# ACADEMY OF ROMANIAN SCIENTISTS

# ANNALS

## SERIES ON MATHEMATICS AND ITS APPLICATIONS

**VOLUME 6** **2014** **NUMBER 2**

ISSN 2066 – 6594

TOPICS:

- **ORDINARY AND PARTIAL DIFFERENTIAL EQUATIONS**
- **OPTIMIZATION, OPTIMAL CONTROL AND DESIGN**
- **NUMERICAL ANALYSIS AND SCIENTIFIC COMPUTING**
- **ALGEBRAIC, TOPOLOGICAL AND DIFFERENTIAL STRUCTURES**
- **PROBABILITY AND STATISTICS**
- **ALGEBRAIC AND DIFFERENTIAL GEOMETRY**
- **MATHEMATICAL MODELLING IN MECHANICS ENGINEERING SCIENCES**
- **MATHEMATICAL ECONOMY AND GAME THEORY**
- **MATHEMATICAL PHYSICS AND APPLICATIONS**

EDITURA

ACADEMIEI OAMENILOR DE ȘTIINȚĂ DIN ROMÂNIA

**Annals of the Academy of Romanian Scientists**

**Series on Mathematics and its Applications**

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 6, No. 2/2014

# CONTENTS

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 6, No. 2/2014

# QUANTUM AND CLASSICAL CORRELATIONS IN GAUSSIAN OPEN QUANTUM SYSTEMS*

Aurelian Isar†

**Abstract**

In the framework of the theory of open systems based on completely positive quantum dynamical semigroups, we give a description of the continuous-variable quantum correlations (quantum entanglement and quantum discord) for a system consisting of two noninteracting bosonic modes embedded in a thermal environment. We solve the Kossakowski-Lindblad master equation for the time evolution of the considered system and describe the entanglement and discord in terms of the covariance matrix for Gaussian input states. For all values of the temperature of the thermal reservoir, an initial separable Gaussian state remains separable for all times. We study the time evolution of logarithmic negativity, which characterizes the degree of entanglement, and show that in the case of an entangled initial squeezed thermal state, entanglement suppression takes place for all temperatures of the environment, including zero temperature. We analyze the time evolution of the Gaussian quantum discord, which is a measure of all quantum correlations in the bipartite state, including entanglement, and show that it decays asymptotically in time under the effect of the thermal bath. This is in contrast with the sudden death of entanglement. Before the suppression of the entanglement, the qualitative evolution of quantum discord is very similar to that of the entanglement. We describe also the time evolution of the degree of

*Accepted for publication in revised form on June 5-th, 2013

†isar@theory.nipne.ro, National Institute of Physics and Nuclear Engineering, P.O.Box MG-6, Bucharest-Magurele, Romania; Academy of Romanian Scientists, 54 Splaiul Independentei, Bucharest 050094, Romania

classical correlations and of quantum mutual information, which measures the total correlations of the quantum system.

**MSC**: 81P40, 81R30, 81S22

**keywords:** quantum correlations, entanglement, discord, squeezed states, open systems, master equations.

## 1 Introduction

The study of quantum correlations is a key issue in quantum information theory [1] and quantum entanglement represents the indispensable physical resource for the description and performance of quantum information processing tasks, like quantum teleportation, cryptography, superdense coding and quantum computation [2]. However, entanglement does not describe all the non-classical properties of quantum correlations. Recent theoretical and experimental results indicate that some non-entangled mixed states can improve performance in some quantum computing tasks [3]. Zurek [4, 5] defined the quantum discord as a measure of quantum correlations which includes entanglement of bipartite systems and it can also exist in separable states. The total amount of correlations contained in a quantum state is given by the quantum mutual information which is equal to the sum of the quantum discord and classical correlations [6].

In recent years there is an increasing interest in using non-classical entangled states of continuous variable systems in applications of quantum information processing, communication and computation [7]. In this respect, Gaussian states, in particular two-mode Gaussian states, play a key role since they can be easily created and controlled experimentally. Due to the unavoidable interaction with the environment, in order to describe realistically quantum information processes it is necessary to take decoherence and dissipation into consideration. Decoherence and dynamics of quantum entanglement in continuous variable open systems have been intensively studied in the last years [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21].

In this review paper we describe, in the framework of the theory of open systems based on completely positive quantum dynamical semigroups, the dynamics of continuous variable quantum entanglement and quantum discord of a subsystem consisting of two uncoupled bosonic modes (harmonic oscillators) interacting with a common thermal environment. We are interested in discussing the correlation effect of the environment, therefore we assume that the two modes are independent, i.e. they do not interact directly. The initial state of the open system is taken of Gaussian form and the evolution under the quantum dynamical semigroup assures the preservation in time of the Gaussian form of the state. In particular, we consider unimodal squeezed states, squeezed vacuum states, and symmetric and non-symmetric squeezed thermal states as initial states [22, 23, 24]. We show that entanglement suppression (entanglement sudden death) takes place for all temperatures of the environment, including zero temperature. We analyze the time evolution of Gaussian quantum discord, which is a measure of all quantum correlations in the bipartite state, including entanglement, and show that discord decays asymptotically in time under the effect of the thermal bath. This is contrast with the sudden death of entanglement. Before the suppression of the entanglement, the qualitative evolution of quantum discord is very similar to that of the entanglement.

The paper is organized as follows. In Sect. 2 the notion of the quantum dynamical semigroup is defined using the concept of a completely positive map. Then we give the general form of the Kossakowski-Lindblad quantum mechanical master equation describing the evolution of open quantum systems in the Markovian approximation. We mention the role of complete positivity in connection with the quantum entanglement of systems interacting with an external environment. In Sec. 3 we write the equations of motion in the Heisenberg picture for two independent bosonic modes interacting with a general environment and give the general solution of the evolution equation for the covariance matrix, i.e. we derive the variances and covariances of coordinates and momenta corresponding to a generic two-mode Gaussian state. Then, by using the Peres-Simon necessary and sufficient condition for separability of two-mode Gaussian states [25, 26], we investigate in Sec. 4 the dynamics of quantum correlations (quantum entanglement and Gaussian quantum discord) for the considered subsystem. We describe also the time evolution of the degree of classical correlations and of quantum mutual information. A summary and conclusions are given in Sec. 5. In Appendix we present some elementary notions and examples of quantum correlations (entanglement) in quantum information theory, and describe the influence of diffusion and dissipation on the dynamics of a harmonic oscillator interacting with an environment, in particular with a thermal bath.

## 2 Axiomatic theory of open quantum systems

The time evolution of a closed physical system is given by a dynamical group $U_t$, uniquely determined by its generator $H$, which is the Hamiltonian operator of the system. The action of the dynamical group $U_t$ on any density matrix $\rho$ from the set $\mathcal{D}(\mathcal{H})$ of all density matrices in the Hilbert space $\mathcal{H}$ of the quantum system is defined by

$$\rho(t) = U_t(\rho) = e^{-\frac{i}{\hbar}Ht}\rho e^{\frac{i}{\hbar}Ht} \tag{1}$$

for all $t \in (-\infty, \infty)$. According to von Neumann, density operators $\rho \in \mathcal{D}(\mathcal{H})$ are trace class (Tr $\rho < \infty$), self-adjoint ($\rho^\dagger = \rho$), positive ($\rho > 0$) operators with Tr $\rho = 1$. All these properties are conserved by the time evolution defined by $U_t$.

In the case of open quantum systems, the time evolution $\Phi_t$ of the density operator $\rho(t) = \Phi_t(\rho)$ has to preserve the von Neumann conditions for all times. It follows that $\Phi_t$ must have the following properties:

(i) $\Phi_{\mathrm{t}}(\lambda_1\rho_1 + \lambda_2\rho_2) = \lambda_1\Phi_{\mathrm{t}}(\rho_1) + \lambda_2\Phi_{\mathrm{t}}(\rho_2)$ for $\lambda_1, \lambda_2 \geq 0$, $\lambda_1 + \lambda_2 = 1$, i. e. $\Phi_t$ must preserve the convex structure of $\mathcal{D}(\mathcal{H})$,

(ii) $\Phi_{\mathrm{t}}(\rho^\dagger) = \Phi_{\mathrm{t}}^\dagger(\rho)$,

(iii) $\Phi_{\mathrm{t}}(\rho) > 0$,

(iv) Tr $\Phi_{\mathrm{t}}(\rho) = 1$.

The time evolution $U_t$ for closed systems must be a group $U_{t+s} = U_t U_s$. We have also $U_0(\rho) = \rho$ and $U_t(\rho) \to \rho$ in the trace norm when $t \to 0$. The dual group $\widetilde{U}_t$ acting on the observables $A \in \mathcal{B}(\mathcal{H})$, i.e. on the bounded operators on $\mathcal{H}$, is given by

$$\widetilde{U}_t(A) = e^{\frac{i}{\hbar}Ht} A e^{-\frac{i}{\hbar}Ht}. \tag{2}$$

Then $\widetilde{U}_t(AB) = \widetilde{U}_t(A)\widetilde{U}_t(B)$ and $\widetilde{U}_t(I) = I$, where $I$ is the identity operator on $\mathcal{H}$. Also, $\widetilde{U}_t(A) \to A$ ultraweakly when $t \to 0$ and $\widetilde{U}_t$ is an ultraweakly continuous mapping [27, 28, 29]. These mappings have a strong positivity property called complete positivity:

$$\sum_{i,j} B_i^\dagger \widetilde{U}_t(A_i^\dagger A_j) B_j \geq 0, \quad A_i, B_i \in \mathcal{B}(\mathcal{H}). \tag{3}$$

In the axiomatic approach to the description of the evolution of open quantum systems [27, 28, 29], one supposes that the time evolution $\Phi_t$ of open systems is not very different from the time evolution of closed systems. The simplest dynamics $\Phi_t$ which introduces a preferred direction in time, characteristic for dissipative processes, is that in which the group condition is replaced by the semigroup condition [27, 30, 31]

$$\Phi_{t+s} = \Phi_t \Phi_s, \ t, s \geq 0. \tag{4}$$

The complete positivity condition has the form:

$$\sum_{i,j} B_i^\dagger \widetilde{\Phi}_t(A_i^\dagger A_j) B_j \geq 0, \quad A_i, B_i \in \mathcal{B}(\mathcal{H}), \tag{5}$$

where $\widetilde{\Phi}_t$ denotes the dual of $\Phi_t$ acting on $\mathcal{B}(\mathcal{H})$, defined by the duality condition

$$\mathrm{Tr}(\Phi_t(\rho)A) = \mathrm{Tr}(\rho \widetilde{\Phi}_t(A)). \tag{6}$$

Then the conditions $\mathrm{Tr}\Phi_t(\rho) = 1$ and $\widetilde{\Phi}_t(I) = I$ are equivalent. Also the conditions $\widetilde{\Phi}_t(A) \to A$ ultraweakly when $t \to 0$ and $\Phi_t(\rho) \to \rho$ in the trace norm when $t \to 0$, are equivalent. For the semigroups with these properties and with a more weak property of positivity than Eq. (5), namely

$$A \geq 0 \to \widetilde{\Phi}_t(A) \geq 0, \tag{7}$$

it is well known that there exists a (generally unbounded) mapping $\widetilde{L}$ – the generator of $\widetilde{\Phi}_t$, and $\widetilde{\Phi}_t$ is uniquely determined by $\widetilde{L}$. The dual generator of the dual semigroup $\Phi_t$ is denoted by $L$:

$$\mathrm{Tr}(L(\rho)A) = \mathrm{Tr}(\rho \widetilde{L}(A)). \tag{8}$$

The evolution equations by which $L$ and $\widetilde{L}$ determine uniquely $\Phi_t$ and $\widetilde{\Phi}_t$, respectively, are given in the Schrödinger and Heisenberg picture by

$$\frac{d\Phi_t(\rho)}{dt} = L(\Phi_t(\rho)) \tag{9}$$

and

$$\frac{d\widetilde{\Phi}_t(A)}{dt} = \widetilde{L}(\widetilde{\Phi}_t(A)). \tag{10}$$

These equations replace in the case of open systems the von Neumann-Liouville equations

$$\frac{dU_t(\rho)}{dt} = -\frac{i}{\hbar}[H, U_t(\rho)] \tag{11}$$

and

$$\frac{d\widetilde{U}_t(A)}{dt} = \frac{i}{\hbar}[H, \widetilde{U}_t(A)], \tag{12}$$

respectively. For applications, Eqs. (9) and (10) are only useful if the detailed structure of the generator $L(\widetilde{L})$ is known and can be related to the concrete properties of the open systems described by such equations. For the class of dynamical semigroups which are completely positive and norm continuous, the generator $\widetilde{L}$ is bounded. In many applications the generator is unbounded.

According to Lindblad [29], the following argument can be used to justify the complete positivity of $\widetilde{\Phi}_t$: if the open system is extended in a trivial way to a larger system described in a Hilbert space $\mathcal{H}\otimes\mathcal{K}$ with the time evolution defined by

$$\widetilde{W}_t(A\otimes B) = \widetilde{\Phi}_t(A)\otimes B, \quad A \in \mathcal{B}(\mathcal{H}),\ B \in \mathcal{B}(\mathcal{K}), \tag{13}$$

then the positivity of the states of the compound system will be preserved by $\widetilde{W}_t$ only if $\widetilde{\Phi}_t$ is completely positive. With this observation a new equivalent definition of the complete positivity is obtained: $\widetilde{\Phi}_t$ is completely positive if $\widetilde{W}_t$ is positive for any finite dimensional Hilbert space $\mathcal{K}$. The physical meaning of complete positivity can mainly be understood in relation to the existence of entangled states, the typical example being given by a vector state with a singlet-like structure that cannot be written as a tensor product of vector states. Positivity property guarantees the physical consistency of evolving states of single systems, while complete positivity prevents inconsistencies in entangled composite systems, and therefore the existence of entangled states makes the request of complete positivity necessary [32].

A bounded mapping $\widetilde{L} : \mathcal{B}(\mathcal{H}) \to \mathcal{B}(\mathcal{H})$ which satisfies $\widetilde{L}(I) = 0$, $\widetilde{L}(A^\dagger) = \widetilde{L}^\dagger(A)$ and

$$\widetilde{L}(A^\dagger A) - \widetilde{L}(A^\dagger)A - A^\dagger\widetilde{L}(A) \geq 0 \tag{14}$$

is called dissipative. The 2-positivity property of the completely positive mapping $\widetilde{\Phi}_t$:

$$\widetilde{\Phi}_t(A^\dagger A) \geq \widetilde{\Phi}_t(A^\dagger)\widetilde{\Phi}_t(A), \tag{15}$$

with equality at $t = 0$, implies that $\widetilde{L}$ is dissipative. Conversely, the dissipativity of $\widetilde{L}$ implies that $\widetilde{\Phi}_t$ is 2-positive. $\widetilde{L}$ is called completely dissipative if all trivial extensions of $\widetilde{L}$ to a compound system described by $\mathcal{H}\otimes\mathcal{K}$

with any finite dimensional Hilbert space $\mathcal{K}$ are dissipative. There exists a one-to-one correspondence between the completely positive norm continuous semigroups $\widetilde{\Phi}_t$ and completely dissipative generators $\widetilde{L}$. The following structural theorem gives the most general form of a completely dissipative mapping $\widetilde{L}$ [29].

**Theorem.** $\widetilde{L}$ is completely dissipative and ultraweakly continuous if and only if it is of the form

$$\widetilde{L}(A) = \frac{i}{\hbar}[H, A] + \frac{1}{2\hbar}\sum_j (V_j^\dagger [A, V_j] + [V_j^\dagger, A]V_j), \tag{16}$$

where $V_j,\ \sum_j V_j^\dagger V_j \in \mathcal{B}(\mathcal{H}),\ H \in \mathcal{B}(\mathcal{H})_{\mathrm{s.a.}}$.

The dual generator on the state space (Schrödinger picture) is of the form

$$L(\rho) = -\frac{i}{\hbar}[H, \rho] + \frac{1}{2\hbar}\sum_j ([V_j \rho, V_j^\dagger] + [V_j, \rho V_j^\dagger]). \tag{17}$$

Eqs. (9) and (17) give the explicit form of the Kossakowski-Lindblad master equation, which is the most general time-homogeneous quantum mechanical Markovian master equation with a bounded Liouville operator [29, 31, 33, 34]:

$$\frac{d\Phi_t(\rho)}{dt} = -\frac{i}{\hbar}[H, \Phi_t(\rho)] + \frac{1}{2\hbar}\sum_j ([V_j \Phi_t(\rho), V_j^\dagger] + [V_j, \Phi_t(\rho) V_j^\dagger]). \tag{18}$$

The assumption of a semigroup dynamics is only applicable in the limit of weak coupling of the subsystem with its environment, i.e. for long relaxation times [35]. We mention that the majority of Markovian master equations found in the literature are of this form after some rearrangement of terms, even for unbounded generators. It is also an empirical fact for many physically interesting situations that the time evolutions $\Phi_t$ drive the system towards a unique final state $\rho(\infty) = \lim_{t\to\infty} \Phi_t(\rho(0))$ for all $\rho(0) \in \mathcal{D}(\mathcal{H})$.

# 3 Time evolution of two independent bosonic modes interacting with an environment

We are interested in the dynamics of quantum correlations in a subsystem composed of two non-interacting (independent) bosonic modes (harmonic oscillators) in weak interaction with a thermal environment, so that their reduced time evolution can be described by a Markovian, completely positive quantum dynamical semigroup. If $\widetilde{\Phi}_t$ is the dynamical semigroup describing the irreversible time evolution of the open quantum system in the Heisenberg representation, then the Kossakowski-Lindblad master equation has the following form for an operator A (see Eqs. (10), (16)) [29, 31, 33, 34]:

$$\frac{d\widetilde{\Phi}_t(A)}{dt} = \frac{i}{\hbar}[H, \widetilde{\Phi}_t(A)] + \frac{1}{2\hbar}\sum_j (V_j^\dagger[\widetilde{\Phi}_t(A), V_j] + [V_j^\dagger, \widetilde{\Phi}_t(A)]V_j). \tag{19}$$

Here, $H$ denotes the Hamiltonian of the open system and the operators $V_j, V_j^\dagger$, defined on the Hilbert space of $H$, represent the interaction of the open system with the environment. We are interested in the set of Gaussian states, therefore we introduce quantum dynamical semigroups which preserve this set and in this case our model represents a Gaussian noise channel. Consequently $H$ is chosen as a polynomial of second degree in the coordinates $x, y$ and momenta $p_x, p_y$ of the two quantum oscillators and $V_j, V_j^\dagger$ are taken polynomials of first degree in these canonical observables. Then in the linear space spanned by the coordinates and momenta there exist only four linearly independent operators $V_{j=1,2,3,4}$ [36]:

$$V_j = a_{xj}p_x + a_{yj}p_y + b_{xj}x + b_{yj}y, \tag{20}$$

where $a_{xj}, a_{yj}, b_{xj}, b_{yj}$ are complex coefficients. The Hamiltonian $H$ of the two uncoupled non-resonant modes of identical mass $m$ and frequencies $\omega_1$ and $\omega_2$ is given by

$$H = \frac{1}{2m}(p_x^2 + p_y^2) + \frac{m}{2}(\omega_1^2x^2 + \omega_2^2y^2). \tag{21}$$

The fact that $\widetilde{\Phi}_t$ is a dynamical semigroup implies the positivity of the following matrix formed by the scalar products of the four vectors $\mathbf{a}_x, \mathbf{a}_y, \mathbf{b}_x, \mathbf{b}_y$, whose entries are the components $a_{xj}, a_{yj}, b_{xj}, b_{yj}$, respectively:

$$\frac{1}{2}\hbar = \begin{pmatrix} (\mathbf{a}_x\mathbf{a}_x) & (\mathbf{a}_x\mathbf{b}_x) & (\mathbf{a}_x\mathbf{a}_y) & (\mathbf{a}_x\mathbf{b}_y) \\ (\mathbf{b}_x\mathbf{a}_x) & (\mathbf{b}_x\mathbf{b}_x) & (\mathbf{b}_x\mathbf{a}_y) & (\mathbf{b}_x\mathbf{b}_y) \\ (\mathbf{a}_y\mathbf{a}_x) & (\mathbf{a}_y\mathbf{b}_x) & (\mathbf{a}_y\mathbf{a}_y) & (\mathbf{a}_y\mathbf{b}_y) \\ (\mathbf{b}_y\mathbf{a}_x) & (\mathbf{b}_y\mathbf{b}_x) & (\mathbf{b}_y\mathbf{a}_y) & (\mathbf{b}_y\mathbf{b}_y) \end{pmatrix} \tag{22}$$

Its matrix elements have to be chosen appropriately to suit various physical models of the environment. For a quite general environment able to induce noise and damping effects, we take this matrix of the following form, where all the coefficients $D_{xx}, D_{xp_x}$,... and $\lambda$ are real quantities, representing the diffusion coefficients and, respectively, the dissipation constant:

$$\left(\begin{array}{cccc} D_{xx} & -D_{xp_x} - i\hbar\lambda/2 & D_{xy} & -D_{xp_y} \\ -D_{xp_x} + i\hbar\lambda/2 & D_{p_xp_x} & -D_{yp_x} & D_{p_xp_y} \\ D_{xy} & -D_{yp_x} & D_{yy} & -D_{yp_y} - i\hbar\lambda/2 \\ -D_{xp_y} & D_{p_xp_y} & -D_{yp_y} + i\hbar\lambda/2 & D_{p_yp_y}. \end{array}\right) \tag{23}$$

It follows that the principal minors of this matrix are positive or zero. From the Cauchy-Schwarz inequality the following relations hold for the coefficients defined in Eq. (23) (from now on we put, for simplicity, $\hbar = 1$):

$$\begin{gathered} D_{xx}D_{p_xp_x} - D^2_{xp_x} \geq \frac{\lambda^2}{4},\ D_{yy}D_{p_yp_y} - D^2_{yp_y} \geq \frac{\lambda^2}{4}, \\ D_{xx}D_{yy} - D^2_{xy} \geq 0,\ D_{p_xp_x}D_{p_yp_y} - D^2_{p_xp_y} \geq 0, \\ D_{xx}D_{p_yp_y} - D^2_{xp_y} \geq 0,\ D_{yy}D_{p_xp_x} - D^2_{yp_x} \geq 0. \end{gathered} \tag{24}$$

The matrix of the coefficients (23) can be conveniently written as (T denotes the transposed matrix)

$$\left(\begin{array}{cc} C_1 & C_3 \\ {C_3}^{\mathrm{T}} & C_2 \end{array}\right), \tag{25}$$

in terms of $2 \times 2$ matrices $C_1 = {C_1}^\dagger$, $C_2 = {C_2}^\dagger$ and $C_3$. This decomposition has a direct physical interpretation: the elements containing the diagonal contributions $C_1$ and $C_2$ represent diffusion and dissipation coefficients corresponding to the first, respectively the second, system in absence of the other, while the elements in $C_3$ represent environment generated couplings between the two modes, taken initially independent.

We introduce the following $4 \times 4$ bimodal covariance matrix:

$$\sigma(t) = \left(\begin{array}{cccc} \sigma_{xx}(t) & \sigma_{xp_x}(t) & \sigma_{xy}(t) & \sigma_{xp_y}(t) \\ \sigma_{xp_x}(t) & \sigma_{p_xp_x}(t) & \sigma_{yp_x}(t) & \sigma_{p_xp_y}(t) \\ \sigma_{xy}(t) & \sigma_{yp_x}(t) & \sigma_{yy}(t) & \sigma_{yp_y}(t) \\ \sigma_{xp_y}(t) & \sigma_{p_xp_y}(t) & \sigma_{yp_y}(t) & \sigma_{p_yp_y}(t) \end{array}\right) \tag{26}$$

where the correlations of operators $R_i$ and $R_j, i, j = 1, .., 4$, with $\mathbf{R} = \{x, p_x, y, p_y\}$, are defined by using the density operator $\rho$ of the initial state of the quantum system, as follows:

$$\sigma_{R_iR_j}(t) = \frac{1}{2}\mathrm{Tr}[\rho(R_iR_j + R_jR_i)(t)] - \mathrm{Tr}[\rho R_i(t)]\mathrm{Tr}[\rho R_j(t)]. \tag{27}$$

The problem of solving the master equation for the operators in Heisenberg representation can be transformed into a problem of solving first-order in time, coupled linear differential equations for the covariance matrix elements. Namely, from Eq. (19) we obtain by direct calculation the following systems of equations for the quantum correlations of the canonical observables [36]:

$$\frac{d\sigma(t)}{dt} = Y\sigma(t) + \sigma(t)Y^{\mathrm{T}} + 2D, \tag{28}$$

where

$$Y = \begin{pmatrix} -\lambda & 1/m & 0 & 0 \\ -m\omega_1^2 & -\lambda & 0 & 0 \\ 0 & 0 & -\lambda & 1/m \\ 0 & 0 & -m\omega_2^2 & -\lambda \end{pmatrix} \tag{29}$$

$$D = \begin{pmatrix} D_{xx} & D_{xp_x} & D_{xy} & D_{xp_y} \\ D_{xp_x} & D_{p_xp_x} & D_{yp_x} & D_{p_xp_y} \\ D_{xy} & D_{yp_x} & D_{yy} & D_{yp_y} \\ D_{xp_y} & D_{p_xp_y} & D_{yp_y} & D_{p_yp_y} \end{pmatrix} \tag{30}$$

Introducing the notation $\sigma(\infty) \equiv \lim_{t\to\infty} \sigma(t)$, the time-dependent solution of Eq. (28) is given by [36]

$$\sigma(t) = M(t)[\sigma(0) - \sigma(\infty)]M^{\mathrm{T}}(t) + \sigma(\infty), \tag{31}$$

where the matrix $M(t) = \exp(Yt)$ has to fulfill the condition $\lim_{t\to\infty} M(t) = 0$. In order that this limit exists, $Y$ must only have eigenvalues with negative real parts. The values at infinity are obtained from the equation

$$Y\sigma(\infty) + \sigma(\infty)Y^{\mathrm{T}} = -2D. \tag{32}$$

# 4 Dynamics of quantum correlations

To describe the dynamics of quantum correlations, we use two types of measures: logarithmic negativity for entanglement, and quantum discord.

## 4.1 Time evolution of entanglement and logarithmic negativity

A well-known sufficient condition for inseparability is the so-called Peres-Horodecki criterion [25, 37], which is based on the observation that the non-completely positive nature of the partial transposition operation of the density matrix for a bipartite system (transposition with respect to degrees of freedom of one subsystem only) may turn an inseparable state into a nonphysical state. The signature of this non-physicality, and thus of quantum entanglement, is the appearance of a negative eigenvalue in the eigenspectrum of the partially transposed density matrix of a bipartite system. The characterization of the separability of continuous variable states using second-order moments of quadrature operators was given in Refs. [26, 38]. For Gaussian states, whose statistical properties are fully characterized by just second-order moments, this criterion was proven to be necessary and sufficient: a Gaussian continuous variable state is separable if and only if the partial transpose of its density matrix is non-negative (positive partial transpose (PPT) criterion).

The two-mode Gaussian state is entirely specified by its covariance matrix (30), which is a real, symmetric and positive matrix with the following block structure:

$$\sigma(t) = \left( \begin{array}{cc} A & C \\ C^{\mathrm{T}} & B \end{array} \right), \tag{33}$$

where $A$, $B$ and $C$ are $2 \times 2$ Hermitian matrices. $A$ and $B$ denote the symmetric covariance matrices for the individual reduced one-mode states, while the matrix $C$ contains the cross-correlations between modes. When these correlations have non-zero values, then the states with $\det C \geq 0$ are separable states, while for $\det C < 0$ it may be possible that the states are entangled.

The $4 \times 4$ covariance matrix (33) (where all first moments can be set to zero by means of local unitary operations which do not affect the entanglement) contains four local symplectic invariants in form of the determinants of the block matrices $A, B, C$ and covariance matrix $\sigma$. Based on the above invariants, Simon [26] derived the following PPT criterion for bipartite Gaussian continuous variable states: the necessary and sufficient condition for separability is $S(t) \geq 0$, where

$$\begin{aligned} S(t) \equiv \det A \det B + (\frac{1}{4} - |\det C|)^2 \\ -\mathrm{Tr}[AJCJBJC^{\mathrm{T}}J] - \frac{1}{4}(\det A + \det B) \end{aligned} \tag{34}$$

and $J$ is the $2 \times 2$ symplectic matrix

$$J = \left( \begin{array}{cc} 0 & 1 \\ -1 & 0 \end{array} \right). \tag{35}$$

For Gaussian states, the measures of entanglement of bipartite systems are based on the invariants constructed from the elements of the covariance matrix [8, 12]. In order to quantify the degree of entanglement of the two-mode states it is suitable to use the logarithmic negativity. For a Gaussian density operator, the logarithmic negativity is completely defined by the symplectic spectrum of the partial transpose of the covariance matrix. It is given by $E_N = \max\{0, -\log_2 2\tilde{\nu}_-\}$, where $\tilde{\nu}_-$ is the smallest of the two symplectic eigenvalues of the partial transpose $\tilde{\sigma}$ of the two-mode covariance matrix $\sigma$ [11]:

$$2\tilde{\nu}_{\mp}^2 = \tilde{\Delta} \mp \sqrt{\tilde{\Delta}^2 - 4\det\sigma} \tag{36}$$

and $\tilde{\Delta}$ is the symplectic invariant (seralian), given by $\tilde{\Delta} = \det A + \det B - 2\det C$.

In our model, the logarithmic negativity is calculated as [39, 40]

$$E_N(t) = \max\{0, -\frac{1}{2}\log_2[4g(\sigma(t))]\}, \tag{37}$$

where

$$\begin{aligned} g(\sigma(t)) &= \frac{1}{2}(\det A + \det B) - \det C \\ &- \left(\left[\frac{1}{2}(\det A + \det B) - \det C\right]^2 - \det\sigma(t)\right)^{1/2}. \end{aligned} \tag{38}$$

It determines the strength of entanglement for $E_N(t) > 0$, and if $E_N(t) \leq 0$, then the state is separable.

We suppose that the asymptotic state of the considered open system is a Gibbs state corresponding to two independent bosonic modes in thermal equilibrium at temperature $T$. Then the quantum diffusion coefficients have the following form [34]:

$$\begin{aligned} m\omega_1 D_{xx} &= \frac{D_{p_xp_x}}{m\omega_1} = \frac{\lambda}{2}\coth\frac{\omega_1}{2kT}, \\ m\omega_2 D_{yy} &= \frac{D_{p_yp_y}}{m\omega_2} = \frac{\lambda}{2}\coth\frac{\omega_2}{2kT}, \\ D_{xp_x} = D_{yp_y} = D_{xy} &= D_{p_xp_y} = D_{xp_y} = D_{yp_x} = 0. \end{aligned} \tag{39}$$

The elements of the covariance matrix can be calculated from Eqs. (31), (32).

In the following, we analyze the dependence of the Simon function $S(t)$ and of the logarithmic negativity $E_N(t)$ on time $t$ and temperature $T$ of the thermal bath, with the diffusion coefficients given by Eqs. (39). We consider two types of the initial Gaussian states: separable and entangled.

1) We consider a separable initial Gaussian state, with the two modes initially prepared in their single-mode squeezed states (unimodal squeezed state) and with its initial covariance matrix taken of the form

$$\sigma_s(0) = \frac{1}{2}\begin{pmatrix} \cosh 2r & \sinh 2r & 0 & 0 \\ \sinh 2r & \cosh 2r & 0 & 0 \\ 0 & 0 & \cosh 2r & \sinh 2r \\ 0 & 0 & \sinh 2r & \cosh 2r \end{pmatrix} \tag{40}$$

where $r$ denotes the squeezing parameter. In this case $S(t)$ becomes strictly positive after the initial moment of time ($S(0) = 0$), so that the initial separable state remains separable for all values of the temperature $T$ and for all times.

2) We take an entangled initial Gaussian state of the form of a two-mode vacuum squeezed state, with the initial covariance matrix given by

$$\sigma_e(0) = \frac{1}{2}\begin{pmatrix} \cosh 2r & 0 & \sinh 2r & 0 \\ 0 & \cosh 2r & 0 & -\sinh 2r \\ \sinh 2r & 0 & \cosh 2r & 0 \\ 0 & -\sinh 2r & 0 & \cosh 2r \end{pmatrix} \tag{41}$$

We observe that for all temperatures $T$, at certain finite moment of time, which depends on $T$, $E_N(t)$ becomes zero and therefore the state becomes separable. This is the so-called phenomenon of entanglement sudden death [23, 41]. It is in contrast to the quantum decoherence, during which the loss of quantum coherence is usually gradual [17, 42].

3) We assume that the initial Gaussian state is a two-mode squeezed thermal state, with the covariance matrix of the form [43]

$$\sigma_{st}(0) = \begin{pmatrix} a & 0 & c & 0 \\ 0 & a & 0 & -c \\ c & 0 & b & 0 \\ 0 & -c & 0 & b \end{pmatrix} \tag{42}$$

with the matrix elements given by

$$
\begin{aligned}
a &= n_1 \cosh^2 r + n_2 \sinh^2 r + \frac{1}{2}\cosh 2r,\\
b &= n_1 \sinh^2 r + n_2 \cosh^2 r + \frac{1}{2}\cosh 2r,\\
c &= \frac{1}{2}(n_1 + n_2 + 1)\sinh 2r,
\end{aligned}
\tag{43}
$$

where $n_1, n_2$ are the average number of thermal photons associated with the two modes and $r$ denotes the squeezing parameter. In the particular case $n_1 = 0$ and $n_2 = 0$, (42) becomes the covariance matrix of the two-mode squeezed vacuum state (41). A two-mode squeezed thermal state is entangled when the squeezing parameter $r$ satisfies the inequality $r > r_s$ [43], where

$$\cosh^2 r_s = \frac{(n_1+1)(n_2+1)}{n_1+n_2+1}. \tag{44}$$

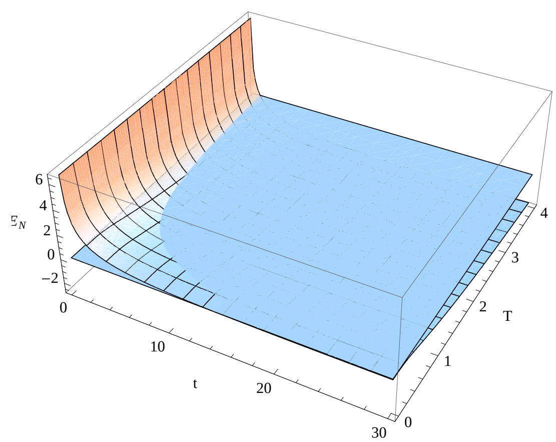



Figure 1: Logarithmic negativity $E_N$ versus time $t$ and temperature $T$ for an entangled initial non-symmetric squeezed thermal state with squeezing parameter $r = 3$, $n_1 = 3, n_2 = 1$ and $\lambda = 0.1, \omega_1 = 1, \omega_2 = 2$. We take $m = \hbar = k = 1$.

The evolution of entangled initial squeezed thermal states with the covariance matrix given by Eq. (42) is illustrated in Fig. 1, where we represent the dependence of the logarithmic negativity $E_N(t)$ on time $t$ and temperature $T$ for the case of an initial non-symmetric Gaussian state ($a \neq b$). For all temperatures $T$, including zero temperature, at certain finite moment of time, which depends on $T$, $E_N(t)$ becomes zero and therefore the state becomes separable. One can show that the dissipation favors the phenomenon of entanglement sudden death – with increasing the dissipation parameter $\lambda$, the entanglement suppression happens earlier. The same qualitative behaviour of the time evolution of entanglement was obtained previously in the particular case $n_1 = 0$ and $n_2 = 0$ corresponding to an initial two-mode squeezed vacuum state and in the case of symmetric initial squeezed thermal states.

One can assert that the asymmetry ($a \neq b$) of the initial Gaussian state favors the suppression of entanglement. The most robust under the influence of the environment is the entanglement of symmetric ($a = b$) initial squeezed thermal states. An even stronger influence on the entanglement has the non-resonant character of the two modes: by increasing the ratio of the frequencies of the two modes, the entanglement sudden death happens earlier in time. The longest surviving entanglement takes place when the modes are resonant ($\omega_1 = \omega_2$). This effect due to the non-resonance of the modes is stronger for small values of the frequencies, and it diminishes, for the same ratio of frequencies, by increasing the values of frequencies.

In our model, in which we suppose that the asymptotic state of the considered open system is a Gibbs state corresponding to two independent bosonic modes in thermal equilibrium, a separable initial state remains separable in time, and it is not possible to generate entanglement. This is in contrast with the possibility of entanglement generation starting, for instance, with a separable state in the case of two non-interacting two-level systems immersed in a common bath [32]. At the same time we notice that in the case of two identical harmonic oscillators interacting with a general environment, characterized by general diffusion and dissipation coefficients, we obtain that for separable initial states and for definite values of these coefficients, entanglement generation or a periodic generation and collapse of entanglement take place [40, 44]. In discussing the entanglement decay, it is interesting to mention that models have been elaborated to realize quantum feedback control of continuous variable entanglement for a system consisting of two interacting bosonic modes plunged into an environment, based on a local technique [45], or on a nonlocal homodyne measurement [46].

The dynamics of entanglement of the two modes strongly depends on the initial states and the coefficients describing the interaction of the system with the thermal environment (dissipation constant and temperature). As expected, the logarithmic negativity has a behaviour similar to that one of the Simon function in what concerns the characteristics of the state of being separable or entangled [39, 40, 42, 44].

### 4.2 Asymptotic entanglement

On general grounds, one expects that the effects of decoherence is dominant in the long-time regime, so that no quantum correlations (entanglement) is expected to be left at infinity. Indeed, using the diffusion coefficients given by Eqs. (39), we obtain from Eq. (32) the following elements of the asymptotic matrices $A(\infty)$ and $B(\infty)$ :

$$\begin{aligned} m\omega_1\sigma_{xx}(\infty) &= \frac{\sigma_{p_xp_x}(\infty)}{m\omega_1} = \frac{1}{2}\coth\frac{\omega_1}{2kT}, \quad \sigma_{xp_x}(\infty) = 0, \\ m\omega_2\sigma_{yy}(\infty) &= \frac{\sigma_{p_yp_y}(\infty)}{m\omega_2} = \frac{1}{2}\coth\frac{\omega_2}{2kT}, \quad \sigma_{yp_y}(\infty) = 0 \end{aligned} \tag{45}$$

and of the entanglement matrix $C(\infty)$ :

$$\sigma_{xy}(\infty) = \sigma_{xp_y}(\infty) = \sigma_{yp_x}(\infty) = \sigma_{p_xp_y}(\infty) = 0. \tag{46}$$

Then the Simon expression (34) takes the following form in the limit of large times:

$$S(\infty) = \frac{1}{16}\left(\coth^2\frac{\omega_1}{2kT} - 1\right)\left(\coth^2\frac{\omega_2}{2kT} - 1\right), \tag{47}$$

and, correspondingly, the equilibrium asymptotic state is always separable in the case of two non-interacting bosonic modes immersed in a common thermal reservoir.

In Refs. [20, 21, 39, 40, 42, 44] we described the dependence of the logarithmic negativity $E_N(t)$ on time and mixed diffusion coefficient for two modes interacting with a general environment. In the present case of a thermal bath, the asymptotic logarithmic negativity is given by (for $\omega_1 \le \omega_2$)

$$E_N(\infty) = -\log_2\coth\frac{\omega_2}{2kT}. \tag{48}$$

It depends only on temperature, and does not depend on the initial Gaussian state. $E_N(\infty) < 0$ for $T \neq 0$ and $E_N(\infty) = 0$ for $T = 0$, and this confirms the previous statement that the asymptotic state is always separable.

### 4.3 Gaussian quantum discord

The separability of quantum states has often been described as a property synonymous with the classicality. However, recent studies have shown that separable states, usually considered as being classically correlated, might also contain quantum correlations. Quantum discord was introduced [4, 5] as a measure of all quantum correlations in a bipartite state, including – but not restricted to – entanglement. Quantum discord has been defined as the difference between two quantum analogues of classically equivalent expression of the mutual information, which is a measure of total correlations in a quantum state. For pure entangled states quantum discord coincides with the entropy of entanglement. Quantum discord can be different from zero also for some mixed separable state and therefore the correlations in such separable states with positive discord are an indicator of quantumness. States with zero discord represent essentially a classical probability distribution embedded in a quantum system.

For an arbitrary bipartite state $\rho_{12}$, the total correlations are expressed by quantum mutual information [47]

$$I(\rho_{12}) = \sum_{i=1,2} S(\rho_i) - S(\rho_{12}), \tag{49}$$

where $\rho_i$ represents the reduced density matrix of subsystem $i$ and $S(\rho) = -\mathrm{Tr}(\rho \ln \rho)$ is the von Neumann entropy. Henderson and Vedral [6] proposed a measure of bipartite classical correlations $C(\rho_{12})$ based on a complete set of local projectors $\{\Pi_2^k\}$ on the subsystem 2: the classical correlation in the bipartite quantum state $\rho_{12}$ can be given by

$$C(\rho_{12}) = S(\rho_1) - \inf_{\{\Pi_2^k\}}\{S(\rho_{1|2})\}, \tag{50}$$

where $S(\rho_{1|2}) = \sum_k p^k S(\rho_1^k)$ is the conditional entropy of subsystem 1 and $\inf\{S(\rho_{1|2})\}$ represents the minimal value of the entropy with respect to a complete set of local measurements $\{\Pi_2^k\}$. Here, $p^k$ is the measurement probability for the $k$th local projector and $\rho_1^k$ denotes the reduced state of subsystem 1 after the local measurements. Then the quantum discord is defined by

$$D(\rho_{12}) = I(\rho_{12}) - C(\rho_{12}). \tag{51}$$

Originally the quantum discord was defined and evaluated mainly for finite dimensional systems. Recently [48, 49] the notion of discord has been extended to the domain of continuous variable systems, in particular to the analysis of bipartite systems described by two-mode Gaussian states. Closed formulas have been derived for bipartite thermal squeezed states [48] and for all two-mode Gaussian states [49].

