ADVANCED CALCULUS

CANNON CLARKE

Advanced Calculus

Advanced Calculus

Cannon Clarke

Advanced Calculus
Cannon Clarke
ISBN: 978-1-9790-0439-0

Published by Academic Studio,
5 Penn Plaza,
19th Floor,
New York, NY 10001, USA

Contents

Washback Effect of University Entrance exams in Applied Mathematics to Social Sciences

Luis J. Rodríguez-Muñiz[1]*, Patricia Díaz[1], Verónica Mier[2], Pedro Alonso[3]

1 Department of Statistics and O.R., and Math. Education, Universidad de Oviedo, Oviedo, Spain, **2** CES San Eutiquio, La Salle, Gijón, Spain, **3** Department of Mathematics, Universidad de Oviedo, Oviedo, Spain

* luisj@uniovi.es

Abstract

Curricular issues of subject Applied Mathematics to Social Sciences are studied in relation to university entrance exams performed in several Spanish regions between 2009–2014. By using quantitative and qualitative analyses, it has been studied how these exams align with curriculum and how they produce a washback on curriculum and teachers' work. Additionally, one questionnaire about teachers' practices has been performed, in order to find out how the exams are influencing teaching methodology development. Main results obtained show that evaluation is producing a bias on the official curriculum, substantially simplifying the specific orientation that should guide applied mathematics. Furthermore, teachers' practices are influenced by the exams, and they usually approach their teaching methodology to the frequent types of exams. Also, slight differences among the teachers lead to distinguish two behavioral subgroups. Results can also be useful in an international context, because of the importance of standardized exit exams in OECD countries.

Editor: J. Alberto Conejero, IUMPA - Universitat Politecnica de Valencia, SPAIN

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Entrance exams to university (from now on, Spanish acronym PAU will be used, from *Pruebas de Acceso a la Universidad*) mean the main way of access to higher education, involving percentages over 70% from the total new freshmen to Spanish university [1]. They are into effect, with slight changes, since 1974. In the last performed model, under the Organic Law of Education [2], Ministry fixes some minimum requirements in Baccalaureate curriculum [3] that is later completed at regional level. Subsequently, the current structure of the PAU [4] was applied for the first time in the academic year 2009/2010. In every Autonomous Community (i.e., region) a committee composed by teachers and University professors designs and implements the exams by choosing the most influential topics in the assessment, the orientation of the questions, the level of domain and other specific characteristics of the exams.

Several years after the implementation of the new model, an important number of assessment units have been released, so that, it is possible to assess which concrete variations they have suffered in the subject Applied Mathematics to Social Sciences 2 (from now on, AMSS2, in the second year of *Bachillerato*, i.e., last year of upper secondary education,), which is the degree of fitting of the exams to the official curriculum, and how it all affects teachers' practices.

Curricular guidelines orientate AMSS2 curriculum towards the need of solving real problems from the field of Social Sciences, within a proper context, and of teaching and learning Mathematics considering its instrumental essence in Social Sciences. Therefore, this kind of problems and exercises should appear in AMSS2 PAU exams. Problem solving and decision making skills are in almost all international curriculum for upper secondary mathematics, especially when they are applied to Social Sciences. The problem of how to reflect this type of problems in exit exams, external assessment or university entrance exams is, therefore, an international trend, as in many OECD countries this type of exams is being implemented (for an international comparative, see [5]).

Two are the main goals in this paper. The first one is to analyse the type of exercises and problems that are proposed in AMSS2 PAU exams and their relationship with the official curricula, in other words, the alignment between exams and the orientations in the official curriculum (note that it is fixed mainly by the national government and slightly completed by regional government, whereas exams composition involve both regional government and public universities in every region). Moreover, it is our interest to determine at what extent the posed questions have a clear relationship with applied problems, as it is stated in the official curricula.

The second goal is to study of the duality between official and real curricula, by analysing teachers' practice. By using an *ad hoc* questionnaire, it is verified if PAU exams produce a washback in teachers' activity.

These targets are addressed with the following research questions:

- How does PAU exams reflect official curricula in AMSS2? Are there biases within the exams?
- Does PAU influence AMSS2 teachers' practice? How does this affect the implementation of the real curricula?

Theoretical framework

Literature describes the effects of high-stake testing programs, defined as "tests whose results are used to trigger actions or decisions such as passing or failing a grade, graduating or not, determining teacher or principal merit or assuming responsibility for a failing district by a state agency" ([6]; cited by [7]). Apple [8,9] was one of the first authors identifying how centralized curriculum and assessment deskills teachers. Runté [10] enumerates these deskilling processes, emphasizing that centralized curriculum limits the range of skills required in making curricular decisions and it implies a shift from student-centred to curriculum-centred instruction. Smith [7] classified the effects of high-stake into 6 different categories, from which the present paper is only focused on types 4 (reducing the time available for instruction) and 5 (narrowing curriculum and reducing teachers' ability to adapt, create, or diverge).

Jones et al. [11] conducted a study with North Carolina teachers concluding that high-stakes testing increase the amount of time that teachers dedicate to practice tests in their lectures. Besides, they confirm that "material that involves higher-order thinking and problem solving often falls by the wayside" [11].

Some authors support benefits from this type of curriculum-based external exit evaluations. Bishop [12] demonstrates how this type of exams improves students' achievement in international assessments. Häkkinen [13] shows how Finnish university system would enlarge the proportion of university students if admission system were based on entrance exams instead of different admission criteria. Ou [14] demonstrates how students barely passing maths exams are more likely to dropout at university in the USA. On the other hand, Jacob [15] showed that

the impact is not positive in general, by using a quantitative model relating scores on exit exams and university dropout. More specifically, other studies explore the field of mathematics [16–18] and also study differences between the official and the real curriculum (in the sense of Perrenoud [19]).

Spanish literature also reflects comparative studies about PAU and curriculum-based exit external evaluations in different countries or university entrance exams [20], some analyses about factors influencing the performance in Spanish exams and their predictive capacity [21–23]. Several studies have pointed out the influence of PAU mathematics exams on the definition of the real curriculum, methodologies, and, barely, teachers' and students' attitudes [24–27]. No references are found in literature considering curricular particularities of subject AMSS2. Some recent research have analysed probability and statistical inference problems in PAU exams in Andalusia [28,29], by using the onto-semiotic approach.

Regarding teaching practices, we follow the Mathematical Knowledge for Teaching framework (MKT) for characterizing teachers' knowledge, developed in the group leaded by Ball [30]. According to this model, teachers' practices belong to the Pedagogical Content Knowledge (PCK) domain, and, particularly, when examining practices in relation with high-stakes exams, we are studying the Knowledge of Content and Curriculum (KCC) and Knowledge of Content and Teaching (KCT). On the other, several authors have pointed out the influence of beliefs on practices, but also the difficulty, even impossibility, of measuring beliefs [31] by themselves, that is: "beliefs are referred to as constructed in the same sense that knowledge is constructed" ([31], p.128).

Assuming this theoretical perspective, it is also necessary to pay attention to literature regarding teachers' practices related to high-stake exams. Bishop [12] showed how Canadian teachers tend to develop more complex tasks in the classroom in order to prepare graduation exams. Many studies have been developed trying to predict students' performance based on several teachers' characteristics such as experience or qualification (see, for instance, [32–34]). Recently, in [35] it is analysed how the type of classroom task and the amount of homework predict the outcomes in the Russian equivalent to PAU exam (USE).

In this research we are centred on the washback effect instead of focussing on students' performance. The term washback has been coined in Educational Sciences to denote the influence of testing on teaching and learning ([36], p.259). It has been coined and especially used in language learning and assessing, but it can also be used in the mathematical context of the present work, since it also affects "curriculum materials, teaching methods, feelings and attitudes, learning" ([37], p.7). The concept itself can be used with different scopes (see, for instance, [38]), but in this paper it will be considered the effect on curriculum and teachers' practices. When studying the effect of German Abitur exams, similar to PAU, [39] determine that: "It can thus be expected that, in the final years of schooling in particular, teachers align the standards of their in-class assessments with those of the upcoming central examinations."

Therefore, our theoretical framework is built by integration of these previous approaches, and it is focused on the curricular analysis of AMSS2 related to the PAU exams, and on analysing the washback from PAU on teaching and learning methodologies. In summary: "It is testing, not the official stated curriculum, that is increasingly determining what is taught, how it is taught, what is learned, and how it is learned" ([40], p.88).

Additionally from previous sources, in the theoretical framework, from the methodological point of view, this work contributes with a novel approach to curricular analysis, paying attention to contents and their frequency appearing in PAU exams, and using a categorization inspired by conceptual focuses, developed in [41]. Since, for the present work, a broader classification than that in [41] is needed, the notion of curricular unit has been developed, and it will be deeply explained in the methodological section.

Methodology

Considering that two different studies are developed, it is necessary to deal with two different samples.

Sample I: PAU exams

Data from PAU exams have been collected in Andalusia, Asturias, the Basque Country and Madrid, from 2009–2010 to 2013–2014. Data were obtained from universities or regional education authorities web sites [42–45]. Asturias has been considered since it is authors' original region, the Basque Country was interesting since it higher degree of autonomy, whereas Andalusia and Madrid have been considered since they have a great number of students in the exams. Therefore, the sample covers over 40% of the total number of students passing PAU exams every year, which illustrate its significance [46].

Each region has some degrees of freedom in the organization of PAU exams, what implies the number of exams is different among them. For instance, some regions as the Basque Country use the same exam for the two phases of PAU (both compulsory and optional ones) where as others as Asturias use different exams for both phases. Additionally, some regions release only the used exams, but other release also emergency exams (those used for extraordinary situations as blackouts, student accidents, etc.). Nevertheless, every exam has two different options, every option consisting in a set of exercises, so that the student has to choose one option and solve it completely, thus, for statistical purposes, we can consider each option as an exam. Hence, summing up, the number of analysed options, and the number of exercises included in options (usually 4 or 5, but it can also vary from one region to another and even from one year to another) are showed in Table 1. Actually, the whole population of exams in these 4 regions is studied, not a sample, since all the released exams have been analysed.

Sample II: teachers

In order to analyse whether teachers' practice in AMSS2 is conditioned by PAU exams, a sample of 51 Mathematics teachers in Secondary Schools has been used. The questionnaire was performed in two different periods: January-April 2013 and September-December 2014. Due to budget constraints, the sample was chosen with teachers from Asturias. They were selected by convenience sampling, by contacting those high schools in which future mathematics teachers were developing their internship. The 51 teachers belong to 21 different high schools (including public, private and state-funded). The University of Oviedo and the Regional Ministry of Education, within its program of Educational Research and Innovation, approved this research. This program implies that all results obtained can be used for research purposes, unless explicit disagreement from any involved agent and, hence, the procedure does not require the written consent, since all the research programs are publicized and approved by a research committee and another academic committee. Thus, teachers were informed about the use of the data and they gave oral consent to it.

Table 1. Number of PAU exams and exercises analysed.

Region	Number of exams	Number of exercises
Andalusia	60	240
Asturias	40	160
Madrid	34	148
Baque Country	20	80
TOTAL	154	628

Instruments

Considering curricular contents, and assessment criteria defined by the Ministry and completed by the regional ministries, three observation tables have been designed in order to analyse content blocks in AMSS2 curriculum: Algebra, Calculus, and Probability & Statistics.

To register the information in an interpretable way, a new tool named Curricular Unit (CU, in the following) has been introduced. CU's are defined as basic observation units for contents, procedures and assessment criteria in the curriculum. CU's are constructed as an *ad hoc* simplification for our problem of the so-called conceptual focuses defined in [33]. CU's have been designed by considering both so-called 'contents' and 'assessment criteria' paragraphs in the official curriculum. Therefore, CU's arrange contents, procedures and assessment criteria into homogeneous curricular structures, allowing a coherent information retrieval and producing meaningful results about the frequency of appearance of each CU. Thus, CU's not only are the tool to systematize the official curricula, but they also allow the observation of appearance frequency in each PAU exam, making it feasible to analyse the whole curriculum. This frequency observation is combined with a qualitative assessment of the type of problems and exercises posed in PAU exams, specifically, whether they propose solving contextualised problems and related with reality or not. Tables 2–4 show the defined CU's for each block and its description.

This classification of the curriculum into CU's is, from authors' point of view, an efficient tool that allows analysing wide curricula, as it is in the case of AMSS2, and a huge number of exams, as presented here.

With respect to the second goal of this paper, to assess teachers' practices one questionnaire has been designed, on the basis of the theoretical framework described above, and following Pajares' statement: "beliefs cannot be directly observed or measured but must be inferred from what people say, intend, and do–fundamental prerequisites that educational researchers have seldom followed" ([47], p. 207). Therefore, questions ask about teachers' practices regarding the teaching methodologies and their relation to PAU exams, the existence of practising tests, the influence of PAU exams on the real curriculum, the type of exercises and problems developed in the classroom, etc. The questionnaire consists of three thematic groups (questions appear in a later section)

- Group 1: Likert-type questions about the influence of PAU exams in their working methodology, selection of topics and assessment methods according to the subject AMSS2. Answers are considered being 1 = Totally disagree, 2 = Disagree, 3 = Neither agree nor disagree, 4 = Agree and 5 = Totally agree.

Table 2. CU's from Algebra block

A1: Matrices	Matrix language. Compile information using matrices.
A2: Matrices Operations	Basic operations. Solving easy linear equations using inverse matrix method
A3: Matrices Problems	Expressing natural language into matrix language.
A4: Range and determinant	Calculation and interpretation of matrix 'range. Calculation and study of determinant´s properties. Relationship between range and determinant.
A5: System of equations	Solving matrix systems. Studying the number of solutions. Compatibility. Gauss' and Cramer's methods.
A6: Inequations	Linear inequations. Linear inequation systems. Graphic interpretation.
A7: Linear programming	Two-dimensional problems. Expressing natural language into linear programming problems. Feasible and optimal solutions.
A8: Applications of linear programming	Applications in solving social, economic and demographic problems.

Table 3. CU's from Calculus block.

C1: Limits	Definition. Relationship between limit and tendency.Graphical interpretation of limits and asymptotic tendency with different kind of functions.
C2: Calculating limits	Rational, irrational, exponential and logarithmic functions.
C3: Real phenomena	Using functions to analyse and interpret real situations in Social Sciences.
C4: Continuity	Definition. Types of continuity. Continuity of polynomial, rational, exponential, logarithmic and piece-wise functions.
C5: Derivative	Definition. Interpretation. Calculus of the tangent line.
C6: Calculating derivative	Derivatives of polynomial, rational, exponential and logarithmic functions (maximum of two combinations). Using derivatives to study function´s local properties.
C7: Optimization	Using derivatives in optimization problems related to Social Sciences and economy.
C8: Studying functions	Studying local and global properties of polynomial and rational functions. Critical analysis of the information.
C9: Calculating integrals	Primitive function. Immediate integrals. Integration methods: integration by parts and by substitution.
C10: Definite integrals	Relationship between definite integral and primitive integral. Calculating easy surfaces and area under a curve by using integrals. Barrow´s rule.

- Group 2: Open-ended questions about contents, competencies and suggestions and ideas to improve current PAU exams.
- Group 3: Questions about personal and professional data.

Procedure and analysis

Data from the CU's have been collected from PAU exams. Each exercise (out of the total 628) has been assigned to one or several CU's. The assignment procedure consists of identifying the CU of every exercise in the whole set of exams.

Regarding the questionnaire about teachers' practice, it was distributed in paper format among mathematics teachers belonging to the Departments of Mathematics at High Schools and having experience in teaching AMSS2. The questionnaire was delivered by the students of

Table 4. CU's from Probability & Statistics block.

S1: Probability	Probability of simple and compound events. Laplace´s rule.
S2: Independent events.	Definitions and rules. Probability of dependent and independent events.
S3: Conditional probability	Definitions and contingency table. Probability of conditional events.
S4: Bayes and Law of total probability	Law of total probability, Bayes´ rule and tree diagram. Posterior probability.
S5: Making decisions with probabilities	Making decisions in Social Sciences problems involving probabilities
S6: Central limit theorem (CLT)	Central limit theorem (CLT). The normal approximation to binomial distribution Law of large numbers. Identifying the normal distribution. Determining the kind of distribution.
S7: Sampling	Choosing the best sample. Studying the representativeness of the sample.
S8: Confidence intervals	Probability distributions: mean and proportion. Confidence intervals for mean, proportion and difference between means.
S9: Hypothesis testing	Hypothesis testing for mean, proportion and difference of means. Studying significant differences between means or proportions of two populations.
S10: Inferring conclusions	Making decisions and obtaining conclusions about different situations.
S11: Distributions	Binomial and Normal distributions, probabilities associated to them.

the Master's Degree in Teaching Training of the University of Oviedo, during their internship period at high schools.

The information collected from the questionnaire has been treated with statistic package R, applying the following analyses:

- Descriptive analyses of each one of the three parts of the questionnaire.
- Quantitative studies about the possible relationship between data from the teachers (group 3) and their answers to the rest of questions (groups 1 and 2). It has been made through different non-parametric test (depending on the characteristics of the analysed variables), as there is no normality in the data, as it will be explained in the correspondent section of results. Correlations among different answers have been also considered.

Results

Curricular units of AMSS2 in PAU exams

The following tables show the percentage of appearance of each CU in the exams, with respect to the total amount of exams in the considered region. It should be noted that within each exam several CU's appear, therefore the different percentages could sum up more than 100.

Table 5 shows clear differences in frequencies of Algebra CU's. For instance, CU A3 (Problems with matrices) appears no more than 10% in three regions but in Asturias it reaches 42.5%. CU A4 only appears in Madrid (20.6% of cases there), whereas in the rest of the exams to obtain the matrix rank is never explicitly posed, neither the meaning of the rank nor its relationship with the determinant.

Despite national AMSS2 curriculum include the specific assessment criteria: 'To transcribe problems expressed in common language into algebraic language' ([3], p.45476), it has been checked that only in the case of Asturias an important number of problems about linear equation system are set out in real contexts (CU A3), whereas in Madrid, Andalusia and the Basque Country most of the exercises are expressed without any context.

This situation appears partially replicated in the case of Linear Programming (CU's A7 and A8). In Asturias all exercises deal with solving social, economic and demographic problems, however, in the rest of regions exercises provide the inequations, asking for their representation and for calculating the maximum or minimum of a given function, but most of the cases, mathematical formulation is not contextualised within a real problem, it is neither used the proper terminology of linear programming (objective function, feasible region, optimum solution, etc.).

Calculus block has less homogeneity, being the most frequent CU's C5, C6 and C8, as it can be seen in Table 6.

Analysing by regions, in Andalusia it was never included any exercise related to integral calculus (CU C9), but in all exams derivate calculus is present (CU's C5 and C6). Moreover, only 17 out of the 60 exercises have a formulation related to real life situations (CU C2). The rest

Table 5. Algebra: Appearance percentage of CU's.

	A1	A2	A3	A4	A5	A6	A7	A8
ANDALUSIA	48.3	45.0	6.7	0	30.0	43.3	33.3	18.3
ASTURIAS	57.5	32.5	42.5	0	72.5	47.5	77.5	77.5
MADRID	44.1	32.4	5.9	20.6	67.6	20.6	38.2	23.5
BASQUE COUNTRY	30.0	40.0	10.0	0	10.0	40.0	50.0	10.0

Table 6. Calculus: Appearance percentage of CU's

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
ANDALUSIA	18.3	28.3	30.0	31.7	83.3	90.0	31.7	46.7	0	0
ASTURIAS	27.5	32.5	40.0	22.5	30.0	55.0	45.0	85.0	50.0	50.0
MADRID	38.2	47.1	14.7	35.5	73.5	82.4	14.7	64.7	47.1	70.6
BASQUE COUNTRY	5.0	10.0	30.0	15.0	45.0	90.0	50.0	80.0	0	35.0

are expressed mathematically without any context, and it is required to study some characteristics of the function (continuity, derivability, trends, etc.). In almost half of the exercises (24 out 60), the function is defined as a piecewise one.

In Asturias, there is an alternation in every exam between an exercise dealing with differential calculus (CU's C5 and C6) and other dealing with integral calculus (CU's C9 and C10). The first one is formulated in real life context (C7), whereas the integral calculus exercise follows always the same structure: calculating the primitive function and calculating the area under the curve, through Barrow's theorem. The functions are usually polynomial.

Regarding Madrid, 22 out of 34 exercises combine differential and integral calculus, but the function is never defined in a real context.

Finally, the Basque Country does not include exercises related to integral calculus (CU C9), however most of the analysed exams present problems about differential calculus (CU's C5, C6 and C7) and the study of functions (CU C8). Social phenomenon analysis only appears in 8 out of the 20 analysed cases.

Probability & Statistics block is the most heterogeneous one, as it is shown in Table 7, being CU's S3 (Conditional Probability) and S4 (Bayes) the most frequent units. Several CU have barely appeared in the exams (CU S6: Central Limit Theorem practical applications, never appeared). The reason could be related to the difficulty to deal with Central Limit Theorem, except when considering it for calculating confidence intervals and hypothesis testing in great samples.

All exercises in Asturias and the Basque Country are related to real life problems that need to be translated into statistic or probabilistic language. There are frequent themes such as alcohol consumption, loan granting, and health and work safety. However, in Madrid and Andalusia these contextualized problems are combined with others out any type of context.

In Asturias most exercises are focused on CU's S3, S4, S9 and S10. But, there is an important number of CU never appearing: S2, S5, S6, S7, S8 and S11.

In Madrid, exercises focused on CU's S7, S8, and S3 are the most frequent, the rest of CU's having frequencies under 40%. In few exercises students have to make decisions about different probabilities or probabilistic scenarios, despite decision-making is one of the basic competencies in AMSS2. On the other hand, hypothesis testing is never posed.

In Andalusia confidence intervals are much less frequent than in Madrid, but, on the other hand, they include exercises about hypothesis and sampling, sampling distribution of the mean or sample representativeness.

Table 7. Probability & Statistics: Appearance percentage of CU's

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
ANDALUSIA	40.0	21.7	71.7	45.0	8.3	0	56.7	46.7	38.3	41.7	0
ASTURIAS	5.0	0	60.0	50.0	0	0	0	0	75.0	75.0	0
MADRID	29.4	14.7	52.5	37.5	5.9	0	85.3	82.4	0	0	11.8
BASQUE COUNTRY	35.0	10.0	10.0	50.0	10.0	0	0	35.0	15.0	0	45.0

In the Basque Country tests problems related to normal distribution usually appear in exams (CU S11) and problems related to total probability law and Bayes' Theorem (CU's S1 and S4). It is also easy to find exercises about calculus of probabilities through product rule, without further reflection about the sense of the operation.

Influence of PAU on AMSS2 teachers' practices

A questionnaire conducted on mathematics teachers in Secondary Schools in Asturias was performed. In this section the most important obtained results are presented.

Teachers' personal and professional data. The questionnaire included a list of questions about social and demographic issues, besides other professional questions. The obtained results are listed below (N = 51):

- Sex: 41.2% of the interviewee were man, 54.9% woman and the 3.9% resting did not specify the sex.
- Age Rank: only one interviewee is between 22 and 35 years old, 19 are between 35 and 50 years old, 29 are over 50 years old and the two remaining have not answered to this question. It clearly reflects that Asturian teachers are quite aged.
- PAU experience: only 4 people have participated in PAU exams, as graders (members of the committee), since 2010.
- 84.3% of the teachers have taught AMSS2. Besides this percentage, 94.1% of teachers in the sample have also taught lessons in the first year of Baccalaureate (AMSS1).
- Regarding the experience teaching AMSS2, 9 people have never give lessons in AMSS2, 19 people have less than 5 years of experience, 12 people have between 5 and 15 years and the 10 remaining people have more than 15 years of experience. One of the teachers has not answered to this question.

Data from Likert-type questions. As it was specified in the description of the questionnaire, a second block contained 17 questions related to teaching methodology and its relationship with PAU, considered in a Likert-type scale from 1 to 5. Some questions are posed in a direct-affirmative way and some other in a negative way to avoid mechanical answering. Results are shown in Table 8.

Results from open questions. The interviewees have given only four answers in this part of the questionnaire:

- In PAU exams students are not assessed on information and communication technologies.
- Students have a narrow basis from compulsory Secondary Education (previous stage to Baccalaureate) in Probability & Statistics.
- PAU exercises are very repetitive, more open problems should be included, instead of solving repeatedly procedural exercises.
- There is a lack of time to develop the official curriculum due to its extension, and this fact affects methodology, avoiding the development of cooperative learning method or researching techniques.

Results from quantitative data analysis. In order to complete the analysis, it have been analysed the existence of statistical relationship among demographic and professional data of

Table 8. Answers to Likert-type questions in the questionnaire for teachers (N = 51).

Question	Median	Mean	SD
1. I do not teach some of the syllabus contents because they usually do not appear in PAU.	1	2.02	1.34
2. I pay more attention to questions that are usually asked in PAU.	4	3.88	1.09
3. I do not check the annual guidelines established by PAU exams coordinators.	1	1.50	0.91
4. I give up questions that are not usually asked in PAU when I do not have time enough to teach the whole syllabus.	4	3.76	1.23
5. I use the same methodology in both Baccalaureate courses.	3	3.59	0.95
6. I prefer teaching all contents within the lectures, instead of using more skill-oriented teaching methods in the second year of Baccalaureate.	3	2.88	1.08
7. PAU exercises determine my methodology in the classroom.	4	3.28	1.16
8. I usually do not use active methodologies in the second course of Baccalaureate.	3	2.67	0.97
9. In order to prepare my lectures I do not carry out an analysis of PAU exercises from previous years.	1	1.63	0.83
10. I consider exercises that usually appear in PAU exams do not influence my students´ learning.	2	2.57	1.31
11. Exercises that I solve in my lectures are usually similar to real problems.	3	3.13	0.91
12. In my exams I use exercises similar to PAU exercises.	4	4.06	0.92
13. I do not make any simulacrum about PAU exams.	2	2.39	1.31
14. If PAU exams did not exist, my students' learning would fit more properly to curriculum	3	2.86	1.06
15. PAU exams do not produce a significant lack of competence in the academic training of students.	4	3.86	1.05
16. I would propose a different kind of exercises in PAU exams.	3	3.12	0.91
17. From my experience, PAU results in AMSS2 are similar to students' results in Baccalaureate.	4	3.76	1.05

the teachers, their answers to the Likert-type questions from the first block and the open ended questions from the third block.

For every Likert-type question and its possible relationship with demographic and professional variables, Kruskal-Wallis non-parametric test was used, as it was previously checked that answers were not normally distributed. This test allows determining if the differences between the different age ranks or years of teaching experience influence the answers of the Likert-type questions.

Results from Kruskal-Wallis test show in all cases that there is no significant relationship between any of the answers to Likert-type questions and the age, or between any of them and the years of teaching in the considered subject. Table 9 shows respective p-values in its last three columns.

Regarding sex variable and each of the Likert-type questions, non-parametric Wilcoxon test used, to detect differences between medians by sex, p-values are shown in the first column of Table 9. It was not found in any case significant relationship among the variables. Therefore, answers to Likert-type questions do not depend on the sex, age or professional experience of the teachers. There is only a p-value that is slightly under 0.05, which will be discussed later.

Relationship between open questions and demographic and professional variables was analysed through Fisher's (in the case of sex) and Barnard's tests (for the rest of variables). It was chosen these kinds of non-parametric tests because results did not have the minimum number of values needed to apply $\chi 2$ test, and regrouping them would lead to nonsense. Obtained p-values are shown in Table 10. Again, results underline the consistency of teachers' answers, which do not depend on the sex, age or years of professional experience.

Table 9. Obtained p-values for tests on Likert-type answers (Wilcoxon test in the first column and Kruskal-Wallis test in the rest).

P	Sex	Age	Teaching 1˚ Baccalaureate	Teaching AMSS2
Question 1	0.94	0.17	0.68	0.20
Question 2	0.89	0.36	0.89	0.51
Question 3	0.21	0.80	0.44	1.00
Question 4	0.84	0.48	0.34	0.24
Question 5	0.92	0.58	0.50	0.69
Question 6	0.52	0.92	0.37	0.82
Question 7	0.50	0.04	0.41	0.28
Question 8	0.48	0.18	0.19	0.50
Question 9	0.85	0.47	1.00	0.27
Question 10	0.09	0.28	0.31	0.52
Question 11	1.00	0.16	0.90	0.98
Question 12	0.23	0.27	0.65	0.33
Question 13	0.54	0.46	0.68	0.18
Question 14	0.58	0.23	0.84	0.38
Question 15	0.11	0.53	0.89	0.85
Question 16	0.54	0.99	0.17	0.75
Question 17	0.57	0.96	0.86	0.41

Additionally, it was made an analysis of the correlations between the answers to Likert-type questions, in order to detect behavioural common patterns. In Table 11, correlation coefficients are shown between the answers to every pair of Likert-type questions. As it can be observed, despite being a sample of relatively small size, it appears significant correlations (over 35%) of direct relation among questions Q1-Q14, Q2-Q4, Q2-Q7, Q2-Q12, Q4-Q7, Q5-Q6, Q5-Q10, Q5-Q13, Q7-Q8, Q7-Q12, Q12-Q15, and Q15-Q17. It also appears important inverse correlations (under –30%) between the questions Q2-Q13, Q3-Q17, Q4-Q10, Q4-Q13, Q7-Q10, Q9-Q17, Q9-Q15, and Q12-Q13. In the discussion section the interpretation of these values will be included with more details.

Analysis of correlations together with answers to Likert-type questions lead to classify questions into several subgroups:

- Considering Q1 and Q14 it is concluded that near 60% of teachers deny leaving parts of the curriculum without teaching when they are not usually asked in PAU exams (Q1), but, on the other hand, around 50% of them acknowledge that, if PAU did not exist, their teaching would be closer to the curriculum (Q14).
- However, more than 70% of the teachers do not deny that in case of being force to renounce to some parts of the curriculum, they would supress the questions that appear less in PAU

Table 10. Obtained p-values from tests on answers to open questions (Fisher test in the first column and Barnard test in the rest).

P	Sex	Age	Teaching 1˚ Baccalaureate	Teaching AMSS2
Open Answer 1	0.31	0.27	0.56	0.31
Open Answer 2	0.14	0.40	0.44	0.46
Open Answer 3	0.29	0,79	0.45	0.18
Open Answer 4	0.50	0.85	0.32	0.54
Open Answer 5	0.75	0.70	0.30	0.30
Open Answer 6	1.00	1.00	0.28	0.52

Table 11. Correlation coefficients among answers to Likert-type questions (in percentage).

Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	
22	-5	30	8	26	29	-8	2	-6	-15	16	-10	35	1	-1	9	Q1
	-3	54	21	13	48	30	0	-31	-7	42	-33	1,8	26	-25	-6	Q2
		-7	20	-28	-3	15	0	22	-13	6	14	20	-14	21	-35	Q3
			-10	-14	36	17	0	-48	-19	27	-34	27	11	4	-4	Q4
				39	-2	22	0	35	-19	8	38	2	16	-11	-20	Q5
					22	30	-3	-9	-26	5	29	18	-12	-13	13	Q6
						42	0	-41	1	54	-25	21	15	-10	-3	Q7
							0	-11	-25	25	4	-11	8	-18	14	Q8
								26	-4	-51	31	-15	-33	-6	-42	Q9
									6	-17	18	-12	25	-1	16	Q10
										27	-15	-20	32	-20	13	Q11
											-43	8	38	-14	21	Q12
												-11	-14	11	-18	Q13
													-27	17	-29	Q14
														-16	35	Q15
															-7	Q16

exams (Q4). Moreover, Q4 has a strong positive correlation with Q2 ('I pay more attention to questions that are usually asked in PAU') and Q7 ('PAU exercises determine my methodology in the classroom'). Additionally, Q7 is positively correlated with Q2, Q8 ('I usually do not use active methodologies in the second course of Baccalaureate') and Q12 ('In my exams I use exercises similar to PAU exercises'). This shows how teachers try to teach the entire curriculum, but in case of lack of time, PAU contents are the chosen ones, keeping a methodology and a kind of exercises similar to PAU exams. Therefore, PAU causes a washback on the official curriculum.

- The inverse correlation between Q2, Q4 and Q12 with Q13 gives consistency to the described model, as this question is posed in a negative way. Therefore, there is a strong relationship between PAU and the day-by-day work in the classroom, establishing a quasi-logical relationship [48] between its preparation and the activity in the classroom.
- Despite previous answers, Q10 offers a not so clear result: about 25% of the interviewees believe that PAU does not influence students' learning. Moreover, this question is negatively correlated with Q4 and Q7. This fact is analysed as the reluctance of teachers to admit a clear influence of PAU through a direct asked question, whereas when it is asked indirectly (Q2, Q3, Q5, Q7, Q9, Q12 or Q13) it is admitted with greater clarity. These answers show an internal conflict between their practices and beliefs, underlying that teachers are sensible in the way they enact their beliefs [49,50]. Teachers are also considering that their students usually perform in AMSS2 exams similarly to Baccalaureate, which is reinforced by Q17, showing a deep teachers' support to the validity of the PAU exams and its prediction capacity with respect to grades obtained by students in the Baccalaureate [23].
- Something similar occurs with Q15, despite previous answers related to PAU influence, more than two thirds of the interviewees affirm that PAU does not cause significant distortion in students' training. Thus, teachers consider that contents than can be avoided are not important in students' training, showing a clear ratification of the coherence of the test.

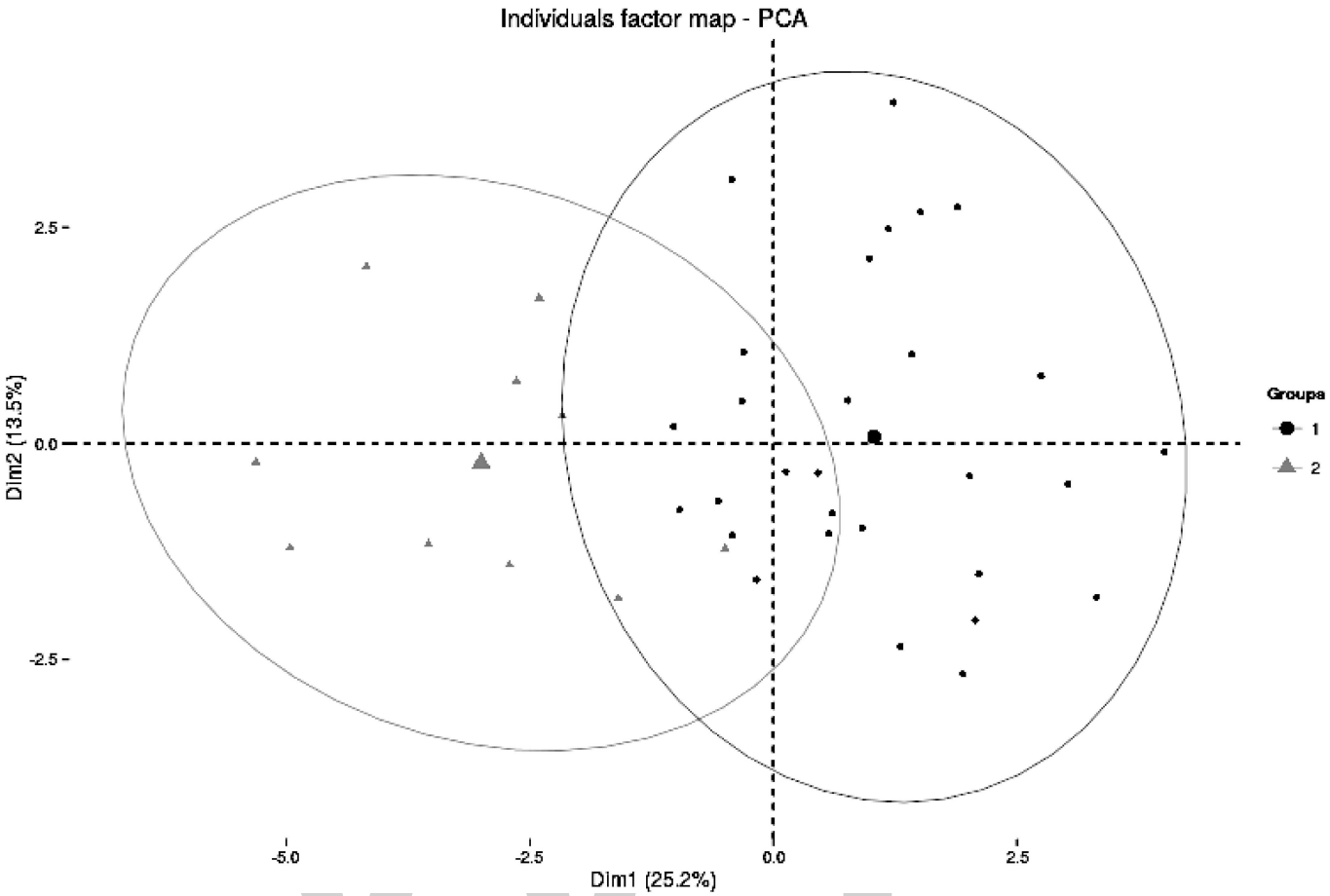

Fig 1. Cluster analysis. Clusters plotted over the two main dimensions in scores (biggest points represent group average)

- Finally, there are five questions (Q6, Q8, Q11, Q14 and Q16) where answers 'Nor agree or disagree' prevail. These are questions about active teaching-learning methodologies, competencies acquisition or using real problems that do not seem to raise a clear opinion among teachers. In the case of Q14, it is observed a small prevalence of the answer 'Disagree', which seems to highlight what has been underlined in the previous point. Q16 shows that most of the teachers do not express their opinion about changing the current PAU exam or maintaining it.

Additionally, a cluster analysis has been performed over data from Likert-type questions. Results show how two main clusters can be distinguished. In Fig 1 the two main dimensions in scores are used as axes to plot the teachers' scores showing both clusters. Nine individuals compose first cluster whereas the rest of the sample (42 members) belong to the second one. Main differences between the two clusters yield on scores in questions Q1, Q2, Q4, Q7, and Q12, with notably lesser scores for cluster 1, and questions Q9, Q10, and Q13, with higher scores for cluster 2 (maximum difference being 1.19 points in Q4). This fact shows that teachers in cluster 1 have a lower agreement degree with questions supporting the influence of PAU in their teaching methodology, whereas they have a higher agreement degree with questions denying that influence. Therefore, we could look at group 1 as teachers acknowledging less influence of PAU on their practice than those in group 2.

Discussion

The first research hypothesis consisted on checking how PAU exams represent the official curriculum and on detecting possible biases on questions posed in PAU exams. Thanks to the study of all the exams released in the four considered regions it can be stated that although there are significant differences between the exams from the regions, there exist substantial parts of the official curriculum that are omitted or that appear underrepresented in all PAU exams. It has especial relevance that several CU's have never appeared in the exams of any region, what implies a clear bias. On the other hand, there are some CU's that have a constant presence in all analysed exams. Thus, it is demonstrated that PAU influences AMSS2 by producing a narrowed curriculum. This statement is consistent with conclusions in [27], but the present study is based on the study of all the released exams in four regions, being the first one doing this in the literature about this topic.

Another question that has been observed in the exercises analysis is the repetitive structure from one year to another within each region. This fact could imply positive effects, as it allows students organising and planning their learning, at this point, Wall [51] pointed out that "It should not be assumed that a 'good' test will automatically produce good effects in the classroom, or that a 'bad' test will necessarily produce negative ones" ([51], p.505-506). But this repetitive structure also has the negative side due to the high predictability of the exam, so students could limit their learning process only to the solving methodology of this kind of exercises with a scarce deepening: only learning to test. This fact opposes the understanding and the analysis of real life situations, as it is stated in the official curriculum. This result is consistent with [23] analysis for the particular case of the defined integral in PAU exams, but also with other research in the general field of high-stake testing. Moreover, this result endorses [10] into the deskilling processes of teachers, particularly math teachers with problem solving. But, on the other hand, PAU exams designing process is not including critical thinking as in Alberta exams ([10], p.173). Subsequently, it is necessary a proper alignment [52] between new curricular standards oriented to problem solving in real contexts and their assessment in central examinations, despite some tasks as critical thinking or decision making in mathematical frameworks have been underlined as very difficult to be evaluated in such type of exams [53]. This result is also consistent with [28], where a lack of contextualized probability problems was pointed out.

Additionally, assuming the lack of open problems and the pre-eminence of algorithmic exercises, there are few exercises including a context of close situations for the students, nor exercises describing phenomenon related to Social Sciences, despite AMSS2 official curriculum considers that a clear priority. Proposed exercises should stimulate reasoning (Jones et al. [11]), searching a solution and making a decision for the mathematical problem. Nevertheless, it should be underlined that PAU exams are limited in time to 1.5 hours, which could hinder to include open problems, which usually are considered without time limitation. This fact makes clear the crucial point about whether this type of assessment is the most adequate to such a curriculum [54]. This is also consistent with Kuhn [55] that underlines the difficulties of introducing context-based tasks in central examinations.

The management of the mathematical language used to model these situations and to express the solution of an exercise constitutes another important skill for the acquisition of a real mathematical competence. In this paper it is demonstrated that, mainly the Statistics & Probability block, some exercises needing translation from verbal to math language appear. This fact supposes an advanced in the direction of a greater didactic suitability of PAU exams, which was suggested in [26]. Nevertheless, research about difficulties and mistakes made by students in certain topics must be taken into account to design this curriculum-based

assessment, for instance, in Statistics [29] points out the high difficulty of hypothesis testing for Baccalaureate students.

Assessment is a crucial point in the teaching-learning process, as it can condition learning processes and, moreover, it requires teachers' adaptation to assess competencies. If extensive and general assessment procedures, as it is PAU, do not reflect this paradigm change from content to competence assessment, teachers' role will be reduced.

The second research hypothesis was to check whether PAU influences AMSS2 teachers' practices and to determine at what extent the real curriculum resulted from this behaviour. The research confirms that, despite there is not any explicit acknowledgment of it, PAU exams produce an influence on teachers' practices in their day-by-day, and, what is more important regarding this paper, they confirm the washback on curriculum. This conclusion is derived from the analysis of the questionnaires.

From author's point of view, results derived from answers to the questionnaire are noticeably different from a similar study in other region [27]. In both cases, teachers point out that their work is not conditioned by PAU, but in the questionnaires employed for the present paper it can be clearly remarked a greater utilitarian use of the teaching-learning processes in the second year of Baccalaureate towards the preparation of PAU exams. Besides, the present work contributes with the novelty of analysing a curriculum that has a defined orientation to the use in real life or Social Sciences frameworks of mathematics and, therefore, it is needed to pay more attention to the notion of mathematical competence, understanding mathematics as a tool to solve Social Science problems. Actually, this reinforces the results obtained in [35], establishing that practice test do not improve the performance, on the contrary, reflexive homework tasks increase it.

Besides, statistical analysis backups the homogeneity of answers among teachers, as there are no significant differences due to the age, sex or the years of professional experience in the subject. Therefore, perception respect to the importance of PAU are strongly consistent among teachers and, thus, results make clear PAU produces a washback in AMSS2 curriculum, not only on contents but also on methodology. Only few teachers answer with certain different responses, as it is demonstrated by the performed cluster analysis, so that, a small group can be distinguished from the rest by a lower degree of disagreement about washback of PAU. Authors are convinced this is a field that really needs an intervention, in order to reinforce teachers' beliefs and about the importance of their practices, as [56] confirmed in the case of Finnish teachers, there is a strong relationship among mathematics teachers between their beliefs and their teaching practices. If a correct alignment is attained between new curriculum and exams, teachers' practices will produce better effects on students' self-regulated learning, as it was pointed out in [57].

Moreover, results open new ways for future research about teachers' practices regarding the new final Baccalaureate exam that will substitute PAU in 2017 and, especially, the outcomes in this paper highlight some improvements that can be considered in designing the new exam. This is the moment to attain a proper alignment, being more faithful to innovations in the curriculum of AMSS2, especially those devoted to solve real (or likely real) problems, to use contextualized mathematics in the field of Social Sciences. It is also important to take into account the results to design a new much more balanced exam, not focused mainly on certain curricular units. Looking at the effect on teachers' practices, results also point out the need of working into more varied type of exams, enhancing teachers' flexibility to prepare and to manage the teaching/learning process, and releasing them from learning-to-test.

Finally, it is necessary to remark two main limitations in the present study. First, it would be recommended to widen the study of the exams to other regions (although the chosen sample considers four regions that suppose an important percentage of the scholar population in

Spain). Second, the sample of teachers that have answered the questionnaire could also be widened to other regions, and be selected by a random sampling.

Acknowledgments

Authors express their gratitude to Tomás Ortega from University of Valladolid, María Gea from University of Granada, and Paloma G. Castro from University of Oviedo for their valuable comments to improve the work.

Author Contributions

Conceptualization: LJRM PD VM PA.

Formal analysis: LJRM PD VM PA.

Investigation: LJRM PD VM PA.

Methodology: LJRM PD VM PA.

Project administration: LJRM PD VM PA.

Supervision: LJRM PD VM PA.

Validation: LJRM PD VM PA.

Visualization: LJRM PD VM PA.

Writing – original draft: LJRM PD VM PA.

Writing – review & editing: LJRM PD VM PA.

References

1. Datos y cifras del sistema universitario español. Curso 2014–2015 [Internet]. Ministerio de Educación, Cultura y Deportes [cited 2015 May]. Available from: http://www.mecd.gob.es/dms/mecd/educacion-mecd/areas-educacion/universidades/estadisticas-informes/estadisticas-informes-documentum/datos-cifras/2012-2013-datos-y-cifras-sistema-universitario-espanol.pdf
2. Jefatura del Estado. Ley Orgánica 2/2006, de 3 de mayo, de Educación. Boletín Oficial del Estado 2006 May 4; 106:17158–17207.
3. Ministerio de Educación y Ciencia. Real Decreto 1467/2007, de 2 de noviembre, por el que se establece la estructura del Bachillerato y se fijan sus enseñanzas mínimas. Boletín Oficial del Estado 2007 Nov 6; 266:45381–45477.
4. Ministerio de la Presidencia. Real Decreto 1892/2008, de 14 de noviembre, por el que se regulan las condiciones para el acceso a las enseñanzas universitarias oficiales de grado y los procedimientos de admisión a las universidades públicas españolas. Boletín Oficial del Estado 2008 Nov 24; 283:46932–46946.
5. Klein ED, van Ackeren I. Challenges and problems for research in the field of statewide exams. A stock taking of differing procedures and standardization levels. Studies in Educational Evaluation 2011; 37:180–188.
6. Popham WJ. The merits of measurement-driven instruction. Phi Delta Kappan 1987; 68:679–682.
7. Smith ML. The effects of external testing on teachers. Educational Researcher 1991; 20(5):8–11.
8. Apple MW. Education and power. New York: Ark Paperbacks; 1982.
9. Apple MW. *Teachers and texts*: *A political economy of class and gender relations in education*. New York: Routledge; 1986.

10. Runté R. The impact of centralized examinations on teacher professionalism. Canadian Journal of Education 1998; 23(2):166–181.
11. Jones MG, Jones BD, Hardin B, Chapman L, Yarbrough T, Davis M. The impact of high-stakes testing on teachers and students in North Carolina. The Phi Delta Kappan 1999; 81(3):199–203.
12. Bishop J. The effect of national standards and curriculum-based exams on achievement. The American Economic Review 1997; 87(2):260–264.
13. Häkkinen I. Do university entrance exams predict academic achievement? Working paper, Department of Economics, Uppsala University, no. 2004:16.
14. Ou D. To leave or not to leave? A regression discontinuity analysis of the impact of failing the high school exit exam. Economics of Education Review 2010; 29:171–186.
15. Jacob BA. The impact of High School Graduation Exams. Educational Evaluation and Policy Analysis 2001; 23(2):99–121.
16. Bergquist E. Types of reasoning required in university exams in mathematics. The Journal of Mathematical Behavior 2007; 26(4):348–370.
17. Luk HS. The gap between secondary school and university mathematics. International Journal of Mathematical Education in Science and Technology 2007; 36(2–3):161–174.
18. Cross DI. Alignment, cohesion, and change: examining mathematics teachers' belief structures and their influence on instructional practices. Journal of Mathematics Teacher Education 2009; 12:325–346.
19. Perrenoud P. La fabrication de l'excellence scolaire: du curriculum aux pratiques d'évaluation. Genève: Droz; 1984.
20. Muñoz Repiso M, Murillo FJ. La selectividad a examen: estudio comparativo del acceso a la universidad en algunos países de Europa. Cuadernos de Pedagogía 1999; 282:91–97.
21. Muñoz Repiso M, Murillo FJ. Los resultados en la selectividad actual: algunas cuestiones a debate. Revista de Educación 1997: 314:29–48.
22. Cuxart A, Martí Recober M, Ferer Juliá F. Algunos factores que inciden en el rendimiento y la evaluación en los alumnos de las pruebas de aptitud de acceso a la universidad (PAAU). Revista de Educación 1997; 314:63–88.
23. Gaviria Soto JL. La equiparación del expediente de Bachillerato en el proceso de selección de alumnos para el acceso a la universidad. Revista de Educación 2005; 337:351–387.
24. Pastor J. Las pruebas de matemáticas en los exámenes de acceso. Enseñanza de las Ciencias 1984: 2(1):17–24.
25. Goberna MA, López MA, Pastor J. La influencia del examen de selectividad en la enseñanza (análisis de una experiencia en matemáticas de COU). Enseñanza de las Ciencias 1985; 3(3):181–184.
26. Contreras de la Fuente A, Ordóñez Cañada L, Wilhemli MR. Influencia de las pruebas de acceso a la universidad en la enseñanza de la integral definida en el Bachillerato. Enseñanza de las Ciencias 2010; 28(3):367–384.
27. Ruiz de Gauna J, Dávila P, Etxebarria J, Sarasua JM. Pruebas de selectividad en matemáticas en la UPV-EHU. Resultados y opiniones de los profesores. Revista de Educación 2013; 362:217–246.
28. López-Martín MM, Contreras JM, Carretero M, Serrano L. Análisis de los problemas de probabilidad propuestos en las pruebas de acceso a la Universidad en Andalucía. Avances de Investigación en Educación Matemática 2015; 9:65–84.
29. López-Martín MM, Batanero C, Díaz-Batanero C, Gea MM. La inferencia estadística en las pruebas de acceso a la universidad en Andalucía. Revista Paranaense de Educaçao Matemática 2016; 5(8):33–59.
30. Ball DL, Thames M, Phelps G. Content knowledge for teaching: what makes it special?. Journal of Teacher Education 2008; 59(5):389–407.
31. Beswick K. Teachers' beliefs about school mathematics and mathematicians' mathematics and their relationship to practice. Educational Studies in Mathematics 2012; 79:127–147.
32. Rockoff J. The impact of individual teachers on student achievement: evidence from panel data. American Economic Review 2004; 94:247–252.
33. Clotfelter CT, Ladd HF, Vigdor JL. Teacher credentials and student achievement in high school: a cross-subject analysis with student fixed effects. Journal of Human Resources 2010; 45(3):655–682.
34. Hill H, Rowan B, Ball DL. Effects of teachers' mathematical knowledge for teaching on student achievement. American Educational Research Journal 2005; 42(2):371–406.

35. Zakharov A, Carnoy M, Loyalka P. Which teaching practices improve student performance on high-stakes exams? Evidence from Russia. International Journal of Educational Development 2014; 36:13–21.
36. Bailey K. Working for washback: a review of the washback concept in language testing. Language Testing 1996; 13(3):257–279.
37. Spratt M. Washback and the classroom: the implications for teaching and learning of studies of washback from exams. Language Teaching Research 2005; 9(1):5–29.
38. Damankesh M, Babaii E. The washback effect of Iranian high school final examinations on students' test-taking and test-preparation strategies. Studies in Educational Evaluation 2015; 45:62–69.
39. Neumann M, Trautwein U, Nagy G. Do central examinations lead to greater grading comparability? A study of frame-of-reference effects on the University entrance qualification in Germany. Studies in Educational Evaluation 2011; 37:206–217.
40. Madaus GF. The influence of testing on the curriculum. In Tanner LN, editor. Critical issues in curriculum: eighty-seventh yearbook of the National Society for the Sutdy of Education (Part 1). Chicago: University of Chicago Press; 1988.
41. Rico L, Marín A, Lupiáñez JL, Gómez P. Planificación de las matemáticas escolares en secundaria. El caso de los números naturales. Suma 2008; 58:7–23.
42. Ejercicios de convocatorias anteriores [Internet]. Universidad Complutense de Madrid [cited May 2015]. Available from: https://www.ucm.es/examenes-de-selectividad-2016
43. Exámenes de convocatorias anteriores [Internet]. Universidad del País Vasco [cited May 2015]. Available from: http://goo.gl/2ide0
44. Exámenes de PAU por año [Internet]. Universidad de Oviedo [cited May 2015]. Available from: http://www.uniovi.es/accesoyayudas/estudios/pau/examenes
45. Exámenes y orientaciones sobre Selectividad [Internet]. Junta de Andalucía [cited May 2015]. Available from: http://www.juntadeandalucia.es/innovacioncienciayempresa/sguit/g_b_examenes_anteriores.php
46. Datos y cifras del curso escolar 2015–2016 [Internet]. Ministerio de Educación, Cultura y Deportes [cited 2015 May]. Available from: http://www.mecd.gob.es/servicios-al-ciudadano-mecd/dms/mecd/servicios-al-ciudadano-mecd/estadisticas/educacion/indicadores-publicaciones-sintesis/datos-cifras/Datosycifras1516.pdf
47. Pajares MF. Teachers' beliefs and educational research: cleaning up a messy construction. Review of Educational Research 1992; 62:307–332.
48. Green TF. The activities of teaching. New York: McGraw Hill; 1971.
49. Jaworski B. Investigating mathematics teaching: a constructivist enquiry. London: Falmer Press; 1994.
50. Leatham KR. Viewing mathematics teachers' beliefs as sensible systems. Journal of Mathematics Teacher Education 2006; 9:91–102.
51. Wall D. The impact of high-stakes testing on teaching and learning: can this be predicted or controlled? System 2000; 28:499–509.
52. Webb NL. Criteria for alignment of expectations and assessments in mathematics and science education (Research monograph no. 6). Madison: National Institute for Science Education, Univeristy of Wisconsin-Madison; 1997.
53. Leung KC, Leung FKS, Zuo H. A study of the alignment of learning targets and assessment to generic skills in the new senior secondary mathematics curriculum in Hong Kong. Studies in Educational Evaluation 2014; 43:115–132.
54. Lithner J. Mathematical reasoning in task solving. Educational Studies in Mathematics 2000; 41 (2):165–190.
55. Kühn SM. Exploring the use of statewide exit exams to spread innovation—The example of Context in science tasks from an international comparative perspective. Studies in Educational Evaluation 2011; 37:189–195.
56. Kupari P. Instructional practices and teachers' beliefs in Finnish mathematics education. Studies in Educational Evaluation 2003; 29:243–257.
57. Maag Merki K. Effects of the implementation of state-wide exit exams on students' self-regulated learning. Studies in Educational Evaluation 2011; 37:196–205.

High-performance fractional order terminal sliding mode control strategy for DC-DC Buck converter

Jianlin Wang[1,2], Dan Xu[1]*, Huan Zhou[1], Anning Bai[2], Wei Lu[1]

1 Department of Mechanical Electrical Engineering, Xi'an Jiaotong University, Xi'an, ShanXi, China,
2 Department of College of Science, Ningxia Medical University, Yinchuan, NingXia, China

* xudan@xjtu.edu.cn

Abstract

This paper presents an adaption of the fractional order terminal sliding mode control (AFTSMC) strategy for DC-DC Buck converter. The following strategy aims to design a novel nonlinear sliding surface function, with a double closed-loop structure of voltage and current. This strategy is a fusion of two characteristics: terminal sliding mode control (TSMC) and fractional order calculation (FOC). In addition, the influence of "the controller parameters" on the "performance of double closed-loop system" is investigated. It is observed that the value of terminal power has to be chosen to make a compromise between start-up and transient response of the converter. Therefore the AFTSMC strategy chooses the value of the terminal power adaptively, and this strategy can lead to the appropriate number of fractional order as well. Furthermore, through the fractional order analysis, the system can reach the sliding mode surface in a finite time. And the theoretical considerations are verified by numerical simulation. The performance of the AFTSMC and TSMC strategies is tested by computer simulations. And the comparison simulation results show that the AFTSMC exhibits a considerable improvement in terms of a faster output voltage response during load changes. Moreover, AFTSMC obtains a faster dynamical response, smaller steady-state error rate and lower overshoot.

Editor: Jun Ma, Lanzhou University of Technology, CHINA

Introduction

DC–DC power converters are widely applied for supplying various output voltage in many electric vehicular systems, such as DC motor drives, the hybrid energy storage system (HESS), battery equalization and so on [1]. Our research team is mainly engaged in the research of HESS. The HESS contains two or more power sources connected by DC-DC converters. In order to improve the efficiency and performance of the HESS, the high-performance control strategy for DC-DC converters is needed. But DC-DC converters are inherently non-linear system with chaotic circuit. So the stability of the DC-DC converters is very important [2]. Therefore, the design of high-performance control strategy is usually a challenging issue.

The DC-DC converters in HESS are almost always multiple topologies including several MOSFET switches, so they are workable in both Buck mode and Boost mode. The power

Funding: This work has been funded by the National Natural Science Foundation of China (grant no. 51275379), http://www.nsfc.gov.cn.

Competing interests: The authors have declared that no competing interests exist.

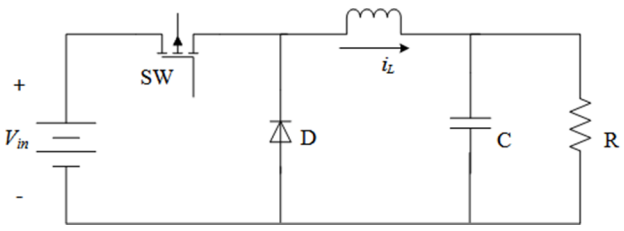

Fig 1. DC-DC Buck converter topology.

consumption of DC-DC converter is an important consideration for the HESS system, when compared with the power consumption of controller [3]. The power consumption of electronic components is a major part of the total power consumption in HESS. Therefore, we created the DC-DC converter structure with simplified design to reduce the power consumption. The control performance of DC-DC converter is studied in this paper. The control strategies of Buck and Boost converters have some similarities, so in this paper we just choose the Buck converter to investigate our novel control strategies.

