_t, t) \dot{B}_\gamma.$$

Consider the function $\phi(x_t) = E(x_t)$, where E (.) is the energy function. Thus the stochastic evolution of the energy function (Sharma and Parthasarathy 2007) can be stated as

$$\frac{dE(x_t)}{dt} = \sum_i \frac{\partial E(x_t)}{\partial x_i} \dot{x}_i + \frac{1}{2} \sum_i (GG^T)_{ii}(x_t, t) \frac{\partial^2 E(x_t)}{\partial x_i^2}$$

$$+ \sum_{i<j} (GG^T)_{ij}(x_t, t) \frac{\partial^2 E(x_t)}{\partial x_i \partial x_j}.$$

The above evolution   for the stochastic differential system of this chapter assumes the following structure:

$$\frac{dE(r,\phi,v_r,\omega)}{dt} = \frac{\partial E(r,\phi,v_r,\omega)}{\partial r} \dot{r} + \frac{\partial E(r,\phi,v_r,\omega)}{\partial \phi} \dot{\phi} + \frac{\partial E(r,\phi,v_r,\omega)}{\partial v_r} \dot{v}_r + \frac{\partial E(r,\phi,v_r,\omega)}{\partial \omega} \dot{\omega}$$

$$+ \frac{1}{2} \sum_{i,j} (GG^T)_{ii}(x_t, t) \frac{\partial^2 E(r,\phi,v_r,\omega)}{\partial x_i^2},$$

where $E(r,\phi,v_r,\omega) = \frac{1}{2}(v_r^2 + r^2\omega^2) + V(r)$. A simple calculation will show that

$$\frac{dE(r,\phi,v_r,\omega)}{dt} = (r\omega^2 + V'(r))v_r + (r\omega^2 - V'(r))v_r + \omega r^2(-\frac{2v_r\omega}{r}) + \sigma_r^2 r^2 + \sigma_\phi$$

$$= \sigma_r^2 r^2 + \sigma_\phi^2.$$

Thus the derivative of the energy function for the stochastic system of concern here will not vanish leading to the non-conservative nature of the energy function.

Appendix 2

The Fokker-Planck equation has received attention in literature and found applications for developing the prediction algorithm for the Itô stochastic differential system. Detailed

discussions on the Fokker-Planck equation, its approximate solutions and applications in sciences can be found in Risken (1984), Stratonovich (1963). The Fokker-Planck equation is also known as the Kolmogorov forward equation. The Fokker-Planck equation is a special case of the stochastic equation (kinetic equation) as well. The stochastic equation is about the evolution of the conditional probability for given initial states for non-Markov processes. The stochastic equation is an infinite series. Here, we explain how the Fokker-Planck equation becomes a special case of the stochastic equation. The conditional probability density

$$p(x_1, x_2, ..., x_n) = p(x_1 | x_2, x_3, ..., x_n) p(x_2 | x_3, x_4, ..., x_n)....p(x_{n-1} | x_n) p(x_n).$$

In the theory of the Markov process, the above can be re-stated as

$$p(x_1, x_2, ..., x_n) = p(x_1 | x_2) p(x_2 | x_3)....p(x_{n-1} | x_n) p(x_n).$$

Thus,

$$p(x_1, x_2, ..., x_n) = q_{t_1, t_2}(x_1, x_2) q_{t_2, t_3}(x_2, x_3)....q_{t_{n-1}, t_n}(x_{n-1}, x_n) q(x_n),$$

where $q_{t_{i-1}, t_i}(x_{i-1}, x_i)$ is the transition probability density, $1 \leq i \leq n$ and $t_{i-1} > t_i$. The transition probability density is the inverse Fourier transform of the conditional characteristic function, i.e.

$$q_{t_{i-1}, t_i}(x_{i-1}, x_i) = \frac{1}{2\pi} \int e^{-iu(x_{i-1}, x_i)} E e^{iu(x_{i-1} - x_i)} du. \tag{11}$$

For deriving the stochastic equation, we consider the conditional probability density $p(x_1 | x_2)$, where

$$p(x_1, x_2) = p(x_1 | x_2) p(x_2).$$

After integrating over the variable x_2, the above equation leads to

$$p(x_1) = \int q_{t_1, t_2}(x_1, x_2) p(x_2) dx_2. \tag{12}$$

Equation (12) in combination with equation (11) leads to

$$p(x_1) = \frac{1}{2\pi} \int e^{-iu(x_1 - x_2)} (E e^{iu(x_1 - x_2)}) p(x_2) dx_2 du. \tag{13}$$

The conditional characteristic function is the conditional moment generating function and the n th order derivative of the conditional characteristic function $Ee^{iu(x_1-x_2)}$ evaluated at the $u = 0$ gives the n th order conditional moment. This can be demonstrated by using the definition of the generating function of mathematical science, i.e.

$$\phi(x,u) = \sum_{0\leq n}\varphi_n(x)u^n,$$

where $\phi(x,u)$ can be regarded as the generating function of the sequence $\{\varphi_n(x)\}$. As a result of this, the characteristic function $Ee^{iu(x_1-x_2)} = \sum_{0\leq n}\frac{(iu)^n}{n!}\langle(x_1-x_2)^n\rangle$. After introducing the definition of the conditional characteristic function, equation (13) can be recast as

$$p(x_1) = \frac{1}{2\pi}\int e^{-iu(x_1-x_2)}(\sum_{0\leq n}\frac{(iu)^n}{n!}\langle(x_1-x_2)^n\rangle)p(x_2)dx_2 du$$

$$= \sum_{0\leq n}\int\frac{1}{n!}(\frac{1}{2\pi}\int(iu)^n e^{-iu(x_1-x_2)}du)\langle(x_1-x_2)^n\rangle p(x_2)dx_2. \qquad (14)$$

The term $\frac{1}{2\pi}\int e^{-iu(x_1-x_2)}(iu)^n du$ within the second integral sign of equation (14) becomes $(-\frac{\partial}{\partial x_1})^n\delta(x_1-x_2)$ and leads to the probability density

$$p(x_1) = \sum_{0\leq n}\int\frac{1}{n!}(-\frac{\partial}{\partial x_1})^n\delta(x_1-x_2)\langle(x_1-x_2)^n\rangle p(x_2)dx_2. \qquad 15)$$

Consider the random variables x_{t_1} and x_{t_2}, where $t_1 > t_2$. The time instants t_1 and t_2 can be taken as $t_1 = t + \tau, t_2 = t$. For the short hand notation, introducing the notion of the stochastic process, taking $x_1 = x_\tau, x_2 = x$, equation (15) can be recast as

$$p(x_\tau) = \sum_{0\leq n}\int\frac{1}{n!}(-\frac{\partial}{\partial x_\tau})^n\delta(x_\tau - x)\langle(x_\tau - x)^n\rangle p(x)dx$$

$$= \sum_{0\leq n}\int\frac{1}{n!}(-\frac{\partial}{\partial x_\tau})^n\delta(x_\tau - x)k_n(x)\tau\, p(x)dx,$$

where $\left\langle \dfrac{(x_\tau - x)^n}{\tau} \right\rangle = k_n(x)$ and the time interval condition $\tau \to 0$ leads to

$$\underset{\tau \to 0}{Lt} \frac{p(x_\tau) - p(x)}{\tau} = \sum_{1 \leq n} \frac{1}{n!} \left(-\frac{\partial}{\partial x_\tau} \right)^n k_n(x) p(x)$$

or

$$\dot{p}(x) = \sum_{1 \leq n} \frac{1}{n!} \left(-\frac{\partial}{\partial x} \right)^n k_n(x\ ) p(x). \tag{16}$$

The above equation describes the evolution of conditional probability density for given initial states for the non-Markovian process. The Fokker-Plank equation is a stochastic equation with $k_i(x) = 0,\ 2 < i$. Suppose the scalar stochastic differential equation of the form

$$dx_t = f(x_t, t)dt + g(x_t, t)dB_t,$$

using the definition of the coefficient $k_n(x)$ of the stochastic equation (16), i.e.

$\left\langle \dfrac{(x_\tau - x)^n}{\tau} \right\rangle = k_n(x),\ \tau \to 0$, we have

$$k_1(x) = f(x, t),$$
$$k_2(x) = g^2(x, t),$$

and the higher-order coefficients of the stochastic equation will vanish as a consequence of the Itô differential rule. Thus, the Fokker-Planck equation

$$\dot{p}(x) = -\frac{\partial}{\partial x} f(x, t) p(x) + \frac{1}{2} \frac{\partial^2 g^2(x, t)}{\partial x^2} p(x).$$

Acknowledgement

5. References

Arnold, V. I. (1995). *Ordinary Differential Equations,* The MIT Press, Cambridge and Massachusetts.

Dacunha-Castelle, D. & Florens-Zmirou, D. (1986). Estimations of the coefficients of a diffusion from discrete observations, *Stochastics,* 19, 263-284.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*, Academic Press, New York and London.

Karatzas, I. & Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus* (graduate text in mathematics), Springer, New York.

Kloeden, P. E. & Platen, E. (1991). *The Numerical Solutions of Stochastic Differential Equations* (applications of mathematics), Springer, New York, 23.

Landau, I. D. & Lifshitz, E. M. (1976). *Mechanics* (Course of Theoretical Physics, Vol 1), Butterworth-Heinemann, Oxford, UK.

Liptser, R. S. & Shiryayev, A. N. (1977). *Statistics of Random Processes* 1, Springer, Berlin.

Protter, Philip E. (2005). *Stochastic Integration and Differential Equations*, Springer, Berlin, Heidelberg, New York.

Pugachev, V. S. & Sinitsyn, I. N. (1977). *Stochastic Differential Systems* (analysis and filtering), John-Wiley and Sons, Chichester and New York.

Revuz, D. & Yor, M. (1991). *Continuous Martingales and Brownian Motion*, Springer-Verlag, Berlin, Heidelberg.

Risken, H. (1984). *The Fokker-Planck Equation: Methods of Solution and Applications,* Springer-Verlag, Berlin.

Sage, A. P. & Melsa, M. L. (1971). *Estimation Theory with Applications to Communications and Control*, Mc-Graw Hill, New York.

Stratonovich, R. L. (1963). *Topics in the Theory of Random Noise* (Vol 1and 2), Gordan and Breach, New York.

Shambhu N. Sharma & Parthasarathy, H. (2007). Dynamics of a stochastically perturbed two-body problem. *Pro. R. Soc. A,* The Royal Society: London, 463, pp.979-1003, (doi: 10.1080/rspa.2006.1801).

Shambhu N. Sharma (2009). A Kushner approach for small random perturbations of a stochastic Duffing-van der Pol system, *Automatica* (a Journal of IFAC, International Federation of Automatic Control), 45, pp. 1097-1099.

Strook, D. W. & Varadhan, S. R. S. (1979). *Multidimensional Diffusion Processes* (classics in mathematics), Springer, Berlin, Heidelberg, New York.

Campen, N. G. van (2007). *Stochastic Processes in Physics and Chemistr*y, Elsevier, Amsterdam, Boston, London.

Wax, N. (ed.) (1954). *Selected Papers on Noise and Stochastic Processes,* Dover Publications, Inc, New York.

Application of coloured noise as a driving force in the stochastic differential equations

W.M.Charles

University of Dar-es-salaam,
College of Natural and Applied sciences,
Department of Mathematics,
P.O.Box 35062 Dar-es-salaam, Tanzania

Abstract

In this chapter we explore the application of coloured noise as a driving force to a set of stochastic differential equations(SDEs). These stochastic differential equations are sometimes called Random flight models as in A. W. Heemink (1990). They are used for prediction of the dispersion of pollutants in atmosphere or in shallow waters e.g Lake, Rivers etc. Usually the advection and diffusion of pollutants in shallow waters use the well known partial differential equations called Advection diffusion equations(ADEs)R.W.Barber et al. (2005). These are consistent with the stochastic differential equations which are driven by Wiener processes as in P.E. Kloeden et al. (2003). The stochastic differential equations which are driven by Wiener processes are called particle models. When the Kolmogorov's forward partial differential equations(Fokker-Planck equation) is interpreted as an advection diffusion equation, the associated set of stochastic differential equations called particle model are derived and are exactly consistent with the advection-diffusion equation as in A. W. Heemink (1990); W. M. Charles et al. (2009). Still, neither the advection-diffusion equation nor the related traditional particle model accurately takes into account the short term spreading behaviour of particles. This is due to the fact that the driving forces are Wiener processes and these have independent increments as in A. W. Heemink (1990); H.B. Fischer et al. (1979). To improve the behaviour of the model shortly after the deployment of contaminants, a particle model forced by a coloured noise process is developed in this chapter. The use of coloured noise as a driving force unlike Brownian motion, enables to us to take into account the short-term correlated turbulent fluid flow velocity of the particles. Furthermore, it is shown that for long-term simulations of the dispersion of particles, both the particle due to Brownian motion and the particle model due to coloured noise are consistent with the advection-diffusion equation.
Keywords: Brownian motion, stochastic differential equations, traditional particle models, coloured noise force, advection-diffusion equation, Fokker-Planck equation.

1. Introduction

Monte carlo simulation is gaining popularity in areas such as oceanographic, atmospheric as well as electricity spot pricing applications. White noise is often used as an important process in many of these applications which involve some error prediction as in A. W. Heemink

(1990); H.B. Fischer et al. (1979); J. R. Hunter et al. (1993); J.W. Stijnen et al. (2003). In these types of applications usually the deterministic models in the form of partial differential equations are available and employed. The solution is in most cases obtained by discretising the partial differential equations as in G.S. Stelling (1983). Processes such as transport of pollutants and sediments can be described by employing partial differential equations(PDEs). These well known PDEs are called advection diffusion equations. In particular when applied in shallow water e.g River, Lakes and Oceans, such effects of turbulence might be considered. However when this happens, it results into a set of partial differential equations. These complicated set of PDEs are difficult to solve and in most cases not easy to get a closed solution. In this chapter we explore the application coloured noise a a driving force to a set of stochastic differential equations(SDEs). These stochastic differential equations are sometimes called Random flight models. They are used for prediction of the dispersion of pollutants in atmosphere or in shallow waters e.g Lake, Rivers J. R. Hunter et al. (1993); R.W.Barber et al. (2005). Usually the advection and diffusion of pollutants in shallow waters use the well known partial differential equations called Advection diffusion equations(ADEs). These are consistent with the stochastic differential equations which are driven by Wiener processes as in C.W. Gardiner (2004); P.E. Kloeden et al. (2003). The stochastic differential equations which are driven by Wiener processes are called particle models. When the Kolmogorov's forward partial differential equations(Fokker-Planck equation) is interpreted as an advection diffusion equation, the associated with this set of stochastic differential equations called particle model are derived and are exactly consistent with the advection-diffusion equation as in W. M. Charles et al. (2009). Still, neither the advection-diffusion equation nor the related traditional particle model accurately takes into account the short term spreading behaviour of particles. This is due to the fact that the driving forces are Wiener processes and these have independent increment. To improve the behaviour of the model shortly after the deployment of contaminants, a particle model forced by a coloured noise process is developed in this article. The use of coloured noise as a driving force unlike Brownian motion, enables to us to take into account the short-term correlated turbulent fluid flow velocity of the particles. Furthermore, it is shown that for long-term simulations of the dispersion of particles, both the particle due to Brownian motion and the particle model due to coloured noise are consistent with the advection-diffusion equation.

To improve the behaviour of the model shortly after the deployment of contaminants, a random flight model forced by a coloured noise process are often used. The scheme in Figure 1, shows that for long term simulation both models, advection diffusion equation and the random flight models have no difference, such situation better to use the well known ADE. The use of coloured noise as a driving force unlike Brownian motion, enables to us to take into account only the short-term correlated turbulent fluid flow velocity of the particles as in A. W. Heemink (1990); W. M. Charles et al. (2009). An exponentially coloured noise process can also be used to mimic well the behaviour of electricity spot prices in the electricity market. Furthermore, when the stochastic numerical models are driven by the white noise, in most cases their order of accuracy is reduced. Such models consider that particles move according to a simple random walk and consequently have independent increment as in A.H. Jazwinski (1970); D.J. Thomson (1987). The reduction of the order of convergence happens because white noise is nowhere differentiable. However, one can develop a stochastic numerical scheme and avoid the reduction of the order of convergence if the coloured noise is employed as a driving force as in A. W. Heemink (1990); J.W. Stijnen et al. (2003); R.W.Barber et al. (2005); P.S. Addison et al. (1997).
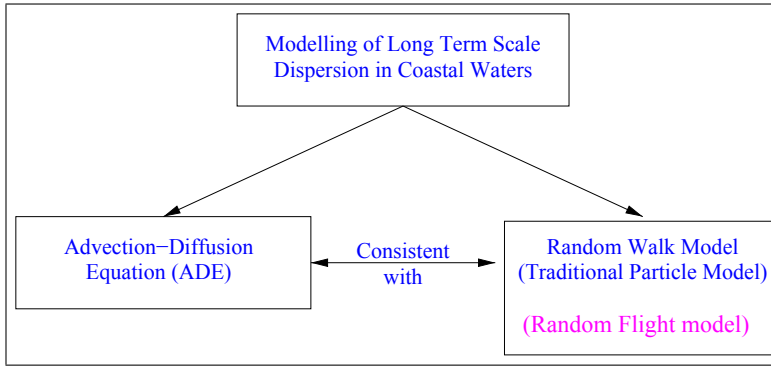
Fig. 1. A schematic diagram showing that for $t >> T_L$ both the ADEs and Random flight models are consistent

The application of coloured noise as a driving force to improve the model prediction of the dispersion of pollutants soon after deployment is discussed in this chapter. For it is well-known that the advection-diffusion equation describes the dispersion of particles in turbulent fluid flow accurately if the diffusing cloud of contaminants has been in the flow longer than a certain Lagrangian time scale and has spread to cover a distance that is larger in size than the largest scale of the turbulent fluid flow as in H.B. Fischer et al. (1979). The Lagrangian time scale (T_L) is a measure of how long it takes before a particle loses memory of its initial turbulent velocity. therefore, both the particle model which is driven by Brownian force and the advection-diffusion model are unable to accurately describe the short time scale correlated behaviour which is available in real turbulent flows at sub-Lagrangian time. Thus, a random flight model have been developed for any length of the coloured noise. This way, the particle model takes correctly into account the diffusion processes over short time scales when the eddy(turbulent) diffusion is less than the molecular diffusion. The inclusion of several parameters in the coloured noise process allows for a better match between the auto-covariance of the model and the underlying physical processes.

2. Coloured noise processes

In this part coloured noise forces are introduced and represent the stochastic velocities of the particles, induced by turbulent fluid flow. It is assumed that this turbulence is isotropic and that the coloured noise processes are stationary and completely described by their zero mean and Lagrangian auto covariance functionH.M. Taylor et al. (1998); W. M. Charles et al. (2009).

2.1 The scalar exponential coloured noise process
The exponentially coloured noise are represented by a linear stochastic differential equation. The exponential coloured noise represent the velocity velocity of the particle;

$$du_1(t) = -\frac{1}{T_L}u_1(t)dt + \alpha_1 dW(t). \tag{1}$$

$$u_1(t) = u_0 e^{\frac{-t}{T_L}} + \alpha_1 \int_0^t e^{-\frac{(t-s)}{T_L}} dW(s) \tag{2}$$

where u_1 is the particle's velocity, $\alpha_1 > 0$ is constant, and T_L is a Lagrangian time scale. For $t > s$ it can be shown as in A.H. Jazwinski (1970), that the scalar exponential coloured noise process in Eqn. (2) has mean, variance and Lagrangian auto-covariance of respectively,

$$\mathbb{E}[u_1(t)] = u_0 e^{\frac{-t}{T_L}}, \ \text{Var}[u_1(t)] = \frac{\alpha_1^2 T_L}{2}\left(1 - e^{-\frac{2t}{T_L}}\right),$$

$$\text{Cov}[u_1(t), u_1(s)] = \frac{\alpha_1^2 T_L}{2} e^{-\frac{|t-s|}{T_L}}. \tag{3}$$

where $\alpha_1 > 0$ is constant, and T_L is a Lagrangian time scale. For $t > s$ it can be shown A.H. Jazwinski (1970), that the scalar exponential coloured noise process in eqn.(2) has mean, variance and Lagrangian auto-covariance of respectively,

$$\mathbb{E}[u_1(t)] = u_0 e^{\frac{-t}{T_L}}, \ \text{Var}[u_1(t)] = \frac{\alpha_1^2 T_L}{2}\left(1 - e^{-\frac{2t}{T_L}}\right),$$

$$\text{Cov}[u_1(t), u_1(s)] = \frac{\alpha_1^2 T_L}{2} e^{-\frac{|t-s|}{T_L}}. \tag{4}$$

2.2 The general vector coloured noise force

The general vector form of a linear stochastic differential equation for coloured noise processes as in A.H. Jazwinski (1970); H.M. Taylor et al. (1998) is given by

$$d\mathbf{u}(t) \quad = \quad \mathbf{F}\mathbf{u}(t)dt + \mathbf{G}(t)dW(t), \ d\mathbf{v}(t) = \mathbf{F}\mathbf{v}(t)dt + \mathbf{G}(t)dW(t). \tag{5}$$

Where $\mathbf{u}(t)$ and $\mathbf{v}(t)$ are vectors of length n, F and G are $n \times n$ respectively $n \times m$ matrix functions in time and $\{W(t); t \geq 0\}$ is an m-vector Brownian process with $\mathbb{E}[dW(t)dW(t)^{\mathsf{T}}] = Q(t)dt$. In this chapter, a special case of the Ornstein-Uhlenbeck process C.W. Gardiner (2004); H.M. Taylor et al. (1998) is extended and repeatedly integrate it to obtain the coloured noise forcing along the x and y-directions:

$$
\begin{array}{rclcrcl}
du_1(t) &=& -\frac{1}{T_L}u_1(t)dt + \alpha_1 dW(t), & \quad & dv_1(t) &=& -\frac{1}{T_L}v_1(t)dt + \alpha_1 dW(t) \\
du_2(t) &=& -\frac{1}{T_L}u_2(t)dt + \frac{1}{T_L}\alpha_2 u_1(t)dt, & & dv_2(t) &=& -\frac{1}{T_L}v_2(t)dt + \frac{1}{T_L}\alpha_2 v_1(t)dt \\
du_3(t) &=& -\frac{1}{T_L}u_3(t)dt + \frac{1}{T_L}\alpha_3 u_2(t)dt, & & dv_3(t) &=& -\frac{1}{T_L}v_3(t)dt + \frac{1}{T_L}\alpha_3 v_2(t)dt \\
du_4(t) &=& -\frac{1}{T_L}u_4(t)dt + \frac{1}{T_L}\alpha_4 u_3(t)dt, & & dv_4(t) &=& -\frac{1}{T_L}v_4(t)dt + \frac{1}{T_L}\alpha_4 v_3(t)dt \\
du_5(t) &=& -\frac{1}{T_L}u_5(t)dt + \frac{1}{T_L}\alpha_5 u_4(t)dt, & & dv_5(t) &=& -\frac{1}{T_L}v_5(t)dt + \frac{1}{T_L}\alpha_5 v_4(t)dt \\
du_6(t) &=& -\frac{1}{T_L}u_6(t)dt + \frac{1}{T_L}\alpha_6 u_5(t)dt, & & dv_6(t) &=& -\frac{1}{T_L}v_6(t)dt + \frac{1}{T_L}\alpha_6 v_5(t)dt \\
\vdots &=& \vdots & & \vdots &=& \vdots \\
du_n(t) &=& -\frac{1}{T_L}u_n(t)dt + \frac{1}{T_L}\alpha_n u_{n-1}(t)dt, & & dv_n(t) &=& -\frac{1}{T_L}v_n(t)dt + \frac{1}{T_L}\alpha_n v_{n-1}(t)dt
\end{array}
\tag{6}
$$

As you keep increasing the length of the coloured noise, an auto-covariance of the velocity processes is modelled more realistically to encompasses the characteristics of an isotropic homogeneous turbulent fluid flow.

Figure 2 in an example of Wiener path and that of a coloured noise process. The sample path of the coloured noise are smoother that that of Wiener process.

The vector Langevin equation (6) generates a stationary, zero-mean, correlated Gaussian process denoted by $(u_n(t), v_n(t))$. The Lagrangian time scale T_L indicates the time over which the process remains significantly correlated in time. The linear system in eqn.(6), is the same in
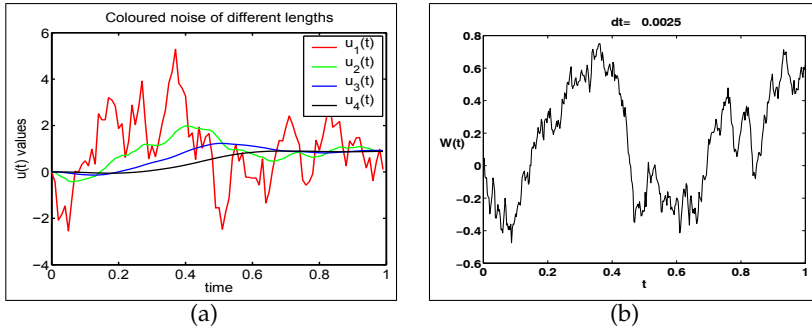
Fig. 2. Sample paths of coloured noise (a) and sample path of Wiener process (b)

the Itô and the Stratonovich sense because the diffusion function is not a function of state but only of time. In order to get more accurate results the stochastic differential equation driven by the coloured is integrated. by the Heun scheme (see e.g., G.N. Milstein (1995); J.W. Stijnen et al. (2003); P.E. Kloeden et al. (2003)).

The main purpose of this chapter is the application of coloured noise forcing in the dispersion of a cloud of contaminants so as to improve the short term behaviour of the model while leaving the long term behaviour unchanged. Being the central part of the model, it is important to study the properties of coloured noise processes in more detail. Coloured noise is a Gaussian process and it is well known that these processes can be completely described by their mean and covariance functions see L. Arnold (1974). From eqn.(2) and from Figure 3(a), it is easily seen that the mean approaches zero throughout and therefore requires little attention. The covariance, however, depends not only on time but also on the initial values of $u_n(0)$ and $v_n(0)$. This immediately gives rise to the question of how to actually choose or determine these values. Let's consider the covariance matrix of the stationary process **u** in the stochastic differential equations of the form (5). It is known (see e.g.,A.H. Jazwinski (1970)) that covariance function can now be described by

$$\frac{dP}{dt} = FP + PF^T + GQG^T. \tag{7}$$

The equation (7) can be equated zero so as to find the steady state covariance matrix $\bar{P}$ which will then be used to generate instances of coloured noise processes. Sampling of instances of u vector by using a steady state matrix, ensures that the process is sampled at its stationary phase thus removing any artefacts due to a certain choice of start values that would otherwise be used. The auto-covariance is depicted in Figure 3(c). Note that the behaviour of a physical process in this case depends on the parameters in the Lagrangian auto-covariance. Of course short term diffusion behaviour is controlled by the auto-covariance function. This provides room for the choice of parameters e.g.,$\alpha_1, \alpha_2 \cdots$. The mean, variance and the auto-covariance are not stationary for a finite time t but as $t \to \infty$,they approach the limiting stationary distribution values as shown in Figure 3(a)–(c).

(a) $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$ (b) $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$

(c) $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 1$
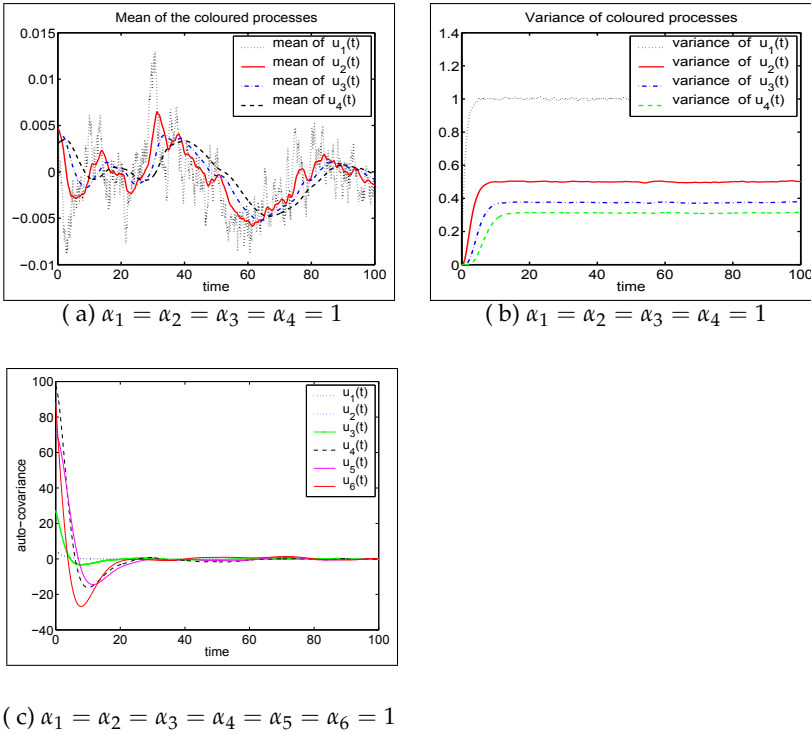
Fig. 3. (a)Shows that the mean goes to zero, while (b)-(c) shows that the variance and auto-covariance of coloured noise processes started from non-stationary to stationary state

2.3 The particle model forced by coloured noise

The prediction of the dispersion of pollutants in shallow waters are modeled by the random flight which is driven by coloured as in A. W. Heemink (1990). In this work, an extension to the work by A. W. Heemink (1990) has been done by generalising the cloured noise to any length that is, to $(u_n(t), v_n(t))$. The coloured noise processes stand for the velocity of the particle at time t in respectively the x and y directions. This way the Lagrangian auto-covariance processes can be modelled more realistically by taking into account the characteristics of the turbulent fluid flow for $t \ll T_L$. By using the following set of equations the random flight model remains consistent with the advection-diffusion equation for $t >> T_L$ while modelling realistically the short term correlation of the turbulent fluid flows. In this application, unlike in W. M. Charles et al. (2005), Longer length of the coloured noise have been chosen, that is $n = 6$ and more experiments are carried out in the whirl pool ideal domain for simulations of the advection and diffusion of pollutants in shallow waters. Thus the following coloured

noise are used.

$$
\begin{array}{llll}
du_1(t) &= -\frac{1}{T_L}u_1(t)dt + \alpha_1 dW(t), & dv_1(t) &= -\frac{1}{T_L}v_1(t)dt + \alpha_1 dW(t) \\
du_2(t) &= -\frac{1}{T_L}u_2(t)dt + \frac{1}{T_L}\alpha_2 u_1(t)dt, & dv_2(t) &= -\frac{1}{T_L}v_2(t)dt + \frac{1}{T_L}\alpha_2 v_1(t)dt \\
du_3(t) &= -\frac{1}{T_L}u_3(t)dt + \frac{1}{T_L}\alpha_3 u_2(t)dt, & dv_3(t) &= -\frac{1}{T_L}v_3(t)dt + \frac{1}{T_L}\alpha_3 v_2(t)dt \\
du_4(t) &= -\frac{1}{T_L}u_4(t)dt + \frac{1}{T_L}\alpha_4 u_3(t)dt, & dv_4(t) &= -\frac{1}{T_L}v_4(t)dt + \frac{1}{T_L}\alpha_4 v_3(t)dt \\
du_5(t) &= -\frac{1}{T_L}u_5(t)dt + \frac{1}{T_L}\alpha_5 u_4(t)dt, & dv_5(t) &= -\frac{1}{T_L}v_5(t)dt + \frac{1}{T_L}\alpha_5 v_4(t)dt \\
du_6(t) &= -\frac{1}{T_L}u_6(t)dt + \frac{1}{T_L}\alpha_6 u_5(t)dt, & dv_6(t) &= -\frac{1}{T_L}v_6(t)dt + \frac{1}{T_L}\alpha_6 v_5(t)dt
\end{array}
\tag{8}
$$

$$
\begin{aligned}
dX(t) &= \left[U + \sigma u_6(t) + (\tfrac{\partial H}{\partial x}D)/H + \tfrac{\partial D}{\partial x} \right] dt \\
dY(t) &= \left[V + \sigma v_6(t) + (\tfrac{\partial H}{\partial y}D)/H + \tfrac{\partial D}{\partial y} \right] dt.
\end{aligned}
\tag{9}
$$

These systems of vector equations are Markovian, this set of equations are referred to as the random flight model. The random flight model(8)–(9) is integrated for many different particles. Note that at the start of the simulation all particles have initial Gaussian velocities $(u_6(0), v_6(0))$ with zero mean and variance that agrees with covariance matrix $\bar{P}$ at a steady state. For instance in this chapter, the following covariance matrix was obtained when the parameters shown in Table 1 were used in the simulation;

$$
\begin{bmatrix}
25000 & 17500 & 2625 & 26.25 & 15.75 & 11.025 \\
17500 & 24500 & 5512 & 73.50 & 55.125 & 46.305 \\
2625 & 5512.5 & 1653.75 & 27.5625 & 24.806250 & 24.310125 \\
26.25 & 73.5 & 27.5625 & 0.551250 & 0.578812 & 0.648270 \\
15.75 & 55.125 & 24.80625 & 0.578812 & 0.694575 & 0.875164 \\
11.025 & 46.305 & 24.310125 & 0.648270 & 0.875164 & 1.22523
\end{bmatrix}
$$

3. The spreading behaviour of a cloud of contaminants

The characteristics of a spreading cloud of contaminants due to Brownian motion and coloured noise processes are discussed in the following sections.

3.1 Long term spreading behaviour of clouds of particles due Brownian motion force

Consider, the following 1 dimensional stochastic differential equation in the Itô sense

$$
dX(t) \stackrel{\text{Itô}}{=} f(t, X_t)dt + g(t, X_t)dW(t)
\tag{10}
$$

where $f(t, X_t)$ is the drift coefficient function and where $g(t, X_t)$ is the diffusion coefficient function. If it assumed that there is no drift term in eqn.(10) that is, $f(X(t), t) = 0$, gives

$$
g(X(t), t) = \sqrt{2D}.
$$

It follows that,

$$
dX(t) \stackrel{\text{Itô}}{=} \sqrt{2D}dW(t).
\tag{11}
$$

By applying following theorem which is found in H.M. Taylor et al. (1998), that for any continuous function the following theorem is applied.

Theorem 1. *Let $g(x)$ be continuous function and $\{W(t), t \geq 0\}$ be the standard Brownian motion process H.M. Taylor et al. (1998). For each $t > 0$, there exits a random variable*

$$\mathcal{F}(g) = \int_0^t g(x) dW(x),$$

which is the limiting of approximating sums

$$\mathcal{F}_n(g) = \sum_{k=1}^{2^n} g(\frac{k}{2^n} t)[W(\frac{k}{2^n} t) - W(\frac{k-1}{2^n} t)],$$

as $n \to \infty$. The random variable $\mathcal{F}(g)$ is normally distributed with mean zero and variance

$$\mathrm{Var}[\mathcal{F}(g)] = \int_0^t g^2(u) du,$$

if $f(x)$ is another continuous function of x then $\mathcal{F}(f)$ and $\mathcal{F}(g)$ have a joint normal distribution with covariance

$$\mathbb{E}[\mathcal{F}(f)\mathcal{F}(g)] = \int_0^t f(x)g(x) dx.$$

to eqn.(11), it can be shown that the variance of a cloud of contaminants grows linearly with time:

$$\mathrm{Var}[X(t)] \overset{\text{Itô}}{=} 2Dt + \text{constant}. \tag{12}$$

For more detailed information as well as the proof of this theorem, the reader is referred to H.M. Taylor et al. (1998) for example.

3.2 Long term spreading behaviour of clouds of contaminants subject to coloured noise forcing

As discussed in earlier, where for example, the first an exponential coloured $u_1(t)$ from eqn (2) is used as forcing coloured noise, if it is assumed that there is no background flow, the position of a particle at time t is given by

$$dX(t) \quad = \quad \sigma u_1(t) dt, \implies X(t) = X(0) + \sigma \int_0^t u_1(m) dm. \tag{13}$$

For simplicity, yet without loss of generality, let $X(0) = u_i(0) = 0$, for $i = 1, 2, \cdots, n$. Now, eqn., (2) leads to $u_1(m) = \alpha_1 \int_0^m e^{-\frac{1}{T_L}(m-k)} dW(k)$, and consequently,

$$X(t) \quad \overset{\text{Itô}}{=} \quad \sigma \alpha_1 T_L \int_0^t (1 - e^{-\frac{1}{T_L}(t-k)}) dW(k). \tag{14}$$

Using Theorem 1, the position of a particle at time t is normally distributed with zero mean and variance:

$$\frac{\mathrm{Var}[X(t)]}{t} = \sigma^2 \alpha_1^2 T_L^2 \left[1 - \frac{2T_L}{t}(1 - e^{\frac{-t}{T_L}}) + \frac{T_L}{2t}(1 - e^{\frac{-2t}{T_L}}) \right].$$

Thus, a position of a particle observed over a long time span as modelled by the coloured noise process $u_1(t)$ behaves much like the one driven by Brownian motion with variance parameter

$\sigma^2 \alpha_1^2 T_L^2$. Hence, the dispersion coefficient is related to variance parameters $\sigma^2 \alpha_1^2 T_L^2 = 2D$. Clarification are done by considering eqn.(14), where the second part is $u_1(t)$ itself;

$$X(t) = \sigma T_L \left[\alpha_1 W(t) - u_1(t) \right], \text{ where } u_1(t) = \alpha_1 \int_0^t e^{-\frac{1}{T_L}(t-k)} dW(k)$$

Let us now rescale the position process in order to better observe the changes over large time spans. By doing so, for $N > 0$, yields,

$$X_N(t) = \frac{1}{\sqrt{N}} X(Nt) = \sigma T_L \left[\alpha_1 \tilde{W}(t) + \frac{1}{\sqrt{N}} u_1(t) \right], \tag{15}$$

where $\tilde{B}(t) = \frac{W(Nt)}{\sqrt{N}}$ remains a standard Brownian motion process. For sufficiently large N it becomes clear that eqn.(15) behaves like Brownian motion as in H.M. Taylor et al. (1998); W. M. Charles et al. (2009):

$$X_N(t) \approx \sigma \alpha_1 T_L \tilde{W}(t).$$

3.3 The analysis of short term spreading behaviour of a cloud of contaminants

The analysis of the coloured noise processes usually starts with a scalar coloured noise, it can be shown using eqn.(4) that

$$\text{Cov}[u_{t+\tau} u_t] = \mathbb{E}[u_{t+\tau} u_t] = \mathbb{E}[v_{t+\tau} v_t] = \frac{1}{2} \alpha_1^2 T_L e^{\frac{-|\tau|}{T_L}} \tag{16}$$

From equation (16), It follows that,

$$\mathbb{E}[u_\tau u_s] = \frac{1}{2} \alpha_1^2 T_L e^{\frac{-|\tau-s|}{T_L}}$$

$$\text{Var}[X_t] = \sigma^2 \int_0^t \int_0^t \frac{1}{2} \alpha_1^2 T_L e^{\frac{-(\tau-s)}{T_L}} d\tau ds \tag{17}$$

The integration of equation (17) can easily be yielded by separately considering the regions $\tau < s$ and $\tau > s$, and it can be shown that

$$\begin{aligned} \text{Var}[X_t] &= \sigma^2 \alpha_1^2 T_L^3 \left(\frac{t^2}{2T_L^2} - \frac{t^3}{6T_L^3} \cdots \right) \\ &= \frac{\sigma^2 \alpha_1^2 T_L t^2}{2} - \frac{\sigma^2 \alpha_1^2 t^3}{6} + \cdots \end{aligned} \tag{18}$$

Since the short time analysis, eqn. (18) are of interest in this section and is considered only for very small values of t in a sense that for $t \ll T_L$ the variance of a cloud of particles shortly after deployment is then given by the following equation:

$$\text{Var}[X_t] = \frac{1}{2} \sigma^2 \alpha_1^2 T_L t^2 \tag{19}$$

With the constant dispersion coefficient $D = \frac{1}{2} \sigma^2 \alpha_1^2 T_L^2$, the variance of the cloud of particles, therefore initially grows with the square of time:

$$\text{Var}[X(t)] = \frac{D}{T_L} t^2 \tag{20}$$

3.4 The general long term behaviour of a cloud of contaminants due to coloured noise

It is assumed that there is no flow in the model and therefore have

$$dX(t) = \sigma u_1(t)dt \longrightarrow X(t) = \int_0^t \sigma u_1(s)ds, \ u_1(s) = \alpha_1 \int_0^s e^{-\frac{1}{T_L}(s-k)}dW(k),$$

$$X(t) = \left(\frac{1}{T_L}\right)^0 \sigma \alpha_1 \int_0^t \int_0^{t-k} e^{-\frac{1}{T_L}(s-k)} \frac{(s-k)^0}{0!}dsdW(k), \ X(0) = 0$$

It can then be shown that

$$u_2(s) \quad = \frac{1}{T_L}\alpha_1\alpha_2 \int_0^s e^{-\frac{1}{T_L}(s-k)}(s-k)dW(k), \tag{21}$$

where $0 < m < s < t$. Since $k < s$, the position of a particle due to coloured noise force eqn.(21) is given by

$$X(t) = \left(\frac{1}{T_L}\right)^1 \sigma \alpha_1 \alpha_2 \int_0^t \int_0^{t-k} e^{-\frac{1}{T_L}(s-k)} \frac{(s-k)^1}{1!}dsdW(k).$$

In general, a position due to $u_n(t)$ force is: $X(t) = \int_0^t \sigma u_n(s)ds$, and it follows that

$$X(t) \stackrel{\text{Itô}}{=} \left(\frac{1}{T_L}\right)^{n-1} \sigma \prod_{i=1}^n \alpha_i \int_0^t \left[\int_0^{t-k} e^{-\frac{1}{T_L}(s-k)} \frac{(s-k)^{n-1}}{(n-1)!}\right] dsdW(k) \tag{22}$$

Careful manipulation using integration by parts of the integral within the square brackets of eqn. (22), yields

$$X(t) \stackrel{\text{Itô}}{=} (T_L)^n \left(\frac{1}{T_L}\right)^{n-1} \sigma \prod_{i=1}^n \alpha_i \int_0^t [1 + \ldots] dW(k), \ \text{ for } \ n \geq 1. \tag{23}$$

Finally, with the aid of Theorem 1 whose proof is found in H.M. Taylor et al. (1998), the variance of a cloud of contaminants can be computed as described in the sections above. The derivation of velocity $v_n(t)$ of the particle along the y direction proceeds completely analogously. Let us now compute the variance of the general equations for position given by eqn.(23)

$$\text{Var}[X(t)] \quad = \sigma^2 (T_L)^2 \prod_{i=1}^n \alpha_i^2 \int_0^t [1 - \ldots]^2 dk \tag{24}$$

For $\sigma > 0$, $\alpha_i > 0$, and $T_L > 0$, the process again behaves like a Brownian process with variance parameters $T_L^2 \sigma^2 \prod_{i=1}^n \alpha_i^2$ as $t \to \infty$. Thus the appropriate diffusion coefficient from eqn.(12)is equals $D = \frac{\sigma^2 T_L^2 \prod_{i=1}^n \alpha_i^2}{2}$. This relation is important because it gives a criterion for various choices of parameters α_i, $i = 1, \cdots, n$, $T_L > 0$. In a simulation the constant dispersion coefficient D often is specified whereas σ must be solved in terms of the other parameters. In the following section we introduce the two dimensional particle model as in W. M. Charles et al. (2009). This model will be used as a comparison with the random flight model during the simulation of the dispersion of pollutants an ideal domain known as whirl pool.

4. Particle model due to Brownian motion force for dispersion of pollutants in shallow waters

The position of particles in water at time t, is designated by $(X(t), Y(t))$. Different random locations of the particle are described with the aid of stochastic differential equation. The integration of the movements of the particle in water is done in two steps. A deterministic step consisting of velocity field of water and a random step known as diffusion modelled by the stochastic process A. W. Heemink (1990);

$$dX(t) \stackrel{\text{Itô}}{=} \left[U + \frac{D}{H} \frac{\partial H}{\partial x} + \frac{\partial D}{\partial x} \right] dt + \sqrt{2D} dW_1(t), \ X(0) = x_0 \tag{25}$$

$$dY(t) \stackrel{\text{Itô}}{=} \left[V + \frac{D}{H} \frac{\partial H}{\partial y} + \frac{\partial D}{\partial y} \right] dt + \sqrt{2D} dW_2(t), \ Y(0) = y_0. \tag{26}$$

Here D is the dispersion coefficient in m^2/s; $U(x, y, t)$, $V(x, y, t)$ are the averaged flow velocities (m/s) in respectively x, y directions; $H(x, y, t)$ is the total depth in m at location (x, y), and $dW(t)$ is a Brownian motion with mean $(0, 0)^T$ and $\mathbb{E}[dW_1(t)dW_2(t)^T] = Idt$ where I is a 2×2 identity matrixP.E. Kloeden et al. (2003). Note that the advective part of the particle model eqns.(25)–(26) is not only containing the averaged water flow velocities but also spatial variations of the diffusion coefficient and the averaged depth. This correction term makes sure that particles are not allowed to be accumulated in regions of low diffusivity as demonstrated by (see e.g., J. R. Hunter et al. (1993); R.W.Barber et al. (2005)). At closed boundaries particle bending is done by halving the time step sizes until the particle no longer crosses closed boundary. As a result there is no loss of mass through such boundaries. The position $(X(t), Y(t))$ process is Markovian and the evolution of its probability density function ($p(x, y, t)$), is described by an advection-diffusion type of the partial differential equation known as the Fokker-Planck equation (see e.g.,A.H. Jazwinski (1970))

5. Discrete version of the particle model driven by Brownian motion

Analytical solutions of stochastic differential equations do not always exist due to their complexity and nonlinearity. Therefore, stochastic numerical integration schemes are often applied as in G.N. Milstein (1995); J.W. Stijnen et al. (2003). An example of a numerical scheme is the Euler scheme which, although not optimal in terms of order of convergence, is easy to implement and requires only $O(\Delta t)$ in the weak sense P.E. Kloeden et al. (2003). Here the time interval $[t_0, T]$ is discretised as $t_0 = 0 < t_1 < t_2 < \cdots < t_{n-1} < t_n = T$, with $\Delta(t_k) = t_{k+1} - t_k$, $\Delta W(t_k) = W(t_{k+1}) - W(t_k)$, for $k = 0, 1, \cdots n$.

$$\bar{X}(t_{k+1}) = \bar{X}(t_k) + \left[U + (\frac{\partial H}{\partial x} D)/H + \frac{\partial D}{\partial x} \right] \Delta(t_k) + \sqrt{2D} \Delta W(t_k) \tag{27}$$

$$\bar{Y}(t_{k+1}) = \bar{Y}(t_k) + \left[V + (\frac{\partial H}{\partial y} D)/H + \frac{\partial D}{\partial y} \right] \Delta(t_k) + \sqrt{2D} \Delta W(t_k) \tag{28}$$

Where $\bar{X}(t_{k+1})$ and $\bar{Y}(t_{k+1})$ are the numerical approximations of the $X(t_{k+1})$ and $Y(t_{k+1})$ positions respectively due to the traditional particle model. The noise increments $\Delta W(t_k)$ are independent and normally distributed N $(0, \Delta(t_k))$ random variables which can be generated using e.g., pseudo-random number generators. The domain information consisting of flow velocities and depth is computed using a hydrodynamic model known as WAQUA see G.S. Stelling (1983). The flow averaged fields are only available on grid points of a rectangularly discretised grid and therefore, interpolation methods are usually used to approximate the values at other positions.

5.1 Boundaries

Numerical schemes such as the Euler scheme often show very poor convergence behaviour G.N. Milstein (1995); P.E. Kloeden et al. (2003). This implies that, in order to get accurate results, small time steps are needed thus requiring much computation. Another problem with the Euler (or any other numerical scheme) is its undesirable behaviour in the vicinity of boundaries; a time step that is too large may result in particles unintentionally crossing boundaries. To tackle this problem two types of boundaries are prescribed. Closed boundaries representing boundaries intrinsic to the domain, and open boundaries which arise from the modeller's decision to artificially limit the domain at that location. Besides these boundary types, the is of what what happens if, during integration, a particle crosses one of these two borders is also considered as in J.W. Stijnen et al. (2003); W. M. Charles et al. (2009);

- In case an open boundary is crossed by a particle, the particle remains in the sea but is now outside the scope of the model and is therefore removed;

- In case a closed boundary is crossed by a particle during the advective step of integration, the step taken is cancelled and the time step halved until the boundary is no longer crossed. However, because of the halving, say n times, the integration time is reduced to $2^{-n}\Delta t$, leaving a remaining $(1 - 2^{-n})\Delta t$ integration time. This means at least another $2^n - 1$ steps need to be taken at the new integration step in order to complete the full time-step Δt. This way, shear along the coastline is modelled;

- If a closed boundary is crossed during the diffusive part of integration, the step size halving procedure described above is maintained with the modification that in addition to the position, the white noise process is also restored to its state prior to the abandoned integration step. Again the process of halving the time step and continuing integration is repeated until no boundaries are crossed and the full Δt time step has been integrated.

6. Numerical Experiments

Before applying both the traditional model(25)–(26) and the part model forced by coloured noise(8)–(9) to a real life pollution problem, A whirl pool have been created as domain for test problem. In this case a whirl pool domain with flow field and a constant total depth of 25 metres is created. In order to compare the spreading behaviour of a cloud of contaminants some experiments using both particle models have been carried out. The table 1 below summarises the simulation parameters that have been used in the experiments:

Summary of the simulation parameters of particle for pollutants dispersion in shallow waters

Whirl pool	Unit	Value
# of steps	-	89999
Δt	s	86.4
Particles	-	5000
$\alpha_1, \alpha_2, \alpha_3$	-	$\alpha_1 = 1, \alpha_2 = 1.4, \alpha_3 = 0.3$
$\alpha_4, \alpha_5, \alpha_6$	-	$\alpha_4 = 0.02, \alpha_5 = 1.2, \alpha_6 = 1.4$
Tracks	-	5
Grid offset	m	$(-20800, -20800)$
Grid size	-	105×105
Cell size	m	400×400
Init. point	m	$(-10000, 14899)$
D	m/s^2	3
T_L	s	50000

Table 1. The simulation parameters of the particle model for the dispersion of pollutants in shallow waters.

From now onwards in this section Brownian motion is denoted by BM and coloured noise by CN. A bunch of 5000 particles are released at the location $(-10000, 14899)$, the simulation starts at time $t_0 = 0$ in the whirl pool domain. The scattering of a cloud of contaminants due to coloured noise or Brownian motions forces is followed at a specified time steps after release. Generally a large number of particles are used P.S. Addison et al. (1997) in numerical simulations. The simulations parameters that have been used for simulations of advection and diffusion of pollutants in shallow waters in this article are summarised in the Table 1. The results in the Figures 4(a)-(b) show that a cloud of 5000 particles have been deployed, while Figure 4(c)-(d) shows that 5000 have spread to cover a certain distance 52 days later for random flight model by CN and particle model by BM respectively. Whereas Figure5(a)-(b) realisations of marked 5 tracks by CN and BM noised respectively while Figure5(c)-(d) show a realisation of a single same marked particle for all the simulation period of 89999 time steps each of 86.4s.

In this chapter a series of experiments are carried out in a stationary homogeneous turbulent flow with zero mean velocity. The Lagrangian time scale as T_L is introduced in the models. Futhermore, an experiment was carried in the empty domain as in W. M. Charles et al. (2009) using random flight model as well as the traditional particle model so as to show the differences between the small scale fluctuations and their similarity in the long scale fluctuations. The simulation of the spreading of a cloud of $20,000$ particles is tracked in an empty domain and its variance is computed. It has been shown that once the particles have been in the flow longer than the time scale T_L, the variance of the spreading cloud grows linearly with time similar to the behaviour of the advection-diffusion equation. Before that time, the variance grows with the square of time (quadratically), creating two different zones of diffusion see Figure 6 as in W. M. Charles et al. (2009). In Section 3.4 it is suggested that for $t \gg TL$ a turbulent mixing coefficient similar to constant dispersion coefficient D such that $D = \frac{\sigma^2 T_L^2 \prod \alpha_i^2}{2}$ can be defined.
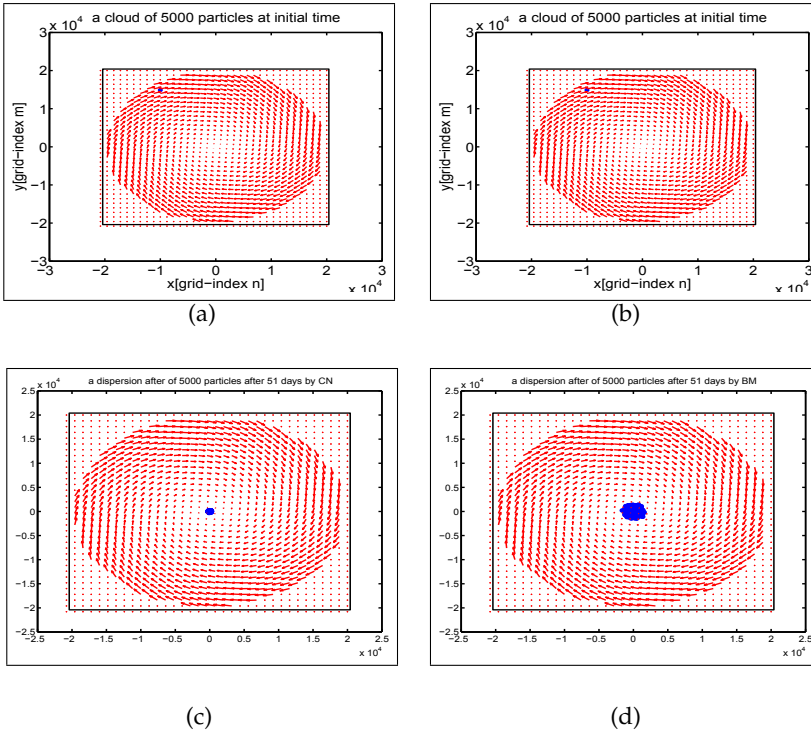
Fig. 4. Dispersion of a cloud of 5000 particles released in the idealised whirl pool domain. (a) Due to coloured noise for $t << T_L$, (b) Due to Brownian motion noise for $t << T_L$, (c) Due to coloured noise for $t >> T_L$, (d) Due to Brownian motion noise for $t >> T_L$.

7. Conclusions

The results obtained in this work suggest that coloured noised can be used to improve the prediction of the dispersion of pollutants. This is possible when a short time correlation is considered which is that case in most cases. Thus, random flight model can provide the modeller with an enhanced tool for the short term simulation of the pollutants by providing more flexibility to account for correlated physical processes of diffusion in the shallow waters. However, in this chapter a general analysis similar to those in W. M. Charles et al. (2009) shows that a process observed over a long time spans as modelled by the coloured noise force behaves much like a Brownian motion model with variance parameter $\sigma^2 T_L^2 \prod_{i=1}^{n} \alpha_i^2$. The use of coloured noise however is more expensive in terms of computation and therefore it is advisable to use the particle model driven by coloured for short term behaviour while adhering to the traditional particle model for long-term simulations.

(a)

(b)

(c)

(d)
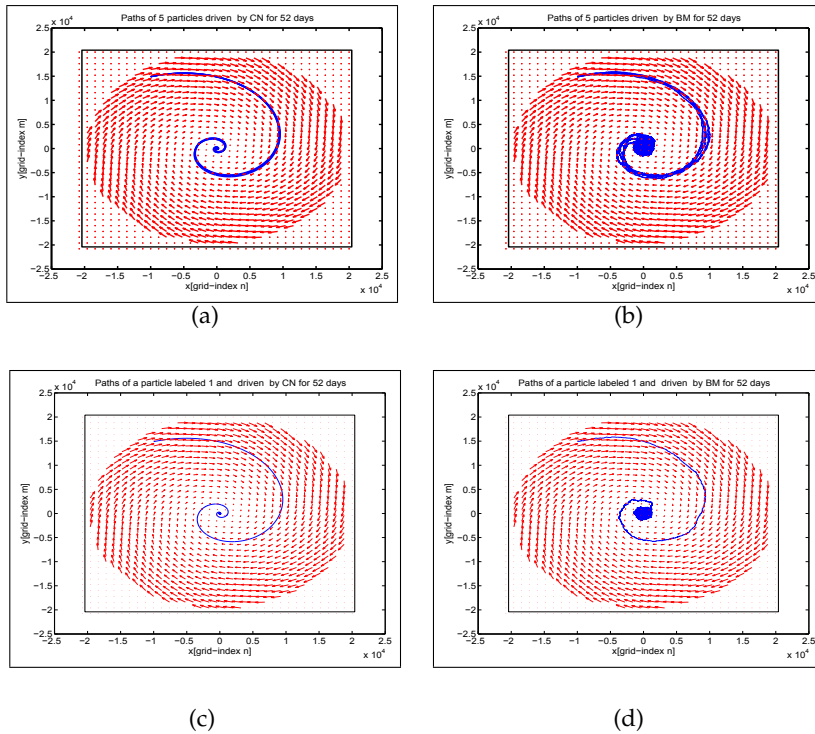
Fig. 5. Tracking of a single marked particle in the whirl pool starting from the location
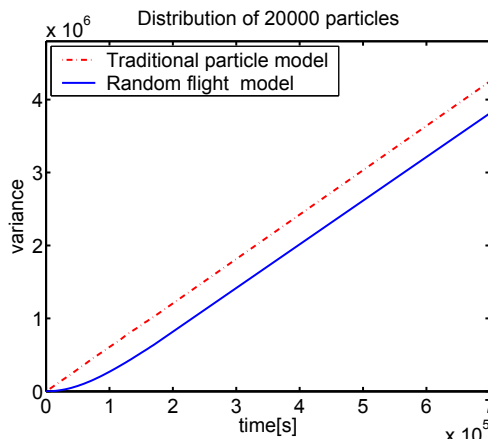
Fig. 6. The variance of a cloud of 20,000 particles in the idealized empty domain. There are two zones, one in which the variance grows quadratically with time for $t << T_L$ and another one it grows linearly with time for $t >> T_L$.

8. Acknowledgement

9. References

A.H. Jazwinski, Stochastic Processes and Filtering Theory, Academic Press, New York, (1970).pp 1-376.

A. W. Heemink, Stochastic modeling of Dispersion in shallow waters, Stochastic Hydrology & hydraulics, No. 4 (1990) pp. 161–174.

C.W. Gardiner,Handbook of Stochastic Methods for physics, Chemistry and the Natural Sciences, Springer-Verlag, Berlin,(2004), pp. 1–409.

D.J. Thomson, Criteria for the selection of stochastic models of particle trajectories in turbulent flows, J. Fluid Mech., No. 180 (1987), pp. 529–556.

G.N. Milstein, Numerical Integration of Stochastic Differential Equations, Kluwer Academic Publishers,(1995).

G.S. Stelling, Communications on construction of computational methods for shallow water flow problems. Delft University of Technology, The Netherlands, Ph.D. Thesis, 1983.

H.B. Fischer and E.J. List and R.C.Y. Koh and J. Imberger and N.H. Brooks, Mixing in Inland and Coastal Waters, Academic Press, New York, (1979) pp. 1–483.

H.M. Taylor and S. Karlin, An Introduction to Stochastic Modelling. Academic Press, San Diego and California USA, 1998

J. R. Hunter and P.D Crais and H.E. Phillips, On the use of random walk models with spatially variable diffusivity, J. of Computation Physics, No. 106 (1993), pp. 366–376.

J.W. Stijnen and A.W. Heemink, Numerical treatment of stochastic river quality models driven by colored noise, J. of water resources research, No. 7 (39)(2003), pp. 1–9.

L. Arnold, Stochastic differential equations. Theory and applications, Wiley, London (1974), pp. 1–228.

P.E. Kloeden and E. PlatenA, Numerical solutions of Stochastic Differential equations, Application of Mathematics, Springer-Verlag, New York,(2003), pp. 1-292

R.W.Barber and N.P. Volakos, Modelling depth-integrated pollution dispersion in the Gulf of Thermaikos using a Lagrangian particle technique, Proceedings of the 3rd International Conference on Water Resources Management, Portugal, WIT Transactions on Ecology and the Environment, N0.80 (2005), pp. 173–184.

P.S. Addison and Bo Qu and A. Nisbet and G.Pender, A non-Fickian particle-tracking diffusion model based on fractional Brownian motion, International Journal for numerical methods in fluids, No. 25 (1997) pp. 1373–1384.

W. M. Charles and A. W. Heemink and E. van den Berg, Stochastic particle models for transport problems in coastal waters, Proc. of the 29th International Conference on Coastal Engineering, WIT Transactions on The Built Environment, Vol 78, Lisbon, Portugal,2005, pp. 69–79 ISSN 1746-4498.

W. M. Charles and E. van den Berg and A. W. Heemink, Coloured noise for dispersion of contaminants in shallow waters, J. of Applied Mathematical Modelling: Volume 33, Issue 2, February 2009, pp. 1158-1171.

Complexity and stochastic synchronization in coupled map lattices and cellular automata

Ricardo López-Ruiz
Universidad de Zaragoza
Spain

Juan R. Sánchez
Universidad Nacional de Mar del Plata
Argentine

1. Introduction

Nowadays the question *'what is complexity?'* is a challenge to be answered. This question is triggering a great quantity of works in the frontier of physics, biology, mathematics and computer science. Even more when this century has been told to be the century of *complexity* (Hawking, 2000). Although there seems to be no urgency to answer the above question, many different proposals that have been developed to this respect can be found in the literature (Perakh, 2004). In this context, several articles concerning statistical complexity and stochastic processes are collected in this chapter.

Complex patterns generated by the time evolution of a one-dimensional digitalized coupled map lattice are quantitatively analyzed in Section 2. A method for discerning complexity among the different patterns is implemented. The quantitative results indicate two zones in parameter space where the dynamics shows the most complex patterns. These zones are located on the two edges of an absorbent region where the system displays spatio-temporal intermittency.

The synchronization of two stochastically coupled one-dimensional cellular automata (CA) is analyzed in Section 3. It is shown that the transition to synchronization is characterized by a dramatic increase of the statistical complexity of the patterns generated by the difference automaton. This singular behavior is verified to be present in several CA rules displaying complex behavior.

In Sections 4 and 5, we are concerned with the stability analysis of patterns in extended systems. In general, it has been revealed to be a difficult task. The many nonlinearly interacting degrees of freedom can destabilize the system by adding small perturbations to some of them. The impossibility to control all those degrees of freedom finally drives the dynamics toward a complex spatio-temporal evolution. Hence, it is of a great interest to develop techniques able to compel the dynamics toward a particular kind of structure. The application of such techniques forces the system to approach the stable manifold of the required pattern, and then the dynamics finally decays to that target pattern.

Synchronization strategies in extended systems can be useful in order to achieve such goal. In Section 4, we implement stochastic synchronization between the present configurations of a cellular automata and its precedent ones in order to search for constant patterns. In Section 5, this type of synchronization is specifically used to find symmetrical patterns in the evolution of a single automaton.

2. Complexity in Two-Dimensional Patterns Generated by Coupled Map Lattices

It should be kept in mind that in ancient epochs, time, space, mass, velocity, charge, color, etc. were only perceptions. In the process they are becoming concepts, different tools and instruments are invented for quantifying the perceptions. Finally, only with numbers the scientific laws emerge. In this sense, if by complexity it is to be understood that property present in all systems attached under the epigraph of 'complex systems', this property should be reasonably quantified by the different measures that were proposed in the last years. This kind of indicators is found in those fields where the concept of information is crucial. Thus, the effective measure of complexity (Grassberger, 1986) and the thermodynamical depth (Lloyd & Pagels, 1988) come from physics and other attempts such as algorithmic complexity (Chaitin, 1966; Kolmogorov, 1965), Lempel-Ziv complexity (Lempel & Ziv, 1976) and ϵ-machine complexity (Cruthfield, 1989) arise from the field of computational sciences.

In particular, taking into account the statistical properties of a system, an indicator called the *LMC (LópezRuiz-Mancini-Calbet) complexity* has been introduced (Lopez-Ruiz, 1994; Lopez-Ruiz et al., 1995). This magnitude identifies the entropy or information stored in a system and its disequilibrium i.e., the distance from its actual state to the probability distribution of equilibrium, as the two basic ingredients for calculating its complexity. If H denotes the *information* stored in the system and D is its *disequilibrium*, the LMC complexity C is given by the formula:

$$C(\bar{p}) = H(\bar{p}) \cdot D(\bar{p}) =$$

$$= -k \left(\sum_{i=1}^{N} p_i \log p_i \right) \cdot \left(\sum_{i=1}^{N} \left(p_i - \frac{1}{N} \right)^2 \right) \tag{1}$$

where $\bar{p} = \{p_i\}$, with $p_i \geq 0$ and $i = 1, \cdots, N$, represents the distribution of the N accessible states to the system, and k is a constant taken as $1/\log N$.

As well as the Euclidean distance D is present in the original LMC complexity, other kinds of disequilibrium measures have been proposed in order to remedy some statistical characteristics considered troublesome for some authors (Feldman & Crutchfield, 1998). In particular, some attention has been focused (Lin, 1991; Martin et al., 2003) on the Jensen-Shannon divergence D_{JS} as a measure for evaluating the distance between two different distributions $(\bar{p}_1, \bar{p}_2)$. This distance reads:

$$D_{JS}(\bar{p}_1, \bar{p}_2) = H(\pi_1 \bar{p}_1 + \pi_2 \bar{p}_2) - \pi_1 H(\bar{p}_1) - \pi_2 H(\bar{p}_2), \tag{2}$$

with π_1, π_2 the weights of the two probability distributions $(\bar{p}_1, \bar{p}_2)$ verifying $\pi_1, \pi_2 \geq 0$ and $\pi_1 + \pi_2 = 1$. The ensuing statistical complexity

$$C_{JS} = H \cdot D_{JS} \tag{3}$$

becomes intensive and also keeps the property of distinguishing among distinct degrees of periodicity (Lamberti et al., 2004). Here, we consider $\bar{p}_2$ the equiprobability distribution and $\pi_1 = \pi_2 = 0.5$.
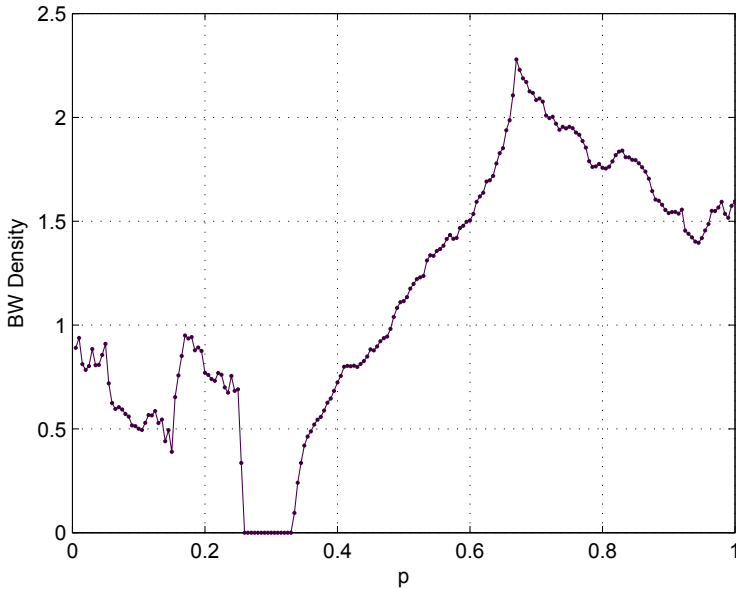
Fig. 1. β versus p. The β-statistics (or BW density) for each p is the rate between the number of *black* and *white* cells depicted by the system in the two-dimensional representation of its after-transient time evolution. (Computations have been performed with $\Delta p = 0.005$ for a lattice of 10000 sites after a transient of 5000 iterations and a running of other 2000 iterations).

As it can be straightforwardly seen, all these LMC-like complexities vanish both for completely ordered and for completely random systems as it is required for the correct asymptotic properties of a such well-behaved measure. Recently, they have been successfully used to discern situations regarded as complex in discrete systems out of equilibrium (Calbet & Lopez-Ruiz, 2001; Lovallo et al., 2005; Rosso et al., 2003; 2005; Shiner et al., 1999; Yu & Chen, 2000).

As an example, the local transition to chaos via intermittency (Pomeau & Manneville, 1980) in the logistic map, $x_{n+1} = \lambda x_n(1 - x_n)$ presents a sharp transition when C is plotted versus the parameter λ in the region around the instability for $\lambda \sim \lambda_t = 3.8284$. When $\lambda < \lambda_t$ the system approaches the laminar regime and the bursts become more unpredictable. The complexity increases. When the point $\lambda = \lambda_t$ is reached a drop to zero occurs for the magnitude C. The system is now periodic and it has lost its complexity. The dynamical behavior of the system is finally well reflected in the magnitude C (see (Lopez-Ruiz et al., 1995)).

When a one-dimensional array of such maps is put together a more complex behavior can be obtained depending on the coupling among the units. Ergo the phenomenon called *spatio-temporal intermittency* can emerge (Chate & Manneville, 1987; Houlrik, 1990; Rolf et al., 1998). This dynamical regime corresponds with a situation where each unit is weakly oscillating around a laminar state that is aperiodically and strongly perturbed for a traveling burst. In this case, the plot of the one-dimensional lattice evolving in time gives rise to complex patterns on the plane. If the coupling among units is modified the system can settle down in an absorbing phase where its dynamics is trivial (Argentina & Coullet, 1997; Zimmermann
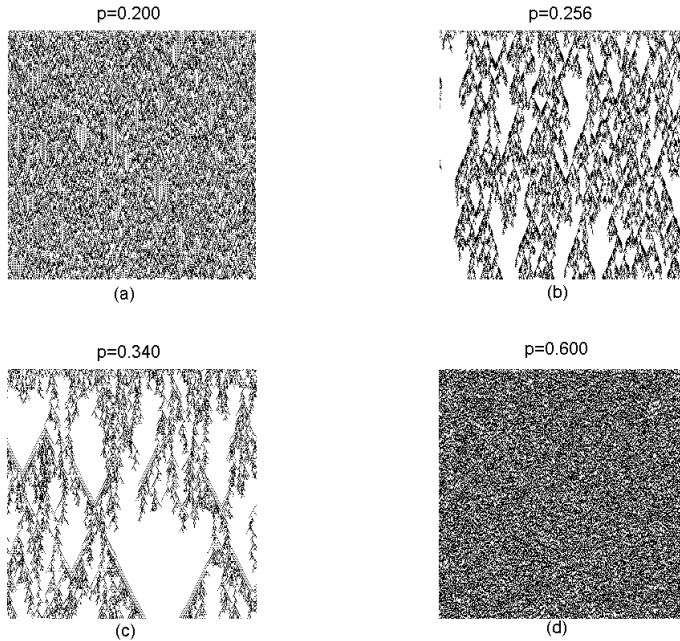
Fig. 2. Digitalized plot of the one-dimensional coupled map lattice (axe OX) evolving in time (axe OY) according to Eq. (4): if $x_i^n > 0.5$ the (i, n)-cell is put in white color and if $x_i^n < 0.5$ the (i, n)-cell is put in black color. The discrete time n is reset to zero after the transitory. (Lattices of 300×300 sites, i.e., $0 < i < 300$ and $0 < n < 300$).

et al., 2000) and then homogeneous patterns are obtained. Therefore an abrupt transition to spatio-temporal intermittency can be depicted by the system (Menon et al., 2003; Pomeau, 1986) when modifying the coupling parameter.

In this section, we are concerned with measuring C and C_{JS} in a such transition for a coupled map lattice of logistic type (Sanchez & Lopez-Ruiz, 2005-a). Our system will be a line of sites, $i = 1, \ldots, L$, with periodic boundary conditions. In each site i a local variable x_i^n evolves in time (n) according to a discrete logistic equation. The interaction with the nearest neighbors takes place via a multiplicative coupling:

$$x_i^{n+1} = (4 - 3pX_i^n)x_i^n(1 - x_i^n), \tag{4}$$

where p is the parameter of the system measuring the strength of the coupling ($0 < p < 1$). The variable X_i^n is the digitalized local mean field,

$$X_i^n = nint\left[\frac{1}{2}\left(x_{i+1}^n + x_{i-1}^n\right)\right], \tag{5}$$

with $nint(.)$ the integer function rounding its argument to the nearest integer. Hence $X_i^n = 0$ or 1.

There is a biological motivation behind this kind of systems (Lopez-Ruiz & Fournier-Prunaret, 2004; Lopez-Ruiz, 2005). It could represent a *colony of interacting competitive individuals*. They
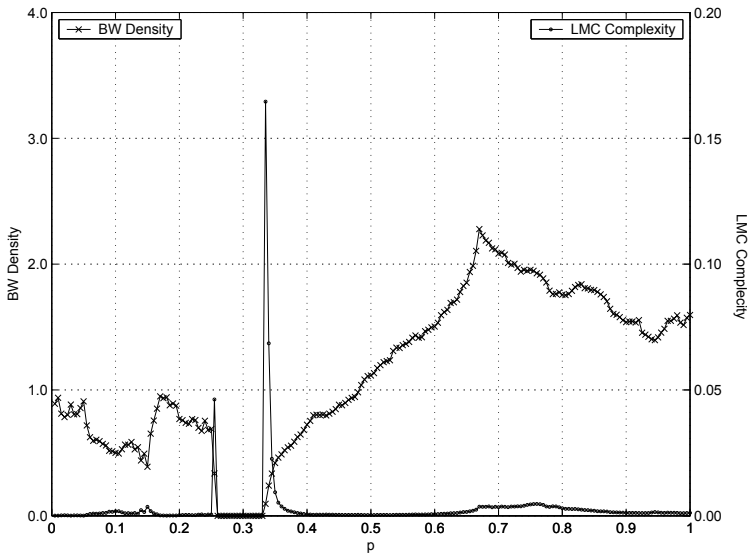
Fig. 3. (●) C versus p. Observe the peaks of the LMC complexity located just on the borders of the absorbent region $0.258 < p < 0.335$, where $\beta = 0$ (×). (Computations have been performed with $\Delta p = 0.005$ for a lattice of 10000 sites after a transient of 5000 iterations and a running of other 2000 iterations).

evolve randomly when they are independent ($p = 0$). If some competitive interaction ($p > 0$) among them takes place the local dynamics loses its erratic component and becomes chaotic or periodic in time depending on how populated the vicinity is. Hence, for bigger X_i^n more populated is the neighborhood of the individual i and more constrained is its free action. At a first sight, it would seem that some particular values of p could stabilize the system. In fact, this is the case. Let us choose a number of individuals for the colony ($L = 500$ for instance), let us initialize it randomly in the range $0 < x_i < 1$ and let it evolve until the asymptotic regime is attained. Then the *black/white* statistics of the system is performed. That is, the state of the variable x_i is compared with the critical level 0.5 for $i = 1, \ldots, L$: if $x_i > 0.5$ the site i is considered *white* (high density cell) and a counter N_w is increased by one, or if $x_i < 0.5$ the site i is considered *black* (low density cell) and a counter N_b is increased by one. This process is executed in the stationary regime for a set of iterations. The *black/white* statistics is then the rate $\beta = N_b/N_w$. If β is plotted versus the coupling parameter p the Figure 1 is obtained. The region $0.258 < p < 0.335$ where β vanishes is remarkable. As stated above, β represents the rate between the number of black cells and the number of white cells appearing in the two-dimensional digitalized representation of the colony evolution. A whole white pattern is obtained for this range of p. The phenomenon of spatio-temporal intermittency is displayed by the system in the two borders of this parameter region (Fig. 2). Bursts of low density (black color) travel in an irregular way through the high density regions (white color). In this case two-dimensional complex patterns are shown by the time evolution of the system (Fig. 2b-c). If the coupling p is far enough from this region, i.e., $p < 0.25$ or $p > 0.4$, the absorbent region loses its influence on the global dynamics and less structured and more random patterns than
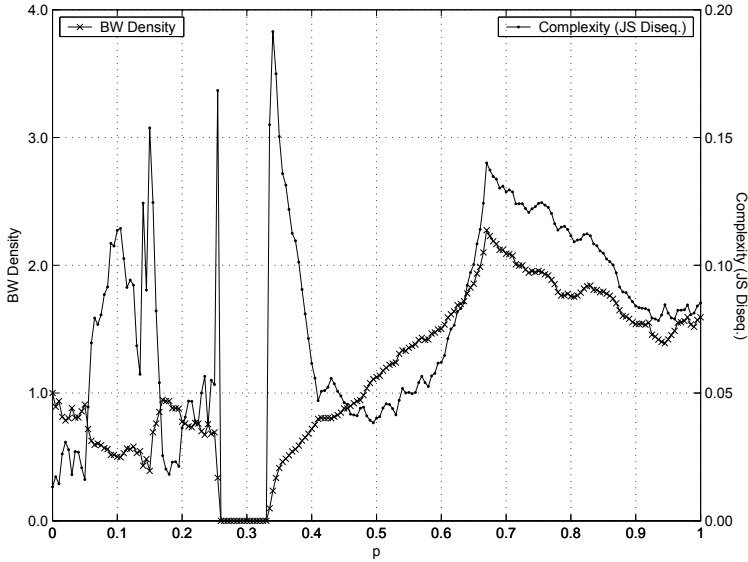
Fig. 4. ($\cdot$) C_{JS} versus p. The peaks of this modified LMC complexity are also evident just on the borders of the absorbent region $0.258 < p < 0.335$, where $\beta = 0$ ($\times$). (Computations have been performed with $\Delta p = 0.005$ for a lattice of 10000 sites after a transient of 5000 iterations and a running of other 2000 iterations).

before are obtained (Fig. 2a-d). For $p = 0$ we have no coupling of the maps, and each map generates so called fully developed chaos, where the invariant measure is well-known to be symmetric around 0.5. From this we conclude that $\beta(p = 0) = 1$. Let us observe that this symmetrical behavior of the invariant measure is broken for small p, and β decreases slightly in the vicinity of $p = 0$.

If the LMC complexities are quantified as function of p, our *intuition* is confirmed. The method proposed in (Lopez-Ruiz et al., 1995) to calculate C is now adapted to the case of two-dimensional patterns. First, we let the system evolve until the asymptotic regime is attained. This transient is discarded. Then, for each time n, we map the whole lattice in a binary sequence: 0 if $x_i^n < 0.5$ and 1 if $x_i^n > 0.5$, for $i = 1, \ldots, L$. This L-binary string is analyzed by blocks of n_o bits, where n_o can be considered the scale of observation. For this scale, there are 2^{n_o} possible states but only some of them are accessible. These accessible states as well as their probabilities are found in the L-binary string. Next, the magnitudes H, D, D_{JS}, C and C_{JS} are directly calculated for this particular time n by applying the formulas (1-3). We repeat this process for a set of successive time units $(n, n+1, \cdots, n+m)$. The mean values of H, D, D_{JS}, C and C_{JS} for these m time units are finally obtained and plotted in Fig. 3-4.

Figures 3,4 show the result for the case of $n_o = 10$. Let us observe that the highest C and C_{JS} are reached when the dynamics displays spatio-temporal intermittency, that is, the *most complex patterns* are obtained for those values of p that are located on the borders of the absorbent region $0.258 < p < 0.335$. Thus the plot of C and C_{JS} versus p shows two tight peaks around the values $p = 0.256$ and $p = 0.34$ (Fig. 3,4). Let us remark that the LMC complexity C can be neglected far from the absorbent region. Contrarily to this behavior, the magnitude C_{JS} also

shows high peaks in some other sharp transition of β located in the region $0 < p < 25$, and an intriguing correlation with the *black/white* statistics in the region $0.4 < p < 1$. All these facts as well as the stability study of the different dynamical regions of system (4) are not the object of the present writing but they deserve attention and a further study.

If the detection of complexity in the two-dimensional case requires to identify some sharp change when comparing different patterns, those regions in the parameter space where an abrupt transition happens should be explored in order to obtain the most complex patterns. Smoothness seems not to be at the origin of complexity. As well as a selected few distinct molecules among all the possible are in the basis of life (McKay, 2004), discreteness and its spiky appearance could indicate the way towards complexity. Let us recall that the distributions with the highest LMC complexity are just those distributions with a spiky-like appearance (Anteneodo & Plastino, 1996; Calbet & Lopez-Ruiz, 2001). In this line, the striking result here exposed confirms the capability of the LMC-like complexities for signaling a transition to complex behavior when regarding two-dimensional patterns (Sanchez & Lopez-Ruiz, 2005-b).

3. Detecting Synchronization in Cellular Automata by Complexity Measurements

Despite all the efforts devoted to understand the meaning of *complexity*, we still do not have an instrument in the laboratories specially designed for quantifying this property. Maybe this is not the final objective of all those theoretical attempts carried out in the most diverse fields of knowledge in the last years (Bennett, 1985; Chaitin, 1966; Cruthfield, 1989; Grassberger, 1986; Kolmogorov, 1965; Lempel & Ziv, 1976; Lloyd & Pagels, 1988; Shiner et al., 1999), but, for a moment, let us think in that possibility.

Similarly to any other device, our hypothetical apparatus will have an input and an output. The input could be the time evolution of some variables of the system. The instrument records those signals, analyzes them with a proper program and finally screens the result in the form of a *complexity measurement*. This process is repeated for several values of the parameters controlling the dynamics of the system. If our interest is focused in the *most complex configuration* of the system we have now the possibility of tuning such an state by regarding the complexity plot obtained at the end of this process.

As a real applicability of this proposal, let us apply it to an *à-la-mode* problem. The clusterization or synchronization of chaotic coupled elements was put in evidence at the beginning of the nineties (Kaneko, 1989; Lopez-Ruiz & Perez-Garcia, 1991). Since then, a lot of publications have been devoted to this subject (Boccaletti et al., 2002). Let us consider one particular of these systems to illuminate our proposal.

(1) SYSTEM: We take two coupled elementary one dimensional cellular automata (CA: see next section in which CA are concisely explained) displaying complex spatio-temporal dynamics (Wolfram, 1983). It has been shown that this system can undergo through a synchronization transition (Morelli & Zanette, 1998). The transition to full synchronization occurs at a critical value p_c of a synchronization parameter p. Briefly the numerical experiment is as follows. Two L-cell CA with the same evolution rule Φ are started from different random initial conditions for each automaton. Then, at each time step, the dynamics of the coupled CA is governed by the successive application of two evolution operators; the independent evolution of each CA according to its corresponding rule Φ and the application of a stochastic operator that compares the states σ_i^1 and σ_i^2 of all the cells, $i = 1, ... L$, in each automaton. If $\sigma_i^1 = \sigma_i^2$, both states are kept invariant. If $\sigma_i^1 \neq \sigma_i^2$, they are left unchanged with probability $1 - p$, but both states are updated either to σ_i^1 or to σ_i^2 with equal probability $p/2$. It is shown
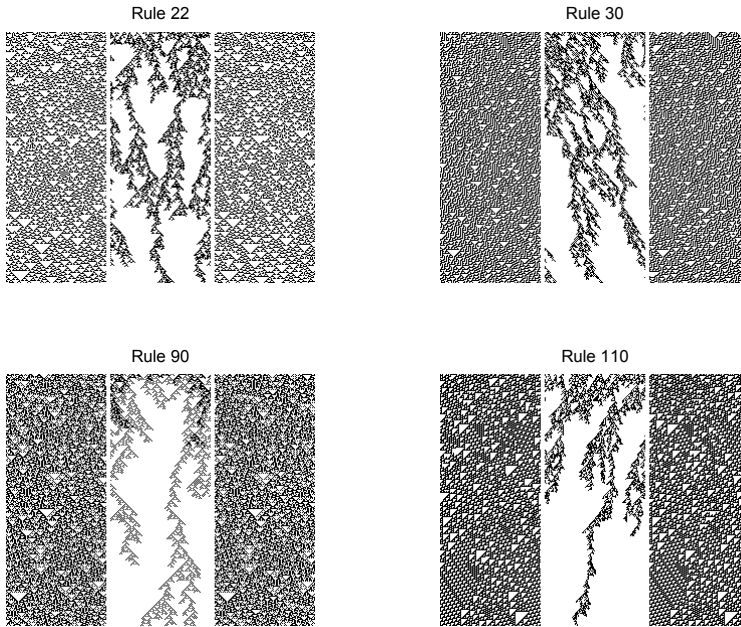
Fig. 5. Spatio-temporal patterns just above the synchronization transition. The left and the right plots show 250 successive states of the two coupled automata and the central plot is the corresponding difference automaton for the rules 22, 30, 90 and 110. The number of sites is $L = 100$ and the coupling probability is $p = 0.23$.

in reference (Morelli & Zanette, 1998) that there exists a critical value of the synchronization parameter ($p_c = 0.193$ for the rule 18) above for which full synchronization is achieved.

(2) DEVICE: We choose a particular instrument to perform our measurements, that is capable of displaying the value of the *LMC complexity* (*C*) (Lopez-Ruiz et al., 1995) defined as in Eq. (1), $C(\{\rho_i\}) = H(\{\rho_i\}) \cdot D(\{\rho_i\})$, where $\{\rho_i\}$ represents the set of probabilities of the N accessible discrete states of the system, with $\rho_i \geq 0$, $i = 1, \cdots , N$, and k is a constant. If $k = 1/logN$ then we have the normalized complexity. C is a statistical measure of complexity that identifies the entropy or information stored in a system and its disequilibrium, i.e., the distance from its actual state to the probability distribution of equilibrium, as the two basic ingredients for calculating the complexity of a system. This quantity vanishes both for completely ordered and for completely random systems giving then the correct asymptotic properties required for a such well-behaved measure, and its calculation has been useful to successfully discern many situations regarded as complex.

(3) INPUT: In particular, the evolution of two coupled CA evolving under the rules 22, 30, 90 and 110 is analyzed. The pattern of the difference automaton will be the input of our device. In Fig. 5 it is shown for a coupling probability $p = 0.23$, just above the synchronization transition. The left and the right plots show 250 successive states of the two automata, whereas the central plot displays the corresponding difference automaton. Such automaton is constructed by comparing one by one all the sites ($L = 100$) of both automata and putting zero when the states σ_i^1 and σ_i^2, $i = 1 \ldots L$, are equal or putting one otherwise. It is worth to observe that
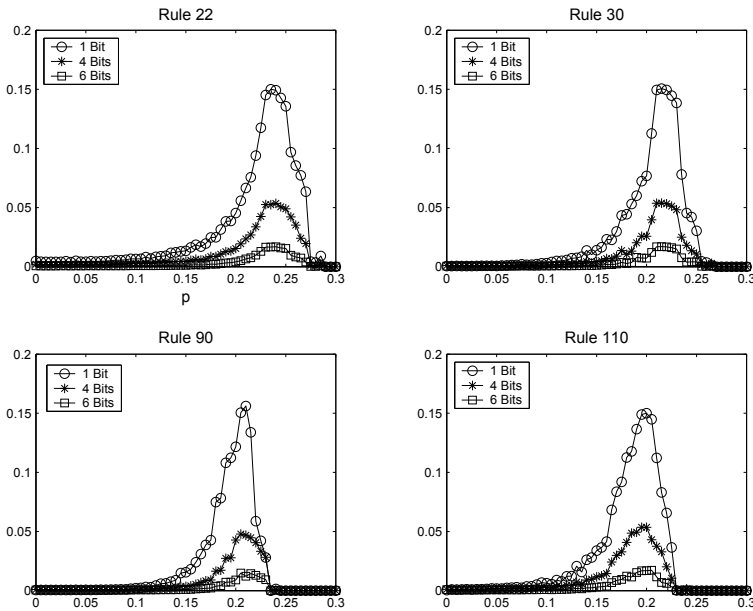
Fig. 6. The normalized complexity C versus the coupling probability p for different scales of observation: $n_o = 1$ (○), $n_o = 4$ (⋆) and $n_o = 6$ (□). C has been calculated over the last 300 iterations of a running of 600 of them for a lattice with $L = 1000$ sites. The synchronization transition is clearly depicted around $p \approx 0.2$ for the different rules.

the difference automaton shows an interesting *complex structure* close to the synchronization transition. This complex pattern is only found in this region of parameter space. When the system if fully synchronized the difference automaton is composed by zeros in all the sites, while when there is no synchronization at all the structure of the difference automaton is completely random.

(4) METHOD OF MEASUREMENT: How to perform the measurement of C for such two-dimensional patterns has been presented in the former section (Sanchez & Lopez-Ruiz, 2005-a). We let the system evolve until the asymptotic regime is attained. The variable σ_i^d in each cell of the difference pattern is successively translated to an unique binary sequence when the variable i covers the spatial dimension of the lattice, $i = 1, \ldots, L$, and the time variable n is consecutively increased. This binary string is analyzed in blocks of n_o bits, where n_o can be considered the scale of observation. The accessible states to the system among the 2^{n_o} possible states is found as well as their probabilities. Then, the magnitudes H, D and C are directly calculated and screened by the device.

(5) OUTPUT: The results of the measurement are shown in Fig. 6. The normalized complexity C as a function of the synchronization parameter p is plotted for different coupled one-dimensional CA that evolve under the rules 22, 30, 90 and 110 , which are known to generate complex patterns. All the plots of Fig. 6 were obtained using the following parameters: number of cell of the automata, $L = 1000$; total evolution time, $T = 600$ steps. For all the cases and scales analyzed, the statistical complexity C shows a dramatic increase close to the syn-

chronization transition. It reflects the complex structure of the difference automaton and the capability of the measurement device here proposed for *clearly signaling* the synchronization transition of two coupled CA.

These results are in agreement with the measurements of C performed in the patterns generated by a one-dimensional logistic coupled map lattice in the former section (Sanchez & Lopez-Ruiz, 2005-a). There the *LMC statistical complexity* (C) also shows a singular behavior close to the two edges of an absorbent region where the lattice displays spatio-temporal intermittency. Hence, in our present case, the synchronization region of the coupled systems can be interpreted as an absorbent region of the difference system. In fact, the highest complexity is reached on the border of this region for $p \approx 0.2$. The parallelism between both systems is therefore complete.

4. Self-Synchronization of Cellular Automata

Cellular automata (CA) are discrete dynamical systems, discrete both in space and time. The simplest one dimensional version of a cellular automaton is formed by a lattice of N sites or cells, numbered by an index $i = 1, \ldots, N$, and with periodic boundary conditions. In each site, a local variable σ_i taking a binary value, either 0 or 1, is asigned. The binary string $\sigma(t)$ formed by all sites values at time t represents a configuration of the system. The system evolves in time by the application of a rule Φ. A new configuration $\sigma(t + 1)$ is obtained under the action of the rule Φ on the state $\sigma(t)$. Then, the evolution of the automata can be writen as

$$\sigma(t + 1) = \Phi\left[\sigma(t)\right]. \tag{6}$$

If coupling among nearest neighbors is used, the value of the site i, $\sigma_i(t + 1)$, at time $t + 1$ is a function of the value of the site itself at time t, $\sigma_i(t)$, and the values of its neighbors $\sigma_{i-1}(t)$ and $\sigma_{i+1}(t)$ at the same time. Then, the local evolution is expressed as

$$\sigma_i(t + 1) = \phi(\sigma_{i-1}(t), \sigma_i(t), \sigma_{i+1}(t)), \tag{7}$$

being ϕ a particular realization of the rule Φ. For such particular implementation, there will be 2^3 different local input configurations for each site and, for each one of them, a binary value can be assigned as output. Therefore there will be 2^8 different rules ϕ, also called the *Wolfram rules*. Each one of these rules produces a different dynamical evolution. In fact, dynamical behavior generated by all 256 rules were already classified in four generic classes. The reader interested in the details of such classification is addressed to the original reference (Wolfram, 1983).

CA provide us with simple dynamical systems, in which we would like to essay different methods of synchronization. A stochastic synchronization technique was introduced in (Morelli & Zanette, 1998) that works in synchronizing two CA evolving under the same rule Φ. The two CA are started from different initial conditions and they are supposed to have partial knowledge about each other. In particular, the CA configurations, $\sigma^1(t)$ and $\sigma^2(t)$, are compared at each time step. Then, a fraction p of the total different sites are made equal (synchronized). The synchronization is stochastic since the location of the sites that are going to be equal is decided at random. Hence, the dynamics of the two coupled CA, $\sigma(t) = (\sigma^1(t), \sigma^2(t))$, is driven by the successive application of two operators:

1. the deterministic operator given by the CA evolution rule Φ, $\Phi[\sigma(t)] = (\Phi[\sigma^1(t)], \Phi[\sigma^2(t)])$, and
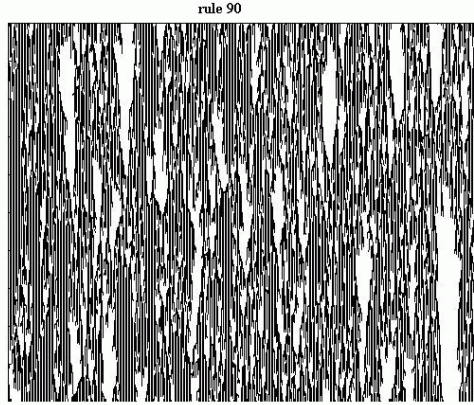
Fig. 7. Rule 90 has two stable patterns: one repeats the 011 string and the other one the 00 string. Such patterns are reached by the first self-synchronization method but there is a dynamical competition between them. In this case $p = 0.9$. Binary value 0 is represented in white and 1 in black. Time goes from top to bottom.

2. the stochastic operator Γ_p that produces the result $\Gamma_p[\sigma(t)]$, in such way that, if the sites are different ($\sigma_i^1 \neq \sigma_i^2$), then Γ_p sets both sites equal to σ_i^1 with the probability $p/2$ or equal to σ_i^2 with the same probability $p/2$. In any other case Γ_p leaves the sites unchanged.

Therefore the temporal evolution of the system can be written as

$$\sigma(t+1) = (\Gamma_p \circ \Phi)[\sigma(t)] = \Gamma_p[(\Phi[\sigma^1(t)], \Phi[\sigma^2(t)])]. \tag{8}$$

A simple way to visualize the transition to synchrony can be done by displaying the evolution of the difference automaton (DA),

$$\delta_i(t) = |\sigma_i^1(t) - \sigma_i^2(t)|. \tag{9}$$

The mean density of active sites for the DA

$$\rho(t) = \frac{1}{N} \sum_{i=1}^{N} \delta_i(t), \tag{10}$$

represents the Hamming distance between the automata and verifies $0 \leq \rho \leq 1$. The automata will be synchronized when $\lim_{t \to \infty} \rho(t) = 0$. As it has been described (Morelli & Zanette, 1998) that two different dynamical regimes, controlled by the parameter p, can be found in the system behavior:

$$p < p_c \quad \rightarrow \quad \lim_{t \to \infty} \rho(t) \neq 0 \text{ (no synchronization)},$$
$$p > p_c \quad \rightarrow \quad \lim_{t \to \infty} \rho(t) = 0 \text{ (synchronization)},$$
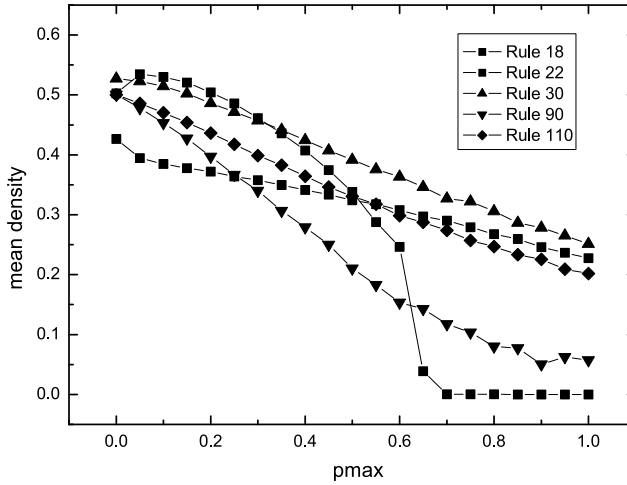
Fig. 8. Mean density ρ vs. $pmax = \tilde{p}$ for different rules evolving under the second synchronization method. The existence of a transition to a synchronized state can be clearly observed for rule 18.

being p_c the parameter for which the transition to the synchrony occurs. When $p \lesssim p_c$ complex structures can be observed in the DA time evolution. In Fig. 5, typical cases of such behavior are shown near the synchronization transition. Lateral panels represent both CA evolving in time where the central strip displays the evolution of the corresponding DA. When p comes close to the critical value p_c the evolution of $\delta(t)$ becomes rare and resembles the problem of structures trying to percolate in the plane (Pomeau, 1986). A method to detect this kind of transition, based in the calculation of a statistical measure of complexity for patterns, has been proposed in the former sections (Sanchez & Lopez-Ruiz, 2005-a), (Sanchez & Lopez-Ruiz, 2005-b).

4.1 First Self-Synchronization Method

Let us now take a single cellular automaton (Ilachinski, 2001; Toffoli & Margolus, 1987). If $\sigma^1(t)$ is the state of the automaton at time t, $\sigma^1(t) = \sigma(t)$, and $\sigma^2(t)$ is the state obtained from the application of the rule Φ on that state, $\sigma^2(t) = \Phi[\sigma^1(t)]$, then the operator Γ_p can be applied on the pair $(\sigma^1(t), \sigma^2(t))$, giving rise to the evolution law

$$\sigma(t+1) = \Gamma_p[(\sigma^1(t), \sigma^2(t))] = \Gamma_p[(\sigma(t), \Phi[\sigma(t)])]. \tag{11}$$

The application of the Γ_p operator is as follows. When $\sigma_i^1 \neq \sigma_i^2$, the sites i of the state $\sigma^2(t)$ are updated to the correspondent values taken in $\sigma^1(t)$ with a probability p. The updated array $\sigma^2(t)$ is the new state $\sigma(t+1)$.
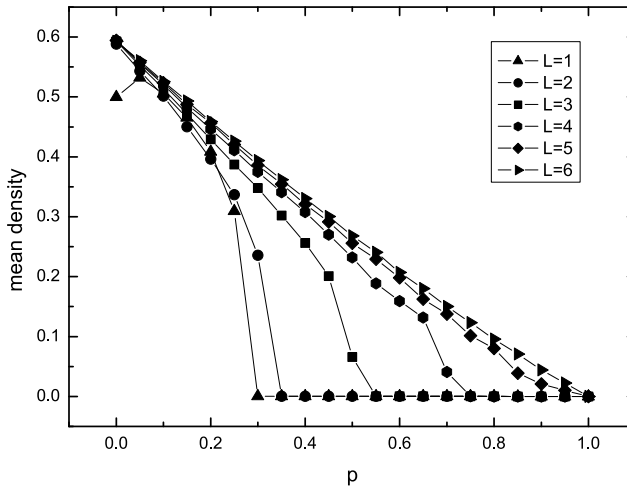
Fig. 9. Mean density ρ vs. p for rule 18 evolving under the third self-synchronization method. The existence of a transition to a synchronized state can be observed despite of the randomness in the election of neighbors within a range L, up to $L = 4$.

It is worth to observe that if the system is initialized with a configuration constant in time for the rule Φ, $\Phi[\sigma] = \sigma$, then this state σ is not modified when the dynamic equation (11) is applied. Hence the evolution will produce a pattern constant in time. However, in general, this stability is marginal. A small modification of the initial condition gives rise to patterns variable in time. In fact, as the parameter p increases, a competition among the different marginally stable structures takes place. The dynamics drives the system to stay close to those states, although oscillating continuously and randomly among them. Hence, a complex spatio-temporal behavior is obtained. Some of these patterns can be seen in Fig. 7. However, in rule 18, the pattern becomes stable and, independently of the initial conditions, the system evolves toward this state, which is the null pattern in this case (Sanchez & Lopez-Ruiz, 2006).

4.2 Second Self-Synchronization Method

Now we introduce a new stochastic element in the application of the operator Γ_p. To differentiate from the previous case we call it $\tilde{\Gamma}_{\tilde{p}}$. The action of this operator consists in applying at each time the operator Γ_p, with p chosen at random in the interval $(0, \tilde{p})$. The evolution law of the automaton is in this case:

$$\sigma(t+1) = \tilde{\Gamma}_{\tilde{p}}[(\sigma^1(t), \sigma^2(t))] = \tilde{\Gamma}_{\tilde{p}}[(\sigma(t), \Phi[\sigma(t)])]. \tag{12}$$

The DA density between the present state and the previous one, defined as $\delta(t) = \mid \sigma(t) - \sigma(t-1) \mid$, is plotted as a function of $\tilde{p}$ for different rules Φ in Fig. 8. Only when the system becomes self-synchronized there will be a fall to zero in the DA density. Let us observe again
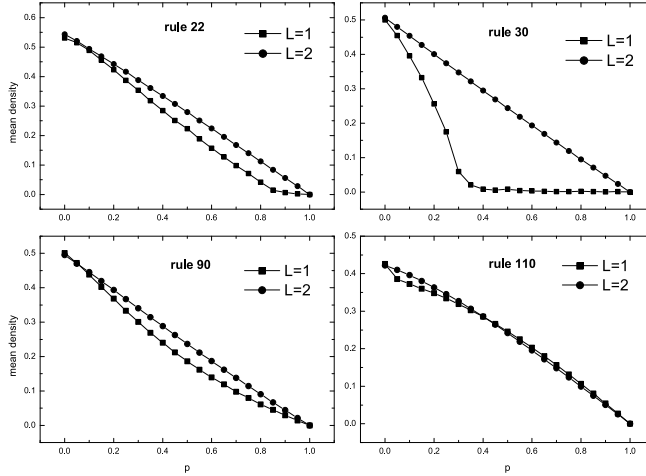
Fig. 10. Mean density ρ vs. p for different rules evolving under the third self-synchronization method. The density of the system decreases linearly with p.

that the behavior reported in the first self-synchronization method is newly obtained in this case. Rule 18 undergoes a phase transition for a critical value of $\tilde{p}$. For $\tilde{p}$ greater than the critical value, the method is able to find the stable structure of the system (Sanchez & Lopez-Ruiz, 2006). For the rest of the rules the freezing phase is not found. The dynamics generates patterns where the different marginally stable structures randomly compete. Hence the DA density decays linearly with $\tilde{p}$ (see Fig. 8).

4.3 Third Self-Synchronization Method

At last, we introduce another type of stochastic element in the application of the rule Φ. Given an integer number L, the surrounding of site i at each time step is redefined. A site i_l is randomly chosen among the L neighbors of site i to the left, $(i - L, \ldots, i - 1)$. Analogously, a site i_r is randomly chosen among the L neighbors of site i to the right, $(i + 1, \ldots, i + L)$. The rule Φ is now applied on the site i using the triplet (i_l, i, i_r) instead of the usual nearest neighbors of the site. This new version of the rule is called Φ_L, being $\Phi_{L=1} = \Phi$. Later the operator Γ_p acts in identical way as in the first method. Therefore, the dynamical evolution law is:

$$\sigma(t+1) = \Gamma_p[(\sigma^1(t), \sigma^2(t))] = \Gamma_p[(\sigma(t), \Phi_L[\sigma(t)])]. \tag{13}$$

The DA density as a function of p is plotted in Fig. 9 for the rule 18 and in Fig. 10 for other rules. It can be observed again that the rule 18 is a singular case that, even for different L, maintains the memory and continues to self-synchronize. It means that the influence of the rule is even more important than the randomness in the election of the surrounding sites. The system self-synchronizes and decays to the corresponding stable structure. Contrary, for the rest of the rules, the DA density decreases linearly with p even for $L = 1$ as shown in Fig. 10.

Rule 18   Rule 150
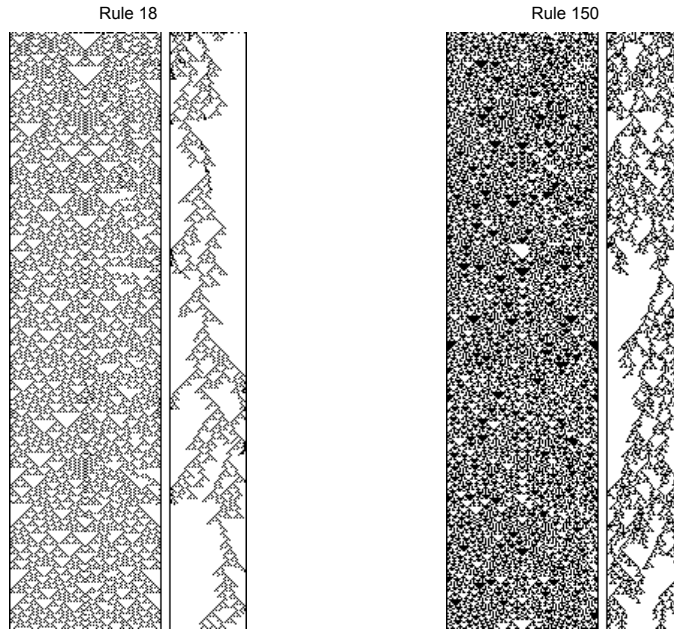
Fig. 11. Space-time configurations of automata with $N = 100$ sites iterated during $T = 400$ time steps evolving under rules 18 and 150 for $p \lesssim p_c$. Left panels show the automaton evolution in time (increasing from top to bottom) and the right panels display the evolution of the corresponding DA.

The systems oscillate randomly among their different marginally stable structures as in the previous methods (Sanchez & Lopez-Ruiz, 2006).

5. Symmetry Pattern Transition in Cellular Automata with Complex Behavior

In this section, the stochastic synchronization method introduced in the former sections (Morelli & Zanette, 1998) for two CA is specifically used to find symmetrical patterns in the evolution of a single automaton. To achieve this goal the stochastic operator, below described, is applied to sites symmetrically located from the center of the lattice. It is shown that a *symmetry* transition take place in the spatio-temporal pattern. The transition forces the automaton to evolve toward complex patterns that have mirror symmetry respect to the central axe of the pattern. In consequence, this synchronization method can also be interpreted as a control technique for stabilizing complex symmetrical patterns.

Cellular automata are extended systems, in our case one-dimensional strings composed of N sites or cells. Each site is labeled by an index $i = 1, \ldots, N$, with a local variable s_i carrying a binary value, either 0 or 1. The set of sites values at time t represents a configuration (state or pattern) σ_t of the automaton. During the automaton evolution, a new configuration σ_{t+1} at time $t + 1$ is obtained by the application of a rule or operator Φ to the present configuration (see former section):

$$\sigma_{t+1} = \Phi \left[\sigma_t \right]. \tag{14}$$

Rule 18   Rule 150
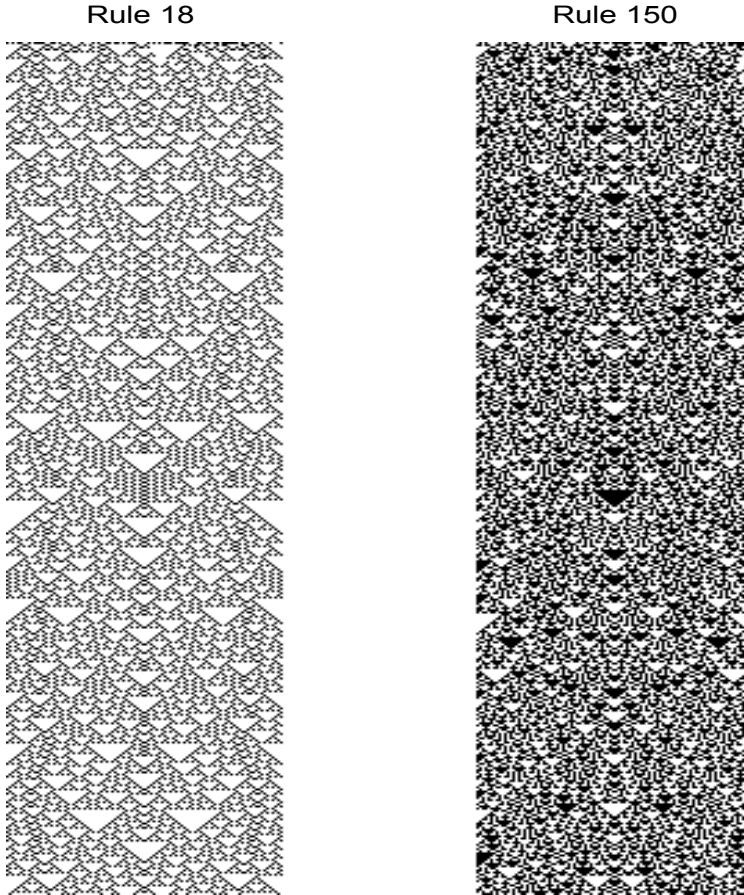
Fig. 12. Time configurations of automata with $N = 100$ sites iterated during $T = 400$ time steps evolving under rules 18 and 150 using $p > p_c$. The space symmetry of the evolving patterns is clearly visible.

5.1 Self-Synchronization Method by Symmetry

Our present interest (Sanchez & Lopez-Ruiz, 2008) resides in those CA evolving under rules capable to show asymptotic complex behavior (rules of class III and IV). The technique applied here is similar to the synchronization scheme introduced by Morelli and Zanette (Morelli & Zanette, 1998) for two CA evolving under the same rule Φ. The strategy supposes that the two systems have a *partial* knowledge one about each other. At each time step and after the application of the rule Φ, both systems compare their present configurations $\Phi[\sigma_t^1]$ and $\Phi[\sigma_t^2]$ along all their extension and they synchronize a percentage p of the total of their different sites. The location of the percentage p of sites that are going to be put equal is decided at random and, for this reason, it is said to be an stochastic synchronization. If we call this stochastic operator Γ_p, its action over the couple $(\Phi[\sigma_t^1], \Phi[\sigma_t^2])$ can be represented by the expression:

$$(\sigma_{t+1}^1, \sigma_{t+1}^2) = \Gamma_p(\Phi[\sigma_t^1], \Phi[\sigma_t^2]) = (\Gamma_p \circ \Phi)(\sigma_t^1, \sigma_t^2). \tag{15}$$
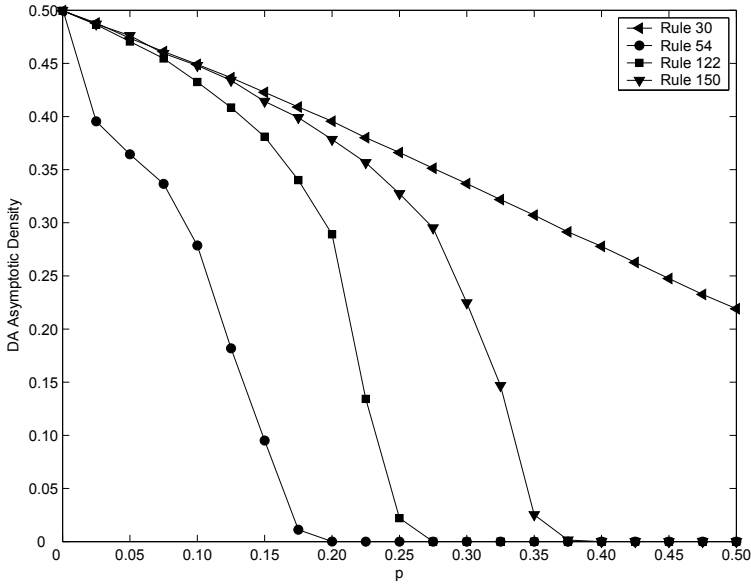
Fig. 13. Asymptotic density of the DA for different rules is plotted as a function of the coupling probability p. Different values of p_c for each rule appear clearly at the points where $\rho \to 0$. The automata with $N = 4000$ sites were iterated during $T = 500$ time steps. The mean values of the last 100 steps were used for density calculations.

Rule	18	22	30	54	60	90	105	110	122	126	146	150	182
p_c	0.25	0.27	1.00	0.20	1.00	0.25	0.37	1.00	0.27	0.30	0.25	0.37	0.25

Table 1. Numerically obtained values of the critical probability p_c for different rules displaying complex behavior. Rules that can not sustain symmetric patterns need fully coupling of the symmetric sites, i.e. ($p_c = 1$).

The same strategy can be applied to a single automaton with a even number of sites (Sanchez & Lopez-Ruiz, 2008). Now the evolution equation, $\sigma_{t+1} = (\Gamma_p \circ \Phi)[\sigma_t]$, given by the successive action of the two operators Φ and Γ_p, can be applied to the configuration σ_t as follows:

1. the deterministic operator Φ for the evolution of the automaton produces $\Phi[\sigma_t]$, and,

2. the stochastic operator Γ_p, produces the result $\Gamma_p(\Phi[\sigma_t])$, in such way that, if sites symmetrically located from the center are different, i.e. $s_i \neq s_{N-i+1}$, then Γ_p equals s_{N-i+1} to s_i with probability p. Γ_p leaves the sites unchanged with probability $1 - p$.

A simple way to visualize the transition to a symmetric pattern can be done by splitting the automaton in two subsystems (σ_t^1, σ_t^2),

- σ_t^1, composed by the set of sites $s(i)$ with $i = 1, \ldots, N/2$ and
- σ_t^2, composed the set of symmetrically located sites $s(N - i + 1)$ with $i = 1, \ldots, N/2$,

and displaying the evolution of the difference automaton (DA), defined as

$$\delta^t = |\sigma_t^1 - \sigma_t^2| . \tag{16}$$

The mean density of active sites for the difference automaton, defined as

$$\rho^t = \frac{2}{N} \sum_{i=1}^{N/2} \delta^t \tag{17}$$

represents the Hamming distance between the sets σ^1 and σ^2. It is clear that the automaton will display a symmetric pattern when $\lim_{t\to\infty} \rho^t = 0$. For class III and IV rules, a symmetry transition controlled by the parameter p is found. The transition is characterized by the DA behavior:

$$\text{when } p < p_c \quad \to \quad \lim_{t\to\infty} \rho^t \neq 0 \text{ (complex non-symmetric patterns)},$$
$$\text{when } p > p_c \quad \to \quad \lim_{t\to\infty} \rho^t = 0 \text{ (complex symmetric patterns)}.$$

The critical value of the parameter p_c signals the transition point.

In Fig. 11 the space-time configurations of automata evolving under rules 18 and 150 are shown for $p \lesssim p_c$. The automata are composed by $N = 100$ sites and were iterated during $T = 400$ time steps. Left panels show the automaton evolution in time (increasing from top to bottom) and the right panels display the evolution of the corresponding DA. For $p \lesssim p_c$, complex structures can be observed in the evolution of the DA. As p approaches its critical value p_c, the evolution of the DA become more stumped and reminds the problem of structures trying to percolate the plane (Pomeau, 1986; Sanchez & Lopez-Ruiz, 2005-a). In Fig. 12 the space-time configurations of the same automata are displayed for $p > p_c$. Now, the space symmetry of the evolving patterns is clearly visible.

Table 1 shows the numerically obtained values of p_c for different rules displaying complex behavior. It can be seen that some rules can not sustain symmetric patterns unless those patterns are forced to it by fully coupling the totality of the symmetric sites ($p_c = 1$). The rules whose local dynamics verify $\phi(s_1, s_0, s_2) = \phi(s_2, s_0, s_1)$ can evidently sustain symmetric patterns, and these structures are induced for $p_c < 1$ by the method here explained.

Finally, in Fig. 13 the asymptotic density of the DA, ρ^t for $t \to \infty$, for different rules is plotted as a function of the coupling probability p. The values of p_c for the different rules appear clearly at the points where $\rho \to 0$.

6. Conclusion

A method to measure statistical complexity in extended systems has been implemented. It has been applied to a transition to spatio-temporal complexity in a coupled map lattice and to a transition to synchronization in two stochastically coupled cellular automata (CA). The statistical indicator shows a peak just in the transition region, marking clearly the change of dynamical behavior in the extended system.

Inspired in stochastic synchronization methods for CA, different schemes for self-synchronization of a single automaton have also been proposed and analyzed. Self-synchronization of a single automaton can be interpreted as a strategy for searching and controlling the structures of the system that are constant in time. In general, it has been found that a competition among all such structures is established, and the system ends up oscillating randomly among them. However, rule 18 is a unique position among all rules because, even with random election of the neighbors sites, the automaton is able to reach the configuration constant in time.

Also a transition from asymmetric to symmetric patterns in time-dependent extended systems has been described. It has been shown that one dimensional cellular automata, started from

fully random initial conditions, can be forced to evolve into complex *symmetrical* patterns by stochastically coupling a proportion *p* of pairs of sites located at equal distance from the center of the lattice. A nontrivial critical value of *p* must be surpassed in order to obtain symmetrical patterns during the evolution. This strategy could be used as an alternative to classify the cellular automata rules -with complex behavior- between those that support time-dependent symmetric patterns and those which do not support such kind of patterns.

7. References

Anteneodo, C. & Plastino, A.R. (1996). Some features of the statistical LMC complexity. *Phys. Lett. A*, Vol. 223, No. 5, 348-354.

Argentina, M. & Coullet, P. (1997). Chaotic nucleation of metastable domains. *Phys. Rev. E*, Vol. 56, No. 3, R2359-R2362.

Bennett, C.H. (1985). Information, dissipation, and the definition of organization. In: *Emerging Syntheses in Science*, David Pines, (Ed.), 297-313, Santa Fe Institute, Santa Fe.

Boccaletti, S.; Kurths, J.; Osipov, G.; Valladares, D.L. & Zhou, C.S. (2002). The synchronization of chaotic systems. *Phys. Rep.*, Vol. 366, No. 1-2, 1-101.

Calbet, X. & López-Ruiz, R. (2001). Tendency toward maximum complexity in a non-equilibrium isolated system. *Phys. Rev. E*, Vol. 63, No.6, 066116(9).

Chaitin, G. (1966). On the length of programs for computing finite binary sequences. *J. Assoc. Comput. Mach.*, Vol. 13, No.4, 547-569.

Chaté, H. & Manneville, P. (1987). Transition to turbulence via spatio-temporal intermittency. *Phys. Rev. Lett.*, Vol. 58, No. 2, 112-115.

Crutchfield, J.P. & Young, K. (1989) Inferring statistical complexity. *Phys. Rev. Lett.*, Vol. 63, No. 2, 105-108.

Feldman D.P. & Crutchfield, J.P. (1998). Measures of statistical complexity: Why?. *Phys. Lett. A*, Vol. 238, No. 4-5, 244-252.

Grassberger, P. (1986). Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.*, Vol. 25, No. 9, 907-915.

Hawking, S. (2000). "I think the next century will be the century of complexity", In *San José Mercury News*, Morning Final Edition, January 23.

Houlrik, J.M.; Webman, I. & Jensen, M.H. (1990). Mean-field theory and critical behavior of coupled map lattices. *Phys. Rev. A*, Vol. 41, No. 8, 4210-4222.

Ilachinski, A. (2001). *Cellular Automata: A Discrete Universe*, World Scientific, Inc. River Edge, NJ.

Kaneko, K. (1989). Chaotic but regular posi-nega switch among coded attractors by cluster-size variation. *Phys. Rev. Lett.*, Vol. 63, No. 3, 219-223.

Kolmogorov, A.N. (1965). Three approaches to the definition of quantity of information. *Probl. Inform. Theory*, Vol. 1, No. 1, 3-11.

Lamberti, W.; Martin, M.T.; Plastino, A. & Rosso, O.A. (2004). Intensive entropic non-triviality measure. *Physica A*, Vol. 334, No. 1-2, 119-131.

Lempel, A. & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Trans. Inform. Theory*, Vol. 22, 75-81.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory*, Vol. 37, No. 1, 145-151.

Lloyd, S. & Pagels, H. (1988). Complexity as thermodynamic depth. *Ann. Phys. (NY)*, Vol. 188, No. 1, 186-213.

López-Ruiz, R. & Pérez-Garcia, C. (1991). Dynamics of maps with a global multiplicative coupling. *Chaos, Solitons and Fractals*, Vol. 1, No. 6, 511-528.

López-Ruiz, R. (1994). *On Instabilities and Complexity*, Ph. D. Thesis, Universidad de Navarra, Pamplona, Spain.

López-Ruiz, R.; Mancini, H.L. & Calbet, X. (1995). A statistical measure of complexity. *Phys. Lett. A*, Vol. 209, No. 5-6, 321-326.

López-Ruiz, R. & Fournier-Prunaret, D. (2004). Complex behaviour in a discrete logistic model for the simbiotic interaction of two species. *Math. Biosc. Eng.*, Vol. 1, No. 2, 307-324.

López-Ruiz, R. (2005). Shannon information, LMC complexity and Rényi entropies: a straightforward approach. *Biophys. Chem.*, Vol. 115, No. 2-3, 215-218.

Lovallo, M.; Lapenna, V. & Telesca, L. (2005) Transition matrix analysis of earthquake magnitude sequences. *Chaos, Solitons and Fractals*, Vol. 24, No. 1, 33-43.

Martin, M.T.; Plastino, A. & Rosso, O.A. (2003). Statistical complexity and disequilibrium. *Phys. Lett. A*, Vol. 311, No. 2-3, 126-132.

McKay, C.P. (2004). What is life?. *PLOS Biology*, Vol. 2, Vol. 9, 1260-1263.

Menon, G.I.; Sinha, S. & Ray, P. (2003). Persistence at the onset of spatio-temporal intermittency in coupled map lattices. *Europhys. Lett.*, Vol. 61, No. 1, 27-33.

Morelli, L.G. & Zanette, D.H. (1998). Synchronization of stochastically coupled cellular automata. *Phys. Rev. E*, Vol. 58, No. 1, R8-R11.

Perakh, M. (2004). Defining complexity. In online site: *On Talk Reason*, paper: www.talkreason.org/articles/complexity.pdf.

Pomeau, Y. & Manneville, P. (1980). Intermittent transition to turbulence in dissipative dynamical systems. *Commun. Math. Phys.*, Vol. 74, No.2, 189-197.

Pomeau, Y. (1986). Front motion, metastability and subcritical bifurcations in hydrodynamics. *Physica D*, Vol. 23, No. 1-3, 3-11.

Rolf, J.; Bohr, T. & Jensen, M.H. (1998). Directed percolation universality in asynchronous evolution of spatiotemporal intermittency. *Phys. Rev. E*, Vol. 57, No. 3, R2503-R2506 (1998).

Rosso, O.A.; Martín, M.T. & Plastino, A. (2003). Tsallis non-extensivity and complexity measures. *Physica A*, Vol. 320, 497-511.

Rosso, O.A.; Martín, M.T. & Plastino, A. (2005). Evidence of self-organization in brain electrical activity using wavelet-based informational tools. *Physica A*, Vol. 347, 444-464.

Sánchez, J.R. & López-Ruiz, R., a, (2005). A method to discern complexity in two-dimensional patterns generated by coupled map lattices. *Physica A*, Vol. 355, No. 2-4, 633-640.

Sánchez, J.R. & López-Ruiz, R., b, (2005). Detecting synchronization in spatially extended discrete systems by complexity measurements. *Discrete Dyn. Nat. Soc.*, Vol. 2005, No. 3, 337-342.

Sánchez, J.R. & López-Ruiz, R. (2006). Self-synchronization of Cellular Automata: an attempt to control patterns. *Lect. Notes Comp. Sci.*, Vol. 3993, No. 3, 353-359.

Sánchez, J.R. & López-Ruiz, R. (2008). Symmetry pattern transition in Cellular Automata with complex behavior. *Chaos, Solitons and Fractals*, vol. 37, No. 3, 638-642.

Shiner, J.S.; Davison, M. & Landsberg, P.T. (1999). Simple measure for complexity. *Phys. Rev. E*, Vol. 59, No. 2, 1459-1464.

Toffoli, T. & Margolus, N. (1987). *Cellular Automata Machines: A New Environment for Modeling*, The MIT Press, Cambridge, Massachusetts.

Wolfram, S. (1983). Statistical mechanics of cellular automata. *Rev. Mod. Phys.*, Vol. 55, No. 3, 601-644.

Yu, Z. & Chen, G. (2000). Rescaled range and transition matrix analysis of DNA sequences. *Comm. Theor. Phys. (Beijing China)*, Vol. 33, No. 4, 673-678.

Zimmermann, M.G.; Toral, R.; Piro, O. & San Miguel, M. (2000). Stochastic spatiotemporal intermittency and noise-induced transition to an absorbing phase. *Phys. Rev. Lett.*, Vol. 85, No. 17, 3612-3615.

Zero-sum stopping game associated with threshold probability

Yoshio Ohtsubo
Kochi University
Japan

Abstract

We consider a zero-sum stopping game (Dynkin's game) with a threshold probability criterion in discrete time stochastic processes. We first obtain fundamental characterization of value function of the game and optimal stopping times for both players as the result of the classical Dynkin's game, but the value function of the game and the optimal stopping time for each player depend upon a threshold value. We also give properties of the value function of the game with respect to threshold value. These are applied to an independent model and we explicitly find a value function of the game and optimal stopping times for both players in a special example.

1. Introduction

In the classical Dynkin's game, a standard criterion function is the expected reward (e.g. DynkinDynkin (1969) and NeveuNeveu (1975)). It is, however, known that the criterion is quite insufficient to characterize the decision problem from the point of view of the decision maker and it is necessary to select other criteria to reflect the variability of risk features for the problem (e.g. WhiteWhite (1988)). In a optimal stopping problem, Denardo and Roth-blumDenardo & Rothblum (1979) consider an optimal stopping problem with an exponential utility function as a criterion function in finite Markov decision chain and use a linear programming to compute an optimal policy. In Kadota et al.Kadota et al. (1996), they investigate an optimal stopping problem with a general utility function in a denumerable Markov chain. They give a sufficient condition for an one-step look ahead (OLA) stopping time to be optimal and characterize a property of an OLA stopping time for risk-averse and risk-seeking utilities. BojdeckiBojdecki (1979) formulates an optimal stopping problem which is concerned with maximizing the probability of a certain event and give necessary and sufficient conditions for existence of an optimal stopping time. He also applies the results to a version of the discrete-time disorder problem. OhtsuboOhtsubo (2003) considers optimal stopping problems with a threshold probability criterion in a Markov process, characterizes optimal values and finds optimal stopping times for finite and infinite horizon cases, and he in Ohtsubo (2003) also investigates optimal stopping problem with analogous objective for discrete time stochastic process and these are applied to a secretary problem, a parking problem and job search problems.

On the other hand, many authors propose a variety of criteria and investigate Markov decision processes for their criteria, instead of standard criteria, that is, the expected discounted total reward and the average expected reward per unit (see WhiteWhite (1988) for survey). Especially, WhiteWhite (1993), Wu and LinWu & Lin (1999), Ohtsubo and ToyonagaOhtsubo & Toyonaga (2002) and OhtsuboOhtsubo (2004) consider a problem in which we minimize a threshold probability. Such a problem is called risk minimizing problem and is available for applications to the percentile of the losses or Value-at-Risk (VaR) in finance (e.g. FilarFilar et al. (1995) and UryasevUryasev (2000)).

In this paper we consider Dynkin's game with a threshold probability in a random sequence. In Section 3 we characterize a value function of game and optimal stopping times for both players and show that the value function of game has properties of a distribution function with respect to a threshold value except a right continuity. In Section 4 we investigate an independent model, as applications of our game, and we explicitly find a value function which is right continuous and optimal stopping times for both players.

2. Formulation of problem

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(\mathcal{F}_n)_{n \in N}$ an increasing family of sub-σ-fields of $\mathcal{F}$, where $N = \{0, 1, 2, \cdots\}$ is a discrete time space. Let $X = (X_n)_{n \in N}, Y = (Y_n)_{n \in N}, W = (W_n)_{n \in N}$ be sequences of random variables defined on $(\Omega, \mathcal{F}, P)$ and adapted to $(\mathcal{F}_n)$ such that $X_n \leq W_n \leq Y_n$ almost surely (a.s.) for all $n \in N$ and $P(\sup_n X_n^+ + \sup_n Y_n^- < \infty) = 1$, where $x^+ = \max(0, x)$ and $x^- = (-x)^+$. The second assumption holds if random variables $\sup_n X_n^+$ and $\sup_n Y_n^-$ are integrable, which are standard conditions given in the classical Dynkin's game. Also let Z be an arbitrary integrable random variable on $(\Omega, \mathcal{F}, P)$. For each $n \in N$, we denote by Γ_n the class of $(\mathcal{F}_n)$–stopping times τ such that $\tau \geq n$ a. s..

We consider the following zero-sum stopping game. There are two players and the first and the second players choose stopping times τ and σ in Γ_0, respectively. Then the reward paid to the first player from the second is equal to

$$g(\tau, \sigma) = X_\tau I_{(\tau < \sigma)} + Y_\sigma I_{(\sigma < \tau)} + W_\tau I_{(\tau = \sigma < \infty)} + Z I_{(\tau = \sigma = \infty)},$$

where I_A is the indicator function of a set A in $\mathcal{F}$. In the classical Dynkin's game the aim of the first player is to maximize the expected gain $E[g(\tau, \sigma)]$ with respect to $\tau \in \Gamma_0$ and that of the second is to minimize this expectation with respect to $\sigma \in \Gamma_0$. In our problem the objective of the first player is to minimize the threshold probability $P[g(\tau, \sigma) \leq r]$ with respect to $\tau \in \Gamma_0$ and the second maximizes the probability with respect to $\sigma \in \Gamma_0$ for a given threshold value r.

We can define processes of minimax and maxmin values corresponding to our problem by

$$\overline{V}_n(r) = \operatorname*{ess\,inf}_{\tau \in \Gamma_n} \operatorname*{ess\,sup}_{\sigma \in \Gamma_n} P[g(\tau, \sigma) \leq r | \mathcal{F}_n],$$

$$\underline{V}_n(r) = \operatorname*{ess\,sup}_{\sigma \in \Gamma_n} \operatorname*{ess\,inf}_{\tau \in \Gamma_n} P[g(\tau, \sigma) \leq r | \mathcal{F}_n],$$

respectively, where $P[g(\tau, \sigma) \leq r | \mathcal{F}_n]$ is a conditional probability of an event $\{g(\tau, \sigma) \leq r\}$ given $\mathcal{F}_n$. See NeveuNeveu (1975) for the definition of ess sup and ess inf. We also define sequences of minimax and maxmin values by

$$\overline{v}_n(r) = \inf_{\tau \in \Gamma_n} \sup_{\sigma \in \Gamma_n} P[g(\tau, \sigma) \leq r], \quad \underline{v}_n(r) = \sup_{\sigma \in \Gamma_n} \inf_{\tau \in \Gamma_n} P[g(\tau, \sigma) \leq r],$$

respectively. For $n \geq 1$ and $\varepsilon \geq 0$, we say that a pair of stopping times $(\tau_\varepsilon, \sigma_\varepsilon)$ in $\Gamma_n \times \Gamma_n$ is ε–saddle point at (n, r) if

$$P[g(\tau_\varepsilon, \sigma) \leq r] - \varepsilon \leq v_n(r) \leq P[g(\tau, \sigma_\varepsilon) \leq r] + \varepsilon$$

for any $\tau \in \Gamma_n$ and any $\sigma \in \Gamma_n$, when $\bar{v}_n(r) = \underline{v}_n(r)$, say $v_n(r)$. .

3. General results

In this section we give fundamental properties of the value function of the game and find a saddle point.

We notice that $P[g(\tau, \sigma) \leq r] = E[I_{(g(\tau,\sigma) \leq r)}]$ and we easily see that

$$I_{(g(\tau,\sigma) \leq r)} = \widetilde{X}_\tau(r) I_{(\tau < \sigma)} + \widetilde{Y}_\sigma(r) I_{(\sigma < \tau)} + \widetilde{W}_\tau(r) I_{(\tau = \sigma < \infty)} + \widetilde{Z}(r) I_{(\tau = \sigma = \infty)},$$

where new sequences $(\widetilde{X}_n(r)), (\widetilde{Y}_n(r)), (\widetilde{W}_n(r))$ and random variable $\widetilde{Z}(r)$ are defined by

$$\widetilde{X}_n(r) = I_{(X_n \leq r)}, \ \widetilde{Y}_n(r) = I_{(Y_n \leq r)}, \ \widetilde{W}_n(r) = I_{(W_n \leq r)}, \ \widetilde{Z}(r) = I_{(Z \leq r)}.$$

Since $X_n \leq W_n \leq Y_n$, we see that $\widetilde{Y}_n(r) \leq \widetilde{W}_n(r) \leq \widetilde{X}_n(r)$ for all r. Thus our problem is just a special version of the classical Dynkin's game for a *fixed* threshold value r.

We first have three propositions below for a fixed r from the result of Dynkin's game (e.g. see NeveuNeveu (1975) and OhtsuboOhtsubo (2000)). In the following proposition, the notation $\text{mid}(a, b, c)$ denotes the middle value among constants a, b and c. For example, when $a < b < c$ then $\text{mid}(a, b, c) = b$. If $a < b$, $\text{mid}(a, b, c) = \max(a, \min(b, c)) = \min(b, \max(a, c))$.

Proposition 3.1. *Let r be arbitrary.*

(a) *For each $n \in N$, $\overline{V}_n(r) = \underline{V}_n(r)$, say $V_n(r)$, and $\bar{v}_n(r) = \underline{v}_n(r) = E[V_n(r)]$, say $v_n(r)$.*

(b) *$(V_n(r))$ is the unique sequence of random variables satisfying the equalities*

$$V_n = \text{mid}(\widetilde{X}_n(r), \widetilde{Y}_n(r), E[V_{n+1} | \mathcal{F}_n]), \ n \in N$$

and the inequalities

$$\widehat{X}_n(r) \leq V_n \leq \widehat{Y}_n(r), \ n \in N,$$

where $(\widehat{X}_n(r))$ is the largest submartingale dominated by $\min(\widetilde{X}_n(r), E[\widetilde{Z}(r) | \mathcal{F}_n])$ and $(\widehat{Y}_n(r))$ is the smallest supermartingale dominating $\max(\widetilde{Y}_n(r), E[\widetilde{Z}(r) | \mathcal{F}_n])$, that is,

$$\widehat{X}_n(r) = \operatorname*{ess\,inf}_{\tau \in \Gamma_n} \ P[g(\tau, \infty) \leq r) | \mathcal{F}_n], \ \widehat{Y}_n(r) = \operatorname*{ess\,sup}_{\sigma \in \Gamma_n} \ P[g(\infty, \sigma) \leq r | \mathcal{F}_n].$$

(c) *For $\varepsilon > 0$, let*

$$\tau_n^\varepsilon(r) = \inf\{k \geq n | V_k(r) \geq \widetilde{X}_k(r) - \varepsilon\},$$
$$\sigma_n^\varepsilon(r) = \inf\{k \geq n | V_k(r) \leq \widetilde{Y}_k(r) + \varepsilon\}$$

Then $(\tau_n^\varepsilon(r), \sigma_n^\varepsilon(r))$ is ε-saddle point at (n, r).

For the value process $\widehat{X}_n(r)$ for the first player, we can obtain it as the following: for $k \geq n$, let

$$\gamma_k^k(r) = \min\left(\widetilde{X}_k(r), E[\widetilde{Z}(r) | \mathcal{F}_k]\right),$$
$$\gamma_n^k(r) = \max\left(\widetilde{X}_n(r), E[\gamma_{n+1}^k(r) | \mathcal{F}_n]\right), \ n < k.$$

Proposition 3.2. *Let r be arbitrary. For each $k, n : k \geq n$, $\gamma_n^k(r) \geq \gamma_n^{k+1}(r)$ and for each $n \in N$,* $\lim_{k \to \infty} \gamma_n^k(r) = \widehat{X}_n(r)$.
For $k \geq n$, let

$$\beta_k^k(r) = \widehat{X}_k(r),$$
$$\beta_n^k(r) = \mathrm{mid}(\widetilde{X}_n(r), \widetilde{Y}_n(r), E[\beta_{n+1}^k(r)|\mathcal{F}_n]), \ n < k,$$

Proposition 3.3. *Let r be arbitrary. For each $k \geq n$, $\beta_n^k(r) \leq \beta_n^{k+1}$ and for each n, $\lim_{k \to \infty} \beta_n^k(r) = V_n(r)$.*

Theorem 3.1. *For each n, $V_n(\cdot)$ has properties of a distribution function on R except for the right continuity.*

Proof. We first notice that $\widetilde{Z}(r) = I_{(Z \leq r)}$ is a nondecreasing function in r. From the definition of a conditional expectation and the dominated convergence theorem, $E[\widetilde{Z}(r)|\mathcal{F}_k]$ for each k is also nondecreasing at r. Since $\widetilde{X}_k(r) = I_{(X_k \leq r)}$ is nondecreasing at r for each $k \in \mathbf{N}$, we see that $\gamma_n^k(r) = \min(\widetilde{X}_k(r), E[\widetilde{Z}(r)|\mathcal{F}_k])$ is a nondecreasing function in r. By induction, $\gamma_n^k(r)$ is nondecreasing in r for each $k \geq n$. Since a sequence $\{\gamma_n^k(r)\}_{k=n}^\infty$ of functions is nonincreasing and $\widehat{X}_n(r) = \lim_{k \to \infty} \gamma_n^k(r)$, it follows that $\beta_n^n(r) = \widehat{X}_n(r)$ is nondecreasing for each n. Similarly, it follows by induction that $\beta_n^k(r)$ is nondecreasing at r for each $n \leq k$, since $\widetilde{Y}_n(r)$ is nondecreasing at r. From Proposition 2.3, the monotonicity of a sequence $\{\beta_n^k(r)\}_{k=n}^\infty$ implies that $V_n(r) = \lim_{k \to \infty} \beta_n^k(r)$ is a nondecreasing function in r.

Next, since we have $V_n(r) \leq \widetilde{X}_n(r)$ and we see that $\widetilde{X}_n(r) = I_{(X_n \leq r)} = 0$ for a sufficiently small r, it follows that $\lim_{r \to -\infty} V_n(r) = 0$. Similarly, we see that $\lim_{r \to \infty} V_n(r) = 1$, since we have $V_n(r) \geq \widetilde{Y}_n(r)$ and we see that $\widetilde{Y}_n(r) = 1$ for a sufficiently large r. Thus this theorem is completely proved.

We give an example below in which the value function $V_n(r)$ is not right continuous at some r.

Example 3.1. Let $X_n = W_n = -1, Y_n = 1/n$ for each n and let $Z = 1$. We shall obtain the value function $V_n(r)$ by Propositions 3.2 and 3.3. Since $\widetilde{X}_k(r) = I_{[-1, \infty)}(r)$ and $\widetilde{Z}(r) = I_{[1, \infty)}(r)$, we have $\gamma_k^k(r) = I_{[1, \infty)}(r)$. By induction, we easily see that $\gamma_n^k(r) = I_{[1, \infty)}(r)$ for each $k \geq n$ and hence $\beta_n^n(r) = \widehat{X}_n(r) = \lim_{k \to \infty} \gamma_n^k = I_{[1, \infty)}(r)$. Next, since $\widetilde{Y}_{k-1}(r) = I_{[1/(k-1), \infty)}(r)$, we have $\beta_{k-1}^k(r) = I_{[1/(k-1), \infty)}(r)$. By induction, we see that $\beta_n^k(r) = I_{[1/(k-1), \infty)}(r)$ for each $k > n$. Thus we have $V_n(r) = \lim_{k \to \infty} \beta_n^k(r) = I_{(0, \infty)}(r)$, which yields that $V_n(r)$ is not right continuous at $r = 0$.

4. Independent model

We shall consider an independent sequences as a special model. Let $(W_n)_{n \in N}$ be a sequence of independent distributed random variables with $P(\sup_n |W_n| < \infty) = 1$, and let Z be a random variable which is independent of $(W_n)_{n \in N}$. For each $n \in N$ let $\mathcal{F}_n$ be the σ-field generated by $\{W_k; k \leq n\}$. Also, for each $n \in N$, let $X_n = W_n - c$ and $Y_n = W_n + d$, where c and d are positive constants.

Since $\mathcal{F}_n$ is independent of $\{W_k; k > n\}$, the relation in Proposition 3.1 (b) is represented as follows:

$$
\begin{aligned}
V_n(r) &= \mathrm{mid}(\widetilde{X}_n(r), \widetilde{Y}_n(r), E[V_{n+1}(r)]) \\
&= \mathrm{mid}(I_{(W_n \leq r+c)}, I_{(W_n \leq r-d)}, E[V_{n+1}(r)]).
\end{aligned}
$$

From Proposition 3.1 (b) and argument analogous to classical optimal stopping problem, we have also

$$\widehat{X}_n(r) = \min(\widetilde{X}_n(r), E[\widetilde{Z}(r)|\mathcal{F}_n], E[\widehat{X}_{n+1}(r)|\mathcal{F}_n]),$$
$$\widehat{Y}_n(r) = \max(\widetilde{Y}_n(r), E[\widetilde{Z}(r)|\mathcal{F}_n], E[\widehat{Y}_{n+1}(r)|\mathcal{F}_n]).$$

Hence we obtain

$$\widehat{X}_n(r) = \min(\widetilde{X}_n(r), P(Z \leq r), E[\widehat{X}_{n+1}(r)]),$$
$$\widehat{Y}_n(r) = \max(\widetilde{Y}_n(r), P(Z \leq r), E[\widehat{Y}_{n+1}(r)]),$$

since $E[\widetilde{Z}(r)|\mathcal{F}_n] = E[\widetilde{Z}(r)] = P(Z \leq r)$.

Example 4.1. Let W be a uniformly distributed random variable on an interval $[0,1]$ and assume that W_n has the same distribution as W for all $n \in N$ and that $0 < c, d < 1/2$. Then since $(W_n)_{n\in N}$ is a sequence of independently and identically distributed random variables, $V_n(r)$ does not depend on n. Hence, letting $V(r) = V_n(r), n \in N$ and $v(r) = E[V(r)]$, we have

$$V(r) = \text{mid}(I_{(W\leq r+c)}, I_{(W\leq r-d)}, v(r)).$$

When $W < r - d$, we have $I_{(W\leq r+c)} = I_{(W\leq r-d)} = 1$, so $V(r) = 1$. When $W \geq r + c$, we have $V(r) = 0$, since $I_{(W\leq r+c)} = I_{(W\leq r-d)} = 0$. Thus we obtain

$$V(r) = I_{(W\leq r-d)} + v(r)I_{(r-d\leq W<r+c)}.$$

Taking the expectation on the both sides, we see that

$$v(r) = P(W \leq r - d) + v(r)P(r - d \leq W < r + c).$$

If $r < d$ then we have $v(r) = v(r)P(0 \leq W < r + c)$. Since $r < d < 1/2 < 1 - c$, $P(0 \leq W < r + c) < 1$ and hence $v(r) = 0$. If $d \leq r < 1 - c$, then we obtain $v(r) = (r - d)/(1 - c - d)$, since $P(W \leq r - d) = r - d$ and $P(r - d \leq W < r + c) = c + d$. Similarly, if $r \geq 1 - c$ then we have $v(r) = 1$. Thus it follows that

$$v(r) = I_{[1-c,\infty)}(r) + (r - d)/(1 - c - d)I_{[d,1-c)}(r).$$

We completely obtained the values $V(r)$ and $v(r)$. By the way we easily see that $\widehat{X}(r) = \widehat{X}_n(r) = E[\widehat{X}(r)]I_{(W\leq r+c)}$, where

$$E[\widehat{X}(r)] = rI_{[1-c,1)}(r) + I_{[1,\infty)}(r),$$

and

$$E[\widehat{Y}(r)] = \widehat{Y}(r) = \widehat{Y}_n(r) = P[Z \leq r]I_{(-\infty,d)}(r) + I_{[d,\infty)}(r).$$

Now $v(r)$ is a distribution function in r. Let U is a random variable corresponding to $v(r)$. Then we see that $E[U] = (1 - c + d))/2$.

We shall next compare our model with the classical Dynkin's game in this example. Let

$$\bar{J}_n = \operatorname*{ess\,inf}_{\tau\in\Gamma_n} \operatorname*{ess\,sup}_{\sigma\in\Gamma_n} E[g(\tau,\sigma)|\mathcal{F}_n],$$
$$\underline{J}_n = \operatorname*{ess\,sup}_{\sigma\in\Gamma_n} \operatorname*{ess\,inf}_{\tau\in\Gamma_n} E[g(\tau,\sigma)|\mathcal{F}_n],$$

be minimax and maxmin value processes, respectively. Then we have $\bar{J}_n = \underline{J}_n = J$, say, since $\bar{J}_n = \underline{J}_n$ does not depend upon n in this example. Also, by solving the relation

$$J = \text{mid}(W - c, W + d, E[J]),$$

we have $E[J] = (1 - c + d)/2$, which coincide with $E[U]$. However, the distribution function of J is represented by

$$P(J \le x) = (x - d)I_{[d,(1-c+d)/2)}(x) + (x + c)I_{[(1-c+d)/2,1-c)}(x) + I_{[1-c,\infty)}(x),$$

which is different from that of U, that is, $v(r)$.

5. Acknowledgments

6. References

Bojdecki, T. (1979). Probability maximizing approach to optimal stopping and its application to a disorder problem. *Stochastics*, Vol.3, 61–71.

Chow, Y. S.; Robbins, H. & Siegmund, D. (1971). *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin, Boston.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw Hill, New York.

Denardo, E. V. & Rothblum, U. G. (1979). Optimal stopping, exponential utility, and linear programming. *Math. Programming*, Vol.16, 228–244.

Dynkin, E. B. (1969). Game variant of a problem on optimal stopping. *Soviet Math. Dokl.*, Vol.10, 270–274.

Filar, J. A., Krass, D. & Ross, K. W. (1995). Percentile performance criteria for limiting average Markov decision processes. *IEEE Trans. Automat. Control*, Vol.40, 2-10.

Kadota, Y., Kurano, M. & Yasuda, M. (1996). Utility-optimal stopping in a denumerable Markov chain. *Bull. Informatics and Cybernetics*, Vol.28, 15-21.

Neveu, J. (1975). *Discrete-Parameter Martingales*. North-Holland, New York.

Ohtsubo, Y. (2000). The values in Dynkin stopping problem with th some constraints. *Mathematica Japonica*, Vol.51, 75-81.

Ohtsubo, Y. & Toyonaga, K. (2002). Optimal policy for minimizing risk models in Markov decision processes. *J. Math. Anal. Appl.*, Vol.271, 66-81.

Ohtsubo, Y. (2003). Value iteration methods in risk minimizing stopping problem. *J. Comput. Appl. Math.*, Vol.152, 427-439.

Ohtsubo, Y. (2003). Risk minimization in optimal stopping problem and applications. *J. Operations Research Society of Japan*, Vol.46, 342-352.

Ohtsubo, Y. (2004). Optimal threshold probability in undiscounted Markov decision processes with a target set. *Applied Math. Computation*, Vol.149, 519-532.

Shiryayev, A. N. (1978). *Optimal Stopping Rules*. Springer, New York.

Uryasev, S. P. (2000). Introduction to theory of probabilistic functions and percentiles (Value-at-Risk). *Probabilistic Constrained Optimization*. Uryasev, S. P., (Ed.), Kluwer Academic Publishers, Dordrecht, pp.1-25.

White, D. J. (1988). Mean, variance and probabilistic criteria in finite Markov decision processes: a review. *J. Optim. Theory Appl.*, Vol.56, 1-29.

White, D. J. (1993). Minimising a threshold probability in discounted Markov decision processes. *J. Math. Anal. Appl.*, Vol.173, 634-646.

Wu, C. & Lin, Y. (1999). Minimizing risk models in Markov decision processes with policies depending on target values. *J. Math. Anal. Appl.* Vol.231, 47-67.

Stochastic independence with respect to upper and lower conditional probabilities defined by Hausdorff outer and inner measures

Serena Doria
University G.d'Annunzio
Italy

1. Introduction

A new model of coherent upper conditional prevision is proposed in a metric space. It is defined by the Choquet integral with respect to the *s*-dimensional Hausdorff outer measure if the conditioning event has positive and finite Hausdorff outer measure in its dimension *s*. Otherwise if the conditioning event has Hausdorff outer measure in its dimension equal to zero or infinity it is defined by a 0-1 valued finitely, but not countably, additive probability.

If the conditioning event has positive and finite Hausdorff outer measure in its dimension the coherent upper conditional prevision is proven to be monotone, comonotonically additive, submodular and continuous from below.

Given a coherent upper conditional prevision the coherent lower conditional prevision is defined as its conjugate.

In Doria (2007) coherent upper and lower conditional probablities are obtained when only 0-1 valued random variables are considered.

The aim of this chapter is to introduce a new definition of stochastic independence with respect to coherent upper and lower conditional probabilities defined by Hausdorff outer and inner measures.

A concept related to the definition of conditional probability is stochastic independence. In a continuous probability space where probability is usually assumed equal to the Lebesgue measure, we have that finite, countable and fractal sets (i.e. the sets with non-integer Hausdorff dimension) have probability equal to zero. For these sets the standard definition of independence given by the factorization property is always satisfied since both members of the equality are zero.

The notion of s-independence with respect to Hausdorff outer and inner measures is introduced to check probabilistic dependence for sets with probability equal to zero, which are always independent according to the standard definition given by the factorization property. Moreover s-independence is compared with the notion of epistemic independence with respect to upper and lower conditional probabilities (Walley, 1991).

The outline of the chapter is the following.

In Section 2 it is proven that a conditional prevision defined by the Radon-Nikodym derivative may be not coherent and examples are given.

In Section 3 coherent upper conditional previsions are defined in a metric space by the Choquet integral with respect to Hausdorff outer measure if the conditioning event has positive and finite Hausdorff outer measure in its dimension. Otherwise they are defined by a 0-1 valued finitely, but not countably, additive probability. Their properties are proven.

In Section 4 the notion of s-irrelevance and s-independence with respect to coherent upper and lower conditional probabilities defined by Hausdorff outer and inner measures are introduced. It is proven that the notions of epistemic irrelevance and s-irrelevance are not always related. In particular we give conditions for which an event B is epistemically irrelevant to an event A, but it is not s-irrelevant. In the Euclidean metric space it is proven that a necessary condition for s-irrelevance between events is that the Hausdorff dimension of the two events and their intersection is equal to the Hausdorff dimension of Ω. Finally sufficient conditions for s-irrelevance between Souslin subsets of $\Re^n$ are given.

In Section 5 some fractal sets are proven to be s-dependent since they do not satisfy the necessary condition for s-independence. In particular the attractor of a finite family of similitudes and its boundary are proven to be s-dependent if the open set condition holds. Moreover a condition for which two middle Cantor sets are s-dependent is given.

It is important to note that all these sets are stochastically independent according the axiomatic definition given by the factorization property if probability is defined by the Lebesgue measure.

In Section 6 curves filling the space, such as Peano curve and Hilbert curve are proven to be s-independent.

2. Conditional expectation and coherent conditional prevision

Partial knowledge is a natural interpretation of conditional probability. This interpretation can be formalized in a different way in the axiomatic approach and in the subjective approach where conditional probability is respectively defined by the Radon-Nikodym derivative or by the axioms of coherence. In both cases conditional probability is obtained as the restriction of conditional expectation or conditional prevision to the class of indicator functions of events.

Some critical situations, which highlight as the axiomatic definition of conditional probability is not always a useful tool to represent partial knoweledge, are proposed in literature and analyzed in this section. In particular the role of the Radon-Nikodym derivative in the assessment of a coherent conditional prevision is investigated.

It is proven that, every time that the σ-field of the conditioning events is properly contained in the σ-field of the probability space and it contains all singletons, the Radon-Nikodym derivative cannot be used as a tool to define coherent conditional previsions. This is due to the fact that one of the defining properties of the Radon-Nikodym derivative, that is to be measurable with respect to the σ-field of the conditioning events, contradicts a necessary condition for the coherence.

Analysis done points out the necessity to introduce a different tool to define coherent conditional previsions.

2.1 Conditional expectation and Radon-Nikodym derivative

In the axiomatic approach Billingsley (1986) conditional expectation is defined with respect to a σ-field $\mathbf{G}$ of conditioning events by the Radon-Nikodym derivative. Let $(\Omega, \mathbf{F}, P)$ be a probability space and let $\mathbf{F}$ and $\mathbf{G}$ be two σ-fields of subsets of Ω with $\mathbf{G}$ contained in $\mathbf{F}$ and let X be an integrable random variable on $(\Omega, \mathbf{F}, P)$. Let P be a probability measure on $\mathbf{F}$; define a measure ν on $\mathbf{G}$ by $\nu(G) = \int_G X dP$. This measure is finite and absolutely continuous with

respect to P. So there exists a function, the Radon-Nikodym derivative denoted by $E[X|\mathbf{G}]$, defined on Ω, $\mathbf{G}$-measurable, integrable and satisfying the functional equation

$$\int_G E[X|\mathbf{G}]dP = \int_G XdP \text{ with } G \text{ in } \mathbf{G}.$$

This function is unique up to a set of P-measure zero and it is a version of the conditional expected value.

If X is the indicator function of any event A belonging to $\mathbf{F}$ then $E[X|\mathbf{G}] = E[A|\mathbf{G}] = P[A|\mathbf{G}]$ is a version of the conditional probability.

Conditional probability can be used to represent partial information (Billingsley, 1986, Section 33).

A probability space $(\Omega, \mathbf{F}, P)$ can be use to represent a random phenomenon or an experiment whose outcome is drawn from according to the probability given by P. Partial information about the experiment can be represented by a sub σ-field $\mathbf{G}$ of $\mathbf{F}$ in the following way: an observer does not know which ω has been drawn but he knows for each $H \in \mathbf{G}$, if ω belongs to H or if ω belongs to H^c. A sub σ-field $\mathbf{G}$ of $\mathbf{F}$ can be identified as partial information about the random experiment, and, fixed A in $\mathbf{F}$, conditional probability can be used to represent partial knowledge about A given the information on $\mathbf{G}$. If conditional probability is defined by the Radon-Nykodim derivative, denoted by $P[A|\mathbf{G}]$, by the standard definition (Billingsley, 1986, p.52) we have that an event A is independent from the σ-field $\mathbf{G}$ if it is independent from each $H \in \mathbf{G}$, that is $P[A|\mathbf{G}] = P(A)$ with probability 1. In (Billingsley, 1986, Example 33.11) it is shown that the interpretation of conditional probability in terms of partial knowledge breaks down in certain cases. Let $\Omega = [0,1]$, let $\mathbf{F}$ be the Borel σ-field of $[0,1]$ and let P be the Lebesgue measure on $\mathbf{F}$. Let $\mathbf{G}$ be the sub σ-field of sets that are either countable or co-countable. Then $P(A)$ is a version of the conditional probability $P[A|\mathbf{G}]$ define by the Radon-Nikodym derivative because $P(G)$ is either 0 or 1 for evey $G \in \mathbf{G}$. So an event A is independent from the information represented by $\mathbf{G}$ and this is a contradiction according to the fact that the information represented by $\mathbf{G}$ is complete since $\mathbf{G}$ contains all the singletons of Ω.

2.2 Coherent upper conditional previsions

In the subjective probabilistic approach (de Finetti 1970, Dubins 1975 and Walley 1991) coherent upper conditional previsions $\overline{P}(\cdot|B)$ are functionals, defined on a linear space of bounded random variables, satisfying the axioms of coherence.

In Walley (1991) coherent upper conditional previsions are defined when the conditioning events are sets of a partition.

Definition 1. *Let Ω be a non-empty set let $\mathbf{B}$ be a partition of Ω. For every $B \in \mathbf{B}$ let $K(B)$ be a linear space of bounded random variables defined on B. Then separately coherent upper conditional previsions are functionals $\overline{P}(\cdot|B)$ defined on $K(B)$, such that the following conditions hold for every X and Y in $K(B)$ and every strictly positive constant λ:*

- *1) $\overline{P}(X|B) \leq \sup(X|B)$;*
- *2) $\overline{P}(\lambda X|B) = \lambda \overline{P}(X|B)$ (positive homogeneity);*
- *3) $\overline{P}(X + Y)|B) \leq \overline{P}(X|B) + \overline{P}(Y|B)$;*
- *4) $\overline{P}(B|B) = 1$.*

Coherent conditional upper previsions can always be extended to coherent upper previsions on the class $\mathbf{L}(B)$ of all bounded random variables defined on B.

Suppose that $\overline{P}(X|B)$ is a coherent upper conditional prevision on $\mathbf{K}$ then its conjugate coherent lower conditional prevision is defined by $\underline{P}(-X|B) = -\overline{P}(X|B)$. If for every X belonging to $\mathbf{K}$ we have $P(X|B) = \underline{P}(X|B) = \overline{P}(X|B)$ then $P(X|B)$ is called a coherent *linear* conditional prevision de Finetti (1970) and it is a linear positive functional on $\mathbf{K}$.

Definition 2. *Let Ω be a non-empty set let $\mathbf{B}$ be a partition of Ω. For every $B \in \mathbf{B}$ let $\mathbf{K}(B)$ be a linear space of bounded random variables defined on B. Then linear coherent conditional previsions are functionals $P(\cdot|B)$ defined on $\mathbf{K}(B)$, such that the following conditions hold for every X and Y in $\mathbf{K}(B)$ and every strictly positive constant λ:*

- *1) if $X > 0$ then $P(X|B) \geq 0$ (positivity);*
- *2) $P(\lambda X|B) = \lambda P(X|B)$ (positive homogeneity);*
- *3) $P(X+Y)|B) = P(X|B) + P(Y|B)$ (linearity);*
- *4) $P(B|B) = 1$.*

Upper conditional probabilities are obtained when only 0-1 valued random variables are considered;

In Dubins (1975) coherent conditional probabilities are defined when the family of the conditioning events is a field of subsets of Ω.

Definition 3. *Let Ω be a non-empty set and let $\mathbf{F}$ and $\mathbf{G}$ be two fields of subsets of Ω, with $\mathbf{G} \subseteq \mathbf{F}$. P is a finitely additive conditional probability on $(\mathbf{F}, \mathbf{G})$ if it is a real function defined on $\mathbf{F} \times \mathbf{G}^0$, where $\mathbf{G}^0 = \mathbf{G} - \oslash$ such that the following conditions hold:*

- *I) given any $H \in \mathbf{G}^0$ and $A_1, ..., A_n \in \mathbf{F}$ and $A_i \cap A_j = \oslash$ for $i \neq j$, the function $P(\cdot|H)$ defined on $\mathbf{F}$ is such that $P(A|H) \geq 0$, $P(\bigcup_{k=1}^n A_k|H) = \sum_{k=1}^n P(A_k|H)$, $P(\Omega|H) = 1$*
- *II) $P(H|H) = 1$ if $H \in \mathbf{G}^0$*
- *III) given $E \in \mathbf{F}$, $H \in \mathbf{F}$ with $A \in \mathbf{G}^0$ and $EA \in \mathbf{G}^0$ then $P(EH|A) = P(E|A)P(H|EA)$.*

From conditions I) and II) we have

II') $P(A|H) = 1$ if $A \in \mathbf{F}$, $H \in \mathbf{G}^0$ and $H \subset A$.

These conditional probabilities are coherent in the sense of de Finetti, since conditions I), II), III) are sufficient for the coherence of P on $\mathbf{C} = \mathbf{F} \times \mathbf{G}^0$ when $\mathbf{F}$ and $\mathbf{G}$ are fields of subsets of Ω with $\mathbf{G} \subseteq \mathbf{F}$ or $\mathbf{G}$ is an additive subclass of $\mathbf{F}$; otherwise if $\mathbf{F}$ and $\mathbf{G}$ are two arbitrary families of subsets of Ω, such that $\Omega \in \mathbf{F}$ the previous conditions are necessary for the coherence but not sufficient.

2.3 Coherent conditional previsions and the Radon-Nikodym derivative

In this subsection the role of the Radon-Nikodym derivative in the assessment of a coherent conditional prevision is analyzed.

The definitions of conditional expectation and coherent linear conditional prevision can be compared when the σ-field $\mathbf{G}$ is generated by the partition $\mathbf{B}$. Let $\mathbf{G}$ be equal or contained in the σ-field generated by a countable class $\mathbf{C}$ of subsets of $\mathbf{F}$ and let $\mathbf{B}$ be the partition generated by the class $\mathbf{C}$. Denote $\Omega' = \mathbf{B}$ and φ_B the function from Ω to Ω' that associates to every $\omega \in \Omega$ the atom B of the partition $\mathbf{B}$ that contains ω; then we have that $P(\mathbf{A}|\mathbf{G}) = P(\mathbf{A}|\mathbf{B}) \circ \varphi_B$ for every $A \in \mathbf{F}$ (Koch, 1997, 262).

The next theorem shows that every time that the σ-field **G** of the conditioning events is properly contained in **F** and it contains all singletons of $[0,1]$ then the conditional prevision, defined by the Radon-Nikodym derivative is not coherent. It occurs because one of the defining properties of conditional expectation that is to be measurable with respect to the σ-field of conditioning events contradicts a necessary condition for coherence of a linear conditional prevision. A bounded random variable is called **B**-measurable or measurable with respect to the partition **B** (Walley, 1991, p.291) if it is constant on the atoms B of the partition. If for every B belonging to **B** $P(X|B)$ are coherent linear conditional previsions and X is **B**-measurable then $P(X|B) = X$ (Walley, 1991, p.292). This necessary condition for coherence is not always satisfied if $P(X|B)$ is defined by the Radon-Nikodym derivative.

Theorem 1. *Let $\Omega = [0,1]$, let **F** be the Borel σ-field of $[0,1]$ and let P be the Lebesgue measure on **F**. Let **G** be a sub σ-field properly contained in **F** and containing all singletons of $[0,1]$. Let **B** be the partition of all singletons of $[0,1]$ and let X be the indicator function of an event A belonging to **F** - **G**. If we define the conditional prevision $P(X|\{\omega\})$ equal to the Radon-Nikodym derivative with probability 1, that is*

$$P(X|\{\omega\}) = E[X|G]$$

except on a subset N of $[0,1]$ of P-measure zero, then the conditional prevision $P(X|\{\omega\})$ is not coherent.

Proof. If the equality $P(X|\{\omega\}) = E[X|G]$ holds with probability 1, then we have that, with probability 1, the linear conditional prevision $P(X|\{\omega\})$ is different from X, the indicator function of A; in fact having fixed A in $F - G$ the indicator function X is not **G**-measurable, it does not verify a property of the Radon-Nikodym derivative and therefore it cannot be assumed as conditional expectation according to axiomatic definition. So the linear conditional prevision $P(X|\{\omega\})$ does not satisfy the necessary condition for being coherent, $P(X|\{\omega\}) = X$ for every singleton $\{\omega\}$ of **G**. ⋄

Example 1. (Billingsley, 1986, Example 33.11) Let $\Omega = [0,1]$, let **F** be the Borel σ-field of Ω, let P be the Lebesgue measure on **F** and let **G** be the sub σ-field of **F** of sets that are either countable or co-countable. Let **B** be the partition of all singletons of Ω; if the linear conditional prevision is defined equal, with probability 1, to conditional expectation defined by the Radon-Nikodym derivative, we have that

$$P(X|\mathbf{B}) = E[X|\mathbf{G}] = P(X).$$

So when X is the indicator function of an event $A = [a,b]$ with $0 < a < b < 1$ then $P(X|B) = P(A)$ and it does not satisfy the necessary condition for coherence that is $P(X|\{\omega\}) = X$ for every singleton $\{\omega\}$ of **G**.
Evident from Theorem 1 and Example 1 is the necessity to introduce a new tool to define coherent linear conditional previsions.

3. Coherent upper conditional previsions defined by Hausdorff outer measures

In this section coherent upper conditional previsions are defined by the Choquet integral with respect to Hausdorff outer measures if the conditioning event B has positive and finite Hausdorff outer measure in its dimension. Otherwise if the conditioning event B has Hausdorff outer measure in its dimension equal to zero or infinity they are defined by a 0-1 valued finitely, but not countably, additive probability.

3.1 Hausdorff outer measures

Given a non-empty set Ω an *outer measure* is a function $\mu^* : \wp(\Omega) \rightarrow [0, +\infty]$ such that $\mu^*(\oslash) = 0$, $\mu^*(A) \leq \mu^*(A')$ if $A \subseteq A'$ and $\mu^*(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu^*(A_i)$.

Examples of outer set functions or outer measures are the Hausdorff outer measures (Falconer 1986, Rogers 1998).

Let (Ω, d) be a metric space. A topology, called the *metric topology*, can be introduced into any metric space by defining the open sets of the space as the sets G with the property:

if x is a point of G, then for some $r > 0$ all points y with $d(x, y) < r$ also belong to G.

It is easy to verify that the open sets defined in this way satisfy the standard axioms of the system of open sets belonging to a topology (Rogers, 1998, p.26).

The diameter of a non empty set U of Ω is defined as $|U| = \sup \{d(x, y) : x, y \in U\}$ and if a subset A of Ω is such that $A \subset \bigcup_i U_i$ and $0 < |U_i| < \delta$ for each i, the class $\{U_i\}$ is called a δ-cover of A.

Let s be a non-negative number. For $\delta > 0$ we define $h_{s,\delta}(A) = \inf \sum_{i=1}^{+\infty} |U_i|^s$, where the infimum is over all δ-covers $\{U_i\}$.

The *Hausdorff s-dimensional outer measure* of A, denoted by $h^s(A)$, is defined as

$$h^s(A) = \lim_{\delta \to 0} h_{s,\delta}(A).$$

This limit exists, but may be infinite, since $h_{s,\delta}(A)$ increases as δ decreases because less δ-covers are available. The *Hausdorff dimension* of a set A, $dim_H(A)$, is defined as the unique value, such that

$$h^s(A) = +\infty \text{ if } 0 \leq s < dim_H(A),$$
$$h^s(A) = 0 \text{ if } dim_H(A) < s < +\infty.$$

We can observe that if $0 < h^s(A) < +\infty$ then $dim_H(A) = s$, but the converse is not true.

Denote by r the Hausdorff dimension of Ω, if an event A is such that $dim_H(A) = s < r$ then the Hausdorff dimension of the complementary set A^c is equal to r since the following relation holds:

$$dim_H(A \cup B) = \max \{dim_H(A), dim_H(B)\}.$$

Hausdorff outer measures are *metric* outer measures, that is
$h^s(E \cup F) = h^s(E) + h^s(F)$ whenever E and F are *positively separated*, i.e.
$d(E, F) = \inf \{d(x, y) : x \in E, y \in F\} > 0$.

A subset A of Ω is called *measurable* with respect to the outer measure h^s if it decomposes every subset of Ω additively, that is if $h^s(E) = h^s(A \cap E) + h^s(E - A)$ for all sets $E \subseteq \Omega$.

All Borel subsets of Ω are measurable with respect to a metric outer measure (Falconer, 1986, Theorem 1.5). So every Borel subset of Ω is measurable with respect to every Hausdorff outer measure h^s since Hausdorff outer measures are metric.

The restriction of h^s to the σ-field of h^s-measurable sets, containing the σ-field of the Borel sets, is called Hausdorff *s*-dimensional measure. The Borel σ-field is the σ-field generated by all open sets. The Borel sets include the closed sets (as complement of the open sets), the F_σ-sets (countable unions of closed sets) the G_σ-sets (countable intersections of open sets), etc. In particular the Hausdorff 0-dimensional measure is the counting measure and the Hausdorff 1-dimensional measure is the Lebesgue measure.

The Hausdorff s-dimensional measures are *modular* on the Borel σ-field, that is $h^s(A \cup B) + h^s(A \cap B) = h^s(A) + h^s(B)$ for every pair of Borelian sets A and B; so that (Denneberg,

1994, Proposition 2.4) the Hausdorff outer measures are *submodular* ($h^s(A \cup B) + h^s(A \cap B) \leq h^s(A) + h^s(B)$).

In (Rogers, 1998, p.50) and (Falconer, 1986, Theorem 1.6 (a)) it has been proven that if A is any subset of Ω there is a G_σ-set G containing A with $h^s(A) = h^s(G)$. In particular h^s is an *outer regular* measure.

Moreover Hausdorff outer measures are *continuous from below* (Falconer, 1986, Lemma 1.3), that is for any increasing sequences of sets $\{A_i\}$ we have $\lim_{i \to \infty} h^s(A_i) = h^s(\lim_{i \to \infty} A_i)$.

3.2 The Choquet integral

We recall the definition of the Choquet integral (Denneberg, 1994) with the aim to define upper conditional previsions by Choquet integral with respect to Hausdorff outer measures and to prove their properties. The Choquet integral is an integral with respect to a monotone set function. Given a non-empty set Ω and denoted by S a set system, containing the empty set and properly contained in $\wp(\Omega)$, the family of all subsets of Ω, a monotone set function μ: $S \to \overline{\Re}_+ = \Re_+ \cup \{+\infty\}$ is such that $\mu(\oslash) = 0$ and if $A, B \in S$ with $A \subseteq B$ then $\mu(A) \leq \mu(B)$. Given a monotone set function μ on S, its *outer set function* is the set function μ^* defined on the whole power set $\wp(\Omega)$ by

$$\mu^*(A) = \inf \{\mu(B) : B \supset A; B \in S\}, A \in \wp(\Omega)$$

The inner set function of μ is the set function μ_* defined on the whole power set $\wp(\Omega)$ by

$$\mu_*(A) = \sup \{\mu(B)|B \subset A; B \in S\}, A \in \wp(\Omega)$$

Let μ be a monotone set function defined on S properly contained in $\wp(\Omega)$ and $X : \Omega \to \overline{\Re} = \Re \cup \{-\infty, +\infty\}$ an arbitrary function on Ω. Then the set function

$$G_{\mu,X}(x) = \mu \{\omega \in \Omega : X(\omega) > x\}$$

is decreasing and it is called *decreasing distribution function* of X with respect to μ. If μ is continuous from below then $G_{\mu,X}(x)$ is right continuous. In particular the decreasing distribution function of X with respect to the Hausdorff outer measures is right continuous since these outer measures are continuous from below. A function $X : \Omega \to \overline{\Re}$ is called upper μ-measurable if $G_{\mu^*,X}(x) = G_{\mu_*,X}(x)$. Given an upper μ-measurable function $X : \Omega \to \overline{R}$ with decreasing distribution function $G_{\mu,X}(x)$, if $\mu(\Omega) < +\infty$, the *asymmetric Choquet integral* of X with respect to μ is defined by

$$\int X d\mu = \int_{-\infty}^0 (G_{\mu,X}(x) - \mu(\Omega))dx + \int_0^\infty G_{\mu,X}(x)dx$$

The integral is in $\Re$, can assume the values $-\infty$, $+\infty$ or is undefined when the right-hand side is $\infty - \infty$.

If $X \geq 0$ or $X \leq 0$ the integral always exists. In particular for $X \geq 0$ we obtain

$$\int X d\mu = \int_0^{+\infty} G_{\mu,X}(x)dx$$

If X is bounded and $\mu(\Omega) = 1$ we have that

- $\int X d\mu = \int_{\inf X}^0 (G_{\mu,X}(x) - 1)dx + \int_0^{\sup X} G_{\mu,X}(x)dx = \int_{\inf X}^{\sup X} G_{\mu,X}(x)dx + \inf X.$

3.3 A new model of coherent upper conditional prevision

A new model of coherent upper conditional prevision is introduced and its properties are proven.

Theorem 2. *Let* (Ω, d) *be a metric space and let B be a partition of Ω. For every $B \in B$ denote by s the Hausdorff dimension of the conditioning event B and by h^s the Hausdorff s-dimensional outer measure. Let $L(B)$ be the class of all bounded random variables on B. Moreover, let m be a 0-1 valued finitely additive, but not countably additive, probability on $\wp(B)$. Then for each $B \in B$ the functionals $\overline{P}(X|B)$ defined on $L(B)$ by*

$$\overline{P}(X|B) = \tfrac{1}{h^s(B)} \int_B X dh^s \text{ if } 0 < h^s(B) < +\infty$$

and by

$$\overline{P}(X|B) = m(XB) \text{ if } h^s(B) = 0, +\infty$$

are coherent upper conditional previsions.

Proof. Since $L(B)$ is a linear space we have to prove that, for every $B \in B$ $\overline{P}(X|B)$ satisfies conditions 1), 2), 3), 4) of Definition 1.

If B has finite and positive Hausdorff outer measure in its dimension s then $\overline{P}(X|B) = \tfrac{1}{h^s(B)} \int_B X dh^s$, so properties 1) and 2) are satisfied since they hold for the Choquet integral (Denneberg, 1994, Proposition 5.1). Property 3) follows from the Subadditivity Theorem (Denneberg, 1994, Theorem 6.3) since Hausdorff outer measures are monotone, submodular and continuous from below. Property 4) holds since $\overline{P}(B|B) = \tfrac{1}{h^s(B)} \int_B dh^s = 1$. If B has Hausdorff outer measure in its dimension equal to zero or infinity we have that the class of all coherent (upper) previsions on $L(B)$ is equivalent to the class of 0-1 valued additive probabilities defined on $\wp(B)$ then $\overline{P}(X|B) = m(XB)$. Then properties 1), 2), 3) are satisfied since m is a 0-1 valued finitely additive probability on $\wp(B)$. Moreover since a different m is choosen for each B we have that $\overline{P}(B|B) = m(B) = 1$. ⋄

The lower conditional previsions $\underline{P}(A|B)$ can be defined as the previous theorem if h_s denotes the Hausdorff s-dimensional inner measure. The unconditional upper prevision is obtained as a particular case when the conditioning event is Ω, that is $\overline{P}(A) = \overline{P}(A|\Omega)$ and $\underline{P}(A) = \underline{P}(A|\Omega)$.

A class of bounded random variables is called a *lattice* if it is closed under point-wise maximum $\vee$ and point-wise minimum $\wedge$.

In the following theorem it is proven that, if the conditioning event has positive and finite Hausdorff outer measure in its dimension s and $L(B)$ is a linear lattice of bounded random variables defined on B, necessary conditions for the functional $\overline{P}(X|B)$ to be represented as Choquet integral with respect to the upper conditional probability μ_B^*, i.e. $\overline{P}(X|B) = \tfrac{1}{h^s(B)} \int X dh^s$, are that $\overline{P}(X|B)$ is monotone, comonotonically additive, submodular and continuous from below.

Theorem 3. *Let* (Ω, d) *be a metric space and let B be a partition of Ω. For every $B \in B$ denote by s the Hausdorff dimension of the conditioning event B and by h^s the Hausdorff s-dimensional outer measure. Let $L(B)$ be a linear lattice of bounded random variables defined on B. If the conditioning event B has positive and finite Hausdorff s-dimensional outer measure in its dimension then the upper conditional prevision $\overline{P}(\cdot|B)$ defined on $L(B)$ as in Theorem 2 satisfies the following properties:*

- *i)* $X \leq Y$ *implies* $\overline{P}(X|B) \leq \overline{P}(Y|B)$ *(monotonicity);*

- *ii) if X and Y are comonotonic, i.e.* $(X(\omega_1) - X(\omega_2))(Y(\omega_1) - (Y(\omega_2)) \geq 0 \; \forall \omega_1, \omega_2 \in B$, *then* $\overline{P}(X + Y|B) = \overline{P}(X|B) + \overline{P}(Y|B)$ *(comonotonic additivity);*
- *iii)* $\overline{P}(X \vee Y|B) + \overline{P}(X \wedge Y|B) \leq \overline{P}(X|B) + \overline{P}(Y|B)$ *(submodularity);*
- *iv)* $\lim_{n \to \infty} \overline{P}(X_n|B) = \overline{P}(X|B)$ *if* X_n *is an increasing sequence of random variables converging to X (continuity from below).*

Proof. Since the conditioning event *B* has positive and finite Hausdorff outer measure in its dimension *s* then the functional $\overline{P}(\cdot|B)$ is defined on $\mathbf{L}(B)$ by the Choquet integral with respect to the upper conditional probability $\mu_B^*(A) = \frac{h^s(AB)}{h^s(B)}$; so conditions *i)* and *ii)* are satisfied because they are properties of the Choquet integral (Denneberg, 1994, Proposition 5.2).
Condition *iii)* is equivalent to require that the monotone set function that represents the functional $\overline{P}(\cdot|B)$ is submodular and it is satisfied since Hausdorff outer measures are submodular. Moreover every *s*-dimensional Hausdorff measure is continuous from below then from the Monotone Convergence Theorem (Denneberg, 1994, Theorem 8.1) we have that the functional $\overline{P}(\cdot|B)$ is continuous from below, that is condition iv). ⋄

Coherent upper conditional probabilities are obtained when only 0-1 valued random variables are considered;

Theorem 4. *Let* (Ω, d) *be a metric space and let* **B** *be a partition of* Ω. *For every* $B \in \mathbf{B}$ *denote by s the Hausdorff dimension of the conditioning event B and by* h^s *the Hausdorff s-dimensional outer measure. Let m be a 0-1 valued finitely additive, but not countably additive, probability on* $\wp(B)$. *Then, for each* $B \in \mathbf{B}$, *the functions defined on* $\wp(B)$ *by*

$$\overline{P}(A|B) = \frac{h^s(AB)}{h^s(B)} \text{ if } 0 < h^s(B) < +\infty$$

and by

$$\overline{P}(A|B) = m(AB) \text{ if } h^s(B) = 0, +\infty$$

are coherent upper conditional probabilities.

Coherent upper conditional probabilities can be defined in the general case where the family of the conditioning events is an additive class of events; they have been defined in Doria (2007):

Theorem 5. *Let* (Ω, d) *be a metric space, let* **F** *be the* σ-*field of all subsets of* Ω *and let* **G** *be an additive subclass of* **F**. *For every* $H \in \mathbf{G}^0 = \mathbf{G} - \oslash$ *and* $A \in \mathbf{F}$ *denote by s the Hausdorff dimension of the conditioning event H, by t the Hausdorff dimension of AH and by* h^s *the Hausdorff s-dimensional outer measure. Let m be a 0-1 valued finitely additive, but not countably additive, probability on* $\wp(H)$ *such that if* $0 < h^t(AH) < +\infty$ *then m(AH) = 0. Then, for each* $H \in \mathbf{G}^0$, *the functions defined on* $\wp(H)$ *by*

$$\overline{P}(A|H) = \frac{h^s(AH)}{h^s(H)} \text{ if } 0 < h^s(H) < +\infty$$

and by

$$\overline{P}(A|H) = m(AH) \text{ if } h^s(H) = 0, +\infty$$

are coherent upper conditional probabilities.

The lower conditional probability $\underline{P}(A|H)$ can be defined as in the previous theorem by the Hausdorff inner measures.

The new model of upper and lower conditional probabilities defined as in Theoren 5 can be used to assess coherent upper and lower conditional probabilities when the extensions of the conditional probability, defined in the axiomatic way, are not coherent.

Example 2. Let $\Omega = [0,1]$, let **F** be the class of all subsets of Ω and let **G** be the σ-field of countable and co-countable subsets of Ω. From Theorem 5 we have that a coherent upper conditional probability on $\mathbf{C} = \mathbf{F} \times \mathbf{G}^0$ can be defined by

$$\overline{P}(A|H) = \frac{h^1(AH)}{h^1(H)} \text{ if } H \text{ is co-countable}$$

$$\overline{P}(A|H) = \frac{h^0(AH)}{h^0(H)} \text{ if } H \text{ is finite}$$

$$\overline{P}(A|H) = m(AH) \text{ if } H \text{ is countable.}$$

4. s-Irrelevance and s-independence

In a recent paper (Doria, 2007) the new definitions of s-irrelevance and s-independence with respect to upper and lower conditional probabilities assigned by outer and inner Hausdorff measures have been proposed. They are based on the fact that epistemic independence and irrelevance, introduced by Walley, must be tested for events A and B, such that they and their intersection AB, have the same Hausdorff dimension. The concept of epistemic independence (Walley, 1991) is based on the notion of irrelevence; given two events A and B, we say that B is irrelevant to A when $\overline{P}(A|B) = \overline{P}(A|B^c) = \overline{P}(A)$ and $\underline{P}(A|B) = \underline{P}(A|B^c) = \underline{P}(A)$.

The events A and B are epistemically independent when B is irrelevant to A and A is irrelevant to B. As a consequence of this definition we can obtain that the factorization property $P(AB) = P(A)P(B)$, which constitutes the standard definition of independence for events, holds either for $P = \overline{P}$ and $P = \underline{P}$. In a continuous probabilistic space (Ω, F, P), where Ω is equal to $[0,1]^n$ and the probability is usually assumed equal to the Lebesgue measure on Ω, we have that the finite, countable and fractal sets (i.e. the sets with Hausdorff dimension non integer) have probability equal to zero. For these sets the standard definition of independence, given by the factorization property, is always satisfied since both members of the equality are zero. In Theorem 6 of this Section we prove that an event B is always irrelevant, according to the definition of Walley, to an event A if $dim_H(A) < dim_H(B) < dim_H(\Omega)$ and A and B have positive and finite Hausdorff outer measures in their dimensions; moreover if A and B are disjoint then they are epistemically independent. Nevertheless B is not s-irrelevant to A.

To avoid these problems the notions of s-irrelevance and s-independence with respect to upper and lower conditional probabilities assigned by a class of Hausdorff outer and inner measures are proposed to test independence. The definitions of s-independence and s-irrelevance are based on the fact that epistemic independence and irrelevance, must be tested for events A and B, such that they and their intersection AB, have the same Hausdorff dimension. According to this approach to independence, sets that represent events can be imagined divided in different layers; in each layer there are sets with the same Hausdorff dimension; two events A and B are s-independent if and only if the events A and B and their intersection AB belong to the same layer and they are epistemically independent.

Definition 4 Let (Ω, d) be a metric space. Denote by **F** the σ-field of all subsets of Ω and by $\mathbf{G}^0 = \mathbf{F} - \oslash$. Denoted by $\overline{P}$ and $\underline{P}$ the upper and lower conditional probabilities defined as in

Theorem 5 and given A and B in $\mathbf{G}^0$, then they are s-independent if the following conditions hold:

- 1s) $dim_H(AB) = dim_H(B) = dim_H(A)$
- 2s) $\overline{P}(A|B) = \overline{P}(A|B^c) = \overline{P}(A)$ and $\underline{P}(A|B) = \underline{P}(A|B^c) = \underline{P}(A)$;
- 3s) $\overline{P}(B|A) = \overline{P}(B|A^c) = \overline{P}(B)$ and $\underline{P}(B|A) = \underline{P}(B|A^c) = \underline{P}(B)$;

B is *s-irrelevant* to A if conditions 1s) and 2s) hold and A is s-irrelevant to B if conditions 1s) and 3s) hold.

Remark 1 Two disjoint events A and B are s-dependent since the Hausdorff dimension of the empty set cannot be equal to that one of any other set so condition 1s) is never satisfied.

Given the Euclidean metric space $([0,1]^n, d)$ in Doria (2007) it is proven that logical independence is a necessary condition for *s*-independence for events with Hausdorff dimension less then n.

Example 3 Let $\Omega = [0,1]$ let A be the Cantor set and let B be a finite subset of Ω such that intersection AB is equal to the empty set. We recall the definition of the Cantor set.

Let $E_0 = [0,1]$, $E_1 = [0,1/3] \cup [2/3,1]$, $E_2 = [0,1/9] \cup [2/9, 1/3] \cup [2/3,7/9] \cup [8/9,1]$, etc., where E_n is obtained by removing the open middle third of each interval in E_{n-1}, so E_n is the union of 2^n intervals, each of length $\frac{1}{3^n}$.

The Cantor's set is the perfect set $E = \bigcap_{n=0}^{\infty} E_n$. The Hausdorff dimension of the Cantor set is $s = \frac{ln2}{ln3}$ and $h^s(E) = 1$.

If $\overline{P}$ and $\underline{P}$ are the upper and lower conditional probabilities defined as in Theorem 5, then they satisfy the factorization property. Moreover B is irrelevant to A according to the definition given by Walley, but B is not s-irrelevant to A since condition 1s) of Definition 4 is not satisfied. The previous example shows that the notion of irrelevance and s-irrelevance are not related if Ω is an infinite set. The next theorem put in evidence this problem in a more general framework.

Theorem 6. *Let Ω be a non-empty set with positive and finite Hausdorff outer measure in its dimension and let $\overline{P}$ and $\underline{P}$ be the upper and lower conditional probabilities defined as in Theorem 5. If A and B are two subsets of Ω such that $dim_H(A) < dim_H(B) < dim_H(\Omega)$ and they have positive and finite Hausdorff outer measures in their dimensions then B is irrelevant to A, but B is not s-irrelevant to A.*

Proof. Denote by t, s and r respectively the Hausdoff dimension of AB, B and Ω; since $dim_H(A) < dim_H(B) < dim_H(\Omega)$ then we have that the Hausdorff dimension of B^c is equal to r and $t < s$. Moreover since A and B and their complements have positive and finite Hausdorff outer measures in their dimensions and upper conditional probability is defined as in Theorem 5 the condition $\overline{P}(A|B) = \overline{P}(A|B^c) = \overline{P}(A)$ becomes $\frac{h^s(AB)}{h^s(B)} = \frac{h^r(AB^c)}{h^r(B^c)} = \frac{h^r(A)}{h^r(\Omega)}$.

These equalities are satisfied since they vanish to $0 = 0 = 0$.

In the same way we can prove the equalities $\underline{P}(A|B) = \underline{P}(A|B^c) = \underline{P}(A)$.

Since $dim_H(A) < dim_H(B)$ then the event B is not s-irrelevant to A since condition 1s) of Definition 4 is not satisfied. $\diamond$

In the sequel the notions of s-irrelevance and s-independence are investigated in the Euclidean metric space.

Given a non-empty subset Ω of $\Re^n$ with Hausdorff dimension equal to n and positive and finite Hausdorff outer measure in its dimension and given two subsets A and B of Ω, we want to find conditions under which B is s-relevant to A. Condition 1s) of the Definition 4

is a necessary condition for s-irrelevance and s-independence. We focus the attention on this condition with the aim to investigate s-relevance.

In (Mattila, 1984, Theorem 6.13) an important result is proven for Souslin sets that is a class of sets, which are defined in terms of unions and intersections of closed sets.

In a metric space the Souslin sets are the sets of the form

$$E = \bigcup_{i_1, i_2, \dots} \bigcap_{k=1}^{\infty} E_{i_1, i_2, \dots, i_k}$$

where $E_{i_1, i_2, \dots, i_k}$ is a closed set or each finite sequence of positive integers.

Every Borelian set is a Souslin set.

Theorem 7 (Mattila, 1984). *Let $(\Re^n, d)$ be the Euclidean metric space and let A and B be two Souslin subsets of $\Re^n$, with $dim_H(A) = s$, $dim_H(B) = t$ and with positive Hausdorff measure in their dimension; denote by $\Psi(x, \delta)$ the closed ball with centre x and radius δ and suppose that the following lower density assumption on B holds*

$$\liminf_{\delta \to 0} \delta^{-t} h^t (B \cap \Psi(x, \delta)) > 0 \text{ for all } x \in B.$$

Then we have that $dim_H(AB) = s + t - n$.

As a consequence of the previous theorem we obtain that a necessary condition for s-irrelevance between Souslin sets of $\Re^n$ is that the two sets and their intersection have the Hausdorff dimension equal to n.

Proposition 1 Let $(\Re^n, d)$ be the Euclidean metric space and let A and B be two Souslin subsets of $\Re^n$ with $\dim_H(A) = s$, $\dim_H(B) = t$, with positive and finite Hausdorff measure in their dimension and such that the lower density assumption on B is satisfied; then B is s-relevant to A if $s \neq n$ or $t \neq n$.

Proof. Since A and B are two Souslin subsets of $\Re^n$ such that lower density assumption on B holds then we have that $\dim_H(AB) = s + t - n$. So condition 1s) of Definition 4 is satisfied if and only if $s = t = s + t - n$, that is $s = t = n$. $\diamond$

In the next section the previous result is used to find examples of s-dependent events.

Given a non-empty subset Ω of $\Re^n$ with Hausdorff dimension equal to n and positive and finite Hausdorff outer measure in its dimension, conditions such that B is s-irrelevant to A are proven. Under regular conditions such as those ones of Theorem 7 we have that condition 1s) of the Definition 4 is satisfied if and only if $dim_H(A) = dim_H(B) = dim_H(AB) = n$. In the next theorem we assume that these equalities hold and sufficient conditions such that B is s-irrelevant to A are proven.

Theorem 8. *Let $(\Re^n, d)$ be the Euclidean metric space and let A and B be two Souslin subsets of $\Re^n$ with positive and finite Hausdorff measure in their dimension, such that the following lower density assumption on B holds*

$$\liminf_{\delta \to 0} \delta^{-n} h^n (B \cap \Psi(x, \delta)) > 0 \text{ for all } x \in B$$

and such that $dim_H(A) = dim_H(B) = dim_H(AB) = n$.

Denoted by t the Hausdorff dimension of B^c then B is s-irrelevant to A in the following cases:

- *a) $t = n$, $h^n(B^c) > 0$ and $h^n(B^c)h^n(AB) - h^n(B)h^n(AB^c) = 0$*
- *b) $t = n$, $h^n(B^c) = 0$ and $h^n(A) = 0$ or*
- *$t = n$, $h^n(B^c) = 0$ and $h^n(A) = h^n(AB) = h^n(\Omega)$*
- *c) $t < n$ and $h^n(A) = 0$ or*
- *$t < n$ and $\frac{h^t(AB^c)}{h^t(B^c)} = \frac{h^n(A)}{h^n(B)}$*

Proof. We have to prove that condition 2s) of Definition 4 is satisfied. We consider the following cases:

- a) $t = n$ and $h^t(B^c) > 0$;
- b) $t = n$ and $h^t(B^c) = 0$;
- c) $t < n$

In the case a) condition 2s) of the definition of s-irrelevance is $\frac{h^n(AB)}{h^n(B)} = \frac{h^n(AB^c)}{h^n(B^c)} = \frac{h^n(A)}{h^n(\Omega)}$.
Since Souslin sets are measurable with respect to every Hausdorff outer measure (Falconer, 1986, p.6) we have that h^n is additive so that $h^n(A) = h^n(AB) + h^n(AB^c)$ and condition 2s) is satisfied if and only if $h^n(B^c)h^n(AB) - h^n(B)h^n(AB^c) = 0$. In the case b) condition 2s) is $\frac{h^n(AB)}{h^n(B)} = m(AB^c) = \frac{h^n(A)}{h^n(\Omega)}$.
Two cases are possible $m(AB^c) = 0$ or $m(AB^c) = 1$.
If $m(AB^c) = 0$ condition 2s) is satisfied if and only if $\frac{h^n(A)}{h^n(\Omega)} = 0$.
If $m(AB^c) = 1$ condition 2s) is satisfied if and only if $h^n(A) = h^n(AB) = h^n(\Omega)$
In the case c) two cases are possible: $0 < h^t(B^c) < +\infty$ or $h^t(B^c) = 0$.
If $0 < h^t(B^c) < \infty$ then:
if $dim_H(AB^c) < t$ then $h^t(AB^c) = 0$ and condition 2s) is satisfied if and only if $h^n(A) = 0$;
if $dim_H(AB^c) = t$ and $h^t(AB^c) = 0$ and condition 2s) is satisfied if and only if $h^n(A) = 0$;
if $dim_H(AB^c) = t$ and $0 < h^t(AB^c) < \infty$ then $h^n(A) = h^n(AB)$ since $h^n(AB^c) = 0$; so condition 2s) is saisfied if and only if $\frac{h^t(AB^c)}{h^t(B^c)} = \frac{h^n(A)}{h^n(B)}$.
If $h^t(B^c) = 0$ then $h^t(AB^c) = 0$ and we obtain the same condition of case b). $\diamond$
Example 4 Let $\Omega = [0, 1]$, let $A = [0, 1] - \frac{1}{2}$ and let B be the complement of the Cantor set. We are in case c) of the previous theorem so B is s-irrelevant to A.

5. Stochastic s-dependence for self-similar sets

In this section some fractal sets are proven to be s-dependent showing that they do not satisfy condition 1s) of Definition 4.
Conditions under which the attractor of a finite family of similitudes and its boundary are s-dependent and two middle third Cantor sets are s-dependent are found. If coherent upper conditional probability is defined as in Theorem 5 we have that all these sets satisfy the factorization property, which is the standard definition of probabilistic dependence.

5.1 s-Dependence of the attractor of a finite family of similitudes on its boundary
Let $(\Re^n, d)$ be the Euclidean metric space. A function $f : \Re^n \to \Re^n$ is called a *contraction* if $d(f(x), f(y)) \leq rd(x, y)$ for all $x, y \in \Re^n$, where $0 < r < 1$ is a constant. The infimum value for which this inequality holds for all x, y is called the *ratio* of the contraction. A contraction that trasforms every subset of $\Re^n$ to a geometrically similar set is called a *similitude*. A similitude is a composition of a dilation, a rotation and translation. A set E is called *invariant* for a finite set of contractions $\{f_1, f_2, ..., f_m\}$ if $E = \bigcup_{i=1}^m f_i(E)$.
If the contractions are similitudes and for some s we have $h^s(E) > 0$ but $h^s(f_i(E) \cap f_j(E)) = 0$ for $i \neq j$ then E is *self similar*. For any finite set of contractions there exists a unique non-empty compact invariant set K (Falconer, 1986, Theorem 8.3), called *attractor*.
Given a finite set of contractions $\{f_1, f_2, ..., f_m\}$ we say that *the open set condition* (OSC) holds if there exists a bounded open set O such that $O \subset \bigcup_{i=1}^m f_i(O)$ and $f_i(O) \cap f_j(O) = \oslash$ for $i \neq j$.

If $\{f_1, f_2, ..., f_m\}$ are similitudes with similarity ratios r_i for $i = 1...m$ the similarity dimension, which has the advantage of being easily calculable, is the unique positive number s for which $\sum_{i=1}^{m} r_i^s = 1$. If the OSC holds then the compact invariant set K is self-similar and the Hausdorff dimension and the similarity dimension of K are equal. If the similarity dimension is equal to n then the interior of K, K^0 is non empty. In Lau (1999) it has been proven that, given a finite family of similitudes and the corresponding attractor K, if K^0 is non-void and Hausdorff dimension of K is equal to n then the Hausdorff dimension of the boundary of K is less than n. Moreover since the lower density assumption holds for a self-similar set, from Proposition 1 of Section 4 we have that K and its boundary are s-dependent. We can observe that if upper and lower probabilities are defined as in Theorem 5 then K and its boundary satisfy the the factorization property.

5.2 s-Dependence for Cantor sets

In this subsection two middle third Cantor sets are proven to be s-dependent.
Let $([0,1], d)$ be the Euclidean metric space and let E be the Cantor set. For every $x \in [0,1]$ we consider the Cantor set $x + E$ that is the the translation of E.
In Davis (1995) it has been proven that for every $\alpha \in [0,1]$ there exists an $x \in (0,1)$ such that the Hausdorff dimension of the intersection $x + E \cap E = (1 - \alpha) \frac{ln2}{ln3}$. So for every $\alpha \in (0,1]$ we have that the two middle third Cantor sets $x + E$ and E are s-dependent since condition 1s) of Definition 4 is not satisfied.
We can observe that if upper conditional probability is defined as in Theorem 5 the factorization property is satisfied and so the two middle Cantor sets are stochastically independent according to the axiomatic definition of independence.

6. s-Independence for curves filling the space

In this section the notions of s-irrelevance and s-independence for events A and B that are represented by curves filling the space are analyzed.In particular Peano curve, Hilbert curve and Peano-Sierpinski curve are proven to be s-independent. Curves filling the space Sagan (1994) can be defined as the limit of a Cauchy sequence of continuous functions f_n, each mapping the unit interval into the unit square. The convergence is uniform so that the limit is a continuous function, i.e. a curve. The definition of irrelevance given by Walley, which is condition 2s) of s-irrelevance, holds when the two events A and B are not the trivial events $(\Omega, \oslash)$. If the conditioning event B is represented by a curve filling the space, we have that the complement of B is the empty-set and so in this case the notion of irrelevance becomes $\overline{P}(A|B) = \overline{P}(A)$; and $\underline{P}(A|B) = \underline{P}(A)$. If A and B are represented by curves filling the space we obtain the following definition of s-independence.

Definition 5 Let (Ω, d) be a metric space and let A and B be two curves filling the space Ω. Then A and B are *s-independent* if the following conditions hold

- 1s) $dim_H(AB) = dim_H(B) = dim_H(A)$
- 2s) $\overline{P}(A|B) = \overline{P}(A)$ and $\underline{P}(A|B) = \underline{P}(A)$;
- 3s) $\overline{P}(B|A) = \overline{P}(B)$ and $\underline{P}(B|A) = \underline{P}(B)$;

Moreover B is *s-irrelevant* to A if conditions 1s) and 2s) are satisfied.

Theorem 9. *Let $\Omega = [0,1]^n$ and let $\overline{P}$ and $\underline{P}$ be the upper and lower conditional probabilities defined as in Theorem 4. If A and B are two curves filling the space then A and B are s-independent.*

Proof. Since A and B are curves filling the space then they and their intersection have Hausdorff dimension equal to n. Moreover since A and B are measurable we have that $\overline{P} = \underline{P} = P$ and conditions 2s) and 3s) of Definition 5 become

$$\frac{h^n(AB)}{h^n(B)} = \frac{h^n(A)}{h^n(\Omega)} \text{ and } \frac{h^n(AB)}{h^n(A)} = \frac{h^n(B)}{h^n(\Omega)}$$

that are satisfied since they vanish to $1 = 1$. $\diamond$

As a consequence of the previous theorem we have that the Peano curve and the Hilbert curve are s-independent. Moreover if B is a curve filling the space $\Omega = [0,1]^n$ and A is any event with Hausdorff dimension equal to n, then B is s-irrelevant to A.

7. Conclusions

In this chapter the notions of s-irrelevance and s-independence with respect to upper and lower conditional probabilities defined by Hausdorff outer and inner measures are introduced. They are used to discover probabilistic dependence for events, which are probabilistic independent with respect to the standard definition given by the factorization property or with respect to the notion of epistemic independence.

Results and examples are given for fractal sets (i.e. sets with non-integer Hausdorff dimension) which often model complex phenomena. In particular the attractor of a finite family of similitudes and its boundary are proven to be s-dependent if the open set condition holds and a sufficient condition is given such that two middle Cantor sets are s-dependent.

Moreover two curves filling the space, such as Peano curve and Hilbert curve, are proven to be s-independent.

8. References

P. Billingsley. (1986). *Probability and measure*, Wiley, USA.

G.J. Davis, Tian-You Hu. On the structure of the intersection of two middle third Cantor sets. *Publications Matemàtiques*, Vol. 39, 43-60, 1995.

B. de Finetti. (1970). *Teoria della Probabilita'*, Einaudi Editore, Torino.

D. Denneberg. (1994). *Non-additive measure and integral*. Kluwer Academic Publishers.

S. Doria. (2007). Probabilistic independence with respect to upper and lower conditional probabilities assigned by Hausdorff outer and inner measures. *International Journal of Approximate Reasoning*, 46, 617-635 .

L.E. Dubins. (1975). Finitely additive conditional probabilities, conglomerability and disintegrations. *The Annals of Probability*, Vol. 3, 89-99.

K.J. Falconer. (1986). *The geometry of fractals sets*. Cambridge University Press.

G. Koch. (1997). *La matematica del probabile* Aracne Editrice.

Ka-Sing Lau, You Xu. (1999). On the boundary of attractors with non-void interior. *Proceedings of the American Mathematica Society*, Vol. 128, N.6, pp.1761-1768.

P. Mattila. (1984). Hausdorff dimension and capacities of intersections of sets in n-space. *Acta Mathematica*, 152, 77-105.

C.A. Rogers. (1998). *Hausdorff measures*. Cambridge University Press.

H. Sagan.(1994). *Space-Filling Curves*. Springer-Verlag.

P. Walley. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London.

Design and experimentation of a large scale distributed stochastic control algorithm applied to energy management problems

Xavier Warin
EDF
France

Stephane Vialle
SUPELEC & AlGorille INRIA Project Team
France

Abstract

The Stochastic Dynamic Programming method often used to solve some stochastic optimization problems is only usable in low dimension, being plagued by the curse of dimensionality. In this article, we explain how to postpone this limit by using *High Performance Computing*: parallel and distributed algorithms design, optimized implementations and usage of large scale distributed architectures (PC clusters and Blue Gene/P).

1. Introduction and objectives

Stochastic optimization is used in many industries to take decisions facing some uncertainties in the future. The asset to optimize can be a network (railway, telecommunication [Charalambous et al. (2005)]), some exotic financial options of american type [Hull (2008)]. In the energy industry, a gaz company may want to optimize the use of a gaz storage [Chen & Forsyth (2009)], [Ludkovski & Carmona (2010, to appear)]. An electricity company may want to optimize the value of a powerplant [Porchet et al. (2009)] facing a price signal and dealing with operational contraints: ramp contraints, minimum on-off times, maximum number of start up during a period.
An integrated energy company may want to maximize an expected revenue coming from many decisions take:

- which thermal assets to use ?
- how should be managed the hydraulic reservoirs ?
- which customers options to exercice ?
- how should the physical portfolio be hedged with future contracts ?

In the previous example, due to the structure of the energy market with limited liquidity, the management of a future position on the market can be seen as the management of a stock of energy available at a given price. So the problem can be seen as an optimization problem

with many stocks to deal with. This example will be taken as a test case for our performance studies.

In order to solve stochastic optimization problems, some methods have be developed in the case of convex continuous optimization: The *Stochastic Dual Dynamic Programming* method [Rotting & Gjelsvik (1992)] is widely used for compagnies having large stocks of water to manage. When the company portfolio is composed of many stocks of water and many power plants a decomposition method can be used [Culioli & Cohen (1990)] and the bundle method may be used for coordination [Bacaud et al. (2001)]. The uncertainty is usually modeled with trees [Heitsch & Romisch (2003)].

In realistic modelization of the previous problem, the convexity is not assured. The constraints may be non linear as for gaz storage for example where injection and withdrawal capacities depend on the position in the stock (and for thermodynamic reason depends on the past controls in accurate model). Most of the time, the problem is not continuous and is in fact a mixed integer stochastic problem: the commands associated to a stock of water can only take some discrete values due to the fact that a turbine has only on-off positions, financial positions are taken for discrete number of stocks... If the constraints and the objectif function are linearized, the stochastic problem can be discretized on the tree and a mixed integer programming solver can be used. In order to be able to use this kind of modelization a non recombining tree has to be build. The explosion of the number of leaves of the tree leads to a huge mixed integer problem to solve.

Therefore when the constraints are non linear or when the problem is non convex, the dynamic programming method developed in 1957 [Bellman (1957)] may be the most attractive. This simple approach faces one flaw: it is an enumerative method and the computational cost goes up exponentially with the number of state variable to manage. This approach is currently used for a number of state variable below 5 or 6. This article introduces the parallelization scheme developed to implement the dynamic programming method, details some improvements required to run large benchmarks on large scale architectures, and presents the serial optimizations achieved to efficiently run on each node of a PC cluster and an IBM Blue Gene/P supercomputer. This approach allows us to tackle a simplified problem with 3 random factors to face and 7 stocks to manage.

2. Stochastic control optimization and simulation

We give a simplified view of a stochastic control optimization problem. Supposing that the problem we propose to solve can be set as:

$$minimize \quad \mathbb{E} \left(\sum_{t=1}^{N} \phi(t, \xi_t, nc_t) \right) \tag{1}$$

where ϕ is a cost function depending on time, the state variable ξ_t (stock and uncertainty,) and depending on the command nc_t realized at date t. For simplicity, we suppose that the control only acts on the deterministic stocks and that the uncertainties are uncontrolled. Some additional constraints are added defining at date t the possible commands nc_t depending on ξ_t.

The software used to manage the energy assets are usually separated into two parts. A first software, an optimization solver is used to calculate the so-called Bellman value until maturity T. The second one will test the Bellman values calculated during the first software run on some scenarios.

2.1 Optimization part

In our implementation of the Bellman method, we store the Bellman values J at a given time step t, for a given uncertainty factor occurring at time t and for some stocks levels. These Bellman values represent the expected gains remaining for the optimal asset management from the date t until the date T starting optimization with a given state. Instead of using usual non recombining trees, we have chosen to use Monte Carlo scenarios to achieve our optimization following [Longstaff & Schwartz (2001)] methodology. The uncertainties are here simplified so that the model is Markovian. The number of scenarios used during this part is rather small (less than a thousand). This part is by far the most time consuming. The algorithm 1 gives the Bellman values J for each time step t calculated by backward recursion. In the algorithm 1, due to the Markovian property of the uncertainty, $s^* = f(s, w)$ is a realization at date $t + \Delta t$ of an uncertainty whose value is equal to s at date t, where w is a random factor independent on s.

> For $t := (nbstep - 1)\Delta t$ to 0
>   For $c \in$ admissible stock levels ($nbstate$ levels)
>   For $s \in$ all uncertainty ($nbtrajectory$)
> $\tilde{J}^*(s, c) = \infty$
>   For $nc \in$ all possible commands for stocks ($nbcommand$)
> $\tilde{J}^*(s, c) = \min(\tilde{J}^*(s, c), \phi(nc) + \mathbb{E}\left(J(t + \Delta t, s^*, c + nc)|s\right))$
> $J^*(t, :, :) := \tilde{J}^*$

Fig. 1. Bellman algorithm, with a backward computation loop.

In our modelization, uncertainties are driven by brownian processus and conditional expectation in the algorithm 1 are calculated by regression methods as explained in [Longstaff & Schwartz (2001)]. Using Monte Carlo, it could have been possible to use Malliavin methods [Bouchard et al. (2004)], or it could have been possible to use a recombining quantization tree [Bally et al. (2005)].

2.2 Simulation part

A second software called a simulator is then used to accurately compute some financial indicators (VaR, EEaR, expected gains on some given periods). The optimization part only gives the Bellman values in each possible state of the system. In the simulation part, the uncertainties are accurately described with using many scenarios (many tens of thousand) to accurately test the previously calculated Bellman values. Besides, the modelization in the optimizer is often a simplified one so that calculation are made possible by a reduction in the number of state variable. In the simulator it is often much more easier to deal with far more complicated constraints so that the modelization is more realistic. In the simulator, all the simulations can be achieved in parallel, so we could think that this part is embarrassingly parallel as shown by algorithm 2. However, we will see in the sequel that the parallelization scheme used during the optimization will bring some difficulties during simulations that will lead to some parallelization task to achieve.

3. Distributed algorithm

Our goal was to develop a distributed application efficiently running both on large PC cluster (using Linux and classic NFS) and on IBM Blue Gene supercomputers. To achieve this goal, we have designed some main mechanisms and sub-algorithms to manage data distribution and load balancing, data routage planning and data routage execution, and file accesses. Next

```
stock(1:nbtrajectory) = initialStock
For t := 0 to (nbstep − 1)Δt
    For s ∈ all uncertainty (nbtrajectory)
        Gain = - ∞
        For nc ∈ all possible commands for stocks (nbcommand)
                GainA = phi(nc) + 𝔼 (J*(t + Δt, s*, stock(s) + nc)|s)
                if GainA > Gain
                    com = nc
                    Gain = GainA
        stock(s) += com
```

Fig. 2. Simulation on some scenarios.

sections introduce our parallelization strategy, detail the most important issues and describe our global distributed algorithm.

3.1 Parallelization overview of the optimization part

As explained in section 2 we use a backward loop to achieve the optimization part of our stochastic control application. This backward loop is applied to calculate the Bellman values at discrete points belonging to a set of N stocks, which form some N-dimensional cube of data, or *data N-cubes*.

Considering one stock X, its stock levels at t_n and t_{n+1} are linked by the equation:

$$X_{n+1} = X_n + Command_n + Supply_n \tag{2}$$

Where:

- X_n and X_{n+1} are possible levels of the X stock, and belong to intervals of possible values ($[X_n^{min}; X_n^{max}]$ and $[X_{n+1}^{min}; X_{n+1}^{max}]$), function of scenarios and physical constraints.

- The *Command* is the change of stock level due to the execution of a *command* on the stock X between t_n and t_{n+1}. It belongs to an interval of values: $[C_n^{min}; C_n^{max}]$, function of scenarios and physical constraints.

- The *Supply$_n$* is the change of stock level due to an external supply (in our test case with hydraulic energy stocks, snow melting and rain represent this supply). Again, it belongs to an interval of values: $[S_n^{min}; S_n^{max}]$, function of scenarios and physical constraints.

Considering the equation 2, the backward loop algorithm introduced in section 2, a set of scenarios and physical constraints, and N stocks, the following 6 sub-steps algorithm is run on each computing node at each time step:

1. When finishing the t_{n+1} computing step and entering t_n one (backward loop), minimal and maximal stock levels of all stocks are computed on each computing node, according to scenarios and physical constraints on each stock. So, each node easily computes N minimal and maximal stock levels that defines the minimal and maximal vertexes of the *N-cube* of points where the Bellman values have to be calculated at date t_n.

2. Each node runs its splitting algorithm of the t_n N-cube to distribute the t_n Bellman values that will be to computed at step t_n on $P = 2^{d_p}$ computing nodes. Each node computes the entire map of this distribution: the t_n *data map*. See section 3.4 for details about the splitting algorithm.

3. Using scenarios and physical constraints set for the application, each node computes the *Commands* and *Supplies* to apply to each stock of each t_n N-subcube of the t_n *data map*. Using equation 2 each node computes the t_{n+1} N-subcube of points where the t_{n+1} Bellman values are required by each node to process the calculation of the Bellman values at the stocks points belonging to its t_n N-subcube. So, each node easily computes the coordinates of the P' t_{n+1} *data influence areas* or t_n *shadow regions*, and builds the entire t_n *shadow region map* without any communication with others nodes. This solution has appeared faster than to compute only local data partition and local shadow region on each node and to exchange messages on the communication network to gather the complete maps on each node.

4. Now each node has the entire t_n *shadow region map* computed at the previous sub-step, and the entire t_{n+1} *data map* computed at the previous time step of the backward loop. Some basic computations of N-cube intersections allow each node to compute the P N-subcubes of points associated to the Bellman values to receive from others nodes and from itself, and the P N-subcubes of points associated to the Bellman values to send to other nodes. Some of these N-subcubes can be empty and have a null size, when some nodes have no N-subcubes of data at time step t_n or t_{n+1}. So, each node builds its t_n *routing plan*, still without any communications with other nodes. See section 3.5 for details about the computation of this routing plan.

5. Using MPI communication routines, each node executes its t_n *routing plan* and brings back the Bellman values associated to points belonging to its t_{n+1} *shadow region* in its local memory. Function of the underlying interconnection network and the machine size, it can be interesting to overlap all communications, or it can be necessary to spread the numerous communications and to achieve several communication sub-steps. See section 3.6 for details about the routing plan execution.

6. Using the t_{n+1} Bellman brought back in its memory, each node can achieve the computation of the optimal commands for all stock points (according to the stochastic control algorithm) and calculate its t_n Bellman value.

7. Then, each node save on disk the t_n Bellman values and some others step results that will be used in the *simulation part* of the application. They are temporary results stored on local disks when exist, or in global storage area, depending of the underlying parallel architecture. Finally, each node cancels its t_{n+1} *data map*, t_n *shadow region map* and t_n *routing plan*. Only its t_n *data map* and t_n *data N-subcube* have to remain to process the following time step.

This time step algorithm is repeated in the backward loop up to time step 0. Then some global results are saved, and the simulation part of the application is run.

3.2 Parallelization overview of the simulation part

In usual sequential software, simulations is achieved scenario by scenario: the stock levels and the commands are calculated from date 0 to date T for each scenario sequentially. This approach is obviously easy to parallelize when the Bellman values are shared by each node. In our case, doing so will mean a lot of time spent in IO. In the algorithm 2, it has been chosen to advance time step by time step and to do the calculation at each time step for all simulations. So Bellman temporary files stored in the optimization part are opened and closed only once by time step to read Bellman values of the next time step.

Similarly to the optimization part, at each time step t_n the following algorithm is achieved by each computing node:

1. Each computing node reads some temporary files of optimization results: the t_{n+1} *data map* and the t_{n+1} *data* (Bellman values). All these reading operations are achieved in parallel from the P computing nodes.

2. For each trajectory (up to the number of trajectories managed by each node):

    (a) Each node simulates the hazard trajectory from time step t_n to time step t_{n+1}.

    (b) From the current N dimensional stock point SP_n, using equation 2, each node computes the t_{n+1} N-subcube of points where the t_{n+1} Bellman values are required to process the calculation of the optimal command at SP_n: the t_n *shadow region* coordinates of the current trajectory.

    (c) All nodes exchange their t_n *shadow region* coordinates using MPI communication routines and achieving a *all_gather* communication scheme. So, each node can build a complete t_{n+1} *shadow region map* in its local memory. In the *optimization* part each node could compute the entire t_{n+1} *shadow region map*, but in the *simulation* part inter-node communications are mandatory.

    (d) Each node computes its *routing plan*, computing N-subcubes intersections of t_{n+1} *data map* and t_{n+1} *shadow region map*. We apply again the 2-step algorithm described on figure 6 and used in the *optimization* part.

    (e) Each node executes its *routing plan* using MPI communication routines, and and brings back the Bellman values associated to points belonging to its t_{n+1} *shadow region* in its local memory. Like in the *optimization* part, depending on the underlying interconnection network and the machine size, it can be interesting to overlap all communications, or it can be necessary to spread the numerous communications and to achieve several communication sub-steps (see section 3.6).

    (f) Using the t_{n+1} Bellman value brought back in its memory, each node can compute the optimal command according to algorithm introduced on figure 2.

3. If required by user, data of the current time step are gathered on computing node 0, and written on disk (see section 3.7).

Finally, some complete results are computed and saved by node 0, like the global gain computed by the entire application.

3.3 Global distributed algorithm

Figure 3 summarizes the main three parts of our complete algorithm to compute optimal commands of a N dimensional optimization problem and to test them in simulation. The first part is the reading of input data files according to the IO strategy introduced in section 3.7. The second part is the *optimization solver* execution, computing some Bellman values in a backward loop (see sections 1 and 3.1). At each step, a N-cube of Bellman values to compute is split on an hypercube of computing nodes to load balance the computations, a shadow region is identified and gathered on each node, some multithreaded local computations of optimal commands are achieved for each point of the N-cube (see section 4.3), and some temporary results are stored on disk. Then, the third part tests the previously computed commands. This *simulation* part runs a forward time step loop (see sections 1 and 3.2) and a Monte-Carlo trajectory sub-loop (see section 4.3), and uses the same previous mechanisms than the second

I - Sequential reading and parallel broadcast
  of initialization data

II - Backward time step loop optimization:

- data and computations distribution
- shadow regions rapatriement with communications
- (multithreaded) optimal command research for all stock points
- temporary data writing on disk

III - Forward time step loop simulation:

- temporary data reading from disk
- Monte-Carlo trajectory sub-loop
    - shadow regions rapatriement with communications
    - trajectory computation
- optimal control result saving

Fig. 3. Main steps of the complete application algorithm.

part. At each time step, each node reads some temporary results it has stored during the *optimization* part, gather some shadow regions, achieves some local computations, and stores the final results on disk.

3.4 Data N-cube splitting and data map setting

During the backward computation loop of the Bellman algorithm (see figure 1) of the *optimization part* of our application, we need to compute a N-dimensional cube of Bellman values at each time step. This computation is long and requires a large amount of memory to store the N-cube data. So we have to *split* this N-cube data on a set of computing nodes both to speedup (using more processors) and to size up (using more memory). Moreover, each dimension of this N-cube represents the *stock levels* of one *stock* that can change from time step t_{n+1} to t_n. Each stock level range can be translated, and/or enlarged or shrunk. So, we have to redistribute our problem at each time step: we have to *split* a new N-cube of stock point when entering a new time step. Our N-cube splitting algorithm is a critical component of our distributed application that must run quickly. During the forward loop of the *simulation part* we reread on disk the maps stored during optimization.