The Gaussian quantum discord of a general two-mode Gaussian state $\rho_{12}$ can be defined as the quantum discord where the conditional entropy is restricted to generalized Gaussian positive operator valued measurements (POVM) on the mode 2 and in terms of symplectic invariants it is given by (the symmetry between the two modes 1 and 2 is broken) [49]

$$D = f(\sqrt{\beta}) - f(\nu_-) - f(\nu_+) + f(\sqrt{\varepsilon}), \tag{52}$$

where

$$f(x) = \frac{x+1}{2}\log\frac{x+1}{2} - \frac{x-1}{2}\log\frac{x-1}{2}, \tag{53}$$

$$\varepsilon = \begin{cases} \dfrac{2\gamma^2 + (\beta-1)(\delta-\alpha) + 2|\gamma|\sqrt{\gamma^2 + (\beta-1)(\delta-\alpha)}}{(\beta-1)^2}, \\ \quad \text{if } \ (\delta - \alpha\beta)^2 \le (\beta+1)\gamma^2(\alpha+\delta) \\ \\ \dfrac{\alpha\beta - \gamma^2 + \delta - \sqrt{\gamma^4 + (\delta-\alpha\beta)^2 - 2\gamma^2(\delta+\alpha\beta)}}{2\beta}, \\ \quad \text{otherwise,} \end{cases} \tag{54}$$

$$\alpha = 4\det A, \quad \beta = 4\det B, \quad \gamma = 4\det C, \quad \delta = 16\det\sigma, \tag{55}$$

and $\nu_\mp$ are the symplectic eigenvalues of the state, given by

$$2\nu_\mp^2 = \Delta \mp \sqrt{\Delta^2 - 4\det\sigma}, \tag{56}$$

where $\Delta = \det A + \det B + 2\det C$. Notice that Gaussian quantum discord only depends on $|\det C|$, i.e., entangled ($\det C < 0$) and separable states are treated on equal footing.

The evolution of the Gaussian quantum discord $D$ is illustrated in Fig. 2, where we represent the dependence of $D$ on time $t$ and temperature $T$ for an entangled initial non-symmetric Gaussian state, taken of the form of a two-mode squeezed thermal state (42), for such values of the parameters which satisfy for all times the first condition in formula (54). The Gaussian discord has nonzero values for all finite times and this fact certifies the existence of non-classical correlations in two-mode Gaussian states, either separable or entangled. Gaussian discord asymptotically decreases in time, compared to the case of logarithmic negativity, which has an evolution leading to a sudden suppression of entanglement. For entangled initial states the Gaussian discord remains strictly positive in time and in the limit of infinite

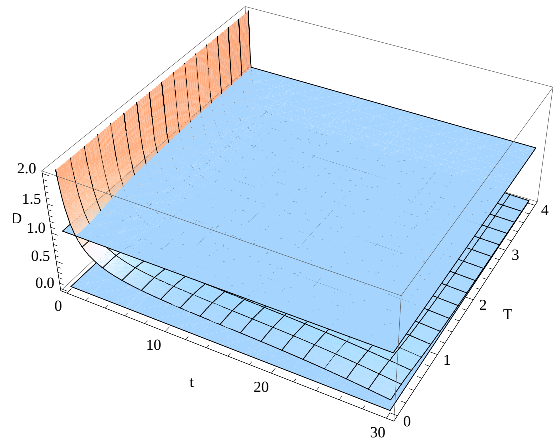



Figure 2: Gaussian quantum discord $D$ versus time $t$ and temperature $T$ for an entangled initial non-symmetric squeezed thermal state with squeezing parameter $r = 3$, $n_1 = 3, n_2 = 1$ and $\lambda = 0.1, \omega_1 = 1, \omega_2 = 2$. We take $m = \hbar = k = 1$.

time it tends asymptotically to zero, corresponding to the thermal product (separable) state, with no correlations at all. One can easily show that for a separable initial Gaussian state with covariance matrix (42) the quantum discord is zero and it keeps this values during the whole time evolution of the state.

From Fig. 2 we notice that, in agreement with the general properties of the Gaussian quantum discord [49], the states can be either separable or entangled for $D \leq 1$ and all the states above the threshold $D = 1$ are entangled. We also notice that the decay of quantum discord is stronger when the temperature $T$ is increasing. It should be remarked that the decay of quantum discord is very similar to that of the entanglement before the time of the sudden death of entanglement. Near the threshold of zero logarithmic negativity ($E_N = 0$), the nonzero values of the discord can quantify the non-classical correlations for separable mixed states and one considers that this fact could make possible some tasks in quantum computation [50]. The discord is increasing with the squeezing parameter $r$ and it is decreasing with increasing the ratio of the frequencies $\omega_1$ and $\omega_2$ of the two modes and the difference of parameters $a$ and $b$.

### 4.4 Classical corellations and quantum mutual information

The measure of classical correlations for a general two-mode Gaussian state $\rho_{12}$ can also be calculated and it is given by [49]

$$C = f(\sqrt{\alpha}) - f(\sqrt{\varepsilon}), \tag{57}$$

while the expression of the quantum mutual information, which measures the total correlations, is given by

$$I = f(\sqrt{\alpha}) + f(\sqrt{\beta}) - f(\nu_-) - f(\nu_+). \tag{58}$$

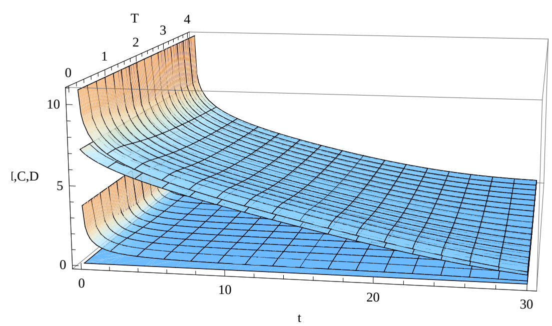



Figure 3: Quantum mutual information $I$ versus time $t$ and temperature $T$ for an entangled initial non-symmetric squeezed thermal state with squeezing parameter $r = 3$, $n_1 = 3, n_2 = 1$ and $\lambda = 0.1, \omega_1 = 1, \omega_2 = 2$. We take $m = \hbar = k = 1$. There are also represented the Gaussian quantum discord and classical correlations.

In Fig. 3 we illustrate the evolution of classical correlations $C$ and quantum mutual information $I$, as functions of time $t$ and temperature $T$ for an entangled initial Gaussian state, taken of the form of a two-mode squeezed thermal state (42). These two quantities manifest a qualitative behaviour similar to that one of the Gaussian discord: they have nonzero values for all finite times and in the limit of infinite time they tend asymptotically to zero, corresponding to the thermal product (separable) state, with no correlations at all. One can also see that the classical correlations and quantum mutual information decrease with increasing the temperature of the thermal bath. One can show that the classical correlations and quantum mutual information increase with increasing the squeezing parameter $r$ and the difference of parameters $a$ and $b$. At the same time classical correlations increase with the ratio of the frequencies $\omega_1$ and $\omega_2$ of the two modes, while quantum mutual information is decreasing with increasing this ratio. For comparison these quantities as well as quantum discord are represented on the same graphic. In the considered case the value of classical correlations is larger than that of quantum correlations, represented by the Gaussian quantum discord.

## 5 Conclusion

We have given a brief review of the theory of open quantum systems based on completely positive quantum dynamical semigroups and mentioned the necessity of the complete positivity for the existence of entangled states of systems interacting with an external environment. In the framework of this theory, by using the Peres-Simon necessary and sufficient condition for separability of two-mode Gaussian states, we investigated the Markovian dynamics of quantum correlations for a subsystem composed of two non-interacting bosonic modes embedded in a thermal bath. We have analyzed the influence of the environment on the dynamics of quantum entanglement and quantum discord for Gaussian initial states. We have described the time evolution of the logarithmic negativity, which characterizes the degree of entanglement of the quantum state, in terms of the covariance matrix for squeezed vacuum states and squeezed thermal states, for the case when the asymptotic state of the considered open system is a Gibbs state corresponding to two independent quantum harmonic oscillators in thermal equilibrium. For all values of the temperature of the thermal reservoir, an initial separable Gaussian state remains separable for all times. The dynamics of the quantum entanglement strongly depends on the initial states and the parameters characterizing the environment (temperature and dissipation constant). For an entangled initial squeezed vacuum state and squeezed thermal state, entanglement suppression (entanglement sudden death) takes place for all values of the temperatures of the environment, including zero temperature. The time when the entanglement is suppressed decreases with increasing the temperature and dissipation.

We described also the time evolution of Gaussian quantum discord, which is a measure of all quantum correlations in the bipartite state, including entanglement. The values of quantum discord decrease asymptotically in time. This is in contrast to the sudden death of entanglement. The time evolution of quantum discord is very similar to that of entanglement before the sudden suppression of the entanglement. Quantum discord is decreasing with increasing the temperature. After the sudden death of entanglement the nonzero values of discord manifest the existence of quantum correlations for separable mixed states. One considers that the robustness of quantum discord could favorize the realization of scalable quantum computing in contrast to the fragility of the entanglement [50]. We described also the time evolution of classical correlations and quantum mutual information, which measures the total correlations of the quantum system.

The existence of quantum correlations between the two bosonic modes interacting with a common environment is the result of the competition between entanglement and quantum decoherence. From the formal point of view, entanglement suppression corresponds to the finite time vanishing of the Simon separability function or, respectively, of the logarithmic negativity.

Presently there is a large debate relative to the physical interpretation existing behind the fascinating phenomena of quantum decoherence and existence of quantum correlations - quantum entanglement and quantum discord. Due to the increased interest manifested towards the continuous variables approach [7, 51] to quantum information theory, the present results, in particular the existence of quantum discord and the possibility of maintaining a bipartite entanglement in a thermal environment for long times, might be useful in controlling entanglement and discord in open systems and also for applications in the field of quantum information processing and communication.

# 6 Appendix

1. **Quantum information** is the study of the information processing tasks that can be accomplished using quantum mechanical systems [1].

Quantum theory, formalized in the first few decades of the 20th century, contains elements that are radically different from the classical description of Nature. An important aspect in these fundamental differences is the existence of quantum correlations in the quantum formalism. In the classical description of Nature, if a system is formed by different subsystems, complete knowledge of the whole system implies that the sum of the information of the subsystems makes up the complete information for the whole system. This is no longer true in the quantum formalism. In the quantum world, there exist states of composite systems for which we might have the complete information, while our knowledge about the subsystems might be completely random. One may reach some paradoxical conclusions if one applies a classical description to states which have characteristic quantum signatures. During the last decades, it was realized that these fundamentally nonclassical states, also denoted as entangled states, can provide us with something else than just paradoxes. They may be used to perform tasks that cannot be achieved with classical states. As benchmarks of this turning point in our view of such nonclassical states, one might mention the spectacular discoveries of (entanglement-based) quantum cryptography, quantum dense coding, and quantum teleportation [52].

Let us consider a bipartite system, which is traditionally supposed to be in possession of Alice (A) and Bob (B), who can be located in distant regions. Let Alice's physical system be described by the Hilbert space $\mathcal{H}_{\mathcal{A}}$ and that of Bob by $\mathcal{H}_{\mathcal{B}}$. Then the joint physical system of Alice and Bob is described by the tensor product Hilbert space $\mathcal{H}_{\mathcal{A}} \otimes \mathcal{H}_{\mathcal{B}}$.

A pure state, i.e., a projector $|\psi_{AB}><\psi_{AB}|$ on a vector $|\psi_{AB}>\in \mathcal{H}_{\mathcal{A}} \otimes \mathcal{H}_{\mathcal{B}}$, is a product state if the states of local subsystems are also pure states, that is, if $|\psi_{AB}>=|\psi_A>\otimes|\psi_B>$. However, there are states that cannot be written in this form. These states are called entangled states.

An example of entangled state is the well-known singlet state $(|01> -|10>)/\sqrt{2}$ (Bell state), where $|0>$ and $|1>$ are two orthonormal states.

A mixed state described by a density operator $\rho_{AB}$ of a two-party system is separable if and only if it can be represented as a convex combination of the product states:

$$\rho_{AB} = \sum_i p_i \rho_A^i \otimes \rho_B^i,$$

where $p_i$ is a probability distribution. Otherwise, the mixed state is said to be inseparable (entangled).

An important operational entanglement criterion is the positive partial transposition (PPT) criterion for detecting entanglement: given a bipartite state $\rho_{AB}$, find the eigenvalues of any of its partial transpositions with respect to one of the subsystems (transposition is equivalent to time reversal, or, expressed in terms of continuous variables, sign change of the momenta). A negative eigenvalue immediately implies that the state is entangled. Examples of states for which the partial transposition has negative eigenvalues include the singlet state.

The notion of entanglement appeared explicitly in the literature first in 1935, long before the dawn of the relatively young field of quantum information, and without any reference to discrete-variable qubit states. In fact, the entangled states treated in this 1935 paper by Einstein, Podolsky, and Rosen (EPR) were two-particle states quantum mechanically correlated with respect to their positions and momenta. The concept of entanglement has played an important role in quantum physics ever since its discovery last century and has now been recognized as a key resource in quantum information science.

The superposition principle leads to the existence of entangled states of two or more quantum systems and such states are characterized by the existence of correlations between the systems, the form of which cannot be satisfactorily accounted for by any classical theory. These have played a central role in the development of quantum theory since early in its development, starting with the famous paradox or dilemma of EPR. No less disturbing than the EPR dilemma is the problem of Schrödinger cat, an example of the apparent absurdity of following entanglement into the macroscopic world. It was Schrödinger who gave us the name entanglement (in German, "Verschränkung"); he emphasized its fundamental significance when he wrote, "I would call this not one but the characteristic trait of quantum mechanics, the one that enforces the entire departure from classical thought".

The prime example for an entangled Gaussian state is the pure two-mode squeezed (vacuum) state, described by the Gaussian Wigner function

$$W_{svs} = \frac{4}{\pi^2} \times \exp\{-e^{-2r}[(x_A + x_B)^2 + (p_A - p_B)^2] - e^{+2r}[(x_A - x_B)^2 + (p_A + p_B)^2]\},$$

where $x_A, p_A, x_B, p_B$ are the coordinates and momenta of the the two-mode system and $r$ is the squeezing parameter.

A unique measure of bipartite entanglement for pure states is given by the partial von Neumann entropy. This is the von Neumann entropy, of the reduced system after tracing out either subsystem: $\mathrm{Tr}\rho_A \ln \rho_A = \mathrm{Tr}\rho_B \ln \rho_B$, where $\rho_A = \mathrm{Tr}_B \rho_{AB},\ \rho_B = \mathrm{Tr}_A \rho_{AB}$.

In order to quantify the degrees of entanglement of an infinite-dimensional bipartite system states it is suitable to use the logarithmic negativity. The logarithmic negativity of a bipartite system consisting of two subsystems A and B is $E_N = \log_2 \|\rho^{\mathrm{T_B}}\|_1$, where $\rho^{\mathrm{T_B}}$ means the partial transpose of a mixed state density matrix operator $\rho_{AB}$ with respect to subsystem B. The operation $\|\cdot\|_1$ denotes the trace norm, which for any Hermitian operator $O$ is defined as $\|O\|_1 \equiv \mathrm{Tr}|O| \equiv \mathrm{Tr}\sqrt{O^\dagger O}$ and it is calculated as the sum of absolute values of the eigenvalues of $O$.

Logarithmic negativity quantifies the degree of violation of PPT criterion for separability, i.e. how much the partial transposition of $\rho$ fails to be positive and it is based on negative eigenvalues of the partial transpose of the subsystem density matrix. For a Gaussian density operator, the negativity is completely defined by the symplectic spectrum of the partial transpose of the covariance matrix.

2. The **damped quantum harmonic oscillator** is considered in the framework of the theory of open systems based on completely positive quantum dynamical semigroups [33, 34]. The basic assumption is that the general form of a bounded mapping $L$ given by Lindblad theorem is also valid for an unbounded completely dissipative mapping $L$:

$$L(\rho) = -\frac{i}{\hbar}[H,\rho] + \frac{1}{2\hbar}\sum_j([V_j\rho, V_j^\dagger] + [V_j, \rho V_j^\dagger]).$$

This assumption gives one of the simplest way to construct an appropriate model for this quantum dissipative system. Another simple condition imposed to the operators $H, V_j, V_j^\dagger$ is that they are functions of the basic observables of the one-dimensional quantum mechanical system $q$ and $p$ with $[q,p] = i\hbar I$, where $I$ is the identity operator on $\mathcal{H}$ of such kind that the obtained model is exactly solvable. A precise version for this last condition is that linear spaces spanned by the first degree (respectively second degree) noncommutative polynomials in $p$ and $q$ are invariant to the action of the completely dissipative mapping $L$. This condition implies that $V_j$ are at most first degree polynomials in $p$ and $q$ and $H$ is at most a second degree polynomial in $p$ and $q$.

Beacause in the linear space of the first degree polynomials in $p$ and $q$ the operators $p$ and $q$ give a basis, there exist only two $C$-linear independent operators $V_1, V_2$ which can be written in the form

$$V_i = a_i p + b_i q,\ \ i = 1, 2,$$

with $a_i, b_i$ complex numbers. The constant term is omitted because its contribution to the generator $L$ is equivalent to terms in $H$ linear in $p$ and $q$ which for simplicity are chosen to be zero. Then $H$ is chosen of the form

$$H = H_0 + \frac{\mu}{2}(pq + qp), \quad H_0 = \frac{1}{2m}p^2 + \frac{m\omega^2}{2}q^2.$$

With these choices the Markovian master equation can be written:

$$\frac{d\rho}{dt} = -\frac{i}{\hbar}[H_0,\rho] - \frac{i}{2\hbar}(\lambda+\mu)[q, \rho p + p\rho] + \frac{i}{2\hbar}(\lambda-\mu)[p, \rho q + q\rho]$$

$$-\frac{D_{pp}}{\hbar^2}[q,[q,\rho]] - \frac{D_{qq}}{\hbar^2}[p,[p,\rho]] + \frac{D_{pq}}{\hbar^2}([q,[p,\rho]] + [p,[q,\rho]]).$$

Here we used the notations:

$$D_{qq}=\frac{\hbar}{2}\sum_{j=1,2}|a_j|^2, D_{pp}=\frac{\hbar}{2}\sum_{j=1,2}|b_j|^2,$$
$$D_{pq}=D_{qp}=-\frac{\hbar}{2}\mathrm{Re}\sum_{j=1,2}a_j^*b_j, \lambda=-\mathrm{Im}\sum_{j=1,2}a_j^*b_j,$$

where $D_{pp}, D_{qq}$ and $D_{pq}$ are the diffusion coefficients and $\lambda$ the friction constant. They satisfy the following fundamental constraints:

*i)* $D_{pp}>0$

*ii)* $D_{qq}>0$

*iii)* $D_{pp}D_{qq}-{D_{pq}}^2\geq\lambda^2\hbar^2/4.$

We introduce the following notations:

$$\sigma_q(t)=\mathrm{Tr}(\rho(t)q),$$
$$\sigma_p(t)=\mathrm{Tr}(\rho(t)p),$$
$$\sigma_{qq}=\mathrm{Tr}(\rho(t)q^2)-\sigma_q^2(t),$$
$$\sigma_{pp}=\mathrm{Tr}(\rho(t)p^2)-\sigma_p^2(t),$$
$$\sigma_{pq}(t)=\mathrm{Tr}(\rho(t)\frac{pq+qp}{2})-\sigma_p(t)\sigma_q(t).$$

In the Heisenberg picture the master equation has the following symmetric form:

$$\frac{d\widetilde{\Phi}_t(A)}{dt}=\widetilde{L}(\widetilde{\Phi}_t(A))=\frac{i}{\hbar}[H_0,\widetilde{\Phi}_t(A)]-\frac{i}{2\hbar}(\lambda+\mu)([\widetilde{\Phi}_t(A),q]p+p[\widetilde{\Phi}_t(A),q])$$
$$+\frac{i}{2\hbar}(\lambda-\mu)(q[\widetilde{\Phi}_t(A),p]+[\widetilde{\Phi}_t(A),p]q)-\frac{D_{pp}}{\hbar^2}[q,[q,\widetilde{\Phi}_t(A)]]$$
$$-\frac{D_{qq}}{\hbar^2}[p,[p,\widetilde{\Phi}_t(A)]]+\frac{D_{pq}}{\hbar^2}([p,[q,\widetilde{\Phi}_t(A)]]+[q,[p,\widetilde{\Phi}_t(A)]]).$$

Denoting by $A$ any selfadjoint operator we have

$$\sigma_A(t)=\mathrm{Tr}(\rho(t)A),\ \sigma_{AA}(t)=\mathrm{Tr}(\rho(t)A^2)-\sigma_A^2(t).$$

It follows that

$$\frac{d\sigma_A(t)}{dt}=\mathrm{Tr}L(\rho(t))A=\mathrm{Tr}\rho(t)\widetilde{L}(A)$$

and

$$\frac{d\sigma_{AA}(t)}{dt} = \mathrm{Tr}L(\rho(t))A^2 - 2\frac{d\sigma_A(t)}{dt}\sigma_A(t) = \mathrm{Tr}\rho(t)\widetilde{L}(A^2) - 2\sigma_A(t)\mathrm{Tr}\rho(t)\widetilde{L}(A).$$

An important consequence of the precise version of solvability condition is the fact that when $A$ is put equal to $p$ or $q$, then $d\sigma_p(t)/dt$ and $d\sigma_q(t)/dt$ are functions only of $\sigma_p(t)$ and $\sigma_q(t)$ and $d\sigma_{pp}(t)/dt, d\sigma_{qq}(t)/dt$ and $d\sigma_{pq}(t)/dt$ are functions only of $\sigma_{pp}(t), \sigma_{qq}(t)$ and $\sigma_{pq}(t)$. This fact allows an immediate determination of the functions of time $\sigma_p(t), \sigma_q(t), \sigma_{pp}(t), \sigma_{qq}(t)$, $\sigma_{pq}(t)$. Indeed we obtain:

$$\frac{d\sigma_q(t)}{dt} = -(\lambda - \mu)\sigma_q(t) + \frac{1}{m}\sigma_p(t),$$

$$\frac{d\sigma_p(t)}{dt} = -m\omega^2\sigma_q(t) - (\lambda + \mu)\sigma_p(t)$$

and

$$\frac{d\sigma_{qq}(t)}{dt} = -2(\lambda - \mu)\sigma_{qq}(t) + \frac{2}{m}\sigma_{pq}(t) + 2D_{qq},$$

$$\frac{d\sigma_{pp}}{dt} = -2(\lambda + \mu)\sigma_{pp}(t) - 2m\omega^2\sigma_{pq}(t) + 2D_{pp},$$

$$\frac{d\sigma_{pq}(t)}{dt} = -m\omega^2\sigma_{qq}(t) + \frac{1}{m}\sigma_{pp}(t) - 2\lambda\sigma_{pq}(t) + 2D_{pq}.$$

The integration of these systems of equations of motion is straightforward. There are two cases: *a)* $\mu > \omega$ (overdamped) and *b)* $\mu < \omega$ (underdamped). In the case *a)* with the notation $\nu^2 = \mu^2 - \omega^2$ we obtain:

$$\sigma_q(t) = e^{-\lambda t}((\cosh\nu t + \frac{\mu}{\nu}\sinh\nu t)\sigma_q(0) + \frac{1}{m\nu}\sinh\nu t\sigma_p(0)),$$

$$\sigma_p(t) = e^{-\lambda t}(-\frac{m\omega^2}{\nu}\sinh\nu t\sigma_q(0) + (\cosh\nu t - \frac{\mu}{\nu}\sinh\nu t)\sigma_p(0)).$$

If $\lambda > \nu$, then $\sigma_q(\infty) = \sigma_p(\infty) = 0$. If $\lambda < \nu$, then $\sigma_q(\infty) = \sigma_p(\infty) \to \infty$. In the case *b)* with the notation $\Omega^2 = \omega^2 - \mu^2$, we obtain:

$$\sigma_q(t) = e^{-\lambda t}((\cos\Omega t + \frac{\mu}{\Omega}\sin\Omega t)\sigma_q(0) + \frac{1}{m\Omega}\sin\Omega t\sigma_p(0)),$$

$$\sigma_p(t) = e^{-\lambda t}(-\frac{m\omega^2}{\Omega}\sin\Omega t\sigma_q(0) + (\cos\Omega t - \frac{\mu}{\Omega}\sin\Omega t)\sigma_p(0))$$

and $\sigma_q(\infty) = \sigma_p(\infty) = 0$.

In order to integrate the system of equations for the covariances it is convenient to consider the vector

$$X(t) = \left( m\omega\sigma_{qq}(t)\frac{1}{m\omega}\sigma_{pp}(t)\sigma_{pq}(t) \right).$$

Introducing the following matrices: in the overdamped case ($\mu > \omega,\ \nu^2 = \mu^2 - \omega^2$)

$$T = \frac{1}{2\nu}\left( \mu + \nu\mu - \nu 2\omega\mu - \nu\mu + \nu 2\omega - \omega - \omega - 2\mu \right),$$

$$K = \left( -2(\lambda - \nu)000 - 2(\lambda + \nu)000 - 2\lambda \right),$$

and in the underdamped case ($\mu < \omega,\ \Omega^2 = \omega^2 - \mu^2$)

$$T = \frac{1}{2i\Omega}\left( \mu + i\Omega\mu - i\Omega 2\omega\mu - i\Omega\mu + i\Omega 2\omega - \omega - \omega - 2\mu \right),$$

$$K = \left( -2(\lambda - i\Omega)000 - 2(\lambda + i\Omega)000 - 2\lambda \right),$$

the solution can be written in the form [33, 34]

$$X(t) = (Te^{Kt}T)(X(0) - X(\infty)) + X(\infty).$$

Between the asymptotic values of $\sigma_{qq}(t), \sigma_{pp}(t), \sigma_{pq}(t)$ and the diffusion coefficients $D_{qq}, D_{pp}, D_{pq}$ there exist the following connection, which is the same for both cases, underdamped and overdamped:

$$D_{qq} = (\lambda - \mu)\sigma_{qq}(\infty) - \frac{1}{m}\sigma_{pq}(\infty),$$

$$D_{pp} = (\lambda + \mu)\sigma_{pp}(\infty) + m\omega^2\sigma_{pq}(\infty),$$

$$D_{pq} = \frac{1}{2}(m\omega^2\sigma_{qq}(\infty) - \frac{1}{m}\sigma_{pp}(\infty) + 2\lambda\sigma_{pq}(\infty)).$$

These relations show that the asymptotic values $\sigma_{qq}(\infty), \sigma_{pp}(\infty), \sigma_{pq}(\infty)$ do not depend on the initial values $\sigma_{qq}(0), \sigma_{pp}(0), \sigma_{pq}(0)$.

If the asymptotic state is a Gibbs state ($T$ denotes the temperature of the thermal bath)

$$\rho_G(\infty) = e^{-\frac{H_0}{kT}}/\mathrm{Tr}(e^{-\frac{H_0}{kT}}),$$

then

$$\sigma_{qq}(\infty) = \frac{\hbar}{2m\omega}\coth\frac{\hbar\omega}{2kT},\ \ \sigma_{pp}(\infty) = \frac{\hbar m\omega}{2}\coth\frac{\hbar\omega}{2kT},\ \ \sigma_{pq}(\infty) = 0$$

and

$$D_{pp} = \frac{\lambda+\mu}{2}\hbar m\omega \coth\frac{\hbar\omega}{2kT},\ D_{qq} = \frac{\lambda-\mu}{2}\frac{\hbar}{m\omega}\coth\frac{\hbar\omega}{2kT},\ D_{pq} = 0$$

and the fundamental constraints are satisfied only if $\lambda > \mu$ and

$$(\lambda^2 - \mu^2)\coth^2\frac{\hbar\omega}{2kT} \geq \lambda^2.$$

If the initial state is the ground state of the harmonic oscillator, then

$$\sigma_{qq}(0) = \frac{\hbar}{2m\omega},\ \sigma_{pp}(0) = \frac{m\hbar\omega}{2},\ \sigma_{pq}(0) = 0.$$

The explicit time dependence of $\sigma_{qq}(t), \sigma_{pp}(t), \sigma_{pq}(t)$ can be given for both under- and overdamped cases if we have the matrix elements of $Te^{Kt}T$. In the overdamped case ($\mu > \omega,\ \nu^2 = \mu^2 - \omega^2$) we have (in this case the restriction $\lambda > \nu$ is necessary):

$$Te^{Kt}T = \frac{e^{-2\lambda t}}{2\nu^2}\left(a_{11}a_{12}a_{13}a_{21}a_{22}a_{23}a_{31}a_{32}a_{33}\right),$$

with
$a_{11} = (\mu^2+\nu^2)\cosh 2\nu t + 2\mu\nu\sinh 2\nu t - \omega^2,$
$a_{12} = (\mu^2-\nu^2)\cosh 2\nu t - \omega^2,$
$a_{13} = 2\omega(\mu\cosh 2\nu t + \nu\sinh 2\nu t - \mu),$
$a_{21} = (\mu^2-\nu^2)\cosh 2\nu t - \omega^2,$
$a_{22} = (\mu^2+\nu^2)\cosh 2\nu t - 2\mu\nu\sinh 2\nu t - \omega^2,$
$a_{23} = 2\omega(\mu\cosh 2\nu t - \nu\sinh 2\nu t - \mu),$
$a_{31} = -\omega(\mu\cosh 2\nu t + \nu\sinh 2\nu t - \mu),$
$a_{32} = -\omega(\mu\cosh 2\nu t - \nu\sinh 2\nu t - \mu),$
$a_{33} = -2(\omega^2\cosh 2\nu t - \mu^2).$

In the underdamped case ($\mu < \omega,\ \Omega^2 = \omega^2 - \mu^2$) we have

$$Te^{Kt}T = -\frac{e^{-2\lambda t}}{2\Omega^2}\left(b_{11}b_{12}b_{13}b_{21}b_{22}b_{23}b_{31}b_{32}b_{33}\right)$$

with
$b_{11} = (\mu^2-\Omega^2)\cos 2\Omega t - 2\mu\Omega\sin 2\Omega t - \omega^2,$
$b_{12} = (\mu^2+\Omega^2)\cos 2\Omega t - \omega^2,$
$b_{13} = 2\omega(\mu\cos 2\Omega t - \Omega\sin 2\Omega t - \mu),$
$b_{21} = (\mu^2+\Omega^2)\cos 2\Omega t - \omega^2,$
$b_{22} = (\mu^2-\Omega^2)\cos 2\Omega t + 2\mu\Omega\sin 2\Omega t - \omega^2,$

$$b_{23} = 2\omega(\mu \cos 2\Omega t + \Omega \sin 2\Omega t - \mu),$$
$$b_{31} = -\omega(\mu \cos 2\Omega t - \Omega \sin 2\Omega t - \mu),$$
$$b_{32} = -\omega(\mu \cos 2\Omega t + \Omega \sin 2\Omega t - \mu),$$
$$b_{33} = -2(\omega^2 \cos 2\Omega t - \mu^2).$$

**Acknowledgments.** The author acknowledges the financial support received from the Romanian Ministry of Education and Research, through the Projects CNCS-UEFISCDI PN-II-ID-PCE-2011-3-0083 and PN 09 37 01 02/2009.

## References

[1] M. A. Nielsen, I. L. Chuang. *Quantum Computation and Quantum Information.* Cambridge University Press, Cambridge, 2010.

[2] R. Horodecki, P. Horodecki, M. Horodecki, K. Horodecki. Quantum entanglement. *Rev. Mod. Phys.* 81:865, 2009.

[3] A. Datta, A. Shaji, C. M. Caves. Quantum discord and the power of one qubit. *Phys. Rev. Lett.* 100:050502, 2008.

[4] W. H. Zurek. Einselection and decoherence from an information theory perspective. *Annalen der Physik (Leipzig).* 9:855, 2000.

[5] H. Ollivier, W. H. Zurek. Quantum discord: a measure of the quantumness of correlations. *Phys. Rev. Lett.* 88:017901, 2001.

[6] L. Henderson, V. Vedral. Classical, quantum and total correlations. *J. Phys. A: Math. Gen.* 34:6899, 2001.

[7] S. L. Braunstein, P. van Loock. Quantum information with continuous variables. *Rev. Mod. Phys.* 77:513, 2005.

[8] S. Olivares, M. G. A. Paris, A. R. Rossi. Optimized teleportation in Gaussian noisy channels. *Phys. Lett. A.* 319:32, 2003.

[9] J. S. Prauzner-Bechcicki. Two-mode squeezed vacuum state coupled to the common thermal reservoir. *J. Phys. A: Math. Gen.* 37:L173, 2004.

[10] P. J. Dodd, J. J. Halliwell. Disentanglement and decoherence by open system dynamics. *Phys. Rev. A.* 69:052105, 2004.

[11] G. Adesso, A. Serafini, F. Illuminati, Extremal entanglement and mixedness in continuous variable systems. *Phys. Rev. A.* 70:022318, 2004.

[12] A. V. Dodonov, V. V. Dodonov, S. S. Mizrahi. Separability dynamics of two-mode Gaussian states in parametric conversion and amplification. *J. Phys. A: Math. Gen.* 38:683, 2005.

[13] F. Benatti, R. Floreanini. Entangling oscillators through environment noise. *J. Phys. A: Math. Gen.* 39:2689, 2006.

[14] D. McHugh, M. Ziman, V. Buzek. Entanglement, purity, and energy: two qubits versus two modes. *Phys. Rev. A.* 74:042303, 2006.

[15] S. Maniscalco, S. Olivares, M. G. A. Paris. Entanglement oscillations in non-Markovian quantum channels. *Phys. Rev. A.* 75:062119, 2007.

[16] J. H. An, W. M. Zhang. Non-Markovian entanglement dynamics of noisy continuous-variable quantum channels. *Phys. Rev. A.* 76:042127, 2007.

[17] A. Isar, W. Scheid. Quantum decoherence and classical correlations of the harmonic oscillator in the Lindblad theory. *Physica A.* 373:298, 2007.

[18] A. Isar. Quantum fidelity for Gaussian states describing the evolution of open systems. *Eur. J. Phys. Special Topics.* 160:225, 2008.

[19] A. Isar. Entanglement dynamics of two-mode Gaussian states in a thermal environment. *J. Russ. Laser Res.* 30:458, 2009.

[20] A. Isar. Entanglement and mixedness in open systems with continuous variables. *J. Russ. Laser Res.* 31:182, 2010.

[21] A. Isar. Dynamics of quantum entanglement in Gaussian open systems. *Phys. Scr.* 82:038116, 2010.

[22] A. Isar. Continuous variable entanglement in open quantum dynamics. *Phys. Scr. T.* 140:014023, 2010.

[23] A. Isar. Quantum entanglement and quantum discord of two-mode Gaussian states in a thermal environment. *Open Sys. Inf. Dynamics.* 18:175, 2011.

[24] A. Isar. Entanglement and discord in two-mode Gaussian open quantum systems. *Phys. Scr. T.* 147:014015, 2012.

[25] A. Peres. Separability criterion for density matrices. *Phys. Rev. Lett.* 77:1413, 1996.

[26] R. Simon. Peres-Horodecki separability criterion for continuous variable systems. *Phys. Rev. Lett.* 84:2726, 2000.

[27] R. S. Ingarden, A. Kossakowski. On the connection of nonequilibrium information thermodynamics with non-hamiltonian quantum mechanics of open systems. *Ann. Phys. (N.Y.).* 89:451,1975.

[28] E. B. Davies. *Quantum Theory of Open Systems.* Academic Press, New York, 1976.

[29] G. Lindblad. On the generators of quantum dynamical semigroups. *Commun. Math. Phys.* 48:119, 1976.

[30] A. Kossakowski. On quantum statistical mechanics of non-Hamiltonian systems. *Rep. Math. Phys.* 3:247, 1972.

[31] V. Gorini, A. Kossakowski, E. C. G. Sudarshan. Completely positive dynamical semigroups of N-level systems. *J. Math. Phys.* 17:821, 1976.

[32] F. Benatti, R. Floreanini, M. Piani. Environment induced entanglement in Markovian dissipative dynamics. *Phys. Rev. Lett.* 91:070402, 2003.

[33] A. Sandulescu, H. Scutaru. Open quantum systems and the damping of collective modes in deep inelastic collisions. *Ann. Phys. (N.Y.).* 173:277, 1987.

[34] A. Isar, A. Sandulescu, H. Scutaru, E. Stefanescu, W. Scheid. Open quantum systems. *Int. J. Mod. Phys. E.* 3:635, 1994.

[35] P. Talkner. The failure of the quantum regression hypothesis. *Ann. Phys. (N.Y.).* 167:390, 1986.

[36] A. Sandulescu, H. Scutaru, W. Scheid. Open quantum system of two coupled harmonic oscillators for application in deep inelastic heavy ion collisions. *J. Phys. A: Math. Gen.* 20:2121, 1987.

[37] M. Horodecki, P. Horodecki, R. Horodecki. Separability of mixed states: necessary and sufficient conditions. *Phys. Lett. A.* 223:1, 1996.

[38] L. M. Duan, G. Giedke, J. I. Cirac, P. Zoller. Inseparability criterion for continuous variable systems. *Phys. Rev. Lett.* 84:2722, 2000.

[39] A. Isar. Asymptotic entanglement in open quantum systems. *Int. J. Quantum Inf.* 6:689, 2008.

[40] A. Isar. Entanglement generation and evolution in open quantum systems. *Open Sys. Inf. Dynamics.* 16:205, 2009.

[41] A. Isar. Entanglement in two-mode continuous variable open quantum systems. *Phys. Scr. T.* 143:014012, 2011.

[42] A. Isar. Decoherence and asymptotic entanglement in open quantum dynamics. *J. Russ. Laser Res.* 28:439, 2007.

[43] P. Marian, T. A. Marian, H. Scutaru. Bures distance as a measure of entanglement for two-mode squeezed thermal states. *Phys. Rev. A.* 68:062309, 2003.

[44] A. Isar. Entanglement in open quantum dynamics. *Phys. Scr. T.* 135:014033, 2009.

[45] S. Mancini. Markovian feedback to control continuous-variable entanglement. *Phys. Rev. A.* 73:010304, 2006.

[46] S. Mancini, H. M. Wiseman. Optimal control of entanglement via quantum feedback. *Phys. Rev. A.* 75:012330, 2007.

[47] B. Groisman, S. Popescu, A. Winter. Quantum, classical, and total amount of correlations in a quantum state. *Phys. Rev. A.* 72:032317, 2005.

[48] P. Giorda, M. G. A. Paris. Gaussian quantum discord. *Phys. Rev. Lett.* 105:020503, 2010.

[49] G. Adesso, A. Datta. Quantum versus classical correlations in Gaussian states. *Phys. Rev. Lett.* 105:030501, 2010.

[50] T. Yu, J. H. Eberly. Finite-time disentanglement via spontaneous emission. *Phys. Rev. Lett.* 93:140404, 2004.

[51] *Quantum Information with Continuous Variables.* Ed. by S. L. Braunstein, A. K. Pati. Kluwer, Dordrecht, 2003.

[52] *Lectures on Quantum Information.* Ed. by D. Bruss, G. Leuchs. Wiley-VCH, Berlin, 2006.

ISSN 2066-6594 Ann. Acad. Rom. Sci.
Ser. Math. Appl.
Vol. 6, No. 2/2014

# GLOBAL SOLUTION FOR THE COAGULATION EQUATION OF WATER DROPS IN FALL WITH THE HORIZONTAL WIND*

Hanane Belhireche † Mohamed Zine Aissaoui ‡
Fateh Ellaggoune §

**Abstract**

We consider the integro-differential equation describing the coagulation process of water drops falling in the air in a three-dimensional domain with presence of a horizontal wind. Under suitable hypothesis and some conditions we prove the existence of the stationary solution thus the global solution using the techniques developed in [10] and [2].

**MSC**: 35Q35, 76N10.

**keywords:** Equation of air motion, integro-differential equation, global solution.

## 1 Introduction

We consider the equation which describes the displacement of drops by the gravitational force and by the horizontal wind as well as the coagulation

*Accepted for publication in revised form on March 21-st, 2014

†hanane.belhireche@gmail.com, Laboratory of Applied Mathematics and Modeling, University 8 mai 1945, P.O. Box 401, Guelma 24000, Algeria.

‡aissaouizine@gmail.com, Laboratory of Applied Mathematics and Modeling, University 8 mai 1945, P.O. Box 401, Guelma 24000, Algeria.

§fellaggoune@gmail.com, Laboratory of Applied Mathematics and Modeling, University 8 mai 1945, P.O. Box 401, Guelma 24000, Algeria.

process. From a mathematical point of view, it is about the Smoluchowski coagulation equation (see [18], [13], [19]) with the displacement of drops determined by their mass; it is an integro-differential equation for an unknown function $\sigma = \sigma(m,t,x,y,z)$, that represents the density (compared to the air volume) of the liquid water contained in the drops of mass $m$ at time $t$ and at position $(x,y,z) \in \mathbb{R}^3$. The air motion in consideration is a horizontal wind in the direction of the $x$ axis which depends on $y$ (i.e $\overline{v} = \overline{v}(y)$). In [10] the authors proved the existence of the stationary solution with presence of a constant horizontal wind whereas in [2] the authors proved the existence and the uniqueness of the global solution of the same equation in a domain with one-dimensional space. In this work, we prove the existence and the uniqueness of the global solution in a three-dimensional domain with presence of a horizontal wind and with initial and boundary conditions (entry conditions) in a suitable spaces.