The sliding mode control (SMC) has many advantages, such as its fast dynamic response, robustness to disturbances, guaranteed stability and simplicity in implementation [4]. There have been a lot of researches on sliding mode control for DC-DC converters. In Ref. [5], Hasan Komurcugil proposed an adaptive terminal sliding mode control strategy for Buck converter, and his sliding surface is a linear one based on linear combination of the system states, using an appropriate time-invariant coefficient. In Ref. [6], Yanmin Wang and her partner designed a double closed-loop structure for DC-DC converter feedback control, and the double closed-loop have a smaller steady-state error than others. In Ref. [7], Junxiao Wang and his partner investigated the performance of the nonlinear disturbance observers with the sliding mode control for Buck converter. In Ref. [8–11], the authors pointed out a fractional order calculation applied in DC-DC converter and control respectively.

In this paper, we focus on the high-performance control strategy for the DC-DC Buck converter, and propose a novel method of fractional order on terminal sliding mode control (FTSMC). Then utilize the method to design a novel nonlinear sliding surface function (based on the double closed-loop structure) that is a fusion of characteristics of TSMC and FOC [12–13].

The rest of the paper is organized as follows: Section 2 introduces the basic principles of the DC-DC Buck converter. Section 3 deals with the design of terminal sliding surfaces for DC-DC Buck converter. Section 4 conducts the design of nonlinear controllers for the Buck converter based on the fractional order calculation and the terminal sliding mode control. Section 5 shows simulation results and the adaptive methods to determine the terminal power parameter values. Section 6 states some conclusions and guidelines for further works.

Modeling the DC-DC Buck converter

The topology of DC-DC Buck converter is shown in Fig 1, and it consists of a DC input voltage source, a MOSFET switch, a diode, an inductor, a capacitor and a load resistor. The average state equations describing the operation of the Buck converter can be written as

$$\frac{di_L}{dt} = \frac{1}{L}(uV_{in} - V_o) \tag{1}$$

$$\frac{dV_o}{dt} = \frac{1}{C}(i_L - \frac{V_o}{R}) \tag{2}$$

where u is the control input that takes 1 for the ON state of the switch, and 0 for the OFF state [14].

Let us define the output voltage error, x_1 is

$$x_1 = V_o - V_{ref} \tag{3}$$

Where V_{ref} is the reference value of the output voltage. By taking the time derivative of (3), x_2 which is the rate of change of voltage error can be expressed as

$$x_2 = \dot{x}_1 = \dot{V}_o - \dot{V}_{ref} \approx \dot{V}_o \tag{4}$$

The state-space model of the Buck converter can be transformed to

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{1}{LC} & -\frac{1}{RC} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{uV_{in} - V_{ref}}{LC} \end{bmatrix} \tag{5}$$

Terminal sliding mode control for the Buck converter

Most of the TSMC strategies are commonly used for the linear sliding surface, which is based on a linear combination of the system states, using an appropriate and time-invariant coefficient [15]. Therefore, the value of coefficient has to be chosen and this is an embarrassment [16–17]. Considering the confusion of this embarrassment, we designed a terminal sliding surface function with a double closed-loop structure of voltage and current. The function S can be defined as

$$S = i_L(t) - k_a {x_1}^{\gamma} - k_b \int_0^t {x_1}^{\gamma} d\tau \tag{6}$$

Where $k_a > 0$, $k_b > 0$, $0 < \gamma < 1$, and they are positive odd integers. When the system is in the terminal sliding mode, it means the Eq (6) is equal to 0 ($S = 0$), assuming

$$y = \int_0^t {x_1}^{\gamma} d\tau \tag{7}$$

The system dynamics can be determined by the following nonlinear differential equation

$$k_a \dot{y} = i_L(t) - k_b y \tag{8}$$

Note that Eq (8) can also be written as

$$dt = \frac{k_a}{i_L - k_b y} dy \tag{9}$$

Taking integral of both side of Eq (9) and evaluating the resulting equation on the closed interval ($x_1(0) \neq 0$, $x_1(t_s) = 0$), the finite time t_s is obtained by

$$t_s = \left| -\frac{k_a}{k_b} \ln\left(i_L - \frac{k_b}{1+\gamma} x_1(0)^{1+\gamma}\right) \right| \tag{10}$$

From Eq (10), it is obvious that the convergence time t_s still depends on the parameters k_a, k_b and γ. Therefore, these parameters must be carefully selected to ensure the desired response.

The sufficient condition for the existence of the terminal sliding mode is given by

$$S\dot{S} < 0 \tag{11}$$

Select the Lyapunov function as

$$V = \frac{1}{2}S^2 \tag{12}$$

The time derivative of Eq (12) can be written as

$$\dot{V} = S\dot{S} \tag{13}$$

In order to satisfy the Lyapunov function, the deviation from the terminal sliding surface S and its time derivative, $\dot{S}$ should be opposite signs in the vicinity of a sliding surface.

When $u = 1$, $S < 0$, so need $\dot{S} > 0$

$$\dot{S} = \frac{1}{L}(V_{in} - V_o) - k_a \cdot \gamma \cdot {x_1}^{\gamma-1} \cdot \frac{1}{C}(i_L - \frac{V_o}{R}) - k_b {x_1}^{\gamma} > 0 \tag{14}$$

When $u = 0$, $S > 0$, so need $\dot{S} < 0$

$$\dot{S} = -\frac{1}{L}V_o - k_a \cdot \gamma \cdot {x_1}^{\gamma-1} \cdot \frac{1}{C}(i_L - \frac{V_o}{R}) - k_b {x_1}^{\gamma} < 0 \tag{15}$$

Make the value of γ, i_L approximately equal to 1 and 0 respectively, the conditions that limit the existence region of the design parameters are obtained as

$$0 < k_a < \frac{CR}{L} \tag{16}$$

$$0 < k_b \leq \frac{k_a}{CR} \tag{17}$$

From Eq (6) and $\dot{S} = 0$, the equivalent control law u_{eq} in this case becomes

$$u_{eq} = \frac{L}{V_{in}}[\frac{k_a \cdot \gamma \cdot {x_1}^{\gamma-1}}{C}(i_L - \frac{V_o}{R}) + k_b \cdot {x_1}^{\gamma} + \frac{V_o}{L}] \tag{18}$$

From expression Eq (18) using the constraint $|u_{eq}| \leq 1$, and considering the aforementioned equilibria conditions, the conditions that limit the existence region of the design parameters are obtained as

$$0 < k_a < \frac{CR}{L}\frac{V_{in} - V_{ref}}{V_{ref}} \tag{19}$$

$$0 < k_b \leq \frac{V_{in}}{LV_{ref}} \tag{20}$$

To solve the inequality Eqs (16), (17), (19) and (20), we should determine the parameters value approximately.

Fractional order terminal sliding mode control for the Buck converter

The terminal sliding surface function is expressed as a fractional order differential equation that is obtained in the form

$$S = i_L(t) - k_a {x_1}^{\gamma} - k_b D_0^{-\lambda} {x_1}^{\gamma} \tag{21}$$

Where $\gamma \in [0,1]$, $\lambda \in [0,1]$, k_a k_b are positive constant. For the Buck converter with FTSMC, the time derivative of Eq (21) can be written as

$$\dot{S} = \dot{i}_L(t) - k_a \cdot \gamma \cdot {x_1}^{\gamma-1} \cdot \dot{x}_1 - k_b D_0^{1-\lambda} {x_1}^{\gamma} \tag{22}$$

Following the procedure of the previous section, the obtained expression for equivalent control is:

$$u_{eq} = \frac{L}{V_{in}} \left[\frac{k_a \cdot \gamma \cdot {x_1}^{\gamma-1}}{C} (i_L - \frac{V_o}{R}) + k_b D_0^{1-\lambda} {x_1}^{\gamma} + \frac{V_o}{L} \right] \tag{23}$$

To obtain the sliding mode dynamics, we insert (21) into (5), and find that the whole closed loop system is in fractional order. Obviously, it is more appropriate to analyze the stability and convergence via the fractional version of Lyapunov by direct method [18–21].

Selecting the Lyapunov function as

$$V = S^2 \tag{24}$$

It follows from the Ref. [22], if 0 is the equilibrium point of system (21) and $x(0) = x_0$, the fractional order derivative of Eq (24), can be written as

$$\begin{aligned} D^{1-\lambda} V &= D^{-\lambda} \dot{V} \leq -KD^{-\lambda} \|x_1\| \\ &= -Kl^{-1} D^{-\lambda} \|S\| \leq -Kl^{-1} \|D^{-\lambda} S\| \\ &= -Kl^{-1} \|x_1\| \end{aligned} \tag{25}$$

Where K is positive constant, l is Lipschitz constant and $l > 0$. So, we can find $V > 0$ and $D^{1-\lambda} V < 0$. In other words, the controlled system satisfies the reaching condition.

When the system reaches the sliding surface, which is $S = 0$, it is in the "terminal sliding" mode. Its dynamics can be determined by the following equation:

$$k_a {x_1}^{\gamma} = i_L(t) - k_b D_0^{-\lambda} {x_1}^{\gamma} \tag{26}$$

We know, several reputed definitions for fractional derivatives are put forward, including Riemann-Liouville definition, Grunwald-Letnikov definition, Caputo definition, Weyl definition, and Marchaud definition [23]. Among them, Riemann-Liouville definition has been well studied. So, we use Riemann-Liouville definition for fractional order differential operation as

$$k_b D_0^{1+\lambda} (D_0^{-1-\lambda} \dot{x}_1^{\gamma}) = D_0^{1+\lambda} (i_L(t) - k_a {x_1}^{\gamma}) \tag{27}$$

Taking fractional integral of both side of Eq (27), the finite time t_s is obtained by

$$t_s = \left| -\frac{k_b \Gamma(\gamma+\lambda)}{k_a \Gamma(\gamma+1)} \ln(i_L - \frac{k_b x_1(0)^{2+\gamma-\lambda}}{(2+\gamma-\lambda)}) \right| \tag{28}$$

Therefore, it can be concluded that system trajectories can reach the equilibrium point in a finite time. When $\lambda = 1$, it is obvious that (28) is equivalent to (10). It means that the finite time taken to attain the equilibrium point of the FTSMC system, is the same as the one of the TSMC system, as given in (10).

Adaptive strategy and simulation results

In order to show the performance of the FTSMC, the DC-DC Buck converter system was subsequently tested by simulations. Simulations are carried out using MATLAB/Simulink. The Simulation framework is shown in S1 Fig, and parameters of Buck converter are given in Table 1.

Table 1. Specifications of Buck converter.

Descriptions	Parameters	Nominal values
Input voltage	V_{in}	25V
Desired output voltage	V_{ref}	10V
Inductance	L	260 uH
Capacitance	C	100 uF
Load resistance	R	1~10Ω

From Table 1 and the Eqs (16), (17), (19) and (20), we chose $k_a = 0.8$ $k_b = 780$, and the value range of terminal power (γ) is between 0 and 1, the performance of the proposed integer-order terminal sliding mode control strategy are showed in Fig 2.

It is clear from Fig 2(A) that the output voltage responses at the start-up become faster with increasing the value of γ. But the large value of γ can make overshoots and take a long time to reach the equilibrium point of the Eq (10). Fig 2(B) shows the responses of the output voltage for step changes in R (from 10 Ω to 1Ω), which are obtained by the SMC method with γ = 1, and the TSMC method with different γ values. Unlike the start-up case, it is interesting to note that the output voltage responses become faster with decreasing the value of γ. Therefore, the value of γ is chosen as some constant, to make a compromise between start up and transient responses of the converter.

When x_1 is near the equilibrium point, it can be seen as $|x_1| < 1$, the γ leads to $|x_1{}^{\gamma}| > |x_1|$. In such a case, the system state with the nonlinear term $x_1{}^{\gamma}$ converges toward equilibrium point faster than the linear term x_1. On the other hand, when $|x_1| > 1$, the γ leads to $|x_1{}^{\gamma}| < |x_1|$, it means the system state with the nonlinear term $x_1{}^{\gamma}$ converges toward equilibrium point slower than the linear term x_1.

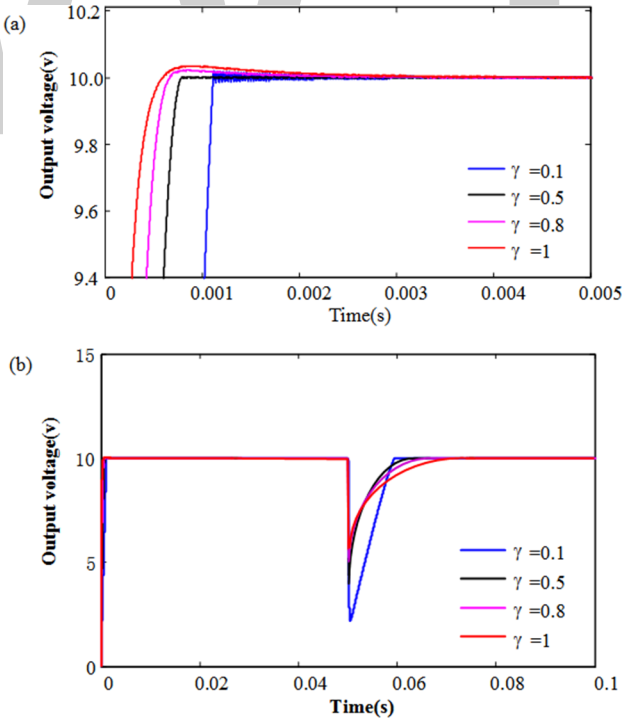

Fig 2. The output voltage dynamic response with different γ. (a) The output voltage dynamic response in start-up; (b) The output voltage dynamic response during load variations.

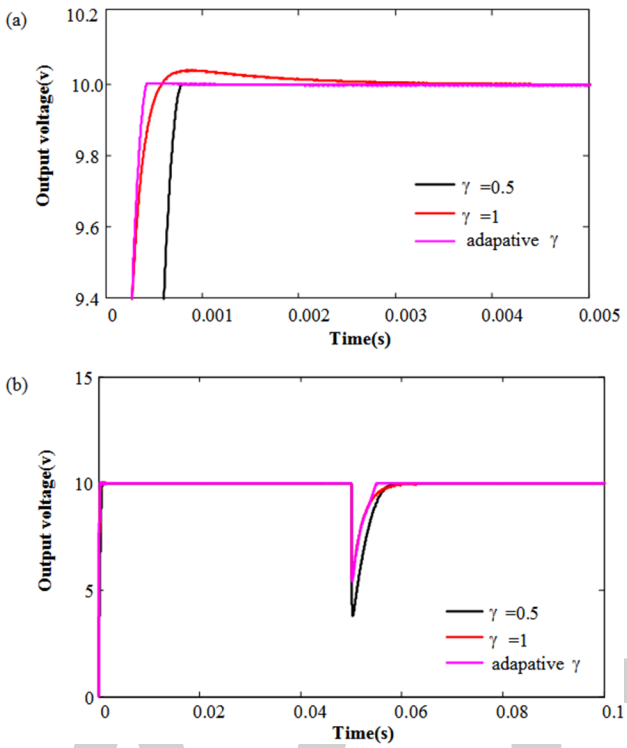

Fig 3. The output voltage dynamic response with adaptive γ. (a) The output voltage dynamic response in start-up; (b) The output voltage dynamic response during load variations.

So, we proposed the adaptive law to choose the value of γ, which builds a monotone increasing function x_1 for γ. This function will choose the value of γ approximately equal to 1 when $|x_1| > 1$, and choose the value of γ much smaller but not less than 0.25 when $|x_1| < 1$. According to the boundary conditions and the Simulink results, we use MATLAB/CFTOOL to fit the function of x_1 for γ, describe it as

$$\gamma = \frac{1}{\pi}\arctan(x_1 - 0.99) + 0.5 \tag{29}$$

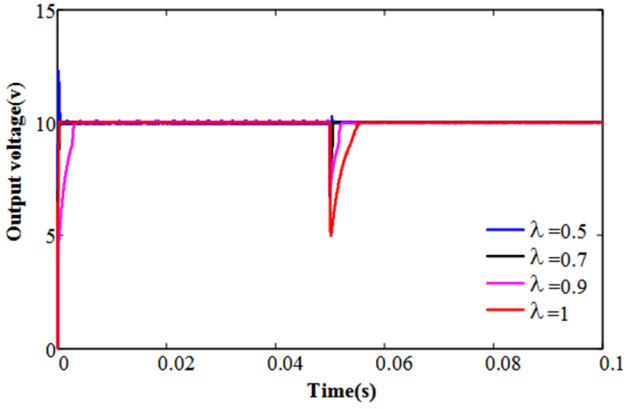

Fig 4. The output voltage dynamic response with different λ.

Table 2. Controller parameters of the proposed methods.

Descriptions	k_a	k_b	γ	λ
SMC	0.8	780	1	1
ATSMC	0.8	780	adaptive	1
AFTSMC	0.8	780	adaptive	0.7

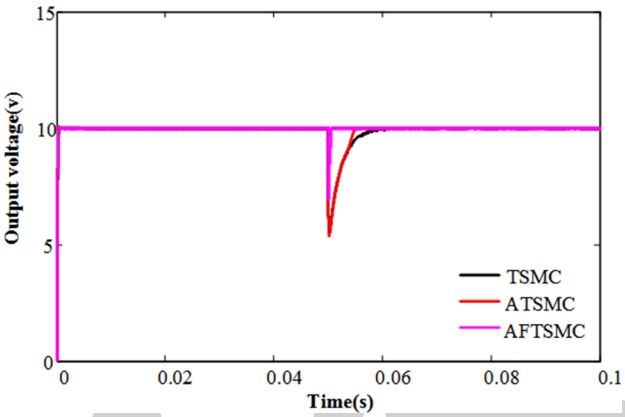

Fig 5. The output voltage dynamic response with different strategy.

From Eq (29), according to the state error, γ is selected adaptively, the simulation result as shown in Fig 3. We can easily observe the dynamic performance of the adaptive strategy better than the other constant terminal power.

Further, we investigate the dynamic response of output voltage with different fractional order (λ) on the basis of the adaptive terminal sliding mode control strategy. Fig 4 shows the simulated start-up and transient responses of the output voltage obtained by AFTSMC strategies with different λ values. It is interesting to note that the output voltage responses become faster with decreasing the value of λ, but when λ = 0.5, the overshoot of the system appears and exceeds 25%. In order to obtain high performance control strategy, we should try to avoid the voltage overshoot and chattering. So choosing λ = 0.7 is our choice for ideal parameter value.

In order to compare the control effect of TSMS, ATSMC, and AFTSMC, we use the parameter selection as shown in Table 2. As shown in Fig 5, the response time of the system with AFTSMC is less than others. At t = 0.05 s, the load resistance is changed from 10Ω to 1 Ω. Therefore, the output current will be increased, and the output voltage has a short step-down. It can be seen that the output voltage returns faster to reference output voltage in AFTSMC.

Conclusions

The fractional order terminal sliding mode control (FTSMC) based on a double closed-loop structure of voltage and current has been proposed. The influence of the controller parameters was investigated. It is observed that the chosen value of terminal power aims to make a compromise between the start-up and the moment when load changes. For this matter, we proposed an adaptive law to choose the terminal power, and the simulation shows that the method is effective. Further, we investigated the dynamic response of output voltage with different fractional orders, on the basis of the adaptive terminal sliding mode control strategy. It is shown that when the fractional order (λ) equal to 0.7, the performance of dynamic responses is better than others. In addition, the simulation results show that the AFTSMC strategy has

the better performance in comparison with the ATSMC and TSMC. The novel fractional terminal sliding mode control exhibits considerable improvement in terms of a faster output voltage response, in the start-up and during load changes.

Supporting information

S1 Fig. Simulation framework diagram. In this simulation framework, the sliding surface function *S* could be the integer or fractional order terminal sliding surface function.

S2 Fig. The output voltage dynamic response with different strategy. The control strategies include terminal sliding mode control (TSMC), adaptive terminal sliding mode control (ATSMC), and adaptive fractional order terminal sliding mode control (AFTSMC) respectively.

Author Contributions

Conceptualization: Huan Zhou.

Methodology: Jianlin Wang, Dan Xu.

Software: Huan Zhou.

Supervision: Anning Bai.

Writing – review & editing: Wei Lu.

References

1. Wu KC, Wu HH, Wei CL. Analysis and Design of Mixed-Mode Operation for Noninverting Buck–Boost DC–DC Converters. IEEE Transactions on Circuits & Systems II Express Briefs. 2015; 62(12):1194–5.
2. Ren GD, Xu Y, Wang CN. Synchronization behavior of coupled neuron circuits composed of memristors. Nonlinear Dynamics. 2017; 88(2):893–901.
3. Wang C, Chu R, Ma J. Controlling a chaotic resonator by means of dynamic track control. Complexity. 2015; 21(1): 370–378
4. Silva FA. Sliding mode control of switching power converters: techniques and implementation. IEEE Industrial Electronics Magazine. 2012; 38(2–3): 203–213.
5. Komurcugil H. Adaptive terminal sliding-mode control strategy for DC-DC Buck converters. Isa Transactions. 2012; 51(6):673–681. https://doi.org/10.1016/j.isatra.2012.07.005 PMID: 22877744
6. Wang YM, Cao YQ, Xia HW. Terminal sliding mode control for Buck converter with structure of voltage and current double closed loop. Electric Machines & Control. 2016; 20(8):92–97. Chinese
7. Wang J, Li S, Yang J, Wu B. Extended state observer-based sliding mode control for PWM-based DC–DC Buck power converter systems with mismatched disturbances. Control Theory & Applications Iet. 2015; 9(4):579–586.
8. Wu C, Si G, Zhang Y, Yang N. The fractional-order state-space averaging modeling of the Buck–Boost DC/DC converter in discontinuous conduction mode and the performance analysis. Nonlinear Dynamics. 2015; 79(1):689–14.
9. Yang N, Wu C, Jia R, Liu C. Fractional-Order Terminal Sliding-Mode Control for Buck DC/DC Converter. Mathematical Problems in Engineering. 2016; (2016-7-31):1–7.
10. Delavari H, Ghaderi R, Ranjbar A, Momani S. Fuzzy fractional order sliding mode controller for nonlinear systems. Communications in Nonlinear Science & Numerical Simulation. 2010; 15(4):963–15.
11. Hosseinnia SH, Tejado I, Vinagre BM, Sierociuk D. Boolean-based fractional order SMC for switching systems application to a DC-DC Buck converter. Signal, Image and Video Processing. 2012; 6(3):445–451.
12. Delghavi MB, Shoja-Majidabad S, Yazdani A. Fractional-Order Sliding-Mode Control of Islanded Distributed Energy Resource Systems. IEEE Transactions on Sustainable Energy. 2016; 7(4):1482–9.

13. Yang N, Wu C, Jia R, Liu C. Modeling and Characteristics Analysis for a Buck-Boost Converter in Pseudo-Continuous Conduction Mode Based on Fractional Calculus. Mathematical Problems in Engineering. 2016; 1:1–11.
14. Lorentz VR, Berberich SE, März M, Bauer AJ, Ryssel H, Poure P, et al. Lossless average inductor current sensor for CMOS integrated DC–DC converters operating at high frequencies. Analog Integrated Circuits & Signal Processing. 2010; 62(3):333–344.
15. Tan SC, Lai YM, Cheung MKH, Tse CK. On the practical design of a sliding mode voltage controlled Buck converter. IEEE Transactions on Power Electronics. 2005; 20(2):425–13.
16. Li H, Wang J, Lam HK, Zhou Q, Du H. Adaptive Sliding Mode Control for Interval Type-2 Fuzzy Systems. IEEE Transactions on Systems Man & Cybernetics Systems. 2016; PP(99):1–10.
17. Ramos R, Biel D, Fossas E, Griño R. Sliding mode controlled multiphase Buck converter with interleaving and current equalization. Control Engineering Practice. 2013; 21(5):737–746.
18. Aguila-Camacho N, Duarte-Mermoud MA, Gallegos JA. Lyapunov functions for fractional order systems. Communications in Nonlinear Science & Numerical Simulation. 2014; 19(9):2951–2957.
19. Gallegos JA, Duarte-Mermoud MA. On the Lyapunov theory for fractional order systems. Applied Mathematics & Computation. 2016; 287:161–170.
20. Wu GC, Baleanu D, Luo WH. Lyapunov functions for Riemann–Liouville-like fractional difference equations. Applied Mathematics & Computation. 2017; 314:228–236.
21. Wu GC, Baleanu D, Xie HP, Chen FL. Chaos synchronization of fractional chaotic maps based on the stability condition. Physica A Statistical Mechanics & Its Applications. 2016; 460:374–383.
22. Li Y, Chen YQ, Podlubny I. Stability of fractional-order nonlinear dynamic systems: Lyapunov direct method and generalized Mittag–Leffler stability. Computers & Mathematics with Applications. 2010; 59 (5):1810–1821.
23. Oldham KB, Spanier J. The fractional calculus. Mathematical Gazette. 1974; 56(247):396–400.

A new implementation for online calculation of manipulator Jacobian

Pramod Chembrammel[1]*, **Thenkurussi Kesavadas**[1,2]

1 Health Care Engineering Systems Center, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America, **2** Industrial and Enterprise Systems Engineering Dept., University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America

* pramodch@illinois.edu

Abstract

This paper describes a new implementation for calculating Jacobian and its time derivative for robot manipulators in real-time. The estimation of Jacobian is the key in the real-time implementation of kinematics and dynamics of complex planar or spatial robots with fixed as well as floating axes in which the Jacobian form changes with the structure. The proposed method is suitable for such implementations. The new method is based on matrix differential calculus. Unlike the conventional methods, which are based on screw theory, the Jacobian calculation in the proposed approach has been reduced to the inner product of two matrices. Use of the new method to derive linear and angular velocity parts of Jacobian and its time derivative is described in detail. We have demonstrated the method using a two-DOF spatial robot and a hyper-redundant spatial robot.

Editor: Jian Huang, Huazhong University of Science and Technology, CHINA

1 Introduction

The agility of a robot depends on how fast it can adapt to an environment. The dynamics of a robot required for such control primarily involves calculation of Jacobian. It relates joint rates to end-effector velocities and relates end-effector forces to joint torques. Also, the columns of the Jacobian are the instantaneous directions in which a desired point on a robot can move [1]. Thus, Jacobian is the key to the analysis and control of robots. Various methods of calculating the Jacobian with a fixed number of links have been explored and successfully demonstrated in [2–4] and in numerous other publications on robot kinematics and dynamics.

Current methods [2, 5] deal with fixed-configuration of the robot in which the Jacobian is calculated a priori with respect to a fixed set of points on the robot. In some cases, Jacobian needs to be calculated with respect to points that vary within a local coordinate system. Such points arise in the case of fixed as well as reconfigurable robots. In the case of reconfigurable robots, these points are generated because of the change in configuration (eg: change in the configuration of a walking robot or addition of intermediate links to a serial manipulator), whereas in the case of fixed-configuration robots, the variation is either due to floating axes [6] or due to external influences. Such variations require that dynamic control be exercised with respect to the varying points, which demand recursive and real-time calculations. Conventional methods of finding Jacobian fail in such cases, these methods involve manual

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

reformulation of the Jacobian when it changes form for the above reasons. In this paper, we propose a new method for the real-time calculation and reformulation of Jacobian. This method can be applied to fixed as well as reconfigurable robots. The scope of this paper is limited to fixed-configuration. The number of matrix operations is $2n + n/(2\delta)$ (see Section 6) as compared to $(3n - 6)$ of the widely used Renaud's method [5]. The performance of the proposed method is comparable to that of Renaud's method for low DOF ($n < 8$) robots and is superior for high-DOF robots (see Section 6). In terms of reformulation, Renaud's method cannot be compared to the proposed method, since Renaud's method involves manual reformulation if the Jacobian changes form.

Our method also permits the autonomous online computation of the time derivative, $\dot{\mathbf{J}}$, of Jacobian, which is required for the robot dynamics as well as to resolve kinematic redundancy using optimal control techniques [7]. $\dot{\mathbf{J}}$ is also required for the estimation of link positions from joint angle measurements (proprioception) using optimal estimation techniques like unscented Kalman filters and particle filters [8–10]. Symbolic computation of Jacobian and its derivative is detailed in [11], but it requires extensive manual intervention and hence is not an autonomous approach. In addition, we are not interested in a symbolic derivation, but in a numerical implementation that can autonomously estimate the Jacobian and its time derivative. Authors could not find any available literature on systematic and autonomous derivation of $\dot{\mathbf{J}}$.

2 Related work on Jacobian estimation

Conventional methods use either the loop closure method [1, 2, 4, 12] or screw theory [2, 5] to calculate Jacobian. The main disadvantage of the loop-closure method is that although it is very useful for planar and spatial mechanisms with a few degrees of freedom (DOF), it is not suitable for complex spatial mechanisms with high DOF and linkages with lower pair joints [13]. Screw-theory-based methods are useful in such cases, but fail to address situations with floating axes of the joints, as demonstrated in [6, 14]. All the reported works efforts have dwelt on finding Jacobian with respect to only one point, i.e., the end-effector. However, there are many situations in which the Jacobian has to be calculated with respect to many points on the robot, which may also vary with time. These methods do not perform automated calculation of the Jacobian matrix and require manual intervention before final solution can be found. That is the main obstacle to real-time implementation.

2.1 Comparison of existing methods

This section compares the computational efficiencies of different existing methods based on screw theory. There are six such methods, as discussed in [5]. Although the methods use the same basic concepts, they yield different forms of the Jacobian matrix. The methods work on the premise of finding the angular (ω) and linear ($\mathbf{v}$) velocities of the desired point on the manipulator by summing the joint rates.

$$\omega = \sum_{i=1}^{n} \dot{q}_i \mathbf{z}_{i-1} \tag{1}$$

$$\mathbf{v} = -\sum_{i=1}^{n} (\dot{q}_i \mathbf{z}_{i-1}) \times \mathbf{p}_{i-1} \tag{2}$$

where $(.) \times (.)$ is a cross product, $\mathbf{p}_{i-1}$ is the coordinate of the desired point in the local coordinate system, and all other $\mathbf{p}_i$ represent the coordinates of the origins of the link coordinate

systems with respect to the preceding coordinate system. These methods differ only in the way the coefficients are extracted out of Eqs (1) and (2) to form the Jacobian. Among the methods discussed in [5], the first three approaches differ from the fourth and fifth in that their computations begin from opposite ends of a multi-link robot. The first three methods begin from the base, while the others begin from the end-effector. Among them, the fourth method, developed by Renaud [15], is computationally superior to the other methods with $(30n - 87)$ multiplications, $(15n - 66)$ additions and subtractions, and $(2n - 2)$ sines and cosines. Here, n is the number of degrees of freedom. The number of matrix operations $(3n - 6)$ is also the least for this method [5]. Although these methods are very efficient in calculating the Jacobian, they require manual intervention to select the axis of rotation or translation ($\mathbf{z}_{i-1}$ in Eq (1)), depending on the type of joint variable. That renders the existing methods incapable of finding the form of Jacobian in real-time, both for re-configurable and for fixed-configuration robots in which the Jacobian with respect to a new point has to be found.

Automatic estimation of Jacobian via zeroing dynamics [16] while the kinematic structure (DH parameters) of a robot is not known, is explained in [17]. In this method, the Jacobian is iteratively estimated by tracking the kinematics of the end-effector using feedback from external sensors like cameras or inertial measurement units (IMUs). However, external feedback makes this approach less attractive for most of the applications for which external tracking is either not required or not possible. In addition, this method is not applicable for the regular (nonredundant) robots; it is applicable only for redundant robots.

In this paper, we propose a new method that facilitates estimation of Jacobian in real-time. Like the loop-closure method, the proposed method is founded on matrix differential calculus; however, the procedure involved in finding the Jacobian is different from those of the loop-closure and screw-theory-based methods. Also, it does not distinguish between joints, either prismatic or revolute. It is therefore possible to use numerical methods for the real-time estimation of Jacobian.

This paper is laid out as following. In section 3, proposed method is described. In section 4, we discuss how we demonstrated it using a simple two-link planar manipulator. The method is summarized as an algorithm in section 5. In section 6 we explain how we demonstrated the real-time application of the method with the help of a redundant spatial robotic arm.

3 The new method

A point, $\mathbf{p}$, in a coordinate system, j, attached to a link of a robot is represented in another coordinate frame, i, or the base coordinate frame, through the use of homogeneous transformation.

$$\mathbf{x}^i(\mathbf{q}) = \mathbf{A}(\mathbf{q})\mathbf{p}^j \tag{3}$$

where $\mathbf{A}$ is the homogeneous transformation matrix, and $\mathbf{x}^i$ is the point after transformation. The superscripts represent the respective coordinate systems. The matrix $\mathbf{A}$ is a function of the generalized coordinates, $\mathbf{q}$, of the robot thus making $\mathbf{x}^i$ also a function of $\mathbf{q}$.

The velocity of the transformed point in the i^{th} coordinate system is given by

$$\dot{\mathbf{X}}^i(\mathbf{q}) = \mathbf{J}(\mathbf{q}, \mathbf{p})\dot{\mathbf{q}} \tag{4}$$

where $\mathbf{J}$ is the Jacobian of the robot with respect to the point $\mathbf{p}^j$. $\dot{\mathbf{X}}^i$ has both global linear velocity, $(\mathbf{v}^i = \dot{\mathbf{x}}^i)$, and angular velocity, ω^i. As is evident from Eq (4), the Jacobian relates the joint rates to the velocities of the desired point $\mathbf{p}^j$ represented in frame i. In the transformation of Eqs (3) to (4), the coordinates of the point $\mathbf{p}$ in the coordinate system j have been absorbed into the $\mathbf{J}$. Thus, if the Jacobian is proved to be a function of independent entities, i.e. as a

function of the derivative of the transformation matrix [18] $\mathbf{A}(\mathbf{q})$ and the constant vector $\mathbf{p}$, then it is only sufficient to perform simple matrix multiplications to find out the Jacobian. This is the key to real-time estimation of $\mathbf{J}(\mathbf{q})$.

$\mathbf{J}$ has two parts; a relative velocity part of dimension $(3 \times n)$, and an angular velocity part of dimension $(3 \times n)$ [2]. The following subsections explain the real-time estimation of both parts of $\mathbf{J}$. (It is to be noted that $\mathbf{x}$ and ω are represented in i and that $\mathbf{p}$ is represented in j. Hence, superscripts are ignored in the following sections for conciseness).

3.1 Calculation of linear velocity part

The velocity equation in Eq (4) is derived as follows. Taking the time derivative of Eq (3),

$$\dot{\mathbf{x}}(\mathbf{q}) = \dot{\mathbf{A}}(\mathbf{q})\mathbf{p} \tag{5}$$

The time derivative of $\mathbf{p}$ is not present in Eq (5) since it is a constant vector in the local coordinate system, j. The time derivative of $\mathbf{A}(\mathbf{q})$ is written as

$$\dot{\mathbf{A}}(\mathbf{q}) = \sum_{i=1}^{n} \frac{\partial \mathbf{A}(q_i)}{\partial q_i} \dot{q}_i \tag{6}$$

Using Eq (6) in Eq (5),

$$\dot{\mathbf{x}}(\mathbf{q}) = \left[\sum_{i=1}^{n} \frac{\partial \mathbf{A}(q_i)}{\partial q_i} \dot{q}_i \right] \mathbf{p} \tag{7}$$

Since the vector $\mathbf{p}$ is common and $\dot{q}_i$ is a scalar, Eq (7) is rewritten as

$$\dot{\mathbf{x}}(\mathbf{q}) = \sum_{i=1}^{n} \left[\frac{\partial \mathbf{A}(q_i)}{\partial q_i} \mathbf{p} \right] \dot{q}_i \tag{8}$$

In matrix notation, Eq (8) is represented by

$$\dot{\mathbf{x}}(\mathbf{q}) = \mathbf{J}_v(\mathbf{q})\dot{\mathbf{q}} \tag{9}$$

where, $\mathbf{J}_v(\mathbf{q})$ is the linear velocity part of $\mathbf{J}(\mathbf{q})$. What is left to perform is to decompose $\mathbf{J}_v(\mathbf{q})$ as a product of independent matrices. From Eq (8), $\mathbf{J}_v(\mathbf{q})$ is given by

$$\mathbf{J}_v(\mathbf{q}) = \left[\frac{\partial \mathbf{A}(q_1)}{\partial q_1}\mathbf{p} \ \ldots \ \frac{\partial \mathbf{A}(q_i)}{\partial q_i}\mathbf{p} \ \ldots \ \frac{\partial \mathbf{A}(q_n)}{\partial q_n}\mathbf{p} \right] \tag{10}$$

Eq (10) is written as the product of two matrices:

$$\mathbf{J}_v(\mathbf{q}) = \mathbf{L}\mathbf{P} \tag{11}$$

where

$$\mathbf{L} = \left[\frac{\partial \mathbf{A}(q_1)}{\partial q_1} \ \ldots \frac{\partial \mathbf{A}(q_i)}{\partial q_i} \ \ldots \ \frac{\partial \mathbf{A}(q_n)}{\partial q_n} \right] \tag{12}$$

and $\mathbf{P} = \mathbf{I}_n \otimes \mathbf{p}$, where $\mathbf{I}_n$ is an $(n \times n)$ identity matrix, $\mathbf{p}$ is in homogeneous form, and $(.) \otimes (.)$ is the Kronecker product.

$$\mathbf{p} = [x \ y \ z \ 1]^{\mathrm{T}} \tag{13}$$

The dimensions of $\mathbf{J}_v$, $\mathbf{L}$, and $\mathbf{P}$ are, respectively, $(4 \times n)$, $(4 \times 4n)$, and $(4n \times n)$. Thus, the linear

velocity part of the Jacobian is

$$\mathbf{J}_v(\mathbf{q}) = (\mathbf{LP})_{(1-3,1-n)} \tag{14}$$

It is to be noted that **P** is a sparse matrix and hence can efficiently store values.

3.2 Calculation of angular velocity part

The matrix **A** in Eq (3) is also called the *transition matrix* from the initial value to the final value [2, 19], and is written as

$$\mathbf{x}(\mathbf{q}, t) = \mathbf{A}(\mathbf{q}(t))\mathbf{x}(\mathbf{q}, 0) \tag{15}$$

Taking the time derivative of Eq (15),

$$\dot{\mathbf{x}}(\mathbf{q}, t) = \dot{\mathbf{A}}(\mathbf{q}(t))\mathbf{x}(\mathbf{q}, 0) \tag{16}$$

Since **A** is invertible, from Eq (15),

$$\dot{\mathbf{x}}(\mathbf{q}, t) = \dot{\mathbf{A}}(\mathbf{q}(t))\mathbf{A}(\mathbf{q}(t))^{-1}\mathbf{x}(\mathbf{q}, t) \tag{17}$$

But,

$$\dot{\mathbf{A}}(\mathbf{q}(t))\mathbf{A}((\mathbf{q}(t))^{-1} = \begin{bmatrix} \dot{\mathbf{R}} & \dot{\mathbf{r}}^{\mathrm{T}} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R}^{\mathrm{T}} & -\mathbf{R}^{\mathrm{T}}\mathbf{r}^{\mathrm{T}} \\ 0 & 1 \end{bmatrix} \tag{18}$$

where **R** is the (3 × 3) orientation part of **A**, and $\mathbf{r}^{\mathrm{T}}$ is a (3 × 1) translation part of **A**. Eq (18) is written as

$$\dot{\mathbf{A}}(\mathbf{q}(t))\mathbf{A}(\mathbf{q}(t))^{-1} = \begin{bmatrix} \dot{\mathbf{R}}\mathbf{R}^{\mathrm{T}} & \dot{\mathbf{r}}^{\mathrm{T}} - \dot{\mathbf{R}}\mathbf{R}^{\mathrm{T}}\mathbf{r}^{\mathrm{T}} \\ 0 & 0 \end{bmatrix} \tag{19}$$

The element $\dot{\mathbf{R}}\mathbf{R}^{\mathrm{T}}$ of Eq (19) is called the *twist matrix* [2] which is the skew-symmetric form of the angular velocity vector [20] of the desired link. Similar to Eq (8), the left hand side of Eq (19) is written using (6) as

$$\dot{\mathbf{A}}(\mathbf{q}(t))\mathbf{A}(\mathbf{q}(t))^{-1} = \sum_{i=1}^{n} \left[\frac{\partial \mathbf{A}(q_i)}{\partial q_i} \mathbf{A}(\mathbf{q}(t))^{-1} \right] \dot{q}_i \tag{20}$$

The right hand side of Eq (20) is thus written as

$$\dot{\mathbf{A}}(\mathbf{q}(t))\mathbf{A}(\mathbf{q}(t))^{-1} = \sum_{i=1}^{n} \mathbf{L}_i\mathbf{P}\dot{q}_i \tag{21}$$

where **L** is given by Eq (12) and $\mathbf{P} \triangleq \mathbf{A}^{-1}$. The subscript *i* on the right hand side represents the i^{th} (4 × 4) block of the matrix **L**. The first (3 × 3) matrix of ($\mathbf{L}_i\mathbf{P}$) represents the twist matrix for each link. Thus, the angular velocity part of the Jacobian is

$$\mathbf{J}_{\omega,i} = [(\mathbf{L}_i\mathbf{P})_{(3,2)} \ (\mathbf{L}_i\mathbf{P})_{(1,3)} \ (\mathbf{L}_i\mathbf{P})_{(2,1)}]^{\mathrm{T}} \tag{22}$$

The subscript *i* on the left hand side of Eq (22) represents the column index of the manipulator Jacobian. The subscripts (3, 2), (1, 3), and (2, 1) of the twist matrix from ($\mathbf{L}_i\mathbf{P}$) represent the *x*, *y*, *z* components of the angular velocity part of the Jacobian. To use the standard notation, the

complete manipulator Jacobian is thus

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_v \\ \\ \mathbf{J}_\omega \end{bmatrix} \tag{23}$$

From Eqs (12), (14), and (22), it is observed that the Jacobian is affected by any variation in the point **p**. With this method. unlike the conventional methods, this variation can be easily incorporated into the calculation of the Jacobian just by recalculating $\mathbf{P} = \mathbf{I}_n \otimes \mathbf{p}$. This method has applications in fixed-configuration robots with varying points.

3.3 Time derivative, $\dot{\mathbf{J}}$, of Jacobian

In order to find out the time derivative of Jacobian, we consider the linear and velocity parts separately because **P** is different in the two cases. The total time derivative of the linear part is

$$\dot{\mathbf{J}}_v = \sum_{i=1}^{n} \frac{\partial \mathbf{J}_v}{\partial q_i} \dot{q}_i \tag{24}$$

Since **P** is a constant from Eq (13),

$$\frac{\partial \mathbf{J}_v}{\partial q_i} = \frac{\partial \mathbf{L}}{\partial q_i} \mathbf{P} \tag{25}$$

Similarly, the total time derivative of the angular part is

$$\dot{\mathbf{J}}_\omega = \sum_{i=1}^{n} \frac{\partial \mathbf{J}_\omega}{\partial q_i} \dot{q}_i \tag{26}$$

Using Eq (22),

$$\frac{\partial \mathbf{J}_{\omega,i}}{\partial q_j} = \frac{\partial}{\partial q_j} \left(\left[(\mathbf{L}_i \mathbf{P})_{(3,2)} \;\; (\mathbf{L}_i \mathbf{P})_{(1,3)} \;\; (\mathbf{L}_i \mathbf{P})_{(2,1)} \right]^{\mathrm{T}} \right) \tag{27}$$

From Eq (21), **P**, in this case, is not a constant. Therefore,

$$\frac{\partial (\mathbf{L}_i \mathbf{P})}{\partial q_j} = \frac{\partial \mathbf{L}_i}{\partial q_j} \mathbf{P} + \mathbf{L}_i \frac{\partial \mathbf{P}}{\partial q_j} \tag{28}$$

But,

$$\frac{\partial \mathbf{P}}{\partial q_j} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial q_j} \mathbf{A}^{-1} \tag{29}$$

where, $\mathbf{L}_i$ is the i^{th} (4×4) block of **L**. Eq (29) is a standard matrix relationship [21]. By using Eq (29) in Eq (28), we obtain

$$\frac{\partial (\mathbf{L}_i \mathbf{P})}{\partial q_j} = \left(\frac{\partial \mathbf{L}_i}{\partial q_j} - \mathbf{L}_i \mathbf{A}^{-1} \mathbf{L}_j \right) \mathbf{A}^{-1} \tag{30}$$

4 Demonstration

In this section, the concepts discussed in the previous section are demonstrated using a two-DOF spatial manipulator shown in Fig 1.

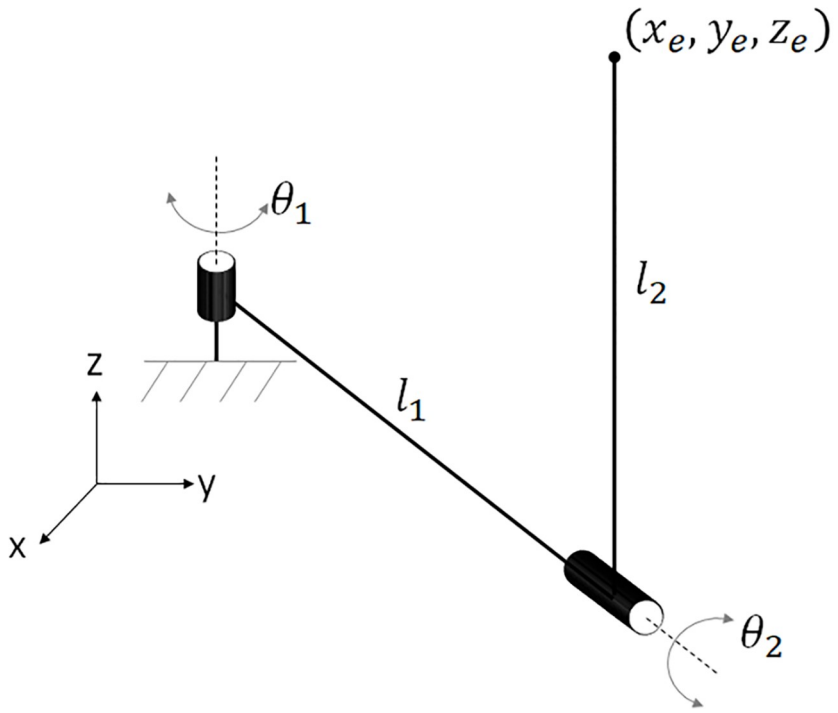

Fig 1. Two-link spatial robot for demonstration of the method.

Considering an arbitrary point (x_e, y_e, z_e) in the local coordinate frame attached to the second link:

$$\mathbf{x}(\mathbf{q}) = \begin{bmatrix} -S\theta_1 C\theta_2 & C\theta_1 & -S\theta_1 S\theta_2 & l_1 C\theta_1 - l_2 S\theta_1 C\theta_2 \\ C\theta_1 C\theta_2 & S\theta_1 & C\theta_1 S\theta_2 & l_1 S\theta_1 + l_2 C\theta_1 C\theta_2 \\ S\theta_2 & 0 & -C\theta_2 & l_2 S\theta_2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_e \\ y_e \\ z_e \\ 1 \end{bmatrix} \tag{31}$$

where C stands for *cosine* and S for *sine*. The transformation matrix in the above equation is recursively calculated using a homogeneous transformation matrix [18]. Taking the time derivative of Eq (31),

$$\dot{\mathbf{x}}(\mathbf{q}, \dot{\mathbf{q}}) = \tilde{\mathbf{J}}\mathbf{X}_e \tag{32}$$

where

$$\tilde{\mathbf{J}} = \mathbf{J}_1\dot{q}_1 + \mathbf{J}_2\dot{q}_2 \tag{33}$$

$$\mathbf{X}_e = [x_e \; y_e \; z_e \; 1]^{\mathrm{T}} \tag{34}$$

$$\mathbf{J}_1 = \begin{bmatrix} -C\theta_1 C\theta_2 & -S\theta_1 & -C\theta_1 S\theta_2 & -l_1 S\theta_1 - l_2 S\theta_1 C\theta_2 \\ -S\theta_1 C\theta_2 & C\theta_1 & -S\theta_1 S\theta_2 & l_1 C\theta_1 - l_2 S\theta_1 C\theta_2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{35}$$

and

$$\mathbf{J}_2 = \begin{bmatrix} S\theta_1 S\theta_2 & 0 & -S\theta_1 C\theta_2 & l_2 S\theta_1 S\theta_2 \\ -C\theta_1 S\theta_2 & 0 & C\theta_1 C\theta_2 & -l_2 C\theta_1 S\theta_2 \\ C\theta_2 & 0 & S\theta_2 & l_2 C\theta_2 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{36}$$

Expanding Eq (32) using Eq (33),

$$\dot{\mathbf{x}}(\mathbf{q}, \dot{\mathbf{q}}) = [\mathbf{J}_1 \dot{q}_1 + \mathbf{J}_2 \dot{q}_2]\mathbf{X}_e \tag{37}$$

Since $\mathbf{X}_e$ is common, rearranging Eq (37),

$$\dot{\mathbf{x}}(\mathbf{q}, \dot{\mathbf{q}}) = \mathbf{J}_1 \dot{q}_1 \mathbf{X}_e + \mathbf{J}_2 \dot{q}_2 \mathbf{X}_e \tag{38}$$

In the matrix form,

$$\dot{\mathbf{x}}(\mathbf{q}, \dot{\mathbf{q}}) = [\mathbf{J}_1 \mathbf{X}_e \quad \mathbf{J}_2 \mathbf{X}_e] \begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \end{bmatrix} \tag{39}$$

In Eq (39), $\mathbf{X}_e$ can be factored out to obtain

$$\dot{\mathbf{x}}(\mathbf{q}, \dot{\mathbf{q}}) = [\mathbf{J}_1 \quad \mathbf{J}_2] (\mathbf{I}_2 \otimes \mathbf{X}_e) \begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \end{bmatrix} \tag{40}$$

If we compare Eq (40) with the standard form of the velocity equation, $\dot{\mathbf{x}} = \mathbf{J}_v \dot{\mathbf{q}}$, the linear velocity part of the Jacobian for the two-link manipulator is,

$$\mathbf{J}_v = [\mathbf{J}_1 \quad \mathbf{J}_2] (\mathbf{I}_2 \otimes \mathbf{X}_e) \tag{41}$$

Eq (41) is the same as Eq (11), in which $\mathbf{L} = [\mathbf{J}_1 \, \mathbf{J}_2]$ and $\mathbf{P} = (\mathbf{I}_2 \otimes \mathbf{X}_e)$. If the end-effector is considered the point of interest, then $\mathbf{X}_e = [0\ 0\ 0\ 1]^T$. Thus, $\mathbf{J}_v$ for the two-DOF spatial manipulator is written as

$$\mathbf{J}_v = \begin{bmatrix} -l_1 S\theta_1 - l_2 C\theta_1 C\theta_2 & l_2 S\theta_1 S\theta_2 \\ l_1 C\theta_1 - l_2 S\theta_1 C\theta_2 & -l_2 C\theta_1 S\theta_2 \\ 0 & l_2 C\theta_2 \\ 0 & 0 \end{bmatrix} \tag{42}$$

As given in Eq (14), the first three rows represent the linear part of the Jacobian. Therefore,

$$\mathbf{J}_v = \begin{bmatrix} -l_1 S\theta_1 - l_2 C\theta_1 C\theta_2 & l_2 S\theta_1 S\theta_2 \\ l_1 C\theta_1 - l_2 S\theta_1 C\theta_2 & -l_2 C\theta_1 S\theta_2 \\ 0 & l_2 C\theta_2 \end{bmatrix} \tag{43}$$

The angular velocity part is obtained as follows. $\mathbf{P}$ is

$$\mathbf{P} = \mathbf{A}^{-1} = \begin{bmatrix} -S\theta_1 C\theta_2 & C\theta_1 C\theta_2 & S\theta_2 & -l_2 \\ C\theta_1 & S\theta_1 & 0 & -l_1 \\ -S\theta_1 S\theta_2 & C\theta_1 S\theta_2 & -C\theta_2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{44}$$

Using Eqs (35) & (36), and following the expression in Eq (22), the block matrix multiplication (**LP**) will yield,

$$\mathbf{J}_{\omega} = \begin{bmatrix} 0 & C\theta_1 \\ 0 & S\theta_1 \\ 1 & 0 \end{bmatrix} \tag{45}$$

From Eqs (43) and (45), the complete Jacobian is

$$\mathbf{J} = \begin{bmatrix} -l_1 S\theta_1 - l_2 C\theta_1 C\theta_2 & l_2 S\theta_1 S\theta_2 \\ l_1 C\theta_1 - l_2 S\theta_1 C\theta_2 & -l_2 C\theta_1 S\theta_2 \\ 0 & l_2 C\theta_2 \\ 0 & C\theta_1 \\ 0 & S\theta_1 \\ 1 & 0 \end{bmatrix} \tag{46}$$

Equation Eq (46) is a standard result. The time derivative $\dot{\mathbf{J}}$ is found as follows.

$$\dot{\mathbf{J}}_v = \left(\frac{\partial \mathbf{L}}{\partial \theta_1}\mathbf{P}\right)\dot{\theta}_1 + \left(\frac{\partial \mathbf{L}}{\partial \theta_2}\mathbf{P}\right)\dot{\theta}_2 \tag{47}$$

where

$$\frac{\partial \mathbf{L}}{\partial \theta_1}\mathbf{P} = \begin{bmatrix} -l_1 C\theta_1 + l_2 S\theta_1 C\theta_2 & l_2 C\theta_1 S\theta_2 \\ -l_1 S\theta_1 - l_2 C\theta_1 C\theta_2 & l_2 S\theta_1 S\theta_2 \\ 0 & 0 \end{bmatrix} \tag{48}$$

$$\frac{\partial \mathbf{L}}{\partial \theta_2}\mathbf{P} = \begin{bmatrix} l_2 C\theta_1 S\theta_2 & l_2 S\theta_1 C\theta_2 \\ l_2 S\theta_1 S\theta_2 & -l_2 C\theta_1 C\theta_2 \\ l_2 C\theta_2 & -l_2 S\theta_2 \end{bmatrix} \tag{49}$$

Using Eq (30),

$$\frac{\partial \mathbf{L}_1\mathbf{P}}{\partial \theta_1} = \frac{\partial \mathbf{L}_1\mathbf{P}}{\partial \theta_2} = \frac{\partial \mathbf{L}_2\mathbf{P}}{\partial \theta_2} = \mathbf{0}_{4\times 4} \tag{50}$$

And,

$$\frac{\partial \mathbf{L}_2\mathbf{P}}{\partial \theta_1} = \begin{bmatrix} 0 & 0 & C\theta_1 & 0 \\ 0 & 0 & S\theta_1 & 0 \\ -C\theta_1 & -S\theta_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{51}$$

Following Eq (27),

$$\frac{\partial \mathbf{J}_{\omega}}{\partial \theta_1} = \begin{bmatrix} 0 & -S\theta_1 \\ 0 & C\theta_1 \\ 0 & 0 \end{bmatrix}, \quad \frac{\partial \mathbf{J}_{\omega}}{\partial \theta_2} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{52}$$

Using Eq (26),

$$\dot{\mathbf{J}}_{\omega} = \begin{bmatrix} 0 & -S\theta_1 \\ 0 & C\theta_1 \\ 0 & 0 \end{bmatrix} \dot{\theta}_1 \tag{53}$$

The time derivative of the complete Jacobian is,

$$\dot{\mathbf{J}} = \begin{bmatrix} \dot{\mathbf{J}}_{v} \\ \dot{\mathbf{J}}_{\omega} \end{bmatrix} \tag{54}$$

For a two-DOF robot the formulation may appear to involve more steps than a conventional method. Computationally, there is no clear advantage in using this method for two- or three-link robots (see section 6). However, the expressions in Eqs (35), (36) and (41) can be autonomously calculated using differential calculus. That procedure is also suitable for the real-time calculation of Jacobian with respect to a point other than the chosen point at the end-effector. Also, it is a very useful method for mechanisms with floating axes and robots with high DOF. In prior work, we have very effectively used this method for floating axes to model the dynamics of lower limbs ($n = 21$) in human walking [14], as well as for modelling flexible guide-wire dynamics [22]. The application for high-DOF robots is explained in section 6 with respect to the real-time implementation of a real 7-DOF spatial robot.

5 Algorithm to find Jacobian

The method discussed in the previous sections can be summarized in the form of an algorithm.

1. Find the final transformation matrix, **A**.
2. Find the (4×4) block matrix,

$$\mathbf{L}_i = \frac{\partial \mathbf{A}(\mathbf{q})}{\partial q_i}$$

3. Form the $(4 \times 4n)$ matrix **L** as given in Eq (12).
4. Find matrix **P** as follows:
   a. for the linear part, use $\mathbf{P} = \mathbf{I}_n \otimes \mathbf{p}$.
   b. for the angular velocity part, use $\mathbf{P} = \mathbf{A}^{-1}$.
5. Calculate the Jacobian matrices $\mathbf{J}_{v/\omega} = \mathbf{LP}$ for both the parts.
6. Find the derivative of **L** as $\frac{\partial \mathbf{L}}{\partial q_i}$.
7. Calculate the time derivative of the linear part of the Jacobian using Eqs (24) and (25) with $\mathbf{P} = \mathbf{I}_n \otimes \mathbf{p}$.
8. Calculate the time derivative of the angular part of the Jacobian using Eqs (30), (27) and (26) with $\mathbf{P} = \mathbf{A}^{-1}$.

From the algorithm and the demonstration in section 4, it is evident that no manual intervention is required. The calculations of matrices **A**, **P** and **L** and their derivatives are

Table 1. Comparison of Renaud's method and the proposed method.

Method	No. of matrix multiplications	
	(3 × 3)	**(4 × 4)**
Renaud	$3n$-6	–
Proposed method	–	$2n + n/(2\delta)$

autonomously carried out. Hence this method is suitable for real-time implementation for any desired point on the robot.