The computation of one Bellman value at one point of the t_n N-cube requires the *influence area* of this value given by equation 2: the t_{n+1} Bellman values at stocks points belonging to a small sub-cube of the t_{n+1} N-cube. Computation of the entire t_n N-sub-cube attached to one computing node requires an *influence area* that can be a large *shadow region*, leading to MPI communication of Bellman values stored on many other computing nodes (see figure 4). To minimize the size of this N-dimensional *shadow region* we favor *cubic* N-sub-cubes in place of *flat* ones. So, we aim to achieve *cubic* split of the N-cube data at each time step.

We decided to split our N-cube data on $P_{max} = 2^{d_{max}}$ computing nodes. We successively split in two equal parts some dimensions of the N-cube, up to obtain $2^{d_{max}}$ sub-cubes, or to have reach the limits of the division of the N-cube. Our algorithm includes 3 sub-steps:
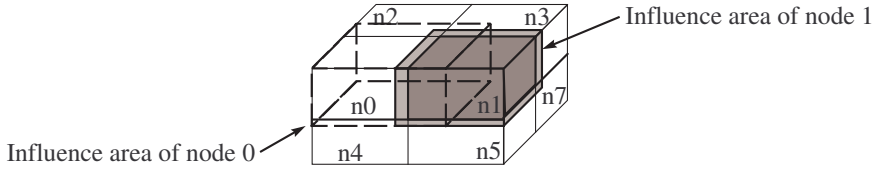
Fig. 4. Example of *cubic* split of the N-Cube of data and computations
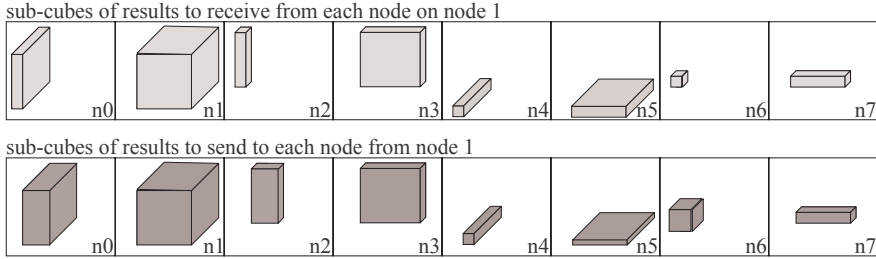
Fig. 5. Example of routing plan established on node 1

1. We split the dimensions of the N-cube in order to obtain sub-cubes with close dimension sizes. We start to sort the N' divisible dimensions in decreasing order, and attempt to split the first one in 2 equal parts with sizes close to size of the second dimension. Then we attempt to split again the size of the 2 first dimensions to reduce their sizes close to the size of the third one. This splitting operation fails if it leads to a sub-cube dimension size smaller than a *minimal size*, set to avoid to process too small data sub-cubes. The splitting operation is repeated up to achieve d_{max} splits, or up to reduce the sizes of the $N' - 1$ first dimensions close to the size of the smallest one. Then, if we have not obtained $2^{d_{max}}$ sub-cubes we run the second sub-step.

2. Previously we have obtained sub-cubes with N' close dimension sizes. Now we sort these N' divisible dimensions in decreasing order, considering their split dimension sizes. We attempt to split again in 2 equal parts each divisible dimension in a round robin way, up to achieve d_{max} splits, or up to reach the limit of the *minimal size* for each divisible dimension. Then, if we have not obtained $2^{d_{max}}$ sub-cubes we run the third sub-step.

3. If it is specified to *exceed* the *minimal size* limit, then we split again in 2 equal parts each divisible dimension in a round robin way, up to achieve d_{max} splits, or up to reach dimension sizes equal to 1. In our application, the *minimal size* value is set before to split a N-cube, and a command line option allows the user to respect or to exceed this limit. So, when processing small problems on large numbers of computing nodes, some experiments are required and can be rapidly conducted to point out the right tuning of our splitting algorithm.

Finally, after running our splitting algorithm at time step t_n we obtain 2^d sub-cubes, and we can give data and work up to $P = 2^d$ computing nodes. When processing small problems on

// Input variable datatypes and declarations on node Me
N-cube-coord_t $DataMap_{t_{n+1}}[P]$
N-cube-coord_t $ShadowRegion_{t_n}[P]$
// Output variable datatypes and declarations on node Me
N-cube-coord_t $LocalRoutingPlan_{t_n}^{Recv}[P]$
N-cube-coord_t $LocalRoutingPlan_{t_n}^{Send}[P]$

// Coordinates computation of the N-subcubes to receive on node Me from all nodes
For $i := 0\ to\ (P-1)$
    $LocalRoutingPlan_{t_n}^{Recv}[i] := DataMap_{t_{n+1}}[i] \cap ShadowRegionMap_{t_n}[Me]$

// Coordinates computation of the N-subcubes to send to all nodes from node Me
For $i := 0\ to\ (P-1)$
    $LocalRoutingPlan_{t_n}^{Send}[i] := DataMap_{t_{n+1}}[Me] \cap ShadowRegionMap_{t_n}[i]$

Fig. 6. Computation of t_n routing plan on computing node Me $(0 \leq Me < P)$

large parallel machines, it is possible not all computing nodes will have some computations to achieve at time step t_n $(P < P_{max})$ (a too fine grained data distribution would lead to inefficient parallelization). This splitting algorithm is run on each computing node at the beginning of time step t_n. They all compute the same N-cube splitting and deduce the same number of provisioned nodes P and the same *data map*.

3.5 Shadow region map and routing plan computations

Different data N-subcubes located on different nodes, or existing at different time steps, can have *shadow regions* with different sizes. Moreover, depending on the time step, the problem size and the number of used computing nodes, the *shadow region* N-subcube of one computing node can reach only its direct neighbors or can encompass these nodes. So, the exact routing plan of each node has to be dynamically established at each time step before to retrieve data from other nodes.

As explained in section 3.1, each node computes the entire *shadow region map*: a table of P coordinates of N-subcubes. In our application these entire maps can be deduced from t_n and t_{n+1} *data maps*, and from scenarios and physical constraints on commands and supplies of each stock. For example, node 1 on figure 4 knows its *shadow region* (light gray cube) in this 3-cube, the *shadow region* of node 0 (dotted line) and of nodes 2 to 7 (not drawn on figure 4).

Then, using both its t_n *shadow region map* and its t_{n+1} *data maps*, each computing node can easily compute its local t_n *routing plan* in two sub-steps:

1. Each node computes the coordinates of the N-subcubes of Bellman values it has to receive from other nodes: the *receive part* of its local *routing plan*. The intersection of the t_n *shadow region* N-subcube of node Me with the t_{n+1} N-subcube of another node gives the t_{n+1} N-subcube of Bellman values the node Me has to receive from this node. So, each node achieve the first loop of the algorithm described on figure 6, and computes P intersections of N-subcubes coordinates, to get the coordinates of the P N-subcube of Bellman values it has to receive. When the *shadow regions* are not too large, many of these P N-subcubes are empty.

2. Each node computes the coordinates of the N-subcubes of Bellman values it has to send to other nodes: the *send part* of its local *routing plan*. The intersection of the t_{n+1} N-subcube of node Me with the t_n *shadow region* N-subcube of another node gives the

t_{n+1} N-subcube of Bellman values the node *Me* has to send to this node. So, each node achieve the second loop of the algorithm described on figure 6, and computes P intersections of N-subcubes coordinates, to get the coordinates of the P N-subcube of Bellman values it has to send. Again, many of these N-subcubes are empty when the *shadow regions* are not too large.

Figure 5 shows an example of local *routing plan* computed on node 1, considering the data distribution partially illustrated on figure 4. This entire routing plan computation consists in $2.P$ intersections of N-subcube coordinates. Finally, this is a very fast integer computation, run at each time step.

3.6 Routing plan execution

Node communications are implemented with non-blocking communications and are overlapped, in order to use the maximal abilities of the interconnection network. However, for large number of nodes we can get small sub-cubes of data on each node, and the influence areas can reach many nodes (not only direct neighbor nodes). Then, the routing plan execution achieves a huge number of communications, and some node interconnexion network could saturate and slow down. So, we have parameterized the routing plan execution with the number of nodes that a node can attempt to contact simultaneously. This mechanism spreads the execution of the communication plan, and the spreading out is controlled by two application options (specified on the command line): one for the *optimization* part, and one for the *simulation* part.

When running our benchmark (see section 5) on our 256 dual-core PC cluster it has been faster not to spread these communications, but on our 8192 quad-core Blue Gene/P it has been really faster to spread the communications of the *simulation* part. Each Blue Gene node has to contact only 128 or 256 other nodes at the same time, to prevent the simulation time to double. When running larger benchmarks (closer to future real case experiments), the size of the data and of the *shadow regions* could increase. Moreover, each *shadow region* could spread on a little bit more nodes. So, the total size and number of communications could increase, and it seems necessary to be able to temporally spread both communications of *optimization* and *simulation* parts, on both our PC-cluster and our Blue Gene/P supercomputer.

So, we have maintained our communication spreading strategy. When running the application, an option on the command line allows to limit the number of simultaneous asynchronous communications a computing node can start. If a saturation of the communication system appears, it is possible to use it sparingly, spreading the communications.

3.7 File IO constraints and adopted solutions

Our application deals with input data files, temporary output and input files, and final result files. These files can be large, and our main target systems have very different file access mechanisms. Computing nodes of IBM Blue Gene supercomputers do not have local disks, but an efficient parallel file system and hardware allows all nodes to concurrently access a global remote disk storage. At the opposite, nodes of our Linux PC cluster have local disks but use basic Linux NFS mechanisms to access global remote disks. All nodes of our cluster can not make their disk accesses at the same time. When increasing the number of used nodes, IO execution times become longer, and finally they freeze.

Temporary files are written and read at each time step. However, each temporary result file is written during the *optimization* part by only one node, and is read during the *simulation* part by only the same node. These files do not require concurrent accesses and their management is

easy. Depending on their path specified on the command line when running the application,
they are stored on local disks (fastest solution on PC cluster), or on a remote global disk (IBM
Blue Gene solution). When using a unique global disk it is possible to store some temporary
index files only once, to reduce the total amount of data stored.

Input data files are read only once at the beginning, but have to be read by each computing
node. Final result files are written at each time step of the *simulation* part and have to store
data from each computing node. In both cases, we have favored the genericity of the file access
mechanism: node 0 opens, accesses and closes files, and sends data to or receives data from
other nodes across the interconnection network (using MPI communication routines). This IO
strategy is an old one and is not always the most efficient, but is highly portable. It has been
implemented in the first version of our distributed application. A new strategy, relying on a
Parallel File System and an efficient hardware, will be designed in future versions.

4. Parallel and distributed implementation issues

4.1 N-cube implementation

Our implementation includes 3 main kinds of arrays: MPI communication buffers, *N-cube
data maps* and *N-cube data*. We have used classic dynamic C arrays to implement the first kind,
and the `blitz++` generic C++ library [Veldhuizen (2001)] to implement the second and third
kinds. However, in order to compile the same source code independently of the number of
energy stocks to process, we have flattened the N-cubes required by our algorithms. Any
N-dimensional array of stock point values becomes a one dimensional array of values.

Our implementation includes the following kind of variables:

- A *stock level range* is a one dimensional array of 2 values, implemented with a
  `blitz::TinyVector` of 2 integer values.

- The coordinates of a *N-cube of stock points* is an array of N *stock level ranges*, implemented
  with a one dimensional `blitz::Array` of N `blitz::TinyVector` of 2 integer values.

- A *map* of *N-cube data* is implemented with a two dimensional array of $P \times N$
  stock level ranges. It is implemented with a two dimensional `blitz::Array` of
  `blitz::TinyVector`).

- A *Bellman value* is depending on the *stock point* considered and on the *alea* considered.
  Our *N-cube data* are arrays of Bellman values function of different *aleas* in a N-cube of
  stock points. A *N-cube data* is implemented with a two dimensional `blitz::Array` of
  `double`: the first dimension index is the flattened N dimensional coordinate of the stock
  point, and the second dimension index is the *alea* index.

- Some one dimensional arrays of `double` are used to store data to send to or to receive
  from another node, and some two dimensional arrays of `double` are used to store data
  to send to or to receive from all computing nodes. Communications are implemented
  with the MPI library and its C API, that was available on all our testbed architectures.
  This API requires addresses of contiguous memory areas, to read data to send or to
  store received data. So, classic C dynamic arrays appeared a nice solution to implement
  communication buffers with sizes updated at each time step.

Finally, `blitz` access mechanism to `blitz` array elements appeared slow. So, inside the com-
puting loop we prefer to get the address of the first element to access using a `blitz` function,
and to access the next elements incrementing a pointer like it is possible for a classic C array.

4.2 MPI communications

Our distributed application consists in loops of local computations and internode communi-
cations, and communications have to be achieved before to run the next local computations.
So, we do not attempt to overlap computations and communications. However, in a com-
munication step each node can exchange messages with many others, so it is important to
attempt to overlap all message exchanges and to avoid to serialize these exchanges.

When routing the Bellman values of the *shadow region* the communication schemes can be
different on each node and at each time step (see sub-steps 5 of section 3.1 and 2.e of sec-
tion 3.2), and data to send is not contiguous in memory. So, we have not used collective
communications (easier to use with regular communication schemes), but asynchronous MPI
point-to-point communication routines. Our communication sub-algorithm is the following:

- compute the size of each message to send or to receive,

- allocate message buffers, for messages to send and to receive,

- make local copy of data to send in the corresponding send buffers,

- start all asynchronous MPI point-to-point *receive* and *send* operations,

- wait until all *receive* operations have completed (synchronization operation),

- store received data in the corresponding application variables (`blitz++` arrays),

- wait until all *send* operations have completed (synchronization operation),

- delete all communication buffers.

As we have chosen to fill *explicit communication buffers* to store data to exchange, we have
used *in place* asynchronous communication routines to exchange these buffers (avoiding to
re-copy data in other buffers with buffered communications). We have used `MPI_Irecv`,
and `MPI_Isend` or `MPI_Issend`, depending on the architecture and MPI library used. The
`MPI_Isend` routines is usually faster but has a non standard behavior, function of the MPI
library and architecture used. The `MPI_Issend` is a little bit longer but has a standardized be-
havior. On Linux PC clusters where different MPI libraries are installed, we use `MPI_Issend`
/ `MPI_Irecv` routines. On IBM Blue Gene supercomputer, with an IBM MPI library, we suc-
cessfully experimented `MPI_Isend` / `MPI_Irecv` routines.

Internode communications required in IO operations to send initial data to each node, or to
save final results on disk in each time step of *simulation part* (see sub-step 7 of section 3.1), are
implemented with some collective MPI communications: `MPI_Bcast`, `MPI_Gather`.

Exchange of *local shadow region coordinates* in each time step of the *simulation part* (see sub-step
2.c of section 3.2) is implemented with a collective `MPI_Allgather` operation. All these com-
munication have very regular schemes and can be efficiently implemented with MPI collective
communication routines.

4.3 Nested loops multithreading

In order to take advantage of multi-core processors we have multithreaded, in order to create
only one MPI process per node and one thread per core in place of one MPI process per core.
Depending on the application and the computations achieved, this strategy can be more or less
efficient. We will see in section 5.4 it leads to serious performance increase of our application.
To achieve multithreading we have split some nested loops using OpenMP standard tool or
the Intel Thread Building Block library (TBB). We maintain these two multithreaded imple-
mentations to improve the portability of our code. For example, in the past we encountered

some problems at execution time using OpenMP with ICC compiler, and TBB was not available on Blue Gene supercomputers. Using OpenMP or Intel TBB, we have adopted an incremental and pragmatic approach to identify the nested loops to parallelize. First, we have multithreaded the *optimization* part of our application (the most time consuming), and second we attempted to multithread the *simulation* part.

In the *optimization* part of our application we have easily multithreaded two nested loops: the first prepares data and the second computes the Bellman values (see section 2). However, only the second has a significant execution time and leads to an efficient multithreaded parallelization. A computing loop in the routing plan execution, packing some data to prepare messages, could be parallelized too. But, it would lead to seriously more complex code while this loop is only $0.15 - 0.20\%$ of the execution time on a 256 dual-core PC cluster and on several thousand nodes of a Blue Gene/P. So, we have not multithreaded this loop.

In the *simulation* part each node processes some independent Monte-Carlo trajectories, and parallelization with multithreading has to be achieved while testing the commands in the algorithm 2. But this application part is not bounded by the amount of computations, but by the amount of data to get back from other nodes and to store in the node memory, because each MC trajectory follows an unpredictable path and requires a specific *shadow region*. So, the impact of multithreading will be limited on the *simulation* part until we improve this part (see section 6).

4.4 Serial optimizations

Beyond the parallel aspects the serial optimization is a critical point to tackle the current and coming processor complexity as well as to exploit the entirely capabilities of the compilers. Three types of serial optimization were carried out to match the processor architecture and to simplify the language complexity, in order to help the compiler to generate the best binary:

1. Substitution or coupling of the main computing parts including blitz++ classes by standard C operations or basic C functions.

2. Loops unrolling with backward technique to ease SIMD or SSE (Streaming SIMD Extension for x86 processor architecture) instructions generation and optimization by the compiler while reducing the number of branches.

3. Moving local data allocations outside the parallel multithreaded sections, to minimize memory fragmentation, to reduce C++ constructor/destructor classes overhead and to control data alignment (to optimize memory bandwidth depending on the memory architecture).

Most of the data are stored and computed within `blitz++` classes. The `blitz++` streamlines the overall implementation by providing arrays operations whatever the data type. Overloading operator is one of the main inhibitor for the compilers to generate an optimal binary. To get round this inhibitor the operations including blitz classes were replaced by standard C pointers and C operations for the most time consuming routines. C pointers and operators of code C are very simple to couple with `blitz++` arrays, and whatever the processor architecture we have got a significant speedup greater than a factor 3 with this technique. See [Vezolle et al. (2009)] for more details about these optimizations.

With the current and future processors it is compulsory to generate vector instructions to reach a good ratio of the serial peak performance. $30 - 40\%$ of the total elapsed time of our software is spent in `while` loops including a `break` test. For a medium case the minimum number of iterations is around 100. A simple look at the assembler code shows that, whatever the level of

the compiler optimization, the structure of the loop and the break test do not allow to unroll techniques and therefore to generate vector instructions. So, we have explicitly loop unrolled these `while`-and-`break` loops, with extra post-computing iterations then unrolling back to get the break point. This method enables vector instructions while reducing the number of branches.

In the shared memory parallel implementation (with Intel TBB library or OpenMP directives) each thread independently allocates local `blitz++` classes (arrays or vectors). The memory allocations are requested concurrently in the heap zone and can generate memory fragmentation as well as potential bank conflicts. In order to reduce the overhead due to memory management between the threads the main local arrays were moved outside the parallel session and indexed per the thread numbers. This optimization decreases the number of memory allocations while allowing a better control of the array alignment between the threads. Moreover, a singleton C++ class was added to `blitz++` library to synchronize the thread memory constructors/destructors and therefore minimize memory fragmentation (this feature can be deactivated depending on the operating system).

5. Experimental performances

5.1 User case introduction
We consider the situation of a power utility that has to satisfy customer load, using the power plants and one reservoir to manage. The utility equally disposes of a trading entity being able to take positions on both the spot market and futures market. We do neither consider the market complete, nor that market-depth is infinite.

5.1.1 Load and price model
The price model will be a two factor model [Clewlow & Strickland (2000)] driven by two brownian motions, and we will use a one factor model for load. In this modelization, the price future $\tilde{F}(t, T)$ corresponding to the price of one MWh seen at date t for delivery at date T evolves around an initial forward curve $\tilde{F}(T_0, T)$ and the load $D(t)$ corresponding to the demand at date t randomly evolves around an average load $D_0(t)$ depending on time. The following SDE describes our uncertainty model for the forward curve $\tilde{F}(t, T)$:

$$\frac{d\tilde{F}(t, T)}{\tilde{F}(t, T)} = \sigma_S(t)e^{-a_S(T-t)}dz_t^S + \sigma_L(t)dz_t^L, \tag{3}$$

with z_t^S and z_t^L two brownian motions, σ_i some volatility parameters.

With the following notations:

$$
\begin{aligned}
V(t1, t2, t3) &= \int_{t_1}^{t_2} \sigma_S(u)^2 e^{-2a_S(t_3-u)} + \sigma_L(u)^2 + 2\rho\sigma_S(u)e^{-a_S(t_3-u)}\sigma_L(u)du, \\
W_S(t_0, t) &= \int_{t_0}^{t} \sigma_S(u)e^{-a_S(t-u)}dz_u^S, \\
W_L(t_0, t) &= \int_{t_0}^{t} \sigma_L(u)dz_u^L,
\end{aligned}
\tag{4}
$$

the integration of the previous equation gives:

$$\tilde{F}(t, T) = \tilde{F}(t_0, T)e^{-\frac{1}{2}V(t_0, t, T) + e^{a_S(T-t)}W_S(t_0, t) + W_L(t_0, t)}. \tag{5}$$

Noting z_t^D a third brownian motion correlated to z_t^S and z_t^L , σ_D the volatility, and noting

$$
\begin{aligned}
V_D(t1, t2) &= \int_{t_1}^{t_2} \sigma_D(u)^2 e^{-2a_D(t_2-u)} du, \\
W_D(t_0, t) &= \int_{t_0}^{t} \sigma_D(u) e^{-a_D(t-u)} dz_u^D,
\end{aligned}
\tag{6}
$$

the load curve follows the following equation:

$$
D(t) = D_0(t) e^{-\frac{1}{2} V_D(t_0, t) + W_D(t_0, t)}.
\tag{7}
$$

With this modelization, the spot price is defined as the limit of the future price:

$$
S(t) = \lim_{T \downarrow t} \tilde{F}(t, T)
\tag{8}
$$

The dynamic of a financial product p for a delivery period of one month $[t_b(p), t_e(p)]$ can be approximated by:

$$
\frac{dF(t, p)}{F(t, p)} = \tilde{\sigma}_S(t, p) e^{-a_S(t_b(p)-t)} dz_t^S + \sigma_L(t) dz_t^L,
\tag{9}
$$

where:

$$
\tilde{\sigma}_S(t, p) = \sigma_S(t) \frac{\sum\limits_{t_i \in [t_b(p), t_e(p)]} e^{-a_S(t_i - t_b(p))}}{\sum\limits_{t_i \in [t_b(p), t_e(p)]} 1}
\tag{10}
$$

5.1.2 Test case

We first introduce some notation for our market products:

$\mathcal{P}(t) = \{p : t < t_b(p)\}$ all futures with delivery after t,

$L(t, p) = \{\tau : \tau < t, p \in \mathcal{P}(\tau)\}$ all time steps τ before t for which the futures product p is available on the market,

$\mathcal{P}^t = \{p : t \in [t_b(p), t_e(p)]\}$ all products in delivery at t,

$\mathcal{P} = \cup_{t \in [0,T]} \mathcal{P}(t)$ all futures products considered.

Now we can write the problem to be solved:

$$
\min \quad \mathbb{E}\left(\sum_{t=0}^{T} [\sum_{i=1}^{npal} c_{i,t} u_{i,t} - v_t S_t + \sum_{p \in \mathcal{P}(t)} (t_e(p) - t_b(p))(q(t,p)F(t,p) + |q(t,p)| \mathcal{B}_t)] \right)
$$

$$
s.t. \quad D_t = \sum_{i=1}^{npal} u_{i,t} - v_t + w_t + \sum_{p \in \mathcal{P}^t} \sum_{s \in L(t,p)} q(s, p)
$$

$$
R_{t+1} = R_t + \Delta t(-w_t + A_t)
\tag{11}
$$

$$
R_{min} \le R_t \le R_{max}
$$

$$
q_{p,min} \le q(s, p) \le q_{p,max} \quad \forall s \in [0, T] \ \forall p \in \mathcal{P}
$$

$$
y_{p,min} \le \sum_{s=0}^{\tau} q(s, p) \le y_{p,max} \quad \forall \tau < t_b(p) \ \forall p \in \mathcal{P}
$$

$$
v_{t,min} \le v_t \le v_{t,max}
$$

$$
0 \le u_{i,t} \le u_{i,t,max},
\tag{12}
$$

where

- D_t is the customer Load at time t in MW

- $u_{i,t}$ is the production of unit i at time t in MW

- v_t is spot transactions in MW (counted positive for sales)

- $q(t, p)$ is the power of the futures product p bought at time t in MW

- $\mathcal{B}_t$ is the spread bid-ask in euros/MWh taking into account the illiquidity of the market: its double value is the price gap purchase/sale of one MWh

- R_t is the level of the reservoir at time t in MWh

- $S_t = F(t, t)$ is the spot price in euros/Mwh

- $F(t, p)$ is the futures price of the product p at time t in euros/MWh

- w_t is the production of the reservoir at time t in MW

- A_t are the reservoir inflows in MW

- Δt the time step in hours

- $q_{p,min}$, $q_{p,max}$ are bounds on what can be bought and sold per time step on the futures market in MW

- $y_{p,min}$, $y_{p,max}$ are the bounds on the size of the portfolio for futures product p

- R_{min}, R_{max} are (natural) bounds on the energy the reservoir can contain.

Some additional values for the initial stocks are also given, and some final values are set for the reservoir stock remaining at date T.

5.1.3 Numerical data

We consider at the begin of a month a four months optimization, where the operator can take position in the future market twice a month using month ahead futures peak and offpeak, two month ahead futures peak and off peak, and three month ahead futures base and peak. So the user has at date 0 6 future products at disposal. The number of trajectories for optimization is 400. The depth of the market for the 6 future products is set to 2000 MW for purchase and sales ($y_{p,min} = -2000$, $y_{p,max} = 2000$). Every two weeks, the company is allowed to change its position in the futures market within the limits of 1000 MW ($q_{p,min} = -1000$, $q_{p,max} = 1000$). All the commands for the futures stocks are tested from -1000 MW to 1000 MW with a step of 1000 MW. The hydraulic command is tested with a step of 1000MW. All the stocks are discretized with a 1000MW step leading to a maximum of $225 * 5^6$ points to explore for the stock state variables. The maximum number of commands tested is $5 * 3^6$ at day 30 for each point stock not saturating the constraints. This discretization is a very accurate one leading to a huge problem to solve. Notice that the number of stocks is decreasing with time. After two months, the two first future delivery periods are past so the problem becomes a 5 stocks problem. After three months , we are left with a three stocks problems and no command to test (delivery of the two last future contracts has begun). The global problem is solved with 6 steps per days, defining the reservoir strategy, and the future commands are tested every two weeks.
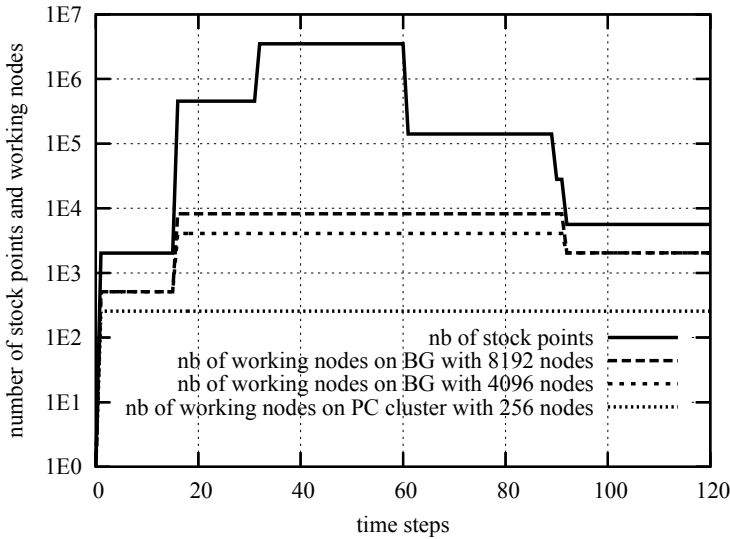
Fig. 7. Evolution of the number of stock points (problem size) and of the number of working nodes (useful size of the machine)

5.2 Testbeds introduction

We used two different parallel machines to test our application and measure its performances: a PC cluster and a supercomputer.

- Our PC cluster was a 256-node cluster of SUPELEC (from CARRI Systems company) with a total of 512 cores. Each node hosts one dual-core processor: INTEL Xeon-3075 at 2.66 GHz, with a front side bus at 1333 MHz. The two cores of each processor share 4 GB of RAM, and the interconnection network is a Gigabit Ethernet network built around a large and fast CISCO 6509 switch.

- Our supercomputer was the IBM Blue Gene/P supercomputer of EDF R&D. It provides up to 8192 nodes and a total of 32768 cores, which communicate through proprietary high-speed networks. Each node hosts one quad-core PowerPC 450 processor at 850 MHz, and the 4 cores share 2 GB of RAM.

5.3 Experimental provisioning of the computing nodes

Figure 7 shows the evolution of the number of stock points of our benchmark application, and the evolution of the number of available nodes that have some work to achieve: the number of provisioned nodes. The number of stock points defines the problem size. It can evolve at each time step of the *optimization* part and the splitting algorithm that distributes the N-cube data and the associated work has to be run at the beginning of each time step (see section 3.1). This algorithm determines the number of available nodes to use at the current time step. The number of stock points of this benchmark increases up to 3 515 625, and we can see on figure 7 the evolution of their distribution on a 256-nodes PC cluster, and on 4096 and 8192 nodes of a Blue Gene supercomputer. Excepted at time step 0 that has only one stock point, it has been possible to use the 256 nodes of our PC cluster at each time step. But it has not been

possible to achieve this efficiency on the Blue Gene. We succeeded to use up to 8192 nodes of
this architecture, but sometimes we used only 2048 or 512 nodes.

However, section 5.4 will introduce the good scalability achieved by the *optimization* part of
our application, both on our 256-nodes PC cluster and our 8192-nodes Blue Gene. In fact, time
steps with small numbers of stock points are not the most time consuming. They do not make
up a significant part of the execution time, and to use a limited number of nodes to process
these time steps does not limit the performances. But it is critical to be able to use a large
number of nodes to process time steps with a great amount of stock points. This dynamic
load balancing and adaptation of the number of working nodes is achieved by our *splitting
algorithm*, as illustrated by figure 7.

Section 3.4 introduces our splitting strategy, aiming to create and distribute *cubic* subcubes and
avoiding *flat* ones. When the backward loop of the *optimization* part leaves step 61 and enters
step 60 the cube of stock points increases a lot (from 140 625 to 3 515 625 stock points) because
dimensions two and five enlarge from 1 to 5 stock levels. In both steps the cube is split in 8192
subcubes, but this division evolves to take advantage of the enlargement of dimensions two
and five. The following equations resume this evolution:

$$step\ 61 : 140625\ stock\ points \quad = \quad 225 \times 1 \times 5 \times 5 \times 1 \times 5 \times 5\ stock\ points \tag{13}$$

$$step\ 60 : 3515625\ stock\ points \quad = \quad 225 \times 5 \times 5 \times 5 \times 5 \times 5 \times 5\ stock\ points \tag{14}$$

$$step\ 61 : 8192\ subcubes \quad = \quad 128 \times 1 \times 4 \times 4 \times 1 \times 2 \times 2\ subcubes \tag{15}$$

$$step\ 60 : 8192\ subcubes \quad = \quad 128 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2\ subcubes \tag{16}$$

$$subcube\ sizes \quad = \quad \begin{bmatrix} min\ nb\ of\ stock\ levels \\ max\ nb\ of\ stock\ levels \end{bmatrix}_{dim1} \times \begin{bmatrix} \ \ \end{bmatrix}_{dim2} \ ...$$

$$step\ 61 : subcube\ sizes \quad = \quad \begin{bmatrix} 1 \\ 2 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} \times \begin{bmatrix} 2 \\ 3 \end{bmatrix} \times \begin{bmatrix} 2 \\ 3 \end{bmatrix} \tag{17}$$

$$step\ 60 : subcube\ sizes \quad = \quad \begin{bmatrix} 1 \\ 2 \end{bmatrix} \times \begin{bmatrix} 2 \\ 3 \end{bmatrix} \times \begin{bmatrix} 2 \\ 3 \end{bmatrix} \times \begin{bmatrix} 2 \\ 3 \end{bmatrix} \times \begin{bmatrix} 2 \\ 3 \end{bmatrix} \times \begin{bmatrix} 2 \\ 3 \end{bmatrix} \times \begin{bmatrix} 2 \\ 3 \end{bmatrix} \tag{18}$$

At time step 61, equations 13 and 15 show the dimension one has a stock level range of size 225
split in 128 subranges. This leads to subcubes with 1 (min) or 2 (max) stock levels in dimension
one on the different nodes, as summarized by equation 17. Similarly, the dimension two has
a stock level range of size 1 split in 1 subrange of size 1, the dimension three has a stock level
range of 5 split in 4 subranges of size 1 or 2... At time step 60, equations 14 and 16 show the
range of dimensions two and five enlarge from 1 to 5 stock levels and their division increases
from 1 to 2 subparts, while the division of dimensions three and four decreases from 4 to
2 subparts. Finally, equation 18 shows the 8192 subcubes are more *cubic*: they have similar
minimal and maximal sizes in their last six dimensions and only their first dimension can
have a smaller size. This kind of data re-distribution can happen each time the global N-cube
of data evolves, even if the number of provisioned nodes remains unchanged, in order to
optimize the computation load balancing and the communication amount.

5.4 Performances function of deployment and optimization mechanisms

Figure 8 shows the different total execution times on the two testbeds introduced in section
5.2 for the following parallelizations:

Design and experimentation of a large scale distributed
stochastic control algorithm applied to energy management problems

121
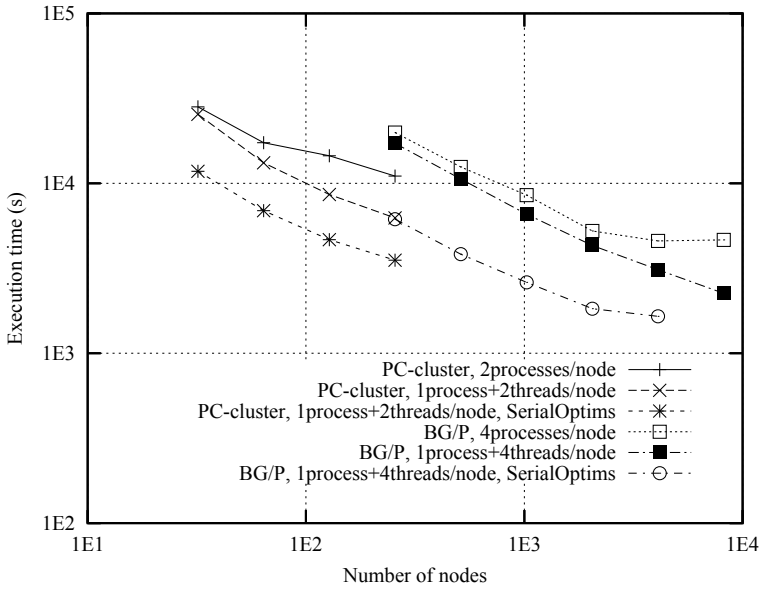
Fig. 8. Total execution times function of the deployment and optimization mechanisms

- implementing no serial optimization and using no thread but running several MPI processes per node (one MPI process per core),

- implementing no serial optimization but using multithreading (one MPI process per node and one thread per core),

- implementing serial optimizations and multithreading (one MPI process per node and one thread per core).

Without multithreading the execution time decreases slowly on the PC-cluster or reaches an asymptote on the Blue Gene/P. When using multithreading the execution time is smaller and decreases regularly up to 256 nodes and 512 cores on PC cluster, and up to 8192 nodes and 32768 cores on Blue Gene/P. So, the *deployment* strategy has a large impact on performances of our application. Performance curves of figure 8 show we have to deploy only one MPI process per node and to run threads to efficiently use the different cores of each node. The multithreading development introduced in section 4.3 has been easy to achieve (parallelizing only some nested loops), and has reduced the execution time and extended the scalability of the application.

These results confirm some previous experiments achieved on our PC cluster and on the Blue Gene/L of EDF without serial optimizations. Multithreading was not available on the Blue Gene/L. Using all cores of each nodes decreased the execution time but did not allowed to reach a good scalability on our Blue Gene/L [(Vialle et al., 2008)].

Serial optimizations introduced in section 4.4 have also an important impact on the performances. We can see on figure 8 they divide the execution time by a factor 1.63 to 2.14 on the PC cluster of SUPELEC, and by a factor 1.88 to 2.79 on the Blue Gene/P supercomputer of EDF (depending on the number of used nodes). Moreover, they lead to reach the scalability limit of our distributed application: the execution time decreases but reaches a new asymptote
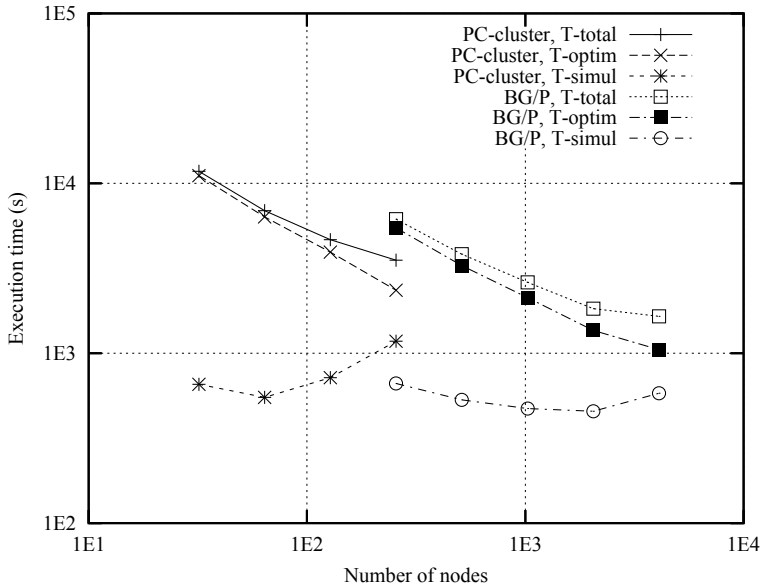
Fig. 9. Details of the best execution times of the application

when using 4096 nodes and 16384 cores on our Blue Gene/P. We can not speedup more this benchmark application with our current algorithms and implementation.

These experiments have allowed to identify the right deployment strategy (running one MPI process per node and multithreading) and the right implementation (using all our serial optimizations). We analyze our best performances in the next section.

5.5 Detailed best performances of the application and its subparts

Figure 9 shows the details of the best execution times (using multithreading and implementing serial optimizations). First, we can observe the *optimization* part of our application scales while the *simulation* part does not speedup and limits the global performances and scaling of the application. So, our N-cube distribution strategy, our *shadow region map* and *routing plan* computations, and our *routing plan* executions appear to be efficient and not to penalize the speedup of the *optimization part*. But our distribution strategy of Monte carlo trajectories in the *simulation part* does not speedup, and limits the performances of the entire application.

Second, we observe on figure 9 our distributed and parallel algorithm, serial optimizations and portable implementation allow to run our complete application on a 7-stocks and 10-state-variables in less than 1*h* on our PC-cluster with 256 nodes and 512 cores, and in less than 30*mn* on our Blue Gene/P supercomputer used with 4096 nodes and 16384 cores. These performances allow to plan some computations we could not run before.

Finally, considering some real and industrial use cases, with bigger data set, the *optimization* part will increase more than the *simulation* part, and our implementation should scale both on our PC cluster and our Blue Gene/P. Our current distributed and parallel implementation is operational to process many of our real problems.

6. Conclusion and perspectives

Our parallel algorithm, serial optimizations and portable implementation allow to run our complete application on a 7-stocks and 10-state-variables in less than $1h$ on our PC-cluster with 256 nodes and 512 cores, and in less than $30mn$ on our Blue Gene/P supercomputer used with 4096 nodes and 16384 cores. On both testbeds, the interest of multithreading and serial optimizations have been measured and emphasized. Then, a detailed analysis has shown the *optimization* part scales while the *simulation part* reaches its limits. These current performances promise high performances for future industrial use cases where the *optimization* part will increase (achieving more computations in one time step) and will become a more significant part of the application.

However, for some high dimension problems, the communications during the *simulation* part could become predominant. We plan to modify this part by reorganizing trajectories so that trajectories with similar stocks levels are treated by the same processor. This will allow us to identify and to bring back the *shadow region* only once per processor at each time step and to decrease the number of communication needed.

Previously our paradigm has been successfully tested too on a smaller case for gaz storage [Makassikis et al. (2008)]. Currently it is used to valuate power plants facing the market prices and for different problems of asset liability management. In order to make easier the development of new stochastic control applications, we aim to develop a generic library to rapidly and efficiently distribute N dimensional cubes of data on large size architectures.

Acknowledgment

7. References

Bacaud, L., Lemarechal, C., Renaud, A. & Sagastizabal, C. (2001). Bundle methods in stochastic optimal power management: A disaggregated approach using preconditioner, *Computational Optimization and Applications* **20**(3).

Bally, V., Pagès, G. & Printems, J. (2005). A quantization method for pricing and hedging multi-dimensional american style options, *Mathematical Finance* **15**(1).

Bellman, R. E. (1957). *Dynamic Programming*, Princeton University Press, Princeton.

Bouchard, B., Ekeland, I. & Touzi, N. (2004). On the malliavin approach to monte carlo approximation of conditional expectations, *Finance and Stochastics* **8**(1): 45–71.

Charalambous, C., Djouadi, S. & Denic, S. Z. (2005). Stochastic power control for wireless networks via sde's: Probabilistic qos measures, *IEEE Transactions on Information Theory* **51**(2): 4396–4401.

Chen, Z. & Forsyth, P. (2009). Implications of a regime switching model on natural gas storage valuation and optimal operation, *Quantitative Finance* **10**: 159–176.

Clewlow, L. & Strickland, C. (2000). *Energy derivatives: Pricing and risk management*, Lacima.

Culioli, J. C. & Cohen, G. (1990). Decomposition-coordination algorithms in stochastic optimization, *SIAM Journal of Control and Optimization* **28**(6).

Heitsch, H. & Romisch, W. (2003). Scenario reduction algorithms in stochastic programming, *Computational Optimization and Applications* **24**.

Hull, J. (2008). *Options, Futures, and Other Derivatives, 7th Economy Edition*, Prentice Hall.

Longstaff, F. & Schwartz, E. (2001). Valuing american options by simulation: A simple least-squares, *Review of Financial Studies* **14**(1).

Ludkovski, M. & Carmona, R. (2010, to appear). Gas storage and supply: An optimal switching approach, *Quantitative Finance* .

Makassikis, C., Vialle, S. & Warin, X. (2008). Large scale distribution of stochastic control algorithms for financial applications, *The First International Workshop on Parallel and Distributed Computing in Finance (PdCoF08)*, Miami, USA.

Porchet, A., Touzi, N. & Warin, X. (2009). Valuation of a powerplant under production constraints and markets incompleteness, *Mathematical Methods of Operations research* **70**(1): 47–75.

Rotting, T. A. & Gjelsvik, A. (1992). Stochastic dual dynamic programming for seasonal scheduling in the norwegian power system, *Transactions on power system* **7**(1).

Veldhuizen, T. (2001). Blitz++ User's Guide, Version 1.2, http://www.oonumerics.org/blitz/manual/blitz.html.

Vezolle, P., Vialle, S. & Warin, X. (2009). Large scale experiment and optimization of a distributed stochastic control algorithm. application to energy management problems, *Workshop on Large-Scale Parallel Processing (LSPP 2009)*, Roma, Italy.

Vialle, S., Warin, X. & Mercier, P. (2008). A N-dimensional stochastic control algorithm for electricity asset management on PC cluster and Blue Gene supercomputer, *9th International Workshop on State-of-the-Art in Scientific and Parallel Computing (PARA08)*, NTNU, Trondheim, Norway.

Exploring Statistical Processes with *Mathematica7*

Fred Spiring
The University of Manitoba
CANADA

1. Introduction

Methods for estimating, assessing and monitoring processes are illustrated using the software package *Mathematica7* (Wolfram (2009)). Graphical techniques that allow the dynamic assessment of underlying distributional properties as well as capabilities are presented and illustrated. In addition, innovative procedures associated with compositional data in the L^3 space are examined and expanded to the L^1 constrained space for two variables and the L^2 space for three variables. Several new conventions are proposed that attempt to provide insights into a variety of processes, all with diagnostic tools useful for, but not limited to the manufacturing sector. Several estimation and inferential techniques are presented with tools for determining associated estimates and the resulting inferences. The manuscript is accompanied by a *Mathematica7* notebook best viewed using *Mathematica7* or *Mathematica7 Player*. *Mathematica7 Player* is a free download available at www.Wolfram.com/products/player/ that allows all features of the notebook to be viewed.

2. Creating Probability Plots

Probability plots are graphical expressions used in examining data structures. Plots provide insights into the suitability of a particular probability density function (pdf) in describing the stochastic behavior of the data and estimates of the unknown parameters of the pdf. Although generally very powerful, the inferences drawn from probability plots are subjective.

The underlying principle behind probability plots is simple and consistent. The order statistics, with $Y_{[i]}$ denoting the ith largest observation, such that

$$Y_{[1]} \leq Y_{[2]} \leq ... \leq Y_{[i]} \leq ... \leq Y_{[n]}$$

are plotted versus their expected values $E(Y_{[i]})$. A linear relationship between the order statistics and their expected values indicates the pdf used in determining the expected values provides a reasonable representation of the behavior of the observed data. A non-linear plot suggests that other pdf(s) may be more suitable in describing the stochastic structure of the data.

The expected value of the ith order statistic is

$$E(Y_{[i]}) = n!/[(i-1)!(n-i)!] \int_0^1 Y_{[i]} \left[F\left(y_{[i]}\right) \right]^{(i-1)} \left[1 - F\left(y_{[i]}\right) \right]^{(n-i)} dF(y_{[i]})$$

where f(y) denotes the pdf being considered, F(y) the associated cumulative distribution function (cdf) and n the size of the dataset under investigation. Because numerical solutions for this equation can be difficult, the approximation $E(Y_{[i]}) = F^{-1}[(i-c)/(n-2c-1)]$, where F^{-1} denotes the inverse cdf and c a constant $(0 \leq c \leq 1)$ is frequently used. Setting c=0.5 (for discussion see Kimball (1960)) results in

$$E(Y_{[i]}) = F^{-1}[(i-0.5)/n]$$

and is the approximation used here. *Mathematica* will be used to evaluate the $E(Y_{[i]})$, create the resulting probability plot, assist in assessing linearity and determine parameter estimates.

Mathematica's Quantile functions are used to find the $E(Y_{[i]})$'s for specific pdfs and create the plot of $Y_{[i]}$ versus $E(Y_{[i]})$. If the resulting plot is considered linear then the pdf used to determine the $E(Y_{[i]})$'s can be used to describe the stochastic structure of the data. Assuming the plot is deemed linear, estimates for the unknown parameters can be determined from the plot.

```
y = {163, 174, 174, 175, 176, 176, 179, 181, 181, 183, 183,
   186, 188, 188, 189, 190, 190, 191, 192, 192, 195, 195,
   197, 197, 208};
pdfs = NormalDistribution[0, 1];
EYpairs[y_, pdfs_]:=
   With[{n = Length[y]},
   Transpose[{Map[Quantile[pdfs, #]&, (Range[n] − 0.5)/n],
   Sort[y]}]]
linePlot:=
   Plot[Evaluate[Fit[EYpairs[y, pdfs], {1, x}, x]],
   {x, −3.0, 3.0}, DisplayFunction->Identity];
r2:=LinearModelFit[EYpairs[y, pdfs], {1, x}, x]
probabilityPlot[yList_, pdfList_]:=
   ListPlot[EYpairs[y, pdfs],
      FrameLabel-> {"E(Y[i]))", "Y[i]"},
      RotateLabel->False,
      PlotRange->{{−3, 3}, Automatic},
      Frame->True,
      GridLines->{{−1, 0, 1}, Automatic},
      DisplayFunction->Identity];
Show[probabilityPlot[yList, pdfList], Prolog->AbsolutePointSize[4]]
```
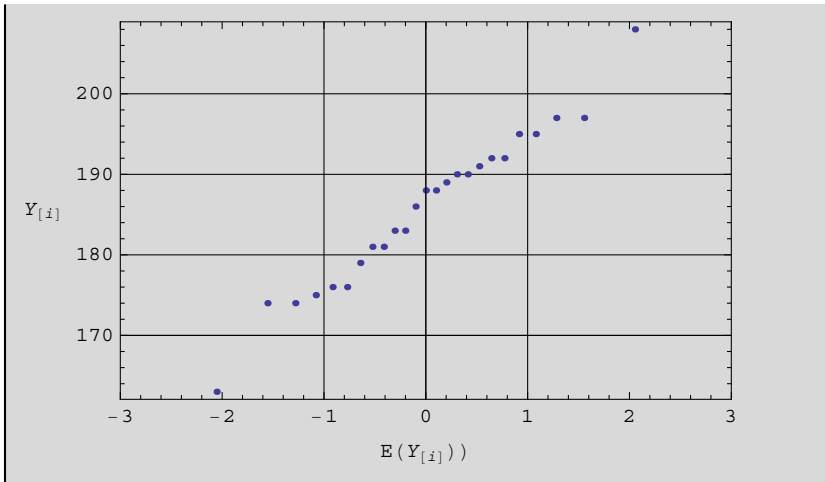
For the normal family of density functions, $f(y) = (2\pi\sigma^2)^{-1/2}exp(-(y-\mu)^2/(2\sigma^2)), -\infty < y < \infty$, the standard pdf is identified in the routine by NormalDistribution[0,1] and the data denoted by y. If the resulting normal probability plot is considered linear, then the intersection of the plot with the $E(Y_{[i]}) = 0$ asymptote provides an estimate for the location parameter μ (in this case 187) and the slope provides an estimate for the scale parameter σ. Using the plot's intersection points with the vertical asymptotes ± 1 and dividing by 2 results in an estimate, in this case, of 10 for σ.

The addition of a least squares line and the resulting coefficient of determination (R^2) provide insights into the linearity of the probability plot. The least squares line provides visual assistance in assessing the linearity, while R^2 provides numerical assessment (as R^2 increases, the more linear the probability plot). The least squares line and R^2 are included in subsequent plots.

```
y = {163, 174, 174, 175, 176, 176, 179, 181, 181, 183, 183,
    186, 188, 188, 189, 190, 190, 191, 192, 192, 195, 195,
    197, 197, 208};
pdfs = NormalDistribution[0, 1];
EYpairs[y_, pdfs_]:=
    With[{n = Length[y]},
    Transpose[{Map[Quantile[pdfs, #]&, (Range[n] − 0.5)/n],
    Sort[y]}]]
linePlot:=
    Plot[Evaluate[Fit[EYpairs[y, pdfs], {1, x}, x]],
    {x, −3.0, 3.0}, DisplayFunction->Identity];
r2:=LinearModelFit[EYpairs[y, pdfs], {1, x}, x]
probabilityPlot[yList_, pdfList_]:=
    ListPlot[EYpairs[y, pdfs],
        FrameLabel-> { "E(Y_{[i]}))", "Y_{[i]}", "= R^2"r2["RSquared"]},
        RotateLabel->False,
```

```
      PlotRange->{{−3, 3}, Automatic},
      Frame->True,
      GridLines->{{−1, 0, 1}, Automatic},
      DisplayFunction->Identity];
Show[probabilityPlot[yList, pdfList], linePlot,
Prolog->AbsolutePointSize[4]]
```


If the resulting probability plot is not considered linear then alternative pdfs may be considered. Simply changing the pdf used in determining the $E(Y_{[i]})$'s will allow different stochastic structures to be examined. Replacing NormalDistribution[0,1] with ExponentialDistribution[1] in the routine determines the $E(Y_{[i]})$'s for the pdf f(y) = (1/θ)exp[-y/θ], 0<y<∞. Altering the PlotRange and position of the asymptotes results in an exponential probability plot.

```
pdfs = ExponentialDistribution[1];
EYpairs[y_, pdfs_]:=
   With[{n = Length[y]},
   Transpose[{Map[Quantile[pdfs, #]&, (Range[n] − 0.5)/n],
   Sort[y]}]]
linePlot:=
   Plot[Evaluate[Fit[EYpairs[y, pdfs], {1, x}, x]],
   {x, 0, 4.0}, DisplayFunction->Identity];
r2:=LinearModelFit[EYpairs[y, pdfs], {1, x}, x]
probabilityPlot[yList_, pdfList_]:=
   ListPlot[EYpairs[y, pdfs],
      FrameLabel-> { "E(Y_{[i]}))", "Y_{[i]} ", "= R^2 "r2["RSquared"]} ,
      RotateLabel->False,
      PlotRange->{{0, 4}, Automatic},
```

```
    Frame->True,
    GridLines->{{Quantile[pdfs, .1],
    Quantile[pdfs, .8]}, Automatic},
    DisplayFunction->Identity];
Show[probabilityPlot[yList, pdfList], linePlot,
Prolog->AbsolutePointSize[4]]
```
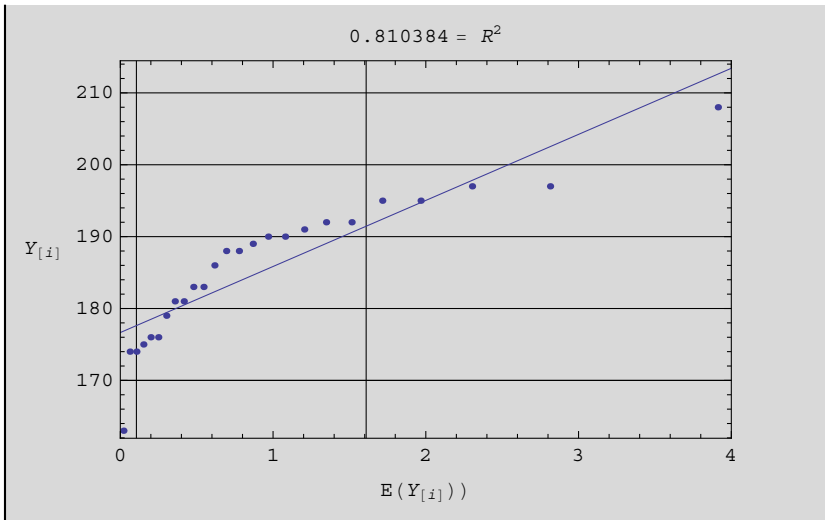

In the case of the standard exponential distribution, if the resulting probability plot is considered linear then an estimate of θ can be determined as $1.504/(y_{.8} - y_{.1})$ (Shapiro (1980)), where $y_{.8}$ is the 80th percentile and $y_{.1}$ is the 10th percentile of the distribution. Vertical asymptotes have been included at the 10th and 80th percentiles to facilitate determining the points of intersection with these asymptotes.

Creating a probability plot for the standard uniform distribution, $f(y) = 1/\theta$, $-\theta/2 \leq y \leq \theta/2$ requires changing the routine to UniformDistribution[0, 1]. In addition the PlotRange is altered to (0, 1) and asymptotes added at .25, .5, .75 . If the resulting uniform probability plot is considered linear, then an estimate of θ is determined using the plot's intersection with the 25th and 75th percentiles (i.e., $y_{.25}, y_{.75}$) as follows $(y_{.75} - y_{.25})/.5$.

```
pdfs = UniformDistribution[{0, 1}];
EYpairs[y_, pdfs_]:=
    With[{n = Length[y]},
    Transpose[{Map[Quantile[pdfs, #]&, (Range[n] − 0.5)/n],
    Sort[y]}]]
linePlot:=
    Plot[Evaluate[Fit[EYpairs[y, pdfs], {1, x}, x]],
    {x, 0, 1.0}, DisplayFunction->Identity];
r2:=LinearModelFit[EYpairs[y, pdfs], {1, x}, x]
```

probabilityPlot[yList_, pdfList_]:=
   ListPlot[EYpairs[y, pdfs],
      FrameLabel-> $\left\{\right.$ **"E($Y_{[i]}$))", "$Y_{[i]}$", "= R^2"r2["RSquared"]** $\left.\right\}$ **,**
      RotateLabel->False,
      PlotRange->{{0, 1}, Automatic},
      Frame->True,
      GridLines->{{.25, .5, .75}, Automatic},
      DisplayFunction->Identity];
Show[probabilityPlot[yList, pdfList], linePlot, Prolog->AbsolutePointSize[4]]
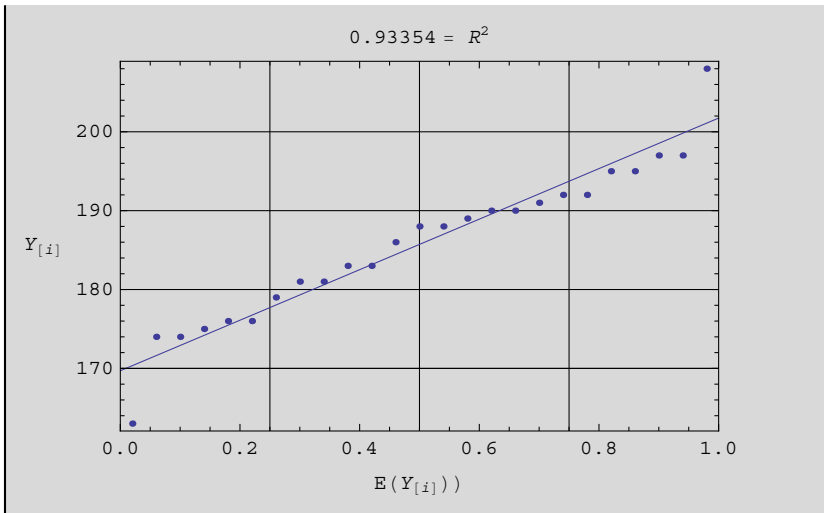
Other pdfs can be examined by changing the distribution function specified in the routine. Probability plots for the LogNormalDistribution[0,1] and WeibullDistribution[1, 3.25] distributions are illustrated.

pdfs = LogNormalDistribution[0, 1];
EYpairs[y_, pdfs_]:=
   With[{n = Length[y]},
   Transpose[{Map[Quantile[pdfs, #]&, (Range[n] − 0.5)/n],
   Sort[y]}]]
linePlot:=
   Plot[Evaluate[Fit[EYpairs[y, pdfs], {1, x}, x]],
   {x, 0, 8}, DisplayFunction->Identity];
r2:=LinearModelFit[EYpairs[y, pdfs], {1, x}, x]
probabilityPlot[yList_, pdfList_]:=
   ListPlot[EYpairs[y, pdfs],
      FrameLabel-> $\left\{\right.$ **"E($Y_{[i]}$))", "$Y_{[i]}$", "= R^2"r2["RSquared"]** $\left.\right\}$ **,**
      RotateLabel->False,
      PlotRange->{{0, 8}, Automatic},

```
    Frame->True,
    GridLines->{{1},Automatic},
    DisplayFunction->Identity];
Show[probabilityPlot[yList,pdfList],linePlot,
Prolog->AbsolutePointSize[4]]
```
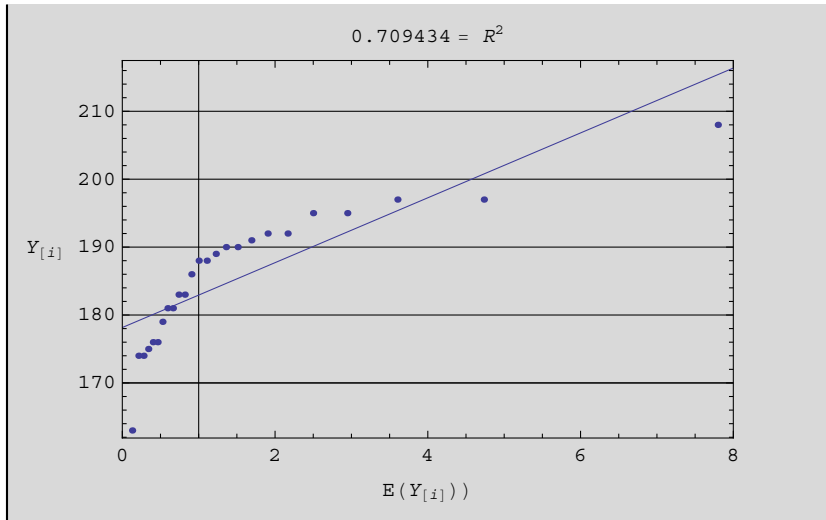


```
pdfs = WeibullDistribution[1,3.25];
EYpairs[y_,pdfs_]:=
   With[{n = Length[y]},
   Transpose[{Map[Quantile[pdfs,#]&,(Range[n] − 0.5)/n],
   Sort[y]}]]
linePlot:=
   Plot[Evaluate[Fit[EYpairs[y,pdfs],{1,x},x]],
   {x,−.5,14},DisplayFunction−>Identity];
r2:=LinearModelFit[EYpairs[y,pdfs],{1,x},x]
probabilityPlot[yList_,pdfList_]:=
   ListPlot[EYpairs[y,pdfs],
     FrameLabel−> { "E(Y[i]))", "Y[i]", "= R² "r2["RSquared"]} ,
     RotateLabel−>False,
     PlotRange−>{{−.5,14},Automatic},
     Frame−>True,
     GridLines−>{{Quantile[pdfs,.1],
     Quantile[pdfs,.9]},Automatic},
     DisplayFunction−>Identity];
Show[probabilityPlot[yList,pdfList],linePlot,Prolog−>AbsolutePointSize[4]]
```
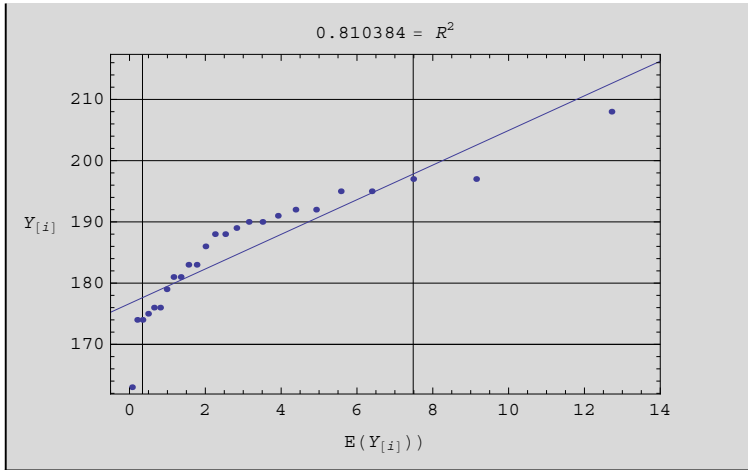
Probability plots are generally restricted to the class of pdfs characterized by location and scale parameters. The Weibull distribution of the form $f(y) = \alpha\beta y^{\beta-1} \exp\left(-\alpha y^\beta\right), \alpha, \beta > 0, 0 \le y < \infty$, is an exception in that α and β are considered scale and shape parameters. In the previous example the values of the parameters were set at 1 and 3.25 respectively. A linear relationship in a Weibull probability plot suggests that the Weibull distribution with specific parameter values is appropriate in describing the stochastic nature of the data. However a non-linear relationship does not necessarily rule out the Weibull family of distributions but may be a reflection on the value(s) of the parameters chosen.

In the case of the Weibull distribution, taking the natural logarithm twice and plotting allows the distribution (with scale and shape parameters) to be examined analogous to those distributions characterized by location and scale parameters (Hahn and Shapiro (1967)). However, in general, probability plots can assess only those distributions with no (or at least a constant) shape parameter. Cheng and Spring (1990) used rotation to illustrate techniques that extend the use of probability plots to a class of pdfs characterized by location, scale and shape parameters. Of particular interest were the Weibull and Tukey's-λ distributions as both are characterized by a location, scale and single shape parameter.

2.1 Creating & Interpreting 3-D Probability Surfaces

Dynamic graphic techniques have opened new frontiers in data display and analysis. With a basic understanding of simple probability plots, subjective interpretation of distributional assumptions can be made for families of distributions that contain a shape parameter. Strong visual results are possible for relatively small sample sizes. In the examples that follow, sample sizes of 25 provide good insights into the distributional properties of the observed data.

Let Y denote a random variable with pdf f(y; μ, σ, λ) and cdf F(y; μ, σ, λ) where μ, σ and λ denote the location, scale and shape parameters of the distribution respectively. Cheng and Spring (1990) defined the X-axis as $E(Y_{[i]}; \lambda)$, scaled the Z-axis arithmetically and defined it as the order statistics $Y_{[i]}$ and let the Y axis denote values of the shape parameter λ, to create

a surface in 3 space. Examination of the resulting surface allowed inferences regarding the stochastic nature of the data as well as estimates for location, scale and shape parameters of the associated pdf.

The resulting surface is essentially an infinite number of traditional probability plots laid side by side. These probability plots are ordered by the value of the shape parameter used in calculating the $E(Y_{[i]})$'s. Slicing the surface along planes parallel to the XZ plane at various points along the Y axis, allows viewing of the "linearity" of the surface by considering the resultant projection on the XZ plane. The projection is a univariate probability plot of the data for a particular value of the shape parameter. The goal then is to slice the surface such that the most linear projection on the XZ plane is found.

Rotation allows viewing of the created surface from several perspectives, enhancing the ability to determine where the surface appears most linear and the associated value of the shape parameter. From the most linear portion of the surface, estimates for the location, scale and shape parameters can be determined. The 50th percentile (or midpoint of the X-axis provides an estimate for the location, the value of the Y-axis where the surface is most linear provides an estimate for the shape parameter and the slope of the surface (in the X-direction) an estimate of the scale.

In practice the order statistics are plotted versus the expected value of the ordered statistics for various values of the shape parameter. Then examining various views of the surface allows one to determine the value of the shape parameter associated with the most linear portion of the curve. From there estimates for the location and scale parameters are possible.

Animation permits a series of univariate probability plots (for specific values of the shape parameter) to be viewed in a sequential fashion, highlighting changes in the probability plots resulting from changes in the shape parameter. This results in a quick and reliable method for determining the most linear portion of the surface. The procedure creates a series of univariate probability plots representing various values of the shape parameter. The observer must determine which of the plots (if any) is most linear. If the surface provides no linear results, then one concludes that the data do not arise from the family of distributions considered.

2.2 Example

Letting y denote the 25 simulated normal distribution results, $E(Y_{[i]}; \lambda)$ the expected value of the associated order statistics and λ the shape parameter, a surface in three space can be created using the following routine.

```
y = {.1, .2025, .3045, .4124, .523, .6433, .7723, .9154, 1.08, 1.282, 1.555, 2.054, −2.054,
  −1.555, −1.282, −1.08, −.9154, −.7723, −.6433, −.523, −.4124, −.3045, −.2025, −.1, 0};
n = Count[y, _]; s = Min[y]; l = Max[y]; d = 4(l − s)/n; x = Sort[y];
t[lambda_] = Table[Point[{x[[j]], ((((j − .5)/n)^lambda − (1 − ((j − .5)/n))^lambda)
  /lambda), lambda}], {j, 1, n}];
Show[Graphics3D[Table[{t[lambda]}, {lambda, 0.05, 1, .05}], Axes−>True,
  AxesLabel−>{Y, "E(Y)", " lambda"}, BoxRatios−>{2, 2, 4},
  ViewPoint−>{1, 0, −2}]]
```
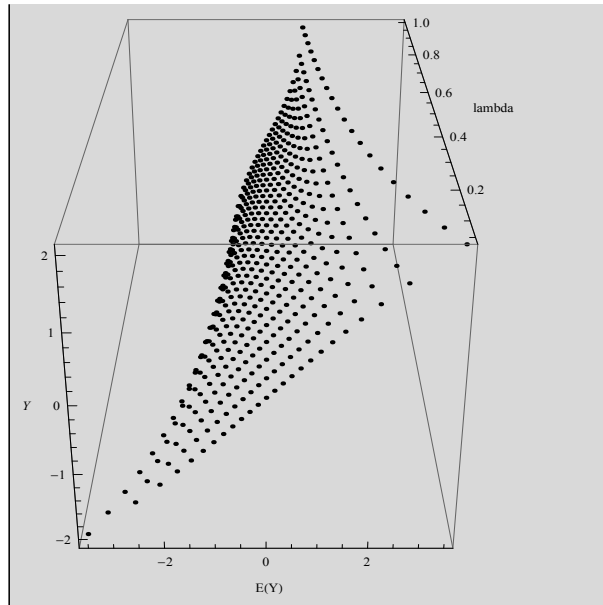
The pdf of the lambda distribution can be determined for specific values of λ, however it is generally given as the pdf of Z under the transformation $Z = ((X^\lambda) - (1 - X)^\lambda))/\lambda$ where $X \sim U[0, 1]$. The transformation is also the percentile function for the distribution and results in the expected value of the order statistics being of the form $E(Y_{[i]}; \lambda) = (((i - 0.5)/n)^\lambda - (1 - (i - 0.5)/n)^\lambda)/\lambda$ While rotation allows different views of this surface, determining the most linear portion can still be difficult. Rather than rotating, slicing and viewing the resulting plots, the following routine creates a series of probability plots enhanced with a regression line and R^2, as well as the associated value of λ that can be viewed using the animation function of *Mathematica*.

$y = \{.1, .2025, .3045, .4124, .523, .6433, .7723, .9154, 1.08, 1.282, 1.555, 2.054,$
    $-2.054, -1.555, -1.282, -1.08, -.9154, -.7723, -.6433, -.523, -.4124, -.3045,$
    $-.2025, -.1, 0\};$
  $n = \text{Length}[y];$
  $s = \text{Min}[y];$
  $t = \text{Max}[y];$
  $d = 4(t - s)/n;$
$l := \text{Text}[\text{Style}[\text{lambda } \text{"}= \lambda\text{"}], \{-3, t\}];$
$\text{pdfs}[j\_] := (((((j - .5)/n)^\wedge \text{lambda} - (1 - ((j - .5)/n))^\wedge \text{lambda})/\text{lambda});$
$\text{EYpairs}[y\_, \text{pdfs}\_] := \text{With}[\{n = \text{Length}[y]\},$
  $\text{Transpose}[\{\text{Map}[\text{pdfs}[\#]\&, \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10,$
  $11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25\}], \text{Sort}[y]\}]];$
$\text{linePlot} := \text{Plot}[\text{Evaluate}[\text{Fit}[\text{EYpairs}[y, \text{pdfs}], \{1, x\}, x]],$
  $\{x, -4.0, 4.0\}, \text{DisplayFunction->Identity}];$
$\text{r2} := \text{LinearModelFit}[\text{EYpairs}[y, \text{pdfs}], \{1, x\}, x];$
$\text{probabilityPlot}[\text{yList}\_, \text{pdfList}\_] :=$

```
   ListPlot[EYpairs[y, pdfs],
FrameLabel-> { "E(Y[i]))", "Y[i]" }, RotateLabel->False,
   PlotRange->{{−4, 4}, {s − d, t + d}}, Frame->True,
 GridLines− > {{−1, 0, 1}, Automatic},
   DisplayFunction->Identity];
rr:=Text [Style ["= R² "r2["RSquared"]] , {3, −t}] ;
SlideView[Table[Show[probabilityPlot[yList, pdfList], linePlot,
Graphics[l], Graphics[rr],
DisplayFunction->$DisplayFunction, ImageSize → Scaled[0.9],
Prolog->AbsolutePointSize[4]], {lambda, 0.05, 1.00, 0.05}],
AppearanceElements → All]
```
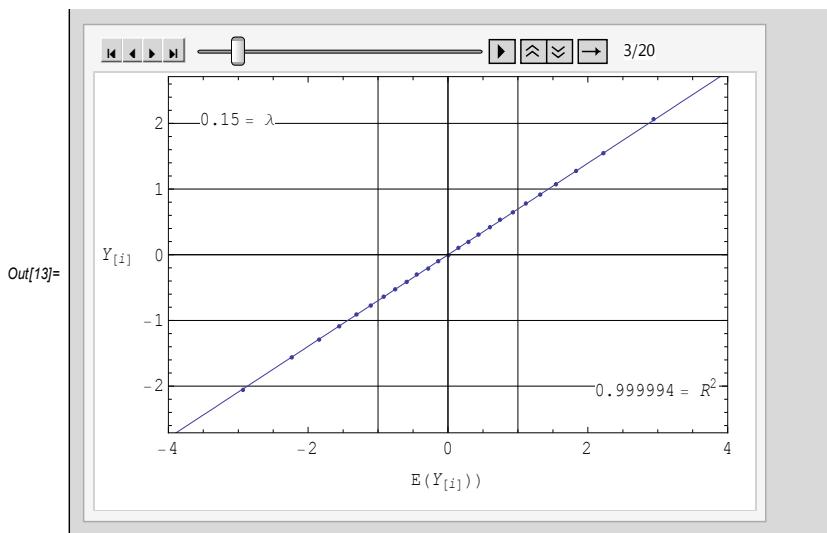

The resulting series of plots represent the projections associated with slices of the surface taken at $\lambda=0.05(.05)1.0$. Using the animation keys to sequentially examine the plots, it quickly becomes apparent that the most linear probability plot occurs at $\lambda=0.15$. The R^2 value supports the visual assessment reaching its maximum of 0.999994 at $\lambda=0.15$. The asymptote $E(Y_{[i]}) = 0$ suggests an estimated mean of 0, while the slope of the plot suggests an estimated standard deviation of (0.8 - (-0.8))/2 = 0.8. This example highlights the relationship that exists between the normal and the symmetric lambda families. The symmetric lambda distribution with $\lambda=0.14$ is used as an approximation to the normal distribution.

The pdf of the standard Weibull distribution is of the form $f(y) = \lambda y^{\lambda-1} e^{-y^\lambda}, 0 < y < \infty$, and the expected value of the order statistics can be approximated by $\left(-\ln\left[\frac{2n+1-2i}{2n}\right]\right)^{\frac{1}{\lambda}}$. The subsequent routine creates a series of univariate probability plots that permits examination of the Weibull family of distributions for values of the shape parameter $\lambda=1(.25)5$. Again the goal is to find the most linear portion of the surface or most linear slice of the surface for the values of the shape parameter considered. In those cases where the "most" linear probability

plot is deemed non-linear then either the Weibull family is inappropriate and/or the value of the shape parameter has not been included.

A Weibull distribution with shape parameter of approximately 3.25 is often cited as a reasonable approximation to the normal distribution. Again using the animation keys to sequentially examine the plots, it quickly becomes apparent that the most linear plot of the series visually appears to occur at λ=3.25 or λ=3.50 while the R^2 value suggests that the most linear plot occurs at λ=3.50.

$y = \{0.1, 0.2025, 0.3045, 0.4124, 0.523, 0.6433, 0.7723, 0.9154, 1.08, 1.282, 1.555, 2.054,$
$-2.054, -1.555, -1.282, -1.08, -0.9154, -0.7723, -0.6433, -0.523, -0.4124,$
$-0.3045, -0.2025, -0.1, 0\};$

$n = \text{Length}[y]; s = \text{Min}[y]; t = \text{Max}[y]; d = \frac{4(t-s)}{n};$
$l := \text{Text}[\text{Style}[\text{lambda "=} \lambda \text{"}], \{\text{"0.5"}, t\}];$

$\text{pdfs}[j\_] := \left(-\text{Log}\left[\frac{2n+1-2j}{2n}\right]\right)^{1/\text{lambda}}$
$\text{EYpairs}[y\_, \text{pdfs}\_] :=$
$\text{With}[\{n = \text{Length}[y]\},$
$\text{Transpose}[$
$\{(\text{pdfs}[\#1]\&) / @\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20,$
$21, 22, 23, 24, 25\}, \text{Sort}[y]\}]]$
$\text{linePlot} := \text{Plot}[\text{Evaluate}[\text{Fit}[\text{EYpairs}[y, \text{pdfs}], \{1, x\}, x]], \{x, 0, \text{"4."}\}, \text{DisplayFunction} \rightarrow \text{Identity}];$
$\text{r2} := \text{LinearModelFit}[\text{EYpairs}[y, \text{pdfs}], \{1, x\}, x];$
$\text{probabilityPlot}[\text{yList}\_, \text{pdfList}\_] :=$
$\text{ListPlot}\left[\text{EYpairs}[y, \text{pdfs}], \text{FrameLabel->}\left\{\text{"E}(Y_{[i]})\text{)"}, \text{"}Y_{[i]}\text{"}\right\}, \text{RotateLabel} \rightarrow \text{False},\right.$
$\text{PlotRange} \rightarrow \{\{0, 4\}, \{s - d, t + d\}\}, \text{Frame} \rightarrow \text{True}, \text{DisplayFunction} \rightarrow \text{Identity}];$
$\text{rr} := \text{Text}\left[\text{Style}\left[\text{"= }R^2\text{"} \text{r2}[\text{"RSquared"}]\right], \{2.5, -2\}\right];$
$\text{SlideView}[$
$\text{Table}[\text{Show}[\text{probabilityPlot}[\text{yList}, \text{pdfList}], \text{linePlot}, \text{Graphics}[\text{rr}], \text{Graphics}[l],$
$\text{DisplayFunction} \rightarrow \$\text{DisplayFunction}, \text{ImageSize} \rightarrow \text{Scaled}[0.9],$
$\text{Prolog} \rightarrow \text{AbsolutePointSize}[4]], \{\text{lambda}, \text{"1."}, \text{"5."}, \text{"0.25"}\}], \text{AppearanceElements} \rightarrow \text{All}]$
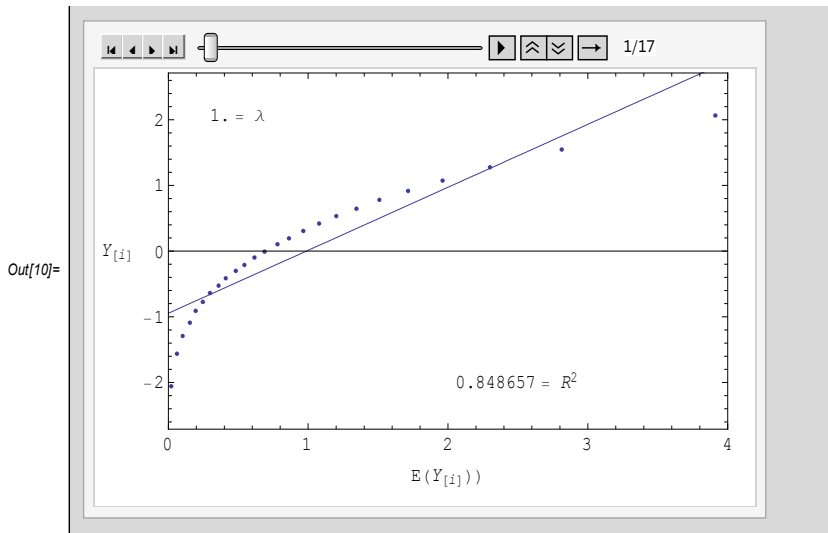
3. Process Capability Paper

Chan, Cheng and Spiring (1988) proposed a graphical technique for examining process capability by combining the concepts that process capability indices assume the underlying distribution is normal and the graphical assessment of normality derived from normal probability plots. The result was Process Capability Paper. An example where 23 observations were gathered from a process with USL=200, T=180 and LSL=160 is illustrated below.
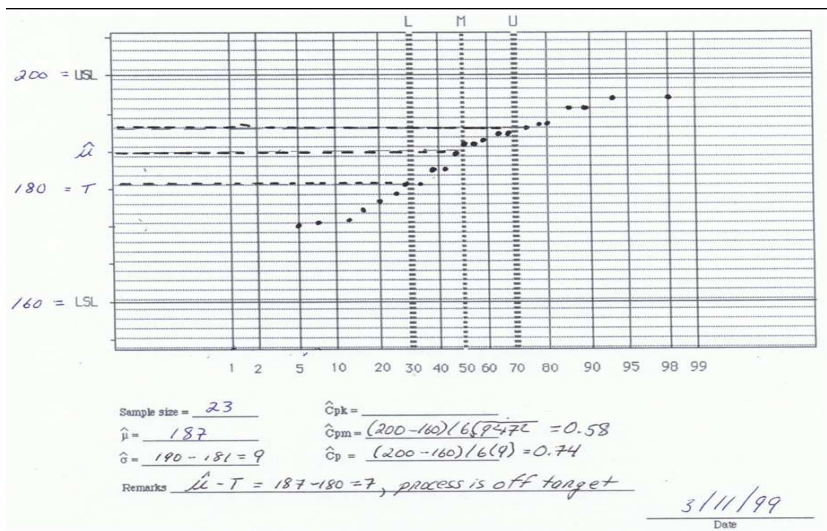
Mathematica can be used to a) evaluate the $E(Y_{[i]})'s$, b) create the resulting Process Capability Paper plot and c) assist in assessing linearity and determining parameter estimates. The

addition of a least squares line and the resulting coefficient of determination (R^2) provide insights into linearity of the probability plot. The least squares line provides visual assistance in assessing the linearity, while R^2 provides numerical assessment. The least squares line and R^2 (i.e., RSquared) are included in subsequent plots. If the resulting capability plot is not considered linear then the various process capability indices may not provide valid indications of process capability.

Mathematica can be used to create the basic format for the Process Capability Paper by inputting the basic information from the process including the study results (data), upper specification limit (USL), lower specification limit (LSL), Target and Target *Cpm* (TCpm). The following *Mathematica* code will create an updated version of the Process Capability Paper.

```
data = {173, 174, 175, 176, 177, 179, 181, 181, 183, 183, 186, 188, 188, 189, 190, 190,
191, 192, 192, 195, 195, 197, 197};

USL:=200; LSL:=160; Target:=185; TCpm:=1.00;
(* code to create plots *)
m:=Mean[data]; s:=StandardDeviation[data];
cpm:=Min[{USL − Target}, {Target − LSL}]/(3 ∗ (s^2 + (m − Target)^2)^(1/2));
sig:=(((Min[{USL − Target}, {Target − LSL}])^2)/(9 ∗ (TCpm^2)))^(1/2);
mut:=(m − Target);
pdfs = NormalDistribution[0, 1];
EYpairs[data_, pdfs_]:=
With[{n = Length[data]},
Transpose[{Map[Quantile[pdfs, #]&, (Range[n] − 0.5)/n], Sort[data]}]]
linePlot:=Plot[Evaluate[Fit[EYpairs[data, pdfs], {1, x}, x]],
{x, −3.0, 3.0}];
r2:=LinearModelFit[EYpairs[data, pdfs], {1, x}, x]
probabilityPlot[dataList_, pdfList_]:=
ListPlot[EYpairs[data, pdfs],
FrameLabel → {"E(Y_{[i]})", "Y_{[i]}", r2["RSquared"] "=", "=Cpm" N[cpm]},
RotateLabel → False,
PlotRange → {{−3.0, 3.0}, {LSL − 1.5 ∗ s, USL + 1.5 ∗ s}}, Frame → True,
GridLines → {{−1, 0, 1}, {Target}}];
Show[probabilityPlot[dataList, pdfList],
Graphics[{RGBColor[.2, .3, 0], Rectangle[{−3, Target}, {0, m}]}],
Graphics[{RGBColor[1, 0, 0], Rectangle[{−3, LSL − 1.5 ∗ s}, {3, LSL}]}],
Graphics[{RGBColor[1, 0, 0], Rectangle[{−3, USL}, {3, USL + 1.5 ∗ s}]}],
Graphics[Text["USL", {−2.5, USL}]], Graphics[Text["LSL", {−2.5, LSL}]],
Graphics[{RGBColor[0, 0, 1], Text["=T.Cpm" N[TCpm], {2.5, USL}]}],
Graphics[{RGBColor[0, 0, 1], AbsoluteThickness[1.5],
Dashing[{.01, .05, .05, .05}],
Line[{{−3, Target − 3 ∗ sig}, {3, Target + 3 ∗ sig}}]}],
probabilityPlot[dataList, pdfList], linePlot, ImageSize → Scaled[1]]
```
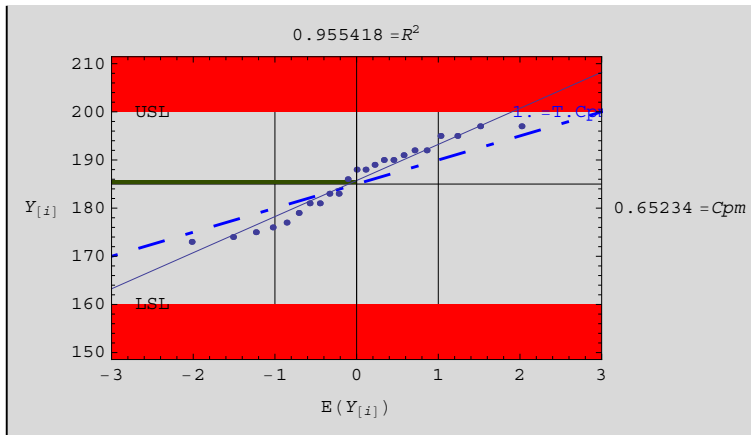
The resulting plot represents 23 observations from a process with a target (T) of 185, upper specification limits (USL) of 200, lower specification limit (LSL) of 160. A Target *Cpm* (TCpm) value of 1 is used to illustrate some of the features included in this enhanced version of Process Capability paper. The enhanced Process Capability paper continues to be a normal probability plot with the y-axis representing the value of the order statistics ($Y_{[i]}$) and the x-axis the expected value of the order statistics ($E(Y_{[i]})$) assuming the underlying distribution is normal.

The resulting plot includes T, USL and LSL with the areas beyond the specification limits highlighted in red. The difference between the process target (T) and the process average (μ) is indicated by the green box. An ordinary least squares (OLS) line (solid line) and the R^2 value (top of the plot frame) are included in order to facilitate the assessment of normality through the linearity of the points and their strength of association. The value of *Cpm* associated with the data is included along with the OLS line (dashed) representing a process that is on target with a *Cpm* = 1.

3.1 Process Capability Paper Enhancements

The *Mathematica*-produced Process Capbility Paper has several enhancements. The output includes the basic features of Process Capability Paper including a normal probability plot of the data including asymptotes at -1, 0 and 1; an OLS line and R^2 in an attempt to enhance the "linearity" assessment of the probability plot; a dashed line reflecting the the slope of the line associated with the target *Cpm*; identified regions beyond the specification limits highlighted in red; and a graphics box indicating the distance the mean is from the target highlighted in green. A second example is illustrated below.

```
data = {0.101, 0.105, 0.099, 0.098, 0.097, 0.101, 0.098, 0.095,
0.099, 0.103, 0.096, 0.104, 0.096, 0.098, 0.097, 0.096, "0.097",
0.099, 0.098, 0.097, 0.095, 0.096, 0.1, 0.097, 0.097};
USL:="0.111"; LSL:="0.089"; Target:="0.1"; TCpm:=2;
m:=Mean[data];
s:=StandardDeviation[data];
```
$$cpm := \frac{Min[\{USL-Target\}, \{Target-LSL\}]}{3\sqrt{s^2+(m-Target)^2}}$$

$$\text{sig} := \sqrt{\frac{\text{Min}[\{\text{USL}-\text{Target}\},\{\text{Target}-\text{LSL}\}]^2}{9\text{TCpm}^2}}$$

mut := $m -$ Target
pdfs = NormalDistribution[0, 1];
EYpairs[data_, pdfs_] :=
With[{$n =$ Length[data]},

$$\text{Transpose}\left[\left\{(\text{Quantile}[\text{pdfs},\#1]\&)/@\frac{\text{Range}[n]-\text{"0.5"}}{n},\text{Sort}[\text{data}]\right\}\right]\right]$$

linePlot := Plot[Evaluate[Fit[EYpairs[data, pdfs], {1, x}, x]],
{x, $-$"4.", "4."}, DisplayFunction $\rightarrow$ Identity];
r2 := LinearModelFit[EYpairs[data, pdfs], {1, x}, x]
probabilityPlot[dataList_, pdfList_] := ListPlot[EYpairs[data, pdfs],

FrameLabel $\rightarrow \left\{ \text{"E}(Y_{[i]})\text{"}, \text{"}Y_{[i]}\text{"}, \text{r2["RSquared"]"="}, \text{"=}Cpm\text{"}N[\text{cpm}] \right\}$,

RotateLabel $\rightarrow$ False,
PlotRange $\rightarrow \{\{-\text{"3."}, \text{"3."}\}, \{\text{LSL} - \text{"1.5"}s, \text{USL} + \text{"1.5"}s\}\}$, Frame $\rightarrow$ True,
Axes $\rightarrow$ None, GridLines $\rightarrow \{\{-1, 0, 1\}, \{\text{Target}\}\}$,
DisplayFunction $\rightarrow$ Identity];
Show[probabilityPlot[dataList, pdfList],
Graphics[{RGBColor["0.2", "0.3", 0], Rectangle[$\{-3$, Target$\}$, $\{0, m\}$]}],
Graphics[{RGBColor[1, 0, 0], Rectangle[$\{-3$, LSL $-$ "1.5"$s\}$, $\{3$, LSL$\}$]}],
Graphics[{RGBColor[1, 0, 0], Rectangle[$\{-3$, USL$\}$, $\{3$, USL $+$ "1.5"$s\}$]}],
Graphics[Text["USL", $\{-$"2.5", USL$\}$]], Graphics[Text["LSL", $\{-$"2.5", LSL$\}$]],
Graphics[{RGBColor[0, 0, 1], Text["=Target Cpm"N[TCpm], {"2.", USL}]}],
Graphics[{RGBColor[0, 0, 1], AbsoluteThickness["1.5"],
Dashing[{"0.01", "0.05", "0.05", "0.05"}],
Line[{$\{-3$, Target $-$ 3sig$\}$, $\{3$, Target $+$ 3sig$\}$}]}],
probabilityPlot[dataList, pdfList], linePlot,
DisplayFunction $\rightarrow$ \$DisplayFunction, Prolog $\rightarrow$ AbsolutePointSize[8],
ImageSize $\rightarrow$ Scaled[1]]
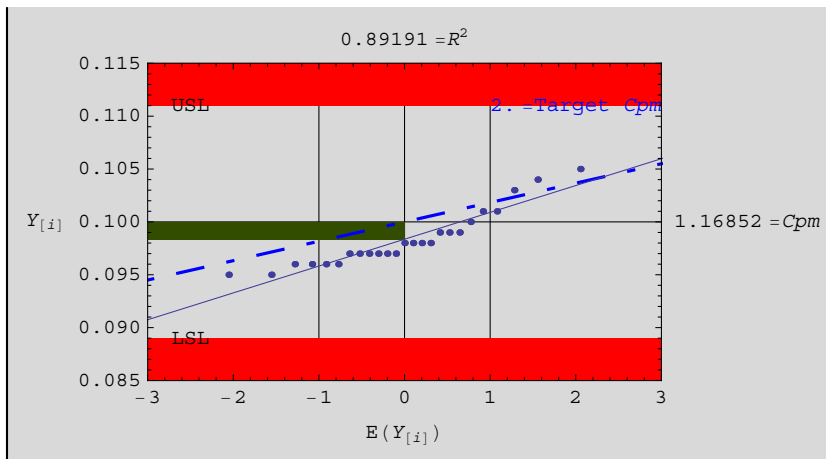
3.2 Additional Views of Process Capability

Mathematica can be used to enhance the inferences from the Enhanced Process Capability Paper by examining departures from the target and incorporating target Capability values with their associated curves.

```
Show[Graphics[{RGBColor[0, 0, 1], Text["Target Cpm", {0, sig}],
Circle [{0, 0}, 1/(3*TCpm) (Min[{USL − Target}, {Target − LSL}]),
{0, 180°}]}],
Graphics[{Black, Text["Observed Cpm", {Abs[mut], s}],
Circle [{0, 0}, 1/(3*cpm) (Min[{USL − Target}, {Target − LSL}]),
{0, 180°}]}], Frame → True,
PlotRange → {{−3 * Abs[mut], 3 * Abs[mut]}, {0, 1.5 * s}},
FrameLabel → {"(μ-T)", "σ"}, RotateLabel → False,
Prolog → {AbsolutePointSize[6], Point[{mut, s}],
Line[{{0, 0}, {0, (1/(3 * cpm))(Min[{USL − Target}, {Target − LSL}])}}]},
ImageSize → Scaled[1]]
```
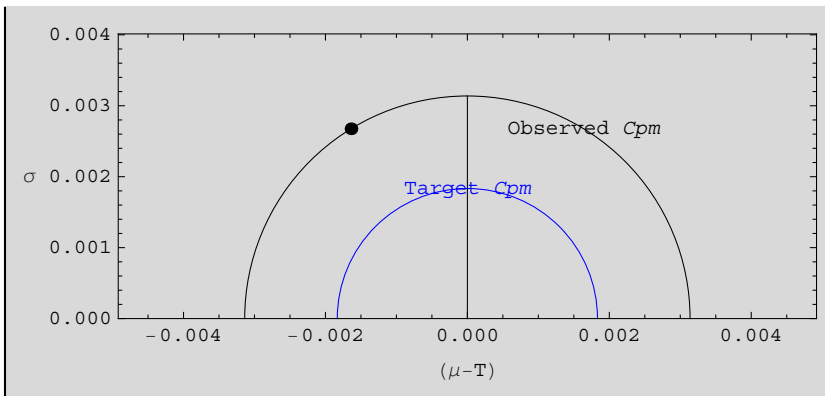

This plot illustrates the various combinations of a process's variability and off-targetness associated with a particular value of *Cpm*. All points lying on the blue (Target *Cpm*) semi-circle represent combinations of off-tagetness (μ - T) and variability (σ) that result in the Target *Cpm*= 2. The black semi-circle represents all combinations of off-targetness and variability that have a *Cpm* value equivalent to that exhibited by the process under investigation. The point represents the observed off-targetness and variability combination associated with the process.

Animation permits a series of plots to be viewed in a sequential fashion resulting in a reliable method for examining a) different views of a single sample or b) multiple samples from comparable processes. The following creates an animated view of multiple samples from a single process. The plot includes the observed *Cpm,* min *Cpm* (red semicircle) and Target *Cpm* for five samples of size five taken from a process.

```
mincpm = 1; TCpm = 2; USL = 0.111; LSL = 0.089; Target = 0.1; groups = 5;
data[1] = {0.101, 0.105, 0.099, 0.098, 0.097};
```

```
data[2] = {0.101, 0.098, 0.095, 0.099, 0.103};
data[3] = {0.096, 0.104, 0.096, 0.098, 0.097};
data[4] = {0.096, 0.097, 0.099, 0.098, 0.097};
data[5] = {0.095, 0.096, 0.1, 0.097, 0.097};
Do[{mu[i] = Mean[data[i]], s[i] = StandardDeviation[data[i]], mut[i] = mu[i] − Target,
cpm[i] = (Min[{USL − Target}, {Target − LSL}]/(3 ∗ (s[i]^2 + (mu[i] − Target)^2)^(1/2))),
n[i] = Length[data[i]]}, {i, groups}];
maxmut:=Max[{Abs[mut[1]], Abs[mut[2]], Abs[mut[3]], Abs[mut[4]], Abs[mut[5]]}];
maxs:=Max[{s[1], s[2], s[3], s[4], s[5]}];
SlideView[
Table[
Show[
Graphics[Circle[{0, 0}, Min[{USL − Target}, {Target − LSL}]/(3cpm[j]),
{0, 180°}]], Frame → True,
Prolog → {AbsolutePointSize[6], Point[{mut[j], s[j]}],
Line[{{0, 0}, {0, 1.5 ∗ maxs}}],
{RGBColor[1, 0, 0], Text["Minimum Cpm", {0.9 ∗ maxmut, 1.45 ∗ maxs}],
Circle[{0, 0}, Min[{USL − Target}, {Target − LSL}]/(3 ∗ mincpm),
{0, 180°}]}, {RGBColor[0, 0, 1],
Text["Target Cpm", {0.9 ∗ maxmut, 1.35 ∗ maxs}],
Circle[{0, 0}, Min[{USL − Target}, {Target − LSL}]/(3 ∗ TCpm),
{0, 180°}]}},
PlotRange → {{−1.5 ∗ maxmut, 1.5 ∗ maxmut}, {0, 1.5 ∗ maxs}},
FrameLabel → {"(μ-T)", "σ", "=Cpm"cpm[j], {"= n"n[j]}},
RotateLabel → False, ImageSize → Scaled[1]], {j, groups}],
AppearanceElements → All]
```
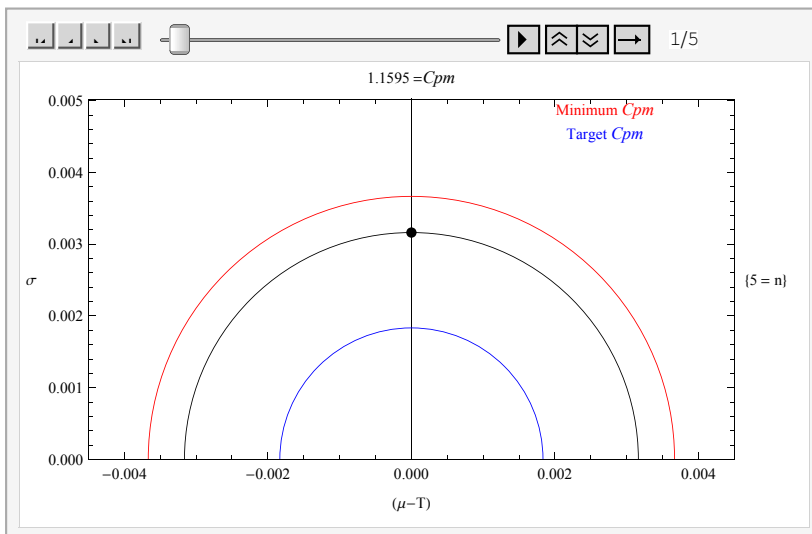
A stationary plot of the observed *Cpm*, min *Cpm* and Target *Cpm* for five samples of size five taken from a process can be created using the following code.

```
Show[
Graphics[
Circle[{0,0},
(((Min[{USL − Target}, {Target − LSL}]))/
(3 ∗ Min[{cpm[1], cpm[2], cpm[3], cpm[4], cpm[5]}])), {0, 180Degree}]],
Graphics[{RGBColor[1, 0, 0], Text["Minimum Cpm", {.9 ∗ maxmut, 1.4 ∗ maxs}],
Circle[{0, 0}, (((Min[{USL − Target}, {Target − LSL}]))/(3 ∗ mincpm)), {0, 180Degree}]}],
Graphics[{RGBColor[0, 0, 1], Text["Target Cpm", {.9 ∗ maxmut, 1.3 ∗ maxs}],
Circle[{0, 0}, (((Min[{USL − Target}, {Target − LSL}]))/(3 ∗ TCpm)), {0, 180Degree}]}],
    Frame → True,
    FrameLabel->{"(μ-T)", "σ", {"= n-max"{Max[n[1], n[2], n[3], n[4], n[5]]},
"= n-min"{Min[n[1], n[2], n[3], n[4], n[5]]}},
"=Smallest Cpm"Min[{cpm[1], cpm[2], cpm[3], cpm[4], cpm[5]}]},
PlotRange->{{−1.5 ∗ maxmut, 1.5 ∗ maxmut}, {0, 1.5 ∗ maxs}}, RotateLabel->False,
Prolog → {AbsolutePointSize[4], Point[{mut[1], s[1]}], Point[{mut[2], s[2]}],
Point[{mut[3], s[3]}], Point[{mut[4], s[4]}], Point[{mut[5], s[5]}],
Line[{{0, 0}, {0, (1.5 ∗ maxs)}}]}], ImageSize → Scaled[1]]
```
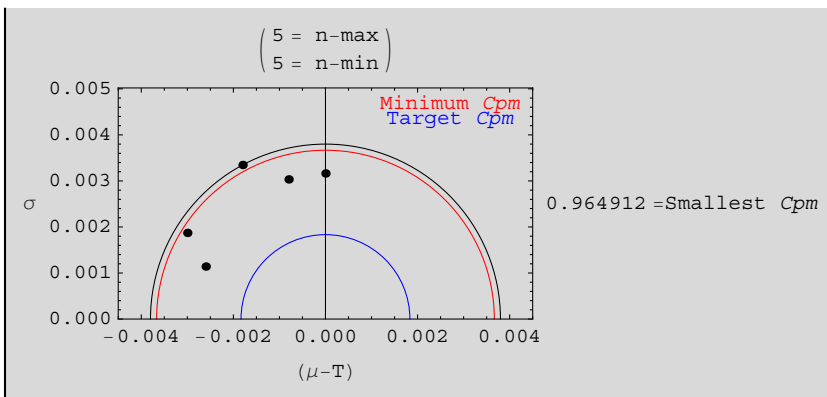

4. Compositional Data

Compositional data refers to the group of constrained space metrics that take the form

$$X_1 + X_2 + X_3 + X_4 + ... + X_d = a$$

where $0 \le X_i \le a$ for all i and each X_i represents a proportion of the total composition a. Setting d=2 results in all possible combinations of R^{+2} (shaded region in Figure 4.1) that satisfy the equation $X_1 + X_2 = a$. Graphically these combinations represent a line in R^{+2} and referred to as the L^1 space.
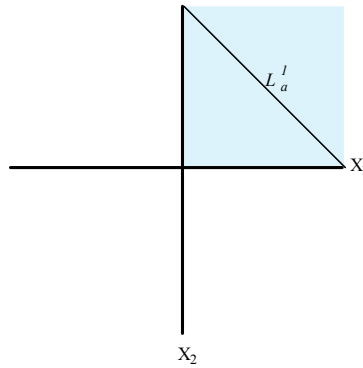
Figure 4.1 The $L^1{}_a$ Constrained Space

All observations in the $L^1{}_a$ space lie on the line $X_1 + X_2 = a$, with different values of a, moving the line either closer of further from the origin. The Euclidean distance along the perpendicular from the origin to the $L^1{}_a$ constrained space is $a\sqrt{2}$. In addition the points where the $L^1{}_a$ space intersects the axes are exactly a units from the origin (see Figure 4.2).
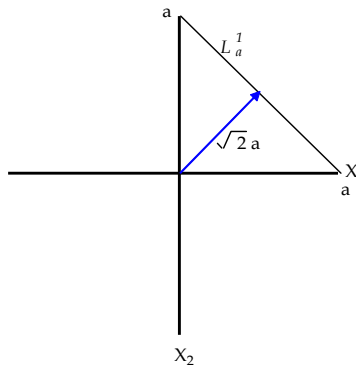
Figure 4.2 Distance from Origin to $L^1{}_a$ along perpindicular

4.1 The L^2 Space

The triple X_1, X_2, X_3 subject to the constraint $X_1 + X_2 + X_3 = a$, represents a point in R^{+3} space. Ternary paper, also referred to as Triangular coordinate paper, is available for observations of the form X_1, X_2, X_3 where $X_1 + X_2 + X_3 = a$ and uses a planar view (see Figure 4.3) of the constrained space. Most commercial ternary paper adds scaling and axes to enhance the plotting procedure with the resulting region referred to as the L^2 space.

An alternative view of the L^2 space rewrites the equation in the form $X_1 + X_2 = a - X_3$ and makes use of the L^1 space. The points X_1, X_2 are located on the $L^1{}_{a-X_3}$ line, which is $\sqrt{2}(a - X_3)$ units (along the perpendicular) from the origin. As the compositional make-up varies (i.e., as we observe different values of the triple X_1, X_2, X_3), the $L^1{}_{a-X_3}$ line will vary as

will the location (i.e., X_1, X_2) on the line. The L^2 space consists of the set of subspaces $L^1{}_{a-X_3}$ where $0 \leq X_3 \leq a$. When using normal arithemetic paper, the values can be determined directly from the plot (see Figure 4.4). X_1 and X_2 are the usual projections onto their appropriate axes, while X_3 is the distance from the intersection of $L^1{}_{a-X_3}$ (with either of the axes) to a on the same axis.
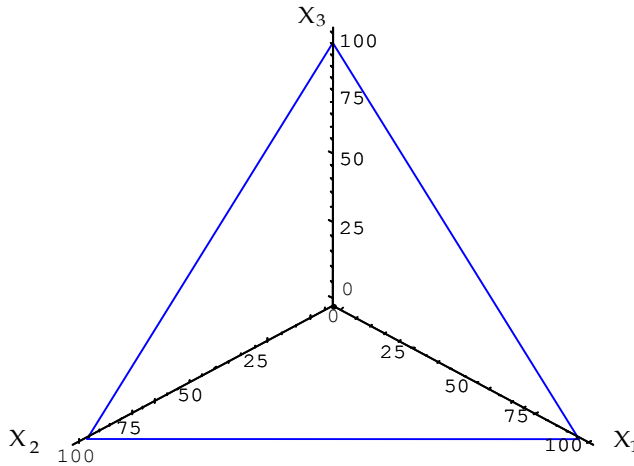
Figure 4.3 The L^2 space as a plane in Three Space
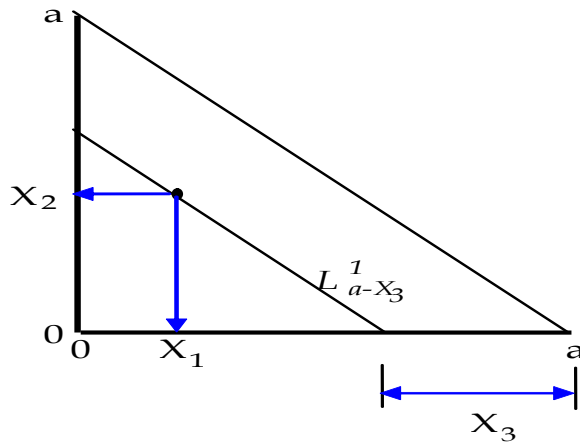
Figure 4.4 Geometric Interpretations of the Constrained Triple

The general $L^2{}_a$ space is easily created using arithmetic graph paper requiring no special scaling or plotting procedures. The plot is easily generalized to allow for various values of a

and the creation of a general form of the constrained space paper. The *Mathematica* code to create the $L^2{}_a$ constrained space paper and plot an observed points (i.e., 0.7, 0.2, 0.1 with a = 1.0) follows.

```
a = 1; x1 = 0.7; x2 = 0.2; x3 = a − x1 − x2;
Show[Graphics[{Point[{0, a}], AbsolutePointSize[7],
Point[{x1, x2}], Point[{a − x3, 0}], Point[{a, 0}],
Line[{{a, 0}, {0, a}}],
{Dashing[{0.05, 0.05}], Line[{{a − x3, 0}, {0, a − x3}}]},
AbsoluteThickness[2],
Line[{{x1, x2}, {x1, 0}}], Line[{{x1, x2}, {0, x2}}],
Line[{{a, 0}, {a − x3, 0}}]}], AxesOrigin->{0.0, 0.0}, Axes->True]]
```
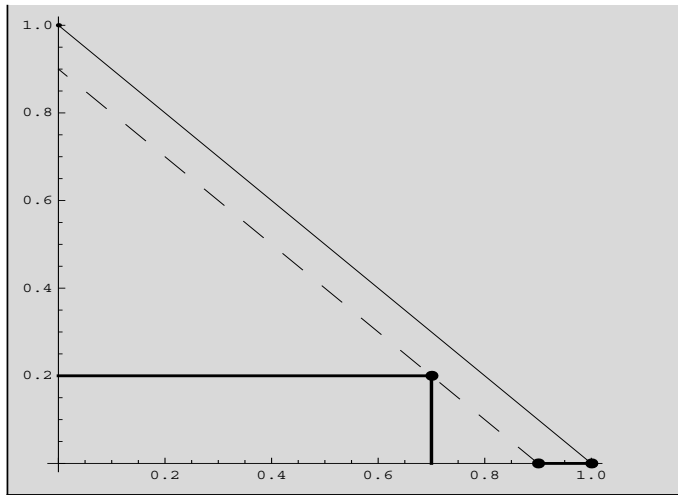

The L^2 space is the area bounded by the X_1, X_2 axes and the solid line that intersects the axes at the value of a=1. The $L^1{}_{a-X_3}$ space is denoted by the dashed line parallel to solid line intersecting the axes exactly X_3 units from a (again 1 in this case). A heavier line has been drawn along the X_1 axis from a towards the origin that is exactly X_3 units in length (0.1 in this case). X_1, X_2, X_3 (in this case 0.7, 0.2, 0.1) has been highlighted at the appropriate point on the solid line (i.e., the $L^1{}_{a-X_3}$ space). In addition the projections onto the axes have been included to facilitate reading the values of X_1, X_2 directly from the plot.

Commercial Ternary paper scales the plane characterized by the points (a, 0, 0), (0, a, 0) and (0, 0, a) in a triangular co-ordinate system. Analogous to the L^1 case where we added a third variable to the mix, the $L^3{}_a$ space can be considered when we add a fourth variable. Similar to the L^2 space development, X_1, X_2, X_3, X_4 where $0 \leq X_i \leq a$ for all i such that $X_1 + X_2 + X_3 + X_4 = a$ can be written as $X_1 + X_2 + X_3 = a − X_4$ and the perpendicular distance between the origin and the L^2 plane to reflect the magnitude of X_4.

Alternatively we could use other techniques to provide the inference regarding Ternary plots in the $L^2 \, xD_1$ domain. Consider the case where X_1, X_2, X_3, X_4 where $0 \leq X_i \leq a$ for all i

such that $X_1 + X_2 + X_3 + X_4 = 100$ and where $X_4 = 0$. This is equivalent to looking at the standard L^2 Ternary plot and in this case would be scaled similar to commercial Ternary paper. The point $(30, 30, 40, 0)$ would appear as follows (see Figure 4.5). This plane would be $\sqrt{2}(100 - 0)$ units along the perpendicular from the origin.
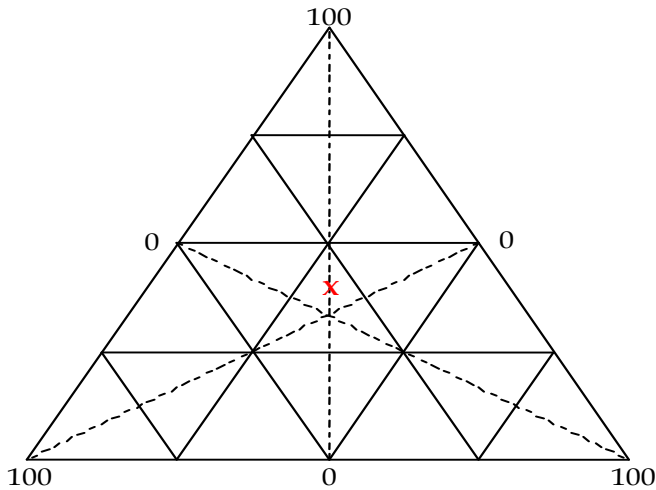
Figure 4.5 Planar view of $L^3{}_a$ Space with point $(30, 30, 40, 0)$

The following *Mathematica* code results in a plane $\sqrt{2}(100 - 0)$ units along the perpindicaular from the origin and point at $(30, 30, 40, 0)$.

```
x = 30; y = 30; z = 40; constraint = 100; a = constraint − x − y − z;
pts = {{0, 100 − a, 0}, {100 − a, 0, 0}, {0, 0, 100 − a}, {0, 100 − a, 0}};
oldpts = {{0, 100, 0}, {100, 0, 0}, {0, 0, 100}, {0, 100, 0}};
oldpts11:={{50 − a/2, 0, 50 − a/2}, {0, 100 − a, 0}}
oldpts12:={{0, 50 − a/2, 50 − a/2}, {100 − a, 0, 0}}
oldpts13:={{50 − a/2, 50 − a/2, 0}, {0, 0, 100 − a}}
Show[Graphics3D[{AbsolutePointSize[7], Point[{x, y, z}],
Line[pts], {{Dashing[{0.01, 0.01}],
Line[oldpts11]}}, {Dashing[{0.01, 0.01}],
{Line[oldpts12]}}, {Dashing[{0.01, 0.01}],
{Line[oldpts13]}}, {Dashing[{0.03, 0.03}], Line[oldpts]}},
{Boxed->False, Ticks->None, Axes->False,
AxesEdge->{{1, 1}, {1, 1}, {1, 1}}, ViewPoint->{−2, −2, −2}}]]
```

Suppose now that we observe the point (30, 30, 20, 20) under the constraint that $X_1 + X_2 + X_3 + X_4 = 100$. The quadruple is represented by a point on the plane that is $\sqrt{2}(100 - 20)$ units form the origin. The point and its associated plane can be depicted as wholly contained within the plane associated with $X_4 = 0$ (dashed triangle below). *Mathematica* produces the above plot of the Planar view of the $L^3{}_a$ space with point (30, 30, 40, 20).

$x = 30; y = 30; z = 20; \text{constraint} = 100; a = \text{constraint} - x - y - z;$
$\text{pts} = \{\{0, 100 - a, 0\}, \{100 - a, 0, 0\}, \{0, 0, 100 - a\}, \{0, 100 - a, 0\}\};$
$\text{oldpts} = \{\{0, 100, 0\}, \{100, 0, 0\}, \{0, 0, 100\}, \{0, 100, 0\}\};$
$\text{oldpts11} := \{\{50 - a/2, 0, 50 - a/2\}, \{0, 100 - a, 0\}\}$
$\text{oldpts12} := \{\{0, 50 - a/2, 50 - a/2\}, \{100 - a, 0, 0\}\}$
$\text{oldpts13} := \{\{50 - a/2, 50 - a/2, 0\}, \{0, 0, 100 - a\}\}$
$\text{Show}[\text{Graphics3D}[\{\text{AbsolutePointSize}[7], \text{Point}[\{x, y, z\}],$
$\text{Line}[\text{pts}], \{\{\text{Dashing}[\{0.01, 0.01\}],$
$\text{Line}[\text{oldpts11}]\}\}, \{\text{Dashing}[\{0.01, 0.01\}],$
$\{\text{Line}[\text{oldpts12}]\}\}, \{\text{Dashing}[\{0.01, 0.01\}],$
$\{\text{Line}[\text{oldpts13}]\}\}, \{\text{Dashing}[\{0.03, 0.03\}], \text{Line}[\text{oldpts}]\}\},$
$\{\text{Boxed->False}, \text{Ticks->None}, \text{Axes->False},$
$\text{AxesEdge->}\{\{1, 1\}, \{1, 1\}, \{1, 1\}\}, \text{ViewPoint->}\{-2, -2, -2\}\}]]$

5. Comments

We have attempted to show how *Mathematica* can provide practitioners with the ability to quickly and simply examine data. In conjunction with functions found in *Mathematica*, the graphical methods developed may provide powerful inferences when assessing distributional forms, estimating parameter values, investigating process capability and examining constrained data.

6. References

Chan, L.K., Cheng, S.W. & Spring, F.A. (1988). A Graphical Technique for Process Capability. ASQ 42nd Annual Quality Congress Transactions.

Cheng, S.W. & Spring, F.A. (1990). Some Applications of 3-D scatter plots in data analysis, Computational Statistics & Data Analysis, 10, pp. 47-61.

Hahn, G.J, & Shapiro, S.S. (1967). Statistical Models in Engineering. John Wiley & Sons, New York.

Kimball, B.F. (1960). On the Choice of Plotting Positions on Probability Paper, Journal of the American Statistical Association, 55, pp. 546-.

Wolfram, S. (2009). *Mathematica7*. Wolfram Media, Illinois and Cambridge University Press, New York.

A learning algorithm based on PSO and L-M for parity problem

Guangyou Yang, Daode Zhang, and Xinyu Hu
School of Mechanical Engineering
Hubei University of Technology
Wuhan,430068
P. R. China

1. Introduction

The Back-propagation network(BP network) is the most representative model and has wide application in artificial neural network(J. L. McClelland, D. E. Rumelhart & the PDP Research Group). Owing to the hidden layer and learning rules in the BP network and the Error Back-propagation algorithm, the BP network can be used to recognize and classify nonlinear pattern(Zhou zhihua, Cao Cungen, 2004). Currently, the applications include handwritings recognition, speech recognition, text - language conversion, image recognition and intelligent control. As the BP algorithm is based on gradient descent learning algorithm, it has some drawbacks such as slow convergence speed and easily falling into local minimum, as well as poor robustness. In the last decade, a series of intelligent algorithms, which is developed from nature simulation, are got wide attention, especially the global stochastic optimization algorithm based on the individual organisms and groups makes a rapid development and gets remarkable achievements in the field of engineering design and intelligent control. The most famous are genetic algorithm, the PSO algorithm (Particle Swarm Optimization, PSO), etc. In this chapter, the research focuses on the integration of the improved PSO algorithm and the Levenberg-Marquardt (L-M) algorithm of neural network, and its application in solving the parity problem, which enhances the optimization property of the algorithm, and solves the problems such as slow convergence speed and easily falling into local minimum.

2. Particle Swarm Optimization (PSO)

2.1 Standard Particle Swarm Optimization

Dr. Kennedy and Dr. Eberhart proposed the PSO algorithm in 1995(Kennedy, & Eberhart, 1995), which derived from the behavior research of flock foraging, and the research found out that the PSO theory can be applied to the function optimization, then it was developed into a universal optimization algorithm gradually. As the concept of PSO is simple and easy to implement, at the same time, it has profound intelligence background, the PSO algorithm attracted extensive attention when it was first proposed and has become a hot topic of

research. The search of PSO spreads all over the solution space, so the global optimal solution can be easily got, what is more, the PSO requires neither continuity nor differentiability for the target function, even doesn't require the format of explicit function, the only requirement is that the problem should be computable. In order to realize the PSO algorithm, a swarm of random particles should be initialized at first, and then get the optimal solution through iteration calculation. For each iteration calculation, the particles found out their individual optimal value of *pbest* through tracking themselves and the global optimal value of *gbest* through tracking the whole swarm. The following formula is used to update the velocity and position.

$$v_{id}^{k+1} = w \cdot v_{id}^{k} + c_1 \cdot rand \cdot (p_{id} - x_{id}^{k}) + c_2 \cdot rand \cdot (p_{gd} - x_{id}^{k}) \tag{1}$$

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \tag{2}$$

In the formula (1) and (2), $i=1, 2, …, m$, m refers to the total number of the particles in the swarm; $d=1, 2,…, n$, d refers to the dimension of the particle; v_{id}^{k} is the No. d dimension component of the flight velocity vector of iteration particle i of the No. k times. x_{id}^{k} is the No. d dimension component of the position vector of iteration particle i of the No. k times; p_{id} is the No. d dimension component of the optimization position (*pbest*) of particle i ; p_{gd} is the No. d dimension component of the optimization position (*gbest*) of the swarm; w is the inertia weight; $c1,c2$ refer to the acceleration constants; rand() refers to the random function, which generates random number between [0, 1]. Moreover, in order to prevent excessive particle velocity, set the speed limit for V_{max}, when accelerating the particle velocity into the level: $v_{id} > V_{max}$, set $v_{id} = V_{max}$; In contrast, on the condition of $v_{id} < -V_{max}$, set $v_{id} = -V_{max}$.
The specific steps of the PSO algorithm are as follows:

(1) Setting the number of particles m, the acceleration constant $c1,c2$，inertia weight coefficient w and the maximum evolution generation $Tmax$，in the n-dimensional space, generating the initial position $X(t)$ and velocity $V(t)$of m-particles at random.
(2) Evaluation of Swarm $X(t)$
   i. Calculating the fitness value *fitness* of each particle.
   ii. Comparing the fitness value of the current particle with its optimal value *fpbest*. If *fitness* < *fpbest*, update *pbest* for the current location, and set the location of *pbest* for the current location of the particle in the n-dimensional space.
   iii. Comparing the fitness value of the current particle with the optimal value *fGbest* of the swarm. If *fitness* < *fGbest*, update *gbest* for the current location, and set the *fGbest* for the optimal fitness value of the swarm, then the current location *gbest* of the particle is referred to as the optimal location of the swarm in the n-dimensional space.
(3) In accordance with the formula (1) and (2), updating the location and velocity of the particles and generating a new swarm $X(t+1)$.

(4) Checking the end condition, if meet the end condition, then stop optimizing; Otherwise, $t=t+1$ and turn to step (2).

In addition, the end condition is referred to as the following two situations: when the optimizing reaches the maximum evolution generation *Tmax* or the fitness value of *gbest* meets the requirement of the given precision.

2.2 Improved Particle Swarm Optimization Algorithm

The PSO algorithm is simple, but research shows that, when the particle swarm is over concentrated, the global search capability of particle swarm will decline and the algorithm is easy to fall into local minimum. If the aggregation degree of the particle swarm can be controlled effectively, the capability of the particle swarm optimizing to the global minimum will be improved. According to the formula (1), the velocity v of the particle will become smaller gradually as the particles move together in the direction of the global optimal location *gbest*. Supposed that both the social and cognitive parts of the velocity become smaller, the velocity of the particles will not become larger, when both of them are close to zero, as $w<1$, the velocity will be rapidly reduced to 0, which leads to the loss of the space exploration ability. When the initial velocity of the particle is not equal to zero, the particles will move away from the global optimal location of *gbest* by inertial movement. When the velocity is close to zero, all the particles will move closer to the location of *gbest* and stop movement. Actually, the PSO algorithm does not guarantee convergence to the global optimal location, but to the optimal location *gbest* of the swarm(LU Zhensu & HOU Zhirong, 2004). Furthermore, as shown in the formula (2), the value of the particle velocity also represents the distance of particle relative to the optimal location *gbest*. When the particles become farther from the *gbest*, the particle velocity will be greater, on the contrary, when the particles become closer to the *gbest*, the velocity will be smaller gradually. Therefore, as shown in the formula (1), by means of the extreme variation of the swarm individual, the velocity of the particles can be controlled in order to prevent the particles from gathering at the location *gbest* quickly, which can control the swarm diversity effectively. Known from the formula (1), when the variability measures are taken, both the social and cognitive parts of each particle velocity are improved, which enhances the particle activity and increases the global search capability of particle swarm to a large extent. The improved PSO(MPSO) is carried out on the basis of standard PSO, which increases the variation operation of optimal location for the swarm individual. The method includes the following steps:

(1) Initializing the position and velocity of particle swarm at random;

(2) The value *pbest* of the particle is set as the current value, and the *gbest* for the optimal particle location of the initial swarm；

(3) Determining whether to meet the convergence criteria or not, if satisfied, turn to step 6; Otherwise, turn to step 4;

(4) In accordance with the formula (1) and (2), updating the location and velocity of the particles, and determining the current location of *pbest* and *gbest*；

(5) Determining whether to meet the convergence criteria or not, if satisfied, turn to step 6; Otherwise, carrying out the optimal location variation operation of swarm individuals according to the formula (3), then turn to step 4;

$$new\_pbest_d = pbest_d(1+\eta\sigma) \tag{3}$$

(6) Outputting the optimization result, and end the algorithm.

In the formula (3), the parameter σ refers to random number which meets the standard Gaussian distribution, the initial value of the parameter η is 1.0, and set $\eta=\beta\eta$ every 50 generations, where the β refers to the random number between [0.01, 0.9]. From above known, the method not only produces a small range of disturbance to achieve the local search with high probability, but also produces a significant disturbance to step out of the local minimum area with large step migration in time.

2.3 Simulation and Result Analysis of the Improved Algorithm

2.3.1 Test Functions

The six frequently used Benchmark functions of the PSO and GA(genetic algorithm) (Wang Xiaoping & Cao Liming, 2002)are selected as the test functions, where the Sphere and Rosenbrock functions are unimodal functions, and the other four functions are multimodal functions. The Table 1 indicates the definition, the value range and the maximum speed limit *Vmax* of these Benchmark functions, where: x refers to real type vector and its dimension is n, x_i refers to the No. i element.

name	Function	Initialization Range	V_{max}
Sphere	$f_1(x) = \sum_{i=1}^{n} x_i^2$	$(-1000,1000)^n$	1000
Rastrigrin	$f_2(x) = \sum_{i=1}^{n} (x_i^2 - 10\cos(2\pi x_i) + 10)$	$(-5.12,5.12)^n$	10
Griewank	$f_3(x) = \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n}\cos(\frac{x_i}{\sqrt{i}}) + 1$	$(-600,600)^n$	600
Rosenbrock	$f_4(x) = \sum_{i=1}^{n}(100x_{i+1} - x_i^2)^2 + (x_i - 1)^2)$	$(-30,30)^n$	100
Ackley	$f_5(x) = -20\exp(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2})$ $-\exp(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)) + 20 + e$	$(-30,30)^n$	30
Schaffer	$f_6(x) = 0.5 + \frac{(\sin\sqrt{x^2 + y^2})^2 - 0.5}{(1.0 + 0.001(x^2 + y^2))}$	$(-5.12,5.12)^n$	1

Table 1. Benchmark functions

2.3.2 Simulation and Analysis of the Algorithm

In order to study the property of the improved algorithm, the different performances are compared between the standard PSO and the improved PSO (mPSO) for Benchmark functions, which adopt linear decreased inertia weight coefficient. The optimal contrast test is performed on the common functions as shown in Table 1. For each algorithm, the maximum evolution generation is 3000, the number of the particles is 30 and the dimension is 10, 20 and 30 respectively, where the dimension of Schaffer function is 2. As for the inertia weight coefficient w, the initial value is 0.9 and the end value is 0.4 in the PSO algorithm, while in the mPSO algorithm, the value of w is fixed and taken to 0.375.The optimum point of the Rosenbrock function is in the position $X=1$ in theory, while for the other functions, the optimum points are in the position $X=0$ and the optimum value are $f(x)=0$. The 50 different optimization search tests are performed on different dimensions of each function. The results are shown in Table 2, where the parameter Avg/Std refers to the average and variance of the optimal fitness value respectively during the 50 tests, iterAvg is the average number of evolution, Ras is the ratio of the number up to target value to the total test number. The desired value of function optimization is set as 1.0e-10, as the fitness value is less than 10e-10, set as 0.

Fun.	Dim	PSO			mPSO		
		Avg/Std	iterAvg	Ras	Avg/Std	iterAvg	Ras
f1	10	0/0	1938.52	50/50	0/0	340.18	50/50
	20	0/0	2597.24	50/50	0/0	397.64	50/50
	30	0/0	3000	1/50	0/0	415.02	50/50
f2	10	2.706/1.407	3000	0/50	0/0	315.24	50/50
	20	15.365/4.491	3000	0/50	0/0	354.10	50/50
	30	41.514/11.200	3000	0/50	0/0	395.22	50/50
f3	10	0.071/0.033	3000	0/50	0/0	294.32	50/50
	20	0.031/0.027	2926.94	8/50	0/0	343.42	50/50
	30	0.013/0.015	2990.52	13/50	0/0	370.06	50/50
f4	10	13.337/25.439	3000	0/50	8.253/0.210	3000	0/50
	20	71.796/175.027	3000	0/50	18.429/0.301	3000	0/50
	30	122.777/260.749	3000	0/50	28.586/0.2730	3000	0/50
f5	10	0/0	2197.40	50/50	0/0	468.08	50/50
	20	0/0	2925.00	47/50	0/0	532.78	50/50
	30	0.027/0.190	3000	0/50	0/0	562.60	50/50
f6	2	0.0001/0.002	857.58	47/50	0/0	67.98	50/50

Table 2. Performance comparison between mPSO and PSO for Benchmark problem

As shown in Table 2, except for the Rosenbrock function, the optimization results of the other functions reach the given target value and the average evolutionary generation is also very little. For the Schaffer function, the optimization test is performed on 2-dimension, while for the other functions, the tests are performed on from 10 dimensions to 30 dimensions. Compared with the standard PSO algorithm, whether the convergence accuracy or the convergence speed of the mPSO algorithm has been significantly improved, and the mPSO algorithm has excellent stability and robustness.

In order to illustrate the relationship between the particle activity and the algorithm performance in different algorithms, the diversity of particle swarm indicates the particle activity. The higher the diversity of particle swarm is, the greater the particle activity is, and the stronger the global search capability of particles is. The diversity of particle swarm is represented as the average distance of the particles, which is defined by Euclidean distance, and the distance L refers to the maximum diagonal length in the search space; The parameters of S and N represent the population size and the solution space dimension, respectively; p_{id} refers to the No.d dimension coordinate of the No.i particle; $\overline{p_d}$ is the average of the No.d dimension coordinate of all particles, so the average distance of the particles is defined as followed:

$$d(t) = \frac{1}{sL} \cdot \sum_{i=1}^{s} \sqrt{\sum_{j=1}^{N}(p_{ij} - \overline{p_d})^2} \tag{4}$$

For the 30-D functions (Schaffer function is 2-D), the optimal fitness value and particles' average distance are shown in Fig.1-6, which indicates the optimization result contrast of the mPSO and PSO algorithm performed on different functions.
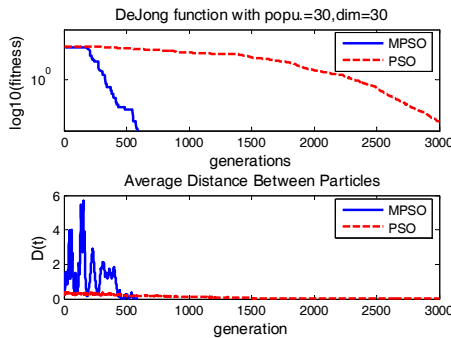
Fig. 1. Minima value and particles' average distance for 30-D Sphere
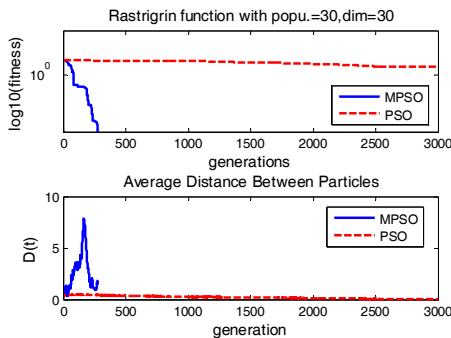
Fig. 2. Minima value and particles' average distance for 30-D Rastrigin
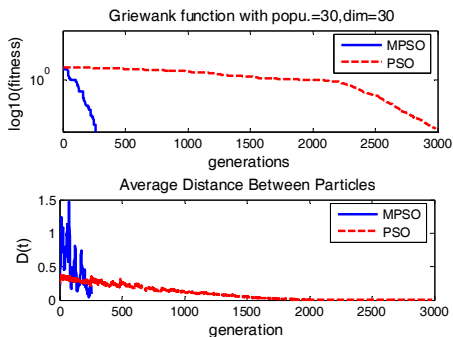
Fig. 3. Minima value and particles' average distance for 30-D Griewank
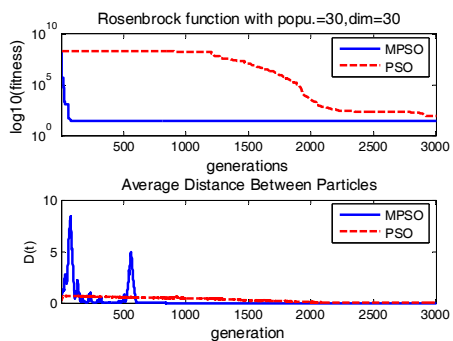
Fig. 4. Minima value and particles' average distance for 30-D Rosenbrock
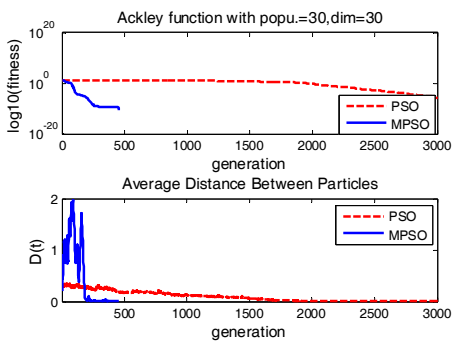
Fig. 5. Minima value and particles' average distance for 30-D Ackley
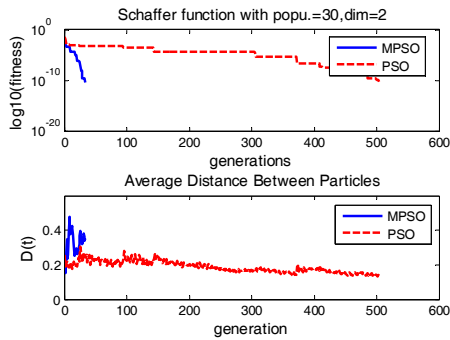
Fig. 6. Minima value and particles' average distance for 2-D Schaffer

As can be seen from the Figure 1-6, except for the Rosenbrock function, the average distance of particle swarm varies considerably, which indicates the particle's high activity as well as the good dynamic flight characteristic, which can also be in favor of the global search due to the avoidance of local minimum. When the particle approaches the global extreme point, the amplitude of its fluctuation reduces gradually, and then the particle converges quickly to the global extreme point. The mPSO algorithm has demonstrated the high accuracy and fast speed of the convergence. Compared with the corresponding graph of PSO algorithm in the chart, the the particles' average distance of the PSO algorithm decreases gradually with the increase of evolution generation, and the fluctuation of the particles is weak, and the activity of the particles disappears little by little, which is the reflection of the algorithm performance, i.e., it means slow convergence speed and the possibility of falling into local minimum. As weak fluctuation means very little diversity of particle swarm, once the particles fall into local minimum, it is quite difficult for them to get out. The above experiments, performed on the test functions, show that: the higher the diversity of particle swarm is, the greater the particle activity is, and the better the dynamic property of particle is, which result in stronger optimization property. Therefore, it is a key step for the PSO to control the activity of the particle swarm effectively. Besides, from the optimization results of mPSO algorithm shown in Table 2, it can be seen that, except for the Rosenbrock function, not only the mean of the other functions has reached the given target value, but also the variance is within the given target value, which shows that the mPSO algorithm has high stability and has better performance than the PSO algorithm. In addition, the chart has also indicated that, for the optimization of Rosenbrock function, whether the mPSO or the PSO algorithm is applied, the particles have high activity at the beginning, then gather around the adaptive value quickly, after which the particle swarm fall into the local minimum with the loss of its activity. Though the optimization result of mPSO for Rosenbrock function is better than the standard PSO algorithm, it has not yet got out of the local minimum. Hence, further study is needed on the optimization of PSO for Rosenbrock function.

3. BP Network Algorithm Based on PSO

3.1 BP Neural Network
Artificial Neural Network (ANN) is an engineering system that can simulate the structure and intelligent activity of human brain, which is based on a good knowledge of the structure

and operation mechanism of the human brain. According to the manner of neuron interconnection, neural network is divided into feedforward neural network and feedback neural network. According to the hierarchical structure, it is separated into single layer and multi-layer neural network. In terms of the manner of information processing, it is separated into continuous and discrete neural network, or definitive and random neural network, or global and local approximation neural network. According to the learning manner, it is separated into supervision and unsupervised learning or weight and structure learning. There are several dozens of neural network structures such as MLP, Adaline, BP, RBF and Hopfield etc. From a learning viewpoint, the feedforward neural network (FNN) is a powerful learning system, which has simple structure and is easy to program. From a systemic viewpoint, the feedforward neural network is a static nonlinear mapping, which has the capability of complex nonlinear processing through the composite mapping of simple nonlinear processing unit.

As the core of feedforward neural network, the BP network is the most essential part of the artificial neural network. Owing to its clear mathematical meaning and steps, Back-Propagation network and its variation form are widely used in more than 80% of artificial neural network model in practice.

3.2 BP Network Algorithm Based on PSO

The BP algorithm is highly dependent on the initial connection weight of the network, therefore, it has the tendency of falling into local minimum with improper initial weight. However, the optimization search of the BP algorithm is under the guidance (in the direction of negative gradient), which is superior to the PSO algorithm and other stochastic search algorithm. There is no doubt that it provides a method for the BP optimization with derivative information. The only problem is how to overcome the BP algorithm for the dependence of the initial weight. The PSO algorithm has strong robustness for the initial weight of neural network (Wang Ling, 2001). By the combination of the PSO and BP algorithm, it could improve the precision, speed and convergence rate of BP algorithm, which makes full use of the advantage of the PSO and BP algorithm, i.e., the PSO has great skill in global search and BP excels in local optimization.

Compared with the traditional optimization algorithm, the feedforward neural network has great differences such as multiple variables, large search space and complex optimized surface. In order to facilitate the PSO algorithm for BP algorithm in certain network structure, the weight vector of NN is used to represent FNN, and each dimension of the particles represents a connection weights or threshold value of FNN, which consists of the individuals of the particle swarm. To take one input layer, a hidden layer and an output layer of FNN as an example, when the number of input nodes was set as R, the number of output nodes was set as $S2$ and the number of hidden nodes was set as $S1$, the dimension N of particles can be obtained from the formula (5):

$$N=S1*(R+1)+ S2*(S1+1)+ S3*(S2+1) \qquad (5)$$

The dimension of the particles and the weight of FNN can be obtained by the following code conversion:

```
    for i=1:S1
        w1(i,:)=Swarm(iPopindex,R*(i-1)+1:R*(i-1)+R);
        c1(i)=Swarm(iPopindex,S1*R+i);
    end
    b1=c1';
    for i=1:S2
      w2(i,:)=Swarm(iPopindex,S1*(R+1)+S1*(i-1)+1:S1*(R+1)+S1*(i-1)+S1);
        c2(i)=Swarm(iPopindex,S1*(R+1)+S2*S1+i);
    end
    b2=c2';
    Where, iPopindex refers to the serial number of the particles.
```

When training the BP network through PSO algorithm, the position vector X of particle swarm is defined as the whole connection weights and threshold value of BP network.. On the basis of the vector X, the individual of the optimization process is formed, and the particle swarm is composed of the individuals. So the method is as follows: at first, initializing the position vector, then minimize the sum of squared errors (adaptive value) between the actual output and ideal output of network, and the optimal position can be searched by PSO algorithm, as shown in the following formula (6):

$$J = \sum_{i=1}^{N} \sum_{k=1}^{c} (T_{ik} - Y_{ik})^2 \qquad (6)$$

Where: N is the sample number of training set; T_{ik} is the ideal output of the No. k output node in the No. i sample; Y_{ik} is the actual output of the No. k output node in the No. i sample; C is the number of output neuron in the network.

The PSO algorithm is used to optimize the BP network weight (PSOBP), the method includes the following main steps:

(1) The position parameter of particle can be determined by the connection weights and the threshold value between the nodes of neural network.

(2) Set the values range [Xmin, Xmax] of the connection weights in neural network, and generate corresponding uniform random numbers of particle swarm, then generate the initial swarm.

(3) Evaluate the individuals in the swarm. Decode the individual and assign to the appropriate connection weights (including the threshold value). Introduce the learning samples to calculate the corresponding network output, then get the learning error E, use it as the individual's adaptive value.

(4) Execute the PSO operation on the individuals of the swarm

(5) Judge the PSO operation whether terminate or not? No, turn to step (3), Otherwise, to step (6).

(6) Decode the optimum individual searched by PSO and assign to the weights of neural network (include the threshold value of nodes).

3.3 FNN Algorithm Based on Improved PSO

The improved PSO (mPSO) is an algorithm based on the optimal location variation for the individual of the particle swarm. Compared with the standard PSO, the mPSO prevents the particles from gathering at the optimal location *gbest* quickly by means of individual extreme variation of the swarm, which enhances the diversity of particle swarm.

The algorithm flow of FNN is as follows:

(1) Setting the number of hidden layers and neurons of neural network. Determining the number m of particles, adaptive threshold e, the maximum number *Tmax* of iterative generation; acceleration constants *c1* and *c2*; inertia weight *w*; Initializing the P and V, which are random number between [-1, 1].

(2) Setting the iteration step $t=0$; Calculating the network error and fitness value of each particle according to the given initial value; Setting the optimal fitness value of individual particles, the individual optimal location, the optimal fitness value and location of the particle swarm.

(3) while($J_g > e$ & $t < T_{max}$)

for $i = 1 : m$

Obtaining the weight and threshold value of the neural network from the decoding of x_i and calculating the output of the neural network, compute the value of J_i according to the formula (6):

if $J_i < J_p(i)$ $J_p(i) = J_i$; $p_i = x_i$; end if

if $J_i < J_g$ $J_g = J_i$; $p_g = x_i$; end if

end for

(4) for i=1:m

Calculating the v_i and x_i of particle swarm according to the PSO;

end for

(5) Execute the variation operation on the individual optimal location of the swarm according to the formula (3).

(6) $t=t+1$;

(7) end while

(8) Result output.

3.4 BP NN Algorithm Based on PSO and L-M

Because the traditional BP algorithm has the following problems: slow convergence speed, uncertainty of system training and proneness to local minimum, the improved BP algorithm is most often used in practice. The Levenberg-Marquardt (L-M for short) optimization algorithm is one of the most successful algorithm among the BP algorithms based on derivative optimization. The L-M algorithm is developed from classical Newton algorithm by calculating the derivative in terms of the nonlinear least squares. The iterative formula of LM algorithm is as follows(Zhang ZX, Sun CZ & Mizutani E, 2000):

$$\theta_{n+1} = \theta_n - \eta (J_n^T J_n + \lambda_n I_n)^{-1} J_n^T r_n \tag{7}$$

Where, I is the unit matrix, λ is a non-negative value. Making use of the changes in the amplitude of λ, the method varies smoothly between two extremes, i.e., the Newton method (when $\lambda \to 0$) and standard gradient method (when $\lambda \to \infty$). So the L-M algorithm is actually the combination of standard Newton method and the gradient descent method, which has the advantages of both the latter two methods.

The main idea of the combination algorithm of PSO and L-M (PSOLM algorithm) is to take the PSO algorithm as the main framework. Firstly, optimize the PSO algorithm, after the evolution of several generations, the optimum individual can be chosen from the particle swarm to carry out the optimization search of L-M algorithm for several steps, which operates the local depth search. The specific steps of the algorithm is as follows:

(1) Generate the initial particle swarm X at random, and $k = 0$.
(2) Operate the optimization search on X with the PSO algorithm.
(3) If the evolution generation k of PSO is greater than the given constant dl, chose the optimal individual of particle swarm to carry out the optimization search of L-M algorithm for several steps.
(4) Based on the returned individual, reassess the new optimal individual and global optimal individual by calculating according to PSO algorithm.
(5) If the target function value meets the requirements of precision ε, then terminate the algorithm and output the result; otherwise, $k = k + 1$, turn to step (2).

The above PSO algorithm is actually the particle swarm optimization algorithm (MPSO) by means of the optimal location variation of individual, and the particle number of particle swarm is 30, $c1=c2=1.45$, $w=0.728$.

4. Research on Neural Network Algorithm for Parity Problem

4.1 XOR Problem

Firstly, taking the XOR problem (2 bit parity problem) as an example to discuss it. The XOR problem is one of the classical questions on the NN learning algorithm research, which includes the irregular optimal curved surface as well as many local minimums. The learning sample of XOR problem is shown in Table 3.

Sample	Input	Output
1	00	0
2	01	1
3	10	1
4	11	0

Table 3. Learning sample of XOR

Different network structures result in different learning generations of given precision 10^{-n} (where: n is the accuracy index). In this part, there is a comparison between the learning generations and the actual learning error. The initial weight ranges among [-1, 1] in BP network and conducted 50 random experiments.

As shown in Table4, it displays the experimental results of 2-2-1 NN structure. The activation functions are S-shaped hyperbolic tangent function (Tansig), S-shaped logarithmic function (Logsig) and linear function (Purelin) respectively, and the learning algorithms include the BP, improved BP (BP algorithm with momentum, BPM) and BP based on the Levenberg-Marquardt (BPLM). Judging from the results for XOR problem, as the number of the neurons in the hidden layer is 2, the BP and improved BP (BPM, BPLM) can't converge completely in 50 experiments.

It can also be seen that the performance of the improved BP is better than that of the basic BP, as for the improved BP, the BPLM performs better than BPM. In addition, the initial value of the algorithm has great influence on the convergence property of BP algorithm, so is the function form of the neurons in the output layer.

XOR Problem		BP		BPM		BPLM	
Activation function	Hidden layer	Tansig	Tangsig	Tansig	Tangsig	Tansig	Tangsig
	Output layer	Purelin	Logsig	Purelin	Logsig	Purelin	Logsig
NN structure：2-2-1		56%	0	60%	0	72%	0

Table 4. Convergence statistics of BP, BPM and BPLM (Accuracy index n=3)

The Table 5 shows the training results under different accuracy indices. The activation functions are Tansig-purelin and tansig-logsig respectively, and the NN algorithms include the BPLM and the PSO with limited factor (cPSO, Clerc, M., 1999). It can be indicated that the basic PSO, which is applied to the BP network for XOR problem, can't converge completely, either. In such circumstance, the number of the neurons in the hidden layer is 2.

XOR learning	Accuracy index	BPLM		cPSO			
		Tansig-purelin		Tansig-purelin		Tansig-logsig	
		Average iteration number	Ras	Average iteration number	Ras	Average iteration number	Ras
Network stucture 2-2-1	3	16.36	36	71.00	35	24.71	26
	6	20.84	40	145.94	36	64.88	25
	10	13.76	38	233.36	36	43.52	25
	20	25.99	8	461.13	38	68.21	26

Table 5. BP training results of BPLM, cPSO, and mPSO

Besides, for the BP and the improved BP algorithm, it has never converged in the given number of experiments when the activation function of output layer in NN is Logsig, while

the form of activation function has relatively minor influence on the PSO algorithm. It can be seen from the table that the form of activation function has certain influence on the learning speed of NN algorithm based on PSO, and the learning algorithm, which adopts Tangsig-Logsig function converges faster than that adopts Tangsig-Purelin function.

The Table 6 shows the optimization results of the PSOBP and PSOBPLM algorithm, which are the combination of MPSO and standard BP (PSOBP) as well as the combination of MPSO and BP algorithm based on L-M (PSOBPLM) respectively. As seen in Table 6, for the given number of experiments, the optimization results of the algorithms have all achieved the specified target value within the given iteration number.

XOR	PSOBP			PSOBPLM		
Accuracy index	Average iteration number		Mean time (s/time)	Average iteration number		Mean time (s/time)
	PSO	BP		PSO	BP	
3	11.06	379.15	3.74	21	2.67	0.72
6	11.12	517.35	5.35	21	3.8	0.75
10	11.46	910.5	8.31	21	4.73	0.73
20	12.3	1578.55	13.37	23.07	20.13	1.97

Table 6. BP optimization results of PSOBP and PSOBPLM algorithm

In addition, the Table 6 has also displayed the average iteration number and the mean time of PSO and BP algorithm under different accuracy indices in 50 experiments. As shown in Table 3, the algorithm of PSO combined with BP or LM has good convergence property, which is hard to realize for single BP (including BPLM) or PSO algorithm. It's especially necessary to notice that the combination of the PSO and LM algorithm brings about very high convergence speed, and the algorithm of PSOBPLM converges much faster than PSOBP algorithm under the condition of high accuracy index. For example, when the network structure is 2-2-1 and the accuracy index is 10 and 20 respectively, the relevant mean time of PSOBP algorithm is 8.31 and 13.37, while for the PSOBPLM algorithm, the mean time is reduced to 0.73 and 1.97. Obviously, the PSOBPLM algorithm has excellent speed performance.

4.2 Parity Problem

The parity problem is one of the famous problems in neural network learning and much more complex than the 2bit XOR problem. The learning sample of parity problem consists of 4-8 bit binary string. When the number of 1 in binary string is odd, the output value is 1; otherwise, the value is 0. When the PSO (including the improved PSO) and PSOBP algorithm are applied to solve the parity problem, the learning speed is quite low and it is impossible to converge to the target value in the given iteration number. The PSOBPLM algorithm, proposed in this article, is applied to test the 4-8bit parity problem. The network structure of 4bit parity problem is 4-4-1, and the activation function of both hidden layer and output layer are Tansig-logsig, the same is with the activation function of NN for 5-8bit

parity problem, and the parameter of NN for 5-8bit parity problem can be got from that of NN for 4bit parity problem by analogy. For each parity problem, 50 random experiments are carried out. The Table 7 shows the experimental result of the PSOBPLM algorithm for 4-8bit parity problem under various accuracy indices. In the Table 7, the Mean, Max and Min represent the average iteration number, the maximum and minimum iteration number, respectively. The number below the PSO and BP column represents the iteration number needed by the corresponding algorithm.

net:4-4-1;							
Accuracy index	Mean		Max		Min		Mean time (s/time)
	PSO	LM	PSO	LM	PSO	LM	
3	21.07	67.60	22	489	21	12	1.15
6	21.10	80.77	22	424	21	11	1.19
10	21.17	114.5	22	699	21	14	1.31
20	25.23	405.73	35	1414	21	18	4.62
net:5-5-1;							
3	50.10	99.27	51	532	50	16	1.49
6	50.07	103.17	52	1019	50	19	1.58
10	50.13	143.57	51	557	50	12	1.84
20	53.77	371.07	65	1960	50	27	5.21
net:6-6-1;							
3	50.23	208.93	52	1103	50	23	2.97
6	50.13	204.47	51	591	50	34	2.58
10	50.50	334.57	53	1281	50	42	3.81
20	53.77	944.73	65	3069	50	49	10.72
net:7-7-1;							
3	50.27	267.5	51	708	50	29	4.66
6	50.27	279.7	51	686	50	35	4.64
10	50.33	278.67	52	1067	50	32	4.53
20	52.77	748.57	59	2206	50	57	11.69
net:8-8-1;							
3	50.23	273.53	52	1066	50	56	7.98
6	50.43	391.63	51	803	50	78	8.29
10	51.63	387.27	54	1388	51	71	10.71
20	54.83	1225.47	63	3560	51	65	30.43

Table 7. Result of PSOBPLM algorithm for 4-8 bit parity problem

As seen in Table 7, the integration of the PSO and L-M algorithm can solve the parity problem. The PSOBPLM algorithm makes full use of the advantage of the PSO and L-M algorithm, i.e., the PSO has great skill in global search and the L-M excels in local optimization, which compensate their own drawback and have complementary advantages. So the PSOLM algorithm has not only a good convergence, but also fast optimization property.

5. Conclusion

As a global evolutionary algorithm, the PSO has simple model and is easy to achieve. The integration of the PSO and L-M algorithm makes full use of their own advantage, i.e., the PSO has great skill in global search and the L-M excels in local fast optimization, which could avoid falling into local minimum and find the global optimal solution for the parity problem effectively. Meanwhile, the PSOBPLM algorithm has better efficiency and robustness. The only shortage of the algorithm is that it needs the derivative information, which increases the algorithm complexity to some extent.

6. References

Clerc, M. (1999). The swarm and the queen: towards a deterministic and adaptive particle swarm optimization. *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 1999)*, pp. 1951-1957, Washington, DC, 1999, IEEE Service Center, Piscataway, NJ.

Eberhart, R. C. & Kennedy, J. (1995). A new optimizer using particle swarm theory. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, pp. 39-43, Nagoya, Japan, 1995, IEEE Service Center, Piscataway, NJ.

James L. McClelland, David E. Rumelhart & the PDP Research Group, (1986). *Parallel Distributed Processing*. MIT press, Cambridge, MA

Kennedy, J. & Eberhart, R. C. (1995). Particle swarm optimization, *Proceedings of IEEE International Conference on Neutral Networks*, IV, pp. 1942-1948, Perth, Australia, 1995, IEEE Service Center, Piscataway, NJ.

LU Zhensu & HOU Zhirong (2004). Particle Swarm Optimization with Adaptive Mutation, *ACATA ELECTRONICA SINICA*, Vol. 33, No. 3, 416-420.ISSN: 3972-2112

Wang Ling (2001). *Intelligent Optimization Algorithms With Applications*, Tsinghua University Press, ISBN: 7-302-04499-6, Beijing

Wang Xiaoping & Cao Liming (2002). *Genetic Algorithm-Its Theory,Application and Software Realization*, Xian Jiaotong University Press, ISBN: 7560514480, Xian

Zhang ZX, Sun CZ & Mizutani E (2000). *Neuro-Fuzzy and Soft Computing*, Xian Jiaotong University Press, ISBN: 7560511872, Xian

Zhou Zhihua, Cao cungen (2004). Neural networks and its application, Tsinghua University Press, ISBN: 7302086508, Beijing

Improved State Estimation of Stochastic Systems via a New Technique of Invariant Embedding

Nicholas A. Nechval and Maris Purgailis
University of Latvia
Latvia

1. Introduction

The state estimation of discrete-time systems in the presence of random disturbances and measurement noise is an important field in modern control theory. A significant research effort has been devoted to the problem of state estimation for stochastic systems. Since Kalman's noteworthy paper (Kalman, 1960), the problem of state estimation in linear and nonlinear systems has been treated extensively and various aspects of the problem have been analyzed (McGarty, 1974; Savkin & Petersen, 1998; Norgaard et al., 2000; Yan & Bitmead, 2005; Alamo et al., 2005; Gillijns & De Moor, 2007; Ko & Bitmead, 2007).

The problem of determining an optimal estimator of the state of stochastic system in the absence of complete information about the distributions of random disturbances and measurement noise is seen to be a standard problem of statistical estimation. Unfortunately, the classical theory of statistical estimation has little to offer in general type of situation of loss function. The bulk of the classical theory has been developed about the assumption of a quadratic, or at least symmetric and analytically simple loss structure. In some cases this assumption is made explicit, although in most it is implicit in the search for estimating procedures that have the "nice" statistical properties of unbiasedness and minimum variance. Such procedures are usually satisfactory if the estimators so generated are to be used solely for the purpose of reporting information to another party for an unknown purpose, when the loss structure is not easily discernible, or when the number of observations is large enough to support Normal approximations and asymptotic results. Unfortunately, we seldom are fortunate enough to be in asymptotic situations. Small sample sizes are generally the rule when estimation of system states and the small sample properties of estimators do not appear to have been thoroughly investigated. Therefore, the above procedures of the state estimation have long been recognized as deficient, however, when the purpose of estimation is the making of a specific decision (or sequence of decisions) on the basis of a limited amount of information in a situation where the losses are clearly asymmetric – as they are here.

There exists a class of control systems where observations are not available at every time due to either physical impossibility and/or the costs involved in taking a measurement. For such systems it is realistic to derive the optimal policy of state estimation with some

constraints imposed on the observation scheme.

It is assumed in this paper that there is a constant cost associated with each observation taken. The optimal estimation policy is obtained for a discrete-time deterministic plant observed through noise. It is shown that there is an optimal number of observations to be taken.

The outline of the paper is as follows. A formulation of the problem is given in Section 2. Section 3 is devoted to characterization of estimators. A comparison of estimators is discussed in Section 4. An invariant embedding technique is described in Section 5. A general problem analysis is presented in Section 6. An example is given in Section 7.

2. Problem Statement

To make the above introduction more precise, consider the discrete-time system, which in particular is described by vector difference equations of the following form:

$$\mathbf{x}(k+1) = \mathbf{A}(k+1,k)\mathbf{x}(k) + \mathbf{B}(k)\mathbf{u}(k), \tag{1}$$

$$\mathbf{z}(k) = \mathbf{H}(k)\mathbf{x}(k) + \mathbf{w}(k), \quad k = 1,2,3,\dots , \tag{2}$$

where $\mathbf{x}(k+1)$ is an n vector representing the state of the system at the (k+1)th time instant with initial condition $\mathbf{x}(1)$; $\mathbf{z}(k)$ is an m vector (the observed signal) which can be termed a measurement of the system at the kth instant; $\mathbf{H}(k)$ is an m × n matrix; $\mathbf{A}(k+1,k)$ is a transition matrix of dimension n × n, and $\mathbf{B}(k)$ is an n × p matrix, $\mathbf{u}(k)$ is a p vector, the control vector of the system; $\mathbf{w}(k)$ is a random vector of dimension m (the measurement noise). By repeated use of (1) we find

$$\mathbf{x}(k) = \mathbf{A}(k,j)\mathbf{x}(j) + \sum_{i=j}^{k-1} \mathbf{A}(k,i+1)\mathbf{B}(i)\mathbf{u}(i), \quad j \le k, \tag{3}$$

where the discrete-time system transition matrix satisfies the matrix difference equation,

$$\mathbf{A}(k+1,j) = \mathbf{A}(k+1,k)\mathbf{A}(k,j), \quad \forall k,j, \tag{4}$$

$$\mathbf{A}(k,k) = \mathbf{I}, \quad \mathbf{A}(k,j) = \prod_{i=j}^{k-1} \mathbf{A}(i+1,i). \tag{5}$$

From these properties, it immediately follows that

$$\mathbf{A}^{-1}(k,j) = \mathbf{A}(j,k), \quad \forall k,j, \tag{6}$$

$$\mathbf{A}(\alpha,\beta)\mathbf{A}(\beta,\gamma) = \mathbf{A}(\alpha,\gamma), \quad \forall \alpha,\beta,\gamma. \tag{7}$$

Thus, for j≤k,

$$\mathbf{x}(j) = \mathbf{A}(j,k)\mathbf{x}(k) - \sum_{i=j}^{k-1} \mathbf{A}(j,i+1)\mathbf{B}(i)\mathbf{u}(i). \tag{8}$$

The problem to be considered is the estimation of the state of the above discrete-time system. This problem may be stated as follows. Given the observed sequence, $\mathbf{z}(1)$, …, $\mathbf{z}(k)$,

it is required to obtain an estimator $\mathbf{d}$ of $\mathbf{x}(l)$ based on all available observed data $\mathbf{Z}^k=\{\mathbf{z}(1), \ldots, \mathbf{z}(k)\}$ such that the expected losses (risk function)

$$R(\boldsymbol{\theta},\mathbf{d}) = E_{\boldsymbol{\theta}}\{r(\boldsymbol{\theta},\mathbf{d})\} \tag{9}$$

is minimized, where r($\boldsymbol{\theta}$,$\mathbf{d}$) is a specified loss function at decision point $\mathbf{d}\equiv\mathbf{d}(\mathbf{Z}^k)$, $\boldsymbol{\theta}=(\mathbf{x}(l),\boldsymbol{\omega})$, $\boldsymbol{\omega}$ is an unknown parametric vector of the probability distribution of $\mathbf{w}(k)$, $k\leq l$.
If it is assumed that a constant cost c>0 is associated with each observation taken, the criterion function for the case of k observations is taken to be

$$r_k(\boldsymbol{\theta},\mathbf{d}) = r(\boldsymbol{\theta},\mathbf{d}) + ck. \tag{10}$$

In this case, the optimization problem is to find

$$\min_k \min_{\mathbf{d}} E_{\boldsymbol{\theta}}\{r_k(\boldsymbol{\theta},\mathbf{d})\}, \tag{11}$$

where the inner minimization operation is with respect to $\mathbf{d}\equiv\mathbf{d}(\mathbf{Z}^k)$, when the k observations have been taken, and where the outer minimization operation is with respect to k.

3. Characterization of Estimators

For any statistical decision problem, an estimator (a decision rule) $\mathbf{d}_1$ is said to be equivalent to an estimator (a decision rule) $\mathbf{d}_2$ if R($\boldsymbol{\theta}$,$\mathbf{d}_1$)=R($\boldsymbol{\theta}$,$\mathbf{d}_2$) for all $\boldsymbol{\theta}\in\Theta$, where R(.) is a risk function, Θ is a parameter space,. An estimator $\mathbf{d}_1$ is said to be uniformly better than an estimator $\mathbf{d}_2$ if R($\boldsymbol{\theta}$,$\mathbf{d}_1$) < R($\boldsymbol{\theta}$,$\mathbf{d}_2$) for all $\boldsymbol{\theta}\in\Theta$. An estimator $\mathbf{d}_1$ is said to be as good as an estimator $\mathbf{d}_2$ if R($\boldsymbol{\theta}$,$\mathbf{d}_1$) $\leq$ R($\boldsymbol{\theta}$,$\mathbf{d}_2$) for all $\boldsymbol{\theta}\in\Theta$. However, it is also possible that we may have "$\mathbf{d}_1$ and $\mathbf{d}_2$ are incomparable", that is, R($\boldsymbol{\theta}$,$\mathbf{d}_1$) <R($\boldsymbol{\theta}$,$\mathbf{d}_2$) for at least one $\boldsymbol{\theta}\in\Theta$, and R($\boldsymbol{\theta}$,$\mathbf{d}_1$) > R($\boldsymbol{\theta}$,$\mathbf{d}_2$) for at least one $\boldsymbol{\theta}\in\Theta$. Therefore, this ordering gives a partial ordering of the set of estimators.
An estimator $\mathbf{d}$ is said to be uniformly non-dominated if there is no estimator uniformly better than $\mathbf{d}$. The conditions that an estimator must satisfy in order that it might be uniformly non-dominated are given by the following theorem.
Theorem 1 (*Uniformly non-dominated estimator*). Let (ξ_τ; τ=1,2, ...) be a sequence of the prior distributions on the parameter space Θ. Suppose that ($\mathbf{d}_\tau$;τ=1,2, ...) and (Q($\xi_\tau,\mathbf{d}_\tau$); τ=1,2, ...) are the sequences of Bayes estimators and prior risks, respectively. If there exists an estimator $\mathbf{d}^*$ such that its risk function R($\boldsymbol{\theta}$,$\mathbf{d}^*$), $\boldsymbol{\theta}\in\Theta$, satisfies the relationship

$$\lim_{\tau\to\infty}\left[Q(\xi_\tau,\mathbf{d}^*) - Q(\xi_\tau,\mathbf{d}_\tau)\right] = 0, \tag{12}$$

where

$$Q(\xi_\tau,\mathbf{d}) = \int_{\Theta} R(\boldsymbol{\theta},\mathbf{d})\xi_\tau(d\boldsymbol{\theta}), \tag{13}$$

then $\mathbf{d}^*$ is an uniformly non-dominated estimator.
Proof. Suppose $\mathbf{d}^*$ is uniformly dominated. Then there exists an estimator $\mathbf{d}^{**}$ such that R($\boldsymbol{\theta}$,$\mathbf{d}^{**}$) < R($\boldsymbol{\theta}$,$\mathbf{d}^*$) for all $\boldsymbol{\theta}\in\Theta$. Let

$$\varepsilon = \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[R(\boldsymbol{\theta}, \mathbf{d}^*) - R(\boldsymbol{\theta}, \mathbf{d}^{**}) \right] > 0. \tag{14}$$

Then

$$Q(\xi_\tau, \mathbf{d}^*) - Q(\xi_\tau, \mathbf{d}^{**}) \geq \varepsilon. \tag{15}$$

Simultaneously,

$$Q(\xi_\tau, \mathbf{d}^{**}) - Q(\xi_\tau, \mathbf{d}_\tau) \geq 0, \tag{16}$$

τ=1,2, ..., and

$$\lim_{\tau \to \infty} \left[Q(\xi_\tau, \mathbf{d}^{**}) - Q(\xi_\tau, \mathbf{d}_\tau) \right] \geq 0. \tag{17}$$

On the other hand,

$$Q(\xi_\tau, \mathbf{d}^{**}) - Q(\xi_\tau, \mathbf{d}_\tau) = \left[Q(\xi_\tau, \mathbf{d}^*) - Q(\xi_\tau, \mathbf{d}_\tau) \right] - \left[Q(\xi_\tau, \mathbf{d}^*) - Q(\xi_\tau, \mathbf{d}^{**}) \right]$$

$$\leq \left[Q(\xi_\tau, \mathbf{d}^*) - Q(\xi_\tau, \mathbf{d}_\tau) \right] - \varepsilon \tag{18}$$

and

$$\lim_{\tau \to \infty} \left[Q(\xi_\tau, \mathbf{d}^{**}) - Q(\xi_\tau, \mathbf{d}_\tau) \right] < 0. \tag{19}$$

This contradiction proves that $\mathbf{d}^*$ is an uniformly non-dominated estimator. □

4. Comparison of Estimators

In order to judge which estimator might be preferred for a given situation, a comparison based on some "closeness to the true value" criteria should be made. The following approach is commonly used (Nechval, 1982; Nechval, 1984). Consider two estimators, say, $\mathbf{d}_1$ and $\mathbf{d}_2$ having risk function $R(\boldsymbol{\theta}, \mathbf{d}_1)$ and $R(\boldsymbol{\theta}, \mathbf{d}_2)$, respectively. Then the relative efficiency of $\mathbf{d}_1$ relative to $\mathbf{d}_2$ is given by

$$\text{rel.eff.}_R \left\{ \mathbf{d}_1, \mathbf{d}_2; \boldsymbol{\theta} \right\} = R(\boldsymbol{\theta}, \mathbf{d}_2) / R(\boldsymbol{\theta}, \mathbf{d}_1). \tag{20}$$

When $\text{rel.eff.}_R \left\{ \mathbf{d}_1, \mathbf{d}_2; \boldsymbol{\theta}_0 \right\} < 1$ for some $\boldsymbol{\theta}_0$, we say that $\mathbf{d}_2$ is more efficient than $\mathbf{d}_1$ at $\boldsymbol{\theta}_0$. If $\text{rel.eff.}_R \left\{ \mathbf{d}_1, \mathbf{d}_2; \boldsymbol{\theta} \right\} \leq 1$ for all $\boldsymbol{\theta}$ with a strict inequality for some $\boldsymbol{\theta}_0$, then $\mathbf{d}_1$ is inadmissible relative to $\mathbf{d}_2$.

5. Invariant Embedding Technique

This paper is concerned with the implications of group theoretic structure for invariant performance indexes. We present an invariant embedding technique based on the constructive use of the invariance principle in mathematical statistics. This technique allows one to solve many problems of the theory of statistical inferences in a simple way. The aim of the present paper is to show how the invariance principle may be employed in the particular case of finding the improved statistical decisions. The technique used here is a special case of more general considerations applicable whenever the statistical problem is invariant under a group of transformations, which acts transitively on the parameter space.

5.1 Preliminaries

Our underlying structure consists of a class of probability models (X, A, P), a one-one mapping ψ taking P onto an index set Θ, a measurable space of actions (U, B), and a real-valued function r defined on $\Theta \times U$. We assume that a group G of one-one A - measurable transformations acts on X and that it leaves the class of models (X, A, P) invariant. We further assume that homomorphic images $\overline{G}$ and $\widetilde{G}$ of G act on Θ and U, respectively. ($\overline{G}$ may be induced on Θ through ψ; $\widetilde{G}$ may be induced on U through r). We shall say that r is invariant if for every $(\boldsymbol{\theta}, \mathbf{u}) \in \Theta \times U$

$$r(\overline{g}\boldsymbol{\theta}, \widetilde{g}\mathbf{u}) = r(\boldsymbol{\theta}, \mathbf{u}), g \in G. \tag{21}$$

Given the structure described above there are aesthetic and sometimes admissibility grounds for restricting attention to decision rules $\boldsymbol{\varphi} : X \to U$ which are (G, $\widetilde{G}$) equivariant in the sense that

$$\boldsymbol{\varphi}(g\mathbf{x}) = \widetilde{g}\boldsymbol{\varphi}(\mathbf{x}), \quad \mathbf{x} \in X, \quad g \in G. \tag{22}$$

If $\overline{G}$ is trivial and (21), (22) hold, we say $\boldsymbol{\varphi}$ is *G-invariant*, or simply invariant (Nechval et al., 2001; Nechval et al., 2003; Nechval & Vasermanis, 2004).

5.2 Invariant Functions

We begin by noting that *r* is invariant in the sense of (21) if and only if r is a $G^{\bullet}$-invariant function, where $G^{\bullet}$ is defined on $\Theta \times U$ as follows: to each $g \in G$, with homomorphic images $\overline{g}, \widetilde{g}$ in $\overline{G}, \widetilde{G}$ respectively, let $g^{\bullet}(\boldsymbol{\theta}, \mathbf{u}) = (\overline{g}\boldsymbol{\theta}, \widetilde{g}\mathbf{u})$, $(\boldsymbol{\theta}, \mathbf{u}) \in (\Theta \times U)$. It is assumed that $\widetilde{G}$ is a homomorphic image of $\overline{G}$.

Definition 1 (*Transitivity*). A transformation group $\overline{G}$ acting on a set Θ is called (uniquely) transitive if for every $\boldsymbol{\theta}$, $\vartheta \in \Theta$ there exists a (unique) $\overline{g} \in \overline{G}$ such that $\overline{g}\boldsymbol{\theta} = \vartheta$. When $\overline{G}$ is transitive on Θ we may index $\overline{G}$ by Θ: fix an arbitrary point $\boldsymbol{\theta} \in \Theta$ and define $\overline{g}_{\boldsymbol{\theta}_1}$ to be the unique $\overline{g} \in \overline{G}$ satisfying $\overline{g}\boldsymbol{\theta} = \boldsymbol{\theta}_1$. The identity of $\overline{G}$ clearly corresponds to $\boldsymbol{\theta}$. An immediate consequence is Lemma 1.

Lemma 1 (*Transformation*). Let $\overline{G}$ be transitive on Θ. Fix $\boldsymbol{\theta} \in \Theta$ and define $\overline{g}_{\boldsymbol{\theta}_1}$ as above. Then $\overline{g}_{\overline{q}\boldsymbol{\theta}_1} = \overline{q}\overline{g}_{\boldsymbol{\theta}_1}$ for $\boldsymbol{\theta} \in \Theta$, $\overline{q} \in \overline{G}$.

Proof. The identity $\overline{g}_{\overline{q}\boldsymbol{\theta}_1}\boldsymbol{\theta} = \overline{q}\boldsymbol{\theta}_1 = \overline{q}\overline{g}_{\boldsymbol{\theta}_1}\boldsymbol{\theta}$ shows that $\overline{g}_{\overline{q}\boldsymbol{\theta}_1}$ and $\overline{q}\overline{g}_{\boldsymbol{\theta}_1}$ both take $\boldsymbol{\theta}$ into $\overline{q}\boldsymbol{\theta}_1$, and the lemma follows by unique transitivity. □

Theorem 2 (*Maximal invariant*). Let $\overline{G}$ be transitive on Θ. Fix a reference point $\boldsymbol{\theta}_0 \in \Theta$ and index $\overline{G}$ by Θ. A maximal invariant M with respect to $G^{\bullet}$ acting on $\Theta \times U$ is defined by

$$M(\boldsymbol{\theta}, \mathbf{u}) = \widetilde{g}_{\boldsymbol{\theta}}^{-1}\mathbf{u}, \quad (\boldsymbol{\theta}, \mathbf{u}) \in \Theta \times U. \tag{23}$$

Proof. For each $(\boldsymbol{\theta}, \mathbf{u}) \in (\Theta \times U)$ and $\overline{g} \in \overline{G}$

$$M(\overline{g}\boldsymbol{\theta}, \widetilde{g}\mathbf{u}) = (\widetilde{g}_{\overline{g}\boldsymbol{\theta}}^{-1})\widetilde{g}\mathbf{u} = (\widetilde{g}\widetilde{g}_{\boldsymbol{\theta}})^{-1}\widetilde{g}\mathbf{u} = \widetilde{g}_{\boldsymbol{\theta}}^{-1}\widetilde{g}^{-1}\widetilde{g}\mathbf{u} = \widetilde{g}_{\boldsymbol{\theta}}^{-1}\mathbf{u} = M(\boldsymbol{\theta}, \mathbf{u}) \tag{24}$$

by Lemma 1 and the structure preserving properties of homomorphisms. Thus M is $G^{\bullet}$-invariant. To see that M is maximal, let $M(\boldsymbol{\theta}_1, \mathbf{u}_1) = M(\boldsymbol{\theta}_2, \mathbf{u}_2)$. Then $\widetilde{g}_{\boldsymbol{\theta}_1}^{-1}\mathbf{u}_1 = \widetilde{g}_{\boldsymbol{\theta}_2}^{-1}\mathbf{u}_2$ or $\mathbf{u}_1 = \widetilde{g}\,\mathbf{u}_2$, where $\widetilde{g} = \widetilde{g}_{\boldsymbol{\theta}_1}\widetilde{g}_{\boldsymbol{\theta}_2}^{-1}$. Since $\boldsymbol{\theta}_1 = \overline{g}_{\boldsymbol{\theta}_1}\boldsymbol{\theta}_0 = \overline{g}_{\boldsymbol{\theta}_1}\overline{g}_{\boldsymbol{\theta}_2}^{-1}\boldsymbol{\theta}_2 = \overline{g}\boldsymbol{\theta}_2$, $(\boldsymbol{\theta}_1, \mathbf{u}_1) = g^{\bullet}(\boldsymbol{\theta}_2, \mathbf{u}_2)$ for some $g^{\bullet} \in G^{\bullet}$, and the proof is complete. □

Corollary 2.1 (*Invariant embedding*). An invariant function, $r(\boldsymbol{\theta}, \mathbf{u})$, can be transformed as follows:

$$r(\boldsymbol{\theta}, \mathbf{u}) = r(\overline{g}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\theta}, \widetilde{g}_{\boldsymbol{\theta}}^{-1}\mathbf{u}) = \ddot{r}(\mathbf{v}, \boldsymbol{\eta}), \tag{25}$$

where $\mathbf{v} = \mathbf{v}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is a function (it is called a pivotal quantity) such that the distribution of $\mathbf{v}$ does not depend on $\boldsymbol{\theta}$; $\boldsymbol{\eta} = \boldsymbol{\eta}(\mathbf{u}, \hat{\boldsymbol{\theta}})$ is an ancillary factor; $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ (or the sufficient statistic for $\boldsymbol{\theta}$).

Corollary 2.2 (*Best invariant decision rule*). If $r(\boldsymbol{\theta}, \mathbf{u})$ is an invariant loss function, the best invariant decision rule is given by

$$\varphi^*(\mathbf{x}) = \mathbf{u}^* = \boldsymbol{\eta}^{-1}(\boldsymbol{\eta}^*, \hat{\boldsymbol{\theta}}), \tag{26}$$

where

$$\boldsymbol{\eta}^* = \arg \inf_{\boldsymbol{\eta}} E_{\boldsymbol{\eta}} \{\ddot{r}(\mathbf{v}, \boldsymbol{\eta})\}. \tag{27}$$

Corollary 2.3 (*Risk*). A risk function (performance index)

$$R(\boldsymbol{\theta}, \varphi(\mathbf{x})) = E_{\boldsymbol{\theta}} \{r(\boldsymbol{\theta}, \varphi(\mathbf{x}))\} = E_{\boldsymbol{\eta}_\circ} \{\ddot{r}(\mathbf{v}_\circ, \boldsymbol{\eta}_\circ)\} \tag{28}$$

is constant on orbits when an invariant decision rule $\varphi(\mathbf{x})$ is used, where $\mathbf{v}_\circ = \mathbf{v}_\circ(\boldsymbol{\theta}, \mathbf{x})$ is a function whose distribution does not depend on $\boldsymbol{\theta}$; $\boldsymbol{\eta}_\circ = \boldsymbol{\eta}_\circ(\mathbf{u}, \mathbf{x})$ is an ancillary factor.

For instance, consider the problem of estimating the location-scale parameter of a distribution belonging to a family generated by a continuous cdf F: P = {$P_{\boldsymbol{\theta}}$: $F((x-\mu)/\sigma)$, $x \in R$, $\boldsymbol{\theta} \in \Theta$}, $\Theta = \{(\mu, \sigma): \mu, \sigma \in R, \sigma > 0\} = U$. The group G of location and scale changes leaves the class of models invariant. Since $\overline{G}$ induced on Θ by $P_{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ is uniquely transitive, we may apply Theorem 1 and obtain invariant loss functions of the form

$$r(\boldsymbol{\theta}, \varphi(x)) = r[(\varphi_1(x) - \mu)/\sigma, \varphi_2(x)/\sigma], \tag{29}$$

where

$$\boldsymbol{\theta} = (\mu, \sigma) \text{ and } \varphi(x) = (\varphi_1(x), \varphi_2(x)). \tag{30}$$

Let $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\sigma})$ and $\mathbf{u} = (u_1, u_2)$, then

$$r(\boldsymbol{\theta}, \mathbf{u}) = \ddot{r}(\mathbf{v}, \boldsymbol{\eta}) = \ddot{r}(v_1 + \eta_1 v_2, \eta_2 v_2), \tag{31}$$

where

$$\mathbf{v}=(v_1,v_2), \; v_1=(\hat{\mu}-\mu)/\sigma \, , \; v_2=\hat{\sigma}/\sigma \; ; \tag{32}$$

$$\boldsymbol{\eta}=(\eta_1,\eta_2), \; \eta_1=(u_1-\hat{\mu})/\hat{\sigma} \, , \; \eta_2=u_2/\hat{\sigma} \, . \tag{33}$$

5.3 Illustrative Example 1

Consider an inventory manager faced with a one-period Christmas-tree stocking problem. Assume the decision maker has demand data on the sale of trees over the last n seasons. For the sake of simplicity, we shall consider the case where the demand data can be measured on a continuous scale. We restrict attention to the case where these demand values constitute independent observations from a distribution belonging to invariant family. In particular, we consider a distribution belonging to location-scale family generated by a continuous cdf F: P ={P_θ: F((x-μ)/σ), x∈R, θ∈Θ}, Θ={(μ,σ): μ,σ∈R, σ>0}, which is indexed by the vector parameter θ=(μ,σ), where μ and σ (>0) are respectively parameters of location and scale. The group G of location and scale changes leaves the class of models invariant. The purpose in restricting attention to such families of distributions is that for such families the decision problem is invariant, and if the estimators of safety stock levels are equivariant (i.e. the group of location and scale changes leaves the decision problem invariant), then any comparison of estimation procedures is independent of the true values of any unknown parameters. The common distributions used in inventory problems are the normal, exponential, Weibull, and gamma distributions.

Let us assume that, for one reason or another, a 100γ% service level is desired (i.e. the decision maker wants to ensure that at least 100γ% of his customers are satisfied). If the demand distribution is completely specified, the appropriate amount of inventory to stock for the season is u satisfying

$$\Pr\{X \le u\} = F\left(\frac{u-\mu}{\sigma}\right) = \gamma \tag{34}$$

or

$$u = \mu + p_\gamma \sigma, \tag{35}$$

where

$$p_\gamma = F^{-1}(\gamma) \tag{36}$$

is the γth percentile of the above distribution. Since the inventory manager does not know μ or σ, the estimator commonly used to estimate u is the maximum likelihood estimator

$$\hat{u} = \hat{\mu} + p_\gamma \hat{\sigma}, \tag{37}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the maximum likelihood estimators of the parameters μ and σ, respectively. This estimator is one possible estimator of u and it may yield poor results.

The correct procedure for estimating u requires establishing a tolerance limit for the percentile. It should be noted that tolerance limits are to percentiles what confidence limits are to parameters. With confidence limits, inferences may be drawn on parameters, whereas with tolerance limits, inferences may be drawn about proportions of a distribution. There

are two criteria for establishing tolerance limits. The first criterion establishes an interval such that the expected percentage of observations falling into the interval just exceeds 100γ% (Hahn & Nelson, 1973). This interval is called the 100γ% expectation interval. The second criterion establishes an interval, which ensures that 100γ% of the population is covered with confidence 1-α (Barlow & Proshan, 1966). Such an interval is called a 100γ% content tolerance interval at level 1-α. The decision as to which interval to construct depends on the nature of the problem. A precision-instrument manufacturer wanting to construct an interval which, with high confidence, contains 90% of the distribution of diameters, for example, would use a 90% content tolerance interval, whereas an inventory manager wanting to stock sufficient items to ensure that in the long run an average of 95% of demand will be satisfied may find expectation intervals more appropriate. Expectation intervals are only appropriate in inventory problems where average service levels are to be controlled. Tolerance limits of the types mentioned above are considered in this subsection. That is, if $f(x;\boldsymbol{\theta})$ denotes the density function of the parent population under consideration and if S is any statistic obtained from a random sample of that population, then $\hat{u}^{\circ} \equiv \hat{u}^{\circ}(S)$ is a lower $100(1-\gamma)\%$ expectation limit if

$$\Pr\left\{X > \hat{u}^{\circ}\right\} = E_{\boldsymbol{\theta}}\left\{\int_{\tilde{u}^{\circ}}^{\infty} f(x;\boldsymbol{\theta})dx\right\} = E_{\boldsymbol{\theta}}\left\{1 - F\left(\frac{\hat{u}^{\circ} - \mu}{\sigma}\right)\right\} = 1 - \gamma. \tag{38}$$

This expression represents a risk of $\hat{u}^{\circ}$, i.e.

$$R^{\circ}(\boldsymbol{\theta}, \hat{u}^{\circ}) = \Pr\left\{X > \hat{u}^{\circ}\right\} = 1 - \gamma. \tag{39}$$

A lower $100(1-\gamma)\%$ content tolerance limit at level 1-α, $\hat{u}^{\bullet} \equiv \hat{u}^{\bullet}(S)$, is defined by

$$\Pr\left\{\int_{\tilde{u}^{\bullet}}^{\infty} f(x;\boldsymbol{\theta})dx \le 1 - \gamma\right\} = \Pr\left\{F\left(\frac{\hat{u}^{\bullet} - \mu}{\sigma}\right) \ge \gamma\right\} = \Pr\left\{\hat{u}^{\bullet} \ge \mu + p_{\gamma}\sigma\right\} = 1 - \alpha. \tag{40}$$

A risk of this limit is

$$R^{\bullet}(\boldsymbol{\theta}, \hat{u}^{\bullet}) = 1 - \Pr\left\{\hat{u}^{\bullet} \ge \mu + p_{\gamma}\sigma\right\} = \alpha. \tag{41}$$

Since it is often desirable to have statistical tolerance limits available for the distributions used to describe demand data in inventory control, the problem is to find these limits. We give below a general procedure for obtaining tolerance limits. This procedure is based on the use of an invariant embedding technique given above.

Lower 100(1-γ)% expectation limit. Suppose $X_1, ..., X_n$ are a random sample from the exponential distribution, with pdf

$$f(x;\sigma) = \frac{1}{\sigma}\exp(-x/\sigma), \quad x \ge 0, \tag{42}$$

where σ>0 is unknown parameter. Let

$$S_n = \sum_{i=1}^{n} X_i. \tag{43}$$

It can be justified by using the factorization theorem that S_n is a sufficient statistic for σ. We wish, on the basis of the sufficient statistic S_n for σ, to construct the lower $100(1-\gamma)\%$ expectation limit for a stock level. It follows from (38) that this limit is defined by

$$\Pr\left\{X > \widehat{u}^\circ\right\} = E_\sigma\left\{\int_{\widehat{u}^\circ}^\infty f(x;\sigma)dx\right\} = E_\sigma\left\{\exp(-\widehat{u}^\circ/\sigma)\right\} = 1 - \gamma. \tag{44}$$

where $\widehat{u}^\circ \equiv \widehat{u}^\circ(S_n)$. Using the technique of invariant embedding of S_n in a maximal invariant

$$M = \widehat{u}^\circ/\sigma, \tag{45}$$

we reduce (44) to

$$E_\sigma\left\{\exp(-\widehat{u}^\circ/\sigma)\right\} = E\left\{\exp(-\eta^\circ V); \eta^\circ\right\} = 1 - \gamma. \tag{46}$$

where

$$V = S_n/\sigma \tag{47}$$

is the pivotal quantity whose distribution does not depend on unknown parameter σ,

$$\eta^\circ = \widehat{u}^\circ/S_n. \tag{48}$$

is an ancillary factor. It is well known that the probability density function of V is given by

$$h(v) = \frac{1}{\Gamma(n)} v^{n-1} \exp(-v), \quad v \geq 0. \tag{49}$$

Thus, for this example, $\widehat{u}^\circ$ can be found explicitly as

$$\widehat{u}^\circ = \eta^\circ S_n, \tag{50}$$

where (see (46))

$$\eta^\circ = \left(\frac{1}{1-\gamma}\right)^n - 1. \tag{51}$$

If the parameters μ and σ were known, it follows from (44) that

$$u = p_\gamma \sigma, \tag{52}$$

where

$$p_\gamma = \ln\left(\frac{1}{1-\gamma}\right). \tag{53}$$

The maximum likelihood estimator of u is given by

$$\widehat{u} = p_\gamma \widehat{\sigma}, \tag{54}$$

where

$$\widehat{\sigma} = S_n/n \tag{55}$$

is the maximum likelihood estimator of the parameter σ.

One can see that each of the above estimators is a member of the class

$$C = \left\{ \hat{d} : \hat{d} = kS_n \right\}, \tag{56}$$

where k is a non-negative real number. A risk of an estimator, which belongs to the class C, is given by

$$R^\circ(\sigma, \hat{d}) = \left(\frac{1}{k+1} \right)^n. \tag{57}$$

Then the relative efficiency of $\hat{d}$ relative to $\hat{u}^\circ$ is given by

$$\text{rel.eff.}_{R^\circ} \left\{ \hat{d}, \hat{u}^\circ ; \sigma \right\} = R^\circ(\sigma, \hat{u}^\circ) / R^\circ(\sigma, \hat{d}) = (1-\gamma)(1+k)^n. \tag{58}$$

If, say,

$$k = p_\gamma / n = n^{-1} \ln(1/(1-\gamma)), \tag{59}$$

n=2 and γ=0.95, then the relative efficiency of the maximum likelihood estimator, $\hat{u}$, relative to $\hat{u}^\circ$ is given by

$$\text{rel.eff.}_{R^\circ} \left\{ \hat{u}, \hat{u}^\circ ; \sigma \right\} = (1-\gamma) \left[1 + n^{-1} \ln(1/(1-\gamma)) \right]^n = 0.312. \tag{60}$$

Lower 100(1-γ)% content tolerance limit at level 1-α. Now we wish, on the basis of a sufficient statistic S_n for σ, to construct the lower 100(1-γ)% content tolerance limit at level 1-α for the size of the stock in order to ensure an adequate service level. It follows from (40) that this tolerance limit is defined by

$$\Pr \left\{ \int_{\hat{u}^\bullet}^{\infty} f(x;\sigma)dx \le 1-\gamma \right\} = \Pr \left\{ F\left(\frac{\hat{u}^\bullet}{\sigma} \right) \ge \gamma \right\} = \Pr \left\{ \hat{u}^\bullet \ge p_\gamma \sigma \right\} = 1-\alpha. \tag{61}$$

By using the technique of invariant embedding of S_n in a maximal invariant

$$M = \hat{u}^\bullet / \sigma, \tag{62}$$

we reduce (61) to

$$\Pr \left\{ \hat{u}^\bullet \ge p_\gamma \sigma \right\} = \Pr \left\{ V \ge p_\gamma / \eta^\bullet \right\} = 1-\alpha. \tag{63}$$

where $\hat{u}^\bullet \equiv \hat{u}^\bullet(S_n)$,

$$\eta^\bullet = \hat{u}^\bullet / S_n \tag{64}$$

is an ancillary factor.

It follows from the above that, in this case, $\hat{u}^\bullet$ can be found explicitly as

$$\hat{u}^\bullet = \eta^\bullet S_n, \tag{65}$$

where

$$\eta^\bullet = \frac{2p_\gamma}{\chi_\alpha^2(2n)} = \frac{2\ln(1/(1-\gamma))}{\chi_\alpha^2(2n)}, \tag{66}$$

$\chi_\alpha^2(2n)$ is the $100\alpha\%$ point of the chi-square distribution with $2n$ degrees of freedom. Since the estimator $\hat{u}^\bullet$ belongs to the class C, then the relative efficiency of $\hat{d} \in \mathcal{C}$ relative to $\hat{u}^\bullet$ is given by

$$\text{rel.eff.}_{R^\bullet}\left\{\hat{d},\hat{u}^\bullet;\sigma\right\}= R^\bullet(\sigma,\hat{u}^\bullet)/R^\bullet(\sigma,\hat{d}) = \alpha\left[1 - \Pr\left\{\chi^2(2n) \geq \frac{2p_\gamma}{k}\right\}\right]^{-1}. \tag{67}$$

If, say, k is given by (59), n=2 and α=0.05, then we have that the relative efficiency of the maximum likelihood estimator, $\hat{u}$, relative to $\hat{u}^\bullet$ is given by

$$\text{rel.eff.}_{R^\bullet}\left\{\hat{u},\hat{u}^\bullet;\sigma\right\}= \alpha\left[1 - \Pr\left\{\chi^2(2n) \geq 2n\right\}\right]^{-1} =0.084. \tag{68}$$

5.4 Illustrative Example 2

Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(k)}$ be the k smallest observations in a sample of size n from the two-parameter exponential distribution, with density

$$f(x;\boldsymbol{\theta}) = \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right), \quad x \geq \mu, \tag{69}$$

where σ>0 and μ are unknown parameters, $\boldsymbol{\theta}$=(μ,σ).

Let $Y_{(r)}$ be the rth smallest observation in a future sample of size m from the same distribution. We wish, on the basis of observed $X_{(1)}$, ..., $X_{(k)}$ to construct prediction intervals for $Y_{(r)}$. Let

$$S_r=(Y_{(r)}-\mu)/\sigma, \quad S_1=(X_{(1)}-\mu)/\sigma \tag{70}$$

and

$$T_1=T/\sigma, \tag{71}$$

where

$$T = \sum_{i=1}^{k}(X_{(i)} - X_{(1)}) + (n-k)(X_{(k)} - X_{(1)}). \tag{72}$$

To construct prediction intervals for $Y_{(r)}$, consider the quantity (invariant statistic)

$$V = n(S_r–S_1)/T_1= n(Y_{(r)}–X_{(1)})/T. \tag{73}$$

It is well known (Epstein & Sobel, 1954) that nS_1 has a standard exponential distribution, that $2T_1 \sim \chi_{2k-2}^2$ and that S_1 and T_1 are independent. Also, S_r is the rth order statistic from a sample of size m from the standard exponential distribution and thus has probability density function (Kendall & Stuart, 1969),

$$f(s_r) = r\binom{m}{r}(1 - e^{-s_r})^{r-1}e^{-s_r(m-r+1)}, \tag{74}$$

if s_r>0, and $f(s_r)$=0 for $s_r\leq0$. Using the technique of invariant embedding, we find after some algebra that

$$F(v)=Pr\{V \le v\} = \begin{cases} 1 - nr\binom{m}{r}\sum_{j=0}^{r-1}\dfrac{\binom{r-1}{j}(-1)^j[1+v(m-r+j+1)/n]^{-k+1}}{(m+n-r+j+1)(m-r+j+1)}, & v > 0, \\ m^{(r)}(1-v)^{-k+1}/(m+n)^{(r)}, & v \le 0, \end{cases}$$ (75)

where $m^{(r)}=m(m–1)\cdots(m–r+1)$.

The special case in which r=1 is worth mentioning, since in this case (75) simplifies somewhat. We find here that we can write

$$F(v)=Pr\{V \le v\} = \begin{cases} 1 - \dfrac{\vartheta}{\vartheta+1}\left(\dfrac{\vartheta}{\vartheta+v}\right)^{k-1}, & v > 0, \\ (\vartheta+1)^{-1}(1-v)^{-k+1}, & v \le 0, \end{cases}$$ (76)

where $\vartheta=n/m$.

Consider the ordered data given by Grubbs (Grubbs, 1971) on the mileages at which nineteen military carriers failed. These were 162, 200, 271, 302, 393, 508, 539, 629, 706, 777, 884, 1008, 1101, 1182, 1463, 1603, 1984, 2355, 2880, and thus constitute a complete sample with k=n=19. We find

$$T = \sum_{i=1}^{19}(X_{(i)} - X_{(1)})) = 15869$$ (77)

and of course $X_{(1)}=162$.

Suppose we wish to set up the shortest-length ($1-\alpha=0.95$) prediction interval for the smallest observation $Y_{(1)}$ in a future sample of size m=5. Consider the invariant statistic

$$V = \frac{n(Y_{(1)} - X_{(1)})}{T}.$$ (78)

Then

$$Pr\left\{v_1 < \frac{n(Y_{(1)} - X_{(1)})}{T} < v_2\right\} = Pr\left\{X_{(1)} + v_1\frac{T}{n} < Y_{(1)} < X_{(1)} + v_2\frac{T}{n}\right\} = Pr\{z_L < Y_{(1)} < z_U\}=1-\alpha,$$ (79)

where

$$z_L=X_{(1)}+v_1T/n$$ (80)

and

$$z_U=X_{(1)}+v_2T/n.$$ (81)

The length of the prediction interval is

$$\Delta_z = z_U - z_L = (T/n)(v_2 - v_1).$$ (82)

We wish to minimize Δ_z subject to

$$F(v_2)-F(v_1)=1-\alpha. \tag{83}$$

It can be shown that the minimum occurs when

$$f(v_1)=f(v_2), \tag{84}$$

where v_1 and v_2 satisfy (83). The shortest-length prediction interval is given by

$$C^*_{Y_{(1)}}(X_{(1)},T) = \left(X_{(1)} + v_1^* \frac{T}{n}, X_{(1)} + v_2^* \frac{T}{n}\right) = (10.78,\ 736.62), \tag{85}$$

where $v_1^* = -0.18105$ and $v_2^* = 0.688$. Thus, the length of this interval is $\Delta_z^* = 736.62 - 10.78$ =725.84.

The equal tails prediction interval at the $1-\alpha=0.95$ confidence level is given by

$$C^\circ_{Y_{(1)}}(X_{(1)},T) = \left(X_{(1)} + v_{\alpha/2} \frac{T}{n}, X_{(1)} + v_{1-\alpha/2} \frac{T}{n}\right) = (57.6,\ 834.34), \tag{86}$$

where $F(v_\alpha)=\alpha$, $v_{\alpha/2}= -0.125$ and $v_{1-\alpha/2}= 0.805$. The length of this interval is $\Delta_z^\circ = 834.34 - 57.6$ = 776.74.

The relative efficiency of $C^\circ_{Y_{(1)}}(X_{(1)},T)$ relative to $C^*_{Y_{(1)}}(X_{(1)},T)$, taking into account Δ_z is given by

$$\text{rel.eff.}_{E_\theta\{\Delta_z\}}\left(C^\circ_{Y_{(1)}}(X_{(1)},T), C^*_{Y_{(1)}}(X_{(1)},T)\right) = \frac{\Delta_z^*}{\Delta_z^\circ} = \frac{v_2^* - v_1^*}{v_{1-\alpha/2} - v_{\alpha/2}} = 0.934. \tag{87}$$

One may also be interested in predicting the mean

$$\overline{Y} = \sum_{j=1}^{m} Y_j / m \tag{88}$$

or total lifetime in a future sample. Consider the quantity

$$V = n(\overline{Y} - X_{(1)})/T. \tag{89}$$

Using the invariant embedding technique, we find after some algebra that

$$F(v)=\Pr\{V\leq v\} = \begin{cases} 1 - \displaystyle\sum_{j=0}^{m-1} \binom{k+j-2}{j} \frac{(v/9)^j[1-(1+9)^{-m+j}]}{(1+v/9)^{k+j-1}}, & v>0, \\ (1+9)^{-m}(1-v)^{-k+1}, & v\leq 0. \end{cases} \tag{90}$$

Probability statements about V lead to prediction intervals for $\overline{Y}$ or

$$\sum_{j=1}^{m} Y_j = m\overline{Y}. \tag{91}$$

5.5 Illustrative Example 3

Suppose that $X_1, ..., X_n$ and $Y_{1i}, ..., Y_{mi}$ (i=1, ..., k) denote n+km independent and identically distributed random variables from a two-parameter exponential distribution with pdf (69), where $\sigma > 0$ and μ are unknown parameters.

Let $X_{(1)}$ be the smallest observation in the initial sample of size n and

$$S_n = \sum_{j=1}^{n} \left(X_j - X_{(1)} \right). \tag{92}$$

It can be justified by using the factorization theorem that $(X_{(1)}, S_n)$ is a sufficient statistic for (μ, σ). Let $Y_{(1i)}$ be the smallest observation in the ith future sample of size m, $\forall i=1(1)k$. We wish, on the basis of a sufficient statistic $(X_{(1)}, S_n)$ for (μ, σ), to construct simultaneous lower one-sided β-content tolerance limits at level γ for $Y_{(1i)}$, i=1, ..., k. It can be shown that this problem is reduced to the problem of constructing a lower one-sided β-content tolerance limit at level γ, $L \equiv L(X_{(1)}, S_n)$, for

$$Y_{(1)} = \min_{1 \le i \le k} Y_{(1i)}. \tag{93}$$

This tolerance limit is defined by

$$\Pr\left\{ \int_{L}^{\infty} f(y_{(1)}; \mu, \sigma) dy_{(1)} \ge \beta \right\} = \Pr\left\{ \frac{L-\mu}{\sigma} \le -\ln \beta^{\frac{1}{km}} \right\} = \gamma. \tag{94}$$

By using the technique of invariant embedding of $(X_{(1)}, S_n)$ into a maximal invariant $M = (L-\mu)/\sigma$, we reduce (94) to

$$\Pr\left\{ V_1 \le -\eta V_2 - \ln \beta^{\frac{1}{km}} \right\} = \gamma, \tag{95}$$

where

$$V_1 = \frac{X_{(1)} - \mu}{\sigma}, \quad V_2 = \frac{S_n}{\sigma} \tag{96}$$

are the pivotal quantities,

$$\eta = \frac{L - X_{(1)}}{S_n} \tag{97}$$

is the ancillary factor. It follows from (95) that

$$1 - \frac{\beta^{\frac{n}{km}}}{(1-\eta n)^{n-1}} = \gamma. \tag{98}$$

Therefore, in this case, L can be found explicitly as

$$L = X_{(1)} + \frac{S_n}{n}\left[1 - \left(\frac{\beta^{\frac{n}{km}}}{1-\gamma}\right)^{\frac{1}{n-1}}\right].$$

(99)

For instance, let us suppose that shipments of a lot of electronic systems of a specified type are made to each of 3 customers. Further suppose each customer selects a random sample of 5 systems and accepts his shipment only if no failures occur before a specified time has elapsed. The manufacturer wishes to take a random sample and to calculate the simultaneous lower one-sided β-content tolerance limits so that all shipments will be accepted with a probability of γ at least for 100β% of the future cases of such k shipments, where β=0.95, γ=0.95, and k=3. The resulting failure times (rounded off to the nearest hour) of an initial sample of size 20 from a population of such electronic systems are: 3149, 3407, 3215, 3296, 3095, 3563, 3178, 3112, 3086, 3160, 3155, 3742, 3143, 3240, 3184, 3621, 3125, 3109, 3118, 3127. It is assumed that the failure times follow a two-parameter exponential distribution with unknown parameters μ and σ. Thus, for this example, n=20, k=3, m=5, β=0.95, γ=0.95, $X_{(1)}$=3086, and S_n=3105.
The manufacturer finds from (99) that

$$L = 3086 + \frac{3105}{20}\left[1 - \left(\frac{(0.95)^{20/15}}{1-0.95}\right)^{\frac{1}{19}}\right] = 3060.$$

(100)

and he has 95% assurance that no failures will occur in each shipment (i.e. each shipment will be accepted) before L=3060 hours at least for 95% of the future cases of such shipments of a lot of electronic systems which will be made to each of three firms.

5.6 Illustrative Example 4
Consider the problem of finding shortest-length confidence interval for system availability. Availability is very important to users of repairable products and systems, such as computer networks, manufacturing systems, power plants, transportation vehicles, and fire-protection systems. Mathematically, the availability of an item is a measure of the fraction of time that the item is in operating condition in relation to total or calendar time, i.e., availability indicates the percent of the time that products are expected to operate satisfactory. There are several measures of availability, namely, inherent availability, achieved availability, and operational availability. For further definition of these availability measures, see (Ireson & Coombs, 1988). Here, we consider inherent availability, which is the most common definition used in the literature. This availability, A, is the designed-in capability of a product and is defined by (Ben-Daya et al., 2000)

$$A = MTBF/(MTBF + MTTR),$$

(101)

where MTTR is the Mean Time To Repair (more generally, the mean time that the process is inoperable when it is down for maintenance or because of a breakdown) and MTBF is the

Mean Time Between Failures (more generally, the mean operating time between one downtime and the next, where each downtime can be due to maintenance or a breakdown). Actually the true inherent availability is rarely known. Usually, it is estimated from the few collected data on the operating (up) times and repair/replace (down) times. The point estimate of the availability is then given by

$$\hat{A} = \hat{MTBF}/(\hat{MTBF} + \hat{MTTR}), \tag{102}$$

where $\hat{A}$ is an estimate of the inherent availability, $\hat{MTBF}$ is an estimate of MTBF from sample data, $\hat{MTTR}$ is an estimate of MTTR from sample data. Obviously, this point estimate is a function of the sample data and the sample size. Different samples will result in different estimates. The sample error affects the quantification of the calculated availability. If the estimates were based on one failure and one repair only, it would be quite risky (Coppola, 1997). We would feel more confident if we had more data (more failures and repairs). The question is how good the estimated inherent availability is. The answer is to attach a confidence level to the calculated availability, or give the confidence limits on the availability at a chosen confidence level. The most interesting confidence limits would be the shortest-length confidence limits on the true availability at a given confidence level.

In a wide variety of inference problems one is not interested in estimating the parameter or testing some hypothesis concerning it. Rather, one wishes to establish a lower or an upper bound, or both, for the real-valued parameter. For example, if X is the time to failure of a piece of equipment, one is interested in a lower bound for the mean of X. If the rv X measures the toxicity of a drug, the concern is to find an upper bound for the mean. Similarly, if the rv X measures the nicotine content of a certain brand of cigarettes, one is interested in determining an upper and a lower bound for the average nicotine content of these cigarettes.

The following result provides a general method of finding shortest-length confidence intervals and covers most cases in practice.

Let $S = s(\mathbf{X})$ be a statistic, based on a random sample $\mathbf{X}$. Let F be the distribution function of the pivotal quantity $V(S,A) \equiv A$ and let v_L, v_U be such that

$$F(v_U) - F(v_L) = Pr\{v_L < V < v_U\} = 1-\alpha. \tag{103}$$

It will be noted that the distribution of V does not depend on any unknown parameter. A $100(1-\alpha)\%$ confidence interval of A is $(A_L(S,v_L,v_U), A_U(S,v_L,v_U))$ and the length of this interval is $\Delta(S,v_L,v_U) = A_U - A_L$. We want to choose v_1, v_2, minimizing $A_U - A_L$ and satisfying (103). Thus, we consider the problem:

Minimize

$$\Delta(S, v_L, v_U) = A_U - A_L, \tag{104}$$

Subject to

$$F(v_U) - F(v_L) = 1-\alpha. \tag{105}$$

The search for the shortest-length confidence interval $\Delta = A_U - A_L$ is greatly facilitated by the use of the following theorem.

Theorem 3 (*Shortest-length confidence interval*). Under appropriate derivative conditions, there will be a pair (v_L, v_U) giving rise to the shortest-length confidence interval $\Delta(S, v_L, v_U) = A_U - A_L$ for A as a solution to the simultaneous equations:

$$\frac{\partial \Delta}{\partial v_L} + \frac{\partial \Delta}{\partial v_U} \frac{F'(v_L)}{F'(v_U)} = 0, \tag{106}$$

$$F(v_U) - F(v_L) = 1-\alpha. \tag{107}$$

Proof. Note that (107) forces v_U to be a function of v_L (or visa-versa). Take $\Delta(S, v_L, v_U)$ as a function of v_L, say $\Delta(S, v_L, v_U(v_L))$. Then, by using the method of Lagrange multipliers, the proof follows immediately.

For instance, consider the problem of constructing the shortest-length confidence interval for system availability from time-to-failure and time-to-repair test data. It is assumed that X_1 (time-to-failure) and X_2 (time-to-repair) are stochastically independent random variables with probability density functions

$$f_1(x_1;\theta_1) = \frac{1}{\theta_1} e^{-x_1/\theta_1}, \quad x_1 \in (0,\infty), \quad \theta_1 > 0, \tag{108}$$

and

$$f_2(x_2;\theta_2) = \frac{1}{\theta_2} e^{-x_2/\theta_2}, \quad x_2 \in (0,\infty), \quad \theta_2 > 0. \tag{109}$$

Availability is usually defined as the probability that a system is operating satisfactorily at any point in time. This probability can be expressed mathematically as

$$A = \theta_1 / (\theta_1 + \theta_2), \tag{110}$$

where θ_1 is a system mean-time-to-failure, θ_2 is a system mean-time-to-repair.

Consider a random sample $\mathbf{X}_1 = (X_{11}, \ldots, X_{1n_1})$ of n_1 times-to-failure and a random sample $\mathbf{X}_2 = (X_{21}, \ldots, X_{2n_2})$ of n_2 times-to-repair drawn from the populations described by (108) and (109) with sample means

$$\overline{X}_1 = \sum_{i=1}^{n_1} X_{1i} / n_1, \qquad \overline{X}_2 = \sum_{i=1}^{n_2} X_{2i} / n_2. \tag{111}$$

It is well known that $2n_1 \overline{X}_1 / \theta_1$ and $2n_2 \overline{X}_2 / \theta_2$ are chi-square distributed variables with $2n_1$ and $2n_2$ degrees of freedom, respectively. They are independent due to the independence of the variables X_1 and X_2.

It follows from (110) that

$$A / (1 - A) = \theta_1 / \theta_2. \tag{112}$$

Using the invariant embedding technique, we obtain from (112) a pivotal quantity

$$V(S,A) = S \frac{A}{1-A} = \frac{\overline{X}_2}{\overline{X}_1} \frac{\theta_1}{\theta_2} = \left(\frac{2n_2 \overline{X}_2 / \theta_2}{2n_2} \right) \bigg/ \left(\frac{2n_1 \overline{X}_1 / \theta_1}{2n_1} \right), \tag{113}$$

which is F-distributed with $(2n_2, 2n_1)$ degrees of freedom, and

$$S = \overline{X}_2 / \overline{X}_1. \tag{114}$$

Thus, (113) allows one to find a $100(1-\alpha)\%$ confidence interval for A from

$$\Pr\{A_L < A < A_U\} = 1 - \alpha, \tag{115}$$

where

$$A_L = v_L / (v_L + S) \quad \text{and} \quad A_U = v_U / (v_U + S). \tag{116}$$

It follows from Theorem 3 that the shortest-length confidence interval for A is given by

$$C_A^* = (A_L, A_U) \tag{117}$$

with

$$\Delta^*(S, v_L, v_U) = A_U - A_L, \tag{118}$$

where v_L and v_U are a solution of

$$(v_L + S)^2 f(v_L) = (v_U + S)^2 f(v_U) \tag{119}$$

(f is the pdf of an F-distributed rv with $(2n_2, 2n_1)$ d.f.) and

$$\Pr\{v_L < V < v_U\} = \Pr\{v_L < F(2n_2, 2n_1) < v_U\} = 1 - \alpha. \tag{120}$$

In practice, the simpler equal tails confidence interval for A,

$$C_A = (A_L, A_U) = \left(v_L / (v_L + S), v_U / (v_U + S)\right) \tag{121}$$

with

$$\Delta(S, v_L, v_U) = A_U - A_L, \tag{122}$$

is employed, where

$$v_L = F_{\alpha/2}(2n_2, 2n_1), \quad v_U = F_{1-\alpha/2}(2n_2, 2n_1), \tag{123}$$

and

$$\Pr\{F(2n_2, 2n_1) > F_{\alpha/2}(2n_2, 2n_1)\} \leq 1 - \alpha/2. \tag{124}$$

Consider, for instance, the following case. A total of 400 hours of operating time with 2 failures, which required an average of 20 hours of repair time, were observed for aircraft air-conditioning equipment. What is the confidence interval for the inherent availability of this equipment at the 90% confidence level?
The point estimate of the inherent availability is

$$\widehat{A} = 200 / (200 + 20) = 0.909, \tag{125}$$

and the confidence interval for the inherent availability, at the 90% confidence level, is found as follows.
From (121), the simpler equal tails confidence interval is

$$C_A = \left(\frac{F_{0.05}(4,4)}{F_{0.05}(4,4) + 1/\widehat{A} - 1}, \frac{F_{0.95}(4,4)}{F_{0.95}(4,4) + 1/\widehat{A} - 1} \right) = (0.61, 0.985), \quad (126)$$

i.e.,

$$\Delta(S, v_L, v_U) = A_U - A_L = 0.375. \quad (127)$$

From (117), the shortest-length confidence interval is

$$C_A^* = \left(\frac{v_L}{v_L + S}, \frac{v_U}{v_U + S} \right) = (0.707, 0.998), \quad (128)$$

where v_L and v_U are a solution of (119) and (120). Thus,

$$\Delta^*(S, v_L, v_U) = A_U - A_L = 0.291. \quad (129)$$

The relative efficiency of C_A relative to C_A^* is given by

$$\text{rel.eff.}_C \left(C_A, C_A^* \right) = \frac{\Delta^*(S, v_L, v_U)}{\Delta(S, v_L, v_U)} = \frac{0.291}{0.375} = 0.776. \quad (130)$$

6. General Problem Analysis

6.1 Inner Minimization

First consider the inner minimization, i.e., k (Section 2) is held fixed for the time being. Then the term ck does not affect the result of this minimization. Consider a situation of state estimation described by one of a family of density functions, indexed by the vector parameter $\boldsymbol{\theta} = (\mu, \sigma)$, where $\mu \equiv x(k)$ and $\sigma \equiv \omega(>0)$ are respectively parameters of location and scale. For this family, invariant under the group of positive linear transformations: $z \rightarrow az + b$ with $a > 0$, we shall assume that there is obtainable from some informative experiment (a random sample of observations $z^k = \{z(0), \ldots, z(k)\}$) a sufficient statistic (m_k, s_k) for (μ, σ) with density function $p_k(m_k, s_k; \mu, \sigma)$ of the form

$$p_k(m_k, s_k; \mu, \sigma) = \sigma^{-2} f_k[(m_k - \mu)/\sigma, s_k/\sigma]. \quad (131)$$

We are thus assuming that for the family of density functions an induced invariance holds under the group G of transformations: $m_k \rightarrow am_k + b$, $s_k \rightarrow as_k$ ($a > 0$). The family of density functions satisfying the above conditions is, of course, the limited one of normal, negative exponential, Weibull and gamma (with known index) density functions.

The loss incurred by making decision d when $\mu \equiv x(l)$ is the true parameter is given by the piecewise-linear loss function

$$r(\boldsymbol{\theta}, d) = \begin{cases} \dfrac{c_1(d - \mu)}{\sigma} & (\mu \leq d), \\[3mm] \dfrac{c_2(\mu - d)}{\sigma} & (\mu > d). \end{cases} \quad (132)$$

The decision problem specified by the informative experiment density function (131) and the loss function (132) is invariant under the group G of transformations. Thus, the problem is to find the best invariant estimator of μ,

$$d^* = \arg \min_{d \in D} R(\boldsymbol{\theta}, d), \tag{133}$$

where D is a set of invariant estimators of μ, $R(\boldsymbol{\theta}, d) = E_{\boldsymbol{\theta}}\{r(\boldsymbol{\theta}, d)\}$ is a risk function.

6.2 Best Invariant Estimator

It can be shown by using the invariant embedding technique that an invariant loss function, $r(\boldsymbol{\theta}, d)$, can be transformed as follows:

$$r(\boldsymbol{\theta}, d) = \ddot{r}(\mathbf{v}, \eta), \tag{134}$$

where

$$\ddot{r}(\mathbf{v}, \eta) = \begin{cases} c_1(v_1 + \eta v_2) & (v_1 \geq -\eta v_2), \\ -c_2(v_1 + \eta v_2) & (v_1 < -\eta v_2), \end{cases} \tag{135}$$

$\mathbf{v} = (v_1, v_2)$, $v_1 = (m_k - \mu)/\sigma$, $v_2 = s_k/\sigma$, $\eta = (d - m_k)/s_k$.

It follows from (134) that the risk associated with d and $\boldsymbol{\theta}$ can be expressed as

$$R(\boldsymbol{\theta}, d) = E_{\boldsymbol{\theta}}\{r(\boldsymbol{\theta}, d)\} = E_k\{\ddot{r}(v, \eta)\} = c_1 \int_0^\infty dv_2 \int_{-\eta v_2}^\infty (v_1 + \eta v_2) f_k(v_1, v_2) dv_1$$

$$- c_2 \int_0^\infty dv_2 \int_{-\infty}^{-\eta v_2} (v_1 + \eta v_2) f_k(v_1, v_2) dv_1, \tag{136}$$

which is constant on orbits when an invariant estimator (decision rule) d is used, where $f_k(v_1, v_2)$ is defined by (131). The fact that the risk (136) is independent of $\boldsymbol{\theta}$ means that a decision rule d, which minimizes (136), is uniformly best invariant. The following theorem gives the central result in this section.

Theorem 4 (*Best invariant estimator of μ*). Suppose that (v_1, v_2) is a random vector having density function

$$v_2 f_k(v_1, v_2) \left[\int_0^\infty v_2 dv_2 \int_{-\infty}^\infty f_k(v_1, v_2) dv_1 \right]^{-1} \quad (v_1 \text{ real}, v_2 > 0), \tag{137}$$

where f_k is defined by (131), and let G_k be the distribution function of v_1/v_2. Then the uniformly best invariant linear-loss estimator of μ is given by

$$d^* = m_k + \eta^* s_k, \tag{138}$$

where

$$G_k(-\eta^*) = c_1 / (c_1 + c_2).\tag{139}$$

Proof. From (136)

$$\frac{\partial E_k\{\ddot{r}(\mathbf{v},\eta)\}}{\partial\eta} = c_1\int_0^\infty v_2 dv_2 \int_{-\eta v_2}^\infty f_k(v_1,v_2)dv_1 - c_2\int_0^\infty v_2 dv_2 \int_{-\infty}^{-\eta v_2} f_k(v_1,v_2)dv_1$$

$$= \int_0^\infty v_2 dv_2 \int_{-\infty}^\infty f_k(v_1,v_2)dv_1[c_1 P_k\{(v_1,v_2):v_1+\eta v_2>0\} - c_2 P_k\{(v_1,v_2):v_1+\eta v_2<0\}]$$

$$= \int_0^\infty v_2 dv_2 \int_{-\infty}^\infty f_k(v_1,v_2)dv_1[c_1(1-G_k(-\eta)) - c_2 G_k(-\eta)].\tag{140}$$

Then the minimum of $E_k\{\ddot{r}(\mathbf{v},\eta)\}$ occurs for η^* being determined by setting $\partial E_k\{\ddot{r}(\mathbf{v},\eta)\}/\partial\eta = 0$ and this reduces to

$$c_1[1-G_k(-\eta^*)] - c_2 G_k(-\eta^*) = 0,\tag{141}$$

which establishes (139). □

Corollary 4.1 (*Minimum risk of the best invariant estimator of μ*). The minimum risk is given by

$$R(\boldsymbol{\theta},d^*) = E_{\boldsymbol{\theta}}\{r(\boldsymbol{\theta},d^*)\} = E_k\{\ddot{r}(\mathbf{v},\eta^*)\} = c_1\int_0^\infty dv_2 \int_{-\eta^* v_2}^\infty v_1 f_k(v_1,v_2)dv_1 - c_2\int_0^\infty dv_2 \int_{-\infty}^{-\eta^* v_2} v_1 f_k(v_1,v_2)dv_1$$

$$\tag{142}$$

with η^* as given by (139).

Proof. These results are immediate from (134) when use is made of $\partial E_k\{\ddot{r}(\mathbf{v},\eta)\}/\partial\eta = 0$. □

6.3 Outer Minimization

The results obtained above can be further extended to find the optimal number of observations. Now

$$E_{\boldsymbol{\theta}}\{\ddot{r}_k(\boldsymbol{\theta},d^*)\} = E_{\boldsymbol{\theta}}\{r(\boldsymbol{\theta},d^*) + ck\} = E_k\{\ddot{r}(\mathbf{v},\eta^*) + ck\} = c_1\int_0^\infty dv_2 \int_{-\eta^* v_2}^\infty v_1 f_k(v_1,v_2)dv_1$$

$$- c_2\int_0^\infty dv_2 \int_{-\infty}^{-\eta^* v_2} v_1 f_k(v_1,v_2)dv_1 + ck\tag{143}$$

is to be minimized with respect to k. It can be shown that this function (which is the constant risk corresponding to taking a sample of fixed sample size k and then estimating x(l) by the expression (108) with k for k*) has at most two minima (if there are two, they are for successive values of k; moreover, there is only one minimum for all but a denumerable set of values of c). If there are two minima, at k* and k*+1, one may randomize in any way between the decisions to take k* or k*+1 observations.