From a technical point of view, this work uses the techniques developed in [10] and [2], in particular the introduction of the curves family on which we consider the coagulation integral operator, and their properties, and on the construction of "cone of dependence" for the solution.

## 2 Position of the problem

Let's consider the domain $\mathbb{R}^2 \times [0,1]$, which represents a "horizontal" area in which the drops move due to the gravitational force and with the wind. Let's indicate by $\sigma(m,t,x,y,z)$ the density of the water liquid contained in the drops of mass $m$ at the point $(x,y,z) \in \mathbb{R}^2 \times ]0,1[$ at the moment $t \in \mathbb{R}_+$.

In the same way to [10] and [2], we suppose that the drops undergo the coagulation process and at the same time move by the gravitational force and the air motion in which they are undergoing the friction effect with this last; these considerations bring us to the following equation (see [1], [16], [10], [2])

$$\partial_t \sigma(m,t,x,y,z) + \nabla_{(x,y,z)} \cdot (\sigma(m,t,x,y,z)u(m)) = \tag{1}$$

$$= \frac{m}{2}\int_0^m \beta(m-m',m')\sigma(m',t,x,y,z)\sigma(m-m',t,x,y,z)dm' +$$

$$-m\int_0^\infty \beta(m,m')\sigma(m,t,x,y,z)\sigma(m',t,x,y,z)dm',$$

where $\nabla_{(x,y,z)} = (\partial_x, \partial_y, \partial_z)$, while $\beta(m_1,m_2)$ represents the probability of meeting between a drop with mass $m_1$ and another with mass $m_2$, and $u(m)$

indicates the velocity of drops with mass $m$. We suppose that

$$\beta(\cdot,\cdot) \in C(\mathbb{R}_+ \times \mathbb{R}_+), \quad \beta(m_1, m_2) \geq 0 \quad \forall (m_1, m_2) \in \mathbb{R}_+ \times \mathbb{R}_+,$$
$$\beta(m_1, m_2) = \beta(m_2, m_1)$$

and we admit that $u = u(m)$ is given by

$$u = u(m) = \Big(\overline{v}(y), 0, -\frac{g}{\alpha(m)}\Big), \tag{2}$$

where $\overline{v}(y)$ is the air velocity, $g$ is a positive constant representing the gravitational acceleration and $\alpha(m)$ is the friction coefficient between drops and air. The relation (2) corresponds, in a good approximation, at the real velocity of drops in the atmosphere (see for example [17], [1], [16]).

As the small drops evaporate immediately due to the very high curve of surface (see [15], [8]) and on the other hand the very large drops fragment due to the friction with surrounding air, we consider that the drops are absent apart from an interval $[\overline{m}_a, \overline{m}_A]$ and consequently the function $\sigma$ verifies

$$\sigma(m) = 0 \qquad \text{for} \quad m \in [0, \overline{m}_a[ \cup ]\overline{m}_A, \infty[.$$

This permit us to define the functions $\alpha(\cdot)$, $\beta(\cdot,\cdot)$ such that

$$0 < \inf_{m \in \mathbb{R}_+} \alpha(m) \leq \sup_{m \in \mathbb{R}_+} \alpha(m) < \infty$$

and

$$\beta(m_1, m_2) = 0 \qquad \text{for} \quad m_1 + m_2 > \overline{m}_A.$$

We pose

$$\overline{\alpha}_0 = \sup_{m \in \mathbb{R}_+} \alpha(m). \tag{3}$$

# 3 Stationary solution

We consider the following stationary equation of (1)

$$\nabla_{(x,y,z)} \cdot (\sigma(m, x, y, z) u(m)) = \tag{4}$$
$$= \frac{m}{2} \int_0^m \beta(m - m', m') \sigma(m', x, y, z) \sigma(m - m', x, y, z) dm' +$$
$$-m \int_0^\infty \beta(m, m') \sigma(m, x, y, z) \sigma(m', x, y, z) dm'$$

with the boundary condition (entry condition)

$$\sigma(m, x, y, 1) = \overline{\sigma}(m, x, y). \tag{5}$$

### 3.1 Preliminaries

To solve the equation (4) with the condition (5), we will use the idea to transform it into an ordinary differential equation, by introducing the change of variables $(m,x,y,z) \mapsto (\tilde{m},\xi,\tilde{y},\tilde{z})$ defined by

$$\begin{cases} \tilde{m} = m, \\ \xi = x - \overline{v}(y)\frac{\alpha(m)}{g}(1-z), \\ \tilde{y} = y, \\ \tilde{z} = z \end{cases} \tag{6}$$

and let us define

$$\tilde{\sigma}(\tilde{m},\xi,\tilde{y},\tilde{z}) = \sigma(m,x,y,z) = \sigma(m,\xi+\overline{v}(y)\frac{\alpha(m)}{g}(1-z),y,z).$$

In the following, we will simply write $m$, $y$, $z$ and $\sigma(m,\xi,y,z)$ instead of $\tilde{m}$, $\tilde{y}$, $\tilde{z}$ and $\tilde{\sigma}(\tilde{m},\xi,\tilde{y},\tilde{z})$, thus, the equation (4) will be

$$\frac{\partial}{\partial z}\sigma(m,\xi,y,z) = \tag{7}$$

$$= -\frac{m\alpha(m)}{2g}\int_0^m \beta(m-m',m')\sigma(m',\eta(m,m',\xi,y,z),y,z)\times$$
$$\times\sigma(m-m',\eta(m,m-m',\xi,y,z),y,z)dm'+$$
$$+\frac{m\alpha(m)}{g}\int_0^\infty \beta(m,m')\sigma(m,\xi,y,z)\sigma(m',\eta(m,m',\xi,y,z),y,z)dm',$$

where

$$\eta(m,m',\xi,y,z) = \xi + \overline{v}(y)\frac{\alpha(m)-\alpha(m')}{g}(1-z)$$

and the condition (5) will be:

$$\sigma(m,\xi,y,1) = \overline{\sigma}(m,\xi,y). \tag{8}$$

Consequently we will reformulate the equation (7) into an ordinary differential equation in a Banach space (or in a Frechet space). To suitably treat the integral operator in a functional framework, we introduce, for each fixed $y \in \mathbb{R}$, $z \in [0,1]$, the curves family given by:

$$\gamma_\tau = \gamma_{\tau,y,z} = \{(m,\xi) \in \mathbb{R}_+ \times \mathbb{R} \mid \xi = \tau - \overline{v}(y)\frac{\alpha(m)}{g}(1-z)\},\ \tau \in \mathbb{R}. \tag{9}$$

This curves family $\gamma_\tau$ is similar to that used in [10], however this last depends of $y$ .

In a similar way to [10] we define a measure $\mu_\gamma$ on the curves $\gamma_\tau$. More precisely, indicating by $P_{\mathbb{R}_+}$ the projection of $\gamma_\tau$ on $\mathbb{R}_+$, we define the measurable sets of $\gamma_\tau$ and the measure $\mu_\gamma$ on $\gamma_\tau$ by the relations

i) $A' \subset \gamma_\tau$ is measurable if and only if $P_{\mathbb{R}_+}A'$ is measurable according to Lebesgue on $\mathbb{R}_+$,

ii) $\mu_\gamma(A') = \mu_{L,\mathbb{R}_+}(P_{\mathbb{R}_+}A')$, where $\mu_{L,\mathbb{R}_+}(\cdot)$ is the Lebesgue's measure on $\mathbb{R}_+$.

As the curves $\gamma_\tau$, $\tau \in \mathbb{R}$, are parallel, it is seen that the projection $P_{\mathbb{R}_+}$ and the measure $\mu_\gamma(\cdot)$ do not depend on $\tau \in \mathbb{R}$.

We remember that the measure $\mu_\gamma(\cdot)$ has the same properties with those proved in [10], indeed we have the following lemmas.

**Lemma 1** *Let $A$ a measurable set (according to Lebesgue) on $\mathbb{R}_+ \times \mathbb{R}$. We pose*

$$A_\tau = \{\, m \in \mathbb{R}_+ \,|\, \exists \xi \in \mathbb{R} \text{ such that } (m,\xi) \in \gamma_\tau \cap A \,\},$$
$$A_m = \{\tau \in \mathbb{R} \,|\, \exists \xi \in \mathbb{R} \text{ such that } (m,\xi) \in \gamma_\tau \cap A\}.$$

*Then we have*

$$\mu_{L,\mathbb{R}_+\times\mathbb{R}}(A) = \tilde{\mu}(A) = \int_{-\infty}^{\infty} \mu_\gamma(A_\tau)d\tau = \int_{\gamma_0} \mu_{L,\mathbb{R}}(A_m)\mu_\gamma(dm) =$$
$$= \int_0^\infty \mu_{L,\mathbb{R}}(A_m)dm. \tag{10}$$

(We indicate by $dm$, $d\tau$, $d\xi$ etc... instead of $\mu_{L,\mathbb{R}_+}(dm)$, $\mu_{L,\mathbb{R}}(d\tau)$, $\mu_{L,\mathbb{R}}(d\xi)$ etc...).

**Lemma 2** *Let $\sigma(m,\xi) \in L^1(\mathbb{R}_+ \times \mathbb{R})$. Then, for almost any $\tau \in \mathbb{R}$ the restriction of $\sigma(m,\xi)$ to $\gamma_\tau$ belongs to $L^1(\gamma_\tau, \mu_\gamma)$.*

**Lemma 3** *Let $\sigma(m,\xi) \in L^1(\mathbb{R}_+ \times \mathbb{R})$. Then we have*

$$\int_{\mathbb{R}_+\times\mathbb{R}} \sigma(m,\xi)dmd\xi = \int_{\mathbb{R}_+\times\mathbb{R}} \sigma(m,\xi)d\tilde{\mu} =$$
$$= \int_{-\infty}^{\infty}\Big(\int_{\gamma_\tau} \sigma(m,\xi)\mu_\gamma(dm)\Big)d\tau = \int_{\gamma_0}\Big(\int_{-\infty}^{\infty} \sigma(m,\xi(m,\tau))d\tau\Big)\mu_\gamma(dm) =$$
$$= \int_0^\infty\Big(\int_{-\infty}^{\infty} \sigma(m,\xi)d\xi\Big)dm = \int_{-\infty}^{\infty}\Big(\int_0^\infty \sigma(m,\xi)dm\Big)d\xi,$$

*where $\xi(m,\tau) = \tau - \overline{v}(y)\,\frac{\alpha(m)}{g}(1-z)$.*

**Lemma 4** *Let $f$ and $g$ two functions belonging in $L^1(\gamma_\tau,\mu_\gamma)$. We pose*

$$(f*g)(m)=\int_{\gamma_\tau} f(m-m')g(m')\mu_\gamma(dm').$$

*Then we have $f*g\in L^1(\gamma_\tau,\mu_\gamma)$ and*

$$\|f*g\|_{L^1(\gamma_\tau,\mu_\gamma)}\le \|f\|_{L^1(\gamma_\tau,\mu_\gamma)}\|g\|_{L^1(\gamma_\tau,\mu_\gamma)}.$$

For the proof of this lemmas see [10].

We pose

$$\tau(m,\xi,y,z)=\xi+\overline{v}(y)\,\frac{\alpha(m)}{g}(1-z),\qquad \gamma_\tau^{[0,m]}=\gamma_\tau\cap[0,m]\times\mathbb{R}. \tag{11}$$

Then we can write the equation (7) in the form

$$\frac{\partial}{\partial z}\sigma(z)=F_z(\sigma(z)),\qquad \sigma(z)=\sigma(\cdot,\cdot,\cdot,z) \tag{12}$$

with

$$F_z(\sigma(z))=F_z(\sigma(z))(m,\xi,y)=$$

$$=-\frac{m\alpha(m)}{2g}\int_{\gamma^{[0,m]}_{\tau(m,\xi,y,z)}}\beta(m-m',m')\sigma(m',\eta',y,z)\sigma(m-m',\eta'',y,z)\mu_\gamma(dm')+$$

$$+\frac{m\alpha(m)}{g}\int_{\gamma_{\tau(m,\xi,y,z)}}\beta(m,m')\sigma(m',\eta',y,z)\sigma(m,\xi,y,z)\mu_\gamma(dm'),$$

where $\eta'$ and $\eta''$ are defined such that

$$(m',\eta')\in\gamma_{\tau(m,\xi,y,z)},\quad (m-m',\eta'')\in\gamma_{\tau(m,\xi,y,z)}.$$

### 3.2 Existence and uniqueness of the solution with the data in $L^1$

To prove the existence and the uniqueness for the solution of the equation (12) with the condition (8), we suppose that:

$$\overline{\sigma}(\cdot,\cdot,\cdot)\in L^1(\mathbb{R}_+\times\mathbb{R}^2)\cap L^\infty(\mathbb{R}_+\times\mathbb{R}^2), \tag{13}$$

$$\overline{\sigma}(m,\xi,y)\ge 0\quad a.e.\quad \text{in } \mathbb{R}_+\times\mathbb{R}^2, \tag{14}$$

$$supp(\overline{\sigma})\subset[\overline{m}_a,\overline{m}_A]\times\mathbb{R}^2, \tag{15}$$

$$\|\overline{\sigma}\|_{L^\infty(\mathbb{R}_+\times\mathbb{R}^2)} < \frac{1}{M_1(\overline{m}_A - \overline{m}_a)}, \tag{16}$$

where

$$M_1 = \sup_{2\overline{m}_a \le m \le \overline{m}_A, \overline{m}_a \le m' \le m - \overline{m}_a} \frac{m\alpha(m)}{2g}\beta(m-m', m'). \tag{17}$$

Then we have the following result.

**Proposition 1** *If* $\overline{\sigma}(m,\xi,y)$ *satisfies the conditions* (13)–(16)*, then the equation* (12) *with the condition* (8) *admits one and only one solution* $\sigma$ *verifying*

$$\sigma \in C([0,1]; L^1(\mathbb{R}_+\times\mathbb{R}^2)) \times L^\infty(\mathbb{R}_+\times\mathbb{R}^2\times[0,1]). \tag{18}$$

**Proof.** As in the equation (12) neither the derivative nor the integral compared to $y$ arise, this circumstance implies that the equation for each fixed $y\in\mathbb{R}$ can be solved independently. That permit us to consider (12), (8) separately for each $y\in\mathbb{R}$. Therefore, we pose $\sigma(m,\xi,z)=\sigma(m,\xi,y,z)$, $\overline{\sigma}(m,\xi)=\overline{\sigma}(m,\xi,y)$ and we write $\overline{v}$ instead of $\overline{v}(y)$, the proposition is proved in the same meaner as proposition 5.1 in [10]. $\square$

### 3.3 Existence and uniqueness of the solution with the data in $L^\infty$

To prove the existence and the uniqueness of the solution for (12), (8) in a general case, we will use the "cone of dependance" property.

Let $\omega$ in $\mathbb{R}_+\times\mathbb{R}^2$ a measurable set such that $0 < mes(\omega) < \infty$, we define

$$D[\omega] = \bigcup_{(m,\xi,y)\in\omega} D_{(m,\xi,y)}, \tag{19}$$

where

$$D_{(m,\xi,y)} = \bigcup_{0\le z\le 1}\Big(\bigcup_{\tau_-(m,\xi,y,z)\le\tau\le\tau_+(m,\xi,y)} \gamma_{\tau,y,z}\Big) = \tag{20}$$

$$= \{(m',\eta',y',z') \in \mathbb{R}_+\times\mathbb{R}^2\times[0,1] \ / \ \eta' = \tau - \overline{v}(y')\frac{\alpha(m')}{g}(1-z'),$$

$$y'=y,\ \tau_-(m,\xi,y',z') \le \tau \le \tau_+(m,\xi,y')\}$$

with

$$\begin{cases} \tau_+(m,\xi,y) = \tau(m,\xi,y,0) = \xi + \overline{v}(y)\frac{\alpha(m)}{g}, \\ \tau_-(m,\xi,y,z) = \tau_+(m,\xi,y) - \overline{v}(y)\frac{\overline{\alpha}_0}{g}z = \xi + \overline{v}(y)\frac{\alpha(m)}{g} - \overline{v}(y)\frac{\overline{\alpha}_0}{g}z. \end{cases} \tag{21}$$

We also define $D_\omega(z)$ by

$$D_\omega(z) = \bigcup_{(m,\xi,y)\in\omega} \Big( \bigcup_{\tau_-(m,\xi,y,z)\le\tau\le\tau_+(m,\xi,y)} \gamma_{\tau,y,z} \Big) = \tag{22}$$

$$= \{(m',\eta',y',z') \in D[\omega] \mid z' = z\}$$

thus $D_\omega(z_1)$ is the intersection of $\bigcup_{(m,\xi,y)\in\omega} D_{(m,\xi,y)}$ with the plan $z = z_1$. According to the definition of the set $D_{(m,\xi,y)}$ we remark that

$$(m',\eta',y',z') \in D_{(m,\xi,y)} \Rightarrow \gamma_{\tau(m',\eta',y',z'),y',z'} \subset D_{(m,\xi,y)},$$

$$\tau(m_1,\xi_1,y_1,0) = \tau(m_2,\xi_2,y_2,0) \Rightarrow D_{(m_1,\xi_1,y_1)} = D_{(m_2,\xi_2,y_2)};$$

consequently, if $(m_1,\xi_1)$ and $(m_2,\xi_2)$ are on a curve $\gamma_{\tau,y,0}$, then they define the same set.

The "cone of dependence" property is given by the following lemma.

**Lemma 5** *Let $\overline{\sigma}^{[1]}$ and $\overline{\sigma}^{[2]}$ be two functions defined on $\mathbb{R}_+ \times \mathbb{R}^2$ satisfying the conditions of proposition 1. Let $\sigma^{[1]}$ (resp. $\sigma^{[2]}$) be the solution of* (12), (8) *with $\overline{\sigma} = \overline{\sigma}^{[1]}$ (resp. $\overline{\sigma} = \overline{\sigma}^{[2]}$). If we have*

$$\overline{\sigma}^{[1]} = \overline{\sigma}^{[2]} \qquad on \ \ D_\omega(1), \tag{23}$$

*then*

$$\sigma^{[1]} = \sigma^{[2]} \qquad a.e. \ \ in \ \ D[\omega]\cdot$$

**Proof.** Writing the equation (12) into an integral form, we have

$$\sigma^{[i]}(m,\xi,y,z) = \overline{\sigma}^{[i]}(m,\xi,y) +$$

$$+\frac{m\alpha(m)}{2g}\int_z^1 \int_{\gamma^{[0,m]}_{\tau(m,\xi,y,z'),y,z'}} \beta(m-m',m')\sigma^{[i]}(m',\eta',y,z')\sigma^{[i]}(m-m',\eta'',y,z')$$

$$\mu_\gamma(dm')dz' - \frac{m\alpha(m)}{g}\int_z^1 \int_{\gamma_{\tau(m,\xi,y,z'),y,z'}} \beta(m,m')\sigma^{[i]}(m',\eta',y,z')\times$$

$$\times\sigma^{[i]}(m,\xi,y,z')\mu_\gamma(dm')dz', \quad i = 1,2.$$

Making the difference for $i = 1$ and $i = 2$, we have

$$\left|\sigma^{[1]}(m,\xi,y,z) - \sigma^{[2]}(m,\xi,y,z)\right| \le \left|\overline{\sigma}^{[1]}(m,\xi,y) - \overline{\sigma}^{[2]}(m,\xi,y)\right| +$$

$$+C_\beta\Big[\int_z^1\int_{\gamma^{[0,m]}_{\tau(m,\xi,y,z'),y,z'}}\Big(\big|\sigma^{[1]}(m-m',\eta'',y,z')-\sigma^{[2]}(m-m',\eta'',y,z')\big|$$

$$\sigma^{[2]}(m',\eta',y,z')+\big|\sigma^{[1]}(m',\eta',y,z')-\sigma^{[2]}(m',\eta',y,z')\big|\sigma^{[1]}(m-m',\eta'',y,z')\Big)$$

$$\mu_\gamma(dm')dz'+\int_z^1\int_{\gamma_{\tau(m,\xi,y,z'),y,z'}}\Big(\big|\sigma^{[1]}(m,\xi,y,z')-\sigma^{[2]}(m,\xi,y,z')\big|\sigma^{[2]}(m',\eta',y,z')+$$

$$+\big|\sigma^{[1]}(m',\eta',y,z')-\sigma^{[2]}(m',\eta',y,z')\big|\sigma^{[1]}(m,\xi,y,z')\Big)\mu_\gamma(dm')dz'\Big],$$

where

$$C_\beta=\max[\sup_{0<m'<m<\infty}\frac{m\alpha(m)}{2g}\beta(m-m',m'),\sup_{m,m'\in\mathbb{R}_+}\frac{m\alpha(m)}{g}\beta(m,m')].$$

We deduce from it that

$$\big|\sigma^{[1]}(m,\xi,y,z)-\sigma^{[2]}(m,\xi,y,z)\big|\le\big|\overline{\sigma}^{[1]}((m,\xi,y)-\overline{\sigma}^{[2]}(m,\xi,y)\big|+ \tag{24}$$

$$+C_\beta\Big[\int_z^1\Big(\|\sigma^{[1]}(\cdot,\cdot,y,z')-\sigma^{[2]}(\cdot,\cdot,y,z')\|_{L^\infty(\gamma_{\tau(m,\xi,y,z'),y,z'})}$$

$$\|\sigma^{[2]}(\cdot,\cdot,y,z')\|_{L^1(\gamma_{\tau(m,\xi,y,z'),y,z'})}+\|\sigma^{[1]}(\cdot,\cdot,y,z')\|_{L^1(\gamma_{\tau(m,\xi,y,z'),y,z'})}$$

$$\|\sigma^{[1]}(\cdot,\cdot,y,z')-\sigma^{[2]}(\cdot,\cdot,y,z')\|_{L^\infty(\gamma_{\tau(m,\xi,y,z'),y,z'})}\Big)dz'+$$

$$+\int_z^1\Big(\|\sigma^{[1]}(\cdot,\cdot,y,z')-\sigma^{[2]}(\cdot,\cdot,y,z')\|_{L^\infty(\gamma_{\tau(m,\xi,y,z'),y,z'})}$$

$$\|\sigma^{[2]}(\cdot,\cdot,y,z')\|_{L^1(\gamma_{\tau(m,\xi,y,z'),y,z'})}+(\overline{m}_A-\overline{m}_a)\times$$

$$\|\sigma^{[1]}(\cdot,\cdot,y,z')-\sigma^{[2]}(\cdot,\cdot,y,z')\|_{L^\infty(\gamma_{\tau(m,\xi,y,z'),y,z'})}$$

$$\|\sigma^{[1]}(\cdot,\cdot,y,z')-\sigma^{[2]}(\cdot,\cdot,y,z')\|_{L^\infty(\gamma_{\tau(m,\xi,y,z'),y,z'})}\Big)dz'\Big].$$

Now let's consider a generic point $(m,\xi,y,z)$ of $D[\omega]$, by virtue of (20)–(21) there exists $(m_0,\xi_0,y_0)\in\omega\subset\mathbb{R}_+\times\mathbb{R}^2$ such that

$$\xi_0+\overline{v}(y_0)\frac{\alpha(m_0)}{g}-\overline{v}(y_0)\frac{\overline{\alpha}_0}{g}z=\tau_-(m_0,\xi_0,y_0,z)\le$$

$$\le\xi+\overline{v}(y)\frac{\alpha(m)}{g}(1-z)\le\tau_+(m_0,\xi_0,y_0)=\xi_0+\overline{v}(y_0)\frac{\alpha(m_0)}{g},$$

$$y=y_0.$$

From this inequalities, the inequality $\overline{v}(y)\frac{\alpha(m)}{g} \leq \overline{v}(y)\frac{\overline{\alpha}_0}{g} < 0$, implies that for $0 \leq z \leq z' \leq 1$, we have

$$\begin{cases} \xi_0 + \overline{v}(y_0)\frac{\alpha(m_0)}{g} - \overline{v}(y_0)\frac{\overline{\alpha}_0}{g} z' \leq \xi + \overline{v}(y)\frac{\alpha(m)}{g}(1-z') \leq \xi_0 + \overline{v}(y_0)\frac{\alpha(m_0)}{g}, \\ y = y_0, \end{cases}$$

by virtue of (11) and (21), we have

$$\begin{cases} \tau_-(m_0,\xi_0,y_0,z') \leq \tau(m,\xi,y,z') \leq \tau_+(m_0,\xi_0,y_0), \\ y = y_0 \end{cases}$$

and, according to the definition (22) of the set $D_\omega(z)$, we prove that

$$\gamma_{\tau(m,\xi,y,z'),y,z'} \subset D_\omega(z') \qquad \text{for} \quad 0 \leq z \leq z' \leq 1.$$

We recall that we have moreover, for $i = 1, 2$

$$\|\sigma^{[i]}(\cdot,\cdot,y,z)\|_{L^1(\gamma_{\tau(m,\xi,y,z),y,z},\mu_\gamma)} \leq (\overline{m}_A - \overline{m}_a)\|\sigma^{[i]}(\cdot,\cdot,y,z)\|_{L^\infty(D_\omega(z))},$$

for almost any $(m,\xi,y) \in \mathbb{R}_+ \times \mathbb{R}^2$.

From (24) we deduce that

$$\|\sigma^{[1]}(\cdot,\cdot,\cdot,z) - \sigma^{[2]}(\cdot,\cdot,\cdot,z)\|_{L^\infty(D_\omega(z))} \leq \|\overline{\sigma}^{[1]} - \overline{\sigma}^{[2]}\|_{L^\infty(D_\omega(1))}$$

$$+C\int_z^1 \Big(\|\sigma^{[1]}(\cdot,\cdot,\cdot,z')\|_{L^\infty(D_\omega(z'))} + \|\sigma^{[2]}(\cdot,\cdot,\cdot,z')\|_{L^\infty(D_\omega(z'))}\Big)\times$$

$$\times\|\sigma^{[1]}(\cdot,\cdot,\cdot,z') - \sigma^{[2]}(\cdot,\cdot,\cdot,z')\|_{L^\infty(D_\omega(z'))}dz',$$

where $C$ is a constant independent of $z$, using the Gronwall's lemma, we obtain

$$\|\sigma^{[1]}(\cdot,\cdot,\cdot,z) - \sigma^{[2]}(\cdot,\cdot,\cdot,z)\|_{L^\infty(D_\omega(z))} \leq \|\overline{\sigma}^{[1]} - \overline{\sigma}^{[2]}\|_{L^\infty(D_\omega(1))}\times \tag{25}$$

$$\times \exp\Big(C\int_z^1 (\|\sigma^{[1]}(\cdot,\cdot,\cdot,z')\|_{L^\infty(D_\omega(z'))} + \|\sigma^{[2]}(\cdot,\cdot,\cdot,z')\|_{L^\infty(D_\omega(z'))})dz'\Big).$$

However, under the assumption (23) we have

$$\|\overline{\sigma}^{[1]} - \overline{\sigma}^{[2]}\|_{L^\infty(D_\omega(1))} = 0,$$

that enables us to deduce from (25) that

$$\|\sigma^{[1]}(\cdot,\cdot,\cdot,z) - \sigma^{[2]}(\cdot,\cdot,\cdot,z)\|_{L^\infty(D_\omega(z))} \leq 0$$

and, taking into account the relation $D[\omega] = \bigcup_{0\le z\le 1} D_\omega(z)$, we have

$$\sigma^{[1]}(m,\xi,y,z) = \sigma^{[2]}(m,\xi,y,z) \qquad \text{a.e. in } \ D[\omega].$$

The lemma is proved. □

Now we can prove the principal theorem.

**Theorem 1** *If $\overline{\sigma}_1 \in L^\infty(\mathbb{R}_+ \times \mathbb{R}^2)$ satisfies the conditions*

$$\overline{\sigma}_1(m,\xi,y) \ge 0 \qquad \text{a.e. on} \quad \mathbb{R}_+ \times \mathbb{R}^2, \tag{26}$$

$$\overline{\sigma}_1(m,\xi,y) = 0 \qquad \text{for} \ \ m \in [0,\overline{m}_a] \cup [\overline{m}_A, \infty[, \tag{27}$$

$$\|\overline{\sigma}_1\|_{L^\infty(\mathbb{R}_+\times\mathbb{R}^2)} < \frac{1}{M_1(\overline{m}_A - \overline{m}_a)}, \tag{28}$$

*then the equation* (12) *with the condition* (8) *admits one and only one solution verifying*

$$\sigma \in L^\infty(\mathbb{R}_+ \times \mathbb{R}^2 \times [0,1])$$

*with*

$$\sigma(m,\xi,y,z) \ge 0 \qquad \text{a.e. in} \ \ \mathbb{R}_+ \times \mathbb{R}^2 \times [0,1],$$

$$\sigma(m,\xi,y,z) = 0 \qquad \text{for} \ \ m \in [0,\overline{m}_a] \cup [\overline{m}_A, \infty[\cdot$$

**Proof.** We consider a measurable and bounded sets family $\omega_i$, $i \in \mathbb{N}^*$, defined by

$$\omega_i = \left\{(m,\xi,y) \in \mathbb{R}_+ \times \mathbb{R}^2 \ / \ \overline{m}_a \le m \le \overline{m}_A, \ -i \le \xi \le i, \ -i \le y \le i\right\}. \tag{29}$$

The definition of $D[\omega]$ permits us to define a number $N$ such that

$$D_{\omega_i}(1) \subset \Big\{(m,\xi,y) \in \mathbb{R}_+ \times \mathbb{R}^2 \ / \ \overline{m}_a \le m \le \overline{m}_A, \ -i-N \le \xi \le i+N,$$

$$-i-1 \le y \le i+1\Big\}.$$

We consider the function $\psi_i \in C^\infty(\mathbb{R}^2)$; $\psi_i \ge 0$ such that

$$\psi_i(\xi,y) = \begin{cases} 1 & \text{if} \quad |\xi| \le i+N \quad \text{and} \quad |y| \le i+1, \\ 0 & \text{if} \quad |\xi| \ge i+N+1 \quad \text{and} \quad |y| \ge i+2, \end{cases} \tag{30}$$

then we have

$$D_{\omega_i}(1) \subset \{(m,\xi,y) \in \mathbb{R}_+ \times \mathbb{R}^2 \ / \ \psi_i(\xi,y) = 1\} \quad \text{for} \quad i \in \mathbb{N}^*. \tag{31}$$

Let the equations family

$$\partial_z \sigma^{[i]}(m,\xi,y,z) = F(\sigma^{[i]}(z))(m,\xi,y), \qquad i \in \mathbb{N}^* \tag{32}$$

(with $F(\cdot)$ defined in (12)), completed by the condition

$$\overline{\sigma}^{[i]} = \psi_i \, \overline{\sigma} \qquad \text{on} \quad \mathbb{R}_+ \times \mathbb{R}^2. \tag{33}$$

According to the proposition 1, the problem (32)–(33) admits one solution

$$\sigma = \sigma^{[i]} \in C([0,1]; L^1(\mathbb{R}_+ \times \mathbb{R}^2) \cap L^\infty(\mathbb{R}_+ \times \mathbb{R}^2 \times [0,1])),$$

such that

$$\sigma^{[i]} \geq 0 \qquad \text{a.e. in} \;\; \mathbb{R}_+ \times \mathbb{R}^2 \times [0,1],$$

$$\sigma^{[i]}(m,\xi,y,z) = 0 \qquad \text{for} \;\; m \in [0, \overline{m}_a] \cup [\overline{m}_A, \infty[.$$

In addition, according to the definition of the sets $\omega_i$, we have

$$D[\omega_i] \subset D[\omega_{i'}] \qquad \text{for} \;\; i \leq i',$$

therefore, by virtue of lemma 5 and of (33), we have

$$\sigma^{[i]} = \sigma^{[i']} \qquad \text{a.e. in} \;\; D[\omega_i] \qquad \text{for} \;\; i \leq i'.$$

Defining $\sigma$ by

$$\sigma = \begin{cases} \sigma^{[1]} & \text{in} \;\; D[\omega_1], \\ \sigma^{[i]} & \text{in} \;\; D[\omega_i] \backslash D[\omega_{i-1}], \quad i = 2, \cdot, \cdot, \cdot, \end{cases}$$

we have

$$\sigma = \sigma^{[i]} \qquad \text{a.e. in} \;\; D_{\omega_i}(1) \quad \forall i \in \mathbb{N}^*$$

and from (32), (61) we obtain

$$\partial_z \sigma(m,\xi,y,z) = F(\sigma(z))(m,\xi,y) \qquad \text{in} \;\; D[\omega_i] \quad \forall i \in \mathbb{N}^*,$$

$$\sigma = \sigma^{[i]} = \overline{\sigma} \qquad \text{on} \;\; D_{\omega_i}(1).$$

Remembering the relations $\mathbb{R}_+ \times \mathbb{R}^2 \times [0,1] \subset \bigcup_{i \in \mathbb{N}^*} D[\omega_i]$ and $\mathbb{R}_+ \times \mathbb{R}^2 \subset \bigcup_{i \in \mathbb{N}^*} D_{\omega_i}(1)$ which result from the definition of $\omega_i$, $D[\omega_i]$, $D_{\omega_i}(1)$, we can conclude that there exists a solution of (12), (8). To prove the uniqueness, let's consider two possible solutions $\sigma_1$ and $\sigma_2$ with $\sigma_1 \neq \sigma_2$ on a set of strictly positive measure, then we can choose a measurable set $\omega$ such that

$0 < mes(\omega) < \infty$ and that $mes(\{(m, \xi, y, z) \in D[\omega] \;/\; \sigma_1 \neq \sigma_2\}) > 0$. However as $\sigma_1$ and $\sigma_2$ are solutions of (12), (8), $\sigma_1 = \sigma_2$ on $\mathbb{R}_+ \times \mathbb{R}^2 \times \{1\}$ and in particular $\sigma_1 = \sigma_2$ on $\mathbb{R}_+ \times \mathbb{R}^2 \times \{1\} \bigcap D[\omega]$; consequently, according to lemma 5, we have $\sigma_1 = \sigma_2$ in $D[\omega]$, this proves that it is not possible to have two solutions $\sigma_1$ and $\sigma_2$ which are different on a set from strictly positive measure. The uniqueness of the solution is proved. □

For the existence and the uniqueness of the solution in the $(m, x, y, z)$ co-ordinates, we have the following theorem.

**Theorem 2** *If $\overline{\sigma} \in L^\infty(\mathbb{R}_+ \times \mathbb{R}^2)$ satisfies the conditions*

$$\overline{\sigma}(m, x, y) \geq 0 \qquad a.e \text{ on } \; \mathbb{R}_+ \times \mathbb{R}^2,$$

$$\overline{\sigma}(m, x, y) = 0 \qquad for \;\; m \in [0, \overline{m}_a] \cup [\overline{m}_A, \infty[,$$

$$\|\overline{\sigma}\|_{L^\infty(\mathbb{R}_+ \times \mathbb{R}^2)} < \frac{1}{M_1(\overline{m}_A - \overline{m}_a)},$$

*then the equation* (4) *with the condition* (5) *admits one solution $\sigma$ and only one verifying*

$$\sigma \in L^\infty(\mathbb{R}_+ \times \mathbb{R}^2 \times [0, 1]),$$

*such that*

$$\sigma(m, x, y, z) \geq 0 \qquad a.e \text{ on } \;\; \mathbb{R}_+ \times \mathbb{R}^2 \times [0, 1],$$

$$\overline{\sigma}(m, x, y, z) = 0 \qquad for \;\; m \in [0, \overline{m}_a] \cup [\overline{m}_A, \infty[.$$

**Proof.** We associate to the problem (4)-(5), where the unknown function to find is $\sigma$, the problem (12), (8) by a bijective mapping defined by the change of variables $(m, x, y, z) \mapsto (\tilde{m}, \xi, \tilde{y}, \tilde{z})$ introduced in (6) with

$$\sigma(m, x, y, z) = \tilde{\sigma}(m, \xi + \overline{v}(y)\frac{\alpha(m)}{g}(1 - z), y, z).$$

If $\tilde{\sigma}(m, \xi, y, z)$ is the solution of the problem (12), (8) in which the existence and the uniqueness have been proved in theorem 1, then, we obtain the existence and the uniqueness of the solution $\sigma$ for (4)-(5) verifying the same conditions. □

## 4 Global solution for the coagulation equation of the drops in fall with a horizontal wind

We will consider the problem to find a function $\sigma(m,t,x,y,z)$, that verifies the equation (1) for

$$(m,t,x,y,z)\in\mathbb{R}_+\times\mathbb{R}_+\times\mathbb{R}^2\times[0,1]$$

and the following boundary condition (entry condition) and initial condition

$$\sigma(m,t,x,y,1)=\overline{\sigma}_1(m,t,x,y), \tag{34}$$

$$\sigma(m,0,x,y,z)=\overline{\sigma}_0(m,x,y,z). \tag{35}$$

In the same way to the stationary case, to solve the equation (1) with the conditions (34)-(35), we will transform it into an ordinary differential equation, by introducing the following variables $(m,t,x,y,z)\mapsto(\tilde{m},\tilde{t},\xi,\tilde{y},\tilde{z})$

$$\begin{cases} \tilde{m}=m, \\ \xi=x-\overline{v}(y)\frac{\alpha(m)}{g}(1-z), \\ \tilde{y}=y, \\ \tilde{z}=z, \\ \tilde{t}=t-\frac{\alpha(m)}{g}(1-z) \end{cases} \tag{36}$$

and the unknown function to find would be

$$\tilde{\sigma}(\tilde{m},\tilde{t},\xi,\tilde{y},\tilde{z})=\sigma(m,t,x,y,z)=$$

$$=\sigma\Big(m,\tilde{t}+\frac{\alpha(m)}{g}(1-z),\xi+\overline{v}(y)\frac{\alpha(m)}{g}(1-z),y,z\Big),$$

we will note by $m$, $y$, $z$ and $\sigma(m,\tilde{t},\xi,y,z)$ instead of $\tilde{m}$, $\tilde{y}$, $\tilde{z}$ and $\tilde{\sigma}(\tilde{m},\tilde{t},\xi,\tilde{y},\tilde{z})$, the equation (1) is changed into

$$\frac{\partial}{\partial z}\sigma\left(m,\widetilde{t},\xi,y,z\right)= \tag{37}$$

$$=-\frac{m\ \alpha(m)}{2g}\int_0^m\beta(m-m',m')\sigma(m',\tilde{t}^*(m,m',\tilde{t},z),\eta(m,m',\xi,y,z),y,z)\times$$

$$\times\sigma(m-m',\tilde{t}^*(m,m-m',\tilde{t},z),\eta(m,m-m',\xi,y,z),y,z)dm'+$$

$$+\frac{m\ \alpha(m)}{g}\int_0^\infty\beta(m,m')\sigma(m,\tilde{t},\xi,y,z)\times$$

$$\times\sigma(m',\tilde{t}^*(m,m',\tilde{t},z),\eta(m,m',\xi,y,z),y,z)dm',$$

where

$$\begin{cases} \tilde{t}^*(m,m',\tilde{t},z)=\tilde{t}+\frac{\alpha(m)-\alpha(m')}{g}(1-z), \\ \eta(m,m',\xi,y,z)=\xi+\overline{v}(y)\frac{\alpha(m)-\alpha(m')}{g}(1-z). \end{cases}$$

We introduce for each fixed $y\in\mathbb{R}$, $z\in[0,1]$, the curves family

$$\gamma_{\tau,\zeta}=\gamma_{\tau,\zeta,y,z}=\Big\{(m,\tilde{t},\xi)\in\mathbb{R}_+\times\mathbb{R}^2\ /\ \tilde{t}=\tau-\frac{\alpha(m)}{g}(1-z), \qquad \xi=\zeta-\overline{v}(y)\frac{\alpha(m)}{g}(1-z)\Big\} \tag{38}$$

with $\tau,\zeta\in\mathbb{R}$.

Let $\tau$, $\zeta$, $\gamma_{\tau,\zeta}^{[0,m]}$ such that

$$\tau(m,\tilde{t},z)=\tilde{t}+\frac{\alpha(m)}{g}(1-z),\quad \zeta(m,\xi,y,z)=\xi+\overline{v}(y)\frac{\alpha(m)}{g}(1-z),$$

$$\gamma_{\tau,\zeta}^{[0,m]}=\gamma_{\tau,\zeta}\cap[0,m]\times\mathbb{R}^2.$$

We note by

$$\kappa=(\tau,\zeta),\quad \vartheta=(\tilde{t},\xi),\quad q=q(y)=(1,\overline{v}(y))^T,$$

then the curves defined in (38) can be written in the following form

$$\gamma_\kappa=\gamma_{\kappa,y,z}=\left\{(m,\vartheta)\in\mathbb{R}_+\times\mathbb{R}^2\ /\ \vartheta=\kappa-q(y)\frac{\alpha(m)}{g}(1-z)\right\} \tag{39}$$

with

$$\kappa(m,\vartheta,y,z)=\vartheta+q(y)\ \frac{\alpha(m)}{g}(1-z),\qquad \gamma_\kappa^{[0,m]}=\gamma_\kappa\cap[0,m]\times\mathbb{R}^2.$$

The curves family $\gamma_\kappa$ is similar to that defined in (9) in the stationary case, so in the same way, we define a measure $\mu_\gamma$ on the curves $\gamma_\kappa$ and the equation (37) will be

$$\frac{\partial}{\partial z}\sigma(z)=F_z(\sigma(z)),\ \sigma(z)=\sigma(\cdot,\cdot,\cdot,\cdot,z), \tag{40}$$

where

$$F_z(\sigma(z))=F_z(\sigma(z))(m,\vartheta,y)= \tag{41}$$

$$= -\frac{m\alpha(m)}{2g}\int_{\gamma^{[0,m]}_{\kappa(m,\vartheta,y,z)}} \beta(m-m',m')\sigma(m',\vartheta',y,z)\sigma(m-m',\vartheta'',y,z)\mu_\gamma(dm')+$$

$$+\frac{m\alpha(m)}{g}\int_{\gamma_{\kappa(m,\vartheta,y,z)}} \beta(m,m')\sigma(m',\vartheta',y,z)\sigma(m,\vartheta,y,z)\mu_\gamma(dm')$$

with $\vartheta'$ and $\vartheta''$ are defined by the relation

$$(m',\vartheta') \in \gamma_{\kappa(m,\vartheta,y,z)}, \qquad (m-m',\vartheta'') \in \gamma^{[0,m]}_{\kappa(m,\vartheta,y,z)}.$$

We remark that this equation is in the same type of the equation (12) in the stationary case and that the integral operator appearing in (41) verifies the same properties in lemmas 1, 2, 3 and 4.