6 Application to high-DOF robots and real-time implementation

The aim of this section is to demonstrate the application to higher-DOF robots for real-time implementation. Following the results given in [5], only the computations that are performed after the computation of the final homogenous transformation matrix, **A**, are considered. The number of matrix operations required is $2n$, which includes matrix-to-vector and matrix-to-matrix multiplications. The matrices and vectors are in the homogenous form (4 × 4) and (4 × 1). A comparison of the computational efficiency of the proposed method with that of Renaud's method is given in Table 1 and Fig 2. We chose to compare our method only to Renaud's method because Renaud's performance is superior to that of any other existing methods. Renaud's method involves multiplication of (3 × 3) matrices, so our method will involve $37n$ more scalar multiplications. Thus, the number of operations is modified to $2n + n/(2\delta)$, where $\delta = 37/64$ is the ratio between the difference and the number of scalar multiplications in our method. From Fig 2, the performance of our proposed method is comparable to Renaud's

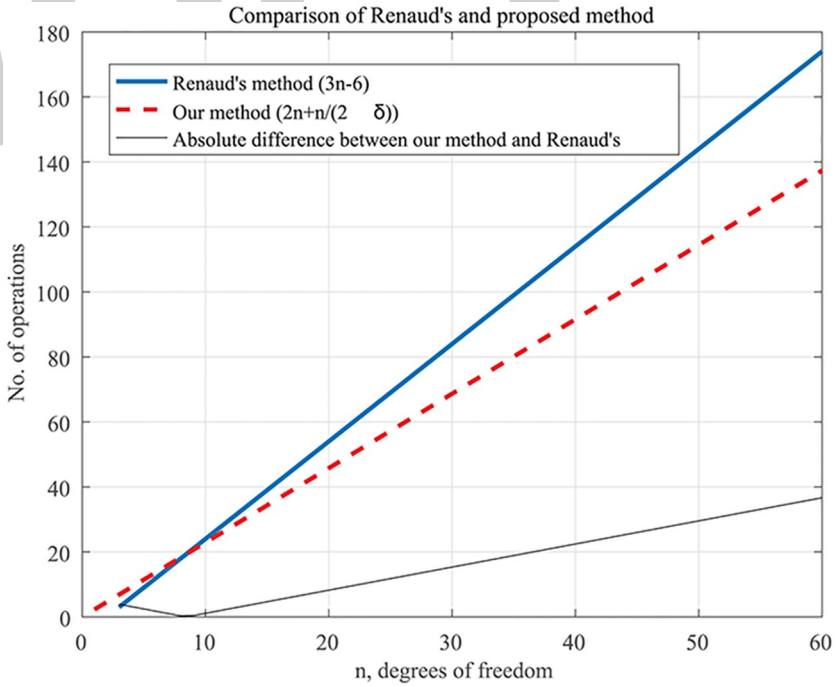

Fig 2. Comparison of computational efficiencies in terms of number of matrix operations in Renaud's and the proposed method. For ($n \geq 6$) the number of matrix operations is lower in the proposed method.

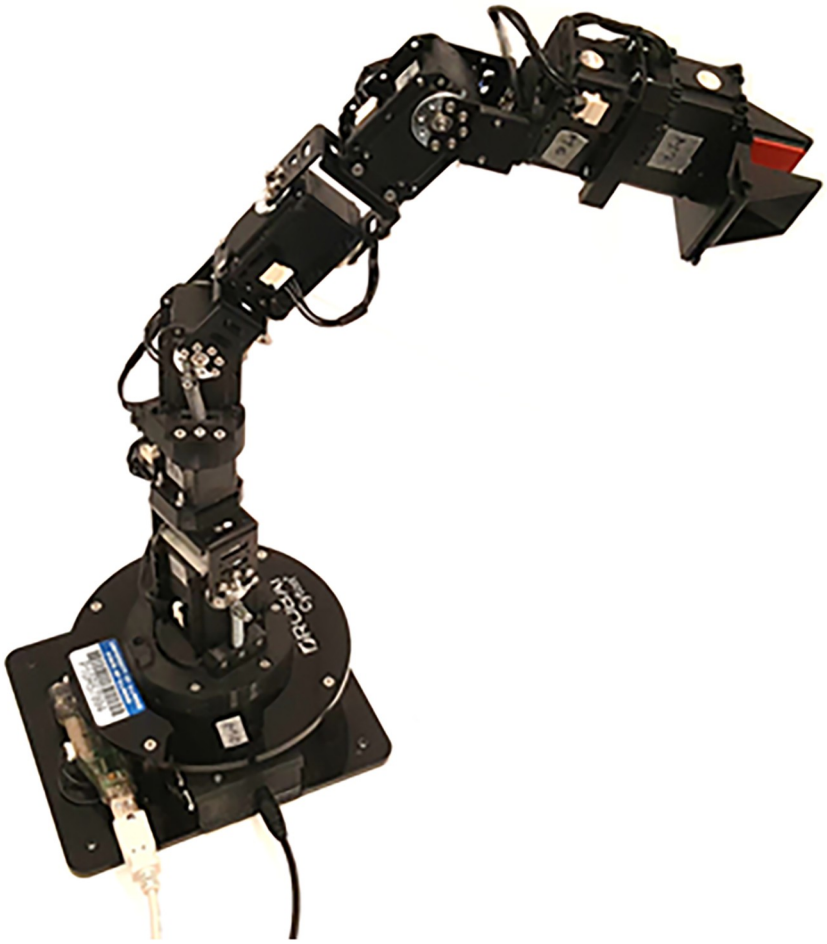

Fig 3. Robai Cyton 7-DOF spatial robot.

for ($n \leq 8$) and superior to it for ($n > 8$). Note that there is no definite way of comparing the performance for $n < 3$) since the expression of Renaud's method is valid only for ($n \geq 3$).

We demonstrated the real-time implementation using a 7-DOF redundant spatial manipulator (Robai Cyton Gamm-300) (Fig 3). We used Jacobian in the redundancy resolution of the manipulator by using an optimal control technique based on Hamiltonian formulation [7] for natural boundary conditions. That involves estimation of $\dot{\mathbf{J}}$. In our implementation, the end-effector of the manipulator is required to follow a circular trajectory with a 0.05 m radius and a frequency of 0.2 rad/s in the y-z plane at a distance of 0.2314 m from the origin along the x-axis. We implemented the differentiations required to find the Jacobian (as in Eq (10)) by using numerical differential based on the central difference method. The differential equations of the motion of the manipulator are numerically integrated using Runge-Kutta fourth order method. The real-time hardware implementation is done by modifying and appending to the open source C++ libraries for matrix operations, TNT and JAMA [23]. The measurments taken by the joint encoders were simultaneously recorded. We compared the results (i.e., the encoder measurements) (see Fig 4) with the simulation results using a second order semi-implicit integrator (ode23s), in MATLAB; ode23s is an implementation of the Bogacki-Shampine method [24]. An error tolerance of 10^{-9} was used in the integrator. We observed that the

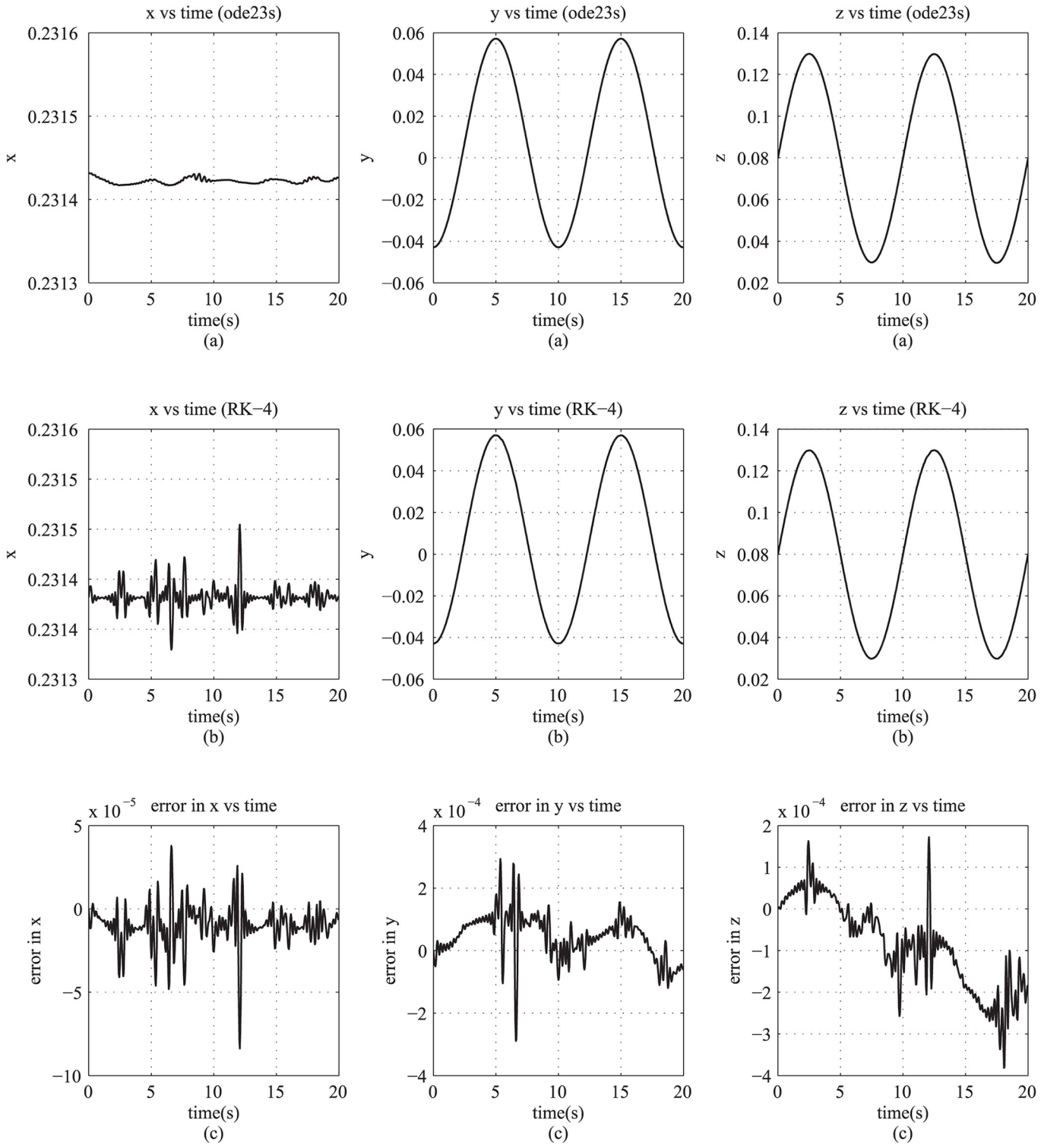

Fig 4. (row-1-(a)) x/y/z components of end-effector trajectory from simulation results using ode23s (Bogacki-Shampine method), (row-2-(b)) x/y/z components of end-effector trajectory from real-time physical implementation using RK-4, and (row-3-(c)) error in x/y/z using Runge-Kutta-4 for integration in real-time implementation.

hardware implementation and the simulation results matched, as is evident from the negigible errors (see third row of Fig 4). Thus, our approach is suitable for real-time implementation and causes negigible delay in the automatic formulation and estimation of the Jacobian.

7 Conclusion

A new implementation for online calculation of manipulator Jaobian has been presented and demonstrated. The method, based on matrix differential calculus, offers a systematic approach to calculating the Jacobian of robotic manipulators. Relative to the conventional methods, the calculation of Jacobian has been reduced to the inner product of two matrices. The matrix differentiations are performed using numerical methods. A real-time implementation of the Jacobian has been demonstrated using a planar two-link robot and a 7-DOF spatial robot. The errors in the hardware implementation of this method have been found to be negligible. Although it is computationally superior only for higher-DOF robots, the method is suitable for autonomous and real-time Jacobian estimations for robots with variable points. Our method is also well-suited for reconfigurable robots, which will be addressed in our future work.

Author Contributions

Conceptualization: Pramod Chembrammel.

Formal analysis: Pramod Chembrammel.

Investigation: Pramod Chembrammel.

Methodology: Pramod Chembrammel.

Resources: Pramod Chembrammel.

Software: Pramod Chembrammel.

Supervision: Thenkurussi Kesavadas.

Writing – review & editing: Thenkurussi Kesavadas.

References

1. Selig J.M. Intrductory Robotics. Prentice Hall; 1992.
2. Spong M.W., Hutchinson S., Vidyasagar M. Robot Modelling and Control John Wiley & Sons, Hoboken, New Jersey; 1970.
3. Craig J.J. Introduction to Robotics: Mechanics and Control. 3rd ed. Pearson Prentice Hall; 2005
4. Tsai L.W. Robot Analysis: The Mechanics of Serial and Parallel Manipulators. John Wiley & Sons, Inc.; 1999
5. Orin D.E., Schrader W.W. Efficient Computation of the Jacobian for Robot Manipulator International Journal of Robotics Research; 1984; 3(4):66–75 https://doi.org/10.1177/027836498400300404
6. Grood E.S., Suntay W. J. A Joint Coordinate System for the Clinical Description of Three-Dimensional Motions: Application to the Knee Transactions of ASME; 1983; 105:136–144

7. Kim S.W., Park K.B., Lee J.J. Redundancy Resolution of Robot Manipulators Using Optimal Kinematic Control Proceedings of International Conference on Robotics and Automation; 1994; 1; p.683-688
8. Gorner M., Stelzer A. A Leg Proprioception Based 6 DOF Oodometry for Statically Stable Walking Robots Autonomous Robots; 2013; 34(4); 311–326 https://doi.org/10.1007/s10514-013-9326-3
9. Komsuoglu H., McMordie D., Saranli U., Moore N., Buehler M., Koditschek D.E. Proprioception Based Behavioral Advances in a Hexapod Robot Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation; 2001; p. 3650-3655 vol.4.
10. Chitta S., Vemaza P., Geykhman R., Lee D.D. Proprioceptive Localization for a Quadrupedal Robot on Known Terrain Proceedings 2007 IEEE International Conference on Robotics and Automation; 2007; p.4582
11. Bruyninckx H., Schutter J.D. Symbolic differentiation of the velocity mapping for a serial kinematic chain Mechanism and Machine Theory; 1996; 31(2):135–148. https://doi.org/10.1016/0094-114X(95)00069-B
12. Ghoshal A. Robotics: Fundamental Concepts and Analysis Oxford University Press, India; 2006.
13. Mallik A.K., Ghosh A., Dittrich G. Kinematic Analysis and Synthesis of Mechanisms CRC Press; 1994.
14. Chembrammel P. Forward Dynamics of Lower Limb Based on Constrained Multibody Dynamics [MS Thesis]. State University of New York; 2013.
15. Renaud M. Geometric and Kinematic Models of a Robot Manipulator: Calculation of the Jacobian Matrix and its Inverse In: Proceedings of 11th International Symposium on Industrial Robots; 1981.
16. Zhang Y., Xiao L., Xiao Z., Mao M. Zeroing Dynamics, Gradient Dynamics, and Newton Iterations CRC Press; 2015.
17. Chen D., Zhang Y., Li S. Tracking Control of Robot Manipulators with Unknown Models: A Jacobian-Matrix-Adaption Method IEEE Transactions on Industrial Informatics; 2018; 14(7):3044–3053. https://doi.org/10.1109/TII.2017.2766455
18. Schilling R.J. Fundamentals of Robotics: Analysis and Control Prentice-Hall, New Jersey: 1990
19. Chen C.T. Linear System Theory and Design 3rd ed. Oxford University Press, New York; 1999.
20. Greenwood D.T. Principles of Dynamics Prentice-Hall, Inc., New Jersey; 1965.
21. Petersen K.B., Pedersen M.S. The Matrix Cookbook 2008
22. Chembrammel P., Younus H.M., Kesavadas T. Modelling and Simulation of Guide-Wire Interaction with Vasculature Using Constrained Multibody Dynamics In: Proceedings of ASME 2013 International Mechanical Engineering Congress and Exposition. ASME; 2013.
23. Template Numerical Toolkit.; http://math.nist.gov/tnt/overview.html
24. Bogacki P., Shampine L.F. A 3(2) Pair of Runge-Kutta Formulas Applied Mathematical Letters. 1989; 2(4):321–325. https://doi.org/10.1016/0893-9659(89)90079-7

Generalized nonlinear Schrödinger equations describing the Second Harmonic Generation of femtosecond pulse, containing a few cycles, and their integrals of motion

Vyacheslav A. Trofimov*, Svetlana Stepanenko, Alexander Razgulin

Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Moscow, Russia

* vatro@cs.msu.ru

Abstract

An interaction of laser pulse, containing a few cycles, with substance is a modern problem, attracting attention of many researches. The frequency conversion is a key problem for a generation of such pulses in various ranges of frequencies. Adequate description of such pulse interaction with a medium is based on a slowly evolving wave approximation (SEWA), which has been proposed earlier for a description of propagation of the laser pulse, containing a few cycles, in a medium with cubic nonlinear response. Despite widely applicability of the frequency conversion for various nonlinear optics problems solutions, SEWA has not been applied and developed for a theoretical investigation of the frequency doubling process until present time. In this study the set of generalized nonlinear Schrödinger equations describing a second harmonic generation of the super-short femtosecond pulse is derived. The equations set contains terms, describing the pulses self-steepening, and the second order dispersion (SOD) of the pulse, a diffraction of the beam as well as mixed derivatives. We propose the transform of the equations set to a type, which does not contain both the mixed derivatives and time derivatives of the nonlinear terms. This transform allows us to derive the integrals of motion of the problem: energy, spectral invariants and Hamiltonian. We show the existence of two specific frequencies (singularities in the Fourier space) inherent to the problem. They may cause an appearance of non-physical absolute instability of the problem solution if the spectral invariants are not taken into account. Moreover, we claim that the energy preservation at the laser pulses propagation may not occur if these invariants do not preserve. Developed conservation laws, in particular, have to be used for developing of the conservative finite-difference schemes, preserving the conservation laws difference analogues, and for developing of adequate theory of the modulation instability of the laser pulses, containing a few cycles.

Editor: Shou-Fu Tian, China University of Mining and Technology, CHINA

Funding: The investigation was made using support of the Russian Science Foundation (Grant No. 19-11-00113) to VAT.

Competing interests: The authors have declared that no competing interests exist.

Introduction

A well-known process of second harmonic generation (SHG) by the laser beam was the first nonlinear optical effect, which was observed by P. Franken et al. [1] in 1961. One year later, J. Armstrong, N. Bloembergen et al. [2] published the fundamental paper, dealing with the optical frequency conversion, in which the coupled nonlinear equations were developed and many explicit solutions of these equations were obtained in the framework of the plane wave approximation. During passed decades the SHG was widely observed: in plasma optics [3–5] and nonlinear optics, including high-harmonic generation [6–12], and in a medium containing nanoparticles [13, 14], and in semiconductor [15]. In [16] the SHG at the boundary between dielectric media is analyzed. Obviously, this list of papers can be easily increased.

Many researchers pay an attention to an investigation of various factors, which restrict the frequency doubling efficiency. For example, in [17] the effect of a group-velocity dispersion was considered. In [18] the authors have considered an influence of interplaying between two types polarizations of the pulse at the fundamental frequency on the frequency doubling efficiency. With increasing of the laser pulse intensity it is necessary to take into account a cubic nonlinear response of a medium under the SHG. An influence of the self-modulation as well as cross-modulation of the laser pulse on the frequency doubling efficiency was investigated experimentally and theoretically (see, for example [2, 19–28]). It should be stressed that the most general analytical theory of this process has been developed recently in [28], taking into account the time-dependent pulse shape and without using the basic wave energy non-depletion approximation on the base of using the integrals of motion (conservation laws) of the problem. We notice that the authors of [22] have obtained the most (but not all) solutions of this problem in the framework of long pulse duration approximation. In [29] an effective way for increasing of SHG efficiency in bulk medium was demonstrated by using of the incident beam with a ring (tubular) profile. This leads to re-profiling of the beams due to their diffraction and, therefore, to decreasing the beam phase distortions.

New opportunities appeared at SHG under the big phase-mismatching, so-called cascading SHG [30–33]. This process allows us to achieve an effective cubic nonlinear response in a medium with a quadratic nonlinear response. This effect was firstly predicted by Yu. N. Karamzin and A. P. Sukhorukov [34] in 1974. Later, it was used for a compression of the soliton with femtosecond duration [30, 34–50], and for the beam self-focusing [51–53], and for using the cascading SHG for suppression of the optical pulse intensity fluctuations, occurring in a sequence of pulses with randomly variations of their maximal intensities [54–57]. Such sequence of the laser pulses is produced by a laser, operating at the free generation mode. The total duration of the pulses sequences is equal to milliseconds. Therefore, the cascading SHG allows realizing the laser generation mode that is similar to the Kerr-locked mode applying for the femtosecond laser system.

There are a lot of other schemes of SHG, using the modern substances (see, for example, [58–84]) and at a present time the problem is still actual and important for various applications. Among these applications, we stress using of SHG for a measurement of pulse parameters and of parameters of a medium as well as for visualization of various fast processes, occurring in the substances. The most famous approaches of parameters measuring are SPIDER (spectral interferometry for direct electric-field reconstruction) [85] and FROG (frequency-resolved optical gating) [86, 87]. Also, in [88], a technique, based on combination of FROG spectra measurements was proposed to completely characterize the amplitude and phase of an ultrashort pulse in space and time. Mix FROG method was implemented in [89]. Thus, the frequency doubling of the optical pulse, containing a few cycles, is of practical interest until present time. However, in this case it is necessary to take into account the self-

steepening of the laser pulses on both fundamental and doubled frequencies. According to our knowledge, the first analysis of this problem was made numerically in [90] at the group velocity synchronism and phase matching, without taking into account the beam diffraction and the second order dispersion influence on the pulses interaction. For writing the corresponding equations, the authors have applied an approach, which was developed first by N. Tzoar and M. Jain [91] under the investigation of the single pulse propagation in a medium with a cubic nonlinear response. Experimental observation of the self-steepening influence on the pulse spectrum deformation was demonstrated in [92]. Then, this approach was successfully used for describing the processes of a laser pulse propagation in the optical fiber [93]. The corresponding nonlinear Schrödinger equation (NLSE) contains the first time derivative of a nonlinear medium response. This equation was derived using the slowly varying envelope approximation (SVEA) and named as generalized NLSE.

Taking into account a laser beam diffraction, T. Brabec and F. Krausz proposed in 1997 a new approach, so-called slowly evolving wave approximation (SEWA). Derived equation was also named as the generalized nonlinear Schrödinger equation (GNLSE) [94]. Let us note that the SEWA requires not only an envelope of a wave packet, but also the phase changing must slowly vary as the pulse covers a distance, which is equal to the wavelength. The main feature of GNLSE consists in its containing of mixed derivatives on time and spatial coordinates, and of the second-order time derivative of a medium nonlinear response (which describes the pulse self-steepening).

It should be emphasized that a differential operator of the GNLSE, written in [91, 93], coincides with a differential operator of the GNLSE, derived in [94] if the optical beam diffraction does not take into account. They differ only by the factor two at the coefficient characterizing the self-steepening term of the equation. Obviously, these equations written in the dimensionless units look the same.

Finding of conservation laws attracts an attention of various authors (see, for example, recently published papers [95–99]). In papers [100, 101] we derived the conservation laws for the frequency doubling process if the self-steepening of the pulses occurs. However, until present time, the set of GNLSEs describing a SHG of super-short femtosecond pulse with taking into account a diffraction of the laser beam (it means that this process is described in the framework of the SEWA) as well as their conservation laws were not derived. We do this in the present paper.

As is well-known, the invariants are very useful for computer simulation of the laser pulse propagation because the equations describing various problems of nonlinear optics are nonlinear ones and as a rule they do not allow us to find their analytical solution. As a consequence, the finite-difference time-domain (FDTD) method [102, 103] for computer simulation of a propagation of the optical pulse, containing a few cycles, is usually employed. For example, in [104] the authors develop the time-transformation method based on FDTD. An alternative method, which permits applying fast numerical algorithms, is a generalized source method (GSM) [105]. Linear GSM was adapted to describe the SHG in diffraction gratings, containing non-centrosymmetric materials [106]. Obviously, there are other papers, in which a computer simulation is applied for an investigation of the SHG problem. As a rule, a split-step method was used at computer simulation.

Another approach for computer simulation is based on developing of the conservative finite-difference schemes [107]. They allow preserving the difference analogues of the corresponding conservation laws. Obviously, it is necessary firstly to write these conservation laws. With this aim, we propose a new transform for the derived GNLSEs, which reduces the equations to the form containing neither mixed derivatives on time and spatial coordinates nor time derivatives of the nonlinear response of a medium. Using a new set of GNLSEs, we obtain

the energy invariant, spectral invariants and Hamiltonian. We add some specific requirements to the problem statement (excluding of the singularities in frequency space) to avoid a development of non-physical instability of the SHG process at taking into account the pulses self-steepening. These requirements, together with the spectral invariants, are also important for a validity of the energy conservation law.

Generalized nonlinear Schrödinger equations derivation

Derivation of the equations for SHG process in the framework of SEWA

Derivation of the equations describing the optical frequency doubling in the framework of SEWA for an electric field strength $E(z, x, t)$ starts from the well-known wave equation:

$$(\partial_z^2 + \nabla_\perp^2)E(z,x,t) - \frac{1}{c^2}\partial_t^2 D(z,x,t) = 0, \tag{1}$$

written above in 2D case, for example. The laser pulse propagates along z coordinate, t is a time, c is the light velocity in a vacuum, x is a transverse coordinate to the optical pulse propagation direction, an operator $\nabla_\perp^2 = \partial_x^2$ is the transversal Laplace operator. We choose only x-coordinate as a transverse one for simplicity. This does not restrict our consideration. The function $D(z, x, t)$ describes an electric field induction and is written in the form:

$$D(z,x,t) = E(z,x,t) + 4\pi P(z,x,t), \tag{2}$$

the function $P(z, x, t)$ is a medium polarization, containing its linear and nonlinear responses:

$$P(z,x,t) = P_{lin}(z,x,t) + P_{nl}(z,x,t). \tag{3}$$

For linear non-instantaneous medium response $P_{lin}(z, x, t)$ the following representation

$$P_{lin}(z,x,t) = \int_0^{+\infty} \chi^{(1)}(t')E(z,x,t-t')dt' \tag{4}$$

is widely used, which yields in the relation for a dielectric permittivity:

$$\varepsilon(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \tilde{\varepsilon}(\omega)\exp(-i\omega t)d\omega, \;\; \tilde{\varepsilon}(\omega) = 1 + 4\pi\tilde{\chi}^{(1)}(\omega), \tag{5}$$

here $\tilde{\chi}^{(1)}(\omega)$ is a linear electric susceptibility of a medium at the frequency ω.

Function $P_{nl}(z, x, t)$ describes a nonlinear response of a medium. If we consider a medium with the quadratic nonlinearity, then this function is defined as $P_{nl} = \chi^{(2)}E^2$, and $\chi^{(2)}$ is a quadratic electric susceptibility.

The next step for the wave equation reduction consists in representation of an electric field strength $E(z, x, t)$ in a following manner:

$$\begin{aligned} E(z,x,t) = E_1(z,x,t) + E_2(z,x,t) = \frac{1}{2}(A_1(z,x,t)\exp[-i(\omega_o t - k_1 z)] + \\ + A_2(z,x,t)\exp[-i(2\omega_o t - k_2 z)] + c.c.), \end{aligned} \tag{6}$$

where $E_1(z, x, t)$, $E_2(z, x, t)$ are the electric field strengths of waves, propagating with carrier frequency ω_o, and with doubled frequency $2\omega_o$ and with wave-numbers k_1, k_2, correspondingly. In (6) we introduce the slowly varying envelopes for both wave packets and denote them as $A_1(z, x, t)$ and $A_2(z, x, t)$. Symbols *c.c.* denote a complex conjugation. The linear polarization of

a medium at these frequencies can be written as

$$P_{lin}(z,x,t) = P_{lin_1}(z,x,t) + P_{lin_2}(z,x,t) = \frac{1}{2}(L_1(z,x,t)\exp[-i(\omega_o t - k_1 z)] + \\ + L_2(z,x,t)\exp[-i(2\omega_o t - k_2 z)] + c.c.). \tag{7}$$

The corresponding nonlinear polarization of a medium at chosen frequencies is written as follows

$$P_{nl}(z,x,t) = \frac{\chi^{(2)}}{4}(A_1^2\exp[-i(2\omega_o t - (2k_1 - k_2)z - k_2 z)] + \\ + 2A_1^* A_2\exp[-i(\omega_o t + (2k_1 - k_2)z - k_1 z)] + c.c.). \tag{8}$$

Let us remind that the following relations for wave-numbers are valid:

$$k_1 = n_\omega(\omega_o/c) = \sqrt{\tilde{\varepsilon}(\omega_o)}(\omega_o/c),$$
$$k_2 = n_{2\omega}(2\omega_o/c) = \sqrt{\tilde{\varepsilon}(2\omega_o)}(2\omega_o/c),$$

where n_ω, $n_{2\omega}$ are the medium refractive indexes at the frequencies ω_o and $2\omega_o$, correspondingly.

Taking into account the expressions (6) and (8) and (1) transforms to the following form:

$$\frac{1}{2}\exp(ik_1 z)\left[\left(\frac{\partial^2 A_1}{\partial z^2} + 2ik_1\frac{\partial A_1}{\partial z} - k_1^2 A_1 + \frac{\partial^2 A_1}{\partial x^2}\right)\exp(-i\omega_o t) - \right. \\ \left. - \frac{1}{c^2}\partial_t^2(D_1(z,x,t) + 4\pi\chi^{(2)}A_1^* A_2\exp[-i(\omega_o t + (2k_1 - k_2)z)])\right] + \\ + \frac{1}{2}\exp(ik_2 z)\left[\left(\frac{\partial^2 A_2}{\partial z^2} + 2ik_2\frac{\partial A_2}{\partial z} - k_2^2 A_2 + \frac{\partial^2 A_2}{\partial x^2}\right)\exp(-2i\omega_o t) - \right. \\ \left. - \frac{1}{c^2}\partial_t^2(D_2(z,x,t) + 2\pi\chi^{(2)}A_1^2\exp[-i(2\omega_o t - (2k_1 - k_2)z)])\right] + c.c. = 0, \tag{9}$$

where $D_j(z, x, t)$, $j = 1, 2$ is a linear part of the electric field induction at the frequencies ω_o and $2\omega_o$, correspondingly:

$$D_j(z,x,t) = (A_j(z,x,t) + 4\pi L_j(z,x,t))\exp(-ij\omega_o t). \tag{10}$$

Let us multiply (9) by $\exp(-ik_1 z)$ or $\exp(-ik_2 z)$. Then, we integrate each of these expressions over the corresponding wave period, taking into account the orthogonal property of the

sine and cosine functions. Then, (9) can be reduced to the following equations:

$$\begin{aligned}\left(\frac{\partial^2 A_1}{\partial z^2}+2ik_1\frac{\partial A_1}{\partial z}-k_1^2A_1+\frac{\partial^2 A_1}{\partial x^2}+\right.\\ \left.+\frac{4\pi\chi^{(2)}\omega_o^2}{c^2}\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial t}\right)^2 A_1^* A_2\exp\left[-i(2k_1-k_2)z\right]\right)\exp\left(-i\omega_o t\right)-\\ -\frac{1}{c^2}\partial_t^2 D_1(z,x,t)=0,\\ \left(\frac{\partial^2 A_2}{\partial z^2}+2ik_2\frac{\partial A_2}{\partial z}-k_2^2A_2+\frac{\partial^2 A_2}{\partial x^2}+\right.\\ \left.+\frac{2\pi\chi^{(2)}(2\omega_o)^2}{c^2}\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial t}\right)^2 A_1^2\exp\left[i(2k_1-k_2)z\right]\right)\exp\left(-2i\omega_o t\right)-\\ -\frac{1}{c^2}\partial_t^2 D_2(z,x,t)=0.\end{aligned} \tag{11}$$

Let us note the Fourier transform property:

$$\begin{aligned}\partial_t^2 D_j(z,x,t)=\frac{1}{2\pi}\partial_t^2\int_{-\infty}^{+\infty}\tilde{D}_j(z,x,\omega)\exp\left(-i\omega t\right)d\omega=\\ =-\frac{1}{2\pi}\int_{-\infty}^{+\infty}\omega^2\tilde{D}_j(z,x,\omega)\exp\left(-i\omega t\right)d\omega,\quad j=1,2.\end{aligned} \tag{12}$$

Therefore, (11) can be written as:

$$\begin{aligned}\left(\frac{\partial^2 A_1}{\partial z^2}+2ik_1\frac{\partial A_1}{\partial z}-k_1^2A_1+\frac{\partial^2 A_1}{\partial x^2}+\right.\\ \left.+\frac{4\pi\chi^{(2)}\omega_o^2}{c^2}\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial t}\right)^2 A_1^* A_2\exp\left[-i(2k_1-k_2)z\right]\right)\exp\left(-i\omega_o t\right)+\\ +\frac{1}{2\pi}\exp\left(-i\omega_o t\right)\int_{-\infty}^{+\infty}\frac{\omega^2}{c^2}\tilde{\varepsilon}(\omega)\tilde{A}_1(z,x,\omega-\omega_o)\exp\left[-i(\omega-\omega_o)t\right]d\omega=0,\\ \left(\frac{\partial^2 A_2}{\partial z^2}+2ik_2\frac{\partial A_2}{\partial z}-k_2^2A_2+\frac{\partial^2 A_2}{\partial x^2}-\right.\\ \left.+\frac{2\pi\chi^{(2)}(2\omega_o)^2}{c^2}\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial t}\right)^2 A_1^2\exp\left[i(2k_1-k_2)z\right]\right)\exp\left(-2i\omega_o t\right)+\\ +\frac{1}{2\pi}\exp\left(-2i\omega_o t\right)\int_{-\infty}^{+\infty}\frac{\omega^2}{c^2}\tilde{\varepsilon}(\omega)\tilde{A}_2(z,x,\omega-2\omega_o)\exp\left[-i(\omega-2\omega_o)t\right]d\omega=\\ =0.\end{aligned} \tag{13}$$

Taking into account the dispersion relation:

$$k^2(\omega)=\frac{\omega^2}{c^2}\tilde{\varepsilon}(\omega), \tag{14}$$

one can write (13) as follows

$$\left(\frac{\partial^2 A_1}{\partial z^2}+2ik_1\frac{\partial A_1}{\partial z}-k_1^2A_1+\frac{\partial^2 A_1}{\partial x^2}+\right.$$
$$\left.+\frac{4\pi\chi^{(2)}\omega_o^2}{c^2}\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial t}\right)^2 A_1^* A_2\exp\left[-i(2k_1-k_2)z\right]\right)\exp\left(-i\omega_o t\right)+ \tag{15}$$
$$+\frac{1}{2\pi}\exp\left(-i\omega_o t\right)\int_{-\infty}^{+\infty}k^2(\omega)\tilde{A}_1(z,x,\omega-\omega_o)\exp\left[-i(\omega-\omega_o)t\right]d\omega=0,$$

$$\left(\frac{\partial^2 A_2}{\partial z^2}+2ik_2\frac{\partial A_2}{\partial z}-k_2^2A_2+\frac{\partial^2 A_2}{\partial x^2}+\right.$$
$$\left.+\frac{2\pi\chi^{(2)}(2\omega_o)^2}{c^2}\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial t}\right)^2 A_1^2\exp\left[i(2k_1-k_2)z\right]\right)\exp\left(-2i\omega_o t\right)+ \tag{16}$$
$$+\frac{1}{2\pi}\exp\left(-2i\omega_o t\right)\int_{-\infty}^{+\infty}k^2(\omega)\tilde{A}_2(z,x,\omega-2\omega_o)\exp\left[-i(\omega-2\omega_o)t\right]d\omega=$$
$$=0.$$

Normally, we expand $k(\omega)$ in a series near the frequency ω_o in (15):

$$k^2(\omega)=\left(k_1+\sum_{n=1}^{\infty}\frac{\partial^{(n)}k(\omega)}{\partial\omega^n}\bigg|_{\omega=\omega_o}\frac{(\omega-\omega_o)^n}{n!}\right)^2=$$
$$=\left(k_1+\beta_1(\omega-\omega_o)+\frac{\beta_2}{2}(\omega-\omega_o)^2+\sum_{n=3}^{\infty}\frac{\beta_n(\omega-\omega_o)^n}{n!}\right)^2 \tag{17}$$

and near the frequency $2\omega_o$ in (16):

$$k^2(\omega)=\left(k_2+\sum_{m=1}^{\infty}\frac{\partial^{(m)}k(\omega)}{\partial\omega^m}\bigg|_{\omega=2\omega_o}\frac{(\omega-2\omega_o)^m}{m!}\right)^2=$$
$$=\left(k_2+s_1(\omega-2\omega_o)+\frac{s_2}{2}(\omega-2\omega_o)^2+\sum_{m=1}^{\infty}\frac{s_m(\omega-2\omega_o)^m}{m!}\right)^2, \tag{18}$$

here β_n and s_m are the n–th or m–th order derivatives of the wave number at the frequencies ω_o and $2\omega_o$, correspondingly.

Let us neglect the terms β_n, s_m, n, m = 3, 4, . . . referring to the third order dispersion and other higher terms in (17) and (18). Also, we neglect the term $(\omega-\omega_o)^4$. Thus, (17) and (18) take the forms:

$$k^2(\omega)=k_1^2+2k_1\beta_1(\omega-\omega_o)+k_1\beta_2(\omega-\omega_o)^2+\beta_1^2(\omega-\omega_o)^2+\beta_1\beta_2(\omega-\omega_o)^3, \tag{19}$$

$$k^2(\omega)=k_2^2+2k_2s_1(\omega-2\omega_o)+k_2s_2(\omega-2\omega_o)^2+$$
$$+s_1^2(\omega-2\omega_o)^2+s_1s_2(\omega-2\omega_o)^3. \tag{20}$$

After Inverse Fourier transform, we obtain the following equations set for describing the frequency doubling process:

$$\begin{aligned}&((ik_1+\frac{\partial}{\partial z})^2+\frac{\partial^2}{\partial x^2})A_1+\\&\quad+\left(k_1^2+2ik_1\beta_1\frac{\partial}{\partial t}-k_1\beta_2\frac{\partial^2}{\partial t^2}-\beta_1^2\frac{\partial^2}{\partial t^2}-i\beta_1\beta_2\frac{\partial^3}{\partial t^3}\right)A_1+\\&\quad+\frac{4\pi\omega_o^2\chi^{(2)}}{c^2}\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial t}\right)^2A_1^*A_2\exp(i\Delta kz)=0,\end{aligned}\tag{21}$$

$$\begin{aligned}&((ik_2+\frac{\partial}{\partial z})^2+\frac{\partial^2}{\partial x^2})A_2+\\&\quad+\left(k_2^2+2ik_2s_1\frac{\partial}{\partial t}-k_2s_2\frac{\partial^2}{\partial t^2}-s_1^2\frac{\partial^2}{\partial t^2}-is_1s_2\frac{\partial^3}{\partial t^3}\right)A_2+\\&\quad+\frac{2\pi(2\omega_o)^2\chi^{(2)}}{c^2}\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial t}\right)^2A_1^2\exp(-i\Delta kz)=0.\end{aligned}\tag{22}$$

Above $\Delta k = k_2 - 2k_1$ is the phase mismatch. (21) and (22) are called as the set of generalized nonlinear Schrödinger equations (GNLSEs).

Coordinate system transform

First type of coordinate system transform. In a coordinate system moving with the first wave packet

$$\xi = z,\quad \tau = t-\beta_1 z,\tag{23}$$

(21) and (22) could be re-written as follows:

$$\begin{aligned}&\left(\left(ik_1+\frac{\partial}{\partial\xi}-\beta_1\frac{\partial}{\partial\tau}\right)^2+\frac{\partial^2}{\partial x^2}\right)A_1+\\&\quad+\left(k_1^2+2ik_1\beta_1\frac{\partial}{\partial\tau}-k_1\beta_2\frac{\partial^2}{\partial\tau^2}-\beta_1^2\frac{\partial^2}{\partial\tau^2}-i\beta_1\beta_2\frac{\partial^3}{\partial\tau^3}\right)A_1+\\&\quad+\frac{4\pi(\omega_o)^2\chi^{(2)}}{c^2}\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial\tau}\right)^2A_1^*A_2\exp(i\Delta k\xi)=0,\end{aligned}\tag{24}$$

$$\begin{aligned}&\left(\left(ik_2+\frac{\partial}{\partial\xi}-\beta_1\frac{\partial}{\partial\tau}\right)^2+\frac{\partial^2}{\partial x^2}\right)A_2+\\&\quad+\left(k_2^2+2ik_2s_1\frac{\partial}{\partial\tau}-k_2s_2\frac{\partial^2}{\partial\tau^2}-s_1^2\frac{\partial^2}{\partial\tau^2}-is_1s_2\frac{\partial^3}{\partial\tau^3}\right)A_2+\\&\quad+\frac{2\pi(2\omega_o)^2\chi^{(2)}}{c^2}\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial\tau}\right)^2A_1^2\exp(-i\Delta k\xi)=0.\end{aligned}\tag{25}$$

Let us raise to the second power the expressions in brackets and then neglect the terms with the second order derivative on ξ – coordinate:

$$\left(2ik_1\frac{\partial}{\partial\xi}-2i\beta_1k_1\frac{\partial}{\partial\tau}-2\beta_1\frac{\partial^2}{\partial\xi\partial\tau}+\frac{\partial^2}{\partial x^2}\right)A_1+$$
$$+\left(2ik_1\beta_1\frac{\partial}{\partial\tau}-k_1\beta_2\frac{\partial^2}{\partial\tau^2}-\beta_1^2\frac{\partial^2}{\partial\tau^2}-i\beta_1\beta_2\frac{\partial^3}{\partial\tau^3}\right)A_1+ \tag{26}$$
$$+\frac{4\pi(\omega_o)^2\chi^{(2)}}{c^2}\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial\tau}\right)^2A_1^*A_2\exp(i\Delta k\xi)=0,$$

$$\left(2ik_2+\frac{\partial}{\partial\xi}-2i\beta_1k_2\frac{\partial}{\partial\tau}-2\beta_1\frac{\partial^2}{\partial\xi\partial\tau}+\frac{\partial^2}{\partial x^2}\right)A_2+$$
$$+\left(k_2^2+2ik_2s_1\frac{\partial}{\partial\tau}-k_2s_2\frac{\partial^2}{\partial\tau^2}-s_1^2\frac{\partial^2}{\partial\tau^2}-is_1s_2\frac{\partial^3}{\partial\tau^3}\right)A_2+ \tag{27}$$
$$+\frac{2\pi(2\omega_o)^2\chi^{(2)}}{c^2}\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial\tau}\right)^2A_1^2\exp(-i\Delta k\xi)=0.$$

By grouping the terms in (26) and (27) and dividing them by $2ik_j$, $j = 1, 2$, we obtain:

$$\left(1+\frac{i\beta_1}{k_1}\frac{\partial}{\partial\tau}\right)\frac{\partial A_1}{\partial\xi}+\frac{i\beta_2}{2}\left(1+\frac{i\beta_1}{k_1}\frac{\partial}{\partial\tau}\right)\frac{\partial^2A_1}{\partial\tau^2}-\frac{i}{2k_1}\frac{\partial^2A_1}{\partial x^2}-$$
$$-i\frac{2\pi\omega_o\chi^{(2)}}{cn_\omega}\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial\tau}\right)^2A_1^*A_2\exp(i\Delta k\xi)=0, \tag{28}$$

$$\left(1+\frac{i\beta_1}{k_2}\frac{\partial}{\partial\tau}\right)\frac{\partial A_2}{\partial\xi}-(\beta_1-s_1)\left(1+\frac{i}{2}\left(\frac{\beta_1+s_1}{k_2}\right)\frac{\partial}{\partial\tau}\right)\frac{\partial A_2}{\partial\tau}+$$
$$+\frac{is_2}{2}\left(1+i\frac{s_1}{k_2}\frac{\partial}{\partial\tau}\right)\frac{\partial^2A_2}{\partial\tau^2}-\frac{i}{2k_2}\frac{\partial^2A_2}{\partial x^2}- \tag{29}$$
$$-i\frac{2\pi\omega_o\chi^{(2)}}{cn_{2\omega}}\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial\tau}\right)^2A_1^2\exp(-i\Delta k\xi)=0.$$

The set of Eqs (28) and (29) describes the SHG process for the pulse with super-short duration, which is about a few femtoseconds.

Further reduction of these equations is a consequence of the following assumptions.

1. The group velocities of pulses and their phase velocities differ insignificantly:

$$\left|\beta_1-\frac{k_1}{\omega_o}\right|\ll1, \tag{30}$$

$$\left|s_1-\frac{k_2}{2\omega_o}\right|\ll1. \tag{31}$$

2. Difference between inverse values of the phase velocities of waves is many times less than unity:

$$\left|\frac{k_2}{2\omega_o}-\frac{k_1}{\omega_o}\right|=\left|\frac{\Delta k}{2\omega_o}\right|\ll 1. \tag{32}$$

As a result, (28) written with respect to the wave for basic frequency ω_o, reduces to

$$\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial\tau}\right)\left(\frac{\partial A_1}{\partial\xi}+\frac{i\beta_2}{2}\frac{\partial^2 A_1}{\partial t^2}\right)- \\ -\frac{i}{2k_1}\frac{\partial^2 A_1}{\partial x^2}-\frac{2\pi\omega_o\chi^{(2)}i}{cn_\omega}\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial\tau}\right)^2 A_1^* A_1\exp\left(i\Delta k\xi\right)=0, \tag{33}$$

and (29) takes the following form:

$$\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial\tau}\right)\left(\frac{\partial A_2}{\partial\xi}-(\beta_1-s_1)\frac{\partial A_2}{\partial t}+\frac{is_2}{2}\frac{\partial^2 A_2}{\partial t^2}\right)- \\ -\frac{i}{2k_2}\frac{\partial^2 A_2}{\partial x^2}-\frac{2\pi\omega_o\chi^{(2)}i}{cn_{2\omega}}\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial\tau}\right)^2 A_1^2\exp\left(-i\Delta k\xi\right)=0. \tag{34}$$

Second type of coordinate system transform. In a coordinate system moving with overage velocity

$$\xi=z,\quad \tau=t-\frac{\beta_1+s_1}{2}z, \tag{35}$$

the set of Eqs (21) and (22) could be re-written as follows:

$$\left(1+\frac{i(\beta_1+s_1)}{2k_1}\frac{\partial}{\partial\tau}\right)\frac{\partial A_1}{\partial\xi}+ \\ +\frac{(\beta_1-s_1)}{2}\left(1+\frac{i(3\beta_1+s_1)}{4k_1}\frac{\partial}{\partial\tau}\right)\frac{\partial A_1}{\partial\tau}+ \\ +\frac{i\beta_2}{2}\left(1+\frac{i\beta_1}{k_1}\frac{\partial}{\partial\tau}\right)\frac{\partial^2 A_1}{\partial\tau^2}-\frac{i}{2k_1}\frac{\partial^2 A_1}{\partial x^2}- \\ -\frac{2\pi\omega_o\chi^{(2)}i}{cn_\omega}\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial\tau}\right)^2 A_1^* A_2\exp\left(i\Delta k\xi\right)=0, \tag{36}$$

$$\left(1+\frac{i(\beta_1+s_1)}{2k_2}\frac{\partial}{\partial\tau}\right)\frac{\partial A_2}{\partial\xi}- \\ -\frac{(\beta_1-s_1)}{2}\left(1+\frac{i(3s_1+\beta_1)}{4k_2}\frac{\partial}{\partial\tau}\right)\frac{\partial A_2}{\partial\tau}+ \\ +\frac{is_2}{2}\left(1+\frac{is_1}{k_2}\frac{\partial}{\partial\tau}\right)\frac{\partial^2 A_2}{\partial\tau^2}-\frac{i}{2k_2}\frac{\partial^2 A_2}{\partial x^2}- \\ -\frac{2\pi\omega_o\chi^{(2)}i}{cn_{2\omega}}\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial\tau}\right)^2 A_1^2\exp\left(-i\Delta k\xi\right)=0. \tag{37}$$

If the assumptions (30)–(32) are valid, then (36), with respect to basic frequency ω_o, reduces to

$$\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial\tau}\right)\frac{\partial A_1}{\partial\xi}+\frac{(\beta_1-s_1)}{2}\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial\tau}\right)\frac{\partial A_1}{\partial t}+ \\ +\frac{i\beta_2}{2}\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial\tau}\right)\frac{\partial^2 A_1}{\partial t^2}-\frac{i}{2k_1}\frac{\partial^2 A_1}{\partial x^2}- \\ -\frac{2\pi\omega_o\chi^{(2)}i}{cn_\omega}\left(1+\frac{i}{\omega_o}\frac{\partial}{\partial\tau}\right)^2 A_1^* A_1\exp(i\Delta k\xi)=0, \tag{38}$$

and (37), with respect to doubled frequency $2\omega_o$. takes the following form:

$$\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial\tau}\right)\frac{\partial A_2}{\partial\xi}-\frac{(\beta_1-s_1)}{2}\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial\tau}\right)\frac{\partial A_2}{\partial t}+ \\ +\frac{is_2}{2}\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial\tau}\right)\frac{\partial^2 A_2}{\partial t^2}-\frac{i}{2k_2}\frac{\partial^2 A_2}{\partial x^2}- \\ -\frac{2\pi\omega_o\chi^{(2)}i}{cn_{2\omega}}\left(1+\frac{i}{2\omega_o}\frac{\partial}{\partial\tau}\right)^2 A_1^2\exp(-i\Delta k\xi)=0. \tag{39}$$

Below we use the First type of Coordinate system transform.

Dimensionless variables and problem statement

Let us introduce dimensionless variables:

$$\xi\to\frac{\xi}{l_{dif}},\quad x\to\frac{x}{a},\quad t\to\frac{t}{\tau_p}, \tag{40}$$

$$A_j\to\frac{A_j}{\sqrt{I_o}},\quad \Delta k\to\Delta k l_{dif},\quad l_{dif}=2k_1a^2. \tag{41}$$

Above the parameter τ_p is a duration of the incident pulse for basic frequency ω_o, I_o is its maximal intensity, the parameter a is a beam radius.

In new variables, the GNLSE set (33) and (34) takes the following form:

$$\left(1+2i\gamma\frac{\partial}{\partial t}\right)\frac{\partial A_1}{\partial\xi}+iD_{21}\left(1+2i\gamma\frac{\partial}{\partial t}\right)\frac{\partial^2 A_1}{\partial t^2}+iD\frac{\partial^2 A_1}{\partial x^2}+ \\ +i\alpha\left(1+2i\gamma\frac{\partial}{\partial t}\right)^2 A_1^* A_2\exp(i\Delta k\xi)=0, \\ \left(1+i\gamma\frac{\partial}{\partial t}\right)\frac{\partial A_2}{\partial\xi}+\nu\left(1+i\gamma\frac{\partial}{\partial t}\right)\frac{\partial A_2}{\partial t}+iD_{22}\left(1+i\gamma\frac{\partial}{\partial t}\right)\frac{\partial^2 A_2}{\partial t^2}+ \\ +\frac{iD}{2}\frac{\partial^2 A_2}{\partial x^2}+i\alpha\left(1+i\gamma\frac{\partial}{\partial t}\right)^2 A_1^2\exp(-i\Delta k\xi), \tag{42}$$

which is considered in the following domain

$$(\xi,x,t)\in\Omega=\Omega_\xi\times\Omega_o,\quad \Omega_\xi=(0,L_\xi], \\ \Omega_o=\Omega_x\times\Omega_t=(0,L_x)\times(0,L_t),$$

with the boundary conditions (BCs) for the functions $A_j(\xi, x, t), j = 1, 2$:

$$A_j\Big|_{\substack{x=0,L_x \\ (\xi,t)\in\Omega_\xi\times\Omega_t}} = A_j\Big|_{\substack{t=0,L_t \\ (\xi,x)\in\Omega_\xi\times\Omega_x}} = \frac{\partial A_j}{\partial t}\Big|_{\substack{t=0 \\ (\xi,x)\in\Omega_\xi\times\Omega_x}} = 0, \tag{43}$$

or

$$A_j\Big|_{\substack{x=0,L_x \\ (\xi,t)\in\Omega_\xi\times\Omega_t}} = A_j\Big|_{\substack{t=0,L_t \\ (\xi,x)\in\Omega_\xi\times\Omega_x}} = \frac{\partial A_j}{\partial t}\Big|_{\substack{t=L_t \\ (\xi,x)\in\Omega_\xi\times\Omega_x}} = 0, \tag{44}$$

and with the initial conditions

$$A_j\Big|_{\substack{\xi=0 \\ (x,t)\in\Omega_x\times\Omega_t}} = A_{oj}(x,t). \tag{45}$$

Above γ is a parameter, which is inversely proportional to the doubled frequency $2\omega_o$ and the pulse duration τ_p:

$$\gamma = \frac{1}{2\omega_o\tau_p}. \tag{46}$$

The parameter ν describes group-velocity mismatch (GVM) normalized by pulse duration τ_p:

$$\nu = \frac{s_1 - \beta_1}{\tau_p} l_{dif}. \tag{47}$$

Parameters D_{21}, D_{22} are equal to a ratio between the diffraction length of the beam and a dispersion length of the pulse for each wave:

$$D_{21} = -\frac{l_{dif}}{l_{dis_1}} sign\left(\frac{\partial^2 k}{\partial\omega^2}\right)\Big|_{\omega=\omega_o} = -\frac{k_1 a^2}{\tau_p^2 \left|\frac{\partial^2 k}{\partial\omega^2}\right|^{-1}\Big|_{\omega=\omega_o}} sign\left(\frac{\partial^2 k}{\partial\omega^2}\right)\Big|_{\omega=\omega_o}, \tag{48}$$

$$D_{22} = -\frac{l_{dif}}{l_{dis_2}} sign\left(\frac{\partial^2 k}{\partial\omega^2}\right)\Big|_{\omega=2\omega_o} = -\frac{k_1 a^2}{\tau_p^2 \left|\frac{\partial^2 k}{\partial\omega^2}\right|^{-1}\Big|_{\omega=2\omega_o}} sign\left(\frac{\partial^2 k}{\partial\omega^2}\right)\Big|_{\omega=2\omega_o}. \tag{49}$$

They characterize the second order dispersion of a wave packet. Parameter D is equal to unity for chosen normalization of spatial coordinate along which the laser pulse propagates. Parameter α describes the nonlinear coupling of waves and is defined in the following form

$$\alpha = \frac{l_{dif}}{l_{nl}} = \frac{4\pi k_1 a^2 \chi^{(2)}\sqrt{I_o}}{c n_\omega}, \tag{50}$$

which is written using the following assumption:

$$\frac{\chi^{(2)}(\omega)}{n_\omega} = \frac{\chi^{(2)}(2\omega)}{n_{2\omega}}, \tag{51}$$

which does not restrict our consideration. Let us note that if the equality (51) is not valid, then one can transform the equations with respect to new variables, in which we obtain only single coefficient (it characterizes the nonlinear coupling of waves for both equations, see [21]).

Usually, at theoretical analysis of the problem (42), the exponentially decaying functions $A_j(\xi, x, t), j = 1, 2$ are used as the initial laser pulse distributions:

$$A_j\Big|_{\substack{x\to\pm\infty \\ (\xi,t)\in\tilde{\Omega}_\xi\times\tilde{\Omega}_t}} = \frac{\partial^{(n)}A_j}{\partial x^n}\Big|_{\substack{x\to\pm\infty \\ (\xi,t)\in\tilde{\Omega}_\xi\times\tilde{\Omega}_t}} \to 0, \qquad A_j\Big|_{\substack{t\to\pm\infty \\ (\xi,x)\in\tilde{\Omega}_\xi\times\tilde{\Omega}_x}} = \frac{\partial^{(n)}A_j}{\partial t^n}\Big|_{\substack{t\to\pm\infty \\ (\xi,x)\in\tilde{\Omega}_\xi\times\tilde{\Omega}_x}} \to 0, \tag{52}$$

and instead of the BCs (43) and (44) the set of equations (42) is solved in the following unbounded (increasing of $|x|$, $|t|$ and z) domain

$$(\xi, x, t) \in \tilde{\Omega} = \tilde{\Omega}_\xi \times \tilde{\Omega}_x \times \tilde{\Omega}_t,$$
$$\tilde{\Omega}_\xi = (0, +\infty), \tilde{\Omega}_x = \tilde{\Omega}_t = (-\infty, +\infty).$$

Below, for definiteness, we use BCs (43). In addition, for convenience, below we use the previous notation for a longitudinal coordinate: z instead of ξ.