7. Example

Consider the one-dimensional discrete-time system, which is described by scalar difference equations of the form (1)-(2), and the case when the measurement noises w(k), k = 1, 2, … (see (2)) are independently and identically distributed random variables drawn from the exponential distribution with the density

$$f(w;\sigma) = (1/\sigma)\exp(-w/\sigma), \quad w \in (0,\infty), \tag{144}$$

where the parameter $\sigma > 0$ is unknown. It is required to find the best invariant estimator of x(l) on the basis of the data sample $\mathbf{z}^k = (z(1), \ldots, z(k))$ relative to the piecewise linear loss function

$$r(\boldsymbol{\theta},d) = \begin{cases} c_1(d-\mu)/\sigma, & d \ge \mu, \\ \\ c_2(\mu-d)/\sigma, & \text{otherwise}, \end{cases} \tag{145}$$

where $\boldsymbol{\theta} = (\mu,\sigma)$, $\mu \equiv x(l)$, $c_1 > 0$, $c_2 = 1$.
The likelihood function of $\mathbf{z}^k$ is

$$L(\mathbf{z}^k;\mu,\sigma) = \frac{1}{\sigma^k}\exp\left[-\sum_{j=1}^{k}(z(j)-H(j)x(j))/\sigma\right] = \frac{1}{\sigma^k}\exp\left[-\sum_{j=1}^{k}a(j)(y(j)-\mu)/\sigma\right], \tag{146}$$

where

$$y(j) = [a(j)]^{-1}\left(z(j)+H(j)\sum_{i=j}^{l-1}A(j,i+1)B(i)u(i)\right), \quad j \le l, \tag{147}$$

$$y(j) = [a(j)]^{-1}\left(z(j)-H(j)\sum_{i=l}^{j-1}A(j,i+1)B(i)u(i)\right), \quad j > l, \tag{148}$$

if l < k (estimation of the past state of the system), and

$$y(j) = \frac{z(j)+b(j)}{a(j)}, \tag{149}$$

$$a(j) = H(j)A(j,k_1), \tag{150}$$

$$b(j) = H(j) \sum_{i=j}^{l-1} A(j, i+1) B(i) u(i), \tag{151}$$

if either $l = k$ (estimation of the current state of the system) or $l > k$ (prediction of the future state of the system).

It can be justified by using the factorization theorem that (m_k, s_k) is a sufficient statistic for $\theta = (\mu, \sigma)$, where

$$m_k = \min_{1 \le j \le k} y(j), \quad s_k = \sum_{j=1}^{k} a(j)[y(j) - m_k]. \tag{152}$$

The probability density function of (m_k, s_k) is given by

$$p_k(m_k, s_k; \mu, \sigma) = \frac{n(k)}{\sigma} e^{-\frac{n(k)[m_k - \mu]}{\sigma}} \frac{1}{\Gamma(k-1)\sigma^{k-1}} s_k^{k-2} e^{-\frac{s_k}{\sigma}}, \quad m_k > \mu, \quad s_k > 0, \tag{153}$$

where

$$n(k) = \sum_{j=1}^{k} a(j). \tag{154}$$

Since the loss function (145) is invariant under the group G of location and scale changes, it follows that

$$r(\theta, d) = \ddot{r}(\mathbf{v}, \eta) = \begin{cases} c_1(v_1 + \eta v_2)/\sigma, & v_1 \ge -\eta v_2, \\ -(v_1 + \eta v_2)/\sigma, & \text{otherwise,} \end{cases} \tag{155}$$

where $\mathbf{v} = (v_1, v_2)$,

$$v_1 = \frac{m_k - \mu}{\sigma}, \quad v_2 = \frac{s_k}{\sigma}, \quad \eta = \frac{d - m_k}{s_k}. \tag{156}$$

Thus, using (138) and (139), we find that the best invariant estimator (BIE) of μ is given by

$$d_{BIE} = m_k + \eta^* s_k, \tag{157}$$

where

$$\eta^* = [1 - (c_1 + 1)^{1/k}]/n(k) = \arg\inf_{\eta} E_k\{\ddot{r}(\mathbf{v}, \eta)\}, \tag{158}$$

$$E_k\{\ddot{r}(\mathbf{v}, \eta)\} = [(c_1 + 1)(1 - \eta n(k))^{-(k-1)} - 1]/n(k) - \eta(k - 1). \tag{159}$$

The risk of this estimator is

$$R(\theta, d_{BIE}) = E_\theta\{r(\theta, d_{BIE})\} = E_k\{\ddot{r}(\mathbf{v}, \eta^*)\} = k[(c_1 + 1)^{1/k} - 1]/n(k). \tag{160}$$

Here the following theorem holds.

Theorem 5 (*Characterization of the estimator d_{BIE}*). For the loss function (145), the best invariant estimator of μ, d_{BIE}, given by (157) is uniformly non-dominated.

Proof. The proof follows immediately from Theorem 1 if we use the prior distribution on the parameter space Θ,

$$\xi_\tau(d\boldsymbol{\theta}) = \frac{1}{\tau\sigma} e^{-\frac{\tau-\mu}{\tau\sigma}} \frac{1}{\Gamma(1/\tau)\sigma^{1/\tau+1}} \left(\frac{1}{\tau}\right)^{1/\tau} e^{-\frac{1}{\tau\sigma}} d\mu d\sigma, \quad \mu \in (-\infty, \tau), \quad \sigma \in (0, \infty). \tag{161}$$

This ends the proof. $\square$

Consider, for comparison, the following estimators of μ (state of the system):
The maximum likelihood estimator (MLE):

$$d_{MLE} = m_k; \tag{162}$$

The minimum variance unbiased estimator (MVUE):

$$d_{MVUE} = m_k - \frac{s_k}{(k-1)n(k)}; \tag{163}$$

The minimum mean square error estimator (MMSEE):

$$d_{MMSEE} = m_k - \frac{s_k}{kn(k)}; \tag{164}$$

The median unbiased estimator (MUE):

$$d_{MUE} = m_k - (2^{1/(k-1)} - 1)\frac{s_k}{n(k)}. \tag{165}$$

Each of the above estimators is readily seen to be of a member of the class

$$C = \{d : d = m_k + \eta s_k\}, \tag{166}$$

where η is a real number. A risk of an estimator, which belongs to the class C, is given by (159). If, say, $k=3$ and $c_1=26$, then we have that

$$\text{rel.eff.}_R\{d_{MLE}, d_{BIE}; \boldsymbol{\theta}\} = 0.231, \tag{167}$$

$$\text{rel.eff.}_R\{d_{MVUE}, d_{BIE}; \boldsymbol{\theta}\} = 0.5, \tag{168}$$

$$\text{rel.eff.}_R\{d_{MMSEE}, d_{BIE}; \boldsymbol{\theta}\} = 0.404, \tag{169}$$

$$\text{rel.eff.}_R\{d_{MUE}, d_{BIE}; \boldsymbol{\theta}\} = 0.45. \tag{170}$$

In this case (143) becomes

$$E_{\boldsymbol{\theta}}\left\{r_k(\boldsymbol{\theta}, d^*)\right\} = E_{\boldsymbol{\theta}}\left\{r(\boldsymbol{\theta}, d_{BIE}) + ck\right\} = E_k\left\{\ddot{r}(\mathbf{v}, \eta^*) + ck\right\} = k[(c_1 + 1)^{1/k} - 1]/n(k) + ck \equiv J_k \; . \tag{171}$$

Now (171) is to be minimized with respect to k. It is easy to see that

$$J_k - J_{k-1} = -\big((k-1)[(c_1 + 1)^{1/(k-1)} - 1]/n(k-1) \; - k[(c_1 + 1)^{1/k} - 1]/n(k)\big) + c. \tag{172}$$

Define

$$\varphi(k) = (k-1)[(c_1 + 1)^{1/(k-1)} - 1]/n(k-1) - k[(c_1 + 1)^{1/k} - 1]/n(k). \tag{173}$$

Thus

$$c \gtrless \varphi(k) \Leftrightarrow J_k \gtrless J_{k-1}. \tag{174}$$

By plotting $\varphi(k)$ versus k the optimal number of observations k^* can be determined.
For each value of c, we can find an equilibrium point of k, i.e., $c=\varphi(k^\bullet)$. The following two cases must be considered:
1) $k^\bullet$ is not an integer. We have $k^{(1)}<k^\bullet<k^{(1)}+1=k^{(2)}$, where $k^{(1)}$ and $k^{(2)}$ are neighboring integers. Since $\varphi(k)$ is monotonically decreasing, we know that $\varphi(k^{(1)})>c$ and $\varphi(k^{(2)})<c$. Then, by using these properties, (172) becomes

$$J_{k^{(1)}} - J_{k^{(1)}-1} = -\varphi(k^{(1)}) + c < 0, \tag{175}$$

$$J_{k^{(2)}} - J_{k^{(1)}} = -\varphi(k^{(2)}) + c > 0, \tag{176}$$

Thus

$$J_{k^{(2)}} > J_{k^{(1)}} < J_{k^{(1)}-1}. \tag{177}$$

Therefore, $k^{(1)}$ is the optimal number of observations. We conclude that the optimal number k^* is equal to the largest integer below the equilibrium point.
2) $k^\bullet$ is an integer. By the same sort of argument, we know that $k^\bullet$ is as good as $k^\bullet$-1. Consequently, both $k^\bullet$ and $k^\bullet$-1 are the optimal number of observations. Notice that in this case, J_{k^*} can be computed directly and precisely from (172).

8. Conclusions and Directions for Future Research

In this paper we construct the minimum risk estimators of state of stochastic systems. The method used is that of the invariant embedding of sample statistics in a loss function in order to form pivotal quantities, which make it possible to eliminate unknown parameters from the problem. This method is a special case of more general considerations applicable whenever the statistical problem is invariant under a group of transformations, which acts transitively on the parameter space.

For a class of state estimation problems where observations on system state vectors are constrained, i.e., when it is not feasible to make observations at every moment, the question of how many observations to take must be answered. This paper models such a class of problems by assigning a fixed cost to each observation taken. The total number of observations is determined as a function of the observation cost.

Extension to the case where the observation cost is an explicit function of the number of observations taken is straightforward. A different way to model the observation constraints should be investigated.

More work is needed, however, to obtain improved decision rules for the problems of unconstrained and constrained optimization under parameter uncertainty when: (i) the observations are from general continuous exponential families of distributions, (ii) the observations are from discrete exponential families of distributions, (iii) some of the observations are from continuous exponential families of distributions and some from discrete exponential families of distributions, (iv) the observations are from multiparameter or multidimensional distributions, (v) the observations are from truncated distributions, (vi) the observations are censored, (vii) the censored observations are from truncated distributions.

9. Acknowledgments

10. References

Alamo, T.; Bravo, J. & Camacho, E. (2005). Guaranteed state estimation by zonotopes. *Automatica*, Vol. 41, pp. 1035 –1043

Barlow, R. E. & Proshan, F. (1966). Tolerance and confidence limits for classes of distributions based on failure rate. *Ann. Math. Stat.*, Vol. 37, pp. 1593 –1601

Ben-Daya, M.; Duffuaa, S.O. & Raouf, A. (2000). *Maintenance, Modeling and Optimization,* Kluwer Academic Publishers, Norwell, Massachusetts

Coppola, A. (1997). Some observations on demonstrating availability. *RAC Journal*, Vol. 6, pp. 17–18

Epstein, B. & Sobel, M. (1954). Some theorems relevant to life testing from an exponential population. *Ann. Math. Statist.*, Vol. 25, pp. 373 – 381

Gillijns, S. & De Moor, B. (2007). Unbiased minimum-variance input and state estimation for linear discrete-time systems. *Automatica*, Vol. 43, pp. 111–116

Grubbs, F. E. (1971). Approximate fiducial bounds on reliability for the two parameter negative exponential distribution. *Technometrics*, Vol. 13, pp. 873 – 876

Hahn, G.J. & Nelson, W. (1973). A survey of prediction intervals and their applications. *J. Qual. Tech.*, Vol. 5, pp. 178 –188

Ireson, W. G. & Coombs, C. F. (1988). *Handbook of Reliability Engineering and Management*, McGraw-Hill, New York

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Trans. ASME, J. Basic Eng.*, Vol. 82, pp. 34 – 45

Kendall, M.G. & Stuart, A. (1969). *The Advanced Theory of Statistics*, Vol. 1 (3rd edition), Griffin, London

Ko, S. & Bitmead, R. R. (2007). State estimation for linear systems with state equality constraints. *Automatica*, Vol. 43, pp. 1363 –1368

McGarty, T. P. (1974). *Stochastic Systems and State Estimation*, John Wiley and Sons, Inc., New York

Nechval, N.A. (1982). *Modern Statistical Methods of Operations Research*, RCAEI, Riga

Nechval, N.A. (1984). *Theory and Methods of Adaptive Control of Stochastic Processes*, RCAEI, Riga

Nechval, N.A.; Nechval, K.N. & Vasermanis, E.K. (2001). Optimization of interval estimators via invariant embedding technique. *IJCAS* (*An International Journal of Computing Anticipatory Systems*), Vol. 9, pp. 241–255

Nechval, N. A.; Nechval, K. N. & Vasermanis, E. K. (2003). Effective state estimation of stochastic systems. *Kybernetes*, Vol. 32, pp. 666 – 678

Nechval, N.A. & Vasermanis, E.K. (2004). *Improved Decisions in Statistics*, SIA "Izglitibas soli", Riga

Norgaard, M.; Poulsen, N. K. & Ravn, O. (2000). New developments in state estimation for nonlinear systems. *Automatica*, Vol. 36, pp. 1627–1638

Savkin, A. & Petersen, L. (1998). Robust state estimation and model validation for discrete-time uncertain system with a deterministic description of noise and uncertainty. *Automatica*, Vol. 34, 1998, pp. 271–274

Yan, J. & Bitmead, R. R. (2005). Incorporating state estimation into model predictive control and its application to network traffic control. *Automatica*, Vol. 41, pp. 595 – 604

Fuzzy identification of Discrete Time Nonlinear Stochastic Systems

Ginalber L. O. Serra
Federal Institute of Education, Science and Technology (IFMA)
Brasil

1. Introduction

System identification is the task of developing or improving a mathematical description of dynamic systems from experimental data (Ljung (1999); Söderström & Stoica (1989)). Depending on the level of a priori insight about the system, this task can be approached in three different ways: *white box modeling, black box modeling* and *gray box modeling*. These models can be used for simulation, prediction, fault detection, design of controllers (*model based control*), and so forth. Nonlinear system identification (Aguirre *et al.* (2005); Serra & Bottura (2005); Sjöberg *et al.* (1995); **?**) is becomming an important tool which can be used to improve control performance and achieve robust behavior (Narendra & Parthasarathy (1990); Serra & Bottura (2006a)). Most processes in industry are characterized by nonlinear and time-varying behavior and are not amenable to conventional modeling approaches due to the lack of precise, formal knowledge about it, its strongly nonlinear behavior and high degree of uncertainty. Methods based on fuzzy models are gradually becoming established not only in academic view point but also because they have been recognized as powerful tools in industrial applications, faciliting the effective development of models by combining information from different sources, such as empirical models, heuristics and data (Hellendoorn & Driankov (1997)). In fuzzy models, the relation between variables are based on if-then rules such as IF $< antecedent >$ THEN $< consequent >$, where antecedent evaluate the model inputs and consequent provide the value of the model output. Takagi and Sugeno, in 1985, developed a new approach in which the key idea was partitioning the input space into fuzzy areas and approximating each area by a linear or a nonlinear model (Takagi & Sugeno (1985)). This structure, so called Takagi-Sugeno (TS) fuzzy model, can be used to approximate a highly nonlinear function of simple structure using a small number of rules. Identification of TS fuzzy model using experimental data is divided into two steps: structure identification and parameter estimation. The former consists of antecedent structure identification and consequent structure identification. The latter consists of antecedent and consequent parameter estimation where the consequent parameters are the coefficients of the linear expressions in the consequent of a fuzzy rule. To be applicable to real world problems, the parameter estimation must be highly efficient. Input and output measurements may be contaminated by noise. For low levels of noise the least squares (LS) method, for example, may produce excellent estimates of the consequent parameters. However, with larger levels of noise, some modifications in this method are required to overcome this inconsistency. Generalized least squares (GLS) method, extended least squares (ELS) method, prediction error (PE) method, are examples of such modifications. A problem

with the use of these methods, in a fuzzy modeling context, is that the inclusion of the prediction error past values in the regression vector, which defines the input linguistic variables, increases the complexity of the fuzzy model structure and are inevitably dependent upon the accuracy of the noise model. To obtain consistent parameter estimates in a noisy environment without modeling the noise, the instrumental variable (IV) method can be used. It is known that by choosing proper instrumental variables, it provides a way to obtain consistent estimates with certain optimal properties (Serra & Bottura (2004; 2006b); Söderström & Stoica (1983)). This paper proposes an approach to nonlinear discrete time systems identification based on instrumental variable method and TS fuzzy model. In the proposed approach, which is an extension of the standard linear IV method (Söderström & Stoica (1983)), the chosen instrumental variables, statistically uncorrelated with the noise, are mapped to fuzzy sets, partitioning the input space in subregions to define valid and unbiased estimates of the consequent parameters for the TS fuzzy model in a noisy environment. From this theoretical background, the *fuzzy instrumental variable* (FIV) concept is proposed, and the main statistical characteristics of the FIV algorithm such as consistency and unbias are derived. Simulation results show that the proposed algorithm is relatively insensitive to the noise on the measured input-output data.

This paper is organized as follow: In Section 2, a brief review of the TS fuzzy model formulation is given. In Section 3, the fuzzy NARX structure is introduced. It is used to formulate the proposed approach. In Section 4, the TS fuzzy model consequent parameters estimate problem in a noisy environment is studied. From this analysis, three Lemmas and one Theorem are proposed to show the consistency and unbias of the parameters estimates in a noisy environment with the proposed approach. The fuzzy instrumental variable concept is also proposed and considerations about how the FIV should be chosen are given. In Section 5, offline and on-line schemes of the fuzzy instrumental variable algorithm are derived. Simulation results showing the efficiency of the FIV approach in a noisy environment are given in Section 6. Finally, the closing remarks are given in Section 7.

2. Takagi-Sugeno Fuzzy Model

The TS fuzzy inference system is composed by a set of IF-THEN rules which partitions the input space, so-called *universe of discourse*, into fuzzy regions described by the rule antecedents in which consequent functions are valid. The consequent of each rule i is a functional expression $y_i = f_i(x)$ (King (1999); Papadakis & Theocaris (2002)). The i-th TS fuzzy rule has the following form:

$$R^{i|i=1,2,\ldots,l} : \text{ IF } x_1 \text{ is } F_1^i \text{ AND } \cdots \text{ AND } x_q \text{ is } F_q^i \text{ THEN } y_i = f_i(\mathbf{x}) \tag{1}$$

where l is the number of rules. The vector $\mathbf{x} \in \Re^q$ contains the antecedent linguistic variables, which has its own universe of discourse partitioned into fuzzy regions by the fuzzy sets representing the linguistic terms. The variable x_j belongs to a fuzzy set F_j^i with a truth value given by a membership function $\mu_{F_j}^i : \Re \to [0,1]$. The truth value h_i for the complete rule i is computed using the aggregation operator, or t-norm, AND, denoted by $\otimes : [0,1] \times [0,1] \to [0,1]$,

$$h_i(x) = \mu_1^i(x_1) \otimes \mu_2^i(x_2) \otimes \ldots \mu_q^i(x_q) \tag{2}$$

Among the different t-norms available, in this work the algebraic product will be used, and

$$h_i(\mathbf{x}) = \prod_{j=1}^{q} \mu_j^i(x_j) \tag{3}$$

The degree of activation for rule i is then normalized as

$$\gamma_i(\mathbf{x}) = \frac{h_i(\mathbf{x})}{\sum_{r=1}^{l} h_r(\mathbf{x})} \tag{4}$$

This normalization implies that

$$\sum_{i=1}^{l} \gamma_i(\mathbf{x}) = 1 \tag{5}$$

The response of the TS fuzzy model is a weighted sum of the consequent functions, i.e., a convex combination of the local functions (models) f_i,

$$y = \sum_{i=1}^{l} \gamma_i(\mathbf{x}) f_i(\mathbf{x}) \tag{6}$$

which can be seen as a linear parameter varying (LPV) system. In this sense, a TS fuzzy model can be considered as a mapping from the antecedent (input) space to a convex region (politope) in the space of the local submodels defined by the consequent parameters, as shown in Fig. 1 (Bergsten (2001)). This property simplifies the analysis of TS fuzzy models in a robust
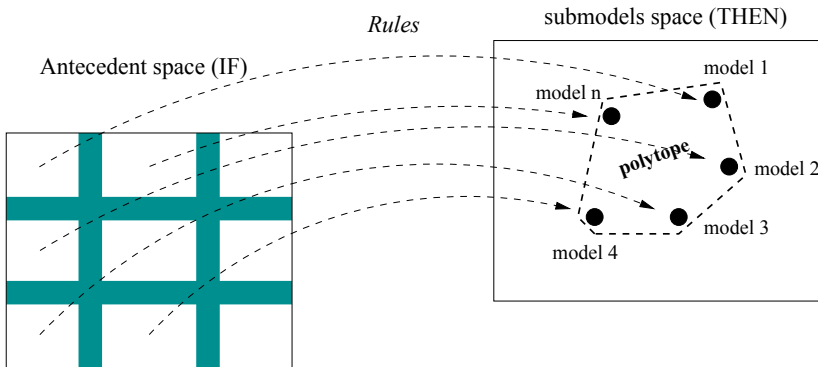

Fig. 1. Mapping to local submodels space.

linear system framework for identification, controllers design with desired closed loop characteristics and stability analysis (Johansen *et al.* (2000); Kadmiry & Driankov (2004); Tanaka *et al.* (1998); Tong & Li (2002)).

3. Fuzzy Structure Model

The nonlinear input-output representation is often used for building TS fuzzy models from data, where the regression vector is represented by a finite number of past inputs and outputs of the system. In this work, the nonlinear autoregressive with exogenous input (NARX) structure model is used. This model is applied in most nonlinear identification methods such as neural networks, radial basis functions, cerebellar model articulation controller (CMAC), and

also fuzzy logic (Brown & Harris (1994)). The NARX model establishes a relation between the collection of past scalar input-output data and the predicted output

$$y(k+1) = F[y(k), \dots, y(k-n_y+1), u(k), \dots, u(k-n_u+1)] \tag{7}$$

where k denotes discrete time samples, n_y and n_u are integers related to the system's order. In terms of rules, the model is given by

$$R^i : \text{ IF } y(k) \text{ is } F_1^i \text{ AND } \cdots \text{ AND } y(k-n_y+1) \text{ is } F_{n_y}^i$$

$$\text{AND } u(k) \text{ is } G_1^i \text{ AND } \cdots \text{ AND } u(k-n_u+1) \text{ is } G_{n_u}^i$$

$$\text{THEN } \hat{y}_i(k+1) = \sum_{j=1}^{n_y} a_{i,j} y(k-j+1) + \sum_{j=1}^{n_u} b_{i,j} u(k-j+1) + c_i \tag{8}$$

where $a_{i,j}$, $b_{i,j}$ and c_i are the consequent parameters to be determined. The inference formula of the TS fuzzy model is a straightforward extension of (6) and is given by

$$y(k+1) = \frac{\sum_{i=1}^{l} h_i(\mathbf{x}) \hat{y}_i(k+1)}{\sum_{i=1}^{l} h_i(\mathbf{x})} \tag{9}$$

or

$$y(k+1) = \sum_{i=1}^{l} \gamma_i(\mathbf{x}) \hat{y}_i(k+1) \tag{10}$$

with

$$\mathbf{x} = [y(k), \dots, y(k-n_y+1), u(k), \dots, u(k-n_u+1)] \tag{11}$$

and $h_i(\mathbf{x})$ is given as (3). This NARX model represents multiple input and single output (MISO) systems directly and multiple input and multiple output (MIMO) systems in a decomposed form as a set of coupled MISO models.

4. Consequent Parameters Estimate

The inference formula of the TS fuzzy model in (10) can be expressed as

$$y(k+1) = \gamma_1(\mathbf{x}_k)[a_{1,1} y(k) + \dots + a_{1,ny} y(k-n_y+1)$$

$$+ b_{1,1} u(k) + \dots + b_{1,nu} u(k-n_u+1) + c_1] + \gamma_2(\mathbf{x}_k)[a_{2,1} y(k)$$

$$+ \dots + a_{2,ny} y(k-n_y+1) + b_{2,1} u(k) + \dots + b_{2,nu} u(k-n_u+1)$$

$$+ c_2] + \dots + \gamma_l(\mathbf{x}_k)[a_{l,1} y(k) + \dots + a_{l,ny} y(k-n_y$$

$$+ 1) + b_{l,1} u(k) + \dots + b_{l,nu} u(k-n_u+1) + c_l] \tag{12}$$

which is linear in the consequent parameters: $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$. For a set of N input-output data pairs $\{(\mathbf{x}_k, y_k) | i = 1, 2, \dots, N\}$ available, the following vetorial form is obtained

$$\mathbf{Y} = [\psi_1 \mathbf{X}, \psi_2 \mathbf{X}, \dots, \psi_l \mathbf{X}] \theta + \Xi \tag{13}$$

where $\psi_i = diag(\gamma_i(\mathbf{x}_k)) \in \Re^{N \times N}$, $\mathbf{X} = [\mathbf{y}_k, \dots, \mathbf{y}_{k-ny+1}, \mathbf{u}_k, \dots, \mathbf{u}_{k-nu+1}, \mathbf{1}] \in \Re^{N \times (n_y+n_u+1)}$, $\mathbf{Y} \in \Re^{N \times 1}$, $\Xi \in \Re^{N \times 1}$ and $\theta \in \Re^{l(n_y+n_u+1) \times 1}$ are the normalized membership degree matrix of (4), the data matrix, the output vector, the approximation error vector and the estimated parameters vector, respectively. If the unknown parameters associated variables are *exactly known* quantities, then the least squares method can be used efficiently. However, in practice, and in the present context, the elements of $\mathbf{X}$ are no exactly known quantities so that its value can be expressed as

$$y_k = \chi_k^T \theta + \eta_k \tag{14}$$

where, at the k-th sampling instant, $\chi_k^T = [\gamma_k^1(\mathbf{x}_k + \xi_k), \dots, \gamma_k^l(\mathbf{x}_k + \xi_k)]$ is the vector of the data with error in variables, $\mathbf{x}_k = [y_{k-1}, \dots, y_{k-n_y}, u_{k-1}, \dots, u_{k-n_u}, 1]^T$ is the vector of the data with exactly known quantities, e.g., free noise input-output data, ξ_k is a vector of noise associated with the observation of $\mathbf{x}_k$, and η_k is a disturbance noise.
The normal equations are formulated as

$$[\sum_{j=1}^{k} \chi_j \chi_j^T]\hat{\theta}_k = \sum_{j=1}^{k} \chi_j y_j \tag{15}$$

and multiplying by $\frac{1}{k}$ gives

$$\{\frac{1}{k}\sum_{j=1}^{k}[\gamma_j^1(\mathbf{x}_j + \xi_j), \dots, \gamma_j^l(\mathbf{x}_j + \xi_j)][\gamma_j^1(\mathbf{x}_j + \xi_j), \dots,$$

$$\gamma_j^l(\mathbf{x}_j + \xi_j)]^T\}\hat{\theta}_k = \frac{1}{k}\sum_{j=1}^{k}[\gamma_j^1(\mathbf{x}_j + \xi_j), \dots, \gamma_j^l(\mathbf{x}_j + \xi_j)]y_j$$

Noting that $y_j = \chi_j^T \theta + \eta_j$,

$$\{\frac{1}{k}\sum_{j=1}^{k}[\gamma_j^1(\mathbf{x}_j + \xi_j), \dots, \gamma_j^l(\mathbf{x}_j + \xi_j)][\gamma_j^1(\mathbf{x}_j + \xi_j), \dots,$$

$$\gamma_j^l(\mathbf{x}_j + \xi_j)]^T\}\hat{\theta}_k = \frac{1}{k}\sum_{j=1}^{k}[\gamma_j^1(\mathbf{x}_j + \xi_j), \dots, \gamma_j^l(\mathbf{x}_j + \xi_j)]$$

$$[\gamma_j^1(\mathbf{x}_j + \xi_j), \dots, \gamma_j^l(\mathbf{x}_j + \xi_j)]^T \theta + \frac{1}{k}\sum_{j=1}^{k}[\gamma_j^1(\mathbf{x}_j + \xi_j), \dots,$$

$$\gamma_j^l(\mathbf{x}_j + \xi_j)]\eta_j \tag{16}$$

and

$$\tilde{\theta}_k = \{\frac{1}{k}\sum_{j=1}^{k}[\gamma_j^1(\mathbf{x}_j + \xi_j), \dots, \gamma_j^l(\mathbf{x}_j + \xi_j)][\gamma_j^1(\mathbf{x}_j + \xi_j), \dots,$$

$$\gamma_j^l(\mathbf{x}_j + \xi_j)]^T\}^{-1}\frac{1}{k}\sum_{j=1}^{k}[\gamma_j^1(\mathbf{x}_j + \xi_j), \dots, \gamma_j^l(\mathbf{x}_j + \xi_j)]\eta_j \tag{17}$$

where $\tilde{\theta}_k = \hat{\theta}_k - \theta$ is the parameter error. Taking the probability in the limit as $k \to \infty$,

$$\text{p.lim } \tilde{\theta}_k = \text{p.lim } \{ \frac{1}{k}\mathbf{C}_k^{-1}\frac{1}{k}\mathbf{b}_k \} \tag{18}$$

with

$$\mathbf{C}_k = \sum_{j=1}^{k}[\gamma_j^1(\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l(\mathbf{x}_j + \xi_j)][\gamma_j^1(\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l(\mathbf{x}_j + \xi_j)]^T$$

$$\mathbf{b}_k = \sum_{j=1}^{k}[\gamma_j^1(\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l(\mathbf{x}_j + \xi_j)]\eta_j$$

Applying Slutsky's theorem and assuming that the elements of $\frac{1}{k}\mathbf{C}_k$ and $\frac{1}{k}\mathbf{b}_k$ converge in probability, we have

$$\text{p.lim } \tilde{\theta}_k = \text{p.lim } \frac{1}{k}\mathbf{C}_k^{-1} \text{ p.lim } \frac{1}{k}\mathbf{b}_k \tag{19}$$

Thus,

$$\text{p.lim } \frac{1}{k}\mathbf{C}_k = \text{p.lim } \frac{1}{k}\sum_{j=1}^{k}[\gamma_j^1(\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l(\mathbf{x}_j + \xi_j)]$$

$$[\gamma_j^1(\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l(\mathbf{x}_j + \xi_j)]^T$$

$$\text{p.lim } \frac{1}{k}\mathbf{C}_k = \text{p.lim } \frac{1}{k}\sum_{j=1}^{k}(\gamma_j^1)^2(\mathbf{x}_j + \xi_j)(\mathbf{x}_j + \xi_j)^T +$$

$$\ldots + \text{p.lim } \frac{1}{k}\sum_{j=1}^{k}(\gamma_j^l)^2(\mathbf{x}_j + \xi_j)(\mathbf{x}_j + \xi_j)^T$$

Assuming $\mathbf{x}_j$ and ξ_j statistically independent,

$$\text{p.lim } \frac{1}{k}\mathbf{C}_k = \text{p.lim } \frac{1}{k}\sum_{j=1}^{k}(\gamma_j^1)^2[\mathbf{x}_j\mathbf{x}_j^T + \xi_j\tilde{\xi}_j^T] + \ldots$$

$$+ \text{p.lim } \frac{1}{k}\sum_{j=1}^{k}(\gamma_j^l)^2[\mathbf{x}_j\mathbf{x}_j^T + \xi_j\tilde{\xi}_j^T]$$

$$\text{p.lim } \frac{1}{k}\mathbf{C}_k = \text{p.lim } \frac{1}{k}\sum_{j=1}^{k}\mathbf{x}_j\mathbf{x}_j^T[(\gamma_j^1)^2 + \ldots + (\gamma_j^l)^2]$$

$$+ \text{p.lim } \frac{1}{k}\sum_{j=1}^{k}\xi_j\tilde{\xi}_j^T[(\gamma_j^1)^2 + \ldots + (\gamma_j^l)^2] \tag{20}$$

with $\sum_{i=1}^{l} \gamma_j^i = 1$. Hence, the asymptotic analysis of the TS fuzzy model consequent parameters estimation is based in a weighted sum of the fuzzy covariance matrices of $\mathbf{x}$ and ξ. Similarly,

$$\text{p.lim } \frac{1}{k}\mathbf{b}_k = \text{p.lim } \frac{1}{k}\sum_{j=1}^{k}[\gamma_j^1(\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l(\mathbf{x}_j + \xi_j)]\eta_j$$

$$\text{p.lim} \frac{1}{k} \mathbf{b}_k = \text{p.lim} \frac{1}{k} \sum_{j=1}^{k} [\gamma_j^1 \xi_j \eta_j, \dots, \gamma_j^l \xi_j \eta_j] \tag{21}$$

Substituting from (20) and (21) in (19), results

$$\text{p.lim} \, \tilde{\theta}_k = \{\text{p.lim} \frac{1}{k} \sum_{j=1}^{k} \mathbf{x}_j \mathbf{x}_j^T [(\gamma_j^1)^2 + \dots + (\gamma_j^l)^2] +$$

$$\text{p.lim} \frac{1}{k} \sum_{j=1}^{k} \xi_j \xi_j^T [(\gamma_j^1)^2 + \dots + (\gamma_j^l)^2]\}^{-1} \text{p.lim} \frac{1}{k} \sum_{j=1}^{k} [\gamma_j^1 \xi_j \eta_j,$$

$$\dots, \gamma_j^l \xi_j \eta_j] \tag{22}$$

with $\sum_{i=1}^{l} \gamma_j^i = 1$. For the case of only one rule ($l = 1$), the analysis is simplified to the linear one, with $\gamma_j^i |_{j=1,\dots,k}^{i=1} = 1$. Thus, this analysis, which is a contribution of this article, is an extension of the standard linear one, from which can result several studies for fuzzy filtering and modeling in a noisy environment, fuzzy signal enhancement in communication channel, and so forth. Provided that the input u_k continues to excite the process and, at the same time, the coefficients in the submodels from the consequent are not all zero, then the output y_k will exist for all k observation intervals. As a result, the fuzzy covariance matrix $\sum_{j=1}^{k} \mathbf{x}_j \mathbf{x}_j^T [(\gamma_j^1)^2 + \dots + (\gamma_j^l)^2]$ will also be non-singular and its inverse will exist. Thus, the only way in which the asymptotic error can be zero is for $\xi_j \eta_j$ identically zero. But, in general, ξ_j and η_j are correlated, the asymptotic error will not be zero and the least squares estimates will be asymptotically biased to an extent determined by the relative ratio of noise to signal variances. In other words, least squares method is not appropriate to estimate the TS fuzzy model consequent parameters in a noisy environment because the estimates will be inconsistent and the bias error will remain no matter how much data can be used in the estimation.

4.1 Fuzzy instrumental variable (FIV)
To overcome this bias error and inconsistence problem, generating a vector of variables which are independent of the noise inputs and correlated with data vetor $\mathbf{x}_j$ from the system is required. If this is possible, then the choice of this vector becomes effective to remove the asymptotic bias from the consequent parameters estimates. The fuzzy least squares estimates is given by:

$$\{\frac{1}{k} \sum_{j=1}^{k} [\gamma_j^1 (\mathbf{x}_j + \xi_j), \dots, \gamma_j^l (\mathbf{x}_j + \xi_j)][\gamma_j^1 (\mathbf{x}_j + \xi_j), \dots,$$

$$\gamma_j^l (\mathbf{x}_j + \xi_j)]^T\} \hat{\theta}_k = \frac{1}{k} \sum_{j=1}^{k} [\gamma_j^1 (\mathbf{x}_j + \xi_j), \dots,$$

$$\gamma_j^l (\mathbf{x}_j + \xi_j)]\{[\gamma_j^1 (\mathbf{x}_j + \xi_j), \dots, \gamma_j^l (\mathbf{x}_j + \xi_j)]^T \theta + \eta_j\}$$

Using a new fuzzy vector of variables of the form $[\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]$, the last equation can be placed as

$$\{\frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j][\gamma_j^1(\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l(\mathbf{x}_j + \xi_j)]^T\}\hat{\theta}_k =$$

$$\frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]\{[\gamma_j^1(\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l(\mathbf{x}_j + \xi_j)]^T\theta +$$

$$\eta_j\} \tag{23}$$

where $\mathbf{z}_j$ is a vector with the order of $\mathbf{x}_j$, associated to the dynamic behavior of the system, and $\beta_j^i \big|_{j=1,\ldots,k}^{i=1,\ldots,l}$ is the normalized degree of activation, as in (4), associated to $\mathbf{z}_j$. For convergence analysis of the estimates, with the inclusion of this new fuzzy vector, the following is proposed:

Lemma 1 *Consider $\mathbf{z}_j$ a vector with the order of $\mathbf{x}_j$, associated to dynamic behavior of the system and independent of the noise input ξ_j; and $\beta_j^i \big|_{j=1,\ldots,k}^{i=1,\ldots,l}$ is the normalized degree of activation, a variable defined as in (4) associated to $\mathbf{z}_j$. Then, at the limit*

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]\xi_j^T = \mathbf{0} \tag{24}$$

Proof: Developing the left side of (24), results

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]\xi_j^T = \lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j \xi_j^T, \ldots,$$

$$\beta_j^l \mathbf{z}_j \xi_j^T]$$

As $\beta_j^i \big|_{j=1,\ldots,k}^{i=1,\ldots,l}$ is a scalar, and, by definition, the chosen variables are independent of the noise inputs, the inner product between $\mathbf{z}_j$ and ξ_j will be zero. Thus, taking the limit, results

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j \xi_j^T, \ldots, \beta_j^l \mathbf{z}_j \xi_j^T] = \mathbf{0}$$

$\square$

Lemma 2 *Under the same conditions as Lemma 1 and $\mathbf{z}_j$ independent of the disturbance noise η_j, then, at the limit*

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]\eta_j = \mathbf{0} \tag{25}$$

Proof: Developing the left side of (25), results

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]\eta_j = \lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j \eta_j, \ldots,$$

$$\beta_j^l \mathbf{z}_j \eta_j]$$

Because the chosen variables are independent of the disturbance noise, the product between $\mathbf{z}_j$ and η_j will be zero in the limit. Hence,

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j \eta_j, \ldots, \beta_j^l \mathbf{z}_j \eta_j] = \mathbf{0}$$

□

Lemma 3 *Under the same conditions as Lemma 1, according to (23), at the limit*

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j][\gamma_j^1 (\mathbf{x}_j + \xi_j), \ldots,$$

$$\gamma_j^l (\mathbf{x}_j + \xi_j)]^T = \mathbf{C_{zx}} \neq 0 \qquad (26)$$

Proof: Developing the left side of (26), results

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j][\gamma_j^1 (\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l (\mathbf{x}_j + \xi_j)]^T =$$

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \gamma_j^1 \mathbf{z}_j (\mathbf{x}_j + \xi_j)^T + \ldots + \beta_j^l \gamma_j^l \mathbf{z}_j (\mathbf{x}_j + \xi_j)^T]$$

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j][\gamma_j^1 (\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l (\mathbf{x}_j + \xi_j)]^T =$$

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \gamma_j^1 (\mathbf{z}_j \mathbf{x}_j^T + \mathbf{z}_j \xi_j^T) + \ldots + \beta_j^l \gamma_j^l (\mathbf{z}_j \mathbf{x}_j^T + \mathbf{z}_j \xi_j^T)]$$

From the **Lemma 1**, this expression is simplifyied as

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j][\gamma_j^1 (\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l (\mathbf{x}_j + \xi_j)]^T =$$

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \gamma_j^1 \mathbf{z}_j \mathbf{x}_j^T + \ldots + \beta_j^l \gamma_j^l \mathbf{z}_j \mathbf{x}_j^T]$$

Due to correlation between $\mathbf{z}_j$ and $\mathbf{x}_j$, this fuzzy covariance matrix has the following property:

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \gamma_j^1 \mathbf{z}_j \mathbf{x}_j^T + \ldots + \beta_j^l \gamma_j^l \mathbf{z}_j \mathbf{x}_j^T] \neq 0 \qquad (27)$$

and

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \gamma_j^1 \mathbf{z}_j \mathbf{x}_j^T + \ldots + \beta_j^l \gamma_j^l \mathbf{z}_j \mathbf{x}_j^T] = \mathbf{C_{zx}} \neq \mathbf{0}$$

□

Theorem 1 *Under suitable conditions outlined from Lemma 1 to 3, the estimation of the parameter vector θ for the model in (12) is strongly consistent, i.e, at the limit*

$$\text{p.lim } \tilde{\theta} = 0 \tag{28}$$

Proof: From the new fuzzy vector of variables of the form $[\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]$, the fuzzy least square estimation can be modifyied as follow:

$$\{\frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j][\gamma_j^1 (\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l (\mathbf{x}_j + \xi_j)]^T\}\hat{\theta}_k =$$

$$\frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]\{[\gamma_j^1 (\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l (\mathbf{x}_j + \xi_j)]^T\theta + \eta_j\}$$

which can be expressed in the form

$$\{\frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j][\gamma_j^1 (\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l (\mathbf{x}_j + \xi_j)]^T\}(\hat{\theta}_k - \theta) =$$

$$\frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]\eta_j$$

and

$$\{\frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j][\gamma_j^1 (\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l (\mathbf{x}_j + \xi_j)]^T\}\tilde{\theta} =$$

$$\frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]\eta_j$$

Taking the probability in the limit as $k \to \infty$, and applying the Slutsky's theorem, we have

$$\text{p.lim } \tilde{\theta}_k = \{\text{p.lim}\frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j][\gamma_j^1 (\mathbf{x}_j + \xi_j), \ldots,$$

$$\gamma_j^l (\mathbf{x}_j + \xi_j)]^T\}^{-1}\{\text{p.lim}\frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]\eta_j\}$$

According to **Lemma 1** and **Lemma 3**, results

$$\text{p.lim } \tilde{\theta}_k = \{\text{p.lim}\mathbf{C_{zx}}\}^{-1}\{\text{p.lim}\frac{1}{k} \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]\eta_j\}$$

where the fuzzy covariance matrix $\mathbf{C_{zx}}$ is non-singular and, as a consequence, the inverse exist. From the **Lemma 2**, we have

$$\text{p.lim } \tilde{\theta}_k = \{\text{p.lim}\mathbf{C_{zx}}\}^{-1}\mathbf{0}$$

Thus, the limit value of the parameter error, in probability, is

$$\text{p.lim } \tilde{\theta} = 0 \tag{29}$$

and the estimates are asymptotically unbiased, as required. ∎

As a consequence of this analysis, the definition of the vector $[\beta_j^1 \mathbf{z}_j, \dots, \beta_j^l \mathbf{z}_j]$ as the *fuzzy instrumental variable vector* or simply the *fuzzy instrumental variable* (FIV) is proposed. Clearly, with the use of the FIV vector in the form suggested, becomes possible to eliminate the asymptotic bias while preserving the existence of a solution. However, the statistical efficiency of the solution is dependent on the degree of correlation between $[\beta_j^1 \mathbf{z}_j, \dots, \beta_j^l \mathbf{z}_j]$ and $[\gamma_j^1 \mathbf{x}_j, \dots, \gamma_j^l \mathbf{x}_j]$. In particular, the lowest variance estimates obtained from this approach occur only when $\mathbf{z}_j = \mathbf{x}_j$ and $\beta_j^i \mid_{j=1,\dots,k}^{i=1,\dots,l} = \gamma_j^i \mid_{j=1,\dots,k}^{i=1,\dots,l}$, i.e., when the $\mathbf{z}_j$ are equal to the dynamic system "free noise" variables, which are unavailable in practice. According to situation, several fuzzy instrumental variables can be chosen. An effective choice of FIV would be the one based on the delayed input sequence

$$\mathbf{z}_j = [u_{k-\tau}, \dots, u_{k-\tau-n}, u_k, \dots, u_{k-n}]^T$$

where τ is chosen so that the elements of the fuzzy covariance matrix $\mathbf{C_{zx}}$ are maximized. In this case, the input signal is considered persistently exciting, e.g., it continuously perturbs or excites the system. Another FIV would be the one based on the delayed input-output sequence

$$\mathbf{z}_j = [y_{k-1-dl}, \cdots, y_{k-n_y-dl}, u_{k-1-dl}, \cdots, u_{k-n_u-dl}]^T$$

where dl is the applied delay. Other FIV could be the one based in the input-output from a "fuzzy auxiliar model" with the same structure of the one used to identify the nonlinear dynamic system. Thus,

$$\mathbf{z}_j = [\hat{y}_{k-1}, \cdots, \hat{y}_{k-n_y}, u_{k-1}, \cdots, u_{k-n_u}]^T$$

where $\hat{y}_k$ is the output of the fuzzy auxiliar model, and u_k is the input of the dynamic system. The inference formula of this fuzzy auxiliar model is given by

$$\hat{y}(k+1) = \beta_1(\mathbf{z}_k)[\alpha_{1,1}\hat{y}(k) + \dots + \alpha_{1,ny}\hat{y}(k - n_y + 1) +$$
$$\rho_{1,1}u(k) + \dots + \rho_{1,nu}u(k - n_u + 1) + \delta_1] + + \beta_2(\mathbf{z}_k)[\alpha_{2,1}\hat{y}(k)$$
$$+ \dots + \alpha_{2,ny}\hat{y}(k - n_y + 1) + \rho_{2,1}u(k) + \dots + \rho_{2,nu}u(k -$$
$$n_u + 1) + \delta_2] + \dots + \beta_l(\mathbf{z}_k)[\alpha_{l,1}\hat{y}(k) + \dots + \alpha_{l,ny}\hat{y}(k -$$
$$n_y + 1) + \rho_{l,1}u(k) + \dots + \rho_{l,nu}u(k - n_u + 1) + \delta_l]$$

which is also linear in the consequent parameters: α, ρ and δ. The closer these parameters are to the actual, but unknown, system parameters ($\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$) as in (12), more correlated $\mathbf{z}_k$ and $\mathbf{x}_k$ will be, and the obtained FIV estimates closer to the optimum.

5. FIV Algorithm

The FIV approach is a simple and attractive technique because it does not require the noise modeling to yield consistent, asymptotically unbiased consequent parameters estimates.

5.1 Off-line scheme
The FIV normal equations are formulated as

$$\sum_{j=1}^{k}[\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j][\gamma_j^1(\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l(\mathbf{x}_j + \xi_j)]^T \hat{\theta}_k - \sum_{j=1}^{k}[\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]y_j = 0 \qquad (30)$$

or, with $\zeta_j = [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]$,

$$[\sum_{j=1}^{k} \zeta_j \chi_j^T]\hat{\theta}_k - \sum_{j=1}^{k} \zeta_j y_j = 0 \qquad (31)$$

so that the FIV estimate is obtained as

$$\hat{\theta}_k = \{\sum_{j=1}^{k}[\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j][\gamma_j^1(\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l(\mathbf{x}_j + \xi_j)]^T\}^{-1} \sum_{j=1}^{k}[\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]y_j \qquad (32)$$

and, in vectorial form, the interest problem may be placed as

$$\hat{\theta} = (\Gamma^T \Sigma)^{-1} \Gamma^T \mathbf{Y} \qquad (33)$$

where $\Gamma^T \in \Re^{l(n_y+n_u+1) \times N}$ is the fuzzy extended instrumental variable matrix with rows given by ζ_j, $\Sigma \in \Re^{N \times l(n_y+n_u+1)}$ is the fuzzy extended data matrix with rows given by χ_j and $\mathbf{Y} \in \Re^{N \times 1}$ is the output vector and $\hat{\theta} \in \Re^{l(n_y+n_u+1) \times 1}$ is the parameters vector. The models can be obtained by the following two approaches:

- *Global approach* : In this approach all linear consequent parameters are estimated simultaneously, minimizing the criterion:

$$\hat{\theta} = \arg\min \| \Gamma^T \Sigma \theta - \Gamma^T \mathbf{Y} \|_2^2 \qquad (34)$$

- *Local approach* : In this approach the consequent parameters are estimated for each rule i, and hence independently of each other, minimizing a set of weighted local criteria $(i = 1, 2, \ldots, l)$:

$$\hat{\theta}_i = \arg\min \| \mathbf{Z}^T \Psi_i \mathbf{X} \theta_i - \mathbf{Z}^T \Psi_i \mathbf{Y} \|_2^2 \qquad (35)$$

where $\mathbf{Z}^T$ has rows given by $\mathbf{z}_j$ and Ψ_i is the normalized membership degree diagonal matrix according to $\mathbf{z}_j$.

5.2 On-line scheme
An on line FIV scheme can be obtained by utilizing the recursive solution to the FIV equations and then updating the fuzzy auxiliar model continuously on the basis of these recursive consequent parameters estimates. The FIV estimate in (32) can take the form

$$\hat{\theta}_k = \mathbf{P}_k \mathbf{b}_k \qquad (36)$$

where

$$\mathbf{P}_k = \sum_{j=1}^{k}[\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j][\gamma_j^1(\mathbf{x}_j + \xi_j), \ldots, \gamma_j^l(\mathbf{x}_j + \xi_j)]^T\}^{-1}$$

and

$$\mathbf{b}_k = \sum_{j=1}^{k} [\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j] y_j$$

which can be expressed as

$$\mathbf{P}_k^{-1} = \mathbf{P}_{k-1}^{-1} + [\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k][\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots, \gamma_k^l(\mathbf{x}_k + \xi_k)]^T \tag{37}$$

and

$$\mathbf{b}_k = \mathbf{b}_{k-1} + [\beta_k^1 \mathbf{z}_k, \ldots, \beta_k^l \mathbf{z}_k] y_k \tag{38}$$

respectively. Pre-multiplying (37) by $\mathbf{P}_k$ and post-multiplying by $\mathbf{P}_{k-1}$ gives

$$\mathbf{P}_{k-1} = \mathbf{P}_k + \mathbf{P}_k [\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k][\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots, \gamma_k^l(\mathbf{x}_k + \xi_k)]^T \mathbf{P}_{k-1} \tag{39}$$

then firstly post-multiplying (39) by the FIV vector $[\beta_j^1 \mathbf{z}_j, \ldots, \beta_j^l \mathbf{z}_j]$, and after that, post-multiplying by $\{1 + [\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots, \gamma_k^l(\mathbf{x}_k + \xi_k)]^T \mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k]\}^{-1}$ $[\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots, \gamma_k^l(\mathbf{x}_k + \xi_k)]^T \mathbf{P}_{k-1}$, results

$$\mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k] = \mathbf{P}_k[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k] +$$

$$\mathbf{P}_k[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k][\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots, \gamma_k^l(\mathbf{x}_k + \xi_k)]^T$$

$$\mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k] \mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k] =$$

$$\mathbf{P}_k[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k]\{1 + [\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots, \gamma_k^l(\mathbf{x}_k + \xi_k)]^T$$

$$\mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k]\} \tag{40}$$

Then, post-multiplying by $\{1 + [\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots, \gamma_k^l(\mathbf{x}_k + \xi_k)]^T \mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k]\}^{-1}$ $[\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots, \gamma_k^l(\mathbf{x}_k + \xi_k)]^T \mathbf{P}_{k-1}$, we obtain

$$\mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k]\{1 + [\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots, \gamma_k^l(\mathbf{x}_k + \xi_k)]^T$$

$$\mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k]\}^{-1}[\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots, \gamma_k^l(\mathbf{x}_k + \xi_k)]^T$$

$$\mathbf{P}_{k-1} = \mathbf{P}_k[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k][\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots, \gamma_k^l(\mathbf{x}_k + \xi_k)]^T$$

$$\mathbf{P}_{k-1} \tag{41}$$

Substituting (39) in (41), we have

$$\mathbf{P}_k = \mathbf{P}_{k-1} - \mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k]\{1 + [\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots,$$

$$\gamma_k^l(\mathbf{x}_k + \xi_k)]^T \mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k]\}^{-1}[\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots,$$

$$\gamma_k^l(\mathbf{x}_k + \xi_k)]^T \mathbf{P}_{k-1} \tag{42}$$

Substituting (42) and (38) in (36), the recursive consequent parameters estimates will be:

$$\hat{\theta}_k = \{\mathbf{P}_{k-1} - \mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k]\{1 + [\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots,$$

$$\gamma_k^l(\mathbf{x}_k + \xi_k)]^T \mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k]\}^{-1}[\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots,$$

$$\gamma_k^l(\mathbf{x}_k + \xi_k)]^T \mathbf{P}_{k-1}\}\{\mathbf{b}_{k-1} + [\beta_k^1 \mathbf{z}_k, \ldots, \beta_k^l \mathbf{z}_k] y_k\}$$

so that finally,

$$\hat{\theta}_k = \hat{\theta}_{k-1} - \mathbf{K}_k \{ [\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots, \gamma_k^l(\mathbf{x}_k + \xi_k)]^T \hat{\theta}_{k-1} - y_k \} \tag{43}$$

where

$$\mathbf{K}_k = \mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_k^l \mathbf{z}_k] \{ 1 + [\gamma_j^1(\mathbf{x}_k + \xi_k), \ldots, \gamma_k^l(\mathbf{x}_k + \xi_k)]^T \mathbf{P}_{k-1}[\beta_k^1 \mathbf{z}_k, \ldots, \beta_j^l \mathbf{z}_k] \}^{-1} \tag{44}$$

The equations (42)-(44) compose the FIV recursive estimation formula, and are implemented to determine unbiased estimates for the TS fuzzy model consequent parameters in a noisy environment.

6. COMPUTATIONAL RESULTS

In the sequel, two examples will be presented to demonstrate the effectiveness and applicability of the proposed algorithm in a noisy environment. Practical application of this method can be seen in (?), where was performed the identification of an aluminium beam, a complex nonlinear time varying plant whose study provides a great background for active vibration control applications in mechanical structures of aircrafts and/or aerospace vehicles.

6.1 Polynomial function approximation
Consider a nonlinear function defined by

$$u_k \quad = u_k^i + v_k \tag{45}$$

$$y_k^i \quad = 1 - 2u_k + u_k^2 \tag{46}$$

$$y_k \quad = y_k^i + c_k - 0.25c_{k-1} \tag{47}$$

In Fig. 2 are shown the true system ($u_k^i \in [0,2], y_k^i$) and the noisy (u_k, y_k) input-output observations with measurements corrupted by normal noise conditions of $\sigma_c = \sigma_v = 0.2$. The results for the TS fuzzy models obtained by applying the proposed FIV algorithm as well as the LS estimation to tune the consequent parameters are shown in Fig. 3. It can be seen, clearly, that the curves for the polynomial function and for the proposed FIV based identification almost cover each other. The fuzzy c-means clustering algorithm was used to criate the antecedent membership functions of the TS fuzzy models, which are shown in Fig. 4. The FIV was based on the filtered output from a "fuzzy auxiliar model" with the same structure of the TS fuzzy model used to identify the nonlinear function. The clusters centers of the membership functions for the LS and FIV estimations were $\mathbf{c} = [-0.0983, 0.2404, 0.6909, 1.1611]^T$ and $\mathbf{c} = [0.1022, 0.4075, 0.7830, 1.1906]^T$, respectively. The TS fuzzy models have the following structure:

$$R^i : \text{ IF } y_k \text{ is } F_i \text{ THEN } \hat{y}_k = a_0 + a_1 u_k + a_2 u_k^2$$

where $i = 1, 2, \ldots, 4$. For the FIV approach, the "fuzzy auxiliar model" has the following structure:

$$R^i : \text{ IF } y_{filt} \text{ is } F_i \text{ THEN } y_{filt} = a_0 + a_1 u_k + a_2 y_{filt}^2$$

where y_{filt} is the filtered output, based on the consequent parameters LS estimation, and used to criate the membership functions, as shown in Fig. 4, as well as the instrumental variable matrix. The resulting TS fuzzy models based on the LS estimation are:
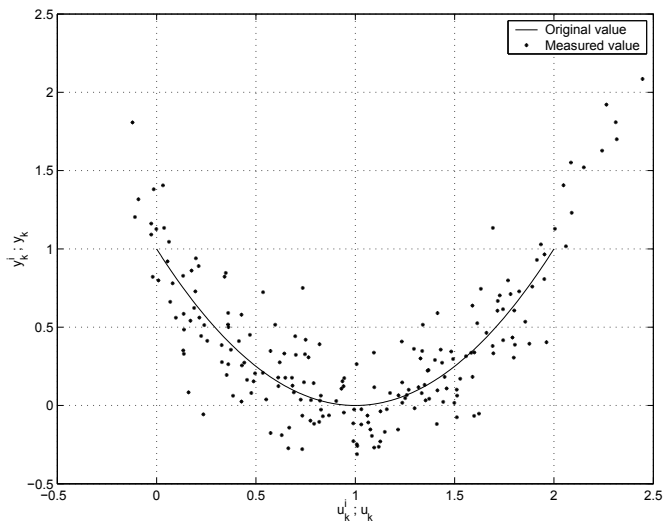
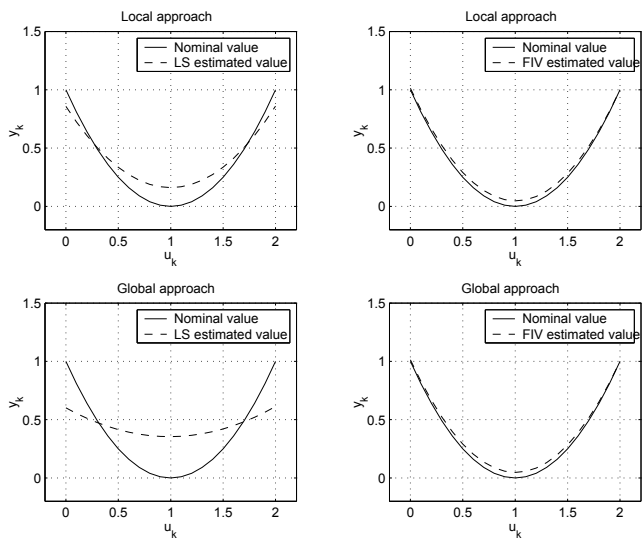Fig. 2. Polynomial function with error in variables.

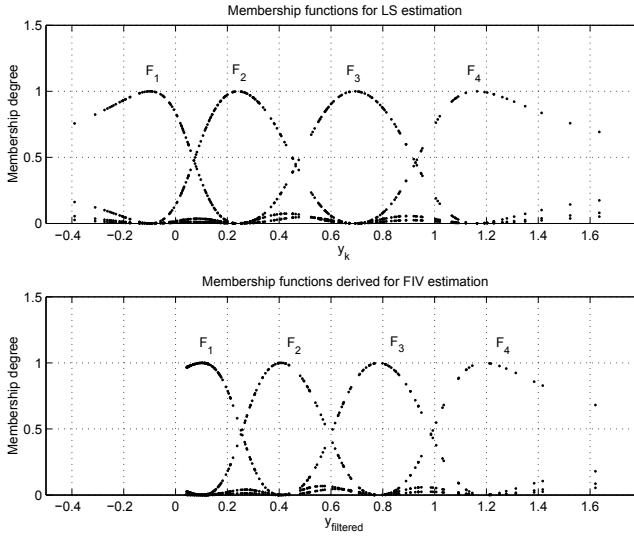Fig. 3. Approximation of the polynomial function.

Fig. 4. Antecedent membership functions.

Local approach:

$$R^1 : \text{ IF } y_k \text{ is } F_1 \text{ THEN } \hat{y}_k = 0.7074 - 1.7120u_k + 0.8717u_k^2$$
$$R^2 : \text{ IF } y_k \text{ is } F_2 \text{ THEN } \hat{y}_k = 0.7466 - 1.2077u_k + 0.5872u_k^2$$
$$R^3 : \text{ IF } y_k \text{ is } F_3 \text{ THEN } \hat{y}_k = 0.8938 - 1.1831u_k + 0.5935u_k^2$$
$$R^4 : \text{ IF } y_k \text{ is } F_4 \text{ THEN } \hat{y}_k = 1.0853 - 1.4776u_k + 0.7397u_k^2$$

Global approach:

$$R^1 : \text{ IF } y_k \text{ is } F_1 \text{ THEN } \hat{y}_k = 0.0621 - 0.4630u_k + 0.2272u_k^2$$
$$R^2 : \text{ IF } y_k \text{ is } F_2 \text{ THEN } \hat{y}_k = 0.3729 - 0.3068u_k + 0.1534u_k^2$$
$$R^3 : \text{ IF } y_k \text{ is } F_3 \text{ THEN } \hat{y}_k = 0.7769 - 0.3790u_k + 0.1891u_k^2$$
$$R^4 : \text{ IF } y_k \text{ is } F_4 \text{ THEN } \hat{y}_k = 1.1933 - 0.8500u_k + 0.4410u_k^2$$

According to Fig. 3, the obtained TS fuzzy models based on LS estimation are very poor and they were not able to aproximate the original nonlinear function data. It shows the influency of noise on the regressors of the data matrix, as explained in section 4, making the consequent parameters estimation biased and inconsistent. On the other hand, the resulting TS fuzzy models based on the FIV estimation are of the form:
Local approach:

$$R^1 : \text{ IF } y_k \text{ is } F_1 \text{ THEN } \hat{y}_k = 1.0130 - 1.9302u_k + 0.9614u_k^2$$
$$R^2 : \text{ IF } y_k \text{ is } F_2 \text{ THEN } \hat{y}_k = 1.0142 - 1.9308u_k + 0.9618u_k^2$$
$$R^3 : \text{ IF } y_k \text{ is } F_3 \text{ THEN } \hat{y}_k = 1.0126 - 1.9177u_k + 0.9555u_k^2$$
$$R^4 : \text{ IF } y_k \text{ is } F_4 \text{ THEN } \hat{y}_k = 1.0123 - 1.9156u_k + 0.9539u_k^2$$

Global approach:

$$R^1 : \text{ IF } y_k \text{ is } F_1 \text{ THEN } \hat{y}_k = 1.0147 - 1.9310u_k + 0.9613u_k^2$$
$$R^2 : \text{ IF } y_k \text{ is } F_2 \text{ THEN } \hat{y}_k = 1.0129 - 1.9196u_k + 0.9570u_k^2$$
$$R^3 : \text{ IF } y_k \text{ is } F_3 \text{ THEN } \hat{y}_k = 1.0125 - 1.9099u_k + 0.9508u_k^2$$
$$R^4 : \text{ IF } y_k \text{ is } F_4 \text{ THEN } \hat{y}_k = 1.0141 - 1.9361u_k + 0.9644u_k^2$$

In this application, to ilustrate the parametric convergence property, the consequent functions have the same structure of the polynomial function. It can be seen that the consequent parameters of the obtained TS fuzzy models based on FIV estimation are close to the nonlinear function parameters in (45)-(47), which shows the robustness of the proposed FIV method in a noisy environment as well as the capability of the identified TS fuzzy models for approximation and generalization of any nonlinear function with error in variables. Two criteria, widely used in analysis of experimental data and fuzzy modeling, can be applied to evaluate the fitness of the obtained TS fuzzy models : Variance Accounted For (VAF)

$$\mathbf{VAF}(\%) = 100 \times \left[1 - \frac{var(\mathbf{Y} - \hat{\mathbf{Y}})}{var(\mathbf{Y})} \right] \qquad (48)$$

where $\mathbf{Y}$ is the nominal output of the plant, $\hat{\mathbf{Y}}$ is the output of the TS fuzzy model and var means signal variance, and Mean Square Error (MSE)

$$\mathbf{MSE} = \frac{1}{N} \sum_{k=1}^{N} (y_k - \hat{y}_k)^2 \qquad (49)$$

where y_k is the nominal output of the plant, $\hat{y}_k$ is the output of the TS fuzzy model and N is the number of points. The obtained TS fuzzy models based on LS estimation presented performance with VAF and MSE of 74.4050% and 0.0226 for the local approach and of 6.0702% and 0.0943 for the global approach, respectively. The obtained TS fuzzy models based on FIV estimation presented performance with VAF and MSE of 99.5874% and 0.0012 for the local approach and of 99.5730% and 0.0013 for the global approach, respectively. The chosen fuzzy instrumental variables satisfied the Lemmas 1-3 as well as the Theorem 1, in section 4.1 and, as a consequence, the proposed algorithm becomes more robust to the noise.

6.2 On-line identification of a second-order nonlinear dynamic system
The plant to be identified consists on a second order highly nonlinear discrete-time system

$$
\begin{aligned}
u_k &= u_k^i + v_k \\
x_{k+1} &= \frac{x_k x_{k-1}(x_k + 2.5)}{1 + x_k^2 + x_{k-1}^2} + u(k) \\
y_{k+1} &= x_{k+1} + c_k - 0.5c_{k-1}
\end{aligned}
\qquad (50)
$$

which is, without noise, a benchmark problem in neural and fuzzy modeling (Narendra & Parthasarathy (1990); Papadakis & Theocaris (2002)), where $x(k)$ is the plant output and $u_k^i = 1.5\sin(\frac{2\pi k}{25})$ is the applied input. In this case v_k and c_k are white noise with zero mean and variance $\sigma_v^2 = \sigma_c^2 = 0.1$ meaning that the noise level applied to outputs takes values between

0 and $\pm20\%$ from its nominal values, which is an acceptable practical percentage of noise. The rule base, for the TS fuzzy model, is of the form:

$$R^i : \text{ IF } y_k \text{ is } F^i_{1,2} \text{ AND } y_{k-1} \text{ is } G^i_{1,2} \text{ THEN}$$

$$\hat{y}_{k+1} = a_{i,1}y_k + a_{i,2}y_{k-1} + b_{i,1}u_k + c_i \tag{51}$$

where $F^i_{1,2}|^{i=1,2,...,l}$ are gaussian fuzzy sets. For the FIV approach, the "fuzzy auxiliar model" has the following structure:

$$R^i : \text{ IF } y^{filt}_k \text{ is } F^i_{1,2} \text{ AND } y^{filt}_{k-1} \text{ is } G^i_{1,2} \text{ THEN}$$

$$\hat{y}^{filt}_{k+1} = a_{i,1}y^{filt}_k + a_{i,2}y^{filt}_{k-1} + b_{i,1}u_k + c_i \tag{52}$$

where $\hat{y}^{filt}$ is the filtered output, based on the consequent parameters LS estimation, and used to criate the membership functions as well as the fuzzy instrumental variable matrix. The number of rules is 4 for the TS fuzzy model, the antecedent parameters are obtained by the ECM method proposed in (Kasabov & Song (2002)). An experimental data set of 500 points is created from (50). The linguistic variables partitions obtained by the ECM method are shown in Fig. 5. The TS fuzzy model consequent parameters recursive estimate result is shown in
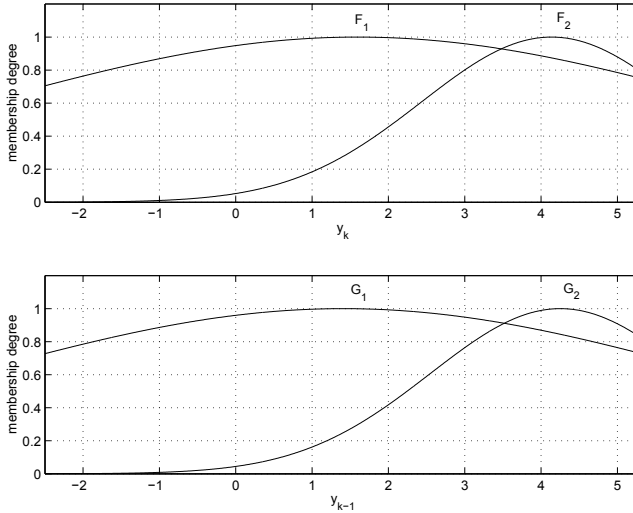
Fig. 5. Antecedent membership functions.

Fig. 6. The coefficient of determination, widely used in analysis of experimental data for time-series modeling, can be applied to evaluate the fitness of the obtained TS fuzzy models:

$$R^2_T = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)}{\sum_{i=1}^{T} y_i^2} \tag{53}$$

where y_i is the nominal output of the plant, $\hat{y}_i$ is the output of the TS fuzzy model and R_T is simply a normalized measure of the degree of explanation of the data. For its experiment the

Fig. 6. Recursive consequent parameters estimate.

Fig. 7. Fuzzy instrumental variable output tracking.

coefficient of determination is 0.9771.

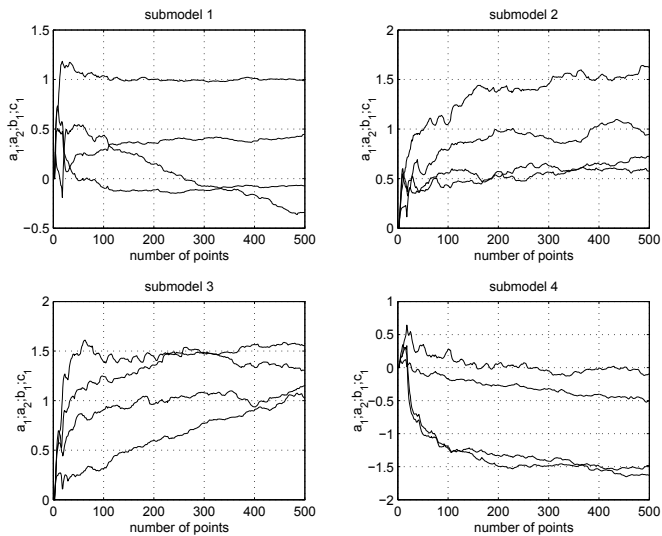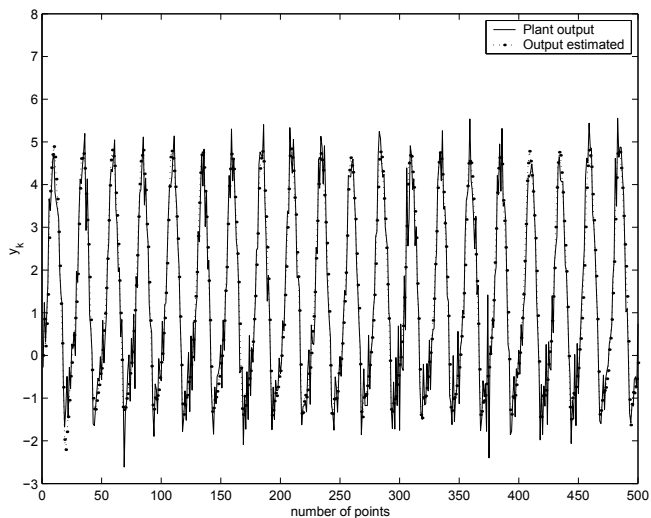According to Fig. 6, it can be seen that the algorithm is sensitive to the nonlinear plant behavior, the parameters estimates are consistent and converge rapidly. As expected, the proposed method provides unbiased and suficiently accurate estimates of the consequent parameters and, as a consequence, high speed of convergence of the TS fuzzy model to the nonlinear plant behavior in a noisy environment. These characteristics are very important in adaptive
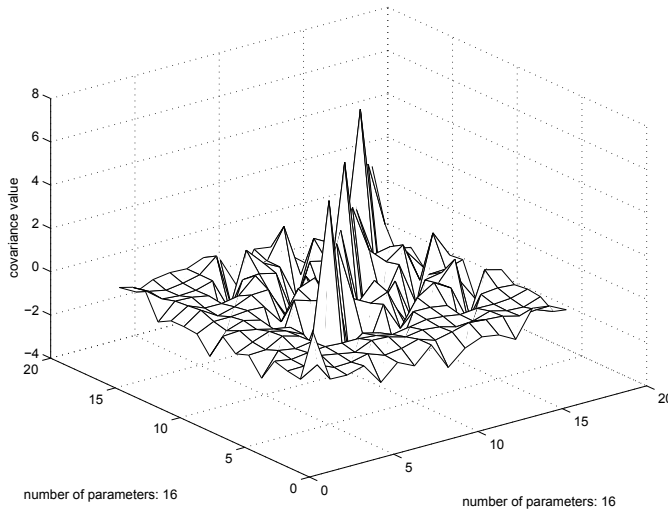
Fig. 8. Fuzzy covariance matrix P_k.

control design applications. The tracking of the nonlinear plant output is shown in Fig. 7. Figure 8 shows the fuzzy covariance matrix $\mathbf{P}_k$ of the recursive parameters estimatesfor the last point. It can be seen that the parametric uncertainty is close to zero and the higher values at this 3-D plot represent the principal diagonal entries, which determine the non-sigular property of this matrix due to fuzzy instrumental variable approach during the estimation process.

7. Conclusions

The concept of fuzzy instrumental variable and an approach for fuzzy identification of nonlinear discrete time systems were proposed. Convergence conditions for identification in a noisy environment in a fuzzy context were studied. Simulation results for off-line and on-line schemes evidence the good quality of this fuzzy instrumental variable approach for identification and function approximation with observation errors in input and output data.

8. References

Aguirre, L.A.; Coelho, M.C.S. & Correa, M.V. (2005). On the interpretation and practice of dynamical differences between Hammerstein and Wiener models, *IEE Proceedings of Control Theory and Applications*, Vol. 152, No. 4, Jul 2005, 349–356, ISSN 1350-2379.

Bergsten, P. (2001). *Observers and controllers for Takagi-Sugeno fuzzy systems*, Thesis, Örebro University.

Brown, M. & Harris, C. (1994). *Neurofuzzy adaptive modelling and control*, Prentice Hall.

Hellendoorn, H. & Driankov, D. (1997). *Fuzzy model identification: selected approaches*, Springer-Verlag.

Johansen, T.A.; Shorten, R. & Murray-Smith, R. (2000). On the interpretation and identification of dynamic Takagi-Sugeno fuzzy models, *IEEE Transactions on Fuzzy Systems*, Vol. 8, No. 3, Jun 2000, 297–313, ISSN 1063-6706.

Kadmiry, B. & Driankov, D. (2004). A fuzzy gain-scheduler for the attitude control of an unmanned helicopter, *IEEE Transactions on Fuzzy Systems*, Vol. 12, No. 3, Aug 2004, 502–515, ISSN 1063-6706.

Kasabov, N.K. & Song, Q. (2002). DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction, *IEEE Transactions on Fuzzy Systems*, Vol. 10, No. 2, Apr 2002, 144–154, ISSN 1063-6706.

King, R.E. (1999). *Computational intelligence in control engineering*, Marcel Dekker.

Ljung, L. (1999). *System Identification: Theory for the user*, Prentice Hall.

Narendra, K.S. & Parthasarathy, K. (1990). Identification and control of dynamical systems using neural networks, *IEEE Transactions on Neural Networks*, Vol. 1, No. 1, Mar 1990, 4–27, ISSN 1045-9227.

Papadakis, S.E. & Theocaris, J.B. (2002). A GA-based fuzzy modeling approach for generating TSK models, *Fuzzy Sets and Systems*, Vol. 131, No. 2, Oct 2002, 121–152, ISSN 0165-0114.

Serra, G.L.O. & Bottura, C.P. (2004). An algorithm for fuzzy identification of nonlinear discrete-time systems, *Proceedings of 43rd IEEE Conference on Decision and Control*, Vol. 5, pp.5521–5526, ISBN 0-7803-8682-5, Bahamas, Dec. 2004, Nassau.

Serra, G. L. O. & Bottura, C. P. (2005). Fuzzy instrumental variable concept and identification algorithm, *Proceedings of 14th IEEE Conference on Fuzzy Systems*, pp.1062–1067, ISBN 0-7803-9159-4, NV, May 2005, Reno.

Serra, G. L. O. & Bottura, C.P. (2006a). Multiobjective evolution based fuzzy PI controller design for nonlinear systems, *Engineering Applications of Artificial Intelligence*, Vol. 19, No. 2, Mar-2006, 157–167.

Serra, G. L. O. & Bottura, C. P. (2006b). An IV-QR algorithm for neuro-fuzzy multivariable on-line identification, *IEEE Transactions on Fuzzy Systems*, Vol. 15; No. 2, Apr-2007, 200–210, ISSN 1063-6706.

Sjöberg, J.; Zhang, Q.; Ljung, L.; Benveniste, A.; Delyon, B.; Glorennec, P.; Hjalmarsson, H. & Juditsky, A. (1995). Nonlinear black-box modeling in system identification : an unified overview, *Automatica: Special issue on trends in system identification*, Vol. 31, No. 12, Dec-1995, 1691–1724, ISSN 0005-1098.

Söderström, T. and Stoica, P. (1989). *System identification*, Prentice Hall.

Söderström, T. and Stoica, P. (1983). *Instrumental variable methods for system identification*, Springer.

Takagi, T. & Sugeno, M. (1985). Fuzzy identification of systems and its aplication to modeling and control, *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 15, No. 1, 116–132, ISSN: 1083-4419.

Tanaka, K.; Ikeda, T. & Wang, H. (1998). Fuzzy regulators and fuzzy observers: relaxed stability conditions and LMI-based designs, *IEEE Transactions on Fuzzy Systems*, Vol. 6, No. 2, May-1998, 250–265, ISSN 1063-6706.

Tong, S. & Li, H. (2002). Observer-based robust fuzzy control of nonlinear systems with parametric uncertainties. *Fuzzy Sets and Systems*, Vol. 131, No. 2, Oct-2002, 165–184, ISSN 0165-0114.

Fuzzy frequency response for stochastic linear parameter varying dynamic systems

Carlos C. T. Ferreira and Ginalber L. O. Serra
Federal Institute of Education, Science and Technology (IFMA)
Brazil

1. Introduction

The design of control systems is currently managed by a large number of requirements such as: increasing competition, environmental requirements, energy and material costs, and demand for robust and fault-tolerant systems which require considerations for effective process control techniques. In this context, the analysis and synthesis of compensators are completely related to each other. In the analysis, the characteristics or dynamic behaviour of the control system are determined. In the synthesis, the compensators are obtained to attend to desired characteristics of the control system based on certain performance criteria. Generally, these criteria may involve disturbance rejection, steady-state errors, transient response characteristics and sensitivity to parameter changes in the plant (Franklin et al. (1986); Ioannou & Sun (1996); Phillips & Harbor (1996)).

Test input signals is one way to analyse the dynamic behaviour of real world system. Many test signals are available, but a simple and useful signal is the sinusoidal wave form because the system output with a sinusoidal wave input is also a sinusoidal wave, but with a different amplitude and phase for a given frequency. This frequency response analysis describes how a dynamic system responds to sinusoidal inputs in a range of frequencies and it has been widely used in academic field and industry, as well as been considered essential for robust control theory (Serra & Bottura (2006; 2009); Serra et al. (2009); Tanaka et al. (1998)).

The frequency response methods were developed during the period $1930 - 1940$ by Harry Nyquist $(1889 - 1976)$ (Nyquist (1932)), Hendrik Bode $(1905 - 1982)$ (Bode (1940)), Nathaniel B. Nichols $(1914 - 1997)$ (James et al. (1947)), and many others. Since then, frequency response methods are among the most useful techniques being available to analyse and synthesise the compensators. In (Jr. (1973)), the U.S. Navy obtains frequency responses for aircraft by applying sinusoidal inputs to the autopilots and measuring its resulting position in flight. In (Lascu et al. (2009)), four current controllers for selective harmonic compensation in parallel with Active Power Filters (APFs) have been analytically compared in terms of frequency response characteristics and maximum operational frequency.

Most real systems, such as circuit components (inductor, resistor, operational amplifier, etc.) are often formulated using differential/integral equations with stochastic parameters (Kolev (1993)). These random variations are most often quantified in terms of boundaries. The classical methods of frequency response do not explore these boundaries for Stochastic Linear Parameter Varying (SLPV) dynamic systems. To overcome this limitation, this chapter pro-

poses the definition of Fuzzy Frequency Response (FFR) and its application for analysis of stochastic linear parameter varying dynamic systems.

2. Formulation Problem

This section presents some essential concepts for the formulation and development of this chapter.

2.1 Stochastic Linear Parameter Varying Dynamic Systems

In terms of transfer function, the general form of SLPV dynamic system is given by Eq. (1), as depicted in Fig.1.
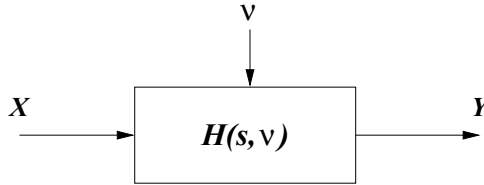

Fig. 1. SLPV dynamic sistem.

$$H(s, \nu) = \frac{Y(s, \nu)}{X(s)} = \frac{b_\alpha(\nu)s^\alpha + b_{\alpha-1}(\nu)s^{\alpha-1} + \ldots + b_\alpha(\nu)s^\alpha + b_1(\nu)s + b_0(\nu)}{s^\beta + a_{\beta-1}(\nu)s^{\beta-1} + \ldots + a_1(\nu)s + a_0(\nu)},$$

(1)

where:

- $H(s, \nu)$ is the transfer function of the SLPV dynamic system;
- $X(s)$ and $Y(s, \nu)$ represent the input and output of SLPV dynamic system, respectively;
- $a_*(\nu)$ and $b_*(\nu)$ are the varying parameters;
- $\nu(t)$ is the time varying scheduling variable;
- s is the Laplace operator and
- α and β are the orders of the numerator and denominator of the transfer function, respectively (with $\beta \geq \alpha$).

The scheduling variable ν belongs to a compact set $\nu \in V$, with its variation limited by $|\dot{\nu}| \leq d^{\max}$, with $d^{\max} \geq 0$.

2.2 Takagi-Sugeno Fuzzy Dynamic Model

The TS inference system, originally proposed in (Takagi (1985)), presents in the consequent a dynamic functional expression of the linguistic variables of the antecedent. The $i \left| \right.^{[i=1,2,\ldots,l]}$ -th rule, where l is the number of rules, is given by:

$$Rule^{(i)} : IF \ \tilde{x}_1 \ is \ F^i_{\{1,2,\ldots,p_{\tilde{x}_1}\}|_{\tilde{x}_1}} \ AND \ldots AND \ \tilde{x}_n \ is \ F^i_{\{1,2,\ldots,p_{\tilde{x}_n}\}|_{\tilde{x}_n}}$$

$$\text{THEN} \quad y_i = f_i(\tilde{\mathbf{x}}), \tag{2}$$

where the total number of rules is $l = p_{\tilde{x}_1} \times \ldots \times p_{\tilde{x}_n}$. The vector $\tilde{\mathbf{x}} = [\tilde{x}_1, \ldots, \tilde{x}_n]^T \in \Re^n$ containing the linguistic variables of antecedent, where T represents the operator for transpose matrix. Each linguistic variable has its own discourse universe $\mathcal{U}_{\tilde{x}_1}, \ldots, \mathcal{U}_{\tilde{x}_n}$, partitioned by fuzzy sets representing its linguistics terms, respectively. In i-th rule, the variable $\tilde{x}_{\{1,2,\ldots,n\}}$ belongs to the fuzzy set $F^i_{\{\tilde{x}_1,\ldots,\tilde{x}_n\}}$ with a membership degree $\mu^i_{F_{\{\tilde{x}_1,\ldots,\tilde{x}_n\}}}$ defined by a membership function $\mu^i_{\{\tilde{x}_1,\ldots,\tilde{x}_n\}} : \Re \to [0,1]$, with $\mu^i_{F_{\{\tilde{x}_1,\ldots,\tilde{x}_n\}}} \in \{\mu^i_{F_{1|\{\tilde{x}_1,\ldots,\tilde{x}_n\}}}, \mu^i_{F_{2|\{\tilde{x}_1,\ldots,\tilde{x}_n\}}}, \cdots, \mu^i_{F_{p|\{\tilde{x}_1,\ldots,\tilde{x}_n\}}}\}$, where $p_{\{\tilde{x}_1,\ldots,\tilde{x}_n\}}$ is the partition number of the discourse universe, associated with the linguistic variable $\tilde{x}_1, \ldots, , \tilde{x}_n$. The TS fuzzy dynamic model output is a convex combination of the dynamic functional expressions of consequent $f_i(\tilde{\mathbf{x}})$, without loss of generality for the bidimensional case, as illustrated in Fig. 2, given by Eq. (3).

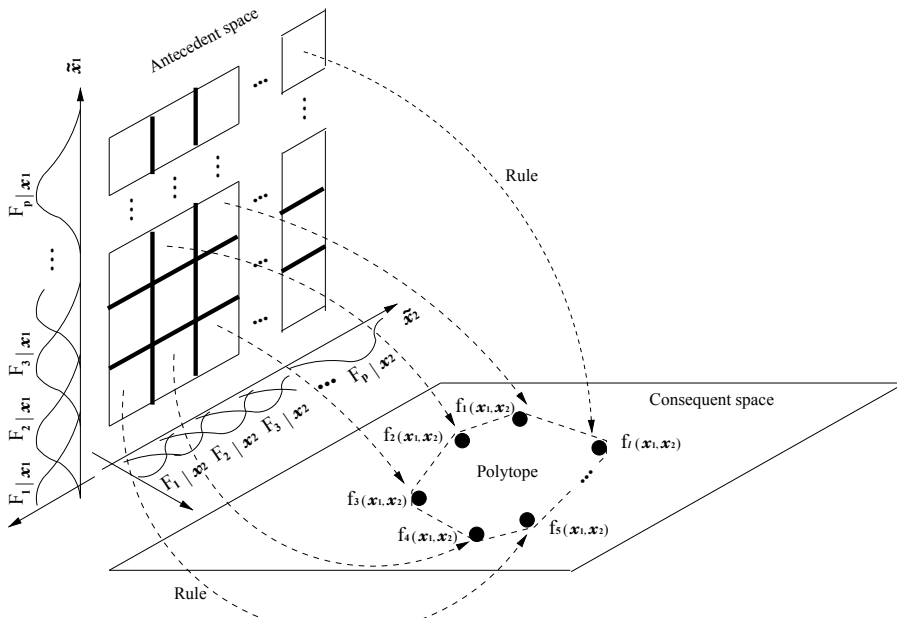Fig. 2. Fuzzy dynamic model: A TS model can be regarded as a mapping from the antecedent space to the consequent parameter one.

$$y(\tilde{\mathbf{x}}, \gamma) = \sum_{i=1}^{l} \gamma_i(\tilde{\mathbf{x}}) f_i(\tilde{\mathbf{x}}), \tag{3}$$

where γ is the scheduling variable of the TS fuzzy dynamic model.
The scheduling variable, well known as normalized activation degree, is given by:

$$\gamma_i(\tilde{\mathbf{x}}) = \frac{h_i(\tilde{\mathbf{x}})}{\sum\limits_{r=1}^{l} h_r(\tilde{\mathbf{x}})}. \tag{4}$$

This normalization implies

$$\sum_{k=1}^{l} \gamma_i(\tilde{\mathbf{x}}) = 1. \tag{5}$$

It can be observed that the TS fuzzy dynamic system, which represents any stochastic dynamic model, may be considered as a class of systems where $\gamma_i(\tilde{\mathbf{x}})$ denotes a decomposition of linguistic variables $[\tilde{x}_1, \ldots, \tilde{x}_n]^T \in \Re^n$ for a polytopic geometric region in the consequent space based on the functional expressions $f_i(\tilde{\mathbf{x}})$.

3. Fuzzy Frequency Response (FFR): Definition

This section demonstrates how a TS fuzzy dynamic model responds to sinusoidal inputs, which is proposed as the definition of fuzzy frequency response. The response of a TS fuzzy dynamic model to a sinusoidal input of frequency ω_1, in both amplitude and phase, is given by the transfer function evaluated at $s = j\omega_1$, as illustrated in Fig. 3.

$$E(s) \longrightarrow \boxed{\widetilde{W}(s) = \sum_{i=1}^{l} \gamma_i W^i(s)} \longrightarrow Y(s)$$
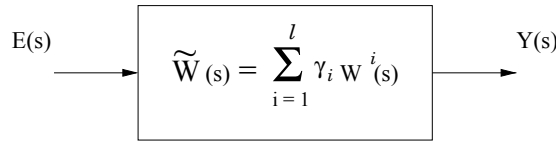
Fig. 3. TS fuzzy transfer function.

For this TS fuzzy dynamic model:

$$Y(s) = \left[\sum_{i=1}^{l} \gamma_i W^i(s) \right] E(s). \tag{6}$$

Consider $\tilde{W}(j\omega) = \sum\limits_{i=1}^{l} \gamma_i W^i(j\omega)$ as a complex number for a given ω:

$$\tilde{W}(j\omega) = \sum_{i=1}^{l} \gamma_i W^i(j\omega) = \left| \sum_{i=1}^{l} \gamma_i W^i(j\omega) \right| e^{j\phi(\omega)} = \left| \sum_{i=1}^{l} \gamma_i W^i(j\omega) \right| \angle \phi(\omega) \tag{7}$$

or

$$\tilde{W}(j\omega) = \sum_{i=1}^{l} \gamma_i W^i(j\omega) = \left| \sum_{i=1}^{l} \gamma_i W^i(j\omega) \right| \angle \arctan \left[\sum_{i=1}^{l} \gamma_i W^i(j\omega) \right]. \tag{8}$$

Then, for the case that the input signal $e(t)$ is sinusoidal, that is:

$$e(t) = A \sin \omega_1 t. \tag{9}$$

The output signal $y_{ss}(t)$, in the steady state, is given by

$$y_{ss}(t) = A \left| \sum_{i=1}^{l} \gamma_i W^i(j\omega) \right| \sin \left[\omega_1 t + \phi(\omega_1) \right]. \tag{10}$$

As a result of the fuzzy frequency response definition, it is then proposed the following *Theorem*:

Theorem 3.1. *Fuzzy frequency response is a region in the frequency domain, defined by the consequent sub-models and based on the operating region of the antecedent space.*

Proof. Considering that $\tilde{v}$ is stochastic and can be represented by linguistic terms, once known its discourse universe, as shown in Fig. 4, the activation degrees, $h_i(\tilde{v})|^{i=1,2,...,l}$ are also stochastic, since it dependes of the dynamic system:

$$h_i(\tilde{v}) = \mu^i_{F_{\tilde{v}_1^*}} \star \mu^i_{F_{\tilde{v}_2^*}} \star \ldots \star \mu^i_{F_{\tilde{v}_n^*}}, \tag{11}$$

where $\tilde{v}^*_{\{1,2,...,n\}} \in \mathcal{U}_{\tilde{v}_{\{1,2,...,n\}}}$, respectively, and $\star$ is a fuzzy logic operator.
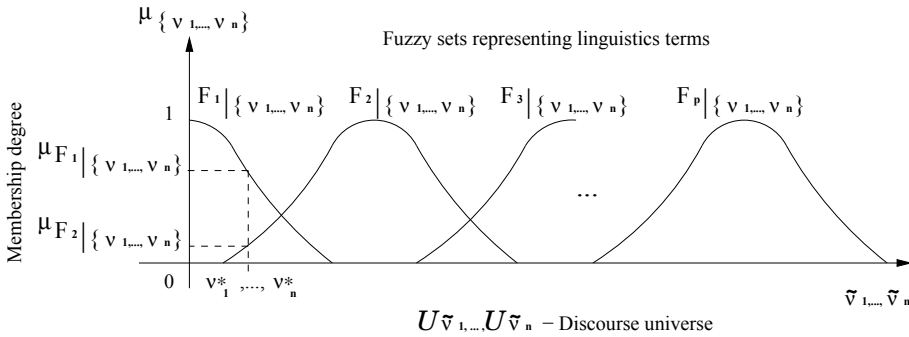

Fig. 4. Functional description of the linguistic variable $\tilde{v}$: linguistic terms, discourse universe and membership degrees.

So, the normalized activation degrees $\gamma_i(\tilde{v})|^{i=1,2,...,l}$, are also stochastics:

$$\gamma_i(\tilde{v}) = \frac{h_i(\tilde{v})}{\sum_{r=1}^{l} h_r(\tilde{v})}. \tag{12}$$

This normalization implies

$$\sum_{k=1}^{l} \gamma_i(\tilde{v}) = 1. \tag{13}$$

Let $F(s)$ be a vectorial space with degree l and $f^1(s), f^2(s), \ldots, f^l(s)$ be transfer functions which belong to this vectorial space. A transfer function $f(s) \in F(s)$ must be a linear convex combination of the vectors $f^1(s), f^2(s), \ldots, f^l(s)$:

$$f(s) \quad = \quad \xi_1 f^1(s) + \xi_2 f^2(s) + \ldots + \xi_l f^l(s), \tag{14}$$

where $\xi_{1,2,\ldots,l}$ are the linear convex combination coefficients. If they are normalized $\left(\sum_{i=1}^{l} \xi_i = 1 \right)$, the vectorial space presents a decomposition of the transfer functions $\left[f^1(s), f^2(s), \ldots, f^l(s) \right]$ in a polytopic geometric shape of the vectorial space $F(s)$. The points of the polytopic geometric shape are defined by the transfer functions $\left[f^1(s), f^2(s), \ldots, f^l(s) \right]$.

The TS fuzzy dynamic model attends to this polytopic property. The sum of the normalized activation degrees is equal to 1, as demonstrated in Eq. (5). To define the points of this fuzzy polytopic geometric shape, each rule of the TS fuzzy dynamic model must be individually activated. This condition is called boundary condition. Thus, the following results are obtained for the Fuzzy Frequency Response (FFR) of the TS fuzzy transfer function:

- *If only the rule* 1 *is activated, it has* $(\gamma_1 = 1, \gamma_2 = 0, \gamma_3 = 0, \ldots, \gamma_l = 0)$. *Hence,*

$$\tilde{W}(j\omega, \tilde{v}) = \left| \sum_{i=1}^{l} \gamma_i(\tilde{v}) W^i(j\omega) \right| \angle \arctan \left[\sum_{i=1}^{l} \gamma_i(\tilde{v}) W^i(j\omega) \right], \tag{15}$$

$$\tilde{W}(j\omega, \tilde{v}) = \left| 1W^1(j\omega) + 0W^2(j\omega) + \ldots + 0W^l(j\omega) \right| \angle \arctan$$
$$\left[1W^1(j\omega) + 0W^2(j\omega) + \ldots + 0W^l(j\omega) \right], \tag{16}$$

$$\tilde{W}(j\omega, \tilde{v}) = \left| W^1(j\omega) \right| \angle \arctan \left[W^1(j\omega) \right]. \tag{17}$$

- *If only the rule* 2 *is activated, it has* $(\gamma_1 = 0, \gamma_2 = 1, \gamma_3 = 0, \ldots, \gamma_l = 0)$. *Hence,*

$$\tilde{W}(j\omega, \tilde{v}) = \left| \sum_{i=1}^{l} \gamma_i(\tilde{v}) W^i(j\omega) \right| \angle \arctan \left[\sum_{i=1}^{l} \gamma_i(\tilde{v}) W^i(j\omega) \right], \tag{18}$$

$$\tilde{W}(j\omega, \tilde{v}) = \left| 0W^1(j\omega) + 1W^2(j\omega) + \ldots + 0W^l(j\omega) \right| \angle \arctan$$
$$\left[0W^1(j\omega) + 1W^2(j\omega) + \ldots + 0W^l(j\omega) \right], \tag{19}$$

$$\tilde{W}(j\omega, \tilde{v}) = \left|W^2(j\omega)\right| \angle \arctan\left[W^2(j\omega)\right]. \tag{20}$$

- *If only the rule l is activated, it has* $(\gamma_1 = 0, \gamma_2 = 0, \gamma_3 = 0, \ldots, \gamma_l = 1)$. *Hence,*

$$\tilde{W}(j\omega, \tilde{v}) = \left|\sum_{i=1}^{l} \gamma_i(\tilde{v})W^i(j\omega)\right| \angle \arctan\left[\sum_{i=1}^{l} \gamma_i(\tilde{v})W^i(j\omega)\right], \tag{21}$$

$$\tilde{W}(j\omega, \tilde{v}) = \left|0W^1(j\omega) + 0W^2(j\omega) + \ldots + 1W^l(j\omega)\right| \angle \arctan$$
$$\left[0W^1(j\omega) + 0W^2(j\omega) + \ldots + 1W^l(j\omega)\right], \tag{22}$$

$$\tilde{W}(j\omega, \tilde{v}) = \left|W^l(j\omega)\right| \angle \arctan\left[W^l(j\omega)\right]. \tag{23}$$

Where $W^1(j\omega), W^2(j\omega), \ldots, W^l(j\omega)$ are the linear sub-models of the uncertain dynamic system.
Note that $\left|W^1(j\omega)\right| \angle \arctan\left[W^1(j\omega)\right]$ and $\left|W^l(j\omega)\right| \angle \arctan\left[W^l(j\omega)\right]$ define a boundary region. Under such circumstances, the fuzzy frequency response converges to a boundary in the frequency domain defined by a surface based on membership degrees. Figure 5 shows the fuzzy frequency response for the bidimensional case, without loss of generality.

□

4. Fuzzy Frequency Response (FFR): Analysis

In this section, the behaviour of the fuzzy frequency response is analysed at low and high frequencies. The idea is to study the magnitude and phase behaviour of the TS fuzzy dynamic model, when ω varies from zero to infinity.

4.1 Low Frequency Analysis
Low frequency analysis of the TS fuzzy dynamic model $\tilde{W}(s)$ can be obtained by:

$$\lim_{\omega \to 0} \sum_{i=1}^{l} \gamma_i W^i(j\omega). \tag{24}$$

The magnitude and phase behaviours at low frequencies, are given by

$$\lim_{\omega \to 0} \left|\sum_{i=1}^{l} \gamma_i W^i(j\omega)\right| \angle \arctan\left[\sum_{i=1}^{l} \gamma_i W^i(j\omega)\right]. \tag{25}$$

Fig. 5. Fuzzy frequency response: mapping from the consequent space to frequency domain region.

4.2 High Frequency Analysis

Likewise, the high frequency analysis of the TS fuzzy dynamic model $\tilde{W}(s)$ can be obtained by:

$$\lim_{\omega \to \infty} \sum_{i=1}^{l} \gamma_i W^i(j\omega). \tag{26}$$

The magnitude and phase behaviours at high frequencies, are given by

$$\lim_{\omega \to \infty} \left| \sum_{i=1}^{l} \gamma_i W^i(j\omega) \right| \angle \arctan \left[\sum_{i=1}^{l} \gamma_i W^i(j\omega) \right]. \tag{27}$$
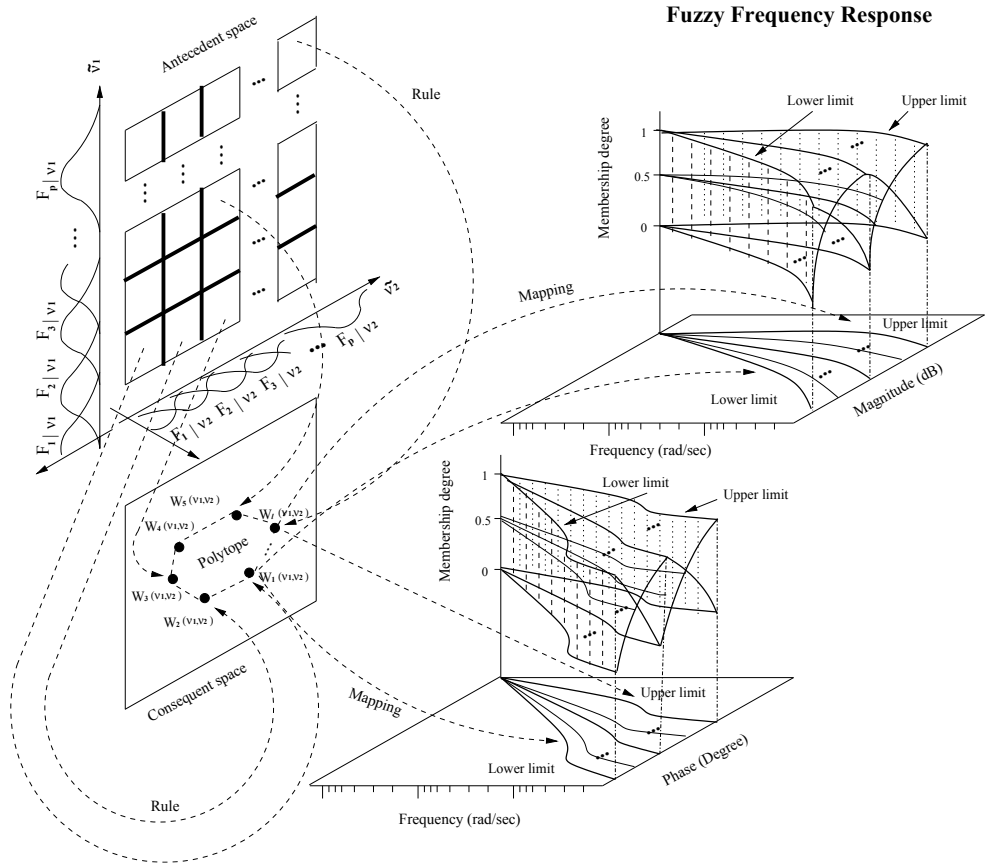
5. Computational Results

To illustrate the FFR: definition and analysis, as shown in sections 3 and 4, consider the following SLPV dynamic system, given by

$$H(s,v) = \frac{Y(s,v)}{U(s)} = \frac{2-v}{[(v+1)s+1]\left[\left(\frac{v}{2}+0.1\right)s+1\right]}, \tag{28}$$

where the scheduling variable is $v = [0,1]$, the gain of the SLPV dynamic system is $K_p = 2-v$, the higher time constant is $\tau = v+1$, and the lower time constant is $\tau' = \frac{v}{2} + 0.1$.

Starting from the SLPV dynamic system in Eq. (28) and assuming the time varying scheduling variable in the range of $[0,1]$, one can obtain the TS fuzzy dynamic model in the following operating points:

<u>Sub-model 1 ($v = 0$):</u>

$$W^1(s,0) = \frac{2}{(s+1)(0.1s+1)} = \frac{2}{0.1s^2 + 1.1s + 1}. \tag{29}$$

<u>Sub-model 2 ($v = 0.5$):</u>

$$W^2(s,0.5) = \frac{1.5}{(1.5s+1)(0.35s+1)} = \frac{1.5}{0.525s^2 + 1.85s + 1}. \tag{30}$$

<u>Sub-model 3 ($v = 1$):</u>

$$W^3(s,1) = \frac{1}{(2s+1)(0.6s+1)} = \frac{1}{1.2s^2 + 2.6s + 1}. \tag{31}$$

The TS fuzzy dynamic model rule base results in:

$$\begin{array}{llllll} Rule^{(1)}: & \text{IF} & v & \text{is} & 0 & \text{THEN} & W^1(s,0) \\ Rule^{(2)}: & \text{IF} & v & \text{is} & 0.5 & \text{THEN} & W^2(s,0.5) \\ Rule^{(3)}: & \text{IF} & v & \text{is} & 1 & \text{THEN} & W^3(s,1), \end{array} \tag{32}$$

and the TS fuzzy dynamic model of the SLPV dynamic system is given by

$$\tilde{W}(s,\tilde{v}) = \sum_{i=1}^{3} \gamma_i(\tilde{v}) W^i(s). \tag{33}$$

Again, starting from Eq. (28), one obtains:

$$Y(s,v) = \cfrac{2-v}{\left(\cfrac{v^2}{2}+0.1v+\cfrac{v}{2}+0.1\right)s^2+\left(v+1+0.1+\cfrac{v}{2}\right)s+1}U(s),$$

$$(34)$$

$$\left(\frac{v^2+1.2v+0.2}{2}\right)s^2Y(s,v)+\left(\frac{3v+2.2}{2}\right)sY(s,v)+Y(s,v)=(2-v)\,U(s)$$

$$(35)$$

and taking the inverse Laplace transform, this yields the differential equation of the SLPV dynamic system:

$$\left(\frac{v^2+1.2v+0.2}{2}\right)\ddot{y}(t)+\left(\frac{3v+2.2}{2}\right)\dot{y}(t)+y(t)=(2-v)\,u(t). \qquad (36)$$

A comparative analysis, via analog simulation between the SLPV dynamic system Eq. (36) and the TS fuzzy dynamic model Eq. (33), can be performed to validate the TS fuzzy dynamic model. A band-limited white noise (normally distributed random signal) was considered as input and the stochastic parameter was based on sinusoidal variation. As shown in Fig. 6, the efficiency of the TS fuzzy dynamic model in order to represent the dynamic behaviour of the SLPV dynamic system in the time domain can be seen.

From Eq. (8) the TS fuzzy dynamic model of the SLPV dynamic system, Eq. (33), can be represented by

$$\tilde{W}(j\omega,\tilde{v}) = \left|\sum_{i=1}^{3}\gamma_i(\tilde{v})W^i(j\omega)\right|\angle\arctan\left[\sum_{i=1}^{3}\gamma_i(\tilde{v})W^i(j\omega)\right] \qquad (37)$$

or

$$\tilde{W}(j\omega,\tilde{v}) = \left|\gamma_1 W^1(j\omega,0)+\gamma_2 W^2(j\omega,0.5)+\gamma_3 W^3(j\omega,1)\right|$$

$$\angle\arctan\left[\gamma_1 W^1(j\omega,0)+\gamma_2 W^2(j\omega,0.5)\ +\gamma_3 W^3(j\omega,1)\right]. \qquad (38)$$

So,

$$\tilde{W}(j\omega,\tilde{v}) = \left|\gamma_1\frac{2}{0.1s^2+1.1s+1}+\gamma_2\frac{1.5}{0.525s^2+1.85s+1}+\right.$$

$$\left.+\gamma_3\frac{1}{1.2s^2+2.6s+1}\right|\angle\arctan\left[\gamma_1\frac{2}{0.1s^2+1.1s+1}+\right.$$
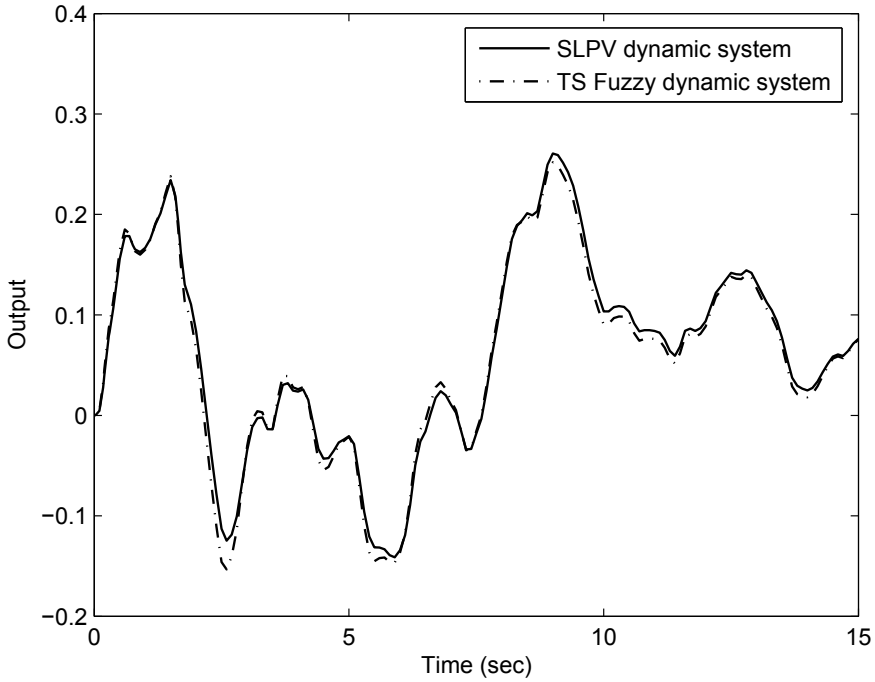
Fig. 6. Validation of the TS fuzzy dynamic model.

$$+\gamma_2 \frac{1.5}{0.525s^2 + 1.85s + 1} + \gamma_3 \frac{1}{1.2s^2 + 2.6s + 1}\Bigg],$$ (39)

$$\tilde{W}(j\omega, \tilde{v}) = \left| 2\gamma_1 \frac{0.6(j\omega)^4 + 3.6(j\omega)^3 + 6.5(j\omega)^2 + 4.5(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega, \tilde{v})\right]}} + \right.$$

$$1.5\gamma_2 \frac{0.1(j\omega)^4 + 1.6(j\omega)^3 + 4.2(j\omega)^2 + 3.7(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega, \tilde{v})\right]}} +$$

$$\left. \gamma_3 \frac{0.1(j\omega)^4 + 0.8(j\omega)^3 + 2.7(j\omega)^2 + 3(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega, \tilde{v})\right]}} \right| \angle \arctan$$

$$\left[2\gamma_1 \frac{0.6(j\omega)^4 + 3.6(j\omega)^3 + 6.5(j\omega)^2 + 4.5(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega, \tilde{v})\right]}} + \right.$$

$$1.5\gamma_2 \frac{0.1(j\omega)^4 + 1.6(j\omega)^3 + 4.2(j\omega)^2 + 3.7(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} +$$

$$\gamma_3 \frac{0.1(j\omega)^4 + 0.8(j\omega)^3 + 2.7(j\omega)^2 + 3(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} \Bigg],$$

(40)

where:

$$Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]} = 0.1(j\omega)^6 + 1.1(j\omega)^5 + 5.2(j\omega)^4 + 11.2(j\omega)^3 + 11.5(j\omega)^2 + 5.6(j\omega) + 1. \quad (41)$$

5.1 Low Frequency Analysis

Starting from the TS fuzzy dynamic model, Eq. (37), and applying the concepts presented in the subsection 4.1, the steady-state response for sinusoidal input at low frequencies for the SLPV dynamic system can be obtained as follows:

$$\lim_{\omega \to 0} \tilde{W}(j\omega,\tilde{v}) = \Bigg| 2\gamma_1 \frac{0.6(j\omega)^4 + 3.6(j\omega)^3 + 6.5(j\omega)^2 + 4.5(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} +$$

$$1.5\gamma_2 \frac{0.1(j\omega)^4 + 1.6(j\omega)^3 + 4.2(j\omega)^2 + 3.7(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} +$$

$$\gamma_3 \frac{0.1(j\omega)^4 + 0.8(j\omega)^3 + 2.7(j\omega)^2 + 3(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} \Bigg| \angle \arctan$$

$$\Bigg[2\gamma_1 \frac{0.6(j\omega)^4 + 3.6(j\omega)^3 + 6.5(j\omega)^2 + 4.5(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} +$$

$$1.5\gamma_2 \frac{0.1(j\omega)^4 + 1.6(j\omega)^3 + 4.2(j\omega)^2 + 3.7(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} +$$

$$+\gamma_3 \frac{0.1(j\omega)^4 + 0.8(j\omega)^3 + 2.7(j\omega)^2 + 3(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} \Bigg].$$

(42)

As ω tends to zero, Eq. (42) can be approximated as follows:

$$\lim_{\omega \to 0} \tilde{W}(j\omega,\tilde{v}) = |2\gamma_1 + 1.5\gamma_2 + \gamma_3| \angle \arctan\left[2\gamma_1 + 1.5\gamma_2 + \gamma_3\right]. \quad (43)$$

Hence

$$\lim_{\omega \to 0} \tilde{W}(j\omega, \tilde{v}) = |2\gamma_1 + 1.5\gamma_2 + \gamma_3| \angle 0^o. \tag{44}$$

Applying the *Theorem 3.1*, proposed in section 3, the obtained boundary conditions at low frequencies are presented in Tab. 1. The fuzzy frequency response of the SLPV dynamic system, at low frequencies, presents a magnitude range in the interval $[0; 6.0206](dB)$ and the phase is 0°.

Table 1. Boundary conditions at low frequencies.

Activated Rule	Boundary Condition	Magnitude (dB)	Phase (Degree)
1	$\gamma_1 = 1; \gamma_2 = 0$ and $\gamma_3 = 0$	6.0206	0°
2	$\gamma_1 = 0; \gamma_2 = 1$ and $\gamma_3 = 0$	3.5218	0°
3	$\gamma_1 = 0; \gamma_2 = 0$ and $\gamma_3 = 1$	0	0°

5.2 High Frequency Analysis

Likewise, starting from the TS fuzzy dynamic model, Eq. (37), and now applying the concepts seen in the subsection 4.2, the steady-state response for sinusoidal input at high frequencies for the SLPV dynamic system can be obtained as follows:

$$\lim_{\omega \to \infty} \tilde{W}(j\omega, \tilde{v}) = \left| 2\gamma_1 \frac{0.6(j\omega)^4 + 3.6(j\omega)^3 + 6.5(j\omega)^2 + 4.5(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} + \right.$$

$$1.5\gamma_2 \frac{0.1(j\omega)^4 + 1.6(j\omega)^3 + 4.2(j\omega)^2 + 3.7(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} +$$

$$\left. \gamma_3 \frac{0.1(j\omega)^4 + 0.8(j\omega)^3 + 2.7(j\omega)^2 + 3(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} \right| \angle \arctan$$

$$\left[2\gamma_1 \frac{0.6(j\omega)^4 + 3.6(j\omega)^3 + 6.5(j\omega)^2 + 4.5(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} + \right.$$

$$1.5\gamma_2 \frac{0.1(j\omega)^4 + 1.6(j\omega)^3 + 4.2(j\omega)^2 + 3.7(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} +$$

$$\left. \gamma_3 \frac{0.1(j\omega)^4 + 0.8(j\omega)^3 + 2.7(j\omega)^2 + 3(j\omega) + 1}{Den_{\left[\tilde{W}(j\omega,\tilde{v})\right]}} \right]. \tag{45}$$

In this analysis, the higher degree terms of the transfer functions in the TS fuzzy dynamic model increase more rapidly than the other ones. Thus,

$$\lim_{\omega \to \infty} \tilde{W}(j\omega, \tilde{v}) = \left| 2\gamma_1 \frac{0.6(j\omega)^4}{0.1(j\omega)^6} + 1.5\gamma_2 \frac{0.1(j\omega)^4}{0.1(j\omega)^6} + \gamma_3 \frac{0.1(j\omega)^4}{0.1(j\omega)^6} \right| \angle \arctan$$

$$\left[2\gamma_1 \frac{0.6(j\omega)^4}{0.1(j\omega)^6} + 1.5\gamma_2 \frac{0.1(j\omega)^4}{0.1(j\omega)^6} + \gamma_3 \frac{0.1(j\omega)^4}{0.1(j\omega)^6} \right]. \tag{46}$$

Hence

$$\lim_{\omega \to \infty} \tilde{W}(j\omega, \tilde{v}) = \left| 2\gamma_1 \frac{0.6}{0.1(j\omega)^2} + 1.5\gamma_2 \frac{0.1}{0.1(j\omega)^2} + \gamma_3 \frac{0.1}{0.1(j\omega)^2} \right| \angle - 180^o. \tag{47}$$

Once again, applying the *Theorem 3.1*, proposed in section 3, the obtained boundary conditions at high frequencies are presented in Tab. 2. The fuzzy frequency response of the SLPV dynamic system, at high frequencies, presents a magnitude range in the interval $\left[\left| \frac{1}{(j\omega)^2} \right|, \left| \frac{12}{(j\omega)^2} \right| \right]$ (dB) and the phase is -180°.

Table 2. Boundary conditions at high frequencies.

Activated Rule	Boundary Condition	Magnitude (dB)	Phase (Degree)
1	$\gamma_1 = 1; \gamma_2 = 0$ and $\gamma_3 = 0$	$\left\| \frac{12}{(j\omega)^2} \right\|$	-180°
2	$\gamma_1 = 0; \gamma_2 = 1$ and $\gamma_3 = 0$	$\left\| \frac{1.5}{(j\omega)^2} \right\|$	-180°
3	$\gamma_1 = 0; \gamma_2 = 0$ and $\gamma_3 = 1$	$\left\| \frac{0.1}{(j\omega)^2} \right\|$	-180°

For comparative analysis, the fuzzy frequency response (boundary conditions at low and high frequencies from Tab. 1-2) and frequency response of the SLPV dynamic system are shown in Fig. 7. For this experiment, the frequency response of the SLPV dynamic system was obtained considering the mean of the stochastic parameter v in the frequency domain as shown in Fig. 8. The proposed structure for determining the frequency response of the SLPV dynamic system is shown in the block diagram (Fig. 9). It can be seen that the fuzzy frequency response is a region in the frequency domain, defined by the consequent linear sub-models $W^i(s)$, starting from the operating region of the antecedent space, as demonstrated by the proposed *Theorem 3.1*. This method highlights the efficiency of the fuzzy frequency response in order to estimate the frequency response of SLPV dynamic systems.
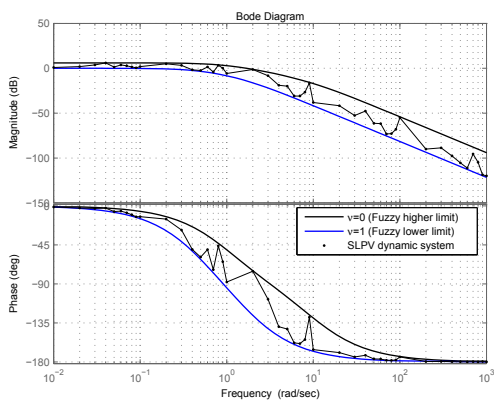
(a) Boundary region.
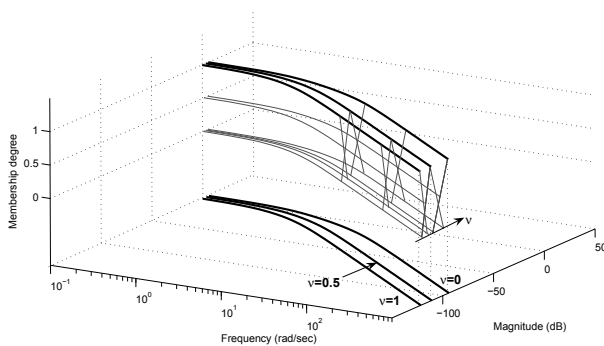
(b) Magnitude.
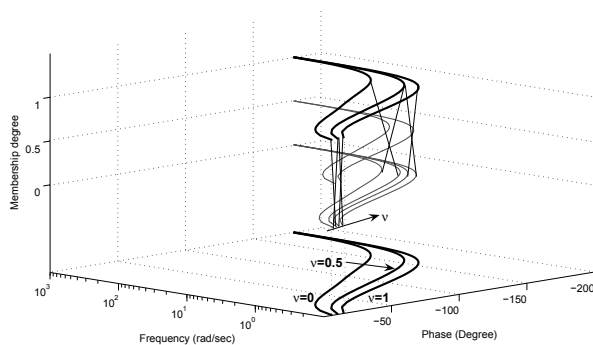
(c) Phase.

Fig. 7. Fuzzy frequency response of the SLPV dynamic system.
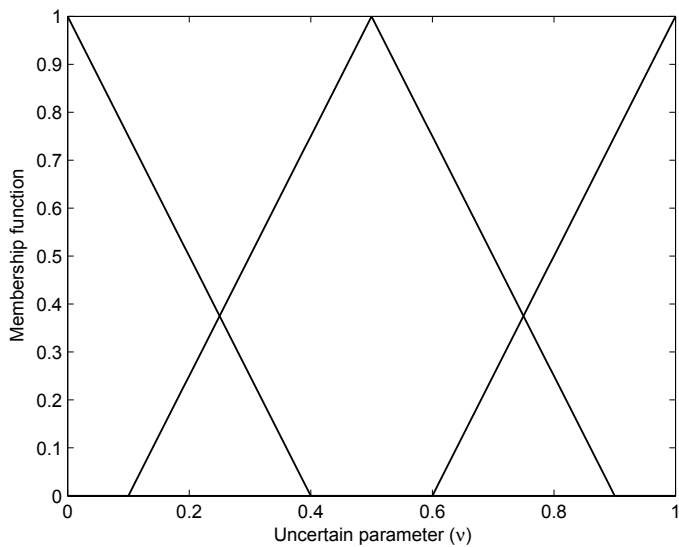
(a) Fuzzy sets of stochastic parameter (v).
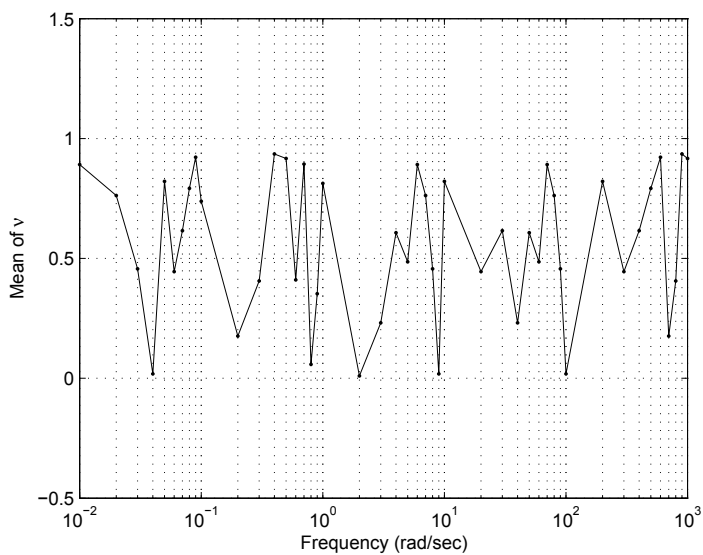
(b) Mean variation of stochastic parameter v in frequency domain.

Fig. 8. Fuzzy and statistic characteristics of the stochastic parameter (v).
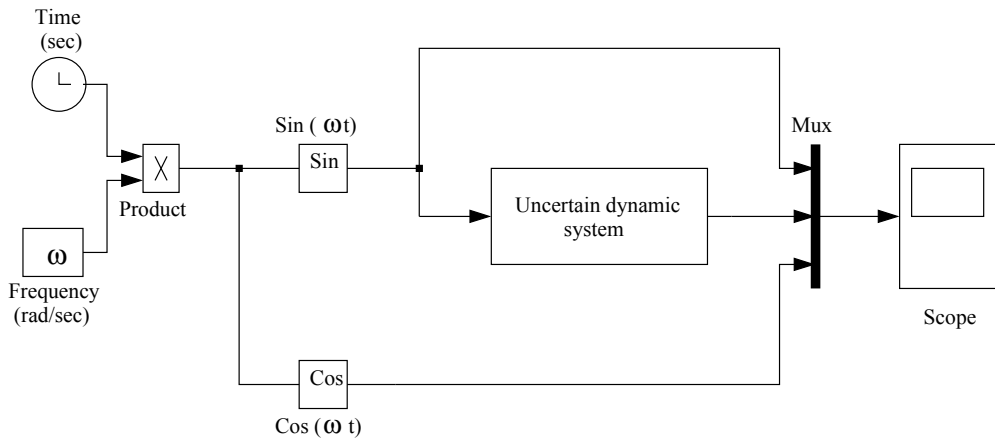
Fig. 9. Simulation diagram in Simulink to obtain Bode diagram of SLPV dynamic system.

6. Final Remarks

The Fuzzy Frequency Response: Definition and Analysis for Stochastic Linear Parameter Varying Dynamic Systems is proposed in this chapter. It was shown that the fuzzy frequency response is a region in the frequency domain, defined by the consequent linear sub-models $W^i(s)$, starting from operating regions of the SLPV dynamic system, according to the proposed *Theorem 3.1*. This formula is very efficient and can be used for robust stability analysis and control design for SLPV dynamic systems.

7. References

Bode, H. W. (1940). Feedback amplifier design. *Bell Systems Technical Journal*, Vol. 19, 42.

Franklin, G. F.; Powell, J. D. & Emami-Naeini, A. (1986). *Feedback Control of Dynamic Systems*, Addison-Wesley, ISBN 0-201-53487-8, USA.

Ioannou, P. A. & Sun, J. (1996). *Robust Adaptive Control*, Prentice Hall, ISBN 0-13-439100-4, New Jersey.

James, H. M.; Nichols, N. B & Phillips, R. S. (1947). *Theory of servomechanisms*, McGraw-Hill, Vol. 25 of MIT Radiation Laboratory Series, New York.

Jr, A. P. Schust (1973). Determination of Aircraft Response Caracteristics in Approach/Landing Configuration for Microwave Landing System Program, Report FT-61R-73. *Naval Air Test Center*, Patuxent River, MD.

Kolev, L. V. (1993). Interval Methodsfor Circuit Analysis. *World Scienhfic*, Singapore.

Lascu, C.; Asiminoaei, L.; Boldea, I. & Blaabjerg, F. (2009). Frequency Response Analysis of Current Controllers for Selective Harmonic Compensation in Active Power Filters. *IEEE Transactions on Industrial Electronics*, Vol. 56, No. 2, Feb 2009, 337-347, ISSN 0278-0046.

Nyquist, H. (1932). Regeneration theory. *Bell Systems Technical Journal*, Vol. 11, 126-147.

Phillips, C. L. & Harbor, R. D. (1996). *Feedback Control Systems*, Prentice Hall, ISBN 0-13-371691-0, New Jersey.

Serra, G. L. O. & Bottura, C. P. (2006). Multiobjective Evolution Based Fuzzy PI Controller Design for Nonlinear Systems. *Engineering Applications of Artificial Intelligence*, Vol. 19, 2, March 2006, 157-167, ISSN 0952-1976.

Serra, G. L. O. & Bottura, C. P. (2007). An IV-QR algorithm for neuro-fuzzy multivariable on-line identification. *IEEE Transactions on Fuzzy Systems*, Vol. 15, 2, April 2007, 200-210, ISSN 1063-6706.

Serra, G. L. O. & Bottura, C. P. (2009). Fuzzy instrumental variable approach for nonlinear discrete-time systems identification in a noisy environment. *Fuzzy Sets and Systems*, Vol. 160, 4, February 2009, 500-520, ISSN 0165-0114.

Serra, G. L. O.; Ferreira, C. C. T. & Silva, J. A. (2009). Development Method for a Robust PID Fuzzy Controller of LPV Systems. *Proceedings of IEEE International Conference on Fuzzy Systems*, pp. 826-830, ISBN 978-1-4244-3596-8, Korea, Outubro 2009, Jeju Island.

Tanaka, K.; Ikeda, T. & Wang, H.O. (1998). Fuzzy regulators and fuzzy observers: relaxed stability conditions and LMI-based designs. *IEEE Transactions on Fuzzy Systems*, Vol. 6, 2, May 1998, 250-265, ISSN 1063-6706.

Takagi, T. & Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 15, 1, 116-132, ISSN 1083-4419.

Delay-dependent exponential stability and filtering for time-delay stochastic systems with nonlinearities

Huaicheng Yan[1], Hao Zhang[2], Hongbo Shi[1] and Max Q.-H. Meng[3]
[1]East China University of Science and Technology,[2]Tongji University, [3]The Chinese University of Hong Kong
PR China

1. Introduction

It is well known that the time-delays are frequently encountered in a variety of dynamic systems such as engineering, biological, and chemical systems, etc., which are very often the main sources of instability and poor performance of systems. Also, in practice, uncertainties are unavoidable since it is very difficult to obtain an exact mathematical model of an object or process due to environmental noise, or slowly varying parameters, etc. Consequently, the problems of robust stability for time-delay systems have been of great importance and have received considerable attention for decades. The developed stability criteria are often classified into two categories according to their dependence on the size of the delays, namely, delay-independent criteria (Park, 2001) and delay-dependent criteria (Wang et al, 1992; Li et al, 1997; Kim, 2001; Moon et al, 2001; Jing et al, 2004; Kwon & Park, 2004; Wu et al, 2004). In general, the latter are less conservative than the former when the size of the time-delay is small. On the other hand, stochastic systems have received much attention since stochastic modelling has come to play an important role in many branches of science and industry. In the past decades, increasing attention has been devoted to the problems of stability of stochastic time-delay systems by a considerable number of researchers (Mao, 1996; Xie & Xie, 2000; Blythe et al, 2001; Xu & Chen, 2002; Lu et al, 2003). Very recently, the problem of exponential stability for delayed stochastic systems with nonlinearities has been extensively investigated by many researchers (Mao, 2002; Yue & Won, 2001; Chen et al, 2005). Motivated by the method for deterministic delayed systems introduced in (Wu et al, 2004), we extend it to uncertain stochastic time-varying delay systems with nonlinearities.

The filter design problem has long been one of the key problems in the areas of control and signal processing. Compared with the Kalman filter, the advantage of H∞ filtering is that the noise sources are arbitrary signals with bounded energy or average power instead of being Gaussian, and no exact statistics are required to be known (Nagpal & Khargonekar, 1991). When parameter uncertainty appears in a system model, the robustness of H∞ filters has to be taken into account. A great number of results on robust H∞ filtering problem have been reported in the literature (Li & Fu, 1997; De Souza et al, 1993), and much attention has been

focused on the robust H∞ filtering problem for time-delay systems (Pila et al, 1999; Wang & Yang, 2002; Xu & Chen, 2004; Gao & Wang, 2003; Fridman et al, 2003; Xu & Van Dooren, 2002; Xu et al, 2003; Zhang et al, 2005; Wang et al, 2006; Wang et al, 2004; Wang et al, 2008; Liu et al, 2008; Zhang & Han, 2008). Depending on whether the existence conditions of filter include the information of delay or not, the existing results on H∞ filtering for time-delay systems can be classified into two types: delay-independent ones (Pila et al, 1999; Wang & Yang, 2002; Xu & Chen, 2004) and delay-dependent ones (Gao & Wang, 2003; Fridman et al, 2003; Xu & Van Dooren, 2002; Xu et al, 2003; Zhang et al, 2005; Wang et al, 2006; Wang et al, 2004; Wang et al, 2008; Liu et al, 2008; Zhang & Han, 2008). On the other hand, since the stochastic systems have gained growing interests recently, H∞ filtering for the time-delay stochastic systems have drawn a lot of attentions from researchers working in related areas (Zhang et al, 2005; Wang et al, 2006; Wang et al, 2008; Liu et al, 2008). It is also known that Markovian jump systems (MJSs) are a set of systems with transitions among the models governed by a Markov chain taking values in a finite set. These systems have the advantages of modeling the dynamic systems subject to abrupt variation in their structures. Therefore, filtering and control for MJSs have drawn much attention recently, see (Xu et al, 2003; Wang et al, 2004). Note that nonlinearities are often introduced in the form of nonlinear disturbances, and exogenous nonlinear disturbances may result from the linearization process of an originally highly nonlinear plant or may be an external nonlinear input, and thus exist in many real-world systems. Therefore, H∞ filtering for nonlinear systems has also been an attractive topic for many years both in the deterministic case (De Souza et al, 1993; Gao & Wang, 2003; Xu & Van Dooren, 2002)) and the stochastic case (Zhang et al, 2005; Wang et al, 2004; Wang et al, 2008; Liu et al, 2008).

Exponential stability is highly desired for filtering processes so that fast convergence and acceptable accuracy in terms of reasonable error covariance can be ensured. A filter is said to be exponential if the dynamics of the estimation error is stochastically exponentially stable. The design of exponential fast filters for linear and nonlinear stochastic systems is also an active research topic; see, e.g. (Wang et al, 2006; Wang et al, 2004). To the best of the authors' knowledge, however, up to now, the problem of delay-range-dependent robust exponential H∞ filtering problem for uncertain *Itô*-type stochastic systems in the simultaneous presence of parameter uncertainties, Markovian switching, nonlinearities, and mode-dependent time-varying delays in a range has not been fully investigated, which still remains open and challenging. This motivates us to investigate the present study.

This chapter is organized as follows. In section 2, the main results are given. Firstly, delay-dependent exponentially mean-square stability for uncertain time-delay stochastic systems with nonlinearities is studied. Secondly, the robust H∞ exponential filtering problem for uncertain stochastic time-delay systems with Markovian switching and nonlinear disturbances is investigated. In section 3, numerical examples and simulations are presented to illustrate the benefits and effectiveness of our proposed theoretical results. Finally, the conclusions are given in section 4.

2. Main results

2.1 Exponential stability of uncertain time-delay nonlinear stochastic systems

Consider the following uncertain stochastic system with time-varying delay and nonlinear stochastic perturbations:

$$\begin{cases} dx(t)=\big[(A+\Delta A(t))x(t)+(B+\Delta B(t))x(t-\tau(t))+f(t,x(t),x(t-\tau(t)))\big]dt+g(t,x(t),x(t-\tau(t)))d\omega(t), \\ x(t)=\phi(t), \quad t\in[-\tau,0], \end{cases} \tag{1}$$

where $x(t)\in\mathbb{R}^n$ is the state vector, A,B,C,D are known real constant matrices with appropriate dimensions, $\omega(t)$ is a scalar Brownian motion defined on a complete probability space (Ω,F,P) with a nature filtration $\{F_t\}_{t\ge0}$. $\phi(t)$ is any given initial data in $L^2_{F_0}\big([-\tau,0];\mathbb{R}^n\big)$. $\tau(t)$ denotes the time-varying delay and is assumed to satisfy either (2a) or (2b):

$$0\le\tau(t)\le\tau, \ \dot{\tau}(t)\le d<1, \tag{2a}$$

$$0\le\tau(t)\le\tau, \tag{2b}$$

where τ and d are constants and the upper bound of $\tau(t)$ and $\dot{\tau}(t)$, respectively. $\Delta A(t)$, $\Delta B(t)$ are all unknown time-varying matrices with appropriate dimensions which represent the system uncertainty and stochastic perturbation uncertainty, respectively. We assume that the uncertainties are norm-bounded and can be described as follows:

$$\begin{bmatrix} \Delta A(t) & \Delta B(t) \end{bmatrix} = EF(t)\begin{bmatrix} G_1 & G_2 \end{bmatrix}, \tag{3}$$

where E,G_1,G_2 are known real constant matrices with appropriate dimensions, $F(t)$ are unknown real matrices with Lebesgue measurable elements bounded by:

$$F^T(t)F(t)\le I. \tag{4}$$

$f(\cdot,\cdot,\cdot):R_+\times R^n\times R^n\to R^n$ and $g(\cdot,\cdot,\cdot):R_+\times R^n\times R^n\to R^{n\times m}$ denote the nonlinear uncertainties which is locally Lipschitz continuous and satisfies the following linear growth conditions

$$\big\|f(t,x(t),x(t-\tau(t)))\big\|\le\big\|F_1x(t)\big\|+\big\|F_2x(t-\tau(t))\big\|, \tag{5}$$

and

$$Trace\big[g^T(t,x(t),x(t-\tau(t)))g(t,x(t),x(t-\tau(t)))\big]\le\big\|H_1x(t)\big\|^2+\big\|H_2x(t-\tau(t))\big\|^2, \tag{6}$$

Throughout this paper, we shall use the following definition for the system (1).
Definition 1 (Chen et al, 2005**).** The uncertain nonlinear stochastic time-delay system (1) is said to be exponentially stable in the mean square sense if there exists a positive scalar $\alpha>0$ such that for all admissible uncertainties

$$\lim_{t\to\infty}\sup\frac{1}{t}\log E\big\|x(t)\big\|^2\le-\alpha. \tag{7}$$

Lemma 1 (Wang et al, 1992**).** For any vectors $x, y \in \mathbb{R}^n$, matrices $A, P \in \mathbb{R}^{n \times n}, D \in \mathbb{R}^{n \times nf}$,

$E \in \mathbb{R}^{nf \times n}$, and $F \in \mathbb{R}^{nf \times nf}$ with $P > 0, F^\mathrm{T} F \leq \mathrm{I}$, and scalar $\varepsilon > 0$, the following inequalities hold:

(i) $2x^\mathrm{T} y \leq x^\mathrm{T} P^{-1} x + y^\mathrm{T} P y$,

(ii) $DFE + E^\mathrm{T} F^\mathrm{T} D^\mathrm{T} \leq \varepsilon^{-1} DD^\mathrm{T} + \varepsilon E^\mathrm{T} E$,

(iii) If $P - \varepsilon DD^\mathrm{T} > 0$, then $(A + DFE)^\mathrm{T} P^{-1} (A + DFE) \leq A^\mathrm{T} (P - \varepsilon DD^\mathrm{T})^{-1} A + \varepsilon E^\mathrm{T} E$.

For convenience, we let

$$y(t) = \left(A + \Delta A(t)\right) x(t) + \left(B + \Delta B(t)\right) x(t - \tau(t)) + f(t, x(t), x(t - \tau(t))), \tag{8}$$

and set

$$f(t) = f(t, x(t), x(t - \tau(t))), \quad g(t) = g(t, x(t), x(t - \tau(t))), \tag{9}$$

then system (1) becomes

$$dx(t) = y(t)dt + g(t)d\omega(t). \tag{10}$$

Then, for any appropriately dimensioned matrices $N_i, M_i, i = 1, 2, 3$, the following equations hold:

$$\Sigma_1 = 2\left[x^\mathrm{T}(t)N_1 + x^\mathrm{T}(t - \tau(t))N_2 + y^\mathrm{T}(t)N_3\right] \times \left[x(t) - x(t - \tau(t)) - \int_{t-\tau(t)}^t y(s)ds - \int_{t-\tau(t)}^t g(s)dw(s)\right] = 0, \tag{11}$$

and

$$\Sigma_2 = 2\left[x^\mathrm{T}(t)M_1 + x^\mathrm{T}(t - \tau(t))M_2 + y^\mathrm{T}(t)M_3\right] \times \left[\left(A + \Delta A(t)\right)x(t) + \left(B + \Delta B(t)\right)x(t - \tau(t)) + f(t) - y(t)\right] = 0, \tag{12}$$

where the free weighting matrices $N_i, M_i, i = 1, 2, 3$ can easily be determined by solving the corresponding LMIs.

On the other hand, for any semi-positive-definite matrix $X = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ * & X_{22} & X_{23} \\ * & * & X_{33} \end{bmatrix} \geq 0$, the following

holds:

$$\Sigma_3 = \tau \xi^\mathrm{T}(t) X \xi(t) - \int_{t-\tau(t)}^t \xi^\mathrm{T}(t) X \xi(t) ds \geq 0, \tag{13}$$

where $\xi^\mathrm{T}(t) = \begin{bmatrix} x^\mathrm{T}(t) & x^\mathrm{T}(t - \tau(t)) & y^\mathrm{T}(t) \end{bmatrix}$.

Theorem 1. When (2a) holds, then for any scalars $\tau > 0, d < 1$, the system (1) is exponentially stable in mean square for all time-varying delays and for all admissible uncertainties, if there exist $P > 0, Q > 0, R > 0, S > 0$, scalars $\rho > 0, \mu > 0, \varepsilon_j > 0, j = 0, 1, ..., 7$, a

Delay-dependent exponential stability
and filtering for time-delay stochastic systems with nonlinearities
239

symmetric semi-positive-definite matrix $X \geq 0$ and any appropriately dimensioned matrices $M_i, N_i, i = 1, 2, 3$, such that the following LMIs hold

$$
\Pi = \begin{bmatrix}
X_{11} & X_{12} & X_{13} & N_1 \\
* & X_{22} & X_{23} & N_2 \\
* & * & X_{33} & N_3 \\
* & * & * & (1-d)Q
\end{bmatrix} \geq 0,
\tag{14}
$$

$$
\Theta = \begin{bmatrix}
\Theta_{11} & \Theta_{12} & \Theta_{13} & N_1 & M_1 & M_1 & M_1E & M_1E & 0 & 0 & 0 & 0 \\
* & \Theta_{22} & \Theta_{23} & N_2 & M_2 & M_2 & 0 & 0 & M_2E & M_2E & 0 & 0 \\
* & * & \Theta_{33} & N_3 & M_3 & M_3 & 0 & 0 & 0 & 0 & M_3E & M_3E \\
* & * & * & -S & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & -\varepsilon_0 I & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & * & -\varepsilon_1 I & 0 & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & * & * & -\varepsilon_2 I & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & * & * & * & -\varepsilon_3 I & 0 & 0 & 0 & 0 \\
* & * & * & * & * & * & * & * & -\varepsilon_4 I & 0 & 0 & 0 \\
* & * & * & * & * & * & * & * & * & -\varepsilon_5 I & 0 & 0 \\
* & * & * & * & * & * & * & * & * & * & -\varepsilon_6 I & 0 \\
* & * & * & * & * & * & * & * & * & * & * & -\varepsilon_7 I
\end{bmatrix} < 0,
\tag{15}
$$

$$
P \leq \rho I,
\tag{16}
$$

$$
S \leq \mu I,
\tag{17}
$$

where

$$\Theta_{11} = R + N_1 + N_1^{\mathrm{T}} + M_1 A + A^{\mathrm{T}} M_1^{\mathrm{T}} + \tau X_{11} + (\varepsilon_2 + \varepsilon_4 + \varepsilon_6) G_1^{\mathrm{T}} G_1 + \varepsilon_0 F_1^{\mathrm{T}} F_1 + (\rho + \frac{\tau \mu}{1-d}) H_1^{\mathrm{T}} H_1,$$

$$\Theta_{12} = -N_1 + N_2^{\mathrm{T}} + M_1 B + A^{\mathrm{T}} M_2^{\mathrm{T}} + \tau X_{12}, \qquad \Theta_{13} = P + N_3^{\mathrm{T}} - M_1 + A^{\mathrm{T}} M_3^{\mathrm{T}} + \tau X_{13},$$

$$\Theta_{22} = -(1-d)R - N_2 - N_2^{\mathrm{T}} + M_2 B + B^{\mathrm{T}} M_2^{\mathrm{T}} + \tau X_{22} + (\varepsilon_3 + \varepsilon_5 + \varepsilon_7) G_2^{\mathrm{T}} G_2 + \varepsilon_1 F_2^{\mathrm{T}} F_2 + (\rho + \frac{\tau \mu}{1-d}) H_2^{\mathrm{T}} H_2,$$

$$\Theta_{23} = -N_3^{\mathrm{T}} - M_2 + B^{\mathrm{T}} M_3^{\mathrm{T}} + \tau X_{23}, \qquad \Theta_{33} = \tau Q - M_3 - M_3^{\mathrm{T}} + \tau X_{33}.$$

Proof. Construct the Lyapunov-Krasovskii functional candidate for system (1) as follows:

$$V(t) = \sum_{i=1}^{4} V_i(t),$$

where

$$V_1(t) = x^{\mathrm{T}}(t) P x(t), \quad V_2(t) = \int_{-\tau(t)}^{0} \int_{t+\theta}^{t} y^{\mathrm{T}}(s) Q y(s) ds d\theta, \quad V_3(t) = \int_{t-\tau(t)}^{t} x^{\mathrm{T}}(s) R x(s) ds,$$

$$V_4(t) = \frac{1}{1-d} \int_{-\tau(t)}^{0} \int_{t+\beta}^{t} trace\left[g^{\mathrm{T}}(s) S g(s) \right] ds d\beta.
\tag{18}$$

Defining x_t by $x_t(s) = x(t+s), -2\tau \leq s \leq 0$, the weak infinitesimal operator L of the stochastic process $\{x_t, t \geq 0\}$ along the evolution of $V_1(t)$ is given by (Blythe et al, 2001):

$$LV_1(t) = 2x^{\mathrm{T}}(t)Py(t) + trace\left[g^{\mathrm{T}}(t)Pg(t) \right]. \tag{19}$$

The weak infinitesimal operator L for the evolution of $V_2(t), V_3(t), V_4(t)$ can be computed directly as follows

$$LV_2(t) = \tau(t)y^{\mathrm{T}}(t)Qy(t) - (1 - \dot{\tau}(t))\int_{t-\tau(t)}^{t} y^{\mathrm{T}}(s)Qy(s)ds, \tag{20}$$

$$LV_3(t) = x^{\mathrm{T}}(t)Rx(t) - (1 - \dot{\tau}(t))x^{\mathrm{T}}(t - \tau(t))Rx(t - \tau(t)), \tag{21}$$

$$LV_4(t) = \frac{1}{1-d}\tau(t)trace\left[g^{\mathrm{T}}(t)Sg(t) \right] - \frac{1}{1-d}(1 - \dot{\tau}(t))\int_{t-\tau(t)}^{t} trace\left[g^{\mathrm{T}}(s)Sg(s) \right]ds. \tag{22}$$

Therefore, using (2a) and adding Eqs. (11)-(13) to Eqs. (19)-(22), then the weak infinitesimal operator of $V(t)$ along the trajectory of system (1) yields

$$LV(t) \leq 2x^{\mathrm{T}}(t)Py(t) + trace\left[g^{\mathrm{T}}(t)Pg(t) \right] + \tau y^{\mathrm{T}}(t)Qy(t) - (1-d)\int_{t-\tau(t)}^{t} y^{\mathrm{T}}(s)Qy(s)ds + x^{\mathrm{T}}(t)Rx(t)$$
$$- (1-d)x^{\mathrm{T}}(t - \tau(t))Rx(t - \tau(t)) + \frac{\tau}{1-d}trace\left[g^{\mathrm{T}}(t)Sg(t) \right] - \int_{t-\tau(t)}^{t} trace\left[g^{\mathrm{T}}(s)Sg(s) \right]ds + \Sigma_1 + \Sigma_2 + \Sigma_3 \tag{23}$$

It follows from (i) of Lemma 1 that

$$-2\left[x^{\mathrm{T}}(t)N_1 + x^{\mathrm{T}}(t - \tau(t))N_2 + y^{\mathrm{T}}(t)N_3 \right]\int_{t-\tau(t)}^{t} g(s)d\omega(s)$$
$$\leq \xi^{\mathrm{T}}(t)NS^{-1}N^{\mathrm{T}}\xi(t) + \left(\int_{t-\tau(t)}^{t} g(s)d\omega(s) \right)^{\mathrm{T}} S\left(\int_{t-\tau(t)}^{t} g(s)d\omega(s) \right), \tag{24}$$

where $N^{\mathrm{T}} = \begin{bmatrix} N_1^{\mathrm{T}} & N_2^{\mathrm{T}} & N_3^{\mathrm{T}} \end{bmatrix}$.

Moreover, from Lemma 1 and (5)

$$2\left[x^{\mathrm{T}}(t)M_1 + x^{\mathrm{T}}(t - \tau(t))M_2 + y^{\mathrm{T}}(t)M_3 \right]f(t)$$
$$\leq \xi^{\mathrm{T}}(t)(\varepsilon_0^{-1} + \varepsilon_1^{-1})MM^{\mathrm{T}}\xi(t) + x^{\mathrm{T}}(t)\varepsilon_0 F_1 F_1^{\mathrm{T}}x(t) + x^{\mathrm{T}}(t - \tau(t))\varepsilon_1 F_2 F_2^{\mathrm{T}}x(t - \tau(t)), \tag{25}$$

where $M^{\mathrm{T}} = \begin{bmatrix} M_1^{\mathrm{T}} & M_2^{\mathrm{T}} & M_3^{\mathrm{T}} \end{bmatrix}$.

Taking note of (6) together with (16) and (17) imply

$$trace[g^T(t)Pg(t)]+\frac{\tau}{1-d}trace[g^T(t)Sg(t)]\le(\rho+\frac{\tau\mu}{1-d})[x^T(t)H_1H_1^Tx(t)+x^T(t-\tau(t))H_2H_2^Tx(t-\tau(t))].$$

(26)

Noting that

$$E\left(\int_{t-\tau(t)}^{t}g(s)d\omega(s)\right)^T S\left(\int_{t-\tau(t)}^{t}g(s)d\omega(s)\right)=E\int_{t-\tau(t)}^{t}trace\left[g^T(s)Sg(s)\right]ds. \quad (27)$$

For the positive scalars $\varepsilon_k>0, k=2,3,\dots,7$, it follows from (3), (4) and Lemma 1 that

$$2x^T(t)M_1\Delta A(t)x(t)\le\varepsilon_2^{-1}x^T(t)M_1EE^TM_1^Tx(t)+\varepsilon_2x^T(t)G_1^TG_1x(t), \quad (28)$$

$$2x^T(t)M_1\Delta B(t)x(t-\tau(t))\le\varepsilon_3^{-1}x^T(t)M_1EE^TM_1^Tx(t)+\varepsilon_3x^T(t-\tau(t))G_2^TG_2x(t-\tau(t)), \quad (29)$$

$$2x^T(t-\tau(t))M_2\Delta A(t)x(t)\le\varepsilon_4^{-1}x^T(t-\tau(t))M_2EE^TM_2^Tx(t-\tau(t))+\varepsilon_4x^T(t)G_1^TG_1x(t), \quad (30)$$

$$2x^T(t-\tau(t))M_2\Delta B(t)x(t-\tau(t))\le\varepsilon_5^{-1}x^T(t-\tau(t))M_2EE^TM_2^Tx(t-\tau(t))+\varepsilon_5x^T(t-\tau(t))G_2^TG_2x(t-\tau(t)), \quad (31)$$

$$2y^T(t)M_3\Delta A(t)x(t)\le\varepsilon_6^{-1}y^T(t)M_3EE^TM_3^Ty(t)+\varepsilon_6x^T(t)G_1^TG_1x(t), \quad (32)$$

$$2y^T(t)M_3\Delta B(t)x(t-\tau(t))\le\varepsilon_7^{-1}y^T(t)M_3EE^TM_3^Ty(t)+\varepsilon_7x^T(t-\tau(t))G_2^TG_2x(t-\tau(t)). \quad (33)$$

Then, taking the mathematical expectation of both sides of (23) and combining (24)-(27) with (28)-(33), it can be concluded that

$$E\{LV(t)\}\le E\{\xi^T(t)\Xi\xi(t)\}-\int_{t-\tau(t)}^{t}E\{\xi^T(t,s)\Pi\xi(t,s)\}ds, \quad (34)$$

where

$$\xi^T(t,s)=\begin{bmatrix}x^T(t) & x^T(t-\tau(t)) & y^T(t) & y^T(s)\end{bmatrix}, \qquad \Xi=\begin{bmatrix}\Xi_{11} & \Xi_{12} & \Xi_{13}\\ * & \Xi_{22} & \Xi_{23}\\ * & * & \Xi_{33}\end{bmatrix},$$

$$\Xi_{11}=\Theta_{11}+N_1S^{-1}N_1^T+(\varepsilon_0^{-1}+\varepsilon_1^{-1})M_1M_1^T+(\varepsilon_2^{-1}+\varepsilon_3^{-1})M_1EE^TM_1^T,$$

$$\Xi_{12}=\Theta_{12}+N_1S^{-1}N_2^T+(\varepsilon_0^{-1}+\varepsilon_1^{-1})M_1M_2^T, \quad \Xi_{13}=\Theta_{13}+N_1S^{-1}N_3^T+(\varepsilon_0^{-1}+\varepsilon_1^{-1})M_1M_3^T,$$

$$\Xi_{22}=\Theta_{22}+N_2S^{-1}N_2^T+(\varepsilon_0^{-1}+\varepsilon_1^{-1})M_2M_2^T+(\varepsilon_4^{-1}+\varepsilon_5^{-1})M_2EE^TM_2^T,$$

$$\Xi_{23}=\Theta_{23}+N_2S^{-1}N_3^T+(\varepsilon_0^{-1}+\varepsilon_1^{-1})M_2M_3^T,$$

$$\Xi_{33}=\Theta_{33}+N_3S^{-1}N_3^T+(\varepsilon_0^{-1}+\varepsilon_1^{-1})M_3M_3^T+(\varepsilon_6^{-1}+\varepsilon_7^{-1})M_3EE^TM_3^T.$$

By applying the Schur complement techniques, $\Xi < 0$ is equivalent to LMI (15). Therefore, if LMIs (14) and (15) are satisfied, one can show that (34) implies

$$E\{LV(t)\} \leq E\{\xi^{\mathrm{T}}(t)\Xi\xi(t)\}. \tag{35}$$

Now we proceed to prove system (1) is exponential stable in mean square, using the similar method of (Chen et al, 2005). Set $\lambda_0 = \lambda_{\min}(-\Xi), \lambda_1 = \lambda_{\min}(P)$, by (35),

$$E\{LV(t)\} \leq -\lambda_0 E\{\xi^{\mathrm{T}}(t)\xi(t)\} \leq -\lambda_0 E\{x^{\mathrm{T}}(t)x(t)\}. \tag{36}$$

From the definitions of $V(t)$ and $y(t)$, there exist positive scalars β_1, β_2 such that

$$\lambda_1 \|x(t)\|^2 \leq V(t) \leq \beta_1 \|x(t)\|^2 + \beta_2 \int_{t-2\tau}^{t} \|x(s)\|^2 ds. \tag{37}$$

Defining a new function as $W(t) = e^{\beta_0 t}V(t),$ its weak infinitesimal operator is given by

$$L\{W(t)\} = \beta_0 e^{\beta_0 t}V(t) + e^{\beta_0 t}L\{V(t)\}, \tag{38}$$

Then, from (36)-(38), by using the generalized $It\hat{o}$ formula, we can obtain that

$$E\{W(t)\} - E\{W(t_0)\} \leq E\int_{t_0}^{t} e^{\beta_0 s}\left[\beta_0\left(\beta_1 \|x(s)\|^2 + \beta_2 \int_{s-2\tau}^{s}\|x(\theta)\|^2 d\theta\right) - \lambda_0 \|x(s)\|^2\right]ds. \tag{39}$$

Since the following inequality holds (Chen et al, 2005)

$$\int_{t_0}^{t} e^{\beta_0 s} ds \int_{s-2\tau}^{s}\|x(\theta)\|^2 d\theta \leq 2\tau e^{2\beta_0 \tau}\int_{t_0-2\tau}^{t}\|x(s)\|^2 e^{\beta_0 s} ds. \tag{40}$$

Therefore, it follows that from (39) and (40),

$$E\{W(t)\} - E\{W(t_0)\} \leq E\int_{t_0}^{t} e^{\beta_0 s}\left[\beta_0(\beta_1 + 2\tau\beta_2 e^{2\beta_0 \tau}) - \lambda_0\right]\|x(s)\|^2 ds + C_0(t_0), \tag{41}$$

where $C_0(t_0) = 2\tau\beta_0\beta_2 e^{2\beta_0 \tau}\int_{t_0-2\tau}^{t_0} E\|x(s)\|^2 e^{\beta_0 s} ds.$

Choose a positive scalar $\beta_0 > 0$ such that (Chen et al, 2005)

$$\beta_0(\beta_1 + 2\tau\beta_2 e^{2\beta_0 \tau}) \leq \lambda_0. \tag{42}$$

Then, by (41) and (42), it is easily obtain

$$\lim_{t\to\infty} \sup \frac{1}{t} \log \mathrm{E}\|x(t)\|^2 \le -\beta_0,$$

which implies that system (1) is exponentially stable in mean square by Definition 1. This completes the proof. $\square$

In the case of the conditon (2b) for system (1), which is derivative-independent, or in the case of $\tau(t)$ is not differentiable. According to the proof of Theorem 1, the following theorem is followed:

Theorem 2. When (2b) holds, then for any scalars $\tau>0$, the stochastic system (1) is exponentially mean-square stable for all admissible uncertainties, if there exist $P>0, Q>0, S>0$, scalars $\rho>0, \mu>0, \varepsilon_j>0, j=0,1,...,7$, matrix $X\ge0$ and any appropriately dimensioned matrices $M_i, N_i, i=1,2,3$, such that (16),(17) and the following LMI holds

$$\tilde{\Pi} = \begin{bmatrix} X_{11} & X_{12} & X_{13} & N_1 \\ * & X_{22} & X_{23} & N_2 \\ * & * & X_{33} & N_3 \\ * & * & * & Q \end{bmatrix} \ge 0, \tag{43}$$

$$\tilde{\Theta} = \begin{bmatrix}
\tilde{\Theta}_{11} & \Theta_{12} & \Theta_{13} & N_1 & M_1 & M_1 & M_1E & M_1E & 0 & 0 & 0 & 0 \\
* & \tilde{\Theta}_{22} & \Theta_{23} & N_2 & M_2 & M_2 & 0 & 0 & M_2E & M_2E & 0 & 0 \\
* & * & \Theta_{33} & N_3 & M_3 & M_3 & 0 & 0 & 0 & 0 & M_3E & M_3E \\
* & * & * & -S & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & -\varepsilon_0 I & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & * & -\varepsilon_1 I & 0 & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & * & * & -\varepsilon_2 I & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & * & * & * & -\varepsilon_3 I & 0 & 0 & 0 & 0 \\
* & * & * & * & * & * & * & * & -\varepsilon_4 I & 0 & 0 & 0 \\
* & * & * & * & * & * & * & * & * & -\varepsilon_5 I & 0 & 0 \\
* & * & * & * & * & * & * & * & * & * & -\varepsilon_6 I & 0 \\
* & * & * & * & * & * & * & * & * & * & * & -\varepsilon_7 I
\end{bmatrix} < 0, \tag{44}$$

Where

$$\tilde{\Theta}_{11} = N_1 + N_1^T + M_1 A + A^T M_1^T + \tau X_{11} + (\varepsilon_2 + \varepsilon_4 + \varepsilon_6) G_1^T G_1 + \varepsilon_0 F_1^T F_1 + (\rho + \tau\mu) H_1^T H_1,$$
$$\tilde{\Theta}_{22} = -N_2 - N_2^T + M_2 B + B^T M_2^T + \tau X_{22} + (\varepsilon_3 + \varepsilon_5 + \varepsilon_7) G_2^T G_2 + \varepsilon_1 F_2^T F_2 + (\rho + \tau\mu) H_2^T H_2.$$

Remark 1. Theorem 1 and 2 provides delay-dependent exponentially stable criteria in mean square for stochastic system (1) in terms of the solvability of LMIs. By using them, one can obtain the MADB τ by solving the following optimization problems:

$$\begin{cases} \max \ \tau \\ s.t. \quad X \geq 0, P > 0, Q > 0, R > 0, Z > 0, \rho > 0, \mu > 0, \varepsilon_j > 0, M_i, N_i, (14)-(17), \ i=1,2,3; \ j=0,1,...,7, \end{cases} \quad (45)$$

or

$$\begin{cases} \max \ \tau \\ s.t. \quad X \geq 0, P > 0, Q > 0, Z > 0, \rho > 0, \mu > 0, \ \varepsilon_j > 0, \ M_i, N_i, (16),(17),(43),(44), \ i=1,2,3; \ j=0,1,...,7. \end{cases} \quad (46)$$

2.2 H∞ exponential filtering for uncertain Markovian switching time-delay stochastic systems with nonlinearities

We consider the following uncertain nonlinear stochastic systems with Markovian jump parameters and mode-dependent time delays

$$(\Sigma): dx(t) = [A(t,r_t)x(t) + A_d(t,r_t)x(t-\tau_{r_t}(t)) + D_1(r_t)f(x(t), x(t-\tau_{r_t}(t)), r_t) + B_1(t,r_t)v(t)]dt$$
$$+ [E(t,r_t)x(t) + E_d(t,r_t)x(t-\tau_{r_t}(t)) + G(t,r_t)v(t)]d\omega(t), \quad (47)$$

$$y(t) = C(t,r_t)x(t) + C_d(t,r_t)x(t-\tau_{r_t}(t)) + D_2(r_t)g(x(t), x(t-\tau_{r_t}(t)), r_t) + B_2(t,r_t)v(t), \quad (48)$$

$$z(t) = L(r_t)x(t), \quad (49)$$

$$x(t) = \phi(t), \quad r(t) = r(0), \ \forall t \in [-\tau_2, 0], \quad (50)$$

where $x(t) \in \mathbb{R}^n$ is the state vector; $v(t) \in \mathbb{R}^p$ is the exogenous disturbance input which belongs to $L_2[0,\infty)$; $y(t) \in \mathbb{R}^q$ is the measurement; $z(t) \in \mathbb{R}^m$ is the signal to be estimated; $\omega(t)$ is a zero-mean one-dimensional Wiener process (Brownian Motion) satisfying $\mathrm{E}[\omega(t)] = 0$ and $\mathrm{E}[\omega^2(t)] = t$; $\{r_t, t \geq 0\}$ is a continuous-time Markovian process with right continuous trajectories and taking values in a finite set $S = \{1,2,\ldots,N\}$ with transition probability matrix $\Pi \triangleq \{\pi_{ij}\}$ given by

$$\Pr\{r_{t+\Delta} = j \mid r_t = i\} = \begin{cases} \pi_{ij}\Delta + o(\Delta), & if \ i \neq j, \\ 1 + \pi_{ii}\Delta + o(\Delta), & if \ i = j, \end{cases} \quad (51)$$

where $\Delta > 0$, $\lim_{\Delta \to 0}(o(\Delta)/\Delta) = 0$; $\pi_{ij} \geq 0$ for $i \neq j$, is the transition rate from mode i at time t to mode j at time $t + \Delta$ and

$$\pi_{ii} = -\sum_{j=1, j\neq i}^{N} \pi_{ij}. \quad (52)$$

In system (Σ), $\tau_{r_t}(t)$ denotes the time-varying delay when the mode is in r_t and satisfies

$$0 \le \tau_{1i} \le \tau_i(t) \le \tau_{2i}, \ \dot{\tau}_i(t) \le d_i < 1, \ \forall r_t = i, i \in S \tag{53}$$

where τ_{1i}, τ_{2i} and d_i are known real constants scalars for any $i \in S$. In (50), $\tau_2 = \max\{\tau_{2i}, i \in S\}$, and $\phi(t)$ is a vector-valued initial continuous function defined on $[-\tau_2, 0]$. $A(t, r_t), A_d(t, r_t), D_1(r_t), B_1(t, r_t), E(t, r_t), E_d(t, r_t), G(t, r_t), C(t, r_t), C_d(t, r_t), D_2(r_t), B_2(t, r_t)$ and $L(r_t)$ are matrix functions governed by Markov process r_t, and

$$
\begin{aligned}
&A(t, r_t) = A(r_t) + \Delta A(t, r_t), \ A_d(t, r_t) = A_d(r_t) + \Delta A_d(t, r_t), \ B_1(t, r_t) = B_1(r_t) + \Delta B_1(t, r_t), \\
&E(t, r_t) = E(r_t) + \Delta E(t, r_t), \ E_d(t, r_t) = E_d(r_t) + \Delta E_d(t, r_t), \ G(t, r_t) = G(r_t) + \Delta G(t, r_t), \\
&C(t, r_t) = C(r_t) + \Delta C(t, r_t), \ C_d(t, r_t) = C_d(r_t) + \Delta C_d(t, r_t), \ B_2(t, r_t) = B_2(r_t) + \Delta B_2(t, r_t).
\end{aligned}
$$

where $A(r_t), A_d(r_t), B_1(r_t), E(r_t), E_d(r_t), G(r_t), C(r_t), C_d(r_t), B_2(r_t)$ and $L(r_t)$ are known real matrices representing the nominal system for all $r_t \in S$, and $\Delta A(t, r_t), \Delta A_d(t, r_t)$, $\Delta E(t, r_t), \Delta E_d(t, r_t), \Delta G(t, r_t), \Delta C(t, r_t), \Delta C_d(t, r_t)$ and $\Delta B_2(t, r_t)$ are unknown matrices representing parameter uncertainties, which are assumed to be of the following form

$$
\begin{bmatrix}
\Delta A(t, r_t) & \Delta A_d(t, r_t) & \Delta B_1(t, r_t) \\
\Delta E(t, r_t) & \Delta E_d(t, r_t) & \Delta G(t, r_t) \\
\Delta C(t, r_t) & \Delta C_d(t, r_t) & \Delta B_2(t, r_t)
\end{bmatrix}
=
\begin{bmatrix}
M_1(r_t) \\
M_2(r_t) \\
M_3(r_t)
\end{bmatrix}
F(t, r_t) \begin{bmatrix} N_1(r_t) & N_2(r_t) & N_3(r_t) \end{bmatrix}, \forall r_t \in S, \tag{54}
$$

where $M_1(r_t), M_2(r_t), N_1(r_t), N_2(r_t)$ and $N_3(r_t)$ are known real constant matrices for all $r_t \in S$, and $F(t, r_t)$ is time-varying matrices with Lebesgue measurable elements satisfying

$$F^{\mathrm{T}}(t, r_t) F(t, r_t) \le I, \quad \forall r_t \in S. \tag{55}$$

Assumption 1: For a fixed system mode $r_t \in S$, there exist known real constant mode-dependent matrices $F_1(r_t) \in \mathbb{R}^{n \times n}$, $F_2(r_t) \in \mathbb{R}^{n \times n}$, $H_1(r_t) \in \mathbb{R}^{n \times n}$ and $H_2(r_t) \in \mathbb{R}^{n \times n}$ such that the unknown nonlinear vector functions $f(\cdot, \cdot, \cdot)$ and $g(\cdot, \cdot, \cdot)$ satisfy the following boundedness conditions:

$$\left| f(x(t), x(t - \tau_{r_t}(t)), r_t) \right| \le \left| F_1(r_t) x(t) \right| + \left| F_2(r_t) x(t - \tau_{r_t}(t)) \right|, \tag{56}$$

$$\left| g(x(t), x(t - \tau_{r_t}(t)), r_t) \right| \le \left| H_1(r_t) x(t) \right| + \left| H_2(r_t) x(t - \tau_{r_t}(t)) \right|. \tag{57}$$

For the sake of notation simplification, in the sequel, for each possible $r_t = i$, $i \in S$, a matrix $M(t, r_t)$ will be denoted by $M_i(t)$; for example, $A(t, r_t)$ is denoted by $A_i(t)$, and $B(t, r_t)$ by B_i, and so on.

For each $i \in S$, we are interested in designing an exponential mean-square stable, Markovian jump, full-order linear filter described by

$$(\Sigma_f): \quad d\hat{x}(t) = A_{fi}\hat{x}(t)dt + B_{fi}y(t)dt, \tag{58}$$

$$\hat{z}(t) = L_{fi}\hat{x}(t), \tag{59}$$

where $\hat{x}(t) \in \mathbb{R}^n$ and $\hat{z}(t) \in \mathbb{R}^q$ for $i \in S$, and the constant matrices A_{fi}, B_{fi} and L_{fi} are filter parameters to be determined.

Denote

$$\tilde{x}(t) = x(t) - \hat{x}(t), \quad \tilde{z}(t) = z(t) - \hat{z}(t), \quad \xi(t) = [x(t) \quad \tilde{x}(t)]^{\mathrm{T}}, \tag{60}$$

Then, for each $r_t = i$, $i \in S$, the filtering error dynamics from the systems (Σ) and (Σ_f) can be described by

$$(\tilde{\Sigma}): d\xi(t) = [\tilde{A}_i(t)\xi(t) + \tilde{A}_{di}(t)H\xi(t - \tau_i(t)) + \tilde{D}_{1i}f(H\xi(t), H\xi(t - \tau_i(t)), i)$$
$$- \tilde{D}_{2i}g(H\xi(t), H\xi(t - \tau_i(t)), i) + \tilde{B}_{1i}(t)v(t)]dt \tag{61}$$
$$+ [\tilde{E}_i(t)H\xi(t) + \tilde{E}_{di}(t)H\xi(t - \tau_i(t)) + \tilde{G}_i(t)v(t)]d\omega(t),$$

$$\tilde{z}(t) = \tilde{L}_i\xi(t), \tag{62}$$

where

$$\tilde{A}_i(t) = \tilde{A}_i + \Delta\tilde{A}_i(t), \qquad \tilde{A}_{di}(t) = \tilde{A}_{di} + \Delta\tilde{A}_{di}(t), \qquad \tilde{B}_i(t) = \tilde{B}_i + \Delta\tilde{B}_i(t),$$

$$\tilde{E}_i(t) = \tilde{E}_i + \Delta\tilde{E}_i(t), \qquad \tilde{E}_{di}(t) = \tilde{E}_{di} + \Delta\tilde{E}_{di}(t), \qquad \tilde{G}_i(t) = \tilde{G}_i + \Delta\tilde{G}_i(t),$$

$$\tilde{A}_i = \begin{bmatrix} A_i & 0 \\ A_i - A_{fi} - B_{fi}C_i & A_{fi} \end{bmatrix}, \quad \Delta\tilde{A}_i(t) = \begin{bmatrix} \Delta A_i(t) & 0 \\ \Delta A_i(t) - B_{fi}\Delta C_i(t) & 0 \end{bmatrix}, \quad \tilde{A}_{di} = \begin{bmatrix} A_{di} \\ A_{di} - B_{fi}C_{di} \end{bmatrix},$$

$$\Delta\tilde{A}_{di}(t) = \begin{bmatrix} \Delta A_{di}(t) \\ \Delta A_{di}(t) - B_{fi}\Delta C_{di}(t) \end{bmatrix}, \quad \tilde{B}_i = \begin{bmatrix} B_{1i} \\ B_{1i} - B_{fi}B_{2i} \end{bmatrix}, \quad \Delta\tilde{B}_i(t) = \begin{bmatrix} \Delta B_{1i}(t) \\ \Delta B_{1i}(t) - B_{fi}\Delta B_{2i}(t) \end{bmatrix},$$

$$\tilde{E}_i = \begin{bmatrix} E_i \\ E_i \end{bmatrix}, \quad \Delta\tilde{E}_i(t) = \begin{bmatrix} \Delta E_i(t) \\ \Delta E_i(t) \end{bmatrix}, \quad \tilde{E}_{di} = \begin{bmatrix} E_{di} \\ E_{di} \end{bmatrix}, \quad \Delta\tilde{E}_{di}(t) = \begin{bmatrix} \Delta E_{di}(t) \\ \Delta E_{di}(t) \end{bmatrix}, \quad \tilde{G}_i = \begin{bmatrix} G_i \\ G_i \end{bmatrix},$$

$$\Delta\tilde{G}_i(t) = \begin{bmatrix} \Delta G_i(t) \\ \Delta G_i(t) \end{bmatrix}, \quad \tilde{D}_{1i} = \begin{bmatrix} D_{1i} \\ D_{1i} \end{bmatrix}, \quad \tilde{D}_{2i} = \begin{bmatrix} 0 \\ B_{fi}D_{2i} \end{bmatrix}, \quad \tilde{L}_i = \begin{bmatrix} L_i - L_{fi} & L_{fi} \end{bmatrix}, \quad H = \begin{bmatrix} I & 0 \end{bmatrix}.$$

Observe the filtering error system (61)-(62) and let $\xi(t;\varsigma)$ denote the state trajectory from the initial data $\xi(\theta) = \varsigma(\theta)$ on $-\tau_2 \leq \theta \leq 0$ in $L^2_{F_0}\left([-\tau_2, 0]; \mathbb{R}^{2n}\right)$. Obviously, the system (61)-(62) admits a trivial solution $\xi(t;0) = 0$ corresponding to the initial data $\varsigma = 0$. Throughout this paper, we adopt the following definition.

Definition 2 (Wang et al, 2004**):** For every $\varsigma \in L^2_{F_0}\left([-\tau_2, 0]; \mathbb{R}^{2n}\right)$, the filtering error system (61)-(62) is said to be robustly exponentially mean-square stable if, when $v(t) = 0$, for every system mode, there exist constant scalars $\alpha > 0$ and $\beta > 0$, such that

$$\mathrm{E}\left|\xi(\mathrm{t};\varsigma)\right|^2 \leq \alpha e^{-\beta t} \sup_{-\tau_2 \leq \theta \leq 0} \mathrm{E}\left|\varsigma(\theta)\right|^2. \tag{63}$$

We are now in a position to formulate the robust H∞ filter design problem to be addressed in this paper as follows: given the system (Σ) and a prescribed $\gamma > 0$, determine an filter (Σ_f) such that, for all admissible uncertainties, nonlinearities as well as delays, the filtering error system $(\tilde{\Sigma})$ is robustly exponentially mean-square stable and

$$\left\|\tilde{z}(t)\right\|_{\mathrm{E}_2} \leq \gamma \left\|v(t)\right\|_2 \tag{64}$$

under zero-initial conditions for any nonzero $v(t) \in L_2[0,\infty)$, where $\left\|\tilde{z}(t)\right\|_{\mathrm{E}_2} = \mathrm{E}\left\{\int_0^\infty \left|\tilde{z}(t)\right|^2 dt\right\}^{1/2}$. The following lemmas will be employed in the proof of our main results.

Lemma 2 (Xie, L., 1996). Let $x \in \mathbb{R}^n, y \in \mathbb{R}^n$ and a scalar $\varepsilon > 0$. Then we have $x^{\mathrm{T}} y + y^{\mathrm{T}} x \leq \varepsilon x^{\mathrm{T}} x + \varepsilon^{-1} y^{\mathrm{T}} y$.

Lemma 3 (Xie, L., 1996). Given matrices $Q = Q^{\mathrm{T}}, H, E$ and $R = R^{\mathrm{T}} > 0$ of appropriate dimensions, $Q + HFE + E^{\mathrm{T}} F^{\mathrm{T}} H^{\mathrm{T}} < 0$ for all F satisfying $F^{\mathrm{T}} F \leq R$, if and only if there exists some $\lambda > 0$ such that $Q + \lambda H H^{\mathrm{T}} + \lambda^{-1} E^{\mathrm{T}} R E < 0$.

To this end, we provide the following theorem to establish a delay-dependent criterion of robust exponential mean-square stability with H∞ performance of system $(\tilde{\Sigma})$, which will be fundamental in the design of the expected H∞ filter.

Theorem 3. Given scalars $\tau_{1i}, \tau_{2i}, d_i$ and $\gamma > 0$, for any delays $\tau_i(t)$ satisfying (7), the filtering error system $(\tilde{\Sigma})$ is robustly exponentially mean-square stable and (64) is satisfied under zero-initial conditions for any nonzero $v(t) \in L_2[0,\infty)$ and all admissible uncertainties if there exist matrices $P_i > 0, i = 1, 2, \ldots, N, Q > 0$ and sclars $\varepsilon_{1i} > 0, \varepsilon_{2i} > 0$ such that the following LMI holds for each $i \in S$

$$\Phi_i = \begin{bmatrix} \Phi_{11} & P_i\tilde{A}_{di}(t) & P_i\tilde{B}_i(t) & H^{\mathrm{T}}\tilde{E}_i^{\mathrm{T}}(t)P_i & P_i\tilde{D}_{1i} & P_i\tilde{D}_{2i} \\ * & \Phi_{22} & 0 & \tilde{E}_{di}^{\mathrm{T}}(t)P_i & 0 & 0 \\ * & * & -\gamma^2 I & \tilde{G}_i^{\mathrm{T}}(t)P_i & 0 & 0 \\ * & * & * & -P_i & 0 & 0 \\ * & * & * & * & -\varepsilon_{1i}I & 0 \\ * & * & * & * & * & -\varepsilon_{2i}I \end{bmatrix} < 0, \tag{65}$$

where

$$\Phi_{11} = \sum_{j=1}^{N}\pi_{ij}P_j + P_i\tilde{A}_i(t) + \tilde{A}_i^{\mathrm{T}}(t)P_i + \mu H^{\mathrm{T}}QH + 2\varepsilon_{1i}H^{\mathrm{T}}F_{1i}^{\mathrm{T}}F_{1i}H + 2\varepsilon_{2i}H^{\mathrm{T}}H_{1i}^{\mathrm{T}}H_{1i}H + \tilde{L}_i^{\mathrm{T}}\tilde{L}_i,$$

$$\Phi_{22} = 2\varepsilon_{1i}F_{2i}^{\mathrm{T}}F_{2i} + 2\varepsilon_{2i}H_{2i}^{\mathrm{T}}H_{2i} - (1-d_i)Q,$$

$$\mu = 1 + \rho(\tau_2 - \tau_1), \quad \rho = \max\left\{\left|\pi_{ii}\right|, i \in S\right\}, \quad \tau_1 = \min\left\{\tau_{1i}, i \in S\right\}, \quad \tau_2 = \max\left\{\tau_{2i}, i \in S\right\}.$$

Proof. Define $x_t(s) = x(t+s), t-\tau_{r_t}(t) \le s \le t$, then $\{(x_t, r_t), t \ge 0\}$ is a Markov process with initial state $(\phi(\cdot), r_0)$. Now, define a stochastic Lyapunov-Krasovskii functional as

$$V(\xi_t, r_t) = \xi^{\mathrm{T}}(t)P(r_t)\xi(t) + \int_{t-\tau_{r_t}(t)}^{t}\xi^{\mathrm{T}}(s)H^{\mathrm{T}}QH\xi(s)ds + \rho\int_{-\tau_2}^{-\tau_1}\int_{t+\theta}^{t}\xi^{\mathrm{T}}(s)H^{\mathrm{T}}QH\xi(s)dsd\theta, \tag{66}$$

Let L be the weak infinitesimal operator of the stochastic process $\{(x_t, r_t), t \ge 0\}$. By Itô differential formula, the stochastic differential of $V(\xi_t, r_t)$ along the trajectory of system $(\tilde{\Sigma})$ with $v(t)=0$ for $r_t = i, i \in S$ is given by

$$dV(\xi_t, i) = L[V(\xi_t, i)] + 2\xi^{\mathrm{T}}(t)P_i[\tilde{E}_i(t)H\xi(t) + \tilde{E}_{di}(t)H\xi(t - \tau_i(t))], \tag{67}$$

where

$$
\begin{aligned}
L[V(\xi_t, i)] = & \xi^{\mathrm{T}}(t)(\sum_{j=1}^{N}\pi_{ij}P_j)\xi(t) + 2\xi^{\mathrm{T}}(t)P_i[\tilde{A}_i(t)\xi(t) + \tilde{A}_{di}(t)H\xi(t - \tau_i(t)) \\
& + \tilde{D}_{1i}f(H\xi(t), H\xi(t-\tau_i(t)), i) - \tilde{D}_{2i}g(H\xi(t), H\xi(t-\tau_i(t)), i)] \\
& + [\tilde{E}_i(t)H\xi(t) + \tilde{E}_{di}(t)H\xi(t-\tau_i(t))]^{\mathrm{T}}P_i[\tilde{E}_i(t)H\xi(t) + \tilde{E}_{di}(t)H\xi(t-\tau_i(t))] \\
& + \sum_{j=1}^{N}\pi_{ij}\int_{t-\tau_j(t)}^{t}\xi^{\mathrm{T}}(s)H^{\mathrm{T}}QH\xi(s)ds + \xi^{\mathrm{T}}(t)H^{\mathrm{T}}QH\xi(t) - (1-\dot{\tau}_i(t))\xi^{\mathrm{T}}(t-\tau_i(t))H^{\mathrm{T}}QH\xi(t-\tau_i(t)) \\
& + \rho(\tau_2 - \tau_1)\xi^{\mathrm{T}}(t)H^{\mathrm{T}}QH\xi(t) - \rho\int_{t-\tau_2}^{t-\tau_1}\xi^{\mathrm{T}}(s)H^{\mathrm{T}}QH\xi(s)ds
\end{aligned}
\tag{68}
$$

Delay-dependent exponential stability
and filtering for time-delay stochastic systems with nonlinearities
249

Noting $\pi_{ij} \geq 0$ for $i \neq j$, and $\pi_{ii} \leq 0$, we have

$$\sum_{j=1}^{N} \pi_{ij} \int_{t-\tau_j(t)}^{t} \xi^{\mathrm{T}}(s) H^{\mathrm{T}} Q H \xi(s) ds \leq -\pi_{ii} \int_{t-\tau_2}^{t-\tau_1} \xi^{\mathrm{T}}(s) H^{\mathrm{T}} Q H \xi(s) ds \leq \rho \int_{t-\tau_2}^{t-\tau_1} \xi^{\mathrm{T}}(s) H^{\mathrm{T}} Q H \xi(s) ds. \quad (69)$$

Noting (56), (57) and using Lemma 2, we have

$$2\xi^{\mathrm{T}}(t) P_i \tilde{D}_{1i} f(H\xi(t), H\xi(t-\tau_i(t)), i)$$
$$\leq \varepsilon_{1i}^{-1} \xi^{\mathrm{T}}(t) P_i \tilde{D}_{1i} \tilde{D}_{1i}^{\mathrm{T}} P \xi(t) + 2\varepsilon_{1i} (\xi^{\mathrm{T}}(t) H^{\mathrm{T}} F_{1i}^{\mathrm{T}} F_{1i} H\xi(t) + \xi^{\mathrm{T}}(t-\tau_i(t)) H F_{2i}^{\mathrm{T}} F_{2i} H\xi(t-\tau_i(t))), \quad (70)$$

and

$$-2\xi^{\mathrm{T}}(t) P_i \tilde{D}_{2i} g(H\xi(t), H\xi(t-\tau_i(t)), i)$$
$$\leq \varepsilon_{2i}^{-1} \xi^{\mathrm{T}}(t) P_i \tilde{D}_{2i} \tilde{D}_{2i}^{\mathrm{T}} P_i \xi(t) + 2\varepsilon_{2i} (\xi^{\mathrm{T}}(t) H^{\mathrm{T}} H_{1i}^{\mathrm{T}} H_{1i} H\xi(t) + \xi^{\mathrm{T}}(t-\tau_i(t)) H H_{2i}^{\mathrm{T}} H_{2i} H\xi(t-\tau_i(t))), \quad (71)$$

Substituting (69)-(71) into (68), then, it follows from (68) that for each $r_t = i, i \in S$

$$L[V(\xi_t, i)] \leq \eta^{\mathrm{T}}(t) \Theta_i \eta(t), \quad (72)$$

where

$$\eta(t) = \begin{bmatrix} \xi^{\mathrm{T}}(t) & \xi^{\mathrm{T}}(t-\tau_i(t)) H^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}, \qquad \Theta_i = \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ * & \Theta_{22} \end{bmatrix},$$

$$\Theta_{11} = \sum_{j=1}^{N} \pi_{ij} P_j + P_i \tilde{A}_i(t) + \tilde{A}_i^{\mathrm{T}}(t) P_i + \varepsilon_{1i}^{-1} P_i \tilde{D}_{1i} \tilde{D}_{1i}^{\mathrm{T}} P + 2\varepsilon_{1i} H^{\mathrm{T}} F_{1i}^{\mathrm{T}} F_{1i} H + \varepsilon_{2i}^{-1} P_i \tilde{D}_{2i} \tilde{D}_{2i}^{\mathrm{T}} P_i$$
$$+ 2\varepsilon_{2i} H^{\mathrm{T}} H_{1i}^{\mathrm{T}} H_{1i} H + H^{\mathrm{T}} \tilde{E}_i^{\mathrm{T}}(t) P_i \tilde{E}_i(t) H + \mu H^{\mathrm{T}} Q H,$$
$$\Theta_{12} = P_i \tilde{A}_{di}(t) + H^{\mathrm{T}} \tilde{E}_i^{\mathrm{T}}(t) P_i \tilde{E}_{di}(t), \Theta_{22} = 2\varepsilon_{1i} F_{2i}^{\mathrm{T}} F_{2i} + 2\varepsilon_{2i} H_{2i}^{\mathrm{T}} H_{2i} + \tilde{E}_{di}^{\mathrm{T}}(t) P_i \tilde{E}_{di}(t) - (1-d_i) Q,$$

By the Schur complement, it is ease to see that LMI in (65) implies that $\Theta_i < 0$. Therefore, from (72) we obtain

$$L[V(\xi_t, i)] \leq -\delta \eta^{\mathrm{T}}(t) \eta(t), \quad (73)$$

where $\delta = \min_{i \in S} \{\lambda_{\min}(-\Theta_i)\}$. By Dynkin's formula, we can obtain

$$E\{V(\xi_t, i)\} - E\{V(\xi_0, r_0)\} = E\left\{\int_0^t L[V(\xi_s, i)] ds\right\} \leq -\delta \int_0^t E\{\xi^{\mathrm{T}}(s) \xi(s)\} ds. \quad (74)$$

On the other hand, it is follows from (66) that

$$E\{V(\xi_t,i)\} \geq \lambda_p E\{\xi^T(t)\xi(t)\}, \qquad (75)$$

where $\lambda_p = \min_{i \in S}\{\lambda_{\min}(P_i)\} > 0.$ Therefore, by (74) and (75),

$$E\{\xi^T(t)\xi(t)\} \leq \lambda_p^{-1}V(\xi_0,r_0) - \delta\lambda_p^{-1}\int_0^t E\{\xi^T(s)\xi(s)\}ds. \qquad (76)$$

Then, applying Gronwall-Bellman lemma to (76) yields

$$E\{\xi^T(t)\xi(t)\} \leq \lambda_p^{-1}V(\xi_0,r_0)e^{-\delta\lambda_p^{-1}t}.$$

Noting that there exists a scalar $\alpha > 0$ such that $\lambda_p^{-1}V(\xi_0,r_0) \leq \alpha \sup_{-\tau_2 \leq \theta \leq 0}|\varsigma(\theta)|^2.$

Defining $\beta = \delta\lambda_p^{-1} > 0$, then we have $\mathrm{E}|\xi(t)|^2 \leq \alpha e^{-\beta t} \sup_{-\tau_2 \leq \theta \leq 0}\mathrm{E}|\varsigma(\theta)|^2,$

and, hence, the robust exponential mean-square stability of the filtering error system $(\tilde{\Sigma})$ with $v(t)=0$ is established.

Now, we shall establish the H∞ performance for the system $(\tilde{\Sigma})$, we introduce

$$J(t) = \mathrm{E}\int_0^t[\tilde{z}^T(s)\tilde{z}(s) - \gamma^2 v^T(s)v(s)]ds, \qquad (77)$$

where $t > 0.$ Noting under the zero initial condition and $\mathrm{E}V(\xi_t,i) \geq 0,$ by the Lyapunov-Krasovskii functional (66), it can be shown that for any nonzero $v(t) \in L_2[0,\infty)$

$$J(t) = \mathrm{E}\left\{\int_0^t[\tilde{z}^T(s)\tilde{z}(s) - \gamma^2 v^T(s)v(s) + LV(\xi_s,i)]ds\right\} - \mathrm{E}V(\xi_t,i) \leq \mathrm{E}\left\{\int_0^t \eta^T(s)\Psi_i\eta(s)ds\right\}, \quad (78)$$

where

$$\eta(s) = \begin{bmatrix}\xi^T(s) & v^T(s)\end{bmatrix}^T, \Psi_i(t) = \begin{bmatrix}\Theta_{11}+\tilde{L}_i^T\tilde{L}_i & \Theta_{12} & P_i\tilde{B}_i(t)+H^T\tilde{E}_i^T(t)P_i\tilde{G}_i(t), \\ * & \Theta_{22} & \tilde{E}_{di}^T(t)P_i\tilde{G}_i(t), \\ * & * & \tilde{G}_i^T(t)P_i\tilde{G}_i(t)-\gamma^2 I\end{bmatrix},$$

Now, applying Schur complement to (65), we have $\Psi_i(t) < 0.$ This together with (78) implies that $J(t) < 0$ for any nonzero $v(t) \in L_2[0,\infty).$ Therefore, under zero conditions and for any nonzero $v(t) \in L_2[0,\infty)$, letting $t \to \infty$, we have $\|\tilde{z}(t)\|_{E_2} \leq \gamma\|v(t)\|_{E2}$ if (65) is satisfied. This completes the proof. □

Now, we are in a position to present a solution to the H∞ exponential filter design problem.

Theorem 4. Consider the uncertain Markovian jump stochastic system (Σ). Given scalars $\tau_{1i}, \tau_{2i}, d_i$ and $\gamma > 0$, for any delays $\tau_i(t)$ satisfying (7), the filtering error system $(\tilde{\Sigma})$ is robustly exponentially mean-square stable and (64) is satisfied under zero-initial conditions for any nonzero $v(t) \in L_2[0,\infty)$ and all admissible uncertainties, if for each $i \in S$ there exist matrices $P_{1i} > 0, P_{2i} > 0, Q > 0, W_i, Z_i$ and sclars $\varepsilon_{1i} > 0, \varepsilon_{2i} > 0, \varepsilon_{3i} > 0, \varepsilon_{4i} > 0$ such that the following LMI holds

$$
\Xi_i =
$$

$$
\begin{bmatrix}
\Xi_{11} & \Xi_{12} & \Xi_{13} & \Xi_{14} & E_i^T P_{1i} & E_i^T P_{2i} & P_{1i}D_{1i} & 0 & P_{1i}M_{1i} & 0 & L_i^T - L_{fi}^T \\
* & \Xi_{22} & \Xi_{23} & \Xi_{24} & 0 & 0 & P_{2i}D_{1i} & Z_iD_{2i} & P_{2i}M_{1i}-Z_iM_{3i} & 0 & L_{fi}^T \\
* & * & \Xi_{33} & \Xi_{34} & E_{di}^T P_{1i} & E_{di}^T P_{2i} & 0 & 0 & 0 & 0 & 0 \\
* & * & * & \Xi_{44} & G_i^T P_{1i} & G_i^T P_{2i} & 0 & 0 & 0 & 0 & 0 \\
* & * & * & * & -P_{1i} & 0 & 0 & 0 & 0 & P_{1i}M_{2i} & 0 \\
* & * & * & * & * & -P_{2i} & 0 & 0 & 0 & P_{2i}M_{2i} & 0 \\
* & * & * & * & * & * & -\varepsilon_{1i}I & 0 & 0 & 0 & 0 \\
* & * & * & * & * & * & * & -\varepsilon_{2i}I & 0 & 0 & 0 \\
* & * & * & * & * & * & * & * & -\varepsilon_{3i}I & 0 & 0 \\
* & * & * & * & * & * & * & * & * & -\varepsilon_{4i}I & 0 \\
* & * & * & * & * & * & * & * & * & * & -I
\end{bmatrix} < 0, (79)
$$

Where

$$
\Xi_{11} = \sum_{j=1}^N \pi_{ij}P_{1j} + P_{1i}A_i + A_i^T P_{1i} + 2\varepsilon_{1i}F_{1i}^T F_{1i} + 2\varepsilon_{2i}H_{1i}^T H_{1i} + \mu Q + \varepsilon_i N_{1i}^T N_{1i},
$$

$$
\Xi_{12} = A_i^T P_{2i} - W_i^T - C_i^T Z_i^T, \quad \Xi_{13} = P_{1i}A_{di} + \varepsilon_i N_{1i}^T N_{2i}, \quad \Xi_{14} = P_{1i}B_{1i} + \varepsilon_i N_{1i}^T N_{3i},
$$

$$
\Xi_{22} = \sum_{j=1}^N \pi_{ij}P_{2j} + W_i + W_i^T, \quad \Xi_{23} = P_{2i}A_{di} - Z_iC_{di}, \qquad \Xi_{24} = P_{2i}B_{1i} - Z_iB_{2i},
$$

$$
\Xi_{33} = 2\varepsilon_{1i}F_{2i}^T F_{2i} + 2\varepsilon_{2i}H_{2i}^T H_{2i} - (1-d_i)Q + \varepsilon_i N_{2i}^T N_{2i}, \quad \Xi_{34} = \varepsilon_i N_{2i}^T N_{3i},
$$

$$
\Xi_{44} = -\gamma^2 I + \varepsilon_i N_{3i}^T N_{3i}, \qquad \varepsilon_i = \varepsilon_{3i} + \varepsilon_{4i}, \qquad \mu = 1 + \rho(\tau_2 - \tau_1).
$$

In this case, a desired robust Markovian jump exponential H∞ filter is given in the form of (58)-(59) with parameters as follows

$$
A_{fi} = P_{2i}^{-1}W_i, \quad B_{fi} = P_{2i}^{-1}Z_i, \quad L_{fi}, \quad i \in S. \tag{80}
$$

Proof. Noting that for $r_t = i, \ i \in S$

$$\begin{bmatrix} \Delta\tilde{A}_i(t) & \Delta\tilde{A}_{di}(t) & \Delta\tilde{B}_i(t) \end{bmatrix} = \tilde{M}_{1i}F_i(t)\begin{bmatrix} \tilde{N}_{1i} & \tilde{N}_{2i} & \tilde{N}_{3i} \end{bmatrix}, \tag{81}$$

and

$$\begin{bmatrix} \Delta\tilde{E}_i(t) & \Delta\tilde{E}_{di}(t) & \Delta\tilde{G}_i(t) \end{bmatrix} = \tilde{M}_{2i}F_i(t)\begin{bmatrix} N_{1i} & N_{2i} & N_{3i} \end{bmatrix}, \tag{82}$$

where

$$\tilde{M}_{1i} = \begin{bmatrix} M_{1i} \\ M_{1i} - B_{fi}M_{3i} \end{bmatrix}, \ \tilde{M}_{2i} = \begin{bmatrix} M_{2i} \\ M_{2i} \end{bmatrix}, \ \tilde{N}_{1i} = \begin{bmatrix} N_{1i} & 0 \end{bmatrix}, \ \tilde{N}_{2i} = N_{2i}, \ \tilde{N}_{3i} = N_{3i}.$$

Then, it is readily to see that (65) can be written in the form as

$$\Phi_i = \Phi_{i0} + \Lambda_{i1}F_i(t)\Gamma_{i1} + \Gamma_{i1}^{\mathrm{T}}F_i^{\mathrm{T}}(t)\Lambda_{i1}^{\mathrm{T}} + \Lambda_{i2}F_i(t)\Gamma_{i2} + \Gamma_{i2}^{\mathrm{T}}F_i^{\mathrm{T}}(t)\Lambda_{i2}^{\mathrm{T}} < 0, \tag{83}$$

where

$$\Phi_{i0} = \begin{bmatrix} \Phi_{110} & P_i\tilde{A}_{di} & P_i\tilde{B}_i & H^{\mathrm{T}}\tilde{E}_i^{\mathrm{T}}P_i & P_i\tilde{D}_{1i} & P_i\tilde{D}_{2i} \\ * & \Phi_{22} & 0 & \tilde{E}_{di}^{\mathrm{T}}P_i & 0 & 0 \\ * & * & -\gamma^2 I & \tilde{G}_i^{\mathrm{T}}P_i & 0 & 0 \\ * & * & * & -P_i & 0 & 0 \\ * & * & * & * & -\varepsilon_{1i}I & 0 \\ * & * & * & * & * & -\varepsilon_{2i}I \end{bmatrix},$$

$$\Phi_{110} = \sum_{j=1}^{N} \pi_{ij}P_j + P_i\tilde{A}_i + \tilde{A}_i^{\mathrm{T}}P_i + 2\varepsilon_{1i}H^{\mathrm{T}}F_{1i}^{\mathrm{T}}F_{1i}H + 2\varepsilon_{2i}H^{\mathrm{T}}H_{1i}^{\mathrm{T}}H_{1i}H + \mu H^{\mathrm{T}}QH + \tilde{L}_i^{\mathrm{T}}\tilde{L}_i,$$

$$\Lambda_{i1} = \begin{bmatrix} \tilde{M}_{1i}^{\mathrm{T}}P_i & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^{\mathrm{T}}, \quad \Gamma_{i1} = \begin{bmatrix} \tilde{N}_{1i} & \tilde{N}_{2i} & \tilde{N}_{3i} & 0 & 0 & 0 \end{bmatrix},$$

$$\Lambda_{i2} = \begin{bmatrix} 0 & 0 & 0 & \tilde{M}_{2i}^{\mathrm{T}}P_i & 0 & 0 \end{bmatrix}^{\mathrm{T}}, \quad \Gamma_{i2} = \begin{bmatrix} N_{1i}H & N_{2i} & N_{3i} & 0 & 0 & 0 \end{bmatrix},$$

From (83) and by using Lemma 3, there exists positive scalars $\varepsilon_{3i} > 0, \varepsilon_{4i} > 0$ such that the following inequality holds

$$\Phi_{i0} + \varepsilon_{3i}^{-1}\Lambda_{i1}\Lambda_{i1}^{\mathrm{T}} + \varepsilon_{3i}\Gamma_{i1}^{\mathrm{T}}\Gamma_{i1} + \varepsilon_{4i}^{-1}\Lambda_{i2}\Lambda_{i2}^{\mathrm{T}} + \varepsilon_{4i}\Gamma_{i2}^{\mathrm{T}}\Gamma_{i2} < 0, \tag{84}$$

then, by applying the Schur complement to (84), we have

$$
\tilde{\Phi}_i = \begin{bmatrix}
\tilde{\Phi}_{11} & \tilde{\Phi}_{12} & \tilde{\Phi}_{13} & H^{\mathrm{T}} \tilde{E}_i^{\mathrm{T}} P_i & P_i \tilde{D}_{1i} & P_i \tilde{D}_{2i} & P_i \tilde{M}_{1i} & 0 \\
* & \tilde{\Phi}_{22} & \tilde{\Phi}_{23} & \tilde{E}_{di}^{\mathrm{T}} P_i & 0 & 0 & 0 & 0 \\
* & * & \tilde{\Phi}_{33} & \tilde{G}_i^{\mathrm{T}} P_i & 0 & 0 & 0 & 0 \\
* & * & * & -P_i & 0 & 0 & 0 & P_i \tilde{M}_{2i} \\
* & * & * & * & -\varepsilon_{1i} I & 0 & 0 & 0 \\
* & * & * & * & * & -\varepsilon_{2i} I & 0 & 0 \\
* & * & * & * & * & * & -\varepsilon_{3i} I & 0 \\
* & * & * & * & * & * & * & -\varepsilon_{4i} I
\end{bmatrix},
\tag{85}
$$

where

$$
\begin{aligned}
\tilde{\Phi}_{11} &= \sum_{j=1}^{N} \pi_{ij} P_j + P_i \tilde{A}_i + \tilde{A}_i^{\mathrm{T}} P_i + 2\varepsilon_{1i} H^{\mathrm{T}} F_{1i}^{\mathrm{T}} F_{1i} H + 2\varepsilon_{2i} H^{\mathrm{T}} H_{1i}^{\mathrm{T}} H_{1i} H + \mu H^{\mathrm{T}} Q H \\
&\quad + \tilde{L}_i^{\mathrm{T}} \tilde{L}_i + \varepsilon_{3i} \tilde{N}_{1i}^{\mathrm{T}} \tilde{N}_{1i} + \varepsilon_{4i} H^{\mathrm{T}} N_{1i}^{\mathrm{T}} N_{1i} H, \\
\tilde{\Phi}_{12} &= P_i \tilde{A}_{di} + \varepsilon_{3i} \tilde{N}_{1i}^{\mathrm{T}} \tilde{N}_{2i} + \varepsilon_{4i} H^{\mathrm{T}} N_{1i}^{\mathrm{T}} N_{2i}, \quad \tilde{\Phi}_{13} = P_i \tilde{B}_i + \varepsilon_{3i} \tilde{N}_{1i}^{\mathrm{T}} \tilde{N}_{3i} + \varepsilon_{4i} H^{\mathrm{T}} N_{1i}^{\mathrm{T}} N_{3i}, \\
\tilde{\Phi}_{22} &= 2\varepsilon_{1i} F_{2i}^{\mathrm{T}} F_{2i} + 2\varepsilon_{2i} H_{2i}^{\mathrm{T}} H_{2i} - (1 - d_i) Q + \varepsilon_{3i} \tilde{N}_{2i}^{\mathrm{T}} \tilde{N}_{2i} + \varepsilon_{4i} N_{2i}^{\mathrm{T}} N_{2i}, \\
\tilde{\Phi}_{23} &= \varepsilon_{3i} \tilde{N}_{2i}^{\mathrm{T}} \tilde{N}_{3i} + \varepsilon_{4i} N_{2i}^{\mathrm{T}} N_{3i}, \qquad\qquad \tilde{\Phi}_{33} = -\gamma^2 I + \varepsilon_{3i} \tilde{N}_{3i}^{\mathrm{T}} \tilde{N}_{3i} + \varepsilon_{4i} N_{3i}^{\mathrm{T}} N_{3i}.
\end{aligned}
$$

For each $r_t = i, \; i \in S$, we define the matrix $P_i > 0$ by

$$
P_i = \begin{bmatrix} P_{1i} & 0 \\ 0 & P_{2i} \end{bmatrix}.
$$

Then, substituting the matrix P_i, the matrices $\tilde{A}_i, \tilde{A}_{di}, \tilde{B}_i, \tilde{E}_i, \tilde{E}_{di}, \tilde{B}_i, \tilde{E}_i, \tilde{E}_{di}, \tilde{G}_i, \tilde{D}_{1i}, \tilde{D}_{2i}, \tilde{L}_i,$ H defined in (61)-(62) into (85) and by introducing some matrices given by $W_i = P_{2i} A_{fi}, Z_i = P_{2i} B_{fi}$, then, we can obtain the results in Theorem 4. This completes the proof. $\square$

3. Numerical Examples and Simulations

Example 1: Consider the uncertain stochastic time-delay system with nonlinearities

$$
dx(t) = \Big[\big(A + \Delta A(t)\big) x(t) + \big(B + \Delta B(t)\big) x(t - \tau(t)) \Big] dt + g(t, x(t), x(t - \tau(t))) d\omega(t), \tag{86}
$$

where

$$A = \begin{bmatrix} -2 & 0 \\ 1 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 0 \\ -0.5 & -1 \end{bmatrix}, \quad \|\Delta A(t)\| \le 0.1, \quad \|\Delta B(t)\| \le 0.1,$$

$$trace\left[g^{\mathrm{T}}(t,x(t),x(t-\tau(t)))g(t,x(t),x(t-\tau(t))) \right] \le 0.1\|x(t)\|^2 + 0.1\|x(t-\tau(t))\|^2.$$

$$E = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}, G_1 = G_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, H_1 = H_2 = \begin{bmatrix} \sqrt{0.1} & 0 \\ 0 & \sqrt{0.1} \end{bmatrix}.$$

For the time-invariant system, applying Theorem 1, it has been found that by using MATLAB LMI Toolbox that system (86) is exponentially stable in mean square for any delay $0 \le \tau \le 1.0898$. It is note that the result of (Yue & Won, 2001) guarantees the exponential stability of (86) when $0 \le \tau \le 0.8635$, whereas by the method of (Mao, 1996) the delay is only allowed 0.1750. According to Theorem 1, the MADB for different d is shown in Table 1. For a comparison with the results of other researchers, a summary is given in the following Table 1. It is obvious that the result in this paper is much less conservative and is an improvement of the results than that of (Mao, 1996) and (Yue & Won, 2001).

The stochastic perturbation of the system is Brownian motion and it can be depicted in Fig.1. The simulation of the state response for system (86) with $\tau = 1.0898$ was depicted in Fig.2.

Methods	$d = 0$	$d = 0.5$	$d = 0.9$
(Mao, 1996)	0.1750	-	-
(Yue & Won, 2001)	0.8635	-	-
Theorem 1	1.0898	0.5335	0.1459

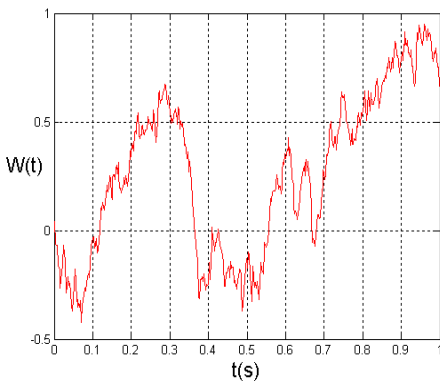Table1. Maximum allowable time delay to different d

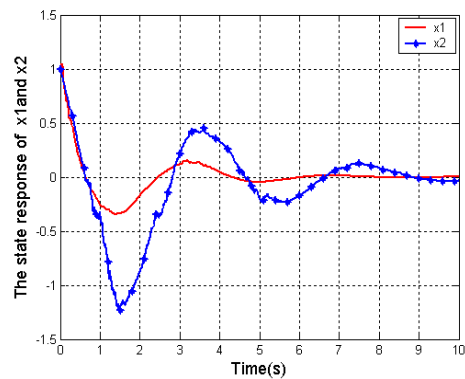
Fig. 1. The trajectory of Brownian motion   Fig. 2. The state response of system (47)

Example 2. Consider the uncertain Markovian jump stochastic systems in the form of (47)-(48) with two modes. For mode 1, the parameters as the following:

$$A_1 = \begin{bmatrix} -3 & 1 & 0 \\ 0.3 & -4.5 & 1 \\ -0.1 & 0.3 & -3.8 \end{bmatrix}, A_{d1} = \begin{bmatrix} -0.2 & 0.1 & 0.6 \\ 0.5 & -1 & -0.8 \\ 0 & 1 & -2.5 \end{bmatrix}, D_{11} = \begin{bmatrix} 0 & 0.1 & 0 \\ 0.1 & 0.1 & 0 \\ 0.1 & 0.2 & 0.2 \end{bmatrix}, B_{11} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix},$$

$$E_1 = \begin{bmatrix} 0.1 & -0.1 & 0.2 \\ 0.3 & 0.3 & -0.4 \\ 0.1 & 0.1 & -0.3 \end{bmatrix}, E_{d1} = \begin{bmatrix} 0.1 & -0.1 & 0.2 \\ 0.3 & 0.3 & -0.4 \\ 0.1 & 0.1 & -0.3 \end{bmatrix}, G_1 = \begin{bmatrix} 0.2 \\ 0 \\ 0.1 \end{bmatrix}, C_1 = \begin{bmatrix} 0.8 & 0.3 & 0 \end{bmatrix},$$

$$C_{d1} = \begin{bmatrix} 0.2 & -0.3 & -0.6 \end{bmatrix}, D_{21} = 0.1, B_{21} = 0.2, L_1 = \begin{bmatrix} 0.5 & -0.1 & 1 \end{bmatrix},$$

$$F_{11} = F_{21} = H_{11} = H_{21} = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}, M_{11} = \begin{bmatrix} 0.1 \\ 0 \\ 0.2 \end{bmatrix}, M_{21} = \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \end{bmatrix}, M_{31} = 0.2,$$

$$N_{11} = \begin{bmatrix} 0.2 & 0 & 0.1 \end{bmatrix}, N_{21} = \begin{bmatrix} 0.1 & 0.2 & 0 \end{bmatrix}, N_{31} = 0.2.$$

and the time-varying delay $\tau(t)$ satisfies (53) with $\tau_{11} = 0.2$, $\tau_{21} = 1.3$, $d_1 = 0.2$.
For mode 2, the dynamics of the system are describe as

$$A_2 = \begin{bmatrix} -2.5 & 0.5 & -0.1 \\ 0.1 & -3.5 & 0.3 \\ -0.1 & 1 & -3.2 \end{bmatrix}, A_{d2} = \begin{bmatrix} 0 & -0.3 & 0.6 \\ 0.1 & 0.5 & 0 \\ -0.6 & 1 & -0.8 \end{bmatrix}, D_{12} = \begin{bmatrix} 0.1 & 0 & 0.1 \\ 0.1 & 0.2 & 0 \\ 0.2 & 0.1 & 0.1 \end{bmatrix}, B_{12} = \begin{bmatrix} -0.6 \\ 0.5 \\ 0 \end{bmatrix},$$

$$E_2 = \begin{bmatrix} 0.1 & -1 & 0.2 \\ 0.3 & 0.3 & -0.4 \\ 1 & 0.1 & 0.3 \end{bmatrix}, E_{d2} = \begin{bmatrix} 0.1 & -0.1 & 0.2 \\ 0.3 & 0.3 & -0.4 \\ 0.1 & 0.1 & 0.3 \end{bmatrix}, G_2 = \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \end{bmatrix}, C_2 = \begin{bmatrix} -0.5 & 0.2 & 0.3 \end{bmatrix},$$

$$C_{d2} = \begin{bmatrix} 0 & -0.6 & 0.2 \end{bmatrix}, D_{22} = 0.1, B_{22} = 0.5, L_2 = \begin{bmatrix} 0 & 1 & 0.6 \end{bmatrix},$$

$$F_{12} = F_{22} = H_{12} = H_{22} = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}, M_{12} = \begin{bmatrix} 0.1 \\ 0.1 \\ 0 \end{bmatrix}, M_{22} = \begin{bmatrix} 0.1 \\ 0.1 \\ 0 \end{bmatrix}, M_{32} = 0.1,$$

$$N_{12} = \begin{bmatrix} 0.1 & 0.1 & 0 \end{bmatrix}, N_{22} = \begin{bmatrix} 0 & -0.1 & 0.2 \end{bmatrix}, N_{32} = 0.1.$$

and the time-varying delay $\tau(t)$ satisfies (53) with $\tau_{12} = 0.1$, $\tau_{22} = 1.1$, $d_2 = 0.3$.

Suppose the transition probability matrix to be $\Pi = \begin{bmatrix} -0.5 & 0.5 \\ 0.3 & -0.3 \end{bmatrix}$.

The objective is to design a Markovian jump H∞ filter in the form of (58)-(59), such that for all admissible uncertainties, the filtering error system is exponentially mean-square stable and (64) holds. In this example, we assume the disturbance attenuation level $\gamma = 1.2$.

By using Matlab LMI Control Toolbox to solve the LMI in (77), we can obtain the solutions as follows:

$$P_{11} = \begin{bmatrix} 0.7952 & 0.0846 & 0.0051 \\ 0.0846 & 0.6355 & -0.1857 \\ 0.0051 & -0.1857 & 0.7103 \end{bmatrix}, P_{21} = \begin{bmatrix} 0.6847 & 0.0549 & -0.0341 \\ 0.0549 & 0.4614 & -0.0350 \\ -0.0341 & -0.0350 & 0.5624 \end{bmatrix}, P_{12} = \begin{bmatrix} 1.0836 & 0.1117 & -0.0355 \\ 0.1117 & 0.9508 & -0.1800 \\ -0.0355 & -0.1800 & 0.7222 \end{bmatrix},$$

$$P_{22} = \begin{bmatrix} 0.5974 & 0.0716 & -0.0536 \\ 0.0716 & 0.5827 & 0.0814 \\ -0.0536 & 0.0814 & 0.3835 \end{bmatrix}, Q = \begin{bmatrix} 1.4336 & -0.0838 & -0.0495 \\ -0.0838 & 2.1859 & -1.1472 \\ -0.0495 & -1.1472 & 1.9649 \end{bmatrix}, W_1 = \begin{bmatrix} -0.9731 & 0.3955 & 0.4457 \\ -0.3560 & -1.1939 & 0.5584 \\ -0.6309 & -0.5217 & -1.2038 \end{bmatrix},$$

$$W_2 = \begin{bmatrix} -0.8741 & -0.0117 & 0.1432 \\ -0.0276 & -1.1101 & -0.0437 \\ -0.1726 & 0.1087 & -0.8501 \end{bmatrix}, Z_1 = \begin{bmatrix} -0.3844 \\ 0.1797 \\ 1.2608 \end{bmatrix}, Z_2 = \begin{bmatrix} 0.0072 \\ -0.0572 \\ -0.0995 \end{bmatrix},$$

$e_{11} = 1.2704, \ e_{21} = 1.1626, \ e_{31} = 1.0887, \ e_{41} = 1.0670, \ e_{12} = 1.2945, \ e_{22} = 1.2173, \ e_{32} = 1.2434, \ e_{42} = 1.2629.$

Then, by Theorem 4, the parameters of desired robust Markovian jump H∞ filter can be obtained as follows

$$A_{f1} = \begin{bmatrix} -1.4278 & 0.7459 & 0.4686 \\ -0.6967 & -2.7564 & 0.9989 \\ -1.2519 & -1.0541 & -2.0500 \end{bmatrix}, B_{f1} = \begin{bmatrix} -0.4989 \\ 0.6196 \\ 2.2503 \end{bmatrix}, L_{f1} = \begin{bmatrix} 0.3042 & 0.0467 & 0.7872 \end{bmatrix};$$

$$A_{f2} = \begin{bmatrix} -1.5571 & 0.2940 & 0.0074 \\ 0.2446 & -2.0474 & 0.2408 \\ -0.7197 & 0.7592 & -2.2669 \end{bmatrix}, B_{f2} = \begin{bmatrix} -0.0024 \\ -0.0635 \\ -0.2463 \end{bmatrix}, L_{f2} = \begin{bmatrix} 0.0037 & 0.5730 & 0.3981 \end{bmatrix}.$$

The simulation result of the state response of the real states $x(t)$ and their estimates $\hat{x}(t)$ are displayed in Fig. 3. Fig. 4 is the simulation result of the estimation error response of $\tilde{z}(t) = z(t) - \hat{z}(t)$. The simulation results demonstrate that the estimation error is robustly exponentially mean-square stable, and thus it can be seen that the designed filter satisfies the specified performance requirements and all the expected objectives are well achieved.
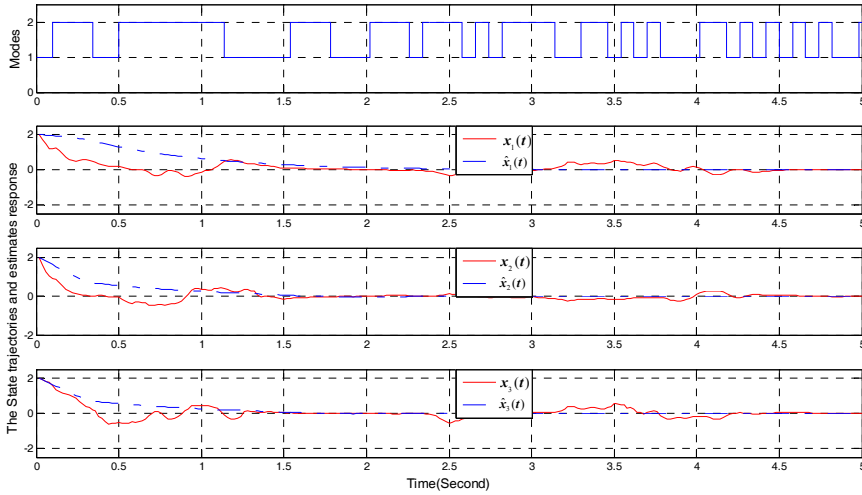
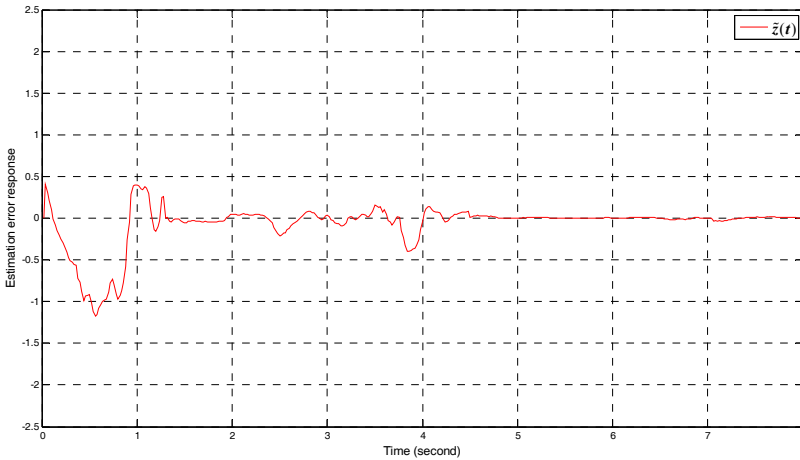Fig. 3. The state trajectories and estimates response

Fig. 4. The estimation error response

4. Conclusion

Both delay-dependent exponential mean-square stability and robust H∞ filtering for time-delay a class of *Itô* stochastic systems with time-varying delays and nonlinearities has addressed in this chapter. Novel stability criteria and H∞ exponential filter design methods are proposed in terms of LMIs. The new criteria are much less conservative than some existing results. The desired filter can be constructed through a convex optimization problem. Numerical examples and simulations have demonstrated the effectiveness and usefulness of the proposed methods.