In the same way, the boundary and the initial conditions will be changed into

$$\sigma\left(m,\tilde{t},\xi,y,1\right) = \overline{\sigma}_1^*(m,\tilde{t},\xi,y) = \overline{\sigma}_1^*(m,\vartheta,y) \tag{42}$$

and

$$\sigma(m,-\frac{\alpha(m)}{g}(1-z),\xi,y,z) = \overline{\sigma}_0^*(m,\xi,y,z), \tag{43}$$

where $\overline{\sigma}_0^*$ and $\overline{\sigma}_1^*$ are the functions obtained of $\overline{\sigma}_0$ and $\overline{\sigma}_1$ by the change of variables introduced in (36).

## 4.1 Solution with an entry condition in class $L^1$

We define the domain in which we will consider the equation (40) by

$$\Omega = \bigcup_{\kappa\in\mathbb{R}_+^*\times\mathbb{R},\ y\in\mathbb{R},\ 0<z<1} \gamma_{\kappa,y,z} = \tag{44}$$

$$= \left\{(m,\vartheta,y,z) = (m,\tilde{t},\xi,y,z) \in \mathbb{R}_+\times\mathbb{R}^3\times]0,1[\ /\ \tilde{t} > \frac{\alpha(m)}{g}(z-1)\right\}$$

and we pose

$$\Gamma_a = \left\{(m,\vartheta,y,z) = (m,\tilde{t},\xi,y,z) \in \mathbb{R}_+\times\mathbb{R}^3\times[0,1]\ /\ \tilde{t} = \frac{\alpha(m)}{g}(z-1)\right\},$$

$$\Gamma_b = \{z=1\}\cap\overline{\Omega}.$$

The conditions (42)–(43) can be written in the form

$$\sigma = \overline{\sigma}_1^* \quad \text{on} \quad \Gamma_b, \qquad \sigma = \overline{\sigma}_0^* \quad \text{on} \quad \Gamma_a. \tag{45}$$

**Proposition 2** *Let $\overline{\sigma}_{(a)} \in L^1(\Gamma_a) \cap L^\infty(\Gamma_a)$ and $\overline{\sigma}_{(b)} \in L^1(\Gamma_b) \cap L^\infty(\Gamma_b)$ such that*

$$\overline{\sigma}_{(a)}(m, \vartheta, y, z) \geq 0 \qquad a.e. \;\; on \quad \Gamma_a, \quad \overline{\sigma}_{(b)}(m, \vartheta, y) \geq 0 \qquad a.e. \;\; on \quad \Gamma_b,$$

$$\overline{\sigma}_{(a)}(m, \vartheta, y, z) = \overline{\sigma}_{(b)}(m, \vartheta, y) = 0 \qquad for \quad m \in [0, \overline{m}_a] \cup [\overline{m}_A, \infty[.$$

*If*

$$\max(\|\overline{\sigma}_{(a)}\|_{L^\infty(\Gamma_a)}, \|\overline{\sigma}_{(b)}\|_{L^\infty(\Gamma_b)}) < \frac{1}{M_1(\overline{m}_A - \overline{m}_a)},$$

*then there exists unique solution $\sigma$ of the equation (40) satisfying to the conditions*

$$\sigma = \overline{\sigma}_{(b)} \quad on \; \Gamma_b, \qquad \sigma = \overline{\sigma}_{(a)} \quad on \; \Gamma_a, \tag{46}$$

*with*

$$\sigma \in C([0,1]; L^1(\Omega_z) \cap L^\infty(\Omega)), \tag{47}$$

*where*

$$\Omega_z = \{(m, \vartheta, y) = (m, \tilde{t}, \xi, y) \in \mathbb{R}_+ \times \mathbb{R}^2 \times \mathbb{R} \; / \; \tilde{t} > \frac{\alpha(m)}{g}(z-1)\}. \tag{48}$$

**Proof.** In (40) and (41), the absence of derivative and integral compared to $y$ is remarked, as in (12), this implies that the equation (40) can be solved separately for each $y \in \mathbb{R}$.

We define for each point $(m, \vartheta) \in \mathbb{R}_+ \times \mathbb{R}^2$ the number $\zeta_1(m, \vartheta) \in [0,1]$ such that

$$\zeta_1(m, \vartheta) = \zeta_1(m, \tilde{t}, \xi) = \zeta_1(m, \tilde{t}) = \begin{cases} \max(0, 1 + \frac{\tilde{t}}{\alpha(m)} g) & \text{if} \;\; \tilde{t} \leq 0, \\ 1 & \text{if} \;\; \tilde{t} > 0 \end{cases} \tag{49}$$

and we have

$$(m, \vartheta, y, \zeta_1(m, \vartheta)) \in \Gamma_b \cup \Gamma_a \quad \forall (m, \vartheta, y) \in \mathbb{R}_+ \times \mathbb{R}^2 \times \mathbb{R}, \quad \tilde{t} \geq -\frac{\alpha(m)}{g},$$

these permit us to replace $(t, x) \in \mathbb{R}^2$ by the time axis, then we find the conditions for the proof of proposition 4.1 in [2], consequently renewing the stages of the proof of this one, we prove the proposition. $\square$

### 4.2 Existence and uniqueness of the global solution in time with a horizontal wind

In the same way to the stationary case, to obtain the existence and the uniqueness of the global solution with an horizontal wind in a general case, we use the "cone of dependence" property and the proposition 2.

We consider a set $\omega \in \mathbb{R}_+ \times \mathbb{R}^3$ such that $0 < mes(\omega) < \infty$ and we define

$$D[\omega] = \bigcup_{(m,\vartheta,y)\in\omega} D_{(m,\vartheta,y)}, \tag{50}$$

where

$$D_{(m,\vartheta,y)} = \Big( \bigcup_{0\le z\le 1} \Big( \bigcup_{\kappa_-(m,\vartheta,y,z)\le\kappa\le\kappa_+(m,\vartheta,y)} \gamma_{\kappa,y,z} \Big)\Big) = \tag{51}$$

$$= \{(m',\vartheta',y',z') \in \mathbb{R}_+ \times \mathbb{R}^3 \times [0,1] / \, \vartheta' = \kappa - q(y')\frac{\alpha(m')}{g}(1-z'),$$
$$y' = y, \ \kappa_-(m,\vartheta,y,z') \le \kappa \le \kappa_+(m,\vartheta,y)\}$$

with

$$\begin{cases} \kappa_+(m,\vartheta,y) = \kappa(m,\vartheta,y,0) = \vartheta + q(y)\frac{\alpha(m)}{g}, \\ \kappa_-(m,\vartheta,y,z) = \kappa_+(m,\vartheta,y) - q(y)\frac{\overline{\alpha}_0}{g}z = \vartheta + q(y)\frac{\alpha(m)}{g} - q(y)\frac{\overline{\alpha}_0}{g}z. \end{cases} \tag{52}$$

We define $D_\omega(z)$ by

$$D_\omega(z) = \bigcup_{(m,\vartheta,y)\in\omega} \Big( \bigcup_{\kappa_-(m,\vartheta,y,z)\le\kappa\le\kappa_+(m,\vartheta,y)} \gamma_{\kappa,y,z} \Big) =$$
$$= \{(m',\vartheta',y',z') \in D[\omega] \mid z' = z\}.$$

We remark that $D[\omega]$ in the evolution case is defined in a similar way to the stationary case (see (19), (20), (21)) then we have the following lemma.

**Lemma 6** *Let $\overline{\sigma}^{[1]}_{(a)}$ and $\overline{\sigma}^{[2]}_{(a)}$ two functions defined on $\Gamma_a$, $\overline{\sigma}^{[1]}_{(b)}$ and $\overline{\sigma}^{[2]}_{(b)}$ two functions defined on $\Gamma_b$. We suppose that $\overline{\sigma}^{[1]}_{(a)}$, $\overline{\sigma}^{[2]}_{(a)}$, $\overline{\sigma}^{[1]}_{(b)}$, $\overline{\sigma}^{[2]}_{(b)}$ satisfy the conditions of the proposition 2. Let $\sigma^{[1]}$ (resp. $\sigma^{[2]}$) the solution of the equation* (40) *with the condition* (46) *and $\overline{\sigma}_{(a)} = \overline{\sigma}^{[1]}_{(a)}$, $\overline{\sigma}_{(b)} = \overline{\sigma}^{[1]}_{(b)}$ (resp. $\overline{\sigma}_{(a)} = \overline{\sigma}^{[2]}_{(a)}$, $\overline{\sigma}_{(b)} = \overline{\sigma}^{[2]}_{(b)}$). If we have*

$$\overline{\sigma}^{[1]}_{(b)} = \overline{\sigma}^{[2]}_{(b)} \quad \text{on } \Gamma_b \cap D[\omega], \qquad \overline{\sigma}^{[1]}_{(a)} = \overline{\sigma}^{[2]}_{(a)} \quad \text{on } \Gamma_a \cap D[\omega], \tag{53}$$

*then*

$$\sigma^{[1]} = \sigma^{[2]} \quad \text{a.e. in } D[\omega].$$

**Proof.** Writing the equation (40) into an integral form, we have

$$\sigma^{[i]}(m,\vartheta,y,z)=\sigma^{[i]}(m,\vartheta,y,\zeta_1(m,\vartheta))+$$

$$+\frac{m\alpha(m)}{2g}\int_z^{\zeta_1}\int_{\gamma^{[0,m]}_{\kappa(m,\vartheta,y,z'),y,z'}}\beta(m-m',m')\sigma^{[i]}(m',\vartheta',y,z')\sigma^{[i]}(m-m',\vartheta'',z')$$

$$\mu_\gamma(dm')dz'-\frac{m\alpha(m)}{g}\int_z^{\zeta_1}\int_{\gamma_{\kappa(m,\vartheta,y,z'),y,z'}}\beta(m,m')\sigma^{[i]}(m',\vartheta',y,z')\sigma^{[i]}(m,\vartheta,y,z')$$

$$\mu_\gamma(dm')dz',\qquad i=1,2.$$

From (45) it results that

$$\sigma^{[i]}(m,\vartheta,y,\zeta_1(m,\vartheta))=\begin{cases}\overline{\sigma}^{[i]}_{(a)} & \text{on } \Gamma_a,\\ \\ \overline{\sigma}^{[i]}_{(b)} & \text{on } \Gamma_b,\end{cases}$$

$\zeta_1(m,\vartheta)$ is the number defined in (49).

Making the difference for $i=1$ and $i=2$, we have

$$|\sigma^{[1]}(m,\vartheta,y,z)-\sigma^{[2]}(m,\vartheta,y,z)|\le|\sigma^{[1]}(m,\vartheta,y,\zeta_1)-\sigma^{[2]}(m,\vartheta,y,\zeta_1)|+$$

$$+C_\beta\Big[\int_z^{\zeta_1}\int_{\gamma^{[0,m]}_{\kappa(m,\vartheta,y,z'),y,z'}}\Big(\big|\sigma^{[1]}(m-m',\vartheta'',y,z')-\sigma^{[2]}(m-m',\vartheta'',y,z')\big|\sigma^{[2]}(m',\vartheta',y,z')+$$

$$+\big|\sigma^{[1]}(m',\vartheta',y,z')-\sigma^{[2]}(m',\vartheta',y,z')\big|\sigma^{[1]}(m-m',\vartheta'',y,z')\Big)\mu_\gamma(dm')dz'+$$

$$+\int_z^{\zeta_1}\int_{\gamma_{\kappa(m,\vartheta,y,z'),y,z'}}\Big(\big|\sigma^{[1]}(m,\vartheta,y,z')-\sigma^{[2]}(m,\vartheta,y,z')\big|\sigma^{[2]}(m',\vartheta',y,z')+$$

$$+\big|\sigma^{[1]}(m',\vartheta',y,z')-\sigma^{[2]}(m',\vartheta',y,z)\big|\sigma^{[1]}(m,\vartheta,y,z')\Big)\mu_\gamma(dm')dz'\Big],$$

we deduce from it that

$$|\sigma^{[1]}(m,\vartheta,y,z)-\sigma^{[2]}(m,\vartheta,y,z)|\le|\sigma^{[1]}(m,\vartheta,y,\zeta_1)-\sigma^{[2]}(m,\vartheta,y,\zeta_1)|+ \tag{54}$$

$$+C_\beta\Big[\int_z^1\Big(\|\sigma^{[1]}(\cdot,\cdot,\cdot,z')-\sigma^{[2]}(\cdot,\cdot,\cdot,z')\|_{L^\infty(\gamma_{\kappa(m,\vartheta,y,z'),y,z'})}$$

$$\|\sigma^{[2]}(\cdot,\cdot,\cdot,z')\|_{L^1(\gamma_{\kappa(m,\vartheta,y,z'),y,z'})}+\|\sigma^{[1]}(\cdot,\cdot,\cdot,z')\|_{L^1(\gamma_{\kappa(m,\vartheta,y,z'),z'})}$$

$$\|\sigma^{[1]}(\cdot,\cdot,\cdot,z') - \sigma^{[2]}(\cdot,\cdot,\cdot,z')\|_{L^\infty(\gamma_{\kappa(m,\vartheta,y,z'),y,z'})}\Big) dz' +$$

$$+ \int_z^1 \Big( \|\sigma^{[1]}(\cdot,\cdot,\cdot,z') - \sigma^{[2]}(\cdot,\cdot,\cdot,z')\|_{L^\infty(\gamma_{\kappa(m,\vartheta,y,z'),y,z'})} \times$$

$$\times \|\sigma^{[2]}(\cdot,\cdot,\cdot,z')\|_{L^1(\gamma_{\kappa(m,\vartheta,y,z'),y,z'})} +$$

$$+(\overline{m}_A - \overline{m}_a)\|\sigma^{[1]}(\cdot,\cdot,z')\|_{L^\infty(\gamma_{\kappa(m,\vartheta,y,z'),y,z'})} \times$$

$$\times \|\sigma^{[1]}(\cdot,\cdot,\cdot,z') - \sigma^{[2]}(\cdot,\cdot,\cdot,z')\|_{L^\infty(\gamma_{\kappa(m,\vartheta,y,z'),y,z'})}\Big) dz' \Big].$$

We remark that this inequality is similar to the inequality (24) in the proof of lemma 5 and by the same way we obtain the result. □

Now we can prove the principal theorem.

**Theorem 3** *If $\overline{\sigma}_0^* \in L^\infty(\Gamma_a)$ and $\overline{\sigma}_1^* \in L^\infty(\Gamma_b)$ satisfy to the conditions*

$$\overline{\sigma}_0^*(m,\xi,y,z) \geq 0 \quad \textit{a.e. on } \Gamma_a, \qquad \overline{\sigma}_1^*(m,\vartheta,y) \geq 0 \quad \textit{a.e. on } \Gamma_b, \tag{55}$$

$$\overline{\sigma}_0^*(m,\xi,y,z) = 0, \quad \overline{\sigma}_1^*(m,\vartheta,y) = 0 \qquad \textit{for } m \in [0,\overline{m}_a] \cup [\overline{m}_A, \infty[, \tag{56}$$

$$\max\left(\|\overline{\sigma}_0^*\|_{L^\infty(\Gamma_a)}; \|\overline{\sigma}_1^*\|_{L^\infty(\Gamma_b)}\right) < \frac{1}{M_1(\overline{m}_A - \overline{m}_a)}, \tag{57}$$

*then the equation* (40) *with the condition* (45) *admits one solution $\sigma$ and only one verifying*

$$\sigma \in L^\infty(\Omega)$$

*with*

$$\sigma(m,\vartheta,y,z) \geq 0 \qquad \textit{a.e. in } \Omega,$$

$$\sigma(m,\vartheta,y,z) = 0, \quad \textit{for } m \in [0,\overline{m}_a] \cup [\overline{m}_A, \infty[.$$

**Proof.** We consider a measurable and bounded sets family $\omega_i$, $i \in \mathbb{N}^*$, defined by

$$\omega_i = \Big\{(m,\vartheta,y) = (m,\tilde{t},\xi,y) \in \mathbb{R}_+ \times \mathbb{R}^3 \ / \ \overline{m}_a \leq m \leq \overline{m}_A, \tag{58}$$

$$-\frac{\alpha(m)}{g} \leq \tilde{t} \leq i,\ -i \leq \xi \leq i,\ -i \leq y \leq i\Big\}$$

$$= \Omega_0 \cap \{(m,\vartheta,y) = (m,\tilde{t},\xi,y) \in [\overline{m}_a,\overline{m}_A] \times \mathbb{R}^3 \ / \ \tilde{t} \leq i\},$$

where $\Omega_0$ is the set defined in (48) with $z = 0$. The definition of $D[\omega]$ (see (50)) permits us to define a number $N$ such that

$$D_{\omega_i}(1) \subset \Big\{(m,\vartheta,y) = (m,\tilde{t},\xi,y) \in \mathbb{R}_+ \times \mathbb{R}^3 \ / \ \overline{m}_a \le m \le \overline{m}_A, \\ \tilde{t} \le i+N, \ -i-N \le \xi \le i+N, \ -i-1 \le y \le i+1\Big\} \tag{59}$$

and we consider a function $\psi_i \in C^\infty(\mathbb{R}^3)$; $\psi_i \ge 0$ such that

$$\psi_i(\vartheta,y) = \psi_i(\tilde{t},\xi,y) = \begin{cases} 1 & \text{if } \ \tilde{t} \le i+N, \ |\xi| \le i+N, \ |y| \le i+1 \\ 0 & \text{if } \ \tilde{t} \ge i+N+1, \ |\xi| \ge i+N+1, \ \ |y| \ge i+2, \end{cases} \tag{60}$$

then we have

$$D_{\omega_i}(1) \subset \{(m,\vartheta,y) \in \mathbb{R}_+ \times \mathbb{R}^3 \ / \ \psi_i(\vartheta,y) = 1\} \quad i \in \mathbb{N}^*. \tag{61}$$

The theorem will be proved in the same way to theorem 5.1 of [2] (see also theorem 1 of the stationary case) by renewing the same stages. □

The existence and the uniqueness of the solution in the $(m,t,x,y,z)$ co-ordinates is given in the following theorem.

**Theorem 4** *If* $\overline{\sigma}_0 \in L^\infty(\mathbb{R}_+ \times \mathbb{R}^2 \times [0,1])$ *and* $\overline{\sigma}_1 \in L^\infty(\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^2)$ *satisfy the conditions*

$$\overline{\sigma}_0(m,x,y,z) \ge 0 \quad a.e. \ on \quad \mathbb{R}_+ \times \mathbb{R}^2 \times [0,1],$$

$$\overline{\sigma}_1(m,t,x,y) \ge 0 \quad a.e. \ on \quad \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^2,$$

$$\overline{\sigma}_0(m,x,y,z) = \overline{\sigma}_1(m,t,x,y) = 0 \qquad for \quad m \in [0,\overline{m}_a] \cup [\overline{m}_A,\infty[,$$

$$\max\left(\|\overline{\sigma}_0\|_{L^\infty(\mathbb{R}_+\times\mathbb{R}^2\times[0,1])}; \|\overline{\sigma}_1\|_{L^\infty(\mathbb{R}_+\times\mathbb{R}_+\times\mathbb{R}^2)}\right) < \frac{1}{M_1(\overline{m}_A - \overline{m}_a)},$$

*then the equation* (1) *with the conditions* (34) *and* (35) *admits one solution* $\sigma$ *and only one verifying*

$$\sigma \in L^\infty(\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^2 \times ]0,1[),$$

*where*

$$\sigma(m,t,x,y,z) \ge 0 \quad a.e. \ in \quad \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^2 \times ]0,1[,$$

$$\sigma(m,t,x,y,z) = 0 \quad for \quad m \in [0,\overline{m}_a] \cup [\overline{m}_A,\infty[.$$

**Proof.** We associate to the problem (1), (34),(35), where the unknown function to find is $\sigma$, the problem (40), (45) by a bijective mapping defined by the change of variables $(m,t,x,y,z) \mapsto (\tilde{m},\tilde{t},\xi,\tilde{y},\tilde{z})$ introduced in (36) with

$$\sigma(m,t,x,y,z) = \tilde{\sigma}\big(m, t - \frac{\alpha(m)}{g}(1-z), x + \overline{v}(y)\frac{\alpha(m)}{g}(1-z), y, z\big).$$

If $\tilde{\sigma}(m,\tilde{t},\xi,y,z)$ is the solution of the problem (40), (45) in which the existence and the uniqueness have been proved in theorem 3, then, we obtain the existence and the uniqueness of the solution $\sigma$ of (1), (34),(35) verifying the same conditions. □

## References


[1] H. Belhireche, M. Z.Aissaoui, H. Fujita Yashima. Equations monodimensionnelles du mouvement de l'air avec la transition de phase de l'eau. *Sci. Techn. Univ. Constantine- A.* 31: 9-17, 2011.

[2] H. Belhireche, M. Z. Aissaoui, H. Fujita Yashima. Solution globale de l'équation de coagulation des gouttelettes en chute. *Ren. Sem. Mat. Univ. Polit. Torino.* 70, 3: 261–278, 2012.

[3] H. Brezis. *Analyse fonctionelle (Théorie et applications).* Masson, 1987.

[4] M. Escobedo, S. Mischler, B. Perthame. Gelation in coagulation and fragmentation models. *Comm. Math. Phys.* 231: 157-188, 2002.

[5] M. Escobedo, J.J.L. Velazquez. On the fundamental solution of a linearized homogeneous coagulation equation. *Comm. Math. Phys.* 297: 759-816, 2010.

[6] M. Escobedo, S. Mischler, M. Rodriguez Ricard. On self-similarity and stationary problem for fragmentation and coagulation models. *Ann. Inst. H. Poincaré, Anal. Non Linéaire.* 22: 99-125, 2005.

[7] H. Fujita Yashima, V. Campana, M. Z. Aissaoui. Système d'équations d'un modèle du mouvement de l'air impliquant la transition de phase de l'eau dans l'atmosphère. *Ann. Math. Afr.* 2: 66-92, 2011.

[8] A. K. Kikoïne, I. K. Kikoïne. *Physique moléculaire (russian traduction).* Mir, Moscou, 1979.

[9] A. N. Kolmogorov, S. V. Fomine. *Eléments de la théorie des fonctions et de l'analyse fonctionelle (Russian traduction).* Mir, Moscou, 1974.

[10] M. Merad, H. Belhireche, H. Fujita Yashima. Solution stationnaire de l'équation de coagulation de gouttelettes en chute avec le vent horizontal. *Rend. Sem. Mat. univ. Padova.* 129: 225-244, 2013.

[11] S. Mischler. Contributions à l'étude mathématique de quelques modèles issus de la physique hors équilibre. *Thesis of enabling.* Univ. Versailles Saint-Quentin, 2001.

[12] S. Mischler, M. Rodriguez Ricard. Existence globale pour l'équation de Smoluchowski continue non homogène et comportement asymptotique des solutions. *C. R. Acad. Sci. Paris, Sér. I, Math.* 336: 407-412, 2003.

[13] H. Müller. Zur allgemeinen Theorie der raschen Koagulation. *Kolloidchem. Beib.* 27: 223-250, 1928.

[14] B. Niethammer, J.J.L. Velazquez. Optimal bounds for self-similar solutions to coagulation equations with multiplicative kernel. To appear on *Commun. PDE.*

[15] F. Prodi, A. Battaglia. *Meteorologia - Parte II, Microfisica.* Grafica Pucci. Roma, 2004. (See also site: http://www.meteo.uni-bonn.de/mitarbeiter/battaglia/teaching.html ).

[16] S. Selvaduray, H. Fujita Yashima. Equazioni del moto dell'aria con la transizione di fase dell'acqua nei tre stati: gassoso, liquido e solido. *Accad. Sci. Torino, Memorie Cl. Sci. Fis., Serie V.* 35: 37-69, 2011.

[17] P.-X. Sheng, J.-T. Mao, J.-G. Li, A.-C. Zhang, J.-G. Sang, N.-X. Pan. *Atmosphere Physique (in Chinese).* Publ. Univ. Pékin, Pékin, 2003.

[18] M. Smoluchowski. Drei Vorträge über Diffusion, Brownische Bewegung und Koagulation von Kolloidteilchen. *Phys. Zeits.* 17: 557-585, 1916.

[19] V. M Voloshtchuk. *Théorie cinétique de coagulation (in Russian).* Gidrometeoizdat, Leningrad, 1984.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 6, No. 2/2014

# Methods and Algorithms for Approximating the Gamma Function and Related Functions. A survey. Part I: Asymptotic Series*

Cristinel Mortici†

**Abstract**

In this survey we present our recent results on analysis of gamma function and related functions. The results obtained are in the theory of asymptotic analysis, approximation of gamma and polygamma functions, or in the theory of completely monotonic functions. The motivation of this first part is the work of C. Mortici [Product Approximations via Asymptotic Integration Amer. Math. Monthly 117 (2010) 434-441] where a simple strategy for constructing asymptotic series is presented. The classical asymptotic series associated to Stirling, Wallis, Glaisher-Kinkelin are rediscovered. In the second section we discuss some new inequalities related to Landau constants and we establish some asymptotic formulas.

**MSC**: 26D15; 11Y60; 41A60; 41A25; 34E05

**Keywords**: gamma function; digamma function; Bernoulli numbers; approximations; asymptotic series; monotonicity; Glaisher-Kinkelin constant; Landau constants; Euler-Mascheroni constant; convergence; speed of convergence

*Accepted for publication in revised form on March 24-th, 2014

†cristinel.mortici@hotmail.com, Valahia University of Târgovişte, Bd. Unirii 18, 130082 Târgovişte, Romania and Academy of Romanian Scientists, Splaiul Independenţei 54, 050094 Bucharest, Romania

# 1 A method for constructing asymptotic series and applications

The problem of approximating the gamma function goes back to Laplace formula which is the continuous version of the Stirling formula. In 1916 Srinivasa Ramanujan (see [4]) proposed a formula which was later studied by E. A. Karatsuba in [13] and Alzer [2].

A method for improving some approximation formulas for large factorials is to consider the corresponding asymptotic series. It is presented in [16] an original approach to the asymptotic evaluation of sums and products. As for usual, to an approximation formula $f(n) \sim g(n)$, it is associated the series

$$f(n) \sim g(n) \exp\left(\sum_{k=1}^{\infty} \frac{a_k}{n^k}\right), \tag{1}$$

also called an asymptotic series. Such series have the advantage that in a truncated form, provides approximations to any accuracy $n^{-k}$.

The strategy in [16] is based on the idea that when series (1) is truncated at the $m$th term, the approximation obtained should be the most precise possible among all approximations

$$f(n) \sim g(n) \exp\left(\sum_{k=1}^{m} \frac{a'_k}{n^k}\right), \tag{2}$$

where $a'_1, a'_2, \ldots, a'_m$ are any real numbers.

The first task is to compare the accuracy of two approximation formulas. We do this by associate to an approximation formula $f(n) \sim g(n)$ the relative error sequence $r_n$ by the relations

$$f(n) = g(n) \exp r_n \ , \qquad n \geq 1.$$

We consider $f(n) \sim g(n)$ as better as $r_n$ converges to zero faster.

Now a new task appears, that is to measure the speed of convergence of the sequence $r_n$. The tool used is the following

**Lemma 1 (Speed of Convergence Lemma).** *If $(r_n)_{n\geq 1}$ is convergent to zero and*

$$\lim_{n\to\infty} n^k(r_n - r_{n+1}) = l \ , \qquad \textit{then} \qquad \lim_{n\to\infty} n^{k-1} r_n = \frac{l}{k-1}, \quad (k \geq 2)\,.$$

*In other words, $r_n$ is of $n^{-(k-1)}$ speed of convergence, in case $r_n - r_{n+1}$ is of order $n^{-k}$.*

We cite from Batir [3]: "This lemma, despite of its simple appearance, is a strong tool to accelerate and measure the speed of convergence of some sequences having limit zero, and has proved by C. Mortici in [16]". As the reviewer of [16] asked, a detailed proof of Lemma 1 was presented for sake of completness.

We introduce the relative error sequence $(\lambda_n)_{n\geq 1}$ by

$$f(n) = g(n) \exp\left(\sum_{k=1}^{m} \frac{a_k}{n^k}\right) \exp \lambda_n \ , \quad n \geq 1.$$

In order to use Lemma 1, we write

$$\lambda_n - \lambda_{n+1} = \sum_{k=2}^{m+1} \frac{x_k - y_k}{n^k} + O\left(\frac{1}{n^{m+2}}\right),$$

where

$$\sum_{k=1}^{m} \frac{a_k}{n^k} - \sum_{k=1}^{m} \frac{a_k}{(n+1)^k} = \sum_{k=2}^{m+1} \frac{y_k}{n^k} + O\left(\frac{1}{n^{m+2}}\right)$$

with

$$a_1 - \binom{k-1}{1} a_2 + \cdots + (-1)^k \binom{k-1}{k-2} a_{k-1} = (-1)^k y_k, \quad 2 \leq k \leq m+1$$

and assuming

$$\ln \frac{f(n)g(n+1)}{g(n)f(n+1)} = \sum_{k=2}^{\infty} \frac{x_k}{n^k}. \tag{3}$$

The following main result is stated in [16].

**Theorem 1.** *Suppose there is some $k$ such that $2 \leq k \leq m+1$ and $x_k \neq y_k$, and let $s = \min\{k \mid 2 \leq k \leq m+1,\ x_k \neq y_k\}$. Then*

$$\lim_{n\to\infty} n^{s-1}\lambda_n = \frac{x_s - y_s}{s-1} \in \mathbb{R} \setminus \{0\},$$

*and therefore the speed of convergence of $(\lambda_n)_{n\geq 1}$ is $n^{-(s-1)}$.*

If $s \geq 3$, conditions $x_k = y_k$, for $2 \leq k \leq s-1$, are equivalent to the triangular system

$$x_k = (-1)^k \left(a_1 - \binom{k-1}{1} a_2 + \cdots + (-1)^k \binom{k-1}{k-2} a_{k-1}\right), \tag{4}$$

which defines uniquely the best coefficients $a_k$, $1 \leq k \leq s-2$.

These theoretical results were applied in [16] to deduce the series associated with some approximation formulas: Stirling, Burnside, Glaisher-Kinkelin, Wallis. Standard construction of these series makes appeal to Bernoulli numbers and Euler-Maclaurin summation formula.

To the Glaisher-Kinkelin constant defined by

$$A = \lim_{n \to \infty} \frac{1^1 2^2 3^3 \cdots n^n}{n^{n^2/2+n/2+1/12} e^{-n^2/4}},$$

the following asymptotic series is considered

$$1^1 2^2 3^3 \cdots n^n \sim A \cdot n^{\frac{n^2+n}{2}+\frac{1}{12}} e^{-n^2/4} \exp\left(\sum_{k=1}^{\infty} \frac{a_k}{n^k}\right).$$

Here we have $f(n) = 1^1 2^2 3^3 \cdots n^n$ and $g(n) = A \cdot n^{\frac{n^2+n}{2}+\frac{1}{12}} e^{-n^2/4}$. The values $x_k$ in (3) are

$$x_k = (-1)^k \left(\frac{1}{2k+2} - \frac{1}{2k+4} - \frac{1}{12k}\right),$$

and the solution of the triangular system (4) is $a_1 = 0$, $a_2 = 1/720$, $a_3 = 0, a_4 = -1/5040$, $a_5 = 0$, $a_6 = 1/10080$, $\ldots$ . Hence

$$1^1 2^2 3^3 \cdots n^n \sim A \cdot n^{\frac{n^2+n}{2}+\frac{1}{12}} e^{-n^2/4} \exp\left(\frac{1}{720n^2} - \frac{1}{5040n^4} + \frac{1}{10\,080n^6} - \cdots\right).$$

The asymptotic series associated to Wallis formula is

$$\frac{\pi}{2} \sim \left(\prod_{j=1}^{n} \frac{4j^2}{4j^2-1}\right) \exp\left(\sum_{k=1}^{\infty} \frac{a_k}{n^k}\right).$$

With $f(n) = \frac{\pi}{2}$ and $g(n) = \prod_{j=1}^{n} \frac{4j^2}{4j^2-1}$ in (3), we get

$$x_k = \frac{(-1)^k}{k} \left(\frac{3^k+1}{2^k} - 2\right).$$

The solution of the triangular system (4) is $a_1 = 1/4$, $a_2 = -1/8$, $a_3 = 5/96$, $a_4 = -1/64$, $\ldots$ . Hence

$$\frac{\pi}{2} \sim \left(\prod_{j=1}^{n} \frac{4j^2}{4j^2-1}\right) \exp\left(\frac{1}{4n} - \frac{1}{8n^2} + \frac{5}{96n^3} - \frac{1}{64n^4} + \cdots\right).$$

By using standard transforms on asymptotic series, it is obtained in [16] the following formula

$$\prod_{j=1}^{n} \frac{4j^2}{4j^2-1} \sim \frac{\pi}{2}\left(1-\frac{1}{4n}+\frac{5}{32n^2}-\frac{11}{128n^3}+\frac{31}{768n^4}-\cdots\right),$$

which is an extension of the following formula presented by Hirschhorn in [12]:

$$\prod_{j=1}^{n} \frac{4j^2}{4j^2-1} \sim \frac{\pi}{2}-\frac{\pi}{8n}+O\left(\frac{1}{n^2}\right) \quad \text{as} \quad n\to\infty.$$

Undoubtedly the most used formula for estimating big factorials is the following

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

now known as Stirling's formula. Classical methods for constructing the corresponding asymptotic series use some equations involving numeric series and improper integrals, Euler-Maclaurin summation formula, Legendre duplication formula, or the analytic definition of Bernoulli numbers. The method proposed in [16] is quite elementary. For the asymptotic series

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left(\sum_{k=1}^{\infty}\frac{a_k}{n^k}\right), \tag{5}$$

with $f(n)=n!$ and $g(n)=\sqrt{2\pi n}\left(\frac{n}{e}\right)^n$, we have in (3)

$$x_k = (-1)^k\,\frac{k-1}{2k(k+1)}.$$

The solution of the triangular system (4) is $a_1 = 1/12$, $a_2 = 0$, $a_3 = -1/360$, $a_4 = 0$, $a_5 = 1/1260$, $a_6 = 0$, $a_7 = -1/1680$, which are coefficients in (5).

It is presented in [17] the following asymptotic expansion in terms of Bernoulli numbers for every $p \in [0,1]$:

$$\Gamma(x+1) \sim \sqrt{2\pi e}\cdot e^{-p}\left(\frac{x+p}{e}\right)^{x+\frac{1}{2}}\cdot \exp\left\{\sum_{k=1}^{\infty}\frac{a_p(x)}{x^k}\right\}, \quad n\to\infty, \tag{6}$$

where

$$a_p(x) = \frac{1}{k(k+1)}\left[B_{k+1}-(-1)^k p^k\left(\left(p-\frac{1}{2}\right)k-\frac{1}{2}\right)\right].$$

The class of approximations (6) was also recently studied by Nemes [22]. Particular case $p = 1/2$ is Burnside series [6]:

$$\Gamma(x+1) \sim \sqrt{2\pi}\left(\frac{x+\frac{1}{2}}{e}\right)^{x+\frac{1}{2}} \exp\left(\sum_{k=1}^{\infty}\left(B_{k+1}+\frac{(-1)^k}{2^{k+1}}\right)\frac{1}{k(k+1)x^k}\right),$$

while $p = 1$ case provides the following formula:

$$\Gamma(x+1) \sim \sqrt{\frac{2\pi}{e}}\left(\frac{x+1}{e}\right)^{x+\frac{1}{2}} \exp\left(\sum_{k=1}^{\infty}\left(B_{k+1}-(-1)^k\frac{k-1}{2}\right)\frac{1}{k(k+1)x^k}\right).$$

As usually truncations of these series provide upper- and lower- estimates. The following double inequalities were presented in [17]:

**Theorem 2.** F*or every $x \geq 1$, we have*

$$\sqrt{2\pi}\left(\frac{x+\frac{1}{2}}{e}\right)^{x+\frac{1}{2}} \exp a(x) < \Gamma(x+1) < \sqrt{2\pi}\left(\frac{x+\frac{1}{2}}{e}\right)^{x+\frac{1}{2}} \exp b(x),$$

*where*

$$a(x) = -\frac{1}{24x}+\frac{1}{48x^2}-\frac{23}{2880x^3}+\frac{1}{640x^4}+\frac{11}{40\,320x^5}+\frac{1}{5376x^6}-\frac{143}{215\,040x^7}$$

*and*

$$b(x) = a(x)+\frac{143}{215\,040x^7}.$$

**Theorem 3.** *For every $x \geq 1$, we have*

$$\sqrt{\frac{2\pi}{e}}\left(\frac{x+1}{e}\right)^{x+\frac{1}{2}} \exp c(x) < \Gamma(x+1) < \sqrt{\frac{2\pi}{e}}\left(\frac{x+1}{e}\right)^{x+\frac{1}{2}} \exp d(x),$$

*where*

$$c(x) = \frac{1}{12x}-\frac{1}{12x^2}+\frac{29}{360x^3}-\frac{3}{40x^4}+\frac{17}{252x^5}-\frac{5}{84x^6}$$

*and*

$$d(x) = c(x)+\frac{89}{1680x^7}.$$

Liu [15] established the following integral version of Stirling's formula

$$\Gamma(n+1) = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot \exp\left(\int_n^{\infty}\frac{\frac{1}{2}-\{t\}}{t}dt\right).$$

An extension to Nemes' family was presented in [18]. The following formula is valid for every $p \in [0, 1]$ :

$$\Gamma(n+1) = \sqrt{2\pi e} \cdot e^{-p} \left(\frac{n+p}{e}\right)^{n+\frac{1}{2}}$$

$$\cdot \exp\left(\int_n^\infty \left(\frac{\frac{3}{2} - p - \{t\}}{t+p} + \frac{p}{p\{t\} + [t]} - \frac{1}{t}\right) dt\right).$$

According to our discussion in general case, the Stirling series in terms of Bernoulli numbers

$$\Gamma(n+1) \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \exp\left\{\sum_{k=1}^\infty \frac{B_{2k}}{2k(2k-1)n^{2k-1}}\right\} \tag{7}$$

is of best performance from the approximation point of view when it is truncated at every term. However better results can be obtained if we consider the truncations in (7) as rational functions of the form

$$\sum_{k=1}^m \frac{B_{2k}}{2k(2k-1)n^{2k-1}} = \frac{R_m(n^2)}{12nT_m(n^2)},$$

where $R_m$, $T_m$ are polynomials of $(m-1)\,th$ degree, with the leading coefficients equal to unity. It is indicated in [19] how can be constructed polynomials $P_m$, $Q_m$ of $(m-1)\,th$ degree such that the approximation

$$\Gamma(n+1) \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \exp \frac{P_m(n^2)}{12nQ_m(n^2)} \tag{8}$$

is the best possible among all approximations of the form

$$\Gamma(n+1) \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \exp \frac{P'_m(n^2)}{12nQ'_m(n^2)},$$

where $P'_m$ and $Q'_m$ are every polynomials of $(m-1)\,th$ degree with leading coefficient equal to unity. New obtained approximations (8) are more accurate than the $m$th approximation of the classical Stirling series (7). Initial approximations

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \exp \frac{n^2 + \frac{53}{210}}{12n\left(n^2 + \frac{2}{7}\right)} =: \rho_1 \tag{9}$$

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\frac{n^4+\frac{2559}{1430}n^2+\frac{22999}{90090}}{12n\left(n^4+\frac{782}{429}n^2+\frac{263}{858}\right)} =: \rho_2 \tag{10}$$

are more accurate than the classical approximations arising from Stirling series truncated at the second, respective at the third term, namely

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\frac{30n^2-1}{360n^3} =: \sigma_1, \tag{11}$$

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\frac{210n^4-7n^2+2}{2520n^5} =: \sigma_2. \tag{12}$$

In order to offer an initial image, we consider a comparison table to prove the superiority of (9)-(10) over (11)-(12).

| $n$ | $\ln(n!/\sigma_1)$ | $\ln(\rho_1/n!)$ |
|---|---|---|
| 10 | $7.8\times10^{-9}$ | $3.6\times10^{-11}$ |
| 100 | $7.9\times10^{-14}$ | $3.6\times10^{-18}$ |
| 250 | $8.1\times10^{-16}$ | $6.0\times10^{-21}$ |

| $n$ | $\ln(\sigma_2/n!)$ | $\ln(n!/\rho_2)$ |
|---|---|---|
| 10 | $5.8\times10^{-11}$ | $5.2\times10^{-15}$ |
| 100 | $5.9\times10^{-18}$ | $5.7\times10^{-26}$ |
| 250 | $9.7\times10^{-21}$ | $3.1\times10^{-27}$ |

Rigorous proofs of these facts are presented in [19]. Remark that the first approximations (8) are the approximations obtained by truncation the classical Stieltjes continued fraction to gamma function, but the proof of this result is left as an open problem in [19].