Equations transform

For transforming the equations set (42) to the form, which does not contain a time derivative of the nonlinear response of medium and mixed derivatives on time and spatial coordinate, we use the following equalities:

$$\left(1 + ij\gamma\frac{\partial}{\partial t}\right)g = ij\gamma\exp\left(\frac{it}{j\gamma}\right)\frac{\partial}{\partial t}\left(g\exp\left(-\frac{it}{j\gamma}\right)\right), \quad j = 1, 2. \tag{53}$$

Also, due to a finiteness of the initial distributions of complex amplitudes and boundedness of the laser pulse propagation distance, let us state the following additional conditions:

$$\frac{\partial^2 A_j}{\partial t^2}\Big|_{\substack{t=0 \\ (z,x)\in\Omega_z\times\Omega_x}} = 0. \tag{54}$$

Using the transforms (53) and condition (54), GNLSE set (42) can be written as follows (see detailed derivation in Appendix [A]):

$$\frac{\partial A_1}{\partial z} + iD_{21}\frac{\partial^2 A_1}{\partial t^2} + \frac{D}{2\gamma}\exp\left(\frac{it}{2\gamma}\right)\frac{\partial^2}{\partial x^2}\left(\int_0^t A_1(z, x, \eta)\exp\left(-\frac{i\eta}{2\gamma}\right)d\eta\right) + i\alpha\left(1 + 2i\gamma\frac{\partial}{\partial t}\right)(A_1^* A_2)\exp(i\Delta kz) = 0, \tag{55}$$

$$\frac{\partial A_2}{\partial z} + v\frac{\partial A_2}{\partial t} + iD_{22}\frac{\partial^2 A_2}{\partial t^2} + \frac{D}{2\gamma}\exp\left(\frac{it}{\gamma}\right)\frac{\partial^2}{\partial x^2}\left(\int_0^t A_2(z, x, \eta)\exp\left(-\frac{i\eta}{\gamma}\right)d\eta\right) + i\alpha\left(1 + i\gamma\frac{\partial}{\partial t}\right)A_1^2\exp(-i\Delta kz) = 0. \tag{56}$$

It should be stressed that these equations will be also used for writing the conservation laws. For further reduction of the equations, let us introduce the functions:

$$P_j(z, x, t) = \int_0^t A_j(z, x, \eta)\exp\left(-\frac{ij\eta}{2\gamma}\right)d\eta, j = 1, 2 \tag{57}$$

or

$$\frac{\partial P_j}{\partial t} = A_j \exp\left(-\frac{ijt}{2\gamma}\right), P_j\Big|_{\substack{t=0 \\ (z,x)\in\Omega_z\times\Omega_x}} = 0. \tag{58}$$

Keeping in mind the relations (53) and (58), the set of Eqs (55) and (56) can be written as follows

$$\frac{\partial P_1}{\partial z} + iD_{21}\int_0^t \frac{\partial^2 A_1}{\partial \eta^2}\exp\left(-\frac{i\eta}{2\gamma}\right)d\eta + \frac{D}{2\gamma}\int_0^t \frac{\partial^2 P_1(z,x,\eta)}{\partial x^2}d\eta - \\ -2\alpha\gamma\exp\left(-\frac{it}{2\gamma}\right)A_1^* A_2\exp(i\Delta kz) = 0, \tag{59}$$

$$\frac{\partial P_2}{\partial z} + \nu\int_0^t \frac{\partial A_2}{\partial t}\exp\left(-\frac{i\eta}{\gamma}\right)d\eta + iD_{22}\int_0^t \frac{\partial^2 A_2}{\partial \eta^2}\exp\left(-\frac{i\eta}{\gamma}\right)d\eta + \\ + \frac{D}{2\gamma}\int_0^t \frac{\partial^2 P_2(z,x,\eta)}{\partial x^2}d\eta - \alpha\gamma\exp\left(-\frac{it}{\gamma}\right)A_1^2\exp(-i\Delta kz) = 0. \tag{60}$$

Further, let us introduce the new functions $E_j(z, x, t)$, $j = 1, 2$ in accordance with a rule:

$$E_j(z,x,t) = \int_0^t A_j(z,x,\eta)\exp\left(\frac{ij(t-\eta)}{2\gamma}\right)d\eta, \tag{61}$$

and the functions $\tilde{E}_j(z, x, t)$, $j = 1, 2$ as

$$\tilde{E}_j = \exp\left(\frac{ijt}{2\gamma}\right)\int_0^t\left(\int_0^\eta A_j(z,x,\tau)\exp\left(-\frac{ij\tau}{2\gamma}\right)d\tau\right)d\eta = \\ = \exp\left(\frac{ijt}{2\gamma}\right)\int_0^t P_j(z,x,\eta)d\eta. \tag{62}$$

Then, (59) and (60) can be written in the following manner:

$$\frac{\partial E_1}{\partial z} + iD_{21}\frac{\partial^2 E_1}{\partial t^2} + \frac{D}{2\gamma}\frac{\partial^2 \tilde{E}_1}{\partial x^2} - 2\alpha\gamma A_1^* A_2\exp(i\Delta kz) = 0, \\ \frac{\partial E_2}{\partial z} + \nu\frac{\partial E_2}{\partial t} + iD_{22}\frac{\partial^2 E_2}{\partial t^2} + \frac{D}{2\gamma}\frac{\partial^2 \tilde{E}_2}{\partial x^2} - \alpha\gamma A_1^2\exp(-i\Delta kz) = 0, \tag{63}$$

$$(z,x,t)\in\Omega = \Omega_z\times\Omega_o.$$

These equations should be solved together with the relaxation equations with respect to the functions $E_j(z, x, t)$, $\tilde{E}_j(z, x, t)$, $A_j(z, x, t)$:

$$\left(\frac{\partial}{\partial t} - i\frac{j}{2\gamma}\right)E_j = A_j, \tag{64}$$

$$\left(\frac{\partial}{\partial t} - i\frac{j}{2\gamma}\right)^2\tilde{E}_j = \left(\frac{\partial}{\partial t} - i\frac{j}{2\gamma}\right)E_j = A_j,\ \ j = 1,2. \tag{65}$$

Taking into account the BCs (43) or (43′), and initial conditions (45) for the functions A_j, $j = 1, 2$, we obtain the following BCs with respect to the functions E_j, $\tilde{E}_j$ on time

coordinate:

$$E_j\Big|_{\substack{t=0\\(z,x)\in\Omega_z\times\Omega_x}} = \frac{\partial E_j}{\partial t}\Big|_{\substack{t=0\\(z,x)\in\Omega_z\times\Omega_x}} = \tilde{E}_j\Big|_{\substack{t=0\\(z,x)\in\Omega_z\times\Omega_x}} = \frac{\partial \tilde{E}_j}{\partial t}\Big|_{\substack{t=0\\(z,x)\in\Omega_z\times\Omega_x}} = 0, \tag{66}$$

$$\left(\frac{\partial E_j}{\partial t} - i\frac{j}{2\gamma}E_j\right)\Big|_{\substack{t=L_t\\(z,x)\in\Omega_z\times\Omega_x}} = 0. \tag{67}$$

As consequence of BCs (43), we obtain the following conditions:

$$\frac{\partial^2 E_j}{\partial t^2}\Big|_{\substack{t=0\\(z,x)\in\Omega_z\times\Omega_x}} = 0, \tag{68}$$

$$\left(\frac{\partial^2 E_j}{\partial t^2} - i\frac{j}{2\gamma}\frac{\partial E_j}{\partial t}\right)\Big|_{\substack{t=L_t\\(z,x)\in\Omega_z\times\Omega_x}} = 0, \tag{69}$$

if we solve the problem with BCs (43′).

Obviously, the BCs for these functions on x-coordinate are

$$E_j\Big|_{\substack{x=0,L_x\\(z,t)\in\Omega_z\times\Omega_t}} = \tilde{E}_j\Big|_{\substack{x=0,L_x\\(z,t)\in\Omega_z\times\Omega_t}} = 0. \tag{70}$$

The initial conditions for the functions $E_j, \tilde{E}_j, j = 1, 2$ are

$$E_j\Big|_{\substack{z=0\\(x,t)\in\Omega_x\times\Omega_t}} = E_{jo}(x,t) = \int_0^t A_{jo}(x,\eta)\exp\left(\frac{i(t-\eta)}{j\gamma}\right)d\eta, \tag{71}$$

$$\tilde{E}_j\Big|_{\substack{z=0\\(x,t)\in\Omega_x\times\Omega_t}} = \tilde{E}_{jo}(x,t) = \int_0^t E_{jo}(x,\eta)\exp\left(\frac{i(t-\eta)}{j\gamma}\right)d\eta. \tag{72}$$

All conditions written above will be used at the construction of the conservation laws.

Problem invariants

Using new variables, some conservation laws can be derived.

The conservation law of energy

Theorem 1. *The problem (42) with BCs (43) and initial conditions (45) possesses the conservation law of energy*

$$I_E(z) = \int_0^{L_x}\int_0^{L_t}(|A_1|^2 + |A_2|^2)dxdt = const. \tag{73}$$

Proof. Let us suppose that the functions A_j satisfy the conditions (43) and (54). Then, we multiply (55) and (56) by A_j^*, and the equations, conjugated to (55) and (56),—by A_j. Then, we sum the obtained equations and integrate the resulted equation with respect to x, t coordinates

in the domain Ω_o. Thus, we obtain the following equation in Appendix [B]:

$$\frac{d}{dz}\int_0^{L_x}\int_0^{L_t}(|A_1|^2+|A_2|^2)dxdt- \\ -\frac{D}{2\gamma}\int_0^{L_x}\left(\left|\frac{\partial P_1}{\partial x}\right|^2\Big|_{\substack{t=L_t\\(z,x)\in\Omega_z\times\Omega_x}}+\left|\frac{\partial P_2}{\partial x}\right|^2\Big|_{\substack{t=L_t\\(z,x)\in\Omega_z\times\Omega_x}}\right)dx=0. \tag{74}$$

For the validity of the invariant (73) one has to show that the last integrals in (74) are equal to zero:

$$\int_0^{L_x}\left|\frac{\partial P_j}{\partial x}\right|^2\Big|_{\substack{t=L_t\\(z,x)\in\Omega_z\times\Omega_x}}dx=0,\ \ j=1,2. \tag{75}$$

In fact, from (55) and (56) at the time moment $t = L_t$ and from the BCs (54), it follows that

$$\frac{\partial^2}{\partial x^2}\left(\int_0^{L_t}A_j(z,x,\eta)\exp\left(-\frac{ij\eta}{2\gamma}\right)d\eta\right)=\frac{\partial^2 P_j(z,x,L_t)}{\partial x^2}=0,\ \ j=1,2. \tag{76}$$

This yields to the relations

$$P_j|_{t=L_t}=C_{1j}(z)+C_{2j}(z)x,\ \ j=1,2. \tag{77}$$

It means that the Fourier harmonic amplitudes at the frequencies $\frac{1}{\gamma}$ and $\frac{1}{2\gamma}$ vary with x-coordinate growing. This property occurs even if we consider the laser pulse linear propagation (in (42) the parameter α equals zero). Obviously, the physical reason of such harmonic amplitude evolution is absent. Therefore, the functions $C_{1j}(z)$, $C_{2j}(z)$ have to be equal zero. If $C_{1j}(z)$ will be equal to nonzero constant then the energy of the laser beam will increase with a propagation distance. Obviously, a physical reason of this is absent for the case under consideration. Consequently, the amplitudes of Fourier harmonics at the frequencies $\frac{1}{\gamma}$ and $\frac{1}{2\gamma}$ must be equal zero:

$$P_j\Big|_{\substack{t=L_t\\(z,x)\in\Omega_z\times\Omega_x}}=0. \tag{78}$$

Thus, the invariant (73) is valid. In a case of the laser pulse nonlinear interaction ($\alpha \neq 0$), the same conditions take place.

Corollary 1. *The Fourier harmonics at the frequencies $\frac{1}{\gamma}$ and $\frac{1}{2\gamma}$ must be absent in the incident pulses spectra the energy conservation law preservation for the set of equations (42)*:

$$\int_0^{L_t}A_{jo}(x,\eta)\exp\left(-\frac{ij\eta}{2\gamma}\right)d\eta=0, j=1,2. \tag{79}$$

Let us note that if the parameter γ tends to zero, then the set of equations (42) reduces to the set of the NLSEs and the energy invariant (73) is valid. If the spectral harmonics $\frac{1}{\gamma}$ or $\frac{1}{2\gamma}$ are absent in the incident pulse spectrum, then these spectral harmonics will be absent in the pulse spectrum computed in any section of a medium. This statement validity is stated by the spectral invariants, which are formulated below.

Spectral invariants

Spectral invariants describe an evolution of spectral harmonic amplitudes at the frequencies $\frac{1}{\gamma}$ and $\frac{1}{2\gamma}$ along z-coordinate and show that the amplitudes should not increase during the laser

pulses interaction. For writing of the Spectral invariants, let us introduce the following additional conditions:

$$\left.\frac{\partial A_j}{\partial x}\right|_{\substack{x=0,L_x\\(z,t)\in\Omega_z\times\Omega_t}}=0,\ \ j=1,2. \tag{80}$$

Actually, in laser physics the propagation distance is boundedness and the domain under consideration can be chosen in such a way that the additional conditions are valid in this domain. We note that these conditions are important only for the invariant obtaining and they are not mandatory for the problem statement. The problem could be solved without additional conditions. Obviously, instead of BCs (80) one can use BCs (52). We stress that the laser sources generate the finite beam distribution in spatial coordinates. Therefore, the statements mentioned above are valid with respect to the domain choice.

Theorem 2. *The problem* (42) *with BCs* (43) *and additional conditions* (78) *and* (80), *and initial conditions* (45) *possesses the spectral invariants*

$$I_{SP_1}(z)=\int_0^{L_x}E_1(z,x,L_t)dx=\ \exp\left(\frac{iD_{21}z}{4\gamma^2}\right)\int_0^{L_x}E_{1o}(x,L_t)dx, \tag{81}$$

$$I_{SP_2}(z)=\int_0^{L_x}E_2(z,x,L_t)dx=\ \exp\left(\frac{iz}{\gamma}\left(\frac{D_{22}}{\gamma}-v\right)\right)\int_0^{L_x}E_{2o}(x,L_t)dx. \tag{82}$$

Proof. Let us consider (63) at a time moment $t=L_t$ and integrate them with respect to x-coordinate:

$$\int_0^{L_x}\left(\frac{\partial E_1}{\partial z}-\frac{iD_{21}}{4\gamma^2}E_1+\frac{D}{2\gamma}\frac{\partial^2\tilde{E}_1}{\partial x^2}\right)dx=0, \tag{83}$$

$$\int_0^{L_x}\left(\frac{\partial E_2}{\partial z}-\frac{iD_{22}}{\gamma^2}E_2+\frac{D}{2\gamma}\frac{\partial^2\tilde{E}_2}{\partial x^2}\right)dx=0. \tag{84}$$

Taking into account the relaxation Eqs (64) and (65) at time moment $t=L_t$

$$\left.\left(\frac{\partial E_j}{\partial t}-i\frac{j}{2\gamma}E_j\right)\right|_{\substack{t=L_t\\(z,x)\in\Omega_z\times\Omega_x}}=0,\ \ j=1,2, \tag{85}$$

$$\left.\left(\frac{\partial\tilde{E}_j}{\partial t}-i\frac{j}{2\gamma}\tilde{E}_j\right)\right|_{\substack{t=L_t\\(z,x)\in\Omega_z\times\Omega_x}}=0,\ \ j=1,2 \tag{86}$$

and additional conditions (78) we obtain:

$$\left.\left(\frac{\partial^2E_j}{\partial t^2}\right)\right|_{\substack{t=L_t\\(z,x)\in\Omega_z\times\Omega_x}}=-\frac{j^2}{4\gamma^2}E_j\Big|_{\substack{t=L_t\\(z,x)\in\Omega_z\times\Omega_x}},\ \ j=1,2. \tag{87}$$

It is easy to see that the third terms in (83) and (84) are equal to zero:

$$\int_0^{L_x} \frac{\partial^2 \tilde{E}_1}{\partial x^2} dx = \exp\left(\frac{iL_t}{2\gamma}\right) \int_0^{L_t} \int_0^{L_x} \frac{\partial^2 P_1}{\partial x^2} dx d\eta = \\ = \exp\left(\frac{iL_t}{2\gamma}\right) \int_0^{L_t} \exp\left(-\frac{i\eta}{2\gamma}\right) \frac{\partial P_1}{\partial x}\Big|_0^{L_x} d\eta. \tag{88}$$

The last integral (88) equals zero due to BCs (43). Thus, we obtain the following problems:

$$\int_0^{L_x} \left(\frac{\partial E_1}{\partial z} - \frac{iD_{21}}{4\gamma^2} E_1\right)\Big|_{\substack{t = L_t \\ (z,x) \in \Omega_z \times \Omega_x}} dx = 0, \tag{89}$$

$$\int_0^{L_x} \left(\frac{\partial E_2}{\partial z} - \left(\frac{iD_{22}}{\gamma^2} - \frac{iv}{\gamma}\right) E_2\right)\Big|_{\substack{t = L_t \\ (z,x) \in \Omega_z \times \Omega_x}} dx = 0 \tag{90}$$

with the initial conditions

$$\int_0^{L_x} E_1(0, x, L_t) dx = \int_0^{L_x} E_{1o}(x, L_t) dx, \tag{91}$$

$$\int_0^{L_x} E_2(0, x, L_t) dx = \int_0^{L_x} E_{2o}(x, L_t) dx. \tag{92}$$

Therefore, the invariants (81) and (82) can be obtained as the solutions of the problems (89)–(92).

Hamiltonian of the problem

Let us use the following substitution for the function $A_2(z, x, t)$:

$$A_2(z, x, t) \to A_2(z, x, t) \exp(-i\Delta k z). \tag{93}$$

Then, the functions $E_2(z, x, t)$, $\tilde{E}_2(z, x, t)$ can be replaced as follows:

$$E_2(z, x, t) \to E_2(z, x, t) \exp(-i\Delta k z), \tag{94}$$

$$\tilde{E}_2(z, x, t) \to \tilde{E}_2(z, x, t) \exp(-i\Delta k z). \tag{95}$$

Using the substitution (93)– (95), the set of equations (42) can be written in the following form:

$$\begin{gathered}\left(1+2i\gamma\frac{\partial}{\partial t}\right)\frac{\partial A_1}{\partial z}+iD_{21}\left(1+2i\gamma\frac{\partial}{\partial t}\right)\frac{\partial^2 A_1}{\partial t^2}+iD\frac{\partial^2 A_1}{\partial x^2}+\\+i\alpha\left(1+2i\gamma\frac{\partial}{\partial t}\right)^2 A_1^* A_2=0,\\ \left(1+i\gamma\frac{\partial}{\partial t}\right)\frac{\partial A_2}{\partial z}-i\Delta k\left(1+i\gamma\frac{\partial}{\partial t}\right)A_2+\\ +v\left(1+i\gamma\frac{\partial}{\partial t}\right)\frac{\partial A_2}{\partial t}+iD_{22}\left(1+i\gamma\frac{\partial}{\partial t}\right)\frac{\partial^2 A_2}{\partial t^2}+\\ +\frac{iD}{2}\frac{\partial^2 A_2}{\partial x^2}+i\alpha\left(1+i\gamma\frac{\partial}{\partial t}\right)^2 A_1^2=0,\end{gathered} \tag{96}$$

which does not contain the terms with $\exp(\pm i\Delta kz)$. Consequently, with respect to the substitutions (94) and (95), the set of equations (96) takes the form:

$$\begin{gathered}\frac{\partial E_1}{\partial z}+iD_{21}\frac{\partial^2 E_1}{\partial t^2}+\frac{D}{2\gamma}\frac{\partial^2 \tilde{E}_1}{\partial x^2}-2\alpha\gamma A_1^* A_2=0,\\ \frac{\partial E_2}{\partial z}+v\frac{\partial E_2}{\partial t}+iD_{22}\frac{\partial^2 E_2}{\partial t^2}+\frac{D}{2\gamma}\frac{\partial^2 \tilde{E}_2}{\partial x^2}-i\Delta kE_2-\alpha\gamma A_1^2=0.\end{gathered} \tag{97}$$

Theorem 3. *The problem* (42) *with BCs* (43) *and additional conditions* (78), *and initial conditions* (45) *possesses a Hamiltonian*

$$\begin{gathered}I_H(z)=\int_0^{L_x}\int_0^{L_t}\left[A_1^*\left(iD_{21}\frac{\partial^2 E_1}{\partial t^2}+\frac{D}{2\gamma}\frac{\partial^2 \tilde{E}_1}{\partial x^2}\right)+A_2^*\left(iD_{22}\frac{\partial^2 E_2}{\partial t^2}+\frac{D}{2\gamma}\frac{\partial^2 \tilde{E}_2}{\partial x^2}\right)+\right.\\ \left.+\frac{\partial E_2^*}{\partial t}(v\tilde{E}_2-i\Delta kE_2)-\alpha\gamma(A_2(A_1^*)^2+A_2^*A_1^2)\right]dxdt=const.\end{gathered} \tag{98}$$

Proof. Let us multiply (63) by $\frac{\partial A^*}{\partial z}$ and the equation, conjugated to it,—by $\frac{\partial A}{\partial z}$. Then, we sum the obtained equations and integrate the resulting expression with respect to x, t coordinates in the domain Ω_o under consideration. Then, we write an equation in the form of the sum of

integrals:

$$\begin{aligned}
&\int_0^{L_x}\int_0^{L_t}\left(\frac{\partial E_1}{\partial z}\frac{\partial A_1^*}{\partial z}+\frac{\partial E_1^*}{\partial z}\frac{\partial A_1}{\partial z}\right)dxdt+\\
&+iD_{21}\int_0^{L_x}\int_0^{L_t}\left(\frac{\partial^2 E_1}{\partial t^2}\frac{\partial A_1^*}{\partial z}-\frac{\partial^2 E_1^*}{\partial t^2}\frac{\partial A_1}{\partial z}\right)dxdt+\\
&+\frac{D}{2\gamma}\int_0^{L_x}\int_0^{L_t}\left(\frac{\partial^2 \tilde{E}_1}{\partial x^2}\frac{\partial A_1^*}{\partial z}+\frac{\partial^2 \tilde{E}_1^*}{\partial x^2}\frac{\partial A_1}{\partial z}\right)dxdt+\\
&+\int_0^{L_x}\int_0^{L_t}\left(\frac{\partial E_2}{\partial z}\frac{\partial A_2^*}{\partial z}+\frac{\partial E_2^*}{\partial z}\frac{\partial A_2}{\partial z}\right)dxdt+\\
&+iD_{22}\int_0^{L_x}\int_0^{L_t}\left(\frac{\partial^2 E_2}{\partial t^2}\frac{\partial A_2^*}{\partial z}-\frac{\partial^2 E_2^*}{\partial t^2}\frac{\partial A_2}{\partial z}\right)dxdt+\\
&+\frac{D}{2\gamma}\int_0^{L_x}\int_0^{L_t}\left(\frac{\partial^2 \tilde{E}_2}{\partial x^2}\frac{\partial A_2^*}{\partial z}+\frac{\partial^2 \tilde{E}_2^*}{\partial x^2}\frac{\partial A_2}{\partial z}\right)dxdt-\\
&-i\Delta k\int_0^{L_x}\int_0^{L_t}\left(E_2\frac{\partial A_2^*}{\partial z}-E_2^*\frac{\partial A_2}{\partial z}\right)dxdt+\\
&+v\int_0^{L_x}\int_0^{L_t}\left(\frac{\partial E_2}{\partial t}\frac{\partial A_2^*}{\partial z}+\frac{\partial E_2^*}{\partial t}\frac{\partial A_2}{\partial z}\right)dxdt-\\
&-\alpha\gamma\int_0^{L_x}\int_0^{L_t}\left(2A_1^*A_2\frac{\partial A_1^*}{\partial z}+A_1A_2^*\frac{\partial A_1}{\partial z}+A_1^2\frac{\partial A_2}{\partial z}+(A_1^*)^2\frac{\partial A_2^*}{\partial z}\right)dxdt=\\
&=I_{11}+iD_{21}I_{21}+\frac{D}{2\gamma}I_{31}+I_{12}+iD_{22}I_{22}+\frac{D}{2\gamma}I_{32}-\\
&-i\Delta kI_4+vI_5-\alpha\gamma I_6=0.
\end{aligned}\tag{99}$$

Analysis of the integrals is made in Appendix [C]. On its base, one can obtain the Hamiltonian of the problem under consideration.

Corollary 2. *Using the inverse transforms for (93)–(95) the Hamiltonian takes the following form*:

$$\begin{aligned}
I_H(z)=\int_0^{L_x}\int_0^{L_t}\Bigg[&A_1^*\left(iD_{21}\frac{\partial^2 E_1}{\partial t^2}+\frac{D}{2\gamma}\frac{\partial^2 \tilde{E}_1}{\partial x^2}\right)+\\
&+A_2^*\left(iD_{22}\frac{\partial^2 E_2}{\partial t^2}+\frac{D}{2\gamma}\frac{\partial^2 \tilde{E}_2}{\partial x^2}\right)+\frac{\partial E_2^*}{\partial t}(v\tilde{E}_2-i\Delta kE_2)-\\
&-\alpha\gamma(A_2\exp(-i\Delta kz)(A_1^*)^2+A_2^*\exp(i\Delta kz)A_1^2)\Bigg]dxdt=const
\end{aligned}\tag{100}$$

Discussion

A few words about applicable range of the laser pulse parameters and a medium at which it is necessary to use the GNLSEs for a description of the process under consideration. It should be noticed that the first experiment, which demonstrated self-steepening of one pulse, was made in paper [91] for the pulse with picosecond duration propagating in optical fiber about 5 km long. The pulse, propagated this distance, possessed non-symmetrical

spectrum. Therefore, a key role for application of GNLSEs plays a relation between the dispersion length of the pulse and the laser pulse propagation distance. However, the pulse self-steepening appears at short length (about a few centimeters) of a medium if the pulse duration is about 10-50 fs.

Another case of using GNLSE for the laser pulse propagation description corresponds to falling of incident laser pulse with sharp intensity distribution on a medium because an influence of a time derivative of the nonlinear response of the medium enhances many times in comparison with the Gaussian pulse propagation. However, contrast temporal nonlinear response can be induced by the laser pulse if an optical bistability occurs. As is well-known, in this case the explosive changing of the characteristics of a medium occurs. As a result, the temporal structure with strong gradient appears. Therefore, a time derivative of the nonlinear response of a medium increases its influence on the phase modulation.

A detail analysis of SHG describing in the framework of the GNLSEs has to be made using computer simulation. Nevertheless, one can suppose some characteristics features of the frequency conversion in such conditions. First, the group velocity of each of wave packets will depend on complex amplitude of another wave packet. Second, obviously that the pulse spectra will distort and become non-symmetrical. Third, the nonlinear length, that defines the distance at which the energy of base wave transfers to the energy of the wave with doubled frequency, will differ for front of the pulse and its trailing part. Fourth, because of the presence of mixed derivatives, the optical beam diffraction will depend on time moment of propagating pulse.

Conclusions

In the framework of the SEWA approximation, we derived the GNLSEs, describing the SHG in a medium with a quadratic nonlinear response for the pulses, containing a few cycles. The main feature of the equations set concludes in a presence of the second order time derivative of the nonlinear response of a medium (dispersion of a medium nonlinear response) as well as the mixed derivatives on time and spatial coordinate.

We proposed the equations transform, which reduced these equations to the other ones, containing neither mixed derivatives of complex amplitudes nor time derivatives of the nonlinear response. These equations are more convenient for the computer simulation and for theoretical analysis of the frequency doubling process of the optical pulses, containing a few cycles.

Based on this transform we derived some conservation laws (invariants) for the SHG problem. We showed an existence of two specific frequencies (singularities in the Fourier space) inherent to the problem. They may cause an appearance of non-physical absolute instability of the problem solution if the spectral invariants are not taken into account. It should be mentioned that a presence of such singularities was also discussed in [32] at analysis of the modulation instability occurring in $\chi^{(2)}$-medium. We showed that the energy conservation law is valid at certain conditions on the incident pulse spectra: the spectra must not contain non-zero spectral amplitudes at two specific frequencies. Their zero-value amplitudes at the pulse propagation are provided by the spectral invariants. We derived also the Hamiltonian (the third invariant) of the problem.

All invariants mentioned above should be taken into account at least for developing of the conservative finite-difference schemes at computer simulation of the problem under consideration.

Appendix

A Equation transform derivation

Taking into account the transforms (53), one can write

$$\begin{aligned} &2i\gamma \exp\left(\frac{it}{2\gamma}\right)\frac{\partial}{\partial z}\left(\frac{\partial}{\partial t}\left(A_1 \exp\left(\frac{-it}{2\gamma}\right)\right)\right)+ \\ &\quad +iD_{21}\cdot 2i\gamma \exp\left(\frac{it}{2\gamma}\right)\frac{\partial}{\partial t}\left(\frac{\partial^2 A_1}{\partial t^2}\exp\left(-\frac{it}{2\gamma}\right)\right)+iD\frac{\partial^2 A_1}{\partial x^2}+ \\ &+i\alpha\cdot 2i\gamma \exp\left(\frac{it}{2\gamma}\right)\frac{\partial}{\partial t}\left[\exp\left(-\frac{it}{2\gamma}\right)\left(1+2i\gamma\frac{\partial}{\partial t}\right)A_1^* A_2 \exp\left(i\Delta kz\right)\right]= \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad =0, \end{aligned} \tag{101}$$

– for the first equation of the set (42), and

$$\begin{aligned} &i\gamma \exp\left(\frac{it}{2\gamma}\right)\frac{\partial}{\partial z}\left(\frac{\partial}{\partial t}\left(A_2 \exp\left(\frac{-it}{\gamma}\right)\right)\right)+iv\gamma \exp\left(\frac{it}{\gamma}\right)\frac{\partial}{\partial t}\left(\frac{\partial A_2}{\partial t}\exp\left(-\frac{it}{\gamma}\right)\right)+ \\ &\quad +iD_{22}\cdot i\gamma \exp\left(\frac{it}{\gamma}\right)\frac{\partial}{\partial t}\left(\frac{\partial^2 A_2}{\partial t^2}\exp\left(-\frac{it}{\gamma}\right)\right)+i\frac{D}{2}\frac{\partial^2 A_2}{\partial x^2}+ \\ &+i\alpha\cdot i\gamma \exp\left(\frac{it}{\gamma}\right)\frac{\partial}{\partial t}\left[\exp\left(-\frac{it}{\gamma}\right)\left(1+i\gamma\frac{\partial}{\partial t}\right)A_1^2 \exp\left(-i\Delta kz\right)\right]=0, \end{aligned} \tag{102}$$

– for the second equation of the set (42).

Then, multiplying (101) by $\exp\left(-\frac{it}{2\gamma}\right)(2i\gamma)^{-1}$, and (102) by $\exp\left(-\frac{it}{\gamma}\right)(i\gamma)^{-1}$, and integrating them with respect to t-coordinate, we obtain:

$$\begin{aligned} &\frac{\partial}{\partial z}\left(A_1 \exp\left(-\frac{it}{2\gamma}\right)\right)+iD_{21}\frac{\partial^2 A_1}{\partial t^2}\exp\left(-\frac{it}{2\gamma}\right)-iD_{21}\frac{\partial^2 A_1}{\partial t^2}\exp\left(-\frac{it}{2\gamma}\right)\Big|_{\substack{t=0\\(z,x)\in\Omega_z\times\Omega_x}}+ \\ &\quad +\frac{D}{2\gamma}\frac{\partial^2}{\partial x^2}\left(\int_0^t A_1(z,x,\eta)\exp\left(-\frac{i\eta}{2\gamma}\right)d\eta\right)+ \\ &\quad +i\alpha\exp\left(-\frac{it}{2\gamma}\right)\left(1+2i\gamma\frac{\partial}{\partial t}\right)\left(A_1^* A_2\exp\left(i\Delta kz\right)\right)=0, \end{aligned} \tag{103}$$

$$\begin{aligned} &\frac{\partial}{\partial z}\left(A_2 \exp\left(-\frac{it}{\gamma}\right)\right)+v\frac{\partial A_2}{\partial t}\exp\left(-\frac{it}{\gamma}\right)+iD_{22}\frac{\partial^2 A_2}{\partial t^2}\exp\left(-\frac{it}{\gamma}\right)- \\ &\quad -iD_{22}\frac{\partial^2 A_2}{\partial t^2}\exp\left(-\frac{it}{\gamma}\right)\Big|_{\substack{t=0\\(z,x)\in\Omega_z\times\Omega_x}}+ \\ &\quad +\frac{D}{2\gamma}\frac{\partial^2}{\partial x^2}\left(\int_0^t A_2(z,x,\eta)\exp\left(-\frac{i\eta}{\gamma}\right)d\eta\right)+ \\ &\quad +i\alpha\exp\left(-\frac{it}{\gamma}\right)\left(1+i\gamma\frac{\partial}{\partial t}\right)\left(A_1^2\exp\left(-i\Delta kz\right)\right)=0, \end{aligned} \tag{104}$$

correspondingly.

Due to a boundedness of the initial distribution and boundedness of the laser pulse propagation distance, let us state the following additional type of the conditions:

$$\left.\frac{\partial^2 A_j}{\partial t^2}\right|_{\substack{t=0 \\ (z,x)\in \Omega_z\times\Omega_x}} = 0, j = 1, 2. \tag{105}$$

Then, (103) and (104) become the following ones:

$$\begin{aligned}\frac{\partial}{\partial z}\left(A_1 \exp\left(-\frac{it}{2\gamma}\right)\right) + iD_{21}\frac{\partial^2 A_1}{\partial t^2}\exp\left(-\frac{it}{2\gamma}\right) + \\ + \frac{D}{2\gamma}\frac{\partial^2}{\partial x^2}\left(\int_0^t A_1(z,x,\eta)\exp\left(-\frac{i\eta}{2\gamma}\right)d\eta\right) + \\ + i\alpha\exp\left(-\frac{it}{2\gamma}\right)\left(1 + 2i\gamma\frac{\partial}{\partial t}\right)(A_1^* A_2\exp(i\Delta kz)) = 0,\end{aligned} \tag{106}$$

$$\begin{aligned}\frac{\partial}{\partial z}\left(A_2 \exp\left(-\frac{it}{\gamma}\right)\right) + v\frac{\partial A_2}{\partial t}\exp\left(-\frac{it}{\gamma}\right) + iD_{22}\frac{\partial^2 A_2}{\partial t^2}\exp\left(-\frac{it}{\gamma}\right) + \\ + \frac{D}{2\gamma}\frac{\partial^2}{\partial x^2}\left(\int_0^t A_2(z,x,\eta)\exp\left(-\frac{i\eta}{\gamma}\right)d\eta\right) + \\ + i\alpha\exp\left(-\frac{it}{\gamma}\right)\left(1 + i\gamma\frac{\partial}{\partial t}\right)(A_1^2\exp(-i\Delta kz)) = 0,\end{aligned} \tag{107}$$

which contain the first derivative of the quadratic nonlinearity. Multiplying (106) by $\exp\left(\frac{it}{2\gamma}\right)$, and (107) by $\exp\left(\frac{it}{\gamma}\right)$, we obtain the following equations:

$$\begin{aligned}\frac{\partial A_1}{\partial z} + iD_{21}\frac{\partial^2 A_1}{\partial t^2} + \frac{D}{2\gamma}\exp\left(\frac{it}{2\gamma}\right)\frac{\partial^2}{\partial x^2}\left(\int_0^t A_1(z,x,\eta)\exp\left(-\frac{i\eta}{2\gamma}\right)d\eta\right) + \\ + i\alpha\left(1 + 2i\gamma\frac{\partial}{\partial t}\right)(A_1^* A_2\exp(i\Delta kz)) = 0,\end{aligned} \tag{108}$$

$$\begin{aligned}\frac{\partial A_2}{\partial z} + v\frac{\partial A_2}{\partial t} + iD_{22}\frac{\partial^2 A_2}{\partial t^2} + \\ + \frac{D}{2\gamma}\exp\left(\frac{it}{\gamma}\right)\frac{\partial^2}{\partial x^2}\left(\int_0^t A_2(z,x,\eta)\exp\left(-\frac{i\eta}{\gamma}\right)d\eta\right) + \\ + i\alpha\left(1 + i\gamma\frac{\partial}{\partial t}\right)(A_1^2\exp(-i\Delta kz)) = 0.\end{aligned} \tag{109}$$

Let us stress that these equations are used also for writing of conservation laws.

B Energy invariant derivation

Using (55) and (56) we obtain the equations for the functions A_j and A_j^*, $j = 1, 2$, which can be re-written as a sum of integrals:

$$\begin{aligned}
&\int_0^{L_x}\int_0^{L_t}\left(\frac{\partial A_1}{\partial z}A_1^* + \frac{\partial A_1^*}{\partial z}A_1 + \frac{\partial A_2}{\partial z}A_2^* + \frac{\partial A_2^*}{\partial z}A_2\right)dxdt + \\
&+v\int_0^{L_x}\int_0^{L_t}\left(\frac{\partial A_2}{\partial t}A_2^* + \frac{\partial A_2^*}{\partial t}A_2\right)dxdt + \\
&+iD_{21}\int_0^{L_x}\int_0^{L_t}\left(\frac{\partial^2 A_1}{\partial t^2}A_1^* - \frac{\partial^2 A_1^*}{\partial t^2}A_1\right)dxdt + \\
&+iD_{22}\int_0^{L_x}\int_0^{L_t}\left(\frac{\partial^2 A_2}{\partial t^2}A_2^* - \frac{\partial^2 A_2^*}{\partial t^2}A_2\right)dxdt + \\
&+\frac{D}{2\gamma}\int_0^{L_x}\int_0^{L_t}\left[\frac{\partial^2}{\partial x^2}\left(\int_0^t A_1(z,x,\eta)\exp\left(-\frac{i\eta}{2\gamma}\right)d\eta\right)\exp\left(\frac{it}{2\gamma}\right)A_1^* + \right. \\
&\left.+\frac{\partial^2}{\partial x^2}\left(\int_0^t A_1^*(z,x,\eta)\exp\left(\frac{i\eta}{2\gamma}\right)d\eta\right)\exp\left(-\frac{it}{2\gamma}\right)A_1\right]dxdt + \\
&+\frac{D}{2\gamma}\int_0^{L_x}\int_0^{L_t}\left[\frac{\partial^2}{\partial x^2}\left(\int_0^t A_2(z,x,\eta)\exp\left(-\frac{i\eta}{\gamma}\right)d\eta\right)\exp\left(\frac{it}{\gamma}\right)A_2^* + \right. \\
&\left.+\frac{\partial^2}{\partial x^2}\left(\int_0^t A_2^*(z,x,\eta)\exp\left(\frac{i\eta}{\gamma}\right)d\eta\right)\exp\left(-\frac{it}{\gamma}\right)A_2\right]dxdt + \\
&+i\alpha\int_0^{L_x}\int_0^{L_t}\left[\left(1+2i\gamma\frac{\partial}{\partial t}\right)((A_1^*)^2A_2)\exp(i\Delta kz) - \right. \\
&-\left(1-2i\gamma\frac{\partial}{\partial t}\right)(A_1^2A_2^*)\exp(-i\Delta kz) + \\
&+\left(1+i\gamma\frac{\partial}{\partial t}\right)(A_1^2A_2^*)\exp(-i\Delta kz) - \\
&\left.-\left(1-i\gamma\frac{\partial}{\partial t}\right)((A_1^*)^2A_2)\exp(i\Delta kz)\right]dxdt = \\
&= I_1 + vI_2 + iD_{21}I_{31} + iD_{22}I_{32} + \frac{D}{2\gamma}I_{41} + \frac{D}{2\gamma}I_{42} + i\alpha I_5.
\end{aligned} \tag{110}$$

Below we discuss transforms of the integrals mentioned above. The first one contains the exact differential of functions A_1, A_2 intensities. Therefore, it is easy to see that the integral I_1 transforms to the following expression:

$$I_1 = \frac{d}{dz}\int_0^{L_x}\int_0^{L_t}(|A_1|^2 + |A_2|^2)dxdt. \tag{111}$$

Also, the integral I_2 contains the time derivative of function A_2 intensity:

$$I_2 = \int_0^{L_x}\int_0^{L_t}\frac{\partial|A_2|^2}{\partial t}dxdt, \tag{112}$$

which equals zero, due to BCs (43). As a consequence of the integration by parts, the equalities

$$I_{3j} = 0, \;\; j = 1, 2,$$

take place under the conditions (43) validity.

Let us show, that the integral of the nonlinear terms I_5 equals zero. The terms in integral I_5 without the parameter γ cancel each other and integral I_5 can be written as

$$I_5 = 3i\gamma \int_0^{L_x} \int_0^{L_t} \frac{\partial}{\partial t} \left((A_1^*)^2 A_2 \exp(i\Delta kz) + A_1^2 A_2^* \exp(-i\Delta kz) \right) dxdt = 0. \tag{113}$$

For the integrals I_{41}, I_{42}, taking into account the functions P_j, $j = 1, 2$ definitions (57) and the BC (43), we obtain:

$$\begin{aligned} I_{4j} &= \int_0^{L_x} \int_0^{L_t} \left(\frac{\partial^2 P_j}{\partial x^2} \frac{\partial P_j^*}{\partial t} + \frac{\partial^2 P_j^*}{\partial x^2} \frac{\partial P_j}{\partial t} \right) dxdt = \\ &= -\int_0^{L_x} \int_0^{L_t} \left[\frac{\partial}{\partial t} \left(\frac{\partial P_j^*}{\partial x} \right) \frac{\partial P_j}{\partial x} + \frac{\partial}{\partial t} \left(\frac{\partial P_j}{\partial x} \right) \frac{\partial P_j^*}{\partial x} \right] dxdt = \\ &= -\int_0^{L_x} \int_0^{L_t} \frac{\partial}{\partial t} \left| \frac{\partial P_j}{\partial x} \right|^2 dxdt = -\int_0^{L_x} \left| \frac{\partial P_j}{\partial x} \right|^2 \Bigg|_{\substack{t = L_t \\ (z,x) \in \Omega_z \times \Omega_x}} dx. \end{aligned} \tag{114}$$

Thus, (110) transforms to the following equation:

$$\begin{aligned} &\frac{d}{dz} \int_0^{L_x} \int_0^{L_t} (|A_1|^2 + |A_2|^2) dxdt - \\ &\quad - \frac{D}{2\gamma} \int_0^{L_x} \left(\left| \frac{\partial P_1}{\partial x} \right|^2 \Bigg|_{\substack{t = L_t \\ (z,x) \in \Omega_z \times \Omega_x}} + \left| \frac{\partial P_2}{\partial x} \right|^2 \Bigg|_{\substack{t = L_t \\ (z,x) \in \Omega_z \times \Omega_x}} \right) dx = 0. \end{aligned} \tag{115}$$

C Hamiltonian derivation

As mentioned in (99), we obtain a sum of the integrals. Let us transform them. First, taking into account (64) and (78) and integrating of the I_{1j} by parts we obtain:

$$\begin{aligned} I_{1j} &= \int_0^{L_x} \int_0^{L_t} \left(\frac{\partial E_j}{\partial z} \left(\frac{\partial^2 E_j^*}{\partial z \partial t} + \frac{i}{\gamma} \frac{\partial E_j^*}{\partial z} \right) + \right. \\ &\quad \left. + \frac{\partial E_j^*}{\partial z} \left(\frac{\partial^2 E_j}{\partial z \partial t} - \frac{ij}{2\gamma} \frac{\partial E_j}{\partial z} \right) \right) dxdt = \int_0^{L_x} \left| \frac{\partial E_j}{\partial z} \right|^2 \Bigg|_{\substack{t = L_t \\ (z,x) \in \Omega_z \times \Omega_x}} dx = \\ &= \int_0^{L_x} \left| \frac{\partial P_j}{\partial z} \right|^2 \Bigg|_{\substack{t = L_t \\ (z,x) \in \Omega_z \times \Omega_x}} dx = 0, \;\; j = 1, 2. \end{aligned} \tag{116}$$

The integrals I_{2j} are transformed into the following expression:

$$\begin{aligned} I_{2j} = \int_0^{L_x} \int_0^{L_t} &\left(\frac{\partial^2 E_j}{\partial t^2} \frac{\partial^2 E_j^*}{\partial z \partial t} - \frac{\partial^2 E_j^*}{\partial t^2} \frac{\partial^2 E_j}{\partial z \partial t} + \right. \\ &\left. + \frac{ij}{2\gamma} \left(\frac{\partial^2 E_j}{\partial t^2} \frac{\partial E_j^*}{\partial z} + \frac{\partial^2 E_j^*}{\partial t^2} \frac{\partial E_j}{\partial z} \right) \right) dxdt = J_{21j} + \frac{ij}{2\gamma} J_{22j}. \end{aligned} \tag{117}$$

It is easy to see that the integrals J_{21j} and J_{22j} are transformed into the following integrals:

$$J_{21j} = \int_0^{L_x} \int_0^{L_t} \frac{\partial}{\partial z} \left(\frac{\partial^2 E_j}{\partial t^2} \frac{\partial E_j^*}{\partial t} \right) dxdt, \tag{118}$$

$$J_{22j} = \int_0^{L_x} \int_0^{L_t} \frac{\partial}{\partial z} \left(\frac{\partial^2 E_j}{\partial t^2} E_j^* \right) dxdt. \tag{119}$$

Thus, the integrals I_{2j} take the form:

$$I_{2j} = \int_0^{L_x} \int_0^{L_t} \frac{\partial}{\partial z} \left(A_j^* \frac{\partial^2 E_j}{\partial t^2} \right) dxdt. \tag{120}$$

The integrals I_{3j} can be written in the following way:

$$\begin{aligned} I_{3j} = \int_0^{L_x} \int_0^{L_t} & \left(\frac{\partial^2 \tilde{E}_j}{\partial x^2} \left(\frac{\partial^2 E_j^*}{\partial z \partial t} + \frac{ij}{2\gamma} \frac{\partial E_j^*}{\partial z} \right) + \right. \\ & \left. + \frac{\partial^2 \tilde{E}_j^*}{\partial x^2} \left(\frac{\partial^2 E_j}{\partial z \partial t} - \frac{ij}{2\gamma} \frac{\partial E_j}{\partial z} \right) \right) dxdt = \\ = \int_0^{L_x} \int_0^{L_t} & \left(\frac{\partial^2 \tilde{E}_j}{\partial x^2} \frac{\partial^3 \tilde{E}_j^*}{\partial z \partial t^2} + \frac{\partial^2 \tilde{E}_j^*}{\partial x^2} \frac{\partial^3 \tilde{E}_j}{\partial z \partial t^2} \right) dxdt + \\ + \frac{2ij}{2\gamma} \int_0^{L_x} \int_0^{L_t} & \left(\frac{\partial^2 \tilde{E}_j}{\partial x^2} \frac{\partial^2 \tilde{E}_j^*}{\partial z \partial t} - \frac{\partial^2 \tilde{E}_j^*}{\partial x^2} \frac{\partial^2 \tilde{E}_j}{\partial z \partial t} \right) dxdt - \\ - \frac{j^2}{4\gamma^2} \int_0^{L_x} \int_0^{L_t} & \left(\frac{\partial^2 \tilde{E}_j}{\partial x^2} \frac{\partial \tilde{E}^*{}_j}{\partial z} + \frac{\partial^2 \tilde{E}^*{}_j}{\partial x^2} \frac{\partial \tilde{E}_j}{\partial z} \right) dxdt = \\ & = J_{31j} + \frac{2ij}{2\gamma} J_{32j} - \frac{j^2}{4\gamma^2} J_{33j}, \end{aligned} \tag{121}$$

where

$$J_{31j} = \int_0^{L_x} \int_0^{L_t} \frac{\partial}{\partial z} \left(\frac{\partial^2 \tilde{E}_j}{\partial x^2} \frac{\partial^2 \tilde{E}_j^*}{\partial t^2} \right) dxdt, \tag{122}$$

and

$$J_{32j} = \int_0^{L_x} \int_0^{L_t} \frac{\partial}{\partial z} \left(\frac{\partial^2 \tilde{E}_j}{\partial x^2} \frac{\partial \tilde{E}_j^*}{\partial t} \right) dxdt + \int_0^{L_x} \frac{\partial^2 \tilde{E}_j^*}{\partial x^2} \frac{\partial \tilde{E}_j}{\partial z} \Big|_{\substack{t = L_t \\ (z,x) \in \Omega_z \times \Omega_x}} dx. \tag{123}$$

Let us note that multiplying equation, conjugated to (63), by $\frac{\partial \tilde{E}_j}{\partial z}$, and taking into account the BCs (43) and additional conditions (78), we obtain, that the second term in (123) equals zero. Thus,

$$J_{32j} = \int_0^{L_x} \int_0^{L_t} \frac{\partial}{\partial z} \left(\frac{\partial^2 \tilde{E}_j}{\partial x^2} \frac{\partial \tilde{E}_j^*}{\partial t} \right) dxdt. \tag{124}$$

The integrals J_{33j} can be written as follows

$$J_{33j} = \int_0^{L_x} \int_0^{L_t} \frac{\partial}{\partial z} \left(\frac{\partial^2 \tilde{E}_j}{\partial x^2} \tilde{E}_j^* \right) dxdt. \tag{125}$$

Also, the integral I_4, by using (64) and integrating by parts, reduces to the following one:

$$\begin{aligned} I_4 = \int_0^{L_x} \int_0^{L_t} \left(E_2 \frac{\partial A_2^*}{\partial z} - E_2^* \frac{\partial A_2}{\partial z} \right) dxdt = \\ = \int_0^{L_x} \int_0^{L_t} \left(E_2 \frac{\partial^2 E_2^*}{\partial z \partial t} + \frac{\partial E_2^*}{\partial t} \frac{\partial E_2}{\partial z} \right) dxdt = \\ = \int_0^{L_x} \int_0^{L_t} \frac{\partial}{\partial z} \left(E_2 \frac{\partial E_2^*}{\partial t} \right) dxdt. \end{aligned} \tag{126}$$

The integrals in I_5 can be written as

$$\begin{aligned} I_5 = \int_0^{L_x} \int_0^{L_t} \left(\frac{\partial E_2}{\partial t} \frac{\partial A_2^*}{\partial z} + \frac{\partial E_2^*}{\partial t} \frac{\partial A_2}{\partial z} \right) dxdt = \\ = \int_0^{L_x} \int_0^{L_t} \left(\frac{\partial E_2}{\partial t} \frac{\partial}{\partial z} \left(\frac{\partial E_2^*}{\partial t} + \frac{i}{\gamma} E_2^* \right) + \right. \\ \left. + \frac{\partial E_2^*}{\partial t} \frac{\partial}{\partial z} \left(\frac{\partial E_2}{\partial t} - \frac{i}{\gamma} E_2 \right) \right) dxdt = \\ = \int_0^{L_x} \int_0^{L_t} \left(\frac{\partial E_2}{\partial t} \frac{\partial^2 E_2^*}{\partial z \partial t} + \frac{\partial E_2^*}{\partial t} \frac{\partial^2 E_2}{\partial z \partial t} \right) dxdt - \\ - \frac{i}{\gamma} \int_0^{L_x} \int_0^{L_t} \left(\frac{\partial^2 E_2^*}{\partial t} \frac{\partial^2 E_2}{\partial z} - \frac{\partial^2 E_2}{\partial t} \frac{\partial^2 E_2^*}{\partial z} \right) dxdt = \\ = I_{51} - \frac{i}{\gamma} I_{52}. \end{aligned} \tag{127}$$

Using the condition (78) and integrating by parts, one can transform the integrals I_{51}, I_{52} to the form

$$I_{51} = \int_0^{L_x} \int_0^{L_t} \frac{\partial}{\partial z} \left| \frac{\partial E_2}{\partial t} \right|^2 dxdt, \tag{128}$$

and

$$I_{52} = \int_0^{L_x} \int_0^{L_t} \frac{\partial}{\partial z} \left(\frac{\partial E_2^*}{\partial t} E_2 \right) dxdt. \tag{129}$$

Obviously, the integral I_6 can be written as follows:

$$\begin{aligned} I_6 = \int_0^{L_x} \int_0^{L_t} \left(2A_1^* A_2 \frac{\partial A_1^*}{\partial z} + A_1 A_2^* \frac{\partial A_1}{\partial z} + \right. \\ \left. + A_1^2 \frac{\partial A_2}{\partial z} + (A_1^*)^2 \frac{\partial A_2^*}{\partial z} \right) dxdt = \\ \int_0^{L_x} \int_0^{L_t} \left(A_2 \frac{\partial (A_1^*)^2}{\partial z} + (A_1^*)^2 \frac{\partial A_2}{\partial z} + A_2^* \frac{\partial A_1^2}{\partial z} + A_1^2 \frac{\partial A_2^*}{\partial z} \right) dxdt = \\ = \int_0^{L_x} \int_0^{L_t} \frac{\partial}{\partial z} \left(A_2 (A_1^*)^2 + A_2^* A_1^2 \right) dxdt. \end{aligned} \tag{130}$$

Substituting the expressions (116), (120), (121), (126), (127) and (130) into (99), we obtain the invariant (98).

Author Contributions

Data curation: Vyacheslav A. Trofimov.

Formal analysis: Vyacheslav A. Trofimov.

Investigation: Vyacheslav A. Trofimov, Svetlana Stepanenko, Alexander Razgulin.

Methodology: Vyacheslav A. Trofimov.

Validation: Vyacheslav A. Trofimov, Svetlana Stepanenko.

Writing – original draft: Vyacheslav A. Trofimov, Svetlana Stepanenko.

Writing – review & editing: Vyacheslav A. Trofimov, Svetlana Stepanenko.

References

1. Franken PA, Hill AE, Peters CW, Weinreich G. Generation of optical harmonics. Phys. Rev. Lett. 1961; 7(4): 118–120. https://doi.org/10.1103/PhysRevLett.7.118
2. Armstrong J, Bloembergen N, Ducuing J, Pershan P. Interactions between Light Waves in a Nonlinear Dielectric. Phys. Rev. 1962; 127(6): 1918–39. https://doi.org/10.1103/PhysRev.127.1918
3. Linde D, Schulz H, Engers T, Schüler H. Second Harmonic Generation in Plasmas Produced by Intense Femtosecond Laser Pulses. IEEE Jour. of Quant. Electr. 1992; 28(10): 2388–97. https://doi.org/10.1109/3.159545
4. Streeter M, Foster PS, Cameron FH, Borghesi M, Brenner C, Carroll DC, et. al. Relativistic plasma surfaces as an efficient second harmonic generator. New J. Phys. 2011; 13: 023041. https://doi.org/10.1088/1367-2630/13/2/023041
5. Singh M, Gupta DN, Suk H. Efficient second- and third-harmonic radiation generation from relativistic laser-plasma interactions. Phys. of Plasm. 2015; 22: 063303. https://doi.org/10.1063/1.4922435
6. Kim S, Jin J, Kim YJ, Park IY, Kim Y, Kim SW. High-harmonic generation by resonant plasmon field enhancement. Nature. 2008; 453: 757–760. https://doi.org/10.1038/nature07012 PMID: 18528390
7. Steingrube D, Schulz E, Binhammer T, Gaarde MB, Couairon A, Morgner U, Kovacev M. High-order harmonic generation directly from a filament. New J. Phys. 2011; 13: 043022. https://doi.org/10.1088/1367-2630/13/4/043022
8. Zheng J, Qiu E, Lin Q. High harmonic generation with sub-cycle pulses. J. Opt. 2011; 13: 075206. https://doi.org/10.1088/2040-8978/13/7/075206
9. Lucchini M, Calegari F, Kim K, Sansone G, Nisoli M. Nonadiabatic quantum path analysis of the high-order harmonic generation in a highly ionized medium. New J. Phys. 2012; 14(2): 023025.
10. Kemper A, Moritz B, Freericks JK, Devereaux TP. Theoretical description of high-order harmonic generation in solids. New J. Phys. 2013; 15: 023003. https://doi.org/10.1088/1367-2630/15/2/023003
11. Vampa G, Hammond TJ, Thiré N, Schmidt BE, Légaé F, McDonald CR, Brabec T, Corkum PB. Linking high harmonics from gases and solids. Nature. 2015; 522: 462–464. https://doi.org/10.1038/nature14517 PMID: 26108855
12. Neyra E, Videla F, Ciappina MF, Pérez-Hernández JA, Roso L, Lewenstein M, Torchia GA. High-order harmonic generation driven by inhomogeneous plasmonics fields spatially bounded: influence on the cut-off law. J. Opt. 2018; 20(3): 034002. https://doi.org/10.1088/2040-8986/aaa6f7
13. Zdanowicz M, Kujala S, Husu H, Kauranen M. Effective medium multipolar tensor analysis of second-harmonic generation from metal nanoparticles. New J. Phys. 2011; 13: 023025. https://doi.org/10.1088/1367-2630/13/2/023025
14. Pavlyukh Y, Berakdar J, Hübner W. Semi-classical approximation for second-harmonic generation in nanoparticles. New J. Phys. 2012; 14: 093044. https://doi.org/10.1088/1367-2630/14/9/093044
15. Kautek W, Sorg N, Krüger J. Femtosecond pulse laser second harmonic generation on semiconductor electrodes. Electrochimica Acta. 1994; 39(8/9): 1245–49. https://doi.org/10.1016/0013-4686(94)E0043-Y

16. Mlejnek M, Wright E, Moloney J, Bloembergen N. Second Harmonic Generation of Femtosecond Pulses at the Boundary of a Nonlinear Dielectric. Phys. Rev. Lett. 1999; 83(15): 2934–37. https://doi.org/10.1103/PhysRevLett.83.2934
17. Sidick E, Knoesen A, Dienes A. Ultrashort-pulse second-harmonic generation. I. Transform-limited fundamental pulses. J. Opt. Soc. Am. B. 1995; 12(9): 1704–12. https://doi.org/10.1364/JOSAB.12.001704
18. Kim D-W, Xiao G-Y, Ma G-B. Temporal properties of the second-harmonic generation of a short pulse. Appl. Opt. 1997; 36(27): 6788–93. https://doi.org/10.1364/ao.36.006788 PMID: 18259545
19. Telegin LS, Chirkin AS. Interaction in frequency doubling of ultrashort laser pulses. Sov. J. Quantum Electron. 1982; 12: 1358. https://doi.org/10.1070/QE1982v012n10ABEH006061
20. Razumikhina TB, Telegin LS, Kholodnykh AI, Chirkin AS. Three-frequency interactions of high-intensity light waves in media with quadratic and cubic nonlinearities. Sov. J. Quantum Electron. 1984; 14(10): 1358–63. https://doi.org/10.1070/QE1984v014n10ABEH006408
21. Ditmire T, Rubenchik AM, Eimerl D, Perry MD. Effects of cubic nonlinearity on frequency doubling of high-power laser pulses. J. Opt. Soc. Am. B. 1996; 13(4): 649–652. https://doi.org/10.1364/JOSAB.13.000649
22. Choe W, Banerjee PP, Caimi FC. Second-harmonic generation in an optical medium with second- and third-order nonlinear susceptibilities. J. Opt. Soc. Am. B. 1991; 8(5): 1013–22. https://doi.org/10.1364/JOSAB.8.001013
23. Komissarova MV, Sukhorukov AP, Trofimov VA. Self-compression of the fundamental and second harmonic pulses in the media with quadratic and cubic nonlinearities. Bulletin of the Russian Academy of Sciences. Physics supplement. Physics of Vibration. 1993; 57: 189–192.
24. Lysak TM, Trofimov VA. The bistable mode of second harmonic generation by femtosecond pulses. Technical Physics. 2001; 46: 1401–06. https://doi.org/10.1134/1.1418503
25. Lysak TM, Trofimov VA. Bistability and uniqueness of solutions in the problem of second harmonic generation of femtosecond pulses. Computational Mathematics and Mathematical Physics. 2001; 41: 1214–26.
26. Lin R, Gao Y. Observation of the modulation instability and frequency-doubling in self-defocusing crystal. Phys. Lett. A. 2011; 375: 3228–31. https://doi.org/10.1016/j.physleta.2011.07.007
27. Kasumova RJ, Safarova GA, Ahmadova AR, Kerimova NV. Influence of self- and cross-phase modulations on an optical frequency doubling process for metamaterials. Appl. Opt. 2018; 57(25): 7385–90. https://doi.org/10.1364/AO.57.007385 PMID: 30182960
28. Trofimov VA, Kharitonov DM, Fedotov MV. Theory of SHG in a medium with combined nonlinear response. J. Opt. Soc. Am. B. 2018; 35(12): 3069–87. https://doi.org/10.1364/JOSAB.35.003069
29. Trofimov VA, Trofimov VV. High effective SHG of femtosecond pulse with ring profile of beam in bulk medium with cubic nonlinear response. Proceedings of SPIE. 2007; 66100: 66100R. https://doi.org/10.1117/12.740023
30. Ashihara S, Nishina J, Shimura T, Kuroda K. Soliton compression of femtosecond pulses in quadratic media. J. Opt. Soc. Am. B. 2002; 19(10) 2505–10. https://doi.org/10.1364/JOSAB.19.002505
31. Liu X, Qian L, Wise F. High-energy pulse compression by use of negative phase shifts produced by the cascade $\chi^{(2)}$: $\chi^{(2)}$ nonlinearity. Opt. Lett. 1999; 24(23): 1777–79. https://doi.org/10.1364/ol.24.001777 PMID: 18079931
32. Wyller J, Królikowski WZ, Bang O, Petersen DE, Rasmussen JJ. Modulational instability in the nonlocal $\chi^{(2)}$-model. Physica D. 2007; 227: 8–25. https://doi.org/10.1016/j.physd.2007.01.002
33. Wang J, Ma Z, Li Y, Lu D, Guo Q, Hu W. Stable quadratic solitons consisting of fundamental waves and oscillatory second harmonics subject to boundary confinement. Phys. Rev. A. 2015; 91: 033801. https://doi.org/10.1103/PhysRevA.91.033801
34. Karamzin YN, Sukhorukov AP. Nonlinear interaction of diffracted light beams in a medium with quadratic nonlinearity: mutual focusing of beams and limitation on the efficiency of optical frequency converters. JETP Lett. 1974; 20: 339.
35. Torruellas WE, Wang Z, Hagan DJ, VanStryland EW, Stegeman GI, Torner L, et. al. Observation of two-dimensional spatial solitary waves in a quadratic medium. Phys. Rev. Lett. 1995; 74: 5036. https://doi.org/10.1103/PhysRevLett.74.5036 PMID: 10058667
36. Schiek R, Baek Y, Stegeman GI. One-dimensional spatial solitary waves due to cascaded second-order nonlinearities in planar waveguides. Phys. Rev. E. 1996; 53: 1138. https://doi.org/10.1103/PhysRevE.53.1138
37. Fuerst RA, Baboiu DM, Lawrence B, Torruellas WE, Stegeman GI, Trillo S, et.al. Spatial modulational instability and multisolitonlike generation in a quadratically nonlinear optical medium. Phys. Rev. Lett. 1997; 78: 2756. https://doi.org/10.1103/PhysRevLett.78.2756

38. Costantini B, De Angelis C, Barthelemy A, Bourliaguet B, Kermene V. Collisions between type II two-dimensional quadratic solitons. Opt. Lett. 1998; 23: 424–426. https://doi.org/10.1364/ol.23.000424 PMID: 18084532

39. Di Trapani P, Caironi D, Valiulis G, Dubietis A, Danielius R, Piskarskas A. Observation of temporal solitons in second-harmonic generation with tilted pulses. Phys. Rev. Lett. 1998; 81: 570. https://doi.org/10.1103/PhysRevLett.81.570

40. Liu X, Qian LJ, Wise FW. Generation of optical spatiotemporal solitons. Phys. Rev. Lett. 1999; 82: 23. https://doi.org/10.1103/PhysRevLett.82.4631

41. Kim DH, Kang JU, Khurgin JB. Cascaded Raman self-frequency shifted soliton generation in an Er/Yb-doped fiber amplifier. Appl. Phys. Lett. 2002; 81: 2695. https://doi.org/10.1063/1.1512823

42. Kharenko DS, Bednyakova AE, Podivilov EV, Fedoruk MP, Apolonski A, Babin SA. Cascaded generation of coherent Raman dissipative solitons. Opt. Lett. 2016; 41: 175. https://doi.org/10.1364/OL.41.000175 PMID: 26696187

43. Buryak AV, Di Trapani P, Skryabin DV, Trillo S. Optical solitons due to quadratic nonlinearities: from basic physics to futuristic applications. Phys. Rep. 2002; 370: 63. https://doi.org/10.1016/S0370-1573(02)00196-5

44. Cheng Z, Fu HY, Li Q. Cascaded photonic crystal fibers for three stage non-integer order soliton compression. Opt. Comm. and Net. 2017; 1–2.