5. Acknowledgments

6. References

Blythe S.; Mao, X. & Liao, X. (2001). Stability of stochastic delay neural networks, *J. Franklin Institute*, 338, 481–495.

Chen, W. H.; Guan, Z. H. & Lu, X. M. (2005). Delay-dependent exponential stability of uncertain stochastic systems with multiple delays: an LMI approach, *Systems Control Lett.*, 54, 547-555.

De Souza, C. E.; Xie, L. & Wang, Y. (1993). H∞ filtering for a class of uncertain nonlinear systems, *Syst. Control Lett.*, 20, 419–426.

Fridman, E.; Shaked, U. & Xie, L. (2003). Robust H∞ filtering of linear systems with time-varying delay, *IEEE Trans. Autom. Control*, 48(1), 159–165.

Gao, H. & Wang, C. (2003). Delay-dependent robust H∞ and $L_2 - L_\infty$ filtering for a class of uncertain nonlinear time-delay systems, *IEEE Trans. Autom. Control*, 48(9), 1661–1666.

Jing, X. J.; Tan,D. L. & Wang, Y.C. (2004). An LMI approach to stability of systems with severe time-delay, *IEEE Trans. Automatic Control*, 49(7), 1192-1195.

Kim, J. H. (2001). Delay and its time-derivative dependent robust stability of time-delayed linear systems with uncertainty, *IEEE Trans. Automatic Control*, 46(5), 789-792.

Kwon, O. M. & Park, J. H. (2004). On Improved delay-dependent robust control for uncertain time-delay systems, *IEEE. Trans. Automatic Control*, 49(11), 1991-1995.

Li, H. & Fu, M. (1997). A linear matrix inequality approach to robust H∞ filtering, *IEEE Trans. Signal Process.*, 45(3), 2338–2350.

Li, X. & de Souza, C. E. (1997). Delay-dependent robust stability and stabilization of uncertain linear delay systems: A linear matrix inequality approach, *IEEE Trans. on Automatic Control*, 42(11), 1144-1148.

Liu,Y.; Wang, Z. & Liu, X. (2008). Robust H∞ filtering for discrete nonlinear stochastic systems with time-varying delay, *J. Math. Anal. Appl.*, 341, 318-336.

Lu, C. Y.; Tsai, J. S. H. ; Jong, G. J. & Su, T. J. (2003). An LMI-based approach for robust stabilization of uncertain stochastic systems with time-varying delays, *IEEE Trans. Automatic Control,* 48(2), 286–289.

Mao, X. (1996). Robustness of exponential stability of stochastic differential delay equation, *IEEE Trans. Automatic Control*, 41(3), 442-447.

Mao, X. (2002). Exponential stability of stochastic delay interval systems with markovian switching, *IEEE Trans. Automatic Control*, 47(10), 1604-1612.

Moon, Y. S.; Park, P.; Kwon, W. H. & Lee, Y.S. (2001). Delay-dependent robust stabilization of uncertain state-delayed systems, *Internat. J. Control*, 74(14), 1447-1455.

Nagpal, K. M. & Khargonekar, P. P. (1991). Filtering and smoothing in an H∞ setting, *IEEE Trans. Automat. Control*, 36(3), 152–166.

Park, J. H. (2001). Robust stabilization for dynamic systems with multiple time varying delays and nonlinear uncertainties, *J. Optim. Theory Appl.*, 108, 155-174.

Pila,A. W.; Shaked, U. & De Souza, C. E. (1999). H∞ filtering for continuous-time linear systems with delay, *IEEE Trans. Automat. Control*, 44(7), 1412–1417.

Wang, Y.; Xie, L. & de Souza, C.E. (1992). Robust control of a class of uncertain nonlinear system, *Systems Control Lett.*, 19, 139–149.

Wang, Z.; Lam, J. & Liu, X. (2004). Exponential filtering for uncertain Markovian jump time-delay systems with nonlinear disturbances, *IEEE Trans. Circuits Syst. II*, 51(5), 262–268.

Wang, Z.; Liu, Y. & Liu, X.(2008). H∞ filtering for uncertain stochastic time-delay systems with sector-bounded nonlinearities, *Automatica*, 44, 1268-1277.

Wang, Z. & Yang, F. (2002). Robust filtering for uncertain linear systems with delayed states and outputs, *IEEE Trans. Circuits Syst. I*, 49(11), 125–130.

Wang, Z.; Yang, F.; Ho, D.W.C. & Liu, X. (2006). Robust H∞ filtering for stochastic time-delay systems with missing measurements, *IEEE Trans. Signal Process.*, 54(7), 2579-2587.

Wu, M.; He, Y.; She, J. H. & Liu, G. P. (2004). Delay-dependent criteria for robust stability of time-varying delays systems, *Automatica*, 40(3), 1435-1439.

Xie, L. (1996). Output feedback H∞ control of systems with parameter uncertainty, *Int. J. Control*, 63(1), 741–750.

Xie, S. & Xie, L. (2000). Stabilization of a class of uncertain large-scale stochastic systems with time delays, *Automatica*, 36, 161–167.

Xu, S. & Chen, T. (2002). Robust H∞ control for uncertain stochastic systems with state delay, *IEEE Trans. Automatic Control*, 47(12), 2089–2094.

Xu, S. & Chen, T. (2004). An LMI approach to the H∞ filter design for uncertain systems with distributed delays, *IEEE Trans. Circuits Syst. II*, 51(4), 195–201.

Xu, S.; Chen, T. & Lam, J. (2003). Robust H∞ filtering for uncertain Markovian jump systems with mode-dependent time delays, *IEEE Trans. Autom. Control*, 48(5), 900-907.

Xu, S. & Van Dooren, P.(2002). Robust H∞ filtering for a class of nonlinear systems with state delay and parameter uncertainty," *Int. J. Control*, 75, 766–774.

Yue,D. & Won, S. (2001). Delay-dependent robust stability of stochastic systems with time delay and nonlinear uncertainties, *Electron. Lett.*, 37(15), 992–993.

Zhang, W.; Chen, B. S. & Tseng, C.-S. (2005). Robust H∞ filtering for nonlinear stochastic systems, *IEEE Trans. Signal Process.*, 53(12), 589–598.

Zhang, X. & Han, Q.-L. (2008). Robust H∞ filtering for a class of uncertain linear systems with time-varying delay, *Automatica*, 44, 157-166.

Optimal Filtering for Linear States over Polynomial Observations

Joel Perez[1], Jose P. Perez[1] and Rogelio Soto[2]
[1] *Department of Physical and Mathematical Sciences, Autonomous University of Nuevo Leon*
[2] *Mechatronics and Automation Department, Monterrey Institute of Technology and Higher Education*
Mexico

1. Introduction

Although the general optimal solution of the filtering problem for nonlinear state and observation equations confused with white Gaussian noises is given by the equation for the conditional density of an unobserved state with respect to observations (see (1–6)), there are a very few known examples of nonlinear systems where that equation can be reduced to a finite-dimensional closed system of filtering equations for a certain number of lower conditional moments (see (7–10) for more details). Some relevant results on filtering for nonlinear stochastic systems can be found in (11–14). There also exists a considerable bibliography on robust filtering for the "general situation" systems (see, for example, (15–23)). Apart form the "general situation," the optimal finite-dimensional filters have recently been designed for certain classes of polynomial system states over linear observations with invertible ((24; 25; 27; 28)) or non-invertible ((26; 29)) observation matrix. However, the cited papers never consider filtering problems with nonlinear, in particular, polynomial observations.

This work presents the optimal finite-dimensional filter for linear system states over polynomial observations, continuing the research in the area of the optimal filtering for polynomial systems, which has been initiated in ((24–27; 29)). Designing the optimal filter over polynomial observations presents a significant advantage in the filtering theory and practice, since it enables one to address some filtering problems with observation nonlinearities, such as the optimal cubic sensor problem (30). The optimal filtering problem is treated proceeding from the general expression for the stochastic Ito differential of the optimal estimate and the error variance (31). As the first result, the Ito differentials for the optimal estimate and error variance corresponding to the stated filtering problem are derived. It is then proved that a closed finite-dimensional system of the optimal filtering equations with respect to a finite number of filtering variables can be obtained for a polynomial observation equation, additionally assuming a conditionally Gaussian initial condition for the higher degree states. This assumption is quite admissible in the filtering framework, since the real distribution of the entire state vector is actually unknown. In this case, the corresponding procedure for designing the optimal filtering equations is established.

As an illustrative example, the closed system of the optimal filtering equations with respect to two variables, the optimal estimate and the error variance, is derived in the explicit form for the particular case of the third degree polynomial observations. This filtering problem generalizes the optimal cubic sensor problem stated in (30), where nonexistence of a closed-form solution is indicated for the "general situation" case, without any assumptions for the third order state distribution. In our paper, taking into account that the real distributions of the first and third degree states are unknown, a conditionally Gaussian initial condition is additionally assumed for the third degree state. The resulting filter yields a reliable and rapidly converging estimate, in spite of a significant difference in the initial conditions between the state and estimate and very noisy observations, in the situation where the unmeasured state itself is a time-shifted Wiener process and the extended Kalman filter (EKF) approach fails.

2. Filtering Problem for Linear States over Polynomial Observations

Let (Ω, F, P) be a complete probability space with an increasing right-continuous family of σ-algebras $F_t, t \geq t_0$, and let $(W_1(t), F_t, t \geq t_0)$ and $(W_2(t), F_t, t \geq t_0)$ be independent Wiener processes. The F_t-measurable random process $(x(t), y(t)$ is described by a linear differential equation for the system state

$$dx(t) = (a_0(t) + a(t)x(t))dt + b(t)dW_1(t), \quad x(t_0) = x_0, \tag{1}$$

and a nonlinear polynomial differential equation for the observation process

$$dy(t) = h(x, t)dt + B(t)dW_2(t). \tag{2}$$

Here, $x(t) \in R^n$ is the state vector and $y(t) \in R^m$ is the observation vector. The initial condition $x_0 \in R^n$ is a Gaussian vector such that x_0, $W_1(t)$, and $W_2(t)$ are independent. It is assumed that $B(t)B^T(t)$ is a positive definite matrix. All coefficients in (1)–(2) are deterministic functions of time of appropriate dimensions. The nonlinear function $h(x, t)$ forms the drift in the observation equation (2).

The nonlinear function $h(x, t)$ is considered a polynomial of n variables, components of the state vector $x(t) \in R^n$, with time-dependent coefficients. Since $x(t) \in R^n$ is a vector, this requires a special definition of the polynomial for $n > 1$. In accordance with (27), a p-degree polynomial of a vector $x(t) \in R^n$ is regarded as a p-linear form of n components of $x(t)$

$$h(x, t) = \alpha_0(t) + \alpha_1(t)x + \alpha_2(t)xx^T + \ldots + \alpha_p(t)x \ldots_{p \ times} \ldots x, \tag{3}$$

where $\alpha_0(t)$ is a vector of dimension n, α_1 is a matrix of dimension $n \times n$, α_2 is a 3D tensor of dimension $n \times n \times n$, α_p is an $(p+1)$D tensor of dimension $n \times \ldots_{(p+1) \ times} \ldots \times n$, and $x \times \ldots_{p \ times} \ldots \times x$ is a pD tensor of dimension $n \times \ldots_{p \ times} \ldots \times n$ obtained by p times spatial multiplication of the vector $x(t)$ by itself (see (27) for more definition). Such a polynomial can also be expressed in the summation form

$$h_k(x, t) = \alpha_{0\,k}(t) + \sum_i \alpha_{1\,ki}(t)x_i(t) + \sum_{ij} \alpha_{2\,kij}(t)x_i(t)x_j(t) + \ldots$$
$$+ \sum_{i_1 \ldots i_p} \alpha_{p\,ki_1 \ldots i_p}(t)x_{i_1}(t) \ldots x_{i_p}(t), \quad k, i, j, i_1, \ldots, i_p = 1, \ldots, n.$$

The estimation problem is to find the optimal estimate $\hat{x}(t)$ of the system state $x(t)$, based on the observation process $Y(t) = \{y(s), 0 \leq s \leq t\}$, that minimizes the Euclidean 2-norm

$$J = E[(x(t) - \hat{x}(t))^T(x(t) - \hat{x}(t)) \mid F_t^Y]$$

at every time moment t. Here, $E[\xi(t) \mid F_t^Y]$ means the conditional expectation of a stochastic process $\xi(t) = (x(t) - \hat{x}(t))^T (x(t) - \hat{x}(t))$ with respect to the σ - algebra F_t^Y generated by the observation process $Y(t)$ in the interval $[t_0, t]$. As known (31), this optimal estimate is given by the conditional expectation

$$\hat{x}(t) = m_x(t) = E(x(t) \mid F_t^Y)$$

of the system state $x(t)$ with respect to the σ - algebra F_t^Y generated by the observation process $Y(t)$ in the interval $[t_0, t]$. As usual, the matrix function

$$P(t) = E[(x(t) - m_x(t))(x(t) - m_x(t))^T \mid F_t^Y]$$

is the estimation error variance.

The proposed solution to this optimal filtering problem is based on the formulas for the Ito differential of the optimal estimate and the estimation error variance (cited after (31)) and given in the following section.

3. Optimal Filter for Linear States over Polynomial Observations

Let us reformulate the problem, introducing the stochastic process $z(t) = h(x, t)$. Using the Ito formula (see (31)) for the stochastic differential of the nonlinear function $h(x, t)$, where $x(t)$ satisfies the equation (1), the following equation is obtained for $z(t)$

$$dz(t) = \frac{\partial h(x, t)}{\partial x}(a_0(t) + a(t)x(t))dt + \frac{\partial h(x, t)}{\partial t}dt + \tag{4}$$

$$\frac{1}{2}\frac{\partial^2 h(x, t)}{\partial x^2}b(t)b^T(t)dt + \frac{\partial h(x, t)}{\partial x}b(t)dW_1(t), \quad z(0) = z_0.$$

Note that the addition $\frac{1}{2}\frac{\partial^2 h(x,t)}{\partial x^2}b(t)b^T(t)dt$ appears in view of the second derivative in x in the Ito formula.

The initial condition $z_0 \in R^n$ is considered a conditionally Gaussian random vector with respect to observations. This assumption is quite admissible in the filtering framework, since the real distributions of $x(t)$ and $z(t)$ are actually unknown. Indeed, as follows from (32), if only two lower conditional moments, expectation m_0 and variance P_0, of a random vector $[z_0, x_0]$ are available, the Gaussian distribution with the same parameters, $N(m_0, P_0)$, is the best approximation for the unknown conditional distribution of $[z_0, x_0]$ with respect to observations. This fact is also a corollary of the central limit theorem (33) in the probability theory.

A key point for further derivations is that the right-hand side of the equation (4) is a polynomial in x. Indeed, since $h(x, t)$ is a polynomial in x, the functions $\frac{\partial h(x,t)}{\partial x}$, $\frac{\partial h(x,t)}{\partial x}x(t)$, $\frac{\partial h(x,t)}{\partial t}$, and $\frac{\partial^2 h(x,t)}{\partial x^2}$ are also polynomial in x. Thus, the equation (4) is a polynomial state equation with a polynomial multiplicative noise. It can be written in the compact form

$$dz(t) = f(x, t)dt + g(x, t)dW_1(t), \quad z(t_0) = z_0, \tag{5}$$

where

$$f(x, t) = \frac{\partial h(x, t)}{\partial x}(a_0(t) + a(t)x(t)) + \frac{\partial h(x, t)}{\partial t} +$$

$$\frac{1}{2}\frac{\partial^2 h(x, t)}{\partial x^2}b(t)b^T(t), \quad g(x, t) = \frac{\partial h(x, t)}{\partial x}b(t).$$

In terms of the process $z(t)$, the observation equation (2) takes the form

$$dy(t) = z(t)dt + B(t)dW_2(t). \tag{6}$$

The reformulated estimation problem is now to find the optimal estimate $[m_z(t), m_x(t)]$ of the system state $[z(t), x(t)]$, based on the observation process $Y(t) = \{y(s), 0 \leq s \leq t\}$. This optimal estimate is given by the conditional expectation

$$m(t) = [m_z(t), m_x(t)] = [E(z(t) \mid F_t^Y), E(x(t) \mid F_t^Y)]$$

of the system state $[z(t), x(t)]$ with respect to the σ - algebra F_t^Y generated by the observation process $Y(t)$ in the interval $[t_0, t]$. The matrix function

$$P(t) = E[([z(t), x(t)] - [m_z(t), m_x(t)]) \times$$

$$([z(t), x(t)] - [m_z(t), m_x(t)])^T \mid F_t^Y]$$

is the estimation error variance for this reformulated problem.

The obtained filtering system includes two equations, (4) (or (5)) and (1), for the partially measured state $[z(t), x(t)]$ and an equation (6) for the observations $y(t)$, where $z(t)$ is a measured polynomial state with polynomial multiplicative noise, $x(t)$ is an unmeasured linear state, and $y(t)$ is a linear observation process directly measuring the state $z(t)$. Hence, the optimal filter for the polynomial system states with unmeasured linear part and polynomial multiplicative noise over linear observations, obtained in (29), can be applied to solving this problem. Indeed, as follows from the general optimal filtering theory (see (31)), the optimal filtering equations take the following particular form for the system (5), (1), (6)

$$dm(t) = E(\bar{f}(x, t) \mid F_t^Y)dt+ \tag{7}$$

$$P(t)[I, 0]^T(B(t)B^T(t))^{-1}(dy(t) - m_z(t)dt),$$
$$dP(t) = (E((([z(t), x(t)] - m(t))(\bar{f}(x, t))^T \mid F_t^Y)+ \tag{8}$$
$$E(\bar{f}(x, t)([z(t), x(t)] - m(t))^T) \mid F_t^Y)+$$
$$E(\bar{g}(x, t)\bar{g}^T(x, t) \mid F_t^Y)-$$
$$P(t)[I, 0]^T(B(t)B^T(t))^{-1}[I, 0]P(t))dt+$$
$$E((([z(t), x(t)] - m(t))([z(t), x(t)] - m(t)) \times$$
$$([z(t), x(t)] - m(t))^T \mid F_t^Y) \times$$
$$[I, 0]^T(B(t)B^T(t))^{-1}(dy(t) - m_z(t)dt),$$

where $\bar{f}(x, t) = [f(x, t), a_0(t) + a(t)x(t)]$ is the polynomial drift term and $\bar{g}(x, t) = [g(x, t), b(t)]$ is the polynomial diffusion (multiplicative noise) term in the entire system of the state equations (4), (1), and the last term should be understood as a 3D tensor (under the expectation sign) convoluted with a vector, which yields a matrix. The matrix $[I, 0]$ is the $m \times (n + m)$ matrix composed of the $m \times m$-dimensional identity matrix and $m \times n$-dimensional zero matrix. The equations (7), (8) should be complemented with the initial conditions $m(t_0) = [m_z(t_0), m_x(t_0)] = E([z_0, x_0] \mid F_{t_0}^Y)$ and $P(t_0) = E[([z_0, x_0] - m(t_0)([z_0, x_0] - m(t_0)^T \mid F_{t_0}^Y]$.

The result given in (27; 29) claims that a closed system of the filtering equations can be obtained for the state $[z(t), x(t)]$ over the observations $y(t)$, in view of the polynomial properties

of the functions in the right-hand side of the equation (4). Indeed, since the observation matrix in (6) is the identity one, i.e., invertible, and the initial condition z_0 is assumed conditionally Gaussian with respect to observations, the random variable $z(t) - m_z(t)$ is conditionally Gaussian with respect to the observation process $y(t)$ for any $t \geq t_0$ ((27; 29)). Moreover, the random variable $x(t) - m_x(t)$ is also conditionally Gaussian with respect to the observation process $y(t)$ for any $t \geq t_0$, because $x(t)$ is Gaussian, in view of (1), and $y(t)$ depends only on $z(t)$, in view of (6), and the assumed conditional Gaussianity of the initial random vector z_0 ((26; 29)). Hence, the entire random vector $[z(t), x(t)] - m(t)$ is conditionally Gaussian with respect to the observation process $y(t)$ for any $t \geq t_0$, and the following considerations outlined in (26; 27; 29) are applicable.

First, since the random variable $x(t) - m(t)$ is conditionally Gaussian, the conditional third moment $E((([z(t), x(t)] - m(t))([z(t), x(t)] - m(t))([z(t), x(t)] - m(t))^T \mid F_t^Y)$ with respect to observations, which stands in the last term of the equation (8), is equal to zero, because the process $[z(t), x(t)] - m(t)$ is conditionally Gaussian. Thus, the entire last term in (8) is vanished and the following variance equation is obtained

$$
dP(t) = (E((([z(t), x(t)] - m(t))(\bar{f}(x, t))^T \mid F_t^Y) +
$$
$$
E(\bar{f}(x, t)([z(t), x(t)] - m(t))^T) \mid F_t^Y) + \tag{9}
$$
$$
E(\bar{g}(x, t)\bar{g}^T(x, t) \mid F_t^Y) -
$$
$$
P(t)[I, 0]^T (\bar{B}(t)\bar{B}^T(t))^{-1}[I, 0]P(t))dt,
$$

with the initial condition $P(t_0) = E[([z_0, x_0] - m(t_0)([z_0, x_0] - m(t_0))^T \mid F_{t_0}^Y]$.

Second, if the functions $\bar{f}(x, t)$ and $\bar{g}(x, t)$ are polynomial functions of the state x with time-dependent coefficients, the expressions of the terms $E(\bar{f}(x, t) \mid F_t^Y)$ in (4) and $E(([z(t), x(t)] - m(t))\bar{f}^T(x, t)) \mid F_t^Y)$ and $E(\bar{g}(x, t)\bar{g}^T(x, t) \mid F_t^Y)$, which should be calculated to obtain a closed system of filtering equations (see (31)), would also include only polynomial terms of x. Then, those polynomial terms can be represented as functions of $m(t)$ and $P(t)$ using the following property of Gaussian random variable $[z(t), x(t)] - m(t)$: all its odd conditional moments, $m_1 = E[([z(t), x(t)] - m(t)) \mid Y(t)], m_3 = E[([z(t), x(t)] - m(t))^3 \mid Y(t)], m_5 = E[([z(t), x(t)] - m(t))^5 \mid Y(t)], \dots$ are equal to 0, and all its even conditional moments $m_2 = E[([z(t), x(t)] - m(t))^2 \mid Y(t)], m_4 = E[([z(t), x(t)] - m(t))^4 \mid Y(t)], \dots$ can be represented as functions of the variance $P(t)$. For example, $m_2 = P, m_4 = 3P^2, m_6 = 15P^3, \dots$ etc. After representing all polynomial terms in (7) and (9), that are generated upon expressing $E(\bar{f}(x, t) \mid F_t^Y)$, $E(([z(t), x(t)] - m(t))\bar{f}^T(x, t)) \mid F_t^Y)$, and $E(\bar{g}(x, t)\bar{g}^T(x, t) \mid F_t^Y)$, as functions of $m(t)$ and $P(t)$, a closed form of the filtering equations would be obtained. The corresponding representations of $E(f(x, t) \mid F_t^Y)$, $E(([z(t), x(t)] - m(t))(f(x, t))^T \mid F_t^Y)$ and $E(\bar{g}(x, t)\bar{g}^T(x, t) \mid F_t^Y)$ have been derived in (24–27; 29) for certain polynomial functions $f(x, t)$ and $g(x, t)$.

In the next example section, a closed form of the filtering equations will be obtained for a particular case of a scalar third degree polynomial function $h(x, t)$ in the equation (2). It should be noted, however, that application of the same procedure would result in designing a closed system of the filtering equations for any polynomial function $h(x, t) \in R^n$ in (2).

4. Example: Third Degree Sensor Filtering Problem

This section presents an example of designing the optimal filter for a linear state over third degree polynomial observations, reducing it to the optimal filtering problem for a second degree

polynomial state with partially measured linear part and second degree polynomial multiplicative noise over linear observations, where a conditionally Gaussian state initial condition is additionally assumed.

Let the unmeasured scalar state $x(t)$ satisfy the trivial linear equation

$$dx(t) = dt + dw_1(t), \quad x(0) = x_0, \tag{10}$$

and the observation process be given by the scalar third degree sensor equation

$$dy(t) = (x^3(t) + x(t))dt + dw_2(t), \tag{11}$$

where $w_1(t)$ and $w_2(t)$ are standard Wiener processes independent of each other and of a Gaussian random variable x_0 serving as the initial condition in (10). The filtering problem is to find the optimal estimate for the linear state (10), using the third degree sensor observations (11).

Let us reformulate the problem, introducing the stochastic process $z(t) = h(x, t) = x^3(t) + x(t)$. Using the Ito formula (see (31)) for the stochastic differential of the cubic function $h(x, t) = x^3(t) + x(t)$, where $x(t)$ satisfies the equation (10), the following equation is obtained for $z(t)$

$$dz(t) = (1 + 3x(t) + 3x^2(t))dt + (3x^2(t) + 1)dw_1(t), \; z(0) = z_0. \tag{12}$$

Here, $\frac{\partial h(x,t)}{\partial x} = 3x^2(t) + 1$, $\frac{1}{2}\frac{\partial^2 h(x,t)}{\partial x^2} = 3x(t)$, and $\frac{\partial h(x,t)}{\partial t} = 0$; therefore, $f(x, t) = 1 + 3x(t) + 3x^2(t)$ and $g(x, t) = 3x^2(t) + 1$. The initial condition $z_0 \in R$ is considered a conditionally Gaussian random vector with respect to observations (see the paragraph following (4) for details). This assumption is quite admissible in the filtering framework, since the real distributions of $x(t)$ and $z(t)$ are unknown. In terms of the process $z(t)$, the observation equation (11) takes the form

$$dy(t) = z(t)dt + dw_2(t). \tag{13}$$

The obtained filtering system includes two equations, (12) and (10), for the partially measured state $[z(t), x(t)]$ and an equation (13) for the observations $y(t)$, where $z(t)$ is a completely measured quadratic state with multiplicative quadratic noise, $x(t)$ is an unmeasured linear state, and $y(t)$ is a linear observation process directly measuring the state $z(t)$. Hence, the designed optimal filter can be applied for solving this problem. The filtering equations (7),(9) take the following particular form for the system (12),(10),(13)

$$dm_1(t) = (1 + 3m_2(t) + 3m_2^2(t) + 3P_{22}(t))dt+ \tag{14}$$

$$P_{11}(t)[dy(t) - m_1(t)dt],$$
$$dm_2(t) = 1 + P_{12}(t)[dy(t) - m_1(t)dt], \tag{15}$$

with the initial conditions $m_1(0) = E(x_0 \mid y(0)) = m_{10}$ and $m_2(0) = E(x_0^3 \mid y(0)) = m_{20}$,

$$\dot{P}_{11}(t) = 12(P_{12}(t)m_2(t)) + 6P_{12}(t) + 27P_{22}^2(t)+ \tag{16}$$

$$54P_{22}(t)m_2^2(t) + 9m_2^4(t) + 6P_{22}(t) + 6m_2^2 + 1 - P_{11}^2(t),$$
$$\dot{P}_{12}(t) = 6(P_{22}(t)m_2(t)) + 3P_{22}(t)+ \tag{17}$$
$$3(m_2^2(t) + P_{22}(t)) + 1 - P_{11}(t)P_{12}(t),$$

$$\dot{P}_{22}(t) = 1 - P_{12}^2(t), \tag{18}$$

with the initial condition $P(0) = E(([x_0, z_0]^T - m(0))([x_0, z_0]^T - m(0))^T \mid y(0)) = P_0$. Here, $m_1(t)$ is the optimal estimate for the state $z(t) = x^3(t) + x(t)$ and $m_2(t)$ is the optimal estimate for the state $x(t)$.

Numerical simulation results are obtained solving the systems of filtering equations (14)–(18). The obtained values of the state estimate $m_2(t)$ satisfying the equation (15) are compared to the real values of the state variable $x(t)$ in (10).

For the filter (14)–(18) and the reference system (12),(10),(13) involved in simulation, the following initial values are assigned: $x_0 = z_0 = 0$, $m_2(0) = 10$, $m_1(0) = 1000$, $P_{11}(0) = 15$, $P_{12}(0) = 3$, $P_{22}(0) = 1$. Gaussian disturbances $dw_1(t)$ and $dw_2(t)$ are realized using the built-in MatLab white noise functions. The simulation interval is $[0, 0.05]$.
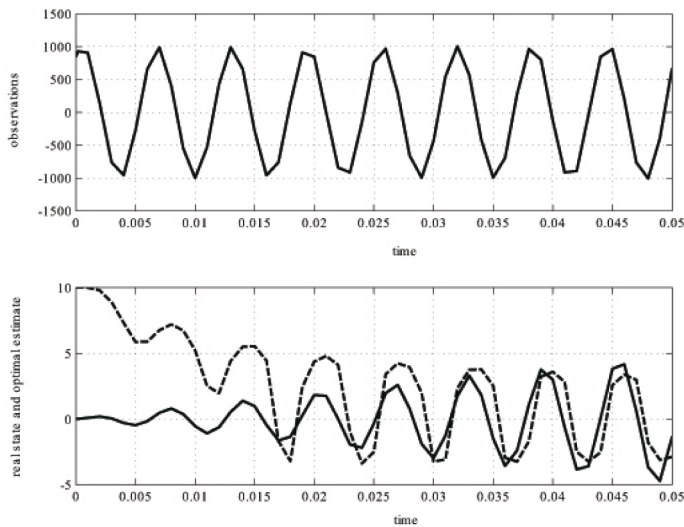

Fig. 1. **Above.** Graph of the observation process $y(t)$ in the interval $[0, 0.05]$. **Below.** Graphs of the real state $x(t)$ (solid line) and its optimal estimate $m_2(t)$ (dashed line) in the interval $[0, 0.05]$.

Figure 1 shows the graphs of the reference state variable $x(t)$ (10) and its optimal estimate $m_2(t)$ (15), as well as the observation process $y(t)$ (11), in the entire simulation interval from $t_0 = 0$ to $T = 0.05$. It can be observed that the optimal estimate given by (14)–(18) converges to the real state (10) very rapidly, in spite of a considerable error in the initial conditions, $m_2(0) - x_0 = 10$, $m_1(0) - z_0 = 1000$, and very noisy observations which do not even reproduce the shape of $z(t) = x^3(t) + x(t)$. Moreover, the estimated signal $x(t)$ itself is a time-shifted Wiener process, i.e., the integral of a white Gaussian noise, which makes the filtering problem even more difficult. It should also be noted that the extended Kalman filter (EKF) approach fails for the system (10),(11), since the linearized value $\partial z / \partial x = 3x^2(t) + 1$ at zero is the unit-valued constant, therefore, the observation process would consist of pure noise.

Thus, it can be concluded that the obtained optimal filter (14)–(18) solves the optimal third degree sensor filtering problem for the system (10),(11) and yields a really good estimate of the unmeasured state in presence of quite complicated observation conditions. Subsequent discussion of the obtained results can be found in Conclusions.

5. Conclusions

This paper presents the optimal filter for linear system states over nonlinear polynomial observations. It is shown that the optimal filter can be obtained in a closed form for any polynomial function in the observation equation. Based on the optimal filter for a bilinear state, the optimal solution is obtained for the optimal third degree sensor filtering problem, assuming a conditionally Gaussian initial condition for the third degree state. This assumption is quite admissible in the filtering framework, since the real distributions of the first and third degree states are unknown. The resulting filter yields a reliable and rapidly converging estimate, in spite of a significant difference in the initial conditions between the state and estimate and very noisy observations, in the situation where the unmeasured state itself is a time-shifted Wiener process and the extended Kalman filter (EKF) approach fails. Although this conclusion follows from the developed theory, the numerical simulation serves as a convincing illustration.

6. References

 [1] Kushner HJ. On differential equations satisfied by conditional probability densities of Markov processes. *SIAM J. Control* 1964; **12**:106-119.
 [2] Duncan TE. *Probability densities for diffusion processes with applications to nonlinear filtering theory*. Technical report, PhD thesis, Standford, CA, USA, 1967.
 [3] Mortensen RE. *Optimal control of continuous-time stochastic systems*. Technical report, PhD thesis, University of California, Berkeley, CA, USA, 1966.
 [4] Zakai M. On the optimal filtering of diffusion processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 1969; **11**: 230Ű243.
 [5] Liptser RS, Shiryayev AN. *Statistics of Random Processes. Vol. I: General Theory*. Springer, 2000 (1st. Ed., 1974).
 [6] Kallianpur G. *Stochastic Filtering Theory*. Springer, 1980.
 [7] Kalman RE, Bucy RS. New results in linear filtering and prediction theory. *ASME Trans., Part D (J. of Basic Engineering)* 1961; **83**: 95-108.
 [8] Wonham WM. Some applications of stochastic differential equations to nonlinear filtering. *SIAM J. Control* 1965; **2**: 347-369.
 [9] Benes VE. Exact finite-dimensional filters for certain diffusions with nonlinear drift. *Stochastics* 1981; **5**: 65-92.
[10] Yau SST. Finite-dimensional filters with nonlinear drift I: a class of filters including both Kalman-Bucy and Benes filters. *J. Math. Systems, Estimation, and Control* 1994; **4**: 181-203.
[11] Germani A, Manes P, Palumbo P. Linear filtering for bilinear stochastic differential systems with unknown inputs, *IEEE Transactions on Automatic Control* 2002; **47**: 1726-1730.
[12] Shin DR, Verriest E. Optimal access control of simple integrated networks with incomplete observations, *Proc. American Control Conf.* 1994; Baltimore, MD, USA, 3487-3488.
[13] Zhang WH, Chen BS, Tseng CS. Robust H_∞ filtering for nonlinear stochastic systems, *IEEE Transactions on Signal Processing* 2005; **53**: 589–598.

[14] Wang Z. Filtering on nonlinear time-delay stochastic systems, *Automatica* 2003, **39**: 101-109.

[15] Shi P. Filtering on sampled-data systems with parametric uncertainty, *IEEE Transactions on Automatic Control* 1998; **43**: 1022-1027.

[16] Xu S., van Dooren PV. Robust H_∞–filtering for a class of nonlinear systems with state delay and parameter uncertainty, *Int. J. Control*, 2002; **75**: 766–774.

[17] Mahmoud M, Shi P. Robust Kalman filtering for continuous time-lag systems with Markovian jump parameters. *IEEE Transactions on Circuits and Systems* 2003; **50**: 98-105.

[18] Sheng J, Chen T, Shah SL. Optimal filtering for multirate systems. *IEEE Transactions on Circuits and Systems* 2005; **52**: 228-232.

[19] Sheng J. Optimal filtering for multirate systems based on lifted models. *Proc. American Control Conf.* 2005; Portland, OR, USA, 3459-3461.

[20] Gao H, Lam J, Xie L, Wang C. New approach to mixed H_2/H_∞–filltering for polytopic discrete-time systems. *IEEE Transactions on Signal Processing* 2005; **53**: 3183–3192.

[21] Boukas EK. Stabilization of stochastic nonlinear hybrid systems, *International Journal of Innovative Computing, Information and Control* 2005; **1**: 131Ű-141.

[22] Zhang H, Basin MV, Skliar M. Optimal state estimation for continuous, stochastic, state-space system with hybrid measurements, *International Journal of Innovative Computing, Information and Control* 2006; **2**: 863–874.

[23] Jeong CS, Yaz E, Bahakeem A, Yaz Y. Nonlinear observer design with general criteria, *International Journal of Innovative Computing, Information and Control* 2006; **2**: 693Ű-704.

[24] Basin MV. On Optimal filtering for polynomial system states. *ASME Trans. J. Dynamic Systems, Measurement, and Control* 2003; **125**: 123-125.

[25] Basin MV, Alcorta-Garcia MA. Optimal filtering and control for third degree polynomial systems. *Dynamics of Continuous, Discrete, and Impulsive Systems* 2003; **10B**: 663-680.

[26] Basin MV, Skliar M. Optimal filtering for partially measured polynomial system states. *Proc. American Control Conf.* 2005; Portland, OR, USA, 4022-4027.

[27] Basin MV, Perez J, Skliar M. Optimal filtering for polynomial system states with polynomial multiplicative noise, *International J. Robust and Nonlinear Control* 2006; **16**: 287-298.

[28] Basin MV, Perez J, Martinez-Zuniga R. Optimal filtering for nonlinear polynomial systems over linear observations with delay, *International Journal of Innovative Computing, Information and Control* 2006; **2**: 357–370.

[29] Basin MV, Perez J, Skliar M. Optimal filtering for polynomial systems with patially measured states and multiplicative noises, *Proc. 45th IEEE Conf. on Decision and Control* 2006; San Diego, CA, USA, 4169-4174.

[30] Hazewinkel M, Marcus SI, Sussmann HJ. Nonexistence of exact finite-dimensional filters for conditional statistics of the cubic sensor problem, *Systems and Control Letters* 1983; **5**: 331-340.

[31] Pugachev VS, Sinitsyn IN. *Stochastic Systems: Theory and Applications*. World Scientific, 2001.

[32] Pugachev VS. *Probability Theory and Mathematical Statistics for Engineers*. Pergamon, 1984.

[33] Tucker HG. *A Graduate Course in Probability*. Academic Press, 1967.

The stochastic matched filter and its applications to detection and de-noising

Philippe Courmontagne
ISEN Toulon - IM2NP
FRANCE

1. Introduction

In several domains of signal processing, such as detection or de-noising, it may be interesting to provide a second-moment characterization of a noise-corrupted signal in terms of uncorrelated random variables. Doing so, the noisy data could be described by its expansion into a weighted sum of known vectors by uncorrelated random variables. Depending on the choice of the basis vectors, some random variables are carrying more signal of interest informations than noise ones. This is the case, for example, when a signal disturbed by a white noise is expanded using the Karhunen-Loève expansion (Karhunen, 1946; Loève, 1955). In these conditions, it is possible either to approximate the signal of interest considering, for the reconstruction, only its associated random variables, or to detect a signal in a noisy environment with an analysis of the random variable power. The purpose of this chapter is to present such an expansion, available for both the additive and multiplicative noise cases, and its application to detection and de-noising. This noisy random signal expansion is known as the stochastic matched filter (Cavassilas, 1991), where the basis vectors are chosen so as to maximize the signal to noise ratio after processing.

At first, we recall some general considerations on a random 1-D discrete-time signal expansion in section 2. In particular, we study the approximation error and the second order statistics of the signal approximation. Then, in section 3, we describe the stochastic matched filter theory for 1-D discrete-time signals and its extension to 2-D discrete-space signals. We finish this section with a study on two different noise cases: the white noise case and the speckle noise case. In the next section, we present the stochastic matched filter in a de-noising context and we briefly discuss the estimator bias. Then, the de-noising being performed by a limitation to order Q of the noisy data expansion, we propose to determine this truncature order using a mean square error criterion. Experimental results on synthetic and real data are given and discussed to evaluate the performances of such an approach. In section 5, we describe the stochastic matched filter in a detection context and we confront the proposed method with signals resulting from underwater acoustics. Finally, some concluding remarks are given in section 6.

2. Random signal expansion

2.1 1-D discrete-time signals

Let $\mathbf{S}$ be a zero mean, stationary, discrete-time random signal, made of M successive samples and let $\{s_1, s_2, \ldots, s_M\}$ be a zero mean, uncorrelated random variable sequence, i.e.:

$$E\{s_n s_m\} = E\left\{s_m^2\right\} \delta_{n,m},\tag{1}$$

where $\delta_{n,m}$ denotes the Kronecker symbol.

It is possible to expand signal $\mathbf{S}$ into series of the form:

$$\mathbf{S} = \sum_{m=1}^{M} s_m \mathbf{\Psi_m},\tag{2}$$

where $\{\mathbf{\Psi_m}\}_{m=1\ldots M}$ corresponds to a M-dimensional deterministic basis. Vectors $\mathbf{\Psi_m}$ are linked to the choice of random variables sequence $\{s_m\}$, so there are many decompositions (2).

These vectors are determined by considering the mathematical expectation of the product of s_m with the random signal $\mathbf{S}$. It comes:

$$\mathbf{\Psi_m} = \frac{1}{E\{s_m^2\}} E\{s_m \mathbf{S}\}.\tag{3}$$

Classically and using a M-dimensional deterministic basis $\{\mathbf{\Phi_m}\}_{m=1\ldots M}$, the random variables s_m can be expressed by the following relation:

$$s_m = \mathbf{S}^T \mathbf{\Phi_m}.\tag{4}$$

The determination of these random variables depends on the choice of the basis $\{\mathbf{\Phi_m}\}_{m=1\ldots M}$. We will use a basis, which provides the uncorrelation of the random variables. Using relations (1) and (4), we can show that the uncorrelation is ensured, when vectors $\mathbf{\Phi_m}$ are solution of the following quadratic form:

$$\mathbf{\Phi_m}^T \mathbf{\Gamma_{SS}} \mathbf{\Phi_n} = E\left\{s_m^2\right\} \delta_{n,m},\tag{5}$$

where $\mathbf{\Gamma_{SS}}$ represents the signal covariance.

There is an infinity of sets of vectors obtained by solving the previous equation. Assuming that a basis $\{\mathbf{\Phi_m}\}_{m=1\ldots M}$ is chosen, we can find random variables using relation (4). Taking into account relations (3) and (4), we obtain as new expression for $\mathbf{\Psi_m}$:

$$\mathbf{\Psi_m} = \frac{1}{E\{s_m^2\}} \mathbf{\Gamma_{SS}} \mathbf{\Phi_m}.\tag{6}$$

Furthermore, using relations (5) and (6), we can show that vectors $\mathbf{\Psi_m}$ and $\mathbf{\Phi_m}$ are linked by the following bi-orthogonality relation:

$$\mathbf{\Phi_m}^T \mathbf{\Psi_n} = \delta_{n,m}.\tag{7}$$

2.2 Approximation error

When the discrete sum, describing the signal expansion (relation (2)), is reduced to Q random variables s_m, only an approximation $\widetilde{\mathbf{S}}_{\mathbf{Q}}$ of the signal is obtained:

$$\widetilde{\mathbf{S}}_{\mathbf{Q}} = \sum_{m=1}^{Q} s_m \mathbf{\Psi_m}. \tag{8}$$

To evaluate the error induced by the restitution, let us consider the mean square error ϵ between signal $\mathbf{S}$ and its approximation $\widetilde{\mathbf{S}}_{\mathbf{Q}}$:

$$\epsilon = E\left\{ \left\| \mathbf{S} - \widetilde{\mathbf{S}}_{\mathbf{Q}} \right\|^2 \right\}, \tag{9}$$

where $\|.\|$ denotes the classical Euclidean norm.
Considering the signal variance σ_S^2, it can be easily shown that:

$$\epsilon = \sigma_S^2 - \sum_{m=1}^{Q} E\left\{ s_m^2 \right\} \|\mathbf{\Psi_m}\|^2, \tag{10}$$

which corresponds to:

$$\epsilon = \sigma_S^2 - \sum_{m=1}^{Q} \frac{\mathbf{\Phi_m}^T \mathbf{\Gamma_{SS}}^2 \mathbf{\Phi_m}}{\mathbf{\Phi_m}^T \mathbf{\Gamma_{SS}} \mathbf{\Phi_m}}. \tag{11}$$

When we consider the whole s_m sequence (i.e. Q equal to M), the approximation error ϵ is weak, and coefficients given by the quadratic form ratio:

$$\frac{\mathbf{\Phi_m}^T \mathbf{\Gamma_{SS}}^2 \mathbf{\Phi_m}}{\mathbf{\Phi_m}^T \mathbf{\Gamma_{SS}} \mathbf{\Phi_m}}$$

are carrying the signal power.

2.3 Second order statistics

The purpose of this section is the determination of the $\widetilde{\mathbf{S}}_{\mathbf{Q}}$ autocorrelation and spectral power density. Let $\mathbf{\Gamma}_{\widetilde{S}_Q \widetilde{S}_Q}$ be the $\widetilde{S}_Q$ autocorrelation, we have:

$$\Gamma_{\widetilde{S}_Q \widetilde{S}_Q}[p] = E\left\{ \widetilde{S}_Q[q] \widetilde{S}_Q^*[p-q] \right\}. \tag{12}$$

Taking into account relation (8) and the uncorrelation of random variables s_m, it comes:

$$\Gamma_{\widetilde{S}_Q \widetilde{S}_Q}[p] = \sum_{m=1}^{Q} E\left\{ s_m^2 \right\} \Psi_m[q] \Psi_m^*[p-q], \tag{13}$$

which leads to, summing all elements of the previous relation:

$$\sum_{q=1}^{M} \Gamma_{\widetilde{S}_Q \widetilde{S}_Q}[p] = \sum_{q=1}^{M} \sum_{m=1}^{Q} E\left\{ s_m^2 \right\} \Psi_m[q] \Psi_m^*[p-q]. \tag{14}$$

So, we have:

$$\Gamma_{\widetilde{S}_Q \widetilde{S}_Q}[p] = \frac{1}{M} \sum_{m=1}^{Q} E\left\{s_m^2\right\} \sum_{q=1}^{M} \Psi_m[q] \Psi_m^*[p-q], \qquad (15)$$

which corresponds to:

$$\Gamma_{\widetilde{S}_Q \widetilde{S}_Q}[p] = \frac{1}{M} \sum_{m=1}^{Q} E\left\{s_m^2\right\} \Gamma_{\Psi_m \Psi_m}[p]. \qquad (16)$$

In these conditions, the $\widetilde{\mathbf{S}}_\mathbf{Q}$ spectral power density is equal to:

$$\gamma_{\widetilde{S}_Q \widetilde{S}_Q}(\nu) = \frac{1}{M} \sum_{m=1}^{Q} E\left\{s_m^2\right\} \gamma_{\Psi_m \Psi_m}(\nu). \qquad (17)$$

3. The Stochastic Matched Filter expansion

Detecting or de-noising a signal of interest $\mathbf{S}$, corrupted by an additive or multiplicative noise $\mathbf{N}$ is a usual signal processing problem. We can find in the literature several processing methods for solving this problem. One of them is based on a stochastic extension of the matched filter notion (Cavassilas, 1991; Chaillan et al., 2007; 2005). The signal of interest pattern is never perfectly known, so it is replaced by a random signal allowing a new formulation of the signal to noise ratio. The optimization of this ratio leads to design a bench of filters and regrouping them strongly increases the signal to noise ratio.

3.1 1-D discrete-time signals: signal-independent additive noise case

Let us consider a noise-corrupted signal $\mathbf{Z}$, made of M successive samples and corresponding to the superposition of a signal of interest $\mathbf{S}$ with a colored noise $\mathbf{N}$. If we consider the signal and noise variances, σ_S^2 and σ_N^2, we have:

$$\mathbf{Z} = \sigma_S \mathbf{S_0} + \sigma_N \mathbf{N_0}, \qquad (18)$$

with $E\left\{\mathbf{S_0}^2\right\} = 1$ and $E\left\{\mathbf{N_0}^2\right\} = 1$. In the previous relation, reduced signals $\mathbf{S_0}$ and $\mathbf{N_0}$ are assumed to be independent, stationary and with zero-mean.

It is possible to expand noise-corrupted signal $\mathbf{Z}$ into a weighted sum of known vectors $\mathbf{\Psi_m}$ by uncorrelated random variables z_m, as described in relation (2). These uncorrelated random variables are determined using the scalar product between noise-corrupted signal $\mathbf{Z}$ and deterministic vectors $\mathbf{\Phi_m}$ (see relation (4)). In order to determine basis $\left\{\mathbf{\Phi_m}\right\}_{m=1\ldots M}$, let us describe the matched filter theory. If we consider a discrete-time, stationary, known input signal $\mathbf{s}$, made of M successive samples, corrupted by an ergodic reduced noise $\mathbf{N_0}$, the matched filter theory consists of finding an impulse response $\mathbf{\Phi}$, which optimizes the signal to noise ratio ρ. Defined as the ratio of the square of signal amplitude to the square of noise amplitude, ρ is given by:

$$\rho = \frac{|\mathbf{s}^T \mathbf{\Phi}|^2}{E\left\{|\mathbf{N_0}^T \mathbf{\Phi}|^2\right\}}. \qquad (19)$$

When the signal is not deterministic, i.e. a random signal $\mathbf{S_0}$, this ratio becomes (Cavassilas, 1991):

$$\rho = \frac{E\left\{|\mathbf{S_0}^T \mathbf{\Phi}|^2\right\}}{E\left\{|\mathbf{N_0}^T \mathbf{\Phi}|^2\right\}}, \qquad (20)$$

which leads to:

$$\rho = \frac{\boldsymbol{\Phi}^T \boldsymbol{\Gamma}_{\mathbf{S_0 S_0}} \boldsymbol{\Phi}}{\boldsymbol{\Phi}^T \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}} \boldsymbol{\Phi}}, \tag{21}$$

where $\boldsymbol{\Gamma}_{\mathbf{S_0 S_0}}$ and $\boldsymbol{\Gamma}_{\mathbf{N_0 N_0}}$ represent signal and noise reduced covariances respectively.

Relation (21) corresponds to the ratio of two quadratic forms. It is a Rayleigh quotient. For this reason, the signal to noise ratio ρ is maximized when the impulse response $\boldsymbol{\Phi}$ corresponds to the eigenvector $\boldsymbol{\Phi_1}$ associated to the greatest eigenvalue λ_1 of the following generalized eigenvalue problem:

$$\boldsymbol{\Gamma}_{\mathbf{S_0 S_0}} \boldsymbol{\Phi_m} = \lambda_m \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}} \boldsymbol{\Phi_m}. \tag{22}$$

Let us consider the signal and noise expansions, we have:

$$\begin{cases} \mathbf{S_0} = \displaystyle\sum_{m=1}^{M} s_m \boldsymbol{\Psi_m} \\ \mathbf{N_0} = \displaystyle\sum_{m=1}^{M} \eta_m \boldsymbol{\Psi_m} \end{cases}, \tag{23}$$

where the random variables defined by:

$$\begin{cases} s_m = \boldsymbol{\Phi_m}^T \mathbf{S_0} \\ \eta_m = \boldsymbol{\Phi_m}^T \mathbf{N_0} \end{cases} \tag{24}$$

are not correlated:

$$\begin{cases} \boldsymbol{\Phi_m}^T \boldsymbol{\Gamma}_{\mathbf{S_0 S_0}} \boldsymbol{\Phi_n} = E\left\{s_m^2\right\} \delta_{n,m} \\ \boldsymbol{\Phi_m}^T \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}} \boldsymbol{\Phi_n} = E\left\{\eta_m^2\right\} \delta_{n,m} \end{cases}. \tag{25}$$

After a normalization step; it is possible to rewrite relations (25) as follows:

$$\begin{cases} \boldsymbol{\Phi_m}^T \boldsymbol{\Gamma}_{\mathbf{S_0 S_0}} \boldsymbol{\Phi_n} = \dfrac{E\left\{s_m^2\right\}}{E\left\{\eta_m^2\right\}} \delta_{n,m} \\ \boldsymbol{\Phi_m}^T \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}} \boldsymbol{\Phi_n} = \delta_{n,m} \end{cases}. \tag{26}$$

Let $\mathbf{P}$ be a matrix made up of column vectors $\boldsymbol{\Phi_m}$, i.e.:

$$\mathbf{P} = (\boldsymbol{\Phi_1}, \boldsymbol{\Phi_2}, \dots, \boldsymbol{\Phi_M}). \tag{27}$$

In these conditions, it comes:

$$\mathbf{P}^T \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}} \mathbf{P} = \mathbf{I}, \tag{28}$$

where $\mathbf{I}$ corresponds to the identity matrix.

This leads to:

$$\begin{aligned} & \left(\mathbf{P}^T \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}} \mathbf{P}\right)^{-1} = \mathbf{I} \\ \Leftrightarrow \quad & \mathbf{P}^{-1} \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}}^{-1} \mathbf{P}^{-T} = \mathbf{I} \\ \Leftrightarrow \quad & \mathbf{P}^T = \mathbf{P}^{-1} \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}}^{-1}. \end{aligned} \tag{29}$$

Let $\mathbf{D}$ be the following diagonal matrix:

$$\mathbf{D} = \begin{pmatrix} E\left\{s_1^2\right\}/E\left\{\eta_1^2\right\} & 0 & \dots & \dots & 0 \\ 0 & E\left\{s_2^2\right\}/E\left\{\eta_2^2\right\} & 0 & \dots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & E\left\{s_{M-1}^2\right\}/E\left\{\eta_{M-1}^2\right\} & 0 \\ 0 & \dots & \dots & 0 & E\left\{s_M^2\right\}/E\left\{\eta_M^2\right\} \end{pmatrix}. \tag{30}$$

It comes:

$$\mathbf{P}^T \boldsymbol{\Gamma}_{\mathbf{S_0 S_0}} \mathbf{P} = \mathbf{D}, \tag{31}$$

which corresponds to, taking into account relation (29):

$$\mathbf{P}^{-1} \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}}{}^{-1} \boldsymbol{\Gamma}_{\mathbf{S_0 S_0}} \mathbf{P} = \mathbf{D}$$
$$\Leftrightarrow \quad \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}}{}^{-1} \boldsymbol{\Gamma}_{\mathbf{S_0 S_0}} \mathbf{P} = \mathbf{PD}$$
$$\Leftrightarrow \quad \boldsymbol{\Gamma}_{\mathbf{S_0 S_0}} \mathbf{P} = \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}} \mathbf{PD}, \tag{32}$$

which leads to:

$$\boldsymbol{\Gamma}_{\mathbf{S_0 S_0}} \boldsymbol{\Phi_m} = \frac{E\left\{s_m^2\right\}}{E\left\{\eta_m^2\right\}} \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}} \boldsymbol{\Phi_m}. \tag{33}$$

This last equation shows, on the one hand, that λ_m equals $E\left\{s_m^2\right\} / E\left\{\eta_m^2\right\}$ and, on the other hand, that the only basis $\left\{\boldsymbol{\Phi_m}\right\}_{m=1...M}$ allowing the simultaneous uncorrelation of the random variables coming from the signal and the noise is made up of vectors $\boldsymbol{\Phi_m}$ solution of the generalized eigenvalue problem (22).

We have $E\left\{\eta_m^2\right\} = 1$ and $E\left\{s_m^2\right\} = \lambda_m$ when the eigenvectors $\boldsymbol{\Phi_m}$ are normalized as follows:

$$\boldsymbol{\Phi_m}^T \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}} \boldsymbol{\Phi_m} = 1, \tag{34}$$

In these conditions and considering relation (6), the deterministic vectors $\boldsymbol{\Psi_m}$ of the noise-corrupted signal expansion are given by:

$$\boldsymbol{\Psi_m} = \boldsymbol{\Gamma}_{\mathbf{N_0 N_0}} \boldsymbol{\Phi_m}. \tag{35}$$

In this context, the noise-corrupted signal expansion is expressed as follows:

$$\mathbf{Z} = \sum_{m=1}^{M} (\sigma_S s_m + \sigma_N \eta_m) \boldsymbol{\Psi_m}, \tag{36}$$

so that, the quadratic moment of the mth coefficient z_m of the noise-corrupted signal expansion is given by:

$$E\left\{z_m^2\right\} = E\left\{(\sigma_S s_m + \sigma_N \eta_m)^2\right\}, \tag{37}$$

which corresponds to:

$$\sigma_S^2 \lambda_m + \sigma_N^2 + \sigma_S \sigma_N \boldsymbol{\Phi_m}^T \left(\boldsymbol{\Gamma}_{\mathbf{S_0 N_0}} + \boldsymbol{\Gamma}_{\mathbf{N_0 S_0}}\right) \boldsymbol{\Phi_m} \tag{38}$$

Signal and noise being independent and one of them at least being zero mean, we can assume that the cross-correlation matrices, $\boldsymbol{\Gamma}_{\mathbf{S_0 N_0}}$ and $\boldsymbol{\Gamma}_{\mathbf{N_0 S_0}}$, are weak. In this condition, the signal to noise ratio ρ_m of component z_m corresponds to the native signal to noise ratio times eigenvalue λ_m:

$$\rho_m = \frac{\sigma_S^2}{\sigma_N^2} \lambda_m. \tag{39}$$

So, an approximation $\widetilde{\mathbf{S}}_{\mathbf{Q}}$ of the signal of interest (the filtered noise-corrupted signal) can be built by keeping only those components associated to eigenvalues greater than a certain threshold. In any case this threshold is greater than one.

3.2 Extension to 2-D discrete-space signals

We consider now a $M \times M$ pixels two-dimensional noise-corrupted signal, $\mathbf{Z}$, which corresponds to a signal of interest $\mathbf{S}$ disturbed by a noise $\mathbf{N}$. The two-dimensional extension of the theory developed in the previous section gives:

$$\mathbf{Z} = \sum_{m=1}^{M^2} z_m \mathbf{\Psi_m}, \tag{40}$$

where $\{\mathbf{\Psi_m}\}_{m=1...M^2}$ is a M^2-dimensional basis of $M \times M$ matrices.

Random variables z_m are determined, using a M^2-dimensional basis $\{\mathbf{\Phi_m}\}_{m=1...M^2}$ of $M \times M$ matrices, as follows:

$$z_m = \sum_{p,q=1}^{M} Z[p,q]\Phi_m[p,q]. \tag{41}$$

These random variables will be not correlated, if matrices $\mathbf{\Phi_m}$ are solution of the two-dimensional extension of the generalized eigenvalue problem (22):

$$\sum_{p_1,q_1=1}^{M} \Gamma_{S_0 S_0}[p_1 - p_2, q_1 - q_2]\Phi_m[p_1,q_1] = \lambda_n \sum_{p_1,q_1=1}^{M} \Gamma_{N_0 N_0}[p_1 - p_2, q_1 - q_2]\Phi_m[p_1,q_1], \tag{42}$$

for all $p_2, q_2 = 1, \ldots, M$.

Assuming that $\mathbf{\Phi_m}$ are normalized as follows:

$$\sum_{p_1,p_2,q_1,q_2=1}^{M} \Gamma_{N_0 N_0}[p_1 - p_2, q_1 - q_2]\Phi_m[p_1,q_1]\Phi_m[p_2,q_2] = 1, \tag{43}$$

the basis $\{\mathbf{\Psi_m}\}_{m=1...M^2}$ derives from:

$$\Psi_m[p_1,q_1] = \sum_{p_2,q_2=1}^{M} \Gamma_{N_0 N_0}[p_1 - p_2, q_1 - q_2]\Phi_m[p_2,q_2]. \tag{44}$$

As for the 1-D discrete-time signals case, using such an expansion leads to a signal to noise ratio of component z_m equal to the native signal to noise ratio times eigenvalue λ_m (see relation (39)). So, all $\mathbf{\Phi_m}$ associated to eigenvalues λ_m greater than a certain level - in any case greater than one - can contribute to an improvement of the signal to noise ratio.

3.3 The white noise case

When $\mathbf{N}$ corresponds to a white noise, its reduced covariance is:

$$\Gamma_{N_0 N_0}[p - q] = \delta[p - q]. \tag{45}$$

Thus, the generalized eigenvalue problem (22) leading to the determination of vectors $\mathbf{\Phi_m}$ and associated eigenvalues is reduced to:

$$\mathbf{\Gamma_{S_0 S_0}}\mathbf{\Phi_m} = \lambda_m \mathbf{\Phi_m}. \tag{46}$$

In this context, we can show that basis vectors $\mathbf{\Psi_m}$ and $\mathbf{\Phi_m}$ are equal. Thus, in the particular case of a white noise, the stochastic matched filter theory is identical to the Karhunen-Loève expansion (Karhunen, 1946; Loève, 1955):

$$\mathbf{Z} = \sum_{m=1}^{M} z_m \mathbf{\Phi_m}. \tag{47}$$

One can show that when the signal covariance is described by a decreasing exponential function ($\Gamma_{S_0 S_0}(t_1, t_2) = e^{-\alpha|t_1 - t_2|}$, with $\alpha \in \mathbb{R}^{+*}$), basis $\{\mathbf{\Phi_m}\}_{m=1...M}$ corresponds to the Fourier basis (Vann Trees, 1968), so that the Fourier expansion is a particular case of the Karhunen-Loève expansion, which is a particular case of the stochastic matched filter expansion.

3.4 The speckle noise case

Some airborne SAR (Synthetic Aperture Radar) imaging devices randomly generate their own corrupting signal, called the speckle noise, generally described as a multiplicative noise (Tur et al., 1982). This is due to the complexity of the techniques developed to get the best resolution of the ground. Given experimental data accuracy and quality, these systems have been used in sonars (SAS imaging device), with similar characteristics.

Under these conditions, we cannot anymore consider the noise-corrupted signal as described in (18), so its expression becomes:

$$\mathbf{Z} = \mathbf{S}.*\mathbf{N}, \tag{48}$$

where $.*$ denotes the term by term product.

In order to fall down in a known context, let consider the Kuan approach (Kuan et al., 1985). Assuming that the multiplicative noise presents a stationary mean ($\bar{N} = E\{\mathbf{N}\}$), we can define the following normalized observation:

$$\mathbf{Z_{norm}} = \mathbf{Z}/\bar{N}. \tag{49}$$

In this condition, we can represent (49) in terms of signal plus signal-dependent additive noise:

$$\mathbf{Z_{norm}} = \mathbf{S} + \left(\frac{\mathbf{N} - \bar{N}}{\bar{N}}\right).*\mathbf{S}. \tag{50}$$

Let $\mathbf{N_a}$ be this signal-dependent additive colored noise:

$$\mathbf{N_a} = (\mathbf{N}/\bar{N} - 1).*\mathbf{S}. \tag{51}$$

Under these conditions, the mean quadratic value of the mth component z_m of the normalized observation expansion is:

$$E\left\{z_m^2\right\} = \sigma_S^2 \lambda_n + \sigma_{N_a}^2 + \sigma_S \sigma_{N_a} \mathbf{\Phi_m}^T \left(\mathbf{\Gamma_{S_0 N_{a_0}}} + \mathbf{\Gamma_{N_{a_0} S_0}}\right) \mathbf{\Phi_m}, \tag{52}$$

where $\mathbf{N_{a_0}}$ corresponds to the reduced noise $\mathbf{N_a}$.

Consequently, the signal to noise ratio ρ_m becomes:

$$\rho_m = \frac{\sigma_S^2 \lambda_m}{\sigma_{N_a}^2 + \sigma_S \sigma_{N_a} \mathbf{\Phi_m}^T \left(\mathbf{\Gamma_{S_0 N_{a_0}}} + \mathbf{\Gamma_{N_{a_0} S_0}}\right) \mathbf{\Phi_m}}. \tag{53}$$

As $\mathbf{S_0}$ and $(\mathbf{N}/\bar{N} - 1)$ are independent, it comes:

$$\Gamma_{S_0 N_{a_0}}[p_1, p_2, q_1, q_2] = E\left\{S_0[p_1, q_1]N_{a_0}[p_2, q_2]\right\}, \tag{54}$$

which is equal to:

$$\frac{1}{\sigma_{N_a}} E\left\{S_0[p_1, q_1]S[p_2, q_2]\right\} \underbrace{\left(\frac{E\{N[p_2, q_2]\}}{\bar{N}} - 1\right)}_{=0} = 0. \tag{55}$$

So that, the cross-correlation matrices between signal $\mathbf{S_0}$ and signal-dependent noise $\mathbf{N_{a_0}}$ vanishes. For this reason, signal to noise ratio in a context of multiplicative noise like the speckle noise, expanded into the stochastic matched filter basis has the same expression than in the case of an additive noise.

4. The Stochastic Matched Filter in a de-noising context

In this section, we present the stochastic matched filtering in a de-noising context for 1-D discrete time signals. The given results can easily be extended to higher dimensions.

4.1 Bias estimator

Let $\mathbf{Z}$ be a M-dimensional noise corrupted observed signal. The use of the stochastic matched filter as a restoring process is based on the decomposition of this observation, into a random variable finite sequence z_m on the $\{\mathbf{\Psi_m}\}_{m=1...M}$ basis. An approximation $\widetilde{\mathbf{S}}_\mathbf{Q}$ is obtained with the z_m coefficients and the Q basis vectors $\mathbf{\Psi_m}$, with Q lower than M:

$$\widetilde{\mathbf{S}}_\mathbf{Q} = \sum_{m=1}^{Q} z_m \mathbf{\Psi_m}. \tag{56}$$

If we examine the M-dimensional vector $\mathbf{E}\left\{\widetilde{\mathbf{S}}_\mathbf{Q}\right\}$, we have:

$$
\begin{aligned}
\mathbf{E}\left\{\widetilde{\mathbf{S}}_\mathbf{Q}\right\} &= E\left\{\sum_{m=1}^{Q} \mathbf{\Psi_m} z_m\right\} \\
&= \sum_{m=1}^{Q} \mathbf{\Psi_m}\mathbf{\Phi_m^T}\mathbf{E}\left\{\mathbf{Z}\right\}
\end{aligned}
\tag{57}
$$

Using the definition of noise-corrupted signal $\mathbf{Z}$, it comes:

$$\mathbf{E}\left\{\widetilde{\mathbf{S}}_\mathbf{Q}\right\} = \sum_{m=1}^{Q} \mathbf{\Psi_m}\mathbf{\Phi_m^T}\left(\mathbf{E}\left\{\mathbf{S}\right\} + \mathbf{E}\left\{\mathbf{N}\right\}\right). \tag{58}$$

Under these conditions, the estimator bias $B_{\widetilde{S}_Q}$ can be expressed as follows:

$$
\begin{aligned}
B_{\widetilde{S}_Q} &= E\left\{\widetilde{\mathbf{S}}_\mathbf{Q} - \mathbf{S}\right\} \\
&= \left(\sum_{m=1}^{Q} \mathbf{\Psi_m}\mathbf{\Phi_m^T} - \mathbf{I}\right)\mathbf{E}\left\{\mathbf{S}\right\} + \sum_{m=1}^{Q} \mathbf{\Psi_m}\mathbf{\Phi_m^T}\mathbf{E}\left\{\mathbf{N}\right\},
\end{aligned}
\tag{59}
$$

where $\mathbf{I}$ denotes the $M \times M$ identity matrix.

Furthermore, if we consider the signal of interest expansion, we have:

$$\mathbf{S} = \left(\sum_{m=1}^{M} \boldsymbol{\Psi}_{\mathbf{m}} \boldsymbol{\Phi}_{\mathbf{m}}^{\mathbf{T}} \right) \mathbf{S}, \tag{60}$$

so that, by identification, it comes:

$$\sum_{m=1}^{M} \boldsymbol{\Psi}_{\mathbf{m}} \boldsymbol{\Phi}_{\mathbf{m}}^{\mathbf{T}} = \mathbf{I}. \tag{61}$$

In this condition, relation (59) can be rewritten as follows:

$$B_{\widetilde{S}_Q} = - \sum_{m=Q+1}^{M} \boldsymbol{\Psi}_{\mathbf{m}} \boldsymbol{\Phi}_{\mathbf{m}}^{\mathbf{T}} E\{\mathbf{S}\} + \sum_{m=1}^{Q} \boldsymbol{\Psi}_{\mathbf{m}} \boldsymbol{\Phi}_{\mathbf{m}}^{\mathbf{T}} E\{\mathbf{N}\}. \tag{62}$$

This last equation corresponds to the estimator bias when no assumption is made on the signal and noise mean values. In our case, signal and noise are both supposed zero-mean, so that the stochastic matched filter allows obtaining an unbiased estimation of the signal of interest.

4.2 De-noising using a mean square error minimization

4.2.1 Problem description

In many signal processing applications, it is necessary to estimate a signal of interest disturbed by an additive or multiplicative noise. We propose here to use the stochastic matched filtering technique as a de-noising process, such as the mean square error between the signal of interest and its approximation will be minimized.

4.2.2 Principle

In the general theory of stochastic matched filtering, Q is chosen so as the Q first eigenvalues, coming from the generalized eigenvalue problem, are greater than one, in order to enhance the mth component of the observation. To improve this choice, let us consider the mean square error ϵ between the signal of interest $\mathbf{S}$ and its approximation $\widetilde{\mathbf{S}}_{\mathbf{Q}}$:

$$\epsilon = E\left\{ \left(\mathbf{S} - \widetilde{\mathbf{S}}_{\mathbf{Q}} \right)^{T} \left(\mathbf{S} - \widetilde{\mathbf{S}}_{\mathbf{Q}} \right) \right\}. \tag{63}$$

It is possible to show that this error, function of Q, can be written as:

$$\epsilon(Q) = \sigma_{S}^{2} \left(1 - \sum_{m=1}^{Q} \lambda_n \|\boldsymbol{\Psi}_{\mathbf{m}}\|^2 \right) + \sigma_{N}^{2} \sum_{m=1}^{Q} \|\boldsymbol{\Psi}_{\mathbf{m}}\|^2. \tag{64}$$

The integer Q is chosen so as to minimize the relation (64). It particularly verifies:

$$(\epsilon(Q) - \epsilon(Q-1)) < 0 \quad \& \quad (\epsilon(Q+1) - \epsilon(Q)) > 0,$$

let us explicit these two inequalities; on the one hand:

$$\epsilon(Q+1) - \epsilon(Q) = \left(\sigma_{N}^{2} - \sigma_{S}^{2} \lambda_{Q+1} \right) \|\boldsymbol{\Psi}_{\mathbf{Q+1}}\|^2 > 0$$

and on the other hand:

$$\epsilon(Q) - \epsilon(Q-1) = \left(\sigma_N^2 - \sigma_S^2 \lambda_Q\right) \left\|\mathbf{\Psi_Q}\right\|^2 < 0.$$

Hence, integer Q verifies:

$$\sigma_S^2 \lambda_Q > \sigma_N^2 > \sigma_S^2 \lambda_{Q+1}.$$

The dimension of the basis $\{\mathbf{\Psi_m}\}_{m=1\ldots Q}$, which minimizes the mean square error between the signal of interest and its approximation, is the number of eigenvalues λ_m verifying:

$$\frac{\sigma_S^2}{\sigma_N^2} \lambda_m > 1, \tag{65}$$

where $\frac{\sigma_S^2}{\sigma_N^2}$ is the signal to noise ratio before processing.

Consequently, if the observation has a high enough signal to noise ratio, many $\mathbf{\Psi_m}$ will be considered for the filtering (so that $\widetilde{\mathbf{S_Q}}$ tends to be equal to $\mathbf{Z}$), and in the opposite case, only a few number will be chosen. In these conditions, this filtering technique applied to an observation $\mathbf{Z}$ with an initial signal to noise ratio $\left.\frac{S}{N}\right|_{\mathbf{Z}}$ substantially enhances the signal of interest perception. Indeed, after processing, the signal to noise ratio $\left.\frac{S}{N}\right|_{\widetilde{\mathbf{S_Q}}}$ becomes:

$$\left.\frac{S}{N}\right|_{\widetilde{\mathbf{S_Q}}} = \left.\frac{S}{N}\right|_{\mathbf{Z}} \frac{\sum\limits_{m=1}^{Q} \lambda_m \left\|\mathbf{\Psi_m}\right\|^2}{\sum\limits_{m=1}^{Q} \left\|\mathbf{\Psi_m}\right\|^2}. \tag{66}$$

4.2.3 The Stochastic Matched Filter

As described in a forthcoming section, the stochastic matched filtering method is applied using a sliding sub-window processing. Therefore, let consider a K-dimensional vector $\mathbf{Z_k}$ corresponding to the data extracted from a window centered on index k of the noisy data, i.e.:

$$\mathbf{Z_k}^T = \left\{ Z\left[k - \frac{K-1}{2}\right], \ldots, Z[k], \ldots, Z\left[k + \frac{K-1}{2}\right] \right\}. \tag{67}$$

This way, M sub-windows $\mathbf{Z_k}$ are extracted to process the whole observation, with $k = 1, \ldots, M$. Furthermore, to reduce the edge effects, the noisy data can be previously completed with zeros or using a mirror effect on its edges.

According to the sliding sub-window processing, only the sample located in the middle of the window is estimated, so that relation (56) becomes:

$$\widetilde{S}_{Q[k]}[k] = \sum_{m=1}^{Q[k]} z_{m,k} \Psi_m \left[\frac{K+1}{2}\right], \tag{68}$$

with:

$$z_{m,k} = \mathbf{Z_k}^T \mathbf{\Phi_m} \tag{69}$$

and where $Q[k]$ corresponds to the number of eigenvalues λ_m times the signal to noise ratio of window $\mathbf{Z_k}$ greater than one, i.e.:

$$\lambda_m \left. \frac{S}{N} \right|_{\mathbf{Z_k}} > 1. \tag{70}$$

To estimate the signal to noise ratio of window $\mathbf{Z_k}$, the signal power is directly computed from the window's data and the noise power is estimated on a part of the noisy data $\mathbf{Z}$, where no useful signal *a priori* occurs. This estimation is generally realized using the maximum likelihood principle.

Using relations (68) and (69), the estimation of the de-noised sample value is realized by a scalar product:

$$\widetilde{S}_{Q[k]}[k] = \mathbf{Z_k}^T \sum_{m=1}^{Q[k]} \mathbf{\Psi}_m \left[\frac{K+1}{2} \right] \mathbf{\Phi_m}. \tag{71}$$

In this case and taking into account the sub-window size, reduced covariances $\mathbf{\Gamma_{S_0 S_0}}$ and $\mathbf{\Gamma_{N_0 N_0}}$ are both $K \times K$ matrices, so that $\{\mathbf{\Phi_n}\}$ and $\{\mathbf{\Psi_n}\}$ are K-dimensional basis.

Such an approach can be completed using the following relation:

$$\widetilde{S}_{Q[k]}[k] = \mathbf{Z_k}^T \mathbf{h_{Q[k]}}, \tag{72}$$

for k taking values between 1 and M, and where:

$$\mathbf{h_{Q[k]}} = \sum_{m=1}^{Q[k]} \mathbf{\Psi}_m \left[\frac{K+1}{2} \right] \mathbf{\Phi_m}. \tag{73}$$

$Q[k]$ taking values between 1 and K, relation (73) permits to compute K vectors $\mathbf{h_q}$, from $\mathbf{h_1}$ ensuring a maximization of the signal to noise ratio, to $\mathbf{h_K}$ whose bandwidth corresponds to the whole useful signal bandwidth. These filters are called the stochastic matched filters for the following.

4.2.4 Algorithm

The algorithm leading to an approximation $\widetilde{\mathbf{S}}_\mathbf{Q}$ of the signal of interest $\mathbf{S}$, by the way of the stochastic extension of the matched filter, using a sliding sub-window processing, is presented below.

1. Modelisation or estimation of reduced covariances $\mathbf{\Gamma_{S_0 S_0}}$ and $\mathbf{\Gamma_{N_0 N_0}}$ of signal of interest and noise respectively.

2. Estimation of the noise power σ_N^2 in an homogeneous area of $\mathbf{Z}$.

3. Determination of eigenvectors $\mathbf{\Phi_m}$ by solving the generalized eigenvalue problem described in (22) or (42).

4. Normalization of $\mathbf{\Phi_m}$ according to (34) or (43).

5. Determination of vectors $\mathbf{\Psi_n}$ (relation (35) or (44)).

6. Computation of the K stochastic matched filters $\mathbf{h_q}$ according to (73).

7. Set to zero M samples approximation $\widetilde{\mathbf{S}}_\mathbf{Q}$.

8. For $k = 1$ to M do:

(a) Sub-window $\mathbf{Z_k}$ extraction.

(b) $\mathbf{Z_k}$ signal to noise ratio estimation.

(c) $Q[k]$ determination according to (70).

(d) Scalar product (72) computation.

Let us note the adaptive nature of this algorithm, each sample being processed with the most adequate filter $\mathbf{h_q}$ depending on the native signal to noise ratio of the processed sub-window.

4.3 Experiments

In this section, we propose two examples of de-noising on synthetic and real data in the case of 2-D discrete-space signals.

4.3.1 2-D discrete-space simulated data

As a first example, consider the Lena image presented in figure 1. This is a 512×512 pixels coded with 8 bits (i.e. 256 gray levels). This image has been artificially noise-corrupted by a zero-mean, Gaussian noise, where the local variance of the noise is a function of the image intensity values (see figure 3.a).

Fig. 1. Lena image, 512×512 pixels, 8 bits encoded (256 gray levels)

The stochastic matched filtering method is based on the assumption of signal and noise stationarity. Generally it is the case for the noise. However, the signal of interest is not necessarily stationary. Obviously, some images can be empirically supposed stationary, it is the case for sea-bed images, for some ocean waves images, in other words for all images able to be assimilated to a texture. But in most cases, an image cannot be considered as the realization of a stationary stochastic process. However after a segmentation operation, it is possible to define textured zones. This way, a particular zone of an image (also called window) can be considered as the realization of a stationary bi-dimensional stochastic process. The dimensions of these windows must be of the same order of magnitude as the texture coherence length. Thus, the stochastic matched filter will be applied on the native image using a windowed processing. The choice of the window dimensions is conditioned by the texture coherence length

mean value.

The implementation of the stochastic matched filter needs to have an *a priori* knowledge of signal of interest and noise covariances. The noise covariance is numerically determined in an homogeneous area of the observation, it means in a zone without any *a priori* information on signal of interest. This covariance is computed by averaging several realizations. The estimated power spectral density associated to the noise covariance is presented on figure 2.a. The signal of interest covariance is modeled analytically in order to match the different textures of the image. In dimension one, the signal of interest autocorrelation function is generally described by a triangular function because its associated power spectral density corresponds to signals with energy contained inside low frequency domain. This is often the case in reality. The model used here is a bi-dimensional extension of the mono-dimensional case. Furthermore, in order to not favor any particular direction of the texture, the model has isotropic property. Given these different remarks, the signal of interest autocorrelation function has been modeled using a Gaussian model, as follows:

$$\Gamma_{S_0 S_0}[n, m] = \exp\left[-\left(n^2 + m^2\right) / (2F_e^2 \sigma^2)\right], \quad \forall (n, m) \in \mathbb{Z}^2, \tag{74}$$

with n and m taking values between $-(K-1)$ and $(K-1)$, where F_e represents the sampling frequency and where σ has to be chosen so as to obtain the most representative power spectral density. $\Gamma_{S_0 S_0}$ being Gaussian, its power spectral density is Gaussian too, with a variance σ_ν^2 equal to $1/(4\pi^2\sigma^2)$. As for a Gaussian signal, 99 % of the signal magnitudes arise in the range $[-3\sigma_\nu; 3\sigma_\nu]$, we have chosen σ_ν such as $6\sigma_\nu = F_e$, so that:

$$\sigma = 3/(\pi F_e). \tag{75}$$

The result power spectral density is presented on figure 2.b.

(a) Estimated noise PSD   (b) Modeled signal PSD
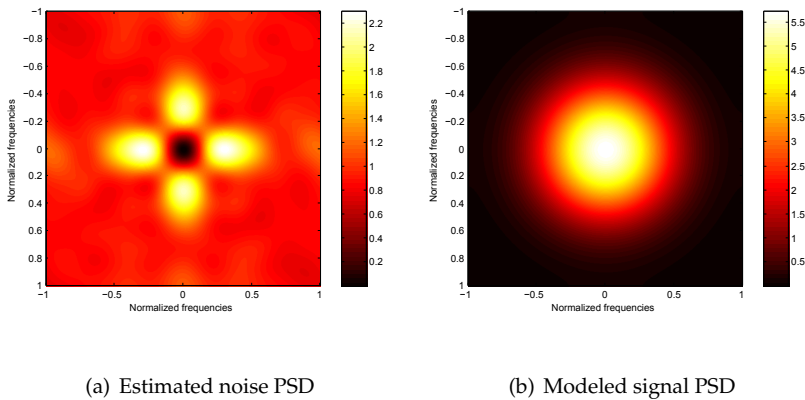
Fig. 2. Signal and noise power spectral densities using normalized frequencies

The dimension of the filtering window for this process is equal to 7×7 pixels, in order to respect the average coherence length of the different textures. For each window, number Q of

eigenvalues has been determined according to relation (70), with:

$$\left.\frac{S}{N}\right|_{\mathbf{Z_k}} = \frac{\sigma_{Z_k}^2 - \sigma_N^2}{\sigma_N^2}, \tag{76}$$

the noise variance σ_N^2 being previously estimated in an homogeneous area of the noise-corrupted data using a maximum likelihood estimator. The resulting image is presented on figure 3.b.

(a) Noisy Lena

(b) De-noised Lena
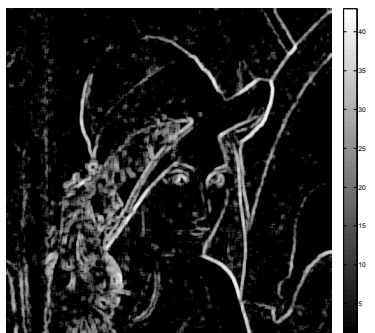
(c) Q values
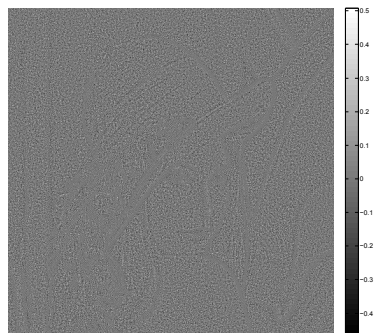
(d) Removed signal

Fig. 3. 1st experiment: Lena image corrupted by a zero-mean Gaussian noise with a local variance dependent of the image intensity values

An analysis of the figure 3.b shows that the stochastic matched filter used as a de-noising process gives some good results in terms of noise rejection and detail preservation. In order to quantify the effectiveness of the process, we propose on figure 3.d an image of the removed signal $\widetilde{\mathbf{N}}$ (i.e. $\widetilde{\mathbf{N}} = \mathbf{Z} - \widetilde{\mathbf{S}}_{\mathbf{Q}}$), where the areas corresponding to useful signal details present an amplitude tending toward zero, the process being similar to an all-pass filter in order to preserve the spatial resolution. Nevertheless, the resulting image is still slightly noise-corrupted locally. It is possible to enhance the de-noising power increasing either the σ value (that corresponds to a diminution of the σ_ν value and so to a smaller signal bandwidth) or the sub-image size, but this would involve a useful signal deterioration by a smoothing effect. In addition, the choice of the number Q of basis vectors by minimizing the mean square error between the signal of interest $\mathbf{S}$ and its approximation $\widetilde{\mathbf{S}}_{\mathbf{Q}}$ implies an image contour well preserved. As an example, we present in figures 3.c and 4 an image of the values of Q used for each window and a curve representative of the theoretical and real improvement of the signal to noise ratio according to these values (relation (66)).
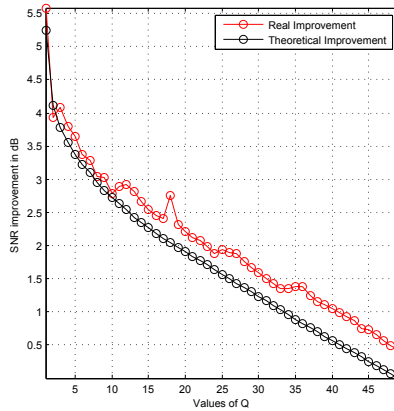
Fig. 4. Theoretical and real SNR improvement in dB of the de-noised data

As previously specified, when the signal to noise ratio is favorable a lot of basis vectors are retained for the filtering. In this case, the stochastic matched filter tends to be an all-pass filter, so that the signal to noise ratio improvement is not significant. On the other hand, when the signal to noise ratio is unfavorable this filtering method allows a great improvement (up to $5\,dB$ when only from 1 up to 2 basis vectors were retained), the stochastic matched filter being similar to a mean filter. Furthermore, the fact that the curves of the theoretical and real improvements are similar reveals the relevance of the signal covariance model.

4.3.2 2-D discrete-space real data

The second example concerns real 2-D discrete-space data acquired by a SAS (Synthetic Aperture Sonar) system. Over the past few years, SAS has been used in sea bed imagery. Active synthetic aperture sonar is a high-resolution acoustic imaging technique that coherently combines the returns from multiple pings to synthesize a large acoustic aperture. Thus, the azimuth resolution of a SAS system does not depend anymore on the length of

the real antenna but on the length of the synthetic antenna. Consequently, in artificially removing the link between azimuth resolution and physical length of the array, it is now possible to use lower frequencies to image the sea bed and keep a good resolution. Therefore, lower frequencies are less attenuated and long ranges can be reached. All these advantages make SAS images of great interest, especially for the detection, localization and eventually classification of objects lying on the sea bottom. But, as any image obtained with a coherent system, SAS images are corrupted by the speckle noise. Such a noise gives a granular aspect to the images, by giving a variance to the intensity of each pixel. This reduces spatial and radiometric resolutions. This noise can be very disturbing for the interpretation and the automatic analysis of SAS images. For this reason a large amount of research works have been dedicated recently to reduce this noise, with as common objectives the strong reduction of the speckle level, coupled to the spatial resolution preservation.

Consider the SAS image[1] presented in figure 5.a. This is a 642×856 pixels image of a wooden barge near Prudence Island. This barge measures roughly 30 meters long and lies in 18 meters of water.

(a) SAS data

(b) De-speckled SAS data

(c) Q values

(d) Removed signal
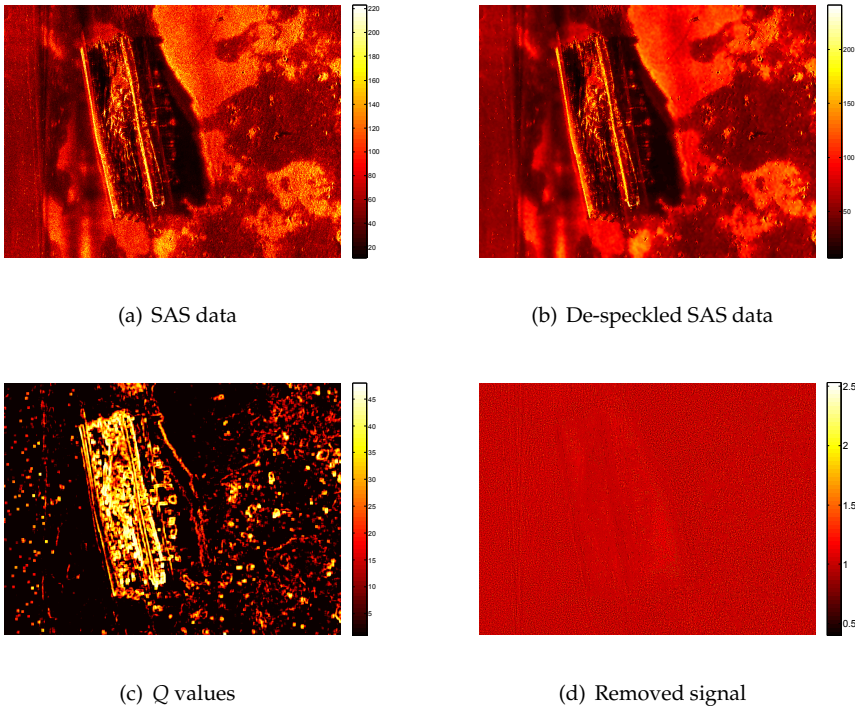
Fig. 5. 2nd experiment: Speckle noise corrupted SAS data: Wooden Barge (Image courtesy of AUVfest 2008)

---

[1] Courtesy of AUVfest 2008: *http://oceanexplorer.noaa.gov*

The same process than for the previous example has been applied to this image to reduce the speckle level. The main differences between the two experiments rest on the computation of the signal and noise statistics. As the speckle noise is a multiplicative noise (see relation (48)), the noise covariance, the noise power and the noise mean value have been estimated on the high-frequency components $\mathbf{Z_{h_{HF}}}$ of an homogeneous area $\mathbf{Z_h}$ of the SAS data:

$$\mathbf{Z_{h_{HF}}} = \mathbf{Z_h}./\mathbf{Z_{h_{BF}}}, \tag{77}$$

where ./ denotes the term by term division and with $\mathbf{Z_{h_{BF}}}$ corresponding to the low-frequency components of the studied area obtained applying a classical low-pass filter. This way, all the low-frequency fluctuations linked to the useful signal are canceled out.

Furthermore, taking into account the multiplicative nature of the noise, to estimate the signal to noise ratio $\left.\frac{S}{N}\right|_{\mathbf{Z_k}}$ of the studied window, the signal variance has been computed as follows:

$$\sigma^2_{S_k} = \frac{\sigma^2_{Z_k} + E\{\mathbf{Z_k}\}^2}{\sigma^2_N + \bar{N}^2} - E\{\mathbf{S_k}\}, \tag{78}$$

where:

$$E\{\mathbf{S_k}\} = \frac{E\{\mathbf{Z_k}\}^2}{\bar{N}^2}. \tag{79}$$

The de-noised SAS data is presented on figure 5.b. An image of the Q values retained for the process and the ratio image $\mathbf{Z}./\widetilde{\mathbf{S}_Q}$ are proposed on figures 5.c and 5.d respectively. These results show that the stochastic matched filter yields good speckle noise reduction, while keeping all the details with no smoothing effect on them (an higher number of basis vectors being retained to process them), so that the spatial resolution seems not to be affected.

4.4 Concluding remarks

In this section, we have presented the stochastic matched filter in a de-noising context. This one is based on a truncation to order Q of the random noisy data expansion (56). To determine this number Q, it has been proposed to minimize the mean square error between the signal of interest and its approximation. Experimental results have shown the usefulness of such an approach. This criterion is not the only one, one can apply to obtain Q. The best method to determine this truncature order may actually depend on the nature of the considered problem. For examples, the determination of Q has been achieved in (Fraschini et al., 2005) considering the Cramér-Rao lower bound and in (Courmontagne, 2007) by the way of a minimization between the speckle noise local statistics and the removal signal local statistics. Furthermore, several stochastic matched filter based de-noising methods exist in the scientific literature, as an example, let cite (Courmontagne & Chaillan, 2006), where the de-noising is achieved using several signal covariance models and several sub-image sizes depending on the windowed noisy data statistics.

5. The Stochastic Matched Filter in a detection context

In this section, the stochastic matched filter is described for its application in the field of short signal detection in a noisy environment.

5.1 Problem formulation

Let consider two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ corresponding to "*there is only noise in the available data*" and "*there is signal of interest in the available data*" respectively and let consider a K-dimensional vector $\mathbf{Z}$ containing the available data. The dimension K is assumed large (i.e. $K >> 100$). Under hypothesis $\mathcal{H}_0$, $\mathbf{Z}$ corresponds to noise only and under hypothesis $\mathcal{H}_1$ to a signal of interest $\mathbf{S}$ corrupted by an additive noise $\mathbf{N}$:

$$\begin{cases} \mathcal{H}_0 & : \quad \mathbf{Z} = \sigma_N \mathbf{N_0} \\ \mathcal{H}_1 & : \quad \mathbf{Z} = \sigma_S \mathbf{S_0} + \sigma_N \mathbf{N_0} \end{cases}, \tag{80}$$

where σ_S and σ_N are signal and noise standard deviation respectively and $E\left\{|\mathbf{S_0}|^2\right\} = E\left\{|\mathbf{N_0}|^2\right\} = 1$. By assumptions, $\mathbf{S_0}$ and $\mathbf{N_0}$ are extracted from two independent, stationary and zero-mean random signals of known autocorrelation functions. This allows us to construct the covariances of $\mathbf{S_0}$ and $\mathbf{N_0}$ denoted $\Gamma_{\mathbf{S_0 S_0}}$ and $\Gamma_{\mathbf{N_0 N_0}}$ respectively.

Using the stochastic matched filter theory, it is possible to access to the set $(\Phi_{\mathbf{m}}, \lambda_m)_{m=1...M}$, with M bounded by K, allowing to compute the uncorrelated random variables z_m associated to observation $\mathbf{Z}$. It comes:

$$\begin{cases} E\left\{z_m^2 / \mathcal{H}_0\right\} = \sigma_N^2 \\ E\left\{z_m^2 / \mathcal{H}_1\right\} = \sigma_S^2 \lambda_m + \sigma_N^2 \end{cases}. \tag{81}$$

Random variables z_m being a linear transformation of a random vector, the central limit theorem can be invoked and we will assume in the sequel that z_m are approximately Gaussian:

$$z_m \hookrightarrow \mathcal{N}\left(0, \left. E\left\{z_m^2 / \mathcal{H}_i\right\}\right|_{i=0,1}\right). \tag{82}$$

Let Γ_0 and Γ_1 be the covariances of the signals in the basis $\{\Phi_{\mathbf{m}}\}_{m=1...M}$, under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, it comes:

$$\Gamma_0 = \sigma_N^2 \mathbf{I}, \tag{83}$$

where $\mathbf{I}$ denotes the $M \times M$ identity matrix and

$$\Gamma_1 = \begin{pmatrix} \sigma_S^2 \lambda_1 + \sigma_N^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_S^2 \lambda_2 + \sigma_N^2 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & \sigma_S^2 \lambda_M + \sigma_N^2 \end{pmatrix}. \tag{84}$$

In these conditions, the probability density functions under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ take for expression:

$$\begin{cases} p(\mathbf{z}/\mathcal{H}_0) = \dfrac{1}{(2\pi)^{\frac{M}{2}} \sqrt{|\Gamma_0|}} \exp\left[\dfrac{-1}{2}\left(\mathbf{z}^T \Gamma_0^{-1} \mathbf{z}\right)\right] \\ p(\mathbf{z}/\mathcal{H}_1) = \dfrac{1}{(2\pi)^{\frac{M}{2}} \sqrt{|\Gamma_1|}} \exp\left[\dfrac{-1}{2}\left(\mathbf{z}^T \Gamma_1^{-1} \mathbf{z}\right)\right] \end{cases}, \tag{85}$$

where $\mathbf{z}$ is a $M-$dimensional vector, whose m^{th} component is z_m:

$$\mathbf{z} = (z_1, z_2, \dots, z_m, \dots, z_M)^T. \tag{86}$$

It is well known that the Neyman-Pearson lemma yields the uniformly most powerful test and allows to obtain the following rule of decision based on the likelihood ratio $\Lambda(\mathbf{z})$:

$$\Lambda(\mathbf{z}) = \frac{p(\mathbf{z}/\mathcal{H}_1)}{p(\mathbf{z}/\mathcal{H}_0)} \underset{<D_0}{\overset{>D_1}{\gtrless}} \lambda, \tag{87}$$

where λ is the convenient threshold.

Taking into account relations (83), (84) and (85), it comes:

$$\underbrace{\sum_{m=1}^{M} \frac{\lambda_m}{\sigma_S^2 \lambda_m + \sigma_N^2}}_{U_M} \underset{<D_0}{\overset{>D_1}{\gtrless}} \underbrace{\frac{\sigma_N^2}{\sigma_S^2} \left[2 \left(\ln \lambda - M \ln \sigma_N \right) + \sum_{m=1}^{M} \ln \left(\sigma_S^2 \lambda_m + \sigma_N^2 \right) \right]}_{T_M}. \tag{88}$$

In these conditions, the detection and the false alarm probabilities are equal to:

$$P_d = \int_{T_M}^{\infty} p_{U_M}(u/\mathcal{H}_1) du \quad \text{and} \quad P_{fa} = \int_{T_M}^{\infty} p_{U_M}(u/\mathcal{H}_0) du. \tag{89}$$

So, the detection problem consists in comparing u to threshold T_M and in finding the most convenient order M for an optimal detection (i.e. a slight false alarm probability and a detection probability quite near one).

5.2 Subspace of dimension one

First, let consider the particular case of a basis $\{\Phi_m\}_{m=1...M}$ restricted to only one vector Φ. In this context, relation (88) leads to:

$$|z| \underset{<D_0}{\overset{>D_1}{\gtrless}} \underbrace{\sqrt{\frac{T_1}{\lambda_1} (\sigma_S^2 \lambda_1 + \sigma_N^2)}}_{z_s} \tag{90}$$

and the detection and false alarm probabilities become:

$$P_d = \int_{D_1} p(z/\mathcal{H}_1) dz \quad \text{and} \quad P_{fa} = \int_{D_1} p(z/\mathcal{H}_0) dz, \tag{91}$$

where $D_1 =]-\infty; -z_s] \cup [z_s; +\infty[$ and with:

$$\begin{cases} \text{under } \mathcal{H}_0 : z \hookrightarrow \mathcal{N} \left(0, \sigma_N^2 \right) \\ \text{under } \mathcal{H}_1 : z \hookrightarrow \mathcal{N} \left(0, \sigma_S^2 \lambda_1 + \sigma_N^2 \right) \end{cases}. \tag{92}$$

From (91), it comes:

$$P_{fa} = 1 - \text{erf} \left(\frac{z_s}{\sqrt{2}\sigma_N} \right), \tag{93}$$

where erf(.) denotes the error function:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp \left[-y^2 \right] dy. \tag{94}$$

In these conditions, the threshold value z_s can be expressed as a function of the false alarm probability:

$$z_s = \sqrt{2}\sigma_N \text{erf}^{-1} (1 - P_{fa}). \tag{95}$$

Furthermore, the detection probability takes the following expression:

$$P_d = 1 - \text{erf}\left(\frac{z_s}{\sqrt{2\left(\sigma_S^2 \lambda_1 + \sigma_N^2\right)}}\right). \tag{96}$$

We deduce from equations (95) and (96), the ROC curve expression:

$$P_d\left(P_{fa}\right) = 1 - \text{erf}\left(\sqrt{\frac{1}{1 + \rho_0 \lambda_1}}\text{erf}^{-1}(1 - P_{fa})\right), \tag{97}$$

where $\rho_0 = \sigma_S^2 / \sigma_N^2$.

One can show that an optimal detection is realized, when λ_1 corresponds to the greatest eigenvalue of the generalized eigenvalue problem (22).

5.3 Subspace of dimension M

Random variable U_M being a weighted sum of square Gaussian random variables, its probability density function, under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, can be approximated by a Gamma law (Kendall & Stuart, 1979; Zhang & Liu, 2002). It comes:

$$p_{U_M}(u/\mathcal{H}_i) \simeq u^{k_i - 1} \frac{\exp\left[\frac{-u}{\theta_i}\right]}{\Gamma(k_i)\theta_i^{k_i}}, \tag{98}$$

for i equal 0 or 1 and where $k_0 \theta_0 = E\{U_M/\mathcal{H}_0\}$, $k_0 \theta_0^2 = \text{VAR}\{U_M/\mathcal{H}_0\}$, $k_1 \theta_1 = E\{U_M/\mathcal{H}_1\}$ and $k_1 \theta_1^2 = \text{VAR}\{U_M/\mathcal{H}_1\}$. In these conditions, it comes under $\mathcal{H}_0$:

$$k_0 = \frac{\left(\displaystyle\sum_{m=1}^{M} \frac{\lambda_m}{1 + \rho_0 \lambda_m}\right)^2}{2 \displaystyle\sum_{m=1}^{M}\left(\frac{\lambda_m}{1 + \rho_0 \lambda_m}\right)^2} \quad \text{and} \quad \theta_0 = 2\frac{\displaystyle\sum_{m=1}^{M}\left(\frac{\lambda_m}{1 + \rho_0 \lambda_m}\right)^2}{\displaystyle\sum_{m=1}^{M} \frac{\lambda_m}{1 + \rho_0 \lambda_m}} \tag{99}$$

and, under $\mathcal{H}_1$:

$$k_1 = \frac{\left(\displaystyle\sum_{m=1}^{M} \lambda_m\right)^2}{2 \displaystyle\sum_{m=1}^{M} \lambda_m^2} \quad \text{and} \quad \theta_1 = 2\frac{\displaystyle\sum_{m=1}^{M} \lambda_m^2}{\displaystyle\sum_{m=1}^{M} \lambda_m}. \tag{100}$$

It has been shown in (Courmontagne et al., 2007) that the use of the stochastic matched filter basis $\{\Phi_\mathbf{m}\}_{m=1...M}$ ensures a maximization of the distance between the maxima of $p_{U_M}(u/\mathcal{H}_0)$ and $p_{U_M}(u/\mathcal{H}_1)$ and so leads to an optimal detection.

The basis dimension M is determined by a numerical way. As the detection algorithm is applied using a sliding sub-window processing, each sub-window containing K samples, we can access to K eigenvectors solution of the generalized eigenvalue problem (22). For each

value of M, bounded by 2 and K, we numerically determine the threshold value T_M allowing a wanted false alarm probability and according to the following relation:

$$P_{fa} = 1 - \Delta u^{k_0} \sum_{q=0}^{Q_M} q^{k_0-1} \frac{\exp\left[\dfrac{-q\Delta u}{\theta_0}\right]}{\Gamma(k_0)\theta_0^{k_0}}, \tag{101}$$

where $T_M = Q_M \Delta u$.
Then for each value of T_M, we compute the detection probability according to:

$$P_d = 1 - \Delta u^{k_1} \sum_{q=0}^{Q_M} q^{k_1-1} \frac{\exp\left[\dfrac{-q\Delta u}{\theta_1}\right]}{\Gamma(k_1)\theta_1^{k_1}}. \tag{102}$$

Finally, the basis dimension will correspond to the M value leading to a threshold value T_M allowing the highest detection probability.

5.4 Experiments
5.4.1 Whale echoes detection
Detecting useful information in the underwater domain has taken an important place in many research works. Whether it is for biological or economical reasons it is important to be able to correctly distinguish the many kinds of entities which belong to animal domain or artificial object domain.

For this reason, we have chosen to confront the proposed process with signals resulting from underwater acoustics. The signal of interest **S** corresponds to an acoustic record of several whale echoes. The sampling rate used for this signal is $44100\,Hz$. Each echo lasts approximatively two seconds. The disturbing signal **N** corresponds to a superposition of various marine acoustic signatures. The simulated received noisy signal **Z** has been constructed as follows:

$$\mathbf{Z} = \mathbf{S} + g\mathbf{N}, \tag{103}$$

where g is a SNR control parameter allowing to evaluate the robustness of the detection processing. Several realizations were built with a SNR taking values from $-12\,dB$ to $12\,dB$ (the SNR corresponds to the ratio of the signal and noise powers in the common spectral bandwidth). As an example, we present on figure 7.a in black lines the noisy data in the case of a SNR equal to $-6\,dB$. On the same graph, in red lines, we have reported the useful signal **S**.

The signal and noise covariances were estimated on signals belonging to the same series of measurement as the signal **S** and the noise **N**. The signal covariance was obtained by fitting a general model based on the average of several realizations of whale clicks while the noise one was estimated from a supposed homogeneous subset of the record.

The ROC curves numerically obtained by the way of relations (101) and (102) with the probability density functions described by relations (98) are presented on figure 6 (for a signal to noise ratio greater than $-8\,dB$ the ROC curves are practically on the graph axes). There are 512 samples in the sub-window, so we can access to 512 eigenvectors (i.e. 512 is the maximal size of the basis). Basis dimension M takes values between 17 and 109 depending on the studied signal to noise ratio ($M = 17$ for $\left.\frac{S}{N}\right|_{\mathbf{Z}} = 12\,dB$ and $M = 109$ for $\left.\frac{S}{N}\right|_{\mathbf{Z}} = -12\,dB$).
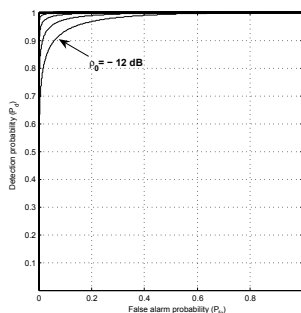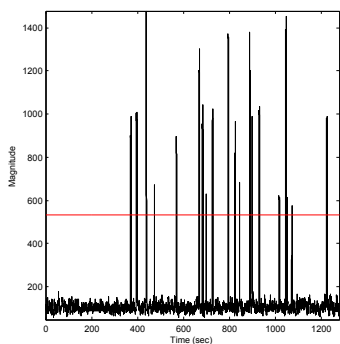
Fig. 6. ROC curves for a SNR taking values in $[-12\,dB; 12\,dB]$
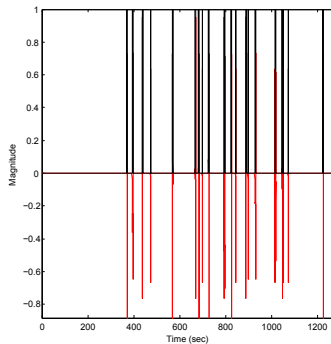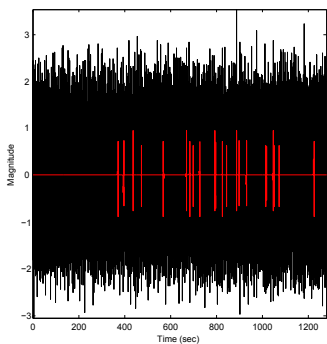
(a) Noisy data with a SNR equal to $-6\,dB$

(b) Result of the detection algorithm
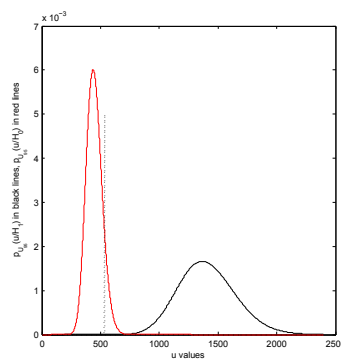
(c) U_M and T_M according to (88)

(d) Probability density functions $p_{U_M}(u/\mathcal{H}_0)$ and $p_{U_M}(u/\mathcal{H}_1)$

Fig. 7. 1st experiment: Whale echoes detection

The false alarm probability has been settled to 10 %. The result to the detection process is proposed on figure 7.b in black lines. On the same graph, the useful signal has been reported in red lines, in order to verify that a detected event coincides with a whale echo. Figure 7.c presents the vector U_M and the automatic threshold T_M: the use of the stochastic matched filter allows to amplify the parts of the noisy data corresponding to whale echoes, while the noise is clearly rejected. As explained in a previous section, the efficiency of the stochastic matched filter is due to its property to dissociate the probability density functions, as shown in figure 7.d (on this graph appears in dashed lines the position of the threshold T_M for a $P_{fa} = 10\%$)).

5.4.2 Mine detection on SAS images

Detection and classification of underwater mines (completely or partially buried) with SAS images is a major challenge to the mine countermeasures community. In this context, experts are looking for more and more efficient detection processes in order to help them in their decisions concerning the use of divers, mines destruction ... As a mine highlight region usually has a corresponding shadow region (see figure 8), most of the methods used to detect and classify objects lying on the seafloor are based on the interpretation of the shadows of the objects. Other methods are focusing on the echo itself. For these approaches, two main problems could occur:

- given the position of the sonar fish and the type of mine encountered, the shape of the echo and its associated shadow zone could vary; but as most of these techniques of detection generally required training, their success can be dependent on the similarity between the training and test data sets,

- given that SAS images are speckle noise corrupted, it is generally necessary to denoise these images before of all; but such a despeckling step could involve miss and/or false detection by an alteration of the echo and/or shadow, given that most of the despeckling methods induce a smoothing effect.

In answer to these problems, we propose to use a one-dimensional detector based on the stochastic matched filter. This detector is applied on each line of the SAS data (considering as a line the data vector in a direction perpendicular to the fish direction). In this context, we construct a very simple model of the signal to be detected (see figure 8), where d, the size of the echo in sight, is a uniform random variable taking values in a range dependent of the mine dimensions. So the problem of mine detection in SAS images is reduced to the one of detecting a one-dimensional signal, such as the model presented in figure 8, in a noisy data vector **Z**.

As the length of the shadow region depends on the fish height, we do not consider the whole shadow for our model, but only its beginning (this corresponds to the length D in figure 8).

The signal covariance is estimated using several realizations of the signal model by making varied the random variable d value. For the noise, its covariance is computed in an area of the data, where no echo is assumed to be present and takes into account the hilly seabed.
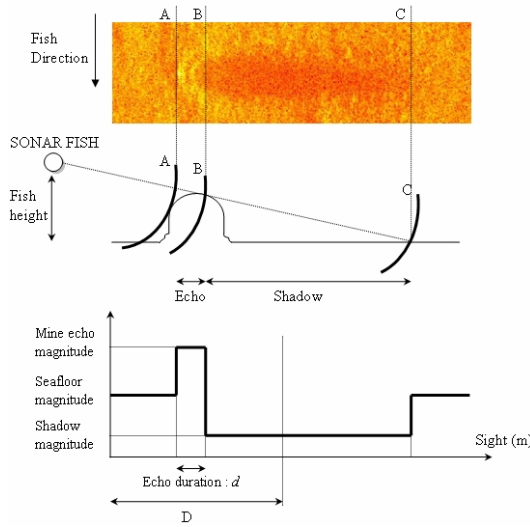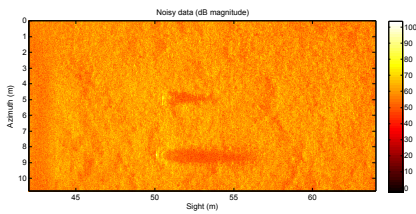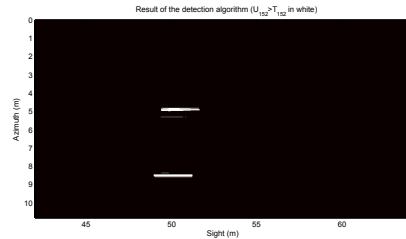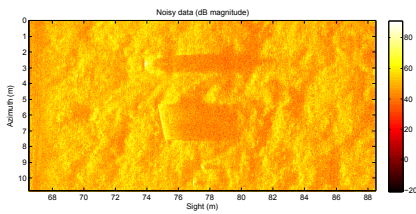
Fig. 8. Signal model: the wave is blocked by objects lying on seafloor and a shadow is generated
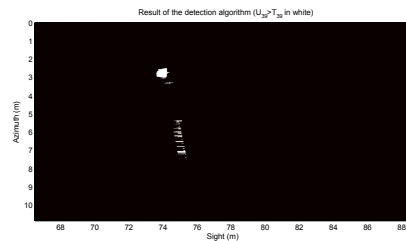

(a) SAS image representing a sphere and a more complex mine like object lying on the seafloor

(b) Result of the detection algorithm

(c) SAS image representing two mine like objects: a sphere and a cylinder lying on its side

(d) Result of the detection algorithm

Fig. 9. 2nd experiment: Mine detection on SAS images

The data set used for this study has been recorded in 1999 during a joint experiment between GESMA (Groupe d'Etudes Sous-Marines de l'Atlantique, France) and DERA (Defence Evaluation and Research Agency, United Kingdom). The sonar was moving along a fixed rail and used a central frequency of $150\,kHz$ with a bandwidth of $64\,kHz$. Figure 9.a and 9.c present two images obtained during this experiment[2]. We recognize the echoes (the bright pixels) in front of the objects and the shadows behind as well as the speckle that gives the granular aspect to the image. Because of the dynamics of the echo compared to the remainder of the image, these figures are represented in dB magnitude.

As the dimensions of the two mine like objects in azimuth and sight are not greater than one meter, the proposed process has been calibrated to detect objects which dimensions in sight are included in the range $[0.5\,m; 1\,m]$ (i.e. the uniform random variable from the signal model takes values in $[0.5; 1]$). For these two experiments, the false alarm probability has been settled to 0.1%, entailing a basis dimension M equal to 152 for the first one and to 39 for the second one. The results obtained applying the detection algorithm are presented on figures 9.b and 9.d. For the two cases, the mine like objects are well detected, without false alarm. These results demonstrate the advantages of such a detection scheme, even in difficult situations, such as the one presented on figure 9.c.

6. Conclusions

This chapter concerned the problem of a noise-corrupted signal expansion and its applications to detection and signal enhancement. The random signal expansion, used here, is the stochastic matched filtering technique. Such a filtering approach is based on the noisy data projection onto a basis of known vectors, with uncorrelated random variables as decomposition coefficients. The basis vectors are chosen such as to maximize the signal to noise ratio after denoising. Several experiments in the fields of short signal detection in a noisy environment and of signals de-noising have shown the efficiency of the proposed expansion, even for unfavorable signal to noise ratio.

7. References

K. Karhunen, Zur spektraltheorie stochastischer prozesse, *Ann. Acad. Sci. Fennicae*, N° 37, 1946, pp. 1-37.

M.M. Loève, *Probability theory*, Princeton, N.J. : Van Nostrand, 1955.

J.-F. Cavassilas, Stochastic matched filter, *Proceedings of the Institute of Acoustics (International Conference on Sonar Signal Processing)*, pp. 194-199, Vol. 13, Part 9, 1991.

F. Chaillan, C. Fraschini and P. Courmontagne, Speckle Noise Reduction in SAS Imagery, *Signal Processing*, Vol. 87, N° 4, 2007, pp. 762-781.

F. Chaillan, C. Fraschini and P. Courmontagne, Stochastic Matched Filtering Method Applied to SAS Imagery, *Proceedings of OCEANS'05*, pp. 233-238, Vol. 1, June 2005, Brest, France.

M. Tur, K.C. Chin, J.W. Goodman, When is speckle noise multiplicative?, *Applied optics*, Vol.21, N° 7, 1982, pp.1157-1159.

D. T. Kuan, A. A. Sawchuk, T. C. Strand and P. Chavel, Adaptive noise smoothing filter for images with signal-dependent noise, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 7, N° 2, 1985, pp. 165-177.

H. L. Vann Trees, *Detection, Estimation, and Modulation Theory, Part I.*, Wiley, New York, 1968.

---

[2] Courtesy of GESMA, France

C. Fraschini, F. Chaillan and P. Courmontagne, An improvement of the discriminating capability of the active SONAR by optimization of a criterion based on the Cramer-Rao lower bound, *Proceedings of OCEANS'05*, pp. 804-809, Vol. 2, June 2005, Brest, France.

P. Courmontagne, SAS Images De-noising: The Jointly Use of an Autoadaptive Mean Filter and the Stochastic Matched Filter, *Proceedings of OCEANS'07*, June 2007, Aberdeen, Scotland.

P. Courmontagne and F. Chaillan, The Adaptive Stochastic Matched Filter for SAS Images Denoising, *Proceedings of OCEANS'06*, September 2006, Boston, USA.

M. Kendall and A. Stuart, *The Advanced Theory of Statistics*, London: Charles Griffin, Vol. 2, 1979.

Q. Zhang and D. Liu, A simple capacity formula for correlated diversity Rician fading channels, *IEEE Communications Letters*, Vol. 6, N° 11, 2002, pp. 481-483.

Ph. Courmontagne, N. Vergnes and C. Jauffret, An optimal subspace projection for signal detection in noisy environment, *Proceedings of OCEANS'07*, September 2007, Vancouver, Canada.

Wireless fading channel models: from classical to stochastic differential equations

Mohammed Olama[1], Seddik Djouadi[2] and Charalambos Charalambous[3]

[1]*Oak Ridge National Laboratory*
[2]*University of Tennessee*
[3]*University of Cyprus*
[1,2]*USA*, [3]*Cyprus*

1. Introduction

The wireless communications channel constitutes the basic physical link between the transmitter and the receiver antennas. Its modeling has been and continues to be a tantalizing issue, while being one of the most fundamental components based on which transmitters and receivers are designed and optimized. The ultimate performance limits of any communication system are determined by the channel it operates in [1]. Realistic channel models are thus of utmost importance for system design and testing.

In addition to exponential power path-loss, wireless channels suffer from stochastic short term fading (STF) due to multipath, and stochastic long term fading (LTF) due to shadowing depending on the geographical area. STF corresponds to severe signal envelope fluctuations, and occurs in densely built-up areas filled with lots of objects like buildings, vehicles, etc. On the other hand, LTF corresponds to less severe mean signal envelope fluctuations, and occurs in sparsely populated or suburban areas [2-4]. In general, LTF and STF are considered as superimposed and may be treated separately [4].

Ossanna [5] was the pioneer to characterize the statistical properties of the signal received by a mobile user, in terms of interference of incident and reflected waves. His model was better suited for describing fading occurring mainly in suburban areas (LTF environments). It is described by the average power loss due to distance and power loss due to reflection of signals from surfaces, which when measured in dB's give rise to normal distributions, and this implies that the channel attenuation coefficient is log-normally distributed [4]. Furthermore, in mobile communications, the LTF channel models are also characterized by their special correlation characteristics which have been reported in [6-8].

Clarke [9] introduced the first comprehensive scattering model describing STF occurring mainly in urban areas. An easy way to simulate Clarke's model using a computer simulation is described in [10]. This model was later expanded to three-dimensions (3D) by Aulin [11]. An indoor STF was first introduced in [12]. Most of these STF models provide information on the frequency response of the channel, described by the Doppler power spectral density

(DPSD). Aulin [11] presented a methodology to find the Doppler power spectrum by computing the Fourier transform of the autocorrelation function of the channel impulse response with respect to time. A different approach, leading to the same Doppler power spectrum relation was presented by Gans [13]. These STF models suggest various distributions for the received signal amplitude such as Rayleigh, Rician, or Nakagami.

Models based on autoregressive and moving averages (AR) are proposed in [14, 15]. However, these models assume that the channel state is completely observable, which in reality is not the case due to additive noise, and requires long observation intervals. First order Markov models for Raleigh fading have been proposed in [16, 17], and the usefulness of a finite-state Markov channel model is argued in [18].

Mobile-to-mobile (or ad hoc) wireless networks comprise nodes that freely and dynamically self-organize into arbitrary and/or temporary network topology without any fixed infrastructure support [19]. They require direct communication between a mobile transmitter and a mobile receiver over a wireless medium. Such mobile-to-mobile communication systems differ from the conventional cellular systems, where one terminal, the base station, is stationary, and only the mobile station is moving. As a consequence, the statistical properties of mobile-to-mobile links are different from cellular ones [20, 21].

Copious ad hoc networking research exists on layers in the open system interconnection (OSI) model above the physical layer. However, neglecting the physical layer while modeling wireless environment is error prone and should be considered more carefully [22]. The experimental results in [23] show that the factors at the physical layer not only affect the absolute performance of a protocol, but because their impact on different protocols is non-uniform, it can even change the relative ranking among protocols for the same scenario. The importance of the physical layer is demonstrated in [24] by evaluating the Medium Access Control (MAC) performance.

Most of the research conducted on wireless channel modeling, such as [1-4, 25, 26], deals mainly with deterministic wireless channel models. In these models, the speeds of the nodes are assumed to be constant and the statistical characteristics of the received signal are assumed to be fixed with time. But in reality, the propagation environment varies continuously due to mobility of the nodes at variable speeds and movement of objects or scatter across transmitters and receivers resulting in appearance or disappearance of existing paths from one instant to the next. As a result, the current models that assume fixed statistics are unable to capture and track complex time variations in the propagation environment. These time variations compel us to introduce more advanced dynamical models based on stochastic differential equations (SDEs), in order to capture higher order dynamics of the wireless channels. The random variables characterizing the instantaneous power in static (deterministic) channel models are generalized to dynamical (stochastic) models including random processes with time-varying statistics [27-31]. The advantage of using SDE methods is due to computational simplicity simply because estimation and identification can be performed recursively and in real time. Parts of the results appearing in this chapter were presented in [27-31].

This chapter is organized as follows. In Section 2, the general time-varying (TV) wireless channel impulse response is introduced. The TV stochastic LTF, STF, and ad hoc wireless channel models are discussed in Sections 3, 4, and 5, respectively. Link performance for cellular and ad hoc channels is presented in Section 6. Finally, Section 7 provides the conclusion.

2. The General Time-Varying Wireless Channel Impulse Response

The impulse response (IR) of a wireless channel is typically characterized by time variations and time spreading [2]. Time variations are due to the relative motion between the transmitter and the receiver and temporal variations of the propagation environment. Time spreading is due to the fact that the emitted electromagnetic wave arrives at the receiver having undergone reflections, diffraction and scattering from various objects along the way, at different delay times. At the receiver, a random number of signal components, copies of a single emitted signal, arrive via different paths thus having undergone different attenuation, phase shifts and time delays, all of which are random and time-varying. This random number of signal components add vectorially giving rise to signal fluctuations, called multipath fading, which are responsible for the degradation of communication system performance.

The general time-varying (TV) model of a wireless fading channel is typically represented by the following multipath *low-pass* equivalent IR [2]

$$C_l(t;\tau) = \sum_{n=1}^{N(t)} r_n(t,\tau) e^{j\Phi_n(t,\tau)} \delta(\tau - \tau_n(t)) = \sum_{n=1}^{N(t)} \left(I_n(t,\tau) + jQ_n(t,\tau) \right) \delta(\tau - \tau_n(t)) \tag{1}$$

where $C_l(t;\tau)$ is the response of the channel at time t, due to an impulse applied at time $t - \tau$, $N(t)$ is the random number of multipath components impinging on the receiver, and the set $\{r_n(t,\tau), \Phi_n(t,\tau), \tau_n(t)\}_{n=1}^{N(t)}$ describes the random TV attenuation, overall phase shift, and arrival time of the different paths, respectively. $\{I_n(t,\tau), Q_n(t,\tau)\}_{n=1}^{N(t)} \triangleq$ $\{r_n(t,\tau)\cos\Phi_n(t,\tau), r_n(t,\tau)\sin\Phi_n(t,\tau)\}_{n=1}^{N(t)}$ are defined as the inphase and quadrature components of each path. Let $s_l(t)$ be the low-pass equivalent representation of the transmitted signal, then the *low-pass* equivalent representation of the received signal is given by

$$y_l(t) = \int_{-\infty}^{\infty} C_l(t;\tau) s_l(t - \tau) d\tau = \sum_{n=1}^{N(t)} r_n(t, \tau_n(t)) e^{j\Phi_n(t,\tau_n(t))} s_l(t - \tau_n(t)) \tag{2}$$

The multipath TV *band-pass* IR is given by [2]

$$C(t;\tau) = \mathrm{Re}\left\{ \sum_{n=1}^{N(t)} \left[r_n(t,\tau) e^{j\Phi_n(t,\tau)} \right] e^{j\omega_c t} \delta(\tau - \tau_n(t)) \right\} = \sum_{n=1}^{N(t)} \left(I_n(t,\tau)\cos\omega_c t - Q_n(t,\tau)\sin\omega_c t \right) \delta(\tau - \tau_n(t)) \tag{3}$$

where ω_c is the carrier frequency, and the *band-pass* representation of the received signal is given by

$$y(t) = \sum_{n=1}^{N(t)} \left(I_n\big(t, \tau_n(t)\big) \cos \omega_c t - Q_n\big(t, \tau_n(t)\big) \sin \omega_c t \right) s_l\big(t - \tau_n(t)\big) \qquad (4)$$

TV LTF, STF, and ad hoc dynamical channel models are considered in this chapter. The stochastic TV LTF channel modeling is discussed first in the next section.

3. Stochastic LTF Channel Modeling

3.1 The Traditional (Static) LTF Channel Model

In this section, we discuss the existing static models and introduce a general approach on how to derive dynamical models. Before introducing the dynamical LTF channel model that captures both space and time variations, we first summarize and interpret the traditional lognormal shadowing model, which serves as a basis in the development of the subsequent TV model. The traditional (time-invariant) power loss (PL) in dB for a given path is given by [4]

$$PL(d)[\text{dB}] := \overline{PL}(d_0)[\text{dB}] + 10\alpha \log\left(\frac{d}{d_0}\right) + \tilde{Z}, \qquad d \geq d_0 \qquad (5)$$

where $\overline{PL}(d_0)$ is the average PL in dB at a reference distance d_0 from the transmitter, the distance d corresponds to the transmitter-receiver separation distance, α is the path-loss exponent which depends on the propagating medium, and $\tilde{Z} \sim \mathcal{N}(0; \sigma^2)$ is a zero-mean Gaussian distributed random variable, which represents the variability of PL due to numerous reflections and possibly any other uncertainty of the propagating environment from one observation instant to the next. The average value of the PL described in (5) is

$$\overline{PL}(d)[\text{dB}] := \overline{PL}(d_0)[\text{dB}] + 10\alpha \log\left(\frac{d}{d_0}\right), \qquad d \geq d_0 \qquad (6)$$

The signal attenuation coefficient, denoted $r(d)$, represents how much the received signal magnitude is attenuated at a distance d with respect to the magnitude of the transmitted signal. It can be represented in terms of the power path loss as [4]

$$r(d) := e^{k \cdot PL(d)[\text{dB}]} \quad \text{where} \quad k = -\ln(10)/20 \qquad (7)$$

Since $PL(d)[\text{dB}]$ is normally distributed, it is clear that the attenuation coefficient, $r(d)$, is log-normally distributed. It can be noticed from (5)-(7) that the statistics of the PL and attenuation coefficient do not depend on time, and therefore these models treat PL as static (time-invariant). They do not take into consideration the relative motion between the transmitter and the receiver, or variations of the propagating environment due to mobility.

Such spatial and time variations of the propagating environment are captured herein by modeling the PL and the envelope of the received signal as random processes that are functions of space and time. Moreover, and perhaps more importantly, traditional models do not take into consideration the correlation properties of the PL in space and at different observation times. In reality, such correlation properties exist, and one way to model them is through stochastic processes, which obey specific type of SDEs.

3.2 Stochastic LTF Channel Models

In transforming the static model to a dynamical model, the random PL in (5) is relaxed to become a random process, denoted by $\{X(t,\tau)\}_{t\geq 0,\tau\geq\tau_0}$, which is a function of both time t and space represented by the time-delay τ, where $\tau = d/c$, d is the path length, c is the speed of light, $\tau_0 = d_0/c$ and d_0 is the reference distance. The signal attenuation is defined by $S(t,\tau) \triangleq e^{kX(t,\tau)}$, where $k = -\ln(10)/20$ [4]. For simplicity, we first introduce the TV lognormal model for a fixed transmitter-receiver separation distance d (or τ) that captures the temporal variations of the propagating environment. Next, we generalize it by allowing both t and τ to vary as the transmitter and receiver, as well as scatters, are allowed to move at variable speeds. This induces spatio-temporal variations in the propagating environment.

When τ is fixed, the proposed model captures the dependence of $\{X(t,\tau)\}_{t\geq 0,\tau\geq\tau_0}$ on time t. This corresponds to examining the time variations of the propagating environment for fixed transmitter-receiver separation distance. The process $\{X(t,\tau)\}_{t\geq 0,\tau\geq\tau_0}$ represents how much power the signal looses at a particular location as a function of time. However, since for a fixed distance d, the PL should be a function of distance, we choose to generate $\{X(t,\tau)\}_{t\geq 0,\tau\geq\tau_0}$ by a mean-reverting version of a general linear SDE given by [29, 30]

$$dX(t,\tau) = \beta(t,\tau)\big(\gamma(t,\tau) - X(t,\tau)\big)dt + \delta(t,\tau)dW(t),$$
$$X(t_0,\tau) \sim N(\overline{PL}(d)[dB];\sigma^2) \tag{8}$$

where $\{W(t)\}_{t\geq 0}$ is the standard Brownian motion (zero drift, unit variance) which is assumed to be independent of $X(t_0,\tau)$, $N(\mu;\kappa)$ denotes a Gaussian random variable with mean μ and variance κ, and $\overline{PL}(d)[dB]$ is the average path-loss in dB. The parameter $\gamma(t,\tau)$ models the average time-varying PL at distance d from the transmitter, which corresponds to $\overline{PL}(d)[dB]$ at d indexed by t. This model tracks and converges to $\gamma(t,\tau)$ as time progresses. The instantaneous drift $\beta(t,\tau)(\gamma(t,\tau) - X(t,\tau))$ represents the effect of pulling the process towards $\gamma(t,\tau)$, while $\beta(t,\tau)$ represents the speed of adjustment towards the mean. Finally, $\delta(t,\tau)$ controls the instantaneous variance or volatility of the process for the instantaneous drift.

Let $\{\theta(t,\tau)\}_{t\geq 0} \triangleq \{\beta(t,\tau),\gamma(t,\tau),\delta(t,\tau)\}_{t\geq 0}$. If the random processes in $\{\theta(t,\tau)\}_{t\geq 0}$ are measurable and bounded [32], then (8) has a unique solution for every $X(t_0,\tau)$ given by [30]

$$X(t,\tau) = e^{-\beta([t,t_0],\tau)}\left(X(t_0,\tau) + \int_{t_0}^{t} e^{\beta([u,t_0],\tau)}\big(\beta(u,\tau)\gamma(u,\tau)du + \delta(u,\tau)dW(u)\big) \right) \qquad (9)$$

where $\beta([t,t_0],\tau) \triangleq \int_{t_0}^{t}\beta(u,\tau)du$. Moreover, using Ito's stochastic differential rule [32] on

$S(t,\tau) = e^{kX(t,\tau)}$ the attenuation coefficient obeys the following SDE

$$dS(t,\tau) = S(t,\tau)\left[\left(k\beta(t,\tau)[\gamma(t,\tau)-\frac{1}{k}\ln S(t,\tau)] + \frac{1}{2}k^2\delta^2(t,\tau) \right)dt + k\delta(t,\tau)dW(t) \right]$$
$$S(t_0,\tau) = e^{kX(t_0,\tau)} \qquad (10)$$

This model captures the temporal variations of the propagating environment as the random parameters $\{\theta(t,\tau)\}_{t\geq 0}$ can be used to model the TV characteristics of the channel for the particular location τ. A different location is characterized by a different set of parameters $\{\theta(t,\tau)\}$.

Now, let us consider the special case when the parameters $\theta(t,\tau)$ are time invariant, i.e., $\theta(\tau) \triangleq \{\beta(\tau),\gamma(\tau),\delta(\tau)\}$. In this case we need to show that the expected value of the dynamic PL $X(t,\tau)$, denoted by $E[X(t,\tau)]$, converges to the traditional average PL in (6). The solution of the SDE model in (8) for the time-invariant case satisfies

$$X(t,\tau) = e^{-\beta(\tau)(t-t_0)}X(t_0,\tau) + \gamma(\tau)\left(1-e^{-\beta(\tau)(t-t_0)}\right) + \delta(\tau)\int_{t_0}^{t}e^{-\beta(\tau)(t-u)}dW(u) \qquad (11)$$

where for a given set of time-invariant parameters $\theta(\tau)$ and if the initial $X(t_0,\tau)$ is Gaussian or fixed, then the distribution of $X(t,\tau)$ is Gaussian with mean and variance given by [32]

$$E\big[X(t,\tau)\big] = \gamma(\tau)\left(1-e^{-\beta(\tau)(t-t_0)}\right) + e^{-\beta(\tau)(t-t_0)}E\{X(t_0,\tau)\}$$
$$Var\big[X(t,\tau)\big] = \delta^2(\tau)\left(\frac{1-e^{-2\beta(\tau)(t-t_0)}}{2\beta(\tau)}\right) + e^{-2\beta(\tau)(t-t_0)}Var\big(X(t_0,\tau)\big) \qquad (12)$$

Expression (12) of the mean and variance shows that the statistics of the communication channel model vary as a function of both time t and space τ. As the observation instant, t, becomes large, the random process $\{X(t,\tau)\}$ converges to a Gaussian random variable with mean $\gamma(\tau) = \overline{PL}(d)[\text{dB}]$ and variance $\delta^2(\tau)/2\beta(\tau)$. Therefore, the traditional lognormal model in (5) is a special case of the general TV LTF model in (8). Moreover, the distribution of $S(t,\tau) = e^{kX(t,\tau)}$ is lognormal with mean and variance

$$E\big[S(t,\tau)\big] = \exp\left(\frac{2kE\big[X(t,\tau)\big] + k^2 Var\big[X(t,\tau)\big]}{2}\right)$$

$$Var\big[S(t,\tau)\big] = \exp\left(2kE\big[X(t,\tau)\big] + 2k^2 Var\big[X(t,\tau)\big]\right) - \exp\left(2kE\big[X(t,\tau)\big] + k^2 Var\big[X(t,\tau)\big]\right)$$

(13)

Now, lets go back to the more general case in which $\{\theta(t,\tau)\}_{t\geq0} \triangleq \{\beta(t,\tau), \gamma(t,\tau), \delta(t,\tau)\}_{t\geq0}$. At a particular location τ, the mean of the PL process $E[X(t,\tau)]$ is required to track the time variations of the average PL. This is illustrated in the following example.

Example 1 [30]: Let

$$\gamma(t,\tau) = \gamma_m(\tau)\left(1 + 0.15e^{-2t/T}\sin\left(\frac{10\pi t}{T}\right)\right)$$

(14)

where $\gamma_m(\tau)$ is the average PL at a specific location τ, T is the observation interval, $\delta(t,\tau) =$ 1400 and $\beta(t,\tau) =$ 225000 (these parameters are determined from experimental measurements), where for simplicity $\delta(t,\tau)$ and $\beta(t,\tau)$ are chosen to be constant, but in general they are functions of both t and τ. The variations of $X(t,\tau)$ as a function of distance and time are represented in Figure 1. The temporal variations of the environment are captured by a TV $\gamma(t,\tau)$ which fluctuates around different average PLs γ_m's, so that each curve corresponds to a different location. It is noticed in Figure 1 that as time progresses, the process $X(t,\tau)$ is pulled towards $\gamma(t,\tau)$. The speed of adjustment towards $\gamma(t,\tau)$ can be controlled by choosing different values of $\beta(t,\tau)$.

Next, the general spatio-temporal lognormal model is introduced by generalizing the previous model to capture both space and time variations, using the fact that $\gamma(t,\tau)$ is a function of both t and τ. In this case, besides initial distances, the motion of mobiles, i.e., their velocities and directions of motion with respect to their base stations are important factors to evaluate TV PLs for the links involved. This is illustrated in a simple way for the case of a single transmitter and a single receiver as follows: Consider a base station (receiver) at an initial distance d from a mobile (transmitter) that moves with a certain constant velocity υ in a direction defined by an arbitrary constant angle θ, where θ is the angle between the direction of motion of the mobile and the distance vector that starts from

the receiver towards the transmitter as shown in Figure 2. At time t, the distance from the transmitter to the receiver, $d(t)$, is given by
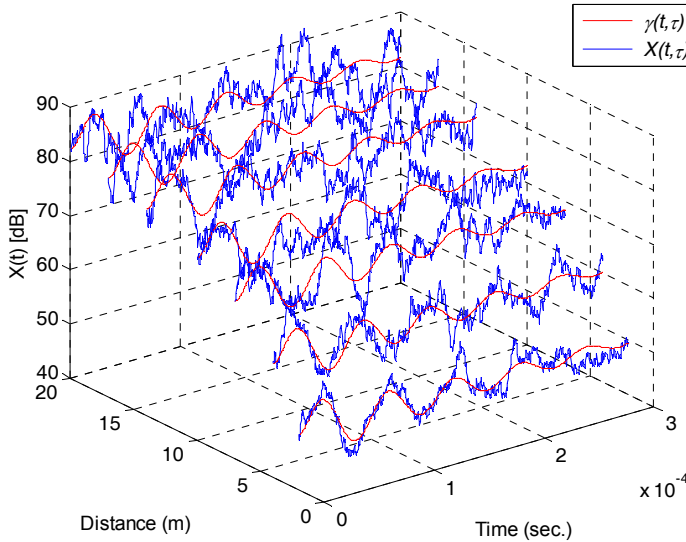

Fig. 1. Mean-reverting power path-loss as a function of t and τ, for the time-varying $\gamma(t,\tau)$ in Example 1.
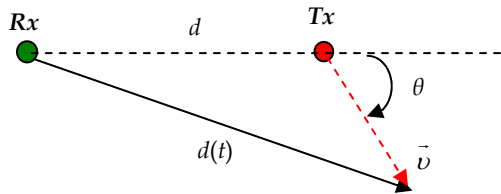

Fig. 2. A transmitter at a distance d from a receiver moves with a velocity υ and in the direction given by θ with respect to the transmitter-receiver axis.

$$d(t) = \sqrt{(d + t\upsilon\cos\theta)^2 + (t\upsilon\sin\theta)^2} = \sqrt{d^2 + (\upsilon t)^2 + 2dt\upsilon\cos\theta} \tag{15}$$

Therefore, the average PL at that location is given by

$$\gamma(t,\tau) = \overline{PL}(d(t))[dB] = \overline{PL}(d_0)[dB] + 10\alpha\log\frac{d(t)}{d_0} + \xi(t), \quad d(t) \geq d_0 \tag{16}$$

where $\overline{PL}(d_0)$ is the average PL in dB at a reference distance d_0, $d(t)$ is defined in (15), α is the path-loss exponent and $\xi(t)$ is an arbitrary function of time representing additional

temporal variations in the propagating environment like the appearance and disappearance of additional scatters.

Now, suppose the mobile moves with an arbitrary velocity, $\left(v_x(t), v_y(t)\right)$, in the x-y plane, where $v_x(t)$, $v_y(t)$ denote the instantaneous velocity components in the x and y directions, respectively. The instantaneous distance from the receiver is thus described by

$$d(t) = \sqrt{\left(d + \int_0^t v_x(t)dt\right)^2 + \left(\int_0^t v_y(t)dt\right)^2} \tag{17}$$

The parameter $\gamma(t,\tau)$ is used in the TV lognormal model (8) to obtain a general spatio-temporal lognormal channel model. This is illustrated in the following example.

Example 2 [30]: Consider a mobile moving at sinusoidal velocity with average speed 80 Km/hr, initial distance $d=50$ meters, $\theta=135$ degrees, and $\xi(t)=0$. Figure 3 shows the mean reverting PL $X(t,\tau)$, $\gamma(t,\tau)$, $E[X(t,\tau)]$, and the velocity of the mobile $\upsilon(t)$ and distance $d(t)$ as a function of time. It can be seen that the mean of $X(t,\tau)$ coincides with the average PL $\gamma(t,\tau)$ and tracks the movement of the mobile. Moreover, the variation of $X(t,\tau)$ is due to uncertainties in the wireless channel such as movements of objects or obstacles between transmitter and receiver that are captured by the spatio-temporal lognormal model (8) and (16). Additional time variations of the propagating environment, while the mobile is moving, can be captured by using the TV PL coefficient $\alpha(t)$ in (16) in addition to the TV parameters $\beta(t,\tau)$ and $\delta(t,\tau)$, or simply by $\xi(t)$. The stochastic STF channel model is discussed in the next section.
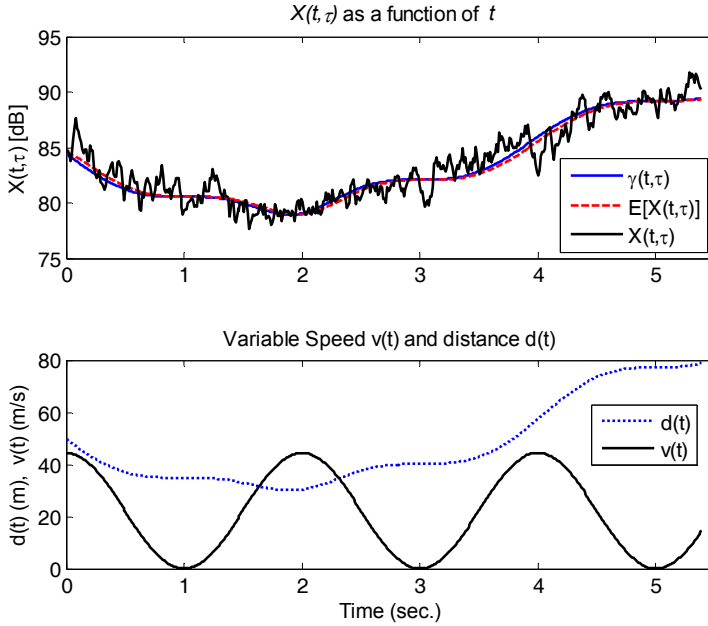
Fig. 3. Mean-reverting power path-loss $X(t,\tau)$ for the TV LTF wireless channel model in Example 2.

4. Stochastic STF Channel Modeling

4.1 The Deterministic DPSD of Wireless Channels

The traditional STF model is based on Ossanna [5] and later Clarke [9] and Aulin's [11] developments. Aulin's model is shown in Figure 4. This model assumes that at each point between a transmitter and a receiver, the total received wave consists of the superposition of N plane waves each having traveled via a different path. The nth wave is characterized by its field vector $E_n(t)$ given by [11]

$$E_n(t) = \text{Re}\left\{r_n(t)e^{j\Phi_n(t)}e^{j\omega_c t}\right\} = I_n(t)\cos\omega_c t - Q_n(t)\sin\omega_c t \tag{18}$$

where $\{I_n(t), Q_n(t)\}$ are the inphase and quadrature components for the nth wave, respectively, $r_n(t) = \sqrt{I_n^2(t) + Q_n^2(t)}$ is the signal envelope, $\Phi_n(t) = \tan^{-1}(Q_n(t)/I_n(t))$ is the phase, and ω_c is the carrier frequency. The total field $E(t)$ can be written as

$$E(t) = \sum_{n=1}^{N} E_n(t) = I(t)\cos\omega_c t - Q(t)\sin\omega_c t \tag{19}$$

where $\{I(t), Q(t)\}$ are inphase and quadrature components of the total wave, respectively, with $I(t) = \sum_{n=1}^{N} I_n(t)$ and $Q(t) = \sum_{n=1}^{N} Q_n(t)$. An application of the central limit theorem
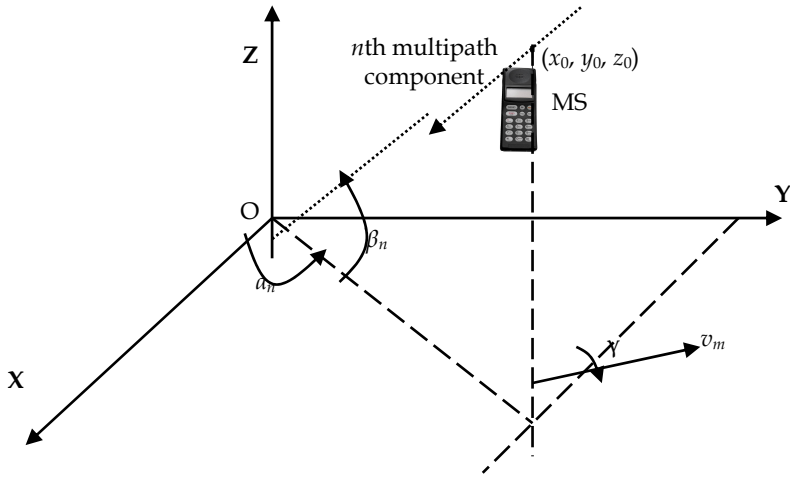
Fig. 4. Aulin's 3D multipath channel model.

states that for large N, the inphase and quadrature components have Gaussian distributions $\mathcal{N}(\overline{x};\sigma^2)$ [9]. The mean is $\overline{x} = \mathrm{E}\{I(t)\} = \mathrm{E}\{Q(t)\}$ and the variance is $\sigma^2 = \mathrm{Var}\{I(t)\} = \mathrm{Var}\{Q(t)\}$. In the case where there is non-line-of-sight (NLOS), then the mean $\overline{x} = 0$ and the received signal amplitude has Rayleigh distribution. In the presence of line-of-sight (LOS) component, $\overline{x} \neq 0$ and the received signal is Rician distributed. Also, it is assumed that $I(t)$ and $Q(t)$ are uncorrelated and thus independent since they are Gaussian distributed [11].

Dependent on the mobile speed, wavelength, and angle of incidence, the Doppler frequency shifts on the multipath rays give rise to a DPSD. The DPSD is defined as the Fourier transform of the autocorrelation function of the channel, and represents the amount of power at various frequencies. Define $\{\alpha_n,\beta_n\}$ as the direction of the incident wave onto the receiver as illustrated in Figure 4. For the case when α_n is uniformly distributed and β_n is fixed, the deterministic DPSD, $S(f)$, is given by [25]

$$S(f) = \begin{cases} \dfrac{E_0}{4\pi} \dfrac{1/f_m}{\sqrt{1-\left(\dfrac{f}{f_m}\right)^2}} \ , & |f| < f_m \\ \\ 0 \ , & \text{otherwise} \end{cases} \tag{20}$$

where f_m is the maximum Doppler frequency, and $E_0/2 = \mathrm{Var}\{I(t)\} = \mathrm{Var}\{Q(t)\}$. A more complex, but realistic, expression for the DPSD, which assumes β_n has probability density function $p_\beta(\beta) = \dfrac{\cos\beta}{2\sin\beta_m}$ where $|\beta| \leq |\beta_m| \leq \dfrac{\pi}{2}$, and for small angles β_m, is given by [11]

$$S(f) = \begin{cases} 0, & |f| > f_m \\ \dfrac{E_0}{4 f_m \sin \beta_m}, & f_m \cos \beta_m \le |f| \le f_m \\ \dfrac{E_0}{4\pi f_m \sin \beta_m} \left[\dfrac{\pi}{2} - \sin^{-1} \left(\dfrac{2\cos^2 \beta_m - 1 - \left(f / f_m \right)^2}{1 - \left(f / f_m \right)^2} \right) \right], & |f| < f_m \cos \beta_m \end{cases} \qquad (21)$$

Expression (21) is illustrated in Figure 5 for different values of mobile speed. Notice that the direction of motion does not play a role because of the uniform scattering assumption, and that the DPSDs described in (20) and (21) are band limited.

The DPSD is the fundamental channel characteristic on which STF dynamical models are based on. The approach presented here is based on traditional system theory using the state space approach [33] while capturing the spectral characteristics of the channel. The main idea in constructing dynamical models for STF channels is to factorize the deterministic DPSD into an approximate nth order even transfer function, and then use a stochastic realization [32] to obtain a state space representation for the inphase and quadrature components.

The wireless channel is considered as a dynamical system for which the input-output map is described in (1) and (3). In practice, one obtains from measurements the power spectral density of the output, and with the knowledge of the power spectral density of the input the power spectral density of the transfer function (wireless channel) can be deduced as

$$S_{yy}\left(f \right) = \left| H\left(f \right) \right|^2 S_{xx}\left(f \right) \qquad (22)$$
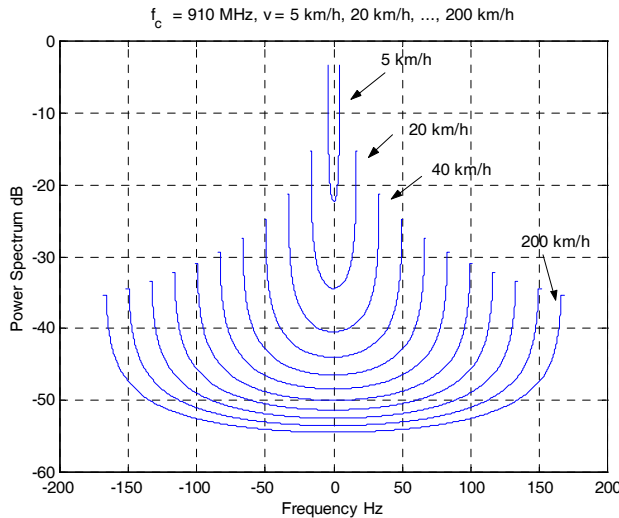
Fig. 5. DPSD for different values of mobile speed (β_m =10 degrees).

where $x(t)$ is a random process with power spectral density $S_{xx}(f)$ representing the input signal to the channel, $y(t)$ is a random process with power spectral density $S_{yy}(f)$ representing the output signal of the channel, and $H(f)$ is the frequency response of the channel, which is the Fourier transform of the channel IR.

In general, in order to identify the random process associated with the DPSD, $S(f)$, in (20) or (21) in the form of an SDE, we need to find a transfer function, $H(f)$ whose magnitude square equals $S(f)$, i.e. $S(f)=|H(f)|^2$. This is equivalent to $S(s)=H(s)H(-s)$, where $s = i2\pi f$ and $i = \sqrt{-1}$. That is, we need to factorize the DPSD. This is an old problem which had been studied by Paley and Wiener [34] and is reformulated here as follows:

Given a non-negative integrable function, $S(f)$, such that the Paley-Wiener condition $\int_{-\infty}^{\infty} \left[\frac{|\log S(f)|}{(1+f^2)} \right] df < \infty$ is satisfied, then there exists a causal, stable, minimum-phase function $H(f)$, such that $|H(f)|^2 = S(f)$, implying that $S(f)$ is factorizable, namely, $S(s)=H(s)H(-s)$. It can be seen that the Paley-Wiener condition is not satisfied when $S(f)$ is band limited (and therefore it is not factorizable), which is the case for wireless links. In order to factorize it, the deterministic DPSD has to be first approximated by a *rational* transfer function, denoted $\tilde{S}(f)$, and is discussed next.

4.2 Approximating the Deterministic DPSD

A number of rational approximation methods can be used to approximate the deterministic DPSD [35], the choice of which depends on the complexity and the required accuracy. The order of approximation dictates how close the approximate curve would be to the actual one. Higher order approximations capture higher order dynamics, and provide better approximations for the DPSD, however computations become more involved. In this section, we consider a simple approximating method which uses a 4th order stable, minimum phase, real, rational approximate transfer function. In Section 5.2, we consider the complex cepstrum approximation algorithm [36], which is based on the Gauss-Newton method for iterative search, and is more accurate than the simple approximating method but requires more computations.

In the simple approximating method, a 4th order even transfer function $\tilde{S}(s)$, is used to approximate the deterministic cellular DPSD, $S(s)$. The approximate function $\tilde{S}(s) = H(s)H(-s)$ is given by [28]

$$\tilde{S}(s) = \frac{K^2}{s^4 + 2\omega_n^2(1 - 2\zeta^2)s^2 + \omega_n^4}, \qquad H(s) = \frac{K}{s^2 + 2\zeta\omega_n s + \omega_n^2} \qquad (23)$$

Equation (23) has three arbitrary parameters $\{\zeta, \omega_n, K\}$, which can be adjusted such that the approximate curve coincides with the actual curve at different points. The reason for presenting 4th order approximation of the DPSD is that we can compute explicit expressions for the constants $\{\zeta, \omega_n, K\}$ as functions of specific points on the data-graphs of the DPSD. In fact, if the approximate density $\tilde{S}(f)$ coincides with the exact density $S(f)$ at $f=0$ and $f = f_{max}$, then the arbitrary parameters $\{\zeta, \omega_n, K\}$ are computed explicitly as

$$\zeta = \sqrt{\frac{1}{2}\left(1 - \sqrt{1 - \frac{S(0)}{S(f_{max})}}\right)}, \ \omega_n = \frac{2\pi f_{max}}{\sqrt{1 - 2\zeta^2}}, \ K = \omega_n^2 \sqrt{S(0)} \tag{24}$$

Figure 6 shows $S(f)$ and its approximation $\tilde{S}(f)$ via a 4th order even function. In the next section, the approximated DPSD is used to develop stochastic STF channel models.

4.3 Stochastic STF Channel Models

A stochastic realization is used here to obtain a state space representation for the inphase and quadrature components [32]. The SDE, which corresponds to $H(s)$ in (23) is

$$d^2 x(t) + 2\zeta\omega_n dx(t) + \omega_n^2 x(t)dt = KdW(t), \quad \dot{x}(0), x(0) \ \text{are given} \tag{25}$$

where $\{dW(t)\}_{t \geq 0}$ is a white-noise process. Equation (25) can be rewritten in terms of inphase and quadrature components as

$$d^2 x_I(t) + 2\zeta\omega_n dx_I(t) + \omega_n^2 x_I(t)dt = KdW_I(t), \quad \dot{x}_I(0), x_I(0) \ \text{are given}$$
$$d^2 x_Q(t) + 2\zeta\omega_n dx_Q(t) + \omega_n^2 x_Q(t)dt = KdW_Q(t), \quad \dot{x}_Q(0), x_Q(0) \ \text{are given} \tag{26}$$

where $\{dW_I(t)\}_{t \geq 0}$ and $\{dW_Q(t)\}_{t \geq 0}$ are two independent and identically distributed (i.i.d.) white Gaussian noises.
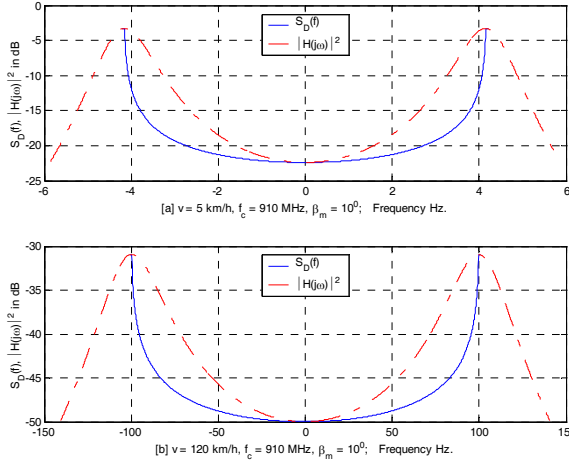
Fig. 6. DPSD, $S_D(f)$, and its approximation $\tilde{S}(\omega) = |H(j\omega)|^2$ via a 4th order transfer function for mobile speed of (a) 5 km/hr and (b) 120 km/hr.

Several stochastic realizations [32] can be used to obtain a state space representation for the inphase and quadrature components of STF channel models. For example, the stochastic observable canonical form (OCF) realization [33] can be used to realize (26) for the inphase and quadrature components for the jth path as

$$
\begin{bmatrix} dX^1_{I,j}(t) \\ dX^2_{I,j}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & -2\xi_n\omega_n \end{bmatrix}\begin{bmatrix} X^1_{I,j}(t) \\ X^2_{I,j}(t) \end{bmatrix}dt + \begin{bmatrix} 0 \\ K \end{bmatrix}dW^I_j(t), \quad \begin{bmatrix} X^1_{I,j}(0) \\ X^2_{I,j}(0) \end{bmatrix},
$$
$$
I_j(t) = \begin{bmatrix} 1 & 0 \end{bmatrix}\begin{bmatrix} X^1_{I,j}(t) \\ X^2_{I,j}(t) \end{bmatrix} + f^I_j(t),
$$
$$
\begin{bmatrix} dX^1_{Q,j}(t) \\ dX^2_{Q,j}(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & -2\xi_n\omega_n \end{bmatrix}\begin{bmatrix} X^1_{Q,j}(t) \\ X^2_{Q,j}(t) \end{bmatrix}dt + \begin{bmatrix} 0 \\ K \end{bmatrix}dW^Q_j(t), \quad \begin{bmatrix} X^1_{Q,j}(0) \\ X^2_{Q,j}(0) \end{bmatrix},
$$
$$
Q_j(t) = \begin{bmatrix} 1 & 0 \end{bmatrix}\begin{bmatrix} X^1_{Q,j}(t) \\ X^2_{Q,j}(t) \end{bmatrix} + f^Q_j(t)
$$

(27)

where $X_{I,j}(t) = [X^1_{I,j}(t)\ X^2_{I,j}(t)]^T$ and $X_{Q,j}(t) = [X^1_{Q,j}(t)\ X^2_{Q,j}(t)]^T$ are state vectors of the inphase and quadrature components. $I_j(t)$ and $Q_j(t)$ correspond to the inphase and quadrature components, respectively, $\{W^I_j(t)\}_{t\geq 0}$ and $\{W^Q_j(t)\}_{t\geq 0}$ are independent standard Brownian motions, which correspond to the inphase and quadrature components of the jth path respectively, the parameters $\{\zeta, \omega_n, K\}$ are obtained from the approximation of the

deterministic DPSD, and $f_j^I(t)$ and $f_j^Q(t)$ are arbitrary functions representing the LOS of the inphase and quadrature components respectively, characterizing further dynamic variations in the environment.

Expression (27) for the jth path can be written in compact form as

$$\begin{bmatrix} dX_I(t) \\ dX_Q(t) \end{bmatrix} = \begin{bmatrix} A_I & 0 \\ 0 & A_Q \end{bmatrix} \begin{bmatrix} X_I(t) \\ X_Q(t) \end{bmatrix} dt + \begin{bmatrix} B_I & 0 \\ 0 & B_Q \end{bmatrix} \begin{bmatrix} dW_I(t) \\ dW_Q(t) \end{bmatrix}$$
$$\begin{bmatrix} I(t) \\ Q(t) \end{bmatrix} = \begin{bmatrix} C_I & 0 \\ 0 & C_Q \end{bmatrix} \begin{bmatrix} X_I(t) \\ X_Q(t) \end{bmatrix} + \begin{bmatrix} f_I(t) \\ f_Q(t) \end{bmatrix} \tag{28}$$

where

$$A_I = A_Q = \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & -2\xi_n\omega_n \end{bmatrix}, \ B_I = B_Q = \begin{bmatrix} 0 \\ K \end{bmatrix}, \ C_I = C_Q = \begin{bmatrix} 1 & 0 \end{bmatrix} \tag{29}$$

$\{W_I(t), W_Q(t)\}_{t \geq 0}$ are independent standard Brownian motions which are independent of the initial random variables $X_I(0)$ and $X_Q(0)$, and $\{f_I(s), f_Q(s); 0 \leq s \leq t\}$ are random processes representing the inphase and quadrature LOS components, respectively. The band-pass representation of the received signal corresponding to the jth path is given as

$$y(t) = \left[\left(C_I X_I(t) + f_I(t) \right) \cos \omega_c t - \left(C_Q X_Q(t) + f_Q(t) \right) \sin \omega_c t \right] s_l \left(t - \tau_j \right) + v(t) \tag{30}$$

where $v(t)$ is the measurement noise. As the DPSD varies from one instant to the next, the channel parameters $\{\zeta, \omega_n, K\}$ also vary in time, and have to be estimated on-line from time domain measurements. Without loss of generality, we consider the case of flat fading, in which the mobile-to-mobile channel has purely multiplicative effect on the signal and the multipath components are not resolvable, and can be considered as a single path [2]. The frequency selective fading case can be handled by including multiple time-delayed echoes. In this case, the delay spread has to be estimated. A sounding device is usually dedicated to estimate the time delay of each discrete path such as Rake receiver [26]. Following the state space representation in (28) and the band pass representation of the received signal in (30), the fading channel can be represented using a general stochastic state space representation of the form [28]

$$dX(t) = A(t)X(t)dt + B(t)dW(t)$$
$$y(t) = C(t)X(t) + D(t)v(t) \tag{31}$$

where

$$X(t) = \begin{bmatrix} X_I(t) & X_Q(t) \end{bmatrix}^T, \ A(t) = \begin{bmatrix} A_I(t) & 0 \\ 0 & A_Q(t) \end{bmatrix}, \ B(t) = \begin{bmatrix} B_I(t) & 0 \\ 0 & B_Q(t) \end{bmatrix},$$

$$C(t) = \begin{bmatrix} \cos(\omega_c t)C_I & -\sin(\omega_c t)C_Q \end{bmatrix}, \ D(t) = \begin{bmatrix} \cos(\omega_c t) & -\sin(\omega_c t) \end{bmatrix}$$

$$v(t) = \begin{bmatrix} v_I(t) & v_Q(t) \end{bmatrix}^T, \ dW(t) = \begin{bmatrix} dW^I(t) & dW^Q(t) \end{bmatrix}^T$$

$$(32)$$

In this case, $y(t)$ represents the received signal measurements, $X(t)$ is the state variable of the inphase and quadrature components, and $v(t)$ is the measurement noise.

Time domain simulation of STF channels can be performed by passing two independent white noise processes through two identical filters, $\tilde{H}(s)$, obtained from the factorization of the deterministic DPSD, one for the inphase and the other for the quadrature component [4], and realized in their state-space form as described in (28) and (29).

Example 3: Consider a flat fading wireless channel with the following parameters: $f_c = 900\,\text{MHz}$, $v = 80 \text{ km/h}$, $\beta_m = 10^\circ$, and $f_j^I(t) = f_j^Q(t) = 0$. Time domain simulation of the inphase and quadrature components, attenuation coefficient, phase angle, input signal, and received signal are shown in Figures 7-9. The inphase and quadrature components have been produced using (28) and (29), while the received signal is reproduced using (30). The simulation of the dynamical STF channel is performed using Simulink in Matlab [37].

4.4 Solution to the Stochastic State Space Model
The stochastic TV state space model described in (31) and (32) has a solution [32, 38]

$$X_L(t) = \Phi_L(t,t_0)\left[X_L(t_0) + \int_{t_0}^{t} \Phi_L^{-1}(u,t_0)B_L(u)dW_L(u) \right]$$

$$(33)$$

where $L = I$ or Q, and $\Phi_L(t,t_0)$ is the fundamental matrix, which satisfies $\dot{\Phi}_L(t,t_0) = A_L(t)\Phi_L(t,t_0)$ and $\Phi_L(t_0,t_0)$ is the identity matrix.
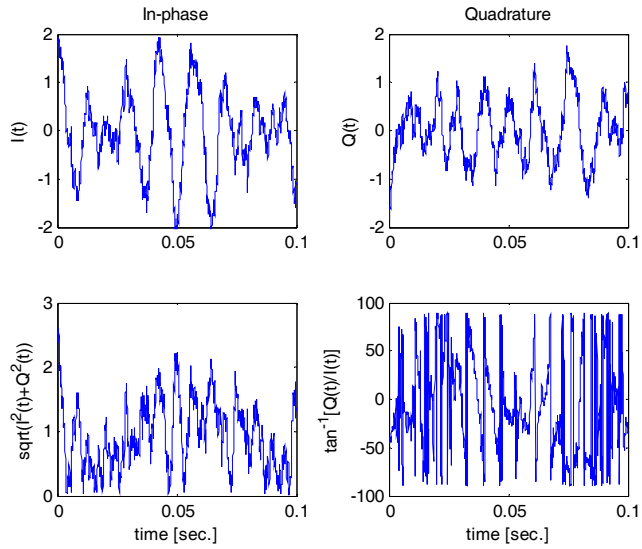
Fig. 7. Inphase and quadrature components, attenuation coefficient, and phase angle of the STF wireless channel in Example 3.
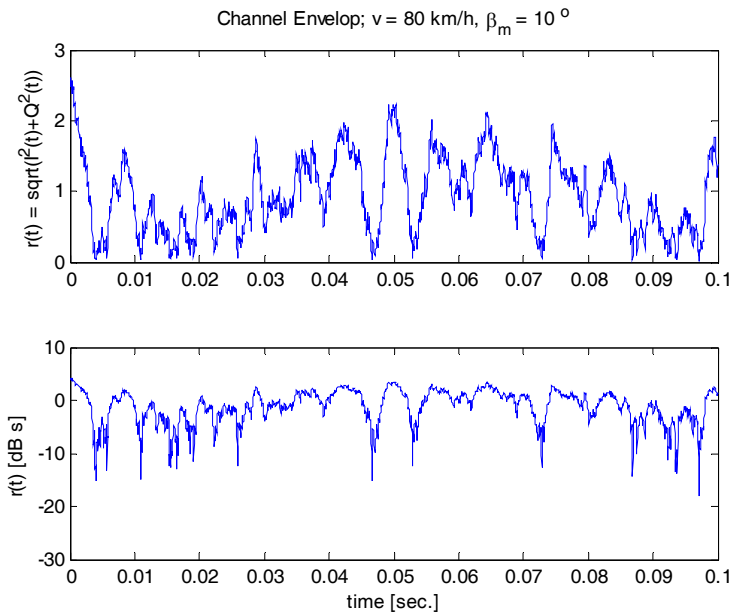

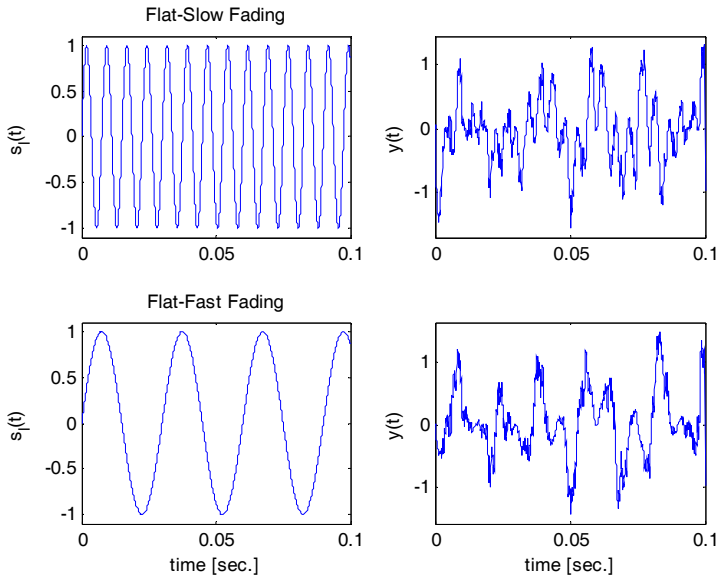Fig. 8. Attenuation coefficient in absolute units and in dB's for the STF wireless channel in Example 3.

Fig. 9. Input signal, $s_l(t)$, and the corresponding received signal, $y(t)$, for flat slow fading (top) and flat fast fading conditions (bottom).

Further computations show that the mean of $X_L(t)$ is given by [32]

$$E\big[X_L(t)\big] = \Phi_L(t,t_0)E\big[X_L(t_0)\big] \tag{34}$$

and the covariance matrix of $X_L(t)$ is given by

$$\Sigma_L(t) = \Phi_L(t,t_0)\left[Var\big[X_L(t_0)\big] + \int_{t_0}^{t}\Phi_L^{-1}(u,t_0)B_L(u)B_L^T(u)\big(\Phi_L^{-1}(u,t_0)\big)^T du\right]\Phi_L^T(t,t_0) \tag{35}$$

Differentiating (35) shows that $\Sigma_L(t)$ satisfies the Riccati equation

$$\dot{\Sigma}_L(t) = A(t)\Sigma_L(t) + \Sigma_L(t)A^T(t) + B(t)B^T(t) \tag{36}$$

For the time-invariant case, $A_L(t) = A_L$ and $B_L(t) = B_L$, equations (33)-(35) simplify to

$$X_L(t) = e^{A_L(t-t_0)} X_L(t_0) + \int_{t_0}^{t} e^{A_L(t-u)} B_L \, dW_L(u)$$

$$E\left[X_L(t)\right] = e^{A_L(t-t_0)} E\left[X_L(t_0)\right] \tag{37}$$

$$\Sigma_L(t) = e^{A_L(t-t_0)} Var\left[X_L(t_0)\right] e^{A_L^T(t-t_0)} + \int_{t_0}^{t} e^{A_L(t-u)} B_L B_L^T e^{A_L^T(t-u)} du$$

It can be seen in (34) and (35) that the mean and variance of the inphase and quadrature components are functions of time. Note that the statistics of the inphase and quadrature components, and therefore the statistics of the STF channel, are time varying. Therefore, these stochastic state space models reflect the TV characteristics of the STF channel. Following the same procedure in developing the STF channel models, the stochastic TV ad hoc channel models are developed in the next section.

5. Stochastic Ad Hoc Channel Modeling

5.1 The Deterministic DPSD of Ad Hoc Channels

Dependent on mobile speed, wavelength, and angle of incidence, the Doppler frequency shifts on the multipath rays give rise to a DPSD. The cellular DPSD for a received fading carrier of frequency f_c is given in (20) and can be described by [25]

$$\frac{S(f)}{pG / \pi f_1} = \begin{cases} \dfrac{1}{\sqrt{1 - \left(\dfrac{f - f_c}{f_1}\right)^2}} & , \ |f - f_c| < f_1 \\[2ex] 0 & , \ \text{otherwise} \end{cases} \tag{38}$$

where f_1 is the maximum Doppler frequency of the mobile , p is the average power received by an isotropic antenna, and G is the gain of the receiving antenna. For a mobile-to-mobile (or ad hoc) link, with f_1 and f_2 as the sender and receiver's maximum Doppler frequencies, respectively, the degree of double mobility, denoted by α is defined by $\alpha = \left[\min(f_1, f_2) / \max(f_1, f_2)\right]$, so $0 \le \alpha \le 1$, where $\alpha = 1$ corresponds to a full double mobility and $\alpha = 0$ to a single mobility like cellular link, implying that cellular channels are a special case of mobile-to-mobile channels. The corresponding deterministic mobile-to-mobile DPSD is given by [39-41]

$$\frac{S(f)}{(pG)^2 / \pi^2 f_m \sqrt{\alpha}} = \begin{cases} K\left(\dfrac{1+\alpha}{2\sqrt{\alpha}} \sqrt{1 - \left(\dfrac{f - f_c}{(1+\alpha)f_m}\right)^2}\right) & , |f - f_c| < (1+\alpha)f_m \\[2ex] 0 & , \text{otherwise} \end{cases} \tag{39}$$

where $K(\cdot)$ is the complete elliptic integral of the first kind, and $f_m = \max(f_1, f_2)$. Figure 10 shows the deterministic mobile-to-mobile DPSDs for different values of a's. Thus, a generalized DPSD has been found where the U-shaped spectrum of cellular channels is a special case.

Here, we follow the same procedure in deriving the stochastic STF channel models in Section 4. The deterministic ad hoc DPSD is first factorized into an approximate nth order even transfer function, and then use a stochastic realization [32] to obtain a state space representation for inphase and quadrature components. The complex cepstrum algorithm [36] is used to approximate the ad hoc DPSD and is discussed next.
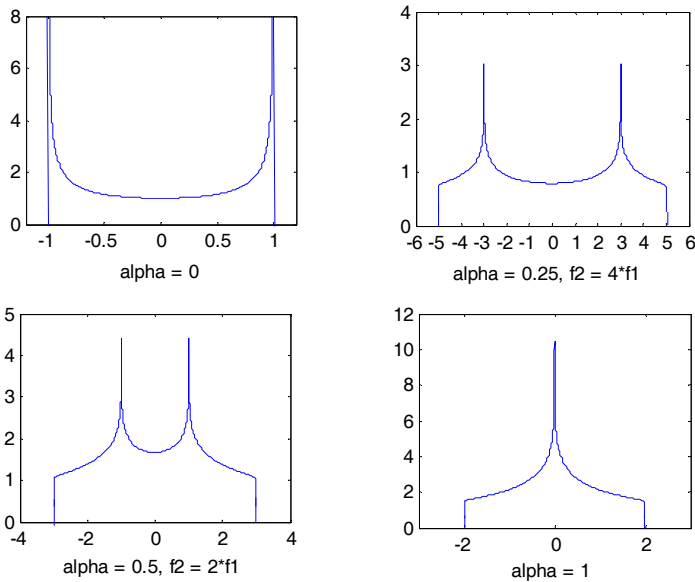
Fig. 10. Ad hoc deterministic DPSDs for different values of α's, with parameters $f_c = 0$, $f_1 = 1$, and $pG = \pi$.

5.2 Approximating the Deterministic Ad Hoc DPSD

Since the ad hoc DPSD is more complicated than the cellular one, we propose to use a more complex and accurate approximation method: The complex cepstrum algorithm [36]. It uses several measured points of the DPSD instead of just three points as in the simple method (described in Section 4.2). It can be explained briefly as follows: On a log-log scale, the magnitude data is interpolated linearly, with a very fine discretization. Then, using the complex cepstrum algorithm [36], the phase, associated with a stable, minimum phase, real, rational transfer function with the same magnitude as the magnitude data is generated.

With the new phase data and the input magnitude data, a real rational transfer function can be found by using the Gauss-Newton method for iterative search [35], which is used to

generate a stable, minimum phase, real rational transfer function, denoted by $\tilde{H}(s)$, to identify the best model from the data of $H(f)$ as

$$\min_{a,b} \sum_{k=1}^{l} wt(f_k)\left|H(f_k)-\tilde{H}(f_k)\right|^2 \tag{40}$$

where

$$\tilde{H}(s)=\frac{b_{m-1}s^{m-1}+...+b_1 s+b_0}{s^m+a_{m-1}s^{m-1}+...+a_1 s+a_0} \tag{41}$$

$b=\{b_{m-1},...,b_0\}$, $a=\{a_{m-1},...,a_0\}$, $wt(f)$ is the weight function, and l is the number of frequency points. Several variants have been suggested in the literature, where the weighting function gives less attention to high frequencies [35]. This algorithm is based on Levi [42]. Figure 11 shows the DPSD, $S(f)$, and its approximation $\tilde{S}(f)$ via different orders using complex cepstrum algorithm. The higher the order of $\tilde{S}(f)$ the better the approximation obtained. It can be seen that approximation with a 4th order transfer function gives a very good approximation.
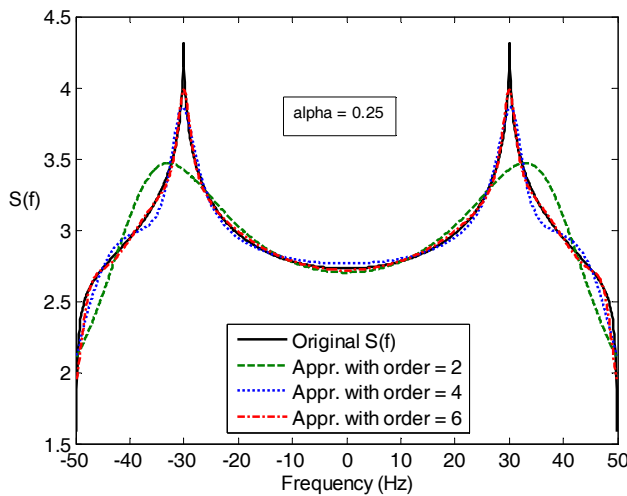
Fig. 11. DPSD, $S(f)$, and its approximations, $\tilde{S}(f)$, using complex cepstrum algorithm for different orders of $\tilde{S}(f)$.

Figure 12(a) and 12(b) show the DPSD, $S(f)$, and its approximation $\tilde{S}(f)$ using the complex cepstrum and simple approximation methods, respectively, for different values of α's via 4th order even function. It can be noticed that the former gives better approximation

than the latter; since it employs all measured points of the DPSD instead of just three points in the simple method.
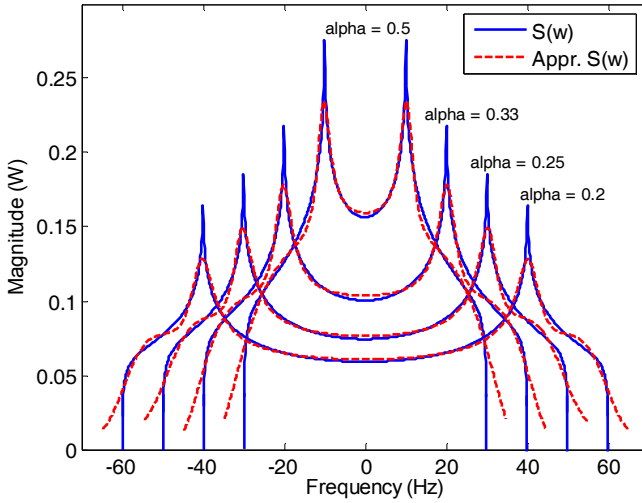
(a)
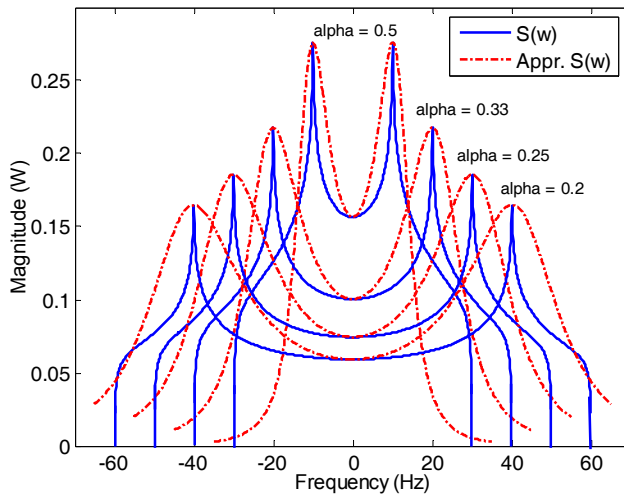
(b)

Fig. 12. DPSD, $S(f)$, and its approximation, $\tilde{S}(f)$, via 4th order function for different a's using (a) the complex cepstrum, and (b) the simple approximation methods.

5.3 Stochastic Ad Hoc Channel Models
The same procedure as in the STF cellular case is used to develop ad hoc channel models. The stochastic OCF is used to realize (41) for the inphase and quadrature components as [28]

$$
\begin{aligned}
dX_{I,j}(t) &= A_I X_{I,j}(t)dt + B_I dW_j^I(t) \\
I_j(t) &= C_I X_{I,j}(t) + f_j^I(t) \\
dX_{Q,j}(t) &= A_Q X_{Q,j}(t)dt + B_Q dW_j^Q(t) \\
Q_j(t) &= C_Q X_{Q,j}(t) + f_j^Q(t)
\end{aligned}
\tag{42}
$$

Where

$$
X_{I,j}(t) = \left[X_{I,j}^1(t), X_{I,j}^2(t), \ldots, X_{I,j}^m(t) \right]^T, \ X_{Q,j}(t) = \left[X_{Q,j}^1(t), X_{Q,j}^2(t), \ldots, X_{Q,j}^m(t) \right]^T
$$

$$
A_I = A_Q = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & & -a_{m-1} \end{bmatrix}, \ B_I = B_Q = \begin{bmatrix} b_{m-1} \\ \vdots \\ \vdots \\ b_1 \\ b_0 \end{bmatrix}, \ C_I = C_Q = \begin{bmatrix} 1 & 0 \cdots 0 \end{bmatrix}
\tag{43}
$$

$X_{I,j}(t)$ and $X_{Q,j}(t)$ are state vectors of the inphase and quadrature components. $I_j(t)$ and $Q_j(t)$ correspond to the inphase and quadrature components, respectively, $\left\{ W_j^I(t) \right\}_{t \geq 0}$ and $\left\{ W_j^Q(t) \right\}_{t \geq 0}$ are independent standard Brownian motions, which correspond to the inphase and quadrature components of the jth path respectively, the parameters $\{ a_{m-1}, \ldots, a_0, b_{m-1}, \ldots, b_0 \}$ are obtained from the approximation of the ad hoc DPSD, and $f_j^I(t)$ and $f_j^Q(t)$ are arbitrary functions representing the LOS of the inphase and quadrature components respectively. Equation (42) for the inphase and quadrature components of the jth path can be described as in (28), and the solution of the ad hoc state space model in (42) is similar to the one for STF model described in Section 4.4. The mean and variance of the ad hoc inphase and quadrature components have the same form as the ones for the STF case in (34) and (35), which show that the statistics are functions of time. The general TV state space representation for the ad hoc channel model is similar to the STF state space representation in (31) and (32).

Example 4: Consider a mobile-to-mobile (ad hoc) channel with parameters $v_1 = 36 \, \text{km/hr} \, (10 \, \text{m/s})$ and $v_2 = 24 \, \text{km/hr} \, (6.6 \, \text{m/s})$, in which $\alpha = 0.66$. Figure 13 shows time domain simulation of the inphase and quadrature components, and the attenuation coefficient. The inphase and quadrature components have been produced using (42) and (43), while the received signal is reproduced using (30). In Figure 13 Gauss-Newton method is used to approximate the deterministic DPSD with 4th order transfer function.
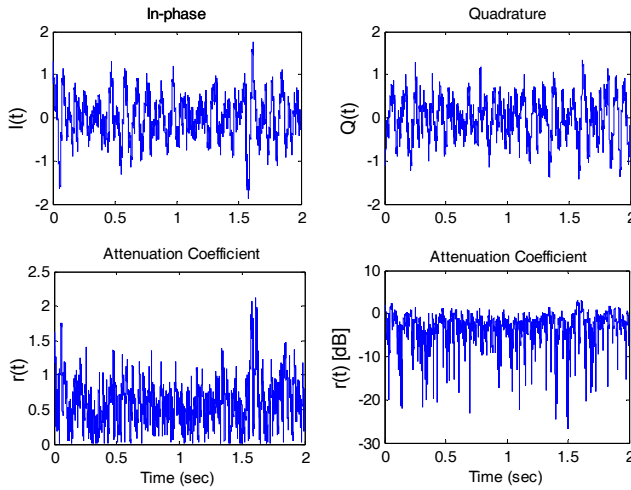
Fig. 13. Inphase and quadrature components $\{I(t), Q(t)\}$, and the attenuation coefficient $r_n(t) = \sqrt{I_n^2(t) + Q_n^2(t)}$, for a mobile-to-mobile channel with $\alpha = 0.66$ in Example 4.

6. Link Performance for Cellular and Ad Hoc Channels

Now, we want to compare the performance of the stochastic mobile-to-mobile link in (42) with the cellular link. We consider BPSK is the modulation technique and the carrier frequency is $f_c = 900\text{MHz}$. We test 10000 frames of $P = 100$ bits each. We assume mobile nodes are vehicles, with the constraint that the average speed over the mobile nodes is 30 km/hr. This implies $v_1 + v_2 = 60\text{km/hr}$, thus for a mobile-to-mobile link with $a = 0$ we get $v_1 = 60\text{km/hr}$ and $v_2 = 0$. The cellular case is defined as the scenario where a link connects a mobile node with speed 30 km/hr to a permanently stationary node, which is the base station. Thus, there is only one mobile node, and the constraint is satisfied. We consider the NLOS case $(f_I = f_Q = 0)$, which represents an environment with large obstructions.

The state space models developed in (27) and (42) are used for simulating the inphase and quadrature components for the cellular and ad hoc channels, respectively. The complex cepstrum approximation method is used to approximate the ad hoc DPSD with a 4th order stable, minimum phase, real, and rational transfer function. The received signal is reproduced using (30). Figure 14 shows the attenuation coefficient, $r(t) = \sqrt{I^2(t) + Q^2(t)}$, for both the cellular case and the worst-case mobile-to-mobile case $(\alpha = 1)$. It can be observed that a mobile-to-mobile link suffers from faster fading by noting the higher frequency components in the worst-case mobile-to-mobile link. Also it can be noticed that deep fading (envelope less than –12 dB) on the mobile-to-mobile link occurs more frequently and less bursty (48 % of the time for the mobile-to-mobile link and 32 % for the cellular link). Therefore, the increased Doppler spread due to double mobility tends to smear the errors out, causing higher frame error rates.

Consider the data rate given by $R_b = P / T_c = 5$ Kbps which is chosen such that the coherence time T_c equals the time it takes to send exactly one frame of length P bits, a condition where variation in Doppler spread greatly impacts the frame error rate (FER). Figure 15 shows the link performance for 10000 frames of 100 bits each. It is clear that the mobile-to-mobile link is worse than the cellular link, but the performance gap decreases as $\alpha \to 1$. This agrees with the main conclusion of [40], that an increase in degree of double mobility mitigates fading by lowering the Doppler spread. The gain in performance is nonlinear with α, as the majority of gain is from $\alpha = 0$ to $\alpha = 0.5$. Intuitively, it makes sense that link performance improves as the degree of double mobility increases, since mobility in the network becomes distributed uniformly over the nodes in a kind of equilibrium.
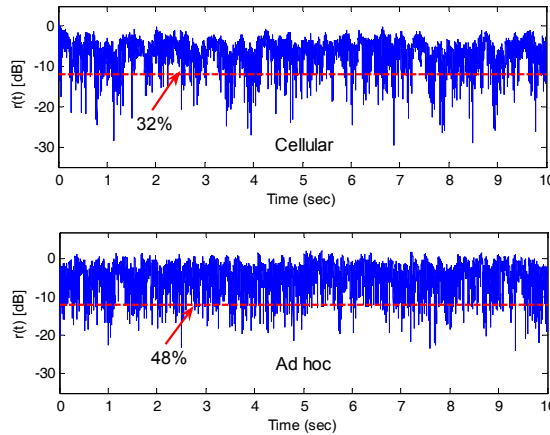
Fig. 14. Rayleigh attenuation coefficient for cellular link and worst-case ad hoc link.
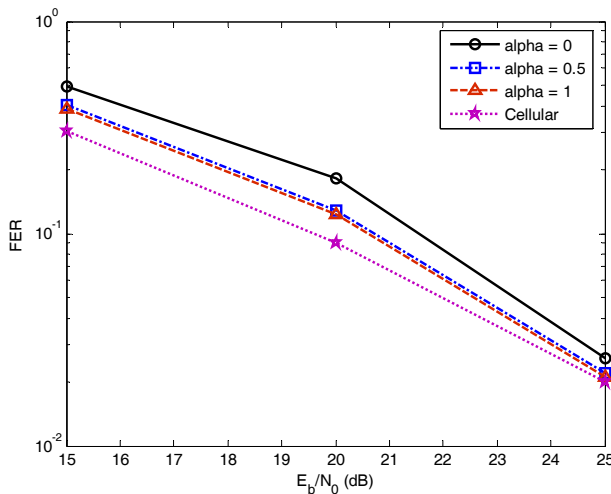
Fig. 15. FER results for Rayleigh mobile-to-mobile link for different a's and compared with cellular link.

7. Conclusion

In this chapter, stochastic models based on SDEs for LTF, STF, and ad hoc wireless channels are derived. These models are useful in capturing nodes mobility and environmental changes in mobile wireless networks. The SDE models described allow viewing the wireless channel as a dynamical system, which shows how the channel evolves in time and space. These models take into consideration the statistical and time variations in wireless communication environments. The dynamics are captured by a stochastic state space model, whose parameters are determined by approximating the deterministic DPSD. Inphase and quadrature components of the channel and their statistics are derived from the proposed models. The state space models have been used to verify the effect of fading on a transmitted signal in wireless fading networks. In addition, since these models are represented in state space form, they allow well-developed tools of estimation and identification to be applied to this class of problems. The advantage of using SDE methods is due to computational simplicity because estimation and identification algorithms can be performed recursively and in real time.

8. Acknowledgments

9. References

[1]   F. Molisch, *Wireless communications*. New York: IEEE Press/ Wiley, 2005.
[2]   J. Proakis, *Digital communications*, McGraw Hill, 4th Edition, 2000.
[3]   G. Stüber, *Principles of mobile communication*, Kluwer, 2nd Edition, 2001.
[4]   T.S. Rappaport, *Wireless communications: Principles and practice*, Prentice Hall, 2nd Edition, 2002.
[5]   J.F. Ossanna, "A model for mobile radio fading due to building reflections, theoretical and experimental waveform power spectra," *Bell Systems Technical Journal*, 43, 2935-2971, 1964.
[6]   F. Graziosi, M. Pratesi, M. Ruggieri, and F. Santucci, "A multicell model of handover initiation in mobile cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 48, no. 3, pp. 802-814, 1999.
[7]   F. Graziosi and F. Santucci, "A general correlation model for shadow fading in mobile systems*," IEEE Communication Letters*, vol. 6, no. 3, pp. 102-104, 2002.
[8]   M. Taaghol and R. Tafazolli, "Correlation model for shadow fading in land-mobile satellite systems," *Electronics Letters*, vol. 33, no. 15, pp.1287-1288, 1997.
[9]   R.H. Clarke, "A statistical theory of mobile radio reception," *Bell Systems Technical Journal*, 47, 957-1000, 1968.

[10] J.I. Smith, "A computer generated multipath fading simulation for mobile radio," *IEEE Trans. on Vehicular Technology*, vol. 24, no. 3, pp. 39-40, Aug. 1975.

[11] T. Aulin, "A modified model for fading signal at a mobile radio channel," *IEEE Trans. on Vehicular Technology*, vol. 28, no. 3, pp. 182-203, 1979.

[12] A. Saleh and R. Valenzula, "A statistical model for indoor multipath propagation," *IEEE Journal on Selected Areas in Communication*, vol. 5, no. 2, pp. 128-137, Feb. 1987.

[13] M. Gans, "A power-spectral theory of propagation in the mobile radio environment," *IEEE Trans. on Vehicular Technology*, vol. 21, no. 1, pp. 27-38, 1972.

[14] A. Duel-Hallen, S. Hu and H. Hallen, "Long-range prediction of fading signals," *IEEE Signal Processing Magazine*, pp. 62-75, May 2000.

[15] K. Baddour and N.C. Beaulieu, "Autoregressive modeling for fading channel simulation," *IEEE Trans. On Wireless Communication*, vol. 4, No. 4, July 2005, pp. 1650-1662.

[16] H.S. Wang and Pao-Chi Chang, "On verifying the first-order Markovian assumption for a Rayleigh fading channel model," *IEEE Transactions on Vehicular Technology*, vol. 45, No. 2, pp. 353-357, May 1996.

[17] C.C. Tan and N.C. Beaulieu, "On first-order Markov modeling for the Rayleigh fading channel," *IEEE Transactions on Communications*, vol. 48, No. 12, pp. December 2000.

[18] H.S. Wang and N. Moayeri, "Finite-state Markov channel: A useful model for radio communication channels," *IEEE Transactions on Vehicular Technology*, Vol. 44, No. 1 , pp. 163-171, February 1995.

[19] I. Chlamtac, M. Conti, and J.J. Liu, "Mobile ad hoc networking: imperatives and challenges," *Ad Hoc Networks*, vol.1, no. 1, 2003.

[20] A.S. Akki and F. Haber, "A statistical model for mobile-to-mobile land communication channel," *IEEE Trans. on Vehicular Technology*, vol. 35, no. 1, pp. 2-7, Feb. 1986.

[21] A.S. Akki, "Statistical properties of mobile-to-mobile land communication channels," *IEEE Trans. on Vehicular Technology*, vol. 43, no. 4, pp. 826-831, Nov. 1994.

[22] J. Dricot, P. De Doncker, E. Zimanyi, and F. Grenez, "Impact of the physical layer on the performance of indoor wireless networks," *Proc. Int. Conf. on Software, Telecommunications and Computer Networks (SOFTCOM)*, pp. 872-876, Split (Croatia), Oct. 2003.

[23] M. Takai, J. Martin, and R. Bagrodia, "Effects of wireless physical layer modeling in mobile ad hoc networks," *Proceedings of the 2nd ACM International Symposium on Mobile Ad Hoc Networking & Computing*, Long Beach, CA, USA, Oct. 2001.

[24] R. Negi and A. Rajeswaran, "Physical layer effect on MAC performance in ad-hoc wireless networks," *Proc. of Commun., Internet and Info. Tech. (CIIT)*, 2003.

[25] W. Jakes, *Microwave mobile communications*, IEEE Inc., NY, 1974.

[26] B. Sklar, *Digital communications: Fundamentals and applications*. Prentice Hall, 2nd Edition, 2001.

[27] C.D. Charalambous and N. Menemenlis, "Stochastic models for long-term multipath fading channels," *Proc. 38th IEEE Conf. Decision Control*, pp. 4947-4952, Phoenix, AZ, Dec. 1999.

[28] M.M. Olama, S.M. Djouadi, and C.D. Charalambous, "Stochastic differential equations for modeling, estimation and identification of mobile-to-mobile communication channels," *IEEE Transactions on Wireless Communications*, vol. 8, no. 4, pp. 1754-1763, 2009.

[29] M.M. Olama, K.K. Jaladhi, S.M. Djouadi, and C.D. Charalambous, "Recursive estimation and identification of time-varying long-term fading channels," *Research Letters in Signal Processing*, Volume 2007 (2007), Article ID 17206, 5 pages, 2007.

[30] M.M. Olama, S.M. Djouadi, and C.D. Charalambous, "Stochastic power control for time varying long-term fading wireless networks," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 89864, 13 pages, 2006.

[31] C.D. Charalambous and N. Menemenlis, "General non-stationary models for short-term and long-term fading channels," *EUROCOMM 2000*, pp. 142-149, April 2000.

[32] B. Oksendal, *Stochastic differential equations: An introduction with applications*, Springer, Berlin, Germany, 1998.

[33] W.J. Rugh, *Linear system theory*, Prentice-Hall, 2nd Edition, 1996.

[34] R.E.A.C. Paley and N. Wiener, "Fourier transforms in the complex domain," *Amer. Math. Soc. Coll., Am. Math.*, vol. 9, 1934.

[35] J.E. Dennis Jr., and R.B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, NJ: Prentice-Hall, 1983.

[36] A.V. Oppenheim and R.W. Schaffer, *Digital signal processing*, Prentice Hall, New Jersey, 1975, pp. 513.

[37] http://www.mathworks.com/

[38] P.E. Caines, *Linear stochastic systems*, New-York Wiley, 1988.

[39] R. Wang and D. Cox, "Channel modeling for ad hoc mobile wireless networks," *Proc. IEEE VTC*, 2002.

[40] R. Wang and D. Cox, "Double mobility mitigates fading in ad hoc wireless networks," *Proc. of the International Symposium on Antennas and Propagation*, vol. 2, pp. 306-309, 2002.

[41] C.S. Patel, G.L. Stuber, and T.G. Pratt, "Simulation of Rayleigh-faded mobile-to-mobile communication channels," *IEEE Trans. on Comm.*, vol. 53, no. 11, pp. 1876-1884, 2005.

[42] E.C. Levi, "Complex curve fitting," *IRE Trans. on Automatic Control*, vol. AC-4, pp. 37-44, 1959.

Information flow and causality quantification in discrete and continuous stochastic systems

X. San Liang

China Institute for Advanced Study,
Central University of Finance and Economics, Beijing,
State Key Laboratory of Satellite Ocean Environment Dynamics, Hangzhou,
and National University of Defense Technology, Changsha,
P. R. China

1. Introduction

Information flow, or information transfer as referred in the literature, is a fundamental physics concept that has applications in a wide variety of disciplines such as neuroscience (e.g., Pereda et al., 2005), atmosphere-ocean science (Kleeman, 2002; 2007; Tribbia, 2005), nonlinear time series analysis (e.g., Kantz & Schreiber, 2004; Abarbanel, 1996), economics, material science, to name several. In control theory, it helps to understand the information structure and hence characterize the cause-effect notion of causality in nonsequential stochastic control systems (e.g., Andersland & Teneketzis, 1992). Given the well-known importance, it has been an active arena of research for several decades (e.g.,Kaneko, 1986; Vastano & Swinney, 1988; Rosenblum et al., 1996; Arnhold et al., 1999; Schreiber, 2000; Kaiser & Schreiber, 2002). However, it was not until recently that the concept is formalized, on a rigorous mathematical and physical footing. In this chapter we will introduce the rigorous formalism initialized in Liang & Kleeman (2005) and established henceforth; we will particularly focus on the part of the studies by Liang (2008) and Liang & Kleeman (2007a,b) that is pertaining to the subjects of this book. For formalisms in a more generic setting or of broader interest the reader should consult and cite the original papers.

The concept of information flow/transfer was originally introduced to overcome the shortcoming of mutual information in reflecting the transfer asymmetry between the transmitter and the recipient. It is well known that mutual information tells the amount of information exchanged (cf. Cove & Thomas, 1991), but does not tell anything about the directionality of the exchange. This is the major thrust that motivates many studies in this field, among which are Vastano & Swinney (1988) and Schreiber (2000). Another thrust, which is also related to the above, is the concern over causality. Traditionally, causality, such as the Granger causality (Granger, 1969), is just a qualitative notion. While it is useful in identifying the causal relation between dynamical events, one would like to have a more accurate measure to quantify this relation. This would be of particular use in characterizing the intricate systems with two-way coupled events, as then we will be able to weigh the relative importance of one event over another. Information flow is expected to function as this quantitative measure.

The third thrust is out of the consideration from general physics. Information flow is a physical concept seen everywhere in our daily life experiences. The renowned baker transformation (cf. section 5 in this chapter), which mimics the kneading of a dough, is such an example. It has been argued intuitively that, as the transformation applies, information flows continuingly from the stretching direction to the folding direction, while no transfer is invoked the other way (e.g., Lasota & Mackey, 1994). Clearly the central issue here is how much the information is transferred between the two directions.

Historically information flow formalisms have been developed in different disciplines (particularly in neuroscience), usually in an empirical or half-empirical way within the context of the problems in question. These include the time-delayed information transfer (Vastano & Swinney, 1988) and the more sophisticated transfer entropy associated with a Markov chain (Schreiber, 2000). Others, though in different appearances, may nevertheless be viewed as the varieties of these two types. Recently, it was observed that even these two are essentially of the same like, in that both deal with the evolution of marginal entropies (Liang & Kleeman, 2005; 2007a). With this observation, Liang & Kleeman realized that actually this important concept can be rigorously formulated, and the corresponding formulas analytically derived rather than empirically proposed. The so-obtained transfer measure possesses nice properties as desired, and has been verified in different applications, with both benchmark systems and real world problems. The objective of this chapter is to give a concise introduction of this formalism. Coming up next is a setup of the mathematical framework, followed by two sections (§3 and §4) where the transfer measures for different systems are derived. In these sections, one will also see a very neat law about entropy production [cf. Eq. (18) in §3.1.2], paralleling the law of energy conservation, and the some properties of the resulting transfer measures (§4.3). Section 5 gives two applications, one about the afore-mentioned baker transformation, the other about a surprisingly interesting causality inference with two highly correlated time series. The final section (section 6) is a brief summary. Through the chapter only two-dimensional systems are considered; for high dimensional formalisms, see Liang & Kleeman (2007)a,b. As a convention in the literature, the terminologies "information flow" and "information transfer" will be used interchangeably throughout.

2. Mathematical formalism

Let Ω be the sample space and $\underline{x} \in \Omega$ the vector of state variables. For convenience, we follow the convention of notation in the physics literature, where random variables and deterministic variables are not distinguished. (In probability theory, they are usually distinguished with lower and upper cases like $\underline{x}$ and $\underline{X}$.) Consider a stochastic process of $\underline{x}$, which may take a continuous time form $\{\underline{x}(t), t \geq 0\}$ or a discrete time form $\{\underline{x}(\tau), \tau\}$, with τ being positive integers signifying discrete time steps. Throughout this chapter, unless otherwise indicated, we limit out discussion within two-dimensional (2D) systems $\underline{x} = (x_1, x_2)^T \in \Omega$ only. The stochastic dynamical systems we will be studying with are, in the discrete time case,

$$\underline{x}(\tau + 1) = \Phi(\underline{x}(\tau)) + \underline{\underline{B}}(\underline{x}, \tau)\underline{v} \tag{1}$$

and, in the continuous time case,

$$d\underline{x} = \underline{F}(\underline{x}, t)dt + \underline{\underline{B}}(\underline{x}, t)d\underline{w}. \tag{2}$$

Here Φ is a 2-dimensional transformation

$$\Phi : \Omega \to \Omega, \qquad (x_1, x_2) \mapsto (\Phi_1(\underline{x}), \Phi_2(\underline{x})), \tag{3}$$

$\underline{\mathbf{F}}$ the vector field, $\underline{\mathbf{v}}$ the white noise, $\underline{\mathbf{w}}$ a standard Wiener process, and $\underline{\mathbf{B}}$ a 2×2 matrix of the perturbation amplitude. The sample space Ω is assumed to be a Cartesian product $\Omega_1 \times \Omega_2$. We therefore just need to examine how information is transferred between the two components, namely x_1 and x_2, of the system in question. Without loss of generality, it suffices to consider only the information transferred from x_2 to x_1, or $T_{2 \to 1}$ for short.

Associated with each state $\underline{\mathbf{x}} \in \Omega$ is a joint probability density function

$$\rho = \rho(\underline{\mathbf{x}}) = \rho(x_1, x_2) \in L^1(\Omega),$$

and two marginal densities

$$\rho_1(x_1) \quad = \quad \int_{\Omega_2} \rho(x_1, x_2)\, dx_2,$$

$$\rho_2(x_2) \quad = \quad \int_{\Omega_1} \rho(x_1, x_2)\, dx_1,$$

with which we have a joint (Shannon) entropy

$$H = -\iint_{\Omega} \rho(\underline{\mathbf{x}}) \log \rho(\underline{\mathbf{x}})\, d\underline{\mathbf{x}}, \tag{4}$$

and marginal entropies

$$H_1 \quad = \quad -\int_{\Omega_1} \rho(x_1) \log \rho(x_1)\, dx_1, \tag{5}$$

$$H_2 \quad = \quad -\int_{\Omega_2} \rho(x_2) \log \rho(x_2)\, dx_2. \tag{6}$$

As $\underline{\mathbf{x}}$ evolves, the densities evolve subsequently. Specifically, corresponding to (2) there is a Fokker-Planck equation that governs the evolution of ρ; if $\underline{\mathbf{x}}$ moves on according to (1), the density is steered forward by the Frobenius-Perron operator (F-P operator henceforth). (Both the Fokker-Planck equation and the F-P operator will be introduced later.) Accordingly the entropies H, H_1, and H_2 also change with time. As reviewed in the introduction, the classical empirical/half-empirical information flow/transfer formalisms, though appearing in different forms, all essentially deal with the evolution of the marginal entropy of the receiving component, i.e., that of x_1 if $T_{2 \to 1}$ is considered. With this Liang & Kleeman (2005) noted that, by carefully classifying the mechanisms that govern the marginal entropy evolution, the concept of information transfer or information flow actually can be put on a rigorous footing. More specifically, the evolution of H_1 can be decomposed into two exclusive parts, according to their driving mechanisms: one is from x_2 only, another with the effect from x_2 excluded. The former, written $T_{2 \to 1}$, is the very information flow or information transfer from x_2 to x_1. Putting the latter as $\frac{dH_{1\not{2}}}{dt}$ for the continuous case, and $\Delta H_{1\not{2}}$ for the discrete case, we therefore have:

(1) For the discrete system (1), the information transferred from x_2 to x_1 is

$$T_{2 \to 1} = \Delta H_1 - \Delta H_{1\not{2}}; \tag{7}$$

(2) For the continuous system (2), the rate of information transferred from x_2 to x_1 is

$$T_{2 \to 1} = \frac{dH_1}{dt} - \frac{dH_{1\not{2}}}{dt}. \tag{8}$$

Likewise, the information flow from x_1 to x_2 can be defined. In the following we will be exploring how these are evaluated.

3. Deterministic systems with random inputs

We begin with the deterministic counterparts of (1) and (2), i.e.,

$$\underline{x}(\tau + 1) = \Phi(\underline{x}(\tau)), \tag{9}$$

and

$$\frac{d\underline{x}}{dt} = \underline{F}(\underline{x}, t), \tag{10}$$

respectively, with randomness limited within initial conditions, and then extend it to generic systems. This is not just because that (9) [resp. (10)] makes a special case of (1) [resp. (2)], but also because historically it is the idiosyncrasy of deterministic systems (Liang & Kleeman, 2005) that stimulates the rigorous formulation for this important physical notion, namely information flow or information transfer.

3.1 Entropy production

We first examine how entropy is produced with the systems (9) and (10). In this subsection, the system dimensionality is not limited to 2, but can be arbitrary.

3.1.1 Entropy evolution with discrete systems

Let $\rho = \rho(\underline{x})$ be the joint density of $\underline{x}$ at step τ, with the dependence on τ suppressed for simplicity. Its evolution is governed by the Frobenius-Perron operator, or F-P operator as will be called,

$$\mathcal{P} : L^1(\Omega) \mapsto L^1(\Omega),$$

which is given by, in a loose sense,

$$\int_\omega \mathcal{P}\rho(\underline{x}) \, d\underline{x} = \int_{\Phi^{-1}(\omega)} \rho(\underline{x}) \, d\underline{x}, \tag{11}$$

for any $\omega \subset \Omega$. [A rigorous definition with measure theory can be seen in Lasota & Mackey (1994).] If Φ is nonsingular and invertible, the right hand side of (11) is

$$\int_{\Phi^{-1}(\omega)} \rho(\underline{x}) \, d\underline{x} \xlongequal{\underline{y}=\Phi(\underline{x})} \int_\omega \rho\left[\Phi^{-1}(\underline{y})\right] \left|J^{-1}\right| \, d\underline{y},$$

where J is the Jacobian of Φ:

$$J = J(\underline{x}) = \det\left[\frac{\partial\Phi(x_1, x_2)}{\partial(x_1, x_2)}\right].$$

and J^{-1} its inverse. So in this case $\mathcal{P}$ can be explicitly written out:

$$\mathcal{P}\rho(\underline{x}) = \rho\left[\Phi^{-1}(\underline{x})\right] \left|J^{-1}\right|. \tag{12}$$

With $\mathcal{P}$, the change of the joint entropy H from time step τ to step $\tau + 1$ is, by (4),

$$\Delta H = H(\tau + 1) - H(\tau)$$

$$= -\iint_\Omega \mathcal{P}\rho(\mathbf{x}) \log \mathcal{P}\rho(\mathbf{x})\, d\mathbf{x} + \iint_\Omega \rho(\mathbf{x}) \log \rho(\mathbf{x})\, d\mathbf{x}. \tag{13}$$

In the case of nonsingular and invertible Φ, the above can be evaluated:

$$\begin{aligned}
\Delta H &= -\iint_\Omega \rho\left[\Phi^{-1}(\mathbf{x})\right]\left|J^{-1}\right| \cdot \log\left(\rho\left[\Phi^{-1}(\mathbf{x})\right]\left|J^{-1}\right|\right)\, d\mathbf{x} + \iint_\Omega \rho \log \rho\, d\mathbf{x} \\
&\overset{\mathbf{y}=\Phi^{-1}(\mathbf{x})}{=\!=\!=\!=\!=} -\iint_\Omega \rho(\mathbf{y})\left[\log \rho(\mathbf{y}) + \log\left|J^{-1}\right|\right]\, d\mathbf{y} + \iint_\Omega \rho \log \rho\, d\mathbf{x} \\
&= \iint_\Omega \rho(\mathbf{y})\left|J\right|\, d\mathbf{y}.
\end{aligned}$$

We hence have the following theorem:

Theorem 3.1. *If the system (9) has a nonsingular and invertible mapping Φ, then the entropy change can be expressed as, in a concise form,*

$$\boxed{\Delta H = E \log |J|,}$$

$$\tag{14}$$

where E is the mathematical expectation with respect to ρ.

Equation (14), which was established in Liang & Kleeman (2005), states that the entropy increase for a discrete system upon one application of an invertible transformation is simply the average logarithm of the rate of area change under the transformation. This extremely concise form of evolution gives us a hint on how the information flow concept may be easily obtained, as will be clear soon.

3.1.2 Entropy evolution with continuous systems

Now consider the continuous system (10). Here the dimensionality is not just limited to 2, but can be any positive integer n. First discretize it on the infinitesimal interval $[t, t + \Delta t]$:

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \mathbf{F}(\mathbf{x}(t), t)\Delta t. \tag{15}$$

This equation defines a mapping $\Phi : \Omega \to \Omega$, $\mathbf{x} \mapsto \mathbf{x} + \mathbf{F}(\mathbf{x}, t)\Delta t$, with a Jacobian

$$\begin{aligned}
J &= \det\left[\frac{\partial \Phi(x_1, x_2, ..., x_n)}{\partial(x_1, x_2, ..., x_n)}\right] \\
&= \det\begin{bmatrix}
1 + \frac{\partial F_1}{\partial x_1}\Delta t & \frac{\partial F_1}{\partial x_2}\Delta t & ... & \frac{\partial F_1}{\partial x_n}\Delta t \\
\frac{\partial F_2}{\partial x_1}\Delta t & 1 + \frac{\partial F_2}{\partial x_2}\Delta t & ... & \frac{\partial F_2}{\partial x_n}\Delta t \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial F_n}{\partial x_1}\Delta t & \frac{\partial F_n}{\partial x_2}\Delta t & ... & 1 + \frac{\partial F_n}{\partial x_n}\Delta t
\end{bmatrix} \\
&= \Delta t \sum_i \frac{\partial F_i}{\partial x_i} + O(\Delta t^2).
\end{aligned} \tag{16}$$

As $\Delta t \to 0$, it is easy to show that Φ is always nonsingular and invertible; in fact, $\Phi^{-1} : \Omega \to \Omega$ can be explicitly found:

$$\Phi^{-1}(\mathbf{x}) = \mathbf{x} - \mathbf{F}(\mathbf{x}, t)\Delta t + O(\Delta t^2). \tag{17}$$

So by (14), as $\Delta t \to 0$,

$$
\begin{aligned}
\frac{dH}{dt} &= \lim_{\Delta t \to 0} \frac{E \log |J|}{\Delta t} \\
&= E \lim_{\Delta t \to 0} \frac{1}{\Delta t} \log \left(1 + \Delta t \sum_i \frac{\partial F_i}{\partial x_i} + O(\Delta t^2) \right) \\
&= E \left(\sum_i \frac{\partial F_i}{\partial x_i} \right).
\end{aligned}
$$

This fulfills the proof of the following important theorem:

Theorem 3.2. *For the deterministic system (10), the entropy H evolves according to*

$$
\boxed{\frac{dH}{dt} = E(\nabla \cdot \underline{\mathbf{F}}).}
$$

(18)

Like (14), Eq. (18) is also in an extremely concise form. It states that the time rate of change of H is totally controlled by the contraction or expansion of the phase space. This important theorem was established by Liang & Kleeman (2005), using the Liouville equation corresponding to (10). But the derivation therein requires some assumption (though very weak) at the boundaries, while here no assumption is invoked.

3.2 Information flow

The elegant formula (18) allows us to obtain with ease the information flow for the continuous system (10). Indeed, this is precisely what Liang & Kleeman (2005) did in establishing the first formalism in a rigorous sense. To be short, consider only the rate of information transfer from x_2 to x_1, namely $T_{2 \to 1}$, which is the difference between the rate of change of the marginal entropy $\frac{dH_1}{dt}$ and that with the effect from x_2 excluded, i.e., $\frac{dH_{1\slashed{2}}}{dt}$. In a 2D system, $\frac{dH_{1\slashed{2}}}{dt}$ is actually equivalent to the rate of H_1 evolution due to x_1 its own, denoted $\frac{dH_1^*}{dt}$. Observing the obvious additivity property of (18), Liang & Kleeman (2005) intuitively argued that

$$
\frac{dH_1^*}{dt} = E \left(\frac{\partial F_1}{\partial x_1} \right).
$$

We hence obtain the following theorem:

Theorem 3.3. *For the 2D system (10),*

$$
\frac{dH_{1\slashed{2}}}{dt} = E \left(\frac{\partial F_1}{\partial x_1} \right) = \iint_\Omega \rho \frac{\partial F_1}{\partial x_1} \, dx_1 dx_2.
$$

(19)

(The proof of this theorem is deferred to later in this subsection.) The information flow from x_2 to x_1 therefore follows easily from (8).

Theorem 3.4. *For the 2D system (10), the rate of information transferred from x_2 to x_1 is*

$$T_{2\to1} = -\mathcal{E}_{2|1}\left(\frac{\partial(F_1\rho_1)}{\partial x_1}\right),\tag{20}$$

where $\mathcal{E}$ is an integration operator defined with respect to the conditional density

$$\rho_{2|1}(x_1|x_1) = \frac{\rho(x_1, x_2)}{\rho_1(x_1)}\tag{21}$$

such that, for any function $f = f(x_1, x_2)$,

$$\mathcal{E}_{2|1}f = \iint_\Omega \rho_{2|1}(x_2|x_1)\cdot f(x_1, x_2)\,dx_1dx_2.\tag{22}$$

Proof
Corresponding to (10) is the Liouville equation

$$\frac{\partial\rho}{\partial t} + \nabla\cdot(\mathbf{F}\rho) = 0\tag{23}$$

that governs the density evolution. Integrating it with respect to x_2 over the subspace Ω_2,

$$\frac{\partial\rho_1}{\partial t} + \frac{\partial}{\partial x_1}\int_{\Omega_2}\rho F_1\,dx_2 = 0.\tag{24}$$

The other term is integrated out with the compact support assumption for ρ. Multiplication by $-(1 + \log\rho_1)$, followed by an integration over Ω_1, gives

$$\begin{aligned}
\frac{dH_1}{dt} &= \iint_\Omega\left[\log\rho_1\frac{\partial(\rho F_1)}{\partial x_1}\right]dx_1dx_2\\
&= -\iint_\Omega\rho\left[\frac{F_1}{\rho_1}\frac{\partial\rho_1}{\partial x_1}\right]dx_1dx_2.
\end{aligned}$$

In the second step integration by parts is used; also used is the compact support assumption for ρ. So

$$\begin{aligned}
T_{2\to1} &= \frac{dH_1}{dt} - \frac{dH_{1\backslash2}}{dt}\\
&= \frac{dH_1}{dt} - E\left(\frac{\partial F_1}{\partial x_1}\right)\\
&= -\iint_\Omega\left(\frac{F_1}{\rho_1}\frac{\partial\rho_1}{\partial x_1} + \frac{\partial F_1}{\partial x_1}\right)\rho\,dx_1dx_2\\
&= -\iint_\Omega\rho_{2|1}(x_2|x_1)\frac{\partial(F_1\rho_1)}{\partial x_1}\,dx_1dx_2.
\end{aligned}$$

Q.E.D.

One may argue that, following the same way with (14), the information flow for the discrete system (9) can be obtained. Indeed this is true, but only in part, as the neat formula (14) requires that the mapping Φ and its components be nonsingular and invertible. Unfortunately,

for many important 2D mappings like the baker transformation we will be introducing in section 5, the requirements are generally not met. We therefore need to consider more generic situations.

By (7), we need to find ΔH_1 and $\Delta H_{1\bar{2}}$ as the system (9) moves forward from step τ to $\tau + 1$. As in the continuous case, it is easy to obtain ΔH_1 from the given mapping Φ. The key is how to find $\Delta H_{1\bar{2}}$, which is the entropy increase in direction 1 as the system goes from τ to $\tau + 1$ under Φ with x_2 frozen instantaneously at step τ, given $x_1(\tau)$. As $\Delta H_{1\bar{2}} = H_{1\bar{2}}(\tau + 1) - H_1(\tau)$, we are done if $H_{1\bar{2}}(\tau + 1)$ is evaluated. This is the marginal entropy for the first component evolved from H_1 with contribution from x_2 excluded from τ to $\tau + 1$. Consider the quantity

$$f \equiv -\log \mathcal{P}_{1\bar{2}}\rho_1(y_1), \tag{25}$$

where $y_1 = \Phi_1(\underline{x})$, and $\mathcal{P}_{1\bar{2}}\rho_1(y_1)$ is the marginal density in direction 1 at step $\tau + 1$, as the density ρ_1 evolves from step τ to step $\tau + 1$ under the transformation:

$$\Phi_{\bar{2}} : y_1 = \Phi_1(x_1, x_2) \tag{26}$$

i.e., the map Φ with x_2 frozen instantaneously at τ as a parameter. Note here we use $y_1 = \Phi_1(\underline{x})$ for the state of component 1 at step $\tau + 1$ (x_1 is for that at step τ); We do not use x_1 with some superscript or subscript in order to avoid any possible confusion in distinguishing the states of x_1 at these two time steps.

With our notation introduced above, $H_{1\bar{2}}(\tau + 1)$ is the mathematical expectation of f. (Recall how Shannon entropy is defined.) In other words, it is equal to the integration of f times some probability density function over the corresponding sample space. The first density to be multiplied is $\mathcal{P}_{1\bar{2}}\rho_1(y_1)$. But f also depends on x_2, we thence need another density for x_2. Recall that the freezing of x_2 is performed on interval $[\tau, \tau + 1]$, given all other components (here only x_1 in this 2D system) at step τ. What we need is therefore the conditional density of x_2 given x_1 at τ, i.e., $\rho(x_2|x_1)$. Put all these together, we therefore have the following result.

Proposition 3.1. *As the system (9) evolves from time step τ to time step $\tau + 1$, if x_2 is instantaneously frozen as a parameter at step τ, the marginal entropy of x_2 at step $\tau + 1$ is*

$$H_{1\bar{2}}(\tau + 1) = -\iint_\Omega \mathcal{P}_{1\bar{2}}\rho_1(y_1) \cdot \log \mathcal{P}_{1\bar{2}}\rho_1(y_1) \cdot \rho(x_2|x_1) \, dy_1 dx_2, \tag{27}$$

where y_1 is given by (26).

Note here we do not do another averaging with respect to x_1, as x_1 is already embedded in y_1.

The information transferred from x_2 to x_1 is now easy to obtain. Since $H_1(\tau)$ is the same, the right hand side of (7) is simply the difference between

$$H_1(\tau + 1) = -\iint_\Omega (\mathcal{P}\rho)_1(y_1) \log(\mathcal{P}\rho)_1(y_1) \, dy_1, \tag{28}$$

where $(\mathcal{P}\rho)_1$ is the marginal density at step $\tau + 1$, and $H_{1\bar{2}}(\tau + 1)$. We hence arrive at the following theorem on information flow.