In order to show our method, let us search the best constants $a_1$, $a_2$ in $m=2$ case:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\frac{n^2+a_1}{12\left(n^3+a_2 n\right)}.$$

For the relative error sequence $z_n$ defined by

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\frac{n^2+a_1}{12\left(n^3+a_2 n\right)} \exp z_n\ , \quad (n\geq 1),$$

we used Maple software for symbolic computation to deduce

$$\begin{aligned} z_n - z_{n+1} &= \left(-\frac{1}{4}a_1+\frac{1}{4}a_2-\frac{1}{120}\right)\frac{1}{n^4} - 2\left(-\frac{1}{4}a_1+\frac{1}{4}a_2-\frac{1}{120}\right)\frac{1}{n^5} \\ &+ \left(\frac{5}{6}a_2-\frac{5}{6}a_1+\frac{5}{12}a_1a_2-\frac{5}{12}a_2^2-\frac{1}{42}\right)\frac{1}{n^6} \\ &+ \left(\frac{5}{4}a_1-\frac{5}{4}a_2-\frac{5}{4}a_1a_2+\frac{5}{4}a_2^2+\frac{5}{168}\right)\frac{1}{n^7} + O\left(\frac{1}{n^8}\right). \end{aligned}$$

Now the fastest sequence $z_n$ is obtained when the first two coefficients in this power series vanish, that is $a_1 = \frac{53}{210}$, $a_2 = \frac{2}{7}$.

In case $m = 3$, we define the sequence $t_n$ by

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\frac{n^4 + b_1 n^2 + b_2}{12\left(n^5 + b_3 n^3 + b_4 n\right)} \exp t_n \ , \quad (n \geq 1)\,.$$

As

$$t_n - t_{n+1} = \left(-\frac{1}{4}b_1 + \frac{1}{4}b_3 - \frac{1}{120}\right)\frac{1}{n^4} - 2\left(-\frac{1}{4}b_1 + \frac{1}{4}b_3 - \frac{1}{120}\right)\frac{1}{n^5}$$
$$+\left(\frac{5}{6}b_3 - \frac{5}{12}b_2 - \frac{5}{6}b_1 + \frac{5}{12}b_4 + \frac{5}{12}b_1 b_3 - \frac{5}{12}b_3^2 - \frac{1}{42}\right)\frac{1}{n^6} + O\left(\frac{1}{n^7}\right),$$

we get $b_1 = \frac{2559}{1430}$, $b_2 = \frac{22\,999}{90\,090}$, $b_3 = \frac{782}{429}$, $b_4 = \frac{263}{858}$. In this case,

$$t_n - t_{n+1} = -\frac{80\,713}{12\,972\,960 n^{12}} + O\left(\frac{1}{n^{13}}\right).$$

The following estimates were stated in [19]:

**Theorem 4.** *For every positive integer $n$, we have*

$$\exp\left(\frac{P_2\left(n^2\right)}{12nQ_2\left(n^2\right)} - \frac{13}{35280n^7}\right) < \frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} < \exp\frac{P_2\left(n^2\right)}{12nQ_2\left(n^2\right)}.$$

We illustrate our method by providing

*The proof of Theorem 4.* We have to prove that $a_n > 0$ and $b_n < 0$, where

$$a_n = \frac{P_2\left(n^2\right)}{12nQ_2\left(n^2\right)} - \ln\frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n},$$

$$b_n = \frac{P_2\left(n^2\right)}{12nQ_2\left(n^2\right)} - \frac{13}{35280n^7} - \ln\frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}.$$

As $a_n$, $b_n$ converge to zero, it suffices to show that $a_n$ is strictly decreasing, while $b_n$ is strictly increasing. In this sense, $a_{n+1} - a_n = f\left(n\right)$, $b_{n+1} - b_n = g\left(n\right)$, where

$$f\left(x\right) = \left(x + \frac{1}{2}\right)\ln\left(1 + \frac{1}{x}\right) - 1 + \frac{P_2\left(\left(x+1\right)^2\right)}{12\left(x+1\right)Q_2\left(\left(x+1\right)^2\right)} - \frac{P_2\left(x^2\right)}{12xQ_2\left(x^2\right)}$$

and

$$g\left(x\right) = f\left(x\right) - \left(\frac{13}{35280\left(x+1\right)^7} - \frac{13}{35280x^7}\right).$$

The function $f$ is strictly concave, while $g$ is strictly convex with $f(\infty) = g(\infty) = 0$, so $f(x) < 0$ and $g(x) > 0$, for every $x \in [1, \infty)$ and the theorem is proved.$\square$

In the same manner, the following result is stated in [19]:

**Theorem 5.** *For every positive integer $n$, we have*

$$\exp\left(\frac{P_3(n^2)}{12nQ_3(n^2)} - \frac{80713}{142702560n^{11}}\right) < \frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} < \exp\frac{P_3(n^2)}{12nQ_3(n^2)}.$$

## 2 Landau constants

E. Landau studied the asymptotic behaviour of the constants

$$G_n = 1 + \left(\frac{1}{2}\right)^2 + \left(\frac{1\cdot 3}{2\cdot 4}\right)^2 + ... + \left(\frac{1\cdot 3\cdot ...\cdot(2n-1)}{2\cdot 4\cdot ...\cdot(2n)}\right)^2,$$

(now known as Landau constants) proving the asymptotic formula $G_n \sim (1/\pi)\ln n$, see e.g. [14]. Then Watson [24] proposed

$$G_n = c_0 + \frac{1}{\pi}\ln(n+1) - \frac{1}{4\pi(n+1)} + O\left(\frac{1}{n^2}\right),$$

where $c_0 = \frac{1}{\pi}(\gamma + 4\ln 2) = 1.06627...$ and $\gamma = 0.577...$ is Euler-Mascheroni constant. Further improvements were presented by Brutman [5]

$$1 + \frac{1}{\pi}\ln(n+1) < G_n < 1.0663 + \frac{1}{\pi}\ln(n+1)$$

and Falaleev [9]

$$1.0662 + \frac{1}{\pi}\ln\left(n+\frac{3}{4}\right) < G_n < 1.0916 + \frac{1}{\pi}\ln\left(n+\frac{3}{4}\right). \tag{13}$$

It is showed in [20] that $3/4$ is the best possible constant that can be used in (13). The proofs are based on inequalities

$$s(x) < \ln\Gamma(x+1) < t(x) \tag{14}$$

where

$$s(x) = \ln\sqrt{2\pi} + \left(x+\frac{1}{2}\right)\ln x - x + \frac{1}{12x} - \frac{1}{360x^3} + \frac{1}{1260x^5} - \frac{1}{1680x^7}$$

and

$$t(x) = \ln\sqrt{2\pi} + \left(x+\frac{1}{2}\right)\ln x - x + \frac{1}{12x} - \frac{1}{360x^3} + \frac{1}{1260x^5} - \frac{1}{1680x^7} + \frac{1}{1188x^9}.$$

They are a consequence of a result of Alzer [2, Theorem 8].

By (14), we get

$$e^{u(x)} < \frac{1}{16^x}\left(\frac{\Gamma(2x+1)}{(\Gamma(x+1))^2}\right)^2 < e^{v(x)}, \tag{15}$$

where

$$u(x) = 2s(2x) - 4t(x) - x\ln 16\ , \quad v(x) = 2t(2x) - 4s(x) - x\ln 16.$$

Mortici [20] used (15) to establish the following

**Theorem 6.** *For every integer $n \geq 1$, we have*

$$c_0 + \frac{1}{\pi}\ln\left(n+\frac{3}{4}\right) < G_n < c_0 + \frac{1}{\pi}\ln\left(n+\frac{3}{4}+\frac{11}{192n}\right). \tag{16}$$

*Proof.* As $n = 1, 2$ cases can be easily proven, we assume $n \geq 3$. The sequence

$$a_n = G_n - c_0 - \frac{1}{\pi}\ln\left(n+\frac{3}{4}\right)$$

converges to zero and it suffices to show that $(a_n)_{n\geq 3}$ is strictly decreasing. As

$$a_n - a_{n-1} = \frac{1}{16^n}\left(\frac{\Gamma(2n+1)}{(\Gamma(n+1))^2}\right)^2 - \frac{1}{\pi}\ln\left(1+\frac{1}{n-\frac{1}{4}}\right) < e^{v(n)} - \frac{1}{\pi}\sum_{k=1}^{4}\frac{(-1)^{k-1}}{k\left(n-\frac{1}{4}\right)^k},$$

we have to prove that $f(x) < 0$, where

$$f(x) = v(x) - \ln\left(\frac{1}{\pi}\sum_{k=1}^{4}\frac{(-1)^{k-1}}{k\left(x-\frac{1}{4}\right)^k}\right).$$

This function has its derivative $f' > 0$ on $[3,\infty)$. Now $f$ is strictly increasing on $[3,\infty)$, with $f(\infty) = 0$, so $f(x) < 0$, for every $x \in [3,\infty)$.

For the right-hand side inequality (16), define the sequence

$$b_n = G_n - c_0 - \frac{1}{\pi}\ln\left(n+\frac{3}{4}+\frac{11}{192n}\right)$$

and proceed as above. We have

$$b_n - b_{n-1} = \frac{1}{16^n}\left(\frac{\Gamma(2n+1)}{(\Gamma(n+1))^2}\right)^2 - \frac{1}{\pi}\ln\left(1+\frac{1+\frac{11}{192n}-\frac{11}{192(n-1)}}{n-\frac{1}{4}+\frac{11}{192(n-1)}}\right)$$

$$> e^{u(n)} - \frac{1}{\pi}\sum_{k=1}^{5}\frac{(-1)^{k-1}}{k\left(\frac{n-\frac{1}{4}+\frac{11}{192(n-1)}}{1+\frac{11}{192n}-\frac{11}{192(n-1)}}\right)^k}.$$

The function

$$g(x) = u(x) - \ln\left(\frac{1}{\pi}\sum_{k=1}^{5}\frac{(-1)^{k-1}}{k\left(\frac{x-\frac{1}{4}+\frac{11}{192(x-1)}}{1+\frac{11}{192x}-\frac{11}{192(x-1)}}\right)^k}\right),$$

is strictly decreasing on $[3,\infty)$, with $g(\infty)=0$, so $g(x)>0$, for every $x\in[3,\infty)$.□

Zhao [25] extended the asymptotic expansion of $G_n$ to

$$G_n = c_0 + \frac{1}{\pi}\ln(n+1) - \frac{1}{4\pi(n+1)} + \frac{5}{192\pi(n+1)^2} + O\left(\frac{1}{(n+1)^3}\right),$$

then Mortici [20] proved the following improvement

**Theorem 7.** *For every integer $n\geq 1$, we have*

$$c_0 + \frac{1}{\pi}\ln(n+1) - \frac{1}{4\pi(n+1)} + \frac{5}{192\pi(n+1)^2} \tag{17}$$

$$+\frac{3}{128\pi(n+1)^3} - \frac{341}{122880\pi(n+1)^4} - \frac{75}{8192\pi(n+1)^5} < G_n$$
$$< c_0 + \frac{1}{\pi}\ln(n+1) - \frac{1}{4\pi(n+1)} + \frac{5}{192\pi(n+1)^2} + \frac{3}{128\pi(n+1)^3} -$$
$$-\frac{341}{122880\pi(n+1)^4}$$

and the following asymptotic formula holds as $n\to\infty$ :

$$G_n = c_0 + \frac{1}{\pi}\ln(n+1) - \frac{1}{4\pi(n+1)} + \frac{5}{192\pi(n+1)^2}$$

$$+\frac{3}{128\pi (n+1)^3} - \frac{341}{122880\pi (n+1)^4} + O\left(\frac{1}{(n+1)^5}\right).$$

Cvijović and Klinowski [8] presented some estimates in terms of the digamma function

$$c_0 + \frac{1}{\pi}\psi\left(n+\frac{5}{4}\right) < G_n < 1.0725 + \frac{1}{\pi}\psi\left(n+\frac{5}{4}\right)$$

and

$$0.9883 + \frac{1}{\pi}\psi\left(n+\frac{3}{2}\right) < G_n < c_0 + \frac{1}{\pi}\psi\left(n+\frac{3}{2}\right) \quad (n \geq 0),$$

as Alzer [1] proved the following double sharp inequality

$$c_0 + \frac{1}{\pi}\psi(n+\alpha) < G_n < c_0 + \frac{1}{\pi}\psi(n+\beta), \quad (n \geq 1),$$

where $\alpha = 5/4$ and $\beta = \psi^{-1}(\pi(1-c_0)) = 1.26621...$ .

Mortici [20] improved the above results of Cvijović, Klinowski and Alzer as follows:

**Theorem 8.** *For every positive integer $n$, we have*

$$c_0 + \frac{1}{\pi}\psi\left(n+\frac{5}{4}\right) + \frac{1}{64\pi n^2} - \frac{3}{128\pi n^3} < G_n$$

$$< c_0 + \frac{1}{\pi}\psi\left(n+\frac{5}{4}\right) + \frac{1}{64\pi n^2} - \frac{3}{128\pi n^3} + \frac{173}{8192\pi n^4}. \tag{18}$$

Cases $n = 1, 2$ are true, so we assume $n \geq 3$. The sequence

$$t_n = G_n - c_0 - \frac{1}{\pi}\psi\left(n+\frac{5}{4}\right) - \frac{1}{64\pi n^2} + \frac{3}{128\pi n^3}$$

is strictly decreasing. As

$$t_n - t_{n-1} = \frac{1}{16^n}\frac{(\Gamma(2n+1))^2}{(\Gamma(n+1))^4} - \frac{1}{\pi\left(n+\frac{1}{4}\right)}$$

$$-\frac{1}{64\pi n^2} + \frac{3}{128\pi n^3} + \frac{1}{64\pi (n-1)^2} - \frac{3}{128\pi (n-1)^3},$$

we have to prove that $m < 0$, where

$$m(x) = v(x) -$$

$$-\ln\left(\frac{1}{\pi\left(x+\frac{1}{4}\right)}+\frac{1}{64\pi x^2}-\frac{3}{128\pi x^3}-\frac{1}{64\pi\left(x-1\right)^2}+\frac{3}{128\pi\left(x-1\right)^3}\right).$$

But $m$ is strictly increasing with $m(\infty)=0$, so $m<0$ on $[3,\infty)$.

For the right-hand side inequality (18), the sequence

$$z_n=G_n-c_0-\frac{1}{\pi}\psi\left(n+\frac{5}{4}\right)-\frac{1}{64\pi n^2}+\frac{3}{128\pi n^3}-\frac{173}{8192\pi n^4}$$

is strictly increasing and the argument is similar.□

Recent studies on Landau and Lebesgue constants were performed by Chen and Choi [7], Granath [10], or Nemes [23].

**Acknowledgement 1** *This work was supported by a grant of the Romanian National Authority for Scientific Research, CNCS-UEFISCDI project number PN-II-ID-PCE-2011-3-0087. Some computations made in this paper were performed using Maple software. Part of this work was included in the habilitation thesis held by the author at the Politehnica University of Bucharest, Romania, September* 2012.

## References

[1] H. Alzer, Inequalities for the constants of Landau and Lebesgue, J. Comput. Appl. Math., 139 (2002) 215–230.

[2] H. Alzer, On some inequalities for the gamma and psi functions, Math. Comp., 66 (1997), no. 217, 373-389.

[3] N. Batir, Sharp bounds for the psi function and harmonic numbers, Math. Inequal. Appl., 14 (2011), 917-925.

[4] Bruce C. Berndt; Robert A. Rankin (2000). "The Books Studied by Ramanujan in India". Amer. Math. Monthly 107 (7): 595–601.

[5] L. Brutman, A sharp estimate of the Landau constants, J. Approx. Theory, 34, 217-220 (1982).

[6] W. Burnside, A rapidly convergent series for log $N!$, Messenger Math., 46(1917), 157-159.

[7] C.-P. Chen and J. Choi, Asymptotic expansions for the constants of Landau and Lebesgue, Adv. Math. 254 (2014), 622-641.

[8] D. Cvijović, J. Klinowski, Inequalities for the Landau constants, Math. Slovaca 50 (2000) 159–164.

[9] L.P. Falaleev, Inequalities for the Landau constants, Siberian Math. J. 32 (1991) 896–897.

[10] On inequalities and asymptotic expansions for the Landau constants Original Research Article

[11] H. Granath, J. Math. Anal. Appl., 386 (2012), 738-743.

[12] M. D. Hirschhorn, Comments on the paper "Wallis sequence estimated through the Euler-Maclaurin formula: even from the Wallis product $\pi$ could be computed fairly accurately" [Austral. Math. Soc. Gaz. 31 (2004) 328–339]" by Lampret, Austral. Math. Soc. Gaz. 32 (2005) 194.

[13] E. A. Karatsuba, On the asymptotic representation of the Euler gamma function by Ramanujan, J. Comput. Appl. Math. 135 (2001) no. 2, 225-240.

[14] E. Landau, Abschätzung der Koeffzientensumme einer Potenzreihe, Arch. Math. Phys. 21 (1913) 42–50, 250–255.

[15] Z. Liu, A new version of the Stirling formula, Tamsui Oxf. J. Math. Sci., 23 (2007), no. 4, 389-392.

[16] C. Mortici, Product Approximations via Asymptotic Integration, Amer. Math. Monthly, 117 (2010), 434-441.

[17] C. Mortici, Asymptotic expansions of the generalized Stirling approximation, Math. Comp. Modelling, 52 (2010), no. 9-10, 1867-1868.

[18] C. Mortici, The asymptotic series of the generalized Stirling formula, Comp. Math. Appl., 60 (2010), no. 3, 786-791.

[19] C. Mortici, The best rational remainders in the Stirling formula, Int. Transf. Spec. Funct., 23 (2012), no. 1, 13-19.

[20] C. Mortici, Sharp bounds of the Landau constants, Math. Comp., 80 (2011), 1011-1018.

[21] C. Mortici, New improvements of the Stirling formula, Appl. Math. Comp., 217 (2010), no. 2, 699-704.

[22] G. Nemes, More accurate approximations for the gamma function, Thai. J. Math., 9 (2011) no. 1, 21-28.

[23] G. Nemes, Proofs of two conjectures on the Landau constants, J. Math. Anal. Appl., 388 (2012), 838-844.

[24] G. N. Watson, The constants of Landau and Lebesgue, Quart. J. Math. Oxford Ser. 1 (2) (1930) 310–318.

[25] D. Zhao, Some sharp estimates of the constants of Landau and Lebesgue, J. Math. Anal. Appl. 349 (2009) 68–73.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 6, No. 2/2014

# Methods and Algorithms for Approximating the Gamma Function and Related Functions. A survey. Part II: Completely monotonic functions*

Cristinel Mortici†

**Abstract**

In this survey we present our recent results on analysis of gamma function and related functions. The results obtained are in the theory of asymptotic analysis, approximation of gamma and polygamma functions, or in the theory of completely monotonic functions. In the second part of this survey we show how the theory of completely monotonic functions can be used to establish sharp bounds for gamma and related functions.

**MSC**: 33B15; 26D15; 11Y60; 41A60; 41A25; 34E05

**Keywords**: gamma function; digamma function; polygamma functions; approximations; asymptotic series; inequalities; monotonicity; complete monotonicity; Stirling formula; Burnside formula; Schuster formula; Wallis' ratio; Kazarinoff's inequality; Minc-Sathre ratio.

## 1 Introduction and Motivation

By a completely monotonic function on an interval $I$ we mean a function $z : I \to \mathbb{R}$ which admits derivatives of any order and satisifies the following

*Accepted for publication in revised form on April 10-th, 2014

†cristinel.mortici@hotmail.com, Valahia University of Târgovişte, Bd. Unirii 18, 130082 Târgovişte, Romania and Academy of Romanian Scientists, Splaiul Independenţei 54, 050094 Bucharest, Romania, cristinel.mortici@hotmail.com

inequalities for every $x \in I$ and integer $n \geq 0$ :

$$(-1)^n z^{(n)}(x) \geq 0.$$

The definition and further properties of other classes of completely monotonic functions (including (almost) completely monotonic, (almost) logarithmically completely monotonic, strongly completely monotonic, completely monotonic of $n$th order) can be found for example in [6], [7], [15], [16], [29], or [34].

Completely monotonic functions are of great help in the problem of approximating the function $z$ itself as well the derivatives $z^{(n)}$. More precisely, if we take into account that the derivative of $z^{(n)}$ keep constant sign and consequently the function $z^{(n)}$ is monotone, $z^{(n)}(x)$ lies between $z^{(n)}(a)$ and $z^{(n)}(b)$, as $x$ runs between $a$ and $b$.

Moreover, completely monotone functions involving gamma function provide sharp bounds for gamma and polygamma functions.

A tool for proving the complete monotonicity of a function is Bernstein-Widder-Hausdorff theorem (see, *e.g.,* [35, p. 161]) which states that a function is completely monotonic on $(0, \infty)$ if and only if the following integral representation is valid for every $x > 0$ :

$$z(x) = \int_0^\infty e^{-xt} d\mu(t). \tag{1}$$

Here $\mu$ is a non-negative measure on $[0, \infty)$ such that the integral converges for all $x > 0$.

The Euler gamma function is defined by the following formula for every real $x > 0$ :

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt,$$

while the logarithmic derivative of $\Gamma$ is called digamma (or psi) function,

$$\psi(x) = \frac{d}{dx}(\ln \Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}.$$

Further derivatives $\psi'$, $\psi''$, $\psi'''$, ... are called tri-, tetra-, penta-gamma function, and in general, $\psi^{(n)}$ with $n = 1, 2, 3, ...$ are polygamma functions.

In order to prove the complete monotonicity of a function involving gamma and polygamma functions on $(0, \infty)$ using (1), the following integral representations are of main help:

$$\frac{1}{x^r} = \frac{1}{\Gamma(r)} \int_0^\infty t^{r-1} e^{-tx} dt \quad (x > 0,\ r > 1) \tag{2}$$

$$\psi(x) = \int_0^\infty \left( \frac{e^{-t}}{t} - \frac{e^{-xt}}{1-e^{-t}} \right) dt \quad (x > 0)$$

and

$$\psi^{(n)}(x) = (-1)^{n+1} \int_0^\infty \frac{t^n}{1-e^{-t}} e^{-tx} dt \quad (x > 0,\ n \geq 1). \tag{3}$$

For further details, please see [1].

Usually to an approximation formula of the gamma function of type

$$\Gamma(x+1) \sim \omega(x) \tag{4}$$

in the sense that the ratio $\Gamma(x+1)/\omega(x)$ tends to 1, as $x$ approaches infinity, the following function is attached:

$$F(x) = \ln \frac{\Gamma(x+1)}{\omega(x)}. \tag{5}$$

If $F$ (sometimes $-F$) is completely monotonic then important results related to approximation formula (4) can be established. Let us assume for example that $F$ is completely monotonic on $[1, \infty)$, possible on $(0, \infty)$. As $F' < 0$, the function $F$ is strictly decreasing on $[1, \infty)$. Thus $F(\infty) < F(x) \leq F(1)$, which can be rearranged in the form of the following double inequality valid for every $x \in [1, \infty)$ :

$$\alpha \cdot \omega(x) < \Gamma(x+1) < \beta \cdot \omega(x).$$

Here the constants $\alpha = \exp F(1)$ and $\beta = \exp F(\infty) = 1$ are the best possible.

Furthermore, we can exploit the monotonicity of $F'$ to obtain sharp bounds for the digamma function. Assuming that $\omega$ is derivable, we get

$$F'(x) = \psi(x+1) - \frac{\omega'(x)}{\omega(x)}.$$

But $F'' > 0$, so $F'$ is strictly increasing on $[1, \infty)$, which can be written as $F'(1) \leq F'(x) < F'(\infty)$. The following sharp inequalities hold true for every real $x \in [1, \infty)$ :

$$\alpha' + \frac{\omega'(x)}{\omega(x)} \leq \psi(x+1) < \beta' + \frac{\omega'(x)}{\omega(x)},$$

where $\alpha' = F'(1)$ and $\beta' = F'(\infty)$.

These are the first illustration of our method for establishing sharp bounds for gamma and digamma functions related to approximation formula (4). In a similar manner inequalities for polygamma functions can be stated using the $n$th derivative of $F$.

In conclusion the study of the monotonicity of the function $F$ associated to an approximation formula (4) is of great importance in the theory of approximation of gamma, polygamma and other related functions.

## 2 The Technique

In order to illustrate the technique, we present the results stated in [18]. Undoubtedly the most used formula for approximating large factorials is Stirling's formula

$$\Gamma(x+1) \sim \sqrt{2\pi x}\left(\frac{x}{e}\right)^{x}.$$

A slightly better result was proposed by Burnside (see, *e.g.* [5]):

$$\Gamma(x+1) \sim \sqrt{2\pi}\left(\frac{x+1/2}{e}\right)^{x+\frac{1}{2}}. \tag{6}$$

It has been proved in [24] that the function

$$F(x) = \ln \frac{\Gamma(x+1)}{\sqrt{2\pi}\left(\frac{x+1/2}{e}\right)^{x+\frac{1}{2}}}.$$

associated to the Burnside formula is completely monotonic.

For sake of completness, we reproduce here a sketch of proof of the above result stated in [24]. As

$$F(x) = \ln\Gamma(x+1) - \ln\sqrt{2\pi} - \left(x+\frac{1}{2}\right)\ln\left(x+\frac{1}{2}\right) + x + \frac{1}{2},$$

we obtain

$$F'(x) = \psi(x+1) - \ln\left(x+\frac{1}{2}\right).$$

Using the recurrence formula

$$\psi(x+1) = \psi(x) + \frac{1}{x}$$

(see, *e.g.,* [1, p. 258]), we obtain

$$F'(x) = \psi(x) + \frac{1}{x} - \ln\left(x + \frac{1}{2}\right),$$

then

$$F''(x) = \psi''(x) - \frac{1}{x^2} - \frac{1}{x + \frac{1}{2}}.$$

Using (2) and (3), we deduce that

$$F''(x) = \int_0^\infty \frac{te^{-xt}}{1 - e^{-t}} dt - \int_0^\infty te^{-xt} dt - \int_0^\infty e^{-\left(x+\frac{1}{2}\right)t} dt,$$

or

$$F''(x) = \int_0^\infty \frac{e^{-(x+1)t}}{1 - e^{-t}} \varphi(t)\, dt,$$

where

$$\varphi(t) = t - e^{-\frac{1}{2}t}\left(e^t - 1\right).$$

The function $\varphi$ is strictly decreasing, since $\varphi'(t) = -\frac{1}{2}e^{-\frac{1}{2}t}\left(e^{\frac{1}{2}t} - 1\right)^2 < 0$. For $t > 0$, we have $\varphi(t) < \varphi(0) = 0$. According to Bernstein- Widder-Hausdorff theorem, $-F''$ is strictly completely monotonic. using the definition, we obtain

$$(-1)^n \left(-F''\right)^{(n)} \geq 0,$$

for every integer $n \geq 0$. By replacing $(-F'')^{(n)}$ by $(-F)^{n+2}$, we deduce

$$(-1)^n F^{(n)} \geq 0, \tag{7}$$

for every integer $n \geq 2$. In order to finalize our proof, we have to show that (7) is valid also for $n = 1$ and $n = 0$.

In this sense, note that $F'$ is strictly decreasing, since $F'' < 0$. But $\lim_{x\to\infty} F'(x) = 0$, so $F'(x) > 0$ and consequently, $F$ is strictly increasing. Using the fact that $\lim_{x\to\infty} F(x) = 0$, we deduce that $F < 0$. This assures the veridicity our assertion that $-F$ is strictly completely monotonic.

As applications of the complete monotonicity of $-F$, the following sharp bounds for the gamma and digamma function were presented in [24] for every real $x \geq 1$ :

$$\omega \cdot \sqrt{2\pi}\left(\frac{x + \frac{1}{2}}{e}\right)^{x+\frac{1}{2}} \leq \Gamma(x+1) < \sqrt{2\pi}\left(\frac{x + \frac{1}{2}}{e}\right)^{x+\frac{1}{2}},$$

where the constant $\omega = \frac{2}{3\sqrt{3\pi}} e^{3/2} = 0.97323\cdots$ is best possible. For every real $x \geq 1$, it holds

$$\ln\left(x+\frac{1}{2}\right)-\frac{1}{x}<\psi\left(x\right)\leq \ln\left(x+\frac{1}{2}\right)-\frac{1}{x}+\zeta,$$

with best possible constant $\zeta = 1 - \ln\frac{3}{2} - \gamma = 0.01731\cdots$.

The same technique was used in [27] to prove the complete monotonicity of a class of functions related to the following inequalities

$$\frac{1}{\sqrt{\pi\left(n+\frac{1}{2}\right)}}<\frac{1\cdot 3\cdot 5\cdots(2n-1)}{2\cdot 4\cdot 6\cdots(2n)}<\frac{1}{\sqrt{\pi\left(n+\frac{1}{4}\right)}},\quad n\geq 1,$$

now called Kazarinoff's inequalities. Please see further details in [4], [8], [9], [12], [13]. Precisely, the function

$$F_a\left(x\right)=\ln\Gamma\left(x+1\right)-\ln\Gamma\left(x+\frac{1}{2}\right)-\frac{1}{2}\ln\left(x+a\right)$$

is completely monotonic when $a \in \left[0, \frac{1}{4}\right]$, while $-F_b$ is completely monotonic when $b \in \left[\frac{1}{2}, \infty\right)$. The following integral representation is valid

$$F_a''\left(x\right)=\int_0^{\infty}\frac{te^{-(x+1+a)t}}{1-e^{-t}}\varphi_a\left(t\right)dt,$$

where $\varphi_a$ admits the following expansion in power series in $t$ :

$$\varphi_a\left(t\right)=\sum_{k=2}^{\infty}w_k t^k,$$

with

$$w_k=a^k-\left(a+\frac{1}{2}\right)^k+\frac{1}{2}.$$

It is stated in [27, Lemma 2.1] that $w_k \geq 0$, if $a \in \left[0, \frac{1}{4}\right]$ and $w_k \leq 0$, if $a \in \left[\frac{1}{2}, \infty\right)$, so the previuos assertions on complete montonicity of functions $F_a$ are now proved. As a consequence, the following inequalities hold true for every $x \geq 1$,

$$\sqrt{x+\frac{1}{4}}<\frac{\Gamma\left(x+1\right)}{\Gamma\left(x+\frac{1}{2}\right)}\leq\omega\sqrt{x+\frac{1}{4}},$$

and

$$\mu\sqrt{x+\frac{1}{2}}\leq\frac{\Gamma\left(x+1\right)}{\Gamma\left(x+\frac{1}{2}\right)}<\sqrt{x+\frac{1}{2}},$$

where the constants $\omega = \frac{4}{\sqrt{5\pi}} = 1.00930\cdots$ and $\mu = \frac{2\sqrt{2}}{\sqrt{3\pi}} = 0.92132\cdots$ are best possible.

It is studied in [19] the following class of approximations for every real parameter $a$ :

$$\Gamma(x+1) \sim \nu_x(a) := \sqrt{2\pi e}e^{-a}\left(\frac{x+a}{e}\right)^{x+\frac{1}{2}}. \tag{8}$$

This class incorporates Stirling's formula $\Gamma(x+1) \sim \nu_x(0)$ , Burnside's formula $\Gamma(x+1) \sim \nu_x\left(\frac{1}{2}\right)$ , but also a recent formula discovered by Schuster [32]

$$\Gamma(x+1) \sim \sqrt{2\pi}e^{-\frac{1}{\sqrt{12}}}\left(\frac{x+\frac{1}{2}+\frac{1}{\sqrt{12}}}{e}\right)^{x+\frac{1}{2}},$$

which can be written as

$$\Gamma(x+1) \sim \nu_x\left(\frac{1}{2}+\frac{1}{\sqrt{12}}\right).$$

Schuster's formula demonstrates the preoccupation of the researchers to find increasingly better approximations of type (8). It is proven in [26] that the best approximations possible (8) are $\Gamma(x+1) \sim \nu_x(\omega)$ and $\Gamma(x+1) \sim \nu_x(\zeta)$ , where

$$\omega = \frac{3-\sqrt{3}}{6}, \quad \zeta = \frac{3+\sqrt{3}}{6}.$$

The following result was presented in [18] relative to the functions associated to (8):

$$G_a(x) = \ln\frac{\Gamma(x+1)}{\sqrt{2\pi e}e^{-a}\left(\frac{x+a}{e}\right)^{x+\frac{1}{2}}}.$$

This function $G_a$ is completely monotonic when $a \in [0,\omega]$ , while $-G_b$ is completely monotonic when $b \in \left[\frac{1}{2},\zeta\right]$ . As a consequence of the complete monotonicity of $G_\omega$ and $-G_\zeta$, the following double inequalities are valid for every $x \geq 0$ :

$$\sqrt{2\pi e}\cdot e^{-\omega}\left(\frac{x+\omega}{e}\right)^{x+1/2} < \Gamma(x+1) \leq \alpha\cdot\sqrt{2\pi e}\cdot e^{-\omega}\left(\frac{x+\omega}{e}\right)^{x+1/2},$$

where $\alpha = 1.07204\cdots$ , and

$$\beta\cdot\sqrt{2\pi e}\cdot e^{-\zeta}\left(\frac{x+\zeta}{e}\right)^{x+1/2} < \Gamma(x+1) \leq \sqrt{2\pi e}\cdot e^{-\zeta}\left(\frac{x+\zeta}{e}\right)^{x+1/2},$$

where $\beta = 0.98850\cdots$. By exploiting the monotonicity of $G'_\omega$ and $-G'_\zeta$, the following sharp inequalities on digamma function were presented in [18, Theorem 2.2]:

$$\ln\left(x+\frac{3-\sqrt{3}}{6}\right)+\frac{\sqrt{3}}{6x+3-\sqrt{3}}-\tau \le \psi(x)+\frac{1}{x} < \ln\left(x+\frac{3-\sqrt{3}}{6}\right)+\frac{\sqrt{3}}{6x+3-\sqrt{3}}$$

and

$$\ln\left(x+\frac{3+\sqrt{3}}{6}\right)-\frac{\sqrt{3}}{6x+3+\sqrt{3}} < \psi(x)+\frac{1}{x} \le \ln\left(x+\frac{3+\sqrt{3}}{6}\right)-\frac{\sqrt{3}}{6x+3+\sqrt{3}}+\sigma,$$

where $\tau = 0.00724\cdots$ and $\sigma = 0.00269\cdots$.

Furthermore using the monotonicity of $G''_\omega$ and $-G''_\zeta$, the following sharp inequalities on trigamma function were estabilshed in [18, Theorem 2.3]:

$$\frac{6}{6x+3-\sqrt{3}}-\frac{6\sqrt{3}}{\left(6x+3-\sqrt{3}\right)^2} < \psi'(x)-\frac{1}{x^2} \le \frac{6}{6x+3-\sqrt{3}}-\frac{6\sqrt{3}}{\left(6x+3-\sqrt{3}\right)^2}+\lambda$$

and

$$\frac{6}{6x+3+\sqrt{3}}+\frac{6\sqrt{3}}{\left(6x+3+\sqrt{3}\right)^2}-\nu \le \psi'(x)-\frac{1}{x^2} < \frac{6}{6x+3+\sqrt{3}}+\frac{6\sqrt{3}}{\left(6x+3+\sqrt{3}\right)^2},$$

where $\lambda = 0.01612\cdots$ and $\nu = 0.00436\cdots$.

As another example we present the following class of lower and upper bounds for gamma function:

$$\frac{n^{n+1}e^{-n}\sqrt{2\pi}}{\sqrt{n-\alpha}} \leq \Gamma(n+1) < \frac{n^{n+1}e^{-n}\sqrt{2\pi}}{\sqrt{n-\beta}}, \tag{9}$$

where $\alpha$, $\beta$ are any real numbers. Sandor and Debnath [31] found (9) with $\alpha = 0$, $\beta = 1$, while Batir [3] proposed better estimates using $\alpha = 1 - 2\pi e^{-2}$ and $\beta = 1/6$.

Motivated by the fact that the double inequality (9) can be rearranged as

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left(\frac{n}{n-\alpha}\right)^{1/2} \leq n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left(\frac{n}{n-\beta}\right)^{1/2}, \tag{10}$$

Mortici [21] introduced the class of approximations

$$\Gamma(n+1) \sim \mu_n(a,b) := \sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left(\frac{n+a}{n+b}\right)^{1/2}. \tag{11}$$

which enclose the previous formulas by Sandor and Debnath and Batir. It is proven that the most accurate approximation (11) is obtained in $a = 1/12$, $b = -1/12$ case. The corresponding approximation is better than those arising in (9)-(10). The next comparison table shows the superiority of (11) over

$$\Gamma(n+1) \sim \kappa_n := \sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left(\frac{n}{n-\frac{1}{6}}\right)^{1/2},$$

which is the best approximation among (9)-(10).

| $n$ | $\ln(\kappa_n/\Gamma(n+1))$ | $\ln(\mu_n/\Gamma(n+1))$ |
|---|---|---|
| 25 | $1.13 \times 10^{-5}$ | $1.90 \times 10^{-7}$ |
| 50 | $2.80 \times 10^{-6}$ | $2.37 \times 10^{-8}$ |
| 100 | $6.97 \times 10^{-7}$ | $2.97 \times 10^{-9}$ |
| 1000 | $6.94 \times 10^{-9}$ | $2.97 \times 10^{-12}$ |

It is considered in [21] the function associated to approximation formula (11):

$$G(x) = \ln\frac{\Gamma(x+1)}{\sqrt{2\pi x}\left(\frac{x}{e}\right)^x\left(\frac{x+\frac{1}{12}}{x-\frac{1}{12}}\right)^{1/2}}$$

and it has been proved that $-G$ is completely monotonic. As a direct consequence of this fact, the following sharp inequalities are valid for every real

$x \geq 1$ :

$$\omega \cdot \sqrt{2\pi x}\left(\frac{x}{e}\right)^x \left(\frac{x+\frac{1}{12}}{x-\frac{1}{12}}\right)^{1/2} < \Gamma(x+1) < \sqrt{2\pi x}\left(\frac{x}{e}\right)^x \left(\frac{x+\frac{1}{12}}{x-\frac{1}{12}}\right)^{1/2},$$

where $\omega = e\sqrt{\frac{11}{26\pi}} = 0.99754\cdots$ , and

$$\frac{1}{2\left(x+\frac{1}{12}\right)} - \frac{1}{2\left(x-\frac{1}{12}\right)} < \psi(x) - \left(\ln x - \frac{1}{2x}\right) \leq \frac{1}{2\left(x+\frac{1}{12}\right)} - \frac{1}{2\left(x-\frac{1}{12}\right)} + \tau,$$

with $\tau = -\gamma + \frac{167}{286} = 0.00670\cdots$ .

## 3 Further completely monotone functions

One of the first estimate for the remainder $\lambda_n$ in the Stirling formula

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\lambda_n}$$

was presented by Robbins [30], who proved

$$\frac{1}{12n+1} < \lambda_n < \frac{1}{12n}.$$

Increasingly better estimates were found by Maria [14], Nanjundiah, [28], or Shi et al [33]. Representations of the form

$$\Gamma(x+1) = \sqrt{2\pi}\left(\frac{x}{e}\right)^x e^{\theta(x)/12x}$$

were introduced in the recent past. Shi et al [33] proved that $\theta(x)$ is monotonically increasing on $[1,\infty)$. This result was extended by Mortici [22], who proved that $\theta$ decreases monotonically on $(0,\beta)$ and increases monotonically on $(\beta,\infty)$, where $\beta = 0.34142...$ is the solution of the equation

$$\ln\Gamma(x+1) + x\psi(x+1) - \ln\sqrt{2\pi} - 2x\ln x + x = 0.$$

Moreover $\theta$ is strictly convex on $(0,\infty)$ and the function $-x^{-1}\theta'''$ is completely monotonic on $(0,\infty)$.

It has been studied in [20] the remainder $w$ of the Burnside formula (6)

$$\Gamma(x+1) = \sqrt{2\pi}\left(\frac{x+1/2}{e}\right)^{x+1/2} e^{w(x)}$$

and stated that $-w$ is completely monotonic, in particular $w$ is concave.

Kečkić and Vasić [10] presented the following double inequality

$$\frac{x^{x-1}e^y}{y^{y-1}e^x} \le \frac{\Gamma(x)}{\Gamma(y)} \le \frac{x^{x-\frac{1}{2}}e^y}{y^{y-\frac{1}{2}}e^x}, \tag{12}$$

for all $x \ge y > 1$, which can be rewritten as

$$\frac{e^x\Gamma(x)}{x^{x-1/2}} \le \frac{e^y\Gamma(y)}{y^{y-1/2}} \quad \text{and} \quad \frac{e^y\Gamma(y)}{y^{y-1}} \le \frac{e^x\Gamma(x)}{x^{x-1}}.$$

This becomes equivalent to the fact that the function

$$f(x) = x + \ln\Gamma(x) - \left(x - \frac{1}{2}\right)\ln x - \ln\sqrt{2\pi}$$

is decreasing and the function

$$g(x) = x + \ln\Gamma(x) - (x-1)\ln x$$

is increasing. It is proved in [23] that the functions $f$ and $g'$ are completely monotonic on $(0,\infty)$. As a direct consequence, Kečkić-Vasić inequality (12) follows and it holds also for every $x \ge y > 0$. By using the monotonicity of $f'$ and $g'$, there are established the following sharp inequalities for every real $x \ge 1$:

$$\ln x - \frac{1}{2x} - \tau \le \psi(x) < \ln x - \frac{1}{2x},$$

where the constant $\tau = \gamma - \frac{1}{2} = 0.07721\cdots$ is the best possible, and for every real $x \ge 1$:

$$\ln x - \frac{1}{x} < \psi(x) \le \ln x - \frac{1}{x} + \sigma,$$

where the constant $\sigma = -\gamma + 1 = 0.42278\cdots$ is the best possible.