45. Bache M, Wise FW. Type-I cascaded quadratic soliton compression in lithium niobate: Compressing femtosecond pulses from high-power fiber lasers. Phys. Rev. A. 2010; 81: 053815. https://doi.org/10.1103/PhysRevA.81.053815

46. Zeng X, Ashihara S, Shimura T, Kuroda K. Adiabatic femtosecond pulse compression and control by using quadratic cascading nonlinearity. Nonlinear Optics: Technologies and Applications.—International Society for Optics and Photonics. 2008; 6839: 68390B.

47. Li Q, Kutz J, Wai P. Cascaded higher-order soliton for non-adiabatic pulse compression. J. Opt. Soc. B. 2010; 27: 2180. https://doi.org/10.1364/JOSAB.27.002180

48. Bache M, Królikowski WZ, Moses J, Wise FW. Limits to compression with cascaded quadratic soliton compressors. Opt. Expr. 2008; 16: 3273. https://doi.org/10.1364/OE.16.003273

49. Moses J, Wise FW. Soliton compression in quadratic media: high-energy few-cycle pulses with a frequency-doubling crystal. Opt. Lett. 2006; 31: 1881. https://doi.org/10.1364/ol.31.001881 PMID: 16729102

50. Šuminas R, Tamošauskas G, Valiulis G, Dubietis A. Spatiotemporal light bullets and supercontinuum generation in β-BBO crystal with competing quadratic and cubic nonlinearities. Opt. Lett. 2016; 41: 2097. https://doi.org/10.1364/OL.41.002097 PMID: 27128083

51. Conti C, Trillo S, Di Trapani P, Kilius J, Bramati A, Minardi S, et. al. Effective lensing effects in parametric frequency conversion. J. Opt. Soc. Am. B. 2002; 19(4): 852–859. https://doi.org/10.1364/JOSAB.19.000852

52. Di Trapani P, Bramati A, Minardi S, Chinaglia W, Trillo S, Conti C, et. al. Focusing versus defocusing nonlinearities in self-trapping due to parametric frequency conversion. Phys. Rev. Lett. 2001; 87: 183902. https://doi.org/10.1103/PhysRevLett.87.183902

53. Trofimov V, Lysak T. Catastrophic self-focusing of axially symmetric laser beams due to cascading SHG. Proceedings of SPIE. 2011; 7822: 78220E–78220E-11. https://doi.org/10.1117/12.891265

54. Lysak TM, Trofimov VA. Achieving high-efficiency second harmonic generation in a sequence of laser pulses with random peak intensity. Part I. Efficient generation in optical fibers. Comp. Math. and Modeling. 2008; 19: 333–342. https://doi.org/10.1007/s10598-008-9012-z

55. Lysak TM, Trofimov VA. Achieving high-efficiency second harmonic generation in a sequence of laser pulses with random peak intensity Part II. Suppression of intensity fluctuations in a quadratic-nonlinearity medium. Comp. Math. and Modeling. 2009; 20: 1–25. https://doi.org/10.1007/s10598-009-9015-4

56. Lysak TM, Trofimov VA. Achieving high-efficiency second harmonic generation in a sequence of laser pulses with random peak intensity Part III. Propagation of pulses in a bulk medium. Comp. Math. and Modeling. 2009; 20: 101–112. https://doi.org/10.1007/s10598-009-9030-5

57. Trofimov VA, Lysak TM. Highly efficient SHG of a sequence of laser pulses with a random peak intensity and duration. Opt. Spectrosc. 2009; 107: 399–406. https://doi.org/10.1134/S0030400X0909015X

58. Huttunen M, Mäkitalo J, Bautista G, Kauranen M. Multipolar second-harmonic emission with focused Gaussian beams. New J. Phys. 2012; 14: 113005.

59. O'Donnell K, Torre R. Characterization of the second-harmonic response of a silver-air interface. New J. Phys. 2005; 7: 154.

60. Rodrigo S, Laliena V, Martin-Moreno L. Second-harmonic generation from metallic arrays of rectangular holes. J. Opt. Soc. Am. B. 2015; 32: 15–25. https://doi.org/10.1364/JOSAB.32.000015

61. Luo M, Liu Q. Extraordinary enhancement of second harmonic generation in a periodically patterned distributed Bragg reflector. J. Opt. Soc. Am. B. 2015; 32: 1193–1201. https://doi.org/10.1364/JOSAB.32.001193

62. Yudovich S, Shwartz S. Second-harmonic generation of focused ultrashort x-ray pulses. J. Opt. Soc. Am. B. 2015; 32: 1894–1900. https://doi.org/10.1364/JOSAB.32.001894

63. Butet J, Gallinet B, Thyagarajan K, Martin OJF. Second-harmonic generation from periodic arrays of arbitrary shape plasmonic nanostructures: a surface integral approach. J. Opt. Soc. Am. B. 2013; 30: 2970–79. https://doi.org/10.1364/JOSAB.30.002970

64. Kolmychek IA, Krutyanskiy VL, Murzina TV, Sapozhnikov MV, Karashtin EA, Rogov VV, et. al. First and second order in magnetization effects in optical second-harmonic generation from a trilayer magnetic structure. J. Opt. Soc. Am. B. 2015; 32: 331–338. https://doi.org/10.1364/JOSAB.32.000331

65. Samim M, Krouglov S, Barzda V. Double Stokes Mueller polarimetry of second-harmonic generation in ordered molecular structures. J. Opt. Soc. Am. B. 2015; 32: 451–461. https://doi.org/10.1364/JOSAB.32.000451

66. Arjmand A, Abolghasem P, Han J, Helmy AS. Interface modes for monolithic nonlinear photonics. J. Opt. Soc. Am. B. 2015; 32: 577–587. https://doi.org/10.1364/JOSAB.32.000577

67. Hardhienata H, Alejo-Molina A, Reitböck C, Prylepa A, Stifter D, Hingerl K. Bulk dipolar contribution to second-harmonic generation in zincblende. J. Opt. Soc. Am. B. 2016; 33: 195–201. https://doi.org/10.1364/JOSAB.33.000195

68. Zhang S, Zhang X. Strong second-harmonic generation from bilayer-graphene embedded in one-dimensional photonic crystals. J. Opt. Soc. Am. B. 2016; 33: 452–460. https://doi.org/10.1364/JOSAB.33.000452

69. Sabouri SG, Khorsandi A. Thermal dephasing compensation in high-power and high-repetition-rate second-harmonic generation using spillover loss. J. Opt. Soc. Am. B. 2016; 33: 1640–48. https://doi.org/10.1364/JOSAB.33.001640

70. Tang D, Wang J, Zhou B, Xie G, Ma J, Yuan P, et. al. Temperature-insensitive second-harmonic generation based on noncollinear phase matching in a lithium triborate crystal. J. Opt. Soc. Am. B. 2017; 34: 1659–68. https://doi.org/10.1364/JOSAB.34.001659

71. Guo S, Ge Y, Han Y, He J, Wang J. Investigation of optical inhomogeneity of MgO:PPLN crystals for frequency doubling of 1560 nm laser. Optics Communications. 2014; 326: 114–120. https://doi.org/10.1016/j.optcom.2014.04.008

72. Yuan J.-H, Zhang Y, Mo H, Chen N, Zhang Z. The second-harmonic generation susceptibility in semi-parabolic quantum wells with applied electric field. Optics Communications. 2015; 356: 405–410. https://doi.org/10.1016/j.optcom.2015.08.030

73. Kanseri B, Bouillard M, Tualle-Brouri R. Efficient frequency doubling of femtosecond pulses with BIBO in an external synchronized cavity. Optics Communications. 2016; 380: 148–153. https://doi.org/10.1016/j.optcom.2016.05.067

74. Kato K, Umemura N, Petrov V. Sellmeier and thermo-optic dispersion formulas for $CdGa_2S_4$ and their application to the nonlinear optics of $Hg_{1-x}Cd_xGa_2S_4$. Optics Communications. 2017; 386: 49–52. https://doi.org/10.1016/j.optcom.2016.10.054

75. Zhang Y, Hyodo M, Okada-Shudo Y, Zhu Y, Wang X. Characteristics of pulse width for an enhanced second harmonic generation. Optics Communications. 2017; 387: 241–244. https://doi.org/10.1016/j.optcom.2016.11.058

76. Cai L, Wang Y, Hu H. Efficient second harmonic generation in $\chi^{(2)}$ profile reconfigured lithium niobate thin film. Optics Communications. 2017; 387: 405–408. https://doi.org/10.1016/j.optcom.2016.10.064

77. Leo N, Meier D, Becker P, Bohatý L, Fiebig M. Magnetically driven second-harmonic generation with phase matching in $MnWO_4$. Optics Express. 2015; 23: 27700. https://doi.org/10.1364/OE.23.027700 PMID: 26480432

78. Zeng J, Li J, Li H, Dai Q, Tie S, Lan S. Effects of substrates on the nonlinear optical responses of two-dimensional materials. Optics Express. 2015; 23: 31817. https://doi.org/10.1364/OE.23.031817 PMID: 26698973

79. Kim S, Qi M. Broadband second-harmonic phase-matching in dispersion engineered slot waveguides. Optics Express. 2016; 24(2): 773. https://doi.org/10.1364/OE.24.000773 PMID: 26832462

80. Stegeman GI, Schiek R, Fang H, Malendevich R, Jankovic L, Torner L, et. al. Beam evolution in quadratically nonlinear one-dimensional media: LiNbO3 slab waveguides. Laser Phys. 2003; 13: 137–147.

81. Su W, Qian L, Luo H, Fu X, Zhu H, Wang T, et. al. Induced group-velocity dispersion in phase-mismatched second-harmonic generation. J. Opt. Soc. Am. B. 2006; 23(1): 51–55. https://doi.org/10.1364/JOSAB.23.000051

82. Bache M, Bang O, Zhou BB, Moses J, Wise FW. Optical Cherenkov radiation by cascaded nonlinear interaction: an efficient source of few-cycle energetic near- to mid-IR pulses. Opt. Expr. 2011; 19: 22557–62. https://doi.org/10.1364/OE.19.022557

83. Cai Y, Xu S, Zeng X, Zou D, Li J. High-efficiency intracavity second-harmonic enhancement for a few-cycle laser pulse train. J. Opt. 2012; 14: 105202.

84. Ota Y, Watanabe K, Iwamoto S, Arakawa Y. Measuring the second-order coherence of a nanolaser by intracavity frequency doubling. Phys. Rev. A. 2014; 89: 023824. https://doi.org/10.1103/PhysRevA.89.023824

85. Iaconis C, Walmsley IA. Self-Referencing Spectral Interferometry for Measuring Ultrashort Optical Pulses. IEEE J. Quantum Electron. 1999; 35: 501–509. https://doi.org/10.1109/3.753654

86. Kane DJ, Trebino R. Characterization of arbitrary femtosecond pulses using frequency-resolved optical gating. IEEE J. Quantum Electron. 1993; 29: 571–579. https://doi.org/10.1109/3.199311

87. Baltuska A, Pshenichnikov M, Wiersma D. Second-harmonic generation frequency-resolved optical gating in the single-cycle regime. IEEE J. Quant. El. 1999; 35(4): 459–478. https://doi.org/10.1109/3.753651

88. Bragheri F, Faccio D, Bonaretti F, Lotti A, Clerici M, Jedrkiewicz O, et. al. Complete retrieval of the field of ultrashort optical pulses using the angle-frequency spectrum. Opt. Lett. 2008; 33(24): 2952–54. https://doi.org/10.1364/ol.33.002952 PMID: 19079503

89. Hause A, Kraft S, Rohrmann P, Mitschke F. Reliable multiple-pulse reconstruction from second-harmonic-generation frequency-resolved optical gating spectrograms. J. Opt. Soc. Am. B. 2015; 32: 868–877. https://doi.org/10.1364/JOSAB.32.000868

90. Pirogova I, Sukhorukov A. Effect of nonlinear-wave coupling dispersion on the frequency doubling of subpicosecond light-pulses. Opt. and Spect. 1985; 59(3): 694–696.

91. Tzoar N, Jain M. Self-phase modulation in long-geometry optical waveguides. Phys. Rev. A. 1981; 23(3): 1266–70. https://doi.org/10.1103/PhysRevA.23.1266

92. Anderson D, Lisak M. Nonlinear asymmetric self-phase modulation and self-steepening of pulses in long optical waveguides. Phys. Rev. A. 1983; 27(3): 1393–98. https://doi.org/10.1103/PhysRevA.27.1393

93. Agrawal G. Nonlinear Fiber Optics. (4th ed. Academic Press). 2007.

94. Brabec T, Krausz F. Nonlinear optical pulse propagation in the single-cycle regime. Phys. Rev. Lett. 1997; 78(17): 3282–85. https://doi.org/10.1103/PhysRevLett.78.3282

95. Dong M-J, Tian S-F, Yan X-W, Zhang T-T. Nonlocal symmetries, conservation laws and interaction solutions for the classical Boussinesq–Burgers equation. Nonlinear Dynamics. 2019; 95(1): 273–291. https://doi.org/10.1007/s11071-018-4563-9

96. Yan X-W, Tian S-F, Dong M-J, Wang X-B, Zhang T-T. Nonlocal symmetries, conservation laws and interaction solutions of the generalised dispersive modified Benjamin–Bona–Mahony equation. Zeitschrift für Naturforschung A. 2018; 73(5): 399–405. https://doi.org/10.1515/zna-2017-0436

97. Wang X-B, Tian S-F, Qin C-Y, Zhang T-T. Lie symmetry analysis, conservation laws and analytical solutions of a time-fractional generalized KdV-type equation. J. Nonlinear Math. Phys. 2017; 24(4): 516–530. https://doi.org/10.1080/14029251.2017.1375688

98. Peng W-Q, Tian S-F, Zhang T-T. Dynamics of breather waves and higher-order rogue waves in a coupled nonlinear Schrödinger equation. Europhysics Letters. 2018; 123(5): 50005.

99. Wang X-B, Tian S-F, Qin C-Y, Zhang T-T. Lie symmetry analysis, analytical solutions, and conservation laws of the generalised Whitham–Broer–Kaup–Like equations. Zeitschrift für Naturforschung A. 2017; 72(3): 269–279. https://doi.org/10.1515/zna-2016-0389

100. Varentsova SA, Trofimov VA. Lagrangian of the process of generation of the second harmonic of femtosecond light pulses. Diff. Eq. 1998; 34(7): 997–999.

101. Varentsova SA, Trofimov VA. Invariants in the process of second harmonic generation by femtosecond light pulses. Moscow University Comput. Math. and Cybern. 1998; 4: 45–47.

102. Hovhannisyan D, Stepanyan K, Avagyan R. Computational modelling of second-harmonic generation by a femtosecond laser pulse of a few optical cycles. J. Mod. Opt. 2005; 52(1): 97–107. https://doi.org/10.1080/09500340410001703832

103. Szarvas T, Kis Z. Numerical simulation of nonlinear second harmonic wave generation by the finite difference frequency domain method. J. Opt. Soc. Am. B. 2018; 35(4): 731–740. https://doi.org/10.1364/JOSAB.35.000731

104. Xiao Y, Maywar DN, Agrawal GP. Propagation of few-cycle pulses in nonlinear Kerr media: harmonic generation. Opt. Lett. 2013; 38(5): 724–726. https://doi.org/10.1364/OL.38.000724 PMID: 23455278

105. Shcherbakov AA, Tishchenko AV. New fast and memory-sparing method for rigorous electromagnetic analysis of 2D periodic dielectric structures. J. Quant. Spectrosc. Radiat. Transfer. 2012; 113: 158–171. https://doi.org/10.1016/j.jqsrt.2011.09.019

106. Weismann M, Gallagher D, Panoiu N. Nonlinear generalized source method for modeling second-harmonic generation in diffraction gratings. J. Opt. Soc. Am. B. 2015; 32: 523–533. https://doi.org/10.1364/JOSAB.32.000523

107. Karamzin YN, Sukhorukov AP, Trofimov VA. Mathematical modeling in nonlinear optics. Moscow: Publishing house of Moscow University [in Russian]; 1989.

Fractional-order quantum particle swarm optimization

Lai Xu, Aamir Muhammad, Yifei Pu, Jiliu Zhou, Yi Zhang*

School of Computer Science, Sichuan University, Chengdu, Sichuan Province, China

* yzhang@scu.edu.cn

Abstract

Motivated by the concepts of quantum mechanics and particle swarm optimization (PSO), quantum-behaved particle swarm optimization (QPSO) was developed to achieve better global search ability. This paper proposes a new method to improve the global search ability of QPSO with fractional calculus (FC). Based on one of the most frequently used fractional differential definitions, the Grünwald-Letnikov definition, we introduce its discrete expression into the position updating of QPSO. Extensive experiments on well-known benchmark functions were performed to evaluate the performance of the proposed fractional-order quantum particle swarm optimization (FQPSO). The experimental results demonstrate its superior ability in achieving optimal solutions for several different optimizations.

Editor: Nicholas Chancellor, Durham University, UNITED KINGDOM

Introduction

Particle swarm optimization (PSO) [1], which is inspired by animal social behaviors, such as birds, was first proposed by Kennedy and Eberhart as a population-based optimization technique. In PSO, the potential solutions, which are called particles, go through the solution space by relying on their own experiences and current best particle. PSO has a competitive performance with the classical Genetic Algorithm (GA) [2], evolutionary programming (EP) [3], evolution strategies (ES) [4], genetic programming (GP)[5] and other classic algorithms. It has attracted increasing attention during recent years thanks to its effectiveness in different optimization problems [6][7][8].

Quantum computer [9] was proposed 30 years ago and the formal definition of the quantum computer was given in the late 1980s. Since the quantum computer has shown its potential in several special problems [10], many efforts were dedicated to this field. Several well-known algorithms were proposed, and Shor's quantum factoring algorithm was the most famous one in these methods [11].

Inspired by a similar idea, the quantum-behaved particle swarm optimization (QPSO) [12] was introduced in 2004 by Sun et al. to improve the convergence of classical PSO. In quantum space, particles search in the complete solution space and the global optimum is guaranteed. In recent decades, fractional calculus has drawn increasing interests and been a strong branch of mathematical analyses. Furthermore, the random variables in the physical process can be regarded as the substitution of real stochastic motion. As a result, the fractional calculus can be

Funding: This work was supported by National Key R&D Program of China, http://www.most.gov.cn/, 2017YFB0802300 to JZ; National Natural Science Foundation of China, http://www.nsfc.gov.cn/, 61671312 to YZ; Science and Technology Support Project of Sichuan Province of China, http://kjt.sc.gov.cn/, 2018HH0070 to YZ; and Science and Technology Support Project of Sichuan Province of China, 2013SZ0071, http://kjt.sc.gov.cn/, to YP. The funders had no role in study design, data

collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

introduced to analyze the physical statuses and procedures of objects in Euclidean space. The fractional differential functions have two features. A fractional differential function is a power function for primary functions, and it is an iterative addition or product of specific functions for the other functions. Meanwhile, it has been proved that many fractional-order models are more suitable to describe the natural phenomena. Based on these observations, fractional calculus has been introduced into many fields such as viscoelastic theory [13], diffusion processing [14] and stochastic fractal dynamics [15]. Most of the researches on fractional-order applications focus on the transient state of physical changes. However, the evolutive procedures of systems are rarely included.

In recent years, QPSO has attracted great attention from many researchers. To balance the global and local searching abilities, Xi et al. proposed a novel QPSO called weighted QPSO (WQPSO)[16]. Jiao et al. proposed a dynamic-context cooperative quantum-behaved particle swarm optimization (CQPSO)[17] for medical image segmentation. Although QPSO and its variants have better performance in some aspects, they do not make full use of the state information during the ergodic process and it is inefficient in hunting global optimum. In this paper, a novel quantum particle swarm optimization with the fractional-order position is proposed. Due to the nonlinear, non-causal and non-stationary characteristics of fractional calculus, searching global optimum can be significantly accelerated [18][19].

The rest of this paper is organized as follows: In section 2, some mathematical background about fractional calculus is introduced. Section 3 presents the basic ideas of PSO and QPSO and the proposed method is also given there. Section 4 demonstrated the experimental results of the proposed method. Finally, Section 5 outlines the conclusion.

Background theory for fractional calculus

Grünwald-Letnikov (GL) [20], Riemann-Liouville (RL) [21], and Caputo [22] definitions are three different definitions for fractional calculus in Euclidean space. Due to its convenient computational form, GL definition for the fractional derivative is commonly used for engineering problems.

The GL derivative with an order of function is defined as:

$$\begin{aligned} {}_{a}^{GL}D_{x}^{\alpha}f(x) &= \frac{d^{\alpha}}{[d(x-a)]^{\alpha}}f(x) \\ &= \lim_{N\to\infty}\left\{\frac{\left(\frac{x-a}{N}\right)^{-\alpha}}{\Gamma(-\alpha)}\sum_{k=0}^{N-1}\frac{\Gamma(k-a)}{\Gamma(k+1)}f\left(x-k\left(\frac{x-a}{N}\right)\right)\right\} \end{aligned}, \quad (1)$$

where $f(x)$ is a differintegrable function, $[a,x]$ is the function duration, and Γ is the gamma function. Here, ${}_{a}^{GL}D_{x}^{\alpha}$ denotes the GL fractional differential operator.

In (1), when N is big enough, the limit symbol can be neglected and we can rewrite (1) as:

$$\frac{d^{\alpha}}{dx^{\alpha}}f(x) \cong \frac{x^{-\alpha}N^{\alpha}}{\Gamma(-\alpha)}\sum_{k=0}^{N-1}\frac{\Gamma(k-\alpha)}{\Gamma(k+1)}f\left(x-\frac{kx}{N}\right), \quad (2)$$

which is a proximate form substituting fractional derivative with multiplication and addition

operations [12]. For 1D signal, it has the following expression:

$$\frac{d^{\alpha}}{dx^{\alpha}}f(x) \cong f(x) + (-\alpha)f(x-1) + \frac{(-\alpha)(-\alpha+1)}{2}f(x-2) + \ldots + \frac{(-\alpha)(-\alpha+1)(-\alpha+2)\ldots(-\alpha+n)}{n!}f(x-n) \quad (3)$$

Particle swarm optimization with fractional -order position

Quantum particle swarm optimization

Trajectory analyses in [23] demonstrated that each particle should converge to the corresponding attractor C_i, which is given as follows:

$$C_{id}(t) = a \cdot pb_{id}(t) + (1-a) \cdot gb_d(t), a \sim U(0,1), \quad (4)$$

where $a = c_1r_1/(c_1r_1+c_2r_2)$. It can be seen that the local attractor is a stochastic attractor of particle i that lies in a hyper-rectangle with pb_{id} and gb_d being two ends of its diagonal.

Based on the convergence analysis of PSO [24], inspired by the theory of quantum physics, Sun et al. studied the convergence behavior of PSO and proposed a novel PSO model from quantum mechanics abbreviated as QPSO [25]. Based on the Delta potential, the quantum behavior of particles are considered. In the framework of quantum time-space, the quantum state of a particle can be defined by a wave function $\psi(x,t)$. In 3-D space, $\psi(x,t)$ is given as

$$|\psi|^2 dxdydz = Qdxdydz, \quad (5)$$

where Q is the probability that measures the particle's location in the 3-D space. As a probability density function, we have

$$\int_{-\infty}^{+\infty} |\psi|^2 dxdydz = \int_{-\infty}^{+\infty} Qdxdydz = 1. \quad (6)$$

The normalized version of ψ can be given as:

$$\psi(y) = \frac{1}{\sqrt{L}} e^{-|y|/L}, \quad (7)$$

As a result, Q and the corresponding distribution function F can be obtained as:

$$Q(y) = |\psi(y)|^2 = \frac{1}{L} e^{-2|y|/L}, \quad (8)$$

And

$$F(X_{id}(t+1)) = 1 - e^{-\frac{2|p_{id}(t) - X_{id}(t+1)|}{L_{id}(t)}}, \quad (9)$$

where $L_{id}(t)$ denotes the standard deviation, which describes the search range of each particle. The position of the particle can be obtained by Monte Carlo method with the following formula:

$$s = \frac{1}{L}u = \frac{1}{L}e^{-2|y|/L}, u = rand(0,1), \quad (10)$$

where s denotes a random constant, which is uniformly distributed on $U(0, {}^{1}/_{L})$.

Then, $u = e^{-2|y|/L}$. Let $y = x-c$, we have

$$x = c \pm \frac{L}{2}\ln\left(\frac{1}{u}\right). \tag{11}$$

The convergence condition of PSO is given by:

$$x \to c, when \quad t \to \infty. \tag{12}$$

Let L be the function of time, we have:

$$L = L(t)$$

$$L \to 0, when \quad t \to 0. \tag{13}$$

With (13), we have the iterative version of i-th multidimensional particle as follows

$$X_{id}(t+1) = C_{id}(t) \pm \frac{L_{id}(t)}{2}\ln\left(\frac{1}{u}\right). \tag{14}$$

A global point called mean best position is introduced to evaluate $L_{id}(t)$. The global point, which is denoted by *mbest*, can be computed as the mean of the *pbest* positions of all particles, which can be given as:

$$\begin{aligned} mbest(t) &= (mbest_1(t), mbest_2(t), \ldots, mbest_d(t)) \\ &= \frac{1}{n}\sum_{i=1}^{n} p_{i1}(t), \frac{1}{n}\sum_{i=1}^{n} p_{i2}(t), \ldots, \frac{1}{n}\sum_{i=1}^{n} p_{id}(t) \end{aligned}. \tag{15}$$

The values of $L_{id}(t)$ is calculated by:

$$L_{id}(t) = 2 \cdot \beta \cdot |m_d(t) - X_{id}(t)|. \tag{16}$$

Finally, the position can be given by:

$$X_{id}(t+1) = C_{id}(t) \pm \beta \cdot |mbest_d - X_{id}(t)|\ln\left(\frac{1}{u}\right), \tag{17}$$

where parameter β is step size, which is utilized to control the convergence speed. *rand* is a random number with a range of 0 to 1, which is the deciding factor of "±" in (17).

Table 1 illustrates the main steps of QPSO.

QPSO with the fractional-order position

It is well known that fractional calculus has a remarkable long-term memory characteristic [26]. From the definition of Grünwald-Letnikov in (1), it can be seen that fractional derivative is computed with all historical states and it is naturally suitable for the iterative procedure of intelligent optimization algorithms. For examples, Pires E.J.S introduced fractional calculus theory into the updated formula of particle swarm optimization algorithm [27].

To further improve the speed and accuracy of convergence of QPSO, in this section, the proposed QPSO with the fractional-order position is detailed. Initially, the original position is

Table 1. The main steps of QPSO.

Alogtihm2
Initialize QPSO parameters;
Repeat
For all particles **do**
compute f
If $f(x_i)<f(pb_i)$
$pb_i = X_i$
End
If $f(pb_i)<f(gb)$
$gb = pb_i$
End
Calculate Q using (19)
If $rand>0.5$
$X_{id}(t+1) = C_{id}(t) + \beta \cdot \lvert mbest_d - X_{id}(t)\rvert \cdot \ln\left(\frac{1}{u}\right)$
Else
$X_{id}(t+1) = C_{id}(t) - \beta \cdot \lvert mbest_d - X_{id}(t)\rvert \cdot \ln\left(\frac{1}{u}\right)$
End
$t = t+1$
Until stopping criteria

rearranged to modify the order of the position derivative, which can be derived as:

$$X_{id}(t+1) = C_{id}(t) + \beta \cdot \ln\left(\frac{1}{u}\right) \cdot (mbest_d - X_{id}(t)) \quad (rand > 0.5, mbest_d > X_{id}(t)), \quad (18)$$

$$X_{id}(t+1) = C_{id}(t) + \beta \cdot \ln\left(\frac{1}{u}\right) \cdot (X_{id}(t) - mbest_d) \quad (rand > 0.5, mbest_d < X_{id}(t)), \quad (19)$$

$$X_{id}(t+1) = C_{id}(t) - \beta \cdot \ln\left(\frac{1}{u}\right) \cdot (mbest_d - X_{id}(t)) \quad (rand < 0.5, mbest_d > X_{id}(t)), \quad (20)$$

$$X_{id}(t+1) = C_{id}(t) - \beta \cdot \ln\left(\frac{1}{u}\right) \cdot (X_{id}(t) - mbest_d) \quad (rand < 0.5, mbest_d < X_{id}(t)), \quad (21)$$

(23) and (26) can be uniformly rewritten as:

$$X_{id}(t+1) - X_{id}(t) = C_{id}(t) + \beta \cdot \ln\left(\frac{1}{u}\right) \cdot mbest_d - \left(\beta \cdot \ln\left(\frac{1}{u}\right) \pm 1\right) X_{id}(t). \quad (22)$$

The left side of (22) is the discrete version of the derivative with $\alpha = 1$ and we can extend (22) to a generalized version, leading to the following fractional-order expression

$$D^{\alpha}(X_{id}(t+1)) = C_{id}(t) + \beta \cdot \ln\left(\frac{1}{u}\right) \cdot mbest_d - \left(\beta \cdot \ln\left(\frac{1}{u}\right) \pm 1\right) X_{id}(t), \quad (23)$$

when $rand>0.5, mbest_d>X_{id}(t)$ and $rand<0.5, mbest_d<X_{id}(t)$.

Similarly, for $rand>0.5, mbest_d<X_{id}(t)$ and $rand<0.5, mbest_d>X_{id}(t)$, we have

$$D^{\alpha}(X_{id}(t+1)) = C_{id}(t) - \beta \cdot \ln\left(\frac{1}{u}\right) \cdot mbest_d + \left(\beta \cdot \ln\left(\frac{1}{u}\right) \pm 1\right) X_{id}(t). \quad (24)$$

Previous researches have demonstrated that while the order α of the derivative is set to [0,1], it will introduce a smoother variation and prolong memory effect, which may lead to a

better performance than original integral-order method [12][13]. To study the behavior of the proposed fractional-order strategy, a set of functions are tested and the order α is set to range from 0 to 1 with step size of $\Delta\alpha = 0.1$. To simplify the computational complexity, we usually truncate (3) and only use the first four terms, so we have

$$\begin{aligned} D^{\alpha}(X_{id}(t+1)) = {} & X_{id}(t+1) - \alpha X_{id}(t) \\ & - \frac{1}{2}\alpha(1-\alpha)X_{id}(t-1) \\ & - \frac{1}{6}\alpha(1-\alpha)(2-\alpha)X_{id}(t-2) \\ & - \frac{1}{24}\alpha(1-\alpha)(2-\alpha)(3-\alpha)X_{id}(t-3) \end{aligned} . \tag{25}$$

Then, (23) can be modified to

$$\begin{aligned} X_{id}(t+1) = {} & C_{id}(t) + \beta \cdot \ln\left(\frac{1}{u}\right) \cdot mbest_d \\ & - \left(\beta \cdot \ln\left(\frac{1}{u}\right) \pm 1 - \alpha\right) X_{id}(t) + XX_{id}(t) \end{aligned} , \tag{26}$$

and (24) can be also rewritten as

$$\begin{aligned} X_{id}(t+1) = {} & C_{id}(t) - \beta \cdot \ln\left(\frac{1}{u}\right) \cdot mbest_d \\ & + \left(\beta \cdot \ln\left(\frac{1}{u}\right) \pm 1 + \alpha\right) X_{id}(t) + XX_{id}(t) \end{aligned} , \tag{27}$$

where

$$\begin{aligned} XX_{id}(t) = {} & \frac{1}{2}\alpha(1-\alpha)X_{id}(t-1) \\ & + \frac{1}{6}\alpha(1-\alpha)(2-\alpha)X_{id}(t-2) \\ & + \frac{1}{24}\alpha(1-\alpha)(2-\alpha)(3-\alpha)X_{id}(t-3) \end{aligned} . \tag{28}$$

It can be seen that from (23) and (24), the position updating of particles depends not only on the position of the previous particle but also on the historical position of the particle in different points in time. The position updating of particles is the result of long-term memory, which can protect the population distribution and diversity to a certain extent. The flowchart of the proposed quantum-behaved swarm optimization with the fractional position (FQPSO) is shown in Table 2.

Experiments

Experimental setup

To validate the performance of the proposed FQPSO, 8 benchmark functions [28–30] listed in Table 3 were used to compare FQPSO with PSO and QPSO under the same maximum function evaluations (FEs). For FQPSO, the order was set to from 0.1 to 0.9 with step 0.1. Firstly, to investigate the impact of a fractional position in the proposed algorithm, we use FQPSO with different fractional-orders to compare to QPSO. Then, the best results of FQPSO were used for comparison with other variants of PSO including PSO [31], QPSO, PSO with both chaotic

Table 2. The main steps of FQPSO.

Alogtihm3
Initialize FQPSO parameters;
Initialize population: random X_i
For each particle $i\in[i,s]$
compute f
If $f(x_i)<f(pb_i)$
$pb_i = X_i$
End
If $f(pb_i)<f(gb)$
$gb = pb_i$
End
Calculate Q using the equation
If $rand>0.5, mbest_d<X_{id}(t)$ or $rand<0.5, mbest_d>X_{id}(t)$
$X_{id}(t+1) = C_{id}(t) + \beta\cdot\ln\left(\frac{1}{u}\right)\cdot mbest_d - \left(\beta\cdot\ln\left(\frac{1}{u}\right)\pm 1-\alpha\right)X_{id}(t) + XX_{id}(t)$
Else
If $rand>0.5, mbest_d>X_{id}(t)$ or $rand<0.5, mbest_d<X_{id}(t)$
$X_{id}(t+1) = C_{id}(t) - \beta\cdot\ln\left(\frac{1}{u}\right)\cdot mbest_d + \left(\beta\cdot\ln\left(\frac{1}{u}\right)\pm 1+\alpha\right)X_{id}(t) + XX_{id}(t)$
End
$t = t+1$
Until termination criterion is satisfied

sequences and crossover operation(CCPSO) [32], naive PSO(NPSO) [33] and moderate-random-search strategy PSO(MRPSO) [34]. The parameters of the compared algorithms were set as recommended in the original references. Since the impact of population size on the performance of PSO-based methods is of the minimum significance [35], all experiments in this research were performed with a population size of 20. [34].

The parameters of the compared algorithms were set as recommended in the original references. Since the impact of population size on the performance of PSO-based methods is of the minimum significance [35], all experiments in this research were performed with a population size of 20. β is computed according to the following formula:

$$\beta(t) = (\beta_0 - \beta_1)(t_{\max} - t)/t_{\max} + \beta_1, \tag{29}$$

where $\beta_0 = 0.8$, $\beta_1 = 0.6$, t is the current number of iterations and $t_{\max}$ is the maximum number of iterations [36].

Testing FQPSO with different fractional-order

Since QPSO is a stochastic algorithm, it will lead to a different trajectory convergence every time. Therefore, the simulations were performed 50 times with each value in the parameter set $\alpha = \{0,0.1,0.2,\ldots,1\}$. In Figs 1 and 2, the result is given for the adopted optimization functions f_j, $j = 1,2,\ldots,8$. To show the gains achieved by our proposed algorithm, three groups of the experiments were performed. In unimodal functions (f_1-f_5, Group 1) and multimodal functions (f_6-f_8, Group 2) tests, the maximum numbers of FEs were set to 10000, 30000 and 100000, for 10-D, 30-D and 100-D problems, respectively. In the results, we provided the best results and the mean results. The final results over 50 runs of FQPSO are summarized in Tables 4–7.

Fig 1 shows the performance of FQPSO with different fractional-orders in Group 1. f_1, a Sphere function, is the most widely used unimodal test function. Compared with algorithms with integer-order position, FQPSO shows the best results for this function. Similar results were obtained for other unimodal functions. The improvements achieved by FQPSO on these unimodal functions suggest that fractional-order methods are better at a fine-gained search

Table 3. Benchmark test functions.

F	Formula	Range	X_{max}	f_{min}	X^*
f_1	$\sum_{i=1}^{n} x_i^2$	[-100,100]	100	0	0
f_2	$\sum_{i=1}^{n} \left(\sum_{j=1}^{i} x_j\right)^2$	[-100,100]	100	0	0
f_3	$\sum_{i=1}^{n} i \cdot x_i^2$	[-100,100]	100	0	0
f_4	$\sum_{i=1}^{n} \lvert x_i \rvert + \prod_{i=1}^{n} \lvert x_i \rvert$	[-10,10]	100	0	0
f_5	$\sum_{i=1}^{n} (x_i)^2 + \prod_{i=1}^{n} (x_i)^2$	[-10,10]	100	0	0
f_6	$\sum_{i=1}^{n} (x_i^2 - 10\cos(2\pi x_i) + 10)$	[-100,100]	100	0	0
f_7	$\sum_{i=1}^{n} (\sum_{k=0}^{20} (0.5)^k \cos(2\pi(3)^k(x_i + 0.5)))$ $-n\sum_{k=0}^{20} ((0.5)^k \cos(2\pi \cdot 3^k \cdot 0.5))$	[-5.12,5.12]	5.12	0	0
f_8	$-20\exp\left(-0.2\left(1/n \sum_{i=1}^{n} x_i^2\right)^{1/2}\right)$ $-\exp\left(\frac{1}{n}\sum_{i=1}^{n} \cos 2\pi x_i\right) + 20 + e$	[-5.12,5.12]	32	0	0

X^* denotes the global optimum.

than integer-order ones. However, it is also worth noting that the performances of FQPSO algorithms with orders 0.1 and 0.2 were not better than integer-order. The reason is that (35) is just an approximation of D^{α} and the approximation accuracy of $D^{0.1}$ and $D^{0.2}$ is not good enough. From Fig 1, we can see that most FQPSO methods' convergence accuracies are better

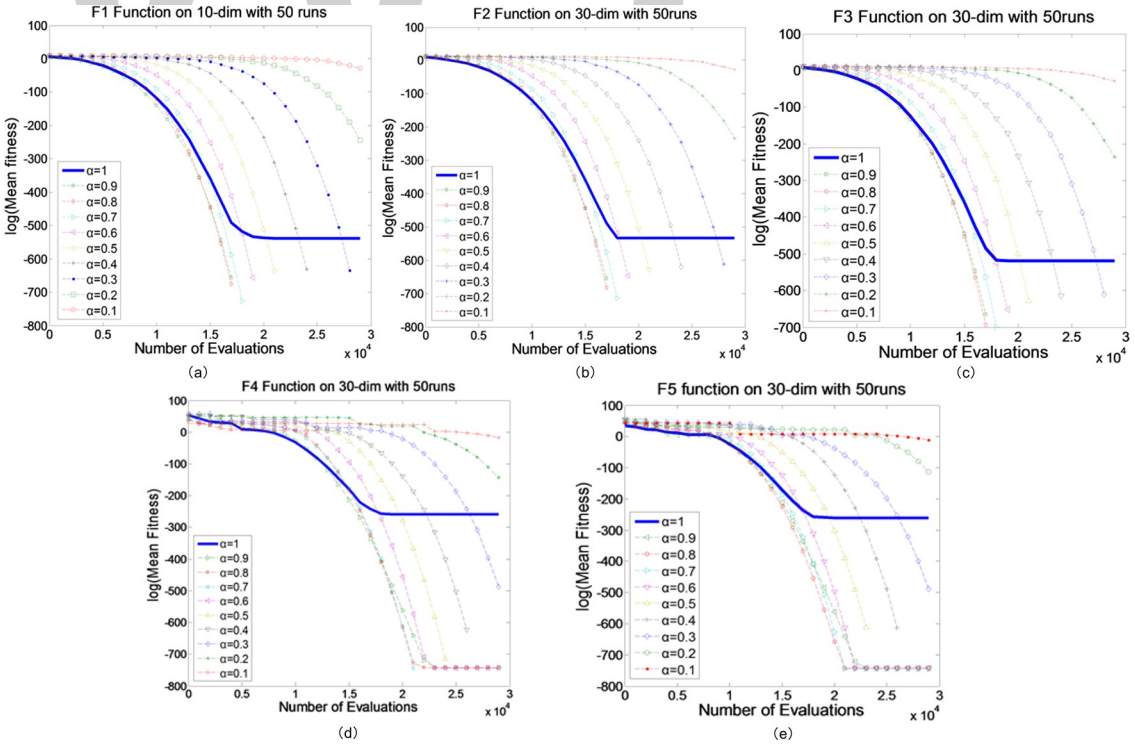

Fig 1. Comparison between FQPSO with different fractional-order on Group 1. (a) f_1, (b) f_2, (c) f_3, (d) f_4, (e) f_5.

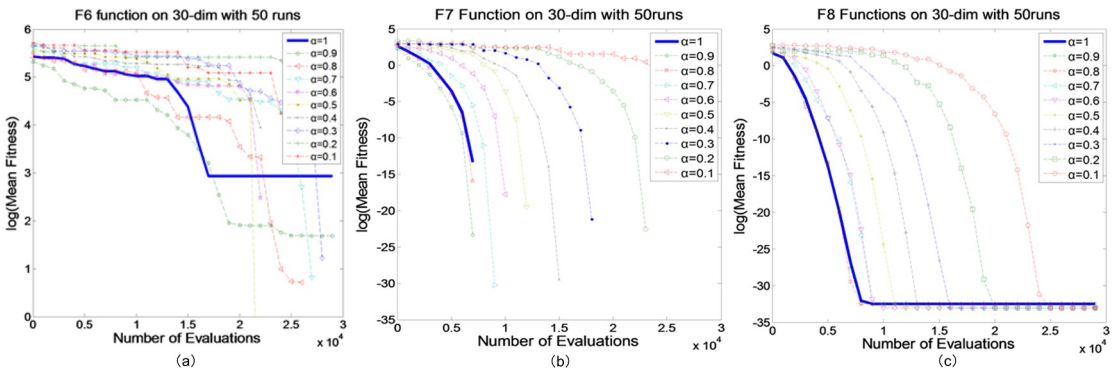

Fig 2. Comparison between FQPSO with different fractional-order on Group 2. (a) f_6, (b) f_7, (c) f_8.

than QPSO. For 10-D and 30-D problems in function 1, 2 and 3 showed in Fig 1A, 1B and 1C and Tables 5 and 6, when $0.3 \leq \alpha \leq 0.9$, the convergence accuracies are better than QPSO. For 10-D and 30-D problems in function 4, the convergence accuracies of are better than QPSO, when $0.4 \leq \alpha \leq 0.9$. For 10-D and 30-D problems in function 5, the convergence accuracies of are better than QPSO, when $0.2 \leq \alpha \leq 0.9$. Tables 5–7 also show that the convergence accuracies are better than QPSO in function 1, 2, 3, 4 and 5 when $0.7 \leq \alpha \leq 0.9$ for100-D problems.

In general, we can always find an appropriate fractional order so that the convergence accuracy of the algorithm is better than the integer order algorithm in Group 1.

In Fig 2, for f_6, f_7 and f_8, the numbers of local minima will increase dramatically as the dimension of the function raises. In this part, we mainly investigated the capability of global searching. f_6 is the generalized Rastrigin's function, which is the most widely used test multimodal functions in PSO algorithm, and tends to be trapped by local minimums. Considering

Table 4. Comparison between FQPSOs with different fractional-order on function 1–2.

Fractional-order		f_1			f_2		
		Dim = 10 FEs = 10000	**Dim = 30 FEs = 30000**	**Dim = 100 FEs = 30000**	**Dim = 10 FEs = 10000**	**Dim = 30 FEs = 30000**	**Dim = 100 FEs = 30000**
$\alpha = 1$	Best	5.6324e-267	2.3453e-241	2.9833e-230	3.7568e-258	1.7033e-237	1.4328e-163
	Mean	4.5321e-265	1.5837e-239	3.5636e-231	1.8379e-255	6.8877e-234	4.5673e-158
$\alpha = 0.9$	Best	0	0	5.3234e-269	0	0	5.3293e-204
	Mean	0	0	4.5639e-267	0	0	6.3214e-201
$\alpha = 0.8$	Best	0	0	7.3535e-248	0	0	8.5313e-198
	Mean	0	0	6.4356e-246	0	0	4.2314e-197
$\alpha = 0.7$	Best	0	0	5.3623e-242	0	0	5.3241e-188
	Mean	0	0	7.5323e-239	0	0	3.1235e-185
$\alpha = 0.6$	Best	0	0	7.4342e-197	0	0	8.4232e-163
	Mean	0	0	6.4329e-194	0	0	5.3123e-161
$\alpha = 0.5$	Best	0	0	5.3252e-146	0	0	4.3242e-141
	Mean	0	0	5.9753e-143	0	0	8.5223e-140
$\alpha = 0.4$	Best	0	0	9.5332e-108	0	0	5.3213e-111
	Mean	0	0	5.4256e-99	0	0	6.4132e-108
$\alpha = 0.3$	Best	0	0	6.5352e-56	0	0	8.5231e-66
	Mean	0	0	5.953e-50	0	0	5.3145e-49
$\alpha = 0.2$	Best	2.1613e-236	5.1257e-106	6.4235e-18	2.1552e-152	1.0355e-105	1.2345e-16
	Mean	6.6966e-250	2.2847e-100	3.4562e-11	7.0445e-145	2.2098e-98	8.5242e-08
$\alpha = 0.1$	Best	2.613e-53	3.5677e-16	4.5712e-08	1.4244e-36	4.5673e-15	5.3113e-06
	Mean	1.2778e-45	3.9080e-12	3.4564e-05	2.0523e-24	2.4097e-11	4.5313e-04

Table 5. Comparison between FQPSOs with different fractional-order on function 3–4.

Fractional-order		f_3			f_4		
		Dim = 10 FEs = 10000	**Dim = 30 FEs = 30000**	**Dim = 100 FEs = 30000**	**Dim = 10 FEs = 10000**	**Dim = 30 FEs = 30000**	**Dim = 100 FEs = 30000**
$\alpha = 1$	Best	6.325e-259	1.0111e-228	4.3529e-194	1.4622e-248	6.2412e-111	6.4353e-56
	Mean	7.424e-255	1.2325e-224	7.5332e-179	1.2011e-215	2.6011e-99	8.4224e-41
$\alpha = 0.9$	Best	0	0	8.4243e-223	0	3.5000e-323	7.5224e-75
	Mean	0	0	6.5324e-215	0	5.9000e-323	6.3243e-71
$\alpha = 0.8$	Best	0	0	9.5352e-245	0	1.0000e-323	8.4242e-77
	Mean	0	0	5.3563e-228	0	4.9000e-324	4.2412e-73
$\alpha = 0.7$	Best	0	0	7.5324e-237	0	0	6.3242e-65
	Mean	0	0	6.4256e-229	0	0	5.4224e-61
$\alpha = 0.6$	Best	0	0	8.5363e-185	0	0	9.4245e-39
	Mean	0	0	6.4353e-178	0	1.000e-323	6.2345e-36
$\alpha =0.5$	Best	0	0	9.5363e-166	0	0	4.2135e-34
	Mean	0	0	7.5352e-161	0	0	1.2356e-33
$\alpha = 0.4$	Best	0	0	8.4256e-154	6.2616e-251	0	5.2214e-27
	Mean	0	0	9.5324e-151	5.5247e-243	0	5.6241e-21
$\alpha = 0.3$	Best	0	0	6.3245e-134	2.6343e-147	3.7503e-228	7.5213e-18
	Mean	0	3.78e-321	5.4242e-131	9.3570e-141	2.8574e-221	5.2134e-15
$\alpha = 0.2$	Best	2.1613e-236	9.7128e-104	5.3523e-67	1.5234e-74	3.9143e-73	5.6231e-12
	Mean	6.6966e-220	4.7649e-100	6.3213e-60	2.0325e-68	5.1233e-69	7.5234e-07
$\alpha = 0.1$	Best	1.2583e-45	7.5296e-15	3.4525e-06	5.9892e-30	6.6300e-13	4.3241e-08
	Mean	5.6267e-39	7.9333e-12	5.3256e-03	2.2309e-25	4.2332e-10	3.3413e-05

more orders to search in the solution space, FQPSO gets more favorable results than the compared algorithms. f_7 is the Ackley function and according to the results in Table 8, the performances of FQPSO have little changes with the variation of dimension and achieve the best

Table 6. Comparison between FQPSOs with different fractional-order on function 5–6.

Fractional-order		f_5			f_6		
		Dim = 10 FEs = 10000	**Dim = 30 FEs = 30000**	**Dim = 100 FEs = 30000**	**Dim = 10 FEs = 10000**	**Dim = 30 FEs = 30000**	**Dim = 100 FEs = 30000**
$\alpha = 1$	Best	9.397e-189	1.1027e-154	5.4324e-109	0.9950	11.9395	32.4524
	Mean	3.5943e-185	1.1027e-155	4.5632e-110	1.5919	16.0250	56.3234
$\alpha = 0.9$	Best	0	0	6.4242e-134	1.7764e-15	0.9950	1.7432
	Mean	0	0	5.6321e-130	0.1090	6.6484	4.5252
$\alpha = 0.8$	Best	0	0	3.4245e-129	0	0.4517	0.9985
	Mean	0	0	5.6324e-126	0	2.2933	10.4245
$\alpha = 0.7$	Best	0	0	8.7432e-119	0	0.0297	0.5943
	Mean	0	0	4.5256e-115	0	2.4696	3.4255
$\alpha = 0.6$	Best	0	0	6.5241e-89	0	0.5342	1.4252
	Mean	0	0	5.3241e-88	0	12.4468	9.3245
$\alpha = 0.5$	Best	0	0	8.5352e-82	2.456e-08	4.5314e-06	0.04255
	Mean	0	0	5.4232e-77	0.01413	6.4245e-04	0.4256
$\alpha = 0.4$	Best	0	0	9.4256e-74	0.1389	6.6789	11.3352
	Mean	0	0	6.6322e-72	1.5126	56.6533	57.3241
$\alpha = 0.3$	Best	4.2206e-251	5.6014e-317	5.3214e-66	0.3100	0.6789	16.4252
	Mean	2.3626e-231	1.6254e-286	5.3242e-65	0.3394	3.3932	54.1343
$\alpha = 0.2$	Best	1.5239e-120	2.4207e-99	1.2345e-49	1.6976e-10	10.2080	26.4952
	Mean	2.9750e-102	1.2928e-77	2.4562e-46	2.1109e-09	185.937	192.345
$\alpha = 0.1$	Best	4.5632e-42	1.2071e-16	9.5224e-08	1.0415	36.4428	86.4211
	Mean	2.9225e-30	1.6344e-06	3.4521e-04	16.0174	66.9430	211.245

Table 7. Comparison between FQPSOs with different fractional-order on function 7–8.

Fractional-order		f_7			f_8		
		Dim = 10 FEs = 10000	**Dim = 30 FEs = 30000**	**Dim = 100 FEs = 30000**	**Dim = 10 FEs = 10000**	**Dim = 30 FEs = 30000**	**Dim = 100 FEs = 30000**
$\alpha = 1$	Best	7.1054e-15	2.5251e-10	8.5322e-08	1.4622e-248	7.9936e-15	4.5231e-06
	Mean	9.9476e-15	4.8798e-07	5.4252e-07	1.2011e-215	7.9936e-15	3.4251e-05
$\alpha = 0.9$	Best	1.9257e-17	5.6302e-11	6.5232e-15	0	4.4409e-15	2.3451e-07
	Mean	1.8609e-13	3.7460e-08	7.5323e-11	0	6.9278e-15	4.4123e-06
$\alpha = 0.8$	Best	4.9016e-25	5.4000e-10	7.4213e-19	0	1.6409e-15	8.4222e-07
	Mean	5.6360e-18	1.1253e-07	9.3421e-15	0	5.1514e-15	5.3245e-06
$\alpha = 0.7$	Best	8.3079e-33	5.4000e-20	6.4231e-24	0	4.4409e-15	6.4134e-07
	Mean	1.1100e-22	8.0348e-15	5.3313e-19	0	4.4708e-15	5.3121e-06
$\alpha = 0.6$	Best	5.0220e-44	3.4005e-12	1.2334e-14	0	4.4409e-15	4.3311e-07
	Mean	2.4418e-26	1.0209e-08	5.4134e-13	0	4.4409e-15	4.2134e-06
$\alpha = 0.5$	Best	6.5183e-43	1.6486e-11	7.4231e-09	0	4.4409e-15	1.2134e-08
	Mean	9.7154e-22	2.2779e-09	9.3134e-06	0	4.4409e-15	5.4131e-06
$\alpha = 0.4$	Best	4.765e-28	3.1412e-17	8.4255e-10	6.2616e-251	4.4409e-15	3.4131e-07
	Mean	8.98017e-13	3.1149e-14	5.3424e-08	5.5247e-243	4.4409e-15	3.4111e-06
$\alpha = 0.3$	Best	2.0456e-16	3.9874e-13	8.4325e-12	2.6343e-147	3.7503e-15	5.3314e-08
	Mean	1.1142e-14	1.0676e-10	4.5231e-10	9.3570e-141	2.8574e-15	4.3141e-07
$\alpha = 0.2$	Best	3.1005e-19	4.2733e-17	3.4112e-05	1.5234e-74	6.7564e-15	6.5131e-03
	Mean	6.01e-13	2.6818e-12	1.2144e-03	2.0325e-68	6.9345e-15	4.5131e-01
$\alpha = 0.1$	Best	0.004777	0.0813	0.1133	5.9892e-30	7.9835e-15	0.1314
	Mean	0.0865	0.7666	1.2134	2.2309e-25	8.2343e-15	1.2144

results on each dimension. Function f_8 is the Weierstrass function, which is continuous everywhere, but differentiable nowhere. In short, FQPSO reaches the global optimum on 10 and 30 dimensions. In Fig 2 and Tables 6 and 7, it can be observed that except FQPSO with orders 0.1 and 0.2, FQPSO can always achieve better results than QPSO. Meanwhile, for function 6, 7 and 8, the convergence accuracies are better than QPSO when $0.3 \leq \alpha \leq 0.9$.

In summary, FQPSO has superior ability in tackling multimodal functions compared with other algorithms. We can always find an appropriate fractional order for the algorithm that has better convergence accuracy than the integer order one in Group 2.

Table 8 shows the time consumption of FQPSO and QPSO in solving function optimization problems. The default time unit is seconds. The experimental results also confirm that the fractional order method only consumes a little more time in each iteration process, and does not cause a lot of waste of time.

Compare with other variants of PSO

In this experiment, the best results of the FQPSO methods were used for comparison with other variants of PSO, including PSO, QPSO, CCPSO, NPSO and MRPSO. The parameters of the compared algorithms were set according to the recommendations in their original papers. The maximum numbers of FEs were respectively set to 10000 and 30000 for solving 10-D and 30-D problems. All experiments were performed with a population size of 20.

Table 8. Time consumption.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
QPSO	0.8123	0.9245	0.8155	0.8934	0.8688	0.9355	0.9642	0.9942
FQPSO	0.8471	0.9334	0.8358	0.9548	0.8899	0.9674	0.9856	1.032

Table 9. Comparison between different PSO algorithms on function 1–3.