Theorem 3.5. *For system (9), the information transferred from x_2 to x_1 is*

$$
\begin{aligned}
T_{2\rightarrow 1} \quad = \quad & -\int_{\Omega_1} (\mathcal{P}\rho)_1(y_1) \cdot \log(\mathcal{P}\rho)_1(y_1) \, dy_1 \\
& + \int_{\Omega} \mathcal{P}_{12}\rho_1(y_1) \cdot \log \mathcal{P}_{12}\rho_1(y_1) \cdot \rho(x_2|x_1) \, dy_1 dx_2.
\end{aligned}
\tag{29}
$$

Likewise the information flow from x_1 to x_2 can be obtained. Note in arriving at (29) no issue about the invertibility of Φ or either of its components is ever invoked. But if invertibility is guaranteed, then the formula may be further simplified.

Corollary 3.1. *In the system (9), if the mapping Φ has a component Φ_1 that is invertible, then*

$$
\Delta H_{1\setminus 2} = E \log |J_1|, \qquad where \quad J_1 = \frac{\partial \Phi_1(\mathbf{x})}{\partial x_1},
\tag{30}
$$

and hence

$$
T_{2\rightarrow 1} = \Delta H_1 - E \log |J_1|.
\tag{31}
$$

Remark: This concise result is just one would expect by the similar heuristic argument in arriving at Theorem 3.3 and Theorem 3.4.

Proof
By (27),

$$
\begin{aligned}
\Delta H_{1\setminus 2} \quad = \quad & -\iint_{\Omega_1 \times \Omega_2} \mathcal{P}_{12}\rho_1(y_1) \cdot \log \mathcal{P}_{12}\rho_1(y_1) \cdot \rho(x_2|x_1) \, dy_1 dx_2 \\
& + \int_{\Omega_1} \rho_1 \log \rho_1 \, dx_1,
\end{aligned}
$$

When Φ_1 is invertible, $J_1 = \frac{\partial \Phi_1}{\partial x_1} \neq 0$, by (12),

$$
\begin{aligned}
\mathcal{P}_{12}\rho_1(y_1) \quad = \quad & \rho \left[\Phi_1^{-1}(y_1, x_2) \right] \left| J_1^{-1} \right| \\
= \quad & \rho_1(x_1) \left| J_1^{-1} \right|.
\end{aligned}
\tag{32}
$$

So

$$
\begin{aligned}
\Delta H_{1\setminus 2} \quad = \quad & -\iint \rho_1(x_1) \left| J_1^{-1} \right| \log \left(\rho_1(x_1) \left| J_1^{-1} \right| \right) \rho(x_2|x_1) |J_1| \, dx_1 dx_2 \\
& + \int \rho_1 \log \rho_1 \, dx_1 \\
= \quad & -\iint \rho_1(x_1) \rho(x_2|x_1) \log \left| J_1^{-1} \right| \, dx_1 dx_2 \\
= \quad & \iint \rho(x_1, x_2) \log |J_1| \, dx_1 dx_2 \\
= \quad & E \log |J_1|,
\end{aligned}
\tag{33}
$$

and the second part follows subsequently.
Q.E.D.

We are now able to prove the first theorem of this subsection, namely Theorem 3.3, which originally was obtained by Liang & Kleeman (2005) through heuristic physical argument.

Proof of Theorem 3.3.
As before, look at an infinitesimal time interval $[t, t + \Delta t]$ and, for clarity, write the state variables at time t and $t + \Delta t$ as, respectively, $\underline{x}$ and $\underline{y}$. Discretization of (10) yields a mapping $\Phi = (\Phi_1, \Phi_2) : \Omega \to \Omega, \underline{x} = (x_1, x_2) \mapsto \underline{y} = (y_1, y_2)$,

$$\Phi : \begin{cases} y_1 = x_1 + \Delta t \cdot F_1(\underline{x}, t), \\ y_2 = x_2 + \Delta t \cdot F_2(\underline{x}, t). \end{cases} \tag{34}$$

As shown before, as $\Delta t \to 0$, Φ is nonsingular and always invertible, so are its components Φ_1 and Φ_2. Moreover, the Jacobian for Φ_1 is

$$J_1 = \frac{\partial y_1}{\partial x_1} = 1 + \Delta t \frac{\partial F_1}{\partial x_1} + O(\Delta t^2). \tag{35}$$

By Corollary 3.1, $\Delta H_{1\underline{2}} = E \log |J_1|$. So

$$
\begin{aligned}
\frac{dH_{1\underline{2}}}{dt} &= \lim_{\Delta t \to 0} \frac{\Delta H_{1\underline{2}}}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} E \left(\log \left| 1 + \Delta t \frac{\partial F_1}{\partial x_1} \right| + O(\Delta t^2) \right) \\
&= E \left(\frac{\partial F_1}{\partial x_1} \right).
\end{aligned}
$$

Q.E.D.

4. Stochastic systems

With the information flow for deterministic systems derived, we now take into account stochasticity and re-consider the problem. We first consider discrete systems in the form of (1), then continuous systems (2).

4.1 Discrete stochastic systems
As our convention, write $\underline{x}(\tau + 1)$ as $\underline{y}$ to avoid confusion. Eq. (1) then defines a mapping sending $\underline{x}$ to $\underline{y}$:

$$\underline{y} = \Phi(\underline{x}) + \underline{\underline{B}}(\underline{x})\underline{v}, \tag{36}$$

where $\underline{v}$ is a vector of white noise defined on $\mathbb{R}^2$, $\underline{\underline{B}} = (b_{ij})$ is a matrix of the perturbation amplitude, and the dependence on τ in the terms is suppressed for notation simplicity. Corresponding to this mapping is a Markov operator $\mathcal{P} : L^1(\Omega) \to L^1(\Omega)$, similar to the F-P operator for the system (9), that sends $\rho(\underline{x}(\tau))$ to $\rho(\underline{x}(\tau + 1))$ or $\rho(\underline{y})$. To find $\mathcal{P}$, we need just find $\rho(\underline{y})$, given $\rho(\underline{x})$, Φ, $\underline{\underline{B}}$, and $\rho(\underline{v})$ which is also written as $\rho_v(\underline{v})$ for clarity. For convenience, $\underline{\underline{B}}$ is assumed to be nonsingular.

Let Π be a transformation of $(\underline{\mathbf{x}}, \underline{\mathbf{v}})$ into $(\underline{\mathbf{z}}, \underline{\mathbf{y}})$ such that

$$\Pi : \left\{ \begin{array}{l} \underline{\mathbf{z}} = \underline{\mathbf{x}}, \\ \underline{\mathbf{y}} = \Phi(\underline{\mathbf{x}}) + \underline{\underline{\mathbf{B}}}(\underline{\mathbf{x}})\underline{\mathbf{v}}. \end{array} \right. \tag{37}$$

Its Jacobian is

$$J = \det \left[\frac{\partial(\underline{\mathbf{z}}, \underline{\mathbf{y}})}{\partial(\underline{\mathbf{x}}, \underline{\mathbf{v}})} \right] = \det \left[\begin{array}{cc} \underline{\mathbf{I}}_2 & \underline{\mathbf{0}}_2 \\ \frac{\partial(y_1, y_2)}{\partial(x_1, x_2)} & \underline{\underline{\mathbf{B}}} \end{array} \right] = \det \underline{\underline{\mathbf{B}}}, \tag{38}$$

where $\underline{\mathbf{I}}_2$ and $\underline{\mathbf{0}}_2$ are 2×2 identity and zero matrices, respectively. Given that $\underline{\underline{\mathbf{B}}}$ is nonsingular, $\det \underline{\underline{\mathbf{B}}}$ is nonzero, and hence Π is invertible:

$$\Pi^{-1} : \left\{ \begin{array}{l} \underline{\mathbf{x}} = \underline{\mathbf{z}}, \\ \underline{\mathbf{v}} = \underline{\underline{\mathbf{B}}}^{-1}(\underline{\mathbf{z}}) \left(\underline{\mathbf{y}} - \Phi(\underline{\mathbf{z}}) \right). \end{array} \right. \tag{39}$$

We now look at how the joint distribution of $(\underline{\mathbf{z}}, \underline{\mathbf{y}})$ is expressed in terms of $(\underline{\mathbf{x}}, \underline{\mathbf{v}})$.

For any $\omega_x \in \Omega$, $\omega_v \in \mathbb{R}^2$,

$$\begin{aligned} \iiiint_{\omega_x \times \omega_v} \rho_{z,y}(\underline{\mathbf{z}}, \underline{\mathbf{y}}) \, d\underline{\mathbf{z}} d\underline{\mathbf{y}} &= \iiiint_{\Pi^{-1}(\omega_x \times \omega_v)} \rho_{x,v}(\underline{\mathbf{x}}, \underline{\mathbf{v}}) \, d\underline{\mathbf{x}} d\underline{\mathbf{v}} \\ &= \iiiint_{\omega_x \times \omega_v} \rho_{x,v} \left(\Pi^{-1}(\underline{\mathbf{z}}, \underline{\mathbf{y}}) \right) \cdot \left| J^{-1} \right| \, d\underline{\mathbf{z}} d\underline{\mathbf{y}}. \end{aligned} \tag{40}$$

As ω_x and ω_v are arbitrarily chosen, the integrands must be equal, and hence

$$\begin{aligned} \rho_{z,y}(\underline{\mathbf{z}}, \underline{\mathbf{y}}) &= \rho_{x,v} \left(\Pi^{-1}(\underline{\mathbf{z}}, \underline{\mathbf{y}}) \right) \cdot \left| J^{-1} \right| \\ &= \rho_{x,v} \left[\underline{\mathbf{z}}, \underline{\underline{\mathbf{B}}}^{-1}(\underline{\mathbf{z}})(\underline{\mathbf{y}} - \Phi(\underline{\mathbf{z}})) \right] \cdot \left| J^{-1} \right| \\ &= \rho_x(\underline{\mathbf{z}}) \cdot \rho_v \left[\underline{\underline{\mathbf{B}}}^{-1}(\underline{\mathbf{z}})(\underline{\mathbf{y}} - \Phi(\underline{\mathbf{z}})) \right] \cdot \left[\det \underline{\underline{\mathbf{B}}}(\underline{\mathbf{z}}) \right]^{-1}. \end{aligned}$$

In the last step, the fact that $\underline{\mathbf{x}}$ and $\underline{\mathbf{v}}$ are independent has been used. Integrate $\underline{\mathbf{z}}$ out and we obtain

$$\begin{aligned} \rho_y(\underline{\mathbf{y}}) &= \iint_\Omega \rho_{z,y}(\underline{\mathbf{z}}, \underline{\mathbf{y}}) \, d\underline{\mathbf{z}} \\ &= \iint_\Omega \rho_x(\underline{\mathbf{z}}) \cdot \rho_v \left[\underline{\underline{\mathbf{B}}}^{-1}(\underline{\mathbf{z}})(\underline{\mathbf{y}} - \Phi(\underline{\mathbf{z}})) \right] \cdot \left[\det \underline{\underline{\mathbf{B}}}(\underline{\mathbf{z}}) \right]^{-1} \, d\underline{\mathbf{z}}. \end{aligned}$$

This equation defines a Markov operator $\mathcal{P}$ (corresponding to the F-P operator in the deterministic case) for system (1):

$$\mathcal{P}\rho(\underline{\mathbf{x}}) = \iint_\Omega \rho(\underline{\mathbf{z}}) \cdot \rho_v \left[\underline{\underline{\mathbf{B}}}^{-1}(\underline{\mathbf{z}})(\underline{\mathbf{y}} - \Phi(\underline{\mathbf{z}})) \right] \cdot \left[\det \underline{\underline{\mathbf{B}}}(\underline{\mathbf{z}}) \right]^{-1} \, d\underline{\mathbf{z}}. \tag{41}$$

In this case ρ_v is a Gaussian distribution with zero mean and an identity covariance matrix, and hence $\mathcal{P}$ can be computed. With it one may calculate the marginal density $(\mathcal{P}\rho)_1$ and hence the marginal entropy at time step $\tau + 1$:

$$H_1(\tau + 1) = - \int_{\Omega_1} (\mathcal{P}\rho)_1(y_1) \cdot \log(\mathcal{P}\rho)_1(y_1) \, dy_1. \tag{42}$$

Next look at $H_{1\bar{2}}(\tau + 1)$. Freezing x_2 at step τ modifies the dynamics to

$$\Phi_{\bar{2}}: \qquad y_1 = \Phi_1(x_1, x_2) + b_{11}v_1 + b_{12}v_2. \qquad (43)$$

Here we distinguish several cases: (1) If $b_{11} = b_{12} = 0$, then this degenerates into a de-terministic system, and the Markov operator is the F-P operator as we derived before; (2) if either of the last two terms vanishes, then follow the same procedure as above and a mod-ified Markov operator $\mathcal{P}_{1\bar{2}}$ is obtained; (3) if b_{11} and b_{12} have no dependence on $\underline{x}_1$, then $b_{11}v_1 + b_{12}v_2 \sim N(0, b_{11}^2 + b_{12}^2)$ can be combined to be one random variable with known distri-bution, and, again, the above procedure applies, and $\mathcal{P}_{1\bar{2}}$ follows accordingly; (4) if neither of the perturbations are zero, then we need to do a transformation from (x_1, v_1, v_2) to (z_1, z_2, y_1) with some random variables z_1 and z_2 as simple as possible. The so-obtained joint density of (z_1, z_2, y_1) is then integrated over the sample spaces of z_1 and z_2, and the resulting marginal entropy is the desired $\mathcal{P}_{1\bar{2}}$. So anyway $\mathcal{P}_{1\bar{2}}$ can be computed, giving

$$H_{1\bar{2}}(\tau + 1) = -\iint_\Omega \mathcal{P}_{1\bar{2}}\rho_1(y_1) \cdot \log \mathcal{P}_{1\bar{2}}\rho_1(y_1) \cdot \rho(x_2|x_1)\, dy_1 dx_2$$

by Proposition 3.1. This subtracted from $H_1(\tau + 1)$ results the information transferred from x_2 to x_1:

$$T_{2\to1} = H_1(\tau + 1) - H_{1\bar{2}}(\tau + 1). \qquad (44)$$

In principle, following the above procedure all the information flows between the system components can be computed. But more often than not this turn out to be very tedious and difficult. In practice, we would like to suggest different approaches, depending on the prob-lem itself.

4.2 Continuous stochastic systems

For the continuous system (2), there is a Fokker-Planck equation governing the density evolu-tion:

$$\frac{\partial \rho}{\partial t} + \frac{\partial(F_1\rho)}{\partial x_1} + \frac{\partial(F_2\rho)}{\partial x_2} = \frac{1}{2}\sum_{i,j=1}^{2}\frac{\partial^2(g_{ij}\rho)}{\partial x_i \partial x_j}, \qquad (45)$$

where

$$g_{ij} = g_{ji} = \sum_{k=1}^{2} b_{ik}b_{jk}, \qquad ij = 1, 2, \qquad (46)$$

and b_{ij} are the entries of the perturbation matrix $\underline{\underline{B}}$. From this it is easy to obtain the evolution of all the entropies, and H_1 in particular.

Proposition 4.1. *For system (2), the marginal entropy of x_1 evolves according to*

$$\frac{dH_1}{dt} = -E\left(F_1\frac{\partial \log\rho_1}{\partial x_1}\right) - \frac{1}{2}E\left(g_{11}\frac{\partial^2 \log\rho_1}{\partial x_1^2}\right), \qquad (47)$$

where E stands for expectation with respect to ρ.

Proof.
Integrate (45) with respect to x_2 over Ω_2 to get

$$\frac{\partial \rho_1}{\partial t} + \int_{\Omega_2} \frac{\partial(F_1\rho)}{\partial x_1}dx_2 = \frac{1}{2}\int_{\Omega_2} \frac{\partial^2(g_{11}\rho)}{\partial x_1^2}\, dx_2. \tag{48}$$

Here we have done integration by parts, and applied the compact support assumption for ρ and its derivatives. For simplicity, hereafter we will suppress the integral domain Ω, unless otherwise noted. Multiplication of (48) by $-(1 + \log \rho_1)$, followed by an integration with respect to x_1 over Ω_1, yields

$$\frac{dH_1}{dt} - \iint \log \rho_1 \frac{\partial(F_1\rho)}{\partial x_1}\, dx_1 dx_2 = -\frac{1}{2}\iint \log \rho_1 \frac{\partial^2(g_{11}\rho)}{\partial x_1^2}\, dx_1 dx_2.$$

Integrate by parts again, and (47) follows. Q.E.D.

As before, the key part is the evaluation of $\frac{dH_{12}}{dt}$. The result is summarized in the following theorem:

Proposition 4.2. *For the system (2), the time rate of change of the marginal entropy of x_1 with x_2 frozen instantaneously is*

$$\frac{dH_{1\backslash 2}}{dt} = E\left(\frac{\partial F_1}{\partial x_1}\right) - \frac{1}{2}E\left(g_{11}\frac{\partial^2 \log \rho_1}{\partial x_1^2}\right) - \frac{1}{2}E\left(\frac{1}{\rho_1}\frac{\partial^2(g_{11}\rho_1)}{\partial x_1^2}\right). \tag{49}$$

Proof.
Examine a small time interval $[t, t + \Delta t]$. We are going to prove the proposition by taking the limit:

$$\frac{dH_{1\backslash 2}}{dt} = \lim_{\Delta t \to 0} \frac{H_{1\backslash 2}(t + \Delta t) - H_1(t)}{\Delta t},$$

which boils down to the derivation of $H_{1\backslash 2}(t + \Delta t)$, namely the marginal entropy of x_1 at time $t + \Delta t$ as x_2 frozen as a parameter instantaneously at t. In principle this may be obtained using the strategy in the preceding subsection, but the evaluation of the convolution proves to be very difficult. To avoid the difficulty, Liang (2008) took a different approach, which we will follow hereafter.

In the stochastic system (2), the state $\underline{\mathbf{x}} = (x_1, x_2)^T$ is carried forth as time goes on. When time reaches t, freeze x_2 instantaneously and see how the state may evolves thenceforth until $t + \Delta t$. For convenience, denote by $x_{1\backslash 2}$ the first component of $\underline{\mathbf{x}}$ with x_2 frozen as a parameter. The system (2) is then modified to

$$dx_{1\backslash 2} = F_1(x_{1\backslash 2}, x_2, t)dt + \sum_k b_{1k}dw_k, \qquad \text{on } [t, t + \Delta t], \tag{50}$$

$$x_{1\backslash 2} = x_1 \qquad \text{at time } t. \tag{51}$$

Just as (45), correspondingly there is a modified Fokker-Planck equation for the density of $x_{1\backslash 2}$, written $\rho_{1\backslash 2}$:

$$\frac{\partial \rho_{1\backslash 2}}{\partial t} + \frac{\partial(F_1\rho_{1\backslash 2})}{\partial x_1} = \frac{1}{2}\frac{\partial^2(g_{11}\rho_{1\backslash 2})}{\partial x_1^2}, \qquad t \in [t, t + \Delta t] \tag{52}$$

$$\rho_{1\backslash 2} = \rho_1 \qquad \text{at } t. \tag{53}$$

Here g_{11} is the same as that defined in (46), i.e., $g_{11} = \sum_k b_{1k}^2$. Eq. (52) divided by $\rho_{1\breve{2}}$ yields

$$\frac{\partial f_t}{\partial t} + \frac{1}{\rho_{1\breve{2}}} \frac{\partial F_1 \rho_{1\breve{2}}}{\partial x_1} = \frac{1}{\rho_{1\breve{2}}} \frac{\partial^2 g_{11} \rho_{1\breve{2}}}{\partial x_1^2},$$

where f_t is a function of x_1,

$$f_t(x_1) = \log \rho_{1\breve{2}}(t, x_1). \tag{54}$$

We are doing this in the hope of obtaining an evolution law for $H_{1\breve{2}}$, as by the definition of Shannon entropy we will just need to consider how the expectation of $-f_t(x_1)$ evolves. Discretizing,

$$f_{t+\Delta t}(x_1) = f_t(x_1) - \frac{\Delta t}{\rho_1} \frac{\partial (F_1 \rho_1)}{\partial x_1} + \frac{\Delta t}{2\rho_1} \frac{\partial^2 (g_{11} \rho_1)}{\partial x_1^2} + O(\Delta t^2),$$

where the fact $\rho_{1\breve{2}} = \rho_1$ at time t has been used. For simplicity, the arguments have been suppressed for functions evaluated at $x_1(t)$, and this convention will be kept throughout this subsection. So

$$f_{t+\Delta t}(x_{1\breve{2}}(t+\Delta t)) = f_t(x_{1\breve{2}}(t+\Delta t)) - \frac{\Delta t}{\rho_1} \frac{\partial (F_1 \rho_1)}{\partial x_1} + \frac{\Delta t}{2\rho_1} \frac{\partial^2 (g_{11} \rho_1)}{\partial x_1^2} + O(\Delta t^2).$$

Using the Euler-Bernstein approximation (e.g., Lasota & Mackey, 1994) of (50), the $x_{1\breve{2}}(t+\Delta t)$ in the argument of f_t on the right hand side can be expanded as

$$x_{1\breve{2}}(t+\Delta t) = x_1(t) + F_1 \Delta t + \sum_k b_{1k} \Delta w_k + O(\Delta t^2).$$

And hence

$$\begin{aligned}
f_{t+\Delta t}&(x_{1\breve{2}}(t+\Delta t)) \\
&= f_t\left(x_1 + F_1 \Delta t + \sum_k b_{1k} \Delta w_k\right) - \frac{\Delta t}{\rho_1} \frac{\partial (F_1 \rho_1)}{\partial x_1} + \frac{\Delta t}{2\rho_1} \frac{\partial^2 (g_{11} \rho_1)}{\partial x_1^2} + O(\Delta t^2) \\
&= f_t(x_1) + \frac{\partial f_t}{\partial x_1}\left(F_1 \Delta t + \sum_k b_{1k} \Delta w_k\right) + \frac{1}{2} \frac{\partial^2 f_t}{\partial x_1^2}\left(F_1 \Delta t + \sum_k b_{1k} \Delta w_k\right)^2 \\
&\quad - \frac{\Delta t}{\rho_1} \frac{\partial (F_1 \rho_1)}{\partial x_1} + \frac{\Delta t}{2\rho_1} \frac{\partial^2 (g_{11} \rho_1)}{\partial x_1^2} + O(\Delta t^2),
\end{aligned} \tag{55}$$

where Taylor series expansion has been performed. Take expectations on both sides with respect to their respective random variables. Recalling how density evolution is defined, these expectations are equal (see Lasota & Mackey, 1994). Thus the left hand side results in $-H_{1\breve{2}}(t+\Delta t)$, and the first term on the right hand side is $-H_1(t)$. Notice that for a Wiener process w_k, $\Delta w_k \sim N(0, \Delta t)$, that is to say,

$$E\Delta w_k = 0, \qquad E(\Delta w_k)^2 = \Delta t;$$

also notice that Δw_k are independent of (x_1, x_2). So

$$E\left(\frac{\partial f_1}{\partial x_1} \sum_k b_{1k} \Delta w_k\right) = E\left(\frac{\partial f_1}{\partial x_1}\right) \sum_k b_{1k} E\Delta w_k = 0.$$

Hence the second term on the right hand side is

$$\Delta t \cdot E\left(F_1\frac{\partial f_t}{\partial x_1}\right).$$

For the same reason, the third term after expansion leaves only one sub-term of order Δt:

$$\frac{1}{2}E\left[\frac{\partial^2 f_t}{\partial x_1^2}\sum_k b_{1k}\Delta w_k\sum_j b_{1j}\Delta w_j\right]$$

$$=\frac{1}{2}E\left[\frac{\partial^2 f_t}{\partial x_1^2}\left(\sum_k b_{1k}^2(\Delta w_k)^2+\sum_{k\neq j}b_{1k}b_{1j}\Delta w_k\Delta w_j\right)\right].$$

Using the independence between the perturbations, the summation over $k\neq j$ inside the parentheses must vanish upon applying expectation. The first summation is equal to $g_{11}\Delta t$, by the definition of g_{ij} and the fact $E(\Delta w_k)^2=\Delta t$. So the whole term is

$$\frac{\Delta t}{2}E\left[g_{11}\frac{\partial^2 f_t}{\partial x_1^2}\right].$$

These, plus the fact that

$$f_t=\log\rho_{1\slashed{2}}(t;x_1)=\log\rho_1,$$

all put together, (55) followed by an expectation on both sides yields

$$H_{1\slashed{2}}(t+\Delta t)=H_1(t)-\Delta t\cdot E\left(F_1\frac{\partial\log\rho_1}{\partial x_1}\right)-\frac{\Delta t}{2}E\left(g_{11}\frac{\partial^2\log\rho_1}{\partial x_1^2}\right)$$

$$+\Delta t\cdot E\left(\frac{1}{\rho_1}\frac{\partial(F_1\rho_1)}{\partial x_1}\right)-\frac{\Delta t}{2}E\left(\frac{1}{\rho_1}\frac{\partial^2(g_{11}\rho_1)}{\partial x_1^2}\right)+O(\Delta t^2).$$

The second and fourth terms on the right hand side can be combined to give

$$\Delta t\cdot E\left(-F_1\frac{\partial\log\rho_1}{\partial x_1}+\frac{1}{\rho_1}\frac{\partial(F_1\rho_1)}{\partial x_1}\right)=\Delta t\cdot E\left(\frac{\partial F_1}{\partial x_1}\right).$$

So

$$\begin{aligned}\frac{dH_{1\slashed{2}}}{dt}&=\lim_{\Delta t\to 0}\frac{H_{1\slashed{2}}(t+\Delta t)-H_1(t)}{\Delta t}\\&=E\left(\frac{\partial F_1}{\partial x_1}\right)-\frac{1}{2}E\left(g_{11}\frac{\partial^2\log\rho_1}{\partial x_1^2}\right)-\frac{1}{2}E\left(\frac{1}{\rho_1}\frac{\partial^2(g_{11}\rho_1)}{\partial x_1^2}\right).\end{aligned}$$

Q.E.D.

With $\frac{dH_{1\slashed{2}}}{dt}$ evaluated, now it is easy to obtain $T_{2\to 1}$, namely, the information flow from x_2 to x_1.

Theorem 4.1. *For the system (2), the time rate of information transferred from x_2 to x_1 is*

$$T_{2\to1} = -\mathcal{E}_{2|1}\left(\frac{\partial(F_1\rho_1)}{\partial x_1}\right) + \frac{1}{2}\mathcal{E}_{2|1}\left(\frac{\partial^2(g_{11}\rho_1)}{\partial x_1^2}\right), \tag{56}$$

where $\mathcal{E}_{2|1}$ is the integration operator defined in Theorem 3.4.

Proof.

Subtracting (49) from (47), one obtains

$$
\begin{aligned}
T_{2\to1} &= -E\left(F_1\frac{\partial\log\rho_1}{\partial x_1}\right) - E\left(\frac{\partial F_1}{\partial x_1}\right) + \frac{1}{2}E\left(\frac{1}{\rho_1}\frac{\partial^2(g_{11}\rho_1)}{\partial x_1^2}\right) \\
&= -E\left(\frac{1}{\rho_1}\frac{\partial(F_1\rho_1)}{\partial x_1}\right) + \frac{1}{2}E\left(\frac{1}{\rho_1}\frac{\partial^2(g_{11}\rho_1)}{\partial x_1^2}\right),
\end{aligned}
\tag{57}
$$

where E is the expectation with respect to $\rho(x_1, x_2)$. Notice that the conditional density of x_2 given x_1 is

$$\rho_{2|1}(x_2|x_1) = \frac{\rho(x_1, x_2)}{\rho_1(x_1)}.$$

The operator

$$E\left(\frac{1}{\rho_1}\cdot\right) = \iint_\Omega \left(\frac{\rho}{\rho_1}\cdot\right)d\underline{\mathbf{x}}$$

is then simply the integration operator $\mathcal{E}_{2|1}$ as defined before in Theorem 3.4. The result thus follows.

Notice that in (56), the first term on the right hand side is precisely that in (20) i.e., the result of Liang & Kleeman (2005) based on intuitive argument for deterministic systems. This derivation supplies an alternative proof of the argument, and hence Theorem 3.4.

Above is the information flow from x_2 to x_1. Likewise, the flow from x_1 to x_2 can be derived. It is

$$T_{1\to2} = -\mathcal{E}_{1|2}\left(\frac{\partial(F_2\rho_2)}{\partial x_2}\right) + \frac{1}{2}\mathcal{E}_{1|2}\left(\frac{\partial^2(g_{22}\rho_2)}{\partial x_2^2}\right), \tag{58}$$

where $\rho_2 = \int \rho\, dx_1$ is the marginal density of x_2, and $\mathcal{E}_{1|2}$ is the operator such that, for any function $f \in L^1(\Omega)$, $\mathcal{E}_{1|2}f = \iint_\Omega \rho_{1|2}(x_1|x_2)f(\mathbf{x})\,d\underline{\mathbf{x}}$.

4.3 Properties

The above-derived information flow between system components possesses a very important property, namely the property of transfer directionality or asymmetry as emphasized in Schreiber (2000). One may have observed that the transfer in one direction need not imply anything about the transfer in the other direction, in contrast to the traditional correlation analysis or mutual information analysis. Particularly, in the extreme case that one component evolves independently from the other, the observation is concretized in the following theorem.

Theorem 4.2. (Causality)
If the evolution of x_1 is independent of x_2, then $T_{2\to1} = 0$.

Proof.
This property holds for formalisms with all the systems, but we here just prove with the continuous case. For the discrete system, the proof is lengthy, and the reader is referred to Liang & Kleeman (2007a) for details.

In (56), if $F_1 = F_1(x_1)$, and g_{11} is independent of x_2, integration can be taken for $\rho_{2|1}$ with respect to x_2 inside the double integrals, which gives

$$\int_{\Omega_2} \rho_{2|1}(x_2|x_1)\, dx_2 = 1.$$

The right hand side hence becomes

$$-\int_{\Omega_1} \frac{\partial(F_1\rho_1)}{\partial x_1}\, dx_1 + \int_{\Omega_1} \frac{\partial(g_{11}\rho_1)}{\partial x_2}\, dx_1.$$

By the compact support assumption, these integrations both vanish, leaving a zero $T_{2\to1}$.

Alternatively, if neither F_1 nor g_{11} has dependency on x_2, the integrals in (48) can be taken within the integrands, making ρ into ρ_1. This way the whole equation becomes a 1D Fokker-Planck equation for ρ_1, and hence x_1 is totally decoupled from the system, behaving like an independent variable. By intuition there should be no information flowing from x_2. Q.E.D.

This theorem shows that, between two evolving state variables x_1 and x_2, evaluation of $T_{2\to1}$ and $T_{1\to2}$ is able to tell which one causes which one and, in a quantitative way, tell how important one is to the other. Our information analysis thus gives a quantitative measure of the causality between two dynamical events. For this reason, this property is also referred to as the property of causality.

Another property holds only for the continuous system (2). Observe that the two terms of (56), the first is the same in form as that in (20), i.e., the corresponding deterministic system. Stochasticity contributes from the second term. An interesting observation is that:

Theorem 4.3. *Given a stochastic system component, if the stochastic perturbation is independent of another component, then the information transfer from the latter is the same in form as that for the corresponding deterministic system.*

Proof.
It suffices to consider only component x_1. If the stochastic perturbation $g_{11} = \sum_k b_{1k}^2$ is independent of x_2, then

$$\mathcal{E}_{2|1}\left(\frac{\partial^2(g_{11}\rho_1)}{\partial x_1^2}\right) = \int \frac{\partial^2(g_{11}\rho_1)}{\partial x_1^2}\, dx_1 = 0.$$

Here we have used the fact $\int \rho_{2|1} dx_2 = 1$. In this case, (56) and (20) have precisely the same form. Q.E.D.

This property is also very interesting since a great deal of noise in real systems appear to be additive; in other words, b_{ij}, and hence g_{ij}, are often constants. By the theorem these

stochastic systems thus function like deterministic in terms of information flow. Of course, the similarity is just in form; they are different in reality. The "deterministic" part of (56) (i.e., the first term) actually need not be deterministic, for stochasticity contributes to the state evolution and hence is embedded in the marginal density. As an illustration of the difference, the differential entropy for deterministic systems may go to minus infinity, e.g., in the case of the attractor of a fixed point or limit cycle, while this does not make an issue for stochastic systems (Ruelle, 1997).

5. Applications

The information flow formalism has been verified with benchmark problems, and applied to the study of several important dynamical system problems. Particularly, in Liang & Kleeman (2007a) we computed the transfers within a Hénon map, and obtained a result unique to our formalism just as one may expect on physical ground. In this section, we present two of these applications/verifications, echoing the challenges initially posed in the introduction.

5.1 Baker transformation

The baker transformation is a 2D mapping $\Phi : \Omega \to \Omega$, $\Omega = [0,1] \times [0,1]$, that mimics the kneading of a dough; it is given by

$$\Phi(x_1, x_2) = \begin{cases} (2x_1, \frac{x_2}{2}) & 0 \le x_1 \le \frac{1}{2},\ 0 \le x_2 \le 1 \\ (2x_1 - 1, \frac{1}{2}x_2 + \frac{1}{2}) & \frac{1}{2} < x_1 \le 1,\ 0 \le x_2 \le 1 \end{cases}. \tag{59}$$

As introduced in the beginning, physicists have observed and intuitively argued that, upon applying the transformation, information flows continuingly from the stretching direction (here x_1) to the folding direction (x_2), while no transfer occurs the other way (see Lasota & Mackey, 1994). However, until Liang & Kleeman (2007a), this important physical phenomenon had not ever been quantitatively studied. In the following, we give a brief presentation of the Liang & Kleeman result.

To start, first look at the F-P operator. It is easy to check that the baker transformation is invertible, and measure preserving (the Jacobian $J = 1$), so by Eq. (14) its joint entropy stays unchanged. (But one of its components is not; see below.) The inverse map is given by

$$\Phi^{-1}(x_1, x_2) = \begin{cases} (\frac{x_1}{2}, 2x_2) & 0 \le x_2 \le \frac{1}{2},\ 0 \le x_1 \le 1 \\ (\frac{x_1+1}{2}, 2x_2 - 1) & \frac{1}{2} \le x_2 \le 1,\ 0 \le x_1 \le 1 \end{cases} \tag{60}$$

Using Φ^{-1}, we can find the counterimage of $[0, x_1] \times [0, x_2]$ to be

1) $0 \le x_2 < \frac{1}{2}$,

$$\Phi^{-1}([0, x_1] \times [0, x_2]) = [0, \frac{x_1}{2}] \times [0, 2x_2]; \tag{61}$$

2) $\frac{1}{2} \le x_2 \le 1$,

$$\begin{aligned} \Phi^{-1}([0, x_1] \times [0, x_2]) &= \Phi^{-1}\left([0, x_1] \times [0, \frac{1}{2}]\right) \cup \Phi^{-1}\left([0, x_1] \times [\frac{1}{2}, x_2]\right) \\ &= [0, \frac{x_1}{2}] \times [0, 1] \cup [\frac{1}{2}, \frac{x_1+1}{2}] \times [0, 2x_2 - 1]. \end{aligned} \tag{62}$$

The F-P operator $\mathcal{P}$ is thus (cf. Lasota & Mackey, 1994)

$$\mathcal{P}\rho(x_1, x_2) = \frac{\partial^2}{\partial x_2 \partial x_1} \iint_{\Phi^{-1}([0,x_1]\times[0,x_2])} \rho(s,t) \, ds dt,$$

which, after a series of transformations, leads to

$$\mathcal{P}\rho(x_1, x_2) = \begin{cases} \rho(\frac{x_1}{2}, 2x_2), & 0 \le x_2 < \frac{1}{2}, \\ \rho(\frac{1+x_1}{2}, 2x_2 - 1), & \frac{1}{2} \le x_2 \le 1. \end{cases} \tag{63}$$

We now prove the following important result:

Theorem 5.1. *For the baker transformation (59),*

  (a) $T_{2\to 1} = 0$,
  (b) $T_{1\to 2} > 0$,

at any time steps.

Proof.
(a) With (63), we know that, upon one transformation, the marginal density of x_1 increases from

$$\rho_1 = \int_0^1 \rho(x_1, x_2) \, dx_2$$

to

$$
\begin{aligned}
\int_0^1 \mathcal{P}\rho(x_1, x_2) \, dx_2 &= \int_0^{1/2} \rho(\frac{x_1}{2}, 2x_2) \, dx_2 + \int_{1/2}^1 \rho(\frac{x_1+1}{2}, 2x_2 - 1) \, dx_2 \\
&= \frac{1}{2} \int_0^1 \left[\rho\left(\frac{x_1}{2}, x_2\right) + \rho\left(\frac{x_1+1}{2}, x_2\right) \right] \, dx_2 \\
&= \frac{1}{2} \left[\rho_1\left(\frac{x_1}{2}\right) + \rho_1\left(\frac{x_1+1}{2}\right) \right].
\end{aligned} \tag{64}
$$

Note that the (59) as a whole is invertible. Its x_1 direction, however, is not. Consider x_1 only, the transformation reduces to a dyadic mapping, $\Phi_1 : [0,1] \to [0,1]$, $\Phi_1(x_1) = 2x_1 \pmod 1$. It is easy to obtain

$$\Phi_1^{-1}([0, x_1]) = [0, \frac{x_1}{2}] \cup [\frac{1}{2}, \frac{1+x_1}{2}]$$

for $x_1 < 1$. So it has an F-P operator

$$
\begin{aligned}
(\mathcal{P}\rho)_{12}(x_1) &= \frac{\partial}{\partial x_1} \int_{\Phi_1^{-1}([0,x_1])} \rho_1(s) \, ds \\
&= \frac{\partial}{\partial x_1} \int_0^{x_1/2} \rho_1(s) \, ds + \frac{\partial}{\partial x_1} \int_{1/2}^{(1+x_1)/2} \rho_1(s) \, ds \\
&= \frac{1}{2} \left[\rho_1\left(\frac{x_1}{2}\right) + \rho_1\left(\frac{1+x_1}{2}\right) \right].
\end{aligned}
$$

This is exactly the same as (64), implying that

$$T_{2\to 1} = 0, \tag{65}$$

which is just as expected.

(b) To compute the transfer in the opposite direction, first compute the marginal distribution

$$
\int_0^1 \mathcal{P}\rho(x_1, x_2)\, dx_1 = \begin{cases} \int_0^1 \rho\left(\frac{x_1}{2}, 2x_2\right)\, dx_1, & 0 \le x_2 < \frac{1}{2}; \\ \int_0^1 \rho\left(\frac{x_1+1}{2}, 2x_2 - 1\right)\, dx_1, & \frac{1}{2} \le x_2 \le 1. \end{cases} \tag{66}
$$

This substituted in

$$
\begin{aligned}
\Delta H_2 \;=\; & -\int_0^1 \int_0^1 \mathcal{P}\rho(x_1, x_2) \cdot \left[\log\left(\int_0^1 \mathcal{P}\rho(\lambda, x_2)d\lambda \right) \right] dx_1 dx_2 \\
& + \int_0^1 \int_0^1 \rho(x_1, x_2) \cdot \left[\log\left(\int_0^1 \rho(\lambda, x_2)d\lambda \right) \right] dx_1 dx_2,
\end{aligned} \tag{67}
$$

after a series of transformation of variables, gives

$$
\Delta H_2 = -\log 2 + (I + I\!\!I), \tag{68}
$$

where

$$
I \;=\; \int_0^1 \int_0^{1/2} \rho(x_1, x_2) \cdot \left[\log \frac{\int_0^1 \rho(\lambda, x_2)d\lambda}{\int_0^{1/2} \rho(\lambda, x_2)d\lambda} \right] dx_1 dx_2, \tag{69}
$$

$$
I\!\!I \;=\; \int_0^1 \int_{1/2}^1 \rho(x_1, x_2) \cdot \left[\log \frac{\int_0^1 \rho(\lambda, x_2)d\lambda}{\int_{1/2}^1 \rho(\lambda, x_2)d\lambda} \right] dx_1 dx_2. \tag{70}
$$

Note both I and $I\!\!I$ are nonnegative, because $\rho(x_1, x_2) \ge 0$ and

$$
\int_0^1 \rho(x_1, x_2)\, dx_1 \ge \int_0^{1/2} \rho(x_1, x_2)\, dx_1 \tag{71}
$$

$$
\int_0^1 \rho(x_1, x_2)\, dx_1 \ge \int_{1/2}^1 \rho(x_1, x_2)\, dx_1. \tag{72}
$$

Moreover, the two equalities cannot hold simultaneously, otherwise ρ will be zero, contradicting to the fact that it is a density distribution. So $I + I\!\!I$ is strictly positive.

On the other hand, in the folding or x_2 direction the transformation is always invertible, and the Jacobian $J_2 = \frac{1}{2}$. By Corollary 3.1,

$$
\Delta H_{2\uparrow} = E \log \frac{1}{2} = -\log 2. \tag{73}
$$

So,

$$
T_{1\to 2} = \Delta H_2 - \Delta H_{2\uparrow} = I + I\!\!I > 0. \tag{74}
$$

Q.E.D.

In plain language, Eqs. (74) and (65) tell that there is always information flowing from x_1 or the stretching direction to x_2 or the folding direction ($T_{1\to 2} > 0$), while no transfer occurs the other way ($T_{2\to 1} = 0$). Illustrated in Fig. 1 is such a scenario, which has been intuitively argued in physics. Our formalism thus yields a result just as one may have expected on physical grounds.
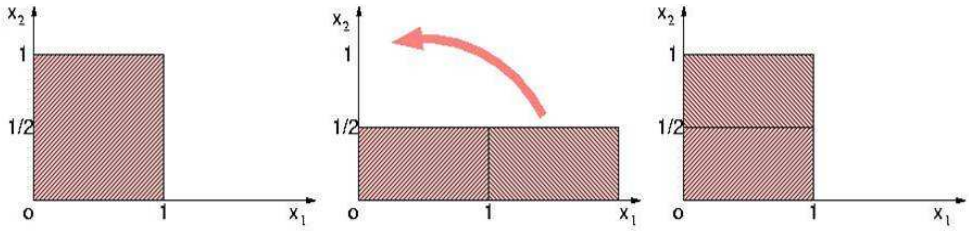
Fig. 1. Illustration of the baker transformation, and the associated information flow (middle) between the components.

5.2 Langevin equation

The formulas (56) and (58) with the stochastic system (2) are expected to be applicable in a wide variety of fields. To help further understand them, Liang (2008) examined a 2D linear system which hereafter we will be using:

$$d\underline{x} = \underline{\underline{A}}\,\underline{x}dt + \underline{\underline{B}}d\underline{w}, \tag{75}$$

where $\underline{w}$ is a Wiener process, and $\underline{\underline{A}} = (a_{ij})$ and $\underline{\underline{B}} = (b_{ij})$ are constant matrices. For convenience, suppose that initially $\underline{x}$ is Gaussian:

$$\underline{x} \sim N(\underline{\mu}, \underline{\underline{C}}).$$

Then it is Gaussian all the time because the system is linear (cf. Gardiner, 1985). Write the mean and covariance as

$$\underline{\mu}(t) = \left(\begin{array}{c} \mu_1(t) \\ \mu_2(t) \end{array}\right), \qquad \underline{\underline{C}}(t) = \left(\begin{array}{cc} c_{11}(t) & c_{12}(t) \\ c_{21}(t) & c_{22}(t) \end{array}\right).$$

It is easy to find the equations according to which they evolve:

$$\frac{d\mu}{dt} = \underline{\underline{A}}\,\underline{\mu}, \tag{76a}$$

$$\frac{d\underline{\underline{C}}}{dt} = \underline{\underline{A}}\,\underline{\underline{C}} + \underline{\underline{C}}\,\underline{\underline{A}}^T + \underline{\underline{B}}\,\underline{\underline{B}}^T. \tag{76b}$$

($\underline{\underline{B}}\,\underline{\underline{B}}^T$ is the matrix (g_{ij}) we have seen before.) Solve them for $\underline{\mu}$ and $\underline{\underline{C}}$, and we obtain the probability density distribution at any time:

$$\rho(\underline{x}) = \frac{1}{2\pi(\det\underline{\underline{C}})^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \underline{\underline{C}}^{-1}(\underline{x}-\underline{\mu})}. \tag{77}$$

Substitute this into (56) and (58), and the transfers $T_{2\to1}$ and $T_{2\to2}$ are obtained accordingly.

As an example, let $\underline{\underline{A}} = \left[\begin{array}{cc} -0.5 & 0.1 \\ a_{21} & -0.5 \end{array}\right]$, $\underline{\underline{B}} = \left[\begin{array}{cc} 1 & 1 \\ 1 & 1 \end{array}\right]$. It is easy to show that both the eigenvalues of $\underline{\underline{A}}$ are negative; the system is hence stable and has an equilibrium solution:

$$\underline{\mu}(\infty) = \left(\begin{array}{c} 0 \\ 0 \end{array}\right), \qquad \underline{\underline{C}}(\infty) = \left(\begin{array}{cc} 2.44 & 2.20 \\ 2.20 & 2.00 \end{array}\right),$$
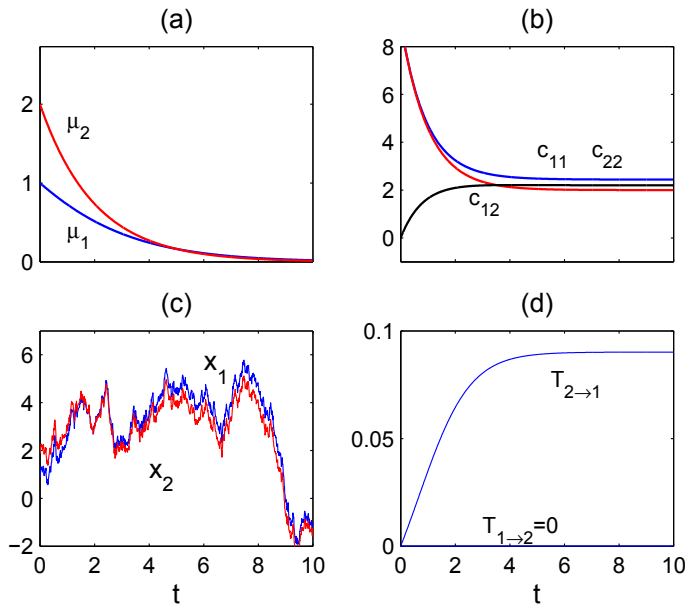
(a)

(b)

(c)

(d)

Fig. 2. A solution of (76) and the corresponding information transfers with the matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ and initial condition as shown in the text. (a) μ_1 and μ_2; (b) c_{11}, $c_{12} = c_{21}$, c_{22}; (c) a sample path starting from (1,2); (d) the computed information transfers $T_{2\to1}$ (upper) and $T_{1\to2} = 0$.

no matter how the system is initialized. Figs. 2a,b give the time evolutions of $\underline{\mu}$ and $\underline{\underline{C}}$ with initial conditions $\underline{\mu}(0) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, and $\underline{\underline{C}}(0) = \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix}$; For reference, in Fig. 2c we also plot a sample path starting from $\underline{x}(0) = \underline{\mu}(0)$. Clearly, though initially x_1 (red line) and x_2 (blue line) they differ by a significant value, soon they begin to merge and thenafter almost follow the same path. To analyze the information transfer, observe that in this case the vector field component

$$F_2 = -0.5x_2,$$

has no dependence on x_1; furthermore,

$$g_{ij} = \sum_k b_{ik}b_{jk}$$

are all constants. So by Theorem 4.2, the information transferred from x_1 to x_2 should vanish at all times:

$$T_{1\to2} = 0.$$

This assertion is confirmed by the computed result. In Fig. 2d, $T_{1\to2}$ is zero through time. The other transfer, $T_{2\to1}$, increases monotonically and eventually approaches to a limit.

Comparing Figs. 2c and 2d one may have more to talk about. Obviously the typical sample paths of x_1 and x_2 in the former are highly correlated—-In fact they are almost the same. This

is in drastic contrast to the zero information flow from x_1 to x_2, namely $T_{1\to2}$, in the latter. The moral here is, even though $x_1(t)$ and $x_2(t)$ are highly correlated, the evolution of x_2 has nothing to do with x_1. To x_1, x_2 is causal, while to x_2, x_1 is not. Through this simple example one sees how information transfer extends the traditional notion of correlation analysis and/or mutual information analysis by including causality.

6. Summary

The past few years have seen a major advance in the formulation of information flow or information transfer, a fundamental general physics and dynamical system concept which has important applications in different disciplines. This advance, beginning with an elegant formula obtained by Liang & Kleeman (2005) for the law of entropy production

$$\frac{dH}{dt} = E(\nabla \cdot \mathbf{F})$$

for system (10), has led to important scientific discoveries in the applied fields such as atmospheric science and oceanography. In this chapter, a concise introduction of the systematic research has been given within the framework of 2D dynamical systems. The resulting transfer is measured by the rate of entropy transferred from one component to another. The measure possesses a property of transfer asymmetry and, if the stochastic perturbation to the receiving component does not rely on the giving component, has a form same as that for the corresponding deterministic system. Explicit formulas, i.e., (56) and (58), have been obtained for generic stochastic systems (2), which we here write down again for easy reference:

$$T_{2\to1} = -E\left[\frac{1}{\rho_1}\frac{\partial(F_1\rho_1)}{\partial x_1}\right] + \frac{1}{2}E\left[\frac{1}{\rho_1}\frac{\partial^2(g_{11}\rho_1)}{\partial x_1^2}\right],$$

$$T_{1\to2} = -E\left[\frac{1}{\rho_2}\frac{\partial(F_2\rho_2)}{\partial x_2}\right] + \frac{1}{2}E\left[\frac{1}{\rho_2}\frac{\partial^2(g_{22}\rho_2)}{\partial x_2^2}\right],$$

where E stands for the mathematical expectation, and $g_{ij} = \sum_{k=1}^{2} b_{ik}b_{jk}$, $i = 1,2$.

We have applied the results to examine the information flow within the baker transformation and a linear system. In the former, it is proved that there is always information flowing from the stretching direction to the folding direction, while no information is transferred the other way. In the latter, one sees that correlation does not necessarily mean causality; for two highly correlated time series, the one-way information transfer could be zero. Information flow analysis thus extends the traditional notion of correlation analysis with causality quantitatively represented, and this quantification is firmly based on a rigorous mathematical and physical footing.

7. References

[1] Abarbanel, H.D.I. (1996). *Analysis of Observed Chaotic Data*, Springer, New York.
[2] Andersland, M.S., and Teneketzis, D. (1992). Information structures, causality, and non-sequential stochastic control I: Design-independent properties, *SIAM J. Control and Optimization*, Vol. 30 (No. 6): 1447-1475.

[3] Arnhold, J., et al (1999). A robust method for detecting interdependences: application to intracranially recorded EEG, *Physica D*, Vol. 134: 419-430.

[4] Cove, T. M., & Thomas, J.A. (1991). *Elements of Information Theory*, Wiley, New York.

[5] Gardiner, C.W. (1985). *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*, Springer-Verlag, New York.

[6] Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, Vol. 37: 424-438.

[7] Kaiser, A. & Schreiber, T. (2002). Information transfer in continuous processes, *Physica D*, Vol. 166: 43-62.

[8] Kaneko, K. (1986). Lyapunov analysis and information flow in coupled map lattices, *Physica D*, Vol. 23: 436-447.

[9] Kantz, H. & Schreiber, Thomas (2004). *Nonlinear Time Series Analysis*, Cambridge University Press.

[10] Kleeman, Richard (2002). Measuring dynamical prediction utility using relative entropy, *J. Atmos. Sci.*, Vol. 59:2057-2072.

[11] Kleeman, Richard (2007). Information flow in ensemble weather predictions, *J. Atmos. Sci.*, Vol. 64(3): 1005-1016.

[12] Lasota, A. & Mackey, M.C. (1994). *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*, Springer, New York.

[13] Liang, X. San (2008). Information flow within stochastic dynamical systems, *Phys. Rev. E*, Vol. 78: 031113.

[14] Liang, X. San. Local predictability and information flow in complex dynamical systems, *Physica D* (in press).

[15] Liang, X. San & Kleeman, Richard (2005). Information transfer between dynamical system components, *Phys. Rev. Lett.*, 95, (No. 24): 244101.

[16] Liang, X. San & Kleeman, Richard (2007a). A rigorous formalism of information transfer between dynamical system components. I. Discrete mapping *Physica D*, Vol 231: 1-9.

[17] Liang, X. San, & Kleeman, Richard (2007b). A rigorous formalism of information transfer between dynamical system components. II. Continuous flow, *Physica D*, Vol. 227: 173-182.

[18] Majda, A.J. & Harlim, J. (2007). Information flow between subspaces of complex dynamical systems, *Proc. Nat'l Acad. Sci.*, Vol. 104: 9558-9563.

[19] Pereda, E., Quiroga, R.Q. & Bhattacharya, J. (2005). Nonlinear multivariate analysis of neurophysiological signals, *Prog. Neurobiol.*, Vol. 77(1-2): 1-37.

[20] Rosenblum, M.G., Pikovsky, A.S. & Kurths, J. (1996). Phase synchronization of chaotic oscillators, *Phys. Rev. Lett.*, Vol. 76:1804-1807.

[21] Ruelle, D. (1997). Positivity of entropy production in the presence of a random thermostat, *J. Stat. Phys.*, Vol. 86(Nos. 5-6):935-951.

[22] Schreiber, Thomas (2000). Measuring information transfer, *Phys. Rev. Lett.*, Vol. 85(2):461.

[23] Tribbia, J.J. (2005). Waves, information and local predictability, *Workshop on Mathematical Issues and Challenges in Data Assimilation for Geophysical Systems: Interdisciplinary perspectives*, IPAM, UCLA, February 22-25, 2005.

[24] Vastano, J.A. & Swinney, H.L. (1988). Information transport in spatiotemporal systems, *Phys. Rev. Lett.*, Vol. 60:1773-1776.

Reduced-Order LQG Controller Design by Minimizing Information Loss*

Suo Zhang[1,2] and Hui Zhang[1,3]

1) State Key Laboratory of Industrial Control Technology,
Institute of Industrial Process Control,
Department of Control Science and Engineering, Zhejiang University, Hangzhou 310027
2) Department of Electrical Engineering,
Zhejiang Institute of Mechanical and Electrical Engineering, Hangzhou, 310053
3) Corresponding author
E-mails: zhangsuo.zju@gmail.com, zhanghui@iipc.zju.edu.cn

Introduction

The problem of controller reduction plays an important role in control theory and has attracted lots of attentions[1-10] in the fields of control theory and application. As noted by Anderson and Liu[2], controller reduction could be done by either direct or indirect methods. In direct methods, designers first constrain the order of the controller and then seek for the suitable gains via optimization. On the other hand, indirect methods include two reduction methodologies: one is firstly to reduce the plant model, and then design the LQG controller based on this model; the other is to find the optimal LQG controller for the full-order model, and then get a reduced-order controller by controller reduction methods. Examples of direct methods include optimal projection theory[3-4] and the parameter optimization approach[5] . Examples of indirect methods include LQG balanced realization[6-8], stable factorization[9] and canonical interactions[10].

In the past, several model reduction methods based on the information theoretic measures were proposed, such as model reduction method based on minimal K-L information distance[11], minimal information loss method(MIL)[12] and minimal information loss based on cross-Gramian matrix(CGMIL)[13]. In this paper, we focus on the controller reduction method based on information theoretic principle. We extend the MIL and CGMIL model reduction methods to the problem of LQG controller reduction.

The proposed controller reduction methods will be introduced in the continuous-time case. Though, they are applicable for both of continuous- and discrete-time systems.

LQG Control

LQG is the most fundamental and widely used optimal control method in control theory. It concerns uncertain linear systems disturbed by additive white noise. LQG compensator is an optimal full-order regulator based on the evaluation states from Kalman filter. The LQG control method can be regarded as the combination of the Kalman filter gain and the optimal control gain based on the separation principle, which guarantees the separated components could be designed and computed independently. In addition, the resulting closed-loop is (under mild conditions) asymptotically stable[14]. The above attractive properties lead to the popularity of LQG design.

The LQG optimal closed-loop system is shown in Fig. 1
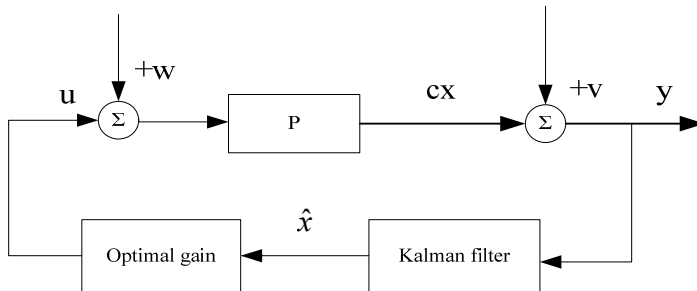
Fig. 1. LQG optimal closed-loop system

Consider the nth-order plant

$$\dot{x}(t) = Ax(t) + B(u(t) + w(t)), x(t_0) = x_0$$
$$y(t) = Cx(t) + v(t),$$

(1)

where $x(t) \in R^n$, $w(t) \in R^m$, $y(t), v(t) \in R^p$. A, B, C are constant matrices with appropriate dimensions. $w(t)$ and $v(t)$ are mutually independent zero-mean white Gaussian random vectors with covariance matrices Q and R ,respectively, and uncorrelated with x0. The performance index is given by

$$J = \lim_{t \to \infty} E\left\{ x^T R_1 x + u^T R_2 u \right\}, R_1 \geq 0, R_2 \geq 0.$$

(2)

While in the latter part, the optimal control law u would be replaced with the reduced-order suboptimal control law, such as u_r and u_G .

The optimal controller is given by

$$\dot{\hat{x}} = A\hat{x} + Bu + L(y - \hat{y}) = (A - BK - LC)\hat{x} + Ly,$$

(3)

$$u = -K\hat{x}. \tag{4}$$

where L and K are Kalman filter gain and optimal control gain derived by two Riccati equations, respectively.

Model Reduction via Minimal Information Loss Method (MIL)[12]

Different from minimal K-L information distance method, which minimizes the information distance between outputs of the full-order model and reduced-order model, the basic idea of MIL is to minimize the state information loss caused by eliminating the state variables with the least contributions to system dynamics.
Consider the n-order plant

$$\dot{x}(t) = Ax(t) + Bw(t), x(t_0) = x_0$$
$$y(t) = Cx(t) + v(t), \tag{5}$$

where $x(t) \in R^n$, $w(t) \in R^m$, $y(t), v(t) \in R^p$. A, B, C are constant matrices with appropriate dimensions. $w(t)$ and $v(t)$ are mutually independent zero-mean white Gaussian random vectors with covariance matrices Q and R ,respectively, and uncorrelated with x0.
To approximate system (5), we try to find a reduced-order plant

$$\dot{x}_r(t) = A_r x_r(t) + B_r w(t), x(t_0) = x_0$$
$$y(t) = C_r x_r(t) + v(t), \tag{6}$$

where $x_r(t) \in R^l, l < n$, $y_r(t) \in R^p$, A_r, B_r, C_r are constant matrices.
Define

$$x_r(t) = \Lambda x(t), \tag{7}$$

where $x_r(t)$ is the aggregation state vector of $x(t)$ and $\Lambda \in R^{l \times n}$ is the aggregation matrix.
From (5), (6) and (7), we obtain

$$A_r = \Lambda A \Lambda^+, B_r = \Lambda B, C_r = C \Lambda^+. \tag{8}$$

In information theory, the information of a stochastic variable is measured by the entropy function[15]. The steady-state entropy of system (5) and (6) are

$$H(x) = \frac{n}{2}\ln(2\pi e) + \frac{1}{2}\ln \det \Pi, \tag{9}$$

$$H(x_r) = \frac{l}{2}\ln(2\pi e) + \frac{1}{2}\ln\det\Pi_r. \tag{10}$$

where

$$\Pi_r = \Lambda\Pi\Lambda^+ \tag{11}$$

The steady-state information loss from (5) and (6) is defined by

$$IL(x; x_r) = H(x) - H(x_r). \tag{12}$$

From (11), (12) can be transformed to

$$H(x) - H(x_r) = \frac{n-l}{2}\ln(2\pi e) + \frac{1}{2}\ln\det(\Pi - \Lambda\Pi\Lambda^+). \tag{13}$$

The aggregation matrix Λ minimizing (13) consists of l eigenvectors corresponding to the l largest eigenvalues of the steady-state covariance matrix Π.

MIL-RCRP: Reduced-order Controller Based-on Reduced-order Plant Model

The basic idea of this method is firstly to find a reduced-order model of the plant, then design the suboptimal LQG controller according to the reduced-order model.
We have obtained the reduced-order model as (6). The LQG controller of the reduced-order model is given by

$$\dot{\hat{x}}_{r1} = A_{c1}\hat{x}_{r1} + B_{c1}y, \tag{14}$$

$$u_{r1} = C_{c1}\hat{x}_{r1}, \tag{15}$$

where $A_{c1} = A_{r1} - B_{r1}K_{r1} - L_{r1}C_{r1}$, $B_{c1} = L_{r1}$, $C_{c1} = -K_{r1}$ .The l-order suboptimal filter gain L_{r1} and suboptimal control gain K_{r1} are given by

$$L_{r1} = S_{r1}(\Lambda_r^+)^T C^T V^{-1}, \ K_{r1} = -R^{-1}\Lambda^T B^T P_{r1}, \tag{16}$$

where S_{r1} and P_{r1} are respectively the non-negative definite solutions to two certain Riccati equations as following:

$$P_{r1}A_{r1} + A_{r1}^T P_{r1} - P_{r1}B_{r1}R^{-1}B_{r1}^T P_{r1} + Q = 0, \tag{17}$$

$$A_{r1}S_{r1} + S_{r1}A_{r1}^T - S_{r1}C_{r1}^T V^{-1}C_{r1}S_{r1} + W = 0. \tag{18}$$

The stability of the closed-loop system is not guaranteed and must be verified.

MIL-RCFP: Reduced-order Controller Based on Full-order Plant Model

In this method , the basic idea is first to find a full-order LQG controller based on the full-order plant model, then get the reduced-order controller by minimizing the information loss between the states of the closed-loop systems with full-order and reduced-order controllers.

The full-order LQG controller is given by as (3) and (4). Then we use MIL method to obtain the reduced-order controller, which approximates the full-order controller.

The *l*-order Kalman filter is given by

$$\dot{\hat{x}}_{r2} = A_{c2}\hat{x}_{r2} + B_{c2}y, \tag{19}$$

where $A_{c2} = \Lambda_c A \Lambda_c^{+} - \Lambda_c BK\Lambda_c^{+} - \Lambda_c LC\Lambda_c^{+},\ B_{c2} = L_{r2} = \Lambda_c L = \Lambda_c SC^T V^{-1}.$

And the *l*-order control gain is given by

$$u_{r2} = C_{c2}\hat{x}_{r2}, \tag{20}$$

where $C_{c2} = -K_{r2} = -K\Lambda_c^{+} = -R^{-1}B^T P\Lambda_c^{+}.\ \Lambda_c$ is the aggregation matrix consists of the *l* eigenvectors corresponding to the *l* largest eigenvalues of the steady-state covariance matrix of the full-order LQG controller.

In what follows, we will propose an alternative approach, the CGMIL method, to the LQG controller-reduction problem. This method is based on the information theoretic properties of the system cross-Gramian matrix[16]. The steady-state entropy function corresponding to the cross-Gramian matrix is used to measure the information loss of the plant system. The two controller-reduction methods based on CGMIL, called CGMIL-RCRP and CGMIL-RCFP, respectively, possess the similar manner as MIL controller reduction methods.

**Model Reduction via Minimal Cross-Gramian
Information Loss Method (CGMIL)[16]**

In the viewpoint of information theory, the steady state information of (5) can be measured by the entropy function $H(x)$, which is defined by the steady-state covariance matrix Π .

Let $\tilde{\Pi}$ denote the steady-state covariance matrix of the state $\tilde{x}$ of the dual system of (5). When Q , the covariance matrix of the zero-mean white Gaussian random noise $w(t)$ is unit matrix I , Π and $\tilde{\Pi}$ are the unique definite solutions to

$$A\Pi + \Pi A^T + BB^T = 0,$$
$$A^T\tilde{\Pi} + \tilde{\Pi}A + C^T C = 0, \tag{21}$$

respectively.

From Linear system theory, the controllability matrix and observability matrix satisfy the following Lyapunov equation respectively:

$$AW_C + W_C A^T + BB^T = 0$$
$$A^T W_O + W_O A + C^T C = 0. \tag{22}$$

By comparing the above equations, we observe that the steady-state covariance matrix is equal to the controllability matrix of (5), and the steady-state covariance matrix of the dual system is equal to the observability matrix. We called $H(x)$ and $H(\tilde{x})$ the "controllability information" and "observability information", respectively. In MIL method, only "controllability information" is involved in deriving the reduced-order model, while the "observability information" is not considered.

In order to improve MIL model reduction method, CGMIL model reduction method was proposed in [13]. By analyzing the information theoretic description of the system, a definition of system "cross-Gramian information" (CGI) was defined based on the information properties of the system cross-Gramian matrix. This matrix indicates the "controllability information" and "observability information" comprehensively.

Fernando and Nicholson first define the cross-Gramian matrix by the step response of the controllability system and observability system. The cross-Gramian matrix of the system is defined by the following equation:

$$\boldsymbol{G}_{\text{cross}} = \int_0^\infty (e^{At}\boldsymbol{b})(e^{A^T t}\boldsymbol{c}^T)^T \, dt = \int_0^\infty e^{At}\boldsymbol{b}\boldsymbol{c}e^{At} \, dt, \tag{23}$$

which satisfies the following Sylvester equation:

$$A\boldsymbol{G}_{\text{cross}} + \boldsymbol{G}_{\text{cross}}A + \boldsymbol{b}\boldsymbol{c} = 0. \tag{24}$$

From [16], the cross-Gramian matrix satisfies the relationship between the controllability matrix and the observability matrix as the following equation:

$$\boldsymbol{G}_{\text{cross}}^2 = W_C W_O. \tag{25}$$

As we know that, the controllability matrix W_C corresponds to the steady-state covariance matrix of the system, while the observability matrix W_O corresponds to the steady-state covariance matrix of the dual system, which satisfy the following equations:

$$\boldsymbol{W}_C = \lim_{t \to \infty} E\{\boldsymbol{x}(t)\boldsymbol{x}^T(t)\}, \tag{26}$$

$$\boldsymbol{W}_O = \lim_{t \to \infty} E\{\tilde{\boldsymbol{x}}(t)\tilde{\boldsymbol{x}}^T(t)\}. \tag{27}$$

Combine equation (25)、(26) and (27), we obtain:

$$\boldsymbol{G}^2{}_{\text{cross}} = W_C W_O = \lim_{t \to \infty} E\{\boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)\} E\{\tilde{\boldsymbol{x}}(t)\tilde{\boldsymbol{x}}^{\mathrm{T}}(t)\}. \tag{28}$$

The cross-Gramian matrix corresponds to the steady-state covariance information of the original system and the steady-state covariance information of the dual system. Here we define a new stochastic state vector $\boldsymbol{\xi}(t)$, and the relationship among $\boldsymbol{\xi}(t)$, $\boldsymbol{x}(t)$ and $\tilde{\boldsymbol{x}}(t)$ satisfies the following equation:

$$\lim_{t \to \infty} E\{\boldsymbol{\xi}(t)\boldsymbol{\xi}^{\mathrm{T}}(t)\} = \lim_{t \to \infty} f(\boldsymbol{x}(t), \tilde{\boldsymbol{x}}(t))$$
$$= \lim_{t \to \infty} E\{\boldsymbol{x}(t)\boldsymbol{x}^{\mathrm{T}}(t)\} E\{\tilde{\boldsymbol{x}}(t)\tilde{\boldsymbol{x}}^{\mathrm{T}}(t)\} = \boldsymbol{G}^2{}_{\text{cross}}. \tag{29}$$

We called $\boldsymbol{\xi}(t)$ as "cross-Gramian stochastic state vector", which denotes the cross-Gramian information of the system.

From the above part, we know that the steady-state covariance matrix of $\boldsymbol{\xi}(t)$ is the cross-Gramian matrix $\boldsymbol{G}^2{}_{\text{cross}}$, the steady information entropy is called cross-Gramian information $I_{\text{cross}}(\boldsymbol{G}^2{}_{\text{cross}})$, which satisfies the following equation:

$$I_{\text{cross}}(\boldsymbol{G}^2{}_{\text{cross}}) = H(\boldsymbol{\xi}). \tag{30}$$

where $\boldsymbol{\xi}$ is the steady form of the stochastic state vector $\boldsymbol{\xi}(t)$, that is $\boldsymbol{\xi} = \lim_{t \to \infty} \boldsymbol{\xi}(t)$, and the information entropy of the steady-state $\boldsymbol{\xi}$ is defined as follows:

$$I_{\text{cross}}(\boldsymbol{G}^2{}_{\text{cross}}) = H(\boldsymbol{\xi}) = \frac{n}{2}\ln(2\pi e) + \frac{1}{2}\ln\det\boldsymbol{G}^2{}_{\text{cross}}. \tag{31}$$

And the following equation can be obtained:

$$I_{\text{cross}}(\boldsymbol{G}^2{}_{\text{cross}}) = \frac{n}{2}\ln(2\pi e) + \frac{1}{2}\ln\det\boldsymbol{PQ}. \tag{32}$$

$$I_{\text{cross}}(\boldsymbol{G}^2{}_{\text{cross}}) = \frac{H(\boldsymbol{x}) + H(\tilde{\boldsymbol{x}})}{2}. \tag{33}$$

From the above, we get that the cross-Gramian matrix indicates the controllability matrix and observability matrix comprehensively.

CGMIL model reduction method is suit for SISO system. The basic idea of the algorithm is

presented as follows, for continuous-time linear system.

The cross-Gramian matrix of the full-order system and the reduced-order system are as follows:

$$A\mathbf{G}_{\text{cross}} + \mathbf{G}_{\text{cross}}A + \boldsymbol{bc} = 0, \tag{34}$$

$$A\mathbf{G}^r_{\text{cross}} + \mathbf{G}^r_{\text{cross}}A + \boldsymbol{bc} = 0. \tag{35}$$

When the system input is zero mean Gaussian white noise signal, the cross-Gramian information of the two systems can be obtained as:

$$I_{\text{cross}}(\boldsymbol{G}^2{}_{\text{cross}}) = H(\boldsymbol{\xi}) = \frac{n}{2}\ln(2\pi e) + \frac{1}{2}\ln\det\boldsymbol{G}^2{}_{\text{cross}}, \tag{36}$$

$$I^r_{\text{cross}}(\boldsymbol{G}^2{}_{\text{cross}}{}^r) = H(\boldsymbol{\xi}_r) = \frac{l}{2}\ln(2\pi e) + \frac{1}{2}\ln\det\boldsymbol{G}^2{}_{\text{cross}}{}^r. \tag{37}$$

The cross-Gramian information loss is:

$$\begin{aligned}\Delta I_{\text{cross}} &= I_{\text{cross}}(\boldsymbol{G}^2{}_{\text{cross}}) - I^r_{\text{cross}}(\boldsymbol{G}^2{}_{\text{cross}}{}^r) = H(\boldsymbol{\xi}) - H(\boldsymbol{\xi}_r) \\ &= \frac{n-l}{2}\ln(2\pi e) + \frac{1}{2}[\ln\det\boldsymbol{G}^2{}_{\text{cross}} - \ln\det\boldsymbol{G}^2{}_{\text{cross}}{}^r].\end{aligned} \tag{38}$$

In order to minimize the information loss, we use the same method with the MIL method:

$$G^2{}_{cross}{}^r = \Lambda G^2{}_{cross}\Lambda^+. \tag{39}$$

where the aggregation matrix Λ is adopted as the l ortho-normal eigenvectors corresponding to the l th largest eigenvalues of the cross-Gramian matrix, then the information loss is minimized.

Theoretical analysis and simulation verification show that, cross-Gramian information is a good information description and CGMIL algorithm is better than the MIL algorithm in the performance of model reduction.

CGMIL-RCRP: Reduced-order Controller Based-on Reduced-order Plant Model By CGMIL

In this section, we apply the similar idea as method 1 of MIL model reduction to obtain the reduced-order controller.

The LQG controller of the reduced-order model consists of Kalman filter and control law as follows:

$$\dot{\hat{x}}_{GC1} = A_{GC1}\hat{x}_{GC1} + B_{GC1}y, \tag{40}$$

$$u_{G1} = C_{GC1}\hat{x}_{G1}. \tag{41}$$

where $A_{GC1} = A_{G1} - B_{G1}K_{G1} - L_{G1}C_{G1}$, $B_{GC1} = L_{G1}$, $C_{GC1} = -K_{G1}$.

The r-order filer gain and control gain are obtained:

$$L_{G1} = S_{G1}C_{G1}{}^T V^{-1} = S_{G1}(\Lambda_{G1}{}^+)^T C^T V^{-1}, \tag{42}$$

$$K_{G1} = -R^{-1}B_{G1}{}^T P_{G1} = -R^{-1}\Lambda_{G1}{}^T B^T P_{G1}. \tag{43}$$

where S_{G1} and P_{G1} satisfy the following Riccati equations

$$P_{G1}A_{G1} + A_{G1}{}^T P_{G1} - P_{G1}B_{G1}R^{-1}B_{G1}{}^T P_{G1} + Q = 0, \tag{44}$$

$$A_{G1}S_{G1} + S_{G1}A_{G1}{}^T - S_{G1}C_{G1}{}^T V^{-1}C_{G1}S_{G1} + W = 0. \tag{45}$$

And the state space equation of the r-order closed-loop system is as follow:

$$\begin{bmatrix} \dot{x} \\ \dot{\hat{x}}_{G1} \end{bmatrix} = \begin{bmatrix} A & BC_{GC1} \\ B_{GC1}C & A_{G1} + B_{G1}C_{GC1} - B_{GC1}C_{G1} \end{bmatrix} \begin{bmatrix} x \\ \hat{x}_{G1} \end{bmatrix} + \begin{bmatrix} w \\ L_{G1}v \end{bmatrix}$$

$$= \begin{bmatrix} A & -BK_{G1} \\ L_{G1}C & A_{G1} - B_{G1}K_{G1} - L_{G1}C_{G1} \end{bmatrix} \begin{bmatrix} x \\ \hat{x}_{G1} \end{bmatrix} + \begin{bmatrix} w \\ L_{G1}v \end{bmatrix}, \tag{46}$$

$$y_{G1} = \begin{bmatrix} C & 0 \end{bmatrix} \begin{bmatrix} x \\ \hat{x}_{G1} \end{bmatrix} + v. \tag{47}$$

CGMIL-RCFP: Reduced-order Controller Based on Full-order Plant Model By CGMIL

Similar to the second method of MIL controller reduction method, the reduced-order controller obtained by the full-order controller using CGMIL method is:

$$\dot{\hat{x}}_{G2} = A_{GC2}\hat{x}_{G2} + B_{GC2}y, \tag{48}$$

$$u_{G2} = C_{GC2}\hat{x}_{G2}. \tag{49}$$

where $A_{GC2} = \Lambda_{G2}A_c\Lambda_{G2}{}^+$, $B_{GC2} = L_{G2}$, $C_{GC2} = -K_{G2}$, Λ_{G2} is the aggregation matrix consists of the l largest eigenvalues corresponding to the lth largest eigenvectors of

the cross-Gramian matrix of the full-order controller. The r-order filter gain and control gain is obtained:

$$L_{G2} = \Lambda_{G2}L = \Lambda_{G2}SC^TV^{-1}, \tag{50}$$

$$K_{G2} = K\Lambda_{G2}^+ = R^{-1}B^TP\Lambda_{G2}^+. \tag{51}$$

The state space equation of the reduced-order controller is then given by:

$$\dot{\hat{x}}_{G2} = A_{GC2}\hat{x}_{G2} + B_{GC2}y = (\Lambda_{G2}A\Lambda_{G2}^+ - \Lambda_{G2}BK\Lambda_{G2}^+ - \Lambda_{G2}LC\Lambda_{G2}^+)\hat{x}_{G2} + \Lambda_{G2}Ly$$
$$u_{G2} = C_{GC2}\hat{x}_{G2} = -K\Lambda_{G2}^+\hat{x}_{G2}. \tag{52}$$

Stability Analysis of the Reduced-Order Controller

Here we present our conclusion in the case of discrete systems.
Suppose the full-order controller is stable, and we analyze the stability of the reduced-order controller obtained by method MIL-RCFP.
Conclusion 1.1 [Lyapunov Criterion] The discrete-time time-invariant linear autonomous system, when the state $x_e = 0$ is asymptotically stable, that is the amplitude of all of the eigenvalues of G $\lambda_i(G)$ $(i = 1, 2, ..., n)$ less than 1. If and only if for any given positive definite symmetric matrix Q, the discrete-time Lyapunov equation:

$$G^TPG + Q = P, \tag{53}$$

has the uniquely positive definite symmetric matrix P.
The system parameter of the full-order controller is: $A_c = A - BK - LC$. From Lyapunov Criterion, the following equation is obtained:

$$A_c\,PA_c^T + Q = P. \tag{54}$$

Multiplying leftly by the aggregation matrix Λ_c and rightly by Λ_c^T, we get:

$$\Lambda_cA_cP(\Lambda_cA_c)^T + \Lambda_cQ\Lambda_c^T = \Lambda_cP\Lambda_c^T. \tag{55}$$

Because $\Lambda_cA_c = A_{c2}\Lambda_c$, the following equation is obtained:

$$A_{c2}\Lambda_cP\Lambda_c^TA_{c2} + \Lambda_cQ\Lambda_c^T = \Lambda_cP\Lambda_c^T. \tag{56}$$

When $\Lambda_c^{'} = [\Lambda_c^{T}, \eta_{l+1}, ..., \eta_n]^T$ is assumed, where $\eta_{l+1}, ..., \eta_n$ is the n-l smallest eigenvectors corresponding to the n-l smallest eigenvalues of the steady-state covariance matrix Π_c. The aggregation matrix $\Lambda_c^{'}$ consists of the orthogonal eigenvectors, when P and Q are positive definite matrix, $\Lambda_c^{'}P(\Lambda_c^{'})^T$ and $\Lambda_c^{'}Q(\Lambda_c^{'})^T$ are positive definite. The matrix $\Lambda_c P(\Lambda_c)^T$ consists of the first $l \times l$ main diagonal elements of matrix $\Lambda_c^{'}P(\Lambda_c^{'})^T$; similarly, the matrix $\Lambda_c Q(\Lambda_c)^T$ consists of the first $l \times l$ main diagonal elements of matrix $\Lambda_c^{'}Q(\Lambda_c^{'})^T$. If $\Lambda_c^{'}P(\Lambda_c^{'})^T$ and $\Lambda_c^{'}Q(\Lambda_c^{'})^T$ are positive definite, then $\Lambda_c P(\Lambda_c)^T$ and $\Lambda_c Q(\Lambda_c)^T$ are positive definite. As a result, the reduced-order controller obtained from method MIL-RCFP is stable.

Illustrative Example

1. Lightly Damped Beam

We applied these two controller-reduction methods to the lightly damped, simply supported beam model described in [11] as (5).
The full-order Kalman filter gain and optimal control gain are given by

$$L = [2.0843 \ 2.2962 \ 0.1416 \ 0.1774 \ -0.2229$$
$$-0.4139 \ -0.0239 \ -0.0142 \ 0.0112 \ -0.0026]^T, \tag{57}$$

$$K = [0.4143 \ 0.8866 \ 0.0054 \ 0.0216 \ -0.0309$$
$$-0.0403 \ 0.0016 \ -0.0025 \ -0.0016 \ 0.0011]. \tag{58}$$

The proposed methods are compared with that given in [11], which will be noted by method 3 later. The order of the reduced controller is 2. We apply the two CGMIL controller reduction methods and the first MIL controller reduction method (MIL-RCRP) to this model. The reduced-order Kalman filter gains and control gains of the reduced-order closed-loop systems are given as follows:

MIL-RCRP: $L_{r1} = [-1.5338; -2.6951]^T, K_{r1} = [-0.1767 \ -0.9624]$

CGMIL-RCRP: $L_{G1} = [-3.0996 \ -0.0904]^T, K_{G1} = [-0.9141 \ -0.3492]$

CGMIL-RCFP: $L_{G2} = [0.4731 \ 0.9706]^T, K_{G2} = [0.4646 \ -0.9785]$

Method 3: $L_{r3} = [2.1564 \ 2.2826]^T, K_{r3} = [0.3916 \ 0.8752]$.

Three kinds of indices are used to illustrate the performances of the reduced-order controllers.

a)   We define the output mean square errors to measure the performances of the reduced-order controllers

$$E_a{}^* = \int_0^T y_*{}^2(t)dt / T,$$ (59)

where $* = 1, 2, 3$ indicates the closed-loop systems obtained from method 1,2,3, respectively. T is the simulation length.

b)   We compare the reduced-order controllers with the full-order one by using relative error indices

$$E_b{}^* = \int_0^T (y(t) - y_*(t))^2 dt / T,$$ (60)

where $y(t)$ is the system output of the full-order closed-loop system.

c)   We also use the LQG performance indices given by following equations, to illustrate the controller performances

$$J^* = \frac{1}{T} \int_0^T \left\{ x^T(t)Qx(t) + u_*{}^T(t)Ru_*(t) \right\} dt .$$ (61)

The performances of the reduced-order controllers are illustrated by simulating the responses of the zero-input and Gaussian white noise, respectively. The simulation results are shown in the following figures and diagrams.

As shown in Fig. 1 (Response to initial conditions), when input noise and observation noise are zero, the system initial states are set as $x_i(0) = 1/i, i = 1, \ldots 10$. The reduced-order closed-loop system derived by method 3 is close to the full-order one.
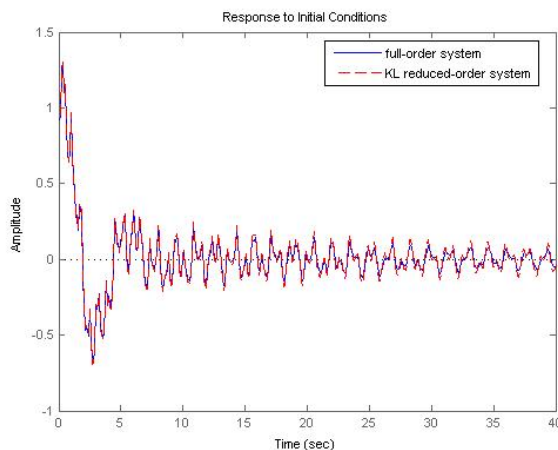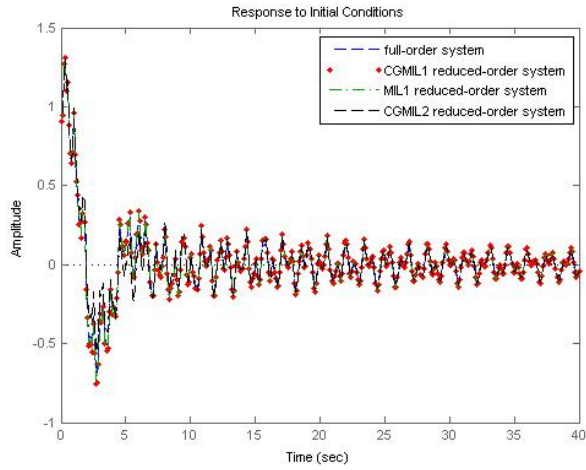
Fig. 1. Zero-input response for full-order system and reduced-order system

In Fig. 2 (Response of Gaussian white noise), almost all the reduced-order closed-loop system are close to the full-order one except the reduced-order system obtained by CGMIL 2.
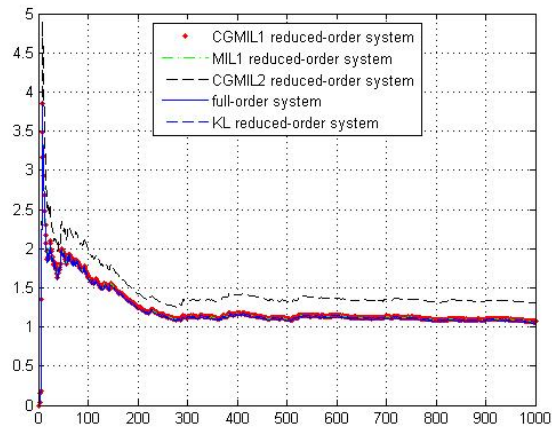

Fig. 2. Gaussian white noise response for full-order system and reduced-order system

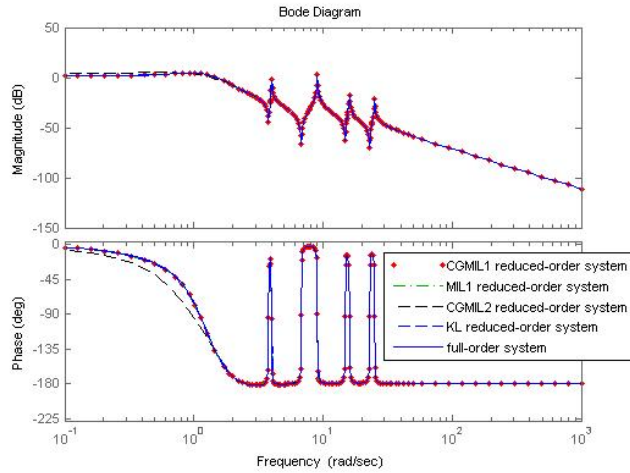As illustrated in Fig. 3 (Bode Plot), the reduced-order closed-loop systems obtained from method 1 and 3 are close to the full-order closed-loop system.

Fig. 3. Bode plots for full-order system and reduced-order system

	CGMIL-RCRP	CGMIL-RCFP	Method 3	MIL-RCRP
E_a^* of the zero-input	0.4139	0.3694	0.3963	0.4139
E_b^* of the zero-input	0.0011	0.0088	9.69e-05	0.0011
E_a^* of the Gaussian white noise	1.0867	1.2382	1.0693	1.0867
E_b^* of the Gaussian white noise	7.7550e-004	0.1367	6.88e-04	7.7550e-004
The LQG performance index J^*	12.5005	16.1723	12.5749	12.5005

Diagram.1 Performances of the reduced-order controllers

2. Deethanizer Model

Distillation column is a common operation unit in chemical industry. We apply these two MIL controller-reduction methods to a 30th-order deethanizer model.

The order of the reduced-order controller is 2. The reduced-order Kalman filter gains and control gains of the reduced-order closed-loop systems are given as follows:

MIL-RCRP: $L_{r1} = [\text{-0.0031 0.0004}]^T$, $K_{r1} = [\text{-0.2289 \quad -0.1007;-0.3751 \quad -0.5665}]^T$;

MIL-RCFP: $L_{r1} = [\text{-0.0054 -0.0082}]^T$, $K_{r2} = [\text{32.8453 \quad 2.0437;-9.4947 \quad 6.6710}]^T$;

We use the same performances as example 1 to measure the reduced-order controller.

Fig. 4 (Impulse Response): When the system input is impulse signal, the reduced-order closed-loop system is close to the full-order system.
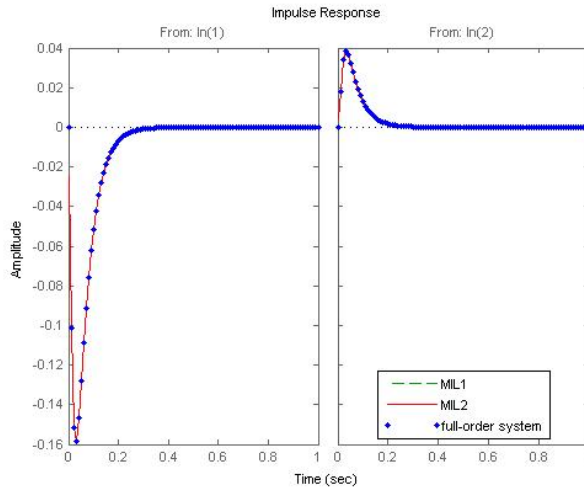

Fig. 4. Impulse response for full-order system and reduced-order system

Fig. 5 (Step Response): When the system input is step signal, the reduced-order closed-loop system is close to the full-order system.
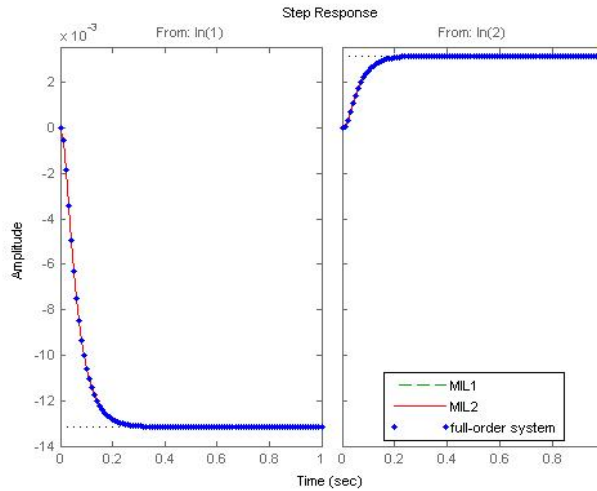

Fig. 5. Step response for full-order system and reduced-order system

Fig. 6 (Gaussian white noise Response): When the system input is Gaussian white noise, the reduced-order closed-loop system is close to the full-order system and outputs are near zero.
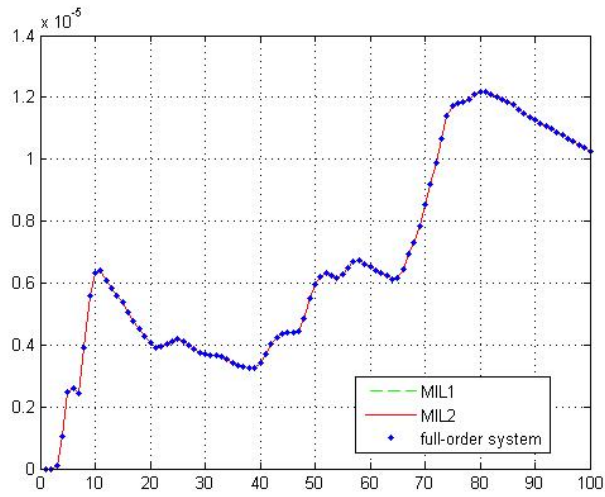
Fig. 6. Gaussian white response for full-order system and reduced-order system
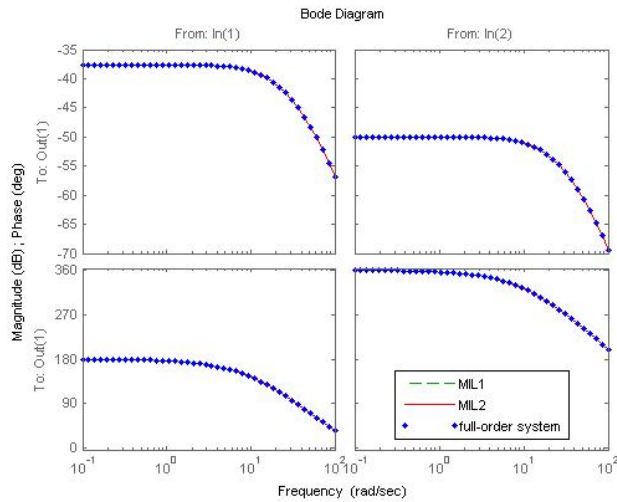
Fig. 7 (Bode Plot):

Fig. 7. Bode plots for full-order system and reduced-order system

	MIL-RCRP	MIL-RCFP	Full-order system
E_a	3.0567e-019	2.4160e-022	0
E_b	2.1658e-005	2.1658e-005	2.1658e-005
J	2.1513e-005	2.1513e-005	2.1513e-005

Diagram.2 Performances of the reduced-order controllers

Conclusion

1. This paper proposed two controller-reduction methods based on the information principle—minimal information loss(MIL). Simulation results show that the reduced-order controllers derived from the proposed two methods can approximate satisfactory performance as the full-order ones.

2. According to the conclusion of literature [17], the closed-loop system with optimal LQG controller is stable. However, its own internal stability can not be guaranteed. If the full-order controller is internal stability, the reduced-order controller is generally stable. We would modify the parameters such as the weighting matrix or noise intensity to avoid the instability of the controller.

3. The performances of the two reduced-order controllers obtained by CGMIL method approximate the full-order one satisfactorily and under certain circumstances. CGMIL method is a better information interpretation instrument of the control system relative to the MIL method, while it is only suit for single-variable stable system.

References

[1] D. C. Hyland and Stephen Richter. On Direct versus Indirect Methods for Reduced-Order Controller Design. IEEE Transactions on Automatic Control, vol. 35, No. 3, pp. 377-379, March 1990.

[2] B. D. O. Anderson and Yi Liu. Controller Reduction: Concepts and Approaches. IEEE Transactions on Automatic Control, vol. 34, No. 8, August, pp. 802-812, 1989.

[3] D. S. Bernstein, D. C. Hyland. The optimal projection equations for fixed-order dynamic compensations. IEEE Transactions on Automatic Control, vol. 29, No. 11, pp. 1034-1037, 1984.

[4] S. Richter. A homotopy algorithm for solving the optimal projection equations for fixed-order dynamic compensation: Existence, convergence and global optimality. In Proc. Amer. Contr. Conf. Minneapolis, MN, June 1987, pp. 1527-1531.

[5] U-L, Ly, A. E. Bryson and R. H. Cannon. Design of low-order compensators using parameter optimization. Automatica, vol. 21, pp. 315-318, 1985.

[6] I. E. Verriest. Suboptimal LQG-design and balanced realization. In Proc. IEEE Conf. Decision Contr. San Diego. CA, Dec. 1981, pp. 686-687.

[7] E. A. Jonckheere and L. M. Silverman. A new set of invariants for linear systems-Application to reduced-order compensator design. IEEE Trans. Automatic. Contr. vol. AC-28, pp. 953-964, 1984.

[8] A. Yousuff and R. E. Skelton. A note on balanced controller reduction. IEEE Trans. Automat. Contr. vol. AC-29, pp. 254-257, 1984.

[9] C. Chiappa, J. F. Magni, Y. Gorrec. A modal multimodel approach for controller order reduction and structuration. Proceedings of the 10th IEEE Conference on Control and Applications, September 2001.

[10] C. De Villemagne and R. E. Skelton. Controller reduction using canonical interactions. IEEE Trans. Automat. Contr. vol. 33, pp. 740-750, 1988.

[11] R. Leland. Reduced-order models and controllers for continuous-time stochastic systems: an information theory approach. IEEE Trans. Automatic Control, 44(9): 1714-1719, 1999.

[12] Hui Zhang, Youxian Sun. Information Theoretic Methods for Stochastic Model Reduction Based on State Projection ,Proceedings of American Control Conference, pp. 2596-2601. June 8-10, Portland, OR, USA, 2005.

[13] Jinbao Fu, Hui Zhang, Youxian Sun. Minimum Information Loss Method based on Cross-Gramian Matrix for Model Reduction (CGMIL). The 7th World Congress on Intelligent Control and Automation WCICA'08, pp. 7339-7343, Chongqing, P. R. China, June.

[14] Yoram Halevi, D. S. Bernstein and M. Haddad. On Stable Full-order and Reduced-order LQG Controllers. Optimal Control Applications and Methods vol.12, pp. 163-172, 1991

[15] S. Ihara. Information Theory for Continuous Systems. Singapore: World Scientific Publishing Co. Pte. Ltd., 1993.

[16] K.V. Fernando and H. Nicholson. On the cross-gramian for symmetric MIMO systems. IEEE Trans. Circuits Systems, CAS—32: 487-489, 1985.

[17] J. C. Doyle and G. Stein. "Robustness with observers", IEEE Trans. Automatic Control, AC-23, 607-611, 1979.

The synthesis problem of the optimum control for nonlinear stochastic structures in the multistructural systems and methods of its solution

Sergey V. Sokolov
Rostov State University of means of communication
Russia

1. The problem statement of the optimum structure selection in nonlinear dynamic systems of random structure

The most interesting and urgent, but unsolved till now, problem in the theory of the dynamic systems of random structure is the synthesis problem for the optimum control of the structure selection in the sense of some known criterion on the basis of the information obtained from the meter. The classical results known in this direction [1] allow to solve the problem of the optimum control only by the system itself (or its specific structure), but not by the selection of the structure.

In this connection the solution of the synthesis problem for the optimum selection control of the structure for the nonlinear stochastic system by the observations of the state vector under the most general assumptions on the character of the criterion applied for the selection optimality is of theoretical and practical interest.

To solve the given problem we formulate it as follows.

For the nonlinear dynamic system of random structure, generally [1] described in the l-th state by the vector equation of form

$$\dot{\xi} = f^{(l)}\left(\xi,t\right) + f_0^{(l)}\left(\xi,t\right)n_t^{(l)}, \qquad \xi\left(t_0\right) = \xi_0 \tag{1}$$

where $l = \overline{1,S}$ is the state number (number of the structure);

$f^{(l)}\left(\xi,t\right), f_0^{(l)}\left(\xi,t\right)$ are nonlinear vector and matrix functions of the appropriate dimension $n^{(l)} \leq N$ and $m^{(l)} \times n^{(l)}$, $N = \max\left(n^{(1)},...,n^{(S)}\right)$;

$\xi(t)$ is the state vector of dimension N in any structure,

$n_t^{(l)}$ is the Gaussian white normalized vector - noise of dimension $m^{(l)}$; the observer of the state vector of which is described, in its turn, by the equation

$$Z = H^{(l)}\left(\xi,t\right) + W_t^{(l)} \tag{2}$$

where Z is the M - dimensional vector of the output signals of the meter;

$H^{(l)}\left(\xi,t\right)$ is the vector - function of the observation of the l-th structure of dimension M;

is the Gaussian white vector - noise with the zero average and matrix of intensities $D_W^{(l)}\left(t\right)$,

we should find such a law of transition $v\left(\xi,t,Z,l = \overline{1,s}\right)$ from one structure into another,

which would provide on the given interval of time $T = [t_0, t_K]$ the optimum of some probabilistic functional J_0 generally nonlinearly dependent on the a posteriori density of distribution $\rho(\xi,t/Z(\tau), \tau \in [t_0, t]) = \rho(\xi,Z,t)$ of the state vector :

$$J_0 = \iint_{T\ \xi_*} \Phi\left[\rho\left(\xi,Z,t\right)\right] d\xi\, dt$$

where ξ_* is the domain of defining argument ξ in which the optimum is searched;
Φ is the given nonlinear analytical function.
Thus the different versions of the form of function Φ allow to cover a wide class of the optimality conditions by accuracy of:
- the probability maximum (minimum) of vector ξ in the area ξ_*: $\Phi(\rho) = \pm\rho$;
- the deviation minimum of the required probability density ρ from the given one g:

$$\Phi\left(\rho\right) = \left(\rho - g\right)^2, \ \Phi\ \left(\rho\right) = \left|\rho - g\right|, \ \Phi\left(\rho\right) = -\rho \ln\left(\frac{g}{\rho}\right) \text{ (the Kulback criterion) etc.;}$$

- the information maximum on the state vector ξ: $\Phi\ \left(\rho\right) = \rho\left[\dfrac{\partial \ln \rho}{\partial \xi}\right]\left[\dfrac{\partial \ln \rho}{\partial \xi}\right]^T$ (the

Fisher criterion) etc.
The similar formulation of the problem covers the selection problem for the optimum structure and in the nonobservable stochastic systems of random structure as well - in this case in the expression for J_0 by ρ we understood the prior density of vector ξ. Thus the form of function Φ should be selected taking into account, naturally, the physical features of the problem solved.
The analysis of the physical contents of the structure selection control providing the optimum of functional J_0 shows that as the vector determining subsequently the control of the structural transitions, it is most expedient to use the vector of intensities of the state change [1, 2]

$$v\left(\xi,Z,t\right) = \left|0 \ \ v_{12}\left(\xi,Z,t\right) \dots \ v_{1s}\left(\xi,Z,t\right) \ v_{21}\left(\xi,Z,t\right) \ 0 \ v_{23}\left(\xi,Z,t\right) \dots \right.$$

$$\left. v_{2s}\left(\xi,Z,t\right) v_{31}\left(\xi,Z,t\right) v_{32}\left(\xi,Z,t\right) \ 0 \ v_{34}\left(\xi,Z,t\right) \dots v_{s(s-1)}\left(\xi,Z,t\right) \ 0\right|^T,$$

where $v_{lr}\left(\xi,Z,t\right)$ is the intensity of transitions from state l into state r, requiring while its forming, for example, in order to prevent the frequent state change, the minimum of its quadratic form on the given interval of time T for $\xi \in \xi_*$, i.e.

$$\min \iint_{T\ \xi_*} v^T\left(\xi,Z,t\right)v\left(\xi,Z,t\right)d\xi\,dt\,.$$

As far as vector v contains the zero components, then in essence, from here on the search not of vector v itself, but vector v_0, related to it by the relationship $v = E_0 v_0$, is carried out, where v_0 is the vector formed from vector v by eliminating the zero components;

E_0 is the matrix formed from the unit one by adding the zero rows to form the appropriate zero elements in vector v.

And finally the minimized criterion J takes the form

$$J = \iint_{T\ \xi_*}\left(\Phi\left[\rho\left(\xi,Z,t\right)\right]+v_0^T\left(\xi,Z,t\right)v_0\left(\xi,Z,t\right)\right)d\xi\,dt \tag{3}$$

In its turn, for process ξ described by equations (1), the density of its a posteriori distribution ρ (DAPD) can be given as

$$\rho\left(\xi,Z,t\right)=\sum_{l=1}^{s}\omega\left(\xi,Z,l,t\right)=\sum_{l=1}^{s}\omega_Z^{(l)}\left(\xi,t\right),$$

where $\omega_Z^{(l)}(\xi,t)$ is the DAPD of the extended vector $\left|\begin{matrix}\xi\\l\end{matrix}\right|$ (l is the state number).

In the case of the continuous process ξ, which is most typical for practice , when the restored values of the l-th state coincide with the final value of the process of the r-th state, functions $\omega_Z^{(l)}(\xi,t)$, $l=\overline{1,S}$, are described by the following system of the Stratonovich generalized equations [1]:

$$\frac{\partial\omega_Z^{(l)}\left(\xi,t\right)}{\partial t}=L\left[\omega_Z^{(l)}\left(\xi,t\right)\right]+Q\left[\omega_Z^{(l)}\left(\xi,t\right)\right]-\sum_{r=1}^{s}v_{lr}\left(\xi,t\right)\omega_Z^{(l)}\left(\xi,t\right)+\sum_{r=1}^{s}v_{rl}\left(\xi,t\right)\omega_Z^{(r)}\left(\xi,t\right),$$

$$l=\overline{1,s},$$

$$Q\left[\omega_Z^{(l)}\left(\xi,t\right)\right]=-\frac{1}{2}\omega_Z^{(l)}\left(\xi,t\right)\left[\gamma^{(l)}\left(\xi,Z,t\right)-\sum_{k=1}^{s}\int_{-\infty}^{\infty}\gamma^{(l)}\left(\xi,Z,t\right)\omega_Z^{(k)}\left(\xi,t\right)d\xi\right],$$

$$\gamma^{(l)}\left(\xi,Z,t\right)=\sum_{p,q=1}^{M}\frac{\hat{D}_{pq}^{(l)}\left(t\right)}{\left|D_W^{(l)}\left(t\right)\right|}\left[Z_p-H_p^{(l)}\left(\xi,t\right)\right]\left[Z_q-H_q^{(l)}\left(\xi,t\right)\right],$$

$\hat{D}_{pq}^{(l)}\left(t\right)$ is the algebraic addition of the pq-th element in the determinant $\left|D_W^{(l)}\left(t\right)\right|$ of matrix $D_W^{(l)}\left(t\right)$;

p,q are indexes of the respective components of vectors;

L is the Fokker –Planck (FP) operator;

or entering vector $v_0\,(\xi,Z,t)$ and vector $\omega_Z\left(\xi,t\right)=\left|\omega_Z^{(1)}\left(\xi,t\right)\;...\;\omega_Z^{(s)}\left(\xi,t\right)\right|^T$, we have in the general form:

$$\frac{\partial\omega_Z\left(\xi,t\right)}{\partial t}=U\Big[\omega_Z\left(\xi,t\right)\Big]-\Big[\Omega\big[\omega_Z\left(\xi,t\right)\big]\big(E_s\otimes I_s\big)-\omega_Z^T\left(\xi,t\right)\otimes E_s\Big]E_0 v_0\left(\xi,Z,t\right),$$

$$U\big[\omega_Z\big]=L\big[\omega_Z\big]+Q\big[\omega_Z\big], \tag{4}$$

where E_S is the unit matrix of dimension S;
I_S is the unit row of dimension S;

$\otimes$ is the symbol of the Kronecker product;

$$\Omega(\omega_Z)=\begin{vmatrix}\omega_Z^{(1)} & & & 0\\ & \omega_Z^{(2)} & & \\ & & ... & \\ 0 & & & \omega_Z^{(s)}\end{vmatrix}.$$

For the nonobservable dynamic systems the FP generalized equations are derived from (4) at $Q = 0$.

Taking into account that introducing vector ω_Z for density ρ the expression has the form

$$\rho\left(\xi,Z,t\right)=I_s\omega_Z\left(\xi,t\right),$$

functional (3) is given as

$$J=\iint_{T\;\xi_*}\Big[\Phi\big[I_S\omega_Z\left(\xi,t\right)\big]+v_0^T\left(\xi,Z,t\right)v_0\left(\xi,Z,t\right)\Big]d\xi\;dt=\int_T W_*\left(t\right)dt, \tag{5}$$

and for the simplification of the subsequent solution the vector equation (4) is rewritten as follows:

$$\frac{\partial\omega_Z}{\partial t}=U(\omega_Z)-\Big[\Omega(\omega_Z)\big(E_S\otimes I_S\big)-\omega_Z^T\otimes E_S\Big]E_0 v_0 =U(\omega_Z)-F(\omega_Z)v_0. \tag{6}$$

Then the problem stated finally can be formulated as the problem of search of vector V_0, that provides the synthesis of such a vector ω_Z described by equation (6), which would deliver the minimum to functional (5). The synthesis of the optimum vector ω_Z allows immediately

to solve the problem of the selection of the optimum structure by defining the maximal components of the vector of the state probabilities $P(t) = \int_{-\infty}^{\infty} \omega_Z(\xi,t)d\xi$ [1].

2. The general solution of the synthesis problem for the stochastic structure control

For the further solution of the problem we use the method of the dynamic programming, according to which by search of the optimum control in the class of the limited piecewise-continuous functions with values from the open area V_* the problem is reduced to the solution of the functional equation [3]

$$\min_{v \in V_*} \left\{ \frac{dV}{dt} + W_* \right\} = 0 \qquad (7)$$

under the final condition $V(t_K) = 0$ with respect to optimum functional V, parametrically dependent on time $t \in T$ and determined on a set of vector - functions ω_Z, satisfying equation (6).
For the linear systems functional V is found as the integrated quadratic form [3]

$$V = \int_{\xi_*} \omega_Z^T(\xi,t) v(\xi,t) \omega_z(\xi,t) d\xi,$$

v is a $S{\times}S$ matrix , whence we have:

$$\frac{dV}{dt} + W_* = \int_{\xi_*} \left\{ \omega_Z^T \frac{\partial v}{\partial t} \omega_Z + \omega_Z^T \left(v^T + v \right) \frac{\partial \omega_Z}{\partial t} + \Phi\left(I_S \omega_Z \right) + v_0^T v_0 \right\} d\xi,$$

and taking into account equation (6) for ω_Z we obtain the initial expression for the subsequent definition of the optimum one V_0^*

$$\frac{dV}{dt} + W_* = \int_{\xi_*} \left\{ \omega_Z^T \frac{\partial v}{\partial t} \omega_Z + \omega_Z^T \left(v^T + v \right)\left(U(\omega_Z) - F(\omega_Z)v_0 \right) + \Phi\left(I_S \omega_Z \right) + v_0^T v_0 \right\} d\xi. \qquad (8)$$

The analysis of the given expression shows that the definition of vector V_0^* from the solution of the functional equation (7) is reduced to the classical problem of search of vector - function realizing the minimum of the certain integral (8). Thus the required vector - function $V_0^*(\xi,Z,t)$ should satisfy the following system of the Euler equations:

$$-F^T\left(\omega_Z \right)\left(v^T + v \right)\omega_Z + 2v_0^* = 0,$$

whence

$$v_0^* = \frac{1}{2} F^T \left(\omega_Z \right) \left(v^T + v \right) \omega_Z \ .$$

The substitution of the found optimum law for the change of state V_0^* into (6) allows to write down the equations for the optimum (in the sense of (5)) vector ω_Z

$$\frac{\partial \omega_Z}{\partial t} = U \left(\omega_Z \right) - \frac{1}{2} F \left(\omega_Z \right) F^T \left(\omega_Z \right) \left(v^T + v \right) \omega_Z \ , \tag{9}$$

the integration of which, in its turn, completes the solution of the selection problem of the optimum structure by defining the maximal component of the vector of the state probabilities

$$P(t) = \int\limits_{-\infty}^{\infty} \omega_z \left(\xi, t \right) d\xi \ .$$

The equations required for defining the matrix function v(ξ,t), included in (9), follow from the constrain

$$\frac{dV}{dt} + W_* = 0 \bigg|_{v_0 = v_0^*}$$

after substituting the found vector V_0^* into (8)

$$\frac{\partial v}{\partial t} = -\frac{1}{s} \left(v^T + v \right) U \left(\omega_Z \right) \omega_0^T - \frac{1}{s^2} \omega_0 \omega_0^T \Phi \left(I_S \omega_Z \right) + \frac{1}{4} \left(v^T + v \right) F \left(\omega_Z \right) F^T \left(\omega_Z \right) \left(v^T + v \right), \tag{10}$$

where $\omega_0 = \left| \dfrac{1}{\omega_Z^{(1)}} \ \dots \ \dfrac{1}{\omega_Z^{(S)}} \right|^T$ is the auxiliary vector introduced for convenience in

transformations and simplification of the record of equation (10).

The joint solution of systems (9, 10) under the boundary-value constrains $\omega_z \left(\xi, t_0 \right) = \omega_{z0}$, $v \left(\xi, t_k \right) = 0$ exhausts, in essence, the theoretical solution of the problem stated, including the prior case as well - for the nonobservable dynamic systems of random structure.

It should be noted that in case of forming V_0^* in the assumption of its independence on ξ, the integrated dependence V_0^* on ω_Z follows from the constrain of minimization of the functional equation (8)

$$v_0^* = \left(2\Delta \right)^{-1} \int\limits_{\xi_*} F^T \left(\omega_z \right) \left(v^T + v \right) \omega_z d\xi, \ \Delta = \int\limits_{\xi_*} d\xi,$$

which after substitution into (6) and (8) results in forming a system already of the integro-differential (as distinct from (9, 10)) equations with partial derivatives, the solution search of which appears to be much more difficult than in the first case.

The synthesis problem of the optimum control for nonlinear
stochastic structures in the multistructural systems and methods of its solution
377

In spite of the fact that the found theoretical solution of the problem stated defines the basic feasibility of the optimum selection of the process structure ξ, the practical solution of the boundary-value problem directly for the conjugated system of equations with the partial derivatives (9, 10) represents a problem now.

Then in this connection as one of the solution methods of this problem (as the most universal one) we consider the method using the expansion of functions ω_z, v into series by some system of the orthonormal functions $\phi = |\phi_1 \ldots \phi_N|^T$ of vector argument:

$$v(\xi,t) = B(t)\phi(\xi), \quad \omega_z(\xi,t) = A(t)\phi(\xi); \quad A(t), B(t)$$ are ordinary and block matrixes of

factors of expansion determined in the course of solving. In this case the problem is reduced to the point-to-point boundary-value problem of integration of the matrix system already of the ordinary differential equations:

$$\frac{\partial A}{\partial t} = \int_{\xi} \left\{ U(A\phi)\phi^T - \frac{1}{2}F(A\phi)F^T(A\phi)(\phi^T B^T + B\phi)(A\phi)\phi^T \right\} d\xi,$$

$$\frac{\partial B}{\partial t} = -\frac{1}{S}\int_{\xi} \left\{ (\phi^T B^T + B\phi)U(A\phi)\omega_0^T(A\phi) - \frac{1}{S^2}\omega_0(A\phi)\omega_0^T(A\phi)\Phi(I_S A\phi) + \right.$$

$$\left. + \frac{1}{4}(\phi^T B^T + B\phi)F(A\phi)F^T(A\phi)(\phi^T B^T + B\phi) \right\}\phi^T d\xi,$$

the solution of which appears to be already much easier and can be carried out by various traditional ways : the ranging short method , the method of the invariant plunge etc. The feature of the practical realization of the solution in this case is the absence of the rigid requirements to its accuracy, as in case of the structure selection the number of the maximal

components of vector $P(t) = A(t)\int_{\xi}\phi(\xi)d\xi$, rather than its value is only defined. For the

illustration of the real feasibility of applying the similar approach we consider the example. For the nonlinear stochastic process of random structure described by the equation

$$\dot{\xi} = f^{(l)}(\xi,t) + n_t,$$

where $l = 1,2$; $f^{(1)}(\xi,t) = -\xi^2$, $f^{(2)}(\xi,t) = -\xi + 0,01\xi^3$,
n_t is the Gaussian normalized white noise,
the equation of the observer has the form

$$Z = H^{(l)}(\xi,t) + W_t,$$

where $l = 1,2$; $H^{(1)}(\xi,t) = 0,5\xi^2$, $H^{(2)}(\xi,t) = 1,2\xi^2 - 0,1\xi^3$;
W_t is the Gaussian normalized white noise.

It is required to carry out the structure selection of process ξ providing the maximum of probability of its occurrence in the given limits $\xi_* = \left[\xi_{min} = -0,5; \xi_{max} = 0,7\right]$ in time interval $T = [0; 300]$ s., i.e. the minimizing criterion

$$J = \int\limits_{T}\int\limits_{\xi_*} \left\{ -\rho(\xi,t) + v_0^T(\xi,t)v_0(\xi,t) \right\} d\xi\, dt,$$

where

$$v_0 = \begin{vmatrix} v_{12} \\ v_{21} \end{vmatrix}; \qquad \rho(\xi,t) = \omega_z^{(1)}(\xi,t) + \omega_z^{(2)}(\xi,t);$$

$$\frac{\partial \omega_z^{(1)}}{\partial t} = \frac{\partial}{\partial \xi}\left(\xi^2 \omega_z^{(1)}\right) + \frac{1}{2}\frac{\partial^2 \omega_z^{(1)}}{\partial \xi^2} -$$

$$-\frac{1}{2}\omega_z^{(1)}(\xi,t)\left[\left(Z-0,5\xi^2\right)^2 - \sum_{k=1}^{2}\int\limits_{-\infty}^{\infty}\left(Z-0,5x^2\right)^2 \omega_z^{(k)}(x,t)\,dx\right] - v_{12}\omega_z^{(1)} + v_{21}\omega_z^{(2)};$$

$$\frac{\partial \omega_z^{(2)}}{\partial t} = \frac{\partial}{\partial \xi}\left[\left(\xi-0,01\xi^3\right)\omega_z^{(2)}\right] + \frac{1}{2}\frac{\partial^2 \omega_z^{(2)}}{\partial \xi^2} - \frac{1}{2}\omega_z^{(2)}(\xi,t)\left[\left(Z-1,2\xi^2+0,1\xi^3\right)^2 -\right.$$

$$\left. -\sum_{k=1}^{2}\int\limits_{-\infty}^{\infty}\left(Z-1,2x^2+0,1x^3\right)^2 \omega_z^{(k)}(x,t)\,dx\right] - v_{21}\omega_z^{(2)} + v_{12}\omega_z^{(1)},$$

or in the vector form

$$\frac{\partial \omega_z}{\partial t} = U(\omega_z) - F(\omega_z)v_0, \qquad \omega_z = \begin{vmatrix} \omega_z^{(1)} & \omega_z^{(2)} \end{vmatrix}^T,$$

$$F(\omega_z) = \begin{vmatrix} \omega_z^{(1)} & -\omega_z^{(2)} \\ -\omega_z^{(1)} & \omega_z^{(2)} \end{vmatrix}.$$

In this case equations (9, 10) for the optimum vector ω_z and conjugated matrix function V have the form:

$$\frac{\partial \omega_z}{\partial t} = U(\omega_z) - \frac{1}{2}\begin{vmatrix} \Omega_0 & -\Omega_0 \\ -\Omega_0 & \Omega_0 \end{vmatrix}\left(V^T + V\right)\omega_z,$$

$$\frac{\partial V}{\partial t} = -\frac{1}{2}\left(V^T + V\right)U(\omega_z)\begin{vmatrix} \dfrac{1}{\omega_z^{(1)}} & \dfrac{1}{\omega_z^{(2)}} \end{vmatrix} +$$

$$+\frac{1}{4}\begin{vmatrix} \dfrac{\omega_z^{(1)}+\omega_z^{(2)}}{\omega_z^{(1)2}} & \dfrac{\omega_z^{(1)}+\omega_z^{(2)}}{\omega_z^{(1)}\omega_z^{(2)}} \\ \dfrac{\omega_z^{(1)}+\omega_z^{(2)}}{\omega_z^{(1)}\omega_z^{(2)}} & \dfrac{\omega_z^{(1)}+\omega_z^{(2)}}{\omega_z^{(2)2}} \end{vmatrix} + \frac{1}{4}\left(V^T+V\right)\begin{vmatrix} \Omega_0 & -\Omega_0 \\ -\Omega_0 & \Omega_0 \end{vmatrix}\left(V^T+V\right),$$

$$\Omega_0 = \omega_z^{(1)2} + \omega_z^{(2)2}.$$

The solution of the given problem was carried out on the basis of the approximation of functions ω_Z, V by the Fourier series in interval [-5; 5] within the accuracy of 4 terms of expansion and integration of the obtained system of equations for factors of expansion by means of the approximated method of the invariant plunge on time interval [0; 300] s. When integration and formation of the approximated values of functions $\omega_{Z^{(1)}}$, $\omega_{Z^{(2)}}$ had been taken place, the numbers of the structures selected by the character of the maximal state probability at the current time, appeared to be distributed in time as follows:

        - in interval [0; 48] s - the second structure;
        - in interval [48; 97] s - the first structure;
        - in interval [97; 300] s - the second structure.

Concurrently the integration of the system of the DAPD equations $\omega_{Z^{(1)}}$, $\omega_{Z^{(2)}}$ was carried out for the traditional case of the uncontrolled change of states with the unit intensity [1] and it was established that in the latter case the value of the minimized criterion J had appeared to be by a factor of 1,35 higher than that of in the optimum control of the structure selection.

Thus it allows to make a conclusion not only on the theoretical solution of the general problem for the optimum structure control of the stochastic dynamic systems, but on the feasibility of the effective practical application of the method developed as well. It is obvious that for the dynamic systems of the great dimension again there arises the traditional problem of the numerical realization of the approach suggested. As far as one of the ways of simplifying the control algorithms lays in the direction of the preliminary solution approximation of the vector equation of the distribution density, then for the multistructural system we apply the approach suggested in [1], allowing to write down the differential equations of the parameters of density, approximating the initial one. Thus it should be taken into account the character of the approximation , which arises only in the multistructural system, - only the normalized densities of the state vectors of each structure (ignoring the probabilities of occurring the structures themselves) are approximated.

In this case the function of the distribution density ξ of the state vector can be given as [1]

$$\rho(\xi, t) = \sum_{l=1}^{S} P_l \rho_l(\xi, t),\qquad(11)$$

where P_l is the probability of the l-th structure;
$\rho_l(\xi, t)$ is the distribution density ξ in the l-th structure.

Then the scheme of the synthesis of the structure control, following from the similar approach, we consider in detail. Thus we investigate the prior case (for the nonobservable structures), as a more adequate one to the practical applications (due to the character of forming the terminal control requiring the whole set of measurements on the given time interval of optimization, that for a number of practical problems is unacceptable). But it should be noted that the synthesis of the a posteriori control on the basis of the approach described below does not practically differ from the prior one - only by the additional component caused by the observation availability in the right-hand part of the equations of parameters (12) [1] given below (not affecting the procedure of forming the control required). So, we investigate the synthesis method for the structure control using the approximating representation of the densities of the state vectors of the dynamic structures.

3. Suboptimum selection control of nonlinear dynamic structures

In solving the practical problems of the analysis and synthesis of systems with random structure to describe densities ρ_l one uses, as a rule, their Gaussian approximation $\tilde{\rho}_l$ which results in specifying the parameters determining $\tilde{\rho}_l$ - the vector of the mathematical expectation $\hat{\xi}^{(l)}$ and the covariance matrix $R^{(l)}$, in the form of the known system of the ordinary differential equations (providing the required trade-off of accuracy against the volume of the computing expenses [4]).

In the case investigated below for the continuous process ξ, when the restored values of the l-th state coincide with the final values of the process of the r-th state, the system of equations for parameters $\tilde{\rho}$, obtained on the basis of the Gaussian approximation ρ, has the following form [1, 4]:

$$\dot{\hat{P}}_l = -\sum_{r=1}^{S}\left(\hat{P}_l v_{lr}\left(\hat{\xi}^{(l)}, R^{(l)}, t\right) - \hat{P}_r v_{rl}\left(\hat{\xi}^{(r)}, R^{(r)}, t\right)\right),$$

$$\dot{\hat{\xi}}^{(l)} = f^{(l)}\left(\hat{\xi}^{(l)}, t\right) + \sum_{r=1}^{S}\frac{\hat{P}_r(t)}{\hat{P}_l(t)}v_{rl}\left(\hat{\xi}^{(r)}, R^{(r)}, t\right)\left[\hat{\xi}^{(r)} - \hat{\xi}^{(l)}\right],$$

$$\dot{R}^{(l)} = R^{(l)}\frac{\partial f^{(l)}}{\partial\hat{\xi}}\left(\hat{\xi}^{(l)}, t\right) + \frac{\partial f^{(l)T}}{\partial\hat{\xi}}\left(\hat{\xi}^{(l)}, t\right)R^{(l)} + f_0^{(l)}\left(\hat{\xi}^{(l)}, t\right)f_0^{(l)^T}\left(\hat{\xi}^{(l)}, t\right) +$$

$$+\sum_{r=1}^{S}\frac{\hat{P}_r(t)}{\hat{P}_l(t)}v_{rl}\left(\hat{\xi}^{(r)}, R^{(r)}, t\right)\left(R^{(r)} - R^{(l)} + \left(\hat{\xi}^{(r)} - \hat{\xi}^{(l)}\right)\left(\hat{\xi}^{(r)} - \hat{\xi}^{(l)}\right)^T\right), \qquad (12)$$

$$l = \overline{1,\, S}\,,$$

$\hat{P}_l$ is the probability of the l-th process structure under the Gaussian approximation;

$v_{lr}\left(\hat{\xi}^{(l)}, R^{(l)}, t\right)$ is the intensity of transitions from state l into state r.

To find the solution required in the general form we present the given system as follows. Introduce vectors

$$\hat{P} = \left|\hat{P}_1 \ldots \hat{P}_S\right|^T, \qquad \hat{\xi} = \left|\hat{\xi}^{(1)T} \ldots \hat{\xi}^{(S)T}\right|^T$$

and using the operation of transformation of matrix A with components a_{ij} ($i,j=m,n$) into vector $A^{(v)}$:

$$A^{(v)} = \left|a_{11}a_{21}\ldots a_{m1}a_{12}a_{22}\ldots a_{m2}\ldots a_{1n}a_{2n}\ldots a_{mn}\right|^T,$$

we previously write down the equations of parameters $\tilde{\rho}$, separating the components into dependent and independent ones on intensities v_{rl}:

$$\hat{P} = \frac{\left| \displaystyle\sum_{r=1}^{S}\left(\hat{P}_r v_{rl} - \hat{P}_l v_{lr}\right) \right|}{\left| \displaystyle\sum_{r=1}^{S}\left(\hat{P}_r v_{rS} - \hat{P}_S v_{Sr}\right) \right|}, \qquad \hat{\xi} = \beta\left(\hat{\xi},t\right) + \frac{\left| \displaystyle\sum_{r=1}^{S} G_{rl} v_{rl} \right|}{\left| \displaystyle\sum_{r=1}^{S} G_{rS} v_{rS} \right|},$$

$$\dot{R}^{(v)} = \Psi\left(\hat{\xi}, R^{(v)}, t\right) + \frac{\left| \displaystyle\sum_{r=1}^{S} Q_{rl} v_{rl} \right|}{\left| \displaystyle\sum_{r=1}^{S} Q_{rS} v_{rS} \right|}, \tag{13}$$

$$R^{(v)} = \left| \frac{R^{(1)(v)}}{R^{(S)(v)}} \right|, \qquad G_{rl} = G_{rl}\left(\hat{\xi}, \hat{P}\right) = \frac{\hat{P}_r}{\hat{P}_l}\left(\hat{\xi}^{(r)} - \hat{\xi}^{(l)}\right),$$

$$Q_{rl} = Q_{rl}\left(\hat{\xi}, R^{(v)}, \hat{P}\right) = \left(R^{(r)} - R^{(l)} + \left(\hat{\xi}^{(r)} - \hat{\xi}^{(l)}\right)\left(\hat{\xi}^{(r)} - \hat{\xi}^{(l)}\right)^T\right)^{(v)} \frac{\hat{P}_r}{\hat{P}_l},$$

and then unifying vectors $\hat{\xi}$ and $R^{(V)}$ into the generalized vector $\hat{X} = \left| \dfrac{\hat{\xi}}{R^{(V)}} \right|$ as well.

Then as the vector, determining the structural transition control in the system examined, we use, similarly to the above-stated, the vector of intensities of the state change

$$v\left(\hat{P}, X, t\right) = \left| 0 \ v_{21}\left(\hat{P}, X, t\right) \dots v_{S1}\left(\hat{P}, X, t\right) \ v_{12}\left(\hat{P}, X, t\right) \ 0 \ v_{32}\left(\hat{P}, X, t\right) \dots \right.$$

$$\left. v_{S2}\left(\hat{P}, X, t\right) v_{13}\left(\hat{P}, X, t\right) \ v_{23}\left(\hat{P}, X, t\right) \ 0 \ v_{43}\left(\hat{P}, X, t\right) \dots v_{(S-1)S}\left(\hat{P}, X, t\right) \ 0 \right|^T,$$

that results in the following form of the minimized criterion J:

$$J = \int_T \int_{\xi_*} \Phi\left[\tilde{\rho}\left(\xi, \hat{P}, X, t\right)\right] d\xi \, dt + \int_T v_0^T\left(\hat{P}, X, t\right) v_0\left(\hat{P}, X, t\right) dt. \tag{14}$$

For the problem solution in view of the designations accepted the system of equations (13) is given as:

$$\hat{P} = \left[\left(E \otimes \hat{P}^T\right) - \hat{P}\left(E \otimes I_S\right)E_1\right]E_0 v_0,$$

$$\dot{X} = \varphi\left(X, t\right) + T\left(\hat{P}, X\right)E_0 v_0, \tag{15}$$

where $\otimes$ is the symbol of the Kronecker product;

I_S is the unit row of dimension S;

$$E_1 = \begin{vmatrix} 100.........................0 \\ \underbrace{0...0}_{S}10.....................0 \\ \underbrace{0......0}_{2S}10...................0 \\ .... \\ \underbrace{0..........0}_{S \times (S-1)}10..............0 \\ 010.........................0 \\ 0\underbrace{0...0}_{S}1.....................0 \\ 0\underbrace{0........0}_{2S}10..............0 \\ .... \\ \underbrace{0............................}_{S \times S-1}01 \end{vmatrix} , \; \varphi(X,t) = \begin{vmatrix} \beta(X,t) \\ \Psi(X,t) \end{vmatrix} , \; E \text{ is the unit matrix,}$$

$$T(\hat{P},X) = \begin{vmatrix} G_{11}...G_{S1} & & & \\ & G_{12}...G_{S2} & 0 & \\ & 0 & & \ddots \\ & & & G_{1S}...G_{SS} \\ Q_{11}...Q_{S1} & & & \\ & Q_{12}...Q_{S2} & 0 & \\ & 0 & & \ddots \\ & & & Q_{1S}...Q_{SS} \end{vmatrix} ,$$

or finally in the canonical form:

$$\begin{vmatrix} \hat{P} \\ \dot{X} \end{vmatrix} = A(X,t) + B(\hat{P},X)v_0 , \qquad A(X,t) = \begin{vmatrix} 0 \\ \varphi(X,t) \end{vmatrix} ,$$

$$B(\hat{P},X) = \begin{vmatrix} (E \otimes \hat{P})^T - \hat{P}(E \otimes I_S)E_1 \\ T(\hat{P},X) \end{vmatrix} E_0 . \tag{16}$$

Thus the solution of the problem stated is reduced, first, to the definition of vector V_0 in (16), providing the minimum of functional (14) (in view of the use in the latter as the density

function its approximation $\tilde{\rho}\left(\xi,\hat{P},X,t\right)$, in this case a Gaussian one), and, secondly, to the

selection of the maximal component in vector $\hat{P}$ the number of which will define the
number of the required structure of the state vector.
The first stage of the solution - the definition of the optimum vector V_0 can be carried out by
means of the principle of the maximum [5]. In this case the Hamiltonian H_* has the form

$$H_*\left(\hat{P},X,t\right)=\int_{\xi_*}\Phi\left[\tilde{\rho}\left(\xi,\hat{P},X,t\right)\right]d\xi+v_0^T\left(\hat{P},X,t\right)v_0\left(\hat{P},X,t\right)+$$

$$+\lambda^T\left\{A\left(X,t\right)+B\left(\hat{P},X\right)v_0\left(\hat{P},X,t\right)\right\},$$

λ is the vector of the conjugate variables;
whence from the stationarity condition H_* the optimum vector V_0^* is defined directly:

$$v_0^*=-\frac{1}{2}B\left(\hat{P},X\right)^T\lambda.$$

The substitution of vector V_0^* into (16) and conjugate system of equations results in the
need for solving the point-to-point boundary-value problem for the following system of
equations:

$$\left|\begin{matrix}\dot{\hat{P}}\\\dot{X}\end{matrix}\right|=A\left(X,t\right)-\frac{1}{2}B\left(\hat{P},X\right)B^T\left(\hat{P},X\right)\lambda,$$

$$\dot{\lambda}=-\int_{\xi_*}\left|\frac{\partial\Phi\left(\xi,\hat{P},X,t\right)}{\partial\hat{P}}\frac{\partial\Phi\left(\xi,\hat{P},X,t\right)}{\partial X}\right|^T d\xi-$$

$$-\left(\left|0\quad\frac{\partial A\left(X,t\right)}{\partial X}\right|^T-\frac{1}{2}\lambda^T B\left(\hat{P},X\right)\left|\frac{\partial B\left(\hat{P},X\right)}{\partial\hat{P}}\frac{\partial B\left(\hat{P},X\right)}{\partial X}\right|^T\right)\lambda,\qquad(17)$$

$$P\left(t_0\right)=\hat{P}_0,\qquad X\left(t_0\right)=X_0,\qquad\lambda\left(t_k\right)=0,$$

the integration of which exhausts the theoretical solution of the problem stated- the
subsequent selection of the maximal component of vector $\hat{P}$, determining the current
number of the required structure of the state vector , does not represent any problem.
Moreover, as above, in selecting the structure the exact value of the maximal component of
vector $\hat{P}$ is not essentially required, only its number is of importance. In practical solving of
problem (17) this allows to use the approximated methods focused on the required trade-off
of accuracy against the volume of the calculations, for example, already approved method of

the approximated invariant plunge, transforming system (17) into the system of the ordinary differential equations solved in the real time [6]:

$$
\begin{vmatrix} \dot{\hat{P}}_* \\ \dot{X}_* \end{vmatrix} = A(X_*,t) - D(t)\int_{\xi_*} \left| \frac{\partial \Phi(\xi,\hat{P}_*,X_*,t)}{\partial \hat{P}_*} \frac{\partial \Phi(\xi,\hat{P}_*,X_*,t)}{\partial X_*} \right|^T d\xi \ ,
$$

$$
\dot{D} = 2 \left| 0 \quad \frac{\partial A(X_*,t)}{\partial X_*} \right| D + D \left| 0 \quad \frac{\partial A(X_*,t)}{\partial X_*} \right|^T +
$$

$$
+ \frac{1}{2} B(\hat{P}_*,X_*) B^T(\hat{P}_*,X_*) - 2D \int_{\xi_*} \left| \frac{\partial^2 \Phi(\xi,\hat{P}_*,X_*,t)}{\partial \hat{P}_* \partial \hat{P}_*} \right.
$$

$$
\left. \frac{\partial^2 \Phi(\xi,\hat{P}_*,X_*,t)}{\partial \hat{P}_* \partial X_*} \frac{\partial^2 \Phi(\xi,\hat{P}_*,X_*,t)}{\partial X_* \partial \hat{P}_*} \frac{\partial^2 \Phi(\xi,\hat{P}_*,X_*,t)}{\partial X_* \partial X_*} \right|^T d\xi D \ ,
$$

where $\hat{P}_*$, X_* are approximated solutions of system (17);
D is the matrix of the weight factors for the deviation of the approximated solution from the required one [6].
The example illustrating the feasibility of the real application of the approach given is considered in [7].
To reduce the computing expenses on the basis of the approximation of the DAPD let us consider the algorithm of the synthesis using the local criterion.

4. The a posteriori local - optimum control for the structure selection

Local criterion J:

$$
J = \int_{\xi_*} \varphi_0 \left[\omega_z(\xi,t) \right] d\xi + \int_{t_0}\int_{\xi_*} v_0^T(\xi,z,t) v_0(\xi,z,t) d\xi \ dt \ , \tag{18}
$$

where φ_0 is the non-negatively defined scalar function generally dependent already directly on vector ω_z,
is typical for the process control in the real time.
In this case the solution is reduced to the search of such vector
$V_0 = V_0(\xi,z,t)$, which would provide the minimum of criterion (18) provided that vector $\omega_Z(\xi,t)$ is described by the known vector equation with partial derivatives, deduced in [1] and transformed in paragraph 1 to the following form (6):

$$
\frac{\partial \omega_z}{\partial t} = U(\omega_z) - F(\omega_z) v_0 \ ,
$$

where $U(\omega_Z)$ is the vector - function representing the vector generalization of the right-hand part of the Stratonovich equation [1];
$F(\omega_Z)$ is the matrix function linearly dependent on the component of vector ω_Z.
In the case considered this allows to obtain the expression of the minimized function $I = \left(-\dot{J}\right)$ for the further definition of the optimum intensity vector for the state change of structures V_0^*

$$I = -\int_{\xi_*} \left(\frac{\partial \varphi_0}{\partial \omega_z} \frac{\partial \omega_z}{\partial t} + v_0^T v_0 \right) d\xi ,$$

or in view of the right-hand part of equation (6):

$$I = -\int_{\xi_*} \left(\frac{\partial \varphi_0}{\partial \omega_z} \left[U(\omega_z) - F(\omega_z) v_0 \right] + v_0^T v_0 \right) d\xi . \tag{19}$$

From the condition of the maximum for expression (19) we have the initial equation for defining vector V_0^*

$$-\frac{\partial \varphi_0}{\partial \omega_z} F(\omega_z) + 2v_0^T = 0 ,$$

whence the vector required

$$v_0^* = \frac{1}{2} F^T(\omega_z) \frac{\partial \varphi_0^T}{\partial \omega_z} . \tag{20}$$

If instead of (18) we use a less general expression, which does not impose any restrictions on the definitional domain of vector V_0, as the criterion of optimization:

$$J_1 = \int_{\xi_*} \varphi_0 \left[\omega_z(\xi,t) \right] d\xi + \int_{t_0}^{t} v_0^T(z,t) v_0(z,t) \, dt , \tag{21}$$

then expression $I_1 = \left(-\dot{J}_1\right)$ as compared to (19) is modified as follows:

$$I_1 = -\left(\int_{\xi_*} \frac{\partial \varphi_0}{\partial \omega_z} \left[U(\omega_z) - F(\omega_z) v_0 \right] d\xi + v_0^T v_0 \right)$$

whence

$$v_0^* = \frac{1}{2} \int_{\xi_*} F^T(\omega_z) \frac{\partial \varphi_0^T}{\partial \omega_z} d\xi \qquad . \tag{22}$$

The substitution of the found optimum vector V_0^* (20) (or (22)) in equation (6) allows to form the equation, describing the required vector ω_Z of the distribution densities of the extended state vectors for the system with the intensity of their change, which provides the optimum of functional (18) (or (21), respectively). So, in case of selecting as the criterion of optimality (18) we have:

$$\frac{\partial \omega_z}{\partial t} = U(\omega_z) - \frac{1}{2} F(\omega_z) F^T(\omega_z) \frac{\partial \varphi_0^T}{\partial \omega_z}, \tag{23}$$

and in case of (21) –

$$\frac{\partial \omega_z}{\partial t} = U(\omega_z) - \frac{1}{2} F(\omega_z) \int\limits_{\xi_*} F^T(\omega_z) \frac{\partial \varphi_0^T}{\partial \omega_z} d\xi. \tag{24}$$

It is obvious that the integration of the equations obtained completes the solution of the selection problem of the optimum structure in the sense of (18) or (21) by the subsequent definition of the maximal component of the vector of the state probabilities

$$P(t) = \int\limits_{-\infty}^{\infty} \omega_z(\xi, t)\, d\xi.$$

It should be noted that from the point of view of the computing expenses the solution of equations (23), (24) appears to be not much more complicated than the initial system of the integro-differential equations (6) (a basic one in the theory of the dynamic systems of random structure [1]) and incommensurably more simple than the solution of the point-to-point boundary-value problem for two systems of the integro-differential equations with partial derivatives, given in paragraph 2. Moreover, the solution of the equations given allows to obtain in the real time the exact solution of the problem stated, while the approach, considered in paragraph 2, provides only the formation of the current suboptimum solution. For the comparative efficiency estimation of the both methods we consider the example.

As the target we choose an observable two-structural nonlinear dynamic system described in paragraph 2, for which functions $U(\omega_Z)$, $F(\omega_Z)$ determining the right-hand part of equation (6) have the form:

$$U(\omega_z) = \left| \begin{array}{c} \dfrac{\partial}{\partial \xi}\left(\xi^2 \omega_z^{(1)}\right) + \dfrac{1}{2}\dfrac{\partial^2 \omega_z^{(1)}}{\partial \xi^2} - \dfrac{1}{2}\omega_z^{(1)}\left[\left(z - 0{,}5\xi^2\right)^2 - \sum\limits_{k=1}^{2}\int\limits_{-\infty}^{\infty}\left(z - 0{,}5x^2\right)^2 \omega_z^{(k)}(x,t)dx\right] \\[4mm] \dfrac{\partial}{\partial \xi}\left[\left(\xi - 0{,}01\xi^3\right)\omega_z^{(2)}\right] + \dfrac{1}{2}\dfrac{\partial^2 \omega_z^{(2)}}{\partial \xi^2} - \dfrac{1}{2}\omega_z^{(2)}\left[\left(z - 1{,}2\xi^2 + 0{,}1\xi^3\right)^2 - \right. \\[4mm] \left. - \sum\limits_{k=1}^{2}\int\limits_{-\infty}^{\infty}\left(z - 1{,}2x^2 + 0{,}1x^3\right)^2 \omega_z^{(k)}(x,t)dx\right] \end{array} \right|,$$

$$F(\omega_z) = \begin{vmatrix} \omega_z^{(1)} & -\omega_z^{(2)} \\ -\omega_z^{(1)} & \omega_z^{(2)} \end{vmatrix},$$

where z is the output signal of the nonlinear observer of process ξ.

As far as according to the constrains of the example given in paragraph 2, the selection of the process structure is required to be carried out, proceeding from the provision of the probability maximum of its occurrence in the given limits $\xi_* = \left[\xi_{min} = -0,5; \ \xi_{max} = 0,7\right]$, then for forming the non-negatively defined criterion function φ_0 in (18), ensuring in minimizing J the performance of the given criterion, we accomplish the following additional constructions.

The condition of providing the maximum of probability $P(t) = \int_{\xi_*} \omega_z(\xi,t)d\xi$ is equivalent to that of the minimum $\left(I_S^T - P\right)$ or $\left(I_S - P^T\right)\left(I_S^T - P\right)$, where I_S is the unit row of dimension S. Evaluating the last expression, we have

$$\left(I_S - P^T\right)\left(I_S^T - P\right) = I_S I_S^T - P^T I_S^T - I_S P + P^T P = S - 2I_S P + P^T P =$$
$$= S \int_{\xi_*} \frac{d\xi}{\xi_{max} - \xi_{min}} - 2I_S \int_{\xi_*} \omega_z(\xi,t)d\xi + \int_{\xi_*} \omega_z^T(\xi,t)d\xi \int_{\xi_*} \omega_z(\xi,t)d\xi . \qquad (25)$$

As far as by virtue of the Cauchy-Bunyakovsky inequality

$$\int_{\xi_*} \omega_z^T(\xi,t)d\xi \int_{\xi_*} \omega_z(\xi,t)d\xi \le \int_{\xi_*} \omega_z^T(\xi,t)\omega_z(\xi,t)d\xi (\xi_{max} - \xi_{min}),$$

then after replacing the last summand by the right-hand part of the given inequality in expression (25), the function (25) remains non-negatively defined. Presenting expression (25) transformed in the form

$$\int_{\xi_*} \left(\frac{S}{\xi_{max} - \xi_{min}} - 2I_S\omega_z(\xi,t) + (\xi_{max} - \xi_{min})\omega_z^T(\xi,t)\omega_z(\xi,t) \right)d\xi ,$$

we obtain the required criterion function φ_0 as follows:

$$\varphi_0(\omega_z) = SA^{-1} - 2I_S\omega_z + A\omega_z^T\omega_z ,$$

where $A = \xi_{max} - \xi_{min}$.

In this case equation (23) for the optimum vector ω_z takes the form

$$\frac{\partial \omega_z}{\partial t} = U(\omega_z) - F(\omega_z) F^T(\omega_z) \left(\left| \begin{matrix} \omega_z^{(1)} \\ \omega_z^{(2)} \end{matrix} \right| A - I_S^T \right) =$$

$$= U(\omega_z) - \left| \begin{matrix} \omega_z^{(1)3} + \omega_z^{(1)} \omega_z^{(2)} \left(\omega_z^{(2)} - \omega_z^{(1)} \right) - \omega_z^{(2)3} \\ -\omega_z^{(1)3} + \omega_z^{(1)} \omega_z^{(2)} \left(\omega_z^{(1)} - \omega_z^{(2)} \right) + \omega_z^{(2)3} \end{matrix} \right| \cdot 1,2.$$

The solution of the given equation was carried out similarly to paragraph 2 on the basis of approximation of functions $\omega_Z^{(1),(2)}$ by the Fourier series on interval [-5; 5] within the accuracy of 4 terms of expansion and integration of the obtained system of equations for the expansion factors on the time interval [0; 300] s. When integration and formation of the approximated values of functions $\omega_Z^{(1)}, \omega_Z^{(2)}$ had been taken place, the numbers of structures, chosen by the character of the maximal state probability at the current time, appeared to be distributed in time as follows:
- in interval [0; 53] s - the second structure;
- in interval [53; 119] s -the first structure;
- in interval [119; 300] s -the second structure.
The comparative analysis of the solution obtained with that of suggested in paragraph 2 shows that with their practically identical accuracy (upon completing the simulation the difference of the probability values for occurring process ξ in the limits ξ_* was less than 7 %) the computing expenses were reduced rather sufficiently in the case considered: the volume of the operative memory, required for the solution, has been decreased by a factor of ~ 3,5 , the solution time - by a factor of ~ 2,1.
Thus, despite of the less generality from the point of view of the theory, the local control in the real multistructural systems has obvious advantages over the terminal one (especially in the systems using the real-time measuring information). In summary we consider the investigated problem of the control synthesis by the local criterion on the basis of applying the approximating representations of the distribution density.

5. Suboptimum structure control on the basis of the local generalized criterions

In this case taking into account the character of the criterions given in paragraphs 3; 4 and preliminary reasoning the minimized local criterion J is written down in the following form:

$$J = \int_{\xi_*} \varphi_0 \left[\tilde{\rho}(\xi, \hat{p}, x, t) \right] d\xi + \int_{t_0}^{t} v_0^T(\hat{p}, x, t) v_0(\hat{p}, x, t) dt , \qquad (26)$$

where φ_0 is the non-negatively defined scalar function.

In this case the solution is reduced to the search of such vector $v_0 = v_0(\hat{p}, x, t)$, which would provide the minimum of criterion (26) given that the extended vector $\begin{vmatrix} \hat{p} \\ x \end{vmatrix}$ is described by the known vector equation derived in [1] and transformed in paragraph 3 to the following form (16):

$$\begin{vmatrix} \dot{\hat{p}} \\ \dot{x} \end{vmatrix} = A(x,t) + B(\hat{p},x)v_0 ,$$

$$A(x,t) = \begin{vmatrix} 0 \\ \varphi(x,t) \end{vmatrix}, \quad B(\hat{p},x) = \begin{vmatrix} (E \otimes \hat{p}^T) - \hat{p}(E \otimes I_S)E_1 \\ T(\hat{p},x) \end{vmatrix} E_0 .$$

The final solution of the problem stated - the selection of the number of the structure, providing the optimum of functional (26) at the real time, is carried out similarly to that of given in paragraph 3 by solving equation (16) in case of the found optimum law v_0^* - for construct vector $\hat{p}$ and define its maximal component, the number of which will define the number of the structure required for the state vector. Following the stated above in paragraphs 3, 4, in this case the optimum function v_0^* is synthesized from the constrain

$$\max_{v_0} \left[-\left(\frac{\partial}{\partial t} \int_{\xi_*} \varphi_0 \left[\tilde{\rho}(\xi, \hat{p}, x, t) \right] d\xi + v_0^T(\hat{p}, x, t) v_0(\hat{p}, x, t) \right) \right],$$

which, in its turn, results in the initial equation with respect to v_0^*:

$$\frac{\partial}{\partial v_0} \int_{\xi_*} \left(\frac{\partial}{\partial t} \varphi_0[\tilde{\rho}] \right) d\xi + 2v_0^{*T} = 0$$

Due to the fact that

$$\frac{\partial}{\partial t} \varphi_0[\tilde{\rho}] = \frac{\partial \varphi_0}{\partial \tilde{\rho}} \left\{ \begin{vmatrix} \frac{\partial \tilde{\rho}}{\partial \hat{p}} & : & \frac{\partial \tilde{\rho}}{\partial x} \end{vmatrix} \begin{vmatrix} \dot{\hat{p}} \\ \dot{x} \end{vmatrix} \right\}$$

and vector $\begin{vmatrix} \hat{p} \\ x \end{vmatrix}$ is described by the system of equations (16), then we obtain the following equation with respect to v_0^*:

$$\frac{\partial}{\partial v_0}\left\{\int_{\xi_*}\frac{\partial\varphi_0}{\partial\tilde{\rho}}\left|\frac{\partial\tilde{\rho}}{\partial\hat{p}}:\frac{\partial\tilde{\rho}}{\partial x}\right|\,d\xi\left(A+Bv_0^*\right)\right\}+2v_0^{*T}=0\,.$$

Finally from the expression given we have

$$\int_{\xi_*}\frac{\partial\varphi_0}{\partial\tilde{\rho}}\left|\frac{\partial\tilde{\rho}}{\partial\hat{p}}:\frac{\partial\tilde{\rho}}{\partial x}\right|\,d\xi\,B+2v_0^{*T}=0$$

whence the optimum law v_0^* is immediately defined:

$$v_0^*=-\frac{1}{2}B^T\int_{\xi_*}\left|\frac{\partial\tilde{\rho}}{\partial\hat{p}}:\frac{\partial\tilde{\rho}}{\partial x}\right|\frac{\partial\varphi_0}{\partial\tilde{\rho}}\,d\xi\,. \tag{27}$$

The substitution (27) into (16) results in the required equations describing the evolution of the parameters of the distribution densities for the state vectors in the structures and the probabilities of occurring of those for the system optimum in the sense of (26):

$$\left|\begin{matrix}\hat{p}\\\dot{x}\end{matrix}\right|=A\left(\hat{p},x,t\right)-\frac{1}{2}B\left(\hat{p},x\right)B^T\left(\hat{p},x\right)\int_{\xi_*}\left|\frac{\partial\tilde{\rho}}{\partial\hat{p}}:\frac{\partial\tilde{\rho}}{\partial x}\right|\frac{\partial\varphi_0}{\partial\tilde{\rho}}\,d\xi\,.$$

Then the structure selection is carried out similarly to that of stated above on the basis of defining the number of the maximal component for vector $\hat{p}$ being sufficiently trivial operation, which does not practically affect the total volume of the computing expenses. Despite some reduction in the generality of the solution considered in comparison, for example, with the similar terminal one, obtained in paragraph 3, its obvious advantage is the absence of the need for the solution of the point-to-point boundary-value problem and, as a consequence, the feasibility to apply the optimum solution simultaneously with the essential reduction of the computing expenses for the real systems as compared to the general case.

So, the comparative analysis of the solution obtained with that of suggested in paragraph 3, which has been carried out on the basis of the numerical simulation, has shown that with their practically identical accuracy (upon completing the simulation the difference of values for the probabilities of occurring process ξ within the limits of interval ξ* was less than 9 %) the computing expenses in the considered case were reduced rather essentially: the volume of the operative memory required for their solution, has decreased by a factor of ~ 3,7 , the solution time -by a factor of 2,4 .

The practical recommendations here are obvious and do not require the additional comments.

6. References

1. Kazakov I.E., Artemiev V.M. Optimization of dynamic systems with random structure.- Moscow, "Science", 1980.
2. Xu X., Antsaklis P.J. Optimal control of switched systems based on parameterization of the switching instants. // IEEE Trans. Automat. Contr., v.49, no.1, pp.2-16, 2004.
3. Sirazetdinov T.K. Optimization of systems with distributional parameters. - Moscow, "Science", 1977.
4. Pugachev V.S., Sinitsyn I.N. Stochastic differential systems. - Moscow, "Science", 2004.
5. Krasovsky A.A. and other. Theory of the automatic control (reference book). - Moscow, "Science", 1987.
6. Pervachev S.V., Perov A.I. Adaptive filtration of messages. - Moscow, "Radio and communication", 1991.
7. S.V. Sokolov, Yu.I. Kolyada, F.V. Mel'nichenko. Optimal control of nonlinear stochastic structures. // Automatic Control and Computer Sciences. Allerton Press Inc., New York, v.32, no.2, pp.18-25, 1998.

Optimal design criteria for isolation devices in vibration control

Giuseppe Carlo Marano[§] and Sara Sgobba[*]
[§]Technical University of Bari, DIASS, viale del Turismo 10,
74100 Taranto (Italy)
[*]International Telematic University UNINETTUNO
Corso Vittorio Emanuele II, 39
00186 Roma (ITALY)

Vibration control and mitigation is an open issue in many engineering applications. Passive strategies was widely studied and applied in many contests, such as automotive, aerospatial, seismic and similar. One open question is how to choose opportunely devices parameters to optimize performances in vibration control. In case of isolators, whose the main scope is decoupling structural elements from the vibrating support, optimal parameters must satisfy both vibration reduction and displacement limitation.

This paper is focused on the a multi-objective optimization criterion for linear viscous-elastic isolation devices, utilised for decreasing high vibration levels induced in mechanical and structural systems, by random loads. In engineering applications base isolator devices are adopted for reducing the acceleration level in the protected system and, consequently, the related damage and the failure probability in acceleration sensitive contents and equipment. However, since these devices act by absorbing a fraction of input energy, they can be subjected to excessive displacements, which can be unacceptable for real applications. Consequently, the mechanical characteristics of these devices must be selected by means of an optimum design criterion in order to attain a better performance control.

The proposed criterion for the optimum design of the mechanical characteristics of the vibration control device is the minimization of a bi-dimensional objective function, which collects two antithetic measures: the first is the index of device efficiency in reducing the vibration level, whereas the second is related to system failure, here associated, as in common applications, to the first exceeding of a suitable response over a given admissible level.

The multi-objective optimization will be carried out by means of a stochastic approach: in detail, the excitation acting at the support of the protected system will be assumed to be a stationary stochastic coloured process.

The design variables of optimization problem, collected in the design vector (DV), are the device frequency and the damping ratio. As cases of study, two different problems will be analysed: the base isolation of a rigid mass and the tuned mass damper positioned on a MDoF structural system, subject to a base acceleration.

The non dominated sorting genetic algorithm in its second version (*NSGA-II*) is be adopted in order to obtain the Pareto sets and the corresponding optimum DV values for different characterizations of system and input.

Keywords: *Random vibrations, multi-objective stochastic optimization, base isolator, tuned mass damper, genetic algorithm.*

Introduction

Dynamic actions are nowadays a wide engineering topic in many applicative and research areas, such as automotive, civil and aerospace. One main problem is how properly model dynamic actions, because of there are many real conditions where it is practically impossible to accurate predict future dynamic actions (i.e. earthquakes, wind pressure, sea waves and rotating machinery induced vibrations). In those cases external loads can be suitably modelled only by using random processes, and as direct consequence, also systems responses are random processes. In these environments, random dynamic analysis seems to be the most suitable method to get practical information concerning systems response and reliability (see for example [1]). It is obvious that also structural optimization methods seem to be practically approached by means of random vibrations theory. Concerning this problem, some recent works have been proposed, typically based on *Standard Optimization Problem* (*SOP*), which finds the optimum solution that coincides with the minimum or the maximum value of a scalar *Objective Function* (*OF*). The first problem definition of structural optimization was proposed by [2], in which constraints were defined by using probabilistic indices of the structural response and the *OF* was defined by the structural weight, leading to a standard nonlinear constrained problem.

In the field of seismic engineering, the use of a stochastic defined *OF* has been proposed for the optimum design of the damping value of a vibrations control device placed on the first story of a building [3], and was defined by the maximum displacement under a white noise excitation. A specific and more complete stochastic approach has also been proposed by [4], aimed to stiffness-damping simultaneous optimization of structural systems. In this work the sum of system response mean squares due to a stationary random excitation was minimized under constraints on total stiffness capacity and total damping capacity.

More recently, an interesting stochastic approach for optimum design of damping devices in seismic protection has been proposed by [5], aimed to minimize the total building life-cycle cost. It was based on a stochastic dynamic approach for failure probability evaluation, and the *OF* was defined in a deterministic way. The optimization problem was formulated by adopting as design variables the location and the amount of the viscous elastic dampers, adopting as constraints the failure probability associated to the crossing of the maximum inter-storey drift over a given allowable value. Reliability analysis was developed by means of the application of the first crossing theory in stationary conditions.

Another interesting work in the field of stochastic structural optimization regards the unconstrained optimization of single [6] and multiple [7] tuned mass dampers, by using as *OF* the structural displacement covariance of the protected system and modelling the input by means of a stationary white noise process.

However, the *SOP* does not usually hold correctly many real structural problems, where often different and conflicting objectives may exist. In these situations, the *SOP* is utilized by selecting a single objective and then incorporating the other objectives as constraints. The

main disadvantage of this approach is that it limits the choices available to the designer, making the optimization process a rather difficult task.

Instead of unique *SOP* solution, a set of alternative solutions can be usually achieved. They are known as the set of *Pareto optimum solutions*, and represent the *best solutions* in a wide sense, that means they are superior to other solutions in the *search space*, when all objectives are considered. If any other information about the choice or preference is given, no one of the corresponding *trade-offs* can be said to be better than the others. Many works in last decade have been done by different authors in the field of multi-objective structural optimization, for systems subject to static or dynamic loads [8].

This work deals with a multi-objective optimization of linear viscous-elastic devices, which are introduced in structural and mechanical systems in order to reduce vibrations level induced by random actions applied at the support. As application, two different problems are considered: first, the vibration base isolation of a rigid mass subject to support acceleration. In detail this is the problem of a vibration absorber for a rigid element isolated from a vibrating support, subject to a random acceleration process. This represents a typical application in many real problems, in mechanical, civil and aeronautics engineering. The main system is a rigid mass linked with the support by means of a linear viscous-elastic element (fig.1). In the multi-objective optimization, the *OF* is a vector which contains two elements: the first one is an index of device performance in reducing the vibration level, here expressed by the acceleration reduction factor. This is assumed to be, in stochastic meaning, the ratio between the mass and the support acceleration variances.

The second objective function is the displacement of the protected mass. In probabilistic meaning it is obtained in terms of the maximum displacement which will not be exceeded in a given time interval and with a given probability. This is achieved by adopting the threshold crossing probability theory. Design variables, which are assumed to be the isolator damping ratio ξ_S and its pulsation ω_s, are collected in the design vector (*DV*). The support acceleration is modelled as a filtered stationary stochastic process.

In order to obtain the Pareto set in the two dimensions space of *OFs*, and the optimum solution in the space of design variables, a specific genetic algorithm approach (the NSGA-II one) is adopted in the two cases of study. A sensitive analysis on the optimum solution is finally performed under different environmental conditions.

Multi-objective stochastic optimization of random vibrating systems

The proposed stochastic multi-objective optimization criterion is adopted in this study in order to define the optimum mechanical parameters in classical problems of vibration control. As before mentioned, two applications are considered which regard, in general, the limitation of vibration effects in mechanical and structural systems subject to base accelerations.

The optimization problem could be formulated as the search of design parameters, collected in the Design Vector (*DV*) $\mathbf{b}$, defined in the admissible domain $\mathbf{\Omega_b}$, able to minimize a given *OF*. This problem, in general, can be formulated in a standard deterministic way, or in a stochastic one, for example by means of response spectral moments. This approach, as before mentioned, has anyway some limits, because when designer looks for the optimum solution, he has to face with the selection of the most suitable criterion for measuring

performance. It is evident that many different quantities, which have a direct influence on the performance, can be considered as efficient criteria. At the same time, those quantities which must satisfy some imposed requirements, and cannot be assumed as criteria, are then used as constraints. It is common in optimization problems, therefore, to use a single *OF* subjected to some probabilistic constraints, as in the first stochastic optimization problem [2]. Usually, inequality constraints on system failure probability are utilised.

In the multi-objective formulation the conflict which may or may not exist between the different criteria is an essential point. Only those quantities which are competing should be considered as independent criteria. The others can be combined into a single criterion, which represents the whole group.

Case of study: protection of a rigid mass from a vibrating support

Let us consider first the case of the isolation of a rigid mass positioned on a vibrating support. In engineering applications the mass can represent a subsystem located on a vibrating mechanical support, as motor device, airplane structure, seismic isolated building and similar. In all these situations, the main goal is to limit the induced accelerations and to control the displacement of the rigid mass with respect to the support. The first objective is related to excessive inertial forces transmitted for example to electronic or mechanical devices, which can be sensitive to this effect (i.e. acceleration sensitive contents and equipment). The second objective is related to an excessive displacement of the protected mass, which can become unacceptable, for example, if the system is located quite closer to other elements, or if the vibration isolator has a limited acceptable lateral deformation over which it will collapse.

The protected element is modelled as a rigid body having a mass m. The isolator device is modelled as a simple viscous-elastic element, which connects the vibrating base with the supported mass (Fig. 1).
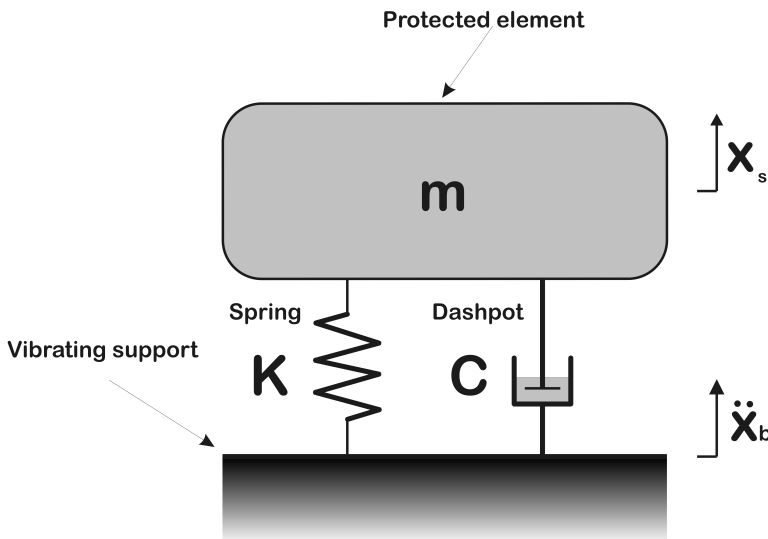
Fig. 1. Schematic Model of a rigid mass isolated from a vibrating support by means of an isolation device.

The stiffness k and the damping c of the isolator device must be optimized in order to minimize the vibration effects on the rigid mass m .

The base acceleration is a stochastic coloured process $\ddot{X}_b(t)$ modelled by means of a second order linear filter [9]:

$$\ddot{X}_b(t) = \ddot{X}_f(t) + w(t) = -\left(2\xi_f \omega_f \dot{X}_f + \omega_f^2 X_f \right) \tag{1}$$

where $w(t)$ is a stationary Gaussian zero mean white noise process, ω_f is the filter pulsation and ξ_f is the filter damping ratio. The motion equations of this combined system are:

$$\ddot{X}_s(t) + 2\xi_s \omega_s \dot{X}_s + \omega_s^2 X_s = -\ddot{X}_b \tag{2}$$

$$\ddot{X}_f(t) + 2\xi_f \omega_f \dot{X}_f + \omega_f^2 X_f = -w(t) \tag{3}$$

$$\ddot{X}_b(t) = \ddot{X}_f + w(t) \tag{4}$$

In the space equations (2)-(4) can be written as:

$$\dot{\mathbf{Z}} = \mathbf{AZ} + \mathbf{F} \tag{5}$$

where the space vector is:

$$\mathbf{Z} = \begin{pmatrix} X_s & X_f & \dot{X}_s & \dot{X}_f \end{pmatrix}^T \tag{6}$$

and the *system matrix* is:

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\omega_s^2 & \omega_f^2 & -2\xi_s \omega_s & 2\xi_f \omega_f \\ 0 & -\omega_f^2 & 0 & -2\xi_f \omega_f \end{pmatrix} \tag{7}$$

where:

$$\omega_s = \sqrt{\frac{k}{m}} \; ; \; \xi_s = \frac{c}{2\sqrt{km}} \tag{8}$$

Finally, the input vector is:

$$\mathbf{F} = -\begin{pmatrix} 0 & 0 & 0 & w(t) \end{pmatrix}^T \tag{9}$$

The space state covariance matrix $\mathbf{R_{ZZ}} = \left\langle \mathbf{ZZ}^T \right\rangle$ is obtained by solving the *Lyapunov* equation:

$$\mathbf{AR_{ZZ}} + \mathbf{R_{ZZ}A}^T + \mathbf{B} = 0 \tag{10}$$

The variance $\sigma^2_{\ddot{y}_s}$ of the absolute mass acceleration $\ddot{y}_s = \ddot{x}_s + \ddot{x}_b$ is:

$$\sigma^2_{\ddot{y}_s} = \mathbf{D}^T \mathbf{R_{ZZ}} \mathbf{D} \tag{11}$$

where:

$$\mathbf{D} = \begin{pmatrix} -\omega_s^2 & 0 & -2\xi_s\omega_s & 0 \end{pmatrix}^T \tag{12}$$

Formulation of multi-objective optimization
of device mechanical characteristics

The multi-objective stochastic optimization problem concerns the evaluation of *DV* $\mathbf{b} = (\omega_s, \xi_s)$ which is able to satisfy the reduction of the transmitted inertial acceleration in the rigid mass and to limit the displacement of this one with respect to the support. These two criteria conflict each others because, when the support rigidity grows at that time the acceleration reduction (i.e. the performance) and the lateral displacement decrease. This situation corresponds for example to the design of a well known vibration control device utilized in the field of seismic engineering: the base isolator. The decoupling between the vibrating support and the protected element, i.e. the effectiveness of vibration control strategy, increases monotonically with the reduction of device stiffness, but at the same time the device displacement grows up. Therefore, in the design of these devices the level of reduction of transmitted acceleration in the protected element, (i.e. the efficiency of control strategy) is related to the allowable maximum value of device displacement, and therefore these two conflicting criteria must be considered in the design.

The multi-objective optimization problem is finally posed:

$$\min\left\{OF_1, OF_2\right\} \tag{13}$$

In detail:

$$OF_1(\mathbf{b}) = \left(\frac{\sigma_{\ddot{y}_s}(\mathbf{b})}{\sigma_{\ddot{x}_b}} \right) \tag{14}$$

where the base vibrating acceleration variance is [11]:

$$\sigma_{\ddot{x}_b}^2 = \frac{\pi}{2} \frac{S_0 \omega_f}{\xi_f} \left(1 + 4\xi^2{}_f\right) \tag{15}$$

being S_0 the power spectral density function of the white noise process.

This OF is a direct protection efficiency index: it tends to a null value for a totally system-base decoupling, and tends to unit for a system rigidly connected with the vibrating base, and so subject to the same acceleration $\sigma_{\ddot{x}_b}$.

In order to make explicit the OF_2, the maximum displacement value X_s^{max} that will not be exceeded with a given probability $\widetilde{P}_f$ in an assigned time interval (assumed to be the duration of random vibration T) is adopted. Therefore:

$$OF_2(\mathbf{b}) = X_S^{max}(\mathbf{b}) : P_f\left(X_S^{max}, \mathbf{b}\right) = P\left\{|X_S| \geq X_S^{max}(\mathbf{b}) \middle| t \in [0,T]\right\} \leq \widetilde{P}_f \tag{16}$$

In case of rare failure events, the Poisson hypothesis could be reasonably utilised and so [1]:

$$P_f\left(X_S^{max}, \mathbf{b}\right) = 1 - e^{-\nu(X_S^{max}, \mathbf{b})T} \tag{17}$$

where the unconditioned mean crossing rate is:

$$\nu = \left(X_S^{max}, \mathbf{b}\right) = \frac{1}{\pi} \frac{\sigma_{\dot{X}_S}}{\sigma_{X_S}} e^{-\left\{\frac{1}{2}\left(\frac{X_S^{max}}{\sigma_{X_S}}\right)^2\right\}} \tag{18}$$

Finally one obtains:

$$OF_2(\mathbf{b}) = X_S^{max} = \sqrt{-2\sigma_{X_S}^2 \ln\left(-\frac{\pi}{T} \frac{\sigma_{X_S}}{\sigma_{\dot{X}_S}} \ln\left(1 - \widetilde{P}_f\right)\right)} \tag{19}$$

The two objective functions are plotted in Figure 2 in terms of ratio ω_s / ω_f and ξ_s.
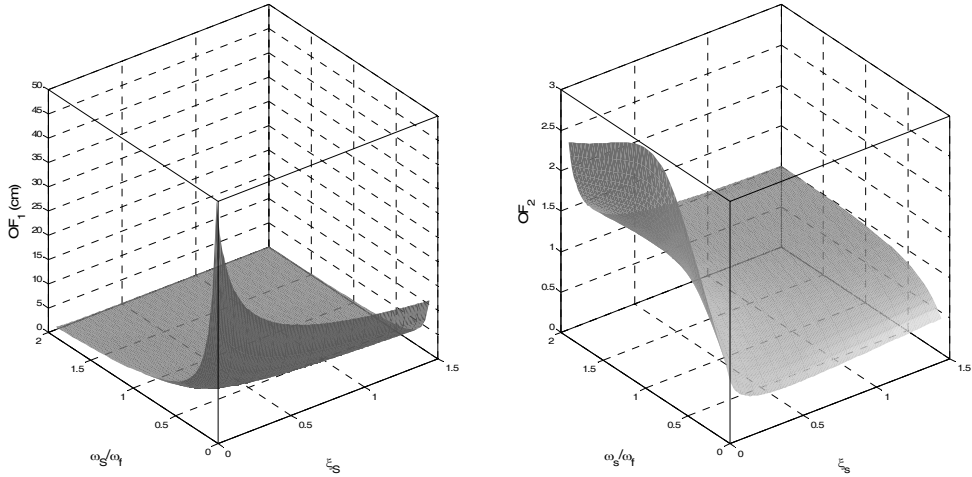
Fig. 2. Conflicting aspect of the two proposed objective functions.

An overview on methods for multi-objective optimization using gas

Many real engineering problems often involve several *OFs* each other in conflict and for them it is not possible to define an universally approved criteria of "optimum" as in single objective optimization. In this field, instead of aiming to find a single solution one can try to produce a set of good compromises. In a typical minimization-based *MOOP*, given two candidate solutions $\left\{ \mathbf{b}_j, \mathbf{b}_k \right\}$, if:

$$\forall i \in \left\{ 1,...,M \right\}, OF_i\left(\mathbf{b}_j\right) \le OF_i\left(\mathbf{b}_k\right) \wedge \exists i \in \left\{ 1,...,M \right\} : OF_i\left(\mathbf{b}_j\right) < OF_i\left(\mathbf{b}_k\right) \tag{20}$$

and defined the two objective vectors:

$$\mathbf{v}\left(\mathbf{b}_j\right) = \left\{ OF_1\left(\mathbf{b}_j\right),...,OF_M\left(\mathbf{b}_j\right) \right\} \tag{21}$$

$$\mathbf{v}\left(\mathbf{b}_k\right) = \left\{ OF_1\left(\mathbf{b}_k\right),...,OF_M\left(\mathbf{b}_k\right) \right\} \tag{22}$$

the vector $\mathbf{v}\left(\mathbf{b}_j\right)$ is said to dominate vector $\mathbf{v}\left(\mathbf{b}_k\right)$ (denoted by $\mathbf{v}\left(\mathbf{b}_j\right) \prec \mathbf{v}\left(\mathbf{b}_k\right)$).

Moreover, if no feasible solution, $\mathbf{v}\left(\mathbf{b}_k\right)$, exists that dominates solution $\mathbf{v}\left(\mathbf{b}_j\right)$, then $\mathbf{v}\left(\mathbf{b}_j\right)$ is classified as a *non-dominated* or *Pareto optimal solution*. The collection of all Pareto optimal solutions are known as the *Pareto optimal set* or *Pareto efficient set*, instead the

corresponding objective vectors are described as the *Pareto front* or *Trade-off surface*. Unfortunately, the Pareto optimum concept almost does not give a single solution, but a set of possible solutions, that cannot be used directly to find the final design solution by an analytic way. On the contrary, usually the decision about the "best solution" to be adopted is formulated by so-called (human) *decision maker* (*DM*), while rarely *DM* doesn't have any role and a generic *Pareto optimal solution* is considered acceptable (*no - preference based methods*). On the other hand, several *preference–based methods* exist in literature. A more general classification of the *preference–based method* is considered when the preference information is used to influence the search [12]. Thus, in *a priori methods*, *DM's* preferences are incorporated before the search begins: therefore, based on the *DM's* preferences, it is possible to avoid producing the whole *Pareto optimal set*. In *progressive methods*, the *DM's* preferences are incorporated during the search: this scheme offers the sure advantage to drive the search process but the *DM* may be unsure of his/her preferences at the beginning of the procedure and may be informed and influenced by information that becomes available during the search. A last class of methods is *a posteriori*: in this case, the optimiser carries out the *Pareto optimal set* and the *DM* chooses a solution ("searches first and decides later"). Many researchers view this last category as standard so that, in the greater part of the circumstances, a *MOOP* is considered resolved once that all *Pareto optimal solutions* are recognized. In the category of *a posteriori approaches*, different Evolutionary Algorithms (*EA*) are presented. In [13] an algorithm for finding constrained Pareto-optimal solutions based on the characteristics of a biological immune system (Constrained Multi-Objective Immune Algorithm, *CMOIA*) is proposed. Other diffused algorithms are the Multiple Objective Genetic Algorithm (*MOGA*) [14] and the Non dominated Sorting in Genetic Algorithm (*NSGA*) [15]. In this work the *NSGA-II* [16] will be adopted in order to obtain the Pareto sets and the correspondent optimum *DV* values for different systems and input configurations, for both the analysed problems (the vibration base isolation of a rigid mass and the TMD positioned on MDoF system subject to a base acceleration). Particularly, the *Real Coded GA* [17], *Binary Tournament Selection* [18], *Simulated Binary Crossover* (SBX) [19] and *polynomial mutation* [17] are used.

Multi-objective optimization of isolator mechanical characteristics

In this section the results of this first optimization problem are analysed. It is assumed that the admissible domain for **b** is the following:

$$\Omega_{\mathbf{b}} = \left\{ \xi_s, \omega_s : 0.01 \le \xi_s \le 2.5 \ \vee \ 1 \ \text{rad/sec} \le \omega_s \le 30 \ \text{rad/sec} \right\}. \tag{23}$$

System parameters are listed in table 1.

Filter damping ratio ξ_f	0.6	
Filter pulsation ω_f	20.94	(rad/sec)
Power spectral density S_0	1000	cm^2/sec^3
T	10^3	sec
Max probability of failure $\widetilde{P}_f$	10^{-2}	

Table 1. System parameters.

Concerning NDGA-II setup, after several try and error analyses, the parameters reported in table 2 have been adopted for the analysis. The selection derives from considerations about the equilibrium of computing cost and solution stability. The population size has been chosen as 500 in order to obtain a continuum Pareto front, and the maximum iteration number here used (100) has been determined after several numerical experiments (type try and error) which indicated that it is the minimum value to obtain stable solutions. This means that adopting a smaller iterations number, some differences in Pareto fronts (obtained for the same input data) take place.

Maximum generation	500
Population size	100
Crossover probability	0.9
Mutation probability	0.1

Table 2. NDGA-II setup.

Symbols	**OF$_2$ (cm)**	**OF$_1$ (cm)**	ω_S^{opt} **(rad/sec)**	ξ_S^{opt}
⭕	171.3159	0.2227	1	0.6256
✳	39.5099	0.3896	2.7629	0.7276
◻	110.5646	0.2624	1.3313	0.6910
◇	1.7741	0.9402	17.8002	2.1599

Table 3. Some numerical data from figure 3.
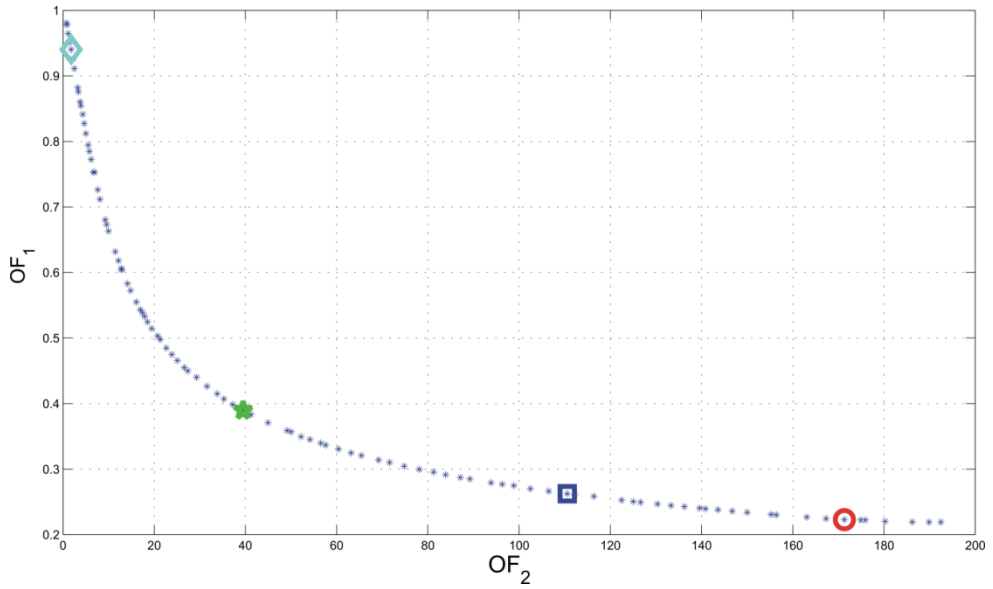
Fig. 3. Pareto front.
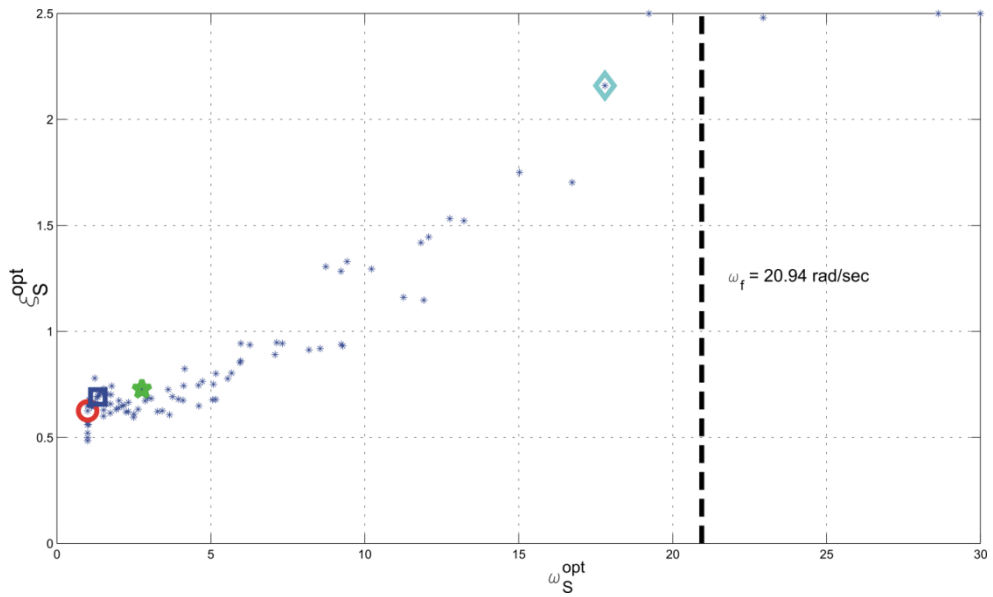
Fig. 4. Space of DV elements.

Figures 3 and 4 show the Pareto front and the space of DV elements, respectively, in this first case of multi-objective optimization problem. More precisely, in figure 4 on X-axis the

optimum frequency of the device ω_s^{opt} is plotted, whereas on the Y-axis the optimum damping ratio ξ_T^{opt} is shown. The vertical line corresponds to the filter frequency ω_f. In Table 3 some numerical data derived from these figures are also reported.

From figure 3 first of all it is possible to notice that a larger level of protection is related to an increase of allowable displacement. Anyway an asymptotic limit value of performance exists, that means that the reduction of transmitted acceleration is in the analysed example at least about 0.2. Moreover, some interesting observations can be carried out by observing the slope of Pareto front, which is not a convex curve. It is possible to distinguish three different portions of the Pareto front, which correspond to different criteria in using the vibration control strategy. In fact, on the left section of the Pareto front, which is related to a low efficiency, by means of a little grow of maximum allowable displacement one can obtain a large increase of performance (the slope is high). Then, in the second portion of Pareto set, the slope of the front reduces and, finally, in the right part an increase of performance is obtained only by means of a large increase of maximum admissible displacement. In this last situation, only little variations of optimum design variables take place (fig. 4). On the contrary, the reduction of maximum displacement is reached by increasing both frequency and damping. The variation is fast as the displacement reduces. Moreover, if the imposed displacement is very low, the control strategy acts by increasing the system frequency and by increasing quickly also the damping, which is associated to energy dissipation.
Figures 5, 7 and 9 show different Pareto fronts obtained for different values of power spectral density, filter damping ratio and filter pulsation. Figures 6, 8 and 10 show the corresponding optimum design variables. All the other parameters adopted are the same of figure 3.
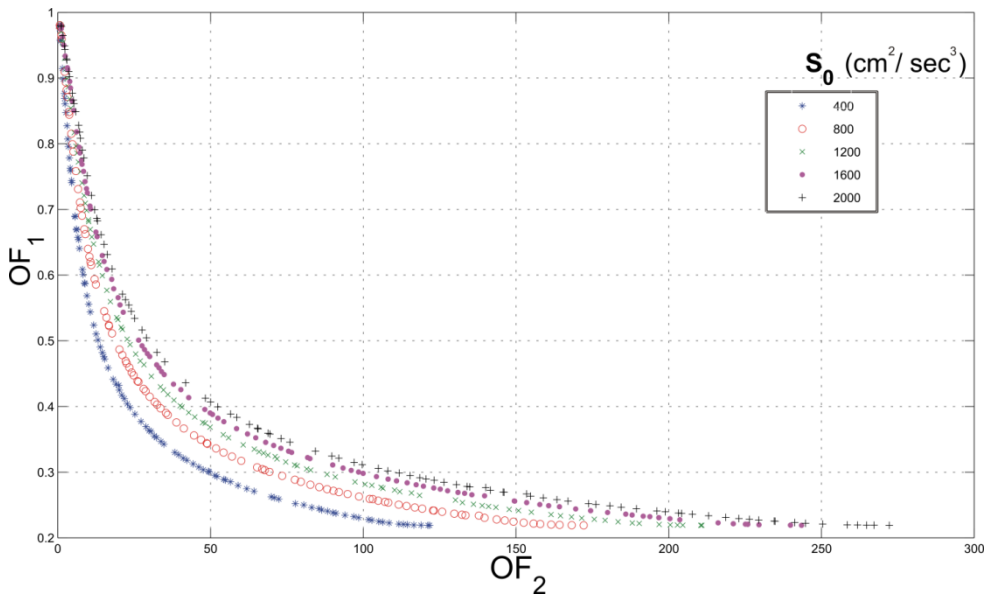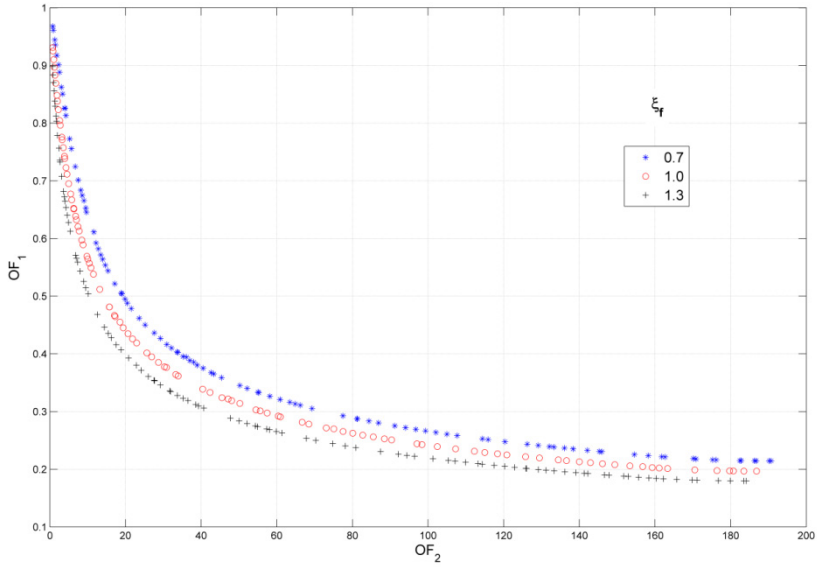

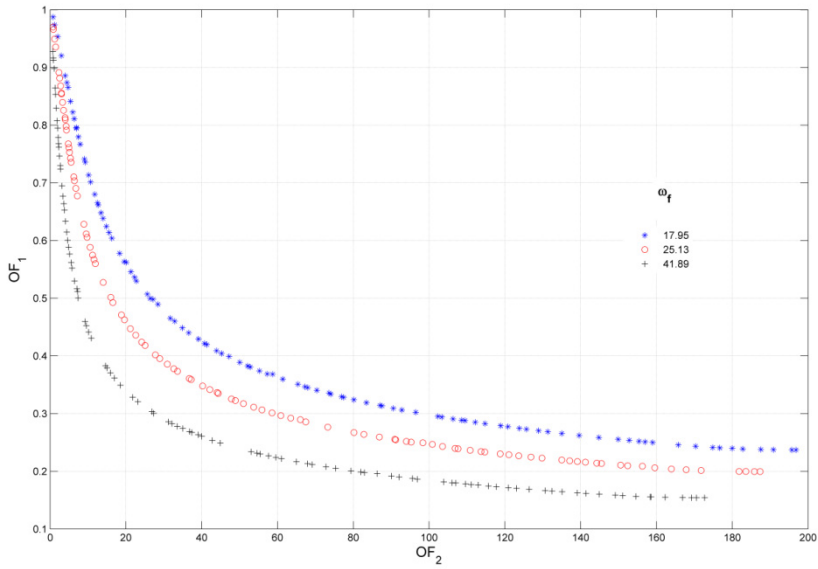Fig. 5. Sensitivity of Pareto front for different values of power spectral density.
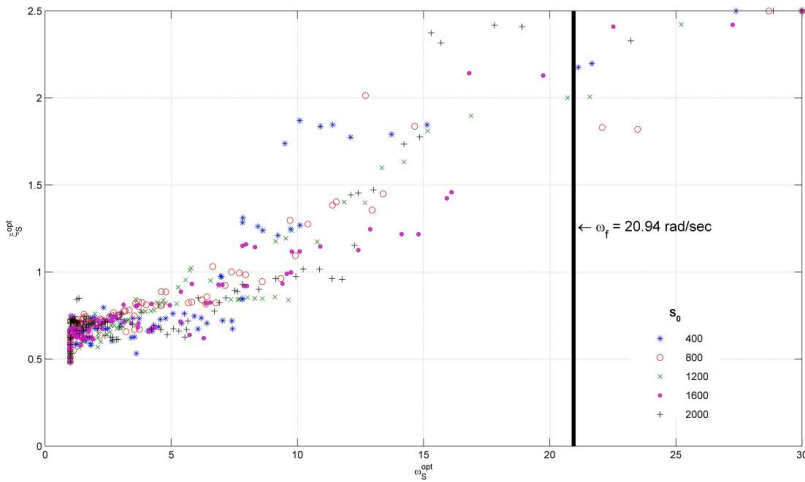
Fig. 6. Space of DV elements of multi-objective problem for different values of power spectral density.

With reference to figure 5 it is possible to notice that a variation of power spectral density induces variation of optimum Pareto front, due to non-linearity of OF_2. It is evident that higher performances are associated with low values of S_0, but the maximum level of vibration reduction (expressed by the asymptotic value of OF_1) is about the same in all cases, also if this situation corresponds to larger displacements for higher values of S_0. This outcome is quite clear, because the requirement on the maximum displacement is associated to S_0 by means of a non-linear formulation; meanwhile the vibration reduction is a linear function of this parameter.

However, the strategy adopted for the optimal solution in terms of design variables are about the same for all values of S_0, as shown in figure 6, where the same variability of the Pareto set for all values of S_0 can be observed.

Fig. 7. Sensitivity of Pareto front for different values of filter damping ratio.
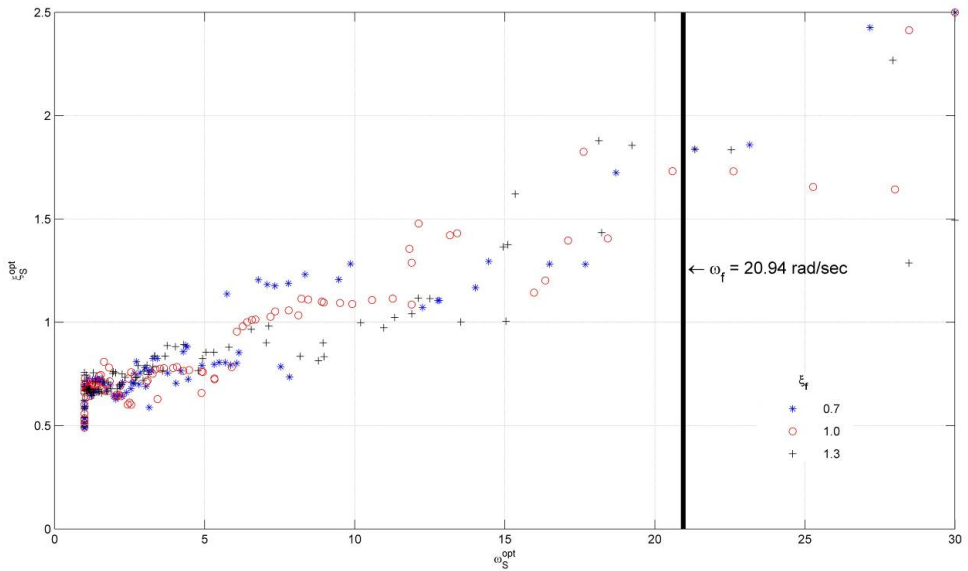
Fig. 8. Space of DV elements of multi-objective problem for different values of filter damping ratio.

Fig. 9. Pareto front for different values of filter pulsation.
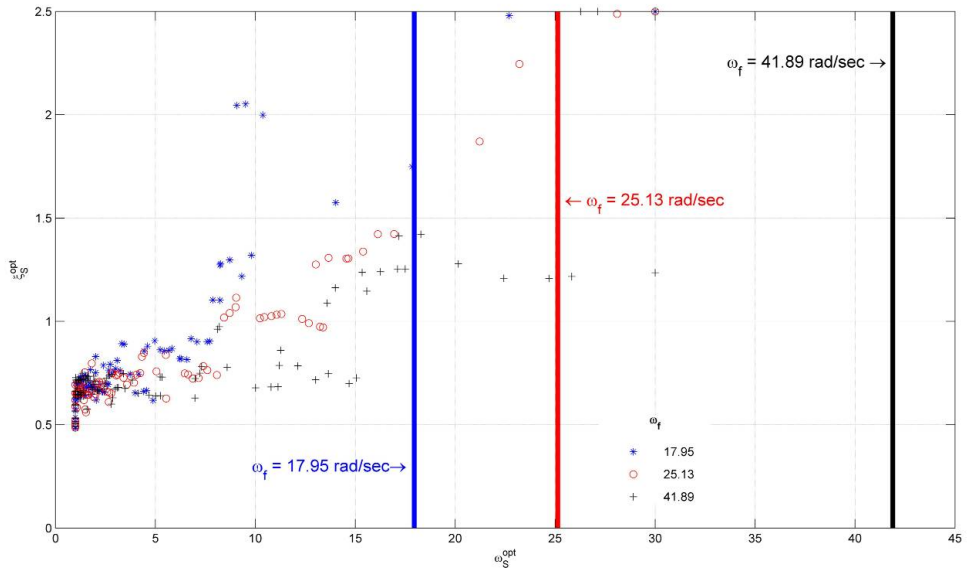
Fig. 10. Space of DV elements of multi-objective problem for different values of filter pulsation.

Moreover, one can deduce that the variability of both input parameters modify the Pareto set, but the excitation frequency ω_f influences the optimum solution more than ξ_f. Actually, from figure 9 it is possible to notice that the maximum performance of *TMD* changes as ω_f varies. Moreover, the initial slopes (for very small admissible displacement) are quite different. In detail, the variation of OF_1 is greater for higher values of ω_f and tends to decrease as this parameter grows up. Also the optimization strategy in terms of optimum design variables changes (fig. 10). On the left portion of DV space only little variations of optimum DV take place, whereas they correspond to the points located at the bottom on the right of Pareto front in figure 9. These values correspond to the asymptotic value of OF_2, where the minimum is attained for each displacement. So that, they tend to be located in a small region of the DV space, quite closer to this unconditional optimum solution point.

Conclusions

In the present study a multi-objective optimization design criterion for linear viscous elastic vibration control devices has been proposed. More in detail, the problem of an isolator device for the vibration control of a single rigid mass have been analysed.

The analysis has been carried out by adopting a stochastic approach, by assuming that the excitations acting on the base of the protected systems are stationary stochastic coloured processes.

In the multi-objective optimization problems two antithetic objectives are considered: the maximization of control strategy performance, expressed in stochastic terms by means of the reduction of transmitted acceleration in the protected systems, and the limitation in stochastic terms of the displacement of the vibrations control device. The design variables are the mechanical characteristics - frequency and damping ratio- of the device.

In order to perform the stochastic multi-objective optimization, the non dominated sorting genetic algorithm in its second version (*NSGA-II*) has been adopted, which supplies the *Pareto set* and the corresponding optimum design variables for different system and input configurations.

The sensitivity analysis carried out has showed that the optimum solution (i.e. the maximization of control strategy, expressed in terms of reduction of the response of the main system, and the limitation of the device displacement) is reached, in the two analysed problems, by adopting different strategies, in function of input and system characterization. These strategies act by varying the optimum frequency and damping ratio of the device differently, in function of the allowable performance.

The novelty of the proposed method is in using a multi-dimensional criterion for the design. Nowadays, this is a very important issue in modern Technical Codes [20], in which several performance requirements, which often can conflict each others, are fixed. In these situations, the designer must select the design variables which make available all objectives and the use of a multi-dimension criterion is very useful in this context.

The validation of the proposed method is demonstrated by developing two applications, in which several parameters involved have been changed. Therefore, results attained by the proposed method can be utilised in order to support the designers in the definition of possible structural solutions in vibration control strategy by using linear viscous-elastic devices.

References

[1] Lutes L. D., Sarkani S., "Random vibrations", Butterworth-Heinemann, Oxford (UK), 2001.

[2] Nigam, N.C., "Structural Optimization in Random Vibration Environment", AIAA, pp.551-553, 1972.

[3] Constantinou M. C., Tadjbakhsh I.G., "Optimum design of a first story damping system", Computer and Structures, Vol.17, pp. 305- 310, 1983.

[4] Takewaki I., "An approach to stiffness-damping simultaneous optimization", Computer Methods in Applied Mechanics and Engineering, Vol. 189(2), pp. 641-650, 2000.

[5] Park K. S., Koh H. M., Hahm D., "Integrated optimum design of viscoelastically damped structural systems", Engineering Structures, Vol.26, pp. 581-591, 2004.

[6] Rundinger F., "Optimum vibration absorber with nonlinear viscous power law damping and white noise excitation", ASCE, Journl of Engineering Mechanics, Vol. 132 (1), pp. 46-53, 2006.

[7] Hoang N., Warnitchai P., "Design of multiple tuned mass dampers by using a numerical optimizer", Earthquake Engineering and Structural Dynamic, Vol. 34(2), pp. 125-144, 2005.

[8] Papadrakakis M., Lagaros N. D., Plevris V., "Multiobjective Optimization of skeletal structures under static and seismic loading conditions, Engineering Optimization, Vol. 34, pp. 645-669, 2002.

[9] Tajimi H., "A statistical method of determining the maximum response of a building during earthquake", Proceedings of 2nd World Conf. on Earthquake Engineering, Tokyo, Japan, 1960.

[10] Lin C.C., Wang J.F., Ueng J.M., "Vibration Control identification of seismically excited m.d.o.f structure-PTMD systems", Journal of Sound and Vibration, Vol.240(1), pp. 87-115, 2001.

[11] Crandal S. H., Mark W. D., "Random vibration in mechanical systems", Accademic Press. NY and London , 1963.

[12] C. A. Coello Coello, "Handling Preferences in Evolutionary Multiobjective Optimization: A Survey", IEEE Neural Networks Council (ed.), Proceedings of the 2000 Congress on Evolutionary Computation (CEC 2000) Vol. 1, IEEE Service Center, Piscataway, New Jersey, pp. 30 -37, 2000.

[13] G. C. Luh, C. H. Chuen, "Multi-Objective optimal design of truss structure with immune algorithm", Computers and Structures, Vol. 82, pp. 829-844, 2004.

[14] C. M. Fonseca, P. J. Fleming, "Genetic Algorithms for Multi-Objective Optimization: Formualtion, Discussion and Generalization", Genetic Algorithms: Proceedings of the 5th International Conference (S. Forrest, ed.) San Mateo, CA: Morgan Kaufmann, 1993.

[15] N. Srinivas, K. Deb, "Multi-objective Optimization Using Nondominated Sorting in Genetic Algorithms", Journal of Evolutionary Computation, Vol. 2 (39), pp. 221-248, 1994.

[16] K. Deb, S. Agrawal, A. Pratap, T. Meyarivan, "A Fast Elitism Multi-Objective Genetic Algorithm: NSGA-II", Proceedings of Parallel Problem Solving from Nature, Sprinter pp. 849-858, 2000.

[17] M. M. Raghuwanshi, O. G. Kakde, "Survey on multiobjective evolutionary and real coded genetic algorithms", Proceedings of the 8th Asia Pacific Symposium on Intelligent and Evolutionary Systems, pp. 150-161, 2004.

[18] T. Blickle, L. Thiele, "A Mathematical Analysis of Tournament Selection", in L. Eschelman, ed., Genetic Algorithms: Proceedings of the 6th International Conference (ICGA95), Morgan Kaufmann, San Francisco, CA, 1995.

[19] K. Deb, R. B. Agrawal, "Simulated binary crossover for continuous search space", Complex System, Vol. 9, pp.115-148, 1995.

[20] SEAOC, "Vision 2000: Performance-Based Seismic Engineering of Buildings, Structural Engineers Association of California, Sacramento, California, 1995.

Sensitivity analysis and stochastic modelling of the effective properties for reinforced elastomers

Marcin Kamiński*,** and Bernd Lauke**

*Technical University of Łódź, Al. Politechniki 6, 90-924 Łódź
Poland
**Leibniz Institute of Polymer Research Dresden, Hohe Strasse 6, 01069 Dresden
Germany

1. Introduction

Determination of the sensitivity gradients as well as probabilistic moments of composite materials and even micro-heterogeneous structures was a subject of many both theoretical and computational analyses reported in (Christensen, 1977; Fu at al., 2009; Kamiński, 2005; Kamiński, 2009). Usually it was assumed that there exists some Representative Volume Element, small in comparison to the entire structure and on the basis of some boundary problem solution on this RVE (like uniform extension for example) the elastic or even inelastic effective tensors were determined. Therefore, using some well established mathematical and numerical methods, sensitivity (via analytical, gradient or Monte-Carlo) or probabilistic (using simulations, spectral analyses or the perturbations) were possible having quite universal character in the sense that the effective tensors formulas are independent of the constituents design. Let us remind also that this cell problem was solved most frequently using rather expensive Finite Element Method based computations (even to determine the hysteretic multi-physics behavior) and did not allow full accounting for the reinforcing particles interactions or the other chemical processes between the components modeling. The challenges in the nanomechanics as one may recognize also below are slightly different – although one needs to predict the effective behavior of the solid reinforced with the nanoparticles, the formulas for effective properties may be addressed through experimental results calibration to the specific components. Such an experimental basis makes it possible to give analytical formulas even for the strain-dependent material models, which was rather impossible in the micromechanics before. Furthermore, it is possible now to account for the particle agglomeration phenomenon, where the dimensionless parameter describing this agglomeration size is directly included into the effective parameter model (Bhowmick, 2008; Heinrich et al., 2002a). Taking this development into consideration, there is a need to answer the question – how the particular models for those effective parameters (like shear modulus here) are sensitive to the design parameters included into the particular model. Moreover, taking into account manufacturing and experimental

statistics it is necessary to determine how this uncertainty (or even stochasticity) propagates and influences probabilistic characteristics of the effective parameters.

Therefore, the main now is to collect various models for the effective shear modulus describing the solids with nanoparticles, group them into some classes considering the similarities in the mathematical form of the physical assumptions. Next, sensitivity gradients of input parameters are determined and the probabilistic characteristics are considered by a randomization of those parameters and, finally, we study some of those theories in the presence of stochastic ageing under non-stationary stochastic processes. Mathematical basis for those studies is given by the stochastic generalized perturbation theory, where all random parameters and functions are expanded via Taylor series with random coefficients. A comparison of the same order quantities and classical integration known from the probability theory allows for a determination of the desired moments and coefficients with a priori assumed accuracy. Now up to fourth order central probabilistic moments as well as the coefficients of variation, asymmetry and concentration are computed – computational part is completed thanks to the usage of symbolic algebra system MAPLE. The main advantage of the perturbation method applied behind the Monte-Carlo simulation is that the preservation of a comparable accuracy is accompanied now by significantly smaller computational time and, further, parametric representation of the resulting moments. Let us mention that the sensitivity analysis is the inherent part of the perturbation approach – since first order partial derivatives are anyway necessary in the equations for the probabilistic moments (up to 10th order derivatives are computed now). Finally, the stochastic ageing phenomenon was modeled, where the output probabilistic moments time fluctuations were obtained. The results obtained and the methods applied in the paper may be further used in optimization of the effective parameters for solids with nanoparticles as well as reliability (and/or durability) analysis for such materials or structures made of them.

2. Comparison of various available theories

As it is known from the homogenization method history, one of the dimensionless techniques leading to the description of the effective parameters is the following relation describing the shear modulus:

$$G^{(eff)} = f\, G_0 \,,\tag{1}$$

where G_0 stands for the virgin, unreinforced material and f means the coefficient of this parameter increase, related to the reinforcement portion applied into it. As it is known, the particular characterization of this coefficient strongly depends on the type of the reinforcement - long or short fibers or reinforcing particles, arrangement of this reinforcement – regular or chaotic, scale of the reinforcement related to the composite specimen (micro or nano, for instance) or, of course, the volumetric ratios of both constituents. It is not necessary to underline that the effective nonlinear behavior of many traditional and nano-composites, not available in the form of simple approximants, needs much more sophisticated techniques based usually on the computer analysis with the use of the Finite Element Method. Let us note also that the elastomers are some specific composite materials, where usually more than two components are analyzed – some interface layers are inserted also (Fukahori, 2004) between them (the so-called SH and GH layers), which

practically makes this specimen 4-component. Therefore, traditional engineering and material-independent theories for the effective properties seem to be no longer valid in this area (Christensen, 1977).

The development of effective shear modulus for the elastomers resulted in various models, which can be generally divided into (a) linear theories based on the volume fractions of the inclusions, (b) linear elastic fractal models as well as (c) stress-softening fractal models. The first group usually obeys the following, the most known approximations, where the coefficient f is modeled using

- Einstein-Smallwood equation

$$f = 1 + 2.5\varphi , \tag{2}$$

- Guth-Gold relation

$$f = 1 + 2.5\varphi + 14.1\varphi^2 , \tag{3}$$

- Pade approximation

$$f = 1 + 2.5\varphi + 5.0\varphi^2 + ... \cong 1 + \frac{2.5\varphi}{1 - 2\varphi} , \tag{4}$$

where φ is the inclusions volume fraction; eqn (2) is introduced under the assumptions on perfect rigidity of the reinforcing particles and the elastomeric matrix incompressibility. The effectiveness of those approximations is presented in Fig. 1, where the parameter φ belongs to the interval [0.0,0.4] resulting in the range of the coefficient f varying from 1 (effective parameter means simply virgin material modulus) up to about 6 at the end of φ variability interval. As it can be expected, the Einstein-Smallwood approximation gives always the lower bound, whereas the upper bound is given by Guth-Gold approach for $0 \leq \varphi \leq 0.325$ and by Padè approximation for $\varphi \geq 0.325$. Taking into account the denominator of eqn (4) one must notice that the singularity is observed for $\varphi=0.5$ and higher volume ratios returns the negative results, which are completely wrong, so that this value is the upper bound of this model availability. The other observation of rather general character is that now the increase of shear modulus is measured not in the range of single percents with respect to the matrix shear modulus value (like the composites with micro-inclusions) but is counted in hundreds of percents, which coincides, for example with the results obtained for the effective viscosities of the fluids with solid nano-particles.

The second class of the homogenized characteristics is proposed for elastomers taking into account the fractal character of the reinforcing particles chains and can be proposed as

$$f \cong 1 + \begin{cases} \Xi^{-\frac{1}{4}} \varphi, & \varphi < \varphi_c \\ \Xi^{-\frac{5}{4}} \varphi^4, & \varphi > \varphi_c \end{cases} \tag{5}$$

where

$$\Xi = \frac{\xi}{b} \tag{6}$$

is also dimensionless parameter relating cluster size ξ to the size of the primary particle of carbon black constituting the reinforcing aggregate diameter (b). The condition that $\varphi < \varphi_c$ means that no aggregate overlap (smaller concentrations coefficients); otherwise the second approximation in eqn (5) is valid. The introduction of parameter Ξ enables to analyze the whole spectra of elastomers without precise definition of their aggregates dimensions in nm. The response surface of the coefficient f with respect to two input quantities $\varphi \in [0.0, 0.4]$ and $\Xi \in [1.0, 10.0]$ is given below – the upper surface is for the non-overlapping situation, while the lower one – for the aggregates overlap. Both criteria return the same, intuitionally clear, result that the larger values of both parameters the larger final coefficient f, however now, under the fractal concept, its value is essentially reduced and is once more counted in percents to the original unreinforced matrix value. Observing the boundary curves for φ_{max} it is apparent that this increase for overlapping and not overlapped cases has quite different character.
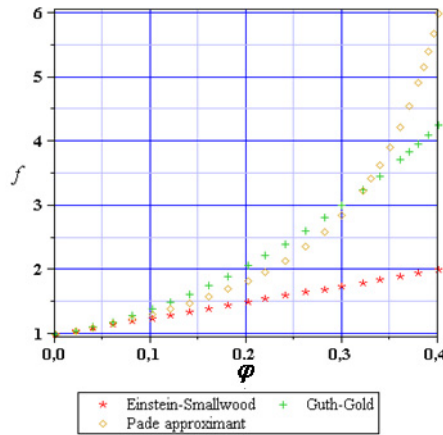
Fig. 1. A comparison of various volumetric approximations for the coefficient f
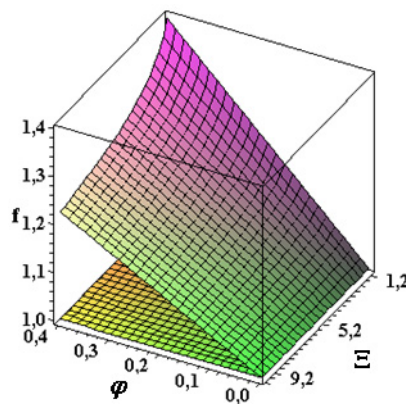
Fig. 2. Coefficient f for the rubbers with the carbon black aggregates by DLA clusters

The model presented above is, as one can compare, the special case of more general approach, where it is assumed

$$f \cong 1 + \begin{cases} \Xi^{2\frac{d_f}{D}-2d_f} \varphi, \ \varphi < \varphi_c \\ \Xi^{2\frac{d_f}{D}-d_f} \varphi^{\frac{2}{3-d_f}}, \ \varphi > \varphi_c \end{cases} \tag{7}$$

d_f means the mass fractal dimension and D is the spectral dimension as a measure of the aggregate connectivity (it is enough to put d_f=2.5 and D=4/3 to obtain eqn (5)). This equation has no parameter visualization according to the larger number of the independent variables.

Finally, the homogenization rules under stress-softening were considered- with Mullins effect (Dorfmann et al., 2004), where for carbon black and silica reinforcements the overlapped configuration $\varphi > \varphi_c$ was noticed. Additionally, the cluster size ξ was considered as the deformation-dependent quantity $\xi = \xi(E)$ with E being some scalar deformation variable related explicitly to the first strain tensor invariant, however, some theories of deformation independent cluster sizes are also available. Those theories are closer to the realistic situations because the function $\xi = \xi(E)$ is recovered empirically and it results in the following formulas describing the coefficient f varying also together with the strain level changes:

- the exponential cluster breakdown

$$f(E) = X_\infty + (X_0 - X_\infty) \exp(-\alpha E) \tag{8}$$

and

- the power-law cluster breakdown

$$f(E) = X_\infty + (X_0 - X_\infty)(1 + E)^{-y}. \tag{9}$$

The following notation is employed here

$$X_\infty = 1 + C(\frac{\xi_0}{b})^{d_w - d_f} \varphi^{\frac{2}{3-d_f}}, \qquad X_0 = 1 + C\varphi^{\frac{2}{3-d_f}}, \tag{10,11}$$

where $C \in \Re$ and d_w is the fractal dimension representing the displacement of the particle from its original position. Because ξ_0 stands for the initial value of the parameter ξ one can rewrite eqn (10) as

$$X_0 = 1 + C\Xi^{d_w - d_f} \varphi^{\frac{2}{3-d_f}}. \tag{12}$$

Below one can find numerical illustration of those parameters variability for some experimentally driven combinations of the input parameters for the specific elastomers. As one may expect, larger volumetric fractions of the reinforcement lead to larger values of the coefficient f; the smaller the values of the strain measure E the more apparent differences

between the values f are computed for various combinations of materials and their volumetric ratios. Comparison of Figs. 3 and 4 shows that independently from the model (exponential or power-law) the smallest values of the parameter f are computed for 40% silica reinforcement, and then in turn – for 40% carbon black, 60% silica and 60% carbon black. So that it can be concluded that, in the context of the coefficient f, the carbon black reinforcement results in larger reinforcement of the elastomer since $G^{(eff)}$ is higher than for the reinforcement by silica for the same volumetric amount of those particles. Comparing the results for all models presented in Figs. 1-4 one can generally notice that the power-law cluster breakdown theory returns the largest values of the studied coefficient f for small values of the stretch of the elastomer analyzed.
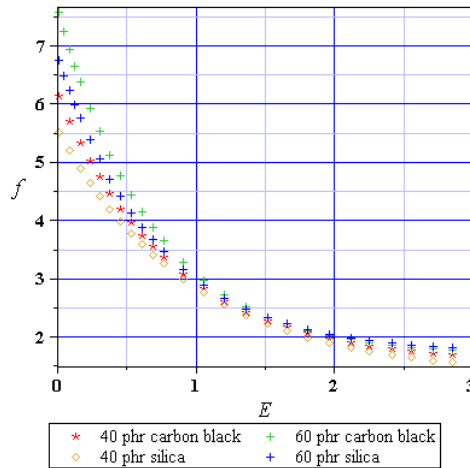
Fig. 3. The curve f=f(E) for the exponential cluster breakdown
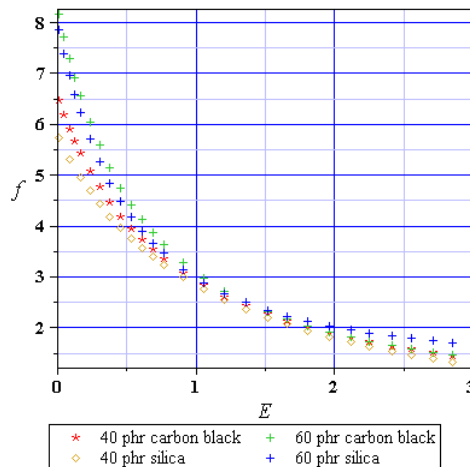
Fig. 4. The curve f=f(E) for the power-law cluster breakdown

3. Design sensitivity analysis

As it is known from the sensitivity and optimization theory, one of the milestones in the optimal design of the elastomers would be numerical (or analytical when available) determination of the sensitivity coefficients for the effective modulus as far as the homogenization theory is employed in the design procedure. Then, by simple partial differentiation of initial eqn (1) with respect to some elastomers design parameter h one obtains

$$\frac{\partial G^{(eff)}}{\partial h} = \frac{\partial f}{\partial h} G_0 + f \frac{\partial G_0}{\partial h} \,.$$ (13)

Considering the engineering aspects of this equation, the second component of the R.H.S. may be neglected because the design parameters are connected in no way with the unreinforced material, so that the only issue is how to determine the partial derivatives of the coefficient f with respect to some design variables like the volumetric ratio of the reinforcement, the cluster size, the exponents and powers as well as the strain rate in stress-dependent models. Further usage of those sensitivities consists in determination of the response functional, like strain energy of the hyperelastic effective medium for the representative stress state on the elastomer specimen, a differentiation of this functional w.r.t. design parameter and, finally, determination of the additional optimal solution.

First, we investigate the sensitivity coefficients as the first partial derivatives of the coefficient f with respect to the reinforcement volumetric ratio, accordingly to the analysis performed at the beginning of Sec. 2. As it could be expected (see Fig. 5), the Einstein-Smallwood returns always positive constant value, which is interpreted obviously that the higher coefficient φ, the larger value of the parameter f. The remaining gradients are also always positive, whereas the upper bounds gives the Guth-Gold model in the interval $\varphi \in [0.0, 0.25]$, for larger volumetric ratios of the reinforcement the Padè approximation exhibit almost uncontrolled growth.
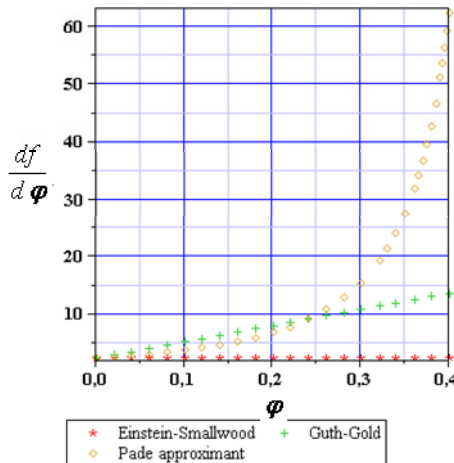

Fig. 5. Sensitivity coefficients for the volumetric coefficient f

The next results computed deal with the sensitivity coefficients of the coefficient f for the theory including the fractal character of the reinforcement for the non-overlapped and overlapped configurations of the elastomer, however now there are two design variables – the volumetric ratio φ as well as the parameter Ξ; the results are given in Figs. 6-7 accordingly, where larger absolute values are obtained in both cases for the elastomer with no overlapping effect.
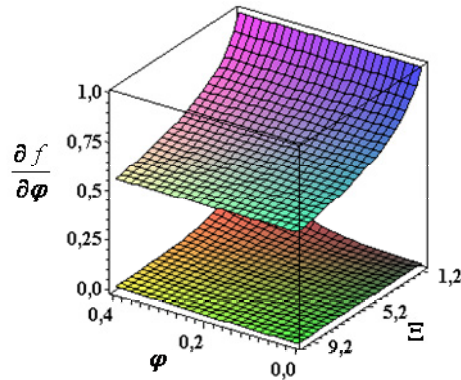

Fig. 6. Sensitivity for rubbers with carbon black aggregates by DLA clusters to volumetric ratio of the reinforcement
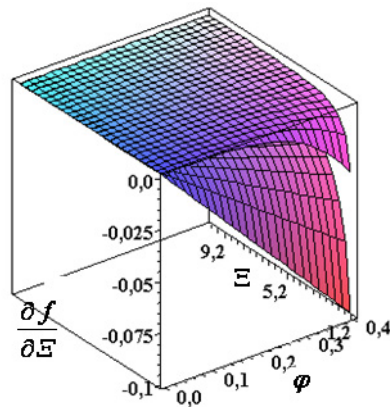
Fig. 7. Sensitivity for rubbers with carbon black aggregates by DLA clusters to Ξ

As it is quite clear from those surfaces variability, when the model with overlapping is considered, the resulting sensitivity gradients are dependent in a comparable way on both parameters φ and Ξ. However, the model with the overlap effect exhibits significant changes to the parameter Ξ, while almost no – with respect to the variable φ. Further, one may find easily that the lower value of Ξ (dimensionless cluster size), the higher are the gradients with respect to φ. Physical interpretation of this result is that the elastomers with the reinforcing particles more independent from each other are more sensitive to this

reinforcement volumetric ratios than the elastomers with larger clusters. Both models return here positive values, so that increasing of those parameters return an increase of the studied gradient value. Fig. 7 contains analogous results for the gradients computed with respect to the cluster size Ξ and now, contrary to the previous results, all combinations of input parameters return negative gradients. This gradient is almost linearly dependent on the parameter φ for the case without overlapping and highly nonlinear w.r.t. Ξ, whereas the overlap effect results in similar dependence of these gradients on both parameters. Quite analogously to the previous figure, the smaller value of the parameter Ξ, the larger output gradient value and opposite interrelation of this gradient to parameter φ.