In 1965, Minc and Sathre [17] have given one of the first estimates of the expression $\phi(r) = (r!)^{1/r}$ and the ratio $\phi(r+1)/\phi(r)$ for every real $r \ge 1$:

$$1 < \frac{\phi(r+1)}{\phi(r)} < 1 + \frac{1}{r}. \tag{13}$$

Inequalities involving the function $\phi(r)$ are of interest in some branches of pure and applied mathematics and they have important applications in the theory of $(0,1)$-matrices.

Mortici [25] improved (13) in the sense of the following inequality for every $x \geq 1$ :

$$\frac{\Gamma(x+2)^{1/(x+1)}}{\Gamma(x+1)^{1/x}} \geq \frac{(4x+4)^{1/(x+1)}}{(4x)^{1/x}}\left(1+\frac{1}{x}\right) > 1.$$

The corresponding function

$$h(x) = x(x+1)\ln\frac{x\Gamma(x+1)^{1/(x+1)}}{(x+1)\Gamma(x)^{1/x}}$$

is considered and the complete monotonicity on $(1,\infty)$ of $h'$ is established. In particular $h'$ is positive, so $h$ is strictly increasing. In consequence, for every $x \geq 1$, we have $h(1) \leq h(x)$. As $h(1) = -\ln 4$, we obtain

$$-\ln 4 \leq x(x+1)\ln\frac{x\Gamma(x+1)^{1/(x+1)}}{(x+1)\Gamma(x)^{1/x}},$$

or

$$\frac{\Gamma(x+1)^{1/(x+1)}}{\Gamma(x)^{1/x}} \geq 4^{\frac{-1}{x(x+1)}}\left(1+\frac{1}{x}\right) > 1,$$

where the constant 4 is best possible. The obtained approximation formula

$$\frac{\Gamma(x+2)^{1/(x+1)}}{\Gamma(x+1)^{1/x}} \sim \frac{(4x+4)^{1/(x+1)}}{(4x)^{1/x}}\left(1+\frac{1}{x}\right),$$

is much better than Minc-Sathre. See [25].

**Acknowledgement 1** *This work was supported by a grant of the Romanian National Authority for Scientific Research, CNCS-UEFISCDI project number PN-II-ID-PCE-2011-3-0087. Some computations made in this paper were performed using Maple software. Part of this work was included in the habilitation thesis held by the author at the Politehnica University of Bucharest, Romania, September* 2012.

# References

[1] M. Abramowitz and I. Stegun, Handbook of Mathematical Functions. Dover Publications, Inc., New-York, 1970.

[2] H. Alzer, On some inequalities for the gamma and psi functions, Math. Comp., 66 (1997), no. 217, 373-389.

[3] N. Batir, Sharp inequalities for factorial $n$, Proyecciones, 27 (2008), no.1, 97-102.

[4] P. S. Bullen, A Dictionary of Inequalities, Pitman Monographs and Surveys in Pure and Applied Mathematics, 97, Addison Wesley Longman Limited, 1998.

[5] W. Burnside, A rapidly convergent series for $\log N!$, Messenger Math., 46(1917), 157-159.

[6] S. Guo and F. Qi, A class of logarithmically completely monotonic functions associated with the gamma function, J. Comput. Appl. Math. 224 (2009), 127-132.

[7] S. Guo and H. M. Srivastava, A certain function class related to the class of logarithmically completely monotonic functions, Math. Comp. Modelling 49 (2009), 2073-2079.

[8] D. K. Kazarinoff, On Wallis' formula, Edinburgh Math. Notes, 40 (1956), 19–21.

[9] N. D. Kazarinoff, Analytic Inequalities, Holt, Rhinehart and Winston, NewYork, 1961.

[10] D. Kečkić and P. M. Vasić, Some inequalities for the gamma function, Publ. Inst. Math. Beograd N. S., 11(1971), 107-114.

[11] S. Koumandos, Remarks on some completely monotonic functions, J. Math. Anal. Appl. 324 (2006), 1458-1461.

[12] [16] J.-Ch. Kuang, Chángyǒng BůdĽengshě (Applied Inequalities), 2nd edition, Hunan Education Press, Changsha, China, 1993. (Chinese)

[13] W. Magnus, F. Oberhettinger and R. P. Soni, Formulas and theorems for the special functions of mathematical physics, Springer, Berlin, 1966.

[14] A. J. Maria, A remark on Stirling's formula, Amer. Math. Monthly, 72 (1965), 1096-1098.

[15] M. Merkle, Completely monotone functions - a digest. arXiv:1211.0900, 2 nov. 2012.

[16] K. S. Miller, S. G. Samko, Completely monotonic functions. Integr. Transf. and Spec. Funct. 12 (2001), No 4, 389-402.

[17] H. Minc and L. Sathre, Some inequalities involving $(r!)^{1/r}$ , Proc. Edinburgh Math. Soc., (2) (14) (1964/1965), 41-46.

[18] C. Mortici, Estimating the digamma and trigamma functions by completely monotonicity arguments, Appl. Math. Comp. 217 (2010), no. 8, 4081-4085.

[19] C. Mortici, The asymptotic series of the generalized Stirling formula, Comp. Math. Appl., 60 (2010), no. 3, 786-791.

[20] C. Mortici, Completely monotonic functions associated with gamma function and applications, Carpathian J. Math., 25 (2009), no. 2, 186-191.

[21] C. Mortici, Some completely monotonic functions relating to the gamma function, Int. Transf. Spec. Fct. 23 (2012), no. 7, 473-479.

[22] C. Mortici, On the monotonicity and convexity of the remainder of the Stirling formula, Appl. Math. Lett., 24 (2011), no. 6, 869-871.

[23] C. Mortici, On an inequality of Keckic and Vasic, Autom. Comp. Appl. Math., 18 (2009) , no. 1, 163-169.

[24] C. Mortici, Sharp bounds for gamma and digamma function arising from Burnside's formula, Rev. Anal. Numer. Th. Approx., 39 (2010), no. 1, 69-72.

[25] C. Mortici, A lower bound for a quotient of roots of factorials, Bul. Acad. Şt. Rep. Moldova Mat., 3 (61) (2009), 64-69.

[26] C. Mortici, An ultimate extremely accurate formula for approximation of the factorial function, Arch. Math. 93 (2009), no. 1, 37-45.

[27] C. Mortici, Sharp inequalities and complete monotonicity for the Wallis ratio, Bull. Belgian Math. Soc. Simon Stevin, 17 (2010), 929-936.

[28] T. S. Nanjundiah, Note on Stirling's formula, Amer. Math. Monthly, 66 (1965), 701-703.

[29] F. Qi and S. Guo, A class of logarithmically completely monotonic functions and the best bounds in the second Kershaw's double inequality, J. Comput. Appl. Math. 212 (2008), 444-456.

[30] H. Robbins, A Remark of Stirling's Formula, Amer. Math. Monthly, 62 (1955), 26-29.

[31] J. Sandor and L. Debnath, On certain inequalities involving the constant $e$ and their applications, J. Math. Anal. Appl. 249 (2000), 569-582.

[32] W. Schuster, Improving Stirling's formula, Arch. Math. 77 (2001) 170–176.

[33] X. Shi, F. Liu and M. Hu, A new asymptotic series for the gamma function, J. Comput. Appl. Math., 195 (2006), 134-154.

[34] H. M. Srivastava, S. Guo and F. Qi, Some properties of a class of functions related to completely monotonic functions, Comp. Math. Appl. 64 (2012), 1649-1654.

[35] D.V. Widder, The Laplace Transform, Princeton Univ. Press, Princeton, 1946.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 6, No. 2/2014

# Coefficient bounds for a subclass of Bi-univalent functions using differential operators*

Chellian Selvaraj[†] Ganapathi Thirupathi[‡]

**Abstract**

In the present paper, we introduce new subclass $ST_{\Sigma}(b, \phi)$ of bi-univalent functions defined in the open disk. Furthermore, we find upper bounds for the second and third coefficients for functions in these new subclass using differential operator.
**MSC**: 30C45

**Keywords**: bi-univalent functions, coefficient estimates, starlike function, convex function, differential operator.

## 1 Introduction. Definitions And Preliminaries

Let $\mathcal{A}$ denote the class of functions $f(z)$ of the form

$$f(z) = z + \sum_{n=2}^{\infty} a_n z^n, \tag{1.1}$$

which are analytic in the open unit disk $\mathbb{U} = \{z \in \mathcal{C} : |z| < 1\}$. Further, by $\mathcal{S}$ we shall denote the class of functions $f \in \mathcal{A}$ which are univalent in $\mathbb{U}$.

Since univalent functions are one-to-one, they are invertible and the inverse functions need not be defined on the entire unit disk $\mathbb{U}$. However, the

*Accepted for publication in revised form on April 5-th, 2014
[†]Presidency College, Chennai-600 005, Tamilnadu, India, pamc9439@yahoo.co.in
[‡]R.M.K.Engineering College, R.S.M.Nagar, Kavaraipettai-601 206, Tamilnadu, India gtvenkat79@gmail.com

famous Koebe one-quarter theorem ensures that the image of the unit disk $\mathbb{U}$ under every function $f \in \mathcal{A}$ contains a disk of radius $1/4$. Thus every univalent function $f$ has an inverse $f^{-1}$ satisfying $f^{-1}(f(z)) = z$, $(z \in \mathbb{U})$ and $f\left(f^{-1}(w)\right) = w$, $\left(|w| < r_0(f), r_0(f) \geq \frac{1}{4}\right)$ where

$$f^{-1}(w) = w - a_2 w^2 + (2a_2^2 - a_3)w^3 - (5a_2^3 - 5a_2 a_3 + a_4)w^4 + \cdots . \quad (1.2)$$

A function $f \in \mathcal{A}$ is said to be bi-univalent in $\mathbb{U}$ if both $f(z)$ and $f^{-1}(z)$ are univalent in $\mathbb{U}$. We let $\Sigma$ to denote the class of bi-univalent functions in $\mathbb{U}$ given by (1.1). If $f(z)$ is bi-univalent, it must be analytic in the boundary of the domain and such that it can be continued across the boundary of the domain so that $f^{-1}(z)$ is defined and analytic throughout $|w| < 1$. Examples of functions in the class $\Sigma$ are

$$\frac{z}{1-z}, -log\,(1-z)$$

and so on.

The coefficient estimate problem for the class $\mathcal{S}$, known as the Bieberbach conjecture, is settled by de-Branges [4], who proved that for a function $f(z) = z + \sum_{n=2}^{\infty} a_n z^n$ in the class $\mathcal{S}$, $|a_n| \leq n$, for $n = 2, 3, \cdots$, with equality only for the rotations of the Koebe function

$$K_0(z) = \frac{z}{(1-z)^2}.$$

In 1967, Lewin [9] introduced the class $\Sigma$ of bi-univalent functions and showed that $|a_2| < 1.51$ for the functions belonging to $\Sigma$. It was earlier believed that for $f \in \Sigma$, the bound was $|a_n| < 1$ for every $n$ and the extremal function in the class was $\frac{z}{1-z}$. E.Netanyahu [11] in 1969, ruined this conjecture by proving that in the set $\Sigma$, $\max_{f\in\Sigma} |a_2| \leq 4/3$. In 1969, Suffridge [15] gave an example of $f \in \Sigma$ for which $a_2 = 4/3$ and conjectured that $|a_2| \leq 4/3$. In 1981, Styer and Wright [14] disproved the conjecture that $|a_2| > 4/3$. Brannan and Clunie [2] conjectured that $|a_2| \leq \sqrt{2}$. Kedzierawski [7] in 1985 proved this conjecture for a special case when the function $f$ and $f^{-1}$ are starlike functions. Brannan and Clunie [2] conjectured that $|a_2| \leq \sqrt{2}$. Tan [16] in proved that $|a_2| \leq 1.485$ which is the best known estimate for functions in the class of bi-univalent functions.

Brannan and Taha [3] introduced certain subclasses of the bi-univalent function class $\Sigma$ similar to the familiar subclasses $S^*(\alpha)$ and $C(\alpha)$ of the univalent function class $\Sigma$. Recently, Ali et al.[1] extended the results of Brannan and Taha [3] by generalising their classes using subordination.

An analytic function $f$ is subordinate to an analytic function $g$, written $f(z) \prec g(z)$, provided there is a Schwarz function $w$ defined on $\mathbb{U}$ with $w(0) = 0$ and $|w(z)| < 1$ satisfying $f(z) = g(w(z))$. Ma and Minda [10], unified various subclasses of starlike and convex functions for which either of the quantity $\dfrac{zf'(z)}{f(z)}$ or $1 + \dfrac{zf''(z)}{f'(z)}$ is subordinate to a more general superordinate function. For this purpose, they considered an analytic function $\phi$ with positive real part in the unit disk $U$, $\phi(0) = 1$, $\phi'(0) > 0$ and $\phi$ maps $U$ onto a region starlike with respect to 1 and symmetric with respect to the real axis. Such a function has a series expansion of the form

$$\phi(z) = 1 + B_1 z + B_2 z^2 + B_3 z^3 + \cdots, (B_1 > 0). \tag{1.3}$$

Recently Selvaraj and Karthikeyan [8] defined the following operator $D_\lambda^m(\alpha_1, \beta_1) f : \mathbb{U} \to \mathbb{U}$ by

$$\begin{aligned} &D_\lambda^0(\alpha_1;\, \beta_1) f(z) = f(z) * \mathcal{G}_{q,s}(\alpha_1,\, \beta_1;\, z), \\ &D_\lambda^1(\alpha_1;\, \beta_1) f(z) = (1-\lambda)(f(z) * \mathcal{G}_{q,s}(\alpha_1,\, \beta_1;\, z)) + \lambda\, z(f(z) * \mathcal{G}_{q,s}(\alpha_1,\, \beta_1;\, z))', \\ &D_\lambda^m(\alpha_1;\, \beta_1) f(z) = D_\lambda^1(D_\lambda^{m-1}(\alpha_1;\, \beta_1) f(z)), \end{aligned} \tag{1.4}$$

where $m \in \mathbb{N}_0$, $\lambda \geq 0$.

If $f \in \mathcal{A}$, then from (1.4) we may easily deduce that

$$D_\lambda^m(\alpha_1;\, \beta_1) f(z) = z + \sum_{n=2}^{\infty} \left[1 + (n-1)\lambda\right]^m \frac{(\alpha_1)_{n-1} \,\ldots\, (\alpha_q)_{n-1}}{(\beta_1)_{n-1} \,\ldots\, (\beta_s)_{n-1}} \frac{a_n z^n}{(n-1)!}. \tag{1.5}$$

Special cases of the operator $D_\lambda^m(\alpha_1;\, \beta_1) f$ includes various other linear operators which were considered in many earlier work on the subject of analytic and univalent functions. If we let $m = 0$ in $D_\lambda^m(\alpha_1;\, \beta_1) f$, we have

$$D_\lambda^0(\alpha_1;\, \beta_1) f(z) = \mathcal{H}_q^1(\alpha_1;\, \beta_1) f(z)$$

where $\mathcal{H}_{q,s}^1(\alpha_1;\, \beta_1)$ is Dziok-Srivastava operator for functions in $\mathcal{A}$ (see [6]) and for $q = 2, s = 1$ $\alpha_1 = \beta_1, \alpha_2 = 1$ and $\lambda = 1$, we get the operator introduced by Salagean([13]). It can be easily verified from the definition of (1.5),

$$z\left(D_\lambda^m(\alpha_1,\, \beta_1) f(z)\right)' = (\alpha_1 + 1) D_\lambda^m(\alpha_1 + 1,\, \beta_1) f(z) - \alpha_1 D_\lambda^m(\alpha_1,\, \beta_1) f(z). \tag{1.6}$$

**Definition 1.1** *Let $b$ be a non-zero complex number. A function $f(z)$ given by (1.1) is said to be in the class $ST_{\Sigma}(b,\phi)$ if the following conditions are satisfied:*

$$f\in\Sigma \quad and \quad 1+\frac{1}{b}\left(\frac{D_{\lambda}^{m+1}(\alpha_1,\beta_1)f(z)}{D_{\lambda}^{m}(\alpha_1,\beta_1)f(z)}-1\right)\prec\phi(z),\quad z\in\mathbb{U} \tag{1.7}$$

$$and \quad 1+\frac{1}{b}\left(\frac{D_{\lambda}^{m+1}(\alpha_1,\beta_1)g(w)}{D_{\lambda}^{m}(\alpha_1,\beta_1)g(w)}-1\right)\prec\phi(z),\quad z\in\mathbb{U} \tag{1.8}$$

*where the function $g$ is given by (1.2).*

**Definition 1.2** *Let $b$ be a non-zero complex number. A function $f(z)$ given by (1.1) is said to be in the class $ST_{\Sigma}(\alpha_1,\beta_1,b,\phi)$ if the following conditions are satisfied:*

$$f\in\Sigma \quad and \quad 1+\frac{1}{b}\left(\frac{D_{\lambda}^{m}(\alpha_1+1,\beta_1)f(z)}{D_{\lambda}^{m}(\alpha_1,\beta_1)f(z)}-1\right)\prec\phi(z),\quad z\in\mathbb{U} \tag{1.9}$$

$$and \quad 1+\frac{1}{b}\left(\frac{D_{\lambda}^{m}(\alpha_1+1,\beta_1)g(w)}{D_{\lambda}^{m}(\alpha_1,\beta_1)g(w)}-1\right)\prec\phi(w),\quad w\in\mathbb{U}, \tag{1.10}$$

*where the function $g$ is given by (1.2).*

## 2 Coefficient estimates

**Lemma 2.1** *[12] If $p\in\wp$, then $|c_k|\le 2$ for each $k$, where $\wp$ is the family of functions $p$ analytic in $\mathbb{U}$ for which $Re p(z)>0$, $p(z)=1+c_1z+c_2z^2+\cdots$ for $z\in\mathbb{U}$.*

**Theorem 2.2** *Let the function $f(z)\in\mathcal{A}$ be given by (1.1). If $f\in ST_{\Sigma}(b,\phi)$, then*

$$|a_2|\le\frac{B_1\sqrt{B_1}\,|b|}{\sqrt{\left(4(1+2\lambda)^m-(1+\lambda)^{2m}\right)B_1^2b\lambda+(B_1-B_2)\lambda^2(1+\lambda)^{2m}}} \tag{2.1}$$

*and*

$$|a_3|\le\frac{(B_1+|B_2-B_1|)\,|b|}{\lambda\left(4(1+2\lambda)^m-(1+\lambda)^{2m}\right)}.$$

**Proof.** Since $f \in ST_{\Sigma}(b,\phi)$, there exists two analytic functions $r, s : \mathbb{U} \to \mathbb{U}$, with $r(0) = 0 = s(0)$, such that

$$1+\frac{1}{b}\left(\frac{D_{\lambda}^{m+1}(\alpha_1,\beta_1)\,f(z)}{D_{\lambda}^{m}(\alpha_1,\beta_1)\,f(z)}-1\right)=\phi(r(z)) \tag{2.2}$$

and

$$1+\frac{1}{b}\left(\frac{D_{\lambda}^{m+1}(\alpha_1,\beta_1)\,g(w)}{D_{\lambda}^{m}(\alpha_1,\beta_1)\,g(w)}-1\right)=\phi(s(z)).$$

It is also written as

$$\begin{aligned}1+\frac{1}{b}\left(\frac{D_{\lambda}^{m+1}(\alpha_1,\beta_1)\,f(z)-D_{\lambda}^{m}(\alpha_1,\beta_1)\,f(z)}{D_{\lambda}^{m}(\alpha_1,\beta_1)\,f(z)}\right)&=\phi(r(z)) \quad \text{and}\\ 1+\frac{1}{b}\left(\frac{D_{\lambda}^{m+1}(\alpha_1,\beta_1)\,g(w)-D_{\lambda}^{m}(\alpha_1,\beta_1)\,g(w)}{D_{\lambda}^{m}(\alpha_1,\beta_1)\,g(w)}\right)&=\phi(s(z)).\end{aligned} \tag{2.3}$$

Define the functions $p$ and $q$ by

$$p(z)=\frac{1+r(z)}{1-r(z)}=1+p_1z+p_2z^2+\cdots \text{ and } q(z)=\frac{1+s(z)}{1-s(z)}=1+q_1z+q_2z^2+\cdots. \tag{2.4}$$

Or equivalently,

$$\begin{aligned}r(z)=\frac{p(z)-1}{p(z)+1}=\frac{1}{2}\Bigg(p_1z+\left(p_2-\frac{p_1^2}{2}\right)z^2+\\ \left(p_3+\frac{p_1}{2}\left(\frac{p_1^2}{2}-p_2\right)-\frac{p_1p_2}{2}\right)z^3+\cdots\end{aligned} \tag{2.5}$$

and

$$\begin{aligned}s(z)=\frac{q(z)-1}{q(z)+1}=\frac{1}{2}\Bigg(q_1z+\left(q_2-\frac{q_1^2}{2}\right)z^2+\\ \left(q_3+\frac{q_1}{2}\left(\frac{q_1^2}{2}-q_2\right)-\frac{q_1q_2}{2}\right)z^3+\cdots.\end{aligned} \tag{2.6}$$

It is clear that $p$ and $q$ are analytic in $\mathbb{U}$ and $p(0) = 1 = q(0)$. Also $p$ and $q$ have positive real part in $\mathbb{U}$ and hence $|p_i| \le 2$ and $|q_i| \le 2$. In the view of (2.3), (2.4)and (2.5), clearly, Using (2.5) and (2.6), one can easily verify that

$$\phi\left(\frac{p(z)-1}{p(z)+1}\right)=1+\frac{B_1p_1}{2}z+\left(\frac{B_1}{2}\left(p_2-\frac{p_1^2}{2}\right)+\frac{1}{4}B_2p_1^2\right)z^2+\cdots \tag{2.7}$$

and

$$\phi\left(\frac{q(w)-1}{q(w)+1}\right)=1+\frac{B_1q_1}{2}w+\left(\frac{B_1}{2}\left(q_2-\frac{q_1^2}{2}\right)+\frac{B_2q_1^2}{4}\right)w^2+\cdots. \tag{2.8}$$

Since $f\in\Sigma$ has the Maclaurin series given by (1.1), computation shows that its inverse $g=f^{-1}$ has the expansion given by (1.2). It follows from (2.6), (2.7) and (2.8) that

$$(1+\lambda)^m a_2=\frac{1}{2\lambda}B_1p_1b, \tag{2.9}$$

$$4\lambda(1+2\lambda)^m a_3-\lambda(1+\lambda)^{2m}a_2^2=\frac{1}{2}bB_1\left(p_2-\frac{1}{2}p_1^2\right)+\frac{1}{4}bB_2p_1^2 \tag{2.10}$$

and

$$-(1+\lambda)^m a_2=\frac{1}{2\lambda}B_1bq_1, \tag{2.11}$$

$$\lambda\left(8\lambda(1+2\lambda)^m-(1+\lambda)^{2m}\right)a_2^2-4\lambda(1+2\lambda)^m a_3=\frac{1}{2}bB_1\left(q_2-\frac{1}{2}q_1^2\right)+\frac{1}{4}bB_2q_1^2. \tag{2.12}$$

From (2.9) and (2.11), it follows that

$$p_1=-q_1. \tag{2.13}$$

Now (2.10), (2.12) and (2.13) gives

$$a_2^2=\frac{B_1^3b^2(p_2+q_2)}{4\left[\left(4(1+2\lambda)^m-(1+\lambda)^{2m}\right)B_1^2b\lambda+(B_1-B_2)\lambda^2(1+\lambda)^{2m}\right]}. \tag{2.14}$$

Using the fact that $|p_2|\le 2$ and $|q_2|\le 2$ gives the desired estimate on $|a_2|$,

$$|a_2| \leq \frac{B_1\sqrt{B_1}\,|b|}{\sqrt{\left(4\left(1+2\lambda\right)^m-\left(1+\lambda\right)^{2m}\right)B_1^2 b\lambda+\left(B_1-B_2\right)\lambda^2\left(1+\lambda\right)^{2m}}}.$$

From (2.10)-(2.12), gives

$$a_3 = \frac{\frac{bB_1}{2}\left[8\left(1+2\lambda\right)^m-\left(1+\lambda\right)^{2m}\right)p_2+(1+\lambda)^{2m}q_2\right]}{8\lambda\left[4(1+2\lambda)^{2m}-(1+\lambda)^{2m}(1+2\lambda)^m\right]}$$

$$+\frac{2(1+2\lambda)^m p_1^2\left(B_2-B_1\right)b}{8\lambda\left[4(1+2\lambda)^{2m}-(1+\lambda)^{2m}(1+2\lambda)^m\right]}$$

Using the inequalities $|p_1| \leq 2$, $|p_2| \leq 2$ and $|q_2| \leq 2$ for functions with positive real part yields the desired estimation of $|a_3|$.

■ For a choice of $\phi(z) = \dfrac{1+Az}{1+Bz}$, $-1 \leq B < A \leq 1$, we have the following corollary.

**Corollary 2.3** *Let* $-1 \leq B < A \leq 1$. *If* $f \in ST_\Sigma\left(b, \frac{1+Az}{1+Bz}\right)$, *then*

$$|a_2| \leq \frac{|b|\left(A-B\right)}{\sqrt{\left(4\left(1+2\lambda\right)^m-\left(1+\lambda\right)^{2m}\right)\left(A-B\right)b\lambda+\left(1+B\right)\lambda^2\left(1+\lambda\right)^{2m}}}$$

*and*

$$|a_3| \leq \frac{|A-B|\left(1+|1+B|\right)|b|}{\lambda\left(4\left(1+2\lambda\right)^m-\left(1+\lambda\right)^{2m}\right)}.$$

**Theorem 2.4** *Let the function* $f(z) \in \mathcal{A}$ *be given by (1.1). If* $ST_\Sigma\left(\alpha_1, \beta_1, b, \phi\right)$, *then*

$$|a_2| \leq \frac{\left(\alpha_1+1\right)B_1\sqrt{B_1}\,|b|}{\sqrt{\left(4\left(1+2\lambda\right)^m-\left(1+\lambda\right)^{2m}\right)B_1^2 b\left(\alpha_1+1\right)+\left(B_1-B_2\right)\left(1+\lambda\right)^{2m}}} \tag{2.15}$$

*and*

$$|a_3| \leq \frac{\left(\alpha_1+1\right)\left(B_1+|B_2-B_1|\right)|b|}{\left(4\left(1+2\lambda\right)^m-\left(1+\lambda\right)^{2m}\right)}.$$

**Proof.** Since $ST_{\Sigma}(\alpha_1, \beta_1, b, \phi)$, there exists two analytic functions $r, s : \mathbb{U} \to \mathbb{U}$, with $r(0) = 0 = s(0)$, such that

$$1 + \frac{1}{b}\left(\frac{D_{\lambda}^{m}(\alpha_1 + 1, \beta_1) f(z)}{D_{\lambda}^{m}(\alpha_1, \beta_1) f(z)} - 1\right) = \phi(r(z)) \tag{2.16}$$

and

$$1 + \frac{1}{b}\left(\frac{D_{\lambda}^{m}(\alpha_1 + 1, \beta_1) g(w)}{D_{\lambda}^{m}(\alpha_1, \beta_1) g(w)} - 1\right) = \phi(s(z)).$$

Using (2.3), (2.4), (2.7) and (2.8), one can easily verified that

$$(1+\lambda)^{m} a_2 = \frac{(\alpha_1 + 1)}{2} B_1 p_1 b, \tag{2.17}$$

$$4(1+2\lambda)^{m} a_3 - (1+\lambda)^{2m} a_2^2 = (\alpha_1 + 1)\left[\frac{1}{2} b B_1 \left(p_2 - \frac{1}{2} p_1^2\right) + \frac{1}{4} b B_2 p_1^2\right] \tag{2.18}$$

and

$$-(1+\lambda)^{m} a_2 = \frac{(\alpha_1 + 1)}{2} B_1 p_1 b, \tag{2.19}$$

$$\left(8(1+2\lambda)^{m} - (1+\lambda)^{2m}\right) a_2^2 - 4(1+2\lambda)^{m} a_3 =$$

$$= (\alpha_1 + 1)\left[\frac{1}{2} b B_1 \left(q_2 - \frac{1}{2} q_1^2\right) + \frac{1}{4} b B_2 q_1^2\right]. \tag{2.20}$$

From (2.17) and (2.19), it follows that

$$p_1 = -q_1. \tag{2.21}$$

Now (2.18), (2.20), (2.21) and using the fact that $|p_2| \leq 2$ and $|q_2| \leq 2$,

$$|a_2| \leq \frac{|\alpha_1 + 1| B_1 \sqrt{B_1} |b|}{\sqrt{\left(4(1+2\lambda)^{m} - (1+\lambda)^{2m}\right) B_1^2 b(\alpha_1 + 1) + (B_1 - B_2)(1+\lambda)^{2m}}}.$$

From (2.18)-(2.20), gives

$$|a_3| \leq \frac{|\alpha_1 + 1| (B_1 + |B_2 - B_1|) |b|}{\left(4(1+2\lambda)^{m} - (1+\lambda)^{2m}\right)}.$$

■

## References

[1] Ali, Rosihan M.; Lee, See Keong; Ravichandran, V.; Supramaniam, Shamani. Coefficient estimates for bi-univalent Ma-Minda starlike and convex functions. Appl. Math. Lett. 25 (2012), no. 3, 344–351.

[2] Brannan, D. A.; Clunie, J.; Kirwan, W. E. Coefficient estimates for a class of star-like functions. Canad. J. Math. 22 1970 476–485.

[3] Brannan, D. A.; Taha, T. S. On some classes of bi-univalent functions. Studia Univ. Babes,-Bolyai Math. 31 (1986), no. 2, 70–77.

[4] L. de Branges, A proof of the Bieberbach conjecture, Acta Math. **154** (1985), no. 1-2, 137–152.

[5] Duren, Peter L. Univalent functions. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 259. Springer-Verlag, New York, 1983.

[6] J. Dziok and H. M. Srivastava, Classes of analytic functions associated with the generalized hypergeometric function, Appl. Math. Comput. **103** (1999), no. 1, 1–13.

[7] Kedzierawski, A.; Waniurski, J. Bi-univalent polynomials of small degree. Complex Variables Theory Appl. 10 (1988), no. 2-3, 97–100.

[8] C. Selvaraj and K. R. Karthikeyan, Differential sandwich theorems for certain subclasses of analytic functions, Math. Commun.**13** (2008), no. 2, 311–319.

[9] Lewin. M , On a coefficient problem for bi-unilvalent functions, Appl.MAth.Lett.24(2011)1569-1573.

[10] Ma, Wan Cang; Minda, David. A unified treatment of some special classes of univalent functions. Proceedings of the Conference on Complex Analysis (Tianjin, 1992), 157–169, Conf. Proc. Lecture Notes Anal., I, Int. Press, Cambridge, MA, 1994.

[11] Netanyahu, E. The minimal distance of the image boundary from the origin and the second coefficient of a univalent function in $z < 1$. Arch. Rational Mech. Anal. 32 1969 100–112.

[12] Pommerenke, Christian. Univalent functions. With a chapter on quadratic differentials by Gerd Jensen. Studia Mathematica/Mathematische Lehrbücher, Band XXV. Vandenhoeck Ruprecht, Göttingen, 1975. 376 pp

[13] Salagean, Grigore Ştefan. Subclasses of univalent functions. Complex analysis fifth Romanian-Finnish seminar, Part 1 (Bucharest, 1981), 362–372, Lecture Notes in Math., 1013, Springer, Berlin, 1983.

[14] Styer, D.; Wright, D. J. Results on bi-univalent functions. Proc. Amer. Math. Soc. 82 (1981), no. 2, 243–248.

[15] Suffridge, T. J. A coefficient problem for a class of univalent functions. Michigan Math. J. 16 1969 33–42.

[16] Tan, De Lin. Coefficient estimates for bi-univalent functions. (Chinese) An English summary appears in Chinese Ann. Math. Ser. B 5 (1984), no. 4, 741742.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 6, No. 2/2014

# OPTIMAL THICKNESS OF A CYLINDRICAL SHELL*

Paul Ziemann†

**Abstract**

In this paper an optimization problem for a cylindrical shell is discussed. The aim is to look for an optimal thickness of a shell to minimize the deformation under an applied external force. As a side condition, the volume of the shell has to stay constant during the optimization process. The deflection is calculated using an approach from shell theory. The resulting control-to-state operator is investigated analytically and a corresponding optimal control problem is formulated. Moreover, necessary conditions for an optimal solution are stated and numerical solutions are presented for different examples.

**MSC**: 49K15, 49J15, 49Q10

**keywords:** Optimal control of PDE, Shape optimization, Linear elasticity

## 1 Introduction

In this paper we discuss an optimization problem in linear elasticity, particularly in shape optimization. In this field, much research has been done in the last years. Some few representative books from Sokolowski [1], Pironneau [2], Haslinger [3] and Delfour [4] should be mentioned here.

*Accepted for publication in revised form on September 25-th, 2014

†paul.ziemann@uni-greifswald.de, Ernst-Moritz-Arndt-Universität Greifswald, Institut für Mathematik und Informatik, Walther-Rathenau-Str. 47, D-17487 Greifswald, Germany

In particular, the book from Neittanmaki, Sprekels and Tiba [5] deals with similar problems, though they are using a different model for calculating the deformation. A related problem is also investigated in a paper by Nestler [6], where a simplified (rotational symmetric) case has been handled. That paper was inspired by works from Lepik, Lepikult, Lellep and Schmidt [7, 8, 9, 10].

The task is to look for an optimal thickness of a cylindrical shell to minimize the deformation under an applied external force. In this paper, the stationary case with a loading applied on the shell's midsurface is treated. As an additional restriction, the volume of the shell has to stay constant during the optimization process. Moreover, the thickness should only vary between specified bounds. The deflection is modelled using the "basic shell model" from Chapelle and Bathe [11] which makes use of the Hypothesis from Mindlin and Reissner. As a main result it is shown, that the resulting control-to-state operator is continuous and Gâteaux-differentiable. Followed by this, a corresponding optimal control problem is formulated and necessary conditions for an optimal solution are deduced. Necessary conditions for similar problems can be found e.g. for the rotational symmetric case in [6] and for elastic beams with piecewise constant thickness in [12]; those restrictions are not necessary in this paper. We also investigate the particular numerical implementation of the problem which makes use of an analytically calculated formula for the objective gradient. Finally, numerical solutions for different examples are presented and investigated in relation to the fulfillment of the necessary conditions and convergence properties on refined grids.

## 2 Geometrical description of the shell

For the geometrical description, we first need a chart describing the midsurface of the shell. Let $\omega \subset \mathbf{R}^2$ be open and connected and $\phi : \overline{\omega} \to \mathbf{R}^3$ be an injective mapping with $\phi \in W^{2,\infty}(\omega)$. We call $\mathbf{S} = \phi(\overline{\omega})$ the midsurface of the shell. We assume that the vectors $a_\alpha := \frac{\partial \phi}{\partial \xi^\alpha}$, $\alpha = 1, 2$ are linearly independent and additionally consider an orthonormal vector $a_3 := \frac{a_1 \times a_2}{\|a_1 \times a_2\|}$. We call $a_1$, $a_2$ a covariant basis of the tangent plane of the midsurface and denote the corresponding contravariant basis by $a^1$, $a^2$. Moreover, denote by $a_{\alpha\beta} := a_\alpha \cdot a_\beta$, $\alpha, \beta = 1, 2$ the covariant components of the first fundamental form.

In our particular case involving a cylindrical shell, the set $\omega$ can be chosen as $\omega = \{(0, L) \times (\psi_a, \psi_b)\}$ and the mapping $\phi$ is defined as $\phi(\xi^1, \xi^2) = (\xi^1, R \cos \xi^2, R \sin \xi^2)$, where $R$ is the radius of the shell.

We introduce $t : \mathbf{S} \to \mathbf{R}^+$, $t \in C^{0,1}(\mathbf{S})$ as the *thickness* of the shell and suppress the parametrization $\phi$ in $t \circ \phi : \omega \to \mathbf{R}^+$ when the context is clear. Let us define the 3D-reference domain

$$\Omega_{(t)} := \left\{ (\xi^1, \xi^2, \xi^3) \in \mathbf{R}^3 \mid (\xi^1, \xi^2) \in \omega, \xi^3 \in \left( \frac{-t(\xi^1, \xi^2)}{2}, \frac{t(\xi^1, \xi^2)}{2} \right) \right\} \quad (1)$$

together with the mapping

$$\Phi_{(t)} : \overline{\Omega}_{(t)} \to \mathbf{R}^3, \quad \Phi_{(t)}(\xi^1, \xi^2, \xi^3) = \phi(\xi^1, \xi^2) + \xi^3 a_3. \quad (2)$$

Note that $\Phi_{(t)}$ depends on the parameter $t$ only via its domain, but not on the right hand side. So the thickness parameter is surpressed for $\Phi$ and the derived geometrical quantities in the following text.