Algorithm		f_1		f_2		f_3	
		Dim = 10 FEs = 10000	**Dim = 30 FEs = 30000**	**Dim = 10 FEs = 10000**	**Dim = 30 FEs = 30000**	**Dim = 10 FEs = 10000**	**Dim = 30 FEs = 30000**
FQPSO	Best	0	0	0	1.6812e-313	0	0
	Mean	0	0	0	1.9619e-296	0	0
	std	0	0	0	5.3432e-307	0	0
QPSO	Best	5.6324e-267	2.3453e-241	3.7568e-258	1.7033e-237	1.5e-256	1.01e-243
	Mean	4.5321e-265	1.5837e-239	1.8379e-255	6.8877e-234	8.1e-254	1.86e-241
	std	3.4523e-266	6.3245e-241	2.5431e-256	2.9832e-237	5.31e-256	7.543e-243
PSO	Best	7.764e-20	6.7954e-14	5.8742e-16	4.3257e-10	5.8734e-21	3.876e-12
	Mean	1.17e-20	1.58e-13	4.9000e-15	1.2264e-09	9.3854e-20	2.086e-06
	std	6.3e-20	4.17e-13	1.3864e-16	2.4987e-10	3.5467e-19	6.4303e-11
CCPSO	Best	3.2341e-97	7.4324e-86	1.2353e-20	9.3557e-16	6.3258e-43	5.2134e-35
	Mean	1,2313e-95	6.0851e-84	8.3483e-20	1.9619e-13	8.5423e-42	3.6880e-33
	std	2.8734e-97	5.3241e-85	5.9834e-20	5.7934e-15	1.5425e-43	7.3424e-34
NPSO	Best	3.4653e-53	5.3789e-38	3.4453e-22	5.2223e-13	1.7431e-14	9.3452e-09
	Mean	5.2356e-52	4.8357e-36	9.5151e-18	1.2580e-11	3.4564e-14	7.2875e-06
	std	2.3456e-53	9.2134e-37	2.3456e-21	9.9863e-13	2.6731e-14	8.4245e-08
MRPSO	Best	1.5677e-110	5.8723e-97	6.3434e-61	4.9053e-45	3.4546e-85	3.4546e-85
	Mean	1.239e-109	6.2391e-93	3.5639e-48	4.4788e-44	9.4671e-84	9.4671e-84
	std	8.2345e-110	6.3394e-97	9.5332e-51	7.3421e-44	2.3546e-85	2.3546e-85

Tables 9 and 10 shows the statistical results of different algorithms on unimodal functions. From the previous results in the last subsection, we can see that FQPSO with $D^{0.8}$ obtained the best results on functions 1–3 and FQPSO with $D^{0.7}$ achieved the best results on functions 4–5. We fixed the orders to compare those results with other variants of PSO. The results from different algorithms on these five unimodal functions suggest that FQPSO is better at a fine-gained search than all the other algorithms. The rapid convergence of the FQPSO can be seen as an evidence for our observation in Fig 3. In summary, FQPSO performs best in solving

Table 10. Comparison between different PSO algorithms on function 4–6.

Algorithm		f_4		f_5		f_6	
		Dim = 10 FEs = 10000	**Dim = 30 FEs = 30000**	**Dim = 10 FEs = 10000**	**Dim = 30 FEs = 30000**	**Dim = 10 FEs = 10000**	**Dim = 30 FEs = 30000**
FQPSO	Best	0	0	0	0	0	7.0106e-06
	Mean	0	0	0	0	0	1.4384
	std	0	0	0	0	0	1.234
QPSO	Best	1.4622e-248	6.2412e-146	9.397e-189	1.1027e-154	1.7764e-10	15.9395
	Mean	1.2011e-215	1.2203e-137	3.5943e-185	1.1027e-155	2.1453e-08	16.0250
	std	5.6231e-235	4.562e-137	6.3423e-187	6.3453e-155	5.4214e-09	5.3453
PSO	Best	5.324e-13	1.0000e-09	5.4234e-19	5.324e-07	4.34e-06	7.3474
	Mean	1.5242e-11	7.5267e-07	2.534e-18	0.0571	3.2e-05	7.8363
	std	2.584e-11	6.3324e-08	3.4532e-19	0.000424	3.2e-05	1.3456
CCPSO	Best	1.324e-30	7.4352e-23	6.4234e-55	5.3256e-44	0.1234	2.9768
	Mean	8.3453e-27	6.3257e-21	4.5313e-51	9.4075e-42	0.1423	4.475
	std	4.5356e-29	5.6485e-22	8.5423e-54	3.4578e-43	0.0542	6.4246
NPSO	Best	5.6354e-09	7.5242e-07	6.4312e-183	4.5353e-175	0.34561	5.323
	Mean	4.3563e-07	1.8913e-04	4.5683e-180	2.2115e-169	0.40593	5.354
	std	2.3446e-09	4.6452e-06	9.4563e-182	3.4532e-172	0.5289	0.0034
MRPSO	Best	6.4232e-154	5.3133e-150	8.6431e-212	4.3534e-209	2.456e-08	4.567
	Mean	7.4323e-151	2.9806e-147	6.4331e-210	1.0762e-203	4.355e-06	4.678
	std	4.2456e-154	4.5624e-149	6.7456e-210	3.5356e-206	5.342e-07	0.543

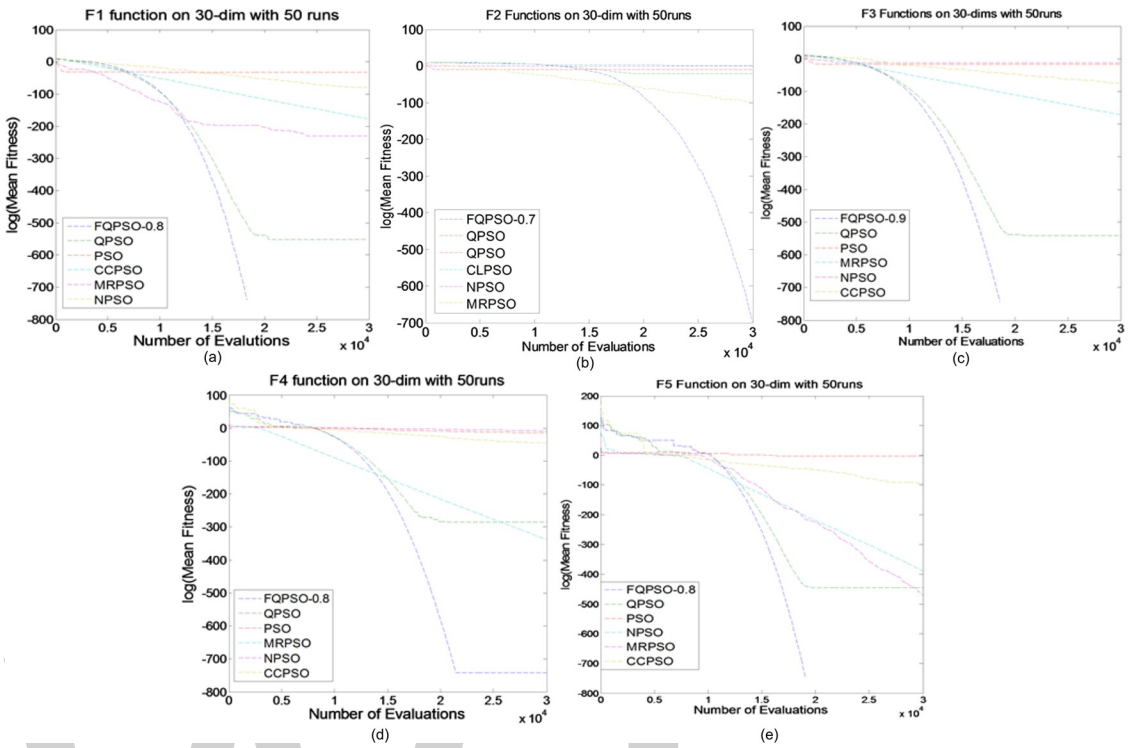

Fig 3. Comparison between different PSO algorithms on Group 1. (a) f_1, (b) f_2, (c) f_3, (d) f_4, (e) f_5.

unimodal functions among all the algorithms. Tables 9 and 10 and Fig 4 show the performances of different algorithms on Group 2.

Tables 9 and 10 shows the statistical results of different algorithms on unimodal functions. From the previous results in the last subsection, we can see that FQPSO with $D^{0.8}$ obtained the best results on functions 1–3 and FQPSO with $D^{0.7}$ achieved the best results on functions 4–5. We fixed the orders to compare those results with other variants of PSO. The results from different algorithms on these five unimodal functions suggest that FQPSO is better at a fine-gained search than all the other algorithms.

The rapid convergence of the FQPSO can be seen as an evidence for our observation in Fig 3.

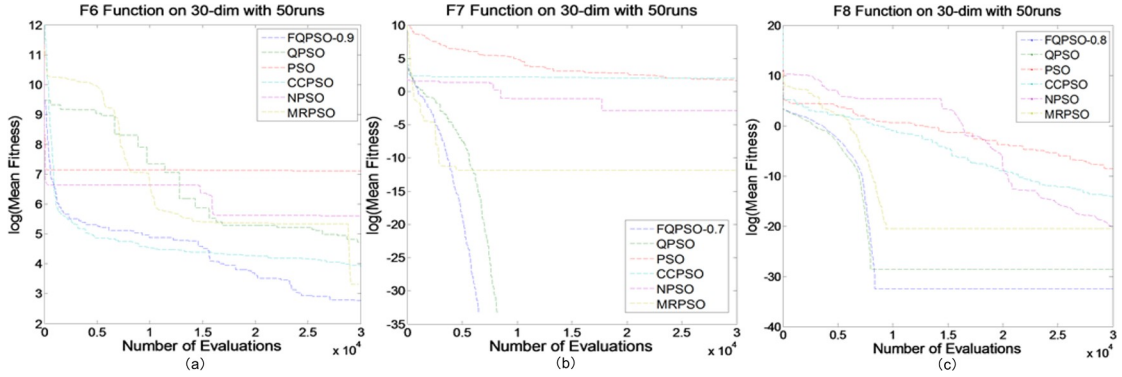

Fig 4. Comparison between different PSO algorithms on Group 2. (a) f_6, (b) f_7, (c) f_8.

Table 11. Comparison between different PSO algorithms on function 7–8.

Algorithm		f_1		f_2	
		Dim = 10 FEs = 10000	**Dim = 30 FEs = 30000**	**Dim = 10 FEs = 10000**	**Dim = 30 FEs = 30000**
FQPSO	Best	7.1054e-15	0	7.6438e-34	1.6409e-15
	Mean	9.9476e-15	0	5.3222e-29	5.1514e-15
	std	8.4534e-14	0	7.4345e-33	3.4356e-15
QPSO	Best	1.9257e-35	2.5251e-26	1.7059e-19	7.9936e-15
	Mean	1.8609e-33	3.6749e-21	6.8877e-20	7.9936e-15
	std	4.3456e-34	5.4245e-26	7.4352e-19	0
PSO	Best	4.9016e-08	41.3	5.6534e-14	5.4326e-05
	Mean	8.345e-06	45.043	6.4523e-13	7.5331e-04
	std	5.4345e-07	10.413	4.5634e-14	4.5674e-05
CCPSO	Best	8.3079e-05	0.0035	6.5341e-19	7.4243e-09
	Mean	1.1100e-04	3.8253	3.2145e-18	4.4509e-08
	std	4.5634e-05	6.4231	5.6313e-19	6.4234e-09
NPSO	Best	1.7431e-14	3.4531e-05	3.4356e-22	5.3563e-13
	Mean	3.4564e-14	0.0801	5.3561e-19	8.4356e-13
	std	2.6731e-14	0.0004	5.5546e-22	6.4325e-13
MRPSO	Best	6.5183e-11	1.6486e-08	5.6356e-27	6.4382e-14
	Mean	9.7154e-09	4.678e-05	1.2343e-24	9.3435e-14
	std	4.5623e-10	6.3245e-07	3.4465e-26	7.4231e-14

In summary, FQPSO performs best in solving unimodal functions among all the algorithms. Tables 10 and 11 and Fig 4 show the performances of different algorithms on Group 2.

From the previous results in the last subsection, it can be noticed that FQPSO with $D^{0.9}$ obtained the best result on function 6, FQPSO with $D^{0.7}$ got the best result on function 7, and FQPSO with $D^{0.9}$ achieved the best result on function 8. We also fixed the orders to compare those results with other variants of PSO. It can be seen that FQPSO obtains the global optimum on 10 and 30 dimensions. FQPSO is better to deal with multimodal functions than other algorithms.

In the results of different PSOs on 30 dimensions also supports our conclusion that FQPSO is suitable for multimodal functions. In summary, FQPSO performs best in solving both unimodal and multimodal functions among all the algorithms.

Conclusion

Inspired by the properties of fractional calculus, we presented a novel QPSO algorithm incorporated with fractional calculus strategy, which is based on the properties of long time memory and non-locality of fractional calculus. The goal is to employ the proposed method to accelerate not only the convergence speed but also avoid the local optimums. Since the property of fractional calculus enables quantum-particles in FQPSO to appear anywhere during iterations, it significantly improves the global searching ability. Furthermore, FQPSO also increases the convergence rate for the quantum particles. As a result, the proposed FQPSO method achieves more favorable results than all the other algorithms.

Author Contributions

Conceptualization: Lai Xu, Yifei Pu.

Data curation: Lai Xu, Aamir Muhammad, Jiliu Zhou.

Formal analysis: Lai Xu, Aamir Muhammad.

Funding acquisition: Jiliu Zhou, Yi Zhang.

Methodology: Lai Xu.

Project administration: Yi Zhang.

Software: Lai Xu, Aamir Muhammad.

Supervision: Yifei Pu, Jiliu Zhou, Yi Zhang.

Validation: Aamir Muhammad, Jiliu Zhou, Yi Zhang.

Visualization: Lai Xu.

Writing – original draft: Lai Xu.

Writing – review & editing: Yi Zhang.

References

1. Eberhart R.; Kennedy J. A new optimizer using particle swarm theory. In Proceedings of International Symposium on MICRO Machine and Human Science.
2. Goldberg D.E. Genetic Algorithms in Search. Optimization and Machine Learning. 1989, 7, 2104–2116.
3. Yuryevich J.; Wong K.P. Evolutionary programming based optimal power flow algorithm. *IEEE Trans Power Syst*. 1999, 14, 1245–1250.
4. Mezuramontes E.; Coello C.A.C. A simple multimembered evolution strategy to solve constrained optimization problems. *IEEE Transactions on Evolutionary Computation*. 2005, 9, 1–17.
5. Nordin P. Genetic Programming III—Darwinian Invention and Problem Solving. *IEEE Transactions on Evolutionary Computation*. 2002, 3, 251–253.
6. Bansal J.C.; Deep K. A Modified Binary Particle Swarm Optimization for Knapsack Problems. Applied Mathematics & Computation. 2012, 218, 11042–11061.
7. Wang K.P.; Huang L,; Zhou C.G. Particle swarm optimization for traveling salesman problem. *Acta Scientiarium Naturalium Universitatis Jilinensis*.2003, 3, 1583–1585.
8. Omran M.; Engelbrecht A.P.; Salman A. Particle Swarm Optimization Method Fof Image Clustering. International Journal of Pattern Recognition & Artificial Intelligence. 2005, 19, 297–321.
9. Benioff P. The computer as a physical system: A microscopic quantum mechanical Hamiltonian model of computers as represented by Turing machines. *Journal of Statistical Physics*. 1980, 22, 563–593.
10. Kane B.E. A silicon-based nuclear spin quantum computer. *Nature*, 1998, 393, 133–137.
11. Steffen M.; Vandersypen L.; Breyta G. Experimental Realization of Shor's quantum factoring algorithm. *American Physical Society*. 2002, 6866.
12. Sun J.; Feng B.; Xu W. Particle swarm optimization with particles having quantum behavior. In Proceedings of Congress on Evolutionary Computation, 2004.
13. Koeller R.C. Applications of Fractional Calculus to the Theory of Viscoelasticity. *Transactions of the Asme Journal of Applied Mechanics*. 1984, 51, 299–307.
14. Ciesielski M.; Leszczynski J. Numerical treatment of an initial-boundary value problem for fractional partial differential equations. Signal Processing, 2006, 86, 2619–2631.
15. Buyukkilic F.; Bayrakdar Z.O.; Demirhan D Investigation of the cumulative diminution process using the Fibonacci method and fractional calculus. Physica A Statistical Mechanics & Its Applications. 2016, 444, 336–344.
16. Xi M.; Sun J.; Xu W. An improved quantum-behaved particle swarm optimization algorithm with weighted mean best position. *Advanced Materials Research*. 2008, 591–593, 376–1380.
17. Jiao L.; Stolkin R.; Shang R. Dynamic-context cooperative quantum-behaved particle swarm optimization based on multilevel thresholding applied to medical image segmentation. Information Sciences. 2015, 294, 408–422.
18. Pu Y.F. Fractional-Order Euler-Lagrange Equation for Fractional-Order Variational Method: A Necessary Condition for Fractional-Order Fixed Boundary Optimization Problems in Signal Processing and Image Processing. IEEE Access. 2016, 99, 1–1.
19. Pu Y.F.; Siarry P.; Chatterjee A.; Wang Z.N.; Zhang Y.; Liu Y.G.; Zhou J.L.; Wang Y. A Fractional-Order Variational Framework for Retinex: Fractional-Order Partial Differential Equation-Based Formulation for

Multi-Scale Nonlocal Contrast Enhancement with Texture Preserving. *IEEE Transactions on Image Processing*. 2017, 27, 1214–1229. https://doi.org/10.1109/TIP.2017.2779601 PMID: 29990194

20. Scherer R.; Kalla S.L.; Tang Y. The Grunwald-Letnikov method for fractional differential equations. Computers & Mathematics with Applications. 2011, 62, 902–917.
21. Abbas S.; Benchohra M. Nonlinear Fractional Order Riemman-Liouville Volterra-Stieltjes Partial Integral Equations on Unbounded Domains. *Communications in Mathematical Analysis*. 2013, 14, 104–117.
22. Abdeljawad T. On Riemann and Caputo fractional differences. Computers & Mathematics with Applications. 2011, 62, 1602–1611.
23. Clerc M.; Kennedy J. The particle swarm: explosion, stability, and convergence in a multi-dimensional complex space. *IEEE Transactions on Evolutionary Computation*. 2002, 6, 58–73.
24. Bonyadi M.R.; Michalewicz Z. Analysis of Stability, Local Convergence, and Transformation Sensitivity of a Variant of the Particle Swarm Optimization Algorithm. *IEEE Transactions on Evolutionary Computation*, 2016, 20, 370–385.
25. Sun J.; Feng B.; Xu W.B. Particle swarm optimization with particles having quantum behavior. In Proceedings of Proc Congress on Evolutionary Computation. 2004.
26. Pu Y.F.; Zhou J.L.; Yuan X. Fractional Differential Mask: A Fractional Differential-Based Approach for Multiscale Texture Enhancement. *IEEE Transactions on Image Processing*. 2010, 19, 491–511. https://doi.org/10.1109/TIP.2009.2035980 PMID: 19933015
27. Pires E.J.S.; Machado J.A.T.; Oliveira P.B.D.M. Particle swarm optimization with fractional-order velocity. Nonlinear Dynamics. 2010, 61, 295–301.
28. Kiranyaz S.; Ince T. Yildirim A. Fractional Particle Swarm Optimization in Multidimensional Search Space. IEEE Transactions on Systems Man & Cybernetics *Part B Cybernetics*. 2010, 40, 298–319.
29. Liang J.J.; Suganthan P.N.; Deb K. Novel composition test functions for global numerical optimization. In Proceedings of Swarm Intelligence Symposium. 2015.
30. Yao X.; Liu Y.; Lin G. Evolutionary programming made faster. *IEEE Transactions on Evolutionary Computation*. 1996, 3, 82–102.
31. Shi Y.; Eberhart R. A modified particle swarm optimizer. In Proceedings of Advances in Natural Computation.
32. Park J.B.; Jeong Y.W.; Shin J.R. An Improved Particle Swarm Optimization for Nonconvex Economic Dispatch Problems. *IEEE Transactions on Power Systems*. 2010, 25, 156–166.
33. Qin J.; Liang Z. A naive Particle Swarm Optimization. In Proceedings of Evolutionary Computation IEEE.
34. Gao H.; Xu W. A New Particle Swarm Algorithm and Its Globally Convergent Modifications. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics *Society*. 2011, 41, 1334.
35. Bergh F.V.D.; Engelbrecht A.P. Effect of swarm size on cooperative particle swarm optimizers. In Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation Conference (GECCO), 2001, 892–899.
36. Sun J.; Wu X.; Palade V.; Fang W,; Lai C.H.; Xu W.B. Convergence analysis and improvements of quantum-behaved particle swarm optimization. Information Sciences, 2012, 193, 81–103.

The structure and existence of solutions of the problem of consumption with satiation in continuous time

Peter Smoczynski[1☯], **Stan Miles**[2☯*]

1 Department of Mathematics and Statistics, Faculty of Science, Thompson Rivers University, Kamloops, BC, Canada, **2** Department of Economics, School of Business and Economics, Thompson Rivers University, Kamloops, BC, Canada

☯ These authors contributed equally to this work.
* stanmiles@tru.ca

Abstract

With the help of the method of Lagrange multipliers and KKT theory, we investigate the structure and existence of optimal solutions of the continuous-time model of consumption with satiation. We show that the differential equations have no solutions in the C^1 class but that solutions exist in a wider space of functions, namely, the space of functions of bounded variation with non-negative Borel measures as controls. We prove our theorems with no additional assumptions about the structure of the control Borel measures. We prove the conjecture made in the earlier literature, that there are only three types of solutions: I-shaped solutions, with a gulp of consumption at the end of the interval and no consumption at the beginning or in the interior; U-shaped solutions, with consumption in the entire interior of the interval and gulps at the beginning and the end; and intermediate (J-shaped) solutions, with an initial interval of abstinence followed by a terminal interval of distributed consumption at rates and a gulp at the end. We also establish the criteria that permit determination of the solution type using the problem's parameters. When the solution structure is known, we reduce the problem of the existence of a solution to algebraic equations and discuss the solvability of these equations. We construct explicit solutions for logarithmic utility and CRRA utility.

Editor: Alfonso Rosa Garcia, Universidad Catolica San Antonio de Murcia, SPAIN

1 Introduction

Recently, Baucells and Sarin [1, 2] described a new and interesting discrete-time model of consumer behavior: the satiation model. The psychological justifications of the model are explained in plain language in [3]. In [4, 5] this model was extended to continuous time, and the solutions were constructed for CRRA utilities.

The satiation model presented in [4, 5] is mathematically described as an optimal control problem (Problem 1) in this paper. In this problem the control (consumption $c(t)$) enters the evolution equation (Eq (2)) in this paper linearly and may be unbounded, hence the classical Pontryagin Maximum Principle (see [6]) is not directly applicable. For problems of this type,

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

often called systems with impulsive controls, more general maximum principles are discussed in [7] and [8], and extensively in [9] and in papers quoted therein. All of these works use a solution-dependent change of the time variable which was introduced in [10]. In [11] the Bellman principle is extended to impulsive systems. In all of these works it is explicitly assumed that the measure controls have atoms. All of these general theories are very complex, hence we convert Problem 1 into a variational problem, and we use the simpler and more straightforward method of Lagrange multipliers. The approach presented in this paper uses very little beyond the finite-dimensional case described in [12]. Our method is simpler than any proof of a maximum principle, especially when unbounded controls are involved. The additional novelties of our approach are as follows: Instead of assuming a priori that the control measures have atoms, we prove that they do. In addition, we use no transformation of the time variable, and we do not require any additional differential equations (such as Eq (4.6) and later equations in [8]) to be satisfied by the atoms of the control measures.

In this work, Problem 1 is transformed into a problem that is easier to handle: Problem 3, with a different utility (see Eq (9)). We first prove that this problem has no solutions with continuously differentiable satiation s. Then we extend the solution space to the BV space (the space of functions of bounded variation). This implies that we must allow the consumption c, the controlling variable, to be a general Borel measure; it could even be as exotic as the derivative of the Cantor function. An important lemma (Lemma 23) is proved using only the assumption that c is a non-negative Borel measure.

We conduct a detailed investigation of the structure of optimal solutions for general utilities. Our general results are proved under the following assumptions: (i) the original utility V and the transformed utility are both concave down, or $-V$ and $-V_S$ are both convex in the classic sense; and (ii) additional barrier conditions (Eqs (36) and (37)) hold. Further, we assume two pairs of inequalities, where one pair (Eqs (32) and (40)) corresponds to a high future discount n, and the other pair (Eqs (33) and (41)) corresponds to a low future discount n. Each pair is sufficient for the existence of solutions; however, the corresponding solutions differ remarkably. In the case of a high future discount, all the consumption must take place in a single gulp, or burst, at time $t = 0$, while in the case of a low future discount a terminal gulp of consumption at $t = T$ is necessary. In [4] the proofs of many theorems use an additional assumption: that there are no gulps of consumption in $(0, T)$. We prove (in Theorems 21 and 25) that this assumption is correct, namely, that the measure c is continuous relative to Lebesgue measure in $(0, T)$, which means there are no gulps of consumption there, and that gulps of consumption can occur only at the endpoints of the interval $[0, T]$. We prove that at most one interval with non-zero consumption is possible and that at most one interval with no consumption is possible. This is done in Theorem 21 for the case of a large future discount and in Theorem 25 for a small future discount. We extend an earlier observation in [4], which was stated only for special cases of utilities and selected parameters, to general utilities and prove that under one of the aforementioned pairs of inequalities, which essentially corresponds to a large future discount, the only type of solution is that in which all the consumption takes place in a single gulp at $t = 0$ (as discussed in Section 6.1), while under the other pair of inequalities, which essentially corresponds to a small future discount, there are three types of solutions: J-shaped ones that correspond to poor or over-satiated agents; U-shaped ones that correspond to rich or under-satiated agents, with two gulps of consumption, one at the beginning and the other at the end; and intermediate solutions with an initial period of abstinence, a terminal gulp of consumption, and a terminal interval of non-gulpy consumption.

The general theory presented in Section 5 is used in Section 8 to reduce the problem of the existence of solutions to simple non-differential equations, and the solvability of these equations is proved. In the case of CRRA utility (Section 8.2), only one of these equations requires

numerical methods, while in the case of logarithmic utility all of them are solved explicitly (Section 8.1).

2 The problem and preliminaries

In this paper, we address the following problem, which is a continuous-time version of the satiation problem in [1]:

Problem 1 *Let V be a concave-down, twice differentiable function of s. Maximize the functional (the sum of instantaneous utilities* $\frac{dV}{ds} \cdot c \cdot dt$*)*

$$U_S = \int_{[0,T]} e^{-n \cdot t} \cdot \frac{dV}{ds}(s(t)) \cdot c(t) \cdot dt \tag{1}$$

under the following constraints:

$$\frac{ds}{dt} = \varphi \cdot c - \gamma \cdot s \tag{2}$$

$$s(0) = s_0 \tag{3}$$

$$\{subscript_0 \ addedd \ here\} s_0 \geq 0 \tag{4}$$

$$c \geq 0 \tag{5}$$

$$\int_{[0,T]} e^{-r \cdot t} \cdot c(t) \cdot dt = W \tag{6}$$

In this model, $s(t)$ is the satiation level caused by consumption $c(t)$, and W is the initial wealth; n, r, γ, and φ are positive parameters representing respectively the agent's discount rate, the risk-less interest rate, the satiation decay rate, and the satiation generation rate caused by consumption c.

Remark 2 *In Problem 1 we deliberately do not specify the space in which we search for solutions. It is more convenient to do this in the transformed Problem 3. However, in Section 3 we argue that the classical space used in Problem 3 does not contain any solutions of the problem. The proper space of BV functions is described in Section 4 and is used in the final version of the problem (Problem 10).*

It is helpful to eliminate $c(t)$ from (1). By constraint (2),

$$c = \frac{1}{f} \cdot \left(\frac{ds}{dt} + \gamma \cdot s \right)$$

Substituting this into (1) and then using integration by parts, we obtain an equation for the functional $f \cdot U_S$ in which the consumption function c does not explicitly appear:

$$f \cdot U_S(s(t)) = \int_{[0,T]} e^{-n \cdot t} \cdot \frac{dV}{ds}(s(t)) \cdot \left(\frac{ds}{dt} + \gamma \cdot s \right) \cdot dt \tag{7}$$

$$= e^{-n \cdot T} \cdot V(s(T)) - V(s_0) + \int_{[0,T]} e^{-n \cdot t} \cdot V_s(s(t)) \cdot dt, \tag{8}$$

where

$$V_S(s) = n \cdot V(s) + \gamma \cdot s \cdot \frac{dV}{ds}(s). \tag{9}$$

The following will be used later, in the proofs of Propositions 8 and 14 and Theorem 21:

$$\frac{dV_S}{ds}(s) = (n + \gamma \cdot (1 - \alpha)) \cdot \frac{dV}{ds}, \tag{10}$$

where

$$\alpha = \alpha(s) = -\frac{s \cdot \frac{d^2V}{ds^2}}{\frac{dV}{ds}} \tag{11}$$

The expression for $\alpha(s)$ looks very similar to the expression for relative risk aversion. The term *relative risk aversion* is appropriate in the context of investments, where every investment carries some risk. However, in this paper the risk of consumption is ignored, hence the term *relative satiation aversion* seems more appropriate. Thus we will call $\alpha(s)$ the relative satiation aversion. The relative satiation aversion for V_S is denoted by α_S:

$$\alpha_S(s) = -\frac{s \cdot \frac{d}{ds}\left((n + \gamma \cdot (1 - \alpha)) \cdot \frac{dV}{ds}\right)}{(n + \gamma \cdot (1 - \alpha)) \cdot \frac{dV}{ds}} = \alpha - \frac{\gamma}{n + \gamma \cdot (1 - \alpha)} \cdot s \cdot \frac{d\alpha}{ds} \tag{12}$$

Similarly to the method used in obtaining (7), we can transform the wealth constraint (6) with the help of integration by parts:

$$f \cdot W = e^{-r \cdot T} \cdot s(T) - s_0 + (\gamma + r) \cdot \int_{[0,T]} e^{-r \cdot t} \cdot s(t) \cdot dt \tag{13}$$

We assume that V_S given by (9) is an increasing, concave-down function of the satiation s. In Section 8.2 it is demonstrated that this is always the case for CRRA utility. Therefore, Problem 1 can be replaced by the following problem:

Problem 3 (The smooth version) *Maximize the functional $f \cdot U_S$ in (8) on the space of all continuously differentiable functions $s \in C^1$ ($[0, T] \to R$) that satisfy the constraints (4), (5), and (3), and constraint (13) with $W > 0$, under the assumption that V and V_S are increasing, concave-down, twice continuously differentiable functions of s.*

Remark 4 *If V is an increasing function, then it follows from (10) that V_S is increasing if*

$$n + \gamma \cdot (1 - \alpha) > 0 \tag{14}$$

Downward concavity of V_S requires that $\alpha_S > 0$, and downward concavity of V requires that $\alpha > 0$. Therefore, these two conditions are assumed to be satisfied in everything that follows.

One of the objectives of this paper is to discuss the solvability conditions for Problem 3. Necessary and sufficient conditions are obtained by generalization of the Lagrange multipliers (Karush–Kuhn–Tucker (KKT) theory for a finite number of dimensions; see [12]). We first note an important consequence of (5):

Proposition 5 *If (4) holds and (5) holds for all $t \in [0, T]$, then $s(t) > 0$ for all $t \in [0, T]$.*

Proof. The solution of Eq (2) is

$$s(t) = e^{-\gamma \cdot t} \cdot \left(s_0 + f \cdot \int_{[0,T]} e^{\gamma \cdot t_1} \cdot c(t_1) \cdot dt_1\right), \tag{15}$$

hence $s(t) > 0$ for $t \geq 0$, thanks to (4) and (5).

Remark 6 *Later it is shown that* $s(t) > 0$ *in* $(0, T]$ *when* $W > 0$ *even if* $s_0 = 0$; *see Remark 36.*

Following the finite-dimensional KKT theory (see [12, 13]), we introduce the Lagrangian functional

$$\begin{aligned} f \cdot \Lambda(s(t)) &= f \cdot U_S(s(t)) + \lambda_W \cdot \left(W - \int_{[0,T]} e^{-r \cdot t} \cdot \left(\frac{ds}{dt} + \gamma \cdot s \right) \cdot dt \right) \\ &\quad + \int_{[0,T]} \lambda_c \cdot \left[\frac{ds}{dt} + \gamma \cdot s \right] \cdot dt \end{aligned} \tag{16}$$

Here $\lambda_W \in R$ is the Lagrange multiplier that corresponds to constraint (13), and $\lambda_c \in C^1$ ($[0, T] \to R$) corresponds to (5). The KKT method also involves additional constraints on $\lambda_c(t)$:

$$0 = \lambda_c \cdot c = \lambda_c \cdot \frac{1}{f} \cdot \left(\frac{ds}{dt} + \gamma \cdot s \right), \ \lambda_c \geq 0. \tag{17}$$

Since $s(t)$ and $\lambda_c(t)$ are assumed to be differentiable on $[0, T]$, they are continuous on that interval. Thus after integration by parts, the functional (16) takes the following form:

$$\begin{aligned} f \cdot \Lambda(s(t)) &= e^{-n \cdot T} \cdot V(s(T)) - V(s_0) - \lambda_W \cdot (e^{-r \cdot T} \cdot s(T) - s_0) \\ &\quad + (\lambda_c(T) \cdot s(T) - \lambda_c(0) \cdot s_0) \\ &\quad + \int_{[0,T]} \left[e^{-n \cdot t} \cdot V_S(s(t)) - \lambda_W \cdot e^{-r \cdot t} \cdot (\gamma + r) \cdot s \right. \\ &\quad \left. + \left(-\frac{d\lambda_c}{dt} + \lambda_c \cdot \gamma \right) \cdot s \right] \cdot dt + \lambda_W \cdot W \end{aligned} \tag{18}$$

The first-order conditions for an extremum are obtained as follows: If $s(t)$ is an optimal solution and $\Delta s(t) \in C^1([0, T] \to R)$ with $\Delta s(0) = 0$ (hence for every real number ε, $s + \varepsilon \cdot \Delta s$ satisfies the initial condition (3): $s(0) + \varepsilon \cdot \Delta s(0) = s_0$), then

$$\begin{aligned} \int_{[0,T]} \frac{\partial \Lambda}{\partial s()} \cdot \Delta s(t) \cdot dt &= \left[e^{-n \cdot T} \cdot \frac{dV}{ds}(s(T)) - \lambda_W \cdot e^{-r \cdot T} + \lambda_c(T) \right] \cdot \Delta s(T) \\ &\quad + \int_{[0,T]} \left[e^{-n \cdot t} \cdot \frac{dV_S}{ds}(s(t)) - \lambda_W \cdot e^{-r \cdot t} \cdot (\gamma + r) \right. \\ &\quad \left. + \left(-\frac{d\lambda_c}{dt} + \lambda_c \cdot \gamma \right) \right] \cdot \Delta s(t) \cdot dt, \end{aligned}$$

which must be 0, hence also $\frac{\partial \Lambda}{\partial s()} = 0$ by the DuBois–Reymond lemma, as $\Delta s(t)$ is arbitrary (except for $\Delta s(0) = 0$, which follows from (3)). This leads to the following set of first-order conditions:

At the boundary (i.e., at $t = T$),

$$e^{-n \cdot T} \cdot \frac{dV}{ds}(s(T)) - \lambda_W \cdot e^{-r \cdot T} + \lambda_c(T) = f \cdot \frac{\partial \Lambda}{\partial s(T)} = 0. \tag{19}$$

In the interior of the interval (i.e., for $t \in (0, T)$),

$$e^{-n \cdot t} \cdot \frac{dV_S}{ds} - \lambda_W \cdot e^{-r \cdot t} \cdot (\gamma + r) + \left(-\frac{d\lambda_c}{dt} + \lambda_c \cdot \gamma \right) = f \cdot \frac{\partial \Lambda}{\partial s()} = 0. \tag{20}$$

KKT theory guarantees that if Problem 3 has a solution s, then that solution is the first of the three entities in $[s(t), \lambda_c(t), \lambda_W]$ that make up the solution of the following problem:

Problem 7 (The C^1 version) *Find* $[s(t), \lambda_c(t), \lambda_W] \in C^1([0, T] \to R) \times C^1([0, T] \to R) \times R$ *that satisfies* (2), (20), *initial condition* (3), *terminal condition* (19), *constraints* (4) *and* (5), *constraint* (13) *with* $W > 0$, *and KKT conditions* (17).

We show in Section 3 that Problem 7 has no (everywhere differentiable) solution. The remedy, discontinuous functions with bounded variation, is described in Section 4. Problem 7 is reformulated as Problem 10 in the BV space. The sufficient conditions are derived in Theorem 11, and in Theorem 13 it is shown that the sufficent conditions are also necessary. The structure of these discontinuous solutions is investigated in greater detail in Section 6. Explicit solutions for logarithmic and CRRA utilities are constructed in Sections 8.1 and 8.2, respectively.

3 Nonexistence of continuously differentiable solutions

By assumption, V and V_S are concave down and all the constraints are linear, hence there can be no more than one solution. We will now show that the existence of a solution of the most tractable form (continuous on $[0, T]$ and continuously differentiable on $(0, T)$) will lead to the contradictions described in the proof of the proposition below.

Proposition 8 *If* $(r - n + \gamma \cdot \alpha(s)) \neq 0$ *and* $\frac{dV}{ds}(s) \neq 0$ *for all* $s \geq 0$, *then there is no continuously differentiable solution* $[s, \lambda_c, \lambda_W]$ *of the boundary value problem (Problem 7).*

Proof. Assume that there is a continuously differentiable solution. First, we assume that consumption $c(T) > 0$. Since $s \in C^1([0, T] \to R)$, (2) implies that c is also continuous and therefore $c(t) > 0$ for all t in an open neighborhood of T. In this case, as a consequence of the first KKT condition in (17), we have $\lambda_c(t) = 0$ for all t close to T, hence the terminal condition (19) and condition (20) with $t = T$ take the following form:

$$0 = e^{-n \cdot T} \cdot \frac{dV}{ds}(s(T)) - \lambda_W \cdot e^{-r \cdot T} \tag{21}$$

$$0 = e^{-n \cdot T} \cdot \frac{dV_S}{ds}(s(T)) - \lambda_W \cdot e^{-r \cdot T} \cdot (\gamma + r). \tag{22}$$

One can eliminate λ_W by subtraction, obtaining, with the help of (10),

$$0 = e^{-n \cdot T} \cdot \left[\frac{dV_S}{ds}(s(T)) - \frac{dV}{ds}(s(T)) \cdot (\gamma + r)\right] = (n - r - \alpha \cdot \gamma) \cdot \frac{dV}{ds}(s(T)).$$

However, $(n - r - \alpha \cdot \gamma) \cdot \frac{dV}{ds}(s(T))$ cannot be 0, by the premises of this proposition. Therefore, the assumption that $s(t)$ is continuously differentiable, and that $c(T) > 0$, is false. The problem seems to be caused by the fact that Eq (20) has to be satisfied with $t \to T$. We could try to get around this problem by assuming that there is no consumption at $t = T$, hence that $\lambda_c(T) \neq 0$, in which case we have one additional variable, $\lambda_c(T)$, that can help us to satisfy Eq (20) at $t = T$. If $c(T) = 0$ and $c(t) > 0$ for all t close to T, then $\lambda_c(t) = 0$ for all these t, hence $\lambda_c(T) = 0$ since λ_c is continuous in $[0, T]$, so we do not have the additional variable $\lambda_c(T)$. In order to obtain it, we need to assume that there is some $T_e < T$ such that $c(t) = 0$ for all $t > T_e$ and $c(t) > 0$ for $t \leq T_e$. But in such a case the solution must be optimal in the interval $[0, T_e]$ and at $t = T_e$ the same boundary condition must be satisfied, hence we obtain the same contradiction. Therefore, the assumption that $s(t)$ is continuously differentiable is false.

Proposition 8 implies that Problem 1 does not have a continuously differentiable solution. Note, however, that all of the results proved in this section for continuous, almost everywhere differentiable functions (except for the nonexistence of a continuously differentiable solution) remain valid for discontinuous solutions with bounded variation, which are

considered in the rest of the paper. This can be deemed as an extreme case of the Lavrentiev phenomenon (see [14]): In one Banach space there is a solution, and in another there are none. In [15] there appears the following remark: "Lavrentiev's phenomenon is related to the existence of singular minimizers, i.e., absolutely continuous minimizers that are not Lipschitz." In the case discussed in this work, the maximizers are of bounded variation but discontinuous, hence we can expect a Lavrentiev phenomenon. However, the optimal value can be approximated by a suitably chosen Lipschitz continuous function when $\alpha > 0$, and probably not when $\alpha < 0$. The functional (7) does not satisfy the coercivity condition (A4) from [14], especially not on any space that imposes restrictions on derivatives (of satiation), for the functional (8) does not even contain the derivative $\frac{ds}{dt}$ explicitly. This suggests that the result on the absence of a Lavrentiev gap in [14] can be proved without the coercivity condition.

4 Functions of bounded variation

If a model has no solutions, as was shown for the satiation model in Section 3, it has to be modified—we can "regularize" it by introducing an artificial viscosity or limiting the analysis to bounded consumption, as in [5],—or it could be abandoned in favor of another model. Alternatively, the definition of a solution could be changed. Intuition would suggest that the satiation model does have a solution and, as a result, an artificial regularization is not necessary. One alternative would be to relax the requirement that the satiation level $s(t)$ be continuous, and instead allow s to have some points of discontinuity. If such discontinuities are permitted, the satiation $s(t)$ is not differentiable in the classical sense at the points of discontinuity. However, $s(t)$ is present in differential Eq (2), hence we need to use non-classical derivatives. In order to accommodate this relaxation, we permit s to be a function of bounded variation (BV); see [16] or [17]. In general, functions belonging to the BV space could be discontinuous, and their derivatives could be Borel measures. The Heaviside function, which is defined by $h(x) = 0$ for $x \leq 0$ and $h(x) = 1$ for $x > 0$, is an example of a BV function. Its derivative is the well-known Borel measure popularly known as the Dirac delta function. For this reason, when satiation $s(t)$ is a BV function, Eq (2) forces consumption c to be a Borel measure. Whenever satiation $s(t)$ has a discontinuity, the singular part of the Borel measure c is the Dirac delta measure. In order to avoid this non-descriptive term, we use a more intuitive term for it: consumption gulp, following [4]. The mathematical description of a "gulp" or a gulp at a point of discontinuity of a function is given in Eq (25).

Functions of a single variable with bounded variation (BV functions) are continuous almost everywhere, and their derivatives are finite Borel measures. The three most important properties of BV functions are as follows:

1. The one-sided limits $\varphi(\tau^-) = \lim_{t\to\tau, t<\tau} \varphi(t)$ and $\varphi(\tau^+) = \lim_{t\to\tau, t>\tau} \varphi(t)$ exist whenever τ is in the closure of the interior of the domain of φ.

2. The derivatives exist even if the functions are discontinuous. The downside is that the derivatives are permitted to be Borel measures.

3. Integration by parts is permitted. If φ is a BV function on $[a, b]$ and g is C^1 on R, then $\frac{d\varphi}{dt}$ is a Borel measure and, if $\varphi(a)$ and $\varphi(b)$ are defined, we can integrate by parts as follows:

$$\int_{[a,b]} \varphi(t) \cdot \frac{dg}{dt}(t) \cdot dt = \varphi(b) \cdot g(b) - \varphi(a) \cdot g(a) - \int_{[a,b]} g(t) \cdot \frac{d\varphi}{dt}(t) \cdot dt \tag{23}$$

If φ is C^1 on R except for some point $\tau \in (a, b)$ and the limits $\varphi(\tau^+)$ and $\varphi(\tau^-)$ exist, then with the help of simple calculus we can proceed as follows:

$$\begin{aligned}\int_{[a,b]} \varphi(t)\cdot\frac{dg}{dt}(t)\cdot dt &= \int_{(a,\tau)} \varphi(t)\cdot\frac{dg}{dt}(t)\cdot dt + \int_{(\tau,b)} \varphi(t)\cdot\frac{dg}{dt}(t)\cdot dt \\ &= \varphi(b)\cdot g(b) - \varphi(a)\cdot g(a) - (\varphi(\tau^+) - \varphi(\tau^-))\cdot g(\tau) \\ &\quad - \int_{(a,\tau)} g(t)\cdot\frac{d\varphi}{dt}(t)\cdot dt - \int_{(\tau,b)} g(t)\cdot\frac{d\varphi}{dt}(t)\cdot dt\end{aligned} \tag{24}$$

In Eq (24), $\frac{d\varphi}{dt}$ denotes the classical derivative, which is defined everywhere except at τ. By comparing (23) to (24), we conclude that the Borel measure $\frac{d\varphi}{dt}$ is given by

$$\left[\frac{d\varphi}{dt}\right]_{\text{generalized}} = \left[\frac{d\varphi}{dt}\right]_{\text{classical}} + (\varphi(\tau^+) - \varphi(\tau^-))\cdot\delta_\tau(t), \tag{25}$$

where $\left[\frac{d\varphi}{dt}(t)\right]_{\text{classical}}$ is the classical derivative and $(\varphi(\tau^+) - \varphi(\tau^-))\cdot\delta_\tau(t)$ is the Dirac delta Borel measure with strength equal to the jump $(\varphi(\tau^+) - \varphi(\tau^-))$ in the value of φ at the discontinuity located at $t = \tau$. This can be interpreted as a decomposition of the Borel measure $\frac{d\varphi}{dt}$ into the sum of the part that is continuous with respect to Lebesgue measure, $\left[\frac{d\varphi}{dt}(t)\right]_{\text{classical}}$, and the part that is singular with respect to Lebesgue measure, $\left[\frac{d\varphi}{dt}\right]_{\text{singular}} = (\varphi(\tau^+) - \varphi(\tau^-))\cdot\delta_\tau(t)$. As mentioned earlier, in this paper the latter is referred to as a "gulp." The theory of BV functions provides us with derivatives of discontinuous functions and formulas for integration by parts which are applicable to discontinuous functions, even if the singular parts of their derivatives are much more complicated than in the example above (e.g., the singular part of the derivative of the Cantor function).

The use of integration by parts to derive formulas (18) and (13), as well as formula (69) in Section 7.3, is justified by the formula in (23) when $s(t)$ is a BV function. Hence all of the identities involving integrals derived in the preceding sections are also valid when C^1 differentiability is relaxed to that in the BV sense. In this paper the discontinuous function is satiation $s(t)$. If $s(t)$ is discontinuous at $t = T$, then Eq (2) implies that this contributes $\frac{s(T)-s(T^-)}{f}\cdot\delta_T(t)$ to consumption. If we know that the discontinuities in satiation can occur only at the endpoints of the interval $[0, T]$, we can avoid BV theory altogether, as was done in [4]. However, in order to prove this, one needs to initially permit the derivative of the satiation to be a general Borel measure on $[0, T]$ and then to show that it is continuous with respect to Lebesgue measure in $(0, T)$. Also, a discontinuity of $s(t)$ at $t = T$ allows us to remove the contradiction obtained in the proof of Proposition 8. This is the approach used in the proof of Proposition 14 in the next section.

Remark 9 Eq (23) *for open sets takes the form*

$$\int_{(a,b)} \varphi(t)\cdot\frac{dg}{dt}(t)\cdot dt = \varphi(b^-)\cdot g(b) - \varphi(a^+)\cdot g(a) - \int_{(a,b)} g(t)\cdot\frac{d\varphi}{dt}(t)\cdot dt. \tag{26}$$

Eqs (26) *and* (23) *imply that*

$$\begin{aligned}\int_{[a,b]} g(t)\cdot\frac{d\varphi}{dt}(t)\cdot dt &= \int_{(a,b)} g(t)\cdot\frac{d\varphi}{dt}(t)\cdot dt \\ &\quad + g(b)\cdot(\varphi(b) - \varphi(b^-)) + g(a)\cdot(\varphi(a^-) - \varphi(a)).\end{aligned}$$

5 Sufficiency and necessity

In Section 3 it was shown that the space $C^1([0, T] \to R) \times C^1([0, T] \to R) \times R$ is inadequate for investigation of solutions of Problem 7, hence we modify that problem as follows:

Problem 10 (BV version) *Find* $[s(t), \lambda_c(t), \lambda_W] \in BV([0, T] \to R) \times BV([0, T] \to R) \times R$ *that satisfies* (2), (20), *initial condition* (3), *terminal condition* (19), *constraints* (4) and (5), *constraint* (13) *with* $W > 0$, *and KKT conditions* (17).

In spite of discontinuities and exotic derivatives, the use of integration by parts is valid for the functions in $BV([0, T] \to R)$, hence all the equations and conclusions obtained in Sections 2 and 3, including Eq (18) and the conditions (19) and (20), remain valid under the premises of Problem 10. Proposition 5 is also valid, and the proof is almost the same: When the consumption c is a non-negative measure, it replaces $c(t_1) \cdot dt_1$, and the second term inside the parentheses in Eq (15) is still positive for all $t \in (0, T]$.

In Section 2 it was proved that if $[s, \lambda_c, \lambda_W] \in C^1([0, T] \to R) \times C^1([0, T] \to R) \times R$ is a stationary point of the Lagrangian functional in (16), then $[s, \lambda_c, \lambda_W]$ must be a solution of Problem 7. With the help of the BV theory, one can easily prove sufficiency of conditions (19) and (20) in Problem 10.

Theorem 11 (on sufficiency) *Suppose* V *and* V_S *are concave-down, continuously differentiable functions, and let* $[s, \lambda_c, \lambda_W] \in BV([0, T] \to R) \times BV([0, T] \to R) \times R$ *be a solution of Problem 10. Then the functional* (8) *attains a maximum value on the set*

$$\{s_1 \in BV([0, T] \to R) \mid s_1 \text{ satisfies all the constraints of Problem 10}\} \tag{27}$$

at s.

Proof. Let $\Delta s \in BV([0, T] \to R)$ be any function such that $s + \Delta s$ satisfies (2), initial condition (3), constraint (4), constraint (13) with $W > 0$, and (5). Then for all $x \in [0, 1]$ the function $s_x = (1 - x) \cdot s + x \cdot (s + \Delta s) = s + x \cdot \Delta s$ also satisfies these conditions. Consider the function

$$\Theta(x) = \varphi \cdot U_S(s_x(t)) = e^{-n \cdot T} \cdot V(s_x(T)) - V(s_0) + \int_{[0,T]} e^{-n \cdot t} \cdot V_S(s_x(t)) \cdot dt,$$

where U_S is given by (8). Since s and Δs are bounded, the bounded convergence theorem ensures that Θ is differentiable with respect to x and

$$\begin{aligned}
\left[\frac{d\Theta}{dx}\right]_{x=0} &= \varphi \cdot \left[\frac{dU_S(s_x(t))}{dx}\right]_{x=0} \\
&= \left[\frac{d}{dx}(e^{-n \cdot T} \cdot V((s + x \cdot \Delta s)(T)) - V(s_0))\right]_{x=0} \\
&\quad + \left[\frac{d}{dx}\left(\int_{[0,T]} e^{-n \cdot t} \cdot V_S((s + x \cdot \Delta s)(t)) \cdot dt\right)\right]_{x=0} \\
&= e^{-n \cdot T} \cdot \frac{dV}{ds}(s(T)) \cdot \Delta s(T) \\
&\quad + \int_{[0,T]} e^{-n \cdot t} \cdot \frac{dV_S}{ds}(s(t)) \cdot \Delta s(t) \cdot dt
\end{aligned}$$

Since $[s, \lambda_c, \lambda_W]$ is a solution of Problem 10, we may use (20) and (19), and we obtain

$$\begin{aligned}\left[\frac{d\Theta}{dx}\right]_{x=0} &= [\lambda_W \cdot e^{-r\cdot T} - \lambda_c(T)] \cdot \Delta s(T) \\ &\quad + \int_{[0,T]} \left[\lambda_W \cdot e^{-r\cdot t} \cdot (\gamma + r) + \frac{d\lambda_c}{dt} - \lambda_c \cdot \gamma\right] \cdot \Delta s(t) \cdot dt \\ &= \lambda_W \cdot e^{-r\cdot T} \cdot \Delta s(T) + \int_{[0,T]} \lambda_W \cdot e^{-r\cdot t} \cdot (\gamma + r) \cdot \Delta s(t) \cdot \\ &\quad -\lambda_c(T) \cdot \Delta s(T) + \int_{[0,T]} \left[\frac{d\lambda_c}{dt} - \lambda_c \cdot \gamma\right] \cdot \Delta s(t) \cdot dt\end{aligned}$$

Now constraint (13) implies that

$$\begin{aligned}&\lambda_W \cdot e^{-r\cdot T} \cdot \Delta s(T) + \int_{[0,T]} \lambda_W \cdot e^{-r\cdot t} \cdot (\gamma + r) \cdot \Delta s(t) \cdot dt \\ &= \frac{d}{dx}\left(\lambda_W \cdot e^{-r\cdot T} \cdot s_x(T) + \lambda_W \cdot \int_{[0,T]} e^{-r\cdot t} \cdot (\gamma + r) \cdot s_x(t) \cdot dt\right) \\ &= \frac{d}{dx}(\varphi \cdot W) = 0,\end{aligned}$$

so

$$\left[\frac{d\Theta}{dx}\right]_{x=0} = -\lambda_c(T) \cdot \Delta s(T) + \int_{[0,T]} \left[\frac{d\lambda_c}{dt} - \lambda_c \cdot \gamma\right] \cdot \Delta s(t) \cdot dt$$

Integration by parts produces

$$\int_{[0,T]} \frac{d\lambda_c}{dt} \cdot \Delta s(t) \cdot dt = \lambda_c(T) \cdot \Delta s(T) - \int_{[0,T]} \frac{d\Delta s}{dt} \cdot \lambda_c(t) \cdot dt$$

since $\Delta s(0) = 0$, hence

$$\begin{aligned}\left[\frac{d\Theta}{dx}\right]_{x=0} &= -\int_{[0,T]} \left[\frac{d\Delta s}{dt} + \Delta s \cdot \gamma\right] \cdot \lambda_c(t) \cdot dt \\ &= -\int_{\text{supp}(\lambda_c)} \left[\frac{d\Delta s}{dt} + \Delta s \cdot \gamma\right] \cdot \lambda_c(t) \cdot dt.\end{aligned}$$

Now note the following: $\lambda_c(t) \geq 0$ and $c(t) = 0$ when $t \in \text{supp}(\lambda_c)$, thanks to KKT conditions (17). Therefore, $\left[\frac{d\Delta s}{dt} + \Delta s \cdot \gamma\right](t) \geq 0$ when $t \in \text{supp}(\lambda_c)$, since $s + \Delta s$ satisfies (5). Hence $\left[\frac{d\Delta s}{dt} + \Delta s \cdot \gamma\right] \cdot \lambda_c(t) \geq 0$ on $\text{supp}(\lambda_c)$, and we conclude that $\frac{d\Theta}{dx}(0) \leq 0$. This implies that $\Theta(x)$ attains a maximum at $x = 0$, that is, at s, since the function Θ is concave down on [0, 1]. But the functional $BV([0, T] \to R) \ni s \to U_S(s)$ is continuous on $BV([0, T] \to R)$, thanks to continuity of V_S, and this ensures that s is also a maximum of the convex functional $U_S(s)$ on the set in (27). Thus s is unique, thanks to the downward concavity of U_S.