Finally, we study the sensitivity coefficients for the exponential and power-law cluster breakdown with respect to the scalar deformation variable E (the results are presented in Figs. 8 and 9). Contrary to the previous sensitivity gradients, all the results are negative as one could predict from Figs. 3 and 4. Significantly larger absolute values are obtained for the exponential cluster breakdown here but, independently from the model, the highest sensitivity is noticed for $E \cong 0$ and then it systematically increases (its absolute values) to almost 0 for $E \cong 3$ and the differences between the elastomers with various reinforcement ratios monotonously vanish. The interrelations between different elastomers sensitivities depends however on the model and the power-law the largest absolute values are obtained for 60% carbon black; then in turn we have 60% silica, 40% carbon black and 40% silica. So, the carbon black reinforcement leads to larger sensitivity of the elastomer for the strain ratio E in the power-law cluster breakdown concept. The exponential model shows somewhat different tendency – 60% carbon black and 60% silica return almost the same gradients values where those first are little larger for intermediate values of E. The sensitivity of the carbon black elastomer apparently prevails, however for smaller amount of the reinforcement.
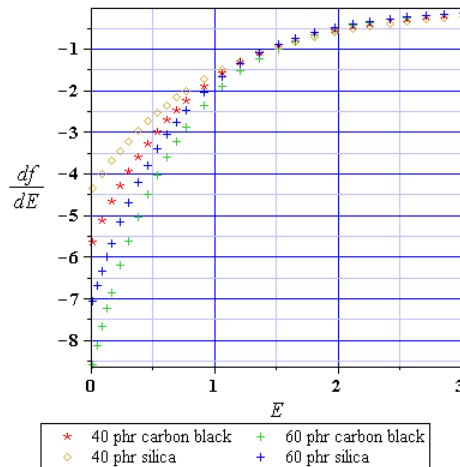

Fig. 8. Sensitivity coefficients for the exponential cluster breakdown via the scalar variable E
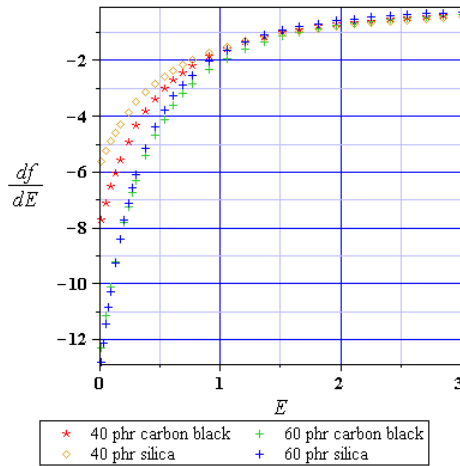
Fig. 9. Sensitivity coefficients for the power-law cluster breakdown to the scalar variable E

4. Effective behaviour for the elastomers with random parameters

The next step towards a more realistic description of the effective modulus for the elastomers reinforced with some fillers is probabilistic analysis, where some composite parameters or even their larger group is considered as the random variable or the vector of random variables. As is it known, there exists a variety of different mathematical approaches to analyze such a problem like determination of the probabilistic moments for $G^{(eff)}(\omega)$. One may use the algebraic transforms following basic probability theory definitions, Monte-Carlo simulation approaches, some spectral methods as well as some of the perturbation methods. Because the statistical description for $G_0(\omega)$ comes from the experiments we will focus here on determination of the random characteristics for $f(\omega)$ only. Because of some algebraic complexity of especially eqns (7-12) the stochastic perturbation technique based on the Taylor series expansion will be employed. To provide this formulation the random variable of the problem is denoted by $b(\omega)$ and the probability density of it as $g(b)$. The expected value of this variable is expressed by

$$E[b] = \int_{-\infty}^{+\infty} b \, g(b) db , \tag{14}$$

while the mth order moment as

$$\mu_m(b) = \int_{-\infty}^{+\infty} (b - E[b])^m g(b) db . \tag{15}$$

The coefficient of variation, asymmetry, flatness and kurtosis are introduced in the form

$$\alpha(b) = \frac{\sqrt{\mu_2(b)}}{E[b]} = \frac{\sqrt{Var(b)}}{E[b]} = \frac{\sigma(b)}{E[b]}, \quad \beta(b) = \frac{\mu_3(b)}{\sigma^3(b)}, \quad \gamma(b) = \frac{\mu_4(b)}{\sigma^4(b)}, \quad \kappa(b) = \gamma(b) - 3. \tag{16}$$

According to the main philosophy of this method, all functions in the basic deterministic problem (heat conductivity, heat capacity, temperature and its gradient as well as the material density) are expressed similarly to the following finite expansion of a random function f:

$$f(b) = f^0(b) + \varepsilon f^{,b}(b)\Delta b + \ldots + \tfrac{1}{n!}\varepsilon^n \frac{\partial^n f(b)}{\partial b^n}\Delta b^n, \tag{17}$$

where ε is a given small perturbation (taken usually as equal to 1), $\varepsilon\Delta b$ denotes the first order variation of b from its expected value

$$\varepsilon\Delta b = \delta b = \varepsilon(b - b^0), \tag{18}$$

while the nth order variation is given as follows:

$$\varepsilon^n \Delta b^n = (\delta b)^n = \varepsilon^n (b - b^0)^n. \tag{19}$$

Using this expansion, the expected values are exactly given by

$$E[f] = f^0(b^0) + \tfrac{1}{2}\varepsilon^2 \frac{\partial^2(f)}{\partial b^2}\mu_2(b) + \tfrac{1}{4!}\varepsilon^4 \frac{\partial^4(f)}{\partial b^4}\mu_4(b) + \tfrac{1}{6!}\varepsilon^6 \frac{\partial^6(f)}{\partial b^6}\mu_6(b) \tag{20}$$

$$+ \ldots + \tfrac{1}{(2m)!}\varepsilon^{2m} \frac{\partial^{2m}(f)}{\partial b^{2m}}\mu_{2m}(b)$$

for any natural m with μ_{2m} being the ordinary probabilistic moment of 2mth order. Usually, according to some previous convergence studies, we may limit this expansion-type approximation to the 10th order. Quite similar considerations lead to the expressions for higher moments, like the variance, for instance

$$Var(f) = \int_{-\infty}^{+\infty}(f^0 + \varepsilon\Delta b f^{,b} + \tfrac{1}{2}\varepsilon^2(\Delta b)^2 f^{,bb} + \tfrac{1}{3!}\varepsilon^3(\Delta b)^3 f^{,bbb} \tag{21}$$

$$+ \tfrac{1}{4!}\varepsilon^4(\Delta b)^4 f^{,bbbb} + \tfrac{1}{5!}\varepsilon^5(\Delta b)^5 f^{,bbbbb} - E[f])^2 p(b)db =$$

$$= \varepsilon^2 \mu_2(b) f^{,b} f^{,b} + \varepsilon^4 \mu_4(b)(\tfrac{1}{4} f^{,bb} f^{,bb} + \tfrac{2}{3!} f^{,b} f^{,bbb})$$

$$+ \varepsilon^6 \mu_6(b)((\tfrac{1}{3!})^2 f^{,bbb} f^{,bbb} + \tfrac{1}{4!} f^{,bbbb} f^{,bb} + \tfrac{2}{5!} f^{,bbbbb} f^{,b})$$

The third probabilistic moment may be recovered from this scheme as

$$\mu_3(f) = \int_{-\infty}^{+\infty}(f - E[f])^3 p(b)db = \int_{-\infty}^{+\infty}(f^0 + \varepsilon f^{,b}\Delta b + \tfrac{1}{2}\varepsilon^2 f^{,bb}\Delta b\Delta b + \ldots - E[f])^3 p(b)db \qquad (22)$$
$$\cong \tfrac{3}{2}\varepsilon^4\mu_4(b)(f^{,b})^2 f^{,bb} + \tfrac{1}{8}\varepsilon^6\mu_6(b)(f^{,bb})^3$$

using the lowest order approximation; the fourth probabilistic moment computation proceeds from the following formula:

$$\mu_4(f) = \int_{-\infty}^{+\infty}(f - E[f])^4 p(b)db = \int_{-\infty}^{+\infty}(\varepsilon f^{,b}\Delta b + \tfrac{1}{2}\varepsilon^2 f^{,bb}\Delta b\Delta b + \ldots)^4 p(b)db \qquad (23)$$
$$\cong \varepsilon^4\mu_4(b)(f^{,b})^4 + \tfrac{3}{2}\varepsilon^6\mu_6(b)(f^{,b}f^{,bb})^2 + \tfrac{1}{16}\varepsilon^8\mu_8(b)(f^{,b})^3(f^{,bb})^4$$

For the higher order moments we need to compute the higher order perturbations which need to be included into all formulas, so that the complexity of the computational model grows non-proportionally together with the precision and the size of the output information needed. This method may be applied as well to determine $G^{(eff)}(\omega)$ - one may apply the Taylor expansion to both components of the R.H.S. of eqn (1), differentiate it symbolically at least up to the given nth order (similarly to eqn (13) w.r.t. variable h) like below

$$\frac{\partial^n G^{(eff)}}{\partial b^n} = \sum_{k=0}^{n}\binom{n}{k}\frac{\partial^k f}{\partial b^k}\frac{\partial^{n-k}G_0}{\partial b^{n-k}} \qquad (24)$$

and include those derivatives into the probabilistic moment equations shown above.
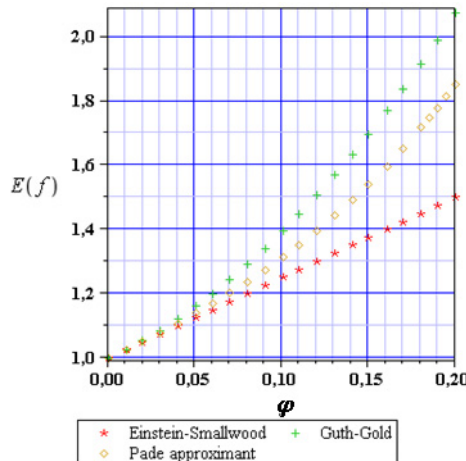

Fig. 10. The expected value of the volumetric coefficient f

The set of equations (20-23) with definitions given by (16) is implemented into the computer algebra system MAPLE, v. 11, as before, to determine the basic probabilistic characteristics for the function $f(\omega)$. The results of numerical analysis are presented in Figs. 10-25, where expected value of the input random variables are marked on the horizontal axes, its standard deviation corresponds to 15% of this expectation, while the output probabilistic moments of the parameter f are given on the vertical axes; the different theories for those coefficient calculations are compared analogously as in the previous sections. The Gaussian input random variables with given first two probabilistic moments are considered in all those computational illustrations
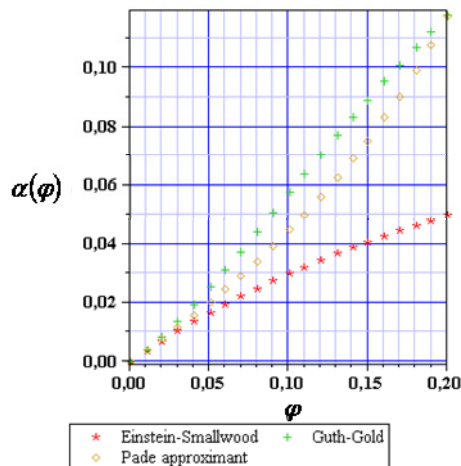
Fig. 11. The coefficient of variation of the volumetric coefficient f

Probabilistic moments and coefficients of up to 4th order of the simple volumetric approximations are given in Figs. 10-13 – there are in turn expected values, standard deviations, asymmetry and kurtosis. All the resulting functions increase here together with an increase of the reinforcement volumetric ratio φ. For the first two order characteristics the largest value is returned by the Guth-Gold model and the smallest - in the case of the Einstein-Smallwood approximation. Let us note also that the random dispersion of the output coefficient *f* is not constant and almost linearly dependent in all the models analyzed on the expected value of φ and is never larger here than the input value α(φ)=0.15. Third and fourth order characteristics demonstrate the maximum for the Guth-Gold theory for φ varying from 0 to the certain critical value, while for higher values the characteristics computed for the Pade approximants prevail significantly. All those characteristics are equal to 0 for the Einstein-Smallwood model because of a linear transform of the parameter φ in this model and it preserves exactly the character of the probability density function in a transform between the input φ and the output *f*. For the two remaining theories (with β>0) larger area of the probability density function remains above the expected value of *f*, while the concentration around this value if higher than for the Gaussian variables.

Now the basic probabilistic characteristics are compared for the homogenization model accounting for the clusters aggregation in the elastomers; the input coefficient φ is the input

Gaussian random parameter here also. Decisively larger values are obtained for the configuration without overlapping effect and all the characteristics are once more positive. In the case of expected values and standard deviations the influence of the coefficient φ on those quantities significantly prevail and has a clear linear character. A nonlinear variability with respect to Ξ is noticed for upper bound on the values of φ and has quite similar character in both Figs. 14 and 15. The maximum value of the coefficient of variation is about 0.02, which is around seven times smaller than the input coefficient, so that the random dispersion significantly decreases in this model.
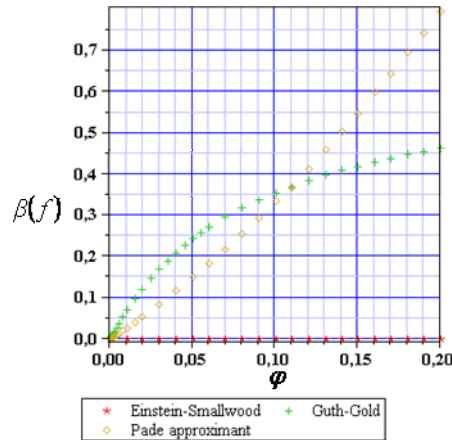
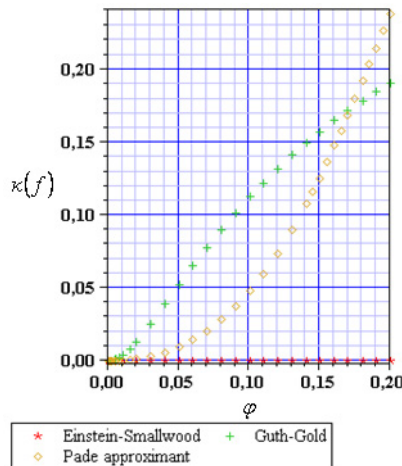Fig. 12. The coefficients of asymmetry of the volumetric coefficient f

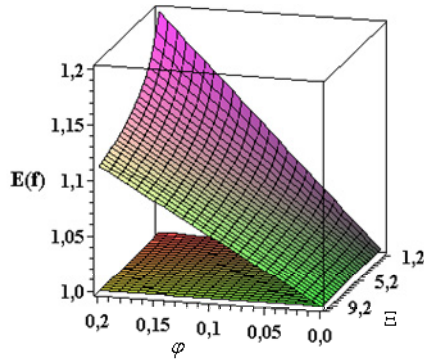Fig. 13. The kurtosis of the volumetric coefficient f

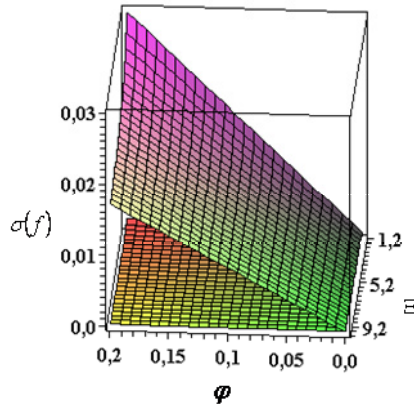Fig. 14. The expected values of the coefficient *f* including aggregation

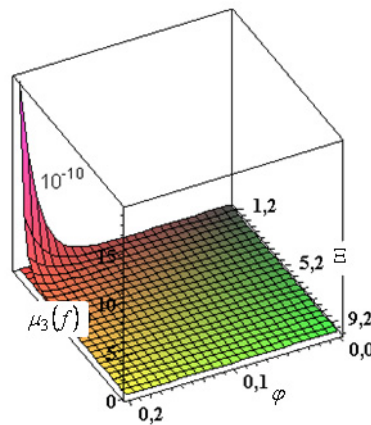Fig. 15. The standard deviations of the coefficient *f* including aggregation

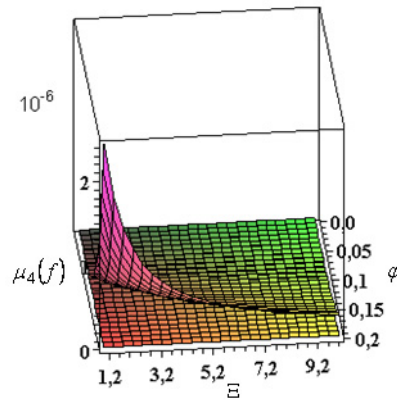Fig. 16. Third central probabilistic moments of the coefficient *f* including aggregation

Fig. 17. Fourth central probabilistic moments of the coefficient *f* including aggregation

Third central probabilistic moments increase rapidly from almost 0 only for the smallest values of Ξ and largest values of φ – it results in $\beta=0$ for the overlapping aggregates and $\beta=1.5$ for the aggregates with no overlap. Fourth moments variations are more apparent for larger values of φ and the entire spectrum of the parameter Ξ. The resulting kurtosis equal 0 and almost 2 – without and with this overlap, respectively. It is seen that the larger values of φ and the smaller Ξ, the larger 3rd and 4th probabilistic moments. So that, analogously to the previous theories, larger part of the resulting PDF is above the median and its concentration is higher than that typical for the Gaussian distribution (for the model without aggregates overlapped). The distribution of the random parameter *f* is almost the same like for the Gaussian input (except the coefficient of variation).

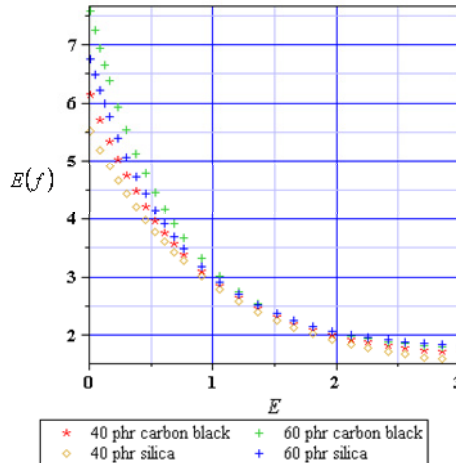|			
∗	40 phr carbon black	+	60 phr carbon black
◇	40 phr silica	+	60 phr silica

Fig. 18. The expected values for the exponential cluster breakdown to the scalar variable E

The probabilistic coefficients for the exponential (Figs. 18-21) and for the power-law (Figs. 22-25) cluster breakdowns are contrasted next; now the overall strain measure E is the Gaussian input variable. As one may predict from the deterministic result, all the expected values decrease together with an increase of the E expectation. The coefficient of variation $\alpha(f)$ (see Fig. 19) behave in a very interesting way – all they increase monotonously from 0 (for $E \cong 0$) to some maximum value (around a half of the considered strains scale) and next, they start to monotonously decrease; maximum dispersion is obtained for 60% of the carbon black here.
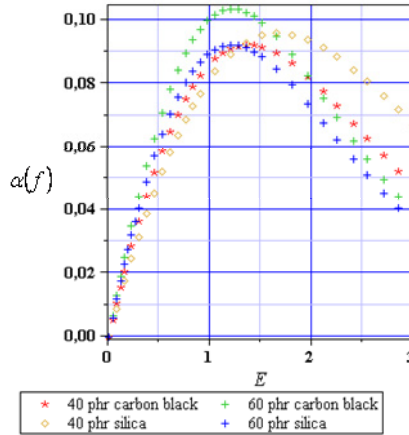

Fig. 19. Coefficients of variation for exponential cluster breakdown to the scalar variable E
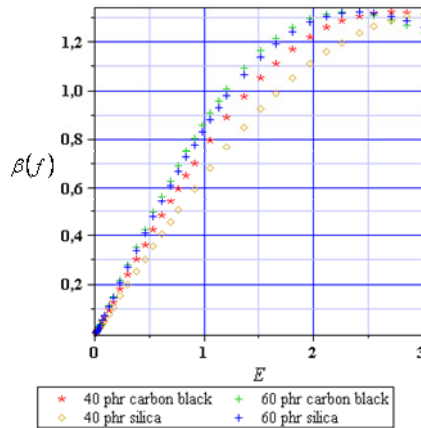

Fig. 20. Asymmetry coefficient for the exponential cluster breakdown to the scalar variable E
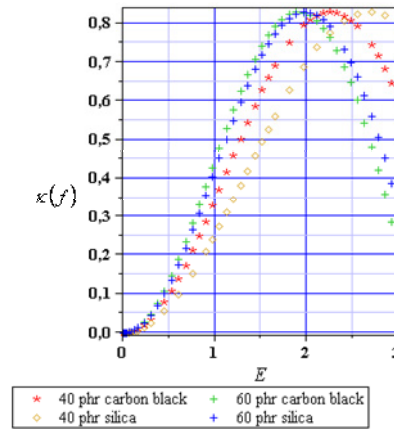
Fig. 21. The kurtosis for the exponential cluster breakdown to the scalar variable E

The asymmetry coefficient $\beta(f)$ and the kurtosis $\kappa(f)$ also behave similarly to $\alpha(f)$, where the additional maxima appear for larger and smaller values of the strain measure E; the coefficient $\beta(f)$ remains positive for all values of the parameter E. Within smaller E values range we notice that larger values of both coefficients are observed for the carbon black and they increase also together with an increase of the reinforcement volumetric ratio. Similarly as before, the PDFs concentration is higher than that for the Gaussian distribution and a right part of resulting distributions prevail. The expected values for the power-law cluster breakdown are shown in Fig. 22; they decrease together with the expectation of the strain measure E and the larger the reinforcement volume is, the larger is the expectation $E[f]$. The coefficients of variation are less predictable here (Fig. 23) – they monotonously increase for 40% of both reinforcing particles, whereas for 60% silica and carbon black they monotonously increase until some maximum and afterwards they both start to decrease; the particular values are close to those presented in Fig. 19.
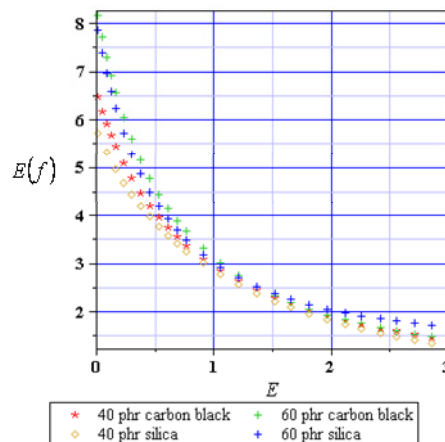

Fig. 22. The expected values for the power-law cluster breakdown to the scalar variable E
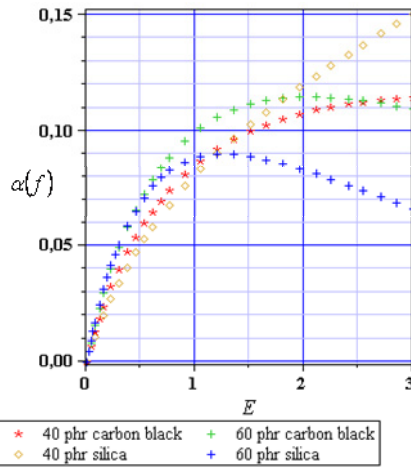
Fig. 23. Coefficients of variation for power-law cluster breakdown to the scalar variable E
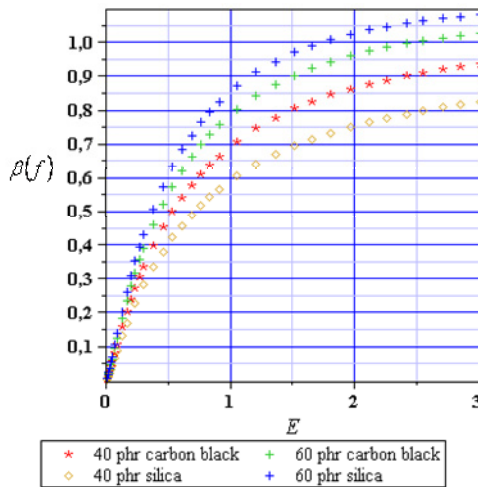
Fig. 24. Asymmetry coefficient for the power-law cluster breakdown to the scalar variable E

The coefficients of asymmetry and kurtosis do not increase as those in Figs. 20-21 – they simply monotonously increase from 0 value typical for E=0 to their maxima for E=3 (the only exception in this rule is kurtosis of the elastomer with 60% of the silica particles). Both coefficients have larger values for the carbon black than for silica and they increase together with the additional reinforcement volumetric ratio increase. Although coefficients of asymmetry exhibit the values quite close to those obtained for the exponential breakdown approach, the kurtosis is approximately two times smaller than before (Fig. 25 vs. Fig. 21).

Fig. 25. The kurtosis for the power-law cluster breakdown to the scalar variable E

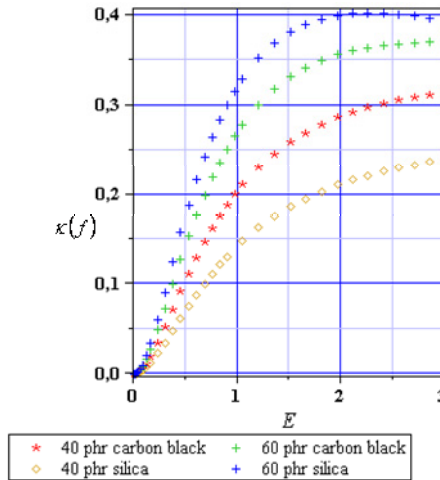## 5. Homogenized parameters for elastomers subjected to the stochastic aging

The engineering practice in many cases leads to the conclusion that the initial values of mechanical parameters decrease stochastically together with the time being. As far as some periodic measurements are available one can approximate in some way those stochastic process moments, however a posteriori analysis is not convenient considering the reliability of designed structures and materials. This stochasticity does not need result from the cyclic fatigue loading (Heinrich et al., 2002b), but may reflect some unpredictable structural accidents, aggressive environmental influences etc. This problem may be also considered in the context of the homogenization method, where the additional formula for effective parameters may include some stochastic processes. Considering above one may suppose for instance the scalar strain variable E as such a process, i.e.

$$E(\omega,t) = E^0(\omega) + \dot{E}(\omega)\, t \tag{25}$$

where superscript 0 denotes here the initial random distribution of the given parameter and dotted quantities stand for the random variations of those parameters (measured in years). From the stochastic point of view it is somewhat similar to the Langevin equation approach (Mark, 2007), where Gaussian fluctuating white noise was applied. It is further assumed that all aforementioned random variables in eqns (25,26) are Gaussian and their first two moments are given; the goal would be to find the basic moments of the process $f(\omega,t)$ to be included in some stochastic counterpart of eqn (1). The plus in eqn (25) suggests that the strain measure with some uncertainty should increase with time (Mark, 2007) according to some unpredictable deformations; introduction of higher order polynomium is also possible here and does not lead to significant computational difficulty. A determination of the first two moments of the process given by eqn (25) leads to the formulas

$$E[E(\omega,t)] = E[E^0(\omega)] + E[\dot{E}(\omega)]\, t \tag{26}$$

and

$$Var(E(\omega,t)) = Var(E^0(\omega,t)) + Var(\dot{E}(\omega,t))\, t^2 . \tag{27}$$

Now four input parameters are effectively needed to provide the analysis for stochastic ageing of any of the models presented above; the additional computational analysis was performed with respect to the exponential and power-law cluster breakdown models below. This part of computational experiments started from the determination of the expected values (Fig. 26), coefficients of variation (Fig. 27), the asymmetry coefficients (Fig. 28) and the kurtosis (Fig. 29) time fluctuations in the power-law model. For this purpose the following input data are adoped: $E[E_0] = 3$, $E[\dot{E}] = 0.03\ year^{-1}$, $Var(E_0) = (0.01\ E[E_0])^2$ and $Var(\dot{E}) = (0.01\ E[\dot{E}])^2$, so that the initial strain measure has extremely large expected value and it still stochastically increases; the time scale for all those experiments marked on the horizontal axis is given of course in years. A general observation is that all of those characteristics decrease together with a time increment, not only the expected value. The elastomer shear modulus become closer to the matrix rather together with the time being and the random distribution of the output coefficient f converges with time to the Gaussian one, however the coefficient of variation also tends to 0 (for at least 60% silica). The interrelations between different elastomers are the same for expectations, asymmetry and kurtosis – larger values are obtained for silica than for the carbon black and the higher volumetric ratio (in percents) the higher values of those probabilistic characteristics; this result remains in the perfect agreement with Figs. 22-25 (showing an initial state to this analysis).
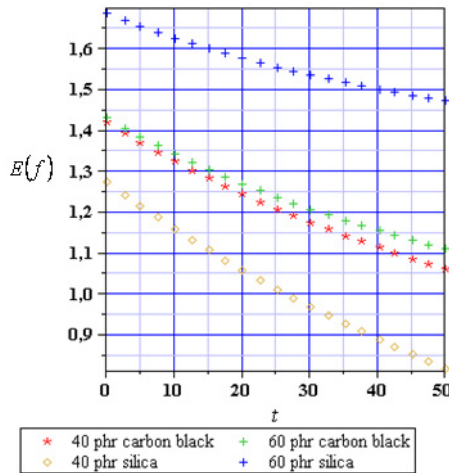
Fig. 26. The expected values for the power-law cluster breakdown to the scalar variable E

The coefficient of variation exhibit exactly the inverse interrelations – higher values are typical for silica reinforcement and for smaller amount of the reinforcing particles in the

elastomer specimen. For 40% silica the expected value of the reinforcement coefficient f becomes smaller than 1 after almost 25 years of such a stochastic ageing. It is apparent that we can determine here the critical age of the elastomer when it becomes too weak for the specific engineering application or, alternatively, determine the specific set of the input data to assure its specific design durability.

Fig. 27. Coefficients of variation for power-law cluster breakdown to the scalar variable E
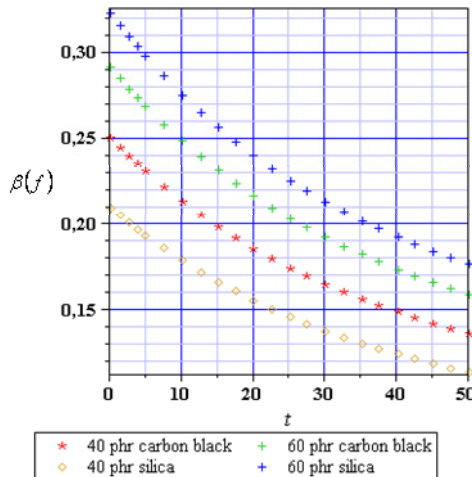
Fig. 28. Asymmetry coefficient for the power-law cluster breakdown to the scalar variable E

The input data set for the stochastic ageing of the elastomer according to the exponential cluster breakdown model is exactly the same as in the power-law approach given above. It results in the expectations (Fig. 30), coefficients of variation (Fig. 31), asymmetry coefficients

(Fig. 32) and kurtosis (Fig. 33) time variations for $t \in [0, 50\ years]$. Their time fluctuations are generally similar qualitatively as before because all of those characteristics decrease in time. The expectations are slightly larger than before and never crosses a limit value of 1, whereas the coefficients are of about three order smaller than those in Fig. 27. The coefficients $\beta(t)$ are now around two times larger than in the case of the power-law cluster breakdown. The interrelations between the particular elastomers are different than those before – although silica dominates and $E[f]$ increases together with the reversed dependence on the reinforcement ratio, the quantitative differences between those elastomers are not similar at all to Figs. 26-27.
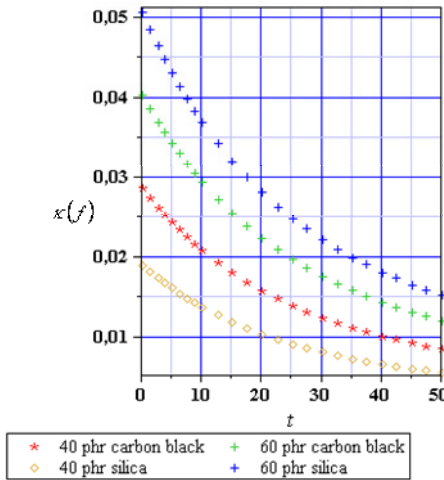
Fig. 29. The kurtosis for the power-law cluster breakdown to the scalar variable E
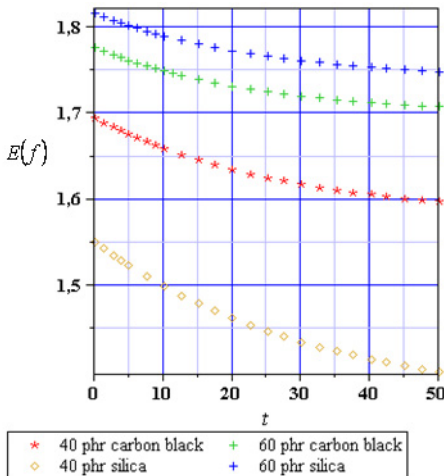
Fig. 30. The expected values for the exponential cluster breakdown to the scalar variable E

The particular elastomers coefficients of asymmetry and kurtosis histories show that larger values are noticed for the carbon black than for the silica and, at the same time, for larger volume fractions of the reinforcements into the elastomer.
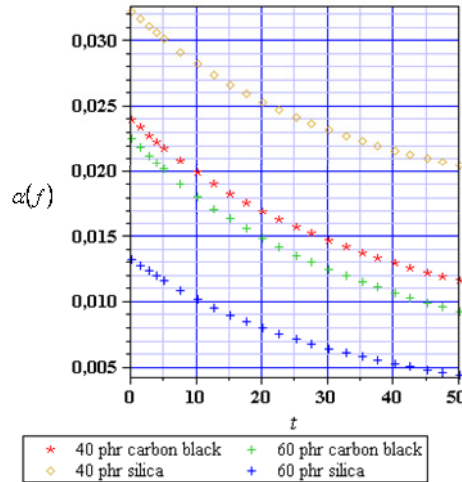

Fig. 31. Coefficients of variation for exponential cluster breakdown to the scalar variable E
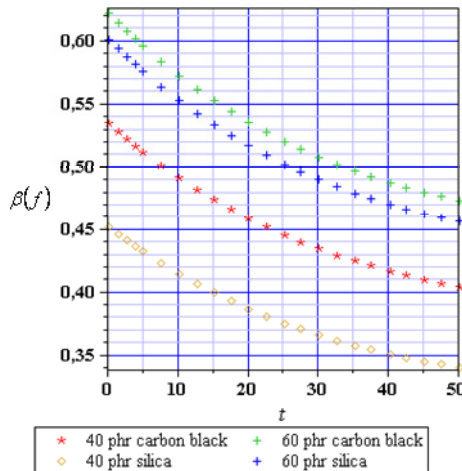
Fig. 32. Asymmetry coefficient for the exponential cluster breakdown to the scalar variable E
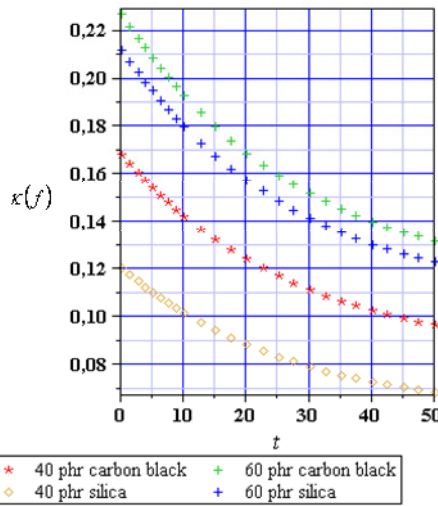
Fig. 33. The kurtosis for the exponential cluster breakdown to the scalar variable E

6. Concluding remarks

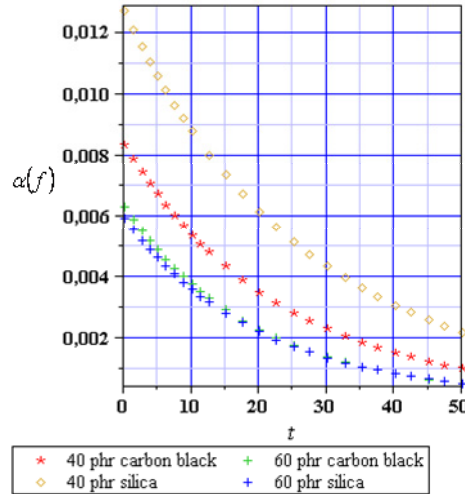1. The computational methodology presented and applied here allows a comparison of various homogenization methods for elastomers reinforced with nanoparticles in terms of parameter variability, sensitivity gradients as well as the resulting probabilistic moments. The most interesting result is the overall decrease of the probabilistic moments for the process $f(\omega;t)$ together with time during stochastic ageing of the elastomer specimen defined as the stochastic increase of the general strain measure E. For further applications an application of the non-Gaussian variables (and processes) is also possible with this model.

2. The results of probabilistic modeling and stochastic analysis are very useful in stochastic reliability analysis of tires, where homogenization methods presented above significantly simplify the computational Finite Element Method model. On the other hand, one may use the stochastic perturbation technique applied here together with the LEFM or EPFM approaches to provide a comparison with the statistical results obtained during the basic impact tests (to predict numerically expected value of the tensile stress at the break) (Reincke et al., 2004).

3. Similarly to other existing and verified homogenization theories, one may use here the energetic approach, where the effective coefficients are found by the equity of strain energies accumulated into the real and the homogenized specimens and calculated from the additional Finite Element Method experiments, similarly to those presented by Fukahori, 2004 and Gehant et al., 2003. This technique, nevertheless giving the relatively precise approximations (contrary to some upper and lower bounds based approaches), needs primary Representative Volume Element consisting of some reinforcing cluster.

7. Acknowledgment

8. References

Bhowmick, A.K., ed. (2008). *Current Topics in Elastomers Research*, CRC Press, ISBN 13: 9780849373176, Boca Raton, Florida

Christensen, R.M. (1979). *Mechanics of Composite Materials*, ISBN 10:0471051675, Wiley

Dorfmann, A. & Ogden, R.W. (2004). A constitutive model for the Mullins effect with permanent set in particle-reinforced rubber, *Int. J. Sol. Struct.*, vol. 41, 1855-1878, ISSN 0020-7683

Fu, S.Y., Lauke, B. & Mai, Y.W. (2009). *Science and Engineering of Short Fibre Reinforced Polymer Composites*, CRC Press, ISBN 9781439810996, Boca Raton, Florida

Fukahori, Y. (2004). The mechanics and mechanism of the carbon black reinforcement of elastomers, *Rubber Chem. Techn.*, Vol. 76, 548-565, ISSN 0035-9475

Gehant, S., Fond, Ch. & Schirrer, R. (2003). Criteria for cavitation of rubber particles: Influence of plastic yielding in the matrix, *Int. J. Fract.*, Vol. 122, 161-175, ISSN 0376-9429

Heinrich, G., Klűppel, M. & Vilgis, T.A. (2002). Reinforcement of elastomers, *Current Opinion in Solid State Mat. Sci.*, Vol. 6, 195-203, ISSN 1359-0286

Heinrich, G., Struve, J. & Gerber, G. (2002). Mesoscopic simulation of dynamic crack propagation in rubber materials, *Polymer*, Vol. 43, 395-401, ISSN 0032-3861

Kamiński, M. (2005). *Computational Mechanics of Composite Materials*, ISBN 1852334274, Springer-Verlag, London-New York

Kamiński, M. (2009). Sensitivity and randomness in homogenization of periodic fiber-reinforced composites via the response function method, *Int. J. Sol. Struct.*, Vol. 46, 923-937, ISSN 0020-7683

Mark, J.E. (2007). *Physical Properties of Polymers Handbook*, 2nd edition, ISBN 13: 9780387312354, Springer-Verlag, New York

Reincke, K., Grellmann, W. & Heinrich, G. (2004). Investigation of mechanical and fracture mechanical properties of elastomers filled with precipitated silica and nanofillers based upon layered silicates, *Rubber Chem. Techn.*, Vol. 77, 662-677, ISSN 0035-9475

Stochastic improvement of structural design

Soprano Alessandro and Caputo Francesco
Second University of Naples
Italy

1. Introduction

It is well understood nowadays that design is not an one-step process, but that it evolves along many phases which, starting from an initial idea, include drafting, preliminary evaluations, trial and error procedures, verifications and so on. All those steps can include considerations that come from different areas, when functional requirements have to be met which pertain to fields not directly related to the structural one, as it happens for noise, environmental prescriptions and so on; but even when that it's not the case, it is very frequent the need to match against opposing demands, for example when the required strength or stiffness is to be coupled with lightness, not to mention the frequently encountered problems related to the available production means.

All the previous cases, and the many others which can be taken into account, justify the introduction of particular design methods, obviously made easier by the ever-increasing use of numerical methods, and first of all of those techniques which are related to the field of mono- or multi-objective or even multidisciplinary optimization, but they are usually confined in the area of deterministic design, where all variables and parameters are considered as fixed in value. As we discuss below, the random, or stochastic, character of one or more parameters and variables can be taken into account, thus adding a deeper insight into the real nature of the problem in hand and consequently providing a more sound and improved design.

Many reasons can induce designers to study a structural project by probabilistic methods, for example because of uncertainties about loads, constraints and environmental conditions, damage propagation and so on; the basic methods used to perform such analyses are well assessed, at least for what refers to the most common cases, where structures can be assumed to be characterized by a linear behaviour and when their complexity is not very great.

Another field where probabilistic analysis is increasingly being used is that related to the requirement to obtain a product which is 'robust' against the possible variations of manufacturing parameters, with this meaning both production tolerances and the settings of machines and equipments; in that case one is looking for the 'best' setting, i.e. that which minimizes the variance of the product against those of design or control variables.

A very usual case – but also a very difficult to be dealt – is that where it is required to take into account also the time variable, which happens when dealing with a structure which degrades because of corrosion, thermal stresses, fatigue, or others; for example, when studying very light structures, such as those of aircrafts, the designer aims to ensure an assigned life to them, which are subjected to random fatigue loads; in advanced age the

aircraft is interested by a WFD (Widespread Fatigue Damage) state, with the presence of many cracks which can grow, ultimately causing failure. This case, which is usually studied by analyzing the behaviour of significant details, is a very complex one, as one has to take into account a large number of cracks or defects, whose sizes and locations can't be predicted, aiming to delay their growth and to limit the probability of failure in the operational life of the aircraft within very small limits (about $10^{-7} \pm 10^{-9}$).

The most widespread technique is a 'decoupled' one, in the sense that a forecast is introduced by one of the available methods about the amount of damage which will probably take place at a prescribed instant and then an analysis in carried out about the residual strength of the structure; that is because the more general study which makes use of the stochastic analysis of the structure is a very complex one and still far away for the actual solution methods; the most used techniques, as the first passage theory, which claim to be the solution, are just a way to move around the real problems.

In any case, the probabilistic analysis of the structure is usually a final step of the design process and it always starts on the basis of a deterministic study which is considered as completed when the other starts. That is also the state that will be considered in the present chapter, where we shall recall the techniques usually adopted and we shall illustrate them by recalling some case studies, based on our experience.

For example, the first case which will be illustrated is that of a riveted sheet structure of the kind most common in the aeronautical field and we shall show how its study can be carried out on the basis of the considerations we introduced above.

The other cases which will be presented in this paper refer to the probabilistic analysis and optimization of structural details of aeronautical as well as of automotive interest; thus, we shall discuss the study of an aeronautical panel, whose residual strength in presence of propagating cracks has to be increased, and with the study of an absorber, of the type used in cars to reduce the accelerations which act on the passengers during an impact or road accident, and whose design has to be improved. In both cases the final behaviour is influenced by design, manufacturing process and operational conditions.

2. General methods for the probabilistic analysis of structures

If we consider the n-dimensional space defined by the random variables which govern a generic problem ("design variables") and which consist of geometrical, material, load, environmental and human factors, we can observe that those sets of coordinates (**x**) that correspond to failure define a domain (the 'failure domain' Ω_f) in opposition to the remainder of the same space, that is known as the 'safety domain' (Ω_s) as it corresponds to survival conditions.

In general terms, the probability of failure can be expressed by the following integral:

$$P_f = \int_{\Omega_f} f(\mathbf{x}) \cdot d\mathbf{x} = \int_{\Omega_f} f_{x_1 x_2 \cdots x_n}(x_1, x_2, \cdots x_n) \cdot dx_1 dx_2 \cdots dx_n \qquad (1)$$

where f_i represents the joint density function of all variables, which, in turn, may happen to be also functions of time. Unfortunately that integral cannot be solved in a closed form in most cases and therefore one has to use approximate methods, which can be included in one of the following typologies:

1) methods that use the limit state surface (LSS, the surface that constitutes the boundary of the failure region) concept: they belong to a group of techniques that model variously the

LSS in both shape and order and use it to obtain an approximate probability of failure; among these, for instance, particularly used are FORM (First Order Reliability Method) and SORM (Second Order Reliability Method), that represent the LSS respectively through the hyper-plane tangent to the same LSS at the point of the largest probability of occurrence or through an hyper-paraboloid of rotation with the vertex at the same point.

2) Simulation methodologies, which are of particular importance when dealing with complex problems: basically, they use Monte-Carlo (MC) technique for the numerical evaluation of the integral above and therefore they define the probability of failure on a frequency basis.

As pointed above, it is necessary to use a simulation technique to study complex structures, but in the same cases each trial has to be carried out through a numerical analysis (for example by FEM); if we couple that circumstance with the need to perform a very large number of trials, which is the case when dealing with very small probabilities of failure, very large runtimes are obtained, which are really impossible to bear. Therefore different means have been introduced in recent years to reduce the number of trials and to make acceptable the simulation procedures.

In this section, therefore, we resume briefly the different methods which are available to carry out analytic or simulation procedures, pointing out the difficulties and/or advantages which characterize them and the particular problems which can arise in their use.

2.1 LSS-based analytical methods

Those methods come from an idea by Cornell (1969), as modified by Hasofer and Lind (1974) who, taking into account only those cases where the design variables could be considered to be normally distributed and uncorrelated, each defined by their mean value μ_I and standard deviation σ_I, modeled the LSS in the standard space, where each variable is represented through the corresponding standard variable, i.e.

$$u_i = \frac{x_i - \mu_i}{\sigma_i} \tag{2}$$

If the LSS can be represented by a hyperplane (fig. 1), it can be shown that the probability of failure is related to the distance β of LSS from the origin in the standard space and therefore is given by

$$P_{fFORM} = 1 - \Phi(\beta) = \Phi(-\beta) \tag{3}$$
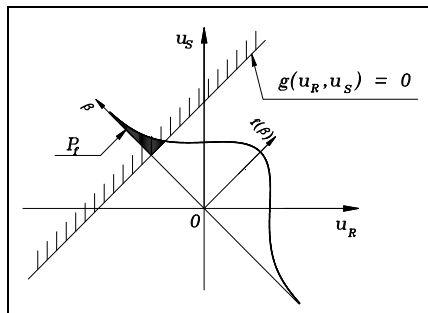
Fig. 1. Probability of failure for a hyperplane LSS
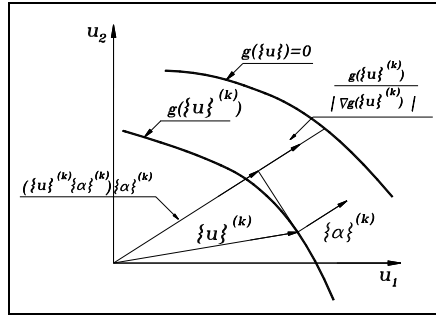
Fig. 2. The search for the design point according to RF's method

It can be also shown that the point of LSS which is located at the least distance β from the origin is the one for which the elementary probability of failure is the largest and for that reason it is called the maximum probability point (MPP) or the design point (DP).

Those concepts have been applied also to the study of problems where the LSS cannot be modeled as an hyperplane; in those cases the basic methods try to approximate the LSS by means of some polynomial, mostly of the first or the second degree; broadly speaking, in both cases the technique adopted uses a Taylor expansion of the real function around some suitably chosen point to obtain the polynomial representation of the LSS and it is quite obvious to use the design point to build the expansion, as thereafter the previous Hasofer and Lind's method can be used.

It is then clear that the solution of such problems requires two distinct steps, i.e. the research of the design point and the evaluation of the probability integral; for example, in the case of FORM (First Order Reliability Method) the most widely applied method, those two steps are coupled in a recursive form of the gradient method (fig. 2), according to a technique introduced by Rackwitz and Fiessler (RF's method). If we represent the LSS through the function $g(\mathbf{x}) = 0$ and indicate with α_I the direction cosines of the inward-pointing normal to the LSS at a point x_0, given by

$$\alpha_i = -\frac{1}{|\nabla g|_0}\left(\frac{\partial g}{\partial u_i}\right)_0 \tag{4}$$

starting from a first trial value of $\mathbf{u}$, the kth n-uple is given by

$$\{\mathbf{u}\}_k = \left[\{\mathbf{u}\}_{k-1}^{T} \cdot \{\alpha\}_k + \frac{g(\{\mathbf{u}\}_{k-1})}{\nabla g(\{\mathbf{u}\}_{k-1})}\right] \cdot \{\alpha\}_k \tag{5}$$

thus obtaining the required design point within an assigned approximation; its distance from the origin is just β and then the probability of failure can be obtained through eq. 3 above.

One of the most evident errors which follow from that technique is that the probability of failure is usually over-estimated and that error grows as curvatures of the real LSS increase; to overcome that inconvenience in presence of highly non-linear surfaces, the SORM

(Second Order Reliability Method) was introduced, but, even with Tved's and Der Kiureghian's developments, its use implies great difficulties. The most relevant result, due to Breitung, appears to be the formulation of the probability of failure in presence of a quadratic LSS via FORM result, expressed by the following expression:

$$P_{fSORM} = \Phi(-\beta) \cdot \prod_{i=1}^{n-1} (1 - \beta \cdot \kappa_i)^{-1/2} = P_{fFORM} \cdot \prod_{i=1}^{n-1} (1 - \beta \cdot \kappa_i)^{-1/2} \qquad (6)$$

where κ_I is the i-th curvature of the LSS; if the connection with FORM is a very convenient one, the evaluation of curvatures usually requires difficult and long computations; it is true that different simplifying assumptions are often introduced to make solution easier, but a complete analysis usually requires a great effort. Moreover, it is often disregarded that the above formulation comes from an asymptotic development and that consequently its result is so more approximate as β values are larger.

As we recalled above, the main hypotheses of those procedures are that the random variables are uncorrelated and normally distributed, but that is not the case in many problems; therefore, some methods have been introduced to overcome those difficulties.

For example, the usually adopted technique deals with correlated variables via an orthogonal transformation such as to build a new set of variables which are uncorrelated, using the well known properties of matrices. For what refers to the second problem, the current procedure is to approximate the behaviour of the real variables by considering dummy gaussian variables which have the same values of the distribution and density functions; that assumption leads to an iterative procedure, which can be stopped when the required approximation has been obtained: that is the original version of the technique, which was devised by Ditlevsen and which is called Normal Tail Approximation; other versions exist, for example the one introduced by Chen and Lind, which is more complex and which, nevertheless, doesn't bring any deeper knowledge on the subject.

At last, it is not possible to disregard the advantages connected with the use of the Response Surface Method, which is quite useful when dealing with rather large problems, for which it is not possible to forecast *a priori* the shape of the LSS and, therefore, the degree of the approximation required. That method, which comes from previous applications in other fields, approximate the LSS by a polynomial, usually of second degree, whose coefficients are obtained by Least Square Approximation or by DOE techniques; the procedure, for example according to Bucher and Burgund, evolves along a series of convergent trials, where one has to establish a center point for the i-th approximation, to find the required coefficients, to determine the design point and then to evaluate the new approximating center point for a new trial.

Beside those here recalled, other methods are available today, such as the Advanced Mean Value or the Correction Factor Method, and so on, and it is often difficult to distinguish their own advantages, but in any case the techniques which we outlined here are the most general and known ones; broadly speaking, all those methods correspond to different degree of approximation, so that their use is not advisable when the number of variables is large or when the expected probabilities of failure is very small, as it is often the case, because of the overlapping of the errors, which can bring results which are very far from the real one.

2.2 Simulation-based reliability assessment

In all those cases where the analytical methods are not to be relied on, for example in presence of many, maybe even not gaussian, variables, one has to use simulation methods to assess the reliability of a structure: about all those methods come from variations or developments of an 'original' method, whose name is Monte-Carlo method and which corresponds to the frequential (or *a posteriori*) definition of probability.
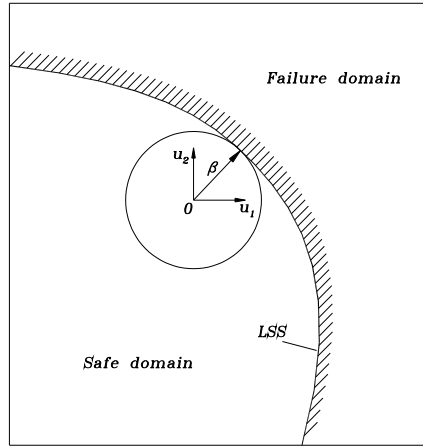
Fig. 3. Domain Restricted Sampling

For a problem with k random variables, of whatever distribution, the method requires the extraction of k random numbers, each of them being associated with the value of one of the variables via the corresponding distribution function; then, the problem is run with the found values and its result (failure of safety) recorded; if that procedure is carried out N times, the required probability, for example that corresponding to failure, is given by P_f = n/N, if the desired result has been obtained n times.

Unfortunately, broadly speaking, the procedure, which can be shown to lead to the 'exact' evaluation of the required probability if N = ∞, is very slow to reach convergence and therefore a large number of trials have to be performed; that is a real problem if one has to deal with complex cases where each solution is to be obtained by numerical methods, for example by FEM or others. That problem is so more evident as the largest part of the results are grouped around the mode of the result distribution, while one usually looks for probability which lie in the tails of the same distribution, i.e. one deals with very small probabilities, for example those corresponding to the failure of an aircraft or of an ocean platform and so on.

It can be shown, by using Bernouilli distribution, that if p is the 'exact' value of the required probability and if one wants to evaluate it with an assigned e_{max} error at a given confidence level defined by the bilateral protection factor k, the minimum number of trials to be carried out is given by

$$N_{min} = \left(\frac{2 \cdot k}{e_{max}} \right)^2 \frac{1-p}{p} \tag{7}$$

for example, if p = 10⁻⁵ and we want to evaluate it with a 10% error at the 95% confidence level, we have to carry out at least N_{min} = 1.537·10⁸ trials, which is such a large number that usually larger errors are accepted, being often satisfied to get at least the order of magnitude of the probability.

It is quite obvious that various methods have been introduced to decrease the number of trials; for example, as we know that no failure point is to be found at a distance smaller than β from the origin of the axis in the standard space, Harbitz introduced the Domain Restricted Sampling (fig. 3), which requires the design point to be found first and then the trials are carried out only at distances from the origin larger than β; the Importance Sampling Method is also very useful, as each of the results obtained from the trials is weighted according to a function, which is given by the analyst and which is usually centered at the design point, with the aim to limit the number of trials corresponding to results which don't lie in the failure region.
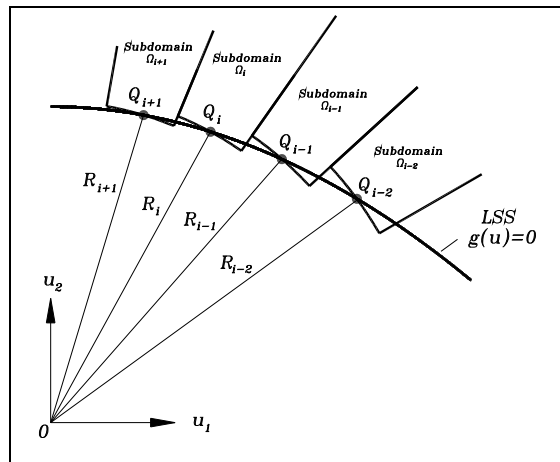
Fig. 4. The method of Directional Simulation

One of the most relevant technique which have been introduced in the recent past is the one known as Directional Simulation; in the version published by Nie and Ellingwood, the sample space is subdivided in an assigned number of sectors through radial hyperplanes (fig. 4); for each sector the mean distance of the LSF is found and the corresponding probability of failure is evaluated, the total probability being given by the simple sum of all results; in this case, not only the number of trials is severely decreased, but a better approximation of the frontier of the failure domain is achieved, with the consequence that the final probability is found with a good approximation.

Other recently appeared variations are related to the extraction of random numbers; those are, in fact, uniformly distributed in the 0-1 range and therefore give results which are rather clustered around the mode of the final distribution. That problem can be avoided if one resorts to use not really random distributions, as those coming from k-discrepancy theory, obtaining points which are better distributed in the sample space.

A new family of techniques have been introduced in the last years, all pertaining to the general family of *genetic algorithms*; that evocative name is usually coupled with an

imaginative interpretation which recalls the evolution of animal settlements, with all its content of selection, marriage, breeding and mutations, but it really covers in a systematic and reasoned way all the steps required to find the design point of an LSS in a given region of space. In fact, one has to define at first the size of the population, i.e. the number of sample points to be used when evaluating the required function; if that function is the distance of the design point from the origin, which is to be minimized, a selection is made such as to exclude from the following steps all points where the value assumed by the function is too large. After that, it is highly probable that the location of the minimum is between two points where the same function shows a small value: that coupling is what corresponds to marriage in the population and the resulting intermediate point represents the breed of the couple. Summing up the previous population, without the excluded points, with the breed, gives a new population which represents a new generation; in order to look around to observe if the minimum point is somehow displaced from the easy connection between parents, some mutation can be introduced, which corresponds to looking around the new-found positions.

It is quite clear that, besides all poetry related to the algorithm, it can be very useful but it is quite difficult to be used, as it is sensitive to all different choices one has to introduce in order to get a final solution: the size of the population, the mating criteria, the measure and the way of the introduction in breed of the parents' characters, the percentage and the amplitude of mutations, are all aspects which are to be the objects of single choices by the analyst and which can have severe consequences on the results, for example in terms of the number of generations required to attain convergence and of the accuracy of the method.

That's why it can be said that a general genetic code which can deal with all reliability problems is not to be expected, at least in the near future, as each problem requires specific cares that only the dedicated attentions of the programmer can guarantee.

3. Examples of analysis of structural details

An example is here introduced to show a particular case of stochastic analysis as applied to the study of structural details, taken from the authors' experience in research in the aeronautical field.

Because of their widespread use, the analysis of the behaviour of riveted sheets is quite common in aerospace applications; at the same time the interest which induced the authors to investigate the problems below is focused on the last stages of the operational life of aircraft, when a large number of fatigue-induced cracks appear at the same time in the sheets, before at least one of them propagates up to induce the failure of the riveted joint: the requirement to increase that life, even in presence of such a population of defects (when we say that a stage of Widespread Fatigue Damage, WFD, is taking place) compelled the authors to investigate such a scenario of a damaged structure.

3.1 Probabilistic behaviour of riveted joints

One of the main scopes of the present activity was devoted to the evaluation of the behaviour of a riveted joint in presence of damage, defined for example as a crack which, stemming from the edge of one of the holes of the joint, propagates toward the nearest one, therefore introducing a higher stress level, at least in the zone adjacent to crack tip.

It would be very appealing to use such easy procedures as compounding to evaluate SIF's for that case, which, as it is now well known, gives an estimate of the stress level which is built by reducing the problem at hand to the combination of simpler cases, for which the solution is known; that procedure is entirely reliable, but for those cases where singularities are so near to each other to develop an interaction effect which the method is not able to take into account.

Unfortunately, even if a huge literature is now available about edge cracks of many geometry, the effect of a loaded hole is not usually treated with the extent it deserves, may be for the particular complexity of the problem; for example, the two well known papers by Tweed and Rooke (1979; 1980) deal with the evaluation of SIF for a crack stemming from a loaded hole, but nothing is said about the effect of the presence of other loaded holes toward which the crack propagates.

Therefore, the problem of the increase of the stress level induced from a propagating crack between loaded holes could be approached only by means of numerical methods and the best idea was, of course, to use the results of FEM to investigate the case. Nevertheless, because of the presence of the external loads, which can alter or even mask the effects of loaded holes, we decided to carry out first an investigation about the behaviour of SIF in presence of two loaded holes.

The first step of the analysis was to choose which among the different parameters of the problem were to be treated as random variables.

Therefore a sort of sensitivity analysis was to be carried out; in our case, we considered a very specific detail, i.e. the space around the hole of a single rivet, to analyze the influence of the various parameters.

By using a Monte-Carlo procedure, some probability parameters were introduced according to experimental evidence for each of the variables in order to assess the required influence on the mean value and the coefficient of variation of the number of cycles before failure of the detail.

In any case, as pitch and diameter of the riveted holes are rather standardized in size, their influence was disregarded, while the sheet thickness was assumed as a deterministic parameter, varying between 1.2 and 4.8 mm; therefore, the investigated parameters were the stress level distribution, the size of the initial defect and the parameters of the propagation law, which was assumed to be of Paris' type.

For what refers to the load, it was supposed to be in presence of traction load cycles with R = 0 and with a mean value which followed a Gaussian probability density function around 60, 90 and 120 MPa, with a coefficient of variation varying according assigned steps; initial crack sizes were considered as normally distributed from 0.2 mm up to limits depending on the examined case, while for what concerns the two parameters of Paris' law, they were considered as characterized by a normal joint pdf between the exponent n and the logarithm of the other one.

Initially, an extensive exploration was carried out, considering each variable in turn as random, while keeping the others as constant and using the code NASGRO® to evaluate the number of cycles to failure; an external routine was written in order to insert the crack code in a M-C procedure. CC04 and TC03 models of NASGRO® library were adopted in order to take into account corner- as well as through-cracks. For all analyses 1,000 trials/point were carried out, as it was assumed as a convenient figure to be accepted to obtain rather stabilized results, while preventing the total runtimes from growing unacceptably long; the said M-C procedure was performed for an assigned statistics of one input variable at the time.

The results obtained can be illustrated by means of the following pictures and first of all of the fig. 5 where the dependence of the mean value of life from the mean amplitude of

remote stress is recorded for different cases where the CV (coefficient of variation) of stress pdf was considered as being constant. The figure assesses the increase of the said mean life to failure in presence of higher CV of stress, as in this case rather low stresses are possible with a relatively high probability and they influence the rate of propagation in a higher measure than large ones.
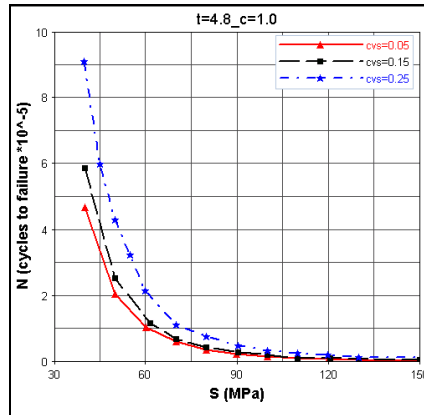
Fig. 5. Influence of the remote stress on the cycles to failure

In fig. 6 the influence of the initial geometry is examined for the case of a corner crack, considered to be elliptical in shape, with length c and depth a; a very interesting aspect of the consequences of a given shape is that for some cases the life for a through crack is longer than the one recorded for some deep corner ones; that case can be explained with the help of the plot of Fig. 7 where the growth of a through crack is compared with those of quarter corner cracks, recording times when a corner crack becomes a through one: as it is clarified in the boxes in the same picture, each point of the dashed curve references to a particular value of the initial depth.
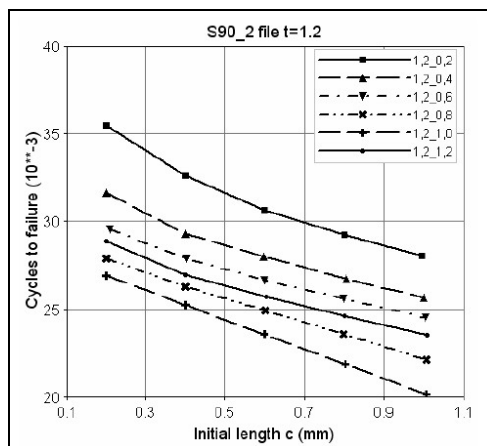
Fig. 6. Influence of the initial length of the crack on cycles to failure
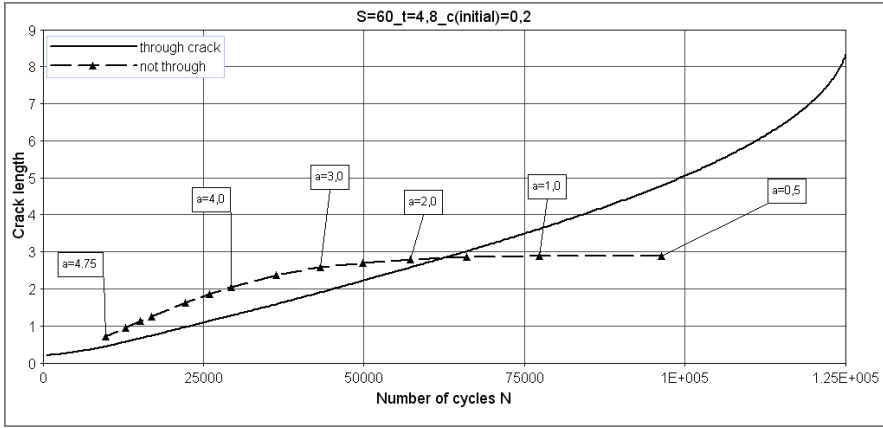
Fig. 7. Propagation behaviour of a corner and a through crack

It can be observed that beyond a certain value of the initial crack depth, depending on the sheet thickness, the length reached when the corner crack becomes a through one is larger than that obtained after the same number of cycles when starting with a through crack, and this effect is presumably connected to the bending effect of corner cracks.

For what concerns the influence exerted by the growth parameters, C and n according to the well known Paris' law, a first analysis was carried out in order to evaluate the influence of spatial randomness of propagation parameters; therefore the analysis was carried out considering that for each stage of propagation the current values of C and n were randomly extracted on the basis of a joint normal pdf between lnC and n. The results, illustrated in Fig. 8, show a strong resemblance with the well known experimental results by Wirkler.

Then an investigation was carried out about the influence of the same ruling parameters on the variance of cycles to failure. It could be shown that the mean value of the initial length has a little influence on the CV of cycles to failure, while on the contrary is largely affected by the CV of the said geometry. On the other hand, both statistical parameters of the distribution of remote stress have a deep influence on the CV of fatigue life.
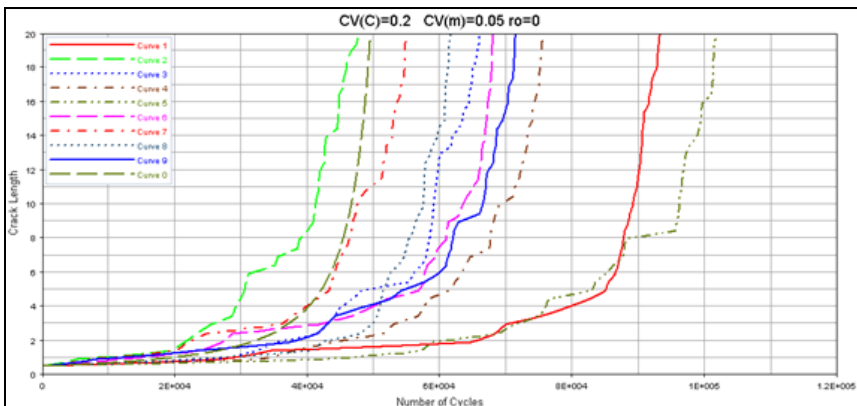
Fig. 8. Crack propagation histories with random parameters

Once the design variables were identified, the attention had to be focused on the type of structure that one wants to use as a reference; in the present case, a simple riveted lap joint for aeronautical application was chosen (fig. 9), composed by two 2024-T3 aluminium sheets, each 1 mm thick, with 3 rows of 10 columns of 5 mm rivets and a pitch of 25 mm. Several reasons suggest to analyze such a structure before beginning a really probabilistic study; for example, the state of stress induced into the component by external loads has to be evaluated and then it is important to know the interactions between existing singularities when a MSD (Multi-Site Damage) or even a WFD (Widespread Fatigue Damage) takes place. Several studies were carried out, in fact (for example, Horst, 2005), considering a probabilistic initiation of cracks followed by a deterministic propagation, on the basis that such a procedure can use very simple techniques, such as compounding (Rooke, 1986). Even if such a possibility is a very appealing one, as it is very fast, at least once the appropriate fundamental solutions have been found and recorded, some doubts arise when one comes to its feasibility.

The fundamental equation of compounding method is indeed as follows:

$$K = K^* + \sum \left(K_i - K^*\right) + K_e \tag{8}$$
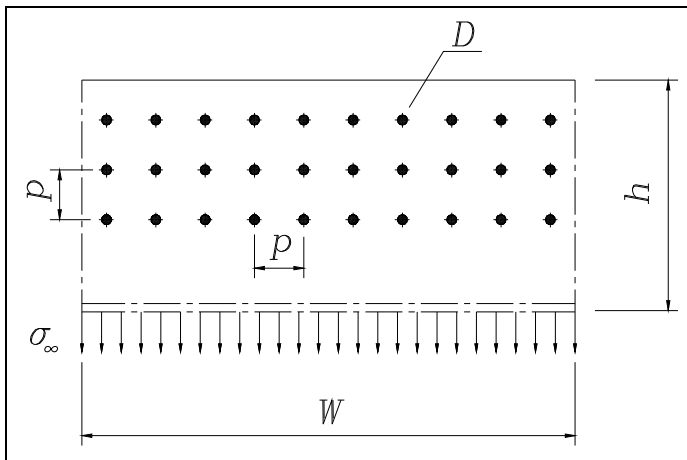
Fig. 9. The model used to study the aeronautical panel in WFD conditions

where the SIF at the crack tip of the crack we want to investigate is expressed by means of the SIF at the same location for the fundamental solution, K^*, plus the increase, with respect to the same 'fundamental' SIF, $(K_i - K^*)$, induced by each other singularity, taken one at a time, plus the effect of interactions between existing singularities, still expressed as a SIF, K_e. As the largest part of literature is related to the case of a few cracks, the K_e term is usually neglected, but that assumption appears to be too weak when dealing with WFD studies, where the singularities approach each other; therefore one of the main reasons to carry out such deterministic analysis is to verify the extent of this approximation. It must be stressed that no widely known result is available for the case of rivet-loaded holes, at least for cases matching with the object of the present analysis; even the most known papers, which we quoted above deal with the evaluation of SIF for cracks which initiate on the edge of a

loaded hole, but it is important to know the consequence of rivet load on cracks which arise elsewhere.

Another aspect, related to the previous one, is the analysis of the load carried by each pitch as damage propagates; as the compliance of partially cracked pitches increases with damage, one is inclined to guess that the mean load carried by those zones decreases, but the nonlinearity of stresses induced by geometrical singularities makes the quantitative measure of such a variation difficult to evaluate; what's more, the usual expression adopted for SIF comes from fundamental cases where just one singularity is present and it is given as a linear function of remote stress. One has to guess if such a reference variable as the stress at infinity is still meaningful in WFD cases.

Furthermore, starting to study the reference structure, an appealing idea to get a fast solution can be to decompose the structure in simple and similar details, each including one pitch, to be analyzed separately and then added together, considering each of them as a finite element or better as a finite strip; that idea induces to consider the problem of the interactions between adjacent details.

In fact, even if the structure is considered to be a two-dimensional one, the propagation of damage in different places brings the consequence of varying interactions, for both normal and shearing stresses. For all reasons above, an extensive analysis of the reference structure is to be carried out in presence of different MSD scenarios; in order to get fast solutions, use can be made of the well known BEASY® commercial code, but different cases are to be verified by means of more complex models.

On the basis of the said controls, a wide set of scenarios could be explored, with two, three and also four cracks existing at a time, using a two-dimensional DBEM model; in the present case, a 100 MPa remote stress was considered, which was transferred to the sheet through the rivets according to a 37%, 26% and 37% distribution of load, as it is usually accepted in literature; that load was applied through an opportune pressure distribution on the edge of each hole. This model, however, cannot take into account two effects, i.e. the limited compliance of holes, due to the presence of rivets and the variations of the load carried by rivets mounted in cracked holes; both those aspects, however, were considered as not very relevant, following the control runs carried out by FEM.
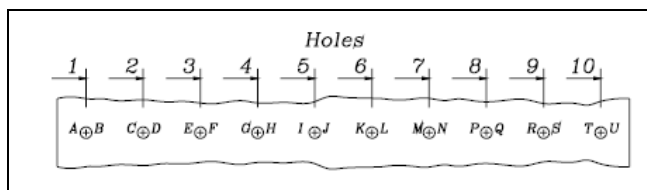
Fig. 10. The code used to represent WFD scenarios

For a better understanding of the following illustrations, one has to refer to fig. 10, where we show the code adopted to identify the cracks; each hole is numbered and each hole side is indicated by a capital letter, followed, if it is the case, by the crack length in mm; therefore, for example, E5J7P3 identifies the case when three cracks are present, the first, 5 mm long, being at the left side of the third hole (third pitch, considering sheet edges), another, 7 mm long, at the right side of the fifth hole (sixth pitch), and the last, 3 mm long, at the left side of the eight hole (eighth pitch).
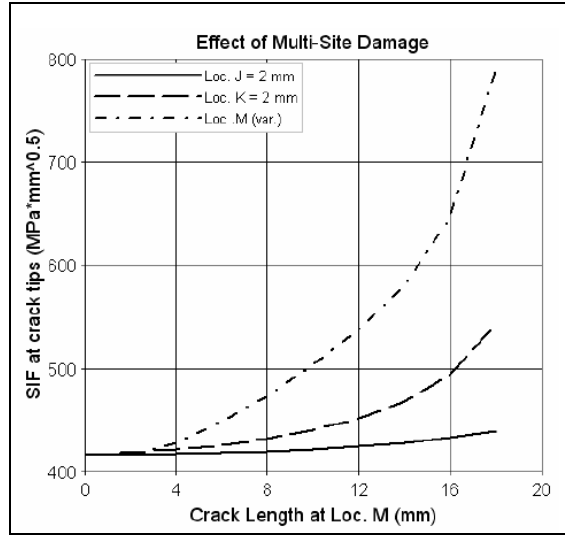
Fig. 11. Behaviour of J2K2Mx scenario
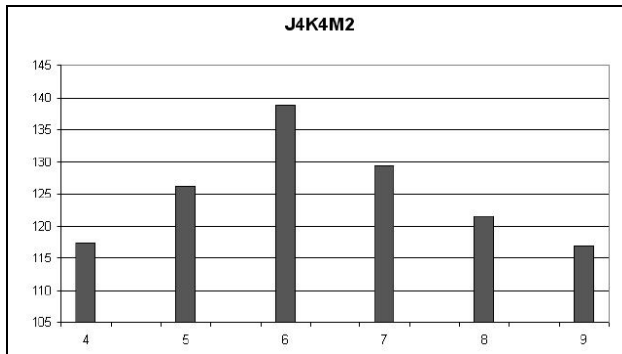
Fig. 12. Mean longitudinal stress loading different pitches for a 2 mm crack in pitch 7
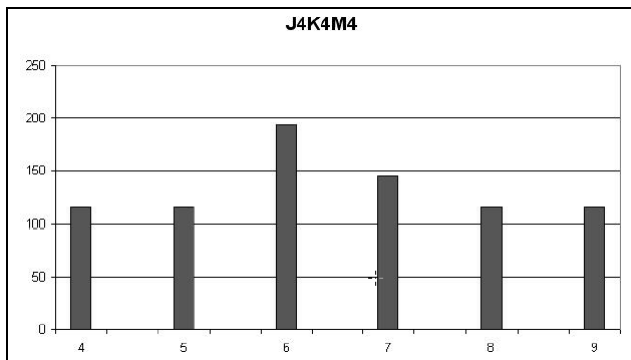
Fig. 13. Mean longitudinal stress loading different pitches for a 4 mm crack in pitch 7

In fig. 11 a three cracks scenario is represented, where in pitch 6 there are two cracks, each 2 mm long and another crack is growing at the right edge of the seventh hole, i.e. in the adjacent seventh pitch; if we consider only LEFM, we can observe that the leftmost crack (at location J) is not much influenced by the presence of the propagating crack at location M, while the central one exhibits an increase in SIF which can reach about 20%.
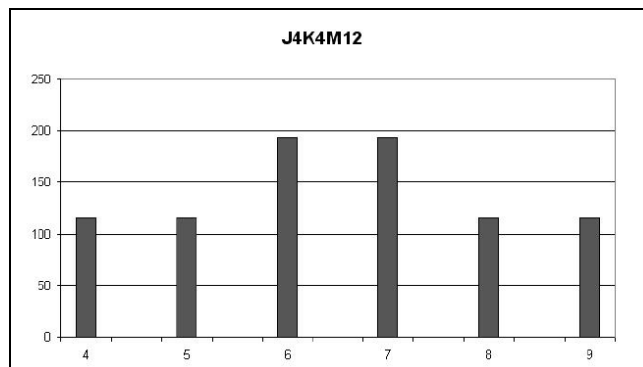
Fig. 14. Mean longitudinal stress loading different pitches for a 12 mm crack in pitch 7

The whole process can be observed by considering the mean longitudinal stress for different scenarios, as illustrated in Fig. 12, 13 and 14; in the first one, we can observe a progressive increase in the mean longitudinal stress around pitch no. 6, which is the most severely reduced and the influence of the small crack at location M is not very high.

As the length of crack in pitch 7 increases, however, the mean longitudinal stresses in both pitches 6 and 7 becomes quite similar and much higher of what is recorded in safe zones, where the same longitudinal stresses are not much increased in respect to what is recorded for a safe structure, because the transfer of load is distributed among many pitches.

The main results obtained through the previously discussed analysis can be summarized by observing that in complex scenarios high interactions exist between singularities and damaged zones, which can prevent the use of simple techniques such as compounding, but that the specific zone to be examined gets up to a single pitch beyond the cracked ones, of course on both sides. At the same time, as expected, we can observe that for WFD conditions, in presence of large cracks, the stress levels become so high that the use of LEFM can be made only from a qualitative standpoint.

If some knowledge about what to expect and how the coupled sheets will behave during the accumulation of damage has been obtained at this point of the analysis, we also realize, as pointed above, that no simple method can be used to evaluate the statistics of failure times, as different aspects will oppose and first of all the amount of the interactions between cracked holes; for that reason the only way which appears to be of some value is the direct M-C interaction as applied to the whole component, i.e. the evaluation of the 'true' history foe the sheets, to be performed the opportune number of times to extract reliable statistics; as the first problem the analyst has to overcome in such cases is the one related to the time consumption, it is of uttermost importance to use the most direct and quick techniques to obtain the desired results; for example, the use of DBEM coupled with an in-house developed code can give, if opportunely built, such guarantees.

In the version we are referring to, the structure was considered to be entirely safe at the beginning of each trial; then a damage process followed, which was considered as to be of Markow type. For the sake of brevity we shall not recall here the characters of such a process, which we consider to be widely known today; we simply mention that we have to define the initial scenario, the damage initiation criterion and the transitional probabilities for damage steps. In any case, we have to point out that other hypothesis could be assumed and first that of an initial damage state as related to EIFS (Equivalent Initial Flaw Size) or to the case of a rogue flaw, for example, don't imply any particular difficulty.

Two possible crack locations were considered at each hole, corresponding to the direction normal to the remote stress; the probability distribution of crack appearance in time was considered as lognormal, given by the following function:

$$f(N_i) = \frac{1}{\sigma_{ln} N_i \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(N_i) - \mu_{ln}}{\sigma_{ln}}\right)^2\right] \tag{10}$$

with an immediate meaning of the different parameters; it has to be noted that in our case the experimental results available in literature were adapted to obtain P-S-N curves, in order to make the statistics dependent on the stress level. At each time of the analysis the extraction of a random number for each of the still safe locations was carried out to represent the probability of damage cumulated locally and compared with the probability coming from eq. (10) above; in the positive case, a new crack was considered as initiated in the opportune location.

In order to save time, the code started to perform the search only at a time where the probability to find at least one cracked location was not less than a quantity p chosen by the user; it is well known that, if p_f is the probability of a given outcome, the probability that the same outcome is found at least for one among n cases happening simultaneously is given by:

$$p = 1 - (1 - p_f)^n ; \tag{11}$$

in our case n is the number of possible locations, thus obtaining the initial analysis time, by inverting the probability function corresponding to eq. (11) above; in our trials it was generally adopted p = 0.005, which revealed to be a conservative choice, but of course other values could also be accepted. A particular choice had also to be made about the kind and the geometry of the initial crack; it is evident that to follow the damage process accurately a defect as small as possible has to be considered, for example a fraction of mm, but in that case some difficulties arise.

For example, such a small crack would fall in the range of *short cracks* and would, therefore, require a different treatment in propagation; in order to limit our analysis to a two-dimensional case we had to consider a crack which was born as a through one and therefore we choose it to be characterized by a length equal to the thickness of the sheet, i.e., 1.0 mm in our case.

Our choice was also justified by the fact that generally the experimental tests used to define the statistics represented in eq. (10) above record the appearance of a crack when the defect reaches a given length or, if carried out on drilled specimens, even match the initiation and the failure times, considering that in such cases the propagation times are very short. Given

an opportune integration step, the same random extraction was performed in correspondence of still safe locations, up to the time (cycle) when all holes were cracked; those already initiated were considered as propagating defects, integrating Paris-Erdogan's law on the basis of SIF values recorded at the previous instant. Therefore, at each step the code looked for still safe locations, where it performed the random extraction to verify the possible initiation of defect, and at the same time, when it met a cracked location, it looked for the SIF value recorded in the previous step and, considering it as constant in the step, carried out the integration of the growth law in order to obtain the new defect length.

The core of the analysis was the coupling of the code with a DBEM module, which in our case was the commercial code BEASY®; a reference input file, representing the safe structure, was prepared by the user and submitted to the code, which analyzed the file, interpreted it and defined the possible crack locations; then, after completing the evaluations needed at the particular step, it would build a new file which contained the same structure, but as damaged as it came from the current analysis and it submitted it to BEASY®; once the DBEM run was carried out, the code read the output files, extracted the SIF values pertaining to each location and performed a new evaluation. For each ligament the analysis ended when the distance between two singularities was smaller than the plastic radius, as given by Irwin

$$r_p = \frac{K_I^2}{\pi \sigma_y^2} \tag{11}$$

where σ_y is the yield stress and K_I the mode-I SIF; that measure is adopted for cracks approaching a hole or an edge, while for the case of two concurrent cracks the limit distance is considered to be given by the sum of the plastic radiuses pertaining to the two defects. Once such limit distance was reached, the ligament was considered as broken, in the sense that no larger cracks could be formed; however, to take into account the capability of the ligament to still carry some load, even in the plastic field, the same net section was still considered in the following steps, thus renouncing to take into account the plastic behaviour of the material. Therefore, the generic M-C trial was considered as ended when one of three conditions are verified, the first being the easiest, i.e. when a limit number of cycles given by the user was reached. The second possibility was that the mean longitudinal stress evaluated in the residual net section reached the yield stress of the material and the third, obviously, was met when all ligaments were broken. Several topics are to be further specified and first of all the probabilistic capabilities of the code, which are not limited to the initiation step. The extent of the probabilistic analysis can be defined by the user, but in the general case, it refers to both loading and propagation parameters.

For the latter, user inputs the statistics of the parameters, considering a joint normal density which couples lnC and n, with a normal marginal distribution for the second parameter; at each propagation step the code extracted at each location new values to be used in the integration of the growth law.

The variation of remote stress was performed in the same way, but it was of greater consequences; first of all we have to mention that a new value of remote stress was extracted at the beginning of each step from the statistical distribution that, for the time being, we considered as a normal one, and then kept constant during the whole step: therefore, variations which occurred for shorter times went unaccounted. The problem which was met when dealing with a variable load concerned the probability of crack initiation, more than

the propagation phase; that's because the variation of stress implies the use of some damage accumulation algorithm, which we used in the linear form of Miner's law, being the most used one.
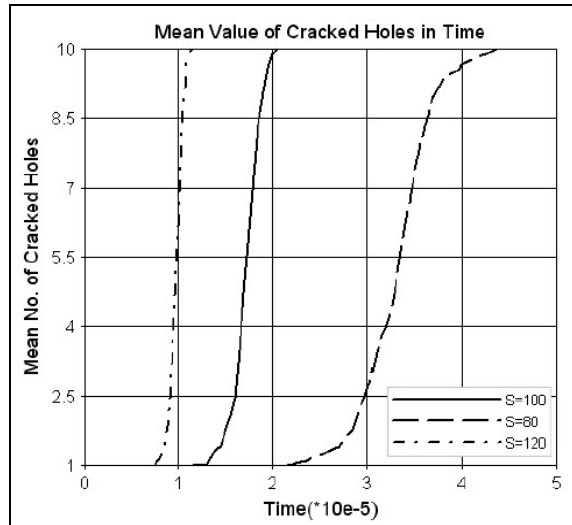
Fig. 15. Cdfs' for a given number of cracked holes in time

However, we have to observe that if the number of cycles to crack initiation is a random variable, as we considered above, the simple sum of deterministic ratios which appears in Miner's law cannot be accepted, as pointed out by Hashin (1980; 1983), the same sum having a probabilistic meaning; therefore, the sum of two random variables, i.e. the damage cumulated and the one corresponding to the next step, has to be carried out by performing the convolution of the two pdfs' involved. This task is carried out by the code, in the present version, by a rather crude technique, recording in a file both the damage cumulated at each location and the new one and then performing the integration by the trapezoidal rule.

At the end of all M-C trials, a final part of our code carried out the statistical analysis of results in such a way as to be dedicated to the kind of problem in hand and to give useful results; for example, we could obtain, as usually, the statistics of initiation and failure times, but also the cumulative density function (cdf) of particular scenarios, as that of cracks longer than a given size, or including an assigned number of holes, as it is illustrated in fig. 15.

4. Multivariate optimization of structures and design

The aim of the previous discussion was the evaluation of the probability of failure of a given structure, with assigned statistics of all the design variables involved, but that is just one of the many aspects which can be dealt within a random analysis of a structural design. In many cases, in fact, one is interested to the combined effects of input variables on some kind of answer or quality of the resulting product, which can be defined as weight, inertia, stiffness, cost, or others; sometimes one wishes to optimize one or several properties of the result, either maximizing or minimizing them, and different parameters can give to the

design opposing tendencies, as it happens for example when one wishes to increase some stiffness of the designed component, while keeping its weight as low as possible.
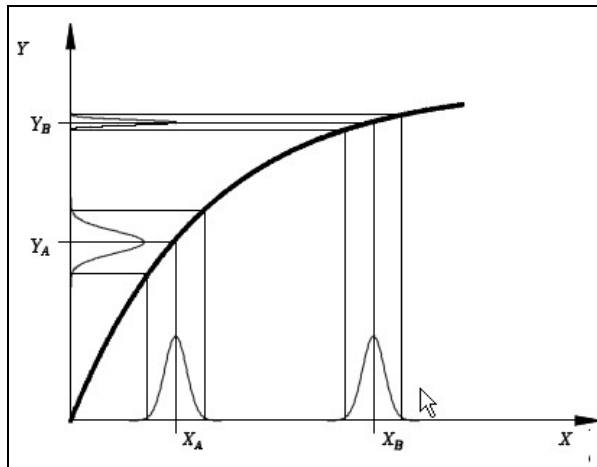
Fig. 16. How the statistics of the result depend on the mean value of the control variables

In any case, one must consider that, at least in the structural field for the case of large deformations, the relationship between the statistic of the response and that of a generic design variable for a complex structure is in general a non-linear one; it is in fact evident from fig. 16 that two different mean values for the random variable x, say x_A and x_B, even in presence of the same standard variation, correspond to responses centered in y_A and y_B, whose coefficients of variation are certainly very different from each other. In those cases, one has to expect that small variations of input can imply large differences for output characteristics, in dependence of the value around which input is centered; that aspect is of relevant importance in all those cases where one has to take into account the influences exerted by manufacturing processes and by the settings of the many input parameters (control variables), as they can give results which mismatch with the prescribed requirements, if not themselves wrong.

Two are the noteworthy cases, among others, i.e. that were one wish to obtain a given result with the largest probability, for example to limit scraps, and the other, where one wishes to obtain a design, which is called 'robust', whose sensitivity to the statistics of control variables is as little as possible.

Usually, that problem can be solved for simple cases by assigning the coefficients of variation of the design variables and looking for the corresponding mean values such as to attain the required result; the above mentioned hypothesis referring to the constancy of the coefficients of variation is usually justified with the connection between variance and quality levels of the production equipments, not to mention the effect of the nowadays probabilistic techniques, which let introduce just one unknown in correspondence of each variable.

Consequently, while in the usual probabilistic problem we are looking for the consequences on the realization of a product arising from the assumption of certain distributions of the design variables, in the theory of optimization and robust design the procedure is reversed,

as we now look for those statistical parameters of the design variables such as to produce an assigned result (target), characterized by a given probability of failure.

It must be considered, however, that no hypothesis can be introduced about the uniqueness of the result, in the sense that more than one design can exist such as to satisfy the assigned probability, and that the result depends on the starting point of the analysis, which is a well known problem also in other cases of probabilistic analysis. Therefore, the most useful way to proceed is to define the target as a function of a given design solution, for example of the result of a deterministic procedure, in order to obtain a feasible or convenient solution.

The main point of multi-objective optimization is the search for the so-called Pareto-set solutions; one starts looking for all feasible solutions, those which don't violate any constraint, and then compare them; in this way, solutions can be classified in two groups, i.e. the dominating ones, which are better than the others for all targets (the 'dominated' solutions) and which are non-dominating among each other. In other words, the Pareto-set is composed by all feasible solutions which are non-dominating each other, i.e. which are not better for at least one condition, while they are all better than the dominated solutions.

As it is clear from above, the search for Pareto-set is just a generalization of the optimization problem and therefore a procedure whatever of the many available ones can be used; for example, genetic algorithm search can be conveniently adopted, even if in a very general way (for example, MOGA, 'Multi-Objective Genetic Algorithm' and all derived kinds), coupled with some comparison technique; it is evident that this procedure can be used at first in a deterministic field, but, if we apply at each search a probabilistic sense, i.e. if we say that the obtained solution has to be a dominating one with a given probability of success (or, in reverse, of failure) we can translate the same problem in a random sense; of course, one has to take into account the large increase of solutions to be obtained in such a way as to build a statistic for each case to evaluate the required probability.

In any case, at the end of the aforesaid procedure one has a number of non-dominating solutions, among which the 'best' one is hopefully included and therefore one has to match against the problem of choosing among them. That is the subject of a 'decision making' procedure, for which several techniques exist, none of them being of general use; the basic procedure is to rank the solutions according to some principle which is formulated by the user, for example setting a 'goal' and evaluating the distance from each solution, to end choosing that whose distance is a minimum. The different commercial codes (for example, Mode-Frontier is well known among such codes) usually have some internal routines for managing decisions, where one can choose among different criteria.

More or less, the same procedure which we have just introduced can be used to obtain a design which exhibits an assigned probability of failure (i.e. of mismatching the required properties) by means of a correct choice of the mean values of the control variables. This problem can be effectively dealt with by an SDI (Stochastic Design Improvement) process, which is carried out through an convenient number of MC (here called runs) as well as of the analysis of the intermediate results. In fact, input - i.e. design variables x - and output - i.e. target y - of an engineering system can be connected by means of a functional relation of the type

$$y = F(x_1, x_2, \ldots, x_n) \tag{12}$$

which in the largest part of the applications cannot be defined analytically, but only rather ideally deduced because of the its complex nature; in practice, it can be obtained by

considering a sample x_i and examining the response y_i, which can be carried out by a simulation procedure and first of all by one of M-C techniques, as recalled above. Considering a whole set of M-C samples, the output can be expressed by a linearized Taylor expansion centered about the mean values of the control variables, as

$$y_j = F\left(\mu_{x_i}\right) + \sum \frac{dF}{dx_i}\left(x_i - \mu_{x_i}\right) = \mu_{y_j} + G\left\{x_i - \mu_{x_i}\right\} \tag{13}$$

where μ_i represents the vector of mean values of input/output variables and where the gradient matrix G can be obtained numerically, carrying out a multivariate regression of y on the x sets obtained by M-C sampling. If y_0 is the required target, we can find the new x_0 values inverting the relation above, i.e. by

$$x_0 = \mu_x + G^{-1}\left\{y_0 - \mu_{y_j}\right\}; \tag{14}$$

as we are dealing with probabilities, the real target is the mean value of the output, which we compare with the mean value of the input, and, considering that, as we shall illustrate below, the procedure will evolve by an iterative technique, it can be stated that the relation above has to be modified as follows, considering the update between the k-th and the (k+1)-th step:

$$\mu_{x_0} = \mu_{x,k+1} = \mu_{x,k} + G^{-1}\left(\mu_{y,k+1} - \mu_{y,k}\right) = \mu_{x,k} + G^{-1}\left(\mu_{y_0} - \mu_{y,k}\right). \tag{15}$$

The SDI technique is based on the assumption that the cloud of points corresponding to the results obtained from a set of MC trials can be moved toward a desired position in the N-dimensional space such as to give the desired result (target) and that the amplitude of the required displacement can be forecast through a close analysis of the points which are in the same cloud (fig 17): in effects, it is assumed that the shape and size of the cloud don't change greatly if the displacement is small enough; it is therefore immediate to realize that an SDI process is composed by several sets of MC trials (runs) with intermediate estimates of the required displacement.
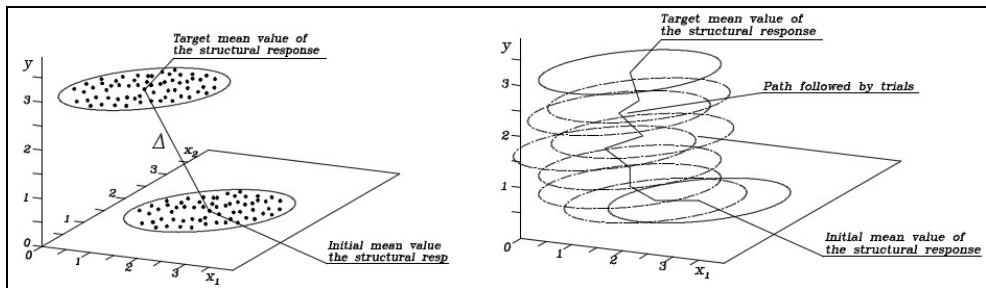
Fig. 17. The principles of SDI processes

It is also clear that the assumption about the invariance of the cloud can be kept just in order to carry out the multivariate regression which is needed to perform a new step - i.e. the

evaluation of the G matrix - but that subsequently a new and correct evaluation of the cloud is needed; in order to save time, the same evaluation can be carried out every k steps, but of course, as k increases, the step amplitude has to be correspondently decreased. It is also immediate that the displacement is obtained by changing the statistics of the design variables and in particular by changing their mean (nominal) values, as in the now available version of the method all distributions are assumed to be uniform, in order to avoid the gathering of results around the mode value. It is also to be pointed out that sometimes the process fails to accomplish its task because of the existing physical limits, but in any case SDI allows to quickly appreciate the feasibility of a specific design, therefore making easier its improvement.

Of course, it may happen that other stochastic variables are present in the problem (the so called background variables): they can be characterized by any type of statistical distribution included in the code library, but they are not modified during the process. Therefore, the SDI process is quite different for example from the classical design optimization, where the designer tries to minimize a given objective function with no previous knowledge of the minimum value, at least in the step of the problem formulation. On the contrary, in the case of the SDI process, it is first stated what is the value that the objective function has to reach, i.e. its target value, according to a particular criterion which can be expressed in terms of maximum displacement, maximum stress, or other. The SDI process gives information about the possibility to reach the objective within the physical limits of the problem and determines which values the project variables must have in order to get it. In other words, the designer specifies the value that an assigned output variable has to reach and the SDI process determines those values of the project variables which ensure that the objective variable becomes equal to the target in the mean sense. Therefore, according to the requirements of the problem, the user defines a set of variables as control variables, which are then characterized from a uniform statistical distribution (natural variability) within which the procedure can let them vary, while observing the corresponding physical (engineering) limits. In the case of a single output variable, the procedure evaluates the Euclidean or Mahalanobis distance of the objective variable from the target after each trial:

$$d_i = \left| y_i - y^* \right| \qquad i = 1, 2, \ldots, N \tag{16}$$

where y_i is the value of the objective variable obtained from the i-th iteration, y^* is the target value and N is the number of trials per run. Then, it is possible to find among the worked trials that one for which the said distance gets the smallest value and subsequently the procedure redefines each project variable according to a new uniform distribution with a mean value equal to that used in such "best" trial. The limits of natural variability are accordingly moved of the same quantity of the mean in such way as to save the amplitude of the physical variability.

If the target is defined by a set of output variables, the displacement toward the condition where each one has a desired (target) value is carried out considering the distance as expressed by:

$$d_i = \sqrt{\Sigma_k \left(y_{i,k} - y_k^* \right)^2} \, , \tag{17}$$

where k represents the generic output variable. If the variables are dimensionally different it is advisable to use a normalized expression of the Euclidean distance:

$$d_i = \sqrt{\sum \omega_k \left(\delta_{i,k} \right)^2} \, , \tag{18}$$

where:

$$\delta_{i,k} = \begin{cases} \dfrac{y_{i,k}}{y_k^*} - 1, & \text{if } y_k^* \neq 0 \\ y_{i,k} & \text{if } y_k^* = 0 \end{cases} \tag{19}$$

but in this case it is of course essential to assign weight factors ω_k to define the relative importance of each variable. Several variations of the basic procedures are available; for example, it is possible to define the target by means of a function which implies an equality or even an inequality; in the latter case the distance is to be considered null if the inequality is satisfied. Once the project variables have been redefined a new run is performed and the process restarts up to the completion of the assigned number of shots. It is possible to plan a criterion of arrest in such way as to make the analysis stop when the distance from the target reaches a given value. In the most cases, it is desirable to control the state of the analysis with a real-time monitoring with the purpose to realize if a satisfactory condition has been obtained.

5. Examples of multivariate optimization

5.1 Study of a riveting operation

The first example we are to illustrate is about the study of a riveting operation; in that case we tried to maximize the residual compression load between the sheets (or, what is the same, the traction load in the stem of the rivet) while keeping the radial stress acting on the wall of the hole as low as possible; the relevant parameters adopted to work out this example are recorded in Tab. 1.

RGR	Hole Radius	Variable	mm	2.030	2.055
RSTEM	Shank Radius	Variable	mm	1.970	2.020
LGR	Shank Length	Variable	mm	7.600	8.400
AVZ	Hammer Stroke	Variable	mm	3.500	4.500
EYG	Young Modulus	Variable	MPa	65,000	75,000
THK	Sheets Thickness	Constant	mm	1.000	
SIZ	Yield Stress	Constant	MPa	215.000	
VLZ	Hammer Speed	Constant	mm/sec	250.000	

Table 1. Relevant parameters for riveting optimization

It is to be said that in this example no relevant result was obtained, because of the ranges of variation of the different parameters were very narrow, but in any case it can be useful to quote it, as it defines a procedure path which is quite general and which shows very clearly the different steps we had to follow. The commercial code used was Mode-Frontier®, which is now very often adopted in the field of multi-objective optimization; that code let the user build his own problem with a logic procedure which makes use of icons, each of them corresponding to a variable or to a step of the procedure, through which the user can readily build his problem as well as the chosen technique of solution; for example, with reference to the table above, in our case the logic tree was that illustrated in fig. 18.
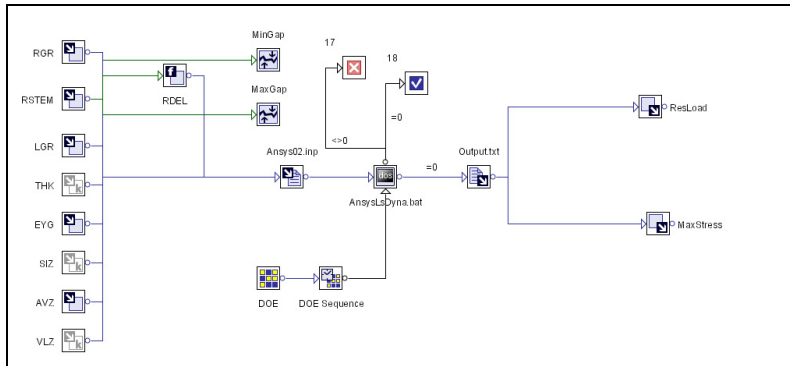
Fig. 18. The building of the problem in Mode-Frontier environment

Summarizing the procedure, after defining all variables and parameters, the work can be set to be run by means of an user-defined script (AnsysLsDyna.bat in fig. 18), in such a way that the code knows that the current values of variables and parameters are to be found somewhere (in Ansys02.inp), to be worked somehow, for example according to a DOE procedure or to a genetic algorithm or other, and that the relevant results will be saved in another file (in Output.txt in our case); those results are to be compared with all the previously obtained ones in order to get the stationary values of interest (in our case, the largest residual load and the smallest residual stress).

The kernel of the procedure, of course, is stored in the script, where the code finds how to pass from input data to output results; in our case, the input values were embedded in an input file for Ansys® preprocessor, which would built a file to be worked by Ls-Dyna® to simulate the riveting operation; as there was no correct correspondence between those two codes, a home-made routine was called to match requirements; another home-made routine would then extract the results of interest from the output files of Ls-Dyna®.

A first pass from Mode-Frontier® was thus carried out, in such a way as to perform a simple 3-levels DOE analysis of the problem; a second task which was asked from the code was to build the response surface of the problem; there was no theoretical reason to behave in such a way, but it was adopted just to spare time, as each Ls-Dyna trial was very time-expensive, if compared with the use of RS: therefore the final results were 'virtual', in the sense that they didn't came from the workout of the real problem, but from its approximate analytic representation.
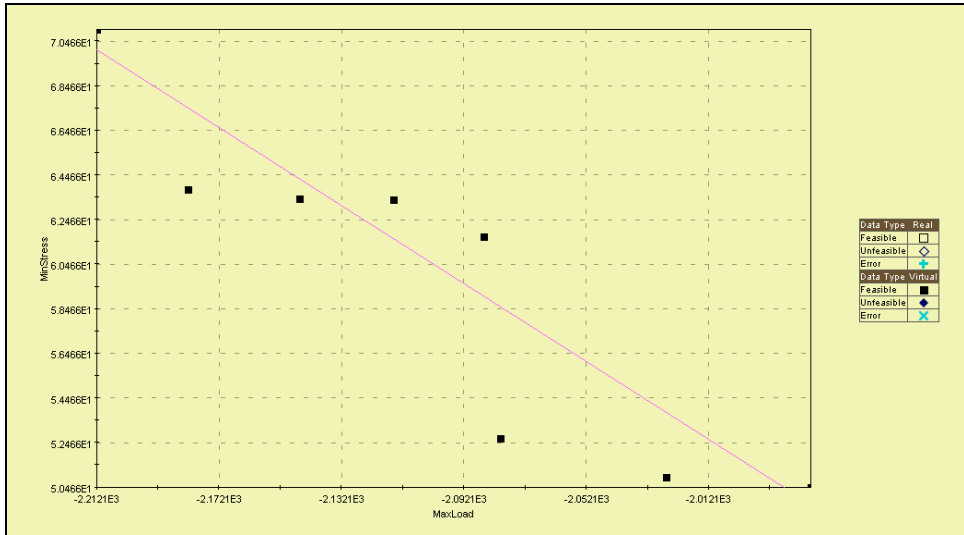
Fig. 19. Pareto-set for the riveting problem

Thus, the Pareto-set for the riveting problem was obtained, as it is shown in fig. 19; it must be realized that the number of useful non dominated results was much larger than it can be shown in the same picture, but, because of the narrow ranges of variance, they overlap and don't appear as distinct points.

The last step was the choice of the most interesting result, which was carried out by means of the Decision Manager routine, which is also a part of Mode-Frontier code.

5.2 The design improvement of a stiffened panel

As a second example we show how a home-made procedure, based on the SDI technique, was used to perform a preliminary robust design of a complex structural component; this procedure is illustrated with reference to the case of a stiffened aeronautical panel, whose residual strength in presence of cracks had to be improved. Numerical results on the reference component had been validated by using experimental results from literature.

To demonstrate the procedure described in the previous section, a stiffened panel constituted by a skin made of Al alloy 2024 T3, divided in three bays by four stiffeners made of Al alloy 7075 T5 (E = 67000 MPa, σ_y = 525 MPa, σ_u = 579 MPa, δ_{ult} = 16%) was considered. The longitudinal size of the panel was 1830 mm, its transversal width 1190 mm, the stringer pitch 340 mm and the nominal thickness 1.27 mm; the stiffeners were 2.06 mm high and 45 mm wide. Each stiffener was connected to the skin by two rows of rivets 4.0 mm diameter.

A finite element model constituted by 8-noded solid elements had been previously developed and analyzed by using the WARP 3D® finite element code. The propagation of two cracks, with the initial lengths of 120 mm and 150 mm respectively, had been simulated by considering the Gurson-Tveergard model, as implemented in the same code, whose parameters were calibrated.
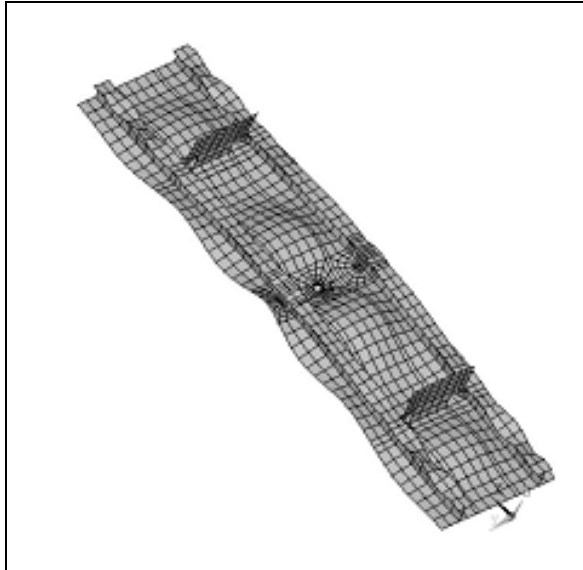
Fig. 20. The model of the stiffened panel

In the proposed application of the SDI procedure, a substructure of the panel was considered (fig. 20), corresponding to its single central bay (the part of the panel within the two central stringers) with a fixed width equal to 680 mm, where a central through-crack was assumed to exist, with an initial length of 20 mm. The pitch between the two stringers and their heights were considered as design variables. As natural variables the stringers pitch (±10.0 range) and the stringers height (±0.4 mm range) were assumed, while the engineering intervals of the variables was considered to be respectively [306 ÷374 mm] and [1.03 ÷ 3.09 mm]. An increment of the maximum value of the residual strength curve (Rmax) of the 10 %, with a success probability greater than 0.80, was assumed as the target.
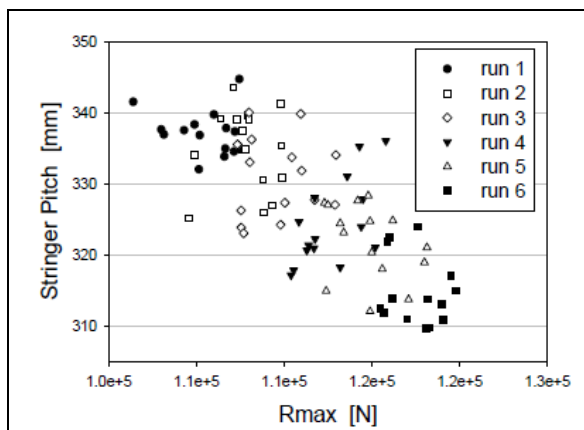
Fig. 21. Path of clouds for R_{max} as a function of the stringer pitch
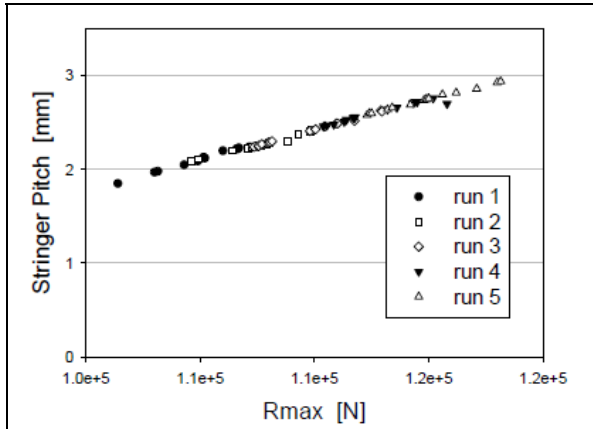
Fig. 22. Stringer pitch vs. Target
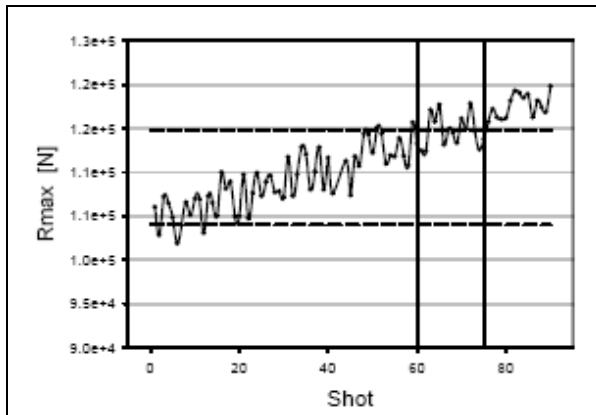
Fig 23. Target vs. shot per run
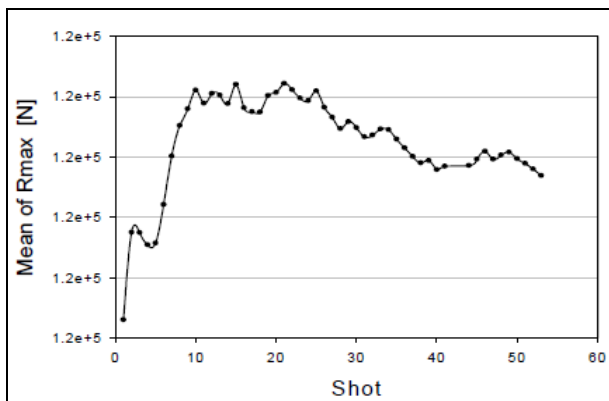
Fig. 24. Mean value of target vs. shot
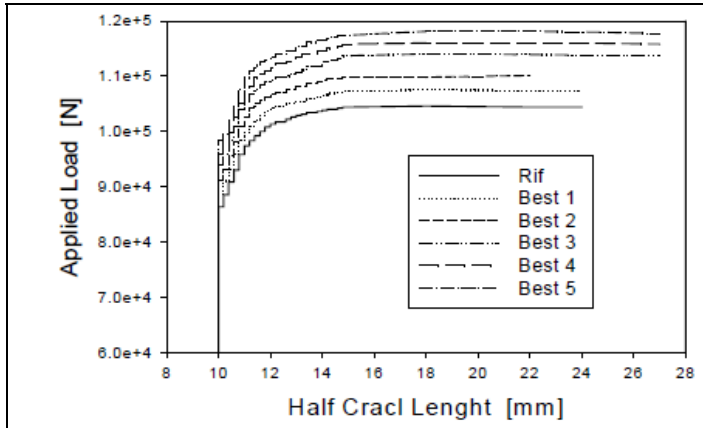
Fig. 25. R-curves obtained for the best shot of each run
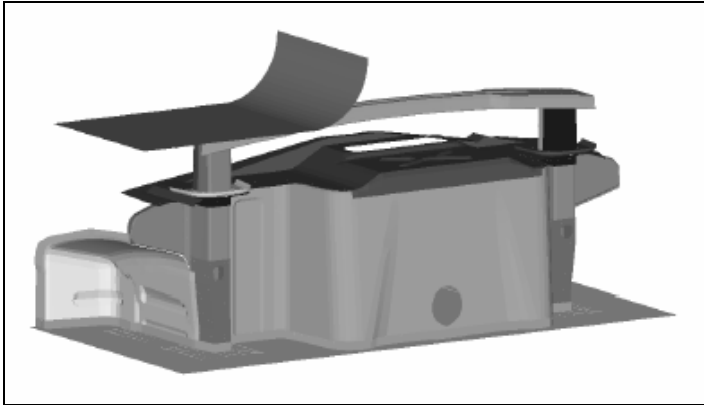
Fig. 26. The global FEM model of the absorber

A total of 6 runs, each one of 15 shots, were considered adequate to satisfy the target, even if at the end of the procedure an extended MC had to be performed in order to assess the obtained results from the 15 shots of the last satisfying run. In the following fig. 21 and 22 the design variables vs. the maximum value of the residual strength are illustrated. In correspondence to these two plots we recorded in fig. 23 the values assumed by the maximum value of the R-curve for each shot. In the same figure the reference value (obtained by considering the initial nominal value of the design variables) of the maximum value of the R-curve is reported together with the target value (dashed line). As it is possible to observe, 9 of the 15 shots of the 5th run overcame the target value; it means that by using the corresponding mean value of the design variable the probability to satisfy the target is of about 0.60.

Therefore, another run (the 6th) was carried out and just one shot didn't overcome the target value, so that the approximate probability to satisfy the target is about 0.93. The mean values of the design variables in the 6th run were respectively 318.8 mm for the stringer pitch and 2.89 mm for the stinger height; the mean value of the output variable was 116000

N. An extended MC (55 trials) was performed on the basis of the statistics of the 6th run and the results showed in the fig. 24 were obtained, where the mean value of the residual strength vs. the number of the trial has been recorded. The new mean of the output variable was 117000 N with a standard deviation of 1800 N and the probability to satisfy the target was exactly 0.80. At the end, in fig. 25, the six R-curves corresponding to the six best shots for each run are reported, together with the reference R-curve.

5.3 Optimization of an impact absorber

As a last example, the SDI procedure was applied to reduce the maximum value of the displacement in time of a rigid barrier that impacted the rear substructure of a vehicle (fig. 26) in a crash test. The reasons which lie behind such a choice are to be found in the increasing interest in numerical analysis of crashworthiness of vehicles because of the more strict regulations concerning the protection of the occupants and related fields. In Europe the present standards to be used in homologation of vehicles are more or less derived by U.S. Code of Federal Regulations, CFR-49.571, but ever-increasing applications are done with reference to other standards, and first of all to EURONCAP. The use of such standards, who are mainly directed to limit biomechanical consequences of the impact - which are controlled by referring the results to standard indexes related to different parts of human body - implies that, besides the introduction of specific safety appliances, as safety belts and airbags, the main strength of car body has to be located in the cell which holds passengers, in order to obtain a sufficient survival volume for the occupants; the other parts of the vehicle are only subsidiary ones, because of the presence of absorbers which reduce the impact energy which is released on the cell.

We can add to all previous considerations that the present case study was adopted as it is well known that vehicle components come from mass production, where geometrical imperfections are to be expected as well as a certain scatter of the mechanical properties of the used materials; therefore, it can't be avoided that the said variations induce some differences of response among otherwise similar components, what can be relevant in particular cases and first of all in impact conditions; the analysis which we carried out was directed to define, through the use of the previously sketched procedure, the general criteria and the methodology required to develop a robust design of those vehicle components which are directly used to limit plastic deformations in impact (impact absorber). In our particular case, the study was carried out with reference to the mentioned substructure, whose effective behaviour in impact (hammer) conditions is known and is associated to those deterministic nominal values of the design variable actually in use, with the immediate objective to obtain a reduction of the longitudinal deformation of the impact absorber.

The substructure is a part of a rear frame of a vehicle, complete with cross-bar and girders, where impact absorbers are inserted; the group is acted upon by a hammer which is constrained to move in the longitudinal direction of the vehicle with an initial impact speed of 16 km/h; the FE model used for the structure consisted of about 23400 nodes and about 21900 elements of beam, shell and contact type, while the hammer was modelled as a "rigid wall". The thicknesses of the internal and external C-shaped plates of the impact absorbers were selected as project variables, with a uniform statistical distribution in the interval [1.7mm±1.9mm]; lower and upper engineering limits were respectively 1.5 mm and 2.1 mm.

This choice was carried out by preliminary performing, by using the probabilistic module of ANSYS® ver 8.0 linked to the explicit FE module of LS-Dyna® included in the same code, a sensitivity analysis of an opportune set of design variables on the objective variable, which is, as already stated before, the maximum displacement of the hammer.

As design variables to be involved in the sensitivity analysis we chose, besides the inner and outer thicknesses of the C-shaped profile of the impact absorbers, the mechanical properties of the three materials constituting the main components of the substructure (the unique young modulus and the three yielding stresses); it was clear from the obtained results that while the relationship existing between the thicknesses of the inner and outer C-shaped profile of the impact absorber and the objective variable is quite linear, as well as the relationship between the yielding stress of the impact absorber material and the same objective variable, a relationship between the other considered variables and the objective variable is undetermined.

Variable	Type	Distribution	Natural variability	Physical limits
Internal plate thick.	Design var.	uniform	1.7-1.9	1.5-2.1
External plate thick.	Design var.	uniform	1.7-1.9	1.5-2.1
material scale factor DC04 strain-rate 0	independent var.	uniform	0.95-1.05	
material scale factor DC04 strain-rate 0.005	dependent var.			
material scale factor DC04 strain-rate 0.05	dependent var.			
material scale factor DC04 strain-rate 0.5	dependent var.			
material scale factor DP600 strain-rate 0	independent var.	uniform	0.95-1.05	
material scale factor DP600 strain-rate 0.005	dependent var.			
material scale factor DP600 strain-rate 0.05	dependent var.			
material scale factor DP600 strain-rate 0.5	dependent var.			
material scale factor S355NC strain-rate 0	independent var.	uniform	0.95-1.05	
material scale factor S355NC strain-rate 0.03	dependent var.			
material scale factor S355NC strain-rate 0.5	dependent var.			

Table 2. The properties of the variables used in the absorber case

It was also quite evident that the only design variables which influence the objective one were the mechanical properties of the material of the impact absorber and the thicknesses of its profiles. A preliminary deterministic run, carried out with the actual design data of the structure gave for the objective variable a 95.94 mm "nominal" value, which was reached after a time of 38.6 ms from the beginning of the impact. The purpose of SDI in our case was

assumed the reduction of that displacement by 10% with respect to this nominal value and therefore an 86.35 mm target value was assigned.

The mechanical properties of the three materials constituting the absorbers and the rear crossbar of the examined substructure were also considered as random; it was assumed that their characteristic stress-strain curves could vary according to a uniform law within 0.5% of the nominal value. This was made possible by introducing a scale factor for the characteristic curves of the materials, which were considered as uniformly distributed in the interval [0.95,1.05].

Moreover, four stress-strain curves were considered for each material, corresponding to as many specific values of the strain-rate. The relationship among those curves and the static one was represented, according to the method used in Ls-Dyna®, by means of a scale factor which let us pass from one curve to another as a function of the strain-rate; also those factors were assumed to be dependent on that applied to the static curve, in order to avoid possible overlapping.

Therefore, the simulation involved 14 random variables, among which only 2 were considered as project variables; in the following Tab. 2 the properties of all the variables are listed. To work out the present case, the commercial well known St-Orm® code was used coupled with the Ls-Dyna® explicit solver for each deterministic FEM analysis and ran on a 2600 MHz bi-processor PC, equipped with a 2 Gb RAM; SDI processing required 9 runs with 25 shots each, with a total of 225 MC trials, and the time required to complete a single LS-Dyna® simulation being of about 2 hours.

As we already pointed out, the stochastic procedure develops through an MC iterative process where the input variables are redefined in each trial in such a way as to move the results toward an assigned target; therefore, we need first to assess what we mean as an attained target.

After every run the statistics of the output variables could be obtained, as well as the number of times that the target was reached, which could be considered as the probability of attainment of the target for the particular design, expressed through the mean values of the input variables. It is noteworthy to specify that these data are only indicative, because the MC procedure developed within a single set of trials is not able to give convergent results due to the low number of iterations.

Therefore, considering the procedure as ended when a run was obtained where all trials gave the desired target value, it was opportune to perform a final MC convergent process to evaluate the extent by which the target had been indeed reached, using the statistical distributions of the variables of the last run. For the same reason, a real-time monitoring could induce the designer to stop the procedure even if not all trials - but "almost" all - of the same run give the target as reached, also to comply with production standard and procedures. As we already pointed out in the previous paragraphs, the stochastic procedure develops through an MC iterative process where the input variables are redefined in each trial in such a way as to move the results toward an assigned target; therefore, we needed first to assess what we meant as an attained target.

For what concerns our case study, the detailed data for every run are recorded in Tab. 3; if compared with the first run, the mean of the displacement distribution in the 9th run is reduced of 8.6% and 23/25 shots respect the target: therefore, the results of the 9th run may be considered as acceptable.

	Thickness of the internal sheet [mm]			Thickness of the external sheet [mm]			Displacement [mm]		Distance from target	
run	left bound	mean value	right bound	left bound	mean value	right bound	mean value	std	mean value	std
1	1.7000	1.8000	1.9000	1.7000	1.8000	1.9000	95.9524	2.4504	0.0585	0.0270
2	1.7903	1.8903	1.9903	1.7909	1.8909	1.9909	91.5568	2.3917	0.0158	0.0190
3	1.7127	1.8127	1.9127	1.8322	1.9322	2.0322	92.5833	2.4903	0.0240	0.0232
4	1.7297	1.8297	1.9297	1.8795	1.9795	2.0795	91.0362	2.3111	0.0132	0.0144
5	1.7624	1.8624	1.9624	1.9475	2.0238	2.1000	88.8615	2.1464	0.0026	0.0061
6	1.7632	1.8632	1.9632	2.0000	2.0500	2.1000	88.0821	1.6937	0.0005	0.0016
7	1.7446	1.8446	1.9447	1.9259	2.01295	2.1	89.6456	2.2662	0.005384	0.01072
8	1.7831	1.8831	1.9831	1.9385	2.0193	2.1	88.5974	2.1700	0.002176	0.00664
9	1.8778	1.9778	2.0778	**1.8846**	**1.9846**	**2.0846**	87.6936	2.1465	0.0009655	0.00367

Table 3. Values of control variables in the different runs
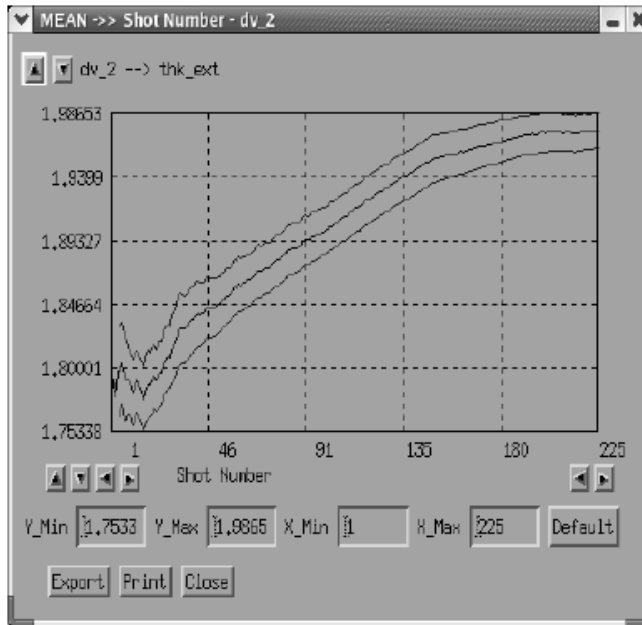
Fig. 27. Thickness of internal plate vs. shot

In the plots of fig. 27 and fig. 28 the values of the thickness of the internal and external plates of the impact absorber versus the current number of shots have been illustrated. The variable which was subjected to the largest modifications in the SDI procedure was the thickness of the external plate of the impact absorber and in fact from an analysis of sensitivity it resulted to influence at the largest extent the distance from the target. For what concerns the other random variables, it resulted from the same analysis of sensitivity that only the material of the absorbers influences in a relevant measure the behaviour of the substructure.
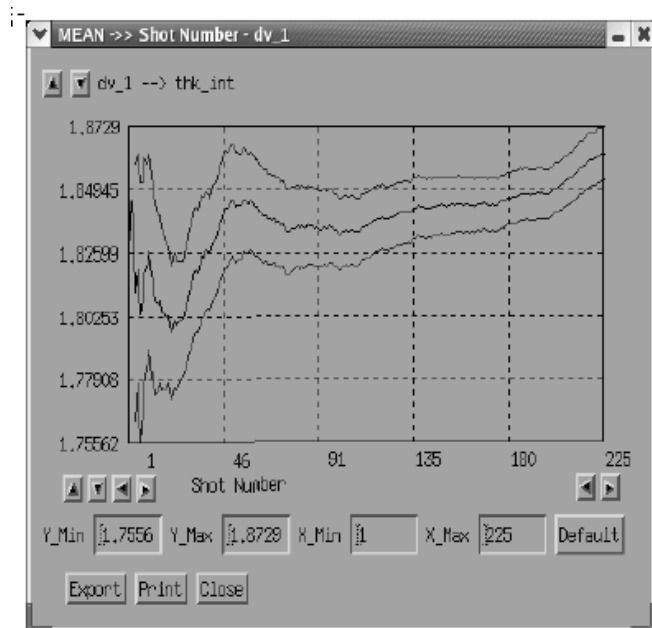
Fig. 28. Thickness of external plate vs. shot
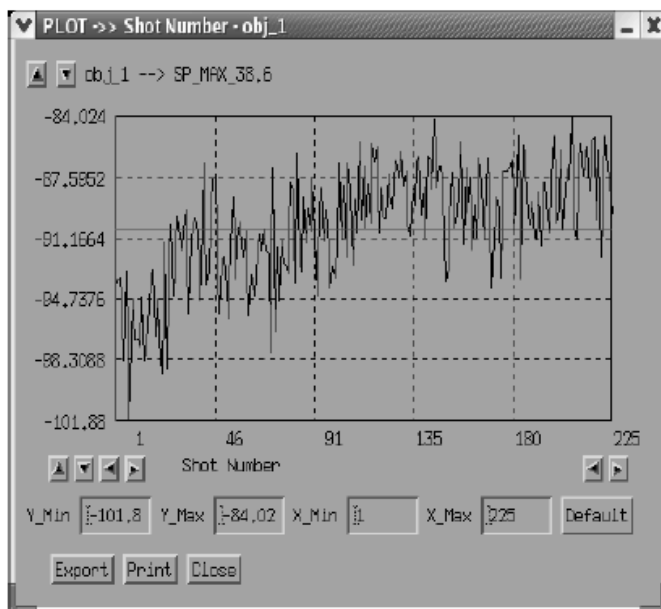
Fig. 29. Scatter plot of the objective variable

It is possible to appreciate from the scatter plots of fig. 29 and fig. 30 how the output variable approached the target value: in the 9th run, only 2 points fall under the line that represents the target and in both cases the distance is less than 0.02 and that is why the 9th run has been considered a good one, in order to save more iterations. In fig. 31 the experimental data related to the displacement of the rigid barrier vs. the time are recorded together with the numerical results obtained before and after the application of the SDI procedure.
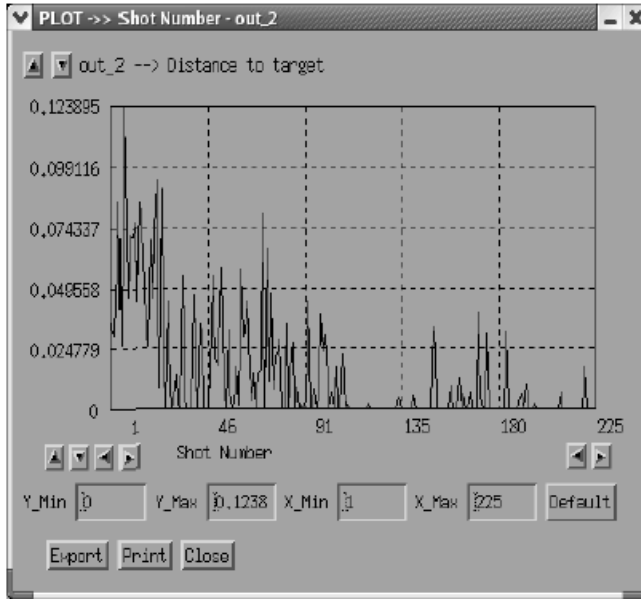
Fig. 30. Scatter plot of the distance to target
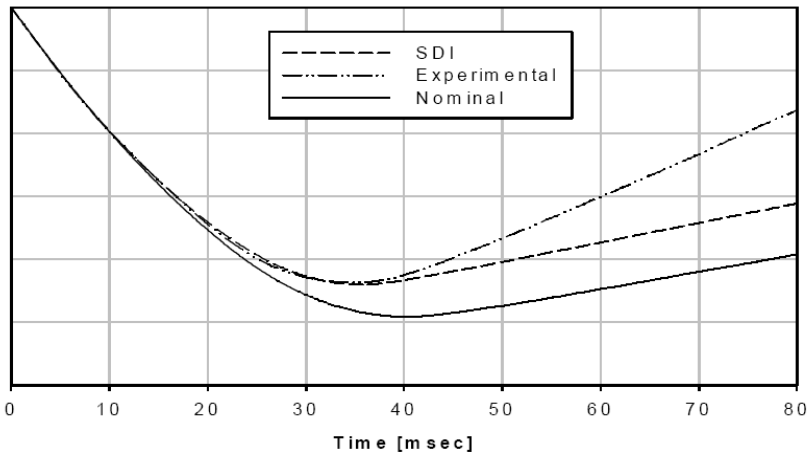
Fig. 31. Output variable vs. Time

The new nominal value of the design variables after the application of SDI procedure is 1.98 mm for both of them. A very good agreement of the numerical solution is observed in comparison to the experimental data in the first part of the curve, where they are practically overlapped and where the attention has been focused during the development of numerical simulations necessary to complete the SDI process. The general conclusion from this study was that the classical numerical simulations based on nominal values of the input variables are not exhaustive of the phenomenon in the case of crash analyses and can bring to incorrect interpretations of the dynamic behaviour of the examined structure. On the contrary, by using an SDI approach, it is possible to have a better understanding of the influence of each input variable on the structural dynamic behaviour and to assign the most appropriate nominal values in order to have results as near as possible to the target values, also in presence of their natural variability.

6. Some useful commercial codes

To fully appreciate the examples above, it may be interesting to summarize briefly the main characteristics of the commercial codes we mentioned in the preceding sections and which have interesting capabilities in the probabilistic analysis of structures; what follows doesn't want to constitute either a complete listing or an assessment of value for those codes, but only a survey of the codes we have used insofar, here published to clarify some of the topics we have just described.

First of all we have to recall that the recent versions of Ansys® couple the well-established deterministic capabilities in FE field with some new routines which work in a probabilistic environment; that innovation is so much interesting because, as we already pointed out, the design refers to structures whose study can't be carried out in a closed form by recourse to an analytical formulation; in those cases we can only hope to obtain the answer of the structure for a given set of loads and boundary conditions and therefore an FE run corresponds just to a single value of the variable set. It is only natural, therefore, that Ansys® extended its capabilities to carry out a Monte-Carlo analysis of the problem, for given statistics of the different variables and parameters.

Therefore, on the basis of a sample file (which Ansys® calls the "analysis file" of the problem) using the well renowned capabilities of its pre-processor, it is possible to characterize each variable with a distribution – to be chosen among a rather limited family of types – and then entrust the code with the task to perform a given amount of trials, from whose results the statistics of the response, as well as its sensitivities, can be recovered. A very useful characteristic of Ansys® is that the code can build a Response Surface on the basis of the obtained results and can carry on new trials using it; therefore it is quite common to carry out a limited number of M-C trials – whose amount depends on the complexity of the structure – maybe using some DOE choice, by which the Response Surface can be derived.

NESSUS®, distributed by SWRI, is a widely known and fully probabilistic code; it includes several probabilistic distributions to represent parameters and variables and is provided with both analytical and simulative methods; even if FORM, SORM, AMV+ and others are present, its main feature is the capability to be interfaced with Ansys®, Abaqus®, Ls-Dyna®, other FE codes and, at last, even with Matlab®, which widens the range of problems it can deal with; under those circumstances, it can work not just with the basic M-C method, but

also with a complete set of numerical simulative methods such as Importance Sampling, Adaptive Importance Sampling, and others. The outputs it can give are such as the cumulative distribution function of the results, the probability of failure or the performance level for a given probability of failure, the probabilistic sensitivity factors and the confidence bounds of the requested result. One important feature of NESSUS® is its capability to deal not only with components but also with systems, via such methods as the Efficient Global Reliability Analysis and the Probabilistic Fault-Tree Analysis.

STRUREL®, distributed by RCP, is a complete package which is similar to NESSUS®, but has many more capabilities, as it can deal, for example, with both time-invariant and time-variant problems. It is really much more difficult to be used, but its advantages are quite evident for the expert user; beside the capabilities we already quoted for the previous code, it can carry out risk and cost analysis, failure mode assessment, reliability assessment for damaged structure, development and optimisation of strategies for inspection and maintenance, reliability oriented structural optimisation. It can also be interfaced with Permas® FE code and with user-made Fortran® routines, in such a way as to make the user able to match with very general and complex problems; a last, but very important feature is the capability to carry out random vibration analysis, with reference, for example, to wave, wind and earthquake loading.

The next two codes are of quite different nature, as they are to be used when one is interested in optimisation and in the building of a robust design. The first one, ST-Orm®, distributed by EASi Engineering, uses the SDI technique to find the setting of the control variables of a design which ensures that the assigned target is reached with a given probability; it uses M-C to obtain a cloud of results and then, applying multilinear regressions and new M-C trials, it generates families of new clouds to reach the desired target value. It claims to be a meta-code, in the sense that its tasks can be subdivided among a number of computers, each one performing a simple task in parallel, in order to save time. An useful characteristic of this code is the possibility to distinguish among the control variables, which are probabilistic variables which can vary in each cloud, and environment parameters which, even if random in character, always exhibit the same distribution, i.e. they are not displaced with clouds. All variables and parameters span in their ranges, which can vary with clouds but cannot go beyond the physical limits which are given by the user, in such a way as to exclude impossible runs.

The last code is the well assessed Mode-Frontier®, whose aim is to carry out a multi-objective optimisation for a given problem; it works with both deterministic and random variables and one of its capabilities is to build the logic of the problem by means of an iconic and very appealing method; as we already discussed, the kernel of the code is formed by a script which can be used to organize all operations and to interface with a large number of external routines. Once the Pareto-set of the problem is obtained, it can be submitted to the Decision Manager of the code, which, following different methods can help the user to choose among the previous results the one which is more convenient.

7. Conclusions and acknowledgments

From all the preceding sections it is apparent how a probabilistic study can contribute to the improvement of a structural design, as it can take into account the uncertainties that are present in all human projects, getting to such an accurate result as to examine also the

manufacturing tolerances and coming to optimize scraps. It is to be understood, however, that such results are not easy to obtain and that it seldom happens that a first-trial analysis is a sound one: a good result is only achieved after many steps have been carried out, and first of all an accurate calibration with experimental data. It happens indeed that one of the more thorny problems the user has to struggle with is the accurate description of the material used in a particular application, as new problems usually require the description of the behaviour of the material in very particular conditions, which is not often available; therefore it happens that new tests have to be created in order to deal with new specifications and that the available tests only apparently match the requirements.

It is quite clear, therefore, that in many cases the use of such new techniques can be justified only in particular conditions, for example when one is dealing with mass production, or when failure involves the loss of many lives or in other similar conditions.

We want to acknowledge the help given by the technicians of many firms, as well by the researchers of our University, first of all by ing. G. Lamanna, to cooperate – and sometimes also to support – the researches which have been quoted in the present chapter.

8. References

Boyd-Lee, A.D., Harrison G.F. & Henderson, M.B. (2001). Evaluation of standard life assessment procedures and life extension methodologies for fracture-critical components, *International Journal of Fatigue,* vol. 23, Supplement no. 1, pp. 11-19, ISSN: 0142-1123

Caputo., F., Soprano, A. & Monacelli, G. (2006). Stochastic design improvement o fan impact absorber, *Latin American Journal of Solids and Structures,* vol. 3, no. 1, pp. 41-58, ISSN: 1679-7817

Deb,K. (2004). *Multi-Objective Optimization using Evolutionary Algorithms,* John Wiley & Sons, ISBN: 0-471-87339-X, Chichester, West Sussex, UK

Deodatis, G., Asada., H. & Ito., S. (1996). Reliability of aircraft structures under non-periodic inspection: a bayesian approach, *Engineering Fracture Mechanics,* vol. 53, no. 5, pp. 789-805, ISSN: 0013-7944

Doltsinis, I., Rau, F. & Werner., M. (1999). Analysis of random systems, in: *Stochastic Analysis of Multivariate Systems in Computational Mechanics and Engineering,* Doltsinis, I. (Ed.), pp. 9-159, CIMNE-International Center for Numerical Methods in Engineering, ISBN: 84-89925-50-X, Barcelona, Spain

Du Bois, P.A. (2004). *Crashworthiness Engineering Course Notes,* Livermore Software Technology Corp., ISBN: ………………, Livermore, Ca. USA

Horst, P. (2005). Criteria for the assessment of multiple site damage in ageing aircraft. *Structural Integrity & Durability,* vol. 1, no. 1, pp. 49-65, ISSN: 1551-3750

Langrand, B.; Deletombe, E.; Markiewicz, E. & Drazétic., P. (1999). Numerical approach for assessment of dynamic strength for riveted joints. *Aerospace Science & technology*, vol. 3, no.7, pp. 431-446, ISSN: 1270-9638

Langrand, B.; Patronelli, L.; Deletombe, E.; Markiewicz, E. & Drazétic., P. (2002). An alternative numerical approach for full scale characterization for riveted joint design, *Aerospace Science & Technology,* vol. 6, no.5, pp. 345-354, ISSN: 1270-9638

Melchers, R.E. (1999). *Structural reliability analysis and prediction,* John Wiley & Sons, ISBN: 0-471-98771-9, Chichester, West Sussex, UK

Murphy, T.E., Tsui, K.L. & Allen, J.K. (2005). A review of robust design for multiple responses, *Research in Engineering Design*, vol. 16, no. 3, pp. 118-132, ISSN: 0934-9839

Soprano, A. & Caputo., F. (2006). Building risk assessment procedures, *Structural Durability & Health Monitoring*, vol. 2, no. 1, pp. 51-68, ISSN: 1930-2983

Tani, I.; Lenoir, D. & Jezequel, L. (2005). Effect of junction stiffness degradation due to fatigue damage ofmetallic structures. *Engineering Structures*, vol. 27, no. 11, pp. 1677-1688, ISSN: 0141-0296

Urban, M.R. (2003). Analysis of the fatigue life of riveted sheet metal helicopter airframe joints, *International Journal of Fatigue*, vol. 25, no. 9-11, pp. 1013-1026, ISSN: 0142-1123

Zang, C. Friswell., M.I. & Mottershead, J.E. (2005). A review of robust optimal design and its application in dynamics, *Computers and Structures*, vol. 83, no. 4-5, pp. 315-326, ISSN: 0045-799

Modelling earthquake ground motions by stochastic method

Nelson Lam
University of Melbourne
Australia

John Wilson
Swinburne University of Technology
Australia

Hing Ho Tsang
University of Hong Kong
China Hong Kong

1. Introduction

The prediction of earthquake ground motions in accordance with recorded observations from past events is the core business of engineering seismology. An attenuation model presents values of parameters characterising the intensities and properties of ground motions estimated of projected earthquake scenarios (which are expressed in terms of magnitude and distance). Empirical attenuation models are developed from regression analysis of recorded strong motion accelerograms. In situations where strong motion data are scarce the database of records has to cover a very large area which may be an entire continent (eg. Ambrasey model for Europe) or a large part of a continent (eg. Toro model for Central & Eastern North America) in order that the size of the database has statistical significance (Toro *et al.*, 1997; Ambrasey, 1995). Thus, attenuation modelling based on regression analysis of instrumental data is problematic when applied to regions of low and moderate seismicity. This is because of insufficient representative data that has been collected and made available for model development purposes.

An alternative approach to attenuation modelling is use of theoretical models. Unlike an empirical model, a theoretical model only makes use of recorded data to help ascertain values of parameters in the model rather than to determine trends from scratch by regression of data. Thus, much less ground motion data is required for the modelling. Data that is available could be used to verify the accuracies of estimates made by the theoretical model. Ground motion simulations by classical wave theory provides comprehensive description of the earthquake ground motions but information that is available would typically not be sufficient as input to the simulations. The heuristic source model of Brune

(1970) which defines the frequency content of seismic waves radiated from a point source is much simpler. The model has only three parameters : seismic moment, distance and the stress parameter. Combining this point source model with a number of filter functions which represent modification effects of the wave travel path and the site provides estimates for the *Fourier* amplitude spectrum of the motion generated by the earthquake on the ground surface. The source model (of Brune) in combination with the various filter functions are collectively known as the seismological model (Boore, 1983). Subsequent research by Atkinson and others provides support for the proposition that simulations from a well calibrated point source model are reasonably consistent with those from the more realistic finite fault models.

The *Fourier* spectrum as defined by the seismological model only provides description of the frequency properties of the ground motions and not the phase angles of the individual frequency components of the waveforms. Thus, details of the wave arrival times which are required for providing a complete description of the ground shaking remain uncertain as they have not been defined by the seismological model. With stochastic modelling, the pre-defined frequency content is combined with random phase angles that are generated by the *Monte Carlo* process. Thus, acceleration time-histories based on randomised wave arrival details are simulated. The simulations can be repeated many times (for the same earthquake scenario and source-path-site conditions) in order that response spectra calculated from every simulated time-histories can be averaged to obtain a smooth, ensemble averaged, response spectrum.

The seismological model has undergone continuous development since its inception in the early 1980's. For example, the original Brune source model has been replaced by the empirical source model of Atkinson (1993) which was developed from seismogram data recorded in *Central and Eastern North America* to represent conditions of intraplate earthquakes. A similar model was subsequently developed by Atkinson & Silva (2000) which was developed from data recorded in *Western North America* to represent conditions of interplate earthquakes. A model to account for the complex spread of energy in space taking into account the wave-guide phenomenon and the dissipation of energy along the wave travel path has also been developed (Atkinson & Boore, 1995). The amplification and attenuation of upward propagating waves taking into account the effects of the shear wave velocity gradient of the earth crust close to the ground surface have also been modelled by Boore & Joyner (1997).

The authors have been making use of the developing seismological model as described for constraining the frequency properties of projected earthquake scenarios for different regions around the world including regions of low-moderate seismicity where strong motion data is scarce (Chandler & Lam, 2002; Lam *et al.*, 2002, 2003, 2006, 2009; Balendra *et al.*, 2001; Yaghmaei_Sabegh & Lam, 2010; Tsang *et al.*, 2010). It is typically assumed in the simulations that the intraplate source model that was originally developed for *Central and Eastern North America* is generally applicable to other intra-plate regions. Values of parameters for defining filter functions of the wave travel path could be ascertained by making references to results of seismic surveys, and in conjunction with Intensity data where necessary. Thus, earthquake ground motions that are recorded locally are not essential for model

development and time-histories simulations. Basic principles of the simulations and an introductory description of the seismological model can be found in the review article written by the authors (Lam *et al.,* 2000a). More detailed descriptions of the techniques for constraining filter functions in the absence of locally recorded ground motions can be found in Tsang & Lam (2010). Operating this modelling procedure is a very involved process. With the view of obtaining quick estimates of the response spectrum without undertaking stochastic simulations, the authors have developed a manual calculation procedure which is known as the *Component Attenuation Model* (CAM). CAM was developed from collating the experience the authors acquired in the development of response spectrum models by the stochastic procedure. The development and application of seismological modelling technique as applied to different countries, which forms the basis of CAM, has been reported in a range of journals spanning a period of ten years since 2000 (eg. Lam *et al.,* 2000a-c; Chandler & Lam, 2004; Lam & Chandler, 2005; Hutchinson *et al.,* 2003 ; Wilson *et al.,* 2003). The writing of this book chapter enables **CAM** to be presented in a coherent, compact, and complete manner.

2. Background to the Component Attenuation Model

A response spectrum for seismic design purposes can be constructed in accordance with parameters characterising the acceleration, velocity and displacement (*A*, *V* and *D*) demand properties of the earthquake. Response spectra presented in different formats are made up of zones representing these entities as shown in Figure 1.

(a) Tripartite Velocity Response Spectrum

(b) Displacement Response Spectrum

(c) Traditional Force-Based Acceleration Response Spectrum

(d) Acceleration-Displacement Response Spectrum
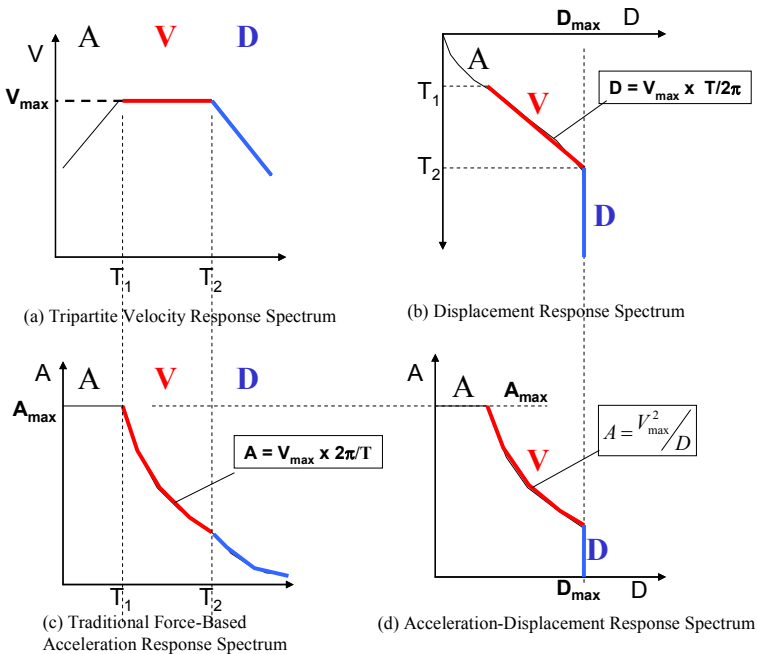
Fig. 1. Earthquake Response Spectra in different formats

The velocity response spectrum in the tri-partite format of Fig.1a in logarithmic scale is the much preferred format to use in the earthquake engineering literature given that spectral values are presented over a wide period range (eg. 0.1s – 10s) and with good resolution. Once the response spectral velocity values have been identified from the spectrum, the corresponding values of the response spectral accelerations and displacements are automatically known by means of the displayed transformation relationships.   The alternative displacement response spectrum format of Fig. 1b which provides a direct indication of the drift demand of the structure in an earthquake was proposed initially by Priestley (1995) when the displacement-based approach of seismic assessment was first introduced. The acceleration-displacement response spectrum (ADRS) diagram format of Fig. 1d is also much preferred by the engineering community given that the spectral acceleration (*A*) values are effectively values of the base shear that have been normalised with respect to the mass of the *single-degree-of-freedom* system. Consequently, the acceleration-displacement (force-displacement) relationship of a structure can be superposed onto the ADRS diagram to identify the *performance point* which represents the estimated seismic response behaviour of the system as shown in Fig. 2. Diagrams representing seismic demand and capacity in this format are also known as the *Capacity Spectrum*. The importance of the velocity and displacement (*V* and *D*) demands as opposed to the acceleration (*A*) demand in the context of protecting lives and lowering the risks of overturning and collapses is evident from Figure 2 in which typical performance points associated with ultimate behaviour of the structure are shown.
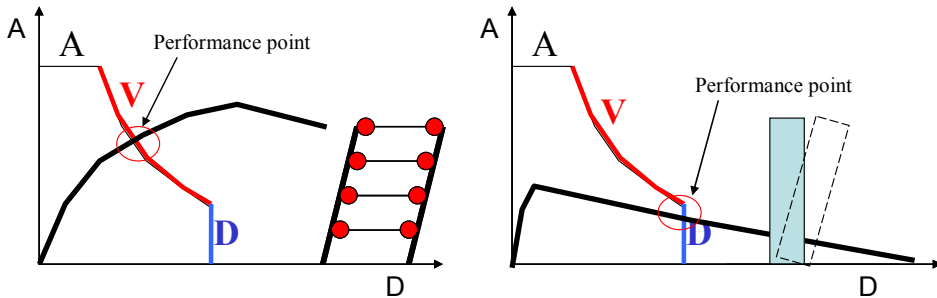
Fig. 2. Use of capacity spectrum for modelling collapse and overturning

The *Component Attenuation Model* (**CAM**) is an innovative framework by which the velocity and displacement demand on SDOF systems are expressed as product of component factors representing conditions of the source, path, local and site. The source factor is generic and hence used across different regions. Other factors that represent the path and local effects can be estimated in accordance with geophysical information of the region. The attenuation relationship is obtained by combining the generic source factor with the area specific factors. Further details of the CAM factors can be found in Sections 3 and 4.

It is shown in the velocity response spectrum of Fig. 3 that predictions of the spectral values by different empirical attenuation models can be highly variable and particularly in the low period range. Clearly, there is much less variability in the estimation of V_{max} in the median period range of 0.5s – 1.0s than that of A_{max} in the lower period range. Predictions by the whole range of attenuation models for the highest point on the velocity spectrum are

conservatively represented by the *Component Attenuation Model* (**CAM**) for rock conditions. The displacement demand behaviour of the earthquake in the high period range is also well constrained by the earthquake magnitude (and hence *seismic moment*). The apparent variability displayed in the high period (low frequency) range by certain models in Fig. 3 is only reflective of the poor resolution of the recorded data and not in the ground motions itself. Thus, the viability of generalising the predictions of the response spectrum parameters (V_{max} and D_{max}) is well demonstrated. Consequently, **CAM** is formulated to provide estimates for these demand parameters.

SD	Sadigh *et al.*	1997
CB	Campbell (Geomatrix)	1997
BJF	Boore, Joyner & Fumal	1997
AS	Abrahamson & Silva	1997
AM	Ambraseys	1995
DA	Dahle *et al.*	1990
TAS	Toro, Abrahamson & Schneider	1997
CAM	Component Attenuation Model (Lam *et al.*)	2000b

Fig. 3. Comparison of response spectra from different attenuation relationships (M7 R=30km on rock)

3. Formulation of the Component Attenuation Model

The Component Attenuation Model which comprises a number of component factors for estimation of the maximum velocity and displacement demand (V_{max} and D_{max}) is represented diagrammatically in Figure 4 and Equations (1) – (10).

$$V_{\max} \text{ or } D_{\max} = \alpha.\beta.G.\gamma.S$$

Reference prediction
based on R=30km
on hard rock

Predicted modification
by energy absorption
along travel path

Predicted modification
by geometrical spread
of energy

Predicted modification
by rock crusts

Predicted
modification
by soil
sediments

Component
Factors
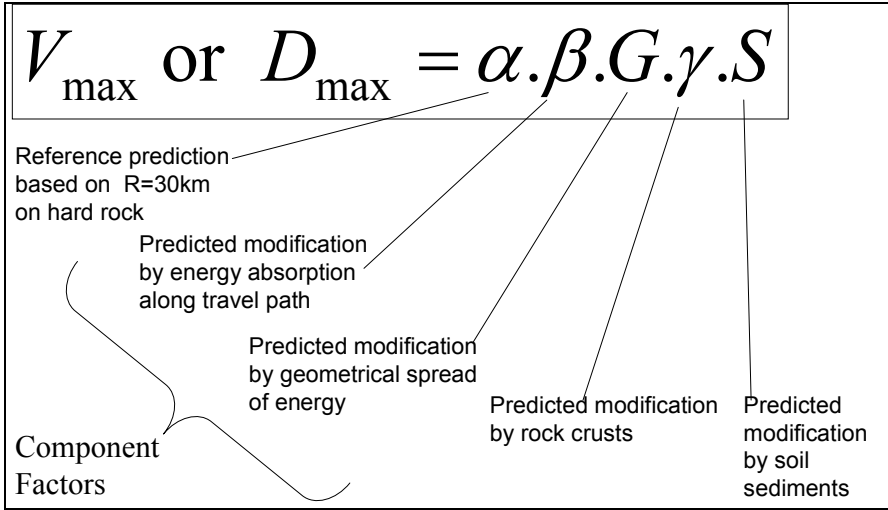
Fig. 4. Use of capacity spectrum for modelling collapse and overturning

Predictions of V_{max}

$$V_{\max} = \alpha_V.G.\beta_V.\gamma_V.S \tag{1}$$

where

$$\alpha_V = 70\{0.35 + 0.65(M-5)^{1.8}\} \qquad M \text{ is moment magnitude} \tag{2}$$

$$\beta_V = \left(\frac{30}{R}\right)^{0.005R} \qquad R \text{ in km for } R < 50\text{km} \tag{3}$$

$$G = \frac{30}{R} \qquad R \text{ in km for } R < 50\text{km} \tag{4}$$

γ_V is crustal factor

(value is typically in the range 1.6 - 2.0 but could be much lower in continental "shield" areas)

S is site factor (value is typically in the range 1.5 - 2.0 for average site)

Predictions of D_{max}

$$D_{max} = \alpha_D.G.\beta_D.\gamma_D.S \tag{5}$$

where

$$\alpha_D = \alpha_V\left(\frac{T_2}{2\pi}\right); \tag{6}$$

$$T_2 = 0.5 + \left(\frac{M-5}{2}\right) \quad for \ M \leq 8 \qquad (Lam \ et \ al., \ 2000b); \tag{7}$$

$$or \ \alpha_D = 10^{M-5} \qquad for \ M \leq 6.5 \quad (Lam \ \& \ Chandler, \ 2004); \tag{8}$$

$$\beta_D = \left(\frac{30}{R}\right)^{0.003R} \qquad R \ in \ km \ for \ R < 50km; \tag{9}$$

$$G = \frac{30}{R} \qquad R \ in \ km \ for \ R < 50km; \tag{10}$$

γ_D is crustal factor

(value is typically in the order of 1.5-1.6 but could be much lower in continental "shield" areas)

S is site factor (value is typically in the range 1.5-2.0 for average site)

4. The Component Factors

4.1 The α_V and α_D factors

The component factors α_V and α_D as defined by equations (2) and (6-8) are for predicting the values of the V_{max} and D_{max} parameters at a reference distance of 30km (Lam *et al.*, 2000b). These equations were obtained from ensemble average response spectra that were simulated in accordance with the seismological source model of Atkinson (1993). The alternative expression of equation (8) for calculation of the value of α_D was derived from a theoretical approach presented by Lam and Chandler (2005). Predictions for the value of α_D from both approaches are very consistent for M < 6.5. For higher moment magnitude, equations (6-7) provide less conservative predictions. Ground motions so simulated have been scaled to a reference distance of 30 km as opposed to the usual 1km. This reference distance value is unique to CAM and is based on conditions of low and moderate seismicity which is characterised by moderate ground shaking with return periods of 500 – 2500 years.

4.2 The β_V and β_D factors

The component factors β_V and β_D are representing reduction in the seismic demand as the result of energy dissipation along the wave travel path. The effects of this form of attenuation, which are known as anelastic attenuation, are only significant to the prediction of the value of the V_{max} and D_{max} parameters at long distances. Thus, simple expressions like equations (3) and (9) have been used to represent its effects at close distances of R < 50 km. At longer distances, the determination of the β_V and β_D factors are expressed as functions of the Quality factor (Q_0) at a reference frequency of 1 hertz. The effects the value of Q_0 have upon the rate of wave attenuation is shown in the schematic diagram of Figure 5. Clearly, the higher the value of Q_0, the better the wave transmission quality of the earth crust. Seismically active regions of young geological formations such as California have the value of Q_0 in the order of 100 – 200. Regions of ancient geological formation (intercontinental shield regions) such as *Central and Eastern North America* and parts of *Central and Western Australia* have the value of Q_0 typically exceeding 500.

High Q value   eg. Q_0= 600   low rate of energy dissipation
schematic diagram – not to scale

Low Q value   eg. Q_0= 100   high rate of energy dissipation
schematic diagram – not to scale
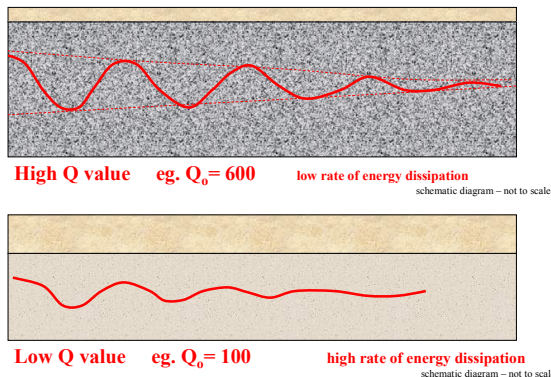
Fig. 5. Schematic representation of the effects of Quality factor on energy dissipation.

In a study by Chandler and Lam (2004) on the attenuation of long distance earthquakes, expressions defining the value of β_V and β_D (ie. rate of decrease in the values of the V_{max} and D_{max} parameters) as function of Q_0 and R have been derived from stochastic simulations of the seismological model. Functions defining the value of β_D is represented graphically by Fig.6 whilst values of β_V can be estimated using equation (11) once the value of β_D has been identified. It is noted that Fig. 6 is restricted to earthquakes with moment magnitude not exceeding 8. The attenuation modelling of (M>8) mega magnitude earthquakes like the subduction earthquakes generated from off-shore of Sumatra would involve stochastic simulations of the seismological model (Lam *et al.*, 2009) and is beyond the scope of this book chapter.
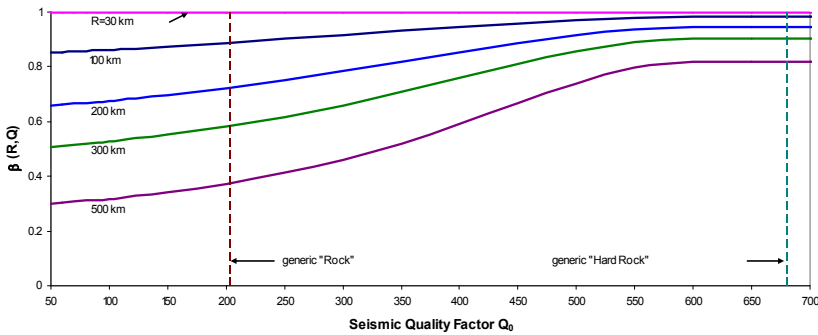
Fig. 6. Chart for determination of the value of β_D as function of Q_0 and R

$$\beta_V \approx 0.8\beta_D \qquad \text{for} \qquad R = 100km \qquad (11a)$$

$$\beta_V \approx 0.6\beta_D \qquad \text{for} \qquad R = 200km \qquad (11b)$$

$$\beta_V \approx 0.4\beta_D \text{ or } 0.5\beta_D \qquad \text{for} \qquad R = 300km \qquad (11c)$$

4.3 The G factor

The G factor represents the effects of the geometrical spread of energy in space as seismic waves are radiated from a point source at the depth of rupture within the earth's crust. At close range to the point source (R < 50km), spherical attenuation applies. The intensity of wave energy decreases in proportion to $1/R^2$ (as area of the surface of a sphere is proportional to the square of its radius). The rate of attenuation of the *Fourier* amplitude of the simulated wave is accordingly proportional to $1/R$ which is consistent with equations (4) and (10). The geometrical attenuation of seismic waves becomes more complex when the value of R is sufficiently large that reflection of waves from the Moho discontinuity and the associated wave-guide effects as shown in Figure 7 needs be taken into account. Thus, the depth of earth crust D in the region (ie. depth to the reflective surface of Moho) is an important modelling parameter. The value of D on land typically varies between 30 km and 60 km. Higher values are found in mountainous regions. Spherical attenuation may be assumed in the range R < 1.5D and cylindrical attenuation in the range R > 2.5D according to Atkinson & Boore (1995). Functions defining the value of the G factor for different values of D in the long distance range are represented graphically by Fig.8.

Geometrical Spread of Energy V_{max} or $D_{max} = \alpha.\beta.G.\gamma.S$
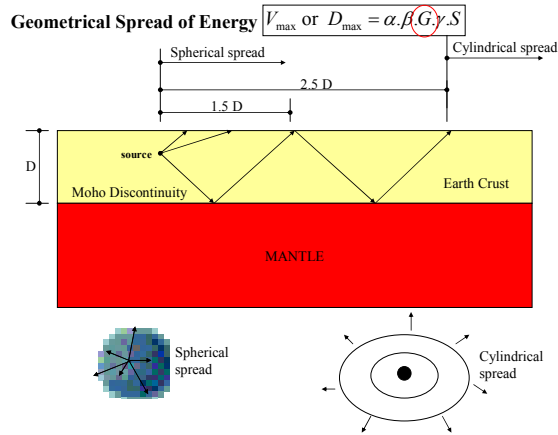
Fig. 7. Schematic representation of geometrical attenuation
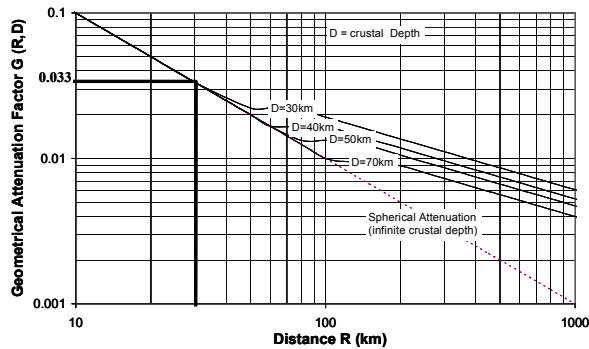
Fig. 8. G factor expressed as function of distance (R) and crustal depth (D)

4.4 The γ_V and γ_D factors

The crustal factor represents effects of modifications to the seismic waves as they propagate up the (rock) crust, and are made up of two components: (i) mid-crustal amplification and (ii) upper crustal modifications.

The amplitude of seismic waves generated at the source of the earthquake is proportional to the shear wave velocity of the earth crusts surrounding the fault rupture raised to the power of 3 (Atkinson & Silva, 1997). The α_V and α_D factors as described in Section 4.1 are both based on shear wave velocity of 3.8 km/s which is representative of conditions of the fault source at depths exceeding 12 km. For most moderate and large magnitude shallow earthquakes of M ≥ 6, the centroid of the ruptured surface is constrained to a depth of around 5 km if the rupture area is of the order of 100 km² or larger. In this depth range, the shear wave velocity is estimated to average at around 3.5 km/s based on models presented by Boore and Joyner (1997). The mid-crustal factor is accordingly equal to 1.3 (being 3.8/3.5 raised to the power of 3).

Upward propagating seismic shear waves can be modified rapidly by the upper 1-2 km of the earth's crust shortly before the wave fronts reaches the ground surface. Much is attributed to the shear wave velocity gradient of the crustal medium. Meanwhile, seismic waves could also be attenuated fairly rapidly through energy dissipation by the typically highly fissured rocks in the upper 3-4 km of the earth's crust. These path effects can be difficult to track if measurements have only been taken from the ground surface. Upper crustal modifications were well demonstrated by the study of Abercrombie (1997) in which seismometer records collected from several km deep boreholes were analysed. Stochastic simulations undertaken the authors based on the generic *rock* profile of Boore and Joyner (1997) and principles of *quarter wave-length method* for the calculation of frequency dependent amplification revealed an upper crustal factor of about 1.2 (Lam *et al*., 2000b) when co-existing attenuation in the upper crust (based on parameters that are consistent with strong ground shaking in active regions like California) had also been taken into account. The attenuation parameter that can be used to characterised upper crustal attenuation is known as *Kappa* (Anderson & Hough, 1984). The value of this parameter for strong ground shaking in generic *rock* is in order of 0.04 – 0.07 (Atkinson & Silva, 1997; Atkinson & Boore, 1998; Tsang & Lam, 2010). For conditions of moderate ground shaking and in regions of older geological formation (which is characterised by a lower *Kappa* value of the order of 0.02-0.03) a higher upper crustal factor of 1.5 in the velocity controlled region of the response spectrum is estimated (Lam & Wilson, 2004). Behaviour of amplification in the displacement controlled region of the response spectrum is more robust and is insensitive to the *Kappa* value.

In summary, the combined crustal factor γ_V for modelling the velocity demand (V_{max}) is accordingly in the range 1.5 – 2.0 (based on the product of "1.3" and "1.2 – 1.5") depending on the intensity of ground shaking and type geological formation, whilst the combined crustal factor γ_D for modelling the displacement demand (D_{max}) is in the order of 1.5 - 1.6. However, much lower values of γ_V or γ_D should be assumed for continental "shield" areas where there are much less modifications of the upward propagating waves by the very hard rock in those areas.

These crustal factor values can be compared with the ratio of ground shaking estimated in regions of very different geological formation but of the same earthquake scenario and source processes. The inferred ratio of ground shaking between *Western Australia* and *Southeastern Australia* has been found to be 1.5 – 1.7 based on the Intensity model of Gaull *et al.* (1990) developed for both regions. Similarly, the inferred ratio of ground shaking between the mid-continental region of *Central and Eastern North America* and that of *Mexican Gulf* (of younger geological formations) has been found to be 1.5 – 1.6 based on the stochastic model of Toro *et al.* (1997). The inferred ratio between *Western North American* and *Central and Eastern North America* has been found to be in between 1.3 – 1.8 based on the stochastic model of Atkinson and Silva (2000). These inferred ratios are all in broad agreement with the values of the γ_V and γ_D factors that have been recommended by CAM.

Recommendations that have been made in the above enable quick estimates of the response spectrum parameters to be made whilst alleviating the need for any rigorous analysis of strong motion or seismological data. Precise evaluation of the crustal factors would involve

measuring and modelling the shear wave velocity gradient of the earth crusts in the region (Chandler *et al.*, 2005a & 2006a; Lam *et al,* 2006; Tsang *et al.*, 2010), constraining *Kappa* values either by analysis of *Coda Wave* data or by making use of generic correlations between values of *Kappa* and shear wave velocity parameters of the earth's crust in the region (Chandler *et al.*, 2005b & 2006b), and calculating filter functions that take into account both the amplification and attenuation effects. Stochastic simulations of the seismological model that have incorporated these developed filter functions can provide direct estimates of the crustal effects on ground shaking in projected earthquake scenarios. However, it is beyond the scope of this book chapter to present details of these modelling processes.

5. Comparison with recorded data and examples

The *Component Attenuation Model* as described is essentially a tool for providing estimates of the response spectrum parameters for rock outcrops. Meanwhile, velocity parameters of ground shaking on average sites can be inferred from Intensity data collected from historical earthquake events. Comparisons of the two sets of data provide estimates of the site factors that represent the difference between ground shaking on rock and that on an average site in pre-determined earthquake scenarios. This calibration process for constraining the site factor is illustrated in the schematic diagram of Figure 9.

Modified Mercalli Intensity (*MMI*) values recorded from observed damage to structures on average sites in an area.

MMI values converted to peak ground velocity (*PGV*) values using well known expressions of Newmark & Hall (82) or the more recently developed model by Atkinson & Kaka (07)

Values of V_{max} and *PGV* predicted for rock sites using CAM $V_{max} = \alpha.\beta.G.\gamma$

The value of the **S** factor is obtained by dividing the value of *PGV* (avg sites) by *PGV* (rock sites)

Fig. 9. Inferring Site factor

Using this calibration approach, the value of S factor for average sites have been found to be 1.5 – 1.8 in a study undertaken for three regions within Central China (Tsang *et al.,* 2010); 1.5 for Australia on average (Lam *et al.,* 2003); 1.7 for Northern Iran (Yaghmaei-Sabegh and Lam, 2010); and a slightly higher value of about 2.0 for the South China region surrounding Hong Kong. Importantly, this range of calibrated site factors obtained from different studies are in broad agreement and consistent with the site factor recommended by *NEHRP* for common shallow soil sites.

Further evaluation of the CAM expressions have been undertaken by Lumantarna et al. (2010) based on comparing response spectrum parameters calculated from the CAM expressions presented in this book chapter and those calculated from some 196 accelerogram records that were made available from data resources provided online by the Pacific Earthquake Engineering Research Centre (PEER). This database of strong motion accelerograms were mainly made up of records taken from California and with a few records from Southern Europe (Italy) and from Turkey. These records which were mainly post 1980 (except for a few taken in the 1970's) were all recorded on Class B sites (soft rock and stiff soil) with shear wave velocity in the range 360 – 750 m/s and from events of magnitude M5 - M7 within epicentral distances 50 – 60 km and thus within the scope of the presented CAM expressions. CAM was then applied using the expressions outlined in Section 3, with γ = 1.5 and S = 1.5 in view of the conditions of strong ground shaking in most of the recorded events. It is shown in the comparative plots of Figs. 10 – 11 that CAM generally provides a conservative estimate for the *Vmax* and *Dmax* values although a large scatter exists. It is important to note that few recorded results exceed 2 times the CAM estimates with less scatter with the recorded values of *Dmax*.
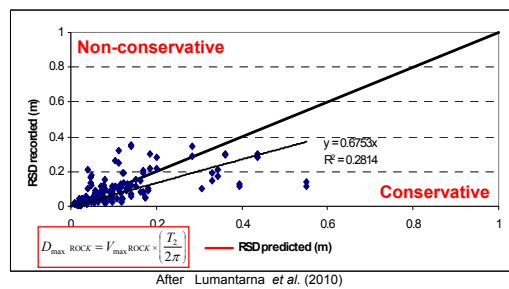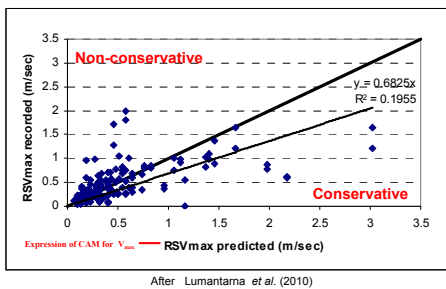
Fig. 10. Recorded and Predicted V_{max} values   Fig. 11. Recorded and Predicted D_{max} values

6. Examples for illustrating the use of CAM

Finally, the use of the CAM expressions for estimating the value of V_{max} and D_{max} are illustrated with two examples: (i) M5.6 event at a distance of 16km and (ii) M7 event at a distance of 100 km. Both earthquake scenarios are assumed to occur in the young geological (sandstone) formation of the Sydney basin. Crustal depth D can be taken as 30 km and value of Q_0 is 200. Example 1 was a real event that occurred in the City of Newcastle in December 1989, but no accelerogram records exist of that event.

6.1 Example 1
Input data is M=5.6, R=16 km

$$V_{\max} = \alpha_V.G.\beta_V.\gamma_V.S \tag{12}$$

where

$$\alpha_V = 70\{0.35 + 0.65(M-5)^{1.8}\} = 70\{0.35 + 0.65(5.6-5)^{1.8}\} = 43mm/s \tag{13}$$

$$\beta_V = \left(\frac{30}{R}\right)^{0.005R} = \left(\frac{30}{16}\right)^{0.005\times16} = 1.05 \tag{14}$$

$$G = \frac{30}{R} = \frac{30}{16} = 1.9 \tag{15}$$

$$\gamma_V = 1.6$$

$$S = 1.5 \text{ for average site} \tag{16}$$

$$V_{\max} = (43mm/s)(1.9)(1.05)(1.6)(1.5) = 205mm/s \tag{17}$$

$$D_{\max} = \alpha_D.G.\beta_D.\gamma_D.S \tag{18}$$

where

$$\alpha_D = \alpha_V\left(\frac{T_2}{2\pi}\right) = 43\times\left(\frac{0.8}{2\pi}\right) = 5.5mm \tag{19}$$

$$T_2 = 0.5 + \left(\frac{M-5}{2}\right) = 0.5 + \left(\frac{5.6-5}{2}\right) = 0.8s \tag{20}$$

$$\beta_D = \left(\frac{30}{R}\right)^{0.003R} = \left(\frac{30}{16}\right)^{0.003\times16} = 1.03 \tag{21}$$

$$G = \frac{30}{R} = \frac{30}{16} = 1.9 \tag{22}$$

$$\gamma_D = S = 1.5 \tag{23}$$

$$D_{\max} = (5.5mm)(1.9)(1.03)(1.5)(1.5) = 24mm \tag{24}$$

Response spectra of two different formats constructed in accordance with the calculated values of V_{max} and D_{max} are shown in Fig. 12 in below.
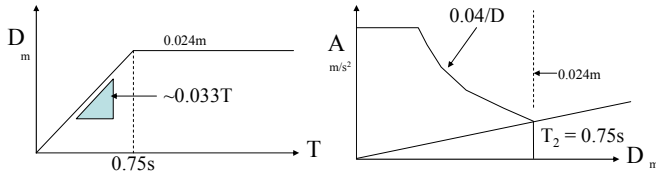
Fig. 12. Response spectra constructed for example 1

6.2 Example 2

Input data is M=7, R=100 km, D=30 km and Q_0 = 200

$$V_{\max} = \alpha.G.\beta.\gamma.S \tag{25}$$

$$\alpha_V = 70\{0.35 + 0.65(M-5)^{1.8}\} = 70\{0.35 + 0.65(7-5)^{1.8}\} = 180mm/s \tag{26}$$

$$G = \frac{0.02}{0.033} = 0.6 \text{ as indicated by chart in Fig.13 below} \tag{27}$$

$$\beta_D = 0.9 \text{ as indicated by the chart in Fig.14 below;} \quad \beta_V \approx 0.8\beta_D \approx 0.7$$

$$V_{\max} = 180mm/s(0.6)(0.7)(1.5)(1.5) \approx 170mm/s \tag{28}$$

$$\alpha_D = \alpha_V\left(\frac{T_2}{2\pi}\right) = \alpha_V\left(\frac{0.5+\left(\frac{M-5}{2}\right)}{2\pi}\right) = 180\left(\frac{0.5+\left(\frac{7-5}{2}\right)}{2\pi}\right) = 180\left(\frac{1.5}{2\pi}\right) = 43mm \tag{29}$$

$$D_{\max} = 43mm(0.6)(0.9)(1.5)(1.5) \approx 55mm \tag{30}$$
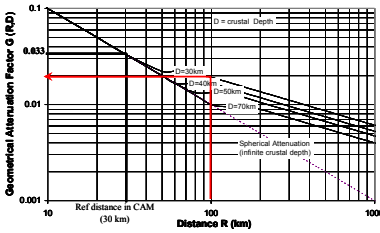
Fig. 13. Identification of the value of G   Fig. 14. Identification of the value of β_D
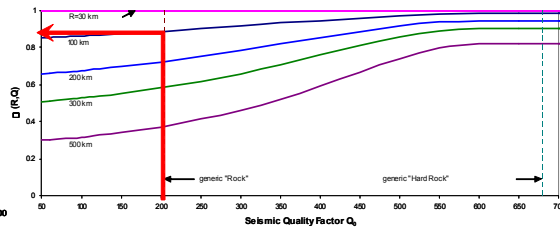
Response spectra of two different formats constructed in accordance with the calculated values of V_{max} and D_{max} for the distant earthquakes are shown in Fig. 15 in below. It is noted that the corner period (T_2) of 1.5s in the source factor has been increased to 2s by the long distance (path) effects which are represented by the β_V and β_D factors.
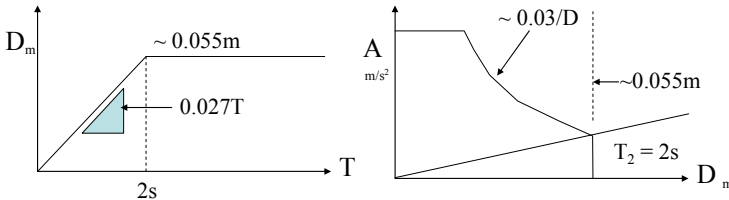
Fig. 15. Response spectra constructed for Example 2

7. Conclusions

This paper introduces the *Component Attenuation Model* (CAM) which is a generalised attenuation model that has been derived from stochastic simulations of the seismological model. The model is made up of a series of component factors representing the effects of the source, the wave travel path, modifications by the earth's crust and that of the site. Expressions and charts have been presented for evaluation of the individual factors. Parameter values calculated by the CAM expressions have been compared with those calculated from some 196 recorded accelerograms obtained from the PEER database. Two examples illustrating the use of CAM have been shown.

8. References

Abercrombie, R.E. (1997), Near-surface attenuation and site effects from comparison of surface and deep borehole recording, *Bulletin of the Seismological Society of America*, Vol.87: 731-744.

Abrahamson, N.A. & Silva, W.J. (1997), Empirical response spectral attenuation relations for shallow crustal earthquakes, *Seismological Research Letters*, Vol. 68(No.1): 94-127.

Ambraseys, N.N. (1995), The prediction of earthquake peak ground acceleration in Europe. *Earthquake Engineering & Structural Dynamics*, Vol. 24: 467-490.

Anderson, J.G. & Hough, S.E. (1984), A Model for the Shape of the Fourier Amplitude Spectrum of Acceleration at High Frequencies, *Bulletin of the Seismological Society of America*, Vol.74 (No.5): 1969-1993.

Atkinson, G.M. (1993), Earthquake source spectra in Eastern North America. *Bulletin of the Seismological Society of America*, Vol.83: 1778-1798.

Atkinson, G.M. & Boore, D.M. (1995), Ground-motion relations for Eastern North America. *Bulletin of the Seismological Society of America*, Vol.85 (No.1): 17-30.

Atkinson, G.M. & Silva, W. (1997), An empirical study of earthquake source spectra for Californian earthquakes, *Bulletin of the Seismological Society of America*, Vol.87: 97-113.

Atkinson, G.M. & Boore, D.M. (1998), Evaluation of models for earthquake source spectra in Eastern North America. *Bulletin of the Seismological Society of America*, Vol.88(No.4): 917-937.

Atkinson, G.M. & Silva, W. (2000), Stochastic modeling of Californian Ground Motions, *Bulletin of the Seismological Society of America*, Vol.90: 255-274.

Atkinson, G. M. & S. I. Kaka (2007), Relationship between Felt Intensity and Instrumental Ground motion in the Central United State and California. *Bulletin of the Seismological Society of America*, Vol.97 (No.2): 497-510.

Balendra, T., Lam, N.T.K., Wilson, J.L. & Kong, K.H. (2001), Analysis of long-distance earthquake tremors and base shear demand for buildings in Singapore, *Journal of Engineering Structures*. Vol.24: 99-108.

Boore, D.M. (1983), Stochastic Simulation of high-frequency ground motions based on seismological model of the radiated spectra, *Bulletin of the American Seismological Society of America*, Vol.73 (No.6): 1865-1894.

Boore, D.M. & Joyner, W.B. (1997), Site amplifications for generic rock sites, *Bulletin of the Seismological Society of America*, Vol.87 (No.2): 327-341.

Boore, D.M., Joyner, W.B. & Fumal, T.E. (1997), Equations for estimating horizontal response spectra and peak acceleration for western North American earthquakes: a summary of recent work, *Seismological Research Letters,* Vol.68: 128-153.

Brune, J.N. (1970), Tectonic stress and the spectra of seismic shear waves from earthquakes, *Journal of Geophysics Research*, Vol.75: 4997-5009.

Campbell, K.W. (1997), Empirical near-source attenuation relationships for horizontal and vertical components of peak ground acceleration, peak ground velocity, and pseudo-absolute acceleration response spectra, *Seismological Research Letters*, Vol.68 (No.1): 154-179.

Chandler, A.M. & Lam, N.T.K. (2002), Intensity Attenuation Relationship for the South China Region and Comparison with the Component Attenuation Model, *Journal of Asian Earth Sciences*. Vol.20: 775-790.

Chandler, A.M. & Lam, N.T.K. (2004), An attenuation model for distant earthquakes, *Earthquake Engineering and Structural Dynamics*. Vol.33 (No.2): 183-210.

Chandler, A.M., Lam, N.T.K., & Tsang, H.H. (2005a), Shear Wave Velocity Modelling in Bedrock for Analysis of Intraplate Seismic Hazard, *Journal of Soil Dynamics & Earthquake Engineering*. Vol.25: 167-185.

Chandler, A.M., Lam, N.T.K., & Tsang, H.H.(2005b), Estimation of Near Surface Attenuation Parameter in Crustal Rock by Calibration Analyses, *International Journal of Seismology and Earthquake Engineering*. Vol. 7 (No.3): 159-172.

Chandler, A.M., Lam, N.T.K., & Tsang, H.H. (2006a), Regional and Local Factors in Attenuation Modelling, *Journal of Asian Earth Sciences*. Vol.27: 892 - 906.

Chandler, A.M., Lam, N.T.K., & Tsang, H.H. (2006b). Near Surface Attenuation Modelling based on Rock Shear Wave Velocity Profile. *Journal of Soil Dynamics & Earthquake Engineering*, Vol.26 (No.11): 1004-1014.

Gaull, B.A., Michael-Leiba, M.O. & Rynn, J.M.W. (1990), Probabilistic earthquake risk maps of Australia, *Australian Journal of Earth Sciences*, Vol.37: 169-187.

Hutchinson, G.L., Lam, N.T.K. & Wilson, J.L. (2003), Determination of earthquake loading and seismic performance in intraplate regions, *Progress in Structural Engineering and Materials*, 2003. Vol.5: 181-194.

Lam, N.T.K., Wilson, J.L. & Hutchinson, G.L. (2000a), Generation of Synthetic Earthquake Accelerograms Using Seismological Modelling : A Review, *Journal of Earthquake Engineering* , Vol.4(No.3): 321-354.

Lam, N.T.K., Wilson, J.L., Chandler, A.M. and Hutchinson, G.L. (2000b), Response Spectral Relationships for Rock Sites Derived from The Component Attenuation Mode, *Earthquake Engineering and Structural Dynamics*, Vol.29 (No.10): 1457-1490.

Lam, N.T.K., Wilson, J.L., Chandler, A.M. and Hutchinson, G.L. (2000c), Response Spectrum Modelling for Rock Sites in Low and Moderate Seismicity Regions Combining Velocity, Displacement and Acceleration Predictions, *Earthquake Engineering and Structural Dynamics*, Vol.29 (No.10): 1491-1526.

Lam, N.T.K., Wilson, J.L., Chandler, A.M. & Hutchinson, G.L.(2002), Response Spectrum Predictions for Potential Near-Field and Far-Field Earthquakes Affecting Hong Kong : Rock Sites. *Soil Dynamics and Earthquake Engineering*. Vol.22: 47-72.

Lam, N.T.K., Sinadinovski, V., Koo, R.C.H. & Wilson, J.L.(2003), Peak Ground Velocity Modelling for Australian Intraplate Earthquakes. *International Journal of Seismology and Earthquake Engineering*. Vol.5(No.2): 11-22.

Lam, N.T.K. & Wilson, J.L.(2004), Displacement Modelling of Intraplate Earthquakes, special issue on Performance Based Seismic Design, *International Seismology and Earthquake Technology Journal*, Vol.41 (No.1): 15-52.

Lam, N.T.K. & Chandler, A.M.(2005). Peak Displacement Demand in Stable Continental Regions, *Earthquake Engineering and Structural Dynamics*. Vol.34: 1047-1072.

Lam, N.T.K., Wilson, J.L. & Srikanth, V.(2005), Accelerograms for dynamic analysis under the New Australian Standard for Earthquake Actions, *Electronic Journal of Structural Engineering*. Vol.5: 10-35.

Lam, N.T.K., Asten, M., Roberts, J., Srikanth, V., Wilson, J.L., Chandler, A.M. & Tsang, H.H. (2006), Generic Approach for Modelling Earthquake Hazard, *Journal of Advances in Structural Engineering*, Vol.9 (No.1): 67-82.

Lam, N.T.K., Balendra, T., Wilson, J.L. & Srikanth, V. (2009), Seismic Load Estimates of Distant Subduction Earthquakes Affecting Singapore, *Engineering Structures,* Vol.31 (No.5): 1230-1240.

Lumantarna, E., Lam, N.T.K. & Wilson, J.L. (2010), Studies of peak displacement demand and second corner period, *Report Infrastructure Group, Department of Civil & Environmental Engineering.*

Newmark, N.M. & Hall, W.J. (1982) *Earthquake spectra and design*. EERI Monograph, Earthquake Engineering Research Institute, California, U.S.A.

Priestley, M.J.N. (1995). Displacement-based seismic assessment of existing reinforced concrete buildings, *Procs. 5th Pacific Conf. on Earthquake Eng.*, Melbourne, 225-244.

Sadigh, K., Chang, C.Y., Egan, J.A., Makdisi, F. & Youngs, R.R. (1997), Attenuation relationships for shallow crustal earthquakes based on Californian strong motion data, *Seismological Research Letters*, Vol.68 (No.1): 180-189.

Toro, G.R., Abrahamson, N.A. & Schneider, J.F. (1997), Model of strong ground motions from earthquakes in Central and Eastern North America: best estimates and uncertainties. *Seismological Research Letters*, Vol.68 (No.1): 41-57.

Tsang, H.H. & Lam, N.T.K. (2010), *Seismic Hazard Assessment in Regions of Low-to-Moderate Seismicity,* Lambert Academic Publishing, Deutsh. ISBN:978-3-8383-3685-5.

Tsang, H.H., Sheikh, N. & Lam, N.T.K. (2010), Regional Differences in Attenuation Modelling for Eastern China, *Journal of Asian Earth* Sciences, Vol. 39(5): 441-459.

Wilson, J.L. & Lam, N.T.K. (2003), A recommended earthquake response spectrum model for Australia, *Australian Journal of Structural Engineering*, Vol.5(No.1): 17-27.

Yaghmaei-Sabegh, S & Lam, N.T.K. (**2010**). Ground motion modelling in Tehran based on the stochastic method, *Soil Dynamics and Earthquake* Engineering, Vol.30 : 525-535.

Quasi-self-similarity for laser-plasma interactions modelled with fuzzy scaling and genetic algorithms

Danilo Rastovic
Control Systems Group
Nehajska 62, 10110 Zagreb, Croatia

1. Introduction

We use the methods of fuzzy fractals description and genetic algorithms for laser-plasma interactions. We work with reduced fractals. The optimization of laser beams and target modelling could be obtained by the methods of artificial intelligence as in a tokamak physics. Normally, the expression for Vlasov equation is different than it is in the case of tokamak. The differences are also true for appropriate Maxwell equations because the design is different The applications of Vlasov-Maxwell equations in the case of inertial confinement fusion are already done by many authors. We shall find with fractal theory the main directions of particles movements and on such a way we will be able to calculate appropriate stochastic differential equations.

An energy source based on inertial fusion has several inherent advantages:
-the underlying target physics can be established on a single shot basis using existing or soon to be completed facilities.
- most, if not all, of the fusion nuclear science and technology can be developed and demonstrated on one repetitively pulsed facility. This includes the target physics, the driver/final optics, the target fabrication and injection, materials and components, and the chamber architecture (Dean, 2008).
By 2010-2012, we should be poised to take full advantage of National Ignition Facility (NIF) ignition with proposals for credible designs of such advanced targets that could be fielded on later phases of NIF.
NIF could also pre-qualify beam injection, tracking, beam slewing requirements for such targets. It mights also entertain the fielding and igniting of targets in „burst-mode", e.g., several targets injected on the fly at a few Hz. This would complement the target injection efforts planned by the high average power laser program on smaller high-rep laser systems.
The objective of the work is to give possibility of applications of the methods of artificial intelligence (fuzzy sets and logic, genetic algorithms) and fractals theory to get better understanding of inertial confinement fusion.

In the theory, the self-similarity of fractals is always an infinite process. But, in practice, as we shall see, this process must be stopped after finite time of steps. Then such a process we call quasi-self-similarity. The second novelty is that in nature, the self-similarity is generally non-symmetric, and it can be described inside the fuzzy sets and systems theory. On such a case each individual object, process etc. has the properties of originality.

New is also the applications of fuzzy scaling on partial differential equations that describe the complexity of plasma behaviour .

2. Laser-plasma interactions

NIF's first four beams were fired into various-sided gold-plated cylinders known as hohlraums, only a few millimeters long. This was a very clear demonstration that NIF is indeed on the path toward ignition. It has been shown analitically that plasma filling by hohlraum wall ablation imposes an upper bound to hohlraum X-ray production. Current simulations indicate that hohlraums used in inertial confinement fusion (ICF) are optimized to drive a fusion capsule to ignition before reaching the x-ray production limits (Dewald et al., 2005).

In the future NIF ignition experiments, the deuterium-tritium fuel capsule will be placed inside a larger hohlraum. All 192 NIF laser beams will heat the interior of the hohlraum, creating x-rays that ablate (burn off) and implode the capsule to ignition. Although the beams of high-power ultraviolet laser light only lasted a maximum of nine nanoseconds, that's considered long to ignition researches.

The Laser MegaJoule and the NIF are mainly designed for indirect drive thermonuclear fusion. Their beam configuration are optimized for x-ray drive. New studies were proposed to employ the x-ray drive configuration for direct-drive fusion. All have focused on irradiation uniformity which is a key parameter for direct-drive and its optimization has been proposed by repointing the beams and by using different pulse shapes. A new solution for high-gains direct-drive fusion with the indirect drive beam irradiation of Laser MegaJoule is presented in the paper (Canaud et al.,2007). Sources of asymmetry are due to facility imperfections such as power imbalance or pointing errors.

By focusing an enormous 1.8 megajoules of energy on a pellet of deuterium-tritium fuel a couple of millimetres in diameter, in a pulse lasting only three nanoseconds, the NIF should make the pellet implode, causing centre to „ignite" in a brief, selfsustaining fusion reaction. Fusion in the pellet could be induced either directly, by placing it in a cylindrical target, or hohlraum, about one centimetre long, and using laser pulse to induce x-rays from the hohlraum which would compress the pellet. Even short of ignition, laser experiments such as the NIF cause some fusion in their targets.

3. The role of fractals theory and fuzzy scaling for ICF design

Quite broadly in ICF design and theory, a salient role has been played by the class of solutions characterized by self-similarity. The similarity property has the highly convenient effect of reducing systems of partial differential equations depending on both space and time variables into ordinary differential equations depending on only a single self-similar variable. In ICF there exist self-similar implosions which transform uniform density solid spheres into uniform density solid spheres of arbitrarily high density.

Any practical ICF implosion begins from an initial shock. There is the inevitably non self-similar character of the flow during the transition period between the (non-self-similar) initial conditions and the (self-similar) asymptotic state. High aspect ratio implosions are the most prone to deviating from self-similarity and stagnating in a non-isochoric configuration. A balance must evidently be struck between competing objectives. The degree of target robustness to deviations in pulse shaping is typical of self-similar implosions (Clark & Tabak, 2007). In this situation are formed different kinds and shapes of fractals.

With over 50 times more energy than present facilities and the ability to produce ignition, NIF will explore new physics regimes. Ignition will allow even larger-parameter space to be accessed as well as new experimental capabilities like high flux neutron experiments. The facility contains a 192-beam Nd-glass laser system that will produce 1.8 MJ, 500 TW of 351-nm light for target experiments (Moses et al., 2008). Some experiments study the effects of microstructure in beryllium and high-density carbon on shock propagation in capsule ablators. Target design efforts continue to study diagnostic signatures in simulated experiments. One result of these studies is understanding the importance of Advanced Radiographic Capability to take high-energy x-ray radiographs of the imploding core for ignition experiments. There are a series of images of an imploding capsule with an imposed asymmetry on the x-ray drive flux. Ignition experiments have stringent requirements for laser energy, power, stability and beam conditioning.

To obtain ignition at ICF laser facilities requires an energetically efficient compression of deuterium and tritium fuel. Ideally this compression should be spherical. However, the cylindrical hohlraum and temporal profiles of the required laser shocks, due to thermodynamic and hydrodynamic constraints, cause the fuel configuration at peak compression to vary. Studies have shown this variation can depend on laser drive conditions and deviate significantly from a sphere. Neutron imaging can be useful diagnostics for determining the nature of the drive conditions (Grim et al., 2008).

Since commercially available metrology equipment is not ideally suited for certifying meso-scale capsules, several unique characterization tools have been developed. This include a very sensitive x-ray transmission radiography system for monitoring the uniformity of these coatings, and quantitative analysis methods for analysing radiographs which allow verification of the distribution (Rastovic, 2005).

One can view the Boltzmann and Vlasov-Poisson-Fokker-Planck(VPFP) equations as providing complementary physics since they both succeed and fail in complementary regimes. The Boltzmann equation gets the short distance physics correct, while the (VPFP) equation captures the long-distance physics (Rastovic, 2008).

In addition to refraction by density gradients, a variety of parametric instabilities exist that convert laser energy into internal plasma waves and scattered electromagnetic waves. Since the irradiation of the fuel pellet requires symmetry for proper implosion and since stray laser light can demage optical systems, understanding the laser-plasma interaction is of critical importance in ICF experiments. For full simulation of ignition-scale geometries and times, it is impractical to use traditional Particle-In-Cell methods (Hittinger & Dorr, 2006).

The change in grid resolution the collor grid requires a nonuniform differencing stencil in the composite grid cells adjacent to the interface.
Within the hohlraum target of indirect drive ICF experiments, the plasma does not have a uniform composition. The fractal picture of the plasma can be obtained if we consider a plasma with N distinct material regions, i.e. as multifluid model.

Piecewise linear reconstruction on these predicted cell-centeres values produces approximate predicted values at the cell interfaces.A fluid plasma model consists of a system of mass, momentum, and energy equations for each electron and ion species, coupled through a Lorentz force term to Maxwell's equations. Nevertheless, due to the wide range of spatial and temporal scales, computational laser-plasma interactions is still in need of major alghorithmic improvements, in order to simulate routinely at NIF-relevant scales.

4. Computational models of inertial confinement

We observe the development of the Rayleigh-Taylor (RT) instability whenever two fluids of different densities are accelerated against the density gradient. The unsteady anisotropic and inhomogeneous turbulent process is a fundamental problem in fluid dynamics.

The model of collisionless plasmas especially in the applied contexts of controlled fusion, and of laser fusion, is a highly idealized one. A way to incorporate collisional effects of a plasma with the background material (e.g. a plasma system in a thermal bath or reservoir) is to consider the motion of an individual particle as Brownian motion caused by collisions with the background medium.
A multi-level approach on the construction of effective partioning of unstructured graphs used in parallel computations is described. The quality of partitioning is estimated by its using in parallel iterative solution of large sparse linear systems arising in discretization of partial differential equations un unstructured grid. Various algorithms of balancing under certain constraints are considered.
Since the nineteenth century, when Boltzmann formalized the concepts of kinetic equations, their range of application has been considerably extended. They are now used also in plasma physics. They all are characterized by a density function that satisfies a partial differential equation in the phase space. Possible singularities of the solution (shock waves for instance) make the chains rule no longer available. Regularity of the solution can be proved using tools as averaging lemmas.
When considering so-called microscopic quantities one discovers that the problem undergoes really complex dynamics. The most noticeable general result is the regularization by averaging on velocities which states that for f(0) element of Lebesque p integrable functions with bounded support in v, macroscopic quantities belong to Sobolev or Besov spaces with positive numbers of derivatives.
The precise gain in regularity, and not only compactness, can be useful for regularity questions. This appears for instance in the topic of nondegenerate hyperbolic scalar balance law, where the kinetic formulation provides a method for proving regularizing effects. Historical progress in the mathematical theory of the Boltzmann equation has been the theory of Di Perna and Lions which proves global existence of weak solutions (so-called

renormalized solutions) in the physical space, i.e. using only a priori estimates. If there is a strong solution, then the normalized solution is unique.

State sensitivitis are partial derivatives describing how the state of a system changes when a design parameter is perturbed. In the context of fluid flows, these states are velocity, pressure, turbulence variables, etc. which are known approximately by a numerical solution of partial differential equations. The continuity sensitivity equation (CSE) is a natural approach to take when using adaptive methods and is useful for developing general purpose software. Both the flow and sensitivity solutions are taken into account for mesh adaptation. If an adaptive mesh generation routine is used for the flow, then the CSE only needs to be computed on the finest mesh. This allows a better control of the sensitivity solution accuracy (Turgeon et al., 2001) .

5. Description of IFC with differential equations

A reduced 1D Vlasov-Maxwell system introduced recently in the physical literature for studying laser-plasma interaction is analyzed (Carrillo & Labrunie, 2006). The electrons move under the effect of an electric field E and magnetic field B. Then, their distribution function f(t,x,v), where x denotes the position variable, is a solution to the Vlasov equation. To achive a high gain in laser thermonuclear targets, deuterium-tritium fuel should be compressed 10000-100000 times with respect to its initial density. In practice, a 100% uniformity of irradiation is impossible due to nonuniform overlaping of the beams, nonuniform amplification in the laser path, and defects in laser amplification channels.

The fields E and B are the sum of three parts:
- a.) the self-consistent fields created by the electrons
- b.) the electromagnetic field of a laser wave which is sent into the medium(called the pump wave)
- c.) The electrostatic field E(x) generated by a background of ions which are considered immobile during the time scale of the wave, and/or by an external, static confinement potential

The model of Vlasov equation and the two remaining Maxwell equations features a strongly nonlinear coupling between the kinetic and electromagnetic variables.

Two reduced models have been defined by physicist: a.) The nonrelativistic model (NR) approximates the relativistic dynamic by the Newtonian one. It is physically justified when the temperature is low enough. b.) the quasi-relativistic model (QR) is acceptable when the proportion of ultra-relativistic electrons is negligible and the pump intensity is moderate. An iterative procedure to solve the 1D Vlasov-Maxwell system for the NR and QR cases is presented.

In the paper (DiPerna & Lions, 1989) the Vlasov-Maxwell system in its classical and relativistic form is studied. The stability of solutions in weak topologies is proven and from this stability result the global existence of a weak solution with large initial data is deduced. The main tools consists of a new regularity result for velocity averages of solutions of some general linear transport equation.

We obtain compactness in L(1) under some extra hypothesis on the initial data. We first recall a regularity result about the Vlasov-Poisson-Fokker-Planck (VPFP) system involving in additional control of entropy. From a probabilistic point of view, the Fokker-Planck equation characterizes the evolution of the probability mass density of particles in phase

space, if we consider the position x(t) and the velocity of the particle v(t) as random variables which satisfies the stochastic differential equation. The VPFP system appears when we consider a great deluge of mutually interacting particles which move in a Brownian way. We know that for initial data small enough and satisfying some suitable integrability conditions, global solutions exist. The plasma region is bordered by a shock. Therefore, to analyze the beam region we return to the Vlasov – Poisson problem and introduce different scalling assumptions (Degond et al., 2003).

We present a collision potential for the Vlasov-Poisson-Boltzmann (VPB) system near vacuum in plasma physics case. This potential measures the future possible collisions between charged particles with different velocities and satisfied a time-decay estimate. The purpose of the paper (Chae et al., 2006) is to study the large-time behaviour of the VPB system via the robust Lyapunov functional D(f) measuring possible future collisions between particles. A generalized collision potential is constructed and the time-asymptotic equivalence between VPB system and linear Vlasov system is established.

In the paper (Carrillo & Toscani, 1998) is intended to study the rate of convergence of homogeneous solutions of the Vlasov-Fokker-Planck(VFP) equation. VFP equation describes a plasma in which the particle change only slightly their momentum during collision events. Assuming that the collision between heavy particles are negligible, the collisions with light particles by means of Brownian motion are approximated. The aim of this work is to study the rate of convergence of solutions. This result can be a first step to reach some results for the VFP system or the non-homogeneous case. It was done for VPFP system in the paper (Carpio, 1998). The VFP equation has been studied in the presence of a external confinant potential in [Bouchut & Dolbeault,1995] proving that the distribution of particles tend to the stationary distribution where in the expression the confinant external potential is also included. To achieve exponential decay, the smoothing effect of the Fokker-Planck term can be used.

The multiscale representation of a function can be compressed with a controlled approximation loss by setting to zero the details with an absolute values less than some given threshold depending on the level. We construct an adaptive mesh. When electromagnetic waves propagate through a plasma layer, they become parametrically unstable. Vlasov simulations provide an exellent description of the small scales of the phase-space mixing are saturated by the numerical dissipation of the numerical scheme. This category of models is motivated by the important problems of the nonlinear interaction of high intensity ultrashort laser pulses with plasmas, with specific application to particle acceleration or inertial confinement fusion purpose. Given the value of the function f at the mesh points at any given time step, we obtain the new value at mesh point. It is semi-Lagrangian Vlasov method.

The change in the fuzzy rule base is done using a variable-structure direct adaptive control algorithm to achieve the pre-defined control objectives. It has a good performance in the training phase as it makes use of initial rule base defined for the fuzzy logic stabilizer. It has a robust estimator since it depends on a variable structure technique. The adaptive nature of the bew controller significantly reduces the rule base size and improves its performance.

In the paper (Besse et al., 2008) is presented a new method for the numerical solution of the relativistic Vlasov-Maxwell system on a phase-grid using an adaptive semi-Lagrangian

method. The multiscale expansion of the distribution function allows to get a sparse representation of the data. Interaction of relativistically strong laser pulses with overdense plasma slabs is investigated. Vlasov codes are powerful tools to study wave-particle interaction with interesting results for trapping and action transfer from particles and waves. Since non-linear kinetic effects are important in laser-plasma interaction, we choose a kinetic description for the plasma.

The algorithm is based on the conservation of the flux of particles, and the distribution function is reconstructed using various techniques that allow control of spurious oscillations or preservation of the positivity. Nonetheless they might be a first step toward more effecient adaptive solvers based on different ideas for the grid refinement or on more efficient implementation.

The main way to improve the efficiency of the adaptive method is to increase the local character in phase-space of the numerical scheme by considering multiscale reconstruction with more compact support and by replacing the semi-Lagrangian method with more local-in space-numerical scheme as compact finite difference schemes, discontinuous-Galerkin method or finite element residual schemes which are well suited for parallel domain decomposition techniques. To overcome the problem of global dependency, we decompose the domain into patches, each path being devoted to a processor. One patch computes its own local cubic spline coefficients by solving reduced linear systems.

Thanks to a restrictive condition on the time step, the inter-processor communications are only done between adjencent processors, which enables us to obtain competitive results from a scalability point of view up to 64 processors. Parallelism is one of the underlying principles of the artificial neural networks. It is known that the neural networks training can be efficiently implemented on parallel computers.

6. Methods of artificial intelligence

Artificial neural networks (ANNs) are computational models implemented in software or specialized hardware devices that attempt to capture the behavioral and adaptive features of biological nervous systems. In order to solve computational or engineering problems with neural networks, learning algorithms are used to find suitable network parameters. At dissipative scales, where the fluid flow is differentiable, the phase-space density of particles is supported an a dynamically evolved fractal set. This attractor is characterized by a non-trivial multiscaling properties. Evolutionary algorithms provide an interesting alternative, or complement, to the commonly used learning algorithms, such as back-propagation. Instead of using a conventional learning algorithm, the characteristics of neural networks can be encoded in artificial genomes and evolved according to a performance criterion.
The evolutionary synthesis of a neural network leads to several design choices. Recent benchmark experiments with evolution strategies, which use a floating-point representation of the synaptic weights, have reported excellent performance with direct encoding of a small, fixed architecture. The topology of a neural network can significantly effect its ability to solve a problem. Direct encoding is typically applied to fixed network topologies, however, it can also be used to evolve the architecture of an ANN (Floreano et al., 2008). Hybrid approaches have attracted considerable attention in the Computational Intelligence community. One of the most popular approaches is the hybridization between fuzzy logic

and genetic algorithms (Herrera, 2008). It is all connected with the phenomena of adaptive behaviour.

Although much effort has been devoted to the fuzzy scaling factors in the past decades, there is still no effective solution. Most of the research works on fuzzy logic controllers have either neglected this issue by directly applying a set of scaling factors (Chopra et al., 2008).We shall apply it to the nonlinear Vlasov-Fokker-Planck equation. The advantage of using the description with IF-AND-THEN rules for VPFP equations is obtaining the possibility for description of anisotropic turbulence. Genetic algorithms can additionally help when there are problems with delays. Theoretical description of non-equilibrium transport is a challenging problem due to singular aspects of governing equations.

Consider the following IF-THEN rules:

IF m(1) is M(i1) and.... IF m(p) is M(ip)
THEN (df/dt)= (A(i) + E(i))f(t) +A(di)f(t-h) + B(i)u(t) + G(i)w(t),
i=1,2,....k   (1)

where M(i,j) are fuzzy sets and m(1),... m(p) are given premise variables, E(i) are the uncertain matrices and Gw(t) is stochastic control. The fuzzy system is hence given by sum of equations

(df/dt) = a(i) ((A(i) + E(i))f(t) + A(di)f(t-h) + B(i)u(t) +G(i)w(t)),   i=1,2...k   (2)

where a(i) are the fuzzy basis functions. The fractional integration is changed by integration on fractals. On these equations if possible apply the standard methods of stochastic control, for example the Monte Carlo method, the method of Riccati equations etc. The delay-dependent condition show the robust stabilizability for some parameters. On such a way are obtained the control systems with fuzzy scaling. With delay expressions is given the influence of genetic algorithms. Under some conditions the system is robustly stable by the Lyapunov theorem (Rastovic, 2009).In the case of instabilities the natural assumption is that the process must be recurrent. Recent technical progress allows to investigate, both experimentally and numerically, the correlation properties of fully developed turbulence in terms of particle trajectories. Non-equilibrium turbulent processes are anisotropic, non-local, multi-scale and multi-phase, and ofted are driven by shocks or acceleration. Their scaling, spectral and invariant properties differ substantially from those of classical Kolmogorov turbulence.

In the paper (Zielinski, 2005) are investigated the problems of computing the interval of possible values of the latest starting times and floats of activities in networks with uncertain durations modeled by fuzzy or interval numbers. There has been provided a possibilistic representation of the problem of determining the fuzzy latest starting times of activities and their floats, a difficulty connected to it has been pointed out. The complexity results for floats are presented (the computation of floats is probably intractable) and some polynomially solvable cases are described. It could be useful in the applications of Kolmogorov-Arnold-Moser theorem for describing the quasi-periodic orbits (Rastovic, 2007). The Kolmogorov-Arnold-Moser theorem uses mainly the fractals that are called Cantori as it is in the case of tokamak theory.

According to Bohr the external conditions of an experiment have to be described in the language of classical physics. The macroscopic arrangement of the measuring part can interact with individual samples of the physical system in such a way that a direct objectively traceable (macroscopic alternative) effect occurs or does not occur. Being thus based operationally on the same kind of objective facts and effects which are already familiar from classical physics, this interpretation avoids the introduction of any subjective element (like knowledge of observers, or human consciousness) into the theory (Cattaneo et al., 2004). A good approximation is a semi-transparent mirror in a path of a photon beam. No doubt this is a certain macroscopic arrangement producing a macroscopic alternative effect (either the photon reaches the plasma „yes" or it does not). In laser plasma interactions the description with fuzzy logic methods can be also be useful.

The finite element method (FEM) is one of the most used techniques for solving partial differential problems. The idea of FEM is to divide the domain of definition of the problem into small regions called elements of the mesh, where an approximation of the solution is searched.
The current numerical approach to the problem of finding the best grid is the mesh adaptation strategy. In the paper (Manevitz & Givoli, 2003), an alternative solution to this problem is obtained using soft computing methods. Fuzzy logic is used for the mesh generation process because it allows reproducing the qualitative reasoning typical of humans, by translating numerical inputs into linguistic values (such as „good", „near", „high") and by evaluating some if-then rules in parallel. Solving the Poisson problem for example, with the FEM in an „intelligent" way requires having an idea of the general behaviour of the solution over the domain. Where it will be smooth, large elements will be required, while elements will be smaller where the solution exibits great changes. FEM experts state that elements must be very small where there is a singularity in the boundary conditions.

The Navier-Stokes equation is supercritical. The nonlinearities become stronger at small distance scales, making it impossible to know (using present techniques) whether solutions remain smooth for all time. Thus it is crucial to understand the scale dependence of non-linearities in fluid mechanics. Renormalization theory, which is the systematic study of short distance limits, is one of the deepast ideas ever to appear in physics. Experience from quantum field theory suggest that we must first replace the Navier-Stokes equations with a „regularized" version, in which there is a short distance cutoff.
A „fuzzy" version of fluid mechanics would describe even larger scale motion, which averages over a fluid elements. Such a „mesoscopic" theory may be what we need to understand many physical phenomena, such as the stability of large vortices. A method that imposes a smallest possible lenght, and a largest possible wavenumber, without breaking symmetries could help us in mathematical,physical and engineering approaches to fluid mechanics.

7. Conclusion

For numerical simulations of laser-plasma interactions we can use the methods of fuzzy scaling and genetic algorithms for obtaining the possibility of description of inertial

controlled fusion phenomena. The same consequences of fractal plasma behaviour can be found as in tokamak physics (Rastovic, 2009). Only, in the case of tokamak we must use the fractals of the type of fuzzy Cantori, but in the case of inertial controlled fusion we must use, for example, the fractals of the type of circle fuzzy Koch curves, i.e. first draw polygon and then on each side of the polygon draw the first step of the Koch curve. At some time the process of self-similarity must be finished . We have got the reduced fractals for fuzzy scaling description. On such a way we obtain the main directions of the particles movements and the possibility for numerical calculations. In different directions of intersections of the sphere should be taken the appropriate polygons.

8. References

1. Dean, S. O. (2008). The rationale for expanded inertial fusion energy program, J. Fusion Energy,27,3,149-153, ISSN 0164-0313
2. Dewald, E. L. ; Suter, L. J. ; Landen, O. L. et al. (2005). Radiation-driven hydrodynamics of high-Z hohlraum on the National Ignition Facility, Phys. Rev. Letters,95,21,215004-215007, ISSN 0031-9007
3. Canaud, B. ; Garaude, F. ; Clique C. et al. (2007). High-gain direct-drive laser fusion with indirect drive beam layout of Laser Megajoule,Nucl. Fusion, 47,12,1652-1655, ISSN 0029-5515
4. Clark, D. S. ; Tabak, M. (2007). A self-similar isochoric implosion for fast ignition, Nucl. Fusion, 47, 9, 1147-1156, ISSN 0029-5515
5. Moses, E. I. (2008). Ignition on the National Ignition Facility,Journal of Physics: Conf. Series, 112, 3, 012003-012008, ISSN 1742-6596
6. Grim, G. P. ; Bradley, P. A. ; Day, R. D. et al. (2008). Neutron imaging development for MegaJoule scale Inertial Confinement Fusion experiments, Journal of Physics: Conf. Series, 112, 3, 032078-032081, ISSN 1742-6596
7. Rastovic, D. (2005). Transport theory and systems theory, Nucl. Technology & Radiation Protection, 20, 1, 50-58, ISSN 0164-0313
8. Rastovic, D. (2008). Fractional Fokker-Planck equations and artificial neural networks for stochastic control of tokamak, J. Fusion Energy, 27, 3, 182-187, ISSN 0164-0313
9. Hittinger, J. A. F. ; Dorr, M. R. (2006). Improving the capabilities of a continuum laser plasma interaction code, Journal of Physics:Conf. Series 46, 1, 422-432, ISSN 1742-6596
10. Turgeon, E. ; Pelletier, D. ; Borggaard, J. (2001). Sensitivity and uncertainty analysis for variable property flows,39th AIAA Aerospace Sci. Meeting and Exibit, Reno, NV, Jan. , AIAA Paper 2001-0139
11. Carrillo, J. A. ; Labrunie, S. (2006). Global solutions for the one-dimensional Vlasov-Maxwell system for laser-plasma interaction, Math. Models Meth. Appl. Sci. , 16, 1, 19-57, ISSN 0170-4214
12. DiPerna, R. J. ; Lions P. L. (1989). Global weak solutions of Vlasov-Maxwell systems, Comm. Pure Appl. Math. , 42, 6, 729-757, ISSN 0010-3640
13. Degond, P. ; Parzani, C. ; Vignal, M. H. (2003) Plasma expansion in vacuum:modeling the breakdown of quasi neutrality, Multiscale Model. Simul. , 2, 1, 158-178, ISSN 1540-3459

14. Chae, M. ; Ha, S. Y. ; Hwang, H. J. (2006) Time-asymptotic behavior of the Vlasov-Poisson-Boltzmann system near vacuum, J. Diff. Equations, 230, 1, 71-85, ISSN 0022-0396

15. Carrillo, J. A. ; Toscani G. (1998). Exponential convergence toward equilibrium for homogeneous Fokker-Planck-type equations, Math. Meth. Appl. Sci. , 21, 12, 1269-1286, ISSN 0170-4214

16. Carpio, A. (1998). Long-time behaviour for solutions of the Vlasov-Poisson-Fokker-Planck equation, Math. Meth. Appl. Sci. , 21, 11, 985-1014, ISSN 0179-4214

17. Bouchut, F. ; Dolbeault, J. (1995). On long asymptotic of the Vlasov-Fokker-Planck equation and of the Vlasov-Poisson-Fokker-Planck system with coulombic and newtonian potentials, Diff. Integ. Eq. , 8, 3, 487-514, ISSN 0893-4983

18. Besse, N. ; Latu, G. ; Ghizzo, A. et al. (2008). A wavelet-MRA-based adaptive semi-Lagrangian method for the relativistic Vlasov-Maxwell system, Jour. Comp. Phys. , 227, 16, 7889-7916, ISSN 0021-9991

19. Floreano, D. ; Durr, P. ; Mattiussi, C. (2008). Neuroevolution: from architectures to learning, Evol. Intel. , 1, 1, 47-62, ISSN 1864-5909

20. Herrera, F. (2008). Genetic fuzzy systems: taxonomy, current research trends and prospects, Evol. Intel. , 1, 1, 27-46, ISSN 1864-5909

21. Chopra, S. ; Mitra, R. ; Kumar, V. (2008). Auto tuning of fuzzy PI type controller using fuzzy logic, Int. Jour. Computational Cognition, 6, 1, 12-18, ISSN 1542-8060

22. Rastovic, D. (2009). Fuzzy scaling and stability of tokamaks, J. Fusion Energy, 28, 1, 101-106, ISSN 0164-0313

23. Zielinski, P. (2005). On computing the latest starting times and floats of activities in an network with imprecise duration, Fuzzy Sets and Systems, 150, 1, 53-76, ISSN 0165-0114

24. Rastovic, D. (2007). On optimal control of tokamak and stellarator plasma behaviour, Chaos, Solitons & Fractals, 32, 2, 676-681, ISSN 0960-0779

25. Cattaneo, G. ; Dalla Chiara, M. L. ; Giuntini, R. et al. (2004). An unsharp logic from quantum computation, Int. J. Theor. Phys. , 43, 7-8, 1803-1817, ISSN 0020-7748

26. Manevitz, L. ; Givoli, D. (2003). Automating the finite element method:a test-bed for soft computing, Appl. Soft Computing Jour. , 3, 1, 37-51, ISSN 1568-4946

27. Rastovic, D. (2009). Fractional variational problems and particle in cell gyrokinetic simulations with fuzzy logic approach for tokamaks, Nuclear Technology & Radiation Protection, 24, 2, 138-144, ISSN 1452-8185

Efficient Stochastic Simulation to Analyze Targeted Properties of Biological Systems

Hiroyuki Kuwahara[1], Curtis Madsen[2], Ivan Mura[3],
Chris Myers[2], Abiezer Tejeda[4] and Chris Winstead[4]
[1]*Carnegie Mellon University, U.S.A.*
[2]*University of Utah, U.S.A.*
[3]*Microsoft Research - University of Trento CoSBi, Italy*
[4]*Utah State University, U.S.A.*

1. Introduction

Randomness in gene expression has been ubiquitously observed from primitive prokaryotes to higher eukaryotes (Elowitz et al., 2002; Johnston & Desplan, 2010; Losick & Desplan, 2008; Maheshri & O'Shea, 2007; Raj & van Oudenaarden, 2008; Raser & O'Shea, 2004; Wernet et al., 2006). Yet, it is widely recognized that living organisms have evolved to control and exploit such underlying stochastic noise to optimize their dynamic characteristics to better cope with and compete in their living environments in a variety of ways (Arkin & Fletcher, 2006). For example, on the one hand, a stochastic switch has been shown to probabilistically regulate expression of an adhesive virulence factor in the main causative pathogen of uncomplicated lower urinary tract infections based on ambient temperature (Gally et al., 1993). On the other hand, marine embryo development seems to work much more deterministically and reliably at different rates over a range of environmental conditions (Istrail et al., 2007). Gaining insights into how living organisms control stochastic effects to achieve specific functions, thus, can have a significant implication in enhancing many aspects of our lives. That is, for example, by understanding the control mechanisms associated with the etiology and stability of complex non-Mendelian diseases, novel and effective therapies for prevention and treatment of such diseases can be developed. However, owing to sheer-size complexity of even a relatively simple biological system, elucidation of stochastic control mechanisms at the molecular level may not be something which can be efficiently and effectively accomplished with the current limitation of controllability and observability in wet-lab experiments alone. This, in turn, makes computational analysis essential to any efforts aimed at understanding of control and information processing mechanisms of intricate biological systems.