We call $\mathbf{B}_{(t)} := \Phi(\Omega_{(t)})$ the shell body, see e.g. figure 1. Let us denote the local covariant and contravariant basis with $g_i$ and $g^i$, $i = 1, 2, 3$ and the covariant and contravariant components of the metric tensor with $g_{ij} = g_i \cdot g_j$, $g^{ij} = g^i \cdot g^j$, $i, j = 1, 2, 3$, resp. Furthermore we assume $t(\xi^1, \xi^2) < 2R$ which is satisfied in general, since for any shell model one assumes that the thickness is much smaller than the principal radii of curvature. In our case, this means $t(\xi^1, \xi^2) \ll R$. Back to our problem, we get

$$\mathbf{B}_{(t)} = \left\{ \begin{pmatrix} \xi^1 \\ (R + \xi^3)\cos(\xi^2) \\ (R + \xi^3)\sin(\xi^2) \end{pmatrix} \mid (\xi^1, \xi^2) \in \omega, \xi^3 \in \left( \frac{-t(\xi^1, \xi^2)}{2}, \frac{t(\xi^1, \xi^2)}{2} \right) \right\} \quad (3)$$

together with the contravariant basis

$$g^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad g^2 = \frac{1}{R + \xi^3} \begin{pmatrix} 0 \\ -\sin(\xi^2) \\ \cos(\xi^2) \end{pmatrix}, \quad g^3 = \begin{pmatrix} 0 \\ \cos(\xi^2) \\ \sin(\xi^2) \end{pmatrix}. \quad (4)$$

The surface and volume element for our shell are given by

$$\begin{aligned} \mathrm{d}S &= \sqrt{a}\,\mathrm{d}\xi^1\,\mathrm{d}\xi^2, & a &= \det(a_{\alpha\beta}) = R^2 \\ \mathrm{d}V &= \sqrt{g}\,\mathrm{d}\xi^1\,\mathrm{d}\xi^2\,\mathrm{d}\xi^3, & g &= \det(g_{mn}) = \sqrt{a}(1 + \tfrac{\xi^3}{R}). \end{aligned} \quad (5)$$

## 3 Modeling the displacement

We consider a *small* displacement $U : \mathbf{B}_{(t)} \to \mathbf{R}^3$ of the shell body. For modeling we use the Reissner-Mindlin kinematical assumptions which

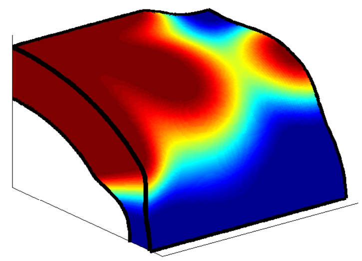


Figure 1: Cylindrical shell with non-constant thickness

state that normals to the midsurface remain straight and unstretched during deformation. This leads to the displacement ansatz

$$U(\xi^1, \xi^2, \xi^3) = u(\xi^1, \xi^2) + \xi^3 \theta(\xi^1, \xi^2) \tag{6}$$

with $u = u_1 a^1 + u_2 a^2 + u_3 a_3$ describing an infinitesimal displacement of all points on a line normal to the midsurface in $\phi(\xi^1, \xi^2)$ and $\theta = \theta_1 a^1 + \theta_2 a^2$ representing a rotation vector. We introduce the space of admissible displacements

$$\mathbf{V} := \left\{ (u, \theta) \mid (u_1, u_2) \in H^1(\mathbf{S})^2, u_3 \in H^1(\mathbf{S}), \theta \in H^1(\mathbf{S})^2 \right\} \cap \mathbf{BC} \tag{7}$$

where $H^1(\mathbf{S})$ and $H^1(\mathbf{S})^2$ are Sobolev-spaces for scalar functions and first order tensors on the midsurface, resp. Again, we suppress the parametrization in $u \circ \phi$ and $\theta \circ \phi$ defined on $\omega$ when the meaning is clear. Let us assume for the boundary conditions $\mathbf{BC}$ that the shell body is softclamped over the whole boundary $\partial\mathbf{S}$, i.e. $u|_{\partial\mathbf{S}} = 0$. We next consider the linear 3D-Green-Lagrange-strain tensor which is given by

$$e_{ij} = \frac{1}{2}(g_i \cdot U_{,j} + g_j \cdot U_{,i}), \quad i, j = 1, 2, 3, \tag{8}$$

where $U_{,i}$ means the partial derivative of $U$ w.r.t. $\xi^i$. By Hooke's Law, we get for the components of the stress tensor

$$\sigma^{ij} = \sum_{k,l=1}^{3} H^{ijkl} e_{kl} \tag{9}$$

with $H^{ijkl} = \tilde{L}_1 g^{ij} g^{kl} + \tilde{L}_2(g^{ik}g^{jl} + g^{il}g^{jk})$ and $\tilde{L}_1,\ \tilde{L}_2$ being the Lamé constants. Using the assumption that the normal stress $\sigma^{33}$ is zero this simplifies to

$$\sigma^{\alpha\beta} = \sum_{\lambda,\mu=1}^{2} C^{\alpha\beta\lambda\mu} e_{\lambda\mu}, C^{\alpha\beta\lambda\mu} = \tfrac{E}{2(1+\nu)} \left( g^{\alpha\lambda} g^{\beta\mu} + g^{\alpha\mu} g^{\beta\lambda} + \tfrac{2\nu}{1-\nu} g^{\alpha\beta} g^{\lambda\mu} \right),$$

$$\sigma^{\alpha 3} = \sum_{\lambda=1}^{2} \frac{1}{2} D^{\alpha\lambda} e_{\lambda 3}, \qquad D^{\alpha\lambda} = \tfrac{2E}{1+\nu} g^{\alpha\lambda}, \quad \alpha, \beta = 1, 2, \tag{10}$$

where $E$ is Young's modulus and $\nu$ is Poisson's ratio. Calculating these quantities for our original problem leads to

$$\mathbf{e} := \begin{pmatrix} e_{11} \\ e_{22} \\ \sqrt{2} e_{12} \end{pmatrix} = \begin{pmatrix} u_{1,1} + \xi^3 \theta_{1,1} \\ u_{2,2} + R u_3 + \xi^3 \left( \theta_{2,2} + \frac{1}{R} u_{2,2} + u_3 \right) + \frac{(\xi^3)^2}{R} \theta_{2,2} \\ \frac{1}{\sqrt{2}} (u_{1,2} + u_{2,1}) + \frac{\xi^3}{\sqrt{2}} \left( \theta_{1,2} + \theta_{2,1} + \frac{1}{R} u_{2,1} \right) + \frac{(\xi^3)^2}{\sqrt{2} R} \theta_{2,1} \end{pmatrix}$$

$$\zeta := \begin{pmatrix} e_{13} \\ e_{23} \end{pmatrix} = \tfrac{1}{2} \begin{pmatrix} \theta_1 + u_{3,1} \\ \theta_2 + u_{3,2} - \frac{1}{R} u_2 \end{pmatrix}. \tag{11}$$

This special vector notation is chosen according to [13] and allows us to rewrite the equilibrium conditions in an elegant way for implementation purposes. Introducing the two-dimensional Lamé-constants

$$L_1 = E \frac{\nu}{(1+\nu)(1-\nu)}, \quad L_2 = \frac{E}{2(1+\nu)} \tag{12}$$

we get for the stress tensor and its matrix representation

$$\mathbf{C} := \begin{pmatrix} L_1 + 2L_2 & \frac{1}{(R+\xi^3)^2} L_1 & 0 \\ \frac{1}{(R+\xi^3)^2} L_1 & \frac{1}{(R+\xi^3)^4} (L_1 + 2L_2) & 0 \\ 0 & 0 & \frac{2}{(R+\xi^3)^2} L_2 \end{pmatrix}$$

$$\mathbf{D} := \begin{pmatrix} 4L_2 & 0 \\ 0 & \frac{4}{(R+\xi^3)^2} L_2 \end{pmatrix}. \tag{13}$$

Now consider a force $f \in L^2(\mathbf{S})$ which is applied orthogonal to the mid-surface and formulate the equilibrium conditions for the stationary case according to the basic shell model from Chapelle and Bathe [11]: Find $(u,\theta) \in \mathbf{V}$ with

$$\begin{aligned}&\int_{\Omega_{(t)}} \sum_{\alpha,\beta,\lambda,\mu=1}^{2} C^{\alpha\beta\lambda\mu} e_{\alpha\beta}(u,\theta) e_{\lambda\mu}(v,\psi) + D^{\alpha\lambda} e_{\alpha 3}(u,\theta) e_{\lambda 3}(v,\psi)\,\mathrm{d}V \\ &= \int_{\omega} f v_3 \,\mathrm{d}S\end{aligned} \tag{14}$$

for all $(v,\psi) \in \mathbf{V}$. We define the bilinear form $A_{(t)}(u,\theta;v,\psi)$ for the left hand side and the linear form $F(v,\psi)$ for the right hand side of (14). The bilinear form can be rewritten using matrix-vector-notation and symmetry properties of the strain and stress tensor as

$$A_{(t)}(u,\theta;v,\psi) = \int_{\Omega_{(t)}} \mathbf{e^T C e} + \zeta^T \mathbf{D} \zeta \,\mathrm{d}V. \tag{15}$$

# 4 Analysis of the model equations

We know from [11] that $A_{(t)}$ is coercive and continuous for fixed $t$, as well as $F$ is continuous. According to the Lax-Milgram-Lemma there is a unique solution to (14). Therefore the control-to-state operator $G$ which maps the control $t$ to the corresponding displacement $(u,\theta)$ is well-defined. Let us define the set

$$U_{\mathrm{reg}} := \{t \in C^{0,1}(\mathbf{S}) \mid 0 < t_{\min} \leq t(\xi^1,\xi^2) \leq t_{\max} < 2R \text{ in } \mathbf{S}\} \tag{16}$$

which is a closed subset of $C^{0,1}(\mathbf{S})$. We now want to investigate the continuity of $G : U_{\mathrm{reg}} \to \mathbf{V}$.

**Lemma 1** *For all $t \in U_{\mathrm{reg}}$ the bilinear forms $A_{(t)}$ have a common coercivity constant, i.e.*

$$A_{(t)}(u,\theta;u,\theta) \geq c \, \|(u,\theta)\|_{\mathbf{V}}^2 \quad \forall\,(u,\theta) \in \mathbf{V}. \tag{17}$$

**Proof.** The proof can easily be derived from the original proof of coercivity in [11].

□

We next consider a sequence $t_n \in U_{\mathrm{reg}}$ that converges strongly to $\bar{t} \in U_{\mathrm{reg}}$ w.r.t. to $\|t\|_{C^{0,1}(\mathbf{S})} = \|t\|_\infty + \mathrm{Lip}(t)$ and denote the corresponding sequence of states by $y_n := (u_n,\theta_n) := G(t_n)$. Using the above Lemma we see that

$$c \, \|(u_n,\theta_n)\|_{\mathbf{V}}^2 \leq A_{(t_n)}(u_n,\theta_n;u_n,\theta_n) = F(u_n,\theta_n) \leq \|f\|_{L^2(\mathbf{S})} \, \|(u_n,\theta_n)\|_{\mathbf{V}}, \tag{18}$$

i.e. the sequence $y_n$ is bounded. Hence there is a weakly convergent subsequence, also denoted by $y_n$, with weak limit $\bar{y} \in \mathbf{V}$.

**Lemma 2** *The weak limit $\bar{y}$ is solution of (14) with thickness $\bar{t}$.*

**Proof.** We first find an alternative form for $A_{(t)}$, namely

$$A_{(t)}(u, \theta; v, \psi) = \int_{\Omega_{(t_{\max})}} \sum_{i,j,k,l=1}^{3} H^{ijkl} e_{ij}(u, \theta) e_{kl}(v, \psi) \chi_{(t)} \, \mathrm{d}V \tag{19}$$

with

$$\chi_{(t)}(\xi^1, \xi^2, \xi^3) := \begin{cases} 1, & \text{if } -\frac{t(\xi^1,\xi^2)}{2} < \xi^3 < \frac{t(\xi^1,\xi^2)}{2} \\ 0, & \text{otherwise} \end{cases} . \tag{20}$$

This can be done since the integrand does not explicitly depend on $t$ and $\chi_{(t)}$ is the characteristic function for $\Omega_{(t)}$. Another expression in terms of the $L^2(\mathbf{B}_{(t_{\max})})$ scalar product for second order tensors is

$$\begin{aligned} A_{(t)}(u, \theta; v, \psi) &= \int_{\Omega_{(t_{\max})}} \langle \sigma(v, \psi) \chi_{(t)}, e(u, \theta) \rangle \, \mathrm{d}V \\ &= \langle \sigma(v, \psi) \chi_{(t)}, e(u, \theta) \rangle_{L^2(\mathbf{B}_{(t_{\max})})}, \end{aligned} \tag{21}$$

where $\sigma(v, \psi)\chi_{(t)}$ has components

$$(\sigma(v, \psi) \chi_{(t)})^{ij} = \sigma(v, \psi)^{ij} \chi_{(t)}. \tag{22}$$

We have $\sigma(v, \psi)\chi_{(t_n)} \to \sigma(v, \psi)\chi_{(\bar{t})}$ in $L^2(\mathbf{B}_{(t_{\max})})$ for fixed $(v, \psi) \in \mathbf{V}$, because

$$\begin{aligned} &\left\| \sigma(v, \psi) \chi_{(t_n)} - \sigma(v, \psi) \chi_{(\bar{t})} \right\|^2_{L^2(\mathbf{B}_{t_{\max}})} \\ &\leq \sum_{i,j,k,l=1}^{3} \left| \int_{\Omega_{(t_{\max})}} g_{ik} g_{jl} \sigma^{ij}(v, \psi) \sigma^{kl}(v, \psi) (\chi_{(t_n)} - \chi_{(\bar{t})}) \, \mathrm{d}V \right| . \end{aligned} \tag{23}$$

It holds $\chi_{(t_n)} - \chi_{(\bar{t})} \to 0$ pointwise a.e., so we can conclude

$$g_{ik} g_{jl} \sigma^{ij}(v, \psi) \sigma^{kl}(v, \psi) (\chi_{(t_n)} - \chi_{(\bar{t})}) \to 0 \quad \text{pointwise a.e.} \tag{24}$$

Furthermore $|g_{ik} g_{jl} \sigma^{ij}(v, \psi) \sigma^{kl}(v, \psi) \chi_{(t_{\max})}|$ is an integrable majorant and we get the convergence of the right hand side from (23) to 0. From $(u_n, \theta_n) \rightharpoonup (\bar{u}, \bar{\theta})$ in $\mathbf{V}$ it follows that all components and covariant derivatives converge weakly to the corresponding limit in $L^2(\mathbf{S})$, and so $e(u_n, \theta_n) \rightharpoonup e(\bar{u}, \bar{\theta})$ in $L^2(\mathbf{B}_{(t_{\max})})$. We get

$$\langle \sigma(v,\psi)\chi_{(t_n)}, e(u_n,\theta_n)\rangle_{L^2(\mathbf{B}_{(t_{\max})})} \to \langle \sigma(v,\psi)\chi_{(\bar{t})}, e(\bar{u},\bar{\theta})\rangle_{L^2(\mathbf{B}_{(t_{\max})})}, \tag{25}$$

and therefore $F(v,\psi) = \lim_{n\to\infty} A_{(t_n)}(u_n,\theta_n;v,\psi) = A_{(\bar{t})}(\bar{u},\bar{\theta};v,\psi)$.

□

From the uniqueness of the limit $(\bar{u}, \bar{\theta})$ we conclude that the whole sequence converges weakly.

**Theorem 1** *The convergence of* $(u_n, \theta_n)$ *to* $(\bar{u}, \bar{\theta})$ *is also strong. Hence the operator* $G : U_{\text{reg}} \to \mathbf{V}$ *is continuous.*

**Proof.** It holds for $(v, \psi) \in \mathbf{V}$

$$0 = \lim_{n\to\infty} \left( A_{(t_n)}(u_n - \bar{u}; \theta_n - \bar{\theta}; v, \psi) + A_{(t_n)}(\bar{u}, \bar{\theta}; v, \psi) - A_{(\bar{t})}(\bar{u}, \bar{\theta}; v, \psi) \right). \tag{26}$$

We now take $(v, \psi) := (u_n - \bar{u}, \theta_n - \bar{\theta})$ and get for the last two terms of (26)

$$\lim_{n\to\infty} \left( A_{(t_n)}(\bar{u}, \bar{\theta}; u_n - \bar{u}, \theta_n - \bar{\theta}) - A_{(\bar{t})}(\bar{u}, \bar{\theta}; u_n - \bar{u}, \theta_n - \bar{\theta}) \right) = 0, \tag{27}$$

because

$$\begin{aligned} &\left| A_{(t_n)}(\bar{u}, \bar{\theta}; u_n - \bar{u}, \theta_n - \bar{\theta}) - A_{(\bar{t})}(\bar{u}, \bar{\theta}; u_n - \bar{u}, \theta_n - \bar{\theta}) \right| \\ &\leq \left| \langle \sigma(\bar{u}, \bar{\theta})(\chi_{(t_n)} - \chi_{(\bar{t})}), e(u_n - \bar{u}, \theta_n - \bar{\theta})\rangle_{L^2(\mathbf{B}_{(t_{\max})})} \right|. \end{aligned} \tag{28}$$

Analog to the proof of the above Lemma we can show

$$\begin{array}{rcll} \sigma(\bar{u}, \bar{\theta})(\chi_{(t_n)} - \chi_{(\bar{t})}) & \to & 0 & \text{in } L^2(\mathbf{B}_{(t_{\max})}) \\ e(u_n - \bar{u}, \theta_n - \bar{\theta}) & \rightharpoonup & 0 & \text{in } L^2(\mathbf{B}_{(t_{\max})}). \end{array} \tag{29}$$

Both statements yield the convergence of the last term from (28) to 0.

From (26) it follows

$$0 = \lim_{n\to\infty} A_{(t_n)}(u_n - \bar{u}, \theta_n - \bar{\theta}; u_n - \bar{u}, \theta_n - \bar{\theta}) \geq \lim_{n\to\infty} c \left\| (u_n - \bar{u}, \theta_n - \bar{\theta}) \right\|_{\mathbf{V}}^2 \geq 0 \tag{30}$$

and therefore the strong convergence $(u_n, \theta_n) \to (\bar{u}, \bar{\theta})$ in $\mathbf{V}$. □

**Theorem 2** *The control-to-state-operator* $G : U_{\text{reg}} \to \mathbf{V}$ *is Gâteaux-differentiable. For a fixed point* $t \in U_{\text{reg}}$ *with state* $(u_{(t)}, \theta_{(t)})$ *and a direction* $q \in C^{0,1}(\mathbf{S})$ *with* $t + \lambda q \in U_{\text{reg}}$ *for all sufficiently small* $\lambda \geq 0$ *it holds* $G'(t)q =$

*$(r,\rho)$, where $(r,\rho)$ is the (unique) solution to the variational problem: Find $(r,\rho)$ in $\mathbf{V}$ such that*

$$A_{(t)}(r,\rho;v,\psi) = Z_q(v,\psi) \tag{31}$$

*holds for all $(v,\psi)\in\mathbf{V}$ where the linear form $Z_q$ is given by*

$$Z_q(v,\psi) = -\int_\omega \sum_{\xi_i^3\in\{\pm\frac{t}{2}\}} \left( \left[ \langle \sigma(v,\psi), e(u_{(t)},\theta_{(t)})\rangle \left(1+\frac{\xi^3}{R}\right)\right]_{\xi^3=\xi_i^3} \right) \frac{q}{2}\,\mathrm{d}S. \tag{32}$$

**Proof.** Consider a direction $q\in C^{0,1}(\mathbf{S})$, $0\le\lambda\in\mathbf{R}$ as well as $t\in U_{\text{reg}}$ like in the theorem statement. For the solutions $(u_{(t+\lambda q)},\theta_{(t+\lambda q)})$ and $u_{(t)},\theta_{(t)}$ of (14) to the thicknesses $t$ and $t+\lambda q$, resp. it holds

$$\begin{array}{rcl} A_{(t)}(u_{(t)},\theta_{(t)};v,\psi) &=& F(v,\psi) \\ A_{(t+\lambda q)}(u_{(t+\lambda q)},\theta_{(t+\lambda q)};v,\psi) &=& F(v,\psi) \end{array} \tag{33}$$

for all $(v,\psi)$ in $\mathbf{V}$. It follows

$$\begin{aligned} 0 &= \frac{1}{\lambda}\left[A_{(t+\lambda q)}(u_{(t+\lambda q)},\theta_{(t+\lambda q)};v,\psi) - A_{(t)}(u_{(t)},\theta_{(t)};v,\psi)\right] \\ &= \int_{\Omega_{(t_{\max})}} \langle\sigma(v,\psi), \frac{e(u_{(t+\lambda q)},\theta_{(t+\lambda q)}) - e(u_{(t)},\theta_{(t)})}{\lambda}\rangle \chi_{(t)}\,\mathrm{d}V \\ &\quad + \int_{\Omega_{(t_{\max})}} \langle\sigma(v,\psi), e(u_{(t+\lambda q)},\theta_{(t+\lambda q)})\rangle \frac{\chi_{(t+\lambda q)}-\chi_{(t)}}{\lambda}\,\mathrm{d}V. \end{aligned} \tag{34}$$

For the last summand from equation (34) the mapping $Z_\lambda:\mathbf{V}\times\mathbf{V}\to\mathbf{R}$,

$$Z_\lambda(u,\theta;v,\psi) := \int_{\Omega_{(t_{\max})}} \langle\sigma(v,\psi), e(u,\theta)\rangle \frac{\chi_{(t+\lambda q)}-\chi_{(t)}}{\lambda}\,\mathrm{d}V \tag{35}$$

is defined.

**Lemma 3** *The limit*

$$-\lim_{\lambda\to 0} Z_\lambda(u_{(t+\lambda q)},\theta_{(t+\lambda q)};v,\psi) =: Z_q(v,\psi) \tag{36}$$

*exists and is in $\mathbf{V}^*$.*

**Proof.** It holds that $Z_\lambda$ can be estimated for fixed $\lambda$, because

$$|Z_\lambda(u,\theta;v,\psi)| \le \frac{1}{\lambda}\int_{\Omega_{(t_{\max})}} |\langle\sigma(v,\psi),e(u,\theta)\rangle|\,\mathrm{d}V \le \frac{1}{\lambda}C\,\|(v,\psi)\|_{\mathbf{V}}\,\|(u,\theta)\|_{\mathbf{V}}. \tag{37}$$

The last inequality comes from the boundedness of $A_{(t_{\max})}$.

We now want to determine the pointwise limit

$$\lim_{\lambda\to 0} Z_\lambda(u,\theta;v,\psi). \tag{38}$$

At first we consider the innermost integral by defining $z : \mathbf{S} \to \mathbf{R}$,

$$z := \int_{-\frac{t_{\max}}{2}}^{\frac{t_{\max}}{2}} \langle \sigma(v,\psi), e(u,\theta)\rangle \frac{\chi_{(t+\lambda q)} - \chi_{(t)}}{\lambda} \sqrt{g}\, \mathrm{d}\xi^3. \tag{39}$$

It holds a.e. in $\mathbf{S}$

$$|z(\xi^1,\xi^2)| \le b(\xi^1,\xi^2) \int_{-\frac{t_{\max}}{2}}^{\frac{t_{\max}}{2}} \left| \frac{\chi_{(t+\lambda q)} - \chi_{(t)}}{\lambda} \right| \mathrm{d}\xi^1 = b(\xi^1,\xi^2)|q(\xi^1,\xi^2)| \tag{40}$$

where $b \in L^1(\mathbf{S})$ because of the boundedness of $H^{ijkl}$ and the polynomial dependence of $e_{ij}$ and $\sqrt{g}$ in $\xi^3$. So there is an integrable majorant and we can write

$$\begin{aligned} \lim_{\lambda\to 0} Z_\lambda(u,\theta;v,\psi) &= \lim_{\lambda\to 0} \int_\omega \int_{-\frac{t_{\max}}{2}}^{\frac{t_{\max}}{2}} \langle \sigma(v,\psi), e(u,\theta)\rangle \frac{\chi_{(t+\lambda q)} - \chi_{(t)}}{\lambda}\, \mathrm{d}V \\ &= \int_\omega \lim_{\lambda\to 0} \int_{-\frac{t_{\max}}{2}}^{\frac{t_{\max}}{2}} \langle \sigma(v,\psi), e(u,\theta)\rangle \frac{\chi_{(t+\lambda q)} - \chi_{(t)}}{\lambda} (1 + \frac{\xi^3}{R})\, \mathrm{d}\xi^3\, \mathrm{d}S. \end{aligned} \tag{41}$$

Now we investigate the limit

$$\begin{aligned} &\lim_{\lambda\to 0} \int_{-\frac{t_{\max}}{2}}^{\frac{t_{\max}}{2}} f(\xi^3) \frac{\chi_{(t+\lambda q)} - \chi_{(t)}}{\lambda}\, \mathrm{d}\xi^3 \\ &= \lim_{\lambda\to\ 0} \frac{1}{\lambda} \left[ \int_{-\frac{(t+\lambda q)}{2}}^{\frac{(t+\lambda q)}{2}} f(\xi^3)\, \mathrm{d}\xi^3 - \int_{-\frac{t}{2}}^{\frac{t}{2}} f(\xi^3)\, \mathrm{d}\xi^3 \right]. \end{aligned} \tag{42}$$

for a continuous function $f \in C([-\frac{t_{\max}}{2}, \frac{t_{\max}}{2}])$. This simplifies to the derivative of the integral bounds with respect to $\lambda$ at $\lambda = 0$ which evaluates according to Leibniz's formula to

$$\left( f\left(\frac{t}{2}\right) + f\left(-\frac{t}{2}\right)\right) \frac{q}{2}. \tag{43}$$

Therefore, we get

$$\lim_{\lambda\to 0} Z_\lambda(u,\theta;v,\psi) = \int_\omega \sum_{\xi_i^3 \in \{\pm\frac{t}{2}\}} \left( \left[ \langle \sigma(v,\psi), e(u,\theta)\rangle (1 + \frac{\xi^3}{R}) \right]_{\xi^3 = \xi_i^3} \right) \frac{q}{2}\, \mathrm{d}S. \tag{44}$$

Because of the convergence for fixed $(u,\theta)$ and $(v,\psi)$, the mappings $Z_\lambda$ are uniformly bounded by the Banach-Steinhaus theorem, i.e.

$$Z_\lambda(u,\theta;v,\psi) \le C \left\|(u,\theta)\right\|_{\mathbf{V}} \left\|(v,\psi)\right\|_{\mathbf{V}}. \tag{45}$$

Now we want to determine for fixed thickness $t$ and direction $q$

$$\lim_{\lambda\to 0} Z_\lambda(u_{(t+\lambda q)},\theta_{(t+\lambda q)};v,\psi). \tag{46}$$

It holds

$$\begin{aligned}
&|Z_\lambda(u_{(t+\lambda q)},\theta_{(t+\lambda q)};v,\psi) - Z(u_{(t)},\theta_{(t)};v,\psi)| \\
&\quad \le \quad |Z_\lambda(u_{(t+\lambda q)}-u_{(t)},\theta_{(t+\lambda q)}-\theta_{(t)};v,\psi)| + |(Z_\lambda - Z)(u_{(t)},\theta_{(t)};v,\psi)| \\
&\quad \le \quad C\left\|(u_{(t+\lambda q)}-u_{(t)},\theta_{(t+\lambda q)}-\theta_{(t)})\right\|_{\mathbf{V}} \left\|(v,\psi)\right\|_{\mathbf{V}} \\
&\qquad\qquad\qquad\qquad\qquad\qquad + |(Z_\lambda - Z)(u_{(t)},\theta_{(t)};v,\psi)| \\
&\quad \to \quad 0, \quad \lambda\to 0
\end{aligned} \tag{47}$$

Since $(u_{(t)},\theta_{(t)})$ is fixed, we write $Z(v,\psi)$ instead of $Z(u_{(t)},\theta_{(t)};v,\psi)$ and consider from now on $Z$ as a mapping $\mathbf{V}\to\mathbf{R}$. Hence we get

$$\lim_{\lambda\to 0} Z_\lambda(u_{(t+\lambda q)},\theta_{(t+\lambda q)};v,\psi) = Z(v,\psi) \tag{48}$$

with $Z\in\mathbf{V}^*$. To indicate the dependence from the initially chosen $q$, we finally define

$$Z_q(v,\psi) := -Z(v,\psi) \tag{49}$$

and we get the linear form $Z_q(v,\psi)\in\mathbf{V}^*$.

□

Back to the proof of theorem 2 we again consider equation (34). Because of the linearity of $e_{ij}$ it follows

$$\begin{aligned}
&\int_{\Omega_{(t_{\max})}} \langle\sigma(v,\psi), \frac{e(u_{(t+\lambda q)},\theta_{(t+\lambda q)}) - e(u_{(t)},\theta_{(t)})}{\lambda}\rangle \chi_{(t)}\,\mathrm{d}V \\
&= A_{(t)}\left(\frac{(u_{(t+\lambda q)},\theta_{(t+\lambda q)}) - (u_{(t)},\theta_{(t)})}{\lambda};v,\psi\right).
\end{aligned} \tag{50}$$

By taking the limit in (34) and using the continuity of $A_{(t)}$ we get

$$A_{(t)}\left(\lim_{\lambda\to 0}\frac{(u_{(t+\lambda q)},\theta_{(t+\lambda q)}) - (u_{(t)},\theta_{(t)})}{\lambda};v,\psi\right) = Z_q(v,\psi) \tag{51}$$

for all $(v,\psi) \in \mathbf{V}$. Now we insert $(v,\psi) = \frac{(u_{(t+\lambda q)},\theta_{(t+\lambda q)})-(u_{(t)},\theta_{(t)})}{\lambda}$ in this equation. This yields

$$\begin{aligned}
&c\left\|\frac{(u_{(t+\lambda q)},\theta_{(t+\lambda q)})-(u_{(t)},\theta_{(t)})}{\lambda}\right\|_{\mathbf{V}}^2 \\
&\le A_{(t)}\left(\tfrac{(u_{(t+\lambda q)},\theta_{(t+\lambda q)})-(u_{(t)},\theta_{(t)})}{\lambda};\tfrac{(u_{(t+\lambda q)},\theta_{(t+\lambda q)})-(u_{(t)},\theta_{(t)})}{\lambda}\right) \\
&= Z_q\left(\tfrac{(u_{(t+\lambda q)},\theta_{(t+\lambda q)}))-(u_{(t)},\theta_{(t)})}{\lambda}\right) \\
&\le C\left\|\frac{(u_{(t+\lambda q)},\theta_{(t+\lambda q)})-(u_{(t)},\theta_{(t)})}{\lambda}\right\|_{\mathbf{V}}
\end{aligned} \tag{52}$$

and we see that the sequence $\frac{(u_{(t+\lambda q)},\theta_{(t+\lambda q)})-(u_{(t)},\theta_{(t)})}{\lambda}$ is bounded. Hence, there is a subsequence $\lambda_n \to 0$ such that

$$\frac{(u_{(t+\lambda_n q)},\theta_{(t+\lambda_n q)})-(u_{(t)},\theta_{(t)})}{\lambda_n} \to (r,\rho), \quad \lambda_n \to 0 \tag{53}$$

weakly in $\mathbf{V}$. Passage to the limit in (51) leads to the variational equation: Find $(r,\rho) \in \mathbf{V}$, such that

$$A_{(t)}(r,\rho;v,\psi) = Z_q(v,\psi) \tag{54}$$

for all $(v,\psi) \in \mathbf{V}$. From the Lax-Milgram-Lemma we know that this equation has a unique solution, so the whole sequence converges to $(r,\rho)$ for $\lambda \to 0$.

□

# 5 Optimization problem

In this section, the actual optimization problem shall be discussed. In our case the minimization uses the compliance functional, where the deformation is weighted with the incoming force. We state the optimization problem as follows:

$$\begin{aligned}
&\min_{t\in C^{0,1}(\mathbf{S}),(u,\theta)\in\mathbf{V}} J(u,\theta;t) := F(u,\theta) + \frac{\lambda}{2}\|t\|_{H^1(\mathbf{S})}^2 \\
&\text{s.t.}: \; A_{(t)}(u,\theta;v,\psi) = F(v,\psi) \quad \forall\,(v,\psi)\in\mathbf{V} \\
&\qquad t_{\min} \le t(\xi^1,\xi^2) \le t_{\max} \qquad \text{in } \mathbf{S} \\
&\qquad \int_\omega t\,\mathrm{d}S = C
\end{aligned} \tag{55}$$

The constant $C$ represents the volume of the shell and $t_{\min}$, $t_{\max}$ are the lower and upper bound for the thickness, resp. The introduction of a regularization term was essential for the quality of the numeric solutions. Additionally, we introduce the set

$$U_{\mathrm{ad}} := \{t \in C^{0,1}(\mathbf{S}) \mid t_{\min} \le t(\xi^1, \xi^2) \le t_{\max} \text{ in } \mathbf{S}, \int_\omega t \,\mathrm{d}S = C\} \subset U_{\mathrm{reg}}. \tag{56}$$

Note that the set $U_{\mathrm{ad}}$ is convex, closed and bounded. By using the control-to-state operator $G$ we can define the reduced objective functional $J_s(t) := J(G(t); t)$ which we will use in this section. The problem (55) can be rewritten in the form

$$\min_{t \in U_{\mathrm{ad}}} J_s(t) = F(G(t)) + \frac{\lambda}{2} \|t\|^2_{H^1(\mathbf{S})} \tag{57}$$

**Theorem 3** *Let the set of admissible thicknesses be restricted to*

$$U^M_{\mathrm{ad}} := \{t \in U_{\mathrm{ad}} \mid \mathrm{Lip}(t) < M\} \tag{58}$$

*for fixed $M > 0$. Then the problem (57) has at least one solution.*

**Proof.** We know from the Arzela-Ascoli theorem that $U^M_{\mathrm{ad}}$ is a closed sequential compact subset of $C(\bar{\omega})$. Moreover, the objective $J_s$ is a composition of continuous mappings. Therefore a minimum exists by the Weierstrass theorem.

□

The aim in this section is to derive necessary conditions for an (locally) optimal solution. As a further result we will also get an expression for the directional derivatives of the objective which is very useful for later numerical calculations. We first define the adjoint state as the solution to:

Find $(p, \eta) \in \mathbf{V}$, such that

$$A_{(t)}(p, \eta; v, \psi) = \nabla_1 J(u, \theta; t)(v, \psi) \tag{59}$$

holds for all $(v, \psi) \in \mathbf{V}$, where $\nabla_1 J(u_0, \theta_0; t_0)$ denotes the Fréchet-derivative of $J$ with respect to $(u, \theta)$ at the point $(u_0, \theta_0; t_0)$. We note that because of $\nabla_1 J(u_0, \theta_0; t_0)(v, \psi) = F(v, \psi)$ in our case the adjoint state is equal to the corresponding original state $(u_{(t)}, \theta_{(t)})$.

**Theorem 4** *The directional derivative of the reduced objective $J_s$ at point $t$ in direction $q$ is given by*

$$J_s'(t)q = Z_q(u_{(t)}, \theta_{(t)}) + \lambda \langle t, q\rangle_{H^1(\mathbf{S})} = Z_q(G(t)) + \lambda \langle t, q\rangle_{H^1(\mathbf{S})}. \tag{60}$$

**Proof.** The proof is straightforward by using the chain rule and symmetry of $A_{(t)}$ and can be found e.g. in [5]. The resulting expression then simplifies because of $\nabla_2 J(u_0, \theta_0; t_0)q = \lambda\langle t_0, q\rangle_{H^1(\mathbf{S})}$ and the equality of adjoint and original state.

□

With the help of the directional derivative (60) of the reduced objective we can state necessary conditions for a (locally) optimal solution:

**Corollary 1** *Let $t^* \in U_{\text{ad}}$ be a (locally) optimal solution for the problem (55) with corresponding state $(u_{(t^*)}, \theta_{(t^*)})$. Then it holds*

$$J_s'(t^*)(q - t^*) = Z_{(q-t^*)}(u_{(t^*)}, \theta_{(t^*)}) + \lambda\langle t^*, q - t^*\rangle_{H^1(\mathbf{S})} \geq 0 \qquad (61)$$

*for all directions $q \in U_{\text{ad}}$.*

Note, that we need the convexity of $U_{\text{ad}}$ for this statement.

## 6 Numerical implementation

For the numerical solution of the optimization problem a Fortran program was written. The state equation is solved using standard FEM-methods. We use an approach with general shell elements based on biquadratic 9-node Lagrange elements that can be found in [11] or [14]. The finite element displacements are thus given by

$$V_h = \sum_{i=1}^{n} h_i(\xi^1, \xi^2)(v^{(i)} + \xi^3\eta^{(i)}), \quad \eta^{(i)} \cdot a_3^{(i)} = 0 \qquad (62)$$

where $v^{(i)} = (v_1^{(i)}, v_2^{(i)}, v_3^{(i)})^T$, $\eta^{(i)} = (\eta_1^{(i)}, \eta_2^{(i)}, \eta_3^{(i)})^T$, $t^{(i)}$ and $a_3^{(i)}$ denote the translational and rotational displacement components in Cartesian coordinates, the thickness and the unit normal vector at node $i$, respectively. The $h_i$ are chosen as the shape functions arising from biquadratic ansatz functions $\lambda_j$, $j = 1, \ldots, 9$ on the reference element. In local element coordinates, this reads

$$V_h|_E = \sum_{j=1}^{9} \lambda_j(r, s)(v^{(j)} + z\frac{t^{(i)}}{2}\eta^{(i)}). \qquad (63)$$

As in (7) we divide into translational and rotational components and consider the finite element displacement space

$$\begin{aligned}\mathbf{V}_h := \{(v_h, \eta_h) \mid v_h = \sum_{i=1}^{n} h_i(\xi^1, \xi^2)v^{(i)}, \eta_h = \pi(\sum_{i=1}^{n} h_i(\xi^1, \xi^2)\eta^{(i)}),\\ \eta^{(i)} \cdot a_3^{(i)} = 0\} \cap \mathbf{BC},\end{aligned} \qquad (64)$$

where $\pi$ denotes the projection operator onto the tangential plane at point $(\xi^1, \xi^2) \in \mathbf{S}$ and $\mathbf{BC}$ imposes appropriate boundary conditions on the displacements. Accordingly, the nodal thicknesses $t^{(i)}$ are interpolated using the shape functions $h_i$ which leads to $t_h = \sum_{i=1}^{n} h_i t^{(i)}$. The strain-vector $\mathbf{e}$ is calculated using strain-displacement matrices arising from (11). The matrices $\mathbf{C}$ and $\mathbf{D}$ are calculated using (13) and the value of the discretized bilinear form $A_{(t_h)}(u_h, \theta_h; v_h, \eta_h)$ is obtained using (15) together with Gaussian quadrature. Accordingly, the discretized linear form $F_h(v_h, \eta_h)$ is calculated. We now have to solve: Find $(u_h, \theta_h) \in \mathbf{V}_h$, such that

$$A_{(t_h)}(u_h, \theta_h; v_h, \psi_h) = F_h(v_h, \psi_h) \quad \text{for all } (v_h, \psi_h) \in \mathbf{V}_h. \tag{65}$$

Let the mapping $G_h : \mathbf{R}^n \to \mathbf{V}_h$, $\vec{t} \mapsto (u_h, \theta_h)$ which maps the vector of the nodal thicknesses $\vec{t}$ via the function $t_h$ to the solution of (65) be defined. Finally, the volume constraint is discretized using Gaussian quadrature for $t_h$ over $\mathbf{S}$. We can now state the finite dimensional optimization problem

$$\begin{aligned} \min_{\vec{t} \in \mathbf{R}^n} \; & J_h(\vec{t}) = F_h(G_h(\vec{t})) \\ s.t.: \quad & B_h \vec{t} = C \\ & \vec{t}_{\min} \le \vec{t} \le \vec{t}_{\max} \end{aligned} \tag{66}$$

where the first constraint represents the volume condition and the second one the pointwise bounds on the thickness.

The linear system arising from (65) is solved using a combination of direct methods (Pardiso solver, see e.g. [15]) and the pcg-method. Namely, the system is solved directly once at the beginning of an iteration of the optimizer, while in the following line search steps the pcg-method is used. As a preconditioner, the LU-factorization obtained from the direct solution does very good work. This combination allows us to benefit from the advantages of both direct and indirect methods.

The actual optimization is done with IpOpt, an Interior-Point algorithm for Large-Scale nonlinear optimization. Here, the expression obtained in (60) for the directional derivative of the objective is used to calculate the gradient of the discrete objective which depends only on the nodal thicknesses. This reduces the running time of the optimizer and raises the accuracy of the solution considerably. The discrete gradient is calculated by evaluating the expression

$$[\nabla J_h(\vec{t})]_i = Z_{h_i}(G_h(\vec{t})) + \lambda \langle t_h, h_i \rangle_{H^1(\mathbf{S})} \tag{67}$$

where we go through all shape functions $h_i$.

Moreover, we can perform an "optimality test" using equation (61). For a computed solution $\vec{t}^*$ this is done by checking

$$\min_{t\in U_{\mathrm{ad}}} J_s'(t^*)t \stackrel{?}{=} J_s'(t^*)t^*, \tag{68}$$

where the optimization problem can be brought into the discrete form

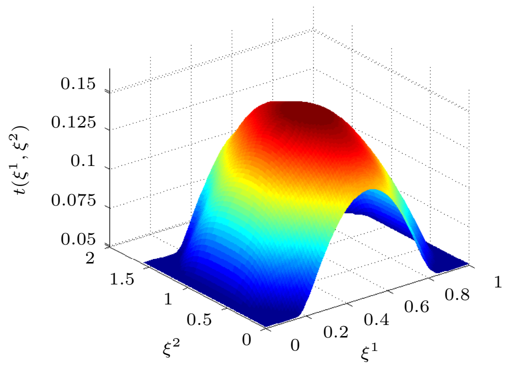
$$\begin{aligned} &\min_{q\in\mathbf{R}^n} \; \nabla J_h(\vec{t}^*)\cdot q \\ &\text{s.t.}: \;\; B_h\vec{q} = C \\ &\qquad\;\; \vec{t}_{\min} \le \vec{q} \le \vec{t}_{\max}. \end{aligned} \tag{69}$$

This is a standard linear program with solution $\vec{q}^*$. In general, our numerical solution $\vec{t}^*$ will not be optimal, so there will be a difference $\nabla J_h(\vec{t}^*)\cdot(\vec{q}^* - \vec{t}^*) = -\varepsilon < 0$, where $\varepsilon$ is an indicator for the precision of $\vec{t}^*$.

# 7 Examples

We discuss a part of a tube where different forces are applied on the midsurface. For this, we consider the domain $\omega = (0,1)\times(0,\frac{\pi}{2})$ together with $\phi(\xi^1,\xi^2) = (\xi^1, \cos(\xi^2), \sin(\xi^2))$ to describe the midsurface. We choose $E = 210$ and $\nu = 0.3$ for the material parameters. The minimal and maximal thickness as well as the volume of the shell are chosen as $t_{\min} = 0.05$ and $t_{\max} = 0.15$ and $C = \frac{\pi}{20}$, respectively. This allows us to start with a constant thickness of 0.1 as a feasible initial solution. The regularization parameter $\lambda$ is chosen in a way that the regularization term is about two to three orders of magnitude smaller than the objective value.

**Example 1** We choose a rotational symmetric force $f(\xi^1,\xi^2) = \xi^1(1-\xi^1)$ which is also symmetric in $\xi^1$. The corresponding optimal thickness profile over the domain $\omega$ is shown in figure 2. We see that the thickness follows approximately the profile of the force and is in particular symmetric in both $\xi^1$ and $\xi^2$. The calculation started on a coarse grid and was refined iteratively on finer grids. In the table from figure 2 the differences between the solution on a very fine grid with step-size of $2^{-8}$ in $\xi^1$-direction ("exact" solution) and on coarser grids are listed, taken in the max-norm on the particular grid. The step size in $\xi^1$-direction for each grid is given in the first column. The step-size in $\xi^2$-direction is chosen to have the same number of nodes. This shows good convergence properties for the thickness when the grid is refined. Moreover, we can see by solution of (69), that the parameter $\varepsilon$ in the optimality test can be chosen as $4.5\cdot 10^{-7}$ which is an indicator for good accuracy of the computed solution $t_h^*$.



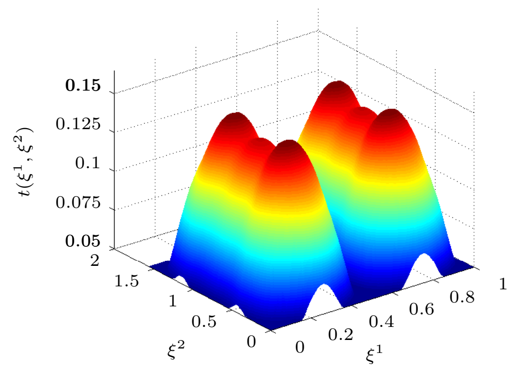

| step − size $h_{\xi^1}$ | $\left\|\vec{t}_{\text{ex}} - \vec{t}_h\right\|_\infty$ |
|---|---|
| $2^{-4}$ | $1.16 \cdot 10^{-2}$ |
| $2^{-5}$ | $1.23 \cdot 10^{-3}$ |
| $2^{-6}$ | $9.92 \cdot 10^{-4}$ |
| $2^{-7}$ | $6.87 \cdot 10^{-4}$ |

Figure 2: Results for example 1, loading $f(\xi^1, \xi^2) = \xi^1(1 - \xi^1)$

**Example 2** We choose a periodic force $f(\xi^1, \xi^2) = \sin(2\pi\xi^1)$. The corresponding optimal thickness profile over the domain $\omega$ is shown in figure 3. Again, the thickness follows approximately the magnitude of the incoming force. It is noticeable for the first two examples that the optimal thickness does depend on $\xi^2$ while the incoming force does not. This could be due to the fact that we consider a part of a tube rather than the full tube. The table from figure 3 again shows good convergence properties for smaller step sizes on the grid, though there is a slightly bigger error than in the first example. The parameter $\varepsilon$ from (69) can be chosen as $2.37 \cdot 10^{-6}$ which indicates good accuracy of the computed solution.