Remark 12 *A proof of necessity of the conditions in Problem 10 is not needed for explicit construction of solutions in the sections of the paper that follow. For the sake of completeness, however, we present a proof of the necessity.*

Theorem 13 (on necessity) *Let V be a twice continuously differentiable function, let c be a non-negative Borel measure on* [0, *T*] *that satisfies* (6), *and let s be the solution of the differential* Eq (2) *that satisfies* (3) *and* (4). *Then* $s \in BV([0, T] \to R)$. *Suppose* [*c*, *s*] *maximizes the functional* (8) *under constraints* (5), (6) *with* $W > 0$, (2), (3), *and* (4). *Then there exist a non-negative*

Lipschitz continuous function $\lambda_c \in Lip([0, T] \to R)$ *and a non-negative number* $\lambda_W \in R$ *such that* $[s, \lambda_c, \lambda_W] \in BV([0, T] \to R) \times Lip([0, T] \to R) \times R$ *is a solution of Problem (10).*

Proof. If c is a Borel measure, then the solution of the initial-value problem (Problem 2) with the initial condition (3) is given by (15), hence the solution is bounded, which means that $\frac{ds}{dt}$ is a Borel measure thanks to (2), and so $s \in BV([0, T] \to R)$. Suppose that under the constraints listed in the Theorem, the functional (8) attains its maximum value on $BV([0, T] \to R)$ at s. We need to construct $\lambda_c \in Lip([0, T] \to R)$ and a $\lambda_W \in R$, and show that $\lambda_c \geq 0$ and $\lambda_W > 0$, and that $[s, \lambda_c, \lambda_W]$ satisfy the differential Eq (20) and the terminal condition (19). Let c_1 be any non-negative Borel measure on $[0, T]$ that satisfies (6), and let s_1 be the corresponding solution of the initial-value problem that consists of Eq (2) together with (3) and (4). In addition, for all $x \in [0, 1]$, let $c_x = (1 - x) \cdot c + x \cdot c_1$ and $s_x = (1 - x) \cdot s + x \cdot s_1$. Then c_x is non-negative and also satisfies (6), and $s_x \in BV([0, T] \to R)$. Consider the function of $x \in [0, 1]$ given by

$$\begin{aligned} U_S(s_x, c_x) &= \int_{[0,T]} e^{-n \cdot t} \cdot \frac{dV}{ds}(s_x(t)) \cdot c_x(t) \cdot dt \\ &= (1 - x) \cdot \int_{[0,T]} e^{-n \cdot t} \cdot \frac{dV}{ds}(s_x(t)) \cdot c(t) \cdot dt \\ &\quad + x \cdot \int_{[0,T]} e^{-n \cdot t} \cdot \frac{dV}{ds}(s_x(t)) \cdot c_1(t) \cdot dt, \end{aligned} \tag{28}$$

where the integrals above are Lebesgue-type integrals with respect to Borel measures. Since s and s_1 are both bounded, with the help of the bounded convergence theorem one can prove that $U_S(s_x, c_x)$ is differentiable with respect to x and that

$$\begin{aligned} \frac{d}{dx} U_S(s_x, c_x) &= \frac{d}{dx} \int_{[0,T]} e^{-n \cdot t} \cdot \frac{dV}{ds}(s_x(t)) \cdot c_x(t) \cdot dt \\ &= \int_{[0,T]} e^{-n \cdot t} \cdot \frac{dV}{ds}(s(t)) \cdot c_1(t) \cdot dt \\ &\quad - \int_{[0,T]} e^{-n \cdot t} \cdot \frac{dV}{ds}(s(t)) \cdot c(t) \cdot dt \\ &\quad + (1 - x) \cdot \int_{[0,T]} e^{-n \cdot t} \cdot \frac{d^2V}{ds^2}(s_x(t)) \cdot (s_1 - s) \cdot c(t) \cdot dt \\ &\quad + x \cdot \int_{[0,T]} e^{-n \cdot t} \cdot \frac{d^2V}{ds^2}(s_x(t)) \cdot (s_1 - s) \cdot c_1(t) \cdot dt \end{aligned}$$

Since $U_S(s_x, c_x)$ attains its maximum at $x = 0$, we have

$$\begin{aligned} 0 &\geq \left(\frac{d}{dx} U_S(s_x, c_x) \right)_{x=0} \\ &= \int_{[0,T]} e^{-n \cdot t} \cdot \frac{dV}{ds}(s_x(t)) \cdot (c_1(t) - c(t)) \cdot dt \\ &\quad + \int_{[0,T]} e^{-n \cdot t} \cdot \frac{d^2V}{ds^2}(s(t)) \cdot (s_1 - s)(t) \cdot c(t) \cdot dt \end{aligned}$$

The second integral may be transformed as follows:

$$\begin{aligned}
& \int_{[0,T]} e^{-n\cdot t}\cdot \frac{d^2V}{ds^2}(s(t))\cdot (s_1-s)(t)\cdot c(t)\cdot dt \\
& \text{(with the help of (2))} \\
= & \int_{[0,T]} e^{-n\cdot t}\cdot \frac{d^2V}{ds^2}(s(t))\cdot (s_1-s)(t)\cdot \frac{1}{\varphi}\cdot \left(\frac{ds}{dt}+\gamma\cdot s\right)\cdot dt \\
= & \frac{1}{\varphi}\cdot \int_{[0,T]} e^{-n\cdot t}\cdot (s_1-s)(t)\cdot \left(\frac{d}{dt}\left(\frac{dV}{ds}(s(t))\right)+\gamma\cdot s\cdot \frac{d^2V}{ds^2}(s(t))\right)\cdot dt
\end{aligned}$$

$$\begin{aligned}
& \text{(integration by parts)} \\
= & \frac{1}{\varphi}\cdot e^{-n\cdot T}\cdot V(s(T))\cdot (s_1-s)(T) \\
& +\frac{1}{\varphi}\cdot \int_{[0,T]} e^{-n\cdot t}\cdot \left(n\cdot \frac{dV}{ds}(s(t))+\gamma\cdot s\cdot \frac{d^2V}{ds^2}(s(t))\right)\cdot (s_1-s)(t)\cdot dt \\
& -\frac{1}{\varphi}\cdot \int_{[0,T]} e^{-n\cdot t}\cdot \frac{dV}{ds}(s(t))\cdot \frac{d}{dt}(s_1-s)(t)\cdot dt \\
& \text{(since } s \text{ and } s_1 \text{ satisfy (2) with } c \text{ and } c_1\text{, respectively)}
\end{aligned}$$

$$\begin{aligned}
= & \frac{1}{\varphi}\cdot e^{-n\cdot T}\cdot \frac{dV}{ds}(s(T))\cdot (s_1-s)(T) \\
& +\frac{1}{\varphi}\cdot \int_{[0,T]} e^{-n\cdot t}\cdot \left(n\cdot \frac{dV}{ds}(s(t))+\gamma\cdot s\cdot \frac{d^2V}{ds^2}(s(t))\right. \\
& \qquad \left.+\gamma\cdot \frac{dV}{ds}(s(t))\right)\cdot (s_1-s)(t)\cdot dt \\
& -\int_{[0,T]} e^{-n\cdot t}\cdot \frac{dV}{ds}(s(t))\cdot (c_1-c)(t)\cdot dt,
\end{aligned}$$

hence thanks to (10),

$$\begin{aligned}
0 \quad \geq & \quad \left(\frac{d}{dx}U_S(s_x,c_x)\right)_{x=0} \\
= & \quad \frac{1}{\varphi}\cdot e^{-n\cdot T}\cdot \frac{dV}{ds}(s(T))\cdot (s_1-s)(T) \\
& +\frac{1}{\varphi}\cdot \int_{[0,T]} e^{-n\cdot t}\cdot \frac{dV_S}{ds}\cdot (s_1-s)(t)\cdot dt.
\end{aligned}$$

Since

$$(s_1-s)(t)=e^{-\gamma\cdot t}\cdot \varphi\cdot \int_{[0,T]} e^{\gamma\cdot t_1}\cdot (c_1(t_1)-c(t_1))\cdot dt_1,$$

we have

$$\begin{aligned} 0 &\geq \left(\frac{d}{dx}U_S(s_x, c_x)\right)_{x=0} \\ &= \frac{1}{\varphi}\cdot e^{-n\cdot T}\cdot\frac{dV}{ds}(s(T))\cdot e^{-\gamma\cdot T}\cdot\varphi\cdot\int_{[0,T]} e^{\gamma\cdot t}\cdot(c_1(t)-c(t))\cdot dt \\ &\quad +\frac{1}{\varphi}\cdot\int_{[0,T]} e^{-n\cdot t}\cdot\frac{dV_S}{ds}\cdot e^{-\gamma\cdot t}\cdot\varphi\cdot\int_{[0,T]} e^{\gamma\cdot t_1}\cdot(c_1(t_1)-c(t_1))\cdot dt_1\cdot dt \end{aligned}$$

Thanks to Fubini's theorem, $\int_{[0,T]}\int_{[0,t]}\ldots\cdot dt_1\cdot dt = \int_{[0,T]}\int_{[t,\,T]}\ldots\cdot dt\cdot dt_1$, hence

$$\begin{aligned} &\frac{1}{\varphi}\cdot\int_{[0,T]} e^{-n\cdot t}\cdot\frac{dV_S}{ds}\cdot e^{-\gamma\cdot t}\cdot\varphi\cdot\int_{[0,T]} e^{\gamma\cdot t_1}\cdot(c_1(t_1)-c(t_1))\cdot dt_1\cdot dt \\ = &\int_{[0,T]}\left[\int_{[t,T]} e^{-(n+\gamma)\cdot t_1}\cdot\frac{dV_S}{ds}(s(t_1))\cdot dt_1\cdot e^{\gamma\cdot t}\cdot(c_1(t)-c(t))\right]\cdot dt, \end{aligned}$$

so

$$\begin{aligned} 0 &\geq \left(\frac{d}{dx}U_S(s_x, c_x)\right)_{x=0} \\ &= \frac{1}{\varphi}\cdot e^{-n\cdot T}\cdot\frac{dV}{ds}(s(T))\cdot e^{-\gamma\cdot T}\cdot\varphi\cdot\int_{[0,T]} e^{\gamma\cdot t}\cdot(c_1(t)-c(t))\cdot dt \\ &\quad +\frac{1}{\varphi}\cdot\int_{[0,T]} e^{-n\cdot t}\cdot\frac{dV_S}{ds}\cdot e^{-\gamma\cdot t}\cdot\varphi\cdot\int_{[0,T]} e^{\gamma\cdot t_1}\cdot(c_1(t_1)-c(t_1))\cdot dt_1\cdot dt \\ &= \frac{dV}{ds}(s(T))\cdot e^{-(n+\gamma)\cdot T}\cdot\int_{[0,T]} e^{\gamma\cdot t}\cdot(c_1(t)-c(t))\cdot dt \\ &\quad +\int_{[0,T]}\int_{[t,T]} e^{-(n+\gamma)\cdot t_1}\cdot\frac{dV_S}{ds}(s(t_1))\cdot dt_1\cdot e^{\gamma\cdot t}\cdot(c_1(t)-c(t))\cdot dt \\ &= \int_{[0,T]} M(t)\cdot(c_1-c)(t)\cdot dt, \end{aligned}$$

where

$$M(t) = e^{\gamma\cdot t}\cdot\left(\frac{dV}{ds}(s(T))\cdot e^{-(n+\gamma)\cdot T}+\int_{[t,T]} e^{-(n+\gamma)\cdot t_1}\cdot\frac{dV_S}{ds}(s(t_1))\cdot dt_1\right) > 0 \tag{29}$$

since $\frac{dV_S}{ds} > 0$ and $\frac{dV}{ds} > 0$. Hence we obtain

$$\int_{[0,T]} M(t)\cdot c(t)\cdot dt \geq \int_{[0,T]} M(t)\cdot c_1(t)\cdot dt,$$

which means that the Borel measure c maximizes $\int_{[0,T]} M(t)\cdot c_1(t)\cdot dt$ on the set of all non-negative Borel measures c_1 that satisfy constraint (6). Since $M > 0$, the intuition behind the solution of this problem is as follows: Let

$$\lambda_W = \sup_{t\in[0,T]}(M(t)\cdot e^{r\cdot t}) > 0$$

Then the optimal c is any non-negative Borel measure that satisfies (6) and

$$\text{supp}\,(c) \subset \{t\in[0,T] \mid M(t)\cdot e^{r\cdot t} = \lambda_W\},$$

in which case

$$\begin{aligned}\int_{[0,T]} M(t)\cdot c(t)\cdot dt &= \int_{\text{supp}(c)} M(t)\cdot c(t)\cdot dt \\ &= \int_{\text{supp}(c)} M(t)\cdot e^{r\cdot t}\cdot e^{-r\cdot t}\cdot c(t)\cdot dt \\ &= \lambda_W\cdot\int_{\text{supp}(c)} e^{-r\cdot t}\cdot c(t)\cdot dt = \lambda_W\cdot W\end{aligned}$$

If c_1 is any non-negative Borel measure on [0, T] that satisfies (6), then

$$\begin{aligned}\int_{[0,T]} M(t)\cdot c_1(t)\cdot dt &= \int_{[0,T]} e^{r\cdot t}\cdot M(t)\cdot e^{-r\cdot t}\cdot c_1(t)\cdot dt \\ &\le \lambda_W\cdot\int_{[0,T]} e^{-r\cdot t}\cdot c_1(t)\cdot dt = \lambda_W\cdot W \\ &= \int_{[0,T]} M(t)\cdot c(t)\cdot dt\end{aligned}$$

Thus c maximizes $\int_{[0,T]} M(t)\cdot c_1(t)\cdot dt$ on the set of all non-negative Borel measures c_1 on [0, T] that satisfy (6). Now let

$$\lambda_c(t) = \lambda_W\cdot e^{-r\cdot t} - M(t) \tag{30}$$

Obviously, $\lambda_c \ge 0$ and $\lambda_c(t) = 0$ for $t \in \text{supp}(c)$, hence λ_c satisfies KKT condition (17). Next, (2), (3), (4), and (5) imply that $s > 0$ on the compact interval [0, T], hence $\inf_{t\in[0,T]}(s(t)) > 0$. Thanks to $\frac{d^2V_S}{ds^2} < 0$ (which is equivalent to $\alpha_S > 0$), this implies that $\frac{dV_S}{ds}(s)$ is bounded on [0, T], which in turn implies that $M \in Lip([0, T] \to R)$, where M is given by (29). Thus $\lambda_c \in Lip([0, T] \to R)$, where λ_c is given by (30). We need to show that λ_c satisfies the boundary condition (19) and the differential Eq (20). Indeed,

$$M(T) = e^{\gamma\cdot T}\cdot\frac{dV}{ds}(s(T))\cdot e^{-(n+\gamma)\cdot T} = \frac{dV}{ds}(s(T))\cdot e^{-n\cdot T},$$

hence

$$\lambda_c(T) = \lambda_W\cdot e^{-r\cdot T} - M(T) = \lambda_W\cdot e^{-r\cdot T} - \frac{dV}{ds}(s(T))\cdot e^{-n\cdot T},$$

so λ_c satisfies the boundary condition (19). Next,

$$\begin{aligned}\frac{d}{dt}M(t) &= \frac{d}{dt}\left[e^{\gamma\cdot t}\cdot\left(\frac{dV}{ds}(s(T))\cdot e^{-(n+\gamma)\cdot T} + \int_{[t,T]} e^{-(n+\gamma)\cdot t_1}\cdot\frac{dV_S}{ds}(s(t_1))\cdot dt_1\right)\right] \\ &= \gamma\cdot M + e^{\gamma\cdot t}\cdot\frac{d}{dt}\left(\frac{dV}{ds}(s(T))\cdot e^{-(n+\gamma)\cdot T} + \int_{[t,T]} e^{-(n+\gamma)\cdot t_1}\cdot\frac{dV_S}{ds}(s(t_1))\cdot dt_1\right) \\ &= \gamma\cdot M + e^{\gamma\cdot t}\cdot\frac{d}{dt}\int_{[t,T]} e^{-(n+\gamma)\cdot t_1}\cdot\frac{dV_S}{ds}(s(t_1))\cdot dt_1 \\ &= \gamma\cdot M - e^{\gamma\cdot t}\cdot e^{-(n+\gamma)\cdot t}\cdot\frac{dV_S}{ds}(s(t)) = \gamma\cdot M - e^{-n\cdot t}\cdot\frac{dV_S}{ds}(s(t)),\end{aligned}$$

so

$$\begin{aligned}\frac{d}{dt}\lambda_c(t) - \lambda_c \cdot \gamma &= \frac{d}{dt}(\lambda_W \cdot e^{-r\cdot t} - M(t)) - (\lambda_W \cdot e^{-r\cdot t} - M(t)) \cdot \gamma \\ &= -(r+\gamma)\cdot \lambda_W \cdot e^{-r\cdot t} - \frac{d}{dt}M(t) + M(t)\cdot\gamma \\ &= -(r+\gamma)\cdot \lambda_W \cdot e^{-r\cdot t} + e^{-n\cdot t}\cdot \frac{dV_S}{ds}(s(t)),\end{aligned}$$

which is the differential Eq (20).

The two proofs above use ideas from [18]. A few minor gaps in that material are covered here.

6 Structure of the optimal solutions

In this section the structure of solutions of Problem (10) is investigated, the main results being Theorems (21) and (25).

Discontinuities complicate the proofs, but they do not destroy the conclusions of Section 2. To a large extent, functions of bounded variation that are not continuously differentiable can be treated formally as if they were (e.g., in terms of differentiation and integration by parts). Note also that condition (5) and Eq (25) require the gulp of consumption $c(t) = \frac{s(t^+)-s(t^-)}{\varphi}$ at each jump in the satiation to be positive, that is, they require the singular part of c to be positive as well as the part that is continuous with respect to Lebesgue measure.

We first show that a discontinuity in the satiation at $t = T$ is essential, as it allows one to resolve the contradiction that arose in the proof of Proposition (8). When functions with bounded variation are considered, Eq (20) reduces to

$$0 = e^{-n\cdot T}\cdot \frac{dV_S}{ds}(s(T^-)) - \lambda_W \cdot e^{-r\cdot T}\cdot(\gamma + r), \tag{31}$$

where $s(T^-) = \lim_{t\to T, t<T} s(t)$. Eq (31) replaces Eq (22). If there is a discontinuity at $t = T$, then $s(T^-) \neq s(T)$; this provides an extra degree of freedom, which removes the contradiction between Eqs (21) and (31) at $t = T$ that eliminated continuous solutions. We can prove even more:

Proposition 14 *Let* $[s, \lambda_c, \lambda_W] \in BV([0, T] \to R) \times BV([0, T] \to R) \times R$ *be a solution of Problem 10. If* $\frac{dV}{ds}(s) > 0$ *and* $\frac{d^2V}{ds^2}(s) < 0$ *for all s, then the following hold*:

1. *There must be a gulp of consumption at t = T if for all s*

$$n < r + \alpha\cdot\gamma. \tag{32}$$

2. *A gulp of consumption at t = T is impossible if for all s*

$$n > r + \alpha\cdot\gamma. \tag{33}$$

Proof. When λ_W is eliminated between (21) and (31) for $t = T$, one obtains

$$0 = e^{-n\cdot T}\cdot\left(\frac{dV_S}{ds}(s(T^-)) - \frac{dV}{ds}(s(T))\cdot(\gamma + r)\right),$$

where $s(T^-) = \lim_{t \to T, t<T} s(t)$. With the help of (10), this can be transformed into

$$\frac{dV}{ds}(s(T)) = \frac{dV_S}{ds}(s(T^-)) \cdot \frac{1}{(\gamma + r)} = \frac{dV}{ds}(s(T^-)) \cdot \left(1 - \frac{r + \gamma \cdot \alpha - n}{(\gamma + r)}\right), \tag{34}$$

hence $\frac{dV}{ds}(s(T)) - \frac{dV}{ds}(s(T^-)) = \frac{dV}{ds}(s(T^-)) \cdot \frac{n-r-\alpha\cdot\gamma}{(\gamma+r)}$. Thus there is a discontinuity, $s(T) \neq s(T^-)$, if $n - r - \alpha(s) \cdot \gamma \neq 0$ for all s.

If (32) holds, then $\frac{dV}{ds}(s(T)) - \frac{dV}{ds}(s(T^-)) < 0$, hence $s(T) - s(T^-) > 0$ because $\frac{d^2V}{ds^2}(s) < 0$.

This discontinuity of $s(t)$ produces the singular part of $\frac{ds}{dt}|_{t=T} = s(T) - s(T^-)$; see Section 4. Now Eq (2) implies that the consumption c has singular part $\frac{s(T)-s(T^-)}{f}$, which we call a consumption gulp. This gulp must be positive, since consumption c must be positive; see (5). However, (33) implies that the gulp of consumption must be negative. Therefore, a gulp of consumption at T is impossible if (33) holds.

Proposition 14 resolves the non-existence contradiction from Section 3: The optimal solutions must be discontinuous at $t = T$ if (32) holds for all s. The next lemma provides the first step in deciphering the structure of the optimal solutions.

Lemma 15 *If* $[s, \lambda_c, \lambda_W] \in BV([0, T] \to R) \times BV([0, T] \to R) \times R$ *is a solution of Problem 10, then* $\lambda_c(t) \in Lip([0, T] \to R)$.

Proof. The satiation s is bounded (although it may be discontinuous if c is not absolutely continuous with respect to Lebesgue measure), because every function of bounded variation is bounded; see [16]. The boundedness of s and condition (19) imply boundedness of $\lambda_c(T)$, which, together with (20) and the help of the Gronwall lemma, implies boundedness of $\lambda_c(t)$ on $[0, T]$. This property, together with (20), implies boundedness of $\frac{d\lambda_c(t)}{dt}$, hence $\lambda_c(t)$ is Lipschitz continuous.

Continuity of λ_c permits us to deduce the following result:

Proposition 16 *Let* $[s, \lambda_c, \lambda_W] \in BV([0, T] \to R) \times Lip([0, T] \to R) \times R$ *be a solution of Problem 10. Then there is a family of at most countably many (relatively) open subintervals of* $[0, T]$ *such that* $\lambda_c > 0$ *and* $c = 0$ *in these subintervals, and* $\lambda_c = 0$ *and* $c \geq 0$ *at the isolated points as well as on the closed subintervals that separate these open subintervals.*

Proof. Thanks to Lemma 15, $\lambda_c(t)$ is continuous, hence the inverse image of $(0, +\infty)$ is relatively open in $[0, T]$. Since the set of rational numbers, which is countable, is dense in the set of all real numbers, there is at most a countable set of (relatively) open subintervals of $[0, T]$ in which $\lambda_c > 0$ and $c = 0$, while $\lambda_c = 0$ and $c > 0$ on each closed subinterval and at each single point that separates any two consecutive (relatively) open subintervals of $[0, T]$ in which $\lambda_c > 0$.

In the remaining part of this section, we prove that there can be at most one (relatively) open subinterval in which $\lambda_c > 0$ and $c = 0$, and at most one closed subinterval on which $\lambda_c = 0$ and $c > 0$; see Theorems 21 and 25. In the next lemma, we construct a basic block of the explicit solution of Problem 10 in open subintervals where $\lambda_c = 0$ and $c > 0$. In this case, Eq (20) simplifies to

$$\frac{e^{-(n-r)t}}{(\gamma + r)} \cdot \frac{dV_S}{ds}(\sigma) = \lambda \tag{35}$$

with $\lambda = \lambda_W \in R$. Eq (35) does not contain $\frac{d\sigma}{dt}$, and this simplifies forthcoming proofs.

Lemma 17 *If* $\frac{dV_S}{ds}(s) > 0$ *and* $\alpha_S(s) > 0$ *for all* $s \geq 0$, *and if both of the following hold,*

$$\lim_{s \to 0^+} \frac{dV_S}{ds} = +\infty \tag{36}$$

$$\lim_{s \to \infty} \frac{dV_S}{ds} = 0, \tag{37}$$

then Eq (35) has a unique bounded, continuous positive solution for all $t \geq 0$ *and* $0 < \lambda \in R$ *that satisfy Eq (39). The corresponding optimal consumption is given by*

$$c(t, \lambda) = \frac{r + \alpha_S \cdot \gamma - n}{f \cdot \alpha_S} \cdot \sigma(t, \lambda) \tag{38}$$

Proof. The conditions $\frac{dV_S}{ds}(s) > 0$ and $\lambda > 0$ ensure existence of a solution $\sigma(t, \lambda)$ of (35), and $\frac{d^2 V_S}{ds^2} < 0$ (which is equivalent to $\alpha_S > 0$) ensures uniqueness and continuity. The two "barrier" requirements, (36) and (37), ensure positivity and boundedness of $\sigma(t, \lambda)$. In order to derive (38), we differentiate the expression on the left-hand side of (35) with respect to t and obtain

$$0 = \frac{d}{dt}\left(\frac{e^{-(n-r)\cdot t}}{(\gamma + r)} \cdot \frac{dV_S}{ds}(\sigma(t))\right) = \frac{e^{-(n-r)\cdot t}}{(\gamma + r)} \cdot \alpha_S(\sigma) \cdot \frac{\frac{dV_S}{ds}}{\sigma} \cdot \left(\frac{(r-n)}{\alpha_S(\sigma)} \cdot \sigma - \frac{d\sigma(t)}{dt}\right)$$

Since $\frac{e^{-(n-r)\cdot t}}{(\gamma+r)} \cdot \alpha_S(\sigma) \cdot \frac{\frac{dV_S}{ds}}{\sigma} \neq 0$, we have

$$\frac{d\sigma(t)}{dt} = \frac{(r-n)}{\alpha_S(\sigma)} \cdot \sigma, \tag{39}$$

and now we transform (2) as follows:

$$c(t) = \frac{1}{f} \cdot \left(\frac{d\sigma}{dt} + \gamma \cdot \sigma\right) = \frac{1}{f} \cdot \left(\frac{(r-n)}{\alpha_S(\sigma)} \cdot \sigma + \gamma \cdot \sigma\right) = \frac{r + \alpha_S \cdot \gamma - n}{f \cdot \alpha_S} \cdot \sigma(t)$$

This completes the derivation of (38).

Remark 18 *In* [4] *a less restrictive assumption is used, namely* $\frac{dV_S}{ds}(s_M) = 0$ *with* $0 < s_M < \infty$ *instead of our* (37). *Our results could be extended, with some modifications, to that case too.*

Definition 19 *The solution of Eq (35),* $\sigma(t, \lambda)$, *whose existence and basic properties are as described in Lemma 17, is called the general solution.*

Now with the help of Lemma 17 we can derive the properties of subintervals of $[0, T]$ in addition to those we derived in Proposition 16.

Proposition 20 *Let* $[s, \lambda_c, \lambda_W] \in BV([0, T] \to R) \times Lip([0, T] \to R) \times R$ *be a solution of Problem 10. In every subinterval of* $[0, T]$ *in which* $\lambda_c > 0$, *the satiation is of the form* $s(t) = const \cdot e^{-\gamma \cdot t}$ *and the following hold*:

1. *In every open subinterval of* $[0, T]$ *in which* $\lambda_c = 0$, *the optimal satiation is* $s(t) = \sigma(t, \lambda_W)$ *if for all* s

$$n < r + \alpha_S \cdot \gamma. \tag{40}$$

2. *Consumption* $c(t)$ *is* 0 *everywhere in* $(0, T)$ *if for all* s

$$n > r + \alpha_s \cdot \gamma. \tag{41}$$

Proof. If $\lambda_c > 0$ in a (relatively) open subinterval, the first KKT condition in (17) implies that $c = 0$ in that subinterval, hence the solution of (2) is of the form $s(t) = const \cdot e^{-\gamma \cdot t}$. If $\lambda_c = 0$ on a closed subinterval, Eq (20) simplifies to (35) with $\lambda = \lambda_W$, hence $s(t) = \sigma(t, \lambda_W)$. If (40) holds for all s, the corresponding consumption is positive. However, if (41) holds for all s, then (38) implies that $c < 0$, which violates requirement (5), hence consumption must be everywhere 0 in $(0, T)$.

Proposition 20 also shows that $s(t)$ must be continuously differentiable everywhere, except possibly at the endpoints of the open subintervals in which $\lambda_c = 0$, where there may be discontinuities in satiation and gulps of consumption; however, as will be proved in Lemma 24, this possibility is excluded. Proposition 20 also shows that the solutions can be divided into two groups. Each of these two groups corresponds to a particular pair of sufficient conditions, one pair being the conditions given in inequalities (32) and (40), and the other pair being the conditions given in inequalities (33) and (41). Hence these two groups of solutions will be investigated separately. The structure of solutions of Problem 10 under (32) and (40) is as described in Theorem 25 (Section 6.1), and the structure of solutions of Problem 10 under (33) and (41) is as described in Theorem 21 (Section 6.2).

6.1 The structure and existence of solutions for large future discount

When the future discount rate n is large, intuition suggests that all the consumption should take place at the beginning of the interval $[0, T]$. This is confirmed by the following theorem:

Theorem 21 (*Explicit solution when the future discount is large*) *If both* V *and* V_S *are increasing, concave-down utilities and* (33) *and* (41) *hold for all* $s \geq 0$, *then the optimal solution* $[s, \lambda_c, \lambda_W] \in BV([0, T] \to R) \times Lip([0, T] \to R) \times R$ *is as follows*: s *is given by*

$$s(t) = \begin{cases} s_0\,, & t = 0 \\ s_+ \cdot e^{-\gamma \cdot t}, & t \in (0, T], \end{cases} \tag{42}$$

where $s_+ = f \cdot W + s_0$ (see (44)); λ_W *is given by* (49), $\lambda_W > 0$; *and* λ_c *is given by* (53), $\lambda_c > 0$. *All the consumption takes place in a single gulp at* $t = 0$.

Proof. By (33), the second part of Proposition 14 implies that there is no gulp of consumption at $t = T$, and Lemma 20 states that there is no consumption in $(0, T)$, hence all the consumption must happen in a gulp at $t = 0$. Therefore, satiation is of the form (42), and the corresponding consumption is given by

$$c(t) = \delta_0(t) \cdot \frac{s_+ - s_0}{f} = \begin{cases} \frac{s_+ - s_0}{f}, & t = 0 \\ 0\,, & t \in (0, T] \end{cases} \tag{43}$$

The constant s_+ can be calculated using the wealth constraint (13):

$$s_+ = f \cdot W + s_0 > s_0 \tag{44}$$

The condition $s_+ > s_0$ ensures that the gulp of consumption at $t = 0$ is positive. We need to show that $\lambda_W > 0$ and $\lambda_c > 0$ in $(0, T)$. For λ_W and λ_c, we have the following equations from Problem 10:

For $t = 0$, by the first KKT condition in (17) and the fact that there is a gulp of consumption ($c(0) > 0$):

$$\lambda_c(0) = 0 \tag{45}$$

For $t \in (0, T)$, from (20):

$$0 = e^{-n\cdot t}\cdot\frac{dV_S}{ds}\left(s_+\cdot e^{-\gamma\cdot t}\right) - \lambda_W\cdot e^{-r\cdot t}\cdot(\gamma + r) + \left(-\frac{d\lambda_c}{dt} + \lambda_c\cdot\gamma\right) \tag{46}$$

For $t = T$, from (19):

$$0 = e^{-n\cdot T}\cdot\frac{dV}{ds}\left(s_+\cdot e^{-\gamma\cdot T}\right) - \lambda_W\cdot e^{-r\cdot T} + \lambda_c(T) \tag{47}$$

We need to prove that $\lambda_c(t) > 0$ in $(0,T]$ and $\lambda_W > 0$. The solution of (46), together with the initial condition (45), is

$$\lambda_c(t) = e^{\gamma\cdot t}\cdot\int_0^t e^{-(n+\gamma)\cdot\tau}\cdot\frac{dV_S}{ds}\left(s_+\cdot e^{-\gamma\cdot\tau}\right)\cdot d\tau - \lambda_W\cdot e^{\gamma\cdot t}\cdot\left(1 - e^{-(r+\gamma)\cdot t}\right) \tag{48}$$

After substitution of this into the boundary condition (47), we obtain an equation for λ_W:

$$\lambda_W = e^{-(n+\gamma)\cdot T}\cdot\frac{dV}{ds}\left(s_+\cdot e^{-\gamma\cdot T}\right) + \int_0^T e^{-(n+\gamma)\cdot\tau}\cdot\frac{dV_S}{ds}\left(s_+\cdot e^{-\gamma\cdot\tau}\right)\cdot d\tau \tag{49}$$

This implies that $\lambda_W > 0$, since $\frac{dV}{ds} > 0$ and $\frac{dV_S}{ds} > 0$. This expression can be simplified. First, note that

$$\frac{d}{d\tau}\left(e^{-(n+\gamma)\cdot\tau}\cdot\frac{dV}{ds}\left(s_+\cdot e^{-\gamma\cdot\tau}\right)\right) = -e^{-(n+\gamma)\cdot\tau}\cdot\frac{dV_S}{ds}\left(s_+\cdot e^{-\gamma\cdot\tau}\right), \tag{50}$$

hence the first term on the right-hand side of (49) can be transformed as follows:

$$e^{-(n+\gamma)\cdot T}\cdot\frac{dV}{ds}\left(s_+\cdot e^{-\gamma\cdot T}\right) = \frac{dV}{ds}\left(s_+\right) - \int_0^T e^{-(n+\gamma)\cdot\tau}\cdot\frac{dV_S}{ds}\left(s(t)\right)\cdot d\tau \tag{51}$$

Therefore, by (10),

$$\begin{aligned}\lambda_W &= \frac{dV}{ds}\left(s_+\right) - \int_0^T e^{-(n+\gamma)\cdot\tau}\cdot(n+\gamma\cdot(1-\alpha))\cdot\frac{dV}{ds}\cdot d\tau \\ &\quad + \int_0^T e^{-(n+\gamma)\cdot\tau}\cdot(n+\gamma\cdot(1-\alpha))\cdot\frac{dV}{ds}\cdot d\tau \\ &= \frac{dV}{ds}\left(s_+\right) > 0\end{aligned}$$

Now (48) takes the form

$$\lambda_c(t) = e^{\gamma\cdot t}\cdot\int_0^t e^{-(n+\gamma)\tau}\cdot\frac{dV_S}{ds}\left(s_+\cdot e^{-\gamma\cdot\tau}\right)\cdot d\tau - \frac{dV}{ds}\left(s_+\right)\cdot e^{\gamma\cdot t} + \frac{dV}{ds}\left(s_+\right)\cdot e^{-r\cdot t} \tag{52}$$

Integration of (50) on $[0, t]$ produces

$$e^{-(n+\gamma)\cdot t}\cdot\frac{dV}{ds}\left(s(t)\right) - \frac{dV}{ds}\left(s_+\right) = -\int_0^t\left(e^{-(n+\gamma)\cdot\tau}\cdot\frac{dV_S}{ds}\left(s_+\cdot e^{-\gamma\cdot\tau}\right)\right)\cdot d\tau,$$

hence

$$\frac{dV}{ds}(s_+) = e^{-(n+\gamma)\cdot t}\cdot\frac{dV}{ds}(s(t)) + \int_0^t e^{-(n+\gamma)\cdot\tau}\cdot\frac{dV_S}{ds}(s_+\cdot e^{-\gamma\cdot\tau})\cdot d\tau,$$

and now (52) takes the form

$$\begin{aligned}\lambda_c(t) &= \frac{dV}{ds}(s_+)\cdot e^{-r\cdot t} - e^{-nt}\cdot\frac{dV}{ds}(s_+\cdot e^{-\gamma t})\\ &= -e^{-r\cdot t}\cdot\left[e^{-(n-r)t}\cdot\frac{dV}{ds}(s_+\cdot e^{-\gamma t}) - \frac{dV}{ds}(s_+)\right].\end{aligned}$$

Using transformations similar to those used in the proof of (51), we can prove the following:

$$e^{-(n-r)\cdot t}\cdot\frac{dV}{ds}(s_+\cdot e^{-\gamma t}) - \frac{dV}{ds}(s_+) = \int_0^t (n-r-\alpha\cdot\gamma)\cdot e^{-(n-r)\cdot\tau}\cdot\frac{dV}{ds}(s_+\cdot e^{-\gamma\cdot\tau})\cdot d\tau,$$

hence

$$\lambda_c(t) = e^{-r\cdot t}\cdot\int_0^t (n-r-\alpha\cdot\gamma)\cdot e^{-(n-r)\cdot\tau}\cdot\frac{dV_S}{ds}(s_+\cdot e^{-\gamma\cdot\tau})\cdot d\tau, \tag{53}$$

which is positive, thanks to (33) and the fact that $\frac{dV_S}{ds} > 0$.

Remark 22 *Since the objective function in Problem 10 is concave down, the sufficient conditions are also necessary, and there is only one solution. This result establishes the existence and uniqueness of solutions of the equivalent of Problem 1 in the class of BV functions, independent of the utility.*

6.2 Small future discount rate and solutions with positive distributed consumption

In this section we discuss the structure of solutions under assumptions (32) and (40). The results of this section permit us to prove the existence of solutions in Section 7. We begin with a lemma, which we will use in the proof of Theorem 25, that permits us to reduce the number of intervals as described in Propositions 16 and 20 (with or without consumption) to no more than two.

Lemma 23 *Let* $[s, \lambda_c, \lambda_W] \in BV([0, T] \to R) \times Lip([0, T] \to R) \times R$ *be a solution of Problem 10, and let* $\alpha_S(s) > 0$. *Suppose there are two times* t_1, t_2 *with* $0 \le t_1 < t_2 \le T$ *at which* $\lambda_c(t_1) = \lambda_c(t_2) = 0$. *Then* $\lambda_c(t) = 0$, $s(t) = \sigma(t, \lambda_W)$ (*the general solution from Definition 19*), *and* $c(t) > 0$ *for all* $t \in (t_1, t_2)$, *hence the non-negative Borel measure c is continuous with respect to Lebesgue measure in* (t_1, t_2).

Proof. In (t_1, t_2), $s(t)$ is a solution of Eq (20), while the general solution $\sigma(t, \lambda_W)$ is a solution of (35). We eliminate λ_W from these two equations by subtraction to obtain

$$0 = e^{-n\cdot t}\cdot\left(\frac{dV_S}{ds}(s(t)) - \frac{dV_S}{ds}(\sigma(t,\lambda_W))\right) + \left(-\frac{d\lambda_c}{dt} + \lambda_c\cdot\gamma\right).$$

Multiplying this equation by $(s(t) - \sigma(t, \lambda_W))$, and with the help of the Fundamental Theorem

of Calculus, we obtain

$$\begin{aligned} 0 &= e^{-n\cdot t}\cdot(s(t)-\sigma(t,\lambda_W))^2\cdot\int_0^1\frac{d^2V_S}{ds^2}((\sigma(t,\lambda_W))+x\cdot(s(t))-(\sigma(t,\lambda_W)))\cdot dx \\ &\quad -\frac{d}{dt}[(s(t)-\sigma(t,\lambda_W))\cdot\lambda_c] \\ &\quad +\frac{d}{dt}(s(t)-\sigma(t,\lambda_W))\cdot\lambda_c+(s(t)-\sigma(t,\lambda_W))\cdot\lambda_c\cdot\gamma \end{aligned}$$

Next, we integrate this over (t_1, t_2) to obtain

$$\begin{aligned} 0 &= \int_{(t_1,t_2)}\left\{\left[e^{-(n+\gamma)\cdot t}\cdot(s(t)-\sigma(t,\lambda_W))^2\right]\right. \\ &\qquad \left.\cdot\int_0^1\left[\frac{d^2V_S}{ds^2}(\sigma(t,\lambda_W)+x\cdot(s(t)-\sigma(t,\lambda_W)))\right]\cdot dx\right\}\cdot dt \\ &\quad +\int_{(t_1,t_2)}\left[\frac{d}{dt}(s(t)-\sigma(t,\lambda_W))+(s(t)-\sigma(t,\lambda_W))\cdot\gamma\right]\cdot\lambda_c\cdot dt, \end{aligned} \tag{54}$$

since the boundary terms $[(s(t)-\sigma(t,\lambda_W))\cdot\lambda_c(t)]_{t=t_1}^{t=t_2}$ vanish, thanks to the fact that $\lambda_c(t_1)=\lambda_c(t_2)=0$. Next, thanks to (2) and (39),

$$\frac{d}{dt}(s(t)-\sigma(t,\lambda_W))=\left(c(t)-\gamma\cdot s(t)-\frac{(r-n)}{\alpha_S(\sigma(t,\lambda_W))}\cdot\sigma(t,\lambda_W)\right),$$

hence the second integral in (54) can be rewritten as follows:

$$\begin{aligned} &\int_{(t_1,t_2)}\left(\frac{d}{dt}(s(t)-\sigma(t,\lambda_W))\cdot\lambda_c+\gamma\cdot(s(t)-\sigma(t,\lambda_W))\cdot\lambda_c\right)\cdot dt \\ &= \int_{(t_1,t_2)}c(t)\cdot\lambda_c\cdot dt-\int_{(t_1,t_2)}\left[\frac{(r-n)}{\alpha_S(\sigma(t,\lambda_W))}+\gamma\right]\cdot\sigma(t,\lambda_W)\cdot\lambda_c\cdot dt \\ &= -\int_{(t_1,t_2)}\frac{r+\gamma\cdot\alpha_S-n}{\alpha_S}\cdot\sigma(t,\lambda_W)\cdot\lambda_c\cdot dt, \end{aligned}$$

since $c(t)\cdot\lambda_c(t)=0$, thanks to the first KKT condition in (17). Hence we obtain

$$\begin{aligned} &\int_{(t_1,t_2)}\left\{\left[e^{-(n+\gamma)\cdot t}\cdot(s(t)-\sigma(t,\lambda_W))^2\right]\right. \\ &\qquad \left.\cdot\left[\int_0^1\frac{d^2V_S}{ds^2}(\sigma(t,\lambda_W)+x\cdot(s(t)-\sigma(t,\lambda_W)))\cdot dx\right]\right\}\cdot dt \\ &= \int_{(t_1,t_2)}\frac{(r+\alpha_S\cdot\gamma-n)}{\alpha_S}\cdot\sigma(t,\lambda_W)\cdot\lambda_c(t)\cdot dt \end{aligned}$$

The left-hand side is less than or equal to 0, thanks to the fact that $\frac{d^2V_S}{ds^2}<0$, while the right-hand side is greater than or equal to 0, thanks to (40). Moreover, $\sigma(t,\lambda_W)>0$ by Lemma 17, and $\lambda_c(t)\geq 0$. Hence both sides must be equal to 0. This is possible only if $(s(t)-\sigma(t,\lambda_W))=0$, hence $c>0$, thanks to (38) and (40), so $\lambda_c(t)=0$ in (t_1, t_2).

Lemma 23 implies that between any two open intervals with $c(t)>0$ and $\lambda_c=0$ there cannot be any interval with $\lambda_c>0$, which in turn implies that there can be at most one interval with $\lambda_c>0$ and at most one interval with $\lambda_c=0$. This fact is used in the proof of Theorem 25. The

next lemma, which is also used in that proof, permits us to reduce the number of gulps of consumption to no more than two.

Lemma 24 *Let* $[s, \lambda_c, \lambda_W] \in BV([0, T] \to R) \times Lip([0, T] \to R) \times R$ *be a solution of Problem 10. If* $\varepsilon > 0$ *and* $T_s \in (0, T)$, *and* T_s *is a switching boundary between an interval* $(T_s - \varepsilon, T_s)$ *in which* $c = 0$ *and* $\lambda_c > 0$ *and an interval* $[T_s, T_s + \varepsilon)$ *in which* $c > 0$ *and* $\lambda_c = 0$, *then the satiation* $s(t)$ *is continuous at the switching time* T_s *and there is no gulp of consumption at* T_s. *The non-negative Borel measure c is continuous with respect to Lebesgue measure in* $(0, T)$.

Proof. If there is a gulp of consumption $\frac{s(T_s^+) - s(T_s^-)}{f} \neq 0$ at T_s, then inequality constraint (5) implies that it must be non-negative. Since $f > 0$, it must be the case that $s(T_s^+) \geq s(T_s^-)$. Therefore, in order to prove that $s(T_s^+) = s(T_s^-)$, we have to prove that $s(T_s^+) \leq s(T_s^-)$. In $(T_s - \varepsilon, T_s)$ satiation $s(t)$ satisfies Eq (2) with $c = 0$, hence $s(t) = const \cdot e^{-\gamma \cdot (t - T_s)} = s(T_s^-) \cdot e^{-\gamma \cdot (t - T_s)}$ in $(T_s - \varepsilon, T_s)$, where $s(T_s^-) = \lim_{t \to T_s, t < T_s} s(t)$. In $[T_s, T_s + \varepsilon)$ the satiation $s(t)$ satisfies Eq (20) with $\lambda_c(t) = 0$, that is, it satisfies Eq (35), hence $\lambda_W = \lim_{t \to T_s, t > T_s} \frac{e^{-(n-r) \cdot t}}{e^{-(n-r) \cdot t}(\gamma + r)} \cdot \frac{dV_S}{ds}(s(t)) = \frac{e^{-(n-r) \cdot T_s}}{(\gamma + r)} \cdot \frac{dV_S}{ds}\left(s\left(T_s^+\right)\right)$, where $s(T_s^+) = \lim_{t \to T_s, t > T_s} s(t)$. Since $\lambda_c(t) \geq 0$ in $(T_s - \varepsilon, T_s)$ and $\lambda_c(T_s) = 0$, we have $0 \geq \frac{d\lambda_c}{dt}\left(T_s^-\right)$. In $(T_s - \varepsilon, T_s)$ the Lagrange multiplier $\lambda_c(t)$ satisfies Eq (20) with $\lambda_c(t) \geq 0$, thanks to the second KKT condition in (17), hence

$$\begin{aligned} 0 &\geq \frac{d\lambda_c}{dt}\left(T_s^-\right) \\ &= e^{-n \cdot T_s} \cdot \frac{dV_S}{ds}\left(s\left(T_s^-\right)\right) - \lambda_W \cdot e^{-r \cdot T_s} \cdot (\gamma + r) + \lambda_c(T_s) \cdot \gamma \\ &= e^{-n \cdot T_s} \cdot \left[\frac{dV_S}{ds}\left(s\left(T_s^-\right)\right) - \frac{dV_S}{ds}\left(s\left(T_s^+\right)\right)\right] \end{aligned}$$

for $\lambda_c(T_s) = 0$, because $\lambda_c(t)$ is continuous and $\lambda_c(t) = 0$ for $t > T_s$; therefore, $\frac{dV_S}{ds}\left(s\left(T_s^+\right)\right) \geq \frac{dV_S}{ds}\left(s\left(T_s^-\right)\right)$. Since $\frac{d^2V_S}{ds^2} < 0$, the above is possible only if $s(T_s^+) \leq s(T_s^-)$, which, together with the opposite inequality $(s(T_s^+) - s(T_s^-)) \geq 0)$, implied by $c > 0$ and $f > 0$, implies that $s(T_s^+) = s(T_s^-)$, and hence that $s(t)$ is continuous at $t = T_s$. The consumption $c(t)$ is 0 in every relatively open interval where $\lambda_c(t) > 0$, hence it is trivially continuous with respect to Lebesgue measure there, and similarly it is continuous with respect to Lebesgue measure in intervals where $c(t) > 0$ by Lemma 23. Now continuity of satiation s at the boundaries between these two types of intervals, which is as described in Proposition 16, implies continuity of consumption c with respect to Lebesgue measure throughout the interval $(0, T)$.

Now we can prove the main result of this section.

Theorem 25 (***The structure of solutions when the future discount is small***) *Let* $[s, \lambda_c, \lambda_W] \in BV([0, T] \to R) \times Lip([0, T] \to R) \times R$ *be a solution of Problem 10. If* (32) *and* (40) *hold, then there is a switching time* $T_s \in [0, T]$ *such that* $c(t) = 0$ *and* $\lambda_c(t) > 0$ *for* $t \in [0, T_s)$; $c(t) = \frac{r + \alpha_S \cdot \gamma - n}{\alpha_S} \cdot s(t) > 0$ *and* $\lambda_c(t) = 0$ *for* $t \in [T_s, T]$; *and* $s(t)$ *is continuous at* $t = T_s$ *if* $T_s < T$. *There is always a gulp of consumption at* $t = T$. *If* $T_s > 0$, *there is no gulp of consumption at* $t = 0$; *if* $T_s = 0$, *a gulp of consumption is possible.*

Proof. Suppose $[s, \lambda_c, \lambda_W] \in BV([0, T] \to R) \times Lip([0, T] \to R) \times R$ is a solution of Problem 10. By Proposition 14, there is a gulp of consumption at $t = T$, and hence the first KKT condition in (17) implies that $\lambda_c(T) = 0$. By Proposition 16, there may be a countable set of closed intervals and isolated points where $\lambda_c = 0$, and any two of these are separated by an open interval where $\lambda_c > 0$. Let $T_s = \inf\{t \in [0, T] | \lambda_c(t) = 0\}$. Since λ_c is continuous (Lemma 15),

$\lambda_c(T_s) = 0$. Now Proposition 23 implies that $c(t) > 0$ for all $t \in [T_s, T]$, hence the first KKT condition in (17) implies that $\lambda_c(t) = 0$ in $[T_s, T]$. Eq (38) implies that $c(t) = \frac{r+\alpha_S \cdot \gamma - n}{\alpha_S} \cdot s(t)$. The definition of T_s implies that $\lambda_c(t) > 0$ and $c(t) = 0$ for all $t \in [0, T_s)$. If $T_s < T$ then the continuity of $s(t)$ at $t = T_s$, and hence the absence of a gulp of consumption at $t = T_s$, follows from Lemma 24. If there is a gulp of consumption at $t = 0$, then T_s is necessarily 0, hence there is no gulp of consumption at $t = 0$ if $T_s > 0$.

Theorem 25 establishes the existence of a unique solution of Problem 10. Since both of the utilities V and V_S are assumed to be concave down, every solution of boundary value Problem 10 is also a maximizer of the functional in (1) in Problem 1 on the BV space. Downward concavity also implies uniqueness; see [13]. In addition, Theorem 25 states that there must be a gulp of consumption at $t = T$; that if there is a gulp of consumption at $t = 0$, then $c(t) > 0$ in all of $[0, T]$; and that if there is no gulp of consumption at $t = 0$, there may be an initial interval $[0, T_s)$ without consumption($c(t) = 0$), but $c(t) > 0$ in $[T_s, T]$, where $T_s = T$ is not excluded. The structure of solutions is more complicated when the following two pairs of assumptions are relaxed: (33) and (41), and (32) and (40).

7 Existence of solutions for small future discount

With the help of the results of Section 6, we can reduce the problem of the existence of solutions to solving non-differential equations, and we prove their solvability here. We prove theorems that, with the help of the diagnostic profiles described in Section 7.1, allow us to use the problem data to determine the type of the solution.

Theorem 25 implies that under assumptions (32) and (40), the original problem (Problem 1) does have a unique solution and there are only three possible types of solutions of Problem 10, described as follows:

Definition 26 *A solution is I-shaped (poor or over-satiated agent) when* $c = 0$ *and* $\lambda_c > 0$ *in* $[0, T)$, *and all consumption is done in a single gulp at* $t = T$; *J-shaped (intermediate agent) when* $c = 0$ *and* $\lambda_c > 0$ *in* $[0, T_s)$, *and* $c > 0$ *and* $\lambda_c = 0$ *in* $[T_s, T]$, *and there is a gulp at* $t = T$; *U-shaped (rich or under-satiated agent) when there is a gulp at* $t = 0$, $c > 0$ *and* $\lambda_c = 0$ *in all of* $[0, T]$, *and there is a gulp at* $t = T$.

Section 6 contains almost a complete proof of existence of a solution of Problem 10. The only missing part is the proof of the second KKT condition in (17): $\lambda_c \geq 0$. We now reduce this question to a simple non-differential inequality.

Lemma 27 *Suppose* (40) *holds for all s and that* $s_0 > 0$. *Let* $[s, \lambda_c, \lambda_W] \in BV([0, T] \to R) \times Lip([0, T] \to R) \times R$ *be an I-shaped solution* ($T_s = T$) *or a J-shaped solution* ($0 < T_s < T$) *that satisfies all the requirements of Problem 10, except possibly the KKT condition* $\lambda_c \geq 0$ (*see* (17)) *in* $[0, T_s)$ *with* $0 < T_s \leq T$. *Then* $\lambda_c(t) > 0$ *in* $[0, T_s)$ *iff*

$$\lambda_W > \frac{1}{(\gamma + r)} \cdot e^{-(n-r)\cdot T_s} \cdot \frac{dV_S}{ds}(s_0 \cdot e^{-\gamma \cdot T_s}) \tag{55}$$

Proof. Eq (20) for λ_c implies that in $[0, T_s)$

$$\frac{d\lambda_c(t)}{dt} - \lambda_c(t) \cdot \gamma = e^{-n\cdot t} \cdot \frac{dV_S}{ds}(s_0 \cdot e^{-\gamma \cdot t}) - \lambda_W \cdot e^{-r\cdot t} \cdot (\gamma + r) \tag{56}$$

We need to prove that $\lambda_c > 0$ in $[0, T_s)$. Multiplying (56) by $e^{-\gamma \cdot t}$, we obtain

$$\frac{d}{dt}(\lambda_c(t) \cdot e^{-\gamma \cdot t}) = e^{-(\gamma + r)\cdot t} \cdot \left[e^{-(n-r)\cdot t} \cdot \frac{dV_S}{ds}(s_0 \cdot e^{-\gamma \cdot t}) - \lambda_W \cdot (\gamma + r) \right],$$

hence it suffices to show that $(\lambda_c(t) \cdot e^{-\gamma \cdot t}) > 0$ in $[0, T_s)$. The derivative of the expression in

square brackets can be written as

$$e^{-(n-r)\cdot t}\cdot[r-n+\gamma\cdot\alpha_S]\cdot\frac{dV_S}{ds}(s_0\cdot e^{-\gamma\cdot t}),$$

which is positive, thanks to $\frac{dV_S}{ds}(s)>0$ and (40), hence the expression in square brackets,

$$e^{-(n-r)\cdot t}\cdot\frac{dV_S}{ds}(s_0\cdot e^{-\gamma\cdot t})-\lambda_W\cdot(\gamma+r),$$

takes its largest value in $[0, T_s)$ at $t = T_s$, and that value is

$$e^{-(n-r)\cdot T_s}\cdot\frac{dV_S}{ds}(s_0\cdot e^{-\gamma\cdot T_s})-\lambda_W\cdot(\gamma+r).$$

If this is positive, then $\frac{d}{dt}(\lambda_c(t)\cdot e^{-\gamma\cdot t})>0$ for all t close to but less than T_s, hence $(\lambda_c(t)\cdot e^{-\gamma\cdot t})<(\lambda_c(T_s)\cdot e^{-\gamma\cdot T_s})=0$ for all such t. Therefore, condition (55) is necessary for $\lambda_c(t)\geq 0$. It remains to show that this is also a sufficient condition. Suppose condition (55) holds. Since $\frac{d^2V_S}{ds^2}<0$, we obtain

$$\begin{aligned}\frac{d}{dt}(\lambda_c(t)\cdot e^{-\gamma\cdot t}) &= e^{-(\gamma+r)\cdot t}\cdot\left[e^{-(n-r)\cdot t}\cdot\frac{dV_S}{ds}(s_0\cdot e^{-\gamma\cdot t})-\lambda_W\cdot(\gamma+r)\right]\\ &\leq e^{-(\gamma+r)\cdot t}\cdot\left[e^{-(n-r)\cdot T_s}\cdot\frac{dV_S}{ds}(s_0\cdot e^{-\gamma\cdot T_s})-\lambda_W\cdot(\gamma+r)\right]<0,\end{aligned}$$

hence $\frac{d}{dt}(\lambda_c(t)\cdot e^{-\gamma\cdot t})<0$ for $t\in[0, T_s)$. Therefore, $(\lambda_c(t)\cdot e^{-\gamma\cdot t})$ is a decreasing function, so $(\lambda_c(t)\cdot e^{-\gamma\cdot t})>(\lambda_c(T_s)\cdot e^{-\gamma\cdot T_s})=0$ for all $t\in[0, T_s)$. Thus condition (55) is necessary and sufficient for $\lambda_c(t)>0$ in $[0, T_s)$.

Remark 28 *If $s_0\to 0$, then condition (55) cannot be satisfied, since that would imply that $\lambda_W>\lim_{s_0\to 0}\frac{dV_S}{ds}(s_0\cdot e^{-\gamma\cdot T_s})=\lim_{s\to 0}\frac{dV_S}{ds}(s)=\infty$, thanks to (36). Hence λ_W would not be real, so the interval $[0, T_s)$ must be empty, that is, $T_s = 0$.*

7.1 The diagnostic profiles

In order to describe how the data of the problem determine the structure of solutions, we introduce two diagnostic profiles; these are the solutions with wealth constraint (6) ignored.

Definition 29 *If $s_0>0$, then the lower diagnostic profile I/J $[s^l, \lambda_c^l, \lambda_W^l]$ (in between I-shaped solutions and J-shaped solutions; see Definition 26) for Problem 10 is*

$$s^l(t)=\begin{cases}s_0\cdot e^{-\gamma\cdot t}, & t\in[0,T)\\ s_T^l, & t=T,\end{cases}$$

where s_T^l is a solution of (21) for $s(T)$,

$$e^{-(n-r)T}\cdot\frac{dV}{ds}(s_T^l)=\lambda_W^l, \tag{57}$$

and where

$$\lambda_W^l=\frac{e^{-(n-r)T}}{(\gamma+r)}\cdot\frac{dV_S}{ds}(s_0\cdot e^{-\gamma\cdot T}) \tag{58}$$

and $\lambda_c^l(t)$ *is the solution of the terminal-value problem*

$$\lambda_c^l(T) = 0, \quad e^{-nt} \cdot \frac{dV_S}{ds}(s_0 e^{-\gamma \cdot t}) - \lambda_W^l \cdot e^{-r \cdot t} \cdot (\gamma + r) + \left(-\frac{d\lambda_c^l}{dt} + \lambda_c^l \cdot \gamma \right) = 0.$$

The corresponding consumption is $c^l(t) = \delta_T(t) \cdot \frac{(s_T^l - s_0 \cdot e^{-\gamma \cdot T})}{f}$, *and the wealth* W^l *necessary to sustain this level of consumption is*

$$\begin{aligned} W^l &= \int_{[0,T]} e^{-r \cdot t} \cdot c(t) \cdot dt = \int_{[0,T]} e^{-r \cdot t} \cdot \delta_T(t) \cdot \frac{(s_T^l - s_0 \cdot e^{-\gamma \cdot T})}{f} \cdot dt \\ &= e^{-r \cdot T} \cdot \frac{(s_T^l - s_0 \cdot e^{-\gamma \cdot T})}{f} \end{aligned} \tag{59}$$

Proposition 30 $\lambda_c^l > 0$ *in* $[0, T)$.

Proof. By Lemma 27, it suffices to prove that condition (55) holds, which follows from (58), since $T_s = T$ in the context of that lemma.

Definition 31 *If* $s_0 > 0$, *then the upper diagnostic profile U/J* $[s^u, \lambda_c^u = 0, \lambda_W^u]$ *(in between U-shaped solutions and J-shaped solutions; see Definition 26) for Problem 10, without a gulp of consumption at* $t = 0$, *is*

$$s^u(t) = \left\{ \begin{matrix} \sigma(t, \lambda_W^u), & t \in [0, T) \\ s_T^u, & t = T \end{matrix} \right\}, \lambda_c^u = 0 \text{ in } [0, T],$$

where $\sigma(t, \lambda)$ *is as described in Definition 19 and* λ_W^u is chosen so that $\sigma(0, \lambda_W^u) = s_0$, *which by* (35) *is equivalent to* $\frac{1}{(\gamma+r)} \cdot \frac{dV_S}{ds}(s_0) = \lambda_W^u$, *and* $s_T^u = s_T$, *where* s_T *is the solution of* Eq (19) *with* $\lambda_c = 0$: $\lambda_W^u = \frac{1}{(\gamma+r)} \cdot \frac{dV_S}{ds}(s_0) = e^{-(n-r)T} \cdot \frac{dV}{ds}(s_T^u)$. *The corresponding consumption is*

$$c^u(t) = \frac{r + \alpha_S \cdot \gamma - n}{\alpha_S} \cdot \sigma(t, \lambda_W^u) + \delta_T(t) \cdot \frac{(s_T^u - \sigma(T, \lambda_W^u))}{f},$$

and the wealth W^u *needed to sustain this level of consumption is*

$$\begin{aligned} f \cdot W^u &= \int_{[0,T]} e^{-r \cdot t} \cdot \left(\frac{ds^u}{dt} + \gamma \cdot s^u \right) \cdot dt \\ &= \int_{[0,T]} \left(\frac{d}{dt}(e^{-r \cdot t} \cdot s^u) + e^{-r \cdot t} \cdot (\gamma + r) \cdot s^u \right) \cdot dt \\ &= e^{-r \cdot T} \cdot s_T^u - s_0 + (\gamma + r) \cdot \int_{[0,T]} e^{-r \cdot t} \cdot \sigma(t, \lambda_W^u) \cdot dt \end{aligned} \tag{60}$$

Remark 32 *If* $s_0 = 0$, *then it is reasonable to accept* $W^u = W^l = 0$, $s_u(t) = s_l(t) = 0$, *and* $\lambda_W^u = \lambda_W^l = +\infty$.

Remark 33 *The discrete analogue of the upper diagnostic profile is the only solution exhibited in* [1]; *other solutions given there were obtained numerically*.

The rest of this section is devoted to proving the following: If $W \leq W^l$, the solution is I-shaped; if $W^l < W \leq W^u$, the solution is J-shaped; and if $W^u < W$, the solution is U-shaped. This is achieved by reduction of Problem 10 to a non-differential equation (Eq (68)) for J-shaped solutions, and to a non-differential equation (Eq (62)) for U-shaped solutions.

Remark 34 *In Proposition 38 it is proved that* $W^u > W^l$ *for* $s_0 > 0$.