Detailed-level stochastic effects in biological systems are—by and large—captured and analyzed by devising the *stochastic chemical kinetics* (SCK) framework (Samoilov & Arkin, 2006). Assuming that the system is spatially homogeneous, the SCK model specifies the time-homogeneous probabilistic reaction rate function of each reaction and discrete changes in the molecular species populations through individual discrete reaction events. While sample trajectories of a SCK model can be accurately realized via Gillespie's *stochastic simula-*

tion algorithm (SSA) (Gillespie, 1976; 1977), the computational requirements of the SSA can be substantial due largely to the fact that it not only requires a potentially large number of simulation runs in order to estimate the system behavior at a reasonable degree of statistical confidence, but it also requires every single elementary-reaction event to be simulated one at a time. As recent advances in experimental techniques have enabled us to unveil more key components and more detailed organization structures of many biological systems, and as we are beginning to address more complex and sophisticated biological questions than ever before, it has become increasingly clear that no single modeling and simulation method can satisfy the needs of a wide spectrum of such complex questions.

One approach to alleviate the computational requirements involved in analysis of SCK models is to speed up the simulation of individual SSA by letting go of exactness. An example of this is τ-leaping method (Gillespie, 2001), which approximates the number of firings of each reaction in a predefined interval rather than executing each reaction individually. Another example is model reduction, which abstracts away dynamically insignificant reactions or species in order to make the overall systems biology analysis more efficient (Kuwahara et al., 2010; 2006).

Another approach to accelerate the analysis of SCK model is to tailor stochastic simulations based on specific dynamical properties of interest and apply a more suitable simulation method than the standard SSA. This chapter describes two such approaches to efficiently analyze various dynamical properties of interest. The rest of this chapter is organized as follows. Section 2 briefly describes SCK and SSA. Section 3 presents a modified SSA to better quantify the normal or typical behavior. Section 4 presents another modified SSA for the analysis of rare deviant events. Section 5 presents a case study analysis of enzymatic futile cycles. Finally, Section 6 presents our conclusions.

2. Stochastic Chemical Kinetics

Stochastic chemical kinetics (SCK) is a theoretical framework that accounts for the statistics of randomly-occurring chemical reactions (Gillespie, 1976; 1977; 2005; 2007). In the SCK framework, a reaction system consists of a liquid volume, Ω, containing a population of randomly-moving molecules. The molecules represent one or more species types. The medium is typically assumed to be "well-stirred," meaning a given reactant molecule may be found at any position in the medium, and may be moving in any direction, with uniform probability. A reaction may occur whenever there is a collision among the respective reactant molecules. When a reaction occurs, the reactants are removed from Ω, and the products are added to Ω. Under the SCK framework, a reaction system's time-evolution is governed by a set of probability laws which can be deduced through combinatorial methods. Suppose a reaction system consists of N chemical species s_i, and the volume Ω contains x_i molecules of species s_i at a specific time t, for $i = 1, ..., N$. Also, suppose the system has M reactions $R_1, R_2, ..., R_M$, and each reaction R_j has a set of reactants $\mathcal{R}_j$. Finally, let r_{ij} be the number of reactants of species i that participate in reaction R_j, and let p_{ij} be the number of products of species i that are produced by reaction R_j. Let $\boldsymbol{\nu}_j$ be the change in $\mathbf{x}$ that results from the occurrence of R_j. The elements of $\boldsymbol{\nu}_j$ are given by $v_{ij} = p_{ij} - r_{ij}$.

Given these definitions, the SSA algorithm computes the following probabilities:

- $a_j(\mathbf{x}, t)\, dt =$ the probability, given state $\mathbf{x}$ at time t, that R_j occurs in a small time-interval of width dt. This is called the *propensity function* of reaction j.

- $P_0\left(\tau\,|\mathbf{x}\right)$ = the probability, given state $\mathbf{x}$ at time t, that *no reaction* occurs in the time-interval $(t, t + \tau)$. It can be shown that $P_0 = \exp\left(-\tau \sum_{j=1}^{M} a_j\right)$, hence P_0 is fully determined by the reactions' propensities.
- The product of these is called the *reaction pdf*, given by $f_R\left(\tau, j\right) dt = a_j\left(\mathbf{x}, t\right) P_0\left(\tau\,|\mathbf{x}\right) dt$. The reaction pdf expresses the probability that the next reaction is R_j, and it occurs at time $t + \tau$.

Stochastic simulation algorithms use the reaction pdf to generate a sequence of reaction events. Because $f_R\left(\tau, j\right)$ is fully determined by the propensities, we may fully characterize the system's time-evolution by computing all the a_j terms. In order to compute the a_j terms, Gillespie proposed the *fundamental hypothesis of SCK*:

$$a_j = c_j \times h_j \tag{1}$$

where

$$
\begin{aligned}
c_j &= \quad \text{the \textit{stochastic reaction constant},} & (2)\\
h_j &= \quad \text{the total number of combinations of reactants in } \mathcal{R}_j. & (3)
\end{aligned}
$$

The stochastic reaction constant c_j is closely related to the traditional reaction-rate constant k_j. It is generally possible to compute c_j from k_j. In many cases, especially with regard to genetic reaction networks, the actual reaction rates are not well known. In these cases, the c_j are estimated by making an educated guess. In some cases, careful experiments have been carried out to determine the reaction constants, but these are in the minority. As a rule of thumb, the c_j constants generally lie between 10^{-4} and 0.1 for "slow" and "fast" reactions, respectively. When the c_j are not precisely known, their estimated values may be tuned within this range to reflect the relative speed expected from each reaction.

The number of reactant combinations, h_j is found by a combinatorial analysis. If reaction R_j involves multiple distinct reactants, as in $s_1 + s_2 \to s_3$, then the number of reactant combinations is the product over the reactant populations: $h_j = x_1 \times x_2$. If R_j has multiple reactants of the same species, as in $2s_1 \to s_2$, then the number of combinations is found by the *n-choose-k* function, in this case $h_j = 0.5x_1\left(x_1 - 1\right)$. In general, the total combinations is given by a product over *n-choose-k* calculations:

$$h_j = \prod_{i \in \mathcal{R}_j} \binom{x_i}{r_{ij}}. \tag{4}$$

For example, the two-reaction system model shown in Figure 1 contains the reaction $s_2 + 2s_3 \to s_1$. The number of combinations for this reaction equals the number of s_2 molecules, times the number of pairs of s_3 molecules, i.e. $h_2 = x_2 \times 0.5x_3\left(x_3 - 1\right)$.

2.1 Gillespie's Stochastic Simulation Algorithm (SSA).

To simulate the time-evolution of a reaction network, one may use the reaction pdf to generate a sequence of random reaction events, starting from a specified initial state $\mathbf{x}\left(t_0\right)$ at start-time t_0. The simulation yields a sequence of system states $\mathbf{x}\left(t_1\right)$, $\mathbf{x}\left(t_2\right)$, ... that occur at non-uniform times t_1, t_2, ..., and so on. This sequence is referred to as a *sample path*, which represents a physically plausible sequence of reactions.

To generate a sample path, we proceed one reaction at a time. For each reaction, we generate two random numbers:

1. τ = the time to the next reaction, and

2. R_μ = the reaction that fires at time $t + \tau$.

We assume the system begins in a specified initial state $\mathbf{x}_0$ at start-time t_0. The SSA is executed by repeating three essential tasks: (1) Generate a time for the next reaction to occur, (2) Generate the reaction that occurs at that time, and (3) Change the state $\mathbf{x}$ to reflect that the reaction has occurred. Algorithm 1 implements these tasks with the proper statistics and produces a physically realistic sample path.

$$\begin{pmatrix} s_1 & + & s_2 & & & \rightarrow & s_3 \\ & & s_2 & + & 2s_3 & \rightarrow & s_1 \end{pmatrix}$$

(a) Reactions

$$\begin{aligned} h_1 &= x_1 \times x_2 \\ h_2 &= x_2 \times 0.5 x_3 (x_3 - 1) \end{aligned}$$

(b) Combinations

$$\begin{aligned} \mathbf{r}_1 &= \langle 1, 1, 0 \rangle \\ \mathbf{r}_2 &= \langle 0, 1, 2 \rangle \end{aligned}$$

(c) Reactants

$$\begin{aligned} \boldsymbol{\nu}_1 &= \langle -1, 1, 1 \rangle \\ \boldsymbol{\nu}_2 &= \langle 1, -1, -2 \rangle \end{aligned}$$

(d) State changes

Fig. 1. An example of a simple two-reaction system model.

Algorithm 1 Gillespie's stochastic simulation algorithm.

1: $t \leftarrow 0$
2: $\mathbf{x} \leftarrow \mathbf{x}_0$.
3: **while** $t < t_{max}$ **do**
4:     **for** $j = 1$ to M **do**
5:         evaluate $h_j(\mathbf{x})$ and $a_j(\mathbf{x})$.
6:     **end for**
7:     Calculate a_0
8:     $r_1 \leftarrow$ a randomly generated number from $\mathcal{U}(0, 1)$.
9:     $r_2 \leftarrow$ a randomly generated number, uniformly distributed in the open interval $(0, 1)$.
10:     $\tau \leftarrow \left(\frac{1}{a_0}\right) \ln\left(\frac{1}{r_1}\right)$.
11:     $\mu \leftarrow$ the least integer for which $r_2 \leq \frac{1}{a_0} \sum_{j=1}^{\mu} a_j$. Then R_μ is the next reaction that occurs
        at time $t + \tau$.
12:     $\mathbf{x} \leftarrow \mathbf{x} + \boldsymbol{\nu}_\mu$.
13:     $t \leftarrow t + \tau$.
14: **end while**

2.2 Approximations of the SSA

A number of researchers have devised methods to reduce the computational complexity of stochastic simulations. In most cases, these methods rely on approximations that are valid

under a restricted set of reaction conditions. Two of the best-known SSA approximations were devised by Gillespie, the τ-leaping method and the chemical Langevin equation (CLE) method (Gillespie, 2001). These methods greatly improve the speed of stochastic simulations in many types of reactions. These methods also help to establish the incremental SSA methods described in section 3.

2.2.1 The τ-leap Method

The τ-leap method improves simulation speed by firing many reactions at the same time. By contrast, the original SSA must compute each separate reaction individually. By computing a bundle of reactions simultaneously, the τ-leap method can run many times faster than the ordinary SSA. The τ-leap method requires a *leap condition* which is that all of the propensity functions a_j remain approximately constant during a sufficiently small time-interval τ. Under this approximation, the propensity for a given reaction R_j during the τ-window is assumed independent of other reactions that may occur during the same time-window. Let k_j be the number of times reaction R_j occurs during the time-window. Then k_j can be shown to be a Poisson distributed random variable. A sample-path can therefore be generated using the modified SSA steps shown in Algorithm 2.

Algorithm 2 The τ-leaping method.

1: $t \leftarrow 0$
2: $\mathbf{x} \leftarrow \mathbf{x}_0$.
3: **while** $t < t_{max}$ **do**
4:     **for** $j = 1$ to M **do**
5:         evaluate $h_j(\mathbf{x})$ and $a_j(\mathbf{x})$.
6:         $k_j \leftarrow$ a random number generated from the Poisson distribution $\mathcal{P}\left(a_j \tau\right)$.
7:     **end for**
8:     $\mathbf{x} \leftarrow \mathbf{x} + \sum_{j=1}^{M} k_j \boldsymbol{\nu}_j$.
9:     $t \leftarrow t + \tau$.
10: **end while**

The leap condition is most likely satisfied in systems with large molecule counts. For these systems, a single reaction produces only a small relative change in the system's state. The change in propensities is correspondingly small. There are convenient run-time methods that test the τ-leap conditions, and adaptively determine the optimal time-step (Gillespie, 2001; Gillespie & Petzold, 2003). In the context of genetic circuits, application of the τ-leap method is complicated by the low count of DNA molecules. Reactions including DNA transcription and translation may induce rapid changes in propensities across the reaction system.

2.2.2 The Chemical Langevin Equation Method

The Chemical Langevin Equation (CLE) method is a further approximation to the τ-leap method that applies in systems with very large molecule counts (Gillespie, 2000; 2001). In addition to the Leap Condition, the CLE method requires that $\tau \gg \max_j \left\{ 1/a_j \right\}$. If these conditions are satisfied, then the discrete Poisson distribution $\mathcal{P}\left(a_j \tau\right)$ approaches the continuous Gaussian (Normal) distribution $\mathcal{N}(\mu, \sigma)$ with $\mu = \sigma^2 = a_j \tau$. The τ-leap method is therefore modified slightly to produce Algorithm 3.

Algorithm 3 The CLE method.

1: $t \leftarrow 0$
2: $\mathbf{x} \leftarrow \mathbf{x}_0$.
3: **while** $t < t_{max}$ **do**
4:     **for** $j = 1$ to M **do**
5:         evaluate $h_j(\mathbf{x})$ and $a_j(\mathbf{x})$.
6:         $k_j \leftarrow$ a random number generated from the Gaussian distribution with mean $a_j\tau$ and variance $a_j\tau$.
7:     **end for**
8:     $\mathbf{x} \leftarrow \mathbf{x} + \sum_{j=1}^{M} k_j \boldsymbol{\nu}_j$.
9:     $t \leftarrow t + \tau$.
10: **end while**

The CLE method is applicable in systems where molecule counts are so large that they may be approximated as continuous values. The CLE method provides a segue between stochastic chemical kinetics and traditional deterministic reaction-rate equation (RRE) models. RRE models use continuous-valued ordinary differential equations (ODEs) to model a reaction system's time-evolution. In the limit as all $a_j\tau \to \infty$, the CLE system converges to a deterministic ODE system. Hence the τ-leap and CLE conditions provide insight into the implicit assumptions that underlie the widely used RRE methods. In systems where the τ-leap and CLE conditions are not satisfied (or are only weakly satisfied), continuous ODE models are invalid and a suitable SCK approach should be used.

3. Determining Typical Behavior

In the analysis of stochastic biochemical systems, one starts with two basic questions. First, what is the system's normal or typical behavior? Second, how robust is that behavior? These questions are especially important for custom-designed biochemical networks, such as synthetic genetic circuits. In this case, the designer is interested in verification that the system's actual behavior matches the designer's intent. This section presents the *incremental* SSA (iSSA) which sets out to answer these questions in small time-increments. This approach has some characteristics in common with the popular SPICE program for simulating electronic circuits (Nagel & Pederson, 1973). The main objective of iSSA is to provide a first-step verification solution for synthetic biochemical systems.

Stochastic simulation algorithms generally provide a single snapshot of a system's possible behavior. If the system exhibits a high level of stochastic activity, the underlying behavior may be obscured by transient "noise". In order to understand the range of typical behaviors, many simulation runs are needed. It is common practice to compute a simulation envelope with the form $\bar{\mathbf{x}} \pm \boldsymbol{\sigma}$, where $\bar{\mathbf{x}}$ and $\boldsymbol{\sigma}$ are the average and standard deviation vectors computed over K SSA sample paths. The average, $\bar{\mathbf{x}}$, is considered to be the system's typical behavior. The standard deviation, $\boldsymbol{\sigma}$, indicates the degree to which the system is expected to deviate from the typical behavior.

The direct average method is suitable for systems that are only weakly stochastic. Direct averaging is not suitable in systems that have multiple operating modes, leading to many divergent behaviors that are all "typical." Unfortunately, the majority of interesting biochemical systems, especially genetic circuits, fall into this category. For example, consider a bi-stable

system that randomly converges toward one of two states. Suppose half of all SSA sample-paths arrive at state 1, and the other half arrive at state 2. By averaging over all SSA runs, one obtains a fictitious middle state that obscures the system's true typical behaviors.

The averaging problem is further compounded in dynamic systems that switch between states, particularly if the state-transitions occur at random times. The simplest example of a dynamic multi-state system is a stochastic oscillator in which the production of some signal is alternately activated and inhibited. One such oscillator is the circadian rhythm model developed by Vilar et al. (2002). Stochastic simulation runs of the circadian rhythm are shown in Fig. 2(a), and the average over all runs is shown in Fig. 2(b). When production is activated, it stimulates a brief but intense production of an output molecule A. As the amount of A increases, it represses its own production and eventually degrades back to zero. This pattern creates "pulses" of A that occur at random times. Because the pulse times are random, they generally do not occur at the same times in different simulation runs. If a direct average is computed, the mis-aligned impulses tend to be masked by the averaging (Samad et al., 2005). The circuit's most relevant and interesting behaviors are consequently concealed by the averaging.

In order to obtain meaningful aggregate information from many SSA runs, the iSSA was proposed by Winstead et al. (2010). In the iSSA, a conventional SSA is executed K times over a short time increment. The time increment is chosen such that the circuit's state changes slightly during the increment, similar to the τ-leaping method described in Section 2. Statistics are gathered at the end of the time increment. A new circuit state is selected from those statistics, and the algorithm is repeated for another increment.

By computing average changes over small time increments, iSSA reveals the typical behavior occurring in each increment. The results of iSSA are stochastic, and repeated iSSA simulations may yield different results. For example, iSSA may be used to simulate a bi-stable stochastic system. The iSSA method follows a cluster of sample paths that are close to each other, and hence tends to arrive at one of the two stable states.

3.1 iSSA Overview

The general steps of the iSSA are shown in Algorithm 4. The iSSA is wrapped around a core SSA algorithm, and may be specialized to perform a variety of incremental analyses. The generic iSSA works by choosing some initial condition for an SSA run, then executing the SSA over a brief interval. Lastly, the iSSA performs some analysis on the incremental SSA results before proceeding to the next increment. The physical interpretation of iSSA results depends on the particular implementation of the `select` function, which define how the SSA simulation conditions are chosen, and the `process` function, which defines how the SSA results are analyzed. This chapter assigns these functions to achieve a *marginal probability density evolution*, and it is known as iSSA-MPDE. The iSSA also allows the use of different SSA methods, such as the τ-leaping or CLE methods, when permitted by the reaction conditions. The following sections examine the derivations and conditions that apply to the iSSA.

3.2 Derivation of iSSA-MPDE

The goal of an iSSA that uses marginal probability density evolution is to provide an alternative approach that reveals the time-evolution of the statistical envelope for each species, under appropriate system conditions. The function definitions for iSSA-MPDE are given in Table 1. In essence, iSSA-MPDE approximates each species as an independent Gaussian-distributed random variable. At the start of each SSA run, the initial molecule counts are randomly generated using each species' marginal Gaussian probability distribution. After all K SSA runs
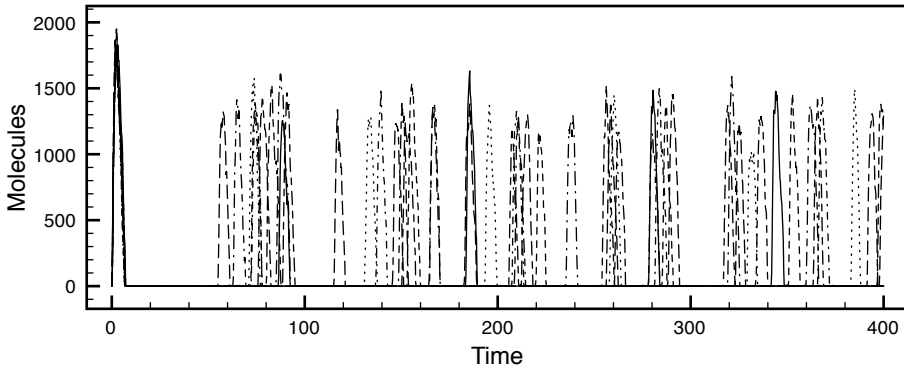
(a)
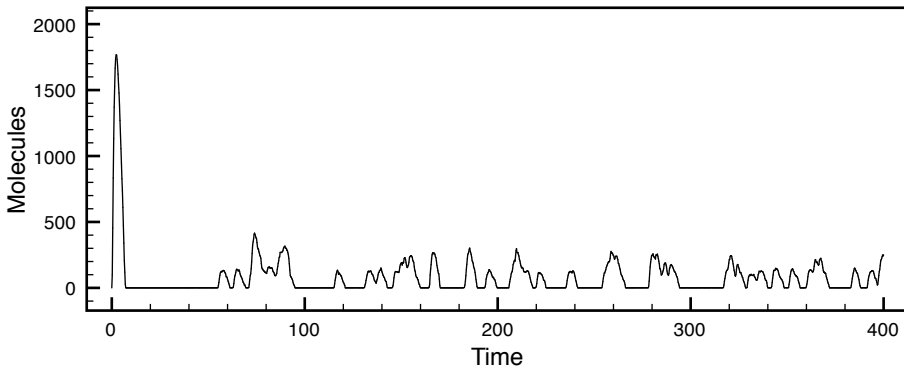
(b)

Fig. 2. (a) SSA simulations of a stochastic oscillator. (b) The average response over all SSA sample-paths, revealing incoherent results due to misaligned SSA events.

Algorithm 4 The general iSSA framework.

1: $t \leftarrow 0$
2: `initialize` the state-information structure S using initial state $\mathbf{x}_0$.
3: **while** $t < t_{max}$ **do**
4:     **for** $k = 1$ to K **do**
5:         `select` a state $\mathbf{x}$ based on the state-information S.
6:         perform one SSA run with start time t, max-time $t + \tau$ and initial state $\mathbf{x}$.
7:         `record` the ending SSA state $\mathbf{x}'$ by appending it to a state-table $\mathbf{X}'$.
8:     **end for**
9:     `process` the state-table $\mathbf{X}'$ to obtain a new state-information structure S.
10:     $t \leftarrow t + \tau$.
11: **end while**

---

are complete, the marginal distributions are estimated by computing the mean and variance for each species. The iSSA-MPDE follows the system's envelope as it evolves from increment to increment, providing an indication of the system's stochastic stability. If the standard deviation remains small relative to the mean, then the envelope may be regarded as a robust indicator of typical behavior.

`struct S:`	contains a mean vector $S.\boldsymbol{\mu}$ and a standard-deviation vector $S.\boldsymbol{\sigma}$.
`initialize:`	$S.\boldsymbol{\mu} \leftarrow \mathbf{x}_0$ for a given initial state $\mathbf{x}_0$, and $S.\boldsymbol{\sigma} \leftarrow \mathbf{0}$.
`select:`	for each species s_j, generate a noise value n_j from the distribution $\mathcal{N}\left(0, S.\sigma_j^2\right)$, and set $x_j \leftarrow S.\mu_j + n_j$.
`record:`	store the kth SSA ending state as $\mathbf{x}_k'$, for $k = 1, ..., K$.
`process:`	compute the sample means and sample variances over all $\mathbf{x}_k'$, and store the results in $S.\boldsymbol{\mu}$ and $S.\boldsymbol{\sigma}$, respectively.

Table 1. Function definitions for iSSA-MPDE.

iSSA-MPDE is derived from the CLE method discussed in Sec. 2, and inherits the τ-leap and CLE conditions[1]. To derive iSSA-MPDE, consider applying the CLE method over a short time-increment τ, beginning at time t with a fixed initial state $\mathbf{x}_0$. At time $t + \tau$, the CLE method returns a state $\mathbf{x}' = \mathbf{x}_0 + \sum_{j=1}^{M} \boldsymbol{\nu}_j$, where each $\boldsymbol{\nu}_j$ is a vector of Gaussian-distributed random values. Because the sum of Gaussians is also Gaussian, the ending state $\mathbf{x}'$ must have a joint Gaussian distribution. Then the distribution of $\mathbf{x}'$ is fully characterized by its mean $\boldsymbol{\mu}$ and its covariance matrix $\boldsymbol{\Gamma}$.

Jointly Gaussian distributions are well understood, and the reaction system's time-evolution can be simulated as the evolution of $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$ using the iSSA function definitions shown in Table 2. We refer to this algorithm as *Gaussian probability density evolution* or iSSA-GPDE. A further simplification is possible if the system is represented as a *linear Gaussian network* (LGN), with the form

$$\mathbf{x}' \approx \mathbf{A}\mathbf{x} + \mathbf{n}, \tag{5}$$

where $\mathbf{A}$ is a linear state-transformation matrix and $\mathbf{n}$ is a vector of zero-mean correlated noise with distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$. This representation is very close to the linear increment approximation used in general-purpose ODE simulators, including SPICE. The linear Gaussian model provides an intuitively convenient "signal plus noise" representation that is familiar to designers in many disciplines, and may be useful for the design and analysis of biochemical systems.

The computational complexity of this method can be significantly reduced by computing only the marginal statistics, rather than the complete covariance matrix. To compute the marginal statistics, only the diagonal entries of the covariance matrix are computed. Ignoring the remaining terms in $\boldsymbol{\Gamma}$ neglects the statistical dependencies among species in the system. To see when this is allowed, let us examine the system's dependency structure using a Bayesian network model, as shown in Fig. 3. The Bayesian network model contains a column of nodes for each time-index. Within each column, there is a node for each species. Two nodes are

---

[1] It is possible to apply iSSA-MPDE under a less restrictive set of conditions, but doing so requires a collection of refinements to the method that are beyond the scope of this chapter.

struct *S*:	contains a mean vector $S.\boldsymbol{\mu}$ and a covariance matrix $S.\boldsymbol{\Gamma}$.
initialize:	$S.\boldsymbol{\mu} \leftarrow \mathbf{x}_0$ for a given initial state $\mathbf{x}_0$, and $S.\boldsymbol{\Gamma} \leftarrow \mathbf{0}$.
select:	generate a correlated noise vector $\mathbf{n}$ from the distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$, and set $\mathbf{x} \leftarrow S.\boldsymbol{\mu} + \mathbf{n}$.
record:	store the kth SSA ending state as $\mathbf{x}'_k$, for $k = 1, ..., K$.
process:	compute the sample mean and sample covariance matrix over all $\mathbf{x}'_k$, and store the results in $S.\boldsymbol{\mu}$ and $S.\boldsymbol{\Gamma}$, respectively.

Table 2. Function definitions for iSSA-GPDE.

connected by an edge if there is a statistical dependency between them. The structure of the Bayesian network is determined by the system's *information matrix*, $\mathbf{J} = \boldsymbol{\Gamma}^{-1}$. An edge (and hence a dependency) exists between nodes x'_a and x'_b if and only if the corresponding entry j_{ab} in J is non-zero (Koller & Friedman, 2009). If $\mathbf{J}$ is approximately diagonal (i.e. if all non-diagonal entries are small relative to the diagonal ones), then the network model contains no edges between any pair $\mathbf{x}'_a$, $\mathbf{x}'_b$. This means that the *marginal* statistics of $\mathbf{x}'$ are fully determined by the statistics of $\mathbf{x}$. This allows for the joint Gaussian probability distribution at time $t + \tau$ to be approximated as a product of marginal Gaussian distributions. Instead of computing the complete covariance matrix $\boldsymbol{\Gamma}$, it is sufficient to compute the diagonal vector $\boldsymbol{\sigma}$. By computing only marginal statistics in iSSA-GPDE, iSSA-MPDE is obtained, with function definitions shown in Table 1.
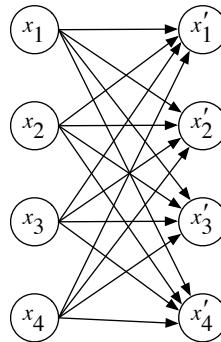
Fig. 3. A linear Gaussian Bayesian network model for a reaction system with four species. Edges in the graph indicate statistical dependencies.

3.3 Conditions and Limitations of iSSA-MPDE

iSSA-MPDE can be interpreted as an instance of belief propagation, with the SSA serving as a Monte Carlo estimate of the species' conditional distributions. When the iSSA-MPDE network is continued over several increments, the corresponding network model is extended, as shown in Fig. 4. When the network is extended in time, loops appear. Some example

loops are indicated by bold edges in Fig. 4. Strictly speaking, belief propagation (and hence iSSA-MPDE) is exact when applied to loop-free Bayesian networks.

(a)

(b)
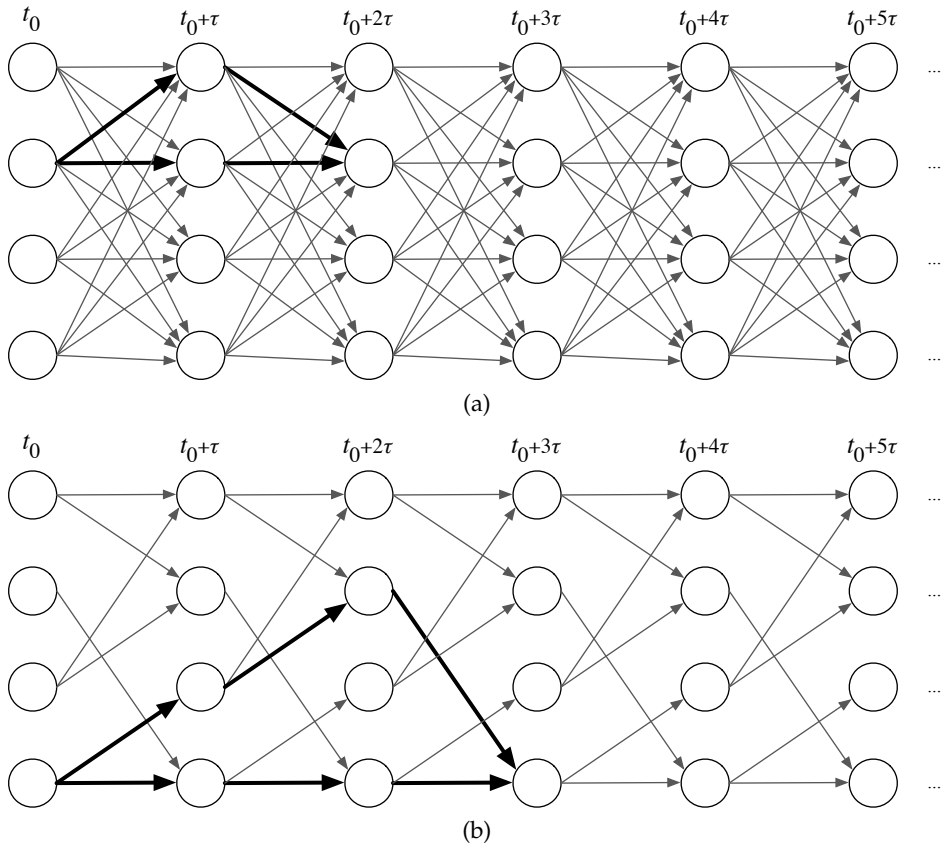
Fig. 4. Loops form when the model is unwrapped across time. (a) A dense reaction model has many short loops. (b) A sparse reaction model has fewer loops, and a larger minimum loop girth.

Loops are unavoidable in reaction network models. As a consequence, iSSA-MPDE corresponds to *loopy* belief propagation, which yields inexact statistical results. Although loopy belief propagation is inexact, it has been shown to provide a close approximation in many application areas (Murphy et al., 1999). The method's accuracy depends on the number of short loops that appear in the graph. An example of a loopy graph is shown in Fig. 4(a). In this graph, there are many loops that allow statistical information to propagate back on top of itself, which distorts the information. A better case is shown in Fig. 4(b), in which there are fewer loops. The highlighted loop in Fig. 4(b) contains six edges. This number is referred to as the loop's girth.

As a general rule, the exactness of loopy belief propagation improves when the minimum loop girth is large. iSSA-MPDE is consequently expected to yield more accurate results for systems with sparse dependencies, as in Fig. 4(b). In networks with dense dependencies, as in Fig. 4(a), iSSA-MPDE may yield distorted results. Large networks of simple reactions (where each reaction contains a small number of reactants and products) tend to be sparse in their dependencies. There are a growing number of *abstraction* methods that reduce the number of effective reactions in a large system and improve the efficiency of simulation. When a system is abstracted in this way, the density of dependencies is unavoidably increased. iSSA-MPDE, therefore, tends to be less attractive for use with abstracted simulation models (Kuwahara et al., 2010; 2006).

3.4 Resolving Variable Dependencies in iSSA-MPDE

In its most basic form, as presented in Table 1, iSSA-MPDE cannot be applied to many important types of reaction systems. This is because many systems have tightly-correlated species which prevent the information matrix from being diagonal. Strong correlations typically arise from conservation constraints, in which the state of one species is completely determined by other states in the system. This section presents a method to identify conservation constraints and correct for their effects in iSSA-MPDE. By resolving conservation constraints, the limitations on iSSA-MPDE can be relaxed considerably, allowing the method to be applied in a broader array of reaction systems.

The circadian rhythm model provides an immediate example of a system with conservation constraints. In this model, the signal molecule A is produced from gene a via transcription/translation reactions. The activity of gene a may be altered by the presence of a repressor molecule R. Hence gene a may be associated with two chemical species, a and a_R, which represent the gene's active and repressed states, respectively. The two states may be represented as distinct species governed by two reactions:

$$a + R \quad \rightarrow \quad a_R, \tag{6}$$

$$a_R \quad \rightarrow \quad a + R \tag{7}$$

In the first of these reactions, the activated gene a is consumed to produce the repressed gene a_R. In the second reaction, the repressed gene is consumed to produce the activated state. At any given time, the gene is in exactly one state. This induces a conservation constraint expressed by the equation $a + a_R = 1$. Since iSSA-MPDE treats a and a_R as independent species, it likely produces states that violate this constraint.

The conservation problem can be resolved if the method is made aware of conservation constraints. Once the constraints are determined, the system may be partitioned into *independent* and *dependent* species. iSSA-MPDE is then executed only on the independent species. The dependent species are determined from the independent ones. This partitioning can be computed automatically at run-time by evaluating the system's stoichiometric matrix, as explained below.

The stoichiometric matrix embodies the network topology of any biochemical system. Several researchers have developed methods for extracting conservation constraints from the stoichiometric matrix (Reder, 1988; Sauro & Ingalls, 2004; Schuster et al., 2002). This section briefly summarizes these techniques and applies them to iSSA-MPDE.

The stoichiometric matrix **N** is defined as follows. If a given reaction network is composed of N species and M reactions, then its stoichiometric matrix is an $M \times N$ matrix in which element

a_{ij} equals the net change in species j due to reaction i. In other words, the columns of $\mathbf{N}$ are the state-change vectors $\boldsymbol{\nu}_j$, as defined in Sec. 2.

$$\mathbf{N} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M,1} & a_{M,2} & \cdots & a_{M,N} \end{pmatrix}$$

Conserved cycles in a chemical reaction network appear as linear dependencies in the row dimensions of the stoichiometric matrix. In systems where conservation constraints appear, the sum of the conserved species must be constant. For example, consider a conservation law of the form $s_1 + s_2 = k$ for some constant k. This law dictates that the rate of appearance of s_1 must equal the rate of disappearance of s_2. Mathematically, this condition is expressed as

$$\frac{dS_1}{dt} + \frac{dS_2}{dt} = 0 \tag{8}$$

When conservation relationships are present in a biochemical network, there are linearly dependent rows in the stoichiometric matrix. Following the notation in Sauro & Ingalls (2004), one can partition the rows of $\mathbf{N}$ into two sections, $\mathbf{N_R}$ and $\mathbf{N_0}$, which represent independent and dependent species, respectively. Thus, one can partition $\mathbf{N}$ as follows:

$$\mathbf{N} = \begin{bmatrix} \mathbf{N_R} \\ \mathbf{N_0} \end{bmatrix} \tag{9}$$

Since $\mathbf{N_0}$ is a function of $\mathbf{N_R}$, the concentrations of the independent species, $\mathbf{N_R}$, can be used to calculate those of the dependent species $\mathbf{N_0}$. This relationship is determined by the *link-zero* matrix, defined as the matrix $\mathbf{L_0}$ which satisfies

$$\mathbf{N_0} = \mathbf{L_0} \times \mathbf{N_R} \tag{10}$$

Equations (9) and (10) can be combined to yield

$$\mathbf{N} = \begin{bmatrix} \mathbf{N_R} \\ \mathbf{L_0 N_R} \end{bmatrix} \tag{11}$$

Equation (11) can be further reduced by combining $\mathbf{L_0}$ with an identity matrix $\mathbf{I}$ and taking $\mathbf{N_R}$ as a common factor outside of the brackets, as shown in Equation (12).

$$\mathbf{N} = \begin{bmatrix} \mathbf{I} \\ \mathbf{L_0} \end{bmatrix} \mathbf{N_R} \tag{12}$$

$$\mathbf{N} = \mathbf{L N_R}, \tag{13}$$

where $\mathbf{L} = [\mathbf{I} \quad \mathbf{L_0}]^{\mathbf{T}}$ is called the *link* matrix. For systems in which conservation relationships do not exist, $\mathbf{N} = \mathbf{N_R}$, thus $\mathbf{L} = \mathbf{I}$.

Based on this analysis, the species are partitioned into independent and dependent state vectors, $\mathbf{s}_i(t)$ and $\mathbf{s}_d(t)$, respectively. Due to the conservation laws, any change in $\mathbf{s}_i$ must be compensated by a corresponding change in $\mathbf{s}_d$, hence

$$\mathbf{s}_d(t) - \mathbf{L_0} \mathbf{s}_i(t) = \mathbf{s}_d(0) - \mathbf{L_0} \mathbf{s}_i(0), \tag{14}$$

If the initial condition is given and the link-zero matrix is known, then the dependent species can always be computed from the independent species. To compute the link-zero matrix, we observe that

$$\begin{bmatrix} -\mathbf{L_0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{N_R} \\ \mathbf{N_0} \end{bmatrix} = \mathbf{0}. \tag{15}$$

This equation reveals that $[-\mathbf{L_0 I}]$ is the left null-space of $\mathbf{N}$. There are a variety of ways to compute the null-space of a matrix, and most numerical tools have built-in functions for this purpose.

iSSA-MPDE can be applied to systems with conservation constraints if the system is suitably partitioned into independent and dependent species. The partitioning is done automatically by identifying the linearly independent rows of the stoichiometric matrix $\mathbf{N}$, which correspond to the independent species in the system. The link-zero matrix is then computed as part of the simulation's initialization. During execution of the iSSA algorithm, the MPDE method is applied only to the independent species. The dependent species are generated using (14). Using this approach, the independent species must satisfy the conditions and limitations discussed above. The dependent species only need to satisfy the conservation constraints expressed by (14).

To demonstrate the MPDE method with constraint resolution, the method was applied to the circadian rhythm model. The results are shown in Fig. 5. The results obtained using this method agree well with the pattern observed in SSA simulations. The MPDE results also reveal the typical characteristics of the circadian rhythm system, which are difficult to discern from the SSA simulation results shown in Fig. 2.
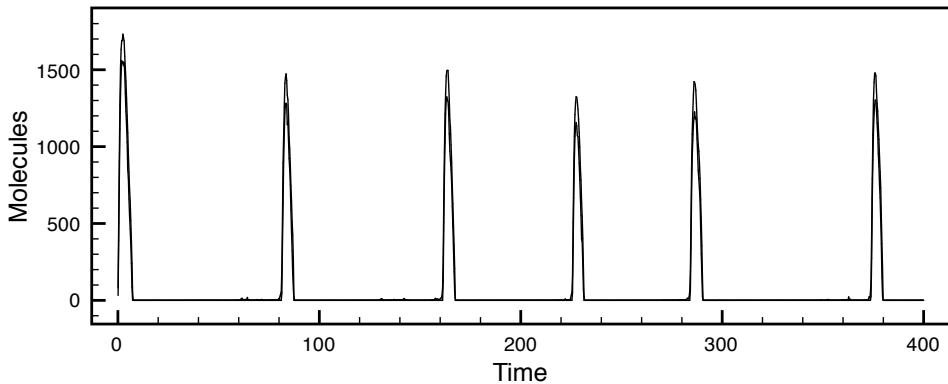
Fig. 5. The circadian rhythm model simulated using iSSA-MPDE with constraint resolution.

4. Rare Deviant Event Analysis

While the previous section discusses how to determine typical behavior, this section describes a method for more efficiently determine the likelihood of rare events. In robust biological systems, wide deviations from highly controlled normal behavior may occur with extremely small probability; nevertheless, they can have significant influences and profound consequences in many systems (Csete & Doyle, 2004). This is particularly true in biochemical and

`struct S:`	contains a mean vector $S.\boldsymbol{\mu}$ and a standard-deviation vector $S.\boldsymbol{\sigma}$.	
`initialize:`	$S.\boldsymbol{\mu} \leftarrow \mathbf{x}_0$ for a given initial state $\mathbf{x}_0$, and $S.\boldsymbol{\sigma} \leftarrow \mathbf{0}$. Independent species are identified from the stoichiometric matrix $\mathbf{N}$. The link-zero matrix $\mathbf{L_0}$ is computed using (15).	
`select:`	for each *independent* species s_j, generate a noise value n_j from the distribution $\mathcal{N}\left(0, S.\sigma_j^2\right)$, and set $x_j \leftarrow S.\mu_j + n_j$. Compute the remaining dependent species using the conservation law (14).	
`record:`	store the kth SSA ending state as $\mathbf{x}'_k$, for $k = 1, ..., K$.	
`process:`	compute the sample means and sample variances for each of the independent species in $\mathbf{x}'$, and store the results in $S.\boldsymbol{\mu}$ and $S.\boldsymbol{\sigma}$, respectively.	

Table 3. Function definitions for the MPDE-iSSA method with resolved conservation constraints.

physiological systems in that, while the occurrence of biochemical events that leads to some abnormal states may be rare, it can have devastating effects. In order to study the underlying mechanisms of such rare yet catastrophic events *in silico*, computational simulation methods may become a useful tool. However, computational analysis of rare events can demand significant computational costs and, even for a relatively small SCK model, computational requirements for a rare event analysis with the SSA may exceed the power of the most current computers. This section presents a simulation method for rare event analysis called *weighted SSA* (wSSA) (Kuwahara & Mura, 2008). Section 4.1 first defines the properties of interest and their computational challenges. Section 4.2 then briefly discusses the theoretical basis of the wSSA. Finally, Section 4.3 presents the algorithm in detail.

4.1 Background
Traditionally, analysis of rare events has been associated with analysis of the first passage time distribution (Gillespie et al., 2009), and considerable attention has been directed towards making the analysis of the first passage time to reach a rare event of interest more efficient (e.g., Allen et al. (2006); Misra & Schwartz (2008)). This section formulates rare event analysis rather differently from the analysis of the first passage time in that the property of interest here is the time-bounded probability of $\mathbf{X}(t)$ reaching a certain subset of states given that the process $\mathbf{X}(t)$ starts from a different state. In other words, our objective is to analyze $P_{t \leq t_{max}}(\mathbf{X} \rightarrow \mathcal{E} \mid \mathbf{x}_0)$, the probability that $\mathbf{X}$ moves to a state in a subset states $\mathcal{E}$ within time limit t_{max}, given $\mathbf{X}(0) = \mathbf{x}_0$ where $\mathbf{x}_0 \notin \mathcal{E}$, specifically when $P_{t \leq t_{max}}(\mathbf{X} \rightarrow \mathcal{E} \mid \mathbf{x}_0)$ is very small. This type of time-bounded rare event analyses may be very useful when it comes to study of specific biological events of interest per cell generation (i.e., before protein and RNA molecules in a mother cell are partitioned via cell division).

A standard way to analyze $P_{t \leq t_{max}}(\mathbf{X} \rightarrow \mathcal{E} \mid \mathbf{x}_0)$ is to define a Boolean random variable Y such that $Y = 1$ if $\mathbf{X}(t)$ moves to some states in $\mathcal{E}$ within the time limit and $Y = 0$ otherwise. Then, the average of Y gives $P_{t \leq t_{max}}(\mathbf{X} \rightarrow \mathcal{E} \mid \mathbf{x}_0)$. Thus, with the SSA, $P_{t \leq t_{max}}(\mathbf{X} \rightarrow \mathcal{E} \mid \mathbf{x}_0)$ can be estimated by generating n samples of Y: $Y_1, \ldots, Y_n$ through n simulation runs of $\mathbf{X}(t)$, and taking the sample average: $1/n \sum_{i=1}^{n} Y_i$. Chief among the problems in this statistical approach to project the probability of a rare event is that it may require a large number of simulation

runs just to observe the first few instances of the rare event of interest. For example, the spontaneous, epigenetic switching rate from the lysogenic state to the lytic state in phage λ-infected *Escherichia coli* (Ptashne, 1992) is experimentally estimated to be in the order of 10^{-7} per cell per generation (Little et al., 1999). Thus, simulation of one cell generation via the SSA would expect to generate sample trajectories of this rare event only once every 10^7 runs, and it would require more than 10^{11} simulation runs to generate an estimated probability with a 95 percent confidence interval with 1 percent relative half-width. This indicates that the computational requirements for obtaining results at a reasonable degree of statistical confidence can be substantial as the number of samples needed for such results may be astronomically high. Furthermore, this highlights the fact that computational requirements involved in rare event analysis of even a relatively simple biological system can far exceed the ability of most computers.

4.2 Theoretical Basis of the wSSA

The wSSA (Kuwahara & Mura, 2008) increases the chance of observing the rare events of interest by utilizing the *importance sampling* technique. Importance sampling manipulates the probability distribution of the sampling so as to observe the events of interest more frequently than it would otherwise with the conventional Monte Carlo sampling. The outcome of each biased sampling is weighted by a likelihood factor to yield the statistically correct and unbiased results. Thus, the importance sampling approach can increase the fraction of samples that result in the events of interest per a given set of simulation runs, and consequently, it can efficiently increase the precision of the estimated probability. An illustrative example of importance sampling is depicted in Figure 6.

By applying importance sampling to simulation of SCK models, hence, the wSSA can substantially increase the frequency of observation of the rare events of interest, allowing reasonable results to be obtained with orders of magnitude smaller simulation runs than the SSA. This can result in a substantial increase in computational efficiency of rare event analysis of biochemical systems.

In order to observe reaction events that can lead to a rare event of interest more often, for each reaction R_j, the wSSA utilizes *predilection function* $b_j(\mathbf{x})$ to select the next reaction instead of utilizing the propensity function $a_j(\mathbf{x})$. The predilection functions are defined such that $b_j(\mathbf{x})dt$ is the probability with which, given $\mathbf{X} = \mathbf{x}$, one R_j reaction event should occur within the next infinitesimal time dt, based on the bias one might have to lead $\mathbf{X}(t)$ towards the events of interest. With the definition of predilection functions, the index of the next reaction selection is sampled with the following probability:

$$Prob\{\text{the next reaction index is } j \text{ given } \mathbf{X} = \mathbf{x}\} = \frac{b_j(\mathbf{x})}{b_0(\mathbf{x})},$$

where $b_0(\mathbf{x}) \equiv \sum_{\mu=1}^{M} b_\mu(\mathbf{x})$. To correct the sampling bias in the reaction selection and yield the statistically unbiased results, each weighted reaction selection is then weighted by the weight function:

$$w(j, \mathbf{x}) = \frac{a_j(\mathbf{x})b_0(\mathbf{x})}{a_0(\mathbf{x})b_j(\mathbf{x})}.$$

Now, consider a k-jump trajectory of $\mathbf{X}(t)$, and let $P_k(j_k, k; \cdots ; j_2, 2; j_1, 1 \mid \mathbf{x_0})$ denote the probability that, given $\mathbf{X} = \mathbf{x_0}$, the first reaction is R_{j_1}, the second reaction is $R_{j_2}, \ldots$, and the k-th
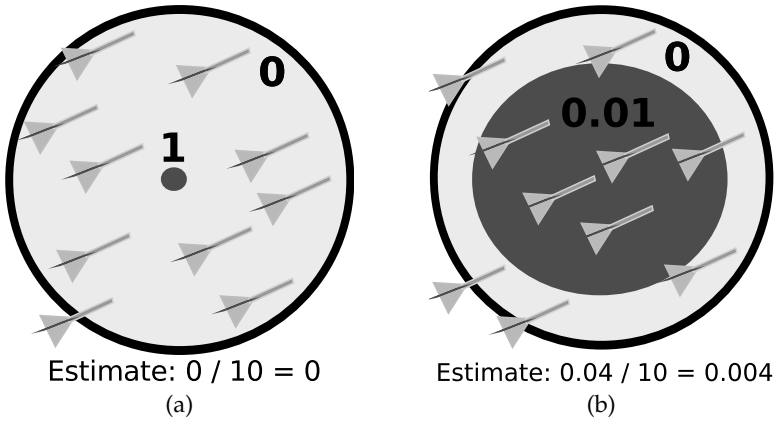
Fig. 6. An illustrative example for importance sampling. Here, the probability of hitting the area of the dart board is uniformly distributed, and the objective is to estimate the fraction of the dark grey area, which is 0.005, by throwing ten darts. (a) With the standard approach, each dart scores 1 if it hits the dark grey area and 0 otherwise. In this example, since no hit is observed in ten darts, the estimate becomes 0. (b) With the importance sampling approach, here, the dark grey area is enlarged 100 times to observe more hits and the score of the dark grey area is reduced by 100 times to correct the unbiased results. In this example, since four among the 10 darts hit the dark grey area, the estimate becomes 0.004, which is substantially closer to the true value than the original estimate.

reaction is R_{j_k}. Then, since $\mathbf{X}(t)$ is Markovian, this joint conditional probability can be expressed as follows:

$$P_k(j_k, k; \cdots; j_2, 2; j_1, 1 \mid \mathbf{x_0}) = \prod_{h=1}^{k} \frac{a_{j_h}(\mathbf{x_{h-1}})}{a_0(\mathbf{x_{h-1}})} \tag{16}$$

where $\mathbf{x_h} = \mathbf{x_0} + \sum_{h'=1}^{h-1} \mathbf{v_{j_{h'}}}$. Equation 16 can also be expressed in terms of the weight functions and the predilection functions as follows:

$$
\begin{aligned}
P_k(j_k, k; \cdots; j_2, 2; j_1, 1 \mid \mathbf{x_0}) &= \prod_{h=1}^{k} \left[\frac{a_{j_h}(\mathbf{x_{h-1}}) b_0(\mathbf{x_{h-1}})}{b_{j_h}(\mathbf{x_{h-1}}) a_0(\mathbf{x_{h-1}})} \right] \frac{b_{j_h}(\mathbf{x_{h-1}})}{b_0(\mathbf{x_{h-1}})} \\
&= \prod_{h=1}^{k} w(j_h, \mathbf{x_{h-1}}) \prod_{h=1}^{k} \frac{b_{j_h}(\mathbf{x_{h-1}})}{b_0(\mathbf{x_{h-1}})}.
\end{aligned}
\tag{17}
$$

Hence, in the wSSA, the estimate of $P_{t \leq t_{max}}(\mathbf{X} \to \mathcal{E} \mid \mathbf{x_0})$ is calculated by first defining the statistical weight of the i-th sample trajectory w_i such that

$$
w_i = \begin{cases} \prod_{h=1}^{k_i} w(j_h, \mathbf{x_{h-1}}) & \text{if } \mathbf{X}(t) \text{ moves to some state in } \mathcal{E} \text{ within the time limit,} \\ 0 & \text{otherwise,} \end{cases}
$$

where k_i is the number of jumps in the i-th sample trajectory. Then, $P_{t \leq t_{max}}(\mathbf{X} \to \mathcal{E} \mid \mathbf{x_0})$ is estimated by taking a sample average of w_i:

$$\frac{1}{n} \sum_{i=1}^{n} w_i.$$

With an adequate choice of the predilection functions, the wSSA can increase the fraction of sample trajectories that result in the rare events of interest. At the same time, it can lower the variance of the estimate by having each w_i smaller than 1.

In Kuwahara & Mura (2008), each predilection function has a restricted form in that each predilection function is proportional to the corresponding propensity function. In other words, for each reaction R_j, $b_j(\mathbf{x})$ is defined as:

$$b_j(\mathbf{x}) = \alpha_j \times a_j(\mathbf{x}), \tag{18}$$

where each $\alpha_j > 0$ is a constant. This restriction can conveniently constrain the predilection functions such that, for each $b_j(\mathbf{x})$, $b_j(\mathbf{x}) = 0$ if and only if $a_j(\mathbf{x}) = 0$, avoiding the case where a possible trajectory of a system is weighted by a factor 0. Clearly, if $\alpha_j = \alpha$ for all j, then $a_j(\mathbf{x})/a_0(\mathbf{x}) = b_j(\mathbf{x})/b_0(\mathbf{x})$. Thus, such a selection of predilection functions may not be useful. Nevertheless, the wSSA can substantially accelerate the analysis of rare events when appropriate predilection functions are used. While optimized selection schemes of the predilection functions require further investigation, it is somewhat intuitive to select predilection functions to alleviate the computational demands in a number of cases. For example, suppose we are interested in analyzing the probability that a species S transitions from θ_1 to θ_2 where $\theta_1 < \theta_2$. Then, most likely, increasing the predilection functions of the production reactions of S and/or decreasing the predilection functions of the degradation reactions of S—even with a small factor—would increase the fraction of the sample trajectories that result in the event of interest. Furthermore, a procedure to choose optimized α_j by running several test runs to compute the variance of the statistical weights has been proposed (Gillespie et al., 2009). However, much work remains to be done in order to more practically select predilection functions.

4.3 Algorithm of the wSSA

Algorithm 5 describes the procedure to estimate $P_{t \leq t_{max}}(\mathbf{X} \to \mathcal{E} \mid \mathbf{x_0})$ with n simulation runs of the wSSA. Note that, while Algorithm 5 is presented in a similar fashion as the counterpart direct method of the SSA, various optimization techniques of the direct method, such as Cao et al. (2004); McCollum et al. (2006), can also be applied to an implementation of the wSSA to further reduce the simulation cost. Furthermore, model abstraction techniques such as Kuwahara & Myers (2007) can be incorporated to further accelerate the simulation process.

First, the algorithm initializes to 0 the variable q, which accumulates statistical weights of each successful sample trajectory (line 1). Then, it generates n sample trajectories of $\mathbf{X}(t)$ via the wSSA. For each simulation run, the initialization is first performed to set the weight of each sample trajectory, w, the time, t, and the system state, $\mathbf{x}$ to 1, 0, and $\mathbf{x_0}$, respectively (line 3). It then evaluates all the propensity functions $a_j(\mathbf{x})$ and all the predilection functions $b_j(\mathbf{x})$, and also calculates $a_0(\mathbf{x})$ and $b_0(\mathbf{x})$ (line 4). Each Monte Carlo simulation is run up to time t_{max}. If, however, a rare event (i.e., $\mathbf{x} \in \mathcal{E}$) occurs within t_{max}, then the current sample trajectory weight w is added to q, and the next simulation run is performed (lines 6-9). Otherwise, the waiting time to the next reaction, τ, is sampled in the same way as in the direct method of the SSA, while the next reaction R_μ is selected using the predilection functions (lines 10-12). Then,

w, t, and $\mathbf{x}$ are updated to reflect the selections of the waiting time and the next reaction (lines 13-15). Any propensity functions and predilection functions that need to be updated based on the firing of one R_μ reaction event are re-evaluated, and $a_0(\mathbf{x})$ and $b_0(\mathbf{x})$ are re-calculated (line 16). After n sample trajectories are generated via the wSSA, the probability that $\mathbf{X}(t)$ reaches some state in $\mathcal{E}$ within t_{max} given $\mathbf{X}(0) = \mathbf{x_0}$ is estimated by q/n (line 19).

Algorithm 5 Estimate of $P_{t \leq t_{max}}(\mathbf{X} \to \mathcal{E} \mid \mathbf{x_0})$ via wSSA

1: $q \leftarrow 0$
2: **for** $k = 1$ to n **do**
3:     $w \leftarrow 1, t \leftarrow 0, \mathbf{x} \leftarrow \mathbf{x_0}$
4:     evaluate all $a_j(\mathbf{x})$ and $b_j(\mathbf{x})$, and calculate $a_0(\mathbf{x})$ and $b_0(\mathbf{x})$
5:     **while** $t \leq t_{max}$ **do**
6:       **if** $\mathbf{x} \in \mathcal{E}$ **then**
7:          $q = q + w$
8:          break out of the while loop
9:       **end if**
10:      $\tau \leftarrow$ a sample of exponential random variable with mean $1/a_0(\mathbf{x})$
11:      $u \leftarrow$ a sample of unit uniform random variable
12:      $\mu \leftarrow$ smallest integer satisfying $\sum_{i=1}^{\mu} b_i(\mathbf{x}) \geq u b_0(\mathbf{x})$
13:      $w \leftarrow w \times \big(a_\mu(\mathbf{x})/b_\mu(\mathbf{x})\big) \times (b_0(\mathbf{x})/a_0(\mathbf{x}))$
14:      $t \leftarrow t + \tau$
15:      $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{v}$-
16:      update $a_j(\mathbf{x})$ and $b_j(\mathbf{x})$, and re-calculate $a_0(\mathbf{x})$ and $b_0(\mathbf{x})$
17:     **end while**
18: **end for**
19: report q/n as the estimated probability

The computational complexity of Algorithm 5 and the counterpart of the standard SSA can be compared by noticing that the multiplication/division operations in the wSSA only increases linearly. Indeed, the operation count in Algorithm 5 differs from the counterpart of the SSA only in the two steps: line 13; and line 16 inside the **while** loop. Line 13 adds a constant number of operations (i.e., 2 multiplications and 2 divisions), while line 16 includes the operations for the update of the predilection functions $b_j(\mathbf{x})$, $j = 1, 2, \ldots, M$ as well as $b_0(\mathbf{x})$. The cost of such updates depends on the specific form of the predilection functions and the network of the model. However, if, as considered in this section, the predilection functions take the form of simple scaling functions of the propensity functions, then these updates require at most M multiplications, which does not change the overall complexity of the presented simulation algorithm between the wSSA and the direct method of the SSA.

5. Case Study: Enzymatic Futile Cycles

This section presents case studies of the two simulation methods described in this chapter to illustrate the usefulness of those methods. Our case studies are based on the analysis of dynamical properties of enzymatic futile cycle models. Section 5.1 introduces the structure of an enzymatic futile cycle model. Section 5.2 shows iSSA results on the futile cycle model. Finally, Section 5.3 shows the results from wSSA-based rare event analysis on this model.

5.1 Enzymatic Futile Cycle Model

The enzymatic futile cycle is composed of two enzymatic reactions running opposite directions, and is ubiquitously seen in biological systems (Voet et al., 1999). In signaling networks, for example, this control motif can be used as a biological network building block that regulates the activity of a protein by representing a phosphorylation-dephosphorylation cycle where the forward enzymatic reaction represents the phosphorylation of a protein via a kinase or an activation of a protein via a small GTP-binding protein, while the backward enzymatic reaction represents the dephosphorylation of the protein via phosphatase (Goldbeter & Koshland, 1981). A three-layered cascade of phosphorylation-dephosphorylation cycles can form the basic structure of the mitogen-activated protein kinase cascade, which facilitates generation of a variety of responses to external stimuli and is ubiquitously seen in eukaryotes to control many biological processes including cell proliferation and apoptosis (Chang & Karin, 2001; Huang & Ferrell, 1996).

The structure of an enzymatic futile cycle model is depicted in Figure 7. This model has six species: S_1 is the enzyme to catalyze the transformation of the protein into the active form; S_2 is the inactive form of the protein; S_3 is the complex of S_1 and S_2; S_4 is the enzyme to catalyze the transformation of the protein into the inactive form; S_5 is the active form of the protein; and S_6 is the complex of S_4 and S_5 (Figure 7(a)). The model has six reactions: R_1 is the formation of S_3; R_2 is the breakup of S_3 into S_1 and S_2; R_3 is the production of S_5; R_4 is the formation of S_6; R_4 is the breakup of S_6 into S_4 and S_5; and R_6 is the production of S_2. A schematic of this model is shown in Figure 7(b).

$$R_1 : \quad S_1 + S_2 \xrightarrow{c_1} S_3$$

$$R_2 : \quad S_3 \xrightarrow{c_2} S_1 + S_2$$

$$R_3 : \quad S_3 \xrightarrow{c_3} S_1 + S_5$$

$$R_4 : \quad S_4 + S_5 \xrightarrow{c_4} S_6$$

$$R_5 : \quad S_6 \xrightarrow{c_5} S_4 + S_5$$

$$R_6 : \quad S_6 \xrightarrow{c_6} S_4 + S_2$$

(a)   (b)
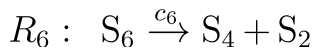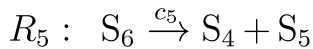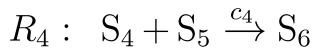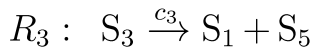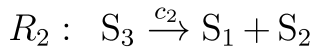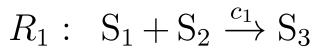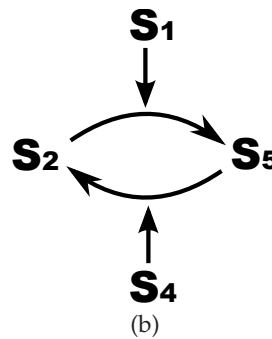
Fig. 7. The structure of an enzymatic futile cycle model. Here, S_1 is the enzyme to catalyze the transformation of S_2 into S_5, while S_4 is the enzyme to catalyze the transformation of S_5 into S_2. S_3 is the complex of S_1 and S_2. S_6 is the complex of S_4 and S_5 (a) A list of the six reactions in the model. (b) A schematic of the enzymatic futile cycle model.

5.2 Bistable Oscillation in Enzymatic Futile Cycles with Noise Driver

To demonstrate the utility of the iSSA, this section considers an enzymatic futile cycle model with a noise driver as shown in Fig. 8 (Samoilov et al., 2005). This model has the same two enzymatic reactions as the original futile cycle model but also includes a species S_7 and four

more reactions that involve S_1 and S_7 in order to simulate noise in the environment. R_7 converts S_1 into S_7; R_8 is the reverse reaction of R_7 and converts S_7 back into S_1; R_9 converts S_1 and S_7 into two S_7; and R_{10} is the reverse reaction of R_9 and converts two S_7 back into S_1 and S_7.

$$R_7: \quad S_1 \xrightarrow{c_7} S_7$$
$$R_8: \quad S_7 \xrightarrow{c_8} S_1$$
$$R_9: \quad S_1 + S_7 \xrightarrow{c_9} 2S_7$$
$$R_{10}: \quad 2S_7 \xrightarrow{c_{10}} S_1 + S_7$$

(a) (b)
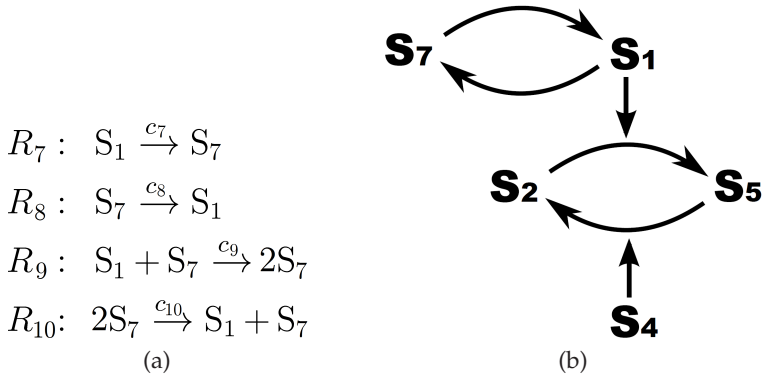
Fig. 8. Model for enzymatic futile cycle with a noise driver. Here the species S_7 has been added to introduce noise on the amount of S_1 available. This model also includes four additional reactions that convert between S_1 and S_7 molecules.

Simulation results for this model are expected to result in random symmetric oscillations of species S_2 and S_5 as depicted in the individual SSA run shown in Figure 9(a). However, Figure 9(b) shows that when 10 SSA runs are averaged together, S_2 and S_5 clearly do not exhibit this behavior and potentially leading to the conclusion that this model does not oscillate. When iSSA is applied to this model, the results reveal the expected oscillatory behavior as shown in Figures 9(c). These plots present the results for 10 runs for each time increment and a time step of 0.01. These results show that drawing conclusions from aggregate SSA statistics is problematic. The iSSA, on the other hand, aggregates stochastic run statistics in small time increments in order to produce typical behavior profiles of genetic circuits.

5.3 Rare Event Analysis in Balanced Enzymatic Futile Cycles

To illustrate the utility of wSSA, this section considers a balanced enzymatic futile cycle model and aims at evaluating $P_{t \leq 100}(X_5 \to 25 \mid \mathbf{x_0})$, the probability that, given $\mathbf{X}(0) = \mathbf{x_0}$, X_5 moves to 25 within 100 time units. In this study, the initial state of the enzymatic futile cycle model is given by

$$X_1(0) = X_4(0) = 1; X_2(0) = X_5(0) = 50; \text{ and } X_3(0) = X_6(0) = 0,$$

and the rate constants are specified as follows:

$$k_1 = k_2 = k_4 = k_5 = 1; \text{ and } k_3 = k_6 = 0.1.$$

Because of the perfect symmetry in the rate constants as well as in the initial molecule counts of the two enzymatic reactions in this setting, $\mathbf{X}(t)$ tends to stay—with high probability—around states in which X_2 and X_5 are balanced from time 0. That is, X_2 and X_5 stay around 50.
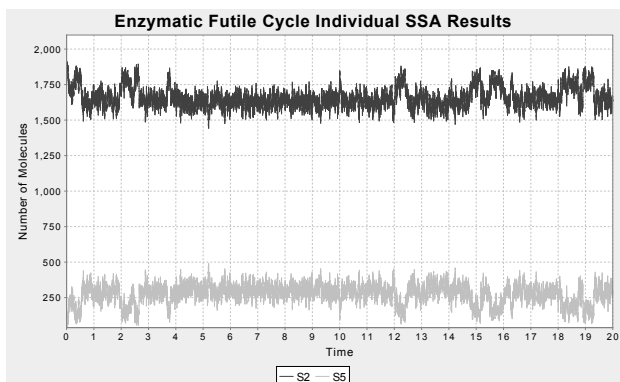
(a)
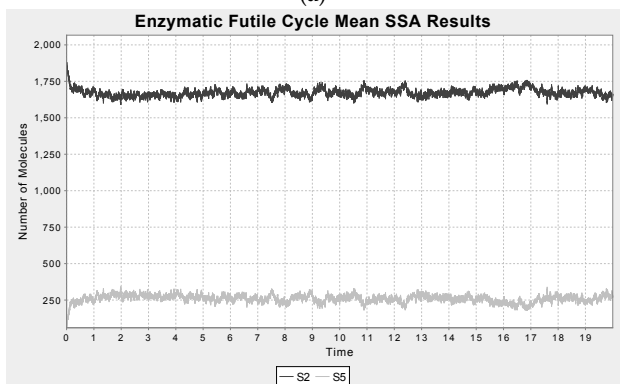
(b)
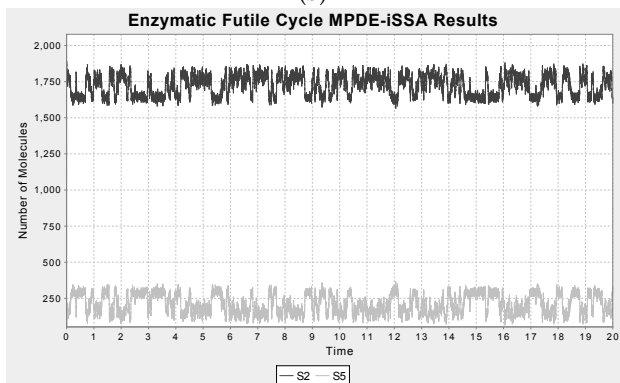
(c)

Fig. 9. SSA simulation results for S_2 and S_5 from the enzymatic futile cycle with noise driver. (a) A single SSA sample path. (b) The mean $\bar{x}(t)$ of 10 independent SSA sample paths. (c) iSSA results using 10 runs for each time increment and a time step of 0.01.

This implies that $X_5 \to 25 \mid \mathbf{x_0}$ is a rare deviant event. As the underlying Markov process has a finite and relatively small number of states, we have computed the exact value of $P_{t \leq 100}(X_5 \to 25 \mid \mathbf{x_0})$ through a numerical solution, which in turn serves as the measure to compare the accuracy of the wSSA and the SSA.

In order to increase the fraction of simulation runs that reach of the states of interest in the wSSA for this analysis, the following predilection functions are used:

$$b_j(\mathbf{x}) = \begin{cases} a_j(\mathbf{x}) & \text{for } j = 1, 2, 4, 5, \\ \gamma a_j(\mathbf{x}) & \text{for } j = 3, \\ \frac{1}{\gamma} a_j(\mathbf{x}) & \text{for } j = 6, \end{cases}$$

where $\gamma = 0.5$. This biasing approach discourages the forward enzymatic reaction while encourages the backward enzymatic reaction, resulting in an increase in the likelihood of X_5 to move to low count states.

Figure 10 depicts the accuracy of the estimates of $P_{t \leq 100}(X_5 \to 25 \mid \mathbf{x_0})$ via the SSA and the wSSA with respect to a number of simulation runs. In the SSA, we did not observe any simulation runs that had resulted in X_5 moving to 25 within 100 time units for the first 10^5 simulation runs, making the estimated probability 0 (Figure 10(a)). On the other hand, wSSA was able to produce a reasonable estimate in the first 100 simulation runs and, throughout, it generated an estimated probability which is in very close agreement with the true probability (Figure 10(a)). Furthermore, the relative distance of the estimate from the true value indicates that the estimate from the wSSA can converge to the true value more rapidly than that from the SSA (Figure 10(b)).
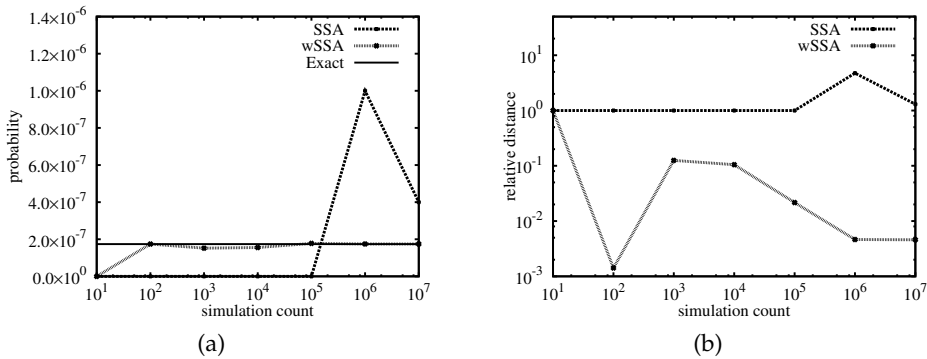
Fig. 10. Comparison of accuracy between SSA and wSSA for the estimate of $P_{t \leq 100}(X_5 \to 25 \mid \mathbf{x_0})$. (a) The estimated probability via the SSA and the wSSA with respect to a number of simulation runs. The solid line represents the true probability. (b) The relative distance of the estimated probability from the true value with respect to a number of simulation runs.

The ratio of the simulation time between the wSSA and the SSA with respect to a number of simulation runs is illustrated in Figure 11(a). This shows that, in the worst case, the run time of wSSA is about 1.2 times slower than the direct method of the SSA. However, since the wSSA achieved orders of magnitude higher accuracy in estimate of $P_{t \leq 100}(X_5 \to 25 \mid \mathbf{x_0})$ than

(a)

(b)
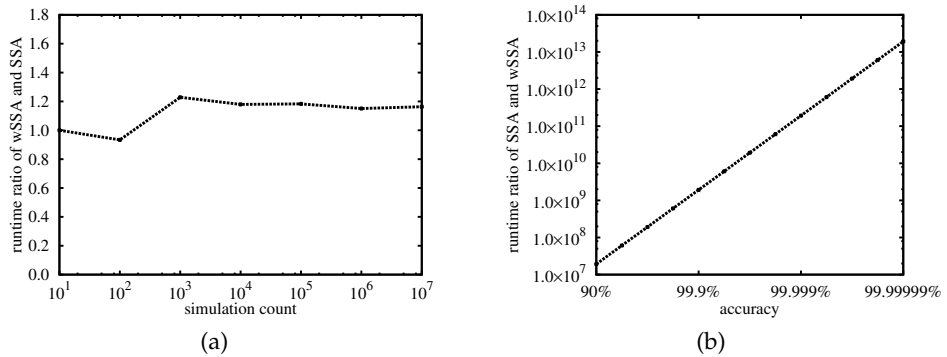
Fig. 11. Comparison of computation efficiency between SSA and wSSA for the estimate of $P_{t \leq 100}(X_5 \to 25 \mid \mathbf{x_0})$. The ratio of the simulation time of the wSSA and the SSA with respect to a number of simulation runs. (b) Ratio of SSA and wSSA computation time for a given level of accuracy.

the SSA per a given number of simulation runs, the wSSA is substantially more efficient than the SSA in computing a high precision estimate of $P_{t \leq 100}(X_5 \to 25 \mid \mathbf{x_0})$.

To better characterize the computational gain obtained with the wSSA over the SSA, we evaluated the number of runs required by SSA to achieve a given accuracy criterion ϵ where ϵ is defined as 1 minus the relative distance of the estimate from the true probability. We then estimated the number of simulation runs required by the SSA through a statistical argument based on confidence intervals (see Appendix of Kuwahara & Mura (2008) for details). By factoring in the estimated number of runs and the average run time, we computed the expected computation time of SSA for given ϵ. Figure 11(b) shows the ratio of the expected computation time between the SSA and wSSA. This illustrates a significant computational gain that is achieved via the wSSA. For instance, while the wSSA can estimate $P_{t \leq 100}(X_5 \to 25 \mid \mathbf{x_0})$ with an accuracy of 99.9999% in 1.7×10^3 seconds, the SSA would need 10^{12} times as much computational time, which is roughly 1.05×10^8 years of computation (i.e., 2.2×10^{19} simulation runs) to achieve that same level of accuracy on the same computer.

6. Conclusions

During stochastic analysis of biological systems, it is important to be able to both determine accurately and efficiently the typical behavior and the probability of rare deviant events. This chapter has introduced two new stochastic simulation algorithms, the iSSA and wSSA, to address these problems. The iSSA has been shown to produce a more stable typical behavior of an oscillatory system than aggregate statistics generated by the traditional SSA. The wSSA has been shown to produce a substantially more accurate estimate of the probability of rare deviant events as compared to same number of runs of the SSA. Taken together, these are powerful tools for the analysis of biological systems.

7. References

Allen, R. J., Frenkel, D. & ten Wolde, P. R. (2006). Forward flux sampling-type schemes for simulating rare events: Efficiency analysis, *The Journal of Chemical Physics* **124**(19): 194111.
    URL: *http://link.aip.org/link/?JCP/124/194111/1*

Arkin, A. & Fletcher, D. (2006). Fast, cheap and somewhat in control, *Genome Biology* **7**(8): 114.
    URL: *http://genomebiology.com/2006/7/8/114*

Cao, Y., Li, H. & Petzold, L. (2004). Efficient formulation of the stochastic simulation algorithm for chemically reacting system, *Journal of Chemical Physics* **121**: 4059–4067.

Chang, L. & Karin, M. (2001). Mammalian MAP kinase signalling cascades, *Nature* **410**(6824): 37–40.
    URL: *http://dx.doi.org/10.1038/35065000*

Csete, M. & Doyle, J. (2004). Bow ties, metabolism and disease, *Trends in Biotechnology* **22**(9): 446–450.

Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. (2002). Stochastic gene expression in a single cell, *Science* **297**: 1183–1186.

Gally, D. L., Bogan, J. A., Eisenstein, B. I. & Blomfield, I. C. (1993). Environmental regulation of the *fim* switch controlling type 1 fimbrial phase variation in *Escherichia coli* K-12: effects of temperature and media, *J. Bacteriol.* **175**: 6186–6193.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, *Journal of Computational Physics* **22**: 403–434.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions, *Journal of Physical Chemistry* **81**(25): 2340–2361.

Gillespie, D. T. (2000). The chemical Langevin equation, *Journal of Chemical Physics* **113**(1).

Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems, *Journal of Chemical Physics* **115**(4): 1716–1733.

Gillespie, D. T. (2005). Stochastic chemical kinetics, *in* S. Yip (ed.), *Handbook of Materials Modeling*, Springer, pp. 1735–1752.

Gillespie, D. T. (2007). Stochastic simulation of chemical kinetics, *Annual Review of Physical Chemistry* **58**(1): 35–55.

Gillespie, D. T. & Petzold, L. R. (2003). Improved leap-size selection for accelerated stochastic simulation, *Journal of Chemical Physics* **119**.

Gillespie, D. T., Roh, M. & Petzold, L. R. (2009). Refining the weighted stochastic simulation algorithm, *The Journal of Chemical Physics* **130**(17): 174103.
    URL: *http://link.aip.org/link/?JCP/130/174103/1*

Goldbeter, A. & Koshland, D. E. (1981). An amplified sensitivity arising from covalent modification in biological systems, *Proceedings of the National Academy of Sciences of the United States of America* **78**(11): 6840–6844.
    URL: *http://www.pnas.org/content/78/11/6840.abstract*

Huang, C. Y. & Ferrell, J. E. (1996). Ultrasensitivity in the mitogen-activated protein kinase cascade, *Proceedings of the National Academy of Sciences of the United States of America* **93**(19): 10078–10083.
    URL: *http://www.pnas.org/content/93/19/10078.abstract*

Istrail, S., De-Leon, S. B.-T. & Davidson, E. H. (2007). The regulatory genome and the computer, *Developmental Biology* **310**(2): 187 – 195.

Johnston, Jr., R. J. & Desplan, C. (2010). Stochastic mechanisms of cell fate specification that yield random or robust outcomes, *Annual Review of Cell and Developmental Biology* **26**(1).

URL: *https://www.annualreviews.orghttps://www.annualreviews.org/doi/abs/10.1146/annurev-cellbio-100109-104113*

Koller, D. & Friedman, N. (2009). *Probabilistic Graphical Models*, MIT Press.

Kuwahara, H. & Mura, I. (2008). An efficient and exact stochastic simulation method to analyze rare events in biochemical systems, *The Journal of Chemical Physics* **129**(16): 165101.
URL: *http://link.aip.org/link/?JCP/129/165101/1*

Kuwahara, H. & Myers, C. (2007). Production-passage-time approximation: A new approximation method to accelerate the simulation process of enzymatic reactions, *The 11th Annual International Conference on Research in Computational Molecular Biology*.

Kuwahara, H., Myers, C. J. & Samoilov, M. S. (2010). Temperature control of fimbriation circuit switch in uropathogenic *Escherichia coli*: Quantitative analysis via automated model abstraction, *PLoS Computational Biology* **6**(3): e1000723.

Kuwahara, H., Myers, C., Samoilov, M., Barker, N. & Arkin, A. (2006). Automated abstraction methodology for genetic regulatory networks, *Trans. on Comput. Syst. Biol.* **VI**: 150–175.

Little, J. W., Shepley, D. P. & Wert, D. W. (1999). Robustness of a gene regulatory circuit, *EMBO Journal* **18**: 4299–4307.

Losick, R. & Desplan, C. (2008). Stochasticity and Cell Fate, *Science* **320**(5872): 65–68.
URL: *http://www.sciencemag.org/cgi/content/abstract/320/5872/65*

Maheshri, N. & O'Shea, E. K. (2007). Living with noisy genes: How cells function reliably with inherent variability in gene expression, *Annual Review of Biophysics and Biomolecular Structure* **36**(1): 413–434.

McCollum, J. M., Peterson, G. D., Cox, C. D., Simpson, M. L. & Samatova, N. F. (2006). The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior, *Computational biology and chemistry* **30**(1): 39–49.

Misra, N. & Schwartz, R. (2008). Efficient stochastic sampling of first-passage times with applications to self-assembly simulations, *The Journal of Chemical Physics* **129**(20): 204109.

Murphy, K., Weiss, Y. & Jordan, M. (1999). Loopy belief propagation for approximate inference: An empirical study, *Uncertainty in AI*.

Nagel, L. W. & Pederson, D. (1973). Spice (simulation program with integrated circuit emphasis), *Technical Report UCB/ERL M382*, EECS Department, University of California, Berkeley.
URL: *http://www.eecs.berkeley.edu/Pubs/TechRpts/1973/22871.html*

Ptashne, M. (1992). *A Genetic Switch*, Cell Press & Blackwell Scientific Publishing.

Raj, A. & van Oudenaarden, A. (2008). Nature, nurture, or chance: Stochastic gene expression and its consequences, *Cell* **135**(2): 216–226.

Raser, J. M. & O'Shea, E. K. (2004). Control of stochasticity in eukaryotic gene expression, *Science* **304**: 1811–1814.

Reder, C. (1988). Metabolic control theory: A structural approach, *Journal of Theoretical Biology* **135**(2): 175 – 201.
URL: *http://www.sciencedirect.com/science/article/B6WMD-4KYW436-3/2/deaa46117df4b026f815bca0af0cbfeb*

Samad, H. E., Khammash, M., Petzold, L. & Gillespie, D. (2005). Stochastic modelling of gene regulatory networks, *International Journal of Robust and Nonlinear Control* **15**: 691–711.

Samoilov, M., Plyasunov, S. & Arkin, A. P. (2005). Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations, *Proceedings of the National Academy of Sciences US* **102**(7): 2310–5.

Samoilov, M. S. & Arkin, A. P. (2006). Deviant effects in molecular reaction pathways, *Nature Biotechnology* **24**: 1235–1240.

Sauro, H. M. & Ingalls, B. (2004). Conservation analysis in biochemical networks: computational issues for software writers., *Biophysical chemistry* **109**(1): 1–15.
   URL: *http://www.ncbi.nlm.nih.gov/pubmed/15059656*

Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I. & Dandekar, T. (2002). Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae* , *Bioinformatics* **18**(2): 351–361.
   URL: *http://bioinformatics.oxfordjournals.org/cgi/content/abstract/18/2/351*

Vilar, J. M. G., Kueh, H. Y., Barkai, N. & Leibler, S. (2002). Mechanisms of noise-resistance in genetic oscillators, *Proceedings of the National Academy of Sciences of the US* **99**(9): 5988–5992.
   URL: *http://www.pnas.org/content/99/9/5988.abstract*

Voet, D., Voet, J. & Pratt, C. (1999). *Fundamentals of biochemistry*, Wiley New York.

Wernet, M. F., Mazzoni, E. O., Celik, A., Duncan, D. M., Duncan, I. & Desplan, C. (2006). Stochastic spineless expression creates the retinal mosaic for colour vision, *Nature* **440**(7081): 174–180.
   URL: *http://dx.doi.org/10.1038/nature04615*

Winstead, C., Madsen, C. & Myers, C. (2010). iSSA: an incremental stochastic simulation algorithm for genetic circuits, *Proc. 2010 IEEE International Symposium on Circuits and Systems (ISCAS 2010)*.

Stochastic Decision Support Models and Optimal Stopping Rules in a New Product Lifetime Testing

Nicholas A. Nechval and Maris Purgailis
University of Latvia
Latvia

1. Introduction

The theory of stopping rules has its roots in the study of the optimality properties of the sequential probability ratio test of Wald and Wolfowitz (1948) and Arrow, Blackwell and Girshick (1949). The essential idea in both of these papers was to create a formal Bayes problem.

The formal Bayes problem is what we would now call an optimal stopping problem. A decision maker observes an adapted sequence $\{R_n, f_n, n \geq I\}$, with $E\{|R_n|\} < \infty$ for all n, where f_n denotes the σ–algebra generated by a sequence of rewards $R_1, \ldots, R_n$. At each time n a choice is to be made, to stop sampling and collect the currently available reward, R_n, or continue sampling in the expectation of collecting a larger reward in the future. An optimal stopping rule N is one that maximizes the expected reward, $E\{R_n\}$. The key to finding an optimal or close to optimal stopping rule is the family of equations

$$Z_n = \max\ (R_n, E\{Z_{n+1} | f_n\}), \quad n = 1, 2, \ldots\ . \tag{1}$$

The informal interpretation of Z_n, is that it is the most one can expect to win if one has already reached stage n; and equations (1) say that this quantity is the maximum of what one can win by stopping at the nth stage and what one can expect to win by taking at least one more observation and proceeding optimally thereafter. The plausible candidate for an optimal rule is to stop with

$$N = \min\{\ n : R_n \geq E\{Z_{n+1} | f_n\}\}, \tag{2}$$

that is, stop as soon as the current reward is at least as large as the most that one can expect to win by continuing. Equations (1) show that $\{Z_n, f_n\}$ is a supermartingale, while $\{Z_{\min(N,n)}, f_n\}$, is a martingale. The equations do not have a unique solution, but in the case where the index n is bounded, say $1 \leq n \leq m$ for some given value of m, the solution of interest satisfies $Z_m, = R_m$. Hence (1) can be solved and the optimal stopping rule can be found by "backward induction". The general strategy of optimal stopping theory is to

approximate the case where no bound m exists by first imposing such a bound, solving the bounded problem and then letting m → ∞. For reviews of the many variations on this problem and the extensive related literature, see Freeman (1983), Petrucelli (1988) and Samuels (1991).

For illustration of the stopping problem, consider the Bayesian sequential estimation problem of a binomial parameter under quadratic loss and constant observation cost. Suppose that the unknown binomial parameter p is assigned a beta prior distribution with integer parameters (a,b) so that

$$\pi(p \mid a, b) = \frac{(b-1)!}{(a-1)!(b-a-1)!} p^{a-1}(1-p)^{b-a-1}, \quad 0 < p < 1. \tag{3}$$

The posterior distribution of p having observed s successes in n trials is simply $\pi(p; s+a, n+b)$ (Raiffa and Schlaifer, 1968); hence the result of sampling may be represented as a plot of s+a against n+b which stops when the stopping boundary is reached. If a=1, b=2, the uniform prior, is taken as the origin, sample paths for any other proper prior will start at the point (a–1,b–2). Consequently stopping boundaries will be obtained using the uniform prior.

Suppose that the loss in estimating p by d is $\vartheta(p-d)^2$ where ϑ is a constant giving loss in terms of cost. Then the Bayes estimator is the current prior mean $(s+l)/(n+2)$ and the Bayes risk is

$$B(s, n) = \frac{\vartheta(s+1)(n-s+1)}{(n+2)^2(n+3)}. \tag{4}$$

At a point (s,n) let D(s,n) be the risk of taking one further observation at a cost c and M(s,n) be the minimum risk, then the dynamic programming equations giving the partition of the (s,n) plane into stopping and continuation points are

$$M(s, n) = \min\{B(s, n), D(s, n)\}, \tag{5}$$

where

$$D(s, n) = c + \frac{s+1}{n+2} M(s+1, n+1) + \frac{n-s+1}{n+2} M(s, n+1). \tag{6}$$

The equations are similar to those of Lindley and Barnett (1965) and Freeman (1970, 1972, 1973). The optimal decision at each point is obtained by working back from a maximum sample size, which is approximately $[(1/2)\sqrt{\vartheta/c}] - 2$. A suboptimal stopping point (s,n) is defined as a first stopping point for fixed s if (s,n–1) is a continuation point, in this case

$$D(s, n-1) = c + \frac{s+1}{n+1} M(s+1, n) + \frac{n-s}{n+1} B(s, n)$$

$$\leq c + \frac{s+1}{n+1} B(s+1, n) + \frac{n-s}{n+1} B(s, n) = D^{\bullet}(s, n-1). \tag{7}$$

A lower bound for the sample size n above may now be found from (7) by setting $B(s,n–1) \geq D^\bullet(s,n–1)$. This leads to

$$[(n + 2)(n + 1)]^2 \leq (\vartheta / c)(s + 1)(n - s). \tag{8}$$

The optimal stopping boundary starts at s = 0 and n, and from (8) it may be shown that this sample size is at least $[(\vartheta/c)^{1/3}]–3$.

The approximate design obtained by (8) will be termed a one step ahead design. Both designs will obviously stop at the same maximum number of observations N, and will give the same decision after (N–1) observations. The one step ahead design gives stopping boundaries, which will lie inside those of the optimal. The one step ahead design is similar to the modified Bayes rule of Amster (1963) and has been used by El-Sayyad and Freeman (1973) to estimate a Poisson process rate.

The present research investigates the frequentist (non-Bayesian) stopping rules. In this paper, stopping rules in fixed-sample testing as well as in sequential-sample testing are discussed.

2. Assumptions and Cost Functions in Fixed-Sample Testing

Let c_1 be the cost per hour of conducting the test, c_2 be the total cost of redesign (including the time required to implement it). The cost of redesign c_2 is undoubtedly the most difficult to estimate. This cost is to include whatever redesigns are necessary to make the probability of failure on rerun negligible. To simplify the mathematics, it is assumed that unnecessary design changes, caused by incorrectly abandoning the test, will also have a beneficial effect on performance. This assumption appears warranted for many electronic and mechanical systems, where the introduction of redundancies, higher-quality components, etc., can always be expected to improve reliability.

It will be assumed in this section that the times of interest to the decision maker are restricted to those where a failure has just occurred.

Let $X_1 \leq X_2 \leq \ldots \leq X_r$ be the first r ordered past observations with lifetime distribution $f(x|\theta)$ from a sample of size n. Let $\hat\theta$ be the maximum-likelihood estimate of θ based upon the first r order statistics $(X_1, \ldots, X_r) \equiv X^r$. Let $g(x_1, x_2, \ldots, x_r|\theta)$ be the joint density of the r observations, $g(x_1, x_2, \ldots, x_r, x_s|\theta)$ be the joint density of the first r and sth order statistics (s > r) and $f(x_s|x^r,\theta)$ be the conditional density of the sth order statistic. If τ_0 is the life specified as acceptable and the product will be accepted if a random sample of n items shows (s-1) or fewer failures in performance testing, then the probability of passing the test after x_r has been observed may be estimated as

$$\hat{p}_{pas} = \int_{\tau_0}^{\infty} f(x_s \mid x^r, \hat\theta)dx_s, \tag{9}$$

where

$$f(x_s \mid x^r, \hat\theta) = \frac{g(x_1, \ldots, x_r, x_s \mid \hat\theta)}{g(x_1, \ldots, x_r \mid \hat\theta)}. \tag{10}$$

The cost of abandoning the test is

$$c_{abandoning} = c_1 \tau_0 + c_2 \tag{11}$$

The estimated cost of continuation of the test is given by

$$\hat{c}_{continuing} = \int_{x_r}^{\tau_0} [c_1(x_s - x_r) + c_1\tau_0 + c_2] f(x_s \mid x^r, \hat{\theta}) dx_s + \int_{\tau_0}^{\infty} c_1(\tau_0 - x_r) f(x_s \mid x^r, \hat{\theta}) dx_s$$

$$= c_1 \int_{x_r}^{\tau_0} x_s f(x_s \mid x^r, \hat{\theta}) dx_s + (1 - \hat{p}_{pas})[c_1(\tau_0 - x_r) + c_2] + \hat{p}_{pas}[c_1(\tau_0 - x_r)]$$

$$= c_1 \int_{x_r}^{\tau_0} x_s f(x_s \mid x^r, \hat{\theta}) dx_s + c_1\tau_0 + c_2 - c_1 x_r - \hat{p}_{pas} c_2. \tag{12}$$

3. Stopping Rule in Fixed-Sample Testing

The decision rule will be based on the relative magnitude of $c_{abandoning}$ and $\hat{c}_{continuing}$. The simplest rule would be:

If $\hat{c}_{continuing} < c_{abandoning}$, i.e., if

$$\int_{x_r}^{\tau_0} x_s f(x_s \mid x^r, \hat{\theta}) dx_s < x_r + \hat{p}_{pas} \frac{c_2}{c_1}, \tag{13}$$

continue the present test;

If $\hat{c}_{continuing} \geq c_{abandoning}$, i.e., if

$$\int_{x_r}^{\tau_0} x_s f(x_s \mid x^r, \hat{\theta}) dx_s \geq x_r + \hat{p}_{pas} \frac{c_2}{c_1}, \tag{14}$$

abandon the present test and initiate a redesign.

4. Estimation of the Probability of Passing the Fixed-sample Test

Evaluation of the cost functions for the lifetime-testing model requires, even for relatively simple probability distributions, the evaluation of some complicated integrals that cannot always be obtained in closed form. For example, using the one-parameter exponential model for lifetime distribution, we have

$$f(x \mid \sigma) = \frac{1}{\sigma} \exp\left(-\frac{x}{\sigma}\right), \quad x \geq 0, \tag{15}$$

$$F(x \mid \sigma) = 1 - \exp\left(-\frac{x}{\sigma}\right). \tag{16}$$

Therefore,

$$g(x_1, \ldots, x_r \mid \sigma) = \frac{n!}{(n-k)!} \frac{1}{\sigma^r} \exp\left(-\sum_{i=1}^{r} \frac{x_i}{\sigma}\right)\left[\exp\left(-\frac{x_r}{\sigma}\right)\right]^{n-r}; \tag{17}$$

$$g(x_1, \ldots, x_r, x_s \mid \sigma) = \frac{n!}{(s-r-1)!(n-s)!} \frac{1}{\sigma^{r+1}} \left[\exp\left(-\frac{x_r}{\sigma}\right) - \exp\left(-\frac{x_s}{\sigma}\right)\right]^{s-r-1}$$

$$\times \exp\left(-\sum_{i=1}^{r} \frac{x_i}{\sigma}\right)\left[\exp\left(-\frac{x_s}{\sigma}\right)\right]^{n-s+1}. \tag{18}$$

The maximum likelihood estimate for σ is

$$\hat{\sigma} = \frac{\sum_{i=1}^{r} x_i + (n-r)x_r}{r}. \tag{19}$$

Replacing σ by $\hat{\sigma}$ in the density functions and simplifying, we obtain

$$\hat{P}_{pas} = \int_{\tau_0}^{\infty} \frac{(n-r)!}{(s-r-1)!(n-s)!} \frac{\left[\exp\left(-\frac{x_r}{\hat{\sigma}}\right) - \exp\left(-\frac{x_s}{\hat{\sigma}}\right)\right]^{s-r-1}}{\left[\exp\left(-\frac{x_r}{\hat{\sigma}}\right)\right]^{n-r}} \frac{1}{\hat{\sigma}}\left[\exp\left(-\frac{x_s}{\hat{\sigma}}\right)\right]^{n-s+1} dx_s. \tag{20}$$

If we write

$$\left[\exp\left(-\frac{x_r}{\hat{\sigma}}\right)\right]^{n-r} = \left[\exp\left(-\frac{x_r}{\hat{\sigma}}\right)\right]^{s-r-1}\left[\exp\left(-\frac{x_r}{\hat{\sigma}}\right)\right]^{n-s+1}, \tag{21}$$

then it is clear that

$$\hat{P}_{pas} = \int_{\tau_0}^{\infty} \frac{(n-r)!}{(s-r-1)!(n-s)!} \frac{1}{\hat{\sigma}} \left[1 - \frac{\exp\left(-\frac{x_s}{\hat{\sigma}}\right)}{\exp\left(-\frac{x_r}{\hat{\sigma}}\right)}\right]^{s-r-1}\left[\frac{\exp\left(-\frac{x_s}{\hat{\sigma}}\right)}{\exp\left(-\frac{x_r}{\hat{\sigma}}\right)}\right]^{n-s+1} dx_s. \tag{22}$$

The change of variable

$$v = \frac{\exp\left(-\dfrac{x_s}{\hat{\sigma}}\right)}{\exp\left(-\dfrac{x_r}{\hat{\sigma}}\right)} \tag{23}$$

leads to

$$\hat{p}_{pas} = \int_0^{\exp\left(-\frac{\tau_0 - x_r}{\hat{\sigma}}\right)} \frac{(n-r)!}{(s-r-1)!(n-s)!} v^{n-s}(1-v)^{s-r-1} dv. \tag{24}$$

Thus, $\hat{p}_{pas}$ is equivalent to the cumulative beta distribution with parameters (n–s+1, s–r). The situation for the Weibull distribution,

$$f(x \mid \sigma, \delta) = \frac{\delta}{\sigma} x^{\delta-1} \exp\left(-\frac{x^\delta}{\sigma}\right), \quad x \ge 0; \quad F(x \mid \sigma, \delta) = 1 - \exp\left(-\frac{x^\delta}{\sigma}\right), \tag{25}$$

is much the same, except that we make the change of variable

$$v = \frac{\exp\left(-\dfrac{x_s^{\hat{\delta}}}{\hat{\sigma}}\right)}{\exp\left(-\dfrac{x_r^{\hat{\delta}}}{\hat{\sigma}}\right)}. \tag{26}$$

The maximum likelihood estimates $\hat{\sigma}$ and $\hat{\delta}$ of the parameters σ and δ, respectively, required in (26), can only be obtained by iterative methods. The appropriate likelihood equations for $X_1, \ldots, X_r$ are

$$\frac{\partial L}{\partial \sigma} = 0 = -\frac{r}{\sigma} + \frac{1}{\sigma^2}\left[\sum_{i=1}^r x_i^\delta + (n-r)x_r^\delta\right], \tag{27}$$

$$\frac{\partial L}{\partial \delta} = 0 = \frac{r}{\delta} + \sum_{i=1}^r x_i - \frac{1}{\sigma}\left[\sum_{i=1}^r x_i^\delta \ln x_i + (n-r)x_r^\delta \ln x_r\right]. \tag{28}$$

Now $\hat{\sigma}$ and $\hat{\delta}$ can be found from solution of

$$\hat{\sigma} = \frac{\displaystyle\sum_{i=1}^r x_i^{\hat{\delta}} + (n-r)x_r^{\hat{\delta}}}{r} \tag{29}$$

and

$$\widehat{\delta} = \left[\left(\sum_{i=1}^{r} x_i^{\widehat{\delta}} \ln x_i + (n-r)x_r^{\widehat{\delta}} \ln x_r \right) \left(\sum_{i=1}^{r} x_i^{\widehat{\delta}} + (n-r)x_r^{\widehat{\delta}} \right)^{-1} - \frac{1}{r} \sum_{i=1}^{r} \ln x_i \right]^{-1}. \tag{30}$$

The method described above is quite general and works well for all closed-form or tabulated cumulative distribution functions, so that numerical integration techniques are not needed for calculating $\widehat{p}_{pas}$. It is easy to see that the general case would involve a change of variable

$$v = \frac{1 - F(x_s \mid \widehat{\theta})}{1 - F(x_r \mid \widehat{\theta})}, \tag{31}$$

where, of course, x_r is a constant.

4.1 Statistical Inferences for Future Order Statistics in the Same Sample
If we deal with small size n of the fixed sample for testing and wish to find the conditional distribution of the sth order statistic to obtain the probability of passing the test after x_r has been observed, then it may be suitable the following results.

Theorem 1 (*Predictive distribution of the sth order statistic X_s on the basis of the past rth order statistic X_r from the exponential distribution of the same sample*). Let $X_1 \leq X_2 \leq ... \leq X_r$ be the first r ordered past observations from a sample of size n from the exponential distribution with the probability density function (PDF) (15), which is characterized by the scale parameter σ. It is assumed that the parameter σ is unknown. Then the predictive probability density function of the sth order statistic X_s may be obtained on the basis of the rth order statistic X_r ($r < s \leq n$) from the same sample as

$$\widetilde{f}(x_s \mid x_r) = \frac{1}{B(s-r, n-s+1)B(r, n-r+1)} \sum_{j=0}^{s-r-1} \sum_{i=0}^{r-1} (-1)^{i+j} \binom{s-r-1}{j} \binom{r-1}{i}$$

$$\times \frac{1}{[w_s(n-s+1+j) + (n-r+1+i)]^2} \frac{1}{x_r}, \quad w_s > 0, \tag{32}$$

where

$$W_s = \frac{X_s - X_r}{X_r}. \tag{33}$$

Proof. It follows readily from standard theory of order statistics (see, for example, Kendall and Stuart (1969)) that the joint distribution of X_r, X_s ($s > r$) is given by

$$f(x_r, x_s \mid \sigma) dx_r dx_s = \frac{1}{B(r, s-r)B(s, n-s+1)}$$

$$\times[F(x_r \mid \sigma)]^{r-1}[F(x_s \mid \sigma) - F(x_r \mid \sigma)]^{s-r-1}[1 - F(x_s \mid \sigma)]^{n-s} dF(x_r \mid \sigma) dF(x_s \mid \sigma), \qquad (34)$$

Making the transformation $z = x_s - x_r$, $x_r = x_r$, and integrating out x_r, we find the density of z as the beta density

$$f(z \mid \sigma) = \frac{1}{B(s-r, n-s+1)}[\exp(-z/\sigma)]^{n-s+1}[1 - \exp(-z/\sigma)]^{s-r-1}\frac{1}{\sigma}. \qquad (35)$$

The distribution of X_r is

$$f(x_r \mid \sigma) dx_r = \frac{1}{B(r, n-r+1)}[F(x_r \mid \sigma)]^{r-1}[1 - F(x_r \mid \sigma)]^{n-r} dF(x_r \mid \sigma), \qquad (36)$$

and since Z, X_r are independent, we have the joint density of Z and X_r as

$$f(z, x_r \mid \sigma) = \frac{1}{B(r, s-r)B(s, n-s+1)}[\exp(-z/\sigma)]^{n-s+1}[1 - \exp(-z/\sigma)]^{s-r-1}$$

$$\times[1 - \exp(-x_r/\sigma)]^{r-1}[\exp(-x_r/\sigma]^{n-r+1}\frac{1}{\sigma^2}. \qquad (37)$$

Making the transformation $w_s = z/x_r$, $x_r = x_r$, we find the joint density of W_s and X_r as

$$f(w_s, x_r \mid \sigma) = \frac{1}{B(r, s-r)B(s, n-s+1)}[\exp(-w_s x_r/\sigma)]^{n-s+1}[1 - \exp(-w_s x_r/\sigma)]^{s-r-1}$$

$$\times[1 - \exp(-x_r/\sigma)]^{r-1}[\exp(-x_r/\sigma]^{n-r+1}x_r\frac{1}{\sigma^2}. \qquad (38)$$

It is then straightforward to integrate out x_r, leaving the density of W_s as

$$f(w_s) = \frac{1}{B(s-r, n-s+1)B(r, n-r+1)}\sum_{j=0}^{s-r-1}\sum_{i=0}^{r-1}(-1)^{i+j}\binom{s-r-1}{j}\binom{r-1}{i}$$

$$\times\frac{1}{[w_s(n-s+1+j) + (n-r+1+i)]^2}, \quad w_s > 0. \qquad (39)$$

It will be noted that the technique of invariant embedding (Nechval, 1982, 1984, 1986, 1988a, 1988b; Nechval et al., 1999, 2000, 2001, 2003a, 2003b, 2004, 2008, 2009) allows one to obtain (39) directly from (34). This ends the proof.

Corollary 1.1.

$$\Pr\{W_s \le w_s\} = \frac{1}{B(s-r,n-s+1)B(r,n-r+1)} \sum_{j=0}^{s-r-1} \sum_{i=0}^{r-1} (-1)^{i+j} \binom{s-r-1}{j}\binom{r-1}{i}$$

$$\times \left(\frac{1}{n-r+1+i} - \frac{1}{w_s(n-s+1+j)+(n-r+1+i)} \right) \frac{1}{(n-s+1+j)}$$

$$= 1 - \frac{1}{B(s-r,n-s+1)B(r,n-r+1)} \sum_{j=0}^{s-r-1} (-1)^j \binom{s-r-1}{j} \left[r(n-s+1+j) \binom{w_s(n-s+1+j)+n}{r} \right]^{-1}.$$

$$(40)$$

For a specified probability level α, w_s can be obtained such that

$$\Pr\{W_s \le w_s \mid X_r = x_r\} = \Pr\left\{ \frac{X_s - x_r}{x_r} \le w_s \right\} = \Pr\{X_s \le (w_s+1)x_r\} = \alpha. \qquad (41)$$

Hence, with confidence α, one could predict X_s to be less than or equal to $(w_s+1)x_r$.

Consider, for instance, the case where n=6 simultaneously tested items have life times following the exponential distribution (15). Two items (r = 2) fail at times 75 and 90 hours. Suppose, say, we are predicting the 4th failure time (s = 4). Using (40), (41), and α = 0.95, we get w_s=10, which yields a predicted value for X_s of 990 hours.

Theorem 2 (*Predictive distribution of the sth order statistic X_s on the basis of the past observations $X_1 \le X_2 \le \ldots \le X_r$ from the exponential distribution of the same sample*). Under conditions of Theorem 1, the predictive probability density function of the sth order statistic X_s (r < s ≤ n) may be obtained on the basis of the past observations $(X_1 \le X_2 \le \ldots \le X_r)$ from the same sample as

$$\widetilde{f}(x_s \mid x^r) = \frac{r}{B(s-r,n-s+1)} \sum_{j=0}^{s-r-1} (-1)^j \binom{s-r-1}{j} \frac{1}{[1+w_s(n-s+1+j)]^{r+1}} \frac{1}{q_r}, \quad w_s > 0, \quad (42)$$

where

$$W_s = \frac{X_s - X_r}{Q_r}, \qquad (43)$$

$$Q_r = \sum_{i=1}^{r} X_i + (n-r)X_r. \qquad (44)$$

Proof. The joint probability density function of $X_1, X_2, \ldots, X_r, X_s$ is given by

$$f(x_1, x_2, ..., x_r, x_s \mid \sigma) = \frac{n!}{(s-r-1)!(n-s)!}$$

$$\times [F(x_s \mid \sigma) - F(x_r \mid \sigma)]^{s-r-1} [1 - F(x_s \mid \sigma)]^{n-s} \prod_{i=1}^{r} f(x_i \mid \sigma) f(x_s \mid \sigma)$$

$$= \frac{n!}{(s-r-1)!(n-s)!} \frac{1}{\sigma^{r+1}} \exp\left(- \frac{\displaystyle\sum_{i=1}^{r} x_i + (n-r)x_r}{\sigma} \right)$$

$$\times \left[1 - \exp\left(- \frac{x_s - x_r}{\sigma} \right) \right]^{s-r-1} \left[\exp\left(- \frac{x_s - x_r}{\sigma} \right) \right]^{n-s+1}. \tag{45}$$

Let

$$V = \frac{Q_r}{\sigma} = \frac{\displaystyle\sum_{i=1}^{r} X_i + (n-r)X_r}{\sigma} \tag{46}$$

and

$$W_s = \frac{X_s - X_r}{Q_r} = \frac{X_s - X_r}{\displaystyle\sum_{i=1}^{r} X_i + (n-r)X_r}. \tag{47}$$

Using the invariant embedding technique (Nechval, 1982, 1984, 1986, 1988a, 1988b; Nechval et al., 1999, 2000, 2001, 2003a, 2003b, 2004, 2008, 2009), we then find in a straightforward manner that the joint density of V, W_s conditional on fixed $x^r = (x_1, x_2, ..., x_r)$, is

$$f(w_s, v \mid x^r) = \vartheta(x^r) \sum_{j=0}^{s-r-1} \binom{s-r-1}{j} (-1)^j \; v^r \exp(-v[1 + w_s(n-s+1+j)]),$$

$$w_s \in (0, \infty), \quad v \in (0, \infty), \tag{48}$$

where

$$\vartheta(x^r) = \left(\int_0^\infty \int_0^\infty \frac{1}{\vartheta(x^r)} f(w_s, v \mid x^r) dw_s dv \right)^{-1} = \frac{1}{B(s-r, n-s+1)\Gamma(r)} \tag{49}$$

is the normalizing constant, which does not depend on x^r. Now v can be integrated out of (48) in a straightforward way to give

$$f(w_s \mid x^r) = \frac{r}{B(s-r,n-s+1)} \sum_{j=0}^{s-r-1} \binom{s-r-1}{j} (-1)^j \frac{1}{[1+w_s(n-s+1+j)]^{r+1}}. \tag{50}$$

Then (42) follows from (50). This completes the proof.

Corollary 2.1.

$$Pr\{W_s \leq w_s\} = 1 - \frac{1}{B(s-r,n-s+1)} \sum_{j=0}^{s-r-1} \binom{s-r-1}{j} (-1)^j \frac{1}{(n-s+1+j)[1+w_s(n-s+1+j)]^r}. \tag{51}$$

For a specified probability level α, w_s can be obtained such that

$$Pr\{W_s \leq w_s \mid X_r = x_r, Q_r = q_r\} = Pr\left\{\frac{X_s - x_r}{q_r} \leq w_s\right\} = Pr\{X_s \leq x_r + q_r w_s\} = \alpha. \tag{52}$$

Hence, with confidence α, one could predict X_s to be less than or equal to $x_r+q_r w_s$.

Consider a life-testing situation similar to that in the above example of Theorem 1, where n = 6 simultaneously tested items have life times following the exponential distribution (15). Two items (r = 2) fail at times 75 and 90 hours. Suppose, say, we are predicting the 4th failure time (s = 4). Using (44), (45), (46), and α = 0.95, we get q_r = 525 and w_s = 1.855, which yield a predicted value for X_s of 1064 hours.

We make two additional remarks concerning evaluation of the above probability (51):

(i) In the important case where s = n, expression (51) simplifies to

$$Pr\{W_s \leq w_s\} = \sum_{j=0}^{n-r} \binom{n-r}{j} (-1)^j \frac{1}{(1+jw_s)^r}. \tag{53}$$

(ii) In the special case where r = s–1, we note that $(s-1)(n-s+1)(X_s-X_{s-1})/Q_{s-1}$ is an F variate with (2, 2s–2) degrees of freedom, so that appropriate probability statements can be read from standard tables of the F distribution.

Theorem 3 (*Predictive distribution of the sth order statistic X_s on the basis of the past order statistics X_r and X_1 from the two-parameter exponential distribution of the same sample*). Let $X_1 \leq X_2 \leq ... \leq X_r$ be the first r ordered past observations from a sample of size n from the exponential distribution with the PDF

$$f(x \mid \sigma) = \frac{1}{\sigma} \exp[-(x-\mu)/\sigma], \quad (\sigma > 0, \ -\infty < \mu < \infty, \ x \geq \mu), \tag{54}$$

which is characterized by the scale parameter σ and the shift parameter μ. It is assumed that these parameters are unknown. Then the predictive PDF of the sth order statistic X_s (s>r) from the same sample may be obtained as

$$\tilde{f}(x_s \mid x_1, x_r) = \frac{1}{B(s-r,n-s+1)B(r-1,n-r+1)} \sum_{j=0}^{s-r-1} \sum_{i=0}^{r-2} (-1)^{i+j} \binom{s-r-1}{j} \binom{r-2}{i}$$

$$\times \frac{1}{[w_s(n-s+1+j)+(n-r+1+i)]^2} \frac{1}{x_r - x_1}, \quad w_s > 0, \tag{55}$$

where

$$W_s = \frac{X_s - X_r}{X_r - X_1}. \tag{56}$$

Proof. It is carried out in the similar way as the proof of Theorem 1.

Corollary 3.1.

$$\Pr\{W_s \le w_s\} = \frac{1}{B(s-r,n-s+1)B(r-1,n-r+1)} \sum_{j=0}^{s-r-1} \sum_{i=0}^{r-2} (-1)^{i+j} \binom{s-r-1}{j}\binom{r-2}{i}$$

$$\times \left(\frac{1}{n-r+1+i} - \frac{1}{w_s(n-s+1+j)+(n-r+1+i)} \right) \frac{1}{(n-s+1+j)}$$

$$= 1 - \frac{1}{B(s-r,n-s+1)B(r-1,n-r+1)} \sum_{j=0}^{s-r-1} \binom{s-r-1}{j}(-1)^j \frac{[(r-1)(n-s+1+j)]^{-1}}{\binom{w_s(n-s+1+j)+n}{r-1}} \tag{57}$$

For a specified probability level α, w_s can be obtained such that

$$\Pr\{W_s \le w_s \mid X_r = x_r, X_1 = x_1\} = \Pr\left\{ \frac{X_s - X_r}{x_r - x_1} \le w_s \right\}$$

$$= \Pr\{X_s \le x_r + w_s(x_r - x_1)\} = \alpha. \tag{58}$$

Hence, with confidence α, one could predict X_s to be less than or equal to $x_r + w_s(x_r - x_1)$.

Theorem 4 (*Predictive distribution of the sth order statistic X_s on the basis of the past order statistics $X_1 \le X_2 \le \dots \le X_r$ from the two-parameter exponential distribution of the same sample*). Under conditions of Theorem 3, the predictive probability density function of the sth order statistic X_s (s>r) from the same sample may be obtained as

$$\widetilde{f}(x_s \mid x^r) = \frac{r-1}{B(s-r,n-s+1)} \sum_{j=0}^{s-r-1} \binom{s-r-1}{j}(-1)^j \frac{1}{[1+w_s(n-s+1+j)]^r} \frac{1}{q}, \quad w_s > 0, \tag{59}$$

where

$$W_s = \frac{X_s - X_r}{Q}, \tag{60}$$

$$Q = \sum_{i=1}^{r}(X_i - X_1) + (n-r)(X_r - X_1). \tag{61}$$

Proof. The proof is carried out in the similar way as the proof of Theorem 2.

Corollary 4.1.

$$\Pr\{W_s \le w_s\} = 1 - \frac{1}{B(s-r,n-s+1)} \sum_{j=0}^{s-r-1} \binom{s-r-1}{j}(-1)^j \frac{1}{(n-s+1+j)[1+w_s(n-s+1+j)]^{r-1}}$$

$$= 1 - \frac{1}{B(s-r,n-s+1)} \sum_{j=r+1}^{s} \binom{s-r-1}{s-j}(-1)^{s-j} \frac{1}{(n+1-j)[1+w_s(n+1-j)]^{r-1}}. \tag{62}$$

For a specified probability level α, w_s can be obtained such that

$$\Pr\{W_s \le w_s \mid X_r = x_r, Q = q\} = \Pr\left\{\frac{X_s - x_r}{q} \le w_s\right\} = \Pr\{X_s \le x_r + q w_s\} = \alpha. \tag{63}$$

Hence, with confidence α, one could predict X_s to be less than or equal to $x_r + q w_s$.
Suppose, for instance, that n = 8 items are put on test simultaneously and that the first r = 4 items have the lifetimes 62, 84, 106 and 144 hours. Let the lifetimes of all n items be distributed according to the two-parameter exponential distribution (47) with the same parameters σ and μ. We wish to find a 95% prediction interval of the type (56) for s=8. We obtain from (55) and (56) that $\Pr\{X_s \le 1408.8\} = 0.95$. Thus, we can be 95% confident that the total elapsed time will not exceed 1409 hours.

Theorem 5 (*Predictive distribution of the sth order statistic X_s on the basis of the past order statistics $X_1 \le X_2 \le ... \le X_r$ from the two-parameter Weibull distribution of the same sample*). Let $X_1 \le X_2 \le ... \le X_r$ be the first r ordered past observations from a sample of size n from the two-parameter Weibull distribution given by

$$f(x \mid \beta, \delta) = \frac{\delta}{\beta}\left(\frac{x}{\beta}\right)^{\delta-1} \exp\left[-\left(\frac{x}{\beta}\right)^{\delta}\right] \quad (x > 0), \tag{64}$$

where $\delta > 0$ and $\beta > 0$ are the shape and scale parameters, respectively, which are unknown. Then the predictive PDF of the sth order statistic X_s (s>r) from the same sample may be obtained as

$$\tilde{f}(x_s, v \mid \mathbf{z})$$

$$
= \left(\sum_{j=0}^{s-r-1} \binom{s-r-1}{j} (-1)^{s-r-1-j} \frac{rv^{r-2} e^{v\hat{\delta} \sum_{i=1}^{r} \ln(x_i/\hat{\beta})} ve^{v\left[w_s + \hat{\delta}\ln(x_r/\hat{\beta})\right]}}{\left((n-r-j)e^{v\left[w_s + \hat{\delta}\ln(x_r/\hat{\beta})\right]} + je^{v\hat{\delta}\ln(x_r/\hat{\beta})} + \sum_{i=1}^{r} e^{v\hat{\delta}\ln(x_i/\hat{\beta})} \right)^{r+1}} \frac{\hat{\delta}}{x_s} \right)
$$

$$
\times \left(\sum_{j=0}^{s-r-1} \binom{s-r-1}{j} (-1)^{s-r-1-j} \int_0^{\infty} \frac{v^{r-2} e^{v\hat{\delta} \sum_{i=1}^{r} \ln(x_i/\hat{\beta})}}{(n-r-j)\left(\sum_{i=1}^{r} e^{v\hat{\delta}\ln(x_i/\hat{\beta})} + (n-r)e^{v\hat{\delta}\ln(x_r/\hat{\beta})} \right)^{r}} dv \right)^{-1} ,
$$

$$
w_s \in (-\infty, \infty), \quad v \in (0, \infty). \tag{65}
$$

where $\hat{\beta}$ and $\hat{\delta}$ are the maximum likelihood estimators of β and δ based on the first r ordered past observations $(X_1, \ldots, X_r)$ from a sample of size n from the Weibull distribution, which can be found from solution of

$$
\hat{\beta} = \left(\frac{\sum_{i=1}^{r} x_i^{\hat{\delta}} + (n-r)x_r^{\hat{\delta}}}{r} \right)^{1/\hat{\delta}} , \tag{66}
$$

and

$$
\hat{\delta} = \left[\left(\sum_{i=1}^{r} x_i^{\hat{\delta}} \ln x_i + (n-r)x_r^{\hat{\delta}} \ln x_r \right) \left(\sum_{i=1}^{r} x_i^{\hat{\delta}} + (n-r)x_r^{\hat{\delta}} \right)^{-1} - \frac{1}{r}\sum_{i=1}^{r} \ln x_i \right]^{-1} , \tag{67}
$$

$$
\mathbf{z} = (z_1, z_2, \ldots, z_r) , \tag{68}
$$

$$
Z_i = \hat{\delta} \ln\left(\frac{X_i}{\hat{\beta}} \right), \quad i = 1, \ldots, r, \tag{69}
$$

$$
W_s = \hat{\delta} \ln\left(\frac{X_s}{X_r} \right). \tag{70}
$$

Proof. The joint density of $Y_1 = \ln(X_1), \ldots, Y_r = \ln(X_r), Y_s = \ln(X_s)$ is given by

$$
f(y_1, \ldots, y_r, y_s \mid \mu, \sigma) = \frac{n!}{(s-r-1)!(n-s)!} \prod_{i=1}^{r} f(y_i \mid \mu, \sigma)[F(y_s \mid \mu, \sigma) - F(y_r \mid \mu, \sigma)]^{s-r-1}
$$

$$\times f(y_s \mid \mu, \sigma)[1 - F(y_s \mid \mu, \sigma)]^{n-s}, \tag{71}$$

where

$$f(y \mid \mu, \sigma) = \frac{1}{\sigma} \exp\left[\frac{y-\mu}{\sigma} - \exp\left(\frac{y-\mu}{\sigma}\right)\right], \tag{72}$$

$$F(y \mid \mu, \sigma) = 1 - \exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right], \tag{73}$$

$$\mu = \ln \beta, \quad \sigma = 1/\delta. \tag{74}$$

Let $\hat{\mu}$, $\hat{\sigma}$ be the maximum likelihood estimators (estimates) of μ, σ based on $Y_1, \ldots, Y_r$ and let

$$V_1 = \frac{\hat{\mu} - \mu}{\hat{\sigma}}, \tag{75}$$

$$V = \frac{\hat{\sigma}}{\sigma}, \tag{76}$$

$$W_s = \frac{Y_s - Y_r}{\hat{\sigma}}, \tag{77}$$

and

$$Z_i = \frac{Y_i - \hat{\mu}}{\hat{\sigma}}, \quad i = 1(1)r. \tag{78}$$

Parameters μ and σ in (64) are location and scale parameters, respectively, and it is well known that if $\hat{\mu}$ and $\hat{\sigma}$ are estimates of μ and σ, possessing certain invariance properties, then the quantities V_1 and V are parameter-free. Most, if not all, proposed estimates of μ and σ possess the necessary properties; these include the maximum likelihood estimates and various linear estimates. Z_i, $i=1(1)r$, are ancillary statistics, any $r-2$ of which form a functionally independent set. For notational convenience we include all of $z_1, \ldots, z_r$ in (68); z_{r-1} and z_r can be expressed as function of $z_1, \ldots, z_r$ only.
Using the invariant embedding technique (Nechval, 1982, 1984, 1986, 1988a, 1988b; Nechval et al., 1999, 2000, 2001, 2003a, 2003b, 2004, 2008, 2009), we then find in a straightforward manner that the joint density of V_1, V, W_s conditional on fixed $\mathbf{z} = (z_1, z_2, \ldots, z_r)$, is

$$f(w_s, v, v_1 \mid \mathbf{z}) = \vartheta(\mathbf{z}) v^{r-1} \exp\left(v[\sum_{i=1}^{r} z_i + z_r + w_s] + (r+1)v_1\right)$$

$$\times \sum_{j=0}^{s-r-1} \binom{s-r-1}{j} (-1)^{s-r-1-j} \exp\left[-e^{v_1}\left((n-r-j)e^{v(z_r+w_s)} + je^{vz_r} + \sum_{i=1}^{r} e^{vz_i} \right) \right],$$

$$w_s \in (0,\infty), \quad v \in (0,\infty), \quad v_1 \in (-\infty,\infty), \tag{79}$$

where

$$\vartheta(\mathbf{z}) = \left(\int_{-\infty}^{\infty} \int_{0}^{\infty} \int_{0}^{\infty} \frac{1}{\vartheta(\mathbf{z})} f(w_s, v_1, v \mid \mathbf{z}) dw_s dv dv_1 \right)^{-1}$$

$$\times \left(\sum_{j=0}^{s-r-1} \binom{s-r-1}{j} (-1)^{s-r-1-j} \int_{0}^{\infty} \frac{v^{r-2} \exp\left(v \sum_{i=1}^{r} z_i \right)}{(n-r-j)\left(\sum_{i=1}^{r} \exp[vz_i] + (n-r)\exp[vz_r] \right)^r} dv \right)^{-1} \tag{80}$$

is the normalizing constant.

Now v_1 can be integrated out of (79) in a straightforward way to give

$$f(w_s, v \mid \mathbf{z})$$

$$= \left(\sum_{j=0}^{s-r-1} \binom{s-r-1}{j} (-1)^{s-r-1-j} \frac{rv^{r-2} \exp\left(v \sum_{i=1}^{r} z_i \right) v \exp[v(w_s + z_r)]}{\left((n-r-j)\exp[v(w_s+z_r)] + j\exp[vz_r] + \sum_{i=1}^{r} \exp[vz_i] \right)^{r+1}} \right)$$

$$\times \left(\sum_{j=0}^{s-r-1} \binom{s-r-1}{j} (-1)^{s-r-1-j} \int_{0}^{\infty} \frac{v^{r-2} \exp\left(v \sum_{i=1}^{r} z_i \right)}{(n-r-j)\left(\sum_{i=1}^{r} \exp[vz_i] + (n-r)\exp[vz_r] \right)^r} dv \right)^{-1} \tag{81}$$

Then (65) follows from (81). This completes the proof.

Corollary 5.1. A lower one-sided conditional $(1-\alpha)$ prediction limit h on the sth order statistic X_s (s>r) from the same sample may be obtained from (73) as

$$\Pr\{X_s \geq h \mid \mathbf{z}\} = \Pr\left\{ \hat{\delta}\ln\left(\frac{X_s}{\hat{\beta}} \right) \geq \hat{\delta}\ln\left(\frac{h}{\hat{\beta}} \right) \mid \mathbf{z} \right\} = \Pr\{W_s \geq w_h \mid \mathbf{z}\}$$

$$= \left(\sum_{j=0}^{s-r-1} \binom{s-r-1}{j} (-1)^{s-r-1-j} \int_0^\infty \frac{(n-r-j)^{-1} v^{r-2} e^{v\hat{\delta}\sum_{i=1}^{r}\ln(x_i/\hat{\beta})}}{\left((n-r-j)e^{v[w_h+\hat{\delta}\ln(x_r/\hat{\beta})]} + je^{v\hat{\delta}\ln(x_r/\hat{\beta})} + \sum_{i=1}^{r} e^{v\hat{\delta}\ln(x_i/\hat{\beta})} \right)^r} dv \right)$$

$$\times \left(\sum_{j=0}^{s-r-1} \binom{s-r-1}{j} (-1)^{s-r-1-j} \int_0^\infty \frac{v^{r-2} e^{v\hat{\delta}\sum_{i=1}^{r}\ln(x_i/\hat{\beta})}}{(n-r-j)\left(\sum_{i=1}^{r} e^{v\hat{\delta}\ln(x_i/\hat{\beta})} + (n-r)e^{v\hat{\delta}\ln(x_r/\hat{\beta})} \right)^r} dv \right)^{-1}$$

$$= 1 - \alpha. \tag{82}$$

Let $X_1 \le X_2 \le \ldots \le X_n$ denote the order statistics in a sample of size n from a continuous parent distribution whose cumulative distribution function $F(x|\theta)$ is a strictly increasing function of x, where θ is an unknown parameter. A number of authors have considered the prediction problem for the future observation X_s based on the observed values $X_1 \le \ldots \le X_r$, $1 \le r < s \le n$. Prediction intervals have been treated by Hewitt (1968), Lawless (1971), Lingappaiah (1973), Likes (1974), and Kaminsky (1977).

Consider, in this section, the case when the parameter θ is known. It can be shown that the predictive distribution of X_n, given $X_i = x_i$ for all $i \le r$, is the same as the predictive distribution of X_n, given only $X_r = x_r$, which is given by

$$\Pr\{X_n \le h \mid \theta, X_r = x_r\} = \left[\frac{F(h|\theta) - F(x_r|\theta)}{1 - F(x_r|\theta)} \right]^{n-r} \tag{83}$$

for $h \ge x_r$. We remark also at this point that

$$\Pr\{X_{r+1} \le h \mid \theta, X_r = x_r\} = 1 - \left[1 - \frac{F(h|\theta) - F(x_r|\theta)}{1 - F(x_r|\theta)} \right]^{n-r} = 1 - \left[\frac{1 - F(h|\theta)}{1 - F(x_r|\theta)} \right]^{n-r} \tag{84}$$

for $h \ge x_r$.

4.2 Statistical Inferences for Order Statistics in the Future Sample

Theorem 6 (*Predictive distribution of the lth order statistic Y_l from a set of m future ordered observations $Y_1 \le \ldots \le Y_l \le \ldots \le Y_m$ on the basis of the past sample from the left-truncated Weibull distribution*). Let $X_1 \le X_2 \le \ldots \le X_r$ be the first r ordered past observations from a sample of size n from the left-truncated Weibull distribution with pdf

$$f(x|a,b,\delta) = \frac{\delta}{\sigma} x^{\delta-1} \exp\left[-\left(x^\delta - \mu\right)/\sigma\right], \quad (x^\delta \ge \mu, \sigma, \delta > 0), \tag{85}$$

which is characterized by being three-parameter (μ, σ, δ) where δ is termed the shape parameter, σ is the scale parameter, and μ is the truncation parameter. It is assumed that the parameter δ is known. Then the non-unbiased predictive density function of the lth order statistic Y_l from a set of m future ordered observations $Y_1 \leq \ldots \leq Y_l \leq \ldots \leq Y_m$ is given by

$$
\widetilde{f}(y_l \mid x^n) =
\begin{cases}
n(r-1)l \binom{m}{l} \sum\limits_{i=0}^{l-1} \dfrac{\binom{l-1}{i}(-1)^i \left[1 + w_l(m-l+i+1)\right]^{-r}}{n+m-l+i+1} \dfrac{\delta}{s} y_l^{\delta-1}, & \text{if } y_l \geq x_1, \\[4mm]
n(r-1)\dfrac{m!(n+m-1)!}{(m-1)!(n+m)!}(1-nw_l)^{-r}\dfrac{\delta}{s}y_l^{\delta-1}, & \text{if } y_l \leq x_1,
\end{cases}
\tag{86}
$$

where

$$
W_l = \left(Y_l^\delta - X_1^\delta\right)\big/ S, \tag{87}
$$

$$
S = \sum_{i=1}^{r}(X_i^\delta - X_1^\delta) + (n-r)(X_r^\delta - X_1^\delta). \tag{88}
$$

Proof. It can be justified by using the factorization theorem that (X_1^δ, S) is a sufficient statistic for (μ, σ). We wish, on the basis of the sufficient statistic (X_1^δ, S) for (μ, σ), to construct the non-unbiased predictive density function of the lth order statistic Y_l from a set of m future ordered observations $Y_1 \leq \ldots \leq Y_l \leq \ldots \leq Y_m$.

By using the technique of invariant embedding (Nechval, 1982, 1984, 1986, 1988a, 1988b; Nechval et al., 1999, 2000, 2001, 2003a, 2003b, 2004, 2008, 2009) of (X_1^δ, S), if $X_1 \leq Y_l$, or (Y_l^δ, S), if $X_1 \geq Y_l$, into a pivotal quantity $(Y_l^\delta - \mu)/\sigma$ or $(X_1^\delta - \mu)/\sigma$, respectively, we obtain an ancillary statistic $W_l = \left(Y_l^\delta - X_1^\delta\right)\big/ S$, whose distribution does not depend on any unknown parameter, and the pdf of W_l given by

$$
f(w_l) =
\begin{cases}
n(r-1)l \binom{m}{l} \sum\limits_{i=0}^{l-1} \dfrac{\binom{l-1}{i}(-1)^i \left[1 + w_l(m-l+i+1)\right]^{-r}}{n+m-l+i+1}, & \text{if } w_l \geq 0, \\[4mm]
n(r-1)\dfrac{m!(n+m-1)!}{(m-1)!(n+m)!}(1-nw_l)^{-r}, & \text{if } w_l \leq 0.
\end{cases}
\tag{89}
$$

This ends the proof.

Corollary 6.1. A lower one-sided $(1-\alpha)$ prediction limit h on the lth order statistic Y_l from a set of m future ordered observations $Y_1 \leq \ldots \leq Y_l \leq \ldots \leq Y_m$ $(\Pr\{Y_l \geq h \mid x^n\} = 1-\alpha)$ may be obtained from (89) as

$$h = \left(x_1^\delta + w_h s\right)^{1/\delta},\tag{90}$$

where

$$w_h = \begin{cases} \arg\left(nl\binom{m}{l}\sum_{i=0}^{l-1} \dfrac{\binom{l-1}{i}(-1)^i[1+w_h(m-l+i+1)]^{-(r-1)}}{(n+m-l+i+1)(m-l+i+1)} = 1-\alpha\right), & \text{if } \alpha \ge \dfrac{m!(n+m-1)!}{(m-1)!(n+m)!}, \\[4mm] \arg\left(1 - \dfrac{m!(m+n-1)!}{(m-1)!(m+n)!}(1-nw_h)^{-(r-1)} = 1-\alpha\right), & \text{if } \alpha \le \dfrac{m!(n+m-1)!}{(m-1)!(n+m)!}. \end{cases}\tag{91}$$

(Observe that an upper one-sided conditional α prediction limit h on the lth order statistic Y_1 may be obtained from a lower one-sided $(1-\alpha)$ prediction limit by replacing $1-\alpha$ by α.)

Corollary 6.2. If $l = 1$, then a lower one-sided $(1-\alpha)$ prediction limit h on the minimum Y_1 of a set of m future ordered observations $Y_1 \le \dots \le Y_m$ is given by

$$h = \begin{cases} \left(x_1^\delta + \dfrac{s}{m}\left[\left(\dfrac{n}{(1-\alpha)(n+m)}\right)^{\frac{1}{r-1}} - 1\right]\right)^{1/\delta}, & \alpha \ge \dfrac{m}{n+m}, \\[6mm] \left(x_1^\delta - \dfrac{s}{n}\left[\left(\dfrac{m}{\alpha(n+m)}\right)^{\frac{1}{r-1}} - 1\right]\right)^{1/\delta}, & \alpha \le \dfrac{m}{n+m}. \end{cases}\tag{92}$$

Consider, for instance, an industrial firm which has the policy to replace a certain device, used at several locations in its plant, at the end of 24-month intervals. It doesn't want too many of these items to fail before being replaced. Shipments of a lot of devices are made to each of three firms. Each firm selects a random sample of 5 items and accepts his shipment if no failures occur before a specified lifetime has accumulated. The manufacturer wishes to take a random sample and to calculate the lower prediction limit so that all shipments will be accepted with a probability of 0.95. The resulting lifetimes (rounded off to the nearest month) of an initial sample of size 15 from a population of such devices are given in Table 1.

Observations														
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
8	9	10	12	14	17	20	25	29	30	35	40	47	54	62
Lifetime (in number of month intervals)														

Table 1. The data of resulting lifetimes

Goodness-of-fit testing. It is assumed that

$$X_i \sim f(x \mid a,b,\delta) = \frac{\delta}{\sigma} x^{\delta-1} \exp\left[-\left(x^\delta - \mu\right)/\sigma\right], \quad (x \geq \mu, \sigma, \delta > 0), \ \ i = 1(1)15, \tag{93}$$

where the parameters μ and σ are unknown; (δ=0.87). Thus, for this example, r = n = 15, k = 3, m = 5, $1-\alpha = 0.95$, $X_1^\delta = 6.1$, and S = 170.8. It can be shown that the

$$U_j = 1 - \left(\frac{\displaystyle\sum_{i=2}^{j+1}(n-i+1)(X_i^\delta - X_{i-1}^\delta)}{\displaystyle\sum_{i=2}^{j+2}(n-i+1)(X_i^\delta - X_{i-1}^\delta)}\right)^j, \quad j = 1(1)n-2, \tag{94}$$

are i.i.d. U(0,1) rv's (Nechval et al., 1998). We assess the statistical significance of departures from the left-truncated Weibull model by performing the Kolmogorov-Smirnov goodness-of-fit test. We use the $_{\bullet}$K statistic (Muller et al., 1979). The rejection region for the α level of significance is {$_{\bullet}$K > $_{\bullet}$K$_{n;\alpha}$}. The percentage points for $_{\bullet}$K$_{n;\alpha}$ were given by Muller et al. (1979). For this example,

$$_{\bullet}K = 0.220 < \ _{\bullet}K_{n=13;\alpha=0.05} = 0.361. \tag{95}$$

Thus, there is not evidence to rule out the left-truncated Weibull model. It follows from (92), for

$$\alpha = 0.05 < \frac{km}{n+km} = 0.5, \tag{96}$$

that

$$h = \left(x_1^\delta - \frac{s}{n}\left[\left(\frac{km}{\alpha(n+km)}\right)^{\frac{1}{n-1}} - 1\right]\right)^{1/\delta} = \left(6.1 - \frac{170.8}{15}\left[\left(\frac{15}{0.05(15+15)}\right)^{\frac{1}{14}} - 1\right]\right)^{1/0.87} = 5. \tag{97}$$

Thus, the manufacturer has 95% assurance that no failures will occur in each shipment before h = 5 month intervals.

5. Examples

5.1 Example 1
An electronic component is required to pass a performance test of 500 hours. The specification is that 20 randomly selected items shall be placed on test simultaneously, and 5 failures or less shall occur during 500 hours. The cost of performing the test is $105 per hour. The cost of redesign is $5000. Assume that the failure distribution follows the one-parameter

exponential model (15). Three failures are observed at 80, 220, and 310 hours. Should the test be continued?
We have from (19) and (20)

$$\hat{\theta} = \frac{80 + 220 + 310 + 17 \times 310}{3} = 1960 \text{ hours};$$ (98)

$$\hat{p}_{pas} = \int_{500}^{\infty} \frac{17!}{2!\,14!} \frac{\left[\exp\left(-\frac{310}{1960}\right) - \exp\left(-\frac{x_6}{1960}\right)\right]^2}{\left[\exp\left(-\frac{310}{1960}\right)\right]^{17}} \frac{1}{1960}\left[\exp\left(-\frac{x_6}{1960}\right)\right]^{15} dx_6 = 0.79665;$$ (99)

Since

$$\int_{x_r}^{\tau_0} x_s f(x_s \mid x^r, \hat{\sigma}) dx_s = 430.05 \text{ hours} > x_k + \hat{p}_{pas}\frac{c_2}{c_1} = 310 + 0.79665\frac{5000}{105} = 347.94 \text{ hours},$$ (100)

abandon the present test and initiate a redesign.

5.2 Example 2

Consider the following problem. A specification for an automotive hood latch is that, of 30 items placed on test simultaneously, ten or fewer shall fall during 3000 cycles of operation. The cost of performing the test is \$2.50 per cycle. The cost of redesign is \$8500. Seven failures, which follow the Weibull distribution with the probability density function (25), are observed at 48, 300, 315, 492, 913, 1108, and 1480 cycles. Shall the test be continued beyond the 1480th cycle?
It follows from (29) and (30) that $\hat{\sigma} = 2766.6$ and $\hat{\delta} = 0.9043$. In turn, these estimates yield $\hat{p}_{pas} = 0.25098$. Since

$$\int_{x_r}^{\tau_0} x_s f(x_s \mid x^r, \hat{\sigma}) dx_s = 1877.6 \text{ hours} < x_k + \hat{p}_{pas}\frac{c_2}{c_1} = 1480 + 0.25098\frac{8500}{2.5} = 2333.33 \text{ hours},$$
(101)

continue the present test.

6. Stopping Rule in Sequential-Sample Testing

At the planning stage of a statistical investigation the question of sample size (n) is critical. For such an important issue, there is a surprisingly small amount of published literature. Engineers who conduct reliability tests need to choose the sample size when designing a test plan. The model parameters and quantiles are the typical quantities of interest. The large-sample procedure relies on the property that the distribution of the t-like quantities is close to the standard normal in large samples. To estimate these quantities the maximum

likelihood method is often used. The large-sample procedure to obtain the sample size relies on the property that the distribution of the above quantities is close to standard normal in large samples. The normal approximation is only first order accurate in general. When sample size is not large enough or when there is censoring, the normal approximation is not an accurate way to obtain the confidence intervals. Thus sample size determined by such procedure is dubious.

Sampling is both expensive and time consuming. Hence, there are situations where it is more efficient to take samples sequentially, as opposed to all at one time, and to define a stopping rule to terminate the sampling process. The case where the entire sample is drawn at one instance is known as "fixed sampling". The case where samples are taken in successive stages, according to the results obtained from the previous samplings, is known as "sequential sampling".

Taking samples sequentially and assessing their results at each stage allows the possibility of stopping the process and reaching an early decision. If the situation is clearly favorable or unfavorable (for example, if the sample shows that a widget's quality is definitely good or poor), then terminating the process early saves time and resources. Only in the case where the data is ambiguous do we continue sampling. Only then do we require additional information to take a better decision.

In this section, the following optimal stopping rule for determining the efficient sample size sequentially under assigning warranty period is proposed.

6.1 Stopping Rule on the Basis of the Expected Beneficial Effect

Suppose the random variables X_1, X_2, …, all from the same population, are observed sequentially and follow the two-parameter Weibull fatigue-crack initiation lifetime distribution (64). After the nth observation ($n \geq n_0$, where n_0 is the initial sample size needful to estimate the unknown parameters of the underlying probability model for the data) the experimenter can stop and receive the beneficial effect on performance,

$$c_1 h_{(1:m);\alpha}^{PL} - cn, \tag{102}$$

where c_1 is the unit value of the lower conditional $(1-\alpha)$ prediction limit (warranty period) $h_{(1:m);\alpha}^{PL} \equiv h_{(1:m);\alpha}^{PL}(x^n)$ (Nechval et al., 2007a, 2007b), $x^n = (x_1, …, x_n)$, and c is the sampling cost.

Below a rule is given to determine if the experimenter should stop in the nth observation, x_n, or if he should continue until the (n+l)st observation, X_{n+1}, at which time he is faced with this decision all over again.

Consider $h_{(1:m);\alpha}^{PL}(X_{n+1}, x^n)$ as a function of the random variable X_{n+1}, when x_1, …, x_n are known, then it can be found its expected value

$$E\left\{h_{(1:m);\alpha}^{PL}(X_{n+1}, x^n) \mid x^n\right\} = \int_0^\infty \int_0^\infty h_{(1:m);\alpha}^{PL}(x_{n+1}, x^n) f(x_{n+1}, v \mid x^n) dx_{n+1} dv. \tag{103}$$

where

$$f(x_{n+1}, v \mid x^n) = \frac{nv^{n-2}e^{v\hat{\delta}\sum\limits_{i=1}^{n}\ln\left(\frac{x_i}{\hat{\beta}}\right)}ve^{v\hat{\delta}\ln\left(\frac{x_{n+1}}{\hat{\beta}}\right)}\hat{\delta}x_{n+1}^{-1}}{\int\limits_0^\infty v^{n-2}e^{v\hat{\delta}\sum\limits_{i=1}^{n}\ln\left(\frac{x_i}{\hat{\beta}}\right)}\left(\sum\limits_{i=1}^{n}e^{v\hat{\delta}\ln\left(\frac{x_i}{\hat{\beta}}\right)}\right)^{-n}dv}\left(e^{v\hat{\delta}\ln\left(\frac{x_{n+1}}{\hat{\beta}}\right)}+\sum\limits_{i=1}^{n}e^{v\hat{\delta}\ln\left(\frac{x_i}{\hat{\beta}}\right)}\right)^{-(n+1)}, \quad (104)$$

the maximum likelihood estimates $\hat{\beta}$ and $\hat{\delta}$ of β and δ, respectively, are determined from the equations (66) and (67), $\int\limits_0^\infty f(x_{n+1}, v \mid x^n)dv$ is the predictive probability density function of X_{n+1}.

Now the optimal stopping rule is to determine the expected beneficial effect on performance for continuing

$$c_1 E\left\{h_{(1:m);\alpha}^{PL}(X_{n+1}, x^n) \mid x^n\right\} - c(n+1) \tag{105}$$

and compare this with (102).

If

$$c_1\left(E\left\{h_{(1:m);\alpha}^{PL}(X_{n+1}, x^n) \mid x^n\right\} - h_{(1:m);\alpha}^{PL}(x^n)\right) > c, \tag{106}$$

it is profitable to continue;

If

$$c_1\left(E\left\{h_{(1:m);\alpha}^{PL}(X_{n+1}, x^n) \mid x^n\right\} - h_{(1:m);\alpha}^{PL}(x^n)\right) \leq c, \tag{107}$$

the experimenter should stop.

7. Conclusions

Determining when to stop a statistical test is an important management decision. Several stopping criteria have been proposed, including criteria based on statistical similarity, the probability that the system has a desired reliability, and the expected cost of remaining faults. This paper presents a new stopping rule in fixed-sample testing based on the statistical estimation of total costs involved in the decision to continue beyond an early failure as well as a stopping rule in sequential-sample testing to determine when testing should be stopped.

The paper considers the problem that can be stated as follows. A new product is submitted for lifetime testing. The product will be accepted if a random sample of n items shows less than s failures in performance testing. We want to know whether to stop the test before it is completed if the results of the early observations are unfavorable. A suitable stopping decision saves the cost of the waiting time for completion. On the other hand, an incorrect stopping decision causes an unnecessary design change and a complete rerun of the test. It

is assumed that the redesign would improve the product to such an extent that it would definitely be accepted in a new lifetime testing. The paper presents a stopping rule based on the statistical estimation of total costs involved in the decision to continue beyond an early failure. Sampling is both expensive and time consuming. The cost of sampling plays a fundamental role and since there are many practical situations where there is a time cost and an event cost, a sampling cost per observed event and a cost per unit time are both included. Hence, there are situations where it is more efficient to take samples sequentially, as opposed to all at one time, and to define a stopping rule to terminate the sampling process. One of these situations is considered in the paper. The practical applications of the stopping rules are illustrated with examples.

8. Acknowledgments

9. References

Amster, S. J. (1963). A modified bayes stopping rule. *The Annals of Mathematical Statistics*, Vol. 34, pp. 1404-1413

Arrow, K. J. ; Blackwell, D. & Girshick, M. A., (1949). Bayes and minimax solutions of sequential decision problems. *Econometrica*, Vol. 17, pp. 213-244

El-Sayyad, G. M. & Freeman, P. R. (1973). Bayesian sequential estimation of a Poisson process rate. *Biometrika*, Vol. 60, pp. 289-296

Freeman, P. R. (1970). Optimal bayesian sequential estimation of the median effective dose. *Biometrika*, Vol. 57, pp. 79-89

Freeman, P. R. (1972). Sequential estimation of the size of a population. *Biometrika*, Vol. 59, pp. 9-17

Freeman, P. R. (1973). Sequential recapture. *Biometrika*, Vol. 60, pp. 141-153

Freeman, P. R. (1983). The secretary problem and its extensions: a review. *International Statistical Review*, Vol. 51, pp. 189-206

Hewitt, J.E. (1968). A note on prediction intervals based on partial observations in certain life test experiments. *Technometrics*, Vol. 10, pp. 850-853

Kaminsky, K.S. (1977). Comparison of prediction intervals for failure times when life is exponential. *Technometrics*, Vol. 19, pp. 83-86

Kendall, M. G. & Stuart, A. S. (1969). *The Advanced Theory of Statistics*, Vol. 1 (3rd edition), Charles Griffin and Co. Ltd, London

Lawless, J.F. (1971). A prediction problem concerning samples from the exponential distribution with applications in life testing. *Technometrics*, Vol. 13, pp. 725-730

Likes, J. (1974). Prediction of sth ordered observation for the two-parameter exponential distribution. *Technometrics*, Vol. 16, pp. 241-244

Lindley, D. V. & Barnett, B.N. (1965). Sequential sampling: two decision problems with linear losses for binomial and normal random variables. *Biometrika*, Vol. 52, pp. 507- 532

Lingappaiah, G.S. (1973). Prediction in exponential life testing. *Canadian Journal of Statistics*, Vol. 1, pp. 113-117 Muller, P.H.; Neumann, P. & Storm, R. (1979). *Tables of Mathematical Statistics*, VEB Fachbuchverlag, Leipzig

Nechval, N. A. (1982). *Modern Statistical Methods of Operations Research*, RCAEI, Riga

Nechval, N. A. (1984). *Theory and Methods of Adaptive Control of Stochastic Processes*, RCAEI, Riga

Nechval, N.A. (1986). Effective invariant embedding technique for designing the new or improved statistical procedures of detection and estimation in signal processing systems, In : *Signal Processing III: Theories and Applications*, Young, I. T. et al. (Eds.), pp. 1051-1054, Elsevier Science Publishers B.V., North-Holland

Nechval, N. A. (1988a). A general method for constructing automated procedures for testing quickest detection of a change in quality control. *Computers in Industry*, Vol. 10, pp. 177-183

Nechval, N. A. (1988b). A new efficient approach to constructing the minimum risk estimators of state of stochastic systems from the statistical data of small samples, In : *Preprint of the 8th IFAC Symposium on Identification and System Parameter Estimation*, pp. 71-76, Beijing, P.R. China

Nechval, N.A. & Nechval, K.N. (1998). Characterization theorems for selecting the type of underlying distribution, In: *Proceedings of the 7th Vilnius Conference on Probability Theory and 22nd European Meeting of Statisticians*, pp. 352-353, TEV, Vilnius

Nechval, N. A. & Nechval, K. N. (1999). Invariant embedding technique and its statistical applications, In : *Conference Volume of Contributed Papers of the 52nd Session of the International Statistical Institute*, Finland, pp. 1-2, ISI – International Statistical Institute, Helsinki, http://www.stat.fi/isi99/proceedings/arkisto/varasto/nech0902.pdf

Nechval, N. A. & Nechval, K. N. (2000). State estimation of stochastic systems via invariant embedding technique, In : *Cybernetics and Systems'2000*, Trappl, R. (Ed.), Vol. 1, pp. 96-101, Austrian Society for Cybernetic Studies, Vienna

Nechval, N. A. ; Nechval, K. N. & Vasermanis, E. K. (2001). Optimization of interval estimators via invariant embedding technique. *IJCAS (An International Journal of Computing Anticipatory Systems)*, Vol. 9, pp. 241-255

Nechval, K. N. ; Nechval N. A. & Vasermanis, E. K. (2003a). Adaptive dual control in one biomedical problem. *Kybernetes (The International Journal of Systems & Cybernetics)*, Vol. 32, pp. 658-665

Nechval, N. A. ; Nechval, K. N. & Vasermanis, E. K. (2003b). Effective state estimation of stochastic systems. *Kybernetes (The International Journal of Systems & Cybernetics)*, Vol. 32, pp. 666-678

Nechval, N. A. & Vasermanis, E. K. (2004). *Improved Decisions in Statistics*, SIA "Izglitibas soli", Riga

Nechval, K. N. ; Nechval, N. A. ; Berzins, G. & Purgailis, M. (2007a). Planning inspections in service of fatigue-sensitive aircraft structure components for initial crack detection. *Maintenance and Reliability*, Vol. 35, pp. 76-80

Nechval, K. N. ; Nechval, N. A. ; Berzins, G. & Purgailis, M. (2007b). Planning inspections in service of fatigue-sensitive aircraft structure components under crack propagation. *Maintenance and Reliability*, Vol. 36, pp. 3-8

Nechval, N. A. ; Berzins, G. ; Purgailis, M. & Nechval, K. N. (2008). Improved estimation of state of stochastic systems via invariant embedding technique. *WSEAS Transactions on Mathematics*, Vol. 7, pp. 141-159

Nechval, N. A. ; Berzins, G. ; Purgailis, M. ; Nechval, K .N. & Zolova, N. (2009). Improved adaptive control of stochastic systems. *Advances In Systems Science and Applications*, Vol. 9, pp. 11-20

Petrucelli, J. D. (1988) Secretary Problem, In : *Encyclopedia of Statistical Sciences*, Kotz, S. & Johnson, N. (Eds.), Vol. 8, pp. 326-329, Wiley, New York

Raiffa, H. & Schlaifer, R. (1968). *Applied Statistical Decision Theory*, Institute of Technology Press, Massachusetts

Samuels, S. M. (1991). Secretary Problems, In : *Handbook of Sequential Analysis*, Ghosh, B. K. & Sen, P. K. (Eds.), pp. 35-60, Dekker, New York

Wald, A. & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, Vol. 19, pp. 326-339

A non-linear double stochastic model of return in financial markets

Vygintas Gontis, Julius Ruseckas and Aleksejus Kononovičius
Institute of Theoretical Physics and Astronomy of Vilnius University
Lithuania

1. Introduction

Volatility clustering, evaluated through slowly decaying auto-correlations, Hurst effect or $1/f$ noise for absolute returns, is a characteristic property of most financial assets return time series Willinger et al. (1999). Statistical analysis alone is not able to provide a definite answer for the presence or absence of long-range dependence phenomenon in stock returns or volatility, unless economic mechanisms are proposed to understand the origin of such phenomenon Cont (2005); Willinger et al. (1999). Whether results of statistical analysis correspond to long-range dependence is a difficult question and subject to an ongoing statistical debate Cont (2005).

Extensive empirical analysis of the financial market data, supporting the idea that the long-range volatility correlations arise from trading activity, provides valuable background for further development of the long-ranged memory stochastic models Gabaix et al. (2003); Plerou et al. (2001). The power-law behavior of the auto-regressive conditional duration process Sato (2004) based on the random multiplicative process and it's special case the self-modulation process Takayasu (2003), exhibiting $1/f$ fluctuations, supported the idea of stochastic modeling with a power-law probability density function (PDF) and long-range memory. Thus the agent based economic models Kirman & Teyssiere (2002); Lux & Marchesi (2000) as well as the stochastic models Borland (2004); Gontis et al. (2008; 2010); Queiros (2007) exhibiting long-range dependence phenomenon in volatility or trading volume are of great interest and remain an active topic of research.

Properties of stochastic multiplicative point processes have been investigated analytically and numerically and the formula for the power spectrum has been derived Gontis & Kaulakys (2004). In the more recent papers Kaulakys et al. (2006); Kaulakys & Alaburda (2009); Ruseckas & Kaulakys (2010) the general form of the multiplicative stochastic differential equation (SDE) was derived in agreement with the model earlier proposed in Gontis & Kaulakys (2004). Since Gontis & Kaulakys (2004) a model of trading activity, based on a SDE driven Poisson-like process, was presented Gontis et al. (2008) and in the most recent paper Gontis et al. (2010) we proposed a double stochastic model, whose return time series yield two power-law statistics, i.e., the PDF and the power spectral density (PSD) of absolute return, mimicking the empirical data for the one-minute trading return in the NYSE.

In this chapter we present theoretical arguments and empirical evidence for the non-linear double stochastic model of return in financial markets. With empirical data from NYSE and Vilnius Stock Exchange (VSE) demonstrating universal scaling of return statistical properties,

which is also present in the double stochastic model of return Gontis et al. (2010). The sections in this chapter follow the chronology of our research papers devoted to the stochastic modeling of financial markets. In the second sections we introduce multiplicative stochastic point process reproducing $1/f^{\beta}$ noise and discuss it's possible application as the stochastic model of financial market. In the section 3 we derive multiplicative SDE statistically equivalent to the introduced point process. Further, in the section 4 we propose a Poisson-like process driven by multiplicative SDE. More sophisticated version of SDE reproducing statistics of trading activity in financial markets is presented in the section 5 and empirical analysis of high frequency trading data from NYSE in the section 6. Section 7 introduces the stochastic model with a q-Gaussian PDF and power spectrum $S(f) \sim 1/f^{\beta}$ and the section 8 the double stochastic model of return in financial market. We present scaled empirical analysis of return in New York and Vilnius stock exchanges in comparison with proposed model in the sections 9. Short conclusions of the most recent research results is presented in the section 10.

2. $1/f$ noise: from physics to financial markets

The PSD of a large variety of different evolutionary systems at low frequencies have $1/f$ behavior. $1/f$ noise is observed in condensed matter, river discharge, DNA base sequence structure, cellular automatons, traffic flow, economics, financial markets and other complex systems with the evolutionary elements of self-organization (see, e.g., a bibliographic list of papers by Li (2009)). Considerable amount of such systems have fractal nature and thus their statistics exhibit scaling. It is possible to define a stochastic model system exhibiting fractal statistics and $1/f$ noise, as well. Such model system may represent the limiting behavior of the dynamical or deterministic complex systems, explaining the evolution of the complexity into chaotic regime.

Let us introduce a multiplicative stochastic model for the time interval between events in time series, defining in such a way the multiplicative point process. This model exhibits the first order and the second order power-law statistics and serves as the theoretical description of the empirical trading activity in the financial markets Gontis & Kaulakys (2004).

First of all we consider a signal $I(t)$ as a sequence of the random correlated pulses

$$I(t) = \sum_k a_k \delta(t - t_k) \tag{1}$$

where a_k is a contribution to the signal of one pulse at the time moment t_k, e.g., a contribution of one transaction to the financial data. Signal (1) represents a point process used in a large variety of systems with the flow of point objects or subsequent actions. When $a_k = \bar{a}$ is constant, the point process is completely described by the set of times of the events $\{t_k\}$ or equivalently by the set of inter-event intervals $\{\tau_k = t_{k+1} - t_k\}$.

Various stochastic models of τ_k can be introduced to define a such stochastic point process. In the papers Kaulakys & Meškauskas (1998); Kaulakys (1999; 2000) it was shown analytically that the relatively slow Brownian fluctuations of the inter-event time τ_k yield $1/f$ fluctuations of the signal (1). In the generalized version of the model Gontis & Kaulakys (2004) we have introduced a stochastic multiplicative process for the inter-event time τ_k,

$$\tau_{k+1} = \tau_k + \gamma \tau_k^{2\mu-1} + \tau_k^{\mu} \sigma \varepsilon_k. \tag{2}$$

Here the inter-event time τ_k fluctuates due to the external random perturbation by a sequence of uncorrelated normally distributed random variable $\{\varepsilon_k\}$ with zero expectation and unit

variance, σ denotes the standard deviation of the white noise and $\gamma \ll 1$ is a damping constant. Note that from the big variety of possible stochastic processes we have chosen the multiplicative one as it yields multifractal intermittency and power-law PDF. Certainly, in Eq. (2) the τ_k diffusion has to be restricted in some area $0 < \tau_{min} < \tau_k < \tau_{max}$. Multiplicativity is specified by μ (pure multiplicativity corresponds to $\mu = 1$, while other values of might be considered as well).

The iterative relation (2) can be rewritten as Langevin SDE in k-space, inter-event space,

$$d\tau_k = \gamma\tau_k^{2\mu-1} + \sigma\tau_k^{\mu}dW_k. \tag{3}$$

Here we interpret k as continuous variable while W_k defines the Wiener noise in inter-event space.

Steady state solution of the stationary Fokker-Planck equation with zero flow, corresponding to (3), gives the probability density function for τ_k in the k-space (see, e.g., Gardiner (1986))

$$P_k(\tau_k) = C\tau_k^{\alpha} = \frac{\alpha+1}{\tau_{max}^{(\alpha+1)} - \tau_{min}^{(\alpha+1)}}\tau_k^{\alpha}, \quad \alpha = 2\gamma/\sigma^2 - 2\mu. \tag{4}$$

The steady state solution (4) assumes Ito convention involved in the relation between expressions (2), (3) and (4) and the restriction for the diffusion $0 < \tau_{min} < \tau_k < \tau_{max}$. In the limit $\tau_{min} \to 0$ and $\tau_{max} \to \infty$ the explicit expression of the signal's $I(t)$ PSD $S_{\mu}(f)$ was derived in Gontis & Kaulakys (2004):

$$S_{\mu}(f) = \frac{C\bar{a}^2}{\sqrt{\pi\bar{\tau}}(3-2\mu)f}\left(\frac{\gamma}{\pi f}\right)^{\frac{\alpha}{3-2\mu}}\frac{\Gamma(\frac{1}{2} + \frac{\alpha}{3-2\mu})}{\cos(\frac{\pi\alpha}{2(3-2\mu)})}. \tag{5}$$

Equation (5) reveals that the multiplicative point process (2) results in the PSD $S(f) \sim 1/f^{\beta}$ with the scaling exponent

$$\beta = 1 + \frac{2\gamma/\sigma^2 - 2\mu}{3 - 2\mu}. \tag{6}$$

Analytical results (5) and (6) were confirmed with the numerical calculations of the PSD according to equations (1) and (2).

Let us assume that $a \equiv 1$ and the signal $I(t)$ counts the transactions in financial markets. In that case the number of transactions in the selected time window τ_d, defined as $N(t) = \int_t^{t+\tau_d} I(t)dt$, measures the trading activity. PDF of N for the pure multiplicative model, with $\mu = 1$, can be expressed as, for derivation see Gontis & Kaulakys (2004),

$$P(N) = \frac{C'\tau_d^{2+\alpha}(1+\gamma N)}{N^{3+\alpha}(1+\frac{\gamma}{2}N)^{3+\alpha}} \sim \begin{cases} \frac{1}{N^{3+\alpha}}, & N \ll \gamma^{-1}, \\ \frac{1}{N^{5+2\alpha}}, & N \gg \gamma^{-1}. \end{cases} \tag{7}$$

Numerical calculations confirms the obtained analytical result (7).

In the case of pure multiplicativity, $\mu = 1$, the model has only one parameter, $2\gamma/\sigma^2$, which defines scaling of the PSD, the power-law distributions of inter-event time and the number of deals N per time window. The model proposed with the adjusted parameter $2\gamma/\sigma^2$ nicely describes the empirical PSD and the exponent of power-law long range distribution of the trading activity N in the financial markets, see Gontis & Kaulakys (2004) for details.

Ability of the model to simulate $1/f$ noise as well as to reproduce long-range power-law statistics of trading activity in financial markets promises wide interpretation and application of the model. Nevertheless, there is an evident need to introduce Poisson-like flow of trades in high frequency time scales of financial markets.

3. Power-law statistics arising from the nonlinear stochastic differential equations

In the previous section we introduced the stochastic multiplicative point process, which was proposed in Gontis & Kaulakys (2004), presented a formula for the PSD and discussed a possible application of the model to reproduce the long-range statistical properties of financial markets. The same long-range statistical properties pertaining to the more general ensemble of stochastic systems can be derived from the SDE or by the related Fokker-Plank equation. Supposing that previously introduced multiplicative point process reflects long-range statistical properties of financial markets, we feel the need to derive multiplicative SDE statistically equivalent to the introduced point process. It would be very nice if the SDE was applicable towards the modeling of financial markets as well.

Transition from the occurrence number, k, space SDE to the actual time, t, space in the SDE(3) can be done according to the relation $dt = \tau_k dk$. This transition yields

$$d\tau = \gamma \tau^{2\mu-2} + \sigma \tau^{\mu-1/2} dW. \tag{8}$$

One can transform variables in the SDE (8) from inter-event time, τ, to the average intensity of the signal, $I(t)$, which itself can be expressed as $x = a/\tau$, or to the number of events per unit time interval $n = 1/\tau$. Applying Ito transform of variables to the SDE (8) gives new SDE for x

$$dx = (\sigma^2 - \gamma)\frac{x^{4-2\mu}}{a^{3-2\mu}} + \frac{\sigma x^{5/2-\mu}}{a^{3/2-\mu}} dW. \tag{9}$$

One can introduce scaled time

$$t_s = \frac{\sigma^2}{a^{3-2\mu}} t, \tag{10}$$

and some new parameters

$$\eta = \frac{5}{2} - \mu, \quad \lambda = 2\left(\frac{\gamma}{\sigma^2} + \frac{3}{2} - \mu\right), \tag{11}$$

in order to obtain the class of Ito SDE

$$dx = (\eta - \frac{\lambda}{2})x^{2\eta-1} + x^\eta dW_s. \tag{12}$$

Eq. (12), as far as it corresponds to the point process discussed in the previous section, should generate the power-law distributions of the signal intensity,

$$P(x) \sim x^{-\lambda}, \tag{13}$$

and $1/f^\beta$ noise,

$$S(f) \sim \frac{1}{f^\beta}, \quad \beta = 1 - \frac{3-\lambda}{2\eta-2}. \tag{14}$$

In some cases time series obtained from SDE (12) may diverge, thus hampering numerical calculations. In the real systems some diffusion restriction mechanisms are present, thus restricting diffusion of SDE solutions seems rather natural. One can introduce the exponential restriction into SDE (12) setting distribution densities Gardiner (1986):

$$P(x) \sim x^{-\lambda} \exp\left\{ -\left(\frac{x_{min}}{x}\right)^m - \left(\frac{x}{x_{max}}\right)^m \right\}. \tag{15}$$

In that case SDE (12) is rewritten as

$$dx = \left(\eta - \frac{\lambda}{2} + \frac{m}{2}\left\{\left(\frac{x_{min}}{x}\right)^m - \left(\frac{x}{x_{max}}\right)^m\right\}\right) x^{2\eta-1} + x^\eta dW_s, \tag{16}$$

where m is parameter responsible for sharpness of restriction.

Many numerical simulations were performed to prove validity of power-law statistics (14)-(15) for the class of SDE (16) Kaulakys et al. (2006). Recently (see Ruseckas & Kaulakys (2010)) it was shown that power-law statistics (14)-(15) can be derived directly from the SDE, without relying on the formalization of point processes (namely model discussed in previous section). This, more general, derivation serves as additional justification of equations and provides further insights into the origin of power-law statistics.

4. Fractal point process driven by the nonlinear stochastic differential equation

In the previous section starting from the point process (1) we derived the class of nonlinear SDE (12) or, with limits towards diffusion, (16) . One can consider the appropriate SDE as an initial model of long-range power-law statistics driving another point process in microscopic level. In Gontis & Kaulakys (2006; 2007) we proposed to model trading activity in financial markets as Poisson-like process driven by nonlinear SDE.

From the SDE class (16) one can draw SDE for the number of point events or trades per unit time interval, n, which would be expressed as

$$dn = \left\{\eta - \frac{\lambda}{2} + \frac{m}{2}\left(\frac{n_0}{n}\right)^m\right\} n^{2\eta-1} dt_s + n^\eta dW_s. \tag{17}$$

The Poisson-like flow of events can be introduced by conditional probability of inter-event time, τ,

$$\varphi(\tau|n) = n \exp(-n\tau). \tag{18}$$

Note that here τ is measured in scaled time, t_s, units and the expectation of instantaneous inter-event time, for instantaneous n, is

$$\langle\tau\rangle_n = \int_0^\infty \tau\varphi(\tau|n)d\tau = \frac{1}{n}. \tag{19}$$

The long-range PDF of n, time series obtained from the with Eq. (17) related Fokker-Plank equation, has an explicit form:

$$P_m^{(t)}(n) = \frac{m}{n_0\Gamma(\frac{\lambda-1}{m})}\left(\frac{n_0}{n}\right)^\lambda \exp\left(-\left(\frac{n_0}{n}\right)^m\right). \tag{20}$$

Similarly the long-range PDF of τ can be written as

$$P_m^{(t)}(\tau) = \tau \int\limits_0^\infty n\varphi(\tau|n)P(n)\mathrm{d}n = \frac{m}{\tau_0\Gamma(\frac{\lambda-1}{m})}\frac{\tau}{\tau_0}\int\limits_0^\infty x^{2-\lambda}\exp\left(-\frac{1}{x^m}-x\frac{\tau}{\tau_0}\right)\mathrm{d}x, \quad (21)$$

here we have performed substitution of parameters $\tau_0 = \frac{1}{n_0}$. The integral in (21) has an explicit form, when $m = 1$

$$P_{m=1}(\tau) = \frac{2}{\tau_0\Gamma(\lambda-1)}\left(\frac{\tau}{\tau_0}\right)^{\frac{\lambda-1}{2}}K_{\lambda-3}\left(2\sqrt{\frac{\tau}{\tau_0}}\right) \qquad (22)$$

Here $K_n(z)$ is the modified Bessel function of the second kind. When $\tau \to \infty$, we get

$$P_{m=1}(\tau) \approx \frac{\sqrt{\pi}}{\tau_0\Gamma(\lambda-1)}\left(\frac{\tau}{\tau_0}\right)^{\frac{\lambda}{2}-\frac{3}{4}}\exp\left(-2\sqrt{\frac{\tau}{\tau_0}}\right). \qquad (23)$$

The integral in (21) can be expressed via special functions, when $m = 2$. However, we can obtain asymptotic behavior for small and large $\frac{\tau}{\tau_0}$. When $\frac{\tau}{\tau_0} \to \infty$, using the method of the steepest descent we get

$$P_{m=2}^{(t)}(\tau) = \frac{2^{\frac{10-\lambda}{3}}}{\tau_0\Gamma(\frac{\lambda-1}{2})}\sqrt{\frac{\pi}{3}}\left(\frac{\tau}{\tau_0}\right)^{\frac{\lambda-1}{3}}\exp\left(-3\left(\frac{\tau}{2\tau_0}\right)^{\frac{2}{3}}\right), \qquad (24)$$

while in case of $\frac{\tau}{\tau_0} \to 0$ one can obtain

$$P_{m=2}^{(t)}(\tau) \to \begin{cases} \frac{2\Gamma(3-\lambda)}{\tau_0\Gamma(\frac{\lambda-1}{2})}\left(\frac{\tau}{\tau_0}\right)^{\lambda-2} & 1 < \lambda < 3 \\ \frac{\Gamma(\frac{\lambda-3}{2})}{\tau_0\Gamma(\frac{\lambda-1}{2})}\frac{\tau}{\tau_0} & \lambda > 3 \end{cases} \qquad (25)$$

5. Fractal trading activity of financial market driven by the nonlinear stochastic differential equation

We will investigate how previously introduced modulated Poisson stochastic point process can be adjusted to the empirical trading activity, defined as number of transactions in the selected time window τ_d. In order to obtain the number of events, N, in the selected time window, τ_d, one has to integrate the stochastic signal Eq. (17) in the corresponding time interval. We denote the integrated number of events, N, as

$$N(t, \tau_\mathrm{d}) = \int_t^{t+\tau_\mathrm{d}} n(t')\mathrm{d}t' \qquad (26)$$

and call it the trading activity in case of the financial markets.

Detrended fluctuation analysis Plerou et al. (2000) is one of the ways to analyze the second order statistics related to the autocorrelation of trading activity. The exponents of the detrended fluctuation analysis, ν, obtained by fits for each of the 1000 US stocks show a relatively narrow spread of ν around the mean value $\nu = 0.85 \pm 0.01$ Plerou et al. (2000). We use relation $\beta = 2\nu - 1$ between the exponents ν of the detrended fluctuation analysis and the exponents β of the PSD Beran (1994) and in this way define the empirical value of the exponent for

the power spectral density $\beta = 0.7$. Our analysis of the Vilnius stock exchange (VSE) data confirmed that the PSD of trading activity is the same for various liquid stocks even for the emerging markets Gontis & Kaulakys (2004). The histogram of exponents obtained by fits to the cumulative distributions of trading activites of 1000 US stocks Plerou et al. (2000) gives the value of exponent $\lambda = 4.4 \pm 0.05$ describing the power-law behavior of the trading activity. Empirical values of $\beta = 0.7$ and $\lambda = 4.4$ confirm that the time series of the trading activity in real markets are fractal with the power law statistics. Time series generated by stochastic process (17) are fractal in the same sense.

Nevertheless, we face serious complications trying to adjust model parameters to the empirical data of the financial markets. For the pure multiplicative model, setting $\mu = 1$ or $\eta = 3/2$, we have to take $\lambda = 2.7$ to get $\beta = 0.7$, while empirical λ value being noticeably different - 4.4, i.e. it is impossible to reproduce the empirical PDF and PSD with the same exponent of multiplicativity η. We have proposed possible solution of this problem in our publications Gontis & Kaulakys (2004) deriving PDF for the trading activity N, see Eq. (7). When $N \gg \gamma^{-1}$ one can obtain exactly the required values of $\lambda = 5 + 2\alpha = 4.4$ and $\beta = 0.7$ for $\gamma_\sigma = \frac{\gamma}{\sigma^2} = 0.85$. Despite model being able to mimic empirical data under certain conditions, we cannot accept it as the sufficiently accurate model of the trading activity since the empirical power law distribution is achieved only for very high values of the trading activity. This discrepancy provides insight to the mechanism of the power law distribution converging to the normal distribution through increasing values of the exponent, though empirically observed power law distribution in wide area of N values cannot be reproduced. Let us notice here that the desirable power law distribution of the trading activity with the exponent $\lambda = 4.4$ may be generated by the model (17) with $\eta = 5/2$. Moreover, only the smallest values of τ or high values of n contribute to the power spectral density of trading activity Kaulakys et al. (2006). Thus we feel incentive to combine the stochastic processes with two values of μ or η: (i) $\mu \simeq 0$ or $\eta \simeq 5/2$ for the main area of τ and n diffusion and (ii) $\mu = 1$ or $\eta \simeq 3/2$ for the lowest values of τ or highest values of n. Therefore, we introduce a new SDE for n, which includes two powers of the multiplicative noise,

$$\mathrm{d}n = \left[(\frac{5}{2} - \frac{\lambda}{2}) + \frac{m}{2} \left(\frac{n_0}{n} \right)^m \right] \frac{n^4}{(n\epsilon + 1)^2} \mathrm{d}t + \frac{n^{5/2}}{(n\epsilon + 1)} \mathrm{d}W, \tag{27}$$

where a new parameter ϵ defines crossover between two areas of n diffusion. The corresponding iterative equation for τ_k in such a case is expressed as

$$\tau_{k+1} = \tau_k + \left[(\frac{\lambda + 2}{2} - \eta) - \frac{m}{2} \left(\frac{\tau}{\tau_0} \right)^m \right] \frac{\tau_k}{(\epsilon + \tau_k)^2} + \frac{\tau_k}{\epsilon + \tau_k} \varepsilon_k, \tag{28}$$

where ε_k denotes uncorrelated normally distributed random variable with the zero expectation and unit variance.

Eqs. (27) and (28) define related stochastic variables n and τ, respectively, and they should reproduce the long-range statistical properties of the trading activity and of waiting time in the financial markets. We verify this by the numerical calculations. In Figure 1 we present the PSD calculated for the equivalent processes (a)-(27) and (b)-(28) (see Gontis & Kaulakys (2004) for details of calculations). This approach reveals the structure of the PSD in wide range of frequencies and shows that the model exhibits not one, but two rather different power laws with the exponents $\beta_1 = 0.34$ and $\beta_2 = 0.74$. In Figure 1 we also present the distributions of trading activity (c) and (d), which now have correct exponents. From many numerical calculations performed with the multiplicative point processes we can conclude that combination

of two power laws of spectral density arise only when the multiplicative noise is a crossover
of two power laws as in Eqs. (27) and (28).
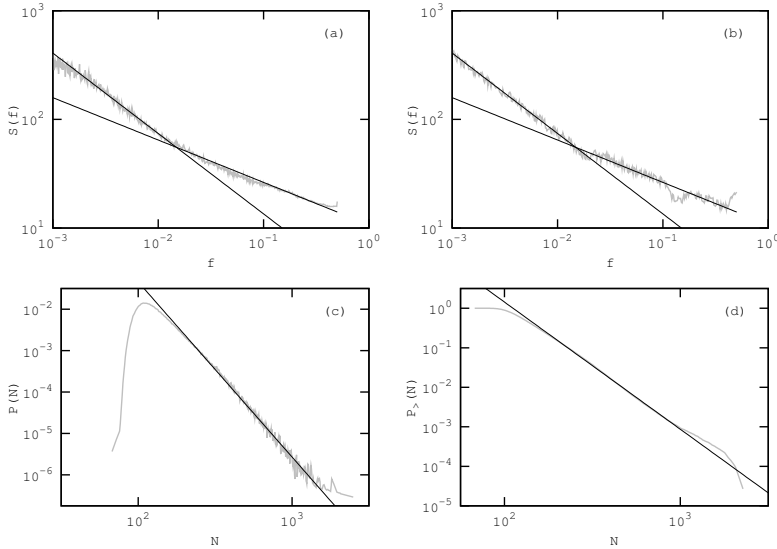
Fig. 1. (a) PSD, $S(f)$, of the τ dependent flow generated by Eq. (28), (b) $S(f)$ calculated from n
time series generated by Eq. (27), (c) PDF, $P(N)$, of number of trades, obtained by integrating
signal, within time window of $\tau_d = 100s$, N, generated by Eq. (28), (d) corresponding inverse
cumulative distribution, $P_>(N)$, of N. Statistical properties are represented by gray curves,
while black lines approximate their power laws: (a) and (b) black lines give approximation
$S(f) \sim 1/f^{\beta_{1,2}}$ with $\beta_1 = 0.34$ and $\beta_2 = 0.74$, (c) black line gives approximation $P(N) \sim N^{-\lambda}$
with exponent $\lambda = 4.3$, (d) black line gives approximation $P_>(N) \sim N^{-\lambda}$ with exponent
$\lambda = 3.3$. Statistical properties were obtained using corresponding model parameter values:
$\lambda = 4.3$; $\epsilon = 0.07$; $\tau_0 = 1$; $m = 6$.

Thus as of now we have introduced the complete set of equations defining the stochastic
model of the trading activity in the financial markets. We have proposed this model follow-
ing our growing interest in the stochastic fractal point processes Gontis & Kaulakys (2004);
Kaulakys et al. (2005; 2006). Our objective to reproduce, in details, statistics of trading activity
is the cause for rather complicated form of the SDE (27) and thus there is low expectation of
analytical results. Therefore we focus on the numerical analysis and direct comparison of the
model with the empirical data.
In order to achieve general description of statistics for different stocks we introduce the scaling
into Eq. (27) utilizing scaled rate $x = n/n_0$ and $\varepsilon' = \varepsilon n_0$. After the substitution Eq. (27)
becomes

$$\mathrm{d}x = \left[\left(\frac{5}{2} - \frac{\lambda}{2} \right) + \frac{m}{2} x^{-m} \right] \frac{x^4}{(x\varepsilon' + 1)^2} \mathrm{d}t + \frac{x^{5/2}}{(x\varepsilon' + 1)} \mathrm{d}W. \tag{29}$$

We have eliminated parameter n_0 as it is specific for each stock. By doing so we also de-
crease number of model parameters to three, which must be defined from the empirical data
of trading activity in the financial markets.

One can solve Eq. (29) using the method of discretization. Thus we introduce variable step of integration $\Delta t = h_k = \kappa^2/x_k$, and the differential equation (29) transforms into the set of difference equations

$$x_{k+1} = x_k + \kappa^2 \left[\left(\frac{5}{2} - \frac{\lambda}{2} \right) + \frac{m}{2} x_k^{-m} \right] \frac{x_k^3}{(x_k \epsilon' + 1)^2} + \kappa \frac{x_k^2}{(x_k \epsilon' + 1)} \varepsilon_k, \tag{30}$$

$$t_{k+1} = t_k + \kappa^2/x_k \tag{31}$$

with $\kappa \ll 1$ being small parameter and ε_k defining Gaussian noise with zero mean and unit variance .

With the substitution of variables, namely $\tau = 1/n$, one can transform Eq. (27) into

$$d\tau = \left[\frac{\lambda - 3}{2} - \frac{m}{2} \left(\frac{\tau}{\tau_0} \right)^m \right] \frac{1}{(\epsilon + \tau)^2} dt + \frac{\sqrt{\tau}}{\epsilon + \tau} dW \tag{32}$$

with limiting time $\tau_0 = 1/n_0$. Further we argue that this form of driving SDE is more suitable for the numerical analysis. First of all, the powers of variables in this equation are lower, but the main advantage is that the Poissonian-like process can be included into the procedure of numerical solution of SDE. As we did with SDE for n we should also introduce a scaling of Eq. (32). It is done by defining the non-dimensional scaled time $t_s = t/\tau_0$, scaled inter-trade time $y = \tau/\tau_0$ and $\epsilon' = \epsilon/\tau_0$. After those transformations Eq. (32) becomes

$$dy = \frac{1}{\tau_0^2} \left[\frac{\lambda - 3}{2} - \frac{m}{2} y^m \right] \frac{1}{(\epsilon' + y)^2} dt_s + \frac{1}{\tau_0} \frac{\sqrt{y}}{\epsilon' + y} dW_s. \tag{33}$$

As in the real discrete market trading we can choose the instantaneous inter-trade time y_k as a step of numerical calculations, $h_k = y_k$, or even more precisely as the random variables with the exponential distribution $P(h_k) = 1/y_k \exp(-h_k/y_k)$. We obtain iterative equation resembling tick by tick trades in the financial markets,

$$y_{k+1} = y_k + \frac{1}{\tau_0^2} \left[\frac{\lambda - 3}{2} - \frac{m}{2} y_k^m \right] \frac{h_k}{(\epsilon' + y_k)^2} + \frac{1}{\tau_0} \frac{\sqrt{y_k h_k}}{\epsilon' + y_k} \varepsilon_k. \tag{34}$$

In this numerical procedure the sequence of $1/y_k$ gives the modulating rate, n, and the sequence of h_k is the Poissonian-like inter-trade times. Seeking higher precision one can use the Milshtein approximation for Eq. (33) instead of Eq. (34).

6. Analysis of empirical stock trading data

Previously, see Gontis et al. (2008), we have analyzed the tick by tick trades of 26 stocks on NYSE traded for 27 months from January, 2005. In this chapter we will briefly discuss main results and valuable conclusions, providing important insights, of the empirical analysis presented in Gontis et al. (2008). Empirical analysis is very important as starting from it we can adjust the parameters of the Poisson-like process driven by SDE Eq. (27) or Eq. (34) to numerically reproduce the empirical trading statistics.

An example of the empirical histogram of τ_k and $N(t, \tau_d)$ and the PSD of IBM trade sequence are shown on Figure 2. The histograms and PSD of the sequences of trades for all 26 stocks are similar to IBM shown on Fig. 2. From the histogram, $P(\tau_k)$, we can obtained model parameter τ_0 value for every stock. One can define the exponent λ' from the power-law tail $P(N) \sim N^{-\lambda'}$
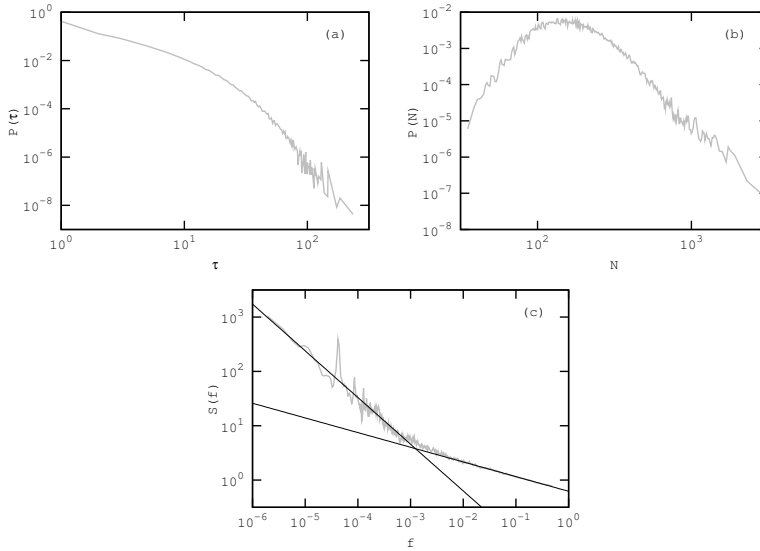
Fig. 2. Trading statistics of IBM stocks. (a) Empirical histogram of the inter-trade time τ_k sequence, $P(\tau)$; (b) histogram of trading activity, $P(N)$, calculated in the time interval $\tau_d = 600$ s; (c) PSD, $S(f)$, of the sequence of trades (gray curve), straight black lines approximate PSD $S(f) \sim 1/f^{\beta_{1,2}}$ with $\beta_1 = 0.33$ and $\beta_2 = 0.94$.

of N histogram. The PSD exhibits two scaling exponents β_1 and β_2 if approximated by power-law $S(f) \sim f^{-\beta_{1,2}}$.

Empirical values of β_1 and β_2 fluctuate around 0.3 and 0.9, respectively, same behavior is observed in different stochastic model realizations. The crossover frequency f_c of two power-laws exhibits some fluctuations around the value $f_c \approx 10^{-3}$ Hz as well. One can observe considerable fluctuations of the exponent λ' around the mean value 4.4. We would like to note that the value of histogram exponent, λ', for integrated trading activity N is higher than for n, as λ' increases with higher values of time scale τ_d.

From the point of view of the proposed model parameter τ_0 is specific for every stock and reflects the average trading intensity in the calm periods of stock exchange. In previous section we have shown that one can eliminate these specific differences in the model by scaling transform of Eq. (32) arriving to the nondimensional SDE (33) and its iterative form (34). These equations and parameters $\sigma' = \sigma/\tau_0$, λ, ϵ' and $m = 2$ define model, which has to reproduce, in details, power-law statistics of the trading activity in the financial markets. From the analysis based on the research of fractal stochastic point processes Gontis & Kaulakys (2004; 2006; 2007); Kaulakys et al. (2005; 2006) and by fitting the numerical calculations to the empirical data we arrive at the conclusion that model parameters should be set as $\sigma' = 0.006$, $\lambda = 4.3$, $\epsilon' = 0.05$ in order to achieve best results. In Figure 3 we have presented statistical properties obtained from our model using aforementioned parameter values - PDF of the sequence of $\tau_k = h_k$, (a), and the PSD of the sequence of trades as point events, (b).

For every selected stock one can easily scale the model sequence of inter-trade times $\tau_k = h_k$ by empirically defined τ_0 to get the model sequence of trades for this stock. One can scale the
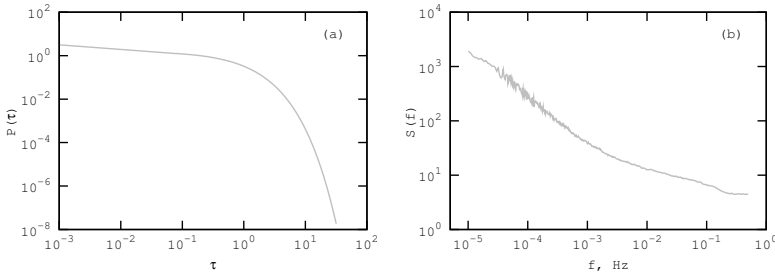
Fig. 3. Distribution of the Poisson-like inter-event times $\tau_k = h_k$, (a), and power spectrum, (b), of the sequence of point events calculated from Eq. (34) with the adjusted parameters $m = 2$, $\sigma' = 0.006$, $\lambda = 4.3$, $\epsilon' = 0.05$, $\tau_0 = 1$.
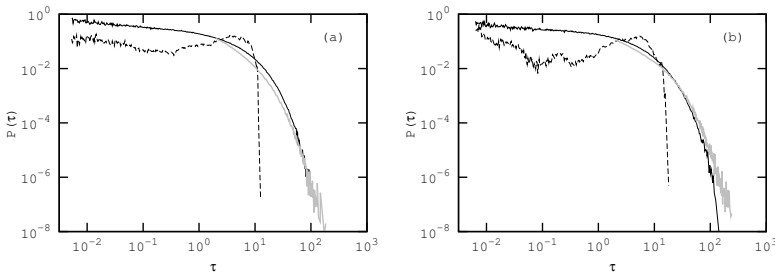
Fig. 4. Distribution of inter-trade times, τ, for (a) IBM and (b) MMM stocks; empirical histogram, gray curve, modeled Poisson-like distribution, black solid curve, distribution of driving $\tau = y_k$ in Eq. (34), black dashed curve. Model parameters are the same as in Fig. 3. $\tau_0 = 5$s for IBM and $\tau_0 = 7.25$s for MMM stocks.

model power spectrum $S(f)$ by $1/\tau_0^2$ for getting the model power spectrum $S_{\text{stock}}(f)$ for the selected stock $S_{\text{stock}}(f) = S(f)/\tau_0^2$.

Previously we have proposed the iterative Eq. (34) as quite accurate stochastic model of trading activity in the financial markets. Nevertheless, one has to admit that real trading activity often has considerable trend as number of shares traded and the whole activity of the markets increases. This might have considerable influence on the empirical long-range distributions and power spectrum of the stocks in consideration. The trend has to be eliminated from the empirical data for the detailed comparison with the model. Only few stocks from the selected list have stable trading activity in the considered period.

In Figure 4, Figure 5 and Figure 6 we compare the model statistical properties with the empirical statistics of the stocks with stable trading activity. As we show in Figure 4, the model Poisson-like distribution can be easily adjusted to the empirical histogram of inter-trade time, with $\tau_0 = 5$ s for IBM trade sequence and with $\tau_0 = 7.25$ s for MMM trading. The comparison with the empirical data is limited by the available accuracy, 1 s, of stock trading time t_k. The probability distribution of driving $\tau = y_k$ Eq. (34), dashed line, illustrates different market behavior in the periods of the low and high trading activity. The Poissonian nature of the stochastic point process hides these differences by considerable smoothing of the PDF. In Figure 5 one can see that the long-range memory properties of the trading activity reflected in
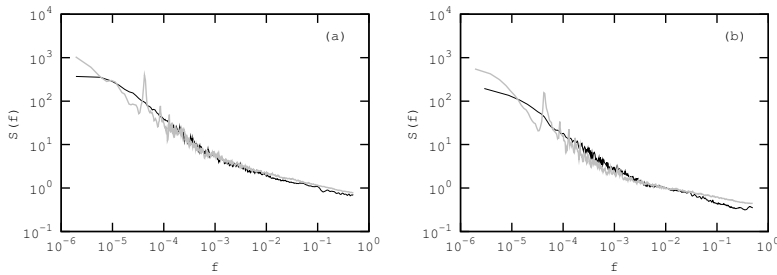
Fig. 5. Modeled, black curves, and empirical, gray curves, PSD of trading activity, N, for (a) IBM and (b) MMM stocks. Parameters are the same as in Fig. 3. $\tau_0 = 5$s for IBM and $\tau_0 = 7.25$s for MMM stocks.
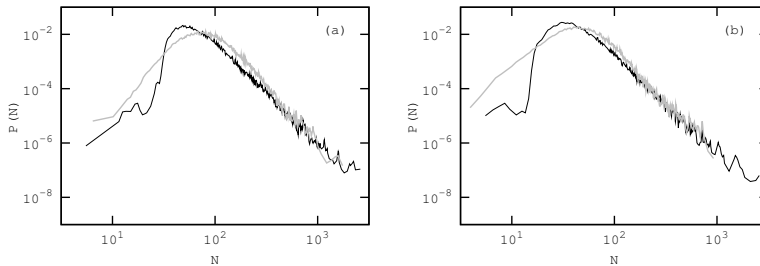
Fig. 6. Modeled, black curve, and empirical, gray curve, PDF of trading activity, N, for (a) IBM and (b) MMM stocks in the time interval $\tau_d = 300$s. Parameters are the same as in Fig. 3. $\tau_0 = 5$s for IBM and $\tau_0 = 7.25$s for MMM stocks.

the PSD are universal and arise from the scaled driving SDE (29) and (33). One can obtain the PSD of the selected stock's trading sequence by scaling model PSD, Figure 3 (b), by $1/\tau_0^2$. The PDF of integrated trading activity N is more sensitive to the market fluctuations. Even the intraday fluctuations of market activity, which are not included in this model, make considerable influence on PDF of N for low values. Nevertheless, as we demonstrate in Figure 6, the model is able to reproduce the power-law tails very well.

In this section we have shown results of the empirical analysis of stocks traded on NYSE. We have used those results as a basis for adjustment of the previously introduced trading activity model parameters. Aforementioned model is based on Poisson-like process, which we have introduced as scalable in previous sections, similar scalability as we see in this section is an inherent feature of actual financial markets.

A new form of scaled equations provides the universal description with the same parameters applicable for all stocks. The proposed new form of the continuous stochastic differential equation enabled us to reproduce the main statistical properties of the trading activity and waiting time, observable in the financial markets. In proposed model the fractured power-law distribution of spectral density with two different exponents arise. This is in agreement with the empirical power spectrum of the trading activity and volatility and implies that the market behavior may be dependent on the level of activity. One can observe at least two stages in market behavior: calm and excited. Ability to reproduce empirical PDF of inter-trade time

and trading activity as well as the power spectrum in very detail for various stocks provides a background for further stochastic modeling of volatility.

7. The stochastic model with a q-Gaussian PDF and power spectrum $S(f) \sim 1/f^\beta$

In section (3) we have introduced the class of SDE (12), (16) exhibiting power-law statistics and proposed Poisson like process modulated by this type of SDE. The latter serves as an appropriate model of trading activity in the financial markets Gontis et al. (2008). In this section we generalize the earlier proposed nonlinear SDE within the non-extensive statistical mechanics framework, Tsallis (2009), to reproduce the long-range statistics with a q-Gaussian PDF and power spectrum $S(f) \sim 1/f^\beta$.

The q-Gaussian PDF of stochastic variable r with variance σ_q^2 can be written as

$$P(r) = A_q \exp_q \left(-\frac{r^2}{(3-q)\sigma_q^2} \right), \tag{35}$$

here A_q is a constant of normalization, while q defines the power law part of the distribution. $P(r)$ is introduced through the variational principle applied to the generalized entropy Tsallis (2009), which is defined as

$$S_q = k\frac{1 - \int [p(r)]^q dr}{1-q}.$$

The q-exponential of variable x is defined as

$$\exp_q(x) = (1 + (1-q)x)^{\frac{1}{1-q}} \tag{36}$$

here we assume that the q-mean $\mu_q = 0$. With some transformation of parameters σ_q and q, namely

$$\lambda = \frac{2}{q-1}, \qquad r_0 = \sigma_q \sqrt{\frac{3-q}{q-1}},$$

we can rewrite the q-Gaussian PDF in a more transparent form:

$$P_{r_0,\lambda}(r) = \frac{\Gamma(\lambda/2)}{\sqrt{\pi}r_0\Gamma(\lambda/2 - 1/2)} \left(\frac{r_0^2}{r_0^2 + r^2} \right)^{\frac{\lambda}{2}}. \tag{37}$$

Looking for the appropriate form of the SDE we start from the general case of a multiplicative equation in the Ito convention with Wiener process W:

$$dr = a(r)dt + b(r)dW. \tag{38}$$

If the stationary distribution of SDE (38) is the q-Gaussian (37), then the coefficients of drift, $a(r)$, and diffusion, $b(r)$, in the SDE are related as follows Gardiner (1986):

$$a(r) = -\frac{\lambda}{2}\frac{r}{r_0^2 + r^2}b(r)^2 + b(r)\frac{db(r)}{dr}. \tag{39}$$

From our previous experience modeling one-over-f noise and trading activity in financial markets Gontis & Kaulakys (2004); Kaulakys et al. (2005), building nonlinear stochastic differential equations exhibiting power law statistics Kaulakys et al. (2006); Kaulakys & Alaburda

(2009), described here in previous sections, we know that processes with power spectrum $S(f) \sim 1/f^\beta$ can be obtained using the multiplicative term $b(r) \sim r^\eta$ or even a slightly modified form $(r_0^2 + r^2)^{\frac{\eta}{2}}$. Therefore, we choose the term $b(r)$ as

$$b(r) = \sigma(r_0^2 + r^2)^{\frac{\eta}{2}} \tag{40}$$

and, consequently, by Eq. (39) we arrive at

$$a(r) = \sigma^2 \left(\eta - \frac{\lambda}{2} \right) (r_0^2 + r^2)^{\eta - 1} r. \tag{41}$$

Having defined drift, Eq. (41), and diffusion, Eq. (40), terms one obtains this SDE:

$$dr = \sigma^2 \left(\eta - \frac{\lambda}{2} \right) (r_0^2 + r^2)^{\eta - 1} r dt + \sigma(r_0^2 + r^2)^{\frac{\eta}{2}} dW. \tag{42}$$

Note that in the simple case $\eta = 1$ Eq. (42) coincides with the model presented in the article Queiros (2007) with

$$b(r) = \sqrt{\frac{\theta}{P(r)^{\frac{2}{\lambda}}}}, \qquad a(r) = -\frac{\theta}{r_0^2} \left(\frac{\lambda}{2} - 1 \right) r \tag{43}$$

Further we will investigate higher values of η in order to cache long-range memory properties of the absolute return in the financial markets. First of all, let us scale our variables

$$x = \frac{r}{r_0}, \qquad t_s = \sigma^2 r_0^{2(\eta - 1)} t \tag{44}$$

to reduce the number of parameters and to get simplified equations. Then SDE

$$dx = \left(\eta - \frac{\lambda}{2} \right) (1 + x^2)^{\eta - 1} x dt_s + (1 + x^2)^{\frac{\eta}{2}} dW_s \tag{45}$$

describes a stochastic process with a stationary q-Gaussian distribution

$$P_\lambda(x) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(\lambda/2)}{\Gamma(\lambda/2 - 1/2)} \left(\frac{1}{1 + x^2} \right)^{\frac{\lambda}{2}} \tag{46}$$

and the power spectral density of the signal $S(f)$

$$S(f) = \frac{A}{f^\beta}, \qquad \beta = 1 + \frac{\lambda - 3}{2(\eta - 1)} \tag{47}$$

$$A = \frac{(\lambda - 1)\Gamma(\beta - 1/2)}{2\sqrt{\pi}(\eta - 1)\sin(\pi\beta/2)} \left(\frac{2 + \lambda - 2\eta}{2\pi} \right)^{\beta - 1} \tag{48}$$

with $0.5 < \beta < 2$, $4 - \eta < \lambda < 1 + 2\eta$ and $\eta > 1$. Eqs. (47-48) were first derived for the multiplicative point process in Gontis & Kaulakys (2004); Kaulakys et al. (2005) and generalized for the nonlinear SDE (42) in Kaulakys et al. (2006); Kaulakys & Alaburda (2009). Although Eq. (42) coincides with Eq. (15) in ref. Kaulakys & Alaburda (2009) only for high values of the variable $r \gg r_0$, these are the values responsible for the PSD. Note that the frequency f in equation (47) is the scaled frequency matching the scaled time t_s (44). The scaled equations (44)-(48) define a stochastic model with two parameters λ and η responsible for the power

law behavior of the signal PDF and power spectrum. Numerical calculations with Eq. (45) confirm analytical formulas (46-48) (see ref. Kaulakys & Alaburda (2009)).

We will need a more sophisticated version of the SDE to reproduce a stochastic process with a fractured PSD of the absolute return, which is observable in financial markets. Having in mind the statistics of the stochastic model (45) defined by Eqs. (46)-(48) and numerical modeling with more sophisticated versions of the SDE (27),(29), we propose an equation combining two powers of multiplicativity

$$\mathrm{d}x = \left(\eta - \frac{\lambda}{2} - \left(\frac{x}{x_{max}}\right)^2\right)\frac{(1+x^2)^{\eta-1}}{((1+x^2)^{\frac{1}{2}}\epsilon + 1)^2}x\mathrm{d}t_s + \frac{(1+x^2)^{\frac{\eta}{2}}}{(1+x^2)^{\frac{1}{2}}\epsilon + 1}\mathrm{d}W_s. \quad (49)$$

In modified SDE (49) model parameter ϵ divides area of x diffusion into two different power law regions to ensure the PSD of $|x|$ with two power law exponents. A similar procedure has been introduced in the model of trading activity Gontis et al. (2008), see previous sections. This procedure provides an approach to the market with behavior dependent on the level of activity and exhibiting two stages: calm and excited. Thus it is not surprising that Eq. (49) models the stochastic return x with two power law statistics, i.e., the PDF and the PSD, reproducing the empirical power law exponents of the trading return in the financial markets. At the same time, via the term $(\frac{x}{x_{max}})^2$ we introduce the exponential diffusion restriction for the high values of x as the markets in the excited stage operate on the limit of non-stationarity. One can solve Eq. (49) in the same manner we did solve trading activity related SDE (29) and (33). Thus we introduce the variable step of numerical integration

$$h_k = \kappa^2\frac{((x_k^2+1)^{\frac{1}{2}}\epsilon + 1)^2}{(x_k^2+1)^{\eta-1}},$$

the differential equation (49) transforms into the set of difference equations

$$x_{k+1} = x_k + \kappa^2\left(\eta - \frac{\lambda}{2} - (x\epsilon^\eta)^2\right)x_k + \kappa(x_k^2+1)^{\frac{1}{2}}\varepsilon_k \quad (50)$$

$$t_{k+1} = t_k + \kappa^2\frac{((x_k^2+1)^{\frac{1}{2}}\epsilon + 1)^2}{(x_k^2+1)^{\eta-1}} \quad (51)$$

The continuous stochastic variable x does not include any time scale as the return defined in a time window τ should. Knowing that the return is an additive variable and depends on the number of transactions in a similar way to trading activity, we define the scaled return X in the time period τ as the integral of the continuous stochastic variable $X = \int_t^{t+\tau} x(t_s)/\tau\,\mathrm{d}t_s$. Note that τ here is measured in scaled time units Eq. (44) though relation between model and empirical time series scales can be established and is very useful then adjusting model and empirical statistical properties.

It is worth recalling that integration of the signal in the time interval τ does not change the behavior of the power spectrum for the frequencies $f \ll \frac{1}{\tau}$. This is just the case we are interested in for the long-range memory analysis of financial variables and we can expect Eqs. (47-48) to work for the stochastic variable X as well.

We have also previously analyzed the influence of signal integration on the PDF in previous modeling of trading activity (see Gontis & Kaulakys (2004)). Integration of the nonlinear stochastic signal increases the exponent of the power law tails in the area of the highest values
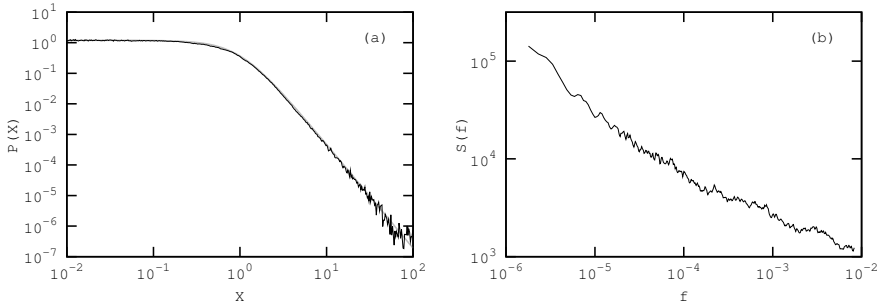
Fig. 7. (a) The numerically calculated PDF of $|X| = |\int_t^{t+\tau} x(t)/\tau\, dt_s|$ from Eq. (51) (black line), in comparison with the theoretical distribution $2P(x)$ Eq. (46) (gray line), and (b) the numerically calculated PSD of $|X|$. Model parameters are set as follows: $\eta = 5/2$, $\lambda = 3.6$, $\tau = 0.0001$ and $\epsilon = 0.01$.

of the integrated signal. This hides fractured behavior of the X PDF, which arises for x as a consequence of the two powers in the multiplicative term of Eq. (49).

In Fig. 12 we demonstrate (a) the numerically calculated PDF of $|X|$ in comparison with the theoretical distribution $2P(x)$ Eq. (46) and (b) the numerically calculated power spectrum of $|X|$ with parameters appropriate for reproducing statistics for the absolute return in financial markets.

8. The double stochastic model of return in financial market

Recently we proposed the double stochastic model of return in financial markets Gontis et al. (2010) based on the nonlinear SDE (49). The main advantage of proposed model is its ability to reproduce power spectral density of absolute return as well as long-term PDF of return. In the model proposed we assume that the empirical return r_t can be written as instantaneous q-Gaussian fluctuations ξ with a slowly diffusing parameter r_0 and constant $\lambda = 5$

$$r_t = \xi\{r_0(t), \lambda\}. \tag{52}$$

q-Gaussian distribution of ξ defining the random instantaneous r_t can be written as follows:

$$P_{r_0,\lambda}(\xi) = \frac{\Gamma(\frac{\lambda}{2})}{r_0\sqrt{\pi}\Gamma(\frac{\lambda}{2} - \frac{1}{2})} \left(\frac{r_0^2}{r_0^2 + \xi^2}\right)^{\lambda/2}, \tag{53}$$

with parameter $r_0(t)$ serving as a measure of instantaneous volatility of return fluctuations. We will model stochastic $r_0(t)$ in the similar way as trading activity in it's stochastic model. In this case nonlinear SDE (49) will serve as modulating one for q-Gaussian return fluctuations. The empirical evidence of this assumption is published in Gontis et al. (2010). The return, $|r|$, we define as absolute difference of logarithms of asset prices, p, in two different time moments separated by time interval τ:

$$r(t, \tau) = |\ln[p(t + \tau)] - \ln[p(t)]|. \tag{54}$$

In empirical analysis we consider dimensionless returns normalized by its dispersion calculated in the whole length of realization. It is worth to notice that $r(\tau)$ is an additive variable,
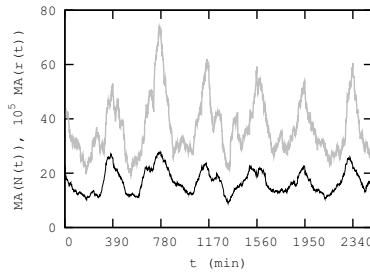
Fig. 8. An example of a moving average for 60 min of empirical absolute returns per minute (gray curve) in comparison with the corresponding moving average of trading activity, number of trades per minute (black curve). Scales are adjusted (magnitude of return time series multiplied by 10^5).

i.e., if $\tau = \sum_i \tau_i$, then $r(\tau) = \sum_i r(\tau_i)$, or in the continuous limit the sum may be replaced by integration. We do propose to model the measure of volatility r_0 by the scaled continuous stochastic variable x, having a meaning of average return per unit time interval. By the empirical analyses of high frequency trading data on NYSE Gontis et al. (2010) we introduced relation:

$$r_0(t, \tau) = 1 + \frac{\bar{r}_0}{\tau_s} \left| \int_{t_s}^{t_s + \tau_s} x(s) \mathrm{d}s \right|, \tag{55}$$

where $\bar{r}_0$ is an empirical parameter and the average return per unit time interval $x(t_s)$ can be modeled by the nonlinear SDE (49), written in a scaled dimensionless time $t_s = \sigma_i^2 t$.

We have performed the empirical analyses (see Gontis et al. (2010)) of the tick by tick trades of 24 stocks, ABT, ADM, BMY, C, CVX, DOW, FNM, GE, GM, HD, IBM, JNJ, JPM, KO, LLY, MMM, MO, MOT, MRK, SLE, PFE, T, WMT, XOM, traded on the NYSE for 27 months from January, 2005, recorded in the Trades and Quotes database. We summed empirical tick by tick returns into one-minute returns to adjust the continuous stochastic model presented. Detailed analysis of the empirical data from the NYSE provides evidence that long-range memory properties of the return strongly depend on fluctuations of trading activity. In Fig. 8 we demonstrate strong correlation of the moving average of absolute returns per minute with the moving average of trading activities (number of trades per minute). Here for the empirical sequences of one-minute returns $\{r_t\}_{t=1}^T$ or trading activities $\{N_t\}_{t=1}^T$ we calculate moving averages MA defined as the centered means for a selected number of minutes n; for example, $\mathrm{MA}(r_t)$ is

$$\mathrm{MA}(r_t) = \frac{1}{n} \sum_{j=t-n/2}^{t+n/2-1} r_j. \tag{56}$$

The best correlation can be achieved when the moving averages are calculated in the period from 60 to 100 minutes.

In order to account for the double stochastic nature of return fluctuations - a hidden slowly diffusing long-range memory process and rapid fluctuations of the instantaneous price changes - we decompose the empirical one-minute return series into two processes: the background

fluctuations and the high amplitude rapid fluctuations dependent on the first one modulating. To perform this empirical decomposition already presented as background idea of the model (52), we assume that the empirical return r_t can be written as instantaneous q-Gaussian fluctuations with a slowly diffusing parameter r_0 dependent on the moving average of the empirical return r_t:

$$r = \xi\{r_0(\mathrm{MA}(r_t)), \lambda_2\}, \tag{57}$$

where $\xi\{r_0, \lambda_2\}$ is a q-Gaussian stochastic variable with the PDF defined by Eq. (53) (the parameter q is $q = 1 + 2/\lambda_2$). In Eq. (57) the parameter r_0 depends on the modulating moving average of returns, $\mathrm{MA}(r_t)$, and the empirically defined power law exponent λ_2. From the empirical time series of the one-minute returns r_t one can draw histograms of r corresponding to defined values of the moving average $\mathrm{MA}(r_t)$. The q-Gaussian PDF is a good approximation to those histograms and the adjusted set of r_0 for selected values of $\mathrm{MA}(r_t)$ gives an empirical definition of the function

$$r_0(\mathrm{MA}(r_t)) = 1 + 2 \times |\mathrm{MA}(r_t)|. \tag{58}$$

The q-Gaussians with $\lambda_2 = 5$ and linear function $r_0(|\mathrm{MA}(r_t)|)$ (58) give a good approximation of r fluctuations for all stocks and values of modulating $\mathrm{MA}(r_t)$. The long-term PDF of moving average $\mathrm{MA}(r_t)$ can be approximated by a q-Gaussian with $\bar{r}_0 = 0.2$ and $\lambda = 3.6$. All these empirically defined parameters form the background for the stochastic model of the return in the financial market.

Consequently, we propose to model the long-range memory stochastic return $\mathrm{MA}(r_t)$ by $X = \frac{\bar{r}_0}{\tau_s}\int_{t_s}^{t_s+\tau_s} x(s) / \, \mathrm{d}s$, where x is a continuous stochastic variable defined by Eq. (49) and $\bar{r}_0 = \bar{r}_0 \times 2 = 0.4$. The remaining parameters ϵ, x_{max} and σ_t^2 can be adjusted for the best model fit to the empirical data and have values $\epsilon = 0.017$, $\sigma_t^2 = 1/3 \times 10^{-6} s^{-1}$ and $\tau_s = \tau \times \sigma_t^2 = 0.00002$; $x_{max} = 1000$.

The parameters of stochastic model were adjusted to the empirical tick by tick one minute returns. An excellent agreement between empirical and model PDF and power spectrum was achieved, see Fig 9 (a,b). Noticeable difference in theoretical and empirical PDFs for small values of return r are related with the prevailing prices of trades expressed in integer values of cents. We do not account for this discreteness in our continuous description. In the empirical PSD one-day resonance - the largest spike with higher harmonics - is present. This seasonality - an intraday activity pattern of the signal - is not included in the model either and this leads to the explicable difference from observed power spectrum.

9. Scaled comparison of model with empirical data

Seeking to discover the universal nature of financial markets we consider that all these parameters are universal for all stocks traded on various exchanges. To prove this we analyze empirical data from very different exchanges New York, one of the most developed with highly liquid stocks, and Vilnius, emerging one with stocks traded rarely. The comparison of model and empirical data scaling with increasing time window of return definition, τ, serves as very significant test for proposed stochastic description of financial markets.

Provided that we use scaled dimensionless equations derived while making very general assumptions, we expect that proposed model should work for various assets traded on different financial markets as well as for various time scales τ. We analyze tick by tick trades of 4 stocks, APG1L, PTR1L, SRS1L, UKB1L, traded on VSE for 50 months since May, 2005, trading data was collected and provided for us by VSE. Stocks traded on VSE in comparison with NYSE are less liquid - mean inter-trade time for analyzed stocks traded on VSE is 362 s, while for

stocks traded on NYSE mean inter-trade time equals 3.02 s. The difference in trading activity exceeds 100 times. This great difference is related with comparatively small number of traders and comparatively small companies participating in the emerging VSE market. Do these different markets have any statistical affinity is an essential question from the theoretical point of market modeling.

First of all we start with returns for very small time scales $\tau = 60$ s. For the VSE up to 95% of one minute trading time intervals elapse without any trade or price change. One can exclude these time intervals from the sequence calculating PDF of return. With such simple procedure calculated PDF of VSE empirical return overlaps with PDF of NYSE empirical return (see Fig 9 (a)).
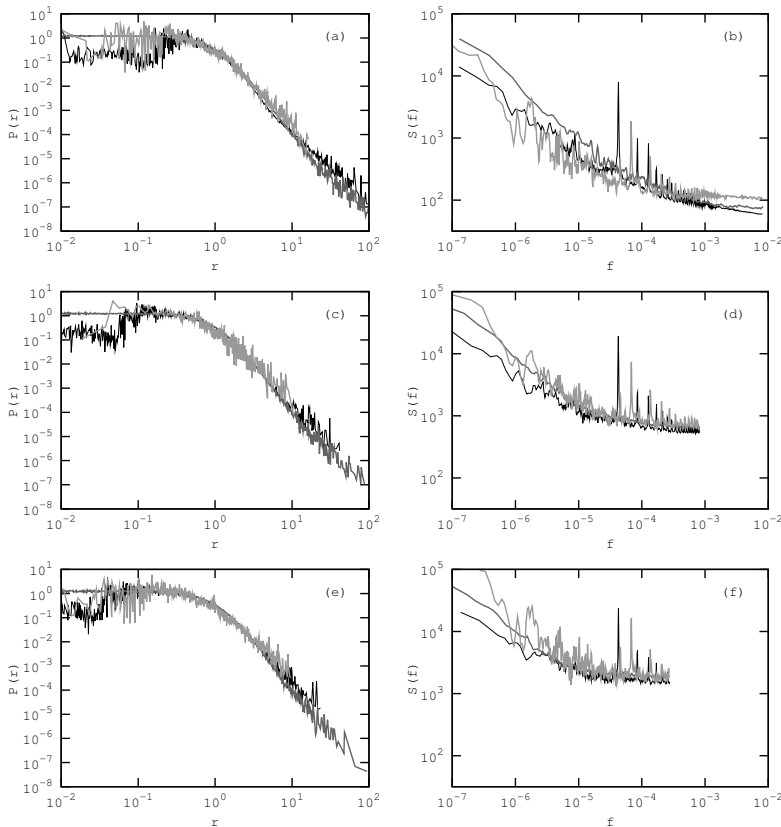
Fig. 9. Comparison of empirical statistics of absolute returns traded on the NYSE (black curves) and VSE (light gray curves) with model, defined by Eq. (49) and Eq. (55), statistics (gray curves). Model parameters are as follows: $\lambda = 5$; $\sigma_t^2 = 1/3 \cdot 10^{-6} s^{-1}$; $\lambda_0 = 3.6$; $\epsilon = 0.017$; $\eta = 2.5$; $\bar{r}_0 = 0.4$; $x_{max} = 1000$. PDF of normalized absolute returns is given on (a),(c),(e) and PSD on (b),(d),(f). (a) and (b) represents results with $\tau = 60s$; (c) and (d) $\tau = 600s$; (e) and (f) $\tau = 1800s$. Empirical data from NYSE is averaged over 24 stocks and empirical data from VSE is averaged over 4 stocks.

One should use full time series of returns calculating the PSD. Nevertheless, despite low VSE liquidity, PSD of VSE and NYSE absolute returns almost overlap. Difference is clearly seen only for higher frequencies and smaller time windows, namely $\tau = 60$ s, and is related directly to the low VSE liquidity, which contributes to the white noise appearance. The different length of trading sessions in financial markets causes different positions of resonant intraday activity spikes. Thus one can conclude that even so marginal market as VSE retains essential statistical features as developed market on NYSE. At the first glance the statistical similarity should be even better for the higher values of return time scale τ.

Therefore further we investigate the behavior of returns on NYSE and VSE for increased values of $\tau = 600$ s and $\tau = 1800$ s with the specific interest to check whether proposed stochastic model scales in the same way as empirical data does. Apparently, as we can see in Fig 9 (d) and (f) PSDs of absolute returns on VSE and on NYSE overlap even better at larger time scale (600 seconds and 1800 seconds). This serves as an additional argument for the very general origin of long range memory properties observed in very different, liquidity-wise, markets. The nonlinear SDE is an applicable model to cache up observed empirical properties. PDFs of absolute return observed in both markets (see Fig 9 (c) and (e)) are practically identical, though we still have to ignore zero returns of VSE to arrive to the same normalization of PDF.

10. Conclusions

In the last sections we introduced a double stochastic process driven by the nonlinear scaled SDE ((49)) reproducing the main statistical properties of the absolute return, observed in the financial markets. Seven parameters of the model enable us to adjust it to the sophisticated power-law statistics of various stocks including long-range behavior. The scaled non-dimensional form of equations gives an opportunity to deal with averaged statistics of various stocks and compare behavior of different markets. All parameters introduced are recoverable from the empirical data and are responsible for the specific statistical features of real markets. Seeking to discover the universal nature of return statistics we have analysed and compared extremely different markets in New York and Vilnius and adjust the model parameters to match statistics of both markets. The most promising result of this research is discovered increasing coincidence of the model with empirical data from the New York and Vilnius markets and between markets, when the time scale of return τ is growing. Further analyses of empirical data and proposed model reasoning by agent behavior is ongoing.