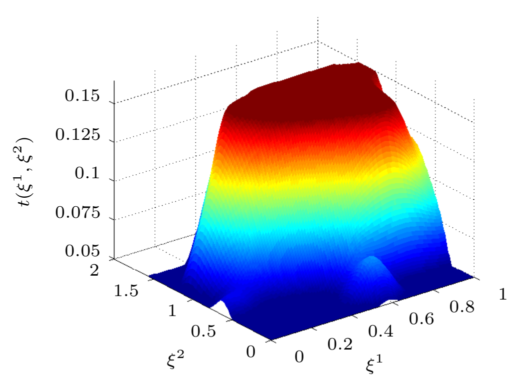
**Example 3** In this example we choose an exponential load $f(\xi^1, \xi^2) = (\exp(\xi^1) - 1)(\exp(\xi^2) - 1)$ which is also asymmetric in $\xi^1$ and $\xi^2$. The corresponding optimal thickness profile over the domain $\omega$ is shown in figure 4. We see that the thickness is maximal in the region with largest incoming force. But also the two elevations with peaks at $(0.578, 0)$ and $(0, 0.626)$ should be mentioned. The values in the table from figure 4 show the convergence properties on finer grids and the parameter $\varepsilon$ from (69) can be chosen as $6.34 \cdot 10^{-6}$ which both indicates good accuracy of the solution.

**Example 4** We choose a discontinuous force $f(\xi^1, \xi^2) = 1_{[\frac{1}{4}, \frac{3}{4}] \times [\frac{\pi}{4} - \frac{1}{4}, \frac{\pi}{4} + \frac{1}{4}]}$. The corresponding optimal thickness profile over the domain $\omega$ is shown in figure 5. It is noticeable that the optimal thickness is maximal in a region shaped like a cross while the incoming force is applied at a region shaped like




| step − size $h_{\xi^1}$ | $\left\|\vec{t}_{\text{ex}} - \vec{t}_h\right\|_\infty$ |
|---|---|
| $2^{-4}$ | $4.22 \cdot 10^{-2}$ |
| $2^{-5}$ | $2.40 \cdot 10^{-2}$ |
| $2^{-6}$ | $7.68 \cdot 10^{-3}$ |
| $2^{-7}$ | $1.55 \cdot 10^{-3}$ |

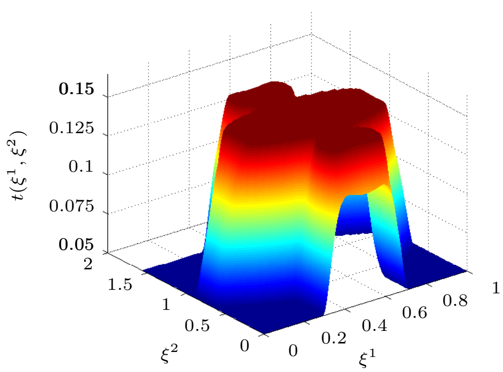
Figure 3: Results for example 2, loading $f(\xi^1, \xi^2) = \sin(2\pi\xi^1)$




| step − size $h_{\xi^1}$ | $\left\|\vec{t}_{\text{ex}} - \vec{t}_h\right\|_\infty$ |
|---|---|
| $2^{-4}$ | $2.48 \cdot 10^{-2}$ |
| $2^{-5}$ | $5.04 \cdot 10^{-3}$ |
| $2^{-6}$ | $1.04 \cdot 10^{-3}$ |
| $2^{-7}$ | $2.12 \cdot 10^{-4}$ |

Figure 4: Results for example 3, loading $f(\xi^1, \xi^2) = (\exp(\xi^1)-1)(\exp(\xi^2)-1)$

a square, see figure 6. The discontinuity of the incoming force is reflected quite good in the optimal thickness. The table from figure 5 shows similar convergence properties as in the second example. The parameter $\varepsilon$ from (69) can be chosen as $3.15 \cdot 10^{-6}$ which indicates good accuracy of the computed solution.




| $\mathrm{step} - \mathrm{size}\ h_{\xi^1}$ | $\left\|\vec{t}_{\mathrm{ex}} - \vec{t}_h\right\|_\infty$ |
|---|---|
| $2^{-4}$ | $4.45 \cdot 10^{-2}$ |
| $2^{-5}$ | $1.96 \cdot 10^{-2}$ |
| $2^{-6}$ | $7.04 \cdot 10^{-3}$ |
| $2^{-7}$ | $1.88 \cdot 10^{-3}$ |

Figure 5: Results for example 4, loading $f(\xi^1, \xi^2) = 1_{[\frac{1}{4}, \frac{3}{4}] \times [\frac{\pi}{4} - \frac{1}{4}, \frac{\pi}{4} + \frac{1}{4}]}$

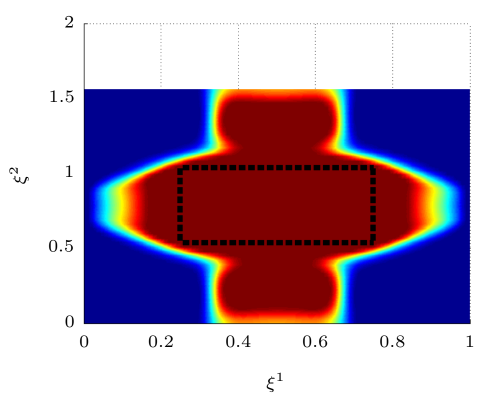



Figure 6: View from above, black line = discontinuity of $f$

## 8 Concluding remarks

In this paper we discussed thickness optimization problems for cylindrical shells where the load is applied to the shell's midsurface. In particular we showed the continuity and Gâteaux-differentiability of the control-to-state operator arising from the model equations. The result was used to deduce an expression for the directional derivative of the objective which was the compliance functional in our case. This allowed us to state necessary conditions for an optimal solution. An effective numerical implementation based on direct methods was possible on quite fine grids by using the discretized expression for the directional derivative together with finite element methods. Different examples were investigated where the optimal thickness followed the incoming force in a reasonable way. The computed thicknesses on refined grids showed a good convergence behaviour as well as the evaluation of the necessary conditions indicated good accuracy of the solutions.

## References

[1] J. Sokolowski, J.-P. Zolesio. *Introduction to Shape Optimization - Shape Sensitivity Analysis.* Springer Series in Computational Mathematics 16, 1992.

[2] O. Pironneau. *Optimal Shape Design for Elliptic Systems (Scientific Computation).* Springer Verlag, 1983.

[3] J. Haslinger, R. A. E. Mäkinen. *Introduction to Shape Optimization: Theory, Approximation and Computation.* SIAM Philadelphia, 2003.

[4] M. C. Delfour, J.-P. Zolesio. *Shapes and geometries. Analysis, differential calculus and optimization.* SIAM Philadelphia, 2001.

[5] P. Neittanmaki, D. Tiba, J. Sprekels. *Optimization of Elliptic Systems.* Springer, 2006.

[6] P. Nestler. Optimal thickness of a cylindrical shell. *Ann. Acad. Rom. Sci., Ser. Math. Appl.* 4:183–208, 2012.

[7] Ü. Lepik, T. Lepikult. Automated calculation and optimal design of rigid-plastic beams under dynamic loading. *Int. J. Impact Engng.* 6:87–99, 1987.

[8] T. Lepikult, W. Schmidt. Optimal design of rigid-plastic beams subjected to dynamical loading. *Structural Optimization.* 18:116–125, 1999.

[9] J. Lellep. Optimization of inelastic cylindrical shells. *Eng. Optimization.* 29:359–375, 1997.

[10] J. Lellep, P. Nestler, W. Schmidt. Optimization of elastic cylindrical shells. In *Proceedings of the Twenty Second Nordic Seminar on Computional Mechanics.* Aalborg, 2009.

[11] D. Chapelle, K. J. Bathe. *The Finite Element Analysis of Shells - Fundamentals.* Springer, 2003.

[12] V. Azhmyakov and W. Schmidt. On the optimal design of elastic beams. *Struct. Multidisc. Optim.* 27:80–88, 2004.

[13] P. Pedersen. *Elasticity - Anisotropy - Laminates, with matrix formulation, finite elements and an index to matrices.* Solid mechanics, 1998.

[14] P. Lee, K. J. Bathe. Insight into finite element shell discretizations by use of the "basic shell mathematical model". *Computers and Structures.* 83:69–90, 2005.

[15] O. Schenk, K. Gartner, W. Fichtner. Efficient Sparse LU Factorization with Left-right Looking Strategy on Shared Memory Multiprocessors. BIT, 40(1):158-176, 2000.

Ann. Acad. Rom. Sci.
Ser. Math. Appl.
ISSN 2066-6594 Vol. 6, No. 2/2014

# BOUNDEDNESS CONDITIONS FOR THE ANISOTROPIC NORM OF STOCHASTIC SYSTEMS WITH MULTIPLICATIVE NOISE*

Adrian–Mihail Stoica† Isaac Yaesh‡

**Abstract**

The aim of this paper is to provide conditions for the boundedness of the anisotropic norm of discrete–time linear stochastic systems with multiplicative noise. It is proved that these conditions can be expressed in terms of the existence of a stabilising solution of a specific Riccati equation satisfying some additional constraints.

**MSC**: 93E03, 93E10, 93E25

**keywords:** anisotropic norm, stochastic systems, multiplicative noise, optimal estimation

## 1 Introduction

The signal filtering problem received much attention over the last seven decades, starting with the early formulation and developments due to E. Hopf and N. Wiener in the 1940's. Two decades later, the well-known results of Kalman and Bucy ([10], [11]) have been successfully implemented in

*Accepted for publication in revised form on November 3-rd, 2014.

†adrian.stoica@upb.ro, University "Politehnica" of Bucharest, Faculty of aerospace engineering, Str. Polizu, No. 1, 011063, Bucharest, Romania

‡iyaesh@imi-israel.com, Control Department, IMI Advanced Systems Div., P.O.B. 1044/77, Ramat-Hasharon, 47100, Israel

many applications including aerospace, signal processing, geophysics, etc. and they have strongly influenced the research in this area. Some surveys on linear filtering and estimation can be found for instance in [9] and in [20]. Many papers devoted to this topics investigate the filtering performances with respect to the uncertainty modelling errors of the system which state is estimated. This interest is motivated by the fact that the filter performance deteriorates in the presence of modelling errors. Some of these papers consider the problem of robust filtering when the system is subject to parametric uncertainty (see *e.g.* [3], [6], [12], and the references therein). There are applications in which the system parameters are corrupted with random perturbations leading to stochastic models with *multiplicative noise*. Such stochastic systems have been intensively studied over the last few decades (see [23] for early references), many of the recent theoretical developments including optimal control and filtering results ([4], [6], [15]). Another important issue arising in filtering applications is related to the input of the systems generating the filtered signals. Besides developments based on Kalman filtering, also known as $H_2$-type filtering since the exogenous input signals are assumed white noises, alternative approaches have been proposed where deterministic bounded energy inputs are considered. Such formulations and developments have been performed in the framework of the $H_\infty$-norm minimisation ([7], [20]). Many practical applications require a compromise between the $H_2$ and the $H_\infty$ filtering since the $H_2$ norm minimisation of the estimation error may not be suitable when the considered signals are strongly coloured (e.g. periodic signals), and that $H_\infty$-optimization may poorly perform when these signals are weakly coloured (e.g. white noise), (see e.g. [1] and [16]). An promising alternative to accomplish such compromise is to use the so-called *a-anisotropic norm* (see e.g. [5], [13], [22]) since it offers and intermediate topology between the $H_2$ and $H_\infty$ norms. More precisely, if the coloured signal is generated by an $m$-dimensional exogenous input, the $a$-anisotropic norm $|||F|||_a$ of a stable system $F$ has the property $1/\sqrt{m}\|F\|_2 \leq |||F|||_a \leq \|F\|_\infty$ (see, for instance [13]).

In [22] a Bounded Real Lemma type result is proved for the anisotropic norm of discrete-time deterministic systems. It is shown that the boundedness norm condition implies to solve a nonconvex optimization in which frequency representation of the filtered signal plays a crucial role.

The aim of the present paper is to determine boundedness conditions for the anisotropic norm of *stochastic systems with multiplicative noise*. By contrast with the above mentioned papers, all the developments of this paper use time representations of the signals and the obtained results provide a generalisation of the ones derived the absence of the multiplicative noise and for the case when the system is subject to state-dependent noise [21].

*Notation.* Throughout the paper the superscript '$T$' stands for matrix transposition, $\mathbf{R}^n$ denotes the $n$ dimensional Euclidean space, $\mathbf{R}^{n\times m}$ is the set of all $n\times m$ real matrices, and the notation $P>0$ ($P\geq 0$), for $P\in\mathbf{R}^{n\times n}$ means that $P$ is symmetric and positive definite (positive semidefinite). The trace of a matrix $Z$ is denoted by $Tr\{Z\}$, and $|v|$ denotes the Euclidian norm of an $n$-dimensional vector $v$.

## 2 Preliminaries and Problem Statement

Consider the stochastic system with multiplicative noise

$$\begin{array}{rcl} x(t+1) & = & (A_0+\sum_{i=1}^{r}\xi_i(t)A_i)\,x(t)+(B_0+\sum_{i=1}^{r}\xi_i(t)B_i)\,w(t) \\ y(t) & = & Cx(t)+Dw(t),\; t=0,1,... \end{array} \tag{1}$$

where $\xi(t)=(\xi_1(t),...,\xi_r(t))^T$ is a sequence of independent random vectors $\xi:\Omega\to\mathbf{R}^r$ on a probability space$(\Omega,\mathcal{F},\mathcal{P})$. It is assumed that $\{\xi(t)\}_{t\geq 0}$ satisfies the conditions $E\left[\xi(t)\right]=0$ and $E\left[\xi(t)\xi^T(t)\right]=I_r$, $t=0,1,...$. The matrices of the state space model (1) have the dimensions $A_i\in\mathbf{R}^{n\times n}$, $B_i\in\mathbf{R}^{n\times m}$, $i=0,1....,r$, $C\in\mathbf{R}^{p\times n}$, $D\in\mathbf{R}^{p\times m}$.

It is assumed that the input $w(t)$ are random variables generated by a linear stochastic filter with multiplicative noise $G$

$$\begin{array}{rcl} \tilde{x}(t+1) & = & (A_{f_0}+\sum_{i=1}^{r}\xi_i(t)A_{f_i})\,\tilde{x}(t)+(B_{f_0}+\sum_{i=1}^{r}\xi_i(t)B_{f_i})\,v(t) \\ w(t) & = & C_f\tilde{x}(t)+D_f v(t),\; t=0,1,... \end{array} \tag{2}$$

where the order $n_f$ and the matrices $A_{f_i}\in\mathbf{R}^{n_f\times n_f}$, $B_{f_i}\in\mathbf{R}^{n_f\times m}$, $i=0,1....,r$, $C_f\in\mathbf{R}^{m\times n_f}$, $D\in\mathbf{R}^{m\times m}$ are not prefixed and $v(t)\in\mathbf{R}^m$ are white noise vectors with the properties $E\left[v(t)\right]=0$ and $E\left[v(t)v^T(t)\right]=I_m$, $t=0,1,...$. It is assumed that $\{\xi(t)\}_{t\geq 0}$ and $\{v(t)\}_{t\geq 0}$ are independent stochastic processes.

**Definition 1** *A stochastic system with multiplicative noise of form (1) with $B_i=0$, $i=0,1,...,r$ is called exponentially stable in mean square (ESMS) if there exist $\beta\geq 1$ and $\rho\in(0,1)$ such that $E\left[|\Phi(t,s)x(0)|^2\right]\leq\beta\rho^{(t-s)}|x(0)|^2$ for all $t\geq s\geq 0$, $x(0)\in\mathbf{R}^n$, where $\Phi(t,s)$ denotes the fundamental matrix solution of (1).*

Throughout the paper it will be assumed that both systems (1) and (2) are ESMS.

**Definition 2** *The $H_2$-type norm of the ESMS system (1) is defined as*

$$\|F\|_2 = \left[ \lim_{\ell \to \infty} \frac{1}{\ell} \sum_{t=0}^{\ell} E\left[ y^T(t) y(t) \right] \right]^{\frac{1}{2}}.$$

The next result provides a method to compute the $H_2$ norm of the stochastic system (1) (see e.g. [4]).

**Lemma 1** *The $H_2$ type norm of the ESMS system (1) is given by $\|F\|_2 = \left( Tr\left( \sum_{i=0}^{r} B_i^T X B_i + D^T D \right) \right)^{\frac{1}{2}}$ where $X \geq 0$ is the solution of the Lyapunov equation $X = \sum_{i=0}^{r} A_i^T X A_i + C^T C$.*

Let $L^2(\mathbf{Z} \times \Omega, \mathbf{R}^m)$ the space of all sequences $w = \{w(t)\}_{t \in \mathbf{Z}_+}$ of $m$-dimensional vectors with $\|w\|^2 := \sum_{t=-\infty}^{\infty} E|w(t)|^2 < \infty$ and by $\tilde{L}^2(\mathbf{Z}_+ \times \Omega, \mathbf{R}^m)$ the space of all $w \in L^2(\mathbf{Z}_+ \times \Omega, \mathbf{R}^m)$ such that $w(t)$ are measurable with respect to $\mathcal{F}_t$ for every $t \in \mathbf{Z}_+$, $\mathcal{F}_t \subset \mathcal{F}$ denoting a family of $\sigma$-algebras associated to the vectors $\xi(t)$. In [14] it is proved that if the system (1) is ESMS, one may define the linear bounded input-output operator

$$(Fw)(t) : \tilde{L}^2(\mathbf{Z}_+ \times \Omega, \mathbf{R}^m) \to \tilde{L}^2(\mathbf{Z}_+ \times \Omega, \mathbf{R}^p)$$

by

$$(Fw)(t) = Cx(t) + Dw(t), \ t \in \mathbf{Z}_+,$$

$x(t)$ being the solution of (1) with zero initial condition. Denoting by $\|F\|_\infty$ the norm of the above operator, one can prove the following Bounded Real Lemma type result for stochastic systems of form (1) with respect to the $H_\infty$ norm [14].

**Lemma 2** *The ESMS system (1) has the norm $\|F\|_\infty < \gamma$ for a certain $\gamma > 0$ if and only if the Riccati equation*

$$P = \sum_{i=0}^{r} A_i^T P A_i + \left( \sum_{i=0}^{r} A_i^T P B_i + C^T D \right) \left( \gamma^2 I - \sum_{i=0}^{r} B_i^T P B_i - D^T D \right)^{-1} \times \left( \sum_{i=0}^{r} A_i^T P B_i + C^T D \right)^T + C^T C$$

*has a stabilizing solution $P \geq 0$ such that $\gamma^2 I - \sum_{i=0}^{r} B_i^T P B_i - D^T D > 0$.*

It is recalled that a symmetric solution $P$ of the above Riccati equation is called a *stabilising solution* if the stochastic system

$$x(t+1) = \left( A_0 + B_0 K + \sum_{i=1}^{r} \xi_i(t) \left( A_i + B_i K \right) \right) x(t)$$

is ESMS, where by definition

$$K := \left(\gamma^2 I - \sum_{i=0}^{r} B_i^T P B_i - D^T D\right)^{-1} \left(\sum_{i=0}^{r} A_i^T P B_i + C^T D\right)^T.$$

Given an ESMS filter of form (2), the *mean anisotropy* of the random variable $w(t)$ generated by $G$ is defined as

$$\bar{A}(G) = -\frac{1}{2} \ln\det\left(\frac{mE\left[\tilde{w}(0)\tilde{w}^T(0)\right]}{\|G\|_2^2}\right) \tag{3}$$

where $\tilde{w}(0) = w(0) - E\left[w(0)|\,(w(k))_{k<0}\right]$ denotes the prediction error of $w(0)$ based on $w(k)$, $k < 0$ (see details in [5]). Then the *a-anisotropic norm* of $F$ is defined as ([5])

$$|||F|||_a = \sup_{G\in\mathcal{G}_a} \frac{\|FG\|_2}{\|G\|_2}, \tag{4}$$

where $\mathcal{G}_a$ denotes the set of all stochastic systems of form (2) with $\bar{A}(G) < a$.

## 3 Main result

**Theorem 1** *The stochastic system with multiplicative noise (1) has the a-anisotropic norm less than a given* $\gamma > 0$ *if there exists* $q \in \left(0, \min\left(\gamma^{-2}, \|F\|_\infty^{-2}\right)\right)$ *such that the Riccati equation*

$$\begin{aligned} X = {} & \textstyle\sum_{i=0}^r A_i^T X A_i + \left(\sum_{i=0}^r A_i^T X B_i + C^T D\right) \\ & \times \left(\tfrac{1}{q} I - \textstyle\sum_{i=0}^r B_i^T X B_i - D^T D\right)^{-1} \left(\textstyle\sum_{i=0}^r A_i^T X B_i + C^T D\right)^T + C^T C \end{aligned} \tag{5}$$

*has a stabilizing solution* $X \geq 0$ *satisfying the following conditions*

$$\Psi_q := \frac{1}{q} I - \sum_{i=0}^{r} B_i^T X B_i - D^T D > 0 \tag{6}$$

*and*

$$\det\left(\frac{1}{q} - \gamma^2\right) \Psi_q^{-1} \leq e^{-2a}. \tag{7}$$

*Proof.* Using the Definition 1 of the $H_2$-type norm it follows that the condition $\sup_{G\in\mathcal{G}_a}\frac{\|FG\|_2}{\|G\|_2}<\gamma$ is equivalent with the condition

$$\lim_{\ell\to\infty}\frac{1}{\ell}\sum_{t=0}^{\ell}E\left[|y(t)|^2-\gamma^2|w(t)|^2\right]<0 \tag{8}$$

for all $w(t)$ generated by filters $G\in\mathcal{G}_a$.

For the sake of simplicity writing, the following notations will be introduced:

$$\begin{array}{rcl}\mathcal{A}(t) & := & A_0+\sum_{i=1}^{r}\xi(t)A_i\\ \mathcal{B}(t) & := & B_0+\sum_{i=1}^{r}\xi(t)B_i\,.\end{array}$$

Using (1) it follows that

$$\begin{gathered}x^T(t+1)Xx(t+1)-x^T(t)Xx(t)=[\mathcal{A}(t)x(t)+\\ \mathcal{B}(t)w(t)]^TX[\mathcal{A}(t)x(t)+\mathcal{B}(t)w(t)]-x^T(t)x(t)-y^T(t)y(t)+x^T(t)C^TCx(t)\\ +x^T(t)C^TDw(t)+w^T(t)D^TCx(t)+w^T(t)D^TDw(t)\end{gathered}$$

where we added the zero term $y^T(t)y(t)-(Cx(t)+Dw(t))^T(Cx(t)+Dw(t))$. Collecting terms we readily obtain

$$\begin{gathered}y^T(t)y(t)=x^T(t)[\mathcal{A}(t)^TX\mathcal{A}(t)-X+C^TC]x(t)+w^T(t)[D^TD+\\ \mathcal{B}(t)^TX\mathcal{B}(t)]w(t)+w^T(t)[D^TC+\mathcal{B}(t)^TX\mathcal{A}(t)]x(t)\\ +x^T(t)[C^TD+\mathcal{A}(t)^TX\mathcal{B}(t)]w(t)+x^T(t)Xx(t)-x^T(t+1)Xx(t+1)\,.\end{gathered}$$

Noting that the properties of the random sequence $\{\xi(t)\}_{t\geq 0}$ imply $E\{\mathcal{A}^TX\mathcal{A}\}=\sum_{i=0}^{r}A_i^TXA_i$, $E\{\mathcal{B}^TX\mathcal{B}\}=\sum_{i=0}^{r}B_i^TXB_i$ and $E\{\mathcal{A}^TX\mathcal{B}\}=\sum_{i=0}^{r}A_i^TXB_i$, it follows from the above equation that

$$\begin{gathered}E\{y^T(t)y(t)\}=E\{x^T(t)[\sum_{i=0}^{r}A_i^TXA_i-X+C^TC]x(t)\\ +w^T(t)[D^TD+\sum_{i=0}^{r}B_i^TXB_i]w(t)\\ +w^T(t)[D^TC+\sum_{i=0}^{r}B_i^TXA_i]x(t)+x^T(t)[C^TD+\sum_{i=0}^{r}A_i^TXB_i]w(t)\\ +x^T(t)Xx(t)-x^T(t+1)Xx(t+1)\}\,.\end{gathered}$$

Substituting from (5) into the first bracket in the above equation, one obtains

$$\begin{gathered}E\left[|y(t)|^2\right]=E\Big[x^T(t)Xx(t)-x^T(t+1)Xx(t+1)\\ +x^T\left(\sum_{i=0}^{r}A_i^TXB_i\right)w(t)+w^T(t)\left(\sum_{i=0}^{r}B_i^TXA_i\right)x(t)\\ -x^T(t)\left(\sum_{i=0}^{r}A_i^TXB_i+C^TD\right)\left(\tfrac{1}{q}I-\sum_{i=0}^{r}B_i^TXB_i-D^TD\right)^{-1}\\ \times\left(\sum_{i=0}^{r}A_i^TXB_i+C^TD\right)^Tx(t)+w^T(t)\left(\sum_{i=0}^{r}B_i^TXB_i\right)w(t)\\ +x^T(t)C^TDw(t)+w^T(t)D^TCx(t)+w^T(t)D^TDw(t)\Big]\,.\end{gathered} \tag{9}$$

Define

$$
\begin{aligned}
\mathcal{P}(t) := \Big[ & w^T(t) - x^T(t)\left(\textstyle\sum_{i=0}^r A_i^T X B_i + C^T D\right) \\
& \times \left(\tfrac{1}{q} I - \textstyle\sum_{i=0}^r B_i^T X B_i - D^T D\right)^{-1} \\
& - v^T(t)\left(\tfrac{1}{q} I - \textstyle\sum_{i=0}^r B_i^T X B_i - D^T D\right)^{-\frac{1}{2}}\Big] \\
& \times \left(\tfrac{1}{q} I - \textstyle\sum_{i=0}^r B_i^T X B_i - D^T D\right) \\
\times \Big[ & w(t) - \left(\tfrac{1}{q} I - \textstyle\sum_{i=0}^r B_i^T X B_i - D^T D\right)^{-1}\left(\textstyle\sum_{i=0}^r B_i^T X A_i + D^T C\right) x(t) \\
& - \left(\tfrac{1}{q} I - \textstyle\sum_{i=0}^r B_i^T X B_i - D^T D\right)^{-\frac{1}{2}} v(t)\Big].
\end{aligned}
\tag{10}
$$

Then, using the properties of $\{v(t)\}_{t\geq 0}$ it follows that

$$
\begin{aligned}
E\left[\mathcal{P}(t)\right] = E\Big[ & w^T(t)\left(\tfrac{1}{q} I - \textstyle\sum_{i=0}^r B_i^T X B_i - D^T D\right) w(t) \\
& - w^T(t)\left(\textstyle\sum_{i=0}^r B_i^T X A_i + D^T C\right) x(t) \\
& - x^T(t)\left(\textstyle\sum_{i=0}^r A_i^T X B_i + C^T D\right) w(t) \\
& + x^T(t)\left(\textstyle\sum_{i=0}^r A_i^T X B_i + C^T D\right)\left(\tfrac{1}{q} I - \textstyle\sum_{i=0}^r B_i^T X B_i - D^T D\right)^{-1} \\
& \times \left(\textstyle\sum_{i=0}^r B_i^T X A_i + D^T C\right) x(t) \\
& - 2Tr\, D_f\left(\tfrac{1}{q} I - \textstyle\sum_{i=0}^r B_i^T X B_i - D^T D\right)^{\frac{1}{2}}\Big] + m\,.
\end{aligned}
\tag{11}
$$

Taking into account (9) and (11) one obtains

$$
\begin{aligned}
& E\left[|y(t)|^2 - \gamma^2 |w(t)|^2\right] = E\Big[x^T(t) X x(t) - x^T(t+1) X x(t+1) - \mathcal{P}(t) \\
& - 2Tr\, D_f\left(\tfrac{1}{q} I - \textstyle\sum_{i=0}^r B_i^T X B_i - D^T D\right)^{\frac{1}{2}} + m + \left(\tfrac{1}{q} - \gamma^2\right) w^T(t) w(t)\Big].
\end{aligned}
\tag{12}
$$

Since the systems (1) and(2) are ESMS

$$
\lim_{\ell\to\infty} \frac{1}{\ell} E\left[x^T(0) X x(0) - x^T(\ell) X x(\ell)\right] = 0,
$$

and then one directly obtains that

$$
\begin{aligned}
& \lim_{\ell\to\infty} \tfrac{1}{\ell} \textstyle\sum_{t=0}^{\ell} E\left[|y(t)|^2 - \gamma^2 |w(t)|^2\right] \\
= & \lim_{\ell\to\infty} \tfrac{1}{\ell} E\left[-\textstyle\sum_{t=0}^{\ell} \mathcal{P}(t) + \textstyle\sum_{t=0}^{\ell}\left(\tfrac{1}{q} - \gamma^2\right) w^T(t) w(t)\right] \\
& - 2Tr\, D_f\left(\tfrac{1}{q} I - \textstyle\sum_{i=0}^r B_i^T X B_i - D^T D\right)^{\frac{1}{2}} + m.
\end{aligned}
\tag{13}
$$

From (10) it follows that $\mathcal{P}(t) \geq 0$ and $\mathcal{P}(t) = 0$ for

$$\begin{array}{l} w(t) = \left(\frac{1}{q}I - \sum_{i=0}^{r} B_i^T X B_i - D^T D\right)^{-1} \left(\sum_{i=0}^{r} B_i^T X A_i + D^T C\right) x(t) \\ + \left(\frac{1}{q}I - \sum_{i=0}^{r} B_i^T X B_i - D^T D\right)^{-\frac{1}{2}} v(t). \end{array} \tag{14}$$

The above condition is fulfilled for a filter $G$ having the state $\tilde{x}(t)$ equal to the state $x(t)$ of $F$ and if the following conditions are accomplished

$$\begin{array}{rcl} C_f & = & \left(\frac{1}{q}I - \sum_{i=0}^{r} B_i^T X B_i - D^T D\right)^{-1} \left(\sum_{i=0}^{r} B_i^T X A_i + D^T C\right) \\ D_f & = & \left(\frac{1}{q}I - \sum_{i=0}^{r} B_i^T X B_i - D^T D\right)^{-\frac{1}{2}}. \end{array} \tag{15}$$

For $w(t)$ given by (14) the first equation (1) becomes

$$\begin{array}{l} x(t+1) = \left(A_0 + \sum_{i=1}^{r} \xi_i(t) A_i\right) x(t) + \left(B_0 + \sum_{i=1}^{r} \xi_i(t) B_i\right) \\ \times \left(\frac{1}{q}I - \sum_{i=0}^{r} B_i^T X B_i - D^T D\right)^{-1} \left(\sum_{i=0}^{r} B_i^T X A_i + D^T C\right) x(t) \\ + \left(B_0 + \sum_{i=1}^{r} \xi_i(t) B_i\right) \left(\frac{1}{q}I - \sum_{i=0}^{r} B_i^T X B_i - D^T D\right)^{-\frac{1}{2}} v(t). \end{array} \tag{16}$$

Since $\tilde{x}(t)$ equals $x(t), t = 0, 1, ...$ from the above equation one obtains

$$\begin{array}{rcl} A_{f_i} & = & A_i + B_i \left(\frac{1}{q}I - \sum_{j=0}^{r} B_j^T X B_j - D^T D\right)^{-1} \\ & & \times \left(\sum_{j=0}^{r} B_j^T X A_j + D^T C\right) \\ B_{f_i} & = & B_i \left(\frac{1}{q}I - \sum_{j=0}^{r} B_j^T X B_j - D^T D\right)^{-\frac{1}{2}}, i = 0, 1, ..., r. \end{array} \tag{17}$$

Since $X$ is the stabilising solution of the Riccati equation (5) it follows that the filter with $A_{f_i}$, $i = 0, 1, ..., r$ given above is ESMS.

Based on the expression (15) of $D_f$ and since $\tilde{x}(t) = x(t)$, from the second equation (2) it follows that

$$E\left[\tilde{w}(0)\tilde{w}^T(0)\right] = \left(\frac{1}{q}I - \sum_{i=0}^{r} B_i^T X B_i - D^T D\right)^{-1}. \tag{18}$$

In the following it will be shown that under the assumption (7) from the statement, for all ESMS filters $G \in \mathcal{G}_a$ having $D_f = \Psi_q^{-\frac{1}{2}}$ the following condition is accomplished

$$-m + \left(\frac{1}{q} - \gamma^2\right) \|G\|_2^2 < 0. \tag{19}$$

Indeed, since $G \in \mathcal{G}_a$ and since $D_f = \Psi_q^{-\frac{1}{2}}$ it follows that

$$\det \frac{m\Psi_q^{-1}}{\|G\|_2^2} > e^{-2a}. \tag{20}$$

Taking into account (7) and the above inequality it follows that

$$\det \frac{m\Psi_q^{-1}}{\|G\|_2^2} > \det \left(\frac{1}{q} - \gamma^2\right) \Psi_q^{-1}$$

from which one directly obtains (19). Using the inequality (19), equations (13), (11), the equation for $D_f$ in (15) and Definition 2, it follows that $||FG||_2/||G||_2 < \gamma$.

Let us consider now the more general case for a certain filter $G \in \mathcal{G}_a$, satisfying therefore the condition

$$-\frac{1}{2} \ln \det \frac{mD_f D_f^T}{\|G\|_2^2} \leq a\,. \tag{21}$$

From the above condition and from the assumption (7) it follows that

$$\det \left(\frac{1}{q} - \gamma^2\right) \Psi_q^{-1} < \det \frac{mD_f D_f^T}{\|G\|_2^2}. \tag{22}$$

Using the general property $\det(A) \leq (Tr(A)/m)^m$, from the above inequality one obtains

$$Tr\left(D_f \Psi_q^{\frac{1}{2}}\right) > \left(\frac{1}{q} - \gamma^2\right)^{\frac{1}{2}} m^{\frac{1}{2}} \|G\|_2 \tag{23}$$

and thus

$$\begin{aligned}
&\left(\tfrac{1}{q} - \gamma^2\right) \|G\|_2^2 - 2Tr\left(D_f \Psi_q^{\frac{1}{2}}\right) + m \\
&< \left(\tfrac{1}{q} - \gamma^2\right) \|G\|_2^2 - 2\left(\tfrac{1}{q} - \gamma^2\right)^{\frac{1}{2}} m^{\frac{1}{2}} \|G\|_2 + m \\
&= \left(\left(\tfrac{1}{q} - \gamma^2\right)^{\frac{1}{2}} \|G\|_2 - m^{\frac{1}{2}}\right)^2
\end{aligned} \tag{24}$$

From the above inequality it follows that if

$$\left(\frac{1}{q} - \gamma^2\right) \|G\|_2^2 = m \tag{25}$$

the left hand side of it is negative and therefore from (12) it follows that $\|FG\|_2/\|G\|_2 < \gamma$. The condition (25) implies that

$$\frac{1}{q} - \gamma^2 = \frac{m}{\|G\|_2^2}.$$

Substituting the above expression into (7) one obtains the condition

$$-\frac{1}{2}\ln\det\frac{m\Psi_q^{-1}}{\|G\|_2^2} \geq a\,. \tag{26}$$

Comparing (26) with the definition of the mean anisotropy one concludes that if for a filter $G \in \mathcal{G}_a$ there exists another filter $\hat{G}$ with $\bar{A}\left(\hat{G}\right) \geq a$ such that $\|\hat{G}\|_2 = \|G\|_2$ and having $\hat{D}_f = \Psi_q^{-\frac{1}{2}}$ for a certain $q$ satisfying the assumptions of Theorem 1, then $\|FG\|_2/\|G\|_2 < \gamma$. A similar conclusion is derived in the deterministic framework in [13]. Such a $\hat{G}$ always may be found. Indeed since $\Psi_q^{-1} \to 0$ for $q \to 0$, it follows that the Riccati equation (5) has a stabilising solution and the condition (26) is fulfilled for a small enough $q > 0$. Then for any $G \in \mathcal{G}_a$, based on Lemma 1 one can easily determine $\hat{A}_{f_i}, \hat{B}_{f_i}, i = 0, ..., r$ and $\hat{C}_f$ such that $\|\hat{G}\|_2 = \|G\|_2$.

Using the inequality (19), (13), (11), the equation for $D_f$ in (15) and Definition 2, it follows that $||FG||_2/||G||_2 < \gamma$. Let us finally notice that according with the Lemma 2, it follows that a necessary condition for the existence of a stabilizing solution of the Riccati equation (5) is $1/q \geq \|F\|_\infty^2$, from which it follows that $q \leq \|F\|_\infty^{-2}$. Thus the proof is complete.

## References

[1] D.S. Bernstein and W.M. Haddad, "LQG Control with and $H_\infty$ Performance Bound a Riccati equation Approach", *IEEE Transactions on Automatic Control*, Vol. 34, pp. 293-305, 1989.

[2] R.R. Bitmead, M. Gevers, I.R. Petersen and R.J. Kaye, "Monotonicity and stabilizability properties of solutions of the Riccati difference equation: propositions, lemmas, theorems, fallacious conjectures and counter-examples", *Systems & Control Letters*, vol. 5, No 5, pp. 309-315, April 1985.

[3] Costa, O. L. V., and Kubrusly, C. S., " State-feedback $H_\infty$-control for discrete-time infinite-dimensional stochastic bilinear systems," *J. Math. Sys. Estim. & Contr.*, Vol. 6, 1996, pp. 1-32.

[4] V. Dragan, T. Morozan and A.-M. Stoica, *Mathematical Methods in Robust Control of Discrete-Time Linear Stochastic Systems*, Springer, 2010.

[5] P. Diamond, I. Vladimirov, A. Kurdyukov and A. Semyonov, *Anisotropy-Based Performance Analysis of Linear Discrete Time Invariant Control Systems*, Int. Journal of Control, Vol. 74, pp. 28-42, 2001.

[6] E. Gershon, U. Shaked and I. Yaesh, $H_\infty$ *Control and Estimation of State-Multiplicative Linear Systems*, Springer, 2005.

[7] M.J. Grimble, "$H_\infty$ design of optimal linear filters", *Linear Circuits, Systems and Signal Processing: Theory and Applications*, C. Byrnes, C. Martin and R. Saeks (Eds.) North-Holland, Amsterdam, The Netherlands, pp. 55-540, 1988.

[8] A. Halanay and T. Morozan, "Optimal stabilizing compensators for linear discrete–time linear systems under independent random perturbations", *Revue Roumaine Math. Pures et Appl.*, 37(3), 1992, pp. 213–224.

[9] T. Kailath, "A view of three decades of linear filtering theory", IEEE Transactions on Information Theory, 20, 1974, pp. 146–181.

[10] R. Kalman, "A new approach to linear filtering and prediction problems", *ASME Trans.–Part D, J. Basic Engineering*, 82, 1960, pp. 34-45.

[11] R. Kalman and R.S. Bucy, "New results in linear filtering and prediction theory", *ASME Trans.–Part D, J. Basic Engineering*, 83, 1961, pp. 95-108.

[12] R.S. Mangoubi, *Robust Estimation and Failure Detection*, Springer, 1998.

[13] A. P. Kurdyukov, E. A. Maksimov and M.M. Tchaikovsky *Anisotropy-Based Bounded Real Lemma*, Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems-MTNS 2010, Budapest, Hungary.

[14] T. Morozan, "Parametrized Riccati equations associated to input–output operators for discrete-time systems with state–dependent noise", *Stochastic Analysis and Applications*, 16(5), 1998, pp. 915-931.

[15] I.R. Petersen, V.A. Ugrinovskii and A.V. Savkin, *Robust Control Design Using $H^\infty$ Methods*. Springer-Verlag, 2000.

[16] H. Rotstein and M. Sznaier, *An Exact Solution to General Four Block Discrete-Time Mixed $H_2/H_\infty$ problems via COnvex Optimization*, IEEE Transactions on Automatic Control, Vol. 43, pp. 1475-1481, 1998.

[17] A.-M Stoica and I. Yaesh, "Markovian Jump Delayed Hopfield Networks with Multiplicative Noise", Automatica, Vol. 44, pp. 49-55, 2008.

[18] A.-M Stoica and I. Yaesh, "Kalman Type Filtering for Discrete-Time Stochastic Systems with State-Dependent Noise", Proceedings of the MTNS 2008, Blacksburg, Virginia, 2008.

[19] W. Li, E. Todorov and R.E. Skelton, "Estimation and Control of Systems with Multiplicative Noise via Linear Matrix Inequalities", Proceedings of the 43rd IEEE Conference on Decision and Control, December 14-17, 2004 Atlantis, Paradise Island, Bahamas.

[20] D. Simon, *Optimal State Estimation. Kalman, $H_\infty$ and Nonlinear Approaches*, Wiley, 2006.

[21] A.-M. Stoica and I. Yaesh, On the anisotropic norm of discrete-time stochastic systems with state dependent noise, *Annals of the Academy of Romanian Scientists. Series on mathematics and its applications*, Vol. 4, No. 2, pp. 209-220, 2012.

[22] M.M. Tchaikovsky, A.P. Kurdyukov and V.N. Timin, *Synthesis of Anisotropic Suboptimal Controllers by Convex Optimization*, arXiv 1108.4982v4[cs.SY], 2011.

[23] W.M. Wonham, "Random differential equations in control theory", *Probabilistic methods in Applied Math.*, 2, 1970.