11. References

Beran, J. (1994). *Statistics for Long-Memory Processes*, Chapman and Hall, New York.

Borland, L. & Bouchaud, J.-Ph. (2004). On a multi-timescale statistical feedback model for volatility fluctuations. E-print, arXiv:physics/0507073.

Cont, R. (2005). Long range dependence in financial markets, In: *Fractals in Engineering*, Lutton, E. & Vehel, J., (Eds.), p. 159-180, Springer, Berlin.

Gabaix, X.; Gopikrishnan, P.; Plerou, V. & Stanley, H.E. (2003). A theory of power-law distributions in financial market fluctuations. *Nature*, 423, p. 267-270.

Gardiner, C.W. (1986). *Handbook of Stochastic Methods*, Springer, Berlin.

Gontis, V. & Kaulakys, B. (2004). Multiplicative point process as a model of trading activity. *Physica A*, 343, p. 505-514. Gontis, V. & Kaulakys, B. (2004). Modeling financial markets by the multiplicative sequence of trades. *Physica A*, 344, p. 128-133. Gontis, V.;

Kaulakys, B.; Alaburda, M. & Ruseckas J. (2004). Evolution of Complex Systems and 1/f Noise: from Physics to Financial Markets. *Solid State Phenomena*, 97-98, p. 65-70.

Gontis, V. & Kaulakys, B. (2006). Long-range memory model of trading activity and volatility. *Journal of Statistical Mechanics*, P10016.

Gontis, V. & Kaulakys, B. (2007). Modeling long-range memory trading activity by stochastic differential equations. *Physica A*, 382, p. 114-120.

Gontis, V.; Kaulakys, B. & Ruseckas, J. (2008). Trading activity as driven Poisson process: comparison with empirical data. *Physica A*, 387, p. 3891-3896.

Gontis, V.; Ruseckas, J. & Kononovičius A. (2010). A long-range memory stochastic model of the return in financial markets. *Physica A*, 389, p. 100-106.

Kaulakys, B. & Meškauskas, T., (1998). Modelling 1/f noise. *Physical Review E*, 58, p. 7013-7019.

Kaulakys, B. (1999). Autoregressive model of 1/f noise. *Physical Letters A*, 257, 1-2, p. 37-42.

Kaulakys, B. (2000). On the intrinsic origin of 1/f noise. *Microelectronics Reliability*, 40, p. 1787-1790.

Kaulakys, B.; Gontis, V. & Alaburda, M. (2005). Point process model of 1/f noise versus a sum of Lorentzians. *Physical Review E*, 71, 051105.

Kaulakys, B.; Ruseckas, J.; Gontis, V. & Alaburda, M. (2006). Nonlinear stochastic models of 1/f noise and power-law distributions. *Physica A*, 365, p. 217-221.

Kaulakys, B. & Alaburda, M. (2009). Modeling scaled processes and $1/f^{\beta}$ noise using non-linear stochastic differential equations. *Journal of Statistical Mechanics*, February 2009, P02051.

Kirman, A. & Teyssiere, G. (2002). Microeconomic models for long-memory in the volatility of financial time series. *Studies in nonlinear dynamics and econometrics*, 5, p. 281-302.

Li, W. (2009). URI: http://www.nslij-genetics.org/wli/1fnoise.

Lux, T. & Marchesi, M. (2000). Volatility clustering in financial markets: a microsimulation of interacting agents. *International Journal of Theoretical and Applied Finance*, 3, p. 675-702.

Plerou, V.; Gopikrishnan, P.; Amaral L.; Gabaix, X. & Stanley, H.E. (2000). Economic fluctuations and anomalous diffusion. *Physical Review E*, 62, R3023-R3026.

Plerou, V.; Gopikrishnan, P.; Gabaix, X. et al. (2001). Price fluctuations, market activity, and trading volume. *Quantitative Finance*, 1, p. 262-269.

Duarte Queiros, S.M. (2007). On a generalised model for time-dependent variance with long-term memory. *Europhysics Letters*, 80, 30005

Ruseckas, J. & Kaulakys B. (2010). 1/f noise from nonlinear stochastic differential equations. *Physical Review E*, 81, 031105.

Sato, A.H. (2004). Explanation of power law behavior of autoregressive conditional duration processes based on the random multiplicative process. *Physical Review E*, 69, 047101.

Takayasu, M. & Takayasu, H. (2003). Self-modulation processes and resulting generic 1/f fluctuations. *Physica A*, 324, p. 101-107.

Tsallis, C. (2009). *Introduction to Nonextensive Statistical Mechanics*. Springer, ISBN 978-0-387-85358-1, New York.

Willinger, W.; Taqqu, M. & Teverovsky, V. (1999). Stock market prices and long-range dependence. *Finance and Stochastics*, 3, p. 1-13.

Mean-variance hedging under partial information

M. Mania[1),2)], R. Tevzadze[1),3)] and T. Toronjadze[1),2)]

[1)]*Georgian American University, Business School*
[2)]*A. Razmadze Mathematical Institute*
[3)]*Institute of Cybernetics*
Georgia

Abstract

We consider the mean-variance hedging problem under partial information. The underlying asset price process follows a continuous semimartingale, and strategies have to be constructed when only part of the information in the market is available. We show that the initial mean-variance hedging problem is equivalent to a new mean-variance hedging problem with an additional correction term, which is formulated in terms of observable processes. We prove that the value process of the reduced problem is a square trinomial with coefficients satisfying a triangle system of backward stochastic differential equations and the filtered wealth process of the optimal hedging strategy is characterized as a solution of a linear forward equation.
2000 Mathematics Subject Classification: 90A09, 60H30, 90C39.
Key words and phrases: Backward stochastic differential equation, semimartingale market model, incomplete markets, mean-variance hedging, partial information.

1. Introduction

In the problem of derivative pricing and hedging it is usually assumed that the hedging strategies have to be constructed by using all market information. However, in reality, investors acting in a market have limited access to the information flow. For example, an investor may observe just stock prices, but stock appreciation rates depend on some unobservable factors; one may think that stock prices can be observed only at some time intervals or up to some random moment before an expiration date, or an investor would like to price and hedge a contingent claim whose payoff depends on an unobservable asset, and he observes the prices of an asset correlated with the underlying asset. Besides, investors may not be able to use all available information even if they have access to the full market flow. In all such cases, investors are forced to make decisions based on only a part of the market information.

We study a mean-variance hedging problem under partial information when the asset price process is a continuous semimartingale and the flow of observable events do not necessarily contain all information on prices of the underlying asset.

We assume that the dynamics of the price process of the asset traded on the market is described by a continuous semimartingale $S = (S_t, t \in [0, T])$ defined on a filtered probability space $(\Omega, \mathcal{A}, (\mathcal{A}_t, t \in [0, T]), P)$, satisfying the usual conditions, where $\mathcal{A} = \mathcal{A}_T$ and $T < \infty$ is the fixed time horizon. Suppose that the interest rate is equal to zero and the asset price

process satisfies the structure condition; i.e., the process S admits the decomposition

$$S_t = S_0 + N_t + \int_0^t \lambda_u d\langle N \rangle_u, \quad \langle \lambda \cdot N \rangle_T < \infty \quad \text{a.s.,} \tag{1.1}$$

where N is a continuous $\mathcal{A}$-local martingale and λ is an $\mathcal{A}$-predictable process.

Let G be a filtration smaller than $\mathcal{A}$: $G_t \subseteq \mathcal{A}_t$ for every $t \in [0, T]$.

The filtration G represents the information that the hedger has at his disposal; i.e., hedging strategies have to be constructed using only information available in G.

Let H be a P-square integrable $\mathcal{A}_T$-measurable random variable, representing the payoff of a contingent claim at time T.

We consider the mean-variance hedging problem

$$\text{to minimize} \quad E[(X_T^{x,\pi} - H)^2] \quad \text{over all} \quad \pi \in \Pi(G), \tag{1.2}$$

where $\Pi(G)$ is a class of G-predictable S-integrable processes. Here $X_t^{x,\pi} = x + \int_0^t \pi_u dS_u$ is the wealth process starting from initial capital x, determined by the self-financing trading strategy $\pi \in \Pi(G)$.

In the case $G = \mathcal{A}$ of complete information, the mean-variance hedging problem was introduced by Föllmer and Sondermann (Föllmer & Sondermann, 1986) in the case when S is a martingale and then developed by several authors for a price process admitting a trend (see, e.g., (Duffie & Richardson, 1991), (Hipp, 1993), (Schweizer, 1992), (Schweizer, 1994), (Schäl, 1994), (Gourieroux et al., 1998), (Heath et al., 2001)).

Asset pricing with partial information under various setups has been considered. The mean-variance hedging problem under partial information was first studied by Di Masi, Platen, and Runggaldier (Di Masi et al., 1995) when the stock price process is a martingale and the prices are observed only at discrete time moments. For general filtrations and when the asset price process is a martingale, this problem was solved by Schweizer (Schweizer, 1994) in terms of G-predictable projections. Pham (Pham, 2001) considered the mean-variance hedging problem for a general semimartingale model, assuming that the observable filtration contains the augmented filtration F^S generated by the asset price process S

$$F_t^S \subseteq G_t \quad \text{for every} \quad t \in [0, T]. \tag{1.3}$$

In this paper, using the variance-optimal martingale measure with respect to the filtration G and suitable Kunita–Watanabe decomposition, the theory developed by Gourieroux, Laurent, and Pham (Gourieroux et al., 1998) and Rheinländer and Schweizer (Rheinlander & Schweizer, 1997) to the case of partial information was extended.

If G is not containing F^S, then S is not a G-semimartingale and the problem is more involved. Let us introduce an additional filtration $F = (F_t, t \in [0, T])$, which is an augmented filtration generated by F^S and G.

Then the price process S is a continuous F-semimartingale, and the canonical decomposition of S with respect to the filtration F is of the form

$$S_t = S_0 + \int_0^t \widehat{\lambda}_u^F d\langle M \rangle_u + M_t, \tag{1.4}$$

where $\widehat{\lambda}^F$ is the F-predictable projection of λ and

$$M_t = N_t + \int_0^t [\lambda_u - \widehat{\lambda}_u^F] d\langle N \rangle_u$$

is a continuous F-local martingale. Besides $\langle M \rangle = \langle N \rangle$, and these brackets are F^S-predictable. Throughout the paper we shall make the following assumptions:

(A) $\langle M \rangle$ is G-predictable and $d\langle M \rangle_t dP$ a.e. $\widehat{\lambda}^F = \widehat{\lambda}^G$; hence P-a.s. for each t

$$E(\lambda_t | F^S_{t-} \vee G_t) = E(\lambda_t | G_t);$$

(B) any G-martingale is an F-local martingale;

(C) the filtration G is continuous; i.e., all G-local martingales are continuous;

(D) there exists a martingale measure for S (on F_T) that satisfies the reverse Hölder condition.

Remark. It is evident that if $F^S \subseteq G$, then $\langle M \rangle$ is G-predictable. Besides, in this case $G = F$, and conditions (A) and (B) are satisfied.

We shall use the notation $\widehat{Y}_t$ for the process of the G-projection of Y (note that under the present conditions, for all processes we consider, the optional projection coincides with the predictable projection, and therefore we use for them the same notation). Condition (A) implies that

$$\widehat{S}_t = E(S_t | G_t) = S_0 + \int_0^t \widehat{\lambda}_u d\langle M \rangle_u + \widehat{M}_t.$$

Let

$$H_t = E(H|F_t) = EH + \int_0^t h_u dM_u + L_t \quad \text{and} \quad H_t = EH + \int_0^t h^G_u d\widehat{M}_u + L^G_t$$

be the Galtchouk–Kunita–Watanabe (GKW) decompositions of $H_t = E(H|F_t)$ with respect to local martingales M and $\widehat{M}$, where h and h^G are F-predictable processes and L and L^G are local martingales strongly orthogonal to M and $\widehat{M}$, respectively.

We show (Theorem 3.1) that the initial mean-variance hedging problem (1.2) is equivalent to the problem to minimize the expression

$$E\left[\left(x + \int_0^T \pi_u d\widehat{S}_u - \widehat{H}_T \right)^2 + \int_0^T \left(\pi_u^2 \left(1 - \rho_u^2 \right) + 2\pi_u \widetilde{h}_u \right) d\langle M \rangle_u \right] \tag{1.5}$$

over all $\pi \in \Pi(G)$, where

$$\widetilde{h}_t = \widehat{h^G_t} \rho_t^2 - \widehat{h}_t \quad \text{and} \quad \rho_t^2 = \frac{d\langle \widehat{M} \rangle_t}{d\langle M \rangle_t}.$$

Thus, the problem (1.5), equivalent to (1.2), is formulated in terms of G-adapted processes. One can say that (1.5) is the mean-variance hedging problem under complete information with an additional correction term.

Let us introduce the value process of the problem (1.5):

$$V^H(t, x) = \operatorname*{ess\,inf}_{\pi \in \Pi(G)} E\left[\left(x + \int_t^T \pi_u d\widehat{S}_u - \widehat{H}_T \right)^2 + \int_t^T \left[\pi_u^2 \left(1 - \rho_u^2 \right) + 2\pi_u \widetilde{h}_u \right] d\langle M \rangle_u | G_t \right]. \tag{1.6}$$

We show in Theorem 4.1 that the value function of the problem (1.5) admits a representation

$$V^H(t, x) = V_t(0) - 2V_t(1)x + V_t(2)x^2,$$

where the coefficients $V_t(0)$, $V_t(1)$, and $V_t(2)$ satisfy a triangle system of backward stochastic differential equations (BSDEs). Besides, the filtered wealth process of the optimal hedging strategy is characterized as a solution of the linear forward equation

$$\widehat{X}_t^* = x - \int_0^t \frac{\rho_u^2 \varphi_u(2) + \widehat{\lambda}_u V_u(2)}{1 - \rho_u^2 + \rho_u^2 V_u(2)} \widehat{X}_u^* d\widehat{S}_u + \int_0^t \frac{\rho_u^2 \varphi_u(1) + \widehat{\lambda}_u V_u(1) + \widetilde{h}_u}{1 - \rho_u^2 + \rho_u^2 V_u(2)} d\widehat{S}_u. \qquad (1.7)$$

Note that if $F^S \subseteq G$, then

$$\rho = 1, \quad \widetilde{h} = 0, \quad \widehat{M} = M, \quad \text{and} \quad \widehat{S} = S. \qquad (1.8)$$

In the case of complete information ($G = \mathcal{A}$), in addition to (1.8) we have $\widehat{\lambda} = \lambda$ and $\widehat{M} = N$, and (1.7) gives equations for the optimal wealth process from (Mania & Tevzadze, 2003).

In section 5 we consider a diffusion market model, which consists of two assets S and η, where S_t is a state of a process being controlled and η_t is the observation process. Suppose that S_t and η_t are governed by

$$dS_t = \mu_t dt + \sigma_t dw_t^0, \qquad d\eta_t = a_t dt + b_t dw_t,$$

where w^0 and w are Brownian motions with correlation ρ and the coefficients μ, σ, a, and b are $\mathcal{F}^\eta$-adapted. In this case $\mathcal{A}_t = \mathcal{F}_t = \mathcal{F}_t^{S,\eta}$, and the flow of observable events is $\mathcal{G}_t = \mathcal{F}_t^\eta$. As an application of Theorem 4.1 we also consider a diffusion market model with constant coefficients and assume that an investor observes the price process S only up to a random moment τ before the expiration date T. In this case we give an explicit solution of (1.2).

2. Main Definitions and Auxiliary Facts

Denote by $\mathcal{M}^e(F)$ the set of equivalent martingale measures for S, i.e., the set of probability measures Q equivalent to P such that S is a F-local martingale under Q.
Let

$$\mathcal{M}_2^e(F) = \{Q \in \mathcal{M}^e(F) : EZ_T^2(Q) < \infty\},$$

where $Z_t(Q)$ is the density process (with respect to the filtration F) of Q relative to P. We assume that $\mathcal{M}_2^e(F) \neq \varnothing$.

Remark 2.1. Note that $\mathcal{M}_2^e(\mathcal{A}) \neq \varnothing$ implies that $\mathcal{M}_2^e(F) \neq \varnothing$ (see Remark 2.1 from Pham (Pham, 2001).
It follows from (1.4) and condition (A), that the density process $Z_t(Q)$ of any element Q of $\mathcal{M}^e(F)$ is expressed as an exponential martingale of the form

$$\mathcal{E}_t(-\widehat{\lambda} \cdot M + L),$$

where L is a F-local martingale strongly orthogonal to M and $\mathcal{E}_t(X)$ is the Doleans–Dade exponential of X.

If the local martingale $Z_t^{\min} = \mathcal{E}_t(-\widehat{\lambda} \cdot M)$ is a true martingale, $dQ^{min}/dP = Z_T^{min}$ defines the minimal martingale measure for S.
Recall that a measure Q satisfies the reverse Hölder inequality $R_2(P)$ if there exists a constant C such that

$$E\left(\frac{Z_T^2(Q)}{Z_\tau^2(Q)} | F_\tau\right) \leq C, \quad P\text{-a.s.}$$

for every F-stopping time τ.

Remark 2.2. If there exists a measure $Q \in \mathcal{M}^e(F)$ that satisfies the reverse Hölder inequality $R_2(P)$, then according to Theorem 3.4 of Kazamaki (Kazamaki, 1994) the martingale $M^Q = -\widehat{\lambda} \cdot M + L$ belongs to the class BMO and hence $-\widehat{\lambda} \cdot M$ also belongs to BMO, i.e.,

$$E\left(\int_\tau^T \widehat{\lambda}_u^2 d\langle M\rangle_u | F_\tau\right) \leq \text{const} \tag{2.1}$$

for every stopping time τ. Therefore, it follows from Theorem 2.3 of (Kazamaki, 1994) that $\mathcal{E}_t(-\widehat{\lambda} \cdot M)$ is a true martingale. So, condition (D) implies that the minimal martingale measure exists (but Z^{min} is not necessarily square integrable).

Let us make some remarks on conditions (B) and (C).

Remark 2.3. Condition (B) is satisfied if and only if the σ-algebras $F_t^S \vee G_t$ and G_T are conditionally independent given G_t for all $t \in [0, T]$ (see Theorem 9.29 from Jacod (Jacod, 1979)).

Remark 2.4. Condition (C) is weaker than the assumption that the filtration F is continuous. The continuity of the filtration F and condition (B) imply the continuity of the filtration G, but the converse is not true in general. Note that filtrations F and F^S can be discontinuous. Recall that the continuity of a filtration means that all local martingales with respect to this filtration are continuous.

By μ^K we denote the Dolean measure of an increasing process K. For all unexplained notations concerning the martingale theory used below, we refer the reader to (Dellacherie & Meyer, 1980), (Liptser & Shiryaev, 1986), (Jacod, 1979).

Let $\Pi(F)$ be the space of all F-predictable S-integrable processes π such that the stochastic integral

$$(\pi \cdot S)_t = \int_0^t \pi_u dS_u, \quad t \in [0, T],$$

is in the $\mathcal{S}^2$ space of semimartingales, i.e.,

$$E\left(\int_0^T \pi_s^2 d\langle M\rangle_s\right) + E\left(\int_0^T |\pi_s \widehat{\lambda}_s| d\langle M\rangle_s\right)^2 < \infty.$$

Denote by $\Pi(G)$ the subspace of $\Pi(F)$ of G-predictable strategies.

Remark 2.5. Since $\widehat{\lambda} \cdot M \in BMO$ (see Remark 2.2), it follows from the proof of Theorem 2.5 of Kazamaki (Kazamaki, 1994) that

$$E\left(\int_0^T |\pi_u \widehat{\lambda}_u| d\langle M\rangle_u\right)^2 = E\langle |\pi| \cdot M, |\widehat{\lambda}| \cdot M\rangle_T^2 \leq 2\|\widehat{\lambda} \cdot M\|_{BMO} E \int_0^T \pi^2 d\langle M\rangle_u < \infty.$$

Therefore, under condition (D) the G-predictable (resp., F-predictable) strategy π belongs to the class $\Pi(G)$ (resp., $\Pi(F)$) if and only if $E \int_0^T \pi_s^2 d\langle M\rangle_s < \infty$.

Define $J_T^2(F)$ and $J_T^2(G)$ as spaces of terminal values of stochastic integrals, i.e.,

$$J_T^2(F) = \{(\pi \cdot S)_T : \pi \in \Pi(F)\}, \qquad J_T^2(G) = \{(\pi \cdot S)_T : \pi \in \Pi(G)\}.$$

For convenience we give some assertions from (Delbaen et al., 1997), which establishes necessary and sufficient conditions for the closedness of the space $J_T^2(F)$ in L^2.

Proposition 2.1. *Let S be a continuous semimartingale. Then the following assertions are equivalent:*

(1) *There is a martingale measure $Q \in \mathcal{M}^e(F)$, and $J_T^2(F)$ is closed in L^2.*

(2) *There is a martingale measure $Q \in \mathcal{M}^e(F)$ that satisfies the reverse Hölder condition $R_2(P)$.*

(3) *There is a constant C such that for all $\pi \in \Pi(F)$ we have*

$$\|\sup_{t \leq T}(\pi \cdot S)_t\|_{L^2(P)} \leq C\|(\pi \cdot S)_T\|_{L^2(P)}.$$

(4) *There is a constant c such that for every stopping time τ, every $A \in \mathcal{F}_\tau$, and every $\pi \in \Pi(F)$, with $\pi = \pi I_{]\tau,T]}$, we have*

$$\|I_A - (\pi \cdot S)_T\|_{L^2(P)} \geq cP(A)^{1/2}.$$

Note that assertion (4) implies that for every stopping time τ and for every $\pi \in \Pi(G)$ we have

$$E\left(\left(1 + \int_\tau^T \pi_u dS_u\right)^2 \Big/ F_\tau\right) \geq c. \tag{2.2}$$

Now we recall some known assertions from the filtering theory. The following proposition can be proved similarly to (Liptser & Shiryaev, 1986)(the detailed proof one can see in (Mania et al., 2009)).

Proposition 2.2. *If conditions* (A), (B), *and* (C) *are satisfied, then for any continuous F-local martingale M, with $M_0 = 0$, and any G-local martingale m^G*

$$\widehat{M}_t = E(M_t|G_t) = \int_0^t \frac{d\widehat{\langle M, m^G \rangle}_u}{d\langle m^G \rangle_u} dm_u^G + L_t^G, \tag{2.3}$$

where L^G is a local martingale orthogonal to m^G.

It follows from this proposition that for any G-predictable, M-integrable process π and any G-martingale m^G

$$\langle \widehat{(\pi \cdot M)}, m^G \rangle_t = \int_0^t \pi_u \frac{d\widehat{\langle M, m^G \rangle}_u}{d\langle m^G \rangle_u} d\langle m^G \rangle_u = \int_0^t \pi_u d\langle \widehat{M}, m^G \rangle_u = \langle \pi \cdot \widehat{M}, m^G \rangle_t.$$

Hence, for any G-predictable, M-integrable process π

$$\widehat{(\pi \cdot M)}_t = E\left(\int_0^t \pi_s dM_s|G_t\right) = \int_0^t \pi_s d\widehat{M}_s. \tag{2.4}$$

Since π, λ, and $\langle M \rangle$ are G-predictable, from (2.4) we have

$$\widehat{(\pi \cdot S)}_t = E\left(\int_0^t \pi_u dS_u|G_t\right) = \int_0^t \pi_u d\widehat{S}_u, \tag{2.5}$$

where

$$\widehat{S}_t = S_0 + \int_0^t \widehat{\lambda}_u d\langle M \rangle_u + \widehat{M}_t.$$

3. Separation Principle: The Optimality Principle

Let us introduce the value function of the problem (1.2) defined as

$$U^H(t, x) = \operatorname*{ess\,inf}_{\pi \in \Pi(G)} E\left(\left(x + \int_t^T \pi_u dS_u - H \right)^2 |G_t \right). \tag{3.1}$$

By the GKW decomposition

$$H_t = E(H|F_t) = EH + \int_0^t h_u dM_u + L_t \tag{3.2}$$

for a F-predictable, M-integrable process h and a local martingale L strongly orthogonal to M. We shall use also the GKW decompositions of $H_t = E(H|F_t)$ with respect to the local martingale $\widehat{M}$

$$H_t = EH + \int_0^t h_u^G d\widehat{M}_u + L_t^G, \tag{3.3}$$

where h^G is a F-predictable process and L^G is a F-local martingale strongly orthogonal to $\widehat{M}$. It follows from Proposition 2.2 (applied for $m^G = \widehat{M}$) and Lemma A.1 that

$$\langle E(H|G_.), \widehat{M} \rangle_t = \int_0^t \widehat{h_u^G} \rho_u^2 d\langle M \rangle_u. \tag{3.4}$$

We shall use the notation

$$\widetilde{h}_t = \widehat{h_t^G} \rho_t^2 - \widehat{h}_t. \tag{3.5}$$

Note that $\widetilde{h}$ belongs to the class $\Pi(G)$ by Lemma A.2.

Let us introduce now a new optimization problem, equivalent to the initial mean-variance hedging problem (1.2), to minimize the expression

$$E\left[\left(x + \int_0^T \pi_u d\widehat{S}_u - \widehat{H}_T \right)^2 + \int_0^T \left(\pi_u^2 \left(1 - \rho_u^2 \right) + 2\pi_u \widetilde{h}_u \right) d\langle M \rangle_u \right] \tag{3.6}$$

over all $\pi \in \Pi(G)$. Recall that $\widehat{S}_t = E(S_t|G_t) = S_0 + \int_0^t \widehat{\lambda}_u d\langle M \rangle_u + \widehat{M}_t$.

Theorem 3.1. *Let conditions* (A), (B), *and* (C) *be satisfied. Then the initial mean-variance hedging problem* (1.2) *is equivalent to the problem* (3.6). *In particular, for any* $\pi \in \Pi(G)$ *and* $t \in [0, T]$

$$E\left[\left(x + \int_t^T \pi_u dS_u - H \right)^2 |G_t \right] = E\left[\left(H - \widehat{H}_T \right)^2 |G_t \right]$$

$$+ E\left[\left(x + \int_t^T \pi_u d\widehat{S}_u - \widehat{H}_T \right)^2 + \int_t^T \left(\pi_u^2 \left(1 - \rho_u^2 \right) + 2\pi_u \widetilde{h}_u \right) d\langle M \rangle_u |G_t \right]. \tag{3.7}$$

Proof. We have

$$E\left[\left(x + \int_t^T \pi_u dS_u - H \right)^2 |G_t \right] = E\left[\left(x + \int_t^T \pi_u d\widehat{S}_u - H + \int_t^T \pi_u d\left(M_u - \widehat{M}_u \right) \right)^2 |G_t \right]$$

$$= E\left[\left(x + \int_t^T \pi_u d\widehat{S}_u - H\right)^2 |G_t\right] + 2E\left[\left(x + \int_t^T \pi_u d\widehat{S}_u - H\right)\left(\int_t^T \pi_u d\left(M_u - \widehat{M}_u\right)\right)|G_t\right]$$

$$+ E\left[\left(\int_t^T \pi_u d\left(M_u - \widehat{M}_u\right)\right)^2 |G_t\right] = I_1 + 2I_2 + I_3. \tag{3.8}$$

It is evident that

$$I_1 = E\left[\left(x + \int_t^T \pi_u d\widehat{S}_u - \widehat{H}_T\right)^2 |G_t\right] + E\left[\left(H - \widehat{H}_T\right)^2 |G_t\right]. \tag{3.9}$$

Since $\pi, \widehat{\lambda}$, and $\langle\widehat{M}\rangle$ are G_T-measurable and the σ-algebras $F_t^S \vee G_t$ and G_T are conditionally independent given G_t (see Remark 2.3), it follows from (2.4) that

$$E\left[\int_t^T \pi_u \widehat{\lambda}_u d\langle M\rangle_u \int_t^T \pi_u d\left(M_u - \widehat{M}_u\right)|G_t\right]$$

$$= E\left[\int_t^T \pi_u \widehat{\lambda}_u d\langle M\rangle_u \int_0^T \pi_u d\left(M_u - \widehat{M}_u\right)|G_t\right]$$

$$- E\left[\int_t^T \pi_u \widehat{\lambda}_u d\langle M\rangle_u \int_0^t \pi_u d\left(M_u - \widehat{M}_u\right)|G_t\right]$$

$$= E\left[\int_t^T \pi_u \widehat{\lambda}_u d\langle M\rangle_u E\left(\int_0^T \pi_u d\left(M_u - \widehat{M}_u\right)|G_T\right)|G_t\right]$$

$$- E\left[\int_t^T \pi_u \widehat{\lambda}_u d\langle M\rangle_u |G_t\right] E\left[\int_0^t \pi_u d\left(M_u - \widehat{M}_u\right)|G_t\right] = 0. \tag{3.10}$$

On the other hand, by using decomposition (3.2), equality (3.4), properties of square characteristics of martingales, and the projection theorem, we obtain

$$E\left[H\int_t^T \pi_u d\left(M_u - \widehat{M}_u\right)|G_t\right] = E\left[H\int_t^T \pi_u dM_u|G_t\right] - E\left[\widehat{H}_T\int_t^T \pi_u d\widehat{M}_u|G_t\right]$$

$$= E\left[\int_t^T \pi_u d\langle M, E(H|F_.)\rangle_u|G_t\right] - E\left[\int_t^T \pi_u d\langle\widehat{H}, \widehat{M}\rangle_u|G_t\right]$$

$$= E\left[\int_t^T \pi_u h_u d\langle M\rangle_u|G_t\right] - E\left[\int_t^T \pi_u \widehat{h}_u^G \rho_u^2 d\langle M\rangle_u|G_t\right]$$

$$= E\left[\int_t^T \pi_u \left(\widehat{h}_u - \widehat{h}_u^G \rho_u^2\right) d\langle M\rangle_u|G_t\right] = -E\left[\int_t^T \pi_u \widetilde{h}_u d\langle M\rangle_u|G_t\right]. \tag{3.11}$$

Finally, it is easy to verify that

$$2E\left[\int_t^T \pi_u \widehat{M}_u \int_t^T \pi_u d\left(M_u - \widehat{M}_u\right)|G_t\right] + E\left[\left(\int_t^T \pi_u d\left(M_u - \widehat{M}_u\right)\right)^2 |G_t\right]$$

$$= E\left[\left(\int_t^T \pi_u^2 d\langle M\rangle_u - \int_t^T \pi_u^2 d\langle\widehat{M}\rangle_u\right)|G_t\right] = E\left[\int_t^T \pi_u^2 \left(1 - \rho_u^2\right) d\langle M\rangle_u|G_t\right]. \tag{3.12}$$

Therefore (3.8), (3.9), (3.10), (3.11), and (3.12) imply the validity of equality (3.7). $\square$

Thus, it follows from Theorem 3.1 that the optimization problems (1.2) and (3.6) are equivalent. Therefore it is sufficient to solve the problem (3.6), which is formulated in terms of G-adapted processes. One can say that (3.6) is a mean-variance hedging problem under complete information with a correction term and can be solved by using methods for complete information.

Let us introduce the value process of the problem (3.6)

$$V^H(t,x) = \operatorname*{ess\,inf}_{\pi \in \Pi(G)} E\left[\left(x + \int_t^T \pi_u d\widehat{S}_u - \widehat{H}_T \right)^2 + \int_t^T \left[\pi_u^2 \left(1 - \rho_u^2\right) + 2\pi_u \tilde{h}_u \right] d\langle M \rangle_u | G_t \right]. \quad (3.13)$$

It follows from Theorem 3.1 that

$$U^H(t,x) = V^H(t,x) + E\left[(H - \widehat{H}_T)^2 | G_t \right]. \quad (3.14)$$

The optimality principle takes in this case the following form.

Proposition 3.1 (optimality principle). *Let conditions* (A), (B) *and* (C) *be satisfied. Then*

(a) *for all $x \in R$, $\pi \in \Pi(G)$, and $s \in [0, T]$ the process*

$$V^H\left(t, x + \int_s^t \pi_u d\widehat{S}_u \right) + \int_s^t \left[\pi_u^2 \left(1 - \rho_u^2\right) + 2\pi_u \tilde{h}_u \right] d\langle M \rangle_u$$

is a submartingale on $[s, T]$, admitting an right continuous with left limits (RCLL) modification.

(b) *π^* is optimal if and only if the process*

$$V^H\left(t, x + \int_s^t \pi_u^* d\widehat{S}_u \right) + \int_s^t \left[(\pi_u^*)^2 \left(1 - \rho_u^2\right) + 2\pi_u^* \tilde{h}_u \right] d\langle M \rangle_u$$

is a martingale.

This assertion can be proved in a standard manner (see, e.g., (El Karoui & Quenez, 1995), (Kramkov, 1996)). The proof more adapted to this case one can see in (Mania & Tevzadze, 2003).

Let

$$V(t,x) = \operatorname*{ess\,inf}_{\pi \in \Pi(G)} E\left[\left(x + \int_t^T \pi_u d\widehat{S}_u \right)^2 + \int_t^T \pi_u^2 \left(1 - \rho_u^2\right) d\langle M \rangle_u | G_t \right]$$

and

$$V_t(2) = \operatorname*{ess\,inf}_{\pi \in \Pi(G)} E\left[\left(1 + \int_t^T \pi_u d\widehat{S}_u \right)^2 + \int_t^T \pi_u^2 \left(1 - \rho_u^2\right) d\langle M \rangle_u | G_t \right].$$

It is evident that $V(t,x)$ (resp., $V_t(2)$) is the value process of the optimization problem (3.6) in the case $H = 0$ (resp., $H = 0$ and $x = 1$), i.e.,

$$V(t,x) = V^0(t,x) \quad \text{and} \quad V_t(2) = V^0(t,1).$$

Since $\Pi(G)$ is a cone, we have

$$V(t,x) = x^2 \operatorname*{ess\,inf}_{\pi \in \Pi(G)} E\left[\left(1 + \int_t^T \frac{\pi_u}{x} d\widehat{S}_u \right)^2 + \int_t^T \left(\frac{\pi_u}{x} \right)^2 \left(1 - \rho_u^2\right) d\langle M \rangle_u | G_t \right] = x^2 V_t(2).$$

$$(3.15)$$

Therefore from Proposition 3.1 and equality (3.15) we have the following.

Corollary 3.1. (a) *The process*

$$V_t(2) \left(1 + \int_s^t \pi_u d\widehat{S}_u \right)^2 + \int_s^t (\pi_u)^2 (1 - \rho_u^2) d\langle M \rangle_u,$$

$t \geq s$, *is a submartingale for all* $\pi \in \Pi(G)$ *and* $s \in [0, T]$.
(b) π^* *is optimal if and only if*

$$V_t(2) \left(1 + \int_s^t \pi_u^* d\widehat{S}_u \right)^2 + \int_s^t (\pi_u^*)^2 (1 - \rho_u^2) d\langle M \rangle_u,$$

$t \geq s$, *is a martingale.*

Note that in the case $H = 0$ from Theorem 3.1 we have

$$E \left[\left(1 + \int_t^T \pi_u dS_u \right)^2 | G_t \right] = E \left[\left(1 + \int_t^T \pi_u d\widehat{S}_u \right)^2 + \int_t^T \pi_u^2 \left(1 - \rho_u^2 \right) d\langle M \rangle_u | G_t \right] \quad (3.16)$$

and, hence,

$$V_t(2) = U^0(t, 1). \tag{3.17}$$

Lemma 3.1. *Let conditions* (A)–(D) *be satisfied. Then there is a constant* $1 \geq c > 0$ *such that* $V_t(2) \geq c$ *for all* $t \in [0, T]$ *a.s. and*

$$1 - \rho_t^2 + \rho_t^2 V_t(2) \geq c \quad \mu^{\langle M \rangle} a.e. \tag{3.18}$$

Proof. Let

$$V_t^F(2) = \operatorname*{ess\,inf}_{\pi \in \Pi(F)} E \left[\left(1 + \int_t^T \pi_u dS_u \right)^2 | F_t \right].$$

It follows from assertion (4) of Proposition 2.1 that there is a constant $c > 0$ such that $V_t^F(2) \geq c$ for all $t \in [0, T]$ a.s. Note that $c \leq 1$ since $V^F \leq 1$. Then by (3.17)

$$V_t(2) = U^0(t, 1) = \operatorname*{ess\,inf}_{\pi \in \Pi(G)} E \left[\left(1 + \int_t^T \pi_u dS_u \right)^2 | G_t \right]$$

$$= \operatorname*{ess\,inf}_{\pi \in \Pi(G)} E \left[E \left(\left(1 + \int_t^T \pi_u dS_u \right)^2 | F_t \right) | G_t \right] \geq E(V_t^F(2) | G_t) \geq c.$$

Therefore, since $\rho_t^2 \leq 1$ by Lemma A.1,

$$1 - \rho_t^2 + \rho_t^2 V_t(2) \geq 1 - \rho_t^2 + \rho_t^2 c \geq \inf_{r \in [0,1]} (1 - r + rc) = c.$$

4. BSDEs for the Value Process

Let us consider the semimartingale backward equation

$$Y_t = Y_0 + \int_0^t f(u, Y_u, \psi_u) d\langle m \rangle_u + \int_0^t \psi_u dm_u + L_t \tag{4.1}$$

with the boundary condition

$$Y_T = \eta, \tag{4.2}$$

where η is an integrable G_T-measurable random variable, $f : \Omega \times [0, T] \times R^2 \to R$ is $\mathcal{P} \times \mathcal{B}(R^2)$ measurable, and m is a local martingale. A solution of (4.1)–(4.2) is a triple (Y, ψ, L), where Y is a special semimartingale, ψ is a predictable m-integrable process, and L a local martingale strongly orthogonal to m. Sometimes we call Y alone the solution of (4.1)–(4.2), keeping in mind that $\psi \cdot m + L$ is the martingale part of Y.

Backward stochastic differential equations have been introduced in (Bismut, 1973) for the linear case as the equations for the adjoint process in the stochastic maximum principle. The semimartingale backward equation, as a stochastic version of the Bellman equation in an optimal control problem, was first derived in (Chitashvili, 1983). The BSDE with more general nonlinear generators was introduced in (Pardoux & Peng, 1990) for the case of Brownian filtration, where the existence and uniqueness of a solution of BSDEs with generators satisfying the global Lifschitz condition was established. These results were generalized for generators with quadratic growth in (Kobylanski, 2000), (Lepeltier & San Martin, 1998) for BSDEs driven by a Brownian motion and in (Morlais, 2009), (Tevzadze, 2008) for BSDEs driven by martingales. But conditions imposed in these papers are too restrictive for our needs. We prove here the existence and uniqueness of a solution by directly showing that the unique solution of the BSDE that we consider is the value of the problem.

In this section we characterize optimal strategies in terms of solutions of suitable semimartingale backward equations.

Theorem 4.1. *Let H be a square integrable F_T-measurable random variable, and let conditions* (A), (B), (C), *and* (D) *be satisfied. Then the value function of the problem* (3.6) *admits a representation*

$$V^H(t, x) = V_t(0) - 2V_t(1)x + V_t(2)x^2, \tag{4.3}$$

where the processes $V_t(0), V_t(1),$ and $V_t(2)$ satisfy the following system of backward equations:

$$Y_t(2) = Y_0(2) + \int_0^t \frac{\left(\psi_s(2)\rho_s^2 + \widehat{\lambda}_s Y_s(2)\right)^2}{1 - \rho_s^2 + \rho_s^2 Y_s(2)} d\langle M \rangle_s + \int_0^t \psi_s(2) d\widehat{M}_s + L_t(2), \ \ Y_T(2) = 1, \tag{4.4}$$

$$Y_t(1) = Y_0(1) + \int_0^t \frac{\left(\psi_s(2)\rho_s^2 + \widehat{\lambda}_s Y_s(2)\right)\left(\psi_s(1)\rho_s^2 + \widehat{\lambda}_s Y_s(1) - \tilde{h}_s\right)}{1 - \rho_s^2 + \rho_s^2 Y_s(2)} d\langle M \rangle_s$$

$$+ \int_0^t \psi_s(1) d\widehat{M}_s + L_t(1), \quad Y_T(1) = E(H|G_T), \tag{4.5}$$

$$Y_t(0) = Y_0(0) + \int_0^t \frac{\left(\psi_s(1)\rho_s^2 + \widehat{\lambda}_s Y_s(1) - \tilde{h}_s\right)^2}{1 - \rho_s^2 + \rho_s^2 Y_s(2)} d\langle M \rangle_s$$

$$+ \int_0^t \psi_s(0) d\widehat{M}_s + L_t(0), \ \ Y_T(0) = E^2(H|G_T), \tag{4.6}$$

where $L(2), L(1),$ and $L(0)$ are G-local martingales orthogonal to $\widehat{M}$.

Besides, the optimal filtered wealth process $\widehat{X}_t^{x,\pi^*} = x + \int_0^t \pi_u^* d\widehat{S}_u$ *is a solution of the linear equation*

$$\widehat{X}_t^* = x - \int_0^t \frac{\rho_u^2 \psi_u(2) + \widehat{\lambda}_u Y_u(2)}{1 - \rho_u^2 + \rho_u^2 Y_u(2)} \widehat{X}_u^* d\widehat{S}_u + \int_0^t \frac{\psi_u(1)\rho_u^2 + \widehat{\lambda}_u Y_u(1) - \tilde{h}_u}{1 - \rho_u^2 + \rho_u^2 Y_u(2)} d\widehat{S}_u. \quad (4.7)$$

Proof. Similarly to the case of complete information one can show that the optimal strategy exists and that $V^H(t,x)$ is a square trinomial of the form (4.3) (see, e.g., (Mania & Tevzadze, 2003)). More precisely the space of stochastic integrals

$$J_{t,T}^2(G) = \left\{ \int_t^T \pi_u dS_u : \pi \in \Pi(G) \right\}$$

is closed by Proposition 2.1, since $\langle M \rangle$ is G-predictable. Hence there exists optimal strategy $\pi^*(t,x) \in \Pi(G)$ and $U^H(t,x) = E[|H - x - \int_t^T \pi_u^*(t,x)dS_u|^2|G_t]$. Since $\int_t^T \pi_u^*(t,x)dS_u$ co-incides with the orthogonal projection of $H - x \in L^2$ on the closed subspace of stochastic integrals, then the optimal strategy is linear with respect to x, i.e., $\pi_u^*(t,x) = \pi_u^0(t) + x\pi_u^1(t)$. This implies that the value function $U^H(t,x)$ is a square trinomial. It follows from the equality (3.14) that $V^H(t,x)$ is also a square trinomial, and it admits the representation (4.3).

Let us show that $V_t(0)$, $V_t(1)$, and $V_t(2)$ satisfy the system (4.4)–(4.6). It is evident that

$$V_t(0) = V^H(t,0) = \operatorname*{ess\,inf}_{\pi \in \Pi(G)} E\left[\left(\int_t^T \pi_u d\widehat{S}_u - \widehat{H}_T \right)^2 + \int_t^T [\pi_u^2(1-\rho_u^2) + 2\pi_u \tilde{h}_u]d\langle M \rangle_u | G_t \right] \quad (4.8)$$

and

$$V_t(2) = V^0(t,1) = \operatorname*{ess\,inf}_{\pi \in \Pi(G)} E\left[\left(1 + \int_t^T \pi_u d\widehat{S}_u \right)^2 + \int_t^T \pi_u^2 \left(1 - \rho_u^2\right) d\langle M \rangle_u | G_t \right]. \quad (4.9)$$

Therefore, it follows from the optimality principle (taking $\pi = 0$) that $V_t(0)$ and $V_t(2)$ are RCLL G-submartingales and

$$V_t(2) \leq E(V_T(2)|G_t) \leq 1, \qquad V_t(0) \leq E(E^2(H|G_T)|G_t) \leq E(H^2|G_t).$$

Since

$$V_t(1) = \frac{1}{2}(V_t(0) + V_t(2) - V^H(t,1)), \quad (4.10)$$

the process $V_t(1)$ is also a special semimartingale, and since $V_t(0) - 2V_t(1)x + V_t(2)x^2 = V^H(t,x) \geq 0$ for all $x \in R$, we have $V_t^2(1) \leq V_t(0)V_t(2)$; hence

$$V_t^2(1) \leq E\left(H^2|G_t\right).$$

Expressions (4.8), (4.9), and (3.13) imply that $V_T(0) = E^2(H|G_T)$, $V_T(2) = 1$, and $V^H(T,x) = (x - E(H|G_T))^2$. Therefore from (4.10) we have $V_T(1) = E(H|G_T)$, and $V(0)$, $V(1)$, and $V(2)$ satisfy the boundary conditions.

Thus, the coefficients $V_t(i)$, $i = 0,1,2$, are special semimartingales, and they admit the decomposition

$$V_t(i) = V_0(i) + A_t(i) + \int_0^t \varphi_s(i)d\widehat{M}_s + m_t(i), \quad i = 0,1,2, \quad (4.11)$$

where $m(0), m(1)$, and $m(2)$ are G-local martingales strongly orthogonal to $\widehat{M}$ and $A(0), A(1)$, and $A(2)$ are G-predictable processes of finite variation.

There exists an increasing continuous G-predictable process K such that

$$\langle M \rangle_t = \int_0^t \nu_u dK_u, \quad A_t(i) = \int_0^t a_u(i) dK_u, \quad i = 0, 1, 2,$$

where ν and $a(i), i = 0, 1, 2,$ are G-predictable processes.

Let $\widehat{X}_{s,t}^{x,\pi} \equiv x + \int_s^t \pi_u d\widehat{S}_u$ and

$$Y_{s,t}^{x,\pi} \equiv V^H\left(t, \widehat{X}_{s,t}^{x,\pi}\right) + \int_s^t \left[\pi_u^2\left(1 - \rho_u^2\right) + 2\pi_u \tilde{h}_u\right] d\langle M \rangle_u.$$

Then by using (4.3), (4.11), and the Itô formula for any $t \geq s$ we have

$$\left(\widehat{X}_{s,t}^{x,\pi}\right)^2 = x + \int_s^t \left[2\pi_u \widehat{\lambda}_u \widehat{X}_{s,u}^{x,\pi} + \pi_u^2 \rho_u^2\right] d\langle M \rangle_u + 2\int_s^t \pi_u \widehat{X}_{s,u}^{x,\pi} d\widehat{M}_u \tag{4.12}$$

and

$$\begin{aligned}
Y_{s,t}^{x,\pi} - V^H(s,x) &= \int_s^t \left[\left(\widehat{X}_{s,u}^{x,\pi}\right)^2 a_u(2) - 2\widehat{X}_{s,u}^{x,\pi} a_u(1) + a_u(0)\right] dK_u \\
&\quad + \int_s^t \left[\pi_u^2\left(1 - \rho_u^2 + \rho_u^2 V_{u-}(2)\right) + 2\pi_u \widehat{X}_{s,u}^{x,\pi}\left(\widehat{\lambda}_u V_{u-}(2) + \varphi_u(2)\rho_u^2\right)\right. \\
&\quad \left. - 2\pi_u\left(V_{u-}(1)\widehat{\lambda}_u + \varphi_u(1)\rho_u^2 - \tilde{h}_u\right)\right] \nu_u dK_u + m_t - m_s, \tag{4.13}
\end{aligned}$$

where m is a local martingale.

Let

$$\begin{aligned}
G(\pi, x) &= G(\omega, u, \pi, x) = \pi^2\left(1 - \rho_u^2 + \rho_u^2 V_{u-}(2)\right) + 2\pi x\left(\widehat{\lambda}_u V_{u-}(2) + \varphi_u(2)\rho_u^2\right) \\
&\quad - 2\pi(V_{u-}(1)\widehat{\lambda}_u + \varphi_u(1)\rho_u^2 - \tilde{h}_u).
\end{aligned}$$

It follows from the optimality principle that for each $\pi \in \Pi(G)$ the process

$$\int_s^t \left[\left(\widehat{X}_{s,u}^{x,\pi}\right)^2 a_u(2) - 2\widehat{X}_{s,u}^{x,\pi} a_u(1) + a_u(0)\right] dK_u + \int_s^t G\left(\pi_u, \widehat{X}_{s,u}^{x,\pi}\right) \nu_u dK_u \tag{4.14}$$

is increasing for any s on $s \leq t \leq T$, and for the optimal strategy π^* we have the equality

$$\int_s^t \left[\left(\widehat{X}_{s,u}^{x,\pi^*}\right)^2 a_u(2) - 2\widehat{X}_{s,u}^{x,\pi^*} a_u(1) + a_u(0)\right] dK_u = -\int_s^t G\left(\pi_u^*, \widehat{X}_{s,u}^{x,\pi^*}\right) \nu_u dK_u. \tag{4.15}$$

Since $\nu_u dK_u = d\langle M \rangle_u$ is continuous, without loss of generality one can assume that the process K is continuous (see (Mania & Tevzadze, 2003) for details). Therefore, by taking in (4.14) $\tau_s(\varepsilon) = \inf\{t \geq s : K_t - K_s \geq \varepsilon\}$ instead of t, we have that for any $\varepsilon > 0$ and $s \geq 0$

$$\frac{1}{\varepsilon}\int_s^{\tau_s(\varepsilon)} \left[\left(\widehat{X}_{s,u}^{x,\pi}\right)^2 a_u(2) - 2\widehat{X}_{s,u}^{x,\pi} a_u(1) + a_u(0)\right] dK_u \geq -\frac{1}{\varepsilon}\int_s^{\tau_s(\varepsilon)} G\left(\pi_u, \widehat{X}_{s,u}^{x,\pi}\right) \nu_u dK_u. \tag{4.16}$$

By passing to the limit in (4.16) as $\varepsilon \to 0$, from Proposition B of (Mania & Tevzadze, 2003) we obtain

$$x^2 a_u(2) - 2x a_u(1) + a_u(0) \geq -G(\pi_u, x)\nu_u, \qquad \mu^K\text{-a.e.,}$$

for all $\pi \in \Pi(G)$. Similarly from (4.15) we have that μ^K-a.e.

$$x^2 a_u(2) - 2x a_u(1) + a_u(0) = -G(\pi_u^*, x) \nu_u$$

and hence

$$x^2 a_u(2) - 2x a_u(1) + a_u(0) = -\nu_u \operatorname*{ess\,inf}_{\pi \in \Pi(G)} G(\pi_u, x). \tag{4.17}$$

The infimum in (4.17) is attained for the strategy

$$\hat{\pi}_t = \frac{V_t(1)\widehat{\lambda}_t + \varphi_t(1)\rho_t^2 - \tilde{h}_t - x(V_t(2)\widehat{\lambda}_t + \varphi_t(2)\rho_t^2)}{1 - \rho_t^2 + \rho_t^2 V_t(2)}. \tag{4.18}$$

From here we can conclude that

$$\operatorname*{ess\,inf}_{\pi \in \Pi(G)} G(\pi_t, x) \geq G(\hat{\pi}_t, x) = -\frac{\left(V_t(1)\widehat{\lambda}_t + \varphi_t(1)\rho_t^2 - \tilde{h}_t - x\left(V_t(2)\widehat{\lambda}_t + \varphi_t(2)\rho_t^2\right)\right)^2}{1 - \rho_t^2 + \rho_t^2 V_t(2)}. \tag{4.19}$$

Let $\pi_t^n = I_{[0,\tau_n[}(t)\hat{\pi}_t$, where $\tau_n = \inf\{t : |V_t(1)| \geq n\}$.
It follows from Lemmas A.2, 3.1, and A.3 that $\pi^n \in \Pi(G)$ for every $n \geq 1$ and hence

$$\operatorname*{ess\,inf}_{\pi \in \Pi(G)} G(\pi_t, x) \leq G(\pi_t^n, x)$$

for all $n \geq 1$. Therefore

$$\operatorname*{ess\,inf}_{\pi \in \Pi(G)} G(\pi_t, x) \leq \lim_{n \to \infty} G(\pi_t^n, x) = G(\hat{\pi}_t, x). \tag{4.20}$$

Thus (4.17), (4.19), and (4.20) imply that

$$\begin{aligned}
x^2 a_t(2) &- 2x a_t(1) + a_t(0) \\
&= \nu_t \frac{(V_t(1)\widehat{\lambda}_t + \varphi_t(1)\rho_t^2 - \tilde{h}_t - x(V_t(2)\widehat{\lambda}_t + \varphi_t(2)\rho_t^2))^2}{1 - \rho_t^2 + \rho_t^2 V_t(2)}, \quad \mu^K\text{-a.e.,}
\end{aligned} \tag{4.21}$$

and by equalizing the coefficients of square trinomials in (4.21) (and integrating with respect to dK) we obtain

$$A_t(2) = \int_0^t \frac{\left(\varphi_s(2)\rho_s^2 + \widehat{\lambda}_s V_s(2)\right)^2}{1 - \rho_s^2 + \rho_s^2 V_s(2)} d\langle M \rangle_s, \tag{4.22}$$

$$A_t(1) = \int_0^t \frac{\left(\varphi_s(2)\rho_s^2 + \widehat{\lambda}_s V_s(2)\right)\left(\varphi_s(1)\rho_s^2 + \widehat{\lambda}_s V_s(1) - \tilde{h}_s\right)}{1 - \rho_s^2 + \rho_s^2 V_s(2)} d\langle M \rangle_s, \tag{4.23}$$

$$A_t(0) = \int_0^t \frac{\left(\varphi_s(1)\rho_s^2 + \widehat{\lambda}_s V_s(1) - \tilde{h}_s\right)^2}{1 - \rho_s^2 + \rho_s^2 V_s(2)} d\langle M \rangle_s, \tag{4.24}$$

which, together with (4.11), implies that the triples $(V(i), \varphi(i), m(i))$, $i = 0, 1, 2$, satisfy the system (4.4)–(4.6).

Note that $A(0)$ and $A(2)$ are integrable increasing processes and relations (4.22) and (4.24) imply that the strategy $\hat{\pi}$ defined by (4.18) belongs to the class $\Pi(G)$.

Let us show now that if the strategy $\pi^* \in \Pi(G)$ is optimal, then the corresponding filtered wealth process $\widehat{X}_t^{\pi^*} = x + \int_0^t \pi_u^* d\widehat{S}_u$ is a solution of (4.7).

By the optimality principle the process

$$Y_t^{\pi^*} = V^H\left(t, \widehat{X}_t^{\pi^*}\right) + \int_0^t \left[(\pi_u^*)^2\left(1 - \rho_u^2\right) + 2\pi_u^* \tilde{h}_u\right] d\langle M\rangle_u$$

is a martingale. By using the Itô formula we have

$$Y_t^{\pi^*} = \int_0^t \left(\widehat{X}_u^{\pi^*}\right)^2 dA_u(2) - 2\int_0^t \widehat{X}_u^{\pi^*} dA_u(1) + A_t(0) + \int_0^t G\left(\pi_u^*, \widehat{X}_u^{\pi^*}\right) d\langle M\rangle_u + N_t,$$

where N is a martingale. Therefore by applying equalities (4.22), (4.23), and (4.24) we obtain

$$Y_t^{\pi^*} = \int_0^t \left(\pi_u^* - \frac{V_u(1)\widehat{\lambda}_u + \varphi_u(1)\rho_u^2 - \tilde{h}_u}{1 - \rho_u^2 + \rho_u^2 V_u(2)}\right.$$
$$\left. + \widehat{X}_u^{\pi^*} \frac{V_u(2)\widehat{\lambda}_u + \varphi_u(2)\rho_u^2}{1 - \rho_u^2 + \rho_u^2 V_u(2)}\right)^2 \left(1 - \rho_u^2 + \rho_u^2 V_u(2)\right) d\langle M\rangle_u + N_t,$$

which implies that $\mu^{\langle M\rangle}$-a.e.

$$\pi_u^* = \frac{V_u(1)\widehat{\lambda}_u + \varphi_u(1)\rho_u^2 - \tilde{h}_u}{1 - \rho_u^2 + \rho_u^2 V_u(2)} - \widehat{X}_u^{\pi^*} \frac{\left(V_u(2)\widehat{\lambda}_u + \varphi_u(2)\rho_u^2\right)}{1 - \rho_u^2 + \rho_u^2 V_u(2)}.$$

By integrating both parts of this equality with respect to $d\widehat{S}$ (and adding then x to the both parts), we obtain that $\widehat{X}^{\pi^*}$ satisfies (4.7). $\qquad\square$

The uniqueness of the system (4.4)–(4.6) we shall prove under following condition (D*), stronger than condition (D).

Assume that

$$(D^*) \qquad\qquad\qquad \int_0^T \frac{\widehat{\lambda}_u^2}{\rho_u^2} d\langle M\rangle_u \le C.$$

Since $\rho^2 \le 1$ (Lemma A.1), it follows from (D*) that the mean-variance tradeoff of S is bounded, i.e.,

$$\int_0^T \widehat{\lambda}_u^2 d\langle M\rangle_u \le C,$$

which implies (see, e.g., Kazamaki (Kazamaki, 1994)) that the minimal martingale measure for S exists and satisfies the reverse Hölder condition $R_2(P)$. So, condition (D*) implies condition (D). Besides, it follows from condition (D*) that the minimal martingale measure $\widehat{Q}^{min}$ for $\widehat{S}$

$$d\widehat{Q}^{min} = \mathcal{E}_T\left(-\frac{\widehat{\lambda}}{\rho^2} \cdot \widehat{M}\right)$$

also exists and satisfies the reverse Hölder condition. Indeed, condition (D*) implies that $\mathcal{E}_t(-2\frac{\widehat{\lambda}}{\rho^2} \cdot \widehat{M})$ is a G-martingale and hence

$$E\left(\mathcal{E}_{tT}^2\left(-\frac{\widehat{\lambda}}{\rho^2} \cdot \widehat{M}\right)|G_t\right) = E\left(\mathcal{E}_{tT}\left(-2\frac{\widehat{\lambda}}{\rho^2} \cdot \widehat{M}\right)e^{\int_t^T \frac{\widehat{\lambda}_u^2}{\rho_u^2}d\langle M\rangle_u}G_t\right) \le e^C.$$

Recall that the process Z belongs to the class D if the family of random variables $Z_\tau I_{(\tau \le T)}$ for all stopping times τ is uniformly integrable.

Theorem 4.2. *Let conditions* (A), (B), (C), *and* (D*) *be satisfied. If a triple* $(Y(0), Y(1), Y(2))$, *where* $Y(0) \in D$, $Y^2(1) \in D$, *and* $c \le Y(2) \le C$ *for some constants* $0 < c < C$, *is a solution of the system* (4.4)–(4.6), *then such a solution is unique and coincides with the triple* $(V(0), V(1), V(2))$.

Proof. Let $Y(2)$ be a bounded strictly positive solution of (4.4), and let

$$\int_0^t \psi_u(2)d\widehat{M}_u + L_t(2)$$

be the martingale part of $Y(2)$.

Since $Y(2)$ solves (4.4), it follows from the Itô formula that for any $\pi \in \Pi(G)$ the process

$$Y_t^\pi = Y_t(2)\left(1 + \int_s^t \pi_u d\widehat{S}_u\right)^2 + \int_s^t \pi_u^2\left(1 - \rho_u^2\right)d\langle M\rangle_u, \tag{4.25}$$

$t \ge s$, is a local submartingale.

Since $\pi \in \Pi(G)$, from Lemma A.1 and the Doob inequality we have

$$E\sup_{t \le T}\left(1 + \int_0^t \pi_u d\widehat{S}\right)^2$$
$$\le \text{const}\left(1 + E\int_0^T \pi_u^2\rho_u^2 d\langle M\rangle_u\right) + E\left(\int_0^T |\pi_u\widehat{\lambda}_u|d\langle M\rangle_u\right)^2 < \infty. \tag{4.26}$$

Therefore, by taking in mind that $Y(2)$ is bounded and $\pi \in \Pi(G)$ we obtain

$$E\left(\sup_{s \le u \le T} Y_u^\pi\right)^2 < \infty,$$

which implies that $Y^\pi \in D$. Thus Y^π is a submartingale (as a local submartingale from the class D), and by the boundary condition $Y_T(2) = 1$ we obtain

$$Y_s(2) \le E\left(\left(1 + \int_s^T \pi_u d\widehat{S}_u\right)^2 + \int_s^T \pi_u^2\left(1 - \rho_u^2\right)d\langle M\rangle_u|G_s\right)$$

for all $\pi \in \Pi(G)$ and hence

$$Y_t(2) \le V_t(2). \tag{4.27}$$

Let

$$\tilde{\pi}_t = -\frac{\widehat{\lambda}_t Y_t(2) + \psi_t(2)\rho_t^2}{1 - \rho_t^2 + \rho_t^2 Y_t(2)}\mathcal{E}_t\left(-\frac{\widehat{\lambda}Y(2) + \psi(2)\rho^2}{1 - \rho^2 + \rho^2 Y(2)} \cdot \widehat{S}\right).$$

Since $1 + \int_0^t \tilde{\pi}_u d\widehat{S}_u = \mathcal{E}_t(-\frac{\widehat{\lambda}Y(2)+\psi(2)\rho^2}{1-\rho^2+\rho^2 Y(2)} \cdot \widehat{S})$, it follows from (4.4) and the Itô formula that the process $Y^{\tilde{\pi}}$ defined by (4.25) is a positive local martingale and hence a supermartingale. Therefore

$$Y_s(2) \geq E\left(\left(1 + \int_s^T \tilde{\pi}_u d\widehat{S}_u\right)^2 + \int_s^T \tilde{\pi}_u^2\left(1 - \rho_u^2\right) d\langle M\rangle_u \Big| G_s\right). \tag{4.28}$$

Let us show that $\tilde{\pi}$ belongs to the class $\Pi(G)$.
From (4.28) and (4.27) we have for every $s \in [0, T]$

$$E\left(\left(1 + \int_s^T \tilde{\pi}_u d\widehat{S}_u\right)^2 + \int_s^T \tilde{\pi}_u^2\left(1 - \rho_u^2\right) d\langle M\rangle_u \Big| G_s\right) \leq Y_s(2) \leq V_s(2) \leq 1 \tag{4.29}$$

and hence

$$E\left(1 + \int_0^T \tilde{\pi}_u d\widehat{S}_u\right)^2 \leq 1, \tag{4.30}$$

$$E\int_0^T \tilde{\pi}_u^2\left(1 - \rho_u^2\right) d\langle M\rangle_u \leq 1. \tag{4.31}$$

By (D*) the minimal martingale measure $\widehat{Q}^{min}$ for $\widehat{S}$ satisfies the reverse Hölder condition, and hence all conditions of Proposition 2.1 are satisfied. Therefore the norm

$$E\left(\int_0^T \tilde{\pi}_s^2 \rho_s^2 d\langle M\rangle_s\right) + E\left(\int_0^T |\tilde{\pi}_s \widehat{\lambda}_s| d\langle M\rangle_s\right)^2$$

is estimated by $E\left(1 + \int_0^T \tilde{\pi}_u d\widehat{S}_u\right)^2$ and hence

$$E\int_0^T \tilde{\pi}_u^2 \rho_u^2 d\langle M\rangle_u < \infty, \quad E\left(\int_0^T |\tilde{\pi}_s \widehat{\lambda}_s| d\langle M\rangle_s\right)^2 < \infty.$$

It follows from (4.31) and the latter inequality that $\tilde{\pi} \in \Pi(G)$, and from (4.28) we obtain

$$Y_t(2) \geq V_t(2),$$

which together with (4.27) gives the equality $Y_t(2) = V_t(2)$.
Thus $V(2)$ is a unique bounded strictly positive solution of (4.4). Besides,

$$\int_0^t \psi_u(2) d\widehat{M}_u = \int_0^t \varphi_u(2) d\widehat{M}_u, \quad L_t(2) = m_t(2) \tag{4.32}$$

for all t, P-a.s.
Let $Y(1)$ be a solution of (4.5) such that $Y^2(1) \in D$. By the Itô formula the process

$$R_t = Y_t(1)\mathcal{E}_t\left(-\frac{\varphi(2)\rho^2 + \widehat{\lambda}V(2)}{1 - \rho^2 + \rho^2 V(2)} \cdot \widehat{S}\right)$$

$$+ \int_0^t \mathcal{E}_u\left(-\frac{\varphi(2)\rho^2 + \widehat{\lambda}V(2)}{1 - \rho^2 + \rho^2 V(2)} \cdot \widehat{S}\right) \frac{(\varphi_u(2)\rho_u^2 + \widehat{\lambda}_u V_u(2))\tilde{h}_u}{1 - \rho_u^2 + \rho_u^2 V_u(2)} d\langle M\rangle_u \tag{4.33}$$

is a local martingale. Let us show that R_t is a martingale.

As was already shown, the strategy

$$\tilde{\pi}_u = \frac{\psi_u(2)\rho_u^2 + \widehat{\lambda}_u Y_u(2)}{1 - \rho^2 + \rho^2 Y_u(2)} \mathcal{E}_u \left(-\frac{\psi(2)\rho^2 + \widehat{\lambda}Y(2)}{1 - \rho^2 + \rho^2 Y(2)} \cdot \widehat{S} \right)$$

belongs to the class $\Pi(G)$.
Therefore (see (4.26)),

$$E \sup_{t \leq T} \mathcal{E}_t^2 \left(-\frac{\psi(2)\rho^2 + \widehat{\lambda}Y(2)}{1 - \rho^2 + \rho^2 Y(2)} \cdot \widehat{S} \right) = E \sup_{t \leq T} \left(1 + \int_0^t \tilde{\pi}_u d\widehat{S} \right)^2 < \infty, \qquad (4.34)$$

and hence

$$Y_t(1)\mathcal{E}_t \left(-\frac{\varphi(2)\rho^2 + \widehat{\lambda}V(2)}{1 - \rho^2 + \rho^2 V(2)} \cdot \widehat{S} \right) \in D.$$

On the other hand, the second term of (4.33) is the process of integrable variation, since $\tilde{\pi} \in \Pi(G)$ and $\tilde{h} \in \Pi(G)$ (see Lemma A.2) imply that

$$E \int_0^T \left| \mathcal{E}_u \left(-\frac{\varphi(2)\rho^2 + \widehat{\lambda}V(2)}{1 - \rho^2 + \rho^2 V(2)} \cdot \widehat{S} \right) \frac{(\varphi_u(2)\rho_u^2 + \widehat{\lambda}_u V_u(2))\tilde{h}_u}{1 - \rho_u^2 + \rho_u^2 V_u(2)} \right| d\langle M \rangle_u$$

$$= E \int_0^T |\tilde{\pi}_u \tilde{h}_u| d\langle M \rangle_u \leq E^{1/2} \int_0^T \tilde{\pi}_u^2 d\langle M \rangle_u E^{1/2} \int_0^T \tilde{h}_u^2 d\langle M \rangle_u < \infty.$$

Therefore, the process R_t belongs to the class D, and hence it is a true martingale. By using the martingale property and the boundary condition we obtain

$$Y_t(1) = E \left(\widehat{H}_T \mathcal{E}_{tT} \left(-\frac{\varphi(2)\rho^2 + \widehat{\lambda}V(2)}{1 - \rho^2 + \rho^2 V(2)} \cdot \widehat{S} \right) \right.$$

$$\left. + \int_t^T \mathcal{E}_{tu} \left(-\frac{\varphi(2)\rho^2 + \widehat{\lambda}V(2)}{1 - \rho^2 + \rho^2 V(2)} \cdot \widehat{S} \right) \frac{(\varphi_u(2)\rho_u^2 + \widehat{\lambda}_u V_u(2))\tilde{h}_u}{1 - \rho_u^2 + \rho_u^2 V_u(2)} d\langle M \rangle_u | G_t \right). \qquad (4.35)$$

Thus, any solution of (4.5) is expressed explicitly in terms of $(V(2), \varphi(2))$ in the form (4.35). Hence the solution of (4.5) is unique, and it coincides with $V_t(1)$.
It is evident that the solution of (4.6) is also unique. $\qquad \square$

Remark 4.1. In the case $F^S \subseteq G$ we have $\rho_t = 1, \tilde{h}_t = 0$, and $\widehat{S}_t = S_t$, and (4.7) takes the form

$$\widehat{X}_t^* = x - \int_0^t \frac{\psi_u(2) + \widehat{\lambda}_u Y_u(2)}{Y_u(2)} \widehat{X}_u^* dS_u + \int_0^t \frac{\psi_u(1) + \widehat{\lambda}_u Y_u(1)}{Y_u(2)} dS_u.$$

Corollary 4.1. *In addition to conditions* (A)–(C) *assume that ρ is a constant and the mean-variance tradeoff $\langle \widehat{\lambda} \cdot M \rangle_T$ is deterministic. Then the solution of* (4.4) *is the triple* $(Y(2), \psi(2), L(2))$, *with* $\psi(2) = 0, L(2) = 0$, *and*

$$Y_t(2) = V_t(2) = \nu \left(\rho, 1 - \rho^2 + \langle \widehat{\lambda} \cdot M \rangle_T - \langle \widehat{\lambda} \cdot M \rangle_t \right), \qquad (4.36)$$

where $\nu(\rho, \alpha)$ is the root of the equation

$$\frac{1 - \rho^2}{x} - \rho^2 \ln x = \alpha. \tag{4.37}$$

Besides,

$$Y_t(1) = E\left(H\mathcal{E}_{tT} \left(-\frac{\widehat{\lambda} V(2)}{1 - \rho^2 + \rho^2 V(2)} \cdot \widehat{S} \right) \right.$$

$$\left. + \int_t^T \mathcal{E}_{tu} \left(-\frac{\widehat{\lambda} V(2)}{1 - \rho^2 + \rho^2 V(2)} \cdot \widehat{S} \right) \frac{\lambda_u V_u(2) \tilde{h}_u}{1 - \rho^2 + \rho^2 V_u(2)} d\langle M \rangle_u \Big| G_t \right) \tag{4.38}$$

uniquely solves (4.5), and the optimal filtered wealth process satisfies the linear equation

$$\widehat{X}_t^* = x - \int_0^t \frac{\widehat{\lambda}_u V_u(2)}{1 - \rho^2 + \rho^2 V_u(2)} \widehat{X}_u^* d\widehat{S}_u + \int_0^t \frac{\varphi_u(1)\rho^2 + \widehat{\lambda}_u V_u(1) - \tilde{h}_u}{1 - \rho^2 + \rho^2 V_u(2)} d\widehat{S}_u. \tag{4.39}$$

Proof. The function $f(x) = \frac{1 - \rho^2}{x} - \rho^2 \ln x$ is differentiable and strictly decreasing on $]0, \infty[$ and takes all values from $]-\infty, +\infty[$. So (4.37) admits a unique solution for all α. Besides, the inverse function $\alpha(x)$ is differentiable. Therefore $Y_t(2)$ is a process of finite variation, and it is adapted since $\langle \widehat{\lambda} \cdot M \rangle_T$ is deterministic.

By definition of $Y_t(2)$ we have that for all $t \in [0, T]$

$$\frac{1 - \rho^2}{Y_t(2)} - \rho^2 \ln Y_t(2) = 1 - \rho^2 + \langle \widehat{\lambda} \cdot M \rangle_T - \langle \widehat{\lambda} \cdot M \rangle_t.$$

It is evident that for $\alpha = 1 - \rho^2$ the solution of (4.37) is equal to 1, and it follows from (4.36) that $Y(2)$ satisfies the boundary condition $Y_T(2) = 1$. Therefore

$$\frac{1 - \rho^2}{Y_t(2)} - \rho^2 \ln Y_t(2) - \left(1 - \rho^2\right) = -\left(1 - \rho^2\right) \int_t^T d\frac{1}{Y_u(2)} + \rho^2 \int_t^T d\ln Y_u(2)$$

$$= \int_t^T \left(\frac{1 - \rho^2}{Y_u^2(2)} + \frac{\rho^2}{Y_u(2)} \right) dY_u(2)$$

and

$$\int_t^T \frac{1 - \rho^2 + \rho^2 Y_u(2)}{Y_u^2(2)} dY_u(2) = \langle \widehat{\lambda} \cdot M \rangle_T - \langle \widehat{\lambda} \cdot M \rangle_t$$

for all $t \in [0, T]$. Hence

$$\int_0^t \frac{1 - \rho^2 + \rho^2 Y_u(2)}{Y_u^2(2)} dY_u(2) = \langle \widehat{\lambda} \cdot M \rangle_t,$$

and, by integrating both parts of this equality with respect to $Y(2)/(1 - \rho^2 + \rho^2 Y(2))$, we obtain that $Y(2)$ satisfies

$$Y_t(2) = Y_0(2) + \int_0^t \frac{Y_u^2(2)\widehat{\lambda}_u^2}{1 - \rho^2 + \rho^2 Y_u(2)} d\langle M \rangle_u, \tag{4.40}$$

which implies that the triple $(Y(2), \psi(2) = 0, L(2) = 0)$ satisfies (4.4) and $Y(2) = V(2)$ by Theorem 4.2. Equations (4.38) and (4.39) follow from (4.35) and (4.7), respectively, by taking $\varphi(2) = 0$. $\qquad \square$

Remark 4.2. In case $F^S \subseteq G$ we have $\widehat{M} = M$ and $\rho = 1$. Therefore (4.40) is linear and $Y_t(2) = e^{\langle \hat{\lambda} \cdot M \rangle_t - \langle \hat{\lambda} \cdot M \rangle_T}$. In the case $\mathcal{A} = G$ of complete information, $Y_t(2) = e^{\langle \lambda \cdot N \rangle_t - \langle \lambda \cdot N \rangle_T}$.

5. Diffusion Market Model

Example 1. Let us consider the financial market model

$$d\tilde{S}_t = \tilde{S}_t \mu_t(\eta) dt + \tilde{S}_t \sigma_t(\eta) dw_t^0,$$
$$d\eta_t = a_t(\eta) dt + b_t(\eta) dw_t,$$

subjected to initial conditions. Here w^0 and w are correlated Brownian motions with $E dw_t^0 dw_t = \rho dt, \rho \in (-1, 1)$.

Let us write

$$w_t = \rho w_t^0 + \sqrt{1 - \rho^2} w_t^1,$$

where w^0 and w^1 are independent Brownian motions. It is evident that $w^\perp = -\sqrt{1 - \rho^2} w^0 + \rho w^1$ is a Brownian motion independent of w, and one can express Brownian motions w^0 and w^1 in terms of w and $w^\perp$ as

$$w_t^0 = \rho w_t - \sqrt{1 - \rho^2} w_t^\perp, \quad w_t^1 = \sqrt{1 - \rho^2} w_t + \rho w_t^\perp. \tag{5.1}$$

Suppose that $b^2 > 0$, $\sigma^2 > 0$, and coefficients μ, σ, a, and b are such that $F_t^{S, \eta} = F_t^{w^0, w}$ and $F_t^\eta = F_t^w$.

We assume that an agent would like to hedge a contingent claim H (which can be a function of S_T and η_T) using only observations based on the process η. So the stochastic basis will be $(\Omega, \mathcal{F}, F_t, P)$, where F_t is the natural filtration of (w^0, w) and the flow of observable events is $G_t = F_t^w$.

Also denote $dS_t = \mu_t dt + \sigma_t dw_t^0$, so that $d\tilde{S}_t = \tilde{S}_t dS_t$ and S is the return of the stock.

Let $\tilde{\pi}_t$ be the number of shares of the stock at time t. Then $\pi_t = \tilde{\pi}_t \tilde{S}_t$ represents an amount of money invested in the stock at the time $t \in [0, T]$. We consider the mean-variance hedging problem

$$\text{to minimize} \quad E\left[\left(x + \int_0^T \tilde{\pi}_t d\tilde{S}_t - H\right)^2\right] \quad \text{over all} \quad \tilde{\pi} \quad \text{for which} \quad \tilde{\pi}\tilde{S} \in \Pi(G), \tag{5.2}$$

which is equivalent to studying the mean-variance hedging problem

$$\text{to minimize} \quad E\left[\left(x + \int_0^T \pi_t dS_t - H\right)^2\right] \quad \text{over all} \quad \pi \in \Pi(G).$$

Remark 5.1. Since S is not G-adapted, $\tilde{\pi}_t$ and $\tilde{\pi}_t \tilde{S}_t$ cannot be simultaneously G-predictable and the problem

$$\text{to minimize} \quad E\left[\left(x + \int_0^T \tilde{\pi}_t d\tilde{S}_t - H\right)^2\right] \quad \text{over all} \quad \tilde{\pi} \in \Pi(G)$$

is not equivalent to the problem (5.2). In this setting, condition (A) is not satisfied, and it needs separate consideration.

By comparing with (1.1) we get that in this case

$$M_t = \int_0^t \sigma_s dw_s^0, \quad \langle M \rangle_t = \int_0^t \sigma_s^2 ds, \quad \lambda_t = \frac{\mu_t}{\sigma_t^2}.$$

It is evident that w is a Brownian motion also with respect to the filtration F^{w^0, w^1} and condition (B) is satisfied. Therefore by Proposition 2.2

$$\widehat{M}_t = \rho \int_0^t \sigma_s dw_s.$$

By the integral representation theorem the GKW decompositions (3.2) and (3.3) take the following forms:

$$c_H = EH, \quad H_t = c_H + \int_0^t h_s \sigma_s dw_s^0 + \int_0^t h_s^1 dw_s^1, \tag{5.3}$$

$$H_t = c_H + \rho \int_0^t h_s^G \sigma_s dw_s + \int_0^t h_s^\perp dw_s^\perp. \tag{5.4}$$

By putting expressions (5.1) for w^0 and w^1 in (5.3) and equalizing integrands of (5.3) and (5.4), we obtain

$$h_t = \rho^2 h_t^G - \sqrt{1-\rho^2} \frac{h_t^\perp}{\sigma_t}$$

and hence

$$\widehat{h}_t = \rho^2 \widehat{h_t^G} - \sqrt{1-\rho^2} \frac{\widehat{h_t^\perp}}{\sigma_t}.$$

Therefore by the definition of $\widetilde{h}$

$$\widetilde{h}_t = \rho^2 \widehat{h_t^G} - \widehat{h}_t = \sqrt{1-\rho^2} \frac{\widehat{h_t^\perp}}{\sigma_t}. \tag{5.5}$$

By using notations

$$Z_s(0) = \rho \sigma_s \varphi_s(0), \ Z_s(1) = \rho \sigma_s \varphi_s(1), \ Z_s(2) = \rho \sigma_s \varphi_s(2), \ \theta_s = \frac{\mu_s}{\sigma_s},$$

we obtain the following corollary of Theorem 4.1.

Corollary 5.1. *Let H be a square integrable F_T-measurable random variable. Then the processes $V_t(0)$, $V_t(1)$, and $V_t(2)$ from (4.3) satisfy the following system of backward equations:*

$$V_t(2) = V_0(2) + \int_0^t \frac{(\rho Z_s(2) + \theta_s V_s(2))^2}{1 - \rho^2 + \rho^2 V_s(2)} ds + \int_0^t Z_s(2) dw_s, \quad V_T(2) = 1, \tag{5.6}$$

$$V_t(1) = V_0(1) + \int_0^t \frac{(\rho Z_s(2) + \theta_s V_s(2))\left(\rho Z_s(1) + \theta_s V_s(1) - \sqrt{1-\rho^2}\,\widehat{h}_s^\perp\right)}{1 - \rho^2 + \rho^2 V_s(2)} ds$$

$$+ \int_0^t Z_s(1) dw_s, \quad V_T(1) = E(H|G_T), \tag{5.7}$$

$$V_t(0) = V_0(0) + \int_0^t \frac{\left(\rho Z_s(1) + \theta_s V_s(1) - \sqrt{1-\rho^2}\,\widehat{h}_s^\perp\right)^2}{1 - \rho^2 + \rho^2 V_s(2)} ds$$

$$+ \int_0^t Z_s(0) dw_s, \quad V_T(0) = E^2(H|G_T). \tag{5.8}$$

Besides, the optimal wealth process $\widehat{X}^*$ satisfies the linear equation

$$\widehat{X}_t^* = x - \int_0^t \frac{\rho Z_s(2) + \theta_s V_s(2)}{1 - \rho^2 + \rho^2 V_s(2)} \widehat{X}_s^* (\theta_s ds + \rho dw_s)$$

$$+ \int_0^t \frac{\rho Z_s(1) + \theta_s V_s(1) - \sqrt{1-\rho^2}\,\widehat{h}_s^\perp}{1 - \rho^2 + \rho^2 V_s(2)} (\theta_s ds + \rho dw_s). \tag{5.9}$$

Suppose now that θ_t and σ_t are deterministic. Then the solution of (5.6) is the pair $(V_t(2), Z_t(2))$, where $Z(2) = 0$ and $V(2)$ satisfies the ordinary differential equation

$$\frac{dV_t(2)}{dt} = \frac{\theta_t^2 V_t^2(2)}{1 - \rho^2 + \rho^2 V_t(2)}, \quad V_T(2) = 1. \tag{5.10}$$

By solving this equation we obtain

$$V_t(2) = \nu\left(\rho, 1 - \rho^2 + \int_t^T \theta_s^2 ds\right) \equiv v_t^{\theta,\rho}, \tag{5.11}$$

where $\nu(\rho,\alpha)$ is the solution of (4.37). From (5.10) it follows that

$$\left(\ln v_t^{\theta,\rho}\right)' = \frac{\theta_t^2 v_t^{\theta,\rho}}{1 - \rho^2 + \rho^2 v_t^{\theta,\rho}} \quad \text{and} \quad \ln \frac{v_s^{\theta,\rho}}{v_t^{\theta,\rho}} = \int_t^s \frac{\theta_r^2 v_r^{\theta,\rho} dr}{1 - \rho^2 + \rho^2 v_r^{\theta,\rho}}. \tag{5.12}$$

If we solve the linear BSDE (5.7) and use (5.12), we obtain

$$V_t(1) = E\left[\widehat{H}_T(w)\mathcal{E}_{tT}\left(-\int_0^{\cdot} \frac{\theta_r v_r^{\theta,\rho}}{1 - \rho^2 + \rho^2 v_r^{\theta,\rho}} (\theta_r dr + \rho dw_r)\right) \middle| G_t\right],$$

$$\int_t^T \frac{\theta_s v_s^{\theta,\rho} \sigma_s}{1 - \rho^2 + \rho^2 v_s^{\theta,\rho}} E\left[\tilde{h}_s(w)\mathcal{E}_{ts}\left(-\int_0^{\cdot} \frac{\theta_r v_r^{\theta,\rho}}{1 - \rho^2 + \rho^2 v_r^{\theta,\rho}} (\theta_r dr + \rho dw_r)\right) \middle| G_t\right] ds$$

$$= v_t^{\theta,\rho} E\left[\widehat{H}_T(w)\mathcal{E}_{tT}\left(-\int_0^{\cdot} \frac{\theta_r v_r^{\theta,\rho}}{1 - \rho^2 + \rho^2 v_r^{\theta,\rho}} \rho dw_r\right) \middle| G_t\right]$$

$$+ v_t^{\theta,\rho} \int_t^T \frac{\mu_s}{1 - \rho^2 + \rho^2 v_s^{\theta,\rho}} E\left[\tilde{h}_s(w)\mathcal{E}_{ts}\left(-\int_0^{\cdot} \frac{\theta_r v_r^{\theta,\rho}}{1 - \rho^2 + \rho^2 v_r^{\theta,\rho}} \rho dw_r\right) \middle| G_t\right] ds.$$

By using the Girsanov theorem we finally get

$$V_t(1) = v_t^{\theta,\rho} E\left[\widehat{H}_T\left(\rho \int_0^{\cdot} \frac{\theta_r v_r^{\theta,\rho}}{1 - \rho^2 + \rho^2 v_r^{\theta,\rho}} dr + w\right) \middle| G_t\right]$$

$$+ v_t^{\theta,\rho} \int_t^T \frac{\mu_s}{1 - \rho^2 + \rho^2 v_s^{\theta,\rho}} E\left[\tilde{h}_s\left(\rho \int_0^{\cdot} \frac{\theta_r v_r^{\theta,\rho}}{1 - \rho^2 + \rho^2 v_r^{\theta,\rho}} dr + w\right) \middle| G_t\right] ds. \tag{5.13}$$

Besides, the optimal strategy is of the form

$$\pi_t^* = -\frac{\theta_t V_t(2)}{(1 - \rho^2 + \rho^2 V_t(2))\sigma_t} \widehat{X}_t^* + \frac{\rho Z_t(1) + \theta_t V_t(1) - \sqrt{1-\rho^2}\,\widehat{h}_t^\perp}{(1 - \rho^2 + \rho^2 V_t(2))\sigma_t}.$$

If in addition μ and σ are constants and the contingent claim is of the form $H = \mathcal{H}(S_T, \eta_T)$, then one can give an explicit expressions also for $\tilde{h}$, $\widehat{h}^\perp$, $\widehat{H}$, and $Z(1)$.

Example 2. In Frey and Runggaldier (Frey & Runggaldier, 1999) the incomplete-information situation arises, assuming that the hedger is unable to monitor the asset continuously but is confined to observations at discrete random points in time $\tau_1, \tau_2, \ldots, \tau_n$. Perhaps it is more natural to assume that the hedger has access to price information on full intervals $[\sigma_1, \tau_1], [\sigma_2, \tau_2], \ldots, [\sigma_n, \tau_n]$. For the models with nonzero drifts, even the case $n = 1$ is non-trivial. Here we consider this case in detail.

Let us consider the financial market model

$$d\widetilde{S}_t = \mu\widetilde{S}_t dt + \sigma\widetilde{S}_t dW_t, \quad S_0 = S,$$

where W is a standard Brownian motion and the coefficients μ and σ are constants. Assume that an investor observes only the returns $S_t - S_0 = \int_0^t \frac{1}{\widetilde{S}_u} d\widetilde{S}_u$ of the stock prices up to a random moment τ before the expiration date T. Let $\mathcal{A}_t = F_t^S$, and let τ be a stopping time with respect to F^S. Then the filtration G_t of observable events is equal to the filtration $F_{t \wedge \tau}^S$. Consider the mean-variance hedging problem

$$\text{to minimize} \quad E\left[\left(x + \int_0^T \pi_t dS_t - H\right)^2\right] \quad \text{over all} \quad \pi \in \Pi(G),$$

where π_t is a dollar amount invested in the stock at time t.
By comparing with (1.1) we get that in this case

$$N_t = M_t = \sigma W_t, \quad \langle M \rangle_t = \sigma^2 t, \quad \lambda_t = \frac{\mu}{\sigma^2}.$$

Let $\theta = \frac{\mu}{\sigma}$. The measure Q defined by $dQ = \mathcal{E}_T(\theta W)dP$ is a unique martingale measure for S, and it is evident that Q satisfies the reverse Hölder condition. It is also evident that any G-martingale is F^S-martingale and that conditions (A)–(C) are satisfied. Besides,

$$E(W_t|G_t) = W_{t \wedge \tau}, \quad \widehat{S}_t = \mu t + \sigma W_{t \wedge \tau} \quad \text{and} \quad \rho_t = I_{\{t \leq \tau\}}. \tag{5.14}$$

By the integral representation theorem

$$E\left(H|F_t^S\right) = EH + \int_0^t h_u \sigma dW_u \tag{5.15}$$

for F-predictable W-integrable process h. On the other hand, by the GKW decomposition with respect to the martingale $W^\tau = (W_{t \wedge \tau}, t \in [0, T])$,

$$E\left(H|F_t^S\right) = EH + \int_0^t h_u^G \sigma dW_u^\tau + L_t^G \tag{5.16}$$

for F^S-predictable process h^G and F^S martingale L^G strongly orthogonal to W^τ. Therefore, by equalizing the right-hand sides of (5.15) and (5.16) and taking the mutual characteristics of both parts with W^τ, we obtain $\int_0^{t \wedge \tau}(h_u^G \rho_u^2 - h_u)du = 0$ and hence

$$\int_0^t \widetilde{h}_u du = \int_0^t \left(\widehat{h}_u^G I_{(u \leq \tau)} - \widehat{h}_u\right) du = -\int_0^t I_{(u > \tau)} E\left(h_u|F_\tau^S\right) du. \tag{5.17}$$

Therefore, by using notations

$$Z_s(0) = \rho\sigma\varphi_s(0), \ Z_s(1) = \rho\sigma\varphi_s(1), \ Z_s(2) = \rho\sigma\varphi_s(2),$$

it follows from Theorem 4.1 that the processes $(V_t(2), Z_t(2))$ and $(V_t(1), Z_t(1))$ satisfy the following system of backward equations:

$$V_t(2) = V_0(2) + \int_0^{t\wedge\tau} \frac{(Z_s(2) + \theta V_s(2))^2}{V_s(2)} ds$$

$$+ \int_{t\wedge\tau}^t \theta^2 V_s^2(2) ds + \int_0^{t\wedge\tau} Z_s(2) dW_s, \quad V_T(2) = 1, \quad (5.18)$$

$$V_t(1) = V_0(1) + \int_0^{t\wedge\tau} \frac{(Z_s(2) + \theta V_s(2))(Z_s(1) + \theta V_s(1))}{V_s(2)} ds$$

$$+ \int_{t\wedge\tau}^t \theta V_s(2) \left(\theta V_s(1) + E\left(h_s|F_\tau^S\right)\right) ds + \int_0^{t\wedge\tau} Z_s(1) dW_s, \quad V_T(1) = E(H|G_T). \quad (5.19)$$

Equation (5.18) admits in this case an explicit solution. To obtain the solution one should solve first the equation

$$U_t = U_0 + \int_0^t \theta^2 U_s^2 ds, \quad U_T = 1, \quad (5.20)$$

in the time interval $[\tau, T]$ and then the BSDE

$$V_t(2) = V_0(2) + \int_0^t \frac{(Z_s(2) + \theta V_s(2))^2}{V_s(2)} ds + \int_0^t Z_s(2) dW_s \quad (5.21)$$

in the interval $[0, \tau]$, with the boundary condition $V_\tau(2) = U_\tau$. The solution of (5.20) is

$$U_t = \frac{1}{1 + \theta^2(T - t)},$$

and the solution of (5.21) is expressed as

$$V_t(2) = \frac{1}{E\big((1 + \theta^2(T - \tau))\mathcal{E}_{t,\tau}^2(-\theta W)|F_t^S\big)}$$

(this can be verified by applying the Itô formula for the process $V_t^{-1}(2)\mathcal{E}_t^2(-\theta W)$ and by using the fact that this process is a martingale). Therefore

$$V_t(2) = \begin{cases} \dfrac{1}{1 + \theta^2(T - t)} & \text{if} \quad t \geq \tau, \\[3mm] \dfrac{1}{E\big((1 + \theta^2(T - \tau))\mathcal{E}_{t,\tau}^2(-\theta W)|F_t^S\big)} & \text{if} \quad t \leq \tau. \end{cases} \quad (5.22)$$

According to (4.37), taking in mind (5.14), (5.17), and the fact that $e^{-\int_t^T \theta^2 V_u(2)du} = \frac{1}{1+\theta^2(T-t)}$ on the set $t \geq \tau$, the solution of (5.19) is equal to

$$V_t(1) = E\left(\frac{H}{1 + \theta^2(T - t)} + \int_t^T \frac{\theta V_u(2)h_u du}{1 + \theta^2(T - u)} \Big| F_t^S\right) I_{(t>\tau)}$$

$$+ E\left(\mathcal{E}_{t,\tau}\left(-\frac{\varphi(2) + \lambda V(2)}{V(2)} \cdot S\right) \left(\frac{H}{1 + \theta^2(T - \tau)} + \int_\tau^T \frac{\theta V_u(2)h_u du}{1 + \theta^2(T - u)}\right) \Big| F_t^S\right) I_{(t\leq\tau)}. \quad (5.23)$$

By Theorem 4.1 the optimal filtered wealth process is a solution of a linear SDE, which takes in this case the following form:

$$\widehat{X}_t^* = x - \int_0^{t \wedge \tau} \frac{\varphi_u(2) + \theta V_u(2)}{V_u(2)} \widehat{X}_u^*(\theta du + dW_u) - \int_{t \wedge \tau}^t \theta^2 V_u(2) \widehat{X}_u^* du$$

$$+ \int_0^{t \wedge \tau} \frac{\varphi_u(1) + \theta V_u(1)}{V_u(2)}(\theta du + dW_u) + \int_{t \wedge \tau}^t \left(\theta^2 V_u(1) + \mu E \left(h_u | F_\tau^S \right) \right) du. \quad (5.24)$$

The optimal strategy is equal to

$$\pi_t^* = \left[-\frac{\varphi_t(2) + \theta V_t(2)}{V_t(2)} I_{(t \leq \tau)} - \theta^2 V_t(2) I_{(t > \tau)} \right] \widehat{X}_t^*$$

$$+ \frac{\varphi_t(1) + \theta V_t(1)}{V_t(2)} I_{(t \leq \tau)} + \left(\theta^2 V_t(1) + \mu E \left(h_t | F_\tau^S \right) \right) I_{(t > \tau)}, \quad (5.25)$$

where $\widehat{X}_t^*$ is a solution of the linear equation (5.24), $V(2)$ and $V(1)$ are given by (5.22) and (5.23), and $\varphi(2)$ and $\varphi(1)$ are integrands of their martingale parts, respectively. In particular the optimal strategy in time interval $[\tau, T]$ (i.e., after interrupting observations) is of the form

$$\pi_t^* = -\theta^2 V_t(2) \widehat{X}_t^* + \theta^2 V_t(1) + \mu E \left(h_t | F_\tau^S \right), \quad (5.26)$$

where

$$\widehat{X}_t^* = \frac{\widehat{X}_\tau^*}{1 + \theta^2 (t - \tau)} - \int_\tau^t \left(\theta^2 V_u(1) - \mu E \left(h_u | F_\tau^S \right) \right) \frac{1}{1 + \theta^2 (t - u)} du.$$

For instance, if τ is deterministic, then $V_t(2)$ is also deterministic:

$$V_t(2) = \begin{cases} \dfrac{1}{1 + \theta^2 (T - t)} & \text{if } t \geq \tau, \\[3mm] \dfrac{1}{1 + \theta^2 (T - t)} e^{-\theta^2 (\tau - t)} & \text{if } t \leq \tau, \end{cases}$$

and $\varphi(2) = 0$.

Note that it is not optimal to do nothing after interrupting observations, and in order to act optimally one should change the strategy deterministically as it is given by (5.26).

Appendix

For convenience we give the proofs of the following assertions used in the paper.

Lemma A.1. *Let conditions* (A)–(C) *be satisfied and* $\widehat{M}_t = E(M_t | G_t)$. *Then* $\langle \widehat{M} \rangle$ *is absolutely continuous w.r.t.* $\langle M \rangle$ *and* $\mu^{\langle M \rangle}$ *a.e.*

$$\rho_t^2 = \frac{d \langle \widehat{M} \rangle_t}{d \langle M \rangle_t} \leq 1.$$

Proof. By (2.4) for any bounded G-predictable process h

$$E \int_0^t h_s^2 d \langle \widehat{M} \rangle_s = E \left(\int_0^t h_s d \widehat{M}_s \right)^2 = E \left(E \left(\int_0^t h_s dM_s | G_t \right) \right)^2$$

$$\leq E \left(\int_0^t h_s dM_s \right)^2 = E \int_0^t h_s^2 d \langle M \rangle_s, \quad (A.1)$$

which implies that $\langle \widehat{M} \rangle$ is absolutely continuous w.r.t. $\langle M \rangle$, i.e.,

$$\langle \widehat{M} \rangle_t = \int_0^t \rho_s^2 d\langle M \rangle_s$$

for a G-predictable process ρ.   □

Moreover (A.1) implies that the process $\langle M \rangle - \langle \widehat{M} \rangle$ is increasing and hence $\rho^2 \leq 1$ $\mu^{\langle M \rangle}$ a.e.

Lemma A.2. *Let $H \in L^2(P, F_T)$, and let conditions* (A)–(C) *be satisfied. Then*

$$E \int_0^T \tilde{h}_u^2 d\langle M \rangle_u < \infty.$$

Proof. It is evident that

$$E \int_0^T (h_u^G)^2 d\langle \widehat{M} \rangle_u < \infty, \qquad E \int_0^T h_u^2 d\langle M \rangle_u < \infty.$$

Therefore, by the definition of $\tilde{h}$ and Lemma A.1,

$$E \int_0^T \tilde{h}_u^2 d\langle M \rangle_u \leq 2E \int_0^T \widehat{h}_u^2 d\langle M \rangle_u + 2E \int_0^T \left(\widehat{h_u^G} \right)^2 \rho_u^4 d\langle M \rangle_u$$

$$\leq 2E \int_0^T h_u^2 d\langle M \rangle_u + 2E \int_0^T \left(h_u^G \right)^2 \rho_u^2 d\langle \widehat{M} \rangle_u < \infty.$$

Thus $\tilde{h} \in \Pi(G)$ by Remark 2.5.   □

Lemma A.3. (a) *Let $Y = (Y_t, t \in [0, T])$ be a bounded positive submartingale with the canonical decomposition*

$$Y_t = Y_0 + B_t + m_t,$$

where B is a predictable increasing process and m is a martingale. Then $m \in BMO$.
(b) *In particular the martingale part of $V(2)$ belongs to BMO. If H is bounded, then martingale parts of $V(0)$ and $V(1)$ also belong to the class BMO, i.e., for $i = 0, 1, 2$,*

$$E \left(\int_\tau^T \varphi_u^2(i) \rho_u^2 d\langle M \rangle_u | G_\tau \right) + E \left(\langle m(i) \rangle_T - \langle m(i) \rangle_\tau | G_\tau \right) \leq C \qquad (A.2)$$

for every stopping time τ.

Proof. By applying the Itô formula for $Y_T^2 - Y_\tau^2$ we have

$$\langle m \rangle_T - \langle m \rangle_\tau + 2 \int_\tau^T Y_u dB_u + 2 \int_\tau^T Y_u dm_u = Y_T^2 - Y_\tau^2 \leq \text{const} \qquad (A.3)$$

Since Y is positive and B is an increasing process, by taking conditional expectations in (A.3) we obtain

$$E(\langle m \rangle_T - \langle m \rangle_\tau | F_\tau) \leq \text{const}$$

for any stopping time τ, and hence $m \in BMO$.
(A.2) follows from assertion (a) applied for positive submartingales $V(0), V(2)$, and $V(0) + V(2) - 2V(1)$. For the case $i = 1$ one should take into account also the inequality

$$\langle m(1) \rangle_t \leq \text{const}(\langle m(0) + m(2) - 2m(1) \rangle_t + \langle m(0) \rangle_t + \langle m(2) \rangle_t).$$

6. Acknowledgments

7. References

Bismut, J. M. (1973) Conjugate convex functions in optimal stochastic control, *J. Math. Anal. Appl.*, Vol. 44, 384–404.

Chitashvili, R. (1983) Martingale ideology in the theory of controlled stochastic processes, In: *Probability theory and mathematical statistics* (Tbilisi, 1982), 73–92, Lecture Notes in Math., 1021, Springer, Berlin.

Di Masi, G. B.; Platen, E.; Runggaldier, W. J. (1995) Hedging of options under discrete observation on assets with stochastic volatility, *Seminar on Stochastic Analysis, Random Fields and Applications*, pp. 359–364, Ascona, 1993, Progr. Probab., 36, Birkhäuser, Basel.

Delbaen, F.: Monat, P.: Schachermayer, W.: Schweizer, W. & Stricker, C. (1997) Weighted norm inequalities and hedging in incomplete markets, *Finance Stoch.*, Vol. 1, 181–227.

Dellacherie, C. & Meyer, P.-A. (1980) *Probabilités et Potentiel*, II, Hermann, Paris.

Duffie, D. & Richardson, H. R. (1991) Mean-variance hedging in continuous time, *Ann. Appl. Probab.*, Vol. 1, No. 1, 1–15.

El Karoui, N. & Quenez, M.-C. (1995) Dynamic programming and pricing of contingent claims in an incomplete market, *SIAM J. Control Optim.*, Vol. 33, No. 1, 29–66.

Föllmer, H. & Sondermann, D. (1986) Hedging of nonredundant contingent claims, In: *Contributions to mathematical economics*, 205–223, North-Holland, Amsterdam.

Frey, R. & Runggaldier, W. J. (1999) Risk-minimizing hedging strategies under restricted information: the case of stochastic volatility models observable only at discrete random times. Financial optimization, *Math. Methods Oper. Res.*, Vol. 50, No. 2, 339–350.

Gourieroux, C.: Laurent, J. P. & Pham, H. (1998) Mean-variance hedging and numeraire, *Math. Finance*, Vol. 8, No. 3, 179–200.

Heath, D.: Platen, E. & Schweizer, M. (2001) A comparison of two quadratic approaches to hedging in incomplete markets, *Math. Finance*, Vol. 11, No. 4, 385–413.

Hipp, C. (1993) Hedging general claims, In: *Proceedings of the 3rd AFIR Colloquium*, 2, pp. 603–613, Rome.

Jacod, J. (1979) *Calcul Stochastique et Problèmes de Martingales*, Lecture Notes in Mathematics, 714. Springer, ISBN 3-540-09253-6, Berlin.

Kazamaki, N. (1994) *Continuous Exponential Martingales and BMO*, Lecture Notes in Mathematics, 1579. Springer-Verlag, ISBN 3-540-58042-5, Berlin.

Kramkov, D. O. (1996) Optional decomposition of supermartingales and hedging contingent claims in incomplete security markets, *Probab. Theory Related Fields*, Vol. 105, No. 4, 459–479.

Kobylanski, M. (2000) Backward stochastic differential equations and partial differential equations with quadratic growth, *Ann. Probab.*, Vol. 28, No. 2, 558–602.

Lepeltier, J.-P. & San Martin, J. (1998) Existence for BSDE with superlinear-quadratic coefficient, *Stochastics Stochastics Rep.*, Vol. 63, No. 3-4, 227–240.

Liptser, R. Sh. & Shiryaev, A. N. (1986) *Martingale Theory*, Probability Theory and Mathematical Statistics, "Nauka", Moscow (in Russian).

Mania, M., Tevzadze, R. & Toronjadze, T. (2009) L^2-approximating Pricing under Restricted Information *Applied Mathematics and Optimization*, Vol.60, 39–70.

Mania, M. & Tevzadze, R. (2003) Backward stochastic PDE and imperfect hedging, *Int. J. Theor. Appl. Finance*, Vol. 6, No. 7, 663–692.

Morlais, M.-A. (2009) Quadratic BSDEs driven by a continuous martingale and applications to the utility maximization problem, *Finance Stoch.*, Vol. 13, No. 1, 121–150.

Pardoux, É. & Peng, S. G. (1990) Adapted solution of a backward stochastic differential equation, *Systems Control Lett.*, Vol. 14, No. 1, 55–61.

Pham, H. (2001) Mean-variance hedging for partially observed drift processes, *Int. J. Theor. Appl. Finance*, Vol. 4, No. 2, 263–284.

Rheinlander, T. & Schweizer, M. (1997) On L^2-projections on a space of stochastic integrals. *Ann. Probab.*, Vol. 25, No. 4, 1810–1831.

Schäl, M. (1994) On quadratic cost criteria for option hedging, *Math. Oper. Res.*, Vol. 19, No. 1, 121–131.

Schweizer, M. (1992) Mean-variance hedging for general claims, *Ann. Appl. Probab.*, Vol. 2, No. 1, 171–179.

Schweizer, M. (1994) Approximating random variables by stochastic integrals, *Ann. Probab.*, Vol. 22, No. 3, 1536–1575.

Schweizer, M. (1994) Risk-minimizing hedging strategies under restricted information, *Math. Finance*, Vol. 4, No. 4, 327–342.

Tevzadze, R. (2008) Solvability of backward stochastic differential equations with quadratic growth, *Stochastic Process. Appl.*, Vol. 118, No. 3, 503–515.

Pertinence and information needs of different subjects on markets and appropriate operative (tactical or strategic) stochastic control approaches

Vladimir Šimović and Vladimir Šimović, j.r.
University of Zagreb
Croatia

1. Short introduction

The main idea of this chapter is that it offers an original scientific discussion with a conclusion concerning the relevance of pertinence and information needs of different subjects on markets (as potential traders on various financial markets, stock markets, bond markets, commodity markets, and currency markets, etc.) and the significance of appropriate operative (tactical or strategic) stochastic control approaches.

The organisation of this chapter is very simple. After a short review of sources used and an overview of completed research, chapter parts with some definitions on the main subjects and research areas follow. Following the above stated, there are chapter sections with relatively short research examples of appropriate operative, tactical and strategic stochastic control approaches. All three approaches fits to adequate pertinence and information needs of different subjects on markets (the operative trading concept example, the tactical concept example as a quantitative approach to tactical asset allocation, and strategic concept examples as technical analysis in financial markets or strategic anti-money laundering analysis). The conclusion to this research is contained in the final chapter segment, before the cited references. In conclusion, this paper proposes quantitative and qualitative models for the right perception of adequate pertinence and information needs of different subjects on markets and the significance of appropriate operative (tactical or strategic) stochastic control approaches and expected results.

2. Important concepts

What was the problem? Even pioneers of information science and older authors (Perry et al., 1956; Perry & Kent, 1958; Taube, 1958; Schultz & Luhn, 1968; Mooers, 1976), which are researching problems considering the data, information or knowledge and document collection and retrieving processes in relation to data, information or knowledge processing, determined at the same time that the main focus should be placed on "real information needs". So, in defined period of time and for all different subjects on markets (as potential

traders on various financial markets, stock markets, bond markets, commodity markets, and currency markets, etc.) we may improve and adjust the activities related to data, information or knowledge collection and retrieving, in order to achieve accurate and useful data, information or knowledge appropriate to operative (tactical or strategic) stochastic control approaches to financial and other markets documentation and results. First, here is only short insight in some definitions of the main terms and subjects of researching area (stochastic, stochastic control, probabilistic and stochastic approaches, modern control and conventional control theory, cybernetics and informatics, pertinence and information needs, subjects on stock, bond, commodity, and currency markets, etc.).

Usually any kind of deterministic or essentially probabilistic time development, in relation to data or information and knowledge processing, which is analyzable in terms of probability, deserves the name of stochastic process. In mathematics, especially in probability theory, the field of stochastic processes has been a major area of research, and stochastic matrix is a matrix that has non-negative real entries that sum to one in each row. Stochastic always means random, and where a stochastic process is one whose behavior is non-deterministic in mathematical sense, in that a system's subsequent state is determined both by the process's predictable actions and by a random element. Also, it is well known from literature (Ĺström, 1970; Bertsekas & Shreve, 1996; Bertsekas, 2005; Bertsekas, 2007; Bertsekas & Tsitsiklis, 2008) that stochastic control is only a subfield of control theory which mainly addresses the design of a control methodology to deal with the probability of uncertainty in the data. In a stochastic control problem, the designer usually assumes that random noise and disturbances exist in both subsystems parts (in the model and in the controller), and the control design always must take into account these random deviations. Also, stochastic control aims to predict and to minimize the effects of these random deviations, by optimizing the design of the controller. Applications of stochastic control solutions are very different, like usage of stochastic control in: artificial intelligence, natural sciences (biology, physics, medicine, creativity, and geomorphology), music, social sciences, teaching and learning, language and linguistics, colour reproduction, mathematical theory and practice, business, manufacturing, finance, insurance, etc. For this research, interesting examples are: usage of stochastic control in insurance (Schmidli, 2008), and usage continuous-time stochastic control and optimization with financial applications (Pham, 2009), or usage stochastic optimal control for researching international finance and debt crises (Stein, 2006), etc. The financial markets use stochastic models to represent the seemingly random behaviour of assets such as stocks, commodities and interest rates, but usually these models are then used by quantitative analysts to value options on stock prices, bond prices, and on interest rates, as it can be seen in Markov models examples and many models examples which exist in the heart of the insurance industry (Schmidli, 2008).

When considering the "real informational needs" in context of relatively limited or different acting time and various interests of different subjects on financial and other markets and their appropriate operative (tactical or strategic) stochastic control approaches, the following facts should be noted:

- − An informational request is different from an information necessity.
- − It is the relevance of the process which connects documents to the informational request.
- − It is the pertinence of the process which connects the documents to the informational need.

In today's turbulent market environment we have different subjects on markets (as potential traders on various markets) with similar or different interests and with relatively limited or even different acting time. Consequently, in order to achieve accurate and useful data, information or knowledge, we have to improve and adjust not only the activities related to retrieving and collecting of data (information or knowledge), but also the tools, techniques and methods appropriate to operative (tactical or strategic) stochastic control approaches to deal with all kind of data, information, knowledge (or documentation) about financial and other markets. Of course, one should always have a clear perception of the documents search algorithm tools which are used in any of research and learning processes, and of possible results of the documents (Data, Information and Knowledge, in short "D,I,K") search. The results of the search can always be (Fig. 1.): relevant and pertinent, relevant and non-pertinent and pertinent and irrelevant. Always, the goal is to have relevant and pertinent results, which can be achieved exclusively by knowing the "real information needs" of the persons (or financial subject). Also, when considering the relevance (Table 1.) of the derived documents (D,I,K): all the derived documents are not always relevant, in other words, all the relevant documents are often not found!
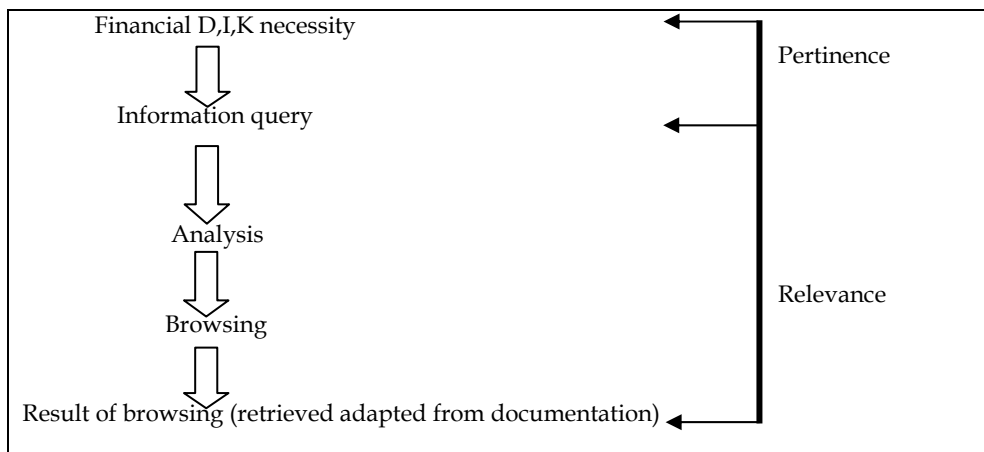

Fig. 1. Algorithm of research of financial documentation (D,I,K), adapted from (Tuđman et al., 1993)

Financial D,I,K	Relevant	Irrelevant	
Found	R_f	I_f	$R_f + I_f$
Not found	R_{nf}	I_{nf}	$R_{nf} + I_{nf}$
	$R_f + R_{nf}$	$I_f + I_{nf}$	

Table 1. The relevance of financial documentation (D,I,K), adapted from (Tuđman et al., 1993)

Relevance can be expressed in percentages (%) through the following terms: exactness or precision, and response or recall. It can also be expressed in the form of the following ratios (Table 2) and equations (1), (2):

Exactness (or precision) = the number of found relevant financial documents (D,I,K) / the number of found financial documents (D,I,K) x 100%
Recall (or response) = the number of found relevant financial documents (D,I,K) / the number of relevant financial documents (D,I,K) in the system x 100%

Table 2. Ratios for exactness or precision, and response or recall

$$E = R_f / N_f \times 100\% \qquad\qquad (1)$$

where E is exactness (or precision); R_f is the number of found relevant financial documents (D,I,K); N_f is the number of found financial documents (D,I,K), and

$$R = R_f / R_s \times 100\% \qquad\qquad (2)$$

where R is recall (or response); R_f is the number of found relevant financial documents (D,I,K); R_s is the number of relevant financial documents (D,I,K) in the system.
Following the above stated, there are chapter sections following with short research examples of appropriate operative, tactical and strategic stochastic control approaches.

3. Operative, tactical and strategic research examples of appropriate stochastic control approaches to various markets

3.1 Example of appropriate operative stochastic control approach
In this chapter we give an operative research example as a relatively original and new stochastic control approach to day trading, and through this approach trader eliminate some of the risks of day trading through market specialization. When we have different subjects on markets, as potential traders on various markets, with similar or different interests, with relatively limited or even different acting time, market specialization help us to improve and adjust not only the activities related to retrieving and collecting data, information or knowledge in turbulent market environment, in order to achieve accurate and useful data, information or knowledge, but also the tools, techniques and methods which are appropriate to operative (tactical or strategic) stochastic control approaches (dealing with relevant data, information, knowledge, or documentation about financial and other markets). The goal of this approach to day trading is to have maximum relevant and pertinent results, which can be achieved exclusively by knowing the "real information needs" of the persons (or financial subject) which we know as day traders. When considering the relevance of the derived financial indicators and documents (D,I,K) referenced to day trading we have to know that all the derived documents are not always relevant, and all the relevant documents are often not found. Market specialization and usage of appropriate stochastic control approach, tools and techniques are necessity.
First question is: what we know about different subjects on markets, as potential traders on various markets, with similar or different interests and with relatively limited or even

Pertinence and information needs of different subjects on markets
and appropriate operative (tactical or strategic) stochastic control approaches
613

different acting time needed for proposed market specialization? The operative concept is that the trader on a specific financial market should specialize him/herself in just one (blue-chip) stock and use existing day trading techniques (trend following, playing news, range trading, scalping, technical analysis, covering spreads…) to make money. Although there is no comprehensive empirical evidence available to answer the question whether individual day-traders gain profits, there is a number of studies (Barber et al., 2005) that point out that only a few are able to consistently earn profits sufficient to cover transaction costs and thus make money. Also, after the US market earned strong returns in 2003, day trading made a comeback and once again became a popular trading method among traders. As an operative concept, the day trading concept of buying and selling stocks on margin alone suggests that it is more risky than the usual "going long" way of making profit. The name, day trading, refers to a practice of buying (selling short) and selling (buying to cover) stocks during the day in such manner, that at the end of the day there has been no net change in position; a complete round – trip trade has been made. A primary motivation of this style of trading is to avoid the risks of radical changes in prices that may occur if a stock is held overnight that could lead to large losses. Traders performing such round – trip trades are called day traders. The U.S. Securities and Exchange Commission adopted a new term in the year 2000, "pattern day trader", referring to a customer who places four or more round-trip orders over a five-day period, provided the number of trades is more than six percent in the account for the five day period. On February 27, 2001, the Securities and Exchange Commission (SEC) approved amendments to National Association of Securities Dealers, Inc. (NASD®) Rule 2520 relating to margin requirements for day traders. Under the approved amendments, a pattern day trader would be required to maintain a minimum equity of $25,000 at all times. If the account falls below the $25,000 requirement, the pattern day trader would not be permitted to day trade until the account is restored.

Second question is: what we know about common techniques and methods used by day traders which represent significant part of different subjects on markets, with similar or different interests, with relatively limited or even different acting time needed for proposed market specialization? There are minimally four common techniques used by day traders: trend following, playing news, range trading and scalping. Playing news and trend following are two techniques that are primarily in the realm of a day trader. When a trader is following a trend, he assumes that the stock which had been rising will continue to rise, and vice versa. One could say he is actually following the stocks "momentum". When a trader is playing news, his basic strategy is to buy a stock which has just announced good news, or sell short a stock which has announced bad news. After its boom during the dotcom frenzy of the late 1990s and the loss in popularity after the Internet bubble burst, day trading is making a comeback. After three years of strong stock market performance, a constantly increasing number of investors use day trading techniques to make profit.

In 2006, a search on the Social Science Service Network reports 395 articles on day trading, with over 40% published in the last 3 years. Similar searches on the most popular online bookstore, Amazon result in more than 400 popular books on day trading. In 2006, according to (Alexa Traffic Rankings, 2006), Amazon was the 15th most popular site and the highest ranked online bookstore on the Top 500 site ranking list. Today, according to (Alexa Traffic Rankings, 2010), Amazon site is the global leader with similar popularity and the highest ranked online bookstore on the Top 500 site ranking list. In 2006, many of the popular news agencies and papers report a surge in day trading popularity while some are

also reporting its negative sides. Associated Press reported a centrepiece "In Japan, day trading surges in popularity" on May 10, 2006 (Associated Press, 2006). The Sunday Times published an article "High-risk day trading makes a comeback" on February 26, 2006 (The Sunday Times, 2006). Searching the most popular World Wide Web searching engine, Google, for the term "day trading" results in over 120,000,000 links. In fact, Google was the most popular search engine according to the last Nielsen NetRatings search engine ratings that were published in November of 2005 (Nielsen NetRatings, 2005).

Fig. 2. Examples of the web reports on day trading

After U.S. Federal Trade Commission warning in 2000, the first one of those links redirects a user's browser to a warning about risks involved in day trading published on the homepage of the U.S. Securities and Exchange Commission.
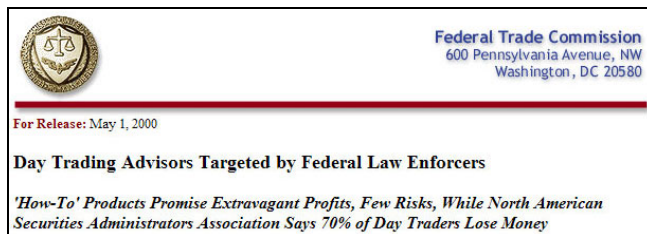
Fig. 3. U.S. Federal Trade Commission warning on day trading

Fig. 4. U.S. Securities and Exchange Commission warning on day trading

New question is: what is the day trading controversy? The day trading controversy is mainly fuelled by its main con, it is risky. The constant usage of margin (borrowed funds) is the strong and the weak point of day trading, because the usage of margin amplifies gains and losses such that substantial losses (and gains) may occur in a short period of time. Because day trading implies a minimum of two trades per business day (buying means selling short, and selling means buying to cover), a part of the day trader's funds are used to pay commissions (the broker's basic fee for purchasing or selling securities as an agent). The higher the number of trades per day is, the bigger the part of day trader's funds is used to pay commissions. Day trading also often requires live quotes which are costly, and therefore also have an impact on the funds of a day trader. For every one of these (main) cons, day trading is, as it was already mentioned, considered risky. An integral part in the day trading controversy is the day trader himself. Claims of easy and fast profits from day trading have attracted a significant number of non experienced and "casual" traders into day trading that do not fully understand the risks they are taking. With its latest comeback, day trading has become a business to people other than traders. Numerous websites offer tips and advices while online bookstores offer books on day trading strategies. With all that in mind, one could wonder do day traders make money. Although that question cannot be answered with certainty, a few existing studies do not paint a pretty picture. A comprehensive analysis of the profitability of all day trading activity in Taiwan over a five year period

(Barber et al., 2005) has shown that in a typical six month period, more than eight out of ten day traders lose money.

Fig. 5. U.S. Securities and Exchange Commission warning on day trading info

Another question is: what is the main concept of specialized day trading? According to (Simovic & Simovic, 2006), the main goal of specialized day trading is to offer a new approach and technique (or method) to day trading, an approach that would dampen or eliminate some of the negative sides of "regular" day trading. The concept is simple. Instead of using the usual day trading techniques on various types of stocks, the trader should specialize himself in using the same, already mentioned techniques, but with just one (blue chip) stock. A blue chip stock is the stock of a well-established company having stable earnings and no extensive liabilities.

And important question is: what is the main concept of proposed specialization? The main reason why specialization is proposed as new stochastic approach is in the fact that trading with different stocks on a daily basis brings a certain element of uncertainty since the day trader often does not have the time to thoroughly "check up" on a stock he is trading with. Focusing on just one stock eliminates the element of uncertainty and gives the day trader the opportunity, to through time better learn about its "behaviour" and how the selected stock reacts to certain events like splits, earning announcements, general (good or bad) news etc. Also, when considering the relevance of the derived documents (D,I,K) needed for day trading, because all the derived documents are not always relevant, and all the relevant documents are often not found, with blue chip stock (which is the stock of a well-established company having stable earnings and no extensive liabilities) stochastically we have lower level of problem. Suppose that our exactness or precision is not higher (E), because we have the same number of found relevant financial documents (R_f) or D,I,K, in relation to the same number of found financial documents (N_f) or D,I,K. But recall or response (R) have to be significantly better (or higher↑), because it represents ratio between the same R_f and now very small (or lower↓) R_s (where R_s is the number of relevant financial documents or D,I,K in our day trading sub-system). Consequently, better pertinence and relevance can be achieved through constant monitoring with same and better tools and techniques, also. A trader could gain knowledge on how the stock reacts on markets ups and downs, better insight on the meaning of afterhours trading activity or the manner how the company releases announcements (according to (DellaVigna & Pollet, 2005), "do worse announcements get announced on Friday to dampen the short-term response of the trader and thus the market?"), etc.

Fig. 6. NASDAQ (National Association of Securities Dealers Automated Quotations) Composite quote data from 1997 to 2006

Because a blue chip stock is the stock of a well-established company having stable earnings and no extensive liabilities the focus was put on blue chip stocks, and consequently they offer stability, which of course translates into low risk. With day trading blue chip stocks one cannot expect great profit in a single day, but the trade-off is that one cannot expect great loss also. For further stochastic analysis of "behaviour" of the blue chip stocks, we've taken Microsoft as an example. The analysis was based on 3 years (756 working days) of daily data, which were obtained on Nasdaq's website (Nasdaq, 2003), from 07/18/2003 till 07/18/2006. That particular period of time was chosen for our analysis because it consists of newest daily quotes from the last MSFT (Symbol that represents a Microsoft stock) stock split (02/18/2003).

The data is consisted of the opening price (o_i), closing price (c_i), daily low (l_i) and the daily high (h_i). Using the following equations we've calculated the values (percentage) of the daily spreads (S_i) between the opening and the closing price. This was done in order to calculate the average spread (S_{av}) between the two already mentioned values which will help us illustrate the potential of this kind of day trading.

$$\left| \frac{c_i - o_i}{o_i} * 100 \right| = S_i \qquad \frac{\sum_{i=1}^{n} S_i}{n} = S_{av} \qquad \begin{array}{l} i \in [1, n] \\ n = 756 \end{array} \qquad (3)$$

The average spread between the opening and the closing price (S_{av}) for a Microsoft stock in that period of time was 0.706059344 % which is a very good indicator of its stability.
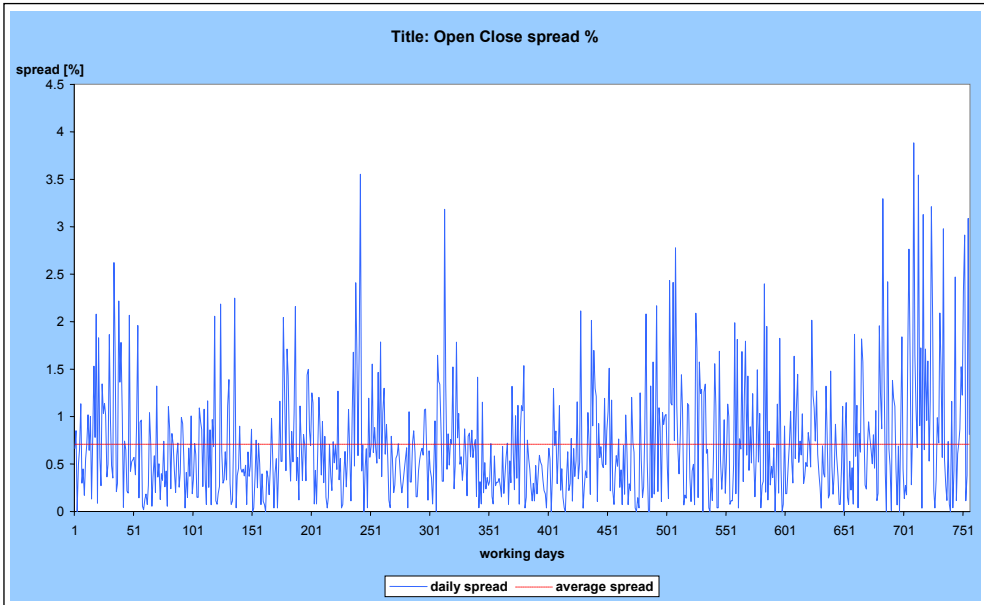
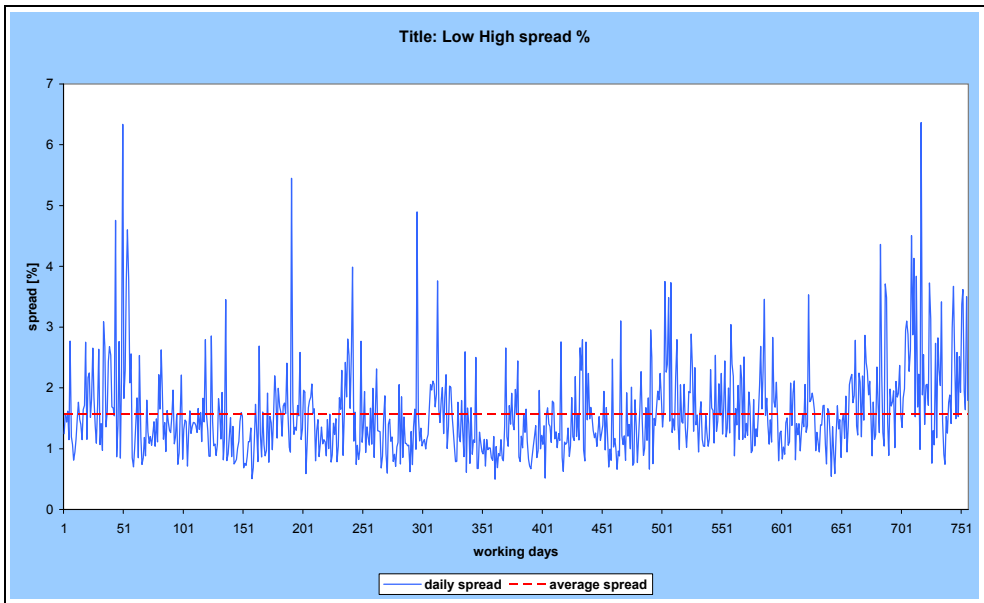Fig. 7. A graph showing the spread between the opening and the closing price

Fig. 8. A graph showing the spread between the daily low and the daily high

Using the exact equations we've also calculated the values (percentage) of the daily spreads (Q_i) between the daily low and the daily high. This was done in order to calculate the

average spread (Q_{av}) between those two values so that we can illustrate the full (but unreachable) potential of this kind of day trading.

$$\left| \frac{h_i - l_i}{l_i} * 100 \right| = Q_i \qquad \frac{\sum_{i=1}^{n} Q_i}{n} = Q_{av} \qquad \begin{array}{l} i \in [1, n] \\ n = 756 \end{array} \qquad (4)$$

The average spread between the daily low and the daily high (Q_{av}) for a Microsoft stock in that period of time was 1.57655549 %.

To show the potential profit for this way of day trading we are going to use the following functions with the already calculated average daily spreads. These functions represent the total return (via percentage) in 30 working days of which a certain number (α) had a "positive" (and the rest "negative") trading outcome.

$$\left(1 + \frac{S_{av}}{100}\right)^{\alpha} * \left(1 - \frac{S_{av}}{100}\right)^{30-\alpha} * 100 = S_p \qquad (5)$$

$$\left(1 + \frac{Q_{av}}{100}\right)^{\alpha} * \left(1 - \frac{Q_{av}}{100}\right)^{30-\alpha} * 100 = Q_p \qquad (6)$$
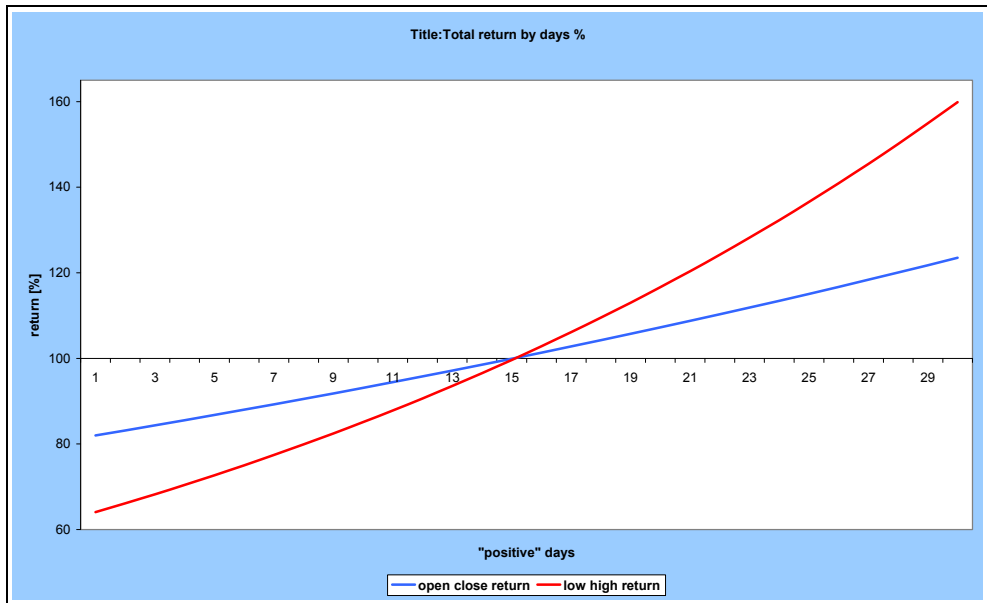

Fig. 9. A graph showing the spread between the daily low and the daily high without the usage of margin
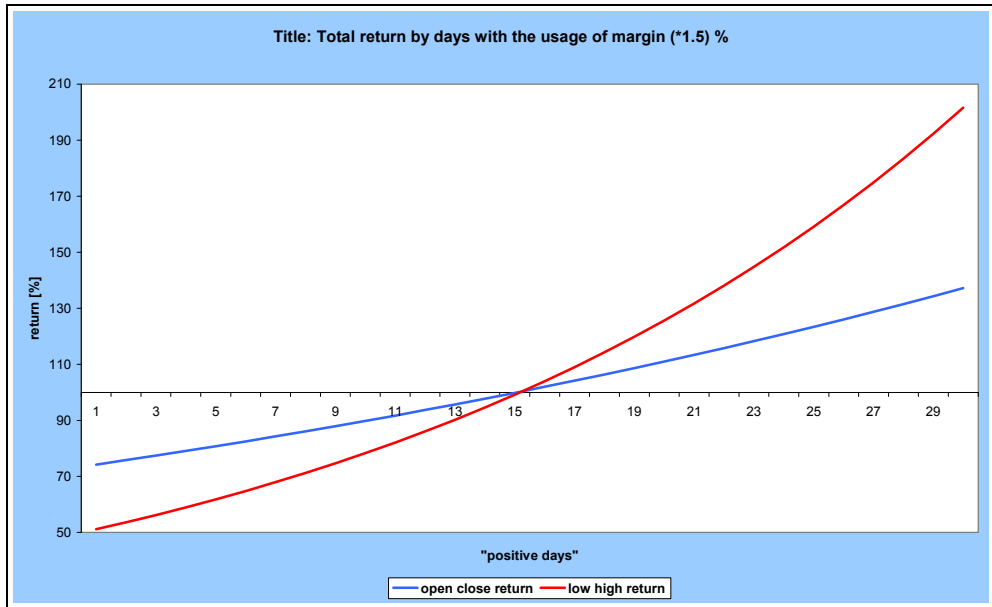
Fig. 10. A graph showing the spread between the daily low and the daily high with the usage of margin

With the usage of margin (M), gains and losses are amplified.

$$\left(1+\frac{S_{av}}{100}*\left(\frac{M}{100}\right)\right)^{\alpha}*\left(1-\frac{S_{av}}{100}*\left(\frac{M}{100}\right)\right)^{30-\alpha}*100 = S_{mp} \qquad (7)$$

$$\left(1+\frac{Q_{av}}{100}*\left(\frac{M}{100}\right)\right)^{\alpha}*\left(1-\frac{Q_{av}}{100}*\left(\frac{M}{100}\right)\right)^{30-\alpha}*100 = Q_{mp} \qquad (8)$$

As Wall Street Gordon Gekko says that "information is the most important commodity when trading" (Kuepper, 2010), through specialized day trading example, we've tried to offers a new approach to day trading information and with it eliminate some of the operative risks of day trading, and to show how important is concept of pertinence and relevance. This operative example tried to explain the reasons behind the concept of specialization model in trading in just one (blue chip) stock with the usage of existing day trading techniques and show that the usage of such concept has potential and can be profitable. With some new stochastic control or optimization approaches we can operatively reduce a level of noise from irrelevant D, I, K, about market, in both of our day trading subsystem parts (model and controller). In our stochastic control solution, the designer have to assume that random noise and disturbances exist in both subsystems parts (in the model and in the controller), and consequently the control design always must take into account these random deviations. This is for further researching.

3.2 Examples of appropriate tactical and strategic stochastic control approaches

In this chapter we give two very short research examples (from practice and relevant literature) of appropriate tactical and strategic stochastic control approaches. The tactical model conceptually represents one example of a quantitative approach to tactical asset allocation (Faber, 2009). In tactical example from relevant literature (Faber, 2009) one can see how to create a simple-to-follow model example as a tactical method for managing risk in a single asset class and, by extension, a portfolio of assets. From the available sources, one can conclude that a non-discretionary trend following model acts as a risk-reduction tactical technique with no adverse impact on return. Here we only try to give adequate references (with original comments and models) that utilizing a monthly system and where an investor would have been able to (avoid massive losses) increase risk adjusted returns and sidestep many of the protracted bear markets in various asset classes. Similar to operative example about day trading this tactical example represent reference model how to analyse and research methods that are used in tactical or even strategic stochastic control approach. There are various technical analysis tools available to tactical level investors, and in defined period of tactical time, for different subjects on markets (as potential tactical traders on various financial markets, stock markets, bond markets, commodity markets, and currency markets, etc.) which may improve and adjust the tactical activities related to collection and retrieving D,I,K, in order to achieve accurate and useful D,I,K, appropriate to tactical (or even strategic) stochastic control approaches to financial and other markets. Mainly the tactical methods used to analyze and predict the performance of a company's stock fall into two broad categories: fundamental and technical analysis. Those who use technical analysis (various tactical level investors, etc.) look for peaks, bottoms, trends, patterns and other factors affecting a stocks, bonds, "forex", futures, options, indexes, currencies and commodities price movement and then make "buy or sell" decisions based on those factors. It is important to notice that this is a tactical level technique many people and companies attempt, but few are truly successful at it. Also, the world of technical analysis is huge because there are literally hundreds of different patterns and indicators that investors and traders claim to have success with.

The main purpose of this tactical example from (Faber, 2009) was to create a simple-to-follow method for managing risk in a single asset class and, by extension, a portfolio of assets. A non-discretionary and trend following model here acts as a risk-reduction tactical technique with no adverse impact on return. Also, notice that when tested on various markets, risk-adjusted returns were almost universally improved, what is tactically and strategically very important. In this example (Faber, 2009) one can see that utilizing a monthly system since 1973, an tactical investor would have been able to increase risk adjusted returns by diversifying portfolio assets and employing a market-timing solution. In addition, the investor would have also been able to sidestep many of the protracted bear markets in various asset classes, and with tactically avoiding these massive losses would have resulted in equity-like returns with bond-like volatility and drawdown.

In this chapter we give a short strategic concept example, because we use specific technical analysis in financial markets as an original strategic concept (with original comments and models) and which can be used for strategic anti-money laundering analysis as another original strategic concept (with original comments and models). What is Technical Analysis (TA)? TA is a tactical and specific strategic method of evaluating the markets value by analyzing statistics generated by market activity, past prices and volume. TA does not

attempt to measure a vehicle's intrinsic value; instead they look at charts for patterns and indicators that will determine future performance. TA has become increasingly popular over the past several years, only when people "believe" that the historical performance is a strong indication of future performance. The use of past performance should come as no surprise, because people (and other market subject) using fundamental analysis have always looked at the past performance of companies by comparing fiscal data from previous quarters and years to determine future growth. The difference lies in the technical analyst's belief that securities move according to very predictable trends and patterns, and that these trends continue until something happens to change the trend, or until this change occurs, price levels are predictable. There are many instances of various investors successfully trading a security using only their past knowledge of the security's chart, without even understanding what the company really does. Although TA is a terrific analytical tool, most of market subjects agree it is much more effective when used in combination with proper money management tools. TA or formula traders on market use mathematical formulae to decide when a stock is going to rise or fall, and most traders use technical indicators; although more experienced traders tend to use fewer of them. Some old traders on market do not even use charts, but buy and sell just from so called approach "reading the tape", that is in fact procedure: watching the bid, then ask and trade with volume numbers from a trading screen. How to "cover spreads" in tactical approach practice? Playing the spread involves buying at the bid price and selling at the ask price, where the numerical difference between these two prices is known as the spread. This procedure allows for profit even when the bid and ask don't move at all, and consequently what the bigger the spread, the more inefficient the market for that particular stock, and the more potential for profit. As opposed to trade commissions, this spread is the mechanism that some large Wall Street firms use to make most of their money since the advent of online discount brokerages.

How to make categorization of tactical and strategic investors and companies by specific trading market? According (Faber, 2009), about 75% of all trades are to the upside - that is, the trader buys an issue hoping its price will rise - because of the stock market's historical tendency to rise and because there are no technical limitations on it. Also, about 25% of equity trades, however, are short sales. The trader borrows stock from his broker and sells the borrowed stock, hoping that the price will fall and he will be able to purchase the shares at a lower price. There are several technical problems with short sales: the broker may not have shares to lend in a specific issue, some short sales can only be made if the stock price or bid has just risen (known as an "uptick"), and the broker can call for return of its shares at any time. When the typical online investor places a market order to buy a stock, his broker submits this order to a market maker (MM), who then fulfills the order at the ask price. Then ask price is the price the MM is asking for the stock, and when the typical online investor places a market order to sell a stock, the broker submits the order to a MM and sells at the bid price, i.e. what the MM is bidding for the stock. Due to the liquidity of the modern market, orders are constantly flowing, and usually a MM will buy a stock just to turn around and sell it to a particular broker. Among all other things one of the main purposes of the MM is to maintain liquidity in the market. How to make categorization of companies by market cap? First, we have to know that market capitalization, often abbreviated to market cap, is a business term that refers to the aggregate value of a firm's outstanding common shares. Market capitalization reflects the total value of a firm's equity currently available on the market, and this measure differs from equity value to the extent that a firm has

outstanding stock options or other securities convertible to common shares. The size and growth of a firm's market cap is often one of the critical measurements of a public company's success or failure. Market cap may increase or decrease for reasons unrelated to performance such as acquisitions, divestitures and stock repurchases, and it is calculated by multiplying the number of outstanding common shares of the firm and the current price of those shares. The term capitalization is sometimes used as a synonym of market cap. It denotes the total amount of funds used to finance a firm's balance sheet and is calculated as market capitalization plus debt, as book or market value, plus preferred stock. The total market cap of all the companies listed on the New York Stock Exchange is greater than the amount of money in the United States. While there are no strong definitions for market cap categorizations, a few terms are frequently used to group companies by capitalization. In the U.S., companies and stocks are often categorized by the following approximate market cap values: micro-cap - market cap under US$100 million; small-cap - market cap below US$1 billion; mid-cap - market cap between US$1 billion and US$5 billion; and large-cap - market cap exceeds US$5 billion. The small-cap definition is far more controversial than those for the mid-cap and large-cap classes. Typical values for the ranges are enumerated and "blue chip" is sometimes used as a synonym for large-cap, while some investors consider any micro-cap or nano-cap issue to be a penny stock, regardless of share price. Examples of share valuation compared to market cap or price, and share ownership, from (Yahoo!® Finance, 2010), and according to: valuation measures and share statistics.

Here is our strategic example. This part is short explanation of the main Operations Research (OR) concepts and results, which are accomplished during the soft computing process (based on fuzzy logic) of the analytical entropy of the modern financial analytical function, what is also a solid base for future simulation modelling works. Here are some remarks about: "soft computing" and "fuzzy logic". The conventional approaches for predicting and understanding the entropy of the modern financial analytical function and the behaviour of various financial markets that are based on well-known analytical techniques can prove to be inadequate. Sometimes, even at the initial stages of establishing an appropriate mathematical model, the computational environment used in such an analytical approach is often too categorical and inflexible in order to cope with the financial intricacy and the complexity of the real world financial systems (like Croatian financial systems and internal financial market are, or can be). The mathematical model of modern financial analytical function is not only based on so called "hard computing" methods (binary logic, crisp systems, numerical analysis, probability theory, differential equations, functional analysis, mathematical programming, approximation theory and crisp software), because the fact that Croatian internal financial market usually has not the attributes of quantitative, precision, formality and categorisation. All mentioned before clearly turns out that in dealing with such systems we have to face with a very high degree of uncertainty and to tolerate very big degree of imprecision. Idea is to exploit a formalisation of the human ability to make rational decision in an uncertain and imprecise environment, also to exploit the soft computing tolerance for imprecision and uncertainty, and to achieve an acceptable model solution at a low cost and tractability. The principle of soft computing, given by Prof. Lotfi A. Zadeh (Zadeh, 1996) is: "Exploit the tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness, low solution cost and better rapport with reality." Soft computing has the attributes of qualitative, dispositional, approximation, and it is oriented towards the analysis based on fuzzy logic, artificial neural

networks and probabilistic reasoning, including genetic algorithms, chaos theory and parts of machine learning. Fuzzy logic is mainly concerned with imprecision and approximate reasoning, neural-computing mainly with learning and curve fitting, genetic-computing with searching and optimisation, and probabilistic reasoning mainly with uncertainty and propagation of belief. The main constituents of soft computing are complementary rather than competitive elements, and usually it can be more effective to use them in a synergetic combination manner (rather than exclusively). A fusion of all three constituents of soft computing is not very common.

When we think about Business Intelligence (BI) in the context of the entropy of the modern financial analytical function it is very important to understand what BI is. BI is process of collecting various business data (financial and other data) and transforming it to BI information that is used to provide better decisions, which improve the organisation performance. For collecting and managing corporate financial and other data many corporate and financial organizations (governmental or not) have more than one operational system. BI (with data warehousing) system is enabling technology that provides some governmental and financial organizations with end-to-end solutions for managing, organizing and exploiting financial and other data throughout the enterprise. This enabling technology provides tools to bring all the pieces of financial business information together in a single organized data repository that is driven by common set of financial and other business definitions. BI systems are used for exploration, analysis and reporting of trends found in the transactional data. They are designed to process inquiries and are vital to creating strategic competitive advantages that can affect an organizations' short-term and long-term financial profitability. There is an urgent need to collate (financial and other) data and provide financial decision-makers with the facility of additional financial reports, facility to explore and analyse data, in different dimensions and arriving at financial and other decisions, strategic to the governmental and financial organization. During the whole BI process, modern financial analytical function is mainly concerned with process of discovering financial knowledge. Maybe it can be the most significant part for the whole BI process. The financial analytical function of the BI was prepared for investigations of various financial events, financial markets, subjects or entities, and for financial business operations controls methods, etc. An application of this model usually increases the investigation group effectiveness, efficiency, and quality of the operational and strategic financial market investigative operations that are in usage during the whole financial knowledge discovery process. During the strategic BI processes financial "knowledge workers" are usually in situation that they have to work with: data marts (for small areas of financial data analysis) or with data warehouses (for larger areas of financial data analysis). Data marts and data warehouses are data collections produced during the analytical mixing of internal and external financial and other data. The analytical mixing is result of logic process that is prepared with detailed or aggregated view on mixing data. Result is data warehouses with synthetic financial view (detailed and aggregated). Synthetic financial view is producing with processes like the strategic BI visualisation & data drilling up and down through the specific time, financial value, financial service, financial product, financial market or combined dimension.

The basic components of the financial BI solution are: multidimensional (financial and other) data store, data extraction tool and front end tool for analysis. In general, an application of modern financial knowledge discovery model may be accomplished through finalisation of

few financial data mining tasks, which can be classified into two categories: descriptive financial data mining (describes the data set in a concise and summary manner and presents interesting general properties of the financial data), and predictive financial data mining (constructs one or a set of financial & other models, performs inference on the available set of financial & other interesting data, and attempts to predict the behaviour of new financial & other interesting data sets). The fast computerisation of the Croatian financial sector and various impacts of data mining should not be under-estimated with usage of the modern financial analytical function.
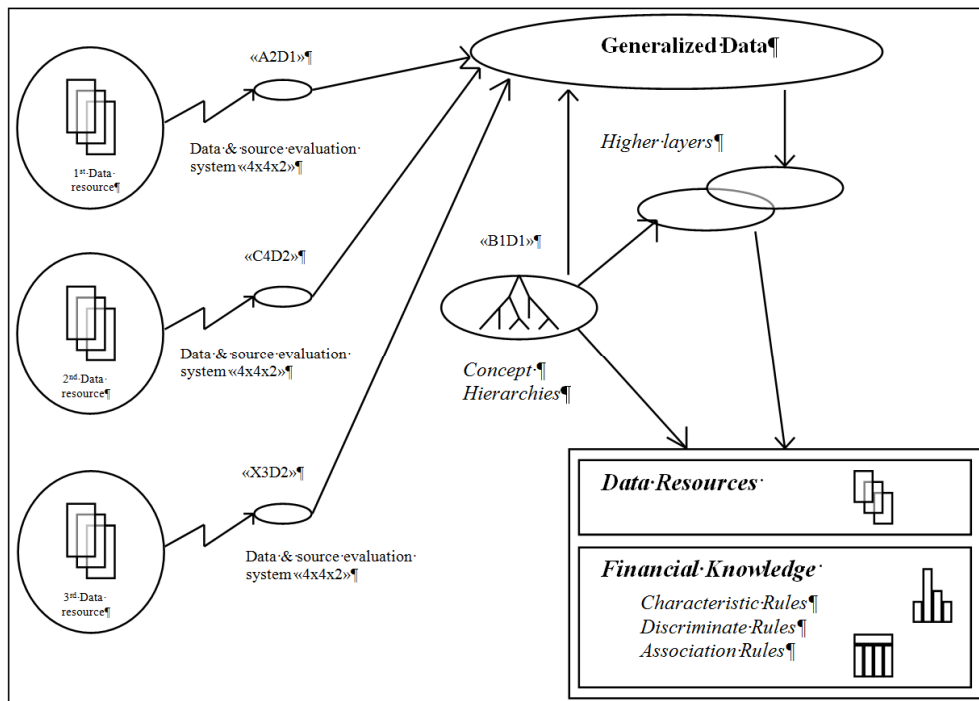

Fig. 11. The BI financial knowledge discovery model

Financial knowledge discovery model is a tool for translating the bits of various (mainly financial) data resources and observations into an understandable pattern of data behaviour (Fig. 11). With application of the financial knowledge discovery model and intelligent software tools in the informational and financial aspects of the BI one can radically change quality of the whole BI process. When a large amount of financial business interrelated data are effectively analysed from different perspectives BI (with data warehousing) system is enabling technology that provides organizations with end-to-end solutions for managing, organizing and exploiting financial and other data. Without BI (& data warehousing) technology this can also pose threats to the goal of protecting data security and guarding against the invasion of financial privacy (throughout the whole enterprise). This enabling technology (BI with data warehousing) provides tools to bring all the pieces of financial business information together in a single security organized data repository that is driven by

common set of financial and other business definitions, which are used for exploration, analysis and reporting of trends found in the transactional data. In the final BI analysis, financial network construction model with financial knowledge discovery concept provides a framework to look at interactions and transactions as a function of both permanent and temporary relations. In an application of profound financial knowledge discovery model dealing with semantic heterogeneity is necessity and only schema level analysis is not sufficient to solve the problem. Because of that the data level analysis with analysis of database (data warehouse) contents in co-operative BI information systems was widely introduced and successfully used. In co-operative BI information systems "On Line Analytical Processing" (OLAM) method/technique is in fact dealing with a multiple-layer database (MLDB) or multidimensional database (MDDB) model and co-operative heterogeneous databases. OLAM deals with generalisation-based financial data mining techniques. For example (Fig. 12), "DBMiner" is a cutting-edge intelligent data mining and data warehousing system and have OLAP and OLAM capabilities (Han, 1999).
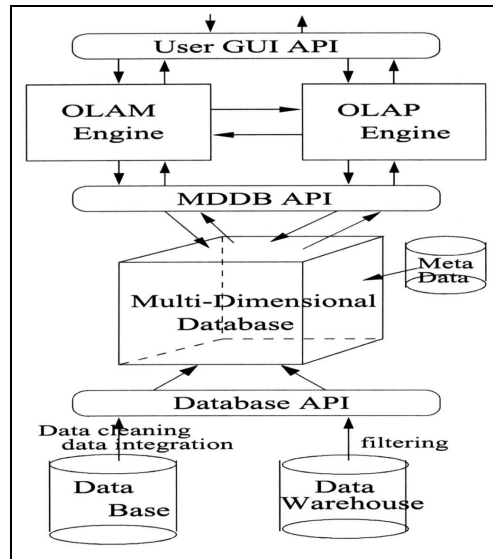
Fig. 12. "DBMiner" software architecture, as integrated OLAM & OLAP architecture (Han, 1999)

Also, BI visualisation techniques (of the modern financial analytical function) are examples that financial data mining is discovery-driven and that financial pattern is automatically extracted from data, what requires substantial search efforts. Prof. Lotfi A. Zadeh founded the soft computing based on fuzzy logic in 1965, with the well-known "theory of fuzzy sets". A fuzzy decision is a special type of fuzzy sets. The decision in a fuzzy environment (depending on the context) can be viewed as the intersection of fuzzy constraints and fuzzy objective function(s), where the fuzzy objective function is characterised by its membership function, and represents constraints. By analogy to no fuzzy environments (where the decision is the selection of activities that simultaneously satisfy objective function(s) and constraints), the decision in a fuzzy environment is defined as the optimal selection of

activities that simultaneously satisfy fuzzy objective function and fuzzy constraints. According to (Simovic et al., 1998), assumptions are that the constraints are no interactive, the logical and corresponds to the intersection. By analogy to crisp (no fuzzy) environments and to crisp decision logic, in fuzzy environments we have slightly different decision logic (usually called "fuzzy decision logic"). A linguistic variable x is a variable whose values are words or sentences in natural or artificial language. For example, if intelligence is interpreted as a linguistic variable, then its term set T(X), as the set of its linguistic values, might be:

T(intelligence) = disinformation + very low information + low information + unknown (or entropy) + high information + very high information + …. ,

where each of the terms in T(intelligence) is a label of fuzzy subset of a universe of discourse, say U = [xmin, xmax], or because of practical reasons usually U = [0, xmax] $\subset$ R. With a linguistic variable are associated two rules: syntactic rule (which defines the well formed sentences in T(X)) and semantic rule (by which the meaning of the terms in T(X) may be determined).

7. Conclusion

In the chapter where we present the conclusion of our research, we also introduce an adequate algorithm (in the form of a graphic and mathematical representation) for the appropriate perception of pertinence and information needs of different subjects on markets. We announce main trends and a great significance of appropriate operative (tactical or strategic) stochastic control (as specific quantitative and qualitative models) approaches in correlation to expected results. Finally, we promote new research areas and suggest future research directions.

8. References

Alexa Traffic Rankings (2006). *http://www.alexa.com/*, (07/18/2006; 04/17/2010), site of The Web Information Company

Associated Press (2006). *http://www.ap.org/*, (05/10/2006; 04/17/2010), site of The Associated Press

Barber, B. M.; Lee, Y. T.; Liu, Y. J. & Odean, T. (2005). Do Individual Day Traders Make Money? Evidence from Taiwan, University of California, Berkeley

Bertsekas, D. P. (2005). *Dynamic Programming and Optimal Control*, *Vol. I, 3rd Edition*, Athena Scientific, ISBN 1-886529-26-4, Nashua

Bertsekas, D. P. (2007). *Dynamic Programming and Optimal Control*, *Vol. II, 3rd Edition*, Athena Scientific, ISBN 1-886529-30-2, Nashua

Bertsekas, D. P. & Shreve, S. E. (1996). *Stochastic Optimal Control: The Discrete-Time Case*, Athena Scientific, ISBN 1-886529-03-5, Nashua

Bertsekas, D. P. & Tsitsiklis, J. N. (2008). *Introduction to Probability, 2nd Edition*, Athena Scientific, ISBN 978-1-886529-23-6, Nashua

DellaVigna, S. & Pollet, J. M. (2005). Strategic Release of Information on Friday: Evidence from Earnings Announcements, University of California, Berkeley

Faber, M. T. (2009). A Quantitative Approach to Tactical Asset Allocation, In: *The Journal of Wealth Management*, (February 2009, Update), Institutional Investor Inc, ISSN 1534-7524, New York

Han, J. (1999). Characteristic Rules, DBMiner, In: *Handbook of Data Mining and Knowledge Discovery*, Kloesgen, W.; Zytkow, J. (Eds.), Oxford University Press

Kuepper, J. (2010). A Day Trading: An Introduction, In: *INVESTOPEDIA A Forbes Digital Company: http://www.investopedia.com/articles/trading/05/011705.asp*, (04/17/2010)

Linnainmaa, J. T. (2003). The Anatomy of Day Traders, In: *American Finance Association 2004 San Diego Meetings*, Rene M. Stulz, (Ed.), AFA, University of California, Berkeley

Ĺström, K. J. (1970). *Introduction to Stochastic Control Theory*, Dover Publications, ISBN 0486445313, New York

Mooers, C. N. (1976). Technology of information handling: A pioneer's view. *Bulletin of the American Society for Information Science*, Vol. 2; No. 8, 18-19

Nasdaq (2003). *http://www.nasdaq.com/*, (from 07/18/2003 till 07/18/2006; 04/17/2010), The NASDAQ Stock Market - Official site of The NASDAQ Stock Market

Nielsen NetRatings (2005). *http://en-us.nielsen.com/tab/product_families/nielsen_netratings*, (04/17/2005; 04/17/2010), Online Measurement Services to Understand, Measure and Respond to Online Consumers

Perry, J. W. & Kent, A. (1958). *Tools for machine literature searching*, Interscience Publishers, New York

Perry, J. W.; Kent, A. & Berry, M. M. (1956). *Machine literature searching*, Western Reserve University Press and Interscience Publishers, Cleveland

Pham, H. (2009). *Continuous-time Stochastic Control and Optimization with Financial Applications*, Springer Berlin Heidelberg, ISBN 978-3-540-89499-5, Berlin Heidelberg

Schmidli, H. (2008). *Stochastic Control in Insurance*, Springer London, ISBN 978-1-84800-002-5, London

Schultz, C. K. & Luhn H. P. (1968). *Pioneer of information science - Selected works*, Spartan Books, New York

Simovic, V.; Radic, D. & Zrinusic, Z. (1998). Operational Model for Analysing and Visualisation of the Interesting and Suspicious Financial Transactions, *Proceedings of the 23rd Meeting of the Euro Working Group on Financial Modelling*, pp. 219-229, Kraków, Jun 1998, EWGFM, Kraków

Simovic, V. & Simovic, V. (2006). Specialized Day Trading - A New View of an Old Game, *Pre-Conference Proceedings of the Special Focus Symposium on 4th Catallactics: Quantitative-Behavioural Modelling Of Human Actions And Interactions On Markets*, pp. 49-56, ISBN 953-99326-3-7, Baden-Baden, Germany, August 2006, ECNSI, Zagreb

Stein, J. L. (2006). *Stochastic Optimal Control, International Finance, and Debt Crises*, Oxford University Press, ISBN 0-19-928057-6 978-0-19-928057-5, Oxford

Taube, M., & Wooster, H. (1958). *Information storage and retrieval: Theory, systems, and devices*, Columbia University Press, New York

The Sunday Times (2006). *http://www.timesonline.co.uk/tol/news/*, (02/26/2006; 04/17/2010), site of The Sunday Times

Tuđman, M.; Boras, D. & Dovedan, Z. (1993). *Introduction to information science, Second edition (in Croatian: Uvod u informacijske znanosti, drugo izdanje)*, Školska knjiga, Zagreb

Zadeh, L. A. (1996), Fuzzy logic, Neural Networks and Soft Computing, In: *Computational Intelligence: Soft Computing and Neuro-Fuzzy Integration with Applications*, Kaynak, O.; Zadeh, L. A.; Tuksen, B.; Rudas I. (Eds.), Springer Verlag, NATO ASI Series

Yahoo!® Finance (2010). *http://finance.yahoo.com*, (04/17/2010), site of The Yahoo! Inc.

Fractional bioeconomic systems: optimal control problems, theory and applications

Darya V. Filatova[a,b], Marek Grzywaczewski[a,c] and Nikolai P. Osmolovskii[c,d,e,f]

[a]Analytical Center RAS
Russian Federation
[b]Jan Kochanowski University in Kielce
Poland
[c]Politechnika Radomska
26-600 Radom, ul. Malczewskiego 20A, Poland
[d]Systems Research Institute,
Polish Academy of Sciences,
ul. Newelska 6, 01-447, Warszawa, Poland
[e]Akademia Podlaska,
ul. 3 Maja 54, 08-110 Siedlce, Poland
[f]This author was supported by grants
RFBR 08-01-00685 and NSh-3233.2008.1

1. Introduction

Exploitation of renewable resources is a task on a global scale inasmuch ecosystems are permanently destroyed by large-scale industrialization and unlimited human population growth. These have made already quit an impact on environment causing climatic destabilization. Thus, prediction of sustainable economic development has to take into account the bioeconomic principles. Although the task is not a new one there is a room for further investigations. It can be explained in the following manner.

It is known that biological systems react on the changes of existence conditions, environment actions and own states. Some of these systems are often utilized in forestry or fishery and therefore human control factor plays a very important role. In order to keep the completeness under uncertain environmental variability and internal transformations the considered biological systems must be in some dynamic equilibrium, which is defined by maximum sustainable yield approach, as a guarantee of the entire system existence. This idea requires removable resource management solved in some optimal sense.

Before the formulation of optimal control problem it is reasonable to notice that despite its popularity maximum sustainable yield (MSY) approach has some obstacles (Clark, 1989). Firstly, it is very sensitive to small errors in population data. Secondly, it does not take into account most of economic aspects of resource exploitation and, at last, it can be hardly used

in "species in interaction" cases. It is clear that the problem solution is strongly connected with a task of appropriate mathematical model selection (Jerry & Raissi, 2005; McDonald et al., 2002).

Initially bioeconomic models contained two main components: one defined dynamics of biological system and second characterized the economic policy of selected system exploitation (Clark, 1989). To make them more realistic different types of uncertainties have been incorporated. It was shown that three sources of uncertainty play an important role in fisheries management: variability in fish dynamics, inaccurate stock size estimates, and inaccurate implementation of harvest quotas, but there is not a unique way of how to include noises in models. To describe environmental noise one can use the following principles (Sethi et al., 2005):

- the variance is proportional to the expected population in the next generation;
- environmental fluctuations affect the population multiplicatively (this holds under a range of conditions - the density-independent or maximum growth rate of individuals are affected);
- demographic and environmental fluctuations can have long-range and/or short-range consequences on biological system.

The goal of this work is to show the ways of problem optimal solution when control object meets the principles mentioned above. The rest of the chapter is organized as follows. In Section 2, we formulate the optimal control problem for given tasks, showing how to convert the stochastic task into non-stochastic one. In Section 3, we derive necessary optimality conditions for short-range and long-range dependences, as it requires the object equation, under certain control and state constraints. Finally, Section 4 provides an application of obtained theoretical results to the problem of maximization of expected utility from the terminal wealth.

2. Fractional bioeconomic systems

2.1 A fishery management model

A renewable resource stock dynamics (or population growth) can be given as growth model of the type

$$dX(t) = s(t, X(t))dt ,\qquad (1)$$

with given initial condition $X(t_0) = X_0$. Here $X(t) \geq 0$ is the size of population at time t, $s(t, X(t))$ is the function, which describes population growth.

Model selection depends on the purpose of the modeling, characteristics of biological model and observed data (Drechsler et al., 2007). Usually one takes

$$s(t, X(t)) = \theta_1 X(t)$$

or

$$s(t, X(t)) = \theta_1 X(t)(1 - \theta_2 X(t)), \tag{2}$$

where $\theta_1 > 0$ is the intrinsic growth rate, $\frac{1}{\theta_2} > 0$ is the carrying capacity, $\theta_2 > 0$.

If in the first case population has an unlimited growth, in the second case we can also show that the population biomass $X(t)$ increases whenever $X(t) < \frac{1}{\theta_2}$, decreases for $X(t) > \frac{1}{\theta_2}$ and is in a state of equilibrium if $X(t) \to \frac{1}{\theta_2}$ as $t \to \infty$ (see Fig. 1).
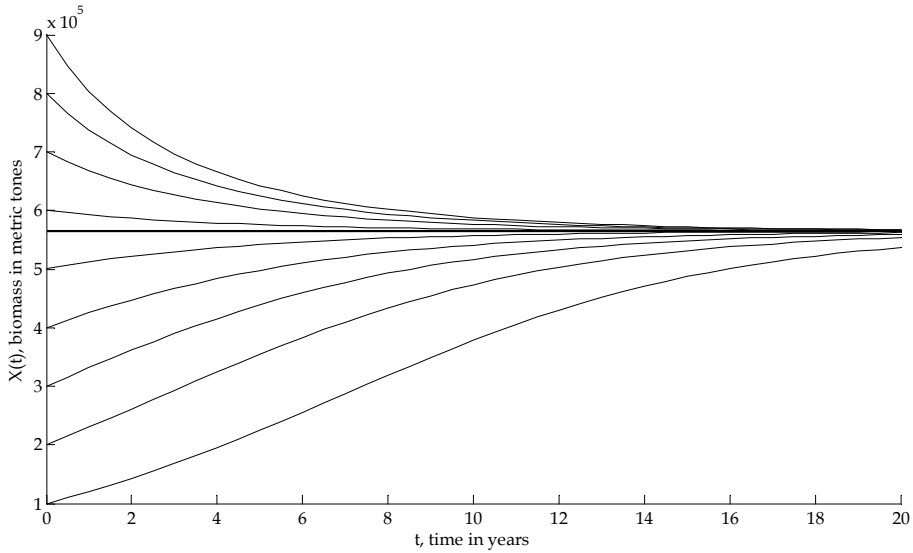
Fig. 1. Changes in population size $X(t)$ predicted by logistic growth function (2) for the southern bluefin tuna (McDonald et al., 2002)

Taking into account continuous harvesting at variable rate $u(t)$ the model (1) can be rewritten as

$$dX(t) = \left[s(t, X(t)) - u(t) \right] dt, \tag{3}$$

where the harvest rate has to be limited, for example

$$0 \le u(t) \le u_{\max}, \tag{4}$$

in order to guarantees the existence of the ecosystem under environmental variability and internal transformations (Edelstein-Keshet, 2005).

Assume that $u(t) = $ constant . In this case the dynamic equation (3) gives a picture of the logistic growth model behavior. So, for $u < \max\left[s(t, X(t))\right]$, the equation has one stable (point B on Fig. 2) and one unstable equilibrium (point A on Fig. 2). For $u > \max\left[s(t, X(t))\right]$, there is not any equilibrium state. If $u = \max\left[s(t, X(t))\right]$, the equation has only a single semistable equilibrium at the point called maximum sustainable yield (point C on Fig. 2). MSY is widely used for finding optimal rates of harvest, however and as it was mentioned before, there are problems with MSY approach (Kugarajh et al., 2006; Kulmala et al., 2008).
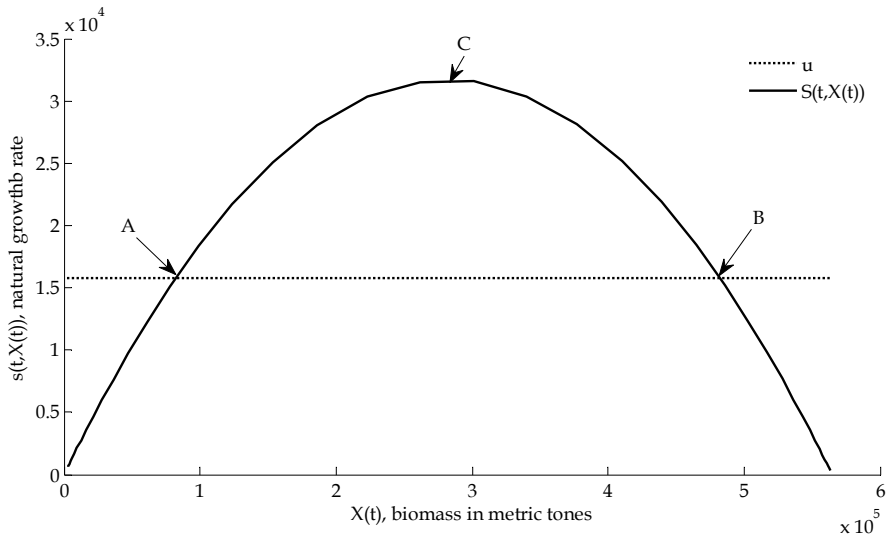
Fig. 2. Population dynamics with constant rate harvesting u for the southern bluefin tuna (McDonald et al., 2002)

To make the model more realistic one has to take into account different types of uncertainties introduced by diverse events as fires, pests, climate changes, government policies, stock prices etc. (Brannstrom & Sumpter, 2006). Very often these events might have long-range or short-range consequences on biological system. To take into account both types of consequences and to describe renewable resource stock dynamics it is reasonable to use stochastic differential equation (SDE) with fractional Brownian motion (fBm):

$$dX(t) = f\left(t, X(t), u(t)\right)dt + \sum_{i=1}^{n} q_i\left(t, X(t)\right)dB_t^{H_i} , \ X(t_0) = X_0 , \tag{5}$$

where $f\left(t, X(t), u(t)\right) := s\left(t, X(t)\right) - u(t)$ and $q_i\left(t, X(t)\right)$ are smooth functions, $dB_t^{H_i}$ are uncorrelated increments of fBm with the Hurst parameters $H_i \in (0,1)$ in the sense that

$$X(t) = X_0 + \int_{t_0}^{t} f(\tau, X(\tau), u(\tau)) d\tau + \sum_{i=1}^{k} \int_{t_0}^{t} q_i(\tau, X(\tau), u(\tau)) dB_\tau^{H_i}, \tag{6}$$

where second integral can be understand as a pathwise integral or as a stochastic Skorokhod integral with respect to the fBm.

An economical component of the bioeconomic model can be introduced as discounted value of utility function or production function, which may involve three types of input, namely labor $L(t)$, capital $C(t)$ and natural resources $X(t)$:

$$F(t, X(t), u(t)) = e^{-\rho t} \Pi(L^{\gamma_L}(t), C^{\gamma_C}(t), X^{\gamma}(t)), \tag{7}$$

where $\Pi(L^{\gamma_L}(t), C^{\gamma_C}(t), X^{\gamma}(t))$ is the multiplicative Cobb-Douglas function with γ_L, γ_C and γ constant of elasticity, which corresponds to the net revenue function at time t from having a resource stock of size $X(t)$ and harvest $u(t)$, ρ is the annual discount rate.

The model (7) was used in (Filatova & Grzywaczewski, 2009) for named task solution, other production function models can be found, for an example in (Kugarajh et al., 2006) or (Gonzalez-Olivares, 2005)):

$$F(t, X(t), u(t)) = e^{-\rho t} \Pi(C(t), X(t)) = e^{-\rho t} \left[p(t, u(t)) - c(t, X(t), u(t)) \right], \tag{8}$$

where $p(\cdot, \cdot)$ is the inverse demand function and $c(\cdot, \cdot, \cdot)$ is the cost function.

In both cases the objective of the management is to maximize the expected utility

$$J(X(\cdot), u(\cdot)) = \max_{u(t)} \mathsf{E} \left[\int_{t_0}^{t_1} F(t, X(t), u(t)) dt \right] \tag{9}$$

on time interval $\left[t_0, t_1 \right]$ **subject to** constraints (4) and (5), where $\mathsf{E}[\cdot]$ is mathematical expectation operator.

The problem (4), (5), (9) could be solved by means of maximum principle staying with the idea of MSY. There are several approaches, which allow find optimal harvest rate. First group operates in terms of stochastic control (Yong, 1999) and (Biagini et al., 2002), second one is based on converting the task (9) to non-random fractional optimal control (Jumarie, 2003). It is also possible to use system of moments equations instead of equation (5) as it was proposed in (Krishnarajaha et al., 2005) and (Lloyd, 2004). Unfortunately, there are some limitations, namely the redefinition of MSY for the model (5) and in a consequence finding an optimal harvest cannot be done by classical approaches (Bousquet et al., 2008) and numerical solution for stochastic control problems is highly complicated even for linear SDEs.

To overcome these obstacles we propose to combine the production functions (7) and (8) using $\mathsf{E}\left[X^{\gamma}(t)\right]$ instead of $\mathsf{E}\left[X(t)\right]$ in the function (8), specifically the goal function (9) takes a form

$$J(X(\cdot),u(\cdot)) = \max_{u(t)} \int_{t_0}^{t_1} F\left(t, \mathsf{E}\left[X^{\gamma}(t)\right], u(t)\right) dt \ , \tag{10}$$

where $\gamma \in (0,1]$.

If the coefficient of elasticity $\gamma = 1$, then the transformation to a non-random task gives a possibility to apply the classical maximum principle. If $0 < \gamma < 1$, then the cost function (8) contains a fractional term, which requires some additional transformations. This allows to introduce an analogue of MSY taking into account multiplicative environmental noises, as it was mentioned in *Introduction*, in the following manner

$$X^* = \max \mathsf{E}\left[X^{\gamma}(t)\right], \tag{11}$$

which can be treated as the state constraint.

Now the optimal harvest task can be summarized as follows. The goal is to maximize the utility function (10) subject to constraints (4), (5), and (11).

2.2 A background of dynamic fractional moment equations

To get an analytical expression for $\mathsf{E}\left[X^{\gamma}(t)\right]$ it is required to complete some transformations. The fractal terms complicate the classical way of the task solution and therefore some appropriate expansion of fractional order is required even if it gives an approximation of dynamic fractional moment equation. In the next reasoning we will use ideas of the fractional difference filters. The basic properties of the fractional Brownian motion can be summarized as follows (Shiryaev, 1998).

Definition. Let $(\Omega, \mathsf{F}, \mathsf{P})$ denotes a probability space and H, $0 < \mathsf{H} < 1$, referred to as the Hurst parameter. A centered Gaussian process $B^{\mathsf{H}} = \left\{B(t,\mathsf{H}), t \geq 0\right\}$ defined on this probability space is a fractional Brownian motion of order H if

$$\mathsf{P}\left\{B(0,\mathsf{H}) = 0\right\} = 1$$

and for any $t, \tau \in \mathsf{R}^+$

$$\mathsf{E}\left\{B(t,\mathsf{H})B(\tau,\mathsf{H})\right\} = \tfrac{1}{2}\left(t^{2\mathsf{H}} + \tau^{2\mathsf{H}} - \left|t - \tau\right|^{2\mathsf{H}}\right).$$

If $\mathsf{H} = \tfrac{1}{2}$, B^{H} is the ordinary Brownian motion.

There are several models of fractional Brownian motion. We will use Maruyama's notation for the model introduced in (Mandelbrot & Van Ness, 1968) in terms of Liouville fractional derivative of order H of Gaussian white noise. In this case, the fBm increment of (5) can be written as

$$dB_t^H = \omega(t)(dt)^H \tag{12}$$

where $\omega(t)$ is the Gaussian random variable.

Now the equation (5) takes a form

$$dX(t) = f\big(t, X(t), u(t)\big)dt + \sum_{p=1}^{n} q_p\big(t, X(t)\big)\omega_p(t)(dt)^{H_p} . \tag{13}$$

The results received in (Jumarie, 2007) allow to obtain the dynamical moments equations

$$m_k := \mathsf{E}\big\{X^k(t)\big\} \equiv \big\langle X^k(t)\big\rangle , \tag{14}$$

where $k \in N^*$.

Using the equality

$$X(t+dt) = X(t) + dX , \tag{15}$$

we get the following relation

$$X^k(t+dt) = X^k(t) + \sum_{j=1}^{k}\binom{k}{j}X^{k-j}(t)(dX)^j , \tag{16}$$

with

$$(dX)^j = \left(f\big(t, X(t), u(t)\big)dt + \sum_{i=1}^{n} q_i\big(t, X(t)\big)dB_i \right)^j ,$$

where $dB_i := dB^{H_i}$.

Taking the mathematical expectation of (16) yields the equality

$$m_k(t+dt) = m_k(t) + \sum_{j=1}^{k}\binom{k}{j}\mathsf{E}\big\{X^{k-j}(t)(dX(t))^j\big\}. \tag{17}$$

In order to obtain the explicit expression of (17) we suppose that random variables ω_i and ω_j are uncorrelated for any $i \neq j$ and denote $\eta(v) = \frac{1}{2n} v^{2\ell} = \frac{1}{2n} \left(\omega(t)(dt)^H \right)^{2\ell}$ for arbitrary integer ℓ. Application of the Ito formula gives

$$\frac{1}{2n} v_t^{2\ell} - \frac{1}{2n} v_0^{2\ell} = \int_0^t v_s^{2\ell-1} dv_s + \frac{1}{2} \int_0^t (2n-1) v_s^{2\ell-2} dv_s .$$

(18)

Taking expectation and solving (18) in iterative manner, we get the following results

$$\mathsf{E}\left(v_t^{2\ell}\right) = \frac{2\ell(2\ell-1)}{2} \int_0^t \mathsf{E}\left(v_s^{2\ell-2}\right) ds$$

$$= \frac{2\ell(2\ell-1)}{2} \cdot \frac{(2\ell-2)(2\ell-3)}{2} \int_0^t \int_0^{t_1} \mathsf{E}\left(v_s^{2\ell-4}\right) ds dt_1$$

$$\vdots$$

$$= \frac{(2\ell)!}{2^\ell} \int_0^t \int_0^{t_1} \dots \int_0^{t_{\ell-1}} 1 \, ds dt_{\ell-1} dt_1$$

Successive solution of this expression brings the sequence $t_{\ell-1}$, $\frac{1}{2!} t_{\ell-2}^2$, $\frac{1}{3!} t_{\ell-3}^3$,...., $\frac{1}{\ell!} t_0^\ell$ and gives the expression for even moments

$$\mathsf{E}\left[\left(\omega(t)(dt)^H \right)^{2\ell} \right] = \frac{(2\ell)!}{\ell! 2^\ell} (dt)^{2\ell H} .$$

The same can be done to get odd moments, namely

$$\mathsf{E}\left\{ \left(\omega(t)(dt)^H \right)^{2\ell+1} \right\} = 0 .$$

Now (17) can be presented in the following way:

$$m_k(t+dt) = m_k(t) + k \left\langle X^{k-1} dX \right\rangle + \frac{k(k-1)}{2} \left\langle X^{k-2} dX^2 \right\rangle + \mathsf{O}\left(dt^{1+\varepsilon} \right),$$

for $k \in N^*$ and $\varepsilon > 0$.

Let L denote the lag operator and γ be the fractional difference parameter. In this case the fractional difference filter $(1-\mathsf{L})^\gamma$ is defined by a hypergeometric function as follows (Tarasov, 2006)

$$(1-\mathsf{L})^\gamma = \sum_{k=0}^{\infty} \frac{\Gamma(k-\gamma)}{\left(\Gamma(-\gamma) \Gamma(k+1) \right)} \mathsf{L}^\gamma ,$$

(19)

where $\Gamma(\cdot)$ is the Gamma function.

Right hand-side of (19) can be also approximated by binominal expansion

$$(1-L)^\gamma \approx 1-\gamma L + \frac{\gamma(\gamma-1)}{2!}L^2 - \frac{\gamma(\gamma-1)(\gamma-2)}{3!}L^3 + \dots$$

This expansion allows to rewrite (17) and finally to get an approximation of dynamic fractional moment equation of order γ

$$dm_\gamma(t) = \gamma f\big(t, m_\gamma(t), u(t)\big)dt + \frac{\gamma(\gamma-1)}{2}\big(q\big(t, m_\gamma^{1/2}(t)\big)\big)^2 (dt)^{2H}, \tag{20}$$

where $m_\gamma(t_0) = E\big[X^\gamma(t_0)\big]$.

To illustrate the dynamic fractional moment equation (20) we will use the following SDE

$$dX(t) = \theta_1 X(t)\big(1-\theta_2 X(t)\big)dt + \theta_3 X(t) dB_t^H, \tag{21}$$

where $X(t_0) = 25000$, $\theta_1 = 0.2246$, $\theta_2 = \frac{1}{564795}$, $\theta_3 = 0.0002$ and $H = 0.5$.

Applying (20) to (21) and using a set of $\gamma \in \{0.25; 0.5; 0.75; 0.95; 1\}$, we can see possible changes in population size (Fig.3) and select the appropriate risk aversion coefficient γ.
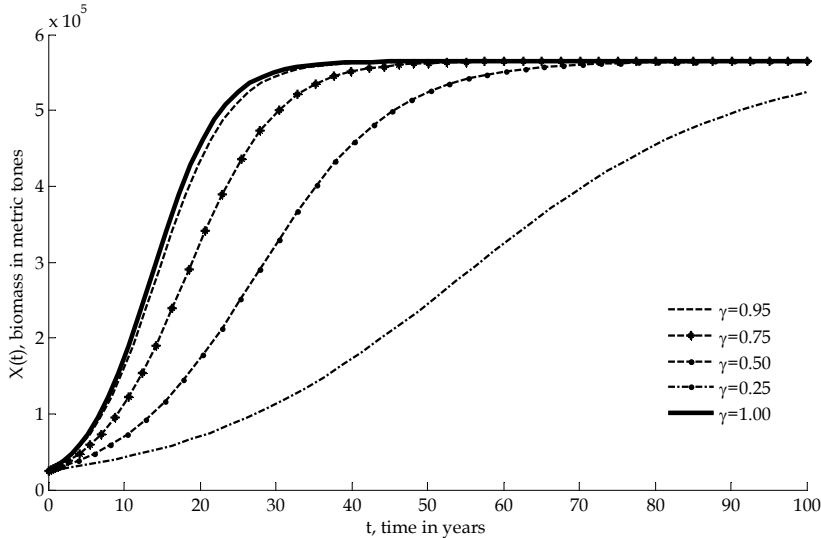

Fig. 3. The dynamic fractional moment equation (20) for equation (21)

2.3 Some required transformations

To get rid of fractional term $(dt)^{2H}$ and to obtain more convenient formulations of the results we replace ordinary fractional differential equation (20) by integral one

$$x(t) - x(t_0) = \int_{t_0}^{t} f(\tau, x(\tau), u(\tau)) d\tau$$

$$+ \int_{t_0}^{t} q(\tau, x(\tau))(d\tau)^{2H} , \qquad (22)$$

where $x(t) := m_\gamma(t)$, $x(t_0) := m_\gamma(t_0)$ for arbitrary selected γ.

Following reasoning is strongly dependent on H value as far as it changes the role of integration with respect to fractional term, namely as in (Jumarie, 2007), denoting the kernel by $\kappa(\tau)$, one has for $0 < H < \frac{1}{2}$

$$\int_{t_0}^{t} \kappa(\tau)(d\tau)^{2H} = 2H \int_{t_0}^{t} (t-\tau)^{2H-1} \kappa(\tau) d\tau , \qquad (23)$$

and for $\frac{1}{2} < H < 1$

$$\int_{t_0}^{t} \kappa(\tau)(d\tau)^{2H} = H^2 \left[\int_{t_0}^{t} (t-\tau)^{H-1} \kappa^{1/2}(\tau) d\tau \right]^2 . \qquad (24)$$

So, if $0 < H < \frac{1}{2}$, then the equation (22) can be rewritten as

$$x(t) - x(t_0) = \int_{t_0}^{t} f(\tau, x(\tau), u(\tau)) d\tau ,$$

$$+ 2H \int_{t_0}^{t} \frac{1}{(t-\tau)^{1-2H}} q(\tau, x(\tau)) d\tau \qquad (25)$$

for $\frac{1}{2} < H < 1$ equation (22) takes the form

$$x(t) - x(t_0) = \int_{t_0}^{t} f(\tau, x(\tau), u(\tau)) d\tau$$

$$+ \left[H \int_{t_0}^{t} \frac{\sqrt{q(\tau, x(\tau))}}{(t-\tau)^{1-H}} d\tau \right]^2 . \qquad (26)$$

3. Local maximum principle

3.1 Statement of the problem

Let the time interval $[t_0, t_1]$ be fixed, $x \in \mathsf{R}$ denote the state variable, and $u \in \mathsf{R}$ denote the control variable. The coast function has the form

$$J(x(\cdot), u(\cdot)) = \left\{ \int_{t_0}^{t_1} F((t, x(t), u(t)) dt + \varphi(x(t_1)) \right\} \to \max_{u(t)}, \tag{27}$$

where F and φ are smooth (C^1) functions, and is subjected to the constraints:

- the object equation (equality constraint)

$$x(t) = x(t_0) + \int_{t_0}^{t} f(\tau, x(\tau), u(\tau)) d\tau$$

$$+ \gamma(1-\gamma) \mathsf{H}_1 \int_{t_0}^{t} \frac{q_1(\tau, x(\tau))}{(t-\tau)^{1-2\mathsf{H}_1}} d\tau + \left[\mathsf{H}_2 \int_{t_0}^{t} \frac{\sqrt{\frac{1}{2}\gamma(1-\gamma)q_2(\tau, x(\tau))}}{(t-\tau)^{1-\mathsf{H}_2}} d\tau \right]^2, \tag{28}$$

where initial condition $x(t_0) = a > 0$ ($a \in \mathsf{R}$), $\mathsf{H}_1 \in (0, 0.5]$ and $\mathsf{H}_2 \in (0.5, 1.0)$,

- the control constraint (inequality constraint)

$$\phi(u(t)) \leq 0, \tag{29}$$

where $\phi(u)$ is a smooth (C^1) vector function of the dimension $\mathbf{p}$,

- the state constraint (inequality constraint)

$$\Phi(x(t)) \leq 0, \tag{30}$$

where $\Phi(x)$ is a smooth (C^1) function of the dimension $\mathbf{q}$.

Consider a more general system of integral equations than (28) with condition (30) (particularly $x(t) \geq 0$)

$$x(t) = \int_{t_0}^{t} f(\tau, x(\tau), u(\tau)) d\tau + \gamma(1-\gamma) \mathsf{H}_1 \int_{t_0}^{t} \frac{\theta_3 x(\tau)}{(t-\tau)^{1-2\mathsf{H}_1}} d\tau + G(y(t)), \tag{31}$$

$$y(t) = b + \int_{t_0}^{t} \frac{g(x(\tau))}{(t-\tau)^{1-\mathsf{H}_2}} d\tau, \tag{32}$$

where $x \in \mathsf{R}^n$, $y \in \mathsf{R}^m$, $u \in \mathsf{R}^r$, $b \in \mathsf{R}^m$, $g(x)$ and $G(x)$ are smooth (C^1) functions. In addition,

$$(t, x, y, u) \in \mathbf{Q}, \tag{33}$$

where $\mathbf{Q}$ is an open set.
So, we study problem (27), (29) - (33).

3.2. Derivation of the local maximum principle

Set $k := \gamma(1-\gamma)H_1\theta_3$, $\gamma_1 := 1 - 2H_1$, and $\gamma_2 := 1 - H_2$. Define a nonlinear operator

$$P : (x, y, u) \in C \times C \times L^\infty \to (z, \zeta) \in C \times C,$$

where

$$z(t) = x(t) - \int_{t_0}^t f(\tau, x(\tau), u(\tau)) d\tau - k \int_{t_0}^t \frac{x(\tau)}{(t-\tau)^{\gamma_1}} d\tau - G(y(t))$$

and

$$\zeta(t) = y(t) - b - \int_{t_0}^t \frac{g(x(\tau))}{(t-\tau)^{\gamma_2}} d\tau.$$

The equation $P(x, y, u) = 0$ is equivalent to the system (31) - (32). Let (x, y, u) be an admissible point in the problem. We assume that $\Phi(x(t_0)) < 0$ and $\Phi(x((t_1)) < 0$. The derivative of P at the point (x, y, u) is a linear operator

$$P'(x, y, u) : (\bar{x}, \bar{y}, \bar{u}) \to (\bar{z}, \bar{\zeta}),$$

where

$$
\begin{aligned}
\bar{z}(t) &= \bar{x}(t) - \int_{t_0}^t f_x(\tau, x(\tau), u(\tau)) \bar{x}(\tau) d\tau \\
&\quad - \int_{t_0}^t f_u(\tau, x(\tau), u(\tau)) \bar{u}(\tau) d\tau \\
&\quad - k \int_{t_0}^t \frac{\bar{x}(\tau)}{(t-\tau)^{\gamma_1}} d\tau - G'(y(t)) \bar{y}(t), \\
\bar{\zeta}(t) &= \bar{y}(t) - \int_{t_0}^t \frac{g'(x(\tau)) \bar{x}(\tau)}{(t-\tau)^{\gamma_2}} d\tau.
\end{aligned}
$$

Set $f_x(\tau) := f_x(\tau, x(\tau), u(\tau))$ $f_u(\tau) := f_u(\tau, x(\tau), u(\tau))$ etc. An arbitrary linear functional ℓ, vanishing on the kernel of the operator $P'(x,y,u)$, has the form

$$\ell(\overline{x},\overline{y},\overline{u}) = \int_{t_0}^{t_1} \overline{x}(t)\,d\sigma_1(t) -$$

$$- \int_{t_0}^{t_1}\left[\int_{t_0}^{t}\left(f_x(\tau)\overline{x}(\tau) + f_u(\tau)\overline{u}(\tau)\right)d\tau\right]d\sigma_1(t) -$$

$$- k\int_{t_0}^{t_1}\left[\int_{t_0}^{t}\frac{\overline{x}(\tau)}{(t-\tau)^{\gamma_1}}d\tau\right]d\sigma_1(t) - \int_{t_0}^{t_1}G'(y(t))\overline{y}(t)\,d\sigma_1(t) +$$

$$+ \int_{t_0}^{t_1}\overline{y}(t)\,d\sigma_2(t) - \int_{t_0}^{t_1}\left[\int_{t_0}^{t}\frac{g'(\tau)\overline{x}(\tau)}{(t-\tau)^{\gamma_2}}d\tau\right]d\sigma_2(t).$$

We change the order of integrating

$$\ell(\overline{x},\overline{y},\overline{u}) = \int_{t_0}^{t_1} \overline{x}(t)\,d\sigma_1(t) -$$

$$- \int_{t_0}^{t_1}\left[\int_{\tau}^{t_1}\left(f_x(\tau)\overline{x}(\tau) + f_u(\tau)\overline{u}(\tau)\right)d\sigma_1(t)\right]d\tau -$$

$$- k\int_{t_0}^{t_1}\left[\int_{\tau}^{t_1}\frac{\overline{x}(\tau)}{(t-\tau)^{\gamma_1}}d\sigma_1(t)\right]d\tau - \int_{t_0}^{t_1}G'(y(t))\overline{y}(t)\,d\sigma_1(t) +$$

$$+ \int_{t_0}^{t_1}\overline{y}(t)\,d\sigma_2(t) - \int_{t_0}^{t_1}\left[\int_{\tau}^{t_1}\frac{g'(\tau)\overline{x}(\tau)}{(t-\tau)^{\gamma_2}}d\sigma_2(t)\right]d\tau.$$

We now replace τ by t and t by τ and get

$$\ell(\overline{x},\overline{y},\overline{u}) = \int_{t_0}^{t_1} \overline{x}(t)\,d\sigma_1(t) -$$

$$- \int_{t_0}^{t_1}\left[\int_{t}^{t_1}\left(f_x(t)\overline{x}(t) + f_u(t)\overline{u}(t)\right)d\sigma_1(\tau)\right]dt -$$

$$- k\int_{t_0}^{t_1}\left[\int_{t}^{t_1}\frac{\overline{x}(t)}{(\tau-t)^{\gamma_1}}d\sigma_1(\tau)\right]dt - \int_{t_0}^{t_1}G'(y(t))\overline{y}(t)\,d\sigma_1(t) +$$

$$+ \int_{t_0}^{t_1}\overline{y}(t)\,d\sigma_2(t) - \int_{t_0}^{t_1}\left[\int_{t}^{t_1}\frac{g'(t)\overline{x}(t)}{(\tau-t)^{\gamma_2}}d\sigma_2(\tau)\right]dt$$

The Euler equation has the form

$$-\alpha_0 \int_{t_0}^{t_1} \left(F_x(t)\overline{x}(t) + F_u(t)\overline{u}(t) \right) dt$$

$$-\alpha_0 \varphi'\left(x(t_1)\right)\overline{x}(t_1)$$

$$+\ell\left(\overline{x},\overline{y},\overline{u}\right) + \left\langle \lambda,\phi'\left(u(\cdot)\right)\overline{u}(\cdot)\right\rangle$$

$$+\int_{t_0}^{t_1} \Phi'\left(x(t)\right)\overline{x}(t)d\mu(t) = 0 ,$$

where $\lambda \in (L^\infty)^*$, $\lambda \geq 0$, $\left\langle \lambda,\phi\left(u(\cdot)\right)\right\rangle = 0$, $d\mu \in C^*$, $d\mu \geq 0$, $\Phi\left(x(t)\right)d\mu(t) = 0$.

Note that the complementary slackness condition $\Phi\left(x(t)\right)d\mu(t) = 0$ combined with the assumptions $\Phi\left(x(t_0)\right) < 0$ and $\Phi\left(x(t_1)\right) < 0$ imply that the measure $d\mu$ is zero in some neighborhoods of the points t_0 and t_1.

Setting in the Euler equation $\overline{x} = 0$, $\overline{y} = 0$, we get

$$-\alpha_0 F_u(t) - f_u(t)\int_t^{t_1} d\sigma_1(\tau) + \lambda^a(t)\phi'\left(u(t)\right) = 0 , \tag{34}$$

where λ^a is an absolutely continuous part of λ. Hence $\lambda^a(\cdot) \geq 0$, $\lambda^a(\cdot)\phi\left(u(\cdot)\right) = 0$.
Setting $\overline{u} = 0$, $\overline{y} = 0$, we get

$$-\alpha_0 F_x(t)dt - \alpha_0\varphi'\left(x(t_1)\right)\delta\left(t - t_1\right)dt + d\sigma_1(t)$$

$$-f_x(t)\left[\int_t^{t_1} d\sigma_1(\tau)\right]dt - k\left[\int_t^{t_1} \frac{d\sigma_1(\tau)}{\left(\tau - t\right)^{\gamma_1}}\right]dt - \tag{35}$$

$$-g'\left(x(t)\right)\left[\int_t^{t_1} \frac{d\sigma_2(\tau)}{\left(\tau - t\right)^{\gamma_2}}\right]dt .$$

$$+\Phi'\left(x(t)\right)d\mu(t) = 0$$

Finally, setting $\overline{u} = 0$, $\overline{x} = 0$, we get

$$-G'\left(y(t)\right)d\sigma_1(t) + d\sigma_2(t) = 0 .$$

From equation (35) we get

$$d\sigma_1(t) = s_1(t)dt + \alpha_0\varphi'\big(x(t_1)\big)\delta(t-t_1)dt - \Phi'\big(x(t)\big)d\mu(t)$$

where $s_1 \in L^1$. Set

$$\psi(t) = \int_t^{t_1} d\sigma_1(\tau).$$

Then

$$\psi(t_1) = \alpha_0\varphi'\big(x(t_1)\big),$$
$$d\sigma_1(t) = -d\psi(t) + \psi(t_1)\delta(t-t_1)dt,$$
$$d\sigma_2(t) = -G'\big(y(t)\big)d\psi(t) + \psi(t_1)G'\big(y(t)\big)\delta(t-t_1)dt.$$

Consequently, we have

$$\int_t^{t_1} \frac{d\sigma_1(\tau)}{(\tau-t)^{\gamma_1}} = -\int_t^{t_1} \frac{d\psi(\tau)}{(\tau-t)^{\gamma_1}} + \psi(t_1)\int_t^{t_1} \frac{\delta(\tau-t_1)d\tau}{(\tau-t)^{\gamma_1}} =$$
$$= -\int_t^{t_1} \frac{d\psi(\tau)}{(\tau-t)^{\gamma_1}} + \frac{\psi(t_1)}{(t_1-t)^{\gamma_1}},$$
$$\int_t^{t_1} \frac{d\sigma_2(\tau)}{(\tau-t)^{\gamma_2}} = -\int_t^{t_1} \frac{G'\big(y(\tau)\big)d\psi(\tau)}{(\tau-t)^{\gamma_2}} + \psi(t_1)\int_t^{t_1} \frac{G'\big(y(\tau)\big)\delta(\tau-t_1)d\tau}{(\tau-t)^{\gamma_2}} =$$
$$= -\int_t^{t_1} \frac{G'\big(y(\tau)\big)d\psi(\tau)}{(\tau-t)^{\gamma_2}} + \psi(t_1)\frac{G'\big(y(t_1)\big)}{(t_1-t)^{\gamma_2}}.$$

Therefore, relation (34) implies the following local maximum principle:

$$-\alpha_0 F_u(t) - f_u(t)\psi(t) + \lambda^a(t)\phi'(u(t)) = 0 \qquad (36)$$

and equation (35) leads to the following adjoint equation

$$-\alpha_0 F_x(t)dt - d\psi(t) - f_x(t)\psi(t)dt \qquad (37)$$
$$+k\left[\int_t^{t_1} \frac{d\psi(\tau)}{(\tau-t)^{\gamma_1}} - \frac{\psi(t_1)}{(t_1-t)^{\gamma_1}}\right]dt$$
$$+g'(x(t))\int_t^{t_1} \frac{G'(y(\tau))}{(\tau-t)^{\gamma_2}}d\psi(\tau) - g'(x(t))\psi(t_1)\frac{G'(y(t_1))}{(t_1-t)^{\gamma_2}} + \Phi'(x(t))d\mu(t) = 0.$$

Thus the following theorem is proved.

Theorem. *Let* $\left(x(t),y(t),u(t)\right)$ *be {the} [an] optimal process on the interval* $\left[t_0,t_1\right]$, *where* $x(\cdot)\in C\left(\left[t_0,t_1\right],\mathsf{R}^n\right)$, $y(\cdot)\in C\left(\left[t_0,t_1\right],\mathsf{R}^m\right)$, $u(\cdot)\in L^\infty\left(\left[t_0,t_1\right],\mathsf{R}^r\right)$. *Then there exists a set of Lagrange multipliers* $\left(\alpha_0,\psi(\cdot),\lambda(\cdot),\mu\right)$ *such that* α_0 *is a scalar,* $\psi(\cdot):\left[t_0,t_1\right]\to\mathsf{R}^n$ *is a function of bounded variation continuous from the left, defining the measure* $d\psi$, $\lambda(\cdot):\left[t_0,t_1\right]\to\mathsf{R}^{r*}$ *is an integrable function,* $\mu(\cdot):\left[t_0,t_1\right]\to\mathsf{R}$ *is a function of bounded variation continuous from the left, defining the measure* $d\mu$, *and the following conditions are fulfilled:*

(a) nonnegativity: $\alpha_0\geq 0$, $\lambda(t)\geq 0$ *a.e. on* $\left[t_0,t_1\right]$, $d\mu\geq 0$;
(b) nontriviality:

$$\alpha_0+\left\|\psi\right\|+\left\|d\mu\right\|>0\,;$$

(c) complementarity:

$$\lambda(t)\phi\left(u(t)\right)=0\quad\text{a.e. on}\quad\left[t_0,t_1\right]\,,$$
$$\Phi\left(x(t)\right)d\mu(t)=0\,;$$

(d) adjoint equation:

$$
\begin{aligned}
-d\psi(t) \;=\;& \psi(t)f_x(t)dt-k\left[\int_t^{t_1}\frac{d\psi(\tau)}{\left(\tau-t\right)^{\gamma_1}}-\frac{\psi(t_1)}{\left(t_1-t\right)^{\gamma_1}}\right]dt\\
&-\left[\int_t^{t_1}\frac{G'\left(y(\tau)\right)d\psi(\tau)}{\left(\tau-t\right)^{\gamma_2}}-\frac{\psi(t_1)G'\left(y(t_1)\right)}{\left(t_1-t\right)^{\gamma_2}}\right]g'\left(x(t)\right)dt+\\
&+\alpha_0 F_x(t,x(t),u(t))dt-\Phi'(x(t))d\mu(t)\,,
\end{aligned}
$$

[where $k:=\gamma(1-\gamma)\mathsf{H}_1\theta_3$, $\gamma_1:=1-2\mathsf{H}_1$, *and* $\gamma_2:=1-\mathsf{H}_2$;]

(e) transversality condition:

$$\psi(t_1)=\alpha_0\varphi'\left(x(t_1)\right);$$

(f) local maximum principle:

$$\psi(t)f_u\left(t,x(t),u(t)\right)+\alpha_0 F_u\left(t,x(t),u(t)\right)-\lambda(t)\phi'\left(u(t)\right)=0\,.$$

4. Example

In this section we will illustrate the theoretical results to get optimal control for the North-East Arctic Cod Fishery, using partly the data presented in (Kugarajh et al., 2006), by means

of the expected utility from terminal wealth maximization and without paying attention on economics and biological aspects of the problem.

Figure 4 shows the biomass time series made by individual vessels. In order to introduce the model for the data description we found the parameters of fBm, using methodology presented in (Filatova, 2008). There was only one significant parameter $H = 0.4501$ (with standard deviation 0.0073), which allowed to select a model of the biomass population, namely

$$dX_t = \theta_1\left(X_t - \theta_2 X_t^2\right)dt + \theta_3 X_t dB_t^H \,, \tag{38}$$
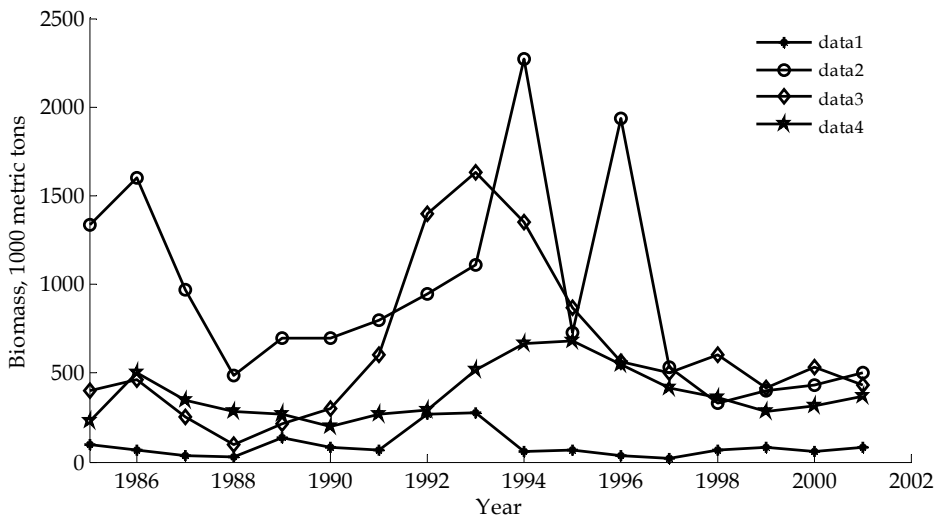
where $H \in (0, 0.5]$ and $X_0 = X(t_0)$.

Fig. 4. The North-East Arctic cod biomass for the years 1985 – 2001.

Next to find estimates of (38) we used ideas of identification methods (Filatova & Grzywaczewski, 2007; Filatova et al., 2007) and got

$$dX_t = 0.6416\left(X_t - \tfrac{1}{1567700.1215} X_t^2\right)dt + 0.0031 X_t dB_t^{0.4501} \,, \tag{39}$$

where initial value $X_0 = 500 \cdot 10^3$.

Applying the goodness-of-fit test for received SDE model (this test can be found in (Allen, 2007)) we calculated for $M = 18$ simulations the test statistics $Q = 5.0912$. Since three parameters were estimated on initial data stock, the number of degree of freedom is $M - 3 = 15$ and the critical value of $\chi^2(0.05;15) = 24.9958$. The probability $p\left(\chi^2(0.05;15) \geq 5.0912\right) < 0.5491$ is greater than the level of significance 0.05. That is, we

cannot reject that SDE model (39) describes the biomass dynamics. Thus, we can use the methodology proposed in this work in order to find the optimal strategy.

The model (39) can be used for the forecast of biomass dynamics (see Fig.5). Since the data had a significant variation, it is reasonable to take $\gamma = 0.8$.
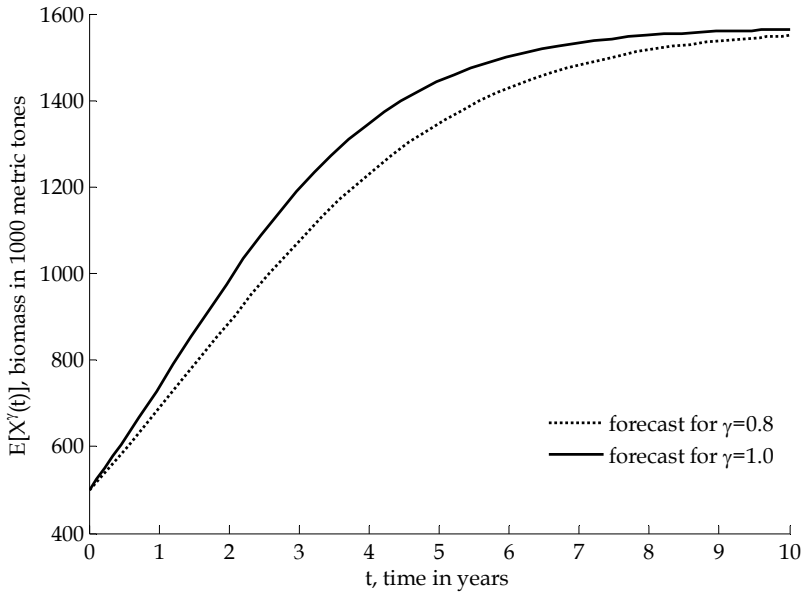

Fig. 5. The 10 years forecast for the North-East Arctic cod biomass for model (39).

Setting $x(t) := \mathsf{E}\left[X^{\gamma}(t) \right]$ and applying transformation (20) we get the object equation (28) in the following form

$$dx_t = \gamma \left[\theta_1 \left(x_t - \theta_2 x_t^2 \right) - u_t \right] dt + \frac{\gamma(\gamma - 1)}{2} \theta_3 x_t dt^{\mathsf{H}},$$ (40)

where $x_0 = \mathsf{E}\left[X^{\gamma}(t_0) \right]$.

Next one can set the constraints (29) and (30) as for an example

$$0 < u(t) < 150 \cdot 10^3,$$ (41)

$$0 < x(t) < 251.44 \cdot 10^3.$$ (42)

Using the integration role (23) the ordinary fractional differential equation (40) can be rewritten as

$$x(t) - x(t_0) = \int_{t_0}^{t} \gamma \Big[\theta_1 \big(x(\tau) - \theta_2 x^2(\tau) \big) - u(\tau) \Big] d\tau$$

$$+ \frac{\gamma(\gamma - 1)}{2} \int_{t_0}^{t} \frac{\beta}{(t-\tau)^{1-\beta}} \theta_3 x(\tau) d\tau , \tag{43}$$

where $\beta = 2H$.

Next we define the goal function (27) with the production function (8)

$$J\big(x(\cdot), u(\cdot)\big) = \left\{ \int_{t_0}^{t_1} e^{-\rho t} \left[au(t) - bu^2(t) - \frac{c}{x(t)} \right] dt + e^{-\rho t_1} \varphi\big(x(t_1)\big) \right\} \to \max_{u(t)} , \tag{44}$$

where $a = 88.25$, $b = 0.0009$ and $c = 1.633 \cdot 10^{11}$ are the parameters of production function (8).

Using (43) and (44) we obtain the adjoint equation (37)

$$-d\psi(t) = \psi(t) \gamma \theta_1 \big(1 - 2\theta_2 x(t) \big) dt + \beta \frac{\gamma(\gamma-1)}{2} \theta_3 \left[\frac{\psi(t_1)}{(t_1-t)^{1-\beta}} - \int_t^{t_1} \frac{\psi'(\tau)}{(\tau-t)^{1-\beta}} d\tau \right] dt$$

$$+ \alpha_0 e^{-\rho t} \frac{c}{x^2(t)} dt - \Phi'\big(x(t)\big) d\mu(t). \tag{45}$$

Finally the local maximum principle (36) is

$$\alpha_0 e^{-\rho t} \Big[a - 2bu(t) \Big] - \gamma \psi(t) + \lambda^a(t) \varphi'(u) = 0 . \tag{46}$$

On the basis of (44) the optimal control function can be defined as

$$u(t) = \frac{1}{2b} \Big(a - \gamma \psi(t) e^{\rho t} + \lambda^a(t) \varphi'(u) e^{\rho t} \Big). \tag{47}$$

Substitution of (47) to (43) gives

$$x(t) - x(t_0) = \int_{t_0}^{t} \gamma \Big[\theta_1 \big(x(\tau) - \theta_2 x^2(\tau) \big) - \frac{1}{2b} \big(a + \gamma \psi(\tau) e^{\rho \tau} + \lambda^a(t) \varphi'(u) e^{\rho t} \big) \Big] d\tau$$

$$+ \frac{\gamma(\gamma-1)}{2} \int_{t_0}^{t} \frac{\beta}{(t-\tau)^{1-\beta}} \theta_3 x(\tau) d\tau. \tag{48}$$

Solution of the system (45), (48) allows to define the solution of adjoint equation $\psi(t)$, optimal control $u(t)$ and as result the expected utility from terminal wealth (44). The ideas

of numerical algorithm for the system (45), (48) are presented in (Filatova et al., 2010), that gives following optimal control (see Fig. 6).
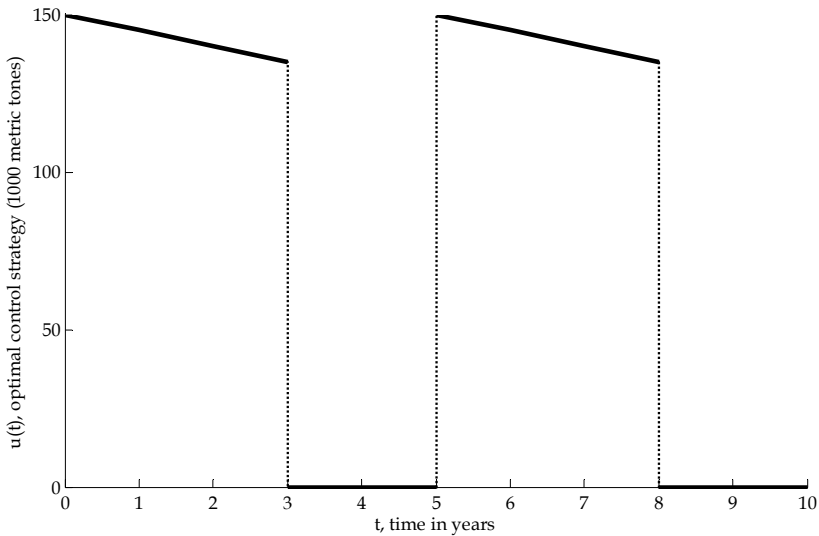

Fig. 6. The optimal control strategy for ten years period for the North-East Arctic cod.

5. Conclusion

In this work we studied stochastic harvest problem, where the biomass dynamics was described by stochastic logarithmic growth model with fractional Brownian motion. Since the data used for the fishery management are not accurate, to maintain existing of the population we proposed to use the risk aversion coefficient for fish stock and added not only control but also state constraints.

This formulation of optimal harvest problem could not be solved by classical methods and required some additional transformations. We used fractional filtration and got the integral object equation, which did not contain stochastic term. As a result stochastic optimization problem was changed to non-random one. Using maximum principle we got necessary optimality conditions, which were used for numerical solution of the North-East Arctic cod fishery problem to set suitable harvest levels.

We hope that to improve the quality of proposed methodology time-varying parameters model can be used as a control object. This requires new parametric identification method from one side and better understanding of economics and biological development of the exploitable ecosystem from the other one.

6. References

Allen, E. (2007). *Modeling with Ito stochastic differential equations*, Springer, ISBN 978-1-4020-5952-0, Dordrecht.

Alvarez, L.H.R. & Koskela, E. (2007). Optimal harvesting under resource stock and price uncertainty. *Journal of Economic Dynamics and Control*, 31, 2461 – 2485, ISSN 0165-1889

Biagini, F.; Hu, Y.; Øksendal, B. & Sulem, A. (2002). A stochastic maximum principle for the processes driven by fractional Brownian motion. *Stochastic processes and their applications*, 100, 233 – 253, ISSN 0304-4149

Biagini, F.; Øksendal, B.; Sulem, A. & Wallner, N. (2004). An introduction to white noise theory and Malliavin calculus for fractional Brownian motion. *The Proceedings of the Royal Society A*, 460, 347–372, ISSN 1364 – 5021.

Bousquet, N.; Duchesne, T. & Rivest, L.-P. (2008). Redefining the maximum sustainable yield for the Schaefer population model including multiplicative environmental noise. *Journal of Theoretical Biology*, 254, 65 – 75, ISSN 0022-5193

Brannstrom, A. & Sumpter, D.J.T (2006). Stochastic analogues of deterministic single-species population models. *Theoretical Population Biology*, 69, 442 – 451, ISSN 0040-5809

Clark, C.W. (1989). Bioeconomic modeling and resource management. In: *Applied mathematical ecology*, Levin, S.A., Hallam, T.G. & Gross, L.J. (Ed), 11 – 57, Springer-Verlag, ISBN 3-540-19465-7, New York.

Gasca-Leyva, E.; Hernandez, J.M. & Veliov, V.M. (2008). Optimal harvesting time in a size-heterogeneous population. *Ecological Modelling*, 210, 161 – 168, ISSN 0304-3800

Gonzalez-Olivares, E.; Saez E.; Stange, E. & Szanto, I. (2005). Topological description of a non-differentiable bioeconomic models. *Rocky Mountain Journal of Mathematics*, 35, 4, 1133 – 1145, ISSN 0035-7596

Drechsler, M.; Grimm, V.; Mysiak, J. & Watzold, F. (2007). Differences and similarities between ecological and economic models for biodiversity conservation. *Ecological Economics*, 62, 232 – 241, ISSN 0921-8009

Edelstein-Keshet, L. (2005). *Mathematical Models in Bioligy*, SIAM, ISBN 978-0-89871-554-5, New York

Filatova, D. & Grzywaczewski, M. (2007). Nonparametric identification methods of stochastic differential equation with fractional Brownian motion. *JAMRIS*, 1, 2, June 2007, 45 – 49, ISSN 1897-8649

Filatova, D.; Grzywaczewski, M.; Shybanova, E. & Zili, M. (2007). Parameter Estimation in Stochastic Differential Equation Driven by Fractional Brownian Motion, *Proceedings of EUROCON 2007: The International Conference on "Computer as a Tool"*, pp. 2316 – 2321, ISBN 978-3-901608-35-3, Poland, September 2007, IEEE TPress, Warsaw.

Filatova, D. (2008). Mixed fractional Brownian motion: some related questions for computer network traffic modeling, *Proceedings of International Conference on Signal and Electronic Systems*, pp. 532 – 538, ISBN 978-83-88309-47-2, Poland, September 2008, IEEE TPress, Krakow

Filatova, D. & Grzywaczewski, M. (2009). Necessary optimality conditions for fractional bioeconomic systems: the problem of optimal harvest rate, In: *Advances in Intelligent and Soft Computing*, 557 – 567, Springer, ISBN. 978-3-642-03201-1, Berlin

Filatova, D.; Grzywaczewski, M. & Osmolovski, N. (2010). Optimal control problem with an integral equation as the control object. *Nonlinear analysis*, 72, February 2010, 1235 – 1246, ISSN 1468-1218

Hosking, J.R.M. (1981). Fractional differencing. *Biometrica*, 68, 1, 165-176, ISSN 1755-8301

Jerry, M. & Raissi, N. (2005). Optimal strategy for structured model of fishing problem. *Comptes Rendus Biologies*, 328, 351 – 356, ISSN 1631-0691

Jumarie, G. (2003). Stochastics of order n in biological systems applications to population dynamics, thermodynamics, nonequilibrium phase and complexity. *Journal of Biological Systems*, 11, 2, 113-137, ISSN 1793-6470

Jumarie, G. (2007). Lagrange mechanics of fractional order, Hamilton-Jacobi fractional PDE and Taylor's series of nondifferentiable functions. *Chaos, solutions and fractals*, 32, 969 – 987, ISSN 0960-0779

Krishnarajaha, I.; Cooka, A.; Marionb, G. & Gibsona, G. (2005). Novel moment closure approximations in stochastic epidemics. *Bulletin of Mathematical Biology*, 67, 855–873, ISSN 1522-9602

Kugarajh, K.; Sandal, L. & Berge, G. (2006). Implementing a stochastic bioeconomic model for the North-East Arctic cod fishery. *Journal of Bioeconomics*, 8, 75 – 87, ISSN 1387-6989

Kulmala, S.; Laukkanen, M. & Michielsens, C. (2008). Reconciling economic and biological modeling of migratory fish stock: Optimal management of the Atlantic salmon fishery in the Baltic Sea. *Ecological Economics*, 64, 716 – 728, ISSN 0921-8009

Lloyd, A.L. (2004). Estimating variability in models for recurrent epidemics: assessing the use of moment closure techniques. *Theoretical Population Biology*, 65, 49–65, ISSN 0040-5809

Loisel, P. & Cartigny, P. (2009). How to model marine reserves? *Nonlinear Analysis*, 10, 1784 – 1796, ISSN 1468-1218

Mandelbrot, B.B. & van Ness, J.W. (1968). Fractional Brownian motions, fractional noises and applications. SIAM Review, 10, 422-437, ISSN 1095-7200

McDonald, A.D.; Sandal, L.K. & Steinshamn, S.I. (2002). Implications of a nested stochastic/deterministic bio-economic model for a pelagic fishery. *Ecological Modelling*, 149, 193-201, ISSN 0304-3800

Milyutin, A.A.; Dmitruk, A.V, & Osmolovskii, N.P. (2004). *Maximum principle in optimal control*, MGU, ISBN , Moscow

Nostbakken, L. (2006). Regime switching in a fishery with stochastic stock and price. *Environmental Economics and Management*, 51, 231 – 241, ISSN 0095-0696

Shiryaev, A.N. (1998). *Essentials of Stochastic Finance: Facts, Models, Theory*. FAZIS, ISBN 5-7036-0043-X, Moscow

Sethi, G.; Costello, C.; Fisher, A.; Hanemann, M. & Karp, L. (2005). Fishery management under multiple uncertainity. *Journal of Environmental Economics and Management*, 50, 300 – 318, ISSN 0095-0696

Tarasov, V.E. (2006). Liouville and Bogoliubov equations with fractional derivatives. *Modern physics letters*, B21, 237 – 248, ISSN 1529-7853.

Yong, J. & Zhou, X.Y. (1999). *Stochastic Control: Hamiltonian Systems and HJB Equations*. Springer, ISBN 0-387-98723-1, Berlin.