7.2 U-shaped solutions for a rich or under-satiated agent with wealth $W \geq W^u$

If $W \geq W^u$ and $s_0 \geq 0$, the solution $[s, \lambda_c, \lambda_W]$ of Problem 10 can be constructed as follows:

$$s(t) = \begin{cases} s_0\,, & t = 0 \\ \sigma(t, \lambda_W), & t \in (0, T) \\ s_T(\lambda_W), & t = T, \end{cases} \tag{61}$$

where $\sigma(0, \lambda)$ is as described in Definition 19, $s(T) = s_T(\lambda_W)$, $s_T(\lambda_W)$ is a solution of Eq (57), and λ_W is to be determined from the wealth constraint (13) reduced to the non-differential equation

$$w^r(\lambda_W) = W, \tag{62}$$

where $w^r(\lambda_W)$ is defined as follows:

$$f \cdot w^r(\lambda_W) = e^{-r \cdot T} \cdot s_T(\lambda_W) - s_0 + (\gamma + r) \cdot \int_{[0,T]} e^{-r \cdot t} \cdot \sigma(t, \lambda_W) \cdot dt.$$

Therefore, the existence of solutions in this case reduces to solving Eq (62), and solvability of this equation is established in the following proposition:

Proposition 35 *If $W \geq W^u$, there is exactly one solution λ_W of Eq (62). In addition, the solution given by (61) is a solution of Problem 10 and also an optimal solution of Problem 1. If $W = W^u$, this solution coincides with the upper diagnostic profile; see Definition 31.*

Proof. We need to prove the existence of a unique solution of Eq (62), and we need to prove that the consumption gulps, $\frac{\sigma(0,\lambda_W)-s_0}{f}$ at $t = 0$ and $\frac{s_T-\sigma(T,\lambda_W)}{f}$ at $t = T$, are positive. In order to prove uniqueness, we find the derivative $\frac{d}{d\lambda_W}(w^r(\lambda_W))$:

$$f \cdot \frac{d}{d\lambda_W}(w^r(\lambda_W)) = e^{-r \cdot T} \cdot \frac{d}{d\lambda_W}(s_T(\lambda_W)) + (\gamma + r) \cdot \int_{[0,T]} e^{-r \cdot t} \cdot \frac{d}{d\lambda_W}(\sigma(t, \lambda_W)) \cdot dt$$

The derivative $\frac{d}{d\lambda_W}(s_T(\lambda_W))$ is calculated using Eq (57):

$$\frac{d}{d\lambda_W}(s_T(\lambda_W)) = \frac{1}{e^{-(n-r)\cdot T} \cdot \frac{d^2V}{ds^2}(s_T)} < 0. \tag{63}$$

The derivative $\frac{d}{d\lambda_W}(\sigma(t, \lambda_W))$ is calculated using Eq (35):

$$\frac{d}{d\lambda_W}(\sigma(t, \lambda_W)) = \frac{e^{(n-r)\cdot t} \cdot (\gamma + r)}{\frac{d^2V_S}{ds^2}(\sigma(t, \lambda_W))} < 0. \tag{64}$$

Thus

$$\frac{d}{d\lambda_W}(w^r(\lambda_W)) < 0 \tag{65}$$

This implies existence and uniqueness of a solution of $w(\lambda_W) = W$ with $\lambda_W \leq \lambda_W^u$ whenever $W \geq W^u$, since $\lim_{\lambda_w \to 0}(\sigma(t, \lambda_w)) = +\infty$ and $\lim_{\lambda_w \to 0}(s_T(\lambda_w)) = +\infty$. Positivity of the gulp of consumption at $t = T$ and the result that $\lambda_W > 0$ follow from Proposition 14. If $W = W^u$, the solution coincides with the upper diagnostic profile from Definition 31, thanks to the uniqueness proved earlier, and there is no gulp of consumption at $t = 0$. Since $\lambda_W < \lambda_W^u$, it follows

that if $W > W^u$, then (64) implies that $s_0 = \sigma(0, \lambda_W^u) < \sigma(0, \lambda_W)$, hence the gulp of consumption at $t = 0$ is positive.

The corresponding consumption is given by

$$c(t) = \delta_0(t) \cdot \frac{\sigma(0,\lambda) - s_0}{f} + \frac{r + \alpha_S \cdot \gamma - n}{\alpha_S} \cdot \sigma(t, \lambda_W) + \delta_T(t) \cdot \frac{s - \sigma(T,\lambda)}{f}$$

The solution $[s, \lambda_c, \lambda_W]$ of Problem 10 which is obtained in Proposition 35 is continuously differentiable in $(0, T)$, consumption is positive in all of $(0, T)$, there is a gulp of consumption at $t = T$, and there is a gulp of consumption at $t = 0$ if $W > W^u$. If $W = W^u$, this solution coincides with the upper diagnostic profile $[s^u, \lambda_c^u = 0, \lambda_W^u]$ given in Definition 31.

Remark 36 *If $s_0 = 0$, only this type of solution is possible and consumption is positive in all of $[0, T]$.*

7.3 J-shaped solutions for an agent with intermediate wealth ($W^l \leq W < W^u$)

If $W^l \leq W \leq W^u$ and $s_0 > 0$, the solution $[s, \lambda_c, \lambda_W]$ of Problem 10 can be constructed as follows:

$$s(t) = \begin{cases} s_0 \cdot e^{-\gamma \cdot t}, & t \in [0, T_s] \\ \sigma(t, \lambda_W), & t \in (T_s, T) \\ s_T, & t = T, \end{cases} \tag{66}$$

where $T_s \in [0, T]$, $\sigma(t, \lambda)$ is as described in Definition 19, and $s(T) = s_T = s_T(\lambda_W(T_s))$ is a solution of Eq (21), where $\lambda_W(T_s)$ is given by (67) below, while λ_c has to be found using Eq (21) for $t \in [0, T_s]$ with the terminal condition $\lambda_c(T_s) = 0$. By Lemma 24, the satiation $s(t)$ is differentiable in $[0, T_s) \cup (T_s, T)$ and continuous at T_s. In addition, it has a discontinuity at $t = T$ and no consumption in $[0, T_s)$. Continuity at $t = T_s$ means that $\sigma(T_s, \lambda_W) = s_0 \cdot e^{-\gamma \cdot T_s}$, which is equivalent to

$$\lambda_W(T_s) = \frac{e^{-(n-r)\cdot T_s}}{(\gamma + r)} \cdot \frac{dV_S}{ds}(s_0 \cdot e^{-\gamma \cdot T_s}) \tag{67}$$

by Lemma 17 and Definition 19. T_s must be determined using the wealth constraint (6), which, by a method similar to the one used to derive (13), can be transformed into

$$W = w^m(T_s), \tag{68}$$

where

$$f \cdot w^m(T_s) = e^{-r\cdot T} \cdot s_T(T_s) - s_0 \cdot e^{-(\gamma + r)\cdot T_s} + (\gamma + r) \cdot \int_{[T_s, T]} e^{-r\cdot t} \cdot \sigma(t, \lambda_W(T_s)) \cdot dt. \tag{69}$$

Therefore, the existence of solutions of Problem 10 in this case reduces to solving a non-differential equation (Eq (68)). The rest of this subsection is devoted to the proof of existence of solutions of Eq (68). We first prove a lemma which will allow us to prove the uniqueness of the solution of (68) and the fact that $W^u > W^l$.

Lemma 37 $\frac{d}{dT_s}(w^m(T_s)) < 0$

Proof. The derivative $\frac{d}{dT_s}(w^m(T_s))$ can be calculated as follows:

$$f \cdot \frac{d}{dT_s} w^m(T_s) = \frac{d}{dT_s}(\lambda_W(T_s)) \cdot \left[e^{-r \cdot T} \cdot \frac{d}{d\lambda_W}(s_T(\lambda_W)) + (\gamma + r) \cdot \int_{[T_s, T]} e^{-r \cdot t} \cdot \frac{d}{d\lambda_W}(\sigma(t, \lambda_W)) \cdot dt \right],$$

since $\sigma(T_s, \lambda_W(T_s)) = s_0 \cdot e^{-\gamma \cdot T_s}$. The expression in square brackets is negative, because (63) and (64) are applicable. The derivative $\frac{d}{dT_s}(\lambda_W(T_s))$ can be calculated using (67):

$$\frac{d}{dT_s}(\lambda_W(T_s)) = \frac{e^{-(n-r) \cdot T_s}}{(\gamma + r)} \cdot \frac{dV_S}{ds}(s_0 \cdot e^{-\gamma \cdot T_s}) \cdot (r - n + \gamma \cdot \alpha_s) > 0,$$

thanks to (40), hence $\frac{d}{dT_s}(w^m(T_s)) < 0$.

Proposition 38 *$W^u > W^l$ if $s_0 > 0$.*

Proof. If $W = W^u = w^m(0)$, the solution constructed in Proposition 39 coincides with the upper diagnostic profile $[s^u, \lambda_c^u = 0, \lambda_W^u]$ from Definition 31. If $W = W^l = w^m(T)$, it coincides with the lower diagnostic profile $[s^l, \lambda_c^l, \lambda_W^l]$ from Definition 29. However, $\frac{d}{dT_s}(w^m(T_s)) < 0$ for all $T_s \in (0, T)$ by Lemma 37, hence $W^u > W^l$.

Proposition 39 *If $s_0 > 0$ and $W^l \leq W \leq W^u$, there is exactly one solution λ_W of Eq (68). Thus the solution given in (61) is a solution of Problem 10 and also a maximizer of the functional in Problem 1.*

Proof. Existence follows from $w^m(T) = W^l \leq W \leq W^u = w^m(0)$ and the Intermediate Value Theorem, and uniqueness follows from $\frac{d}{dT_s}(w^m(T_s)) < 0$. The fact that $\lambda_c(t) > 0$ in $[0, T_s)$ has already been proved in Lemma 27.

The corresponding consumption is

$$c(t) = \left\{ \begin{array}{ll} 0, & t \in [0, T_s) \\ \frac{r + \alpha_S \cdot \gamma - n}{\alpha_S} \cdot \sigma(t, \lambda_W), & t \in (T_s, T) \end{array} \right\} + \delta_T(t) \cdot \frac{(s_T - \sigma(T, \lambda_W))}{f}$$

7.4 I-shaped solutions for a poor or over-satiated agent with wealth $W < W^l$

If $W < W^l$ and $s_0 > 0$, an explicit solution $[s, \lambda_c, \lambda_W]$ can be constructed as follows:

$$s(t) = \left\{ \begin{array}{ll} s_0 \cdot e^{-\gamma \cdot t}, & t \in [0, T) \\ s_T = f \cdot W \cdot e^{r \cdot T} + s_0 \cdot e^{-\gamma \cdot T}, & t = T, \end{array} \right.$$

where s_T can be determined using the wealth constraint (13):

$$\begin{aligned} f \cdot W &= e^{-r \cdot T} \cdot s_T - s_0 \cdot e^{-(\gamma + r) \cdot T} = e^{-r \cdot T} \cdot [s_T - s_0 \cdot e^{-\gamma \cdot T}] \\ s_T &= f \cdot W \cdot e^{r \cdot T} + s_0 \cdot e^{-\gamma \cdot T} > s_0 \cdot e^{-\gamma \cdot T}, \end{aligned} \tag{70}$$

and λ_W can be found using (21):

$$\lambda_W = e^{-(n-r)T} \cdot \frac{dV}{ds}(s_T) = e^{-(n-r)T} \cdot \frac{dV}{ds}(f \cdot W \cdot e^{r \cdot T} + s_0 \cdot e^{-\gamma \cdot T}),$$

while λ_c has to be found using Eq (21) and the terminal condition $\lambda_c(T) = 0$.

Proposition 40 $\lambda_c > 0$ *in* $[0, T)$.

Proof. By Lemma 27, it suffices to prove that condition (55) holds, but this is equivalent to $W < W^l$. Indeed, thanks to (59) and (70), the inequality $W < W^l$ is equivalent to $s_T < s_T^l$. Since $\frac{d^2V}{ds^2} < 0$, this is in turn equivalent to $\frac{dV}{ds}(s_T) > \frac{dV}{ds}(s_T^l)$, which implies that

$$\lambda_W = e^{-(n-r)\cdot T} \cdot \frac{dV}{ds}(s_T) > e^{-(n-r)\cdot T} \cdot \frac{dV}{ds}(s_T^l) = \frac{e^{-(n-r)\cdot T}}{(\gamma + r)} \cdot \frac{dV_S}{ds}(s_0 \cdot e^{-\gamma \cdot T}),$$

thanks to (57) and (58). This proves that condition (55) holds, and hence that $\lambda_c^l > 0$ in $[0, T)$ by Lemma 27.

The corresponding consumption is $c(t) = \delta_T(t) \cdot \frac{(s_T - s_0 \cdot e^{-\gamma \cdot T})}{f}$. There is no consumption in $[0, T)$, but satiation has a discontinuity at $t = T$, and all the consumption takes place in a gulp which is caused by this discontinuity if $W > 0$. If $W = W^l$, this solution coincides with the lower diagnostic profile $[s^l, \lambda_c^l = 0, \lambda_W^l]$ given in Definition 29.

8 Construction of explicit optimal solutions

In the previous section we described how to reduce the problem of finding optimal solutions of Problem 1 to solving non-differential equations. With the help of this theory, in this section we construct explicit solutions for logarithmic and CRRA utilities.

8.1 Logarithmic utility

In this section we solve the appropriate non-differential equations in the case of logarithmic utility: $V(s) = \ln(s)$, $\frac{dV(s)}{ds} = s^{-1}$, $\alpha = 1$. By (9), the derived utility is

$$V_S(s) = V(s) \cdot n + \gamma \cdot s \cdot \frac{dV}{ds}(s) = \ln(s) \cdot n + \gamma \cdot s \cdot s^{-1} = \ln(s) \cdot n + \gamma,$$

$\frac{dV_S}{ds}(s) = \frac{n}{s}$, $\alpha_S = 1$. We first establish several useful results. Conditions (32) and (40) coincide and reduce to $n < r + \gamma$. Here we discuss only the solutions that satisfy this condition.

8.1.1 The diagnostic profiles. Eq (35) for the diagnostic profile of satiation (see definition 29) and the boundary condition (21) for s_T take the following form:

$$0 = e^{-n\cdot T} \cdot \frac{dV}{ds}(s_T) - \lambda_W \cdot e^{-r\cdot T} = e^{-n\cdot T} \cdot \left[(s_T)^{-1} - \lambda_W \cdot e^{(n-r)\cdot T}\right].$$

Their solutions are

$$\sigma(t, \lambda_W) = \frac{1}{\lambda_W} \cdot \frac{n}{\gamma + r} \cdot e^{-(n-r)\cdot t}, \quad s_T = \frac{1}{\lambda_W} \cdot e^{-(n-r)\cdot T} \tag{71}$$

From the initial condition $s_0 = \sigma(0, \lambda_W) = (\lambda_W^u)^{-1} \cdot \frac{n}{\gamma + r}$, we obtain $\lambda_W^u = \frac{n}{(\gamma + r)\cdot s_0}$, hence

$$s_T^u = \frac{1}{\lambda_W^u} \cdot e^{-(n-r)\cdot T} = s_0 \cdot \frac{\gamma + r}{n} \cdot e^{-(n-r)\cdot T}.$$

Note that $\lambda_c^u = 0$, since consumption $c(t) > 0$ in all of $[0, T]$, hence the upper diagnostic profile $s^u(t)$ from Definition 31 is

$$s^u(t) = \left\{ \begin{array}{ll} \sigma(t, \lambda_W^u) = s_0 \cdot e^{-(n-r)\cdot t}, & t \in [0, T) \\ s_T^u = s_0 \cdot \frac{\gamma + r}{n} \cdot e^{-(n-r)\cdot T}, & t = T \end{array} \right\}, \ \lambda_c^u = 0 \text{ in } [0, T]$$

The corresponding consumption is

$$\begin{aligned} c^u(t) &= \frac{r+\alpha_S\cdot\gamma-n}{\alpha_S}\cdot\sigma\left(t,\lambda_W^u\right)+\delta_T(t)\cdot\frac{\left(s_T^u-\sigma(T,\lambda_W^u)\right)}{f} \\ &= s_0\cdot(r+\gamma-n)\cdot e^{-(n-r)\cdot t}+\delta_T(t)\cdot s_0\cdot\frac{\gamma+r-n}{n\cdot f}\cdot e^{-(n-r)\cdot T} \end{aligned}$$

By (60), the wealth necessary for s^u is

$$\begin{aligned} f\cdot W^u &= e^{-r\cdot T}\cdot s_T^u-s_0+\int_{[0,T]}e^{-r\cdot t}\cdot(\gamma+r)\cdot\sigma(t,\lambda_W^u)\cdot dt \\ &= s_0\cdot\left(\frac{\gamma+r}{n}-1\right) \\ &= s_0\cdot\frac{\gamma+r-n}{n} \end{aligned}$$

By Eqs (58) and (57), s_T is a solution of $e^{-(n-r)T}\cdot\frac{dV}{ds}\left(s_T^l\right)=\lambda_W^l=\frac{e^{-(n-r)T}}{(\gamma+r)}\cdot\frac{dV_S}{ds}\left(s_0\cdot e^{-\gamma\cdot T}\right)$, hence $s_T^l=s_0\cdot e^{-\gamma\cdot T}\cdot\frac{(\gamma+r)}{n}$. Then the lower diagnostic profile from Definition 29 is

$$s^l(t)=\begin{cases} s_0\cdot e^{-\gamma\cdot t}, & t\in[0,T) \\ s_T=s_0\cdot e^{-\gamma\cdot T}\cdot\frac{(\gamma+r)}{n}, & t=T \end{cases}$$

The corresponding consumption is $c^l(t)=\delta_T(t)\cdot\frac{\left(s_T^l-s_0\cdot e^{-\gamma\cdot T}\right)}{f}=\delta_T(t)\cdot s_0\cdot e^{-\gamma\cdot T}\cdot\frac{\gamma+r-n}{n\cdot f}$. By (59), the wealth necessary for s^l is

$$f\cdot W^l=e^{-r\cdot T}\cdot\left(s_T-s_0\cdot e^{-\gamma\cdot T}\right)=s_0\cdot\frac{\gamma+r-n}{n}\cdot e^{-(\gamma+r)\cdot T}=f\cdot W^u\cdot e^{-(\gamma+r)\cdot T}<f\cdot W^u$$

Here we will analyze only two cases: that of a normally satiated agent and that of a rich or over-satiated agent. These cases were discussed for general utilities in Section 7.4.

8.1.2 J-shaped solutions for an agent with intermediate wealth ($W^l\le W<W^u$). If $W^l\le W\le W^u$ and $s_0>0$, the solution $[s,\lambda_c,\lambda_W]$ is

$$s(t)=\begin{cases} s_0\cdot e^{-\gamma\cdot t}, & t\in[0,T_s] \\ \frac{1}{\lambda_W}\cdot e^{-(n-r)\cdot t}\cdot\frac{n}{\gamma+r}=s_0\cdot e^{-\gamma\cdot T_s}\cdot e^{-(n-r)\cdot(t-T_s)}, & t\in(T_s,T) \\ \frac{1}{\lambda_W}\cdot e^{-(n-r)\cdot T}=s_0\cdot e^{-\gamma\cdot T_s}\cdot e^{-(n-r)\cdot(T-T_s)}\cdot\frac{\gamma+r}{n}, & t=T, \end{cases}$$

where $\frac{1}{\lambda_W}=s_0\cdot e^{-\gamma\cdot T_s}\cdot e^{(n-r)\cdot T}\cdot\frac{\gamma+r}{n}$ and T_s is to be determined from the wealth constraint (68), hence $f\cdot W=s_0\cdot e^{-(\gamma+r)\cdot T_s}\cdot\frac{\gamma+r-n}{n}$. We deem the above to be an equation for T_s, hence the solution is $T_s=\frac{\ln\left(\frac{W^u}{W}\right)}{(\gamma+r)}=\frac{\ln\left(\frac{s_0}{W}\frac{\gamma+r-n}{f\cdot n}\right)}{(\gamma+r)}$. The corresponding consumption is

$$\begin{aligned} c(t) &= \left\{\begin{matrix} 0, & t\in[0,T_s) \\ s_0\cdot\frac{r+\gamma-n}{n}\cdot e^{-\gamma\cdot T_s}\cdot e^{-(n-r)\cdot(t-T_s)}, & t\in(T_s,T) \end{matrix}\right\} \\ &\quad+\delta_T(t)\cdot s_0\cdot\frac{r+\gamma-n}{f\cdot n}\cdot e^{-\gamma\cdot T_s}\cdot e^{-(n-r)\cdot(T-T_s)} \end{aligned}$$

8.1.3 U-shaped solutions for a rich or under-satiated agent with wealth $W \geq W^u$. If $W \geq W^u$ and $s_0 \geq 0$, the solution $[s, \lambda_c, \lambda_W]$ is

$$s(t) = \left\{ \begin{array}{ll} s_0\,, & t = 0 \\ s_T \cdot \frac{n}{\gamma+r} \cdot e^{(n-r)\cdot(T-t)}, & t \in (0,T) \\ s_T\,, & t = T \end{array} \right\}, \quad \lambda_c = 0 \text{ in } [0,T],$$

where s_T is to be determined from the wealth constraint (62) $f \cdot W + s_0 = s_T \cdot e^{(n-r)\cdot T}$, hence

$$s(t) = \begin{cases} s_0\,, & t = 0 \\ (f \cdot W + s_0) \cdot \frac{n}{\gamma+r} \cdot e^{-(n-r)\cdot t}, & t \in (0,T) \\ (f \cdot W + s_0) \cdot e^{-(n-r)\cdot T}, & t = T \end{cases}$$

The corresponding consumption is

$$\begin{aligned} c(t) &= \delta_0(t) \cdot \frac{(f \cdot W + s_0) \cdot \frac{n}{\gamma+r} - s_0}{f} \\ &+ (r+\gamma-n) \cdot (f \cdot W + s_0) \cdot e^{-(n-r)\cdot t} \cdot \frac{n}{\gamma+r} \\ &+ \delta_T(t) \cdot (f \cdot W + s_0) \cdot e^{-(n-r)\cdot T} \cdot \frac{\gamma+r-n}{f \cdot (\gamma+r)} \end{aligned}$$

8.2 Explicit optimal solutions for CRRA utility

In Section 7 we described how to reduce the problem of finding optimal solutions of Problem 1 to solving non-differential equations. In this section, we solve these equations for the case of CRRA utility $V(s) = \frac{s^{1-\alpha}-1}{1-\alpha}$, $\frac{dV(s)}{ds} = s^{-\alpha}$, where α is a constant, $\alpha \neq 1$. In this case we can use Eq (9) to find the derived utility V_S:

$$\begin{aligned} V_S &= \frac{n+\gamma\cdot(1-\alpha)}{(1-\alpha)} \cdot s^{1-\alpha} - \frac{n}{1-\alpha}, \\ \frac{dV_S}{ds} &= (n+\gamma\cdot(1-\alpha)) \cdot \frac{dV}{ds} \\ &= (n+\gamma\cdot(1-\alpha)) \cdot s^{-\alpha}. \end{aligned}$$

We first establish several useful results. Conditions (32) and (40) coincide and reduce to $n < r + \gamma \cdot \alpha$. The requirements on V_S and α_S in Remark 4 are satisfied. In addition, we need to assume that $n - r \cdot (1-\alpha) > 0$. Eq (35) for the diagnostic profile of satiation and the boundary condition (21) on s_T take the following form:

$$\begin{aligned} \frac{dV_S}{ds}(\sigma(t)) &= (n+\gamma\cdot(1-\alpha)) \cdot \frac{dV}{ds}(\sigma(t)) \\ &= (n+\gamma\cdot(1-\alpha)) \cdot (\sigma(t))^{-\alpha} \\ &= \lambda_W \cdot (\gamma+r) \cdot e^{(n-r)\cdot t}, \\ 0 &= e^{-n\cdot T} \cdot \frac{dV}{ds}(s_T) - \lambda_W \cdot e^{-r\cdot T} \\ &= e^{-n\cdot T} \cdot [(s_T)^{-\alpha} - \lambda_W \cdot e^{(n-r)\cdot T}] \end{aligned}$$

Their solutions are $\sigma(t, \lambda_W) = (\lambda_W^{-1} \cdot Z)^{\frac{1}{\alpha}} \cdot e^{-\frac{n-r}{\alpha}t}$, $s_T = \lambda_W^{-\frac{1}{\alpha}} \cdot e^{-\frac{n-r}{\alpha}T} = \sigma(T, \lambda_W) \cdot Z^{-\frac{1}{\alpha}}$, where

$Z = \frac{(n+\gamma\cdot(1-\alpha))}{(\gamma+r)}$, hence we can also write

$$\sigma(t, s_T) = s_T \cdot Z^{\frac{1}{\alpha}} \cdot e^{\frac{n-r}{\alpha}\cdot(T-t)} \tag{72}$$

8.2.1 The diagnostic profiles. The upper diagnostic profile $s^u(t)$ from Definition 31 is

$$s^u(t) = \left\{ \begin{array}{ll} \sigma(t, \lambda_W) = s_0 \cdot e^{-\frac{n-r}{\alpha}\cdot t}, & t \in [0, T) \\ s_T^u = s_0 \cdot Z^{-\frac{1}{\alpha}} \cdot e^{-\frac{n-r}{\alpha}T}, & t = T \end{array} \right\}, \ \lambda_c^u = 0 \text{ in } [0, T]$$

The corresponding consumption is $c^u(t) = \frac{r+\alpha\cdot\gamma-n}{\alpha} \cdot s_0 \cdot e^{-\frac{n-r}{\alpha}\cdot t} + \delta_T(t) \cdot \frac{s_0}{f} \cdot e^{-\frac{(n-r)}{\alpha}\cdot T} \cdot \left(Z^{-\frac{1}{\alpha}} - 1\right)$. By (60), the wealth necessary for s^u is

$$f \cdot W^u = s_0 \cdot \left[\frac{\gamma\cdot\alpha + r - n}{n + r\cdot\alpha - r} + \left(Z^{-\frac{1}{\alpha}} - \alpha \cdot \frac{\gamma + r}{n - r\cdot(1-\alpha)}\right) \cdot e^{-\left(r+\frac{n-r}{\alpha}\right)\cdot T}\right]$$

The lower diagnostic profile from Definition 29 is

$$s^l(t) = \left\{ \begin{array}{ll} s_0 \cdot e^{-\gamma\cdot t}, & t \in [0, T) \\ s_T^l = s_0 \cdot e^{-\gamma\cdot T} \cdot Z^{-\frac{1}{\alpha}}, & t = T \end{array} \right\}, \ \lambda_W^l = s_0^{-\alpha} \cdot e^{(\gamma\cdot\alpha+r-n)\cdot T} \cdot Z^{-1}$$

The corresponding consumption is $c^l(t) = \delta_T(t) \cdot s_0 \cdot \frac{e^{-\gamma\cdot T}}{f} \cdot \left(\frac{\gamma\cdot\alpha+r-n}{(n+\gamma\cdot(1-\alpha))} + Z^{-\frac{1}{\alpha}} - Z^{-1}\right)$. By (59), the wealth necessary for s^l is

$$f \cdot W^l = e^{-r\cdot T} \cdot \left(s_0 \cdot e^{-\gamma\cdot T} \cdot Z^{-\frac{1}{\alpha}} - s_0 \cdot e^{-\gamma\cdot T}\right) = f \cdot W^u \cdot e^{-(\gamma+r)\cdot T} < f \cdot W^u$$

Below, we discuss the cases of a normally satiated agent and a rich or under-satiated agent. The case of a poor or over-satiated agent was solved for general utilities in Section 7.4.

8.2.2 J-shaped solutions for an agent with intermediate wealth ($W^l \le W < W^u$). If $W^l \le W \le W^u$ and $s_0 > 0$, the solution $[s, \lambda_c, \lambda_W]$ is

$$s(t) = \left\{ \begin{array}{ll} s_0 \cdot e^{-\gamma\cdot t}, & t \in [0, T_s] \\ s_0 \cdot e^{-\gamma\cdot T_s} \cdot e^{-\frac{(n-r)}{\alpha}\cdot(t-T_s)}, & t \in (T_s, T) \\ s_0 \cdot e^{-\gamma\cdot T_s} \cdot e^{-\frac{(n-r)}{\alpha}\cdot(T-T_s)} \cdot Z^{-\frac{1}{\alpha}}, & t = T, \end{array} \right.$$

where T_s is to be determined using the wealth constraint (68), which can be transformed into

$$f \cdot W = s_0 \cdot e^{-(\gamma+r)\cdot T_s} \cdot \left[\frac{r + \gamma\cdot\alpha - n}{(n - r\cdot(1-\alpha))} + e^{-\left(r+\frac{n-r}{\alpha}\right)\cdot(T-T_s)} \cdot E_2\right],$$

and $E_2 = \left(\left(\frac{\gamma+r}{n+\gamma\cdot(1-\alpha)}\right)^{\frac{1}{\alpha}} - \alpha \cdot \frac{\gamma+r}{n-r\cdot(1-\alpha)}\right)$. Note that $E_2 = 0$ when $\alpha = 1$. Thus we can use (13) to obtain an equation for T_s:

$$f \cdot W = s_0 \cdot e^{-(\gamma+r)\cdot T_s} \cdot \left[\frac{r + \gamma\cdot\alpha - n}{n - r\cdot(1-\alpha)} + e^{-\left(r+\frac{n-r}{\alpha}\right)\cdot(T-T_s)} \cdot E_2\right]$$

The corresponding consumption is

$$c(t) = s_0 \cdot e^{-\gamma \cdot T_s} \cdot \left[\left\{ \begin{array}{ll} 0, & t \in [0, T_s) \\ \frac{r+\alpha\cdot\gamma-n}{\alpha} \cdot e^{-\frac{(n-r)}{\alpha}(t-T_s)}, & t \in (T_s, T) \end{array} \right\} + \delta_T(t) \cdot e^{-\frac{(n-r)}{\alpha}(T-T_s)} \cdot \frac{1}{f} \cdot \left(Z^{-\frac{1}{\alpha}} - 1 \right) \right]$$

8.2.3 U-shaped solutions for a rich or under-satiated agent with wealth $W \geq W^u$. If $W \geq W^u$ and $s_0 \geq 0$, it is better to use (72) and write

$$s(t) = \left\{ \begin{array}{ll} s_0\,, & t = 0 \\ s_T \cdot Z^{\frac{1}{\alpha}} \cdot e^{+\frac{n-r}{\alpha}(T-t)}, & t \in (0, T) \\ s_T\,, & t = T \end{array} \right\}, \quad \lambda_c = 0 \text{ in } [0, T],$$

where $s_T = (f \cdot W + s_0) \cdot B^{-1} \cdot e^{-\frac{n-r}{\alpha}T}$ is determined using the wealth constraint (62), $B = 1 - \left(1 - e^{-\left(r+\frac{\delta-r}{\alpha}\right)\cdot T}\right) \cdot E_3$ (note that $B = 1$ when $\alpha = 1$), and $E_3 = 1 - Z^{\frac{1}{\alpha}} \cdot (\gamma + r) \cdot \left(r + \frac{n-r}{\alpha}\right) = Z^{\frac{1}{\alpha}} \cdot E_2$, hence

$$s(t) = \left\{ \begin{array}{ll} s_0\,, & t = 0 \\ (f \cdot W + s_0) \cdot B^{-1} \cdot Z^{\frac{1}{\alpha}} \cdot e^{-\frac{n-r}{\alpha}t}, & t \in (0, T) \\ (f \cdot W + s_0) \cdot B^{-1} \cdot e^{-\frac{n-r}{\alpha}T}, & t = T \end{array} \right.$$

The corresponding consumption is

$$\begin{aligned} c(t) &= \delta_0(t) \cdot \frac{1}{f} \cdot \left((f \cdot W + s_0) \cdot B^{-1} \cdot Z^{\frac{1}{\alpha}} - s_0 \right) \\ &+ \frac{r + \alpha \cdot \gamma - n}{\alpha} \cdot (f \cdot W + s_0) \cdot B^{-1} \cdot Z^{\frac{1}{\alpha}} \cdot e^{-\frac{n-r}{\alpha}t} \\ &+ \delta_T(t) \cdot \frac{1}{f} \cdot (f \cdot W + s_0) \cdot B^{-1} \cdot e^{-\frac{n-r}{\alpha}T} \cdot \left(1 - Z^{\frac{1}{\alpha}}\right) \end{aligned}$$

9 Conclusions and closing remarks

We investigated the structure and existence of solutions of the problem of consumption with satiation in continuous time for general utilities. We proved that the satiation function $s(t)$ cannot be continuously differentiable, hence we assumed s to be a function of bounded variation. We proved that $\frac{ds}{dt}$ and c must be continuous with respect to Lebesgue measure in $(0, T)$ but that there may be a gulp of consumption at either or both endpoints. Since such gulps take place in continuous time, they are not artifacts of discrete time and could be observed experimentally.

We described how to construct optimal solutions.

In Section 6.1, we considered solutions under the assumptions of a high discount rate (assumptions (33) and (41)). This happens in societies characterized by high conflict. The solution we obtained for that case can be described as follows: It is optimal to consume all of one's wealth at once and not wait, no matter what one's utility function is. Indeed, in societies with

low safety levels, when life expectancy is short (e.g., during famines in ancient times), or when an agent expects to lose his wealth to robbers, it seems reasonable to consume all of the wealth at once. This conclusion is independent of the time horizon.

The analysis of consumption with satiation when the discount rate is low (conditions (32) and (40)) is more complicated. It turns out that a gulp of consumption at the end of the consumption period is always optimal (see Proposition 14), because no further penalty can be imposed by the growth of satiation. In fact, there exists anecdotal evidence of addicts having final binges before going into rehab (as in the movie "The Wonderful Whites of Virginia"), smokers taking the last puff before quitting (as in the movie "A Good Woman"), and death-row inmates accepting the offer of the last wish. According to [19], empirical data indicate that such gulps of consumption take place when people consume cultural goods, such as music. Detailed analysis in Section 7.4 shows that when the wealth to be consumed is small and/or satiation is high, the agent consumes all of his wealth at the very end of the consumption period. When the wealth is large and/or satiation is low, the agent consumes some portion in a gulp of consumption at the beginning of his consumption period, as shown in Section 6.2. This can be explained as follows: Wealthy agents consume a lot at the beginning in order to curtail their desires by increasing their satiation level, and then they consume at a moderate rate throughout the rest of their consumption period. They also leave some wealth to be consumed at the very end. There is also an intermediate pattern, where an agent begins consumption after some delay and then enjoys a gulp of consumption at the very end of his consumption interval. Although the terminal gulp of consumption is non-negotiable in our model, we rarely observe it in the real world. This is especially true when non-cultural goods are consumed. A possible explanation is that agents are not aware of the exact time at which their consumption life will end, and as a result they optimize beyond the actual time of termination of their consumption. The finding regarding the optimality of the terminal gulp of consumption may explain the large sums some people spend on funerals and funerary monuments. One extreme example of such behavior is that of Caterina Campodonico, who lived in Italy in the 19th century and is said to have saved for most of her life in order to afford an elaborate funerary monument in the famous Staglieno Cemetery in Genoa; see [20]. Leaving an inheritance to descendants could also be interpreted as a gulp of consumption at the end of the consumption period.

Author Contributions

Conceptualization: Peter Smoczynski, Stan Miles.

Formal analysis: Peter Smoczynski, Stan Miles.

Writing – original draft: Peter Smoczynski, Stan Miles.

Writing – review & editing: Peter Smoczynski, Stan Miles.

References

1. Baucells M, Sarin RK. Satiation in discounted utility. Oper. Res. 2007; 55(1): 170–181. https://doi.org/10.1287/opre.1060.0322
2. Baucells M, Sarin RK. Predicting utility under satiation and habit formation, Manage Sci. 2010; 56: 286–301. https://doi.org/10.1287/mnsc.1090.1113
3. Baucells M, Sarin R. Engineering Happiness. Los Angeles: University of California Press; 2012.
4. Baucells, M. Satiation Preferences—the case of certainty; 2015 [cited 2019 March 10]. Database: SSRN [Internet]. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3321802
5. Baucells M, Zhao L. It is time to get some rest. Manage Sci. Forthcoming.

6. Pontryagin LS, Boltyanskii VG, Gamkrelidze RV, Mishchenko EF. The mathematical theory of optimal processes. New York, NY: John Wiley and Sons, Inc.; 1962.
7. Karamzin DY, de Oliveira VA, Pereira FL, Silva GN. (2014). On some extension of optimal control theory. Eur J Control. 2014; 20: 284–291. https://doi.org/10.1016/j.ejcon.2014.09.003
8. Silva GN, Vinter RB. Necessary conditions for optimal impulsive control problems. SIAM J Control Optim. 1997; 35: 1829–1846. https://doi.org/10.1137/S0363012995281857
9. Miller BM, Rubinovich EY. Impulsive control in continuous and discrete-continuous systems. New York, NY: Springer; 2003.
10. Rishel RW. An extended Pontryagin principle for control systems whose control laws contain measures. J Soc Ind Appl Math, Ser A: Control. 1965; 3: 191–205. https://doi.org/10.1137/0303016
11. Motta M, Rampazzo F. (1996). Dynamic programming for nonlinear systems driven by ordinary and impulsive controls. SIAM J Control Optim. 1996; 34: 199–225. https://doi.org/10.1137/S036301299325493X
12. Boyd S, Vandenberghe L. Convex optimization. Cambridge: Cambridge University Press; 2004.
13. Caputo MR. Foundations of dynamic economic analysis. Cambridge: Cambridge University Press; 2005.
14. Zaslavski A. (2008). Generic nonoccurrence of the Lavrentiev phenomenon for a class of optimal control problems. J Dyn Control Syst. 2008; 14: 95–119. https://doi.org/10.1007/s10883-007-9032-6
15. Dani K, Hrusa WJ, Mizel VJ. Lavrentiev's phenomenon for totally unconstrained variational problems in one dimension. Nonlinear Differ Equ Appl. 2000; 7: 435–446. https://doi.org/10.1007/PL00001434
16. Vol'pert AI, Hudjaev SI. Analysis in classes of discontinuous functions and equations of mathematical physics. Boston, MA: Martinus Nijhoff Publishers; 1985.
17. Hudjaev SI, Vol'pert AI. Analysis in classes of discontinuous functions and equations of mathematical physics, volume 8 of Mechanics: Analysis. Boston, MA: Martinus Nijhoff Publishers; 1985.
18. Bank P, Riedel F, Non-time additive utility optimization—the case of certainty. J Math Econ. 2000; 33 (3): 271–290. https://doi.org/10.1016/S0304-4068(99)00026-9
19. Smith DP. Intertemporal utility in continuous time: Theoretical foundations and empirical validation. Ph. D. Thesis, Heidelberg University. 2014.
20. Campodonico Memorial, 1881. [cited 24 April 2018]. In: American Friends of Italian Monumental Sculpture. Available from: https://staglieno.com/campodonico-memorial-1881.

New nonlinear model of population growth

Badr Saad T. Alkahtani[1], Abdon Atangana[2]*, Ilknur Koca[3]

1 Department of Mathematical, Colle of Science, King Saud University, P.O.Box 1142, Riyadh, 11989, Saudi Arabia, **2** Institute for Groundwater Studies, Faculty of Natural and Agricultural Sciences University of the Free State, Bloemfontein 9300, South Africa, **3** Department of Mathematics, Faculty of Sciences, Mehmet Akif Ersoy University, 15100, Burdur, Turkey

* abdonatangana@yahoo.fr

Abstract

The model of population growth is revised in this paper. A new model is proposed based on the concept of fractional differentiation that uses the generalized Mittag-Leffler function as kernel of differentiation. The new model includes the choice of sexuality. The existence of unique solution is investigated and numerical solution is provided.

Editor: Jun Ma, Lanzhou University of Technology, CHINA

Introduction

Researchers within the field of biology and mathematical biology are interested to know whether or not the certain specie will be instinct or not. This study has fascinated many researchers around the world in recent passed years. For instance to control the spread of a given infectious diseases researchers are interested in their reproductive number, that help to know whether or not the disease will be extinct [1-7]. However if the model is accurate enough they can give reliable predictions, if the predictions show the extinction of a given species, then laws-makers can take some decisions to protect the specie. We can find many examples of this in developed countries, for instance in South Africa, the government gave strict law against the killing of rhinos. In China, we have the protection of the tigers. A global protection of whale in all oceans and Africans elephants that are nowadays consider as rare species [4-7]. In case of infectious diseases, the aim is to end the spread of the virus that can considered as a specie, in this case also the control can be done via mathematical predictions. It is therefore important that in both cased the mathematical formulas should be able to portray more accurately the dynamic of the specie in time [8-9]. Generally speaking mathematical models allows a better thoughtful of how the complex interfaces and processes work. Indeed exhibiting of dynamic interactions in nature can provide a manageable way of understanding how numbers alter in excess of time or relation to each other. The aim of this paper is to provide a new model that will be able to describe the population growth more accurately.

Funding: The authors extend their sincere appreciations to the Deanship of Science Research at King Saud University for funding this prolific research group PRG-1437-35.

Competing interests: The authors have declared that no competing interests exist.

Model of population growth

Biological population demonstrating is worried with the changes in populace size and age spreading within a population as a significance of collaborations of creatures with the physical setting, with individuals of their own species, and with bacteria of other kinds. The biosphere is full of interfaces that varies from modest to dynamic. Earth's processes affect human life and

are momentously stochastic and seem disordered to naked eye. Nevertheless, a embarrassment of patterns can be perceived and are brought forth by using inhabitants demonstrating as an instrument. Population reproductions are employed to control maximum fruitage for agriculturists, to comprehend the changing aspects of biotic annexations, and have plentiful environmental safeguarding insinuations [8-9]. Thomas Malthus was one of the former to demonstrate that, inhabitants evolved with symmetrical configuration despite the fact that envisioning the providence of humankind [8-9]. Nonetheless, Nurgaliev's law is a mathematical equation that portrays the rate of change of proportions of a population at a given time, in terms of the current population size. It is a deterministic conventional discrepancy equation in which the rate change is articulated as a quadratic function of the population size and this equation is given as:

$$\frac{dn(t)}{dt} = an^2(t) - bn(t) \tag{1}$$

In this equation, the size of a population is denoted by n time is measured in years, a is half of the average probability of a birth of male also for females, of a potential arbitrary parents pair within a is year. b is an average probability of a death of a person within a year. The above model has some limitations, the variation of growth of population is an average between two given times, which is not naturally true because the averaging is not the same at the different interval of time. The second problem is that the model does not take into account the choice of sex and also the memory effect. In this work we shall introduce new parameters to the model and also consider the memory effect induces by the fractional differentiation based on the Mittag-Leffler function.

Fractional differentiation

The topic of fractional differentiation is one of the hot topics nowadays in almost all the fields of science, technology and engineering due to its wide applicability and also its ability of model real world problems more accurately than the classical differentiation. The first definition was proposed by Riemann and Liouville is given below as [10-14]

$${}^{RL}D_t^{\alpha}[f(t)] = \frac{1}{\Gamma(1-\alpha)}\frac{d}{dt}\int_0^t (t-y)^{-\alpha} f(y)dy, \;\; 0 < \alpha \leq 1. \tag{2}$$

Caputo when working with a real world problem later modified this definition, as he was unable to recover the usual initial conditions, then Caputo modified Eq (2) by putting the derivative inside the integral to obtain:

$${}^{C}D_t^{\alpha}[f(t)] = \frac{1}{\Gamma(1-\alpha)}\int_0^t (t-y)^{-\alpha}\frac{d}{dy} f(y)dy, \;\; 0 < \alpha \leq 1. \tag{3}$$

Definition (2) and (3) have been intensively used and misused in the last decades in many fields. However, when looking at the definition, we can see that, they are convolution of functions and the power law, or the kernel of transformation is the well-known power law. However, it is clear when looking at the behaviour of some physical phenomena that, not all of them follow the power law. Recently Caputo and Fabrizio suggested a step a head in fractional differentiation when they replaced the power law with exponential decay law as presented

below [10-14]

$$ {}^{CF}_{0}D^{\alpha}_{t}[f(t)] = \frac{M(\alpha)}{1-\alpha}\int_{0}^{t}\frac{d}{dy}f(y)\exp\left[-\frac{\alpha}{1-\alpha}(t-y)\right]dy,\ \ 0<\alpha<1. \tag{4} $$

And Goufo and Atangana proposed the modified version in several research papers and it is given as follows [10-14]

$$ {}^{CFR}_{0}D^{\alpha}_{t}[f(t)] = \frac{M(\alpha)}{1-\alpha}\frac{d}{dt}\int_{0}^{t}f(y)\exp\left[-\frac{\alpha}{1-\alpha}(t-y)\right]dy,\ \ 0<\alpha<1. \tag{5} $$

But their proposition was rejected due to the criteria that need to be satisfied for an operator to be called fractional derivative. However their idea was great because a new kernel was introduced with no singularity. Atangana and Baleanu, to solve the problem in Caputo and Fabrizio operator, they suggested a new kernel based on the generalized Mittag-Leffler function, that is the more suitable function that was introduced to solve some problems of disc of convergence of power law. The function is also considered as the queen of fractional calculus and is more natural than power law. Their definitions are given below as follow [15-20]:

$$ {}^{ABC}_{0}D^{\alpha}_{t}[f(t)] = \frac{B(\alpha)}{1-\alpha}\int_{0}^{t}\frac{d}{dy}f(y)E_{\alpha}\left[-\frac{\alpha}{1-\alpha}(t-y)^{\alpha}\right]dy,\ \ 0<\alpha<1, \tag{6} $$

Also

$$ {}^{ABR}_{0}D^{\alpha}_{t}[f(t)] = \frac{B(\alpha)}{1-\alpha}\frac{d}{dt}\int_{0}^{t}f(y)E_{\alpha}\left[-\frac{\alpha}{1-\alpha}(t-y)^{\alpha}\right]dy,\ \ 0<\alpha<1. \tag{7} $$

This new version was applied in the theory of Chaos with great success therefore in this paper, we make use of this fractional differentiation to provide new model of population growth.

New model of population growth

Many physical observed facts are said to follow the power law evolution. The use of fractional differentiation to predict the population growth was investigated before with Caputo power law fractional derivative in the following [20-23]. More importantly the expansion of mankind of population growth is obviously one of those one of those. However the chose of power law used to model such dynamical system must be chosen with care. The growth of population does take place exponentially as indicated by several classical models, or this does not take place with the trend of power law of x^{α} (the power law population growth can be observed in less developed countries were the rate of birth is very high). Additionally this does not occur with only a fading local memory (the real world situation for fading memory population can be found in developed countries where the rate of birth is very small as time goes on) as in the process of the diffusion within a porous media but rather combine both fading memory and power law. The only fractional operator that can with care and accurately replicate this dynamical system is perhaps the fractional differentiation with generalized Mittag-Leffler kernel. In this paper to accurately include into mathematical formulation the effect of fading memory and also power law, we convert the classical derivative with Atangana-Baleanu fractional derivative, which takes into account the power-law population growth together with fading local memory population growth. In this section, a new model of population growth is suggested using the concept of fractional, in addition to this, a new parameter taking into account the choice of partner will be introduced to well represent the physical investigation into mathematical formulas. Let assume $N(t)$ to be the size of density population in a given period of time. Let p be the probability of an adult to choose a partner with same sex then the following

equation is suggested:

$$ {}^{ABC}_{0}D^{\alpha}_{t}N(t) = aN^2(t) - bN(t) - (1-p)v(t)N^2(t) \tag{8} $$

The new function $v(t)$ is the selection function that a given individual will be convince to choose a partner with same sex. The new model induces also the memory effect due to the fractional differentiation.

We shall first present the equilibrium points of this dynamical system. To obtain them, we assume that the function is time independent therefore Eq (8) is transformed to

$$ \begin{aligned} aN^2 - bN - (1-p)vN^2 &= 0 \\ N &= 0,\ N = \frac{b}{a-(1-p)v} \\ a &\neq (1-p)v \end{aligned} \tag{9} $$

Therefore the realistic equilibrium point is when the proportion of the rate of death with the difference between the birth contribution and the factor of choice of partner. However if the following inequality holds $a-(1-p)v<0$ then mankind specie will die out. If $a=(1-p)v$ then in a near future also mankind will vanish. If the quantity is big enough then mankind will survive.

Existence of solution

The conditions within which the new equation admits a positive solution will be discussed in this section. To do this, we consider $X = C[a, b]$ the Banach space of every continuous real functions defined in the closed set $[a, b]$, which bestowed with the sub norm and Z be the shaft defined as: $Z = \{N \in X, N(t) \geq 0, a \leq t \leq b\}$. We shall present the following Banach fixed-point theorem that will be used for the existence demonstration.

Definition 1: Let E be a real Banach space with a cone H. H initiates a restricted order $\leq$ in E in the succeeding approach [18]

$$ x \leq y \Rightarrow y - x \in H. $$

For every $x, y \in E$ the order interval is defined as $\langle a, b\rangle = \{f \in E: a \leq f \leq b\}$. A cone K is denoted normal, if one can find a positive constant j such that $h, d \in K, \Phi < h < d \Rightarrow \|h\| \leq j\|d\|$, where Φ denotes the zero element of K.

Theorem 1 [18]: Let H be a closed set subspace of a Banach space of D. let G be a contraction mapping with Lipschitz constant $g < 1$ from H to H. Thus G possesses a fixed-point t^* in H. In addition, if t_0 is a random point in H and $\{t_n\}$ is a sequence defined by $t_{n+1} = Gt_n (n = 0, 1, 2\ldots)$, then for a large number n, t_n tends to t^* in H and $d(t_n, t^*) \leq \frac{g^n}{(1-g)} d(t_1, t_0)$.

We present also some properties of Atangana-Baleanu derivative in Caputo sense.

Theorem 2 [19]: Let $f(t) \in H^1(a, b)$, $b > a$ such that the Atangana-Baleanu fractional derivative exists, then the following relationship holds:

$$ {}^{AB}_{0}I^{\alpha}_{t}\{{}^{ABR}_{0}D^{\alpha}_{t}f(t)\} = f(t) \tag{10} $$

$$ {}^{AB}_{0}I^{\alpha}_{t}\{{}^{ABC}_{0}D^{\alpha}_{t}f(t)\} = f(t) - f(0) \tag{11} $$

Proof: By definition we establish the above relation (10) using the Laplace transform operator as follow:

$$ {}_{0}^{AB}I_{t}^{\alpha}\{{}_{0}^{ABR}D_{t}^{\alpha}f(t)\} = \frac{1-\alpha}{B(\alpha)}\ {}_{0}^{ABR}D_{t}^{\alpha}f(t) + \frac{\alpha}{B(\alpha)\Gamma(\alpha)}\int_{0}^{t}(t-y)^{\alpha-1}\ {}_{0}^{ABR}D_{y}^{\alpha}f(y)dy. \tag{12} $$

Applying on both side of Eq (12) the Laplace transform, we obtain the following expression

$$ \begin{aligned} L\{{}_{0}^{AB}I_{t}^{\alpha}\{{}_{0}^{ABR}D_{t}^{\alpha}f(t)\}\} &= \frac{1-\alpha}{B(\alpha)}L\{{}_{0}^{ABR}D_{t}^{\alpha}f(t)\} \\ &\quad + L\left\{\frac{\alpha}{B(\alpha)\Gamma(\alpha)}\int_{0}^{t}(t-y)^{\alpha-1}\ {}_{0}^{ABR}D_{y}^{\alpha}f(y)dy\right\} \end{aligned} \tag{13} $$

$$ \begin{aligned} L\{{}_{0}^{AB}I_{t}^{\alpha}\{{}_{0}^{ABR}D_{t}^{\alpha}f(t)\}\} &= \frac{1-\alpha}{B(\alpha)}\frac{B(\alpha)}{1-\alpha}\frac{s^{\alpha}F(s)}{s^{\alpha}+\frac{\alpha}{1-\alpha}} \\ &\quad + \frac{\alpha}{B(\alpha)}\frac{B(\alpha)}{1-\alpha}s^{-\alpha}\frac{s^{\alpha}F(s)}{s^{\alpha}+\frac{\alpha}{1-\alpha}} \end{aligned} $$

$$ L\{{}_{0}^{AB}I_{t}^{\alpha}\{{}_{0}^{ABR}D_{t}^{\alpha}f(t)\}\} = \frac{s^{\alpha}F(s)}{s^{\alpha}+\frac{\alpha}{1-\alpha}} + \frac{\alpha}{1-\alpha}\frac{F(s)}{s^{\alpha}+\frac{\alpha}{1-\alpha}} $$

$$ L\{{}_{0}^{AB}I_{t}^{\alpha}\{{}_{0}^{ABR}D_{t}^{\alpha}f(t)\}\} = F(s) $$

By the inverse Laplace transform we obtain

$$ {}_{0}^{AB}I_{t}^{\alpha}\{{}_{0}^{ABR}D_{t}^{\alpha}f(t)\} = f(t) \tag{14} $$

The prove Eq (10b) we use another method that consists of solving the following time fractional ordinary differential equation with Atangana-Baleanu derivative in Caputo sense

$$ {}_{0}^{ABC}D_{t}^{\alpha}f(t) = u(t) \tag{15} $$

With the aim to find the function $f(t)$, to do this we employ again the Laplace transform on both sides we obtain

$$ \frac{B(\alpha)}{1-\alpha}\frac{s^{\alpha}F(s)-s^{\alpha-1}f(0)}{s^{\alpha}+\frac{\alpha}{1-\alpha}} = U(s) \tag{16} $$

Rearranging we obtain

$$ \begin{aligned} F(s) &= \frac{1-\alpha}{B(\alpha)}\frac{\left(s^{\alpha}+\frac{\alpha}{1-\alpha}\right)}{s^{\alpha}}U(s) + s^{-1}f(0) \\ f(t) &= \frac{1-\alpha}{B(\alpha)}u(t) + \frac{\alpha}{B(\alpha)\Gamma(\alpha)}\int_{0}^{t}u(y)(t-y)^{\alpha-1}dy + f(0) \\ f(t)-f(0) &= \frac{\alpha}{B(\alpha)\Gamma(\alpha)}\int_{0}^{t}u(y)(t-y)^{\alpha-1}dy + \frac{1-\alpha}{B(\alpha)}u(t) \\ f(t)-f(0) &= {}_{0}^{AB}I_{t}^{\alpha}u(t) \end{aligned} \tag{17} $$

This completes the proof.

Here applying the AB-fractional integral on Eq (8), we obtain the following

$$\begin{aligned} N(t)-N(0) &= \frac{1-\alpha}{AB(\alpha)}\{aN^2(t)-bN(t)-(1-p)v(t)N^2(t)\} \\ &\quad +\frac{\alpha}{AB(\alpha)\Gamma(\alpha)}\int_0^t (t-y)^{\alpha-1}\left\{\begin{array}{c} aN^2(y)-bN(y) \\ -(1-p)v(y)N^2(y) \end{array}\right\}dy \end{aligned} \tag{18}$$

It is important to note that, Eq (17) is equivalent to Eq (8), in this work, we will use Eq (17) to show the existence of Eq (8).

Lemma 1: The mapping G: $H \to H$ defined as

$$\begin{aligned} GN(t) &= \frac{1-\alpha}{AB(\alpha)}V(t,N(t)) \\ &\quad +\frac{\alpha}{AB(\alpha)\Gamma(\alpha)}\int_0^t (t-y)^{\alpha-1}V(y,N(y))dy \\ V(t,N(t)) &= aN^2(t)-bN(t)-(1-p)v(t)N^2(t) \end{aligned} \tag{19}$$

Lemma 2: Let $M \subset H$ be bounded implying, we can find $l>0$ such that,

$$\| N(a)-N(b) \| \leq l(a-b), \ \forall N \in M.$$

Then $G(\bar{M})$ is compact.

Proof: Let $I = \max\{\frac{1-\alpha}{AB(\alpha)} + V(t,N(t)\}, 0 \leq N \leq L$. For $N \in M$, we have the following

$$\begin{aligned} \| GN(t) \| &\leq \frac{1-\alpha}{AB(\alpha)}\| V(t,N(t)) \| \\ &\quad +\frac{\alpha}{AB(\alpha)\Gamma(\alpha)}\int_0^t (t-y)^{\alpha-1}\| V(y,N(y)) \| dy \\ &\leq \frac{1-\alpha}{AB(\alpha)}I+\frac{\alpha}{AB(\alpha)}I\ \frac{t^\alpha}{\Gamma(\alpha+1)} \end{aligned} \tag{20}$$

This implies the function G is bounded.

Let us now consider $N \in M$, t_1, t_2, $t_1 < t_2$, then for a given $\epsilon > 0$, if $|t_2 - t_1| < \Lambda$.

Then

$$\begin{aligned} \| GN(t_2)-GN(t_1) \| &\leq \frac{1-\alpha}{AB(\alpha)}\| V(t_2,N(t_2))-V(t_1,N(t_1)) \| \\ &\quad +\left\| \begin{array}{l} \frac{\alpha}{AB(\alpha)\Gamma(\alpha)}\int_0^{t_2}(t_2-y)^{\alpha-1}\| V(y,N(y)) \| dy \\ -\frac{\alpha}{AB(\alpha)\Gamma(\alpha)}\int_0^{t_1}(t_1-y)^{\alpha-1}\| V(y,N(y)) \| dy \end{array} \right\| \\ &\leq \frac{1-\alpha}{AB(\alpha)}\| V(t_2,N(t_2))-V(t_1,N(t_1)) \| \\ &\quad +\frac{\alpha L}{AB(\alpha)\Gamma(\alpha)}\left\{\int_0^{t_2}(t_2-y)^{\alpha-1}dy-\int_0^{t_1}(t_1-y)^{\alpha-1}dy\right\} \end{aligned} \tag{21}$$

We will treat the above inequality piece by piece we first start with the integral part.

$$\begin{aligned} & \int_0^{t_2} (t_2 - y)^{\alpha-1} dy - \int_0^{t_1} (t_1 - y)^{\alpha-1} dy \\ = & \int_0^{t_1} \{(t_1 - y)^{\alpha-1} - (t_2 - y)^{\alpha-1}\} dy \\ + \int_{t_1}^{t_2} (t_2 - y)^{\alpha-1} dy = & \frac{2}{\Gamma(\alpha+1)} (t_2 - t_1)^{\alpha} \end{aligned} \tag{22}$$

We next investigate the following

$$\begin{aligned} & \| V(t_2, N(t_2)) - V(t_1, N(t_1)) \| \\ = & \left\| \begin{array}{l} a(N^2(t_2) - N^2(t_1)) - b(N(t_2) - N(t_1)) \\ -(1-p)(v(t_2)N^2(t_2) - v(t_1)N^2(t_1)) \end{array} \right\| \\ \leq & |a| \| N^2(t_2) - N^2(t_1) \| + |b| \| N(t_2) - N(t_1) \| \\ & + (1-p) \| N^2(t_2) - N^2(t_1) \| \\ \leq & \{2aL + b + 2L(1-p)\} \| N(t_2) - N(t_1) \| \\ \leq & \{2aL + b + 2L(1-p)\} l \| (t_2 - t_1) \| \\ \leq & J \| (t_2 - t_1) \| \end{aligned} \tag{23}$$

Now putting Eqs (22) and (21) into (20) we obtain:

$$\| GN(t_2) - GN(t_1) \| \leq \frac{1-\alpha}{AB(\alpha)} J \| (t_2 - t_1) \| + \frac{2\alpha}{AB(\alpha)\Gamma(\alpha+1)} \| (t_2 - t_1) \|^{\alpha} \tag{24}$$

Therefore for each $\epsilon > 0$, we can find

$$\Lambda = \frac{\varepsilon}{\frac{1-\alpha}{AB(\alpha)}\{\{2aL + b + 2L(1-p)\}l\} + \frac{2\alpha}{AB(\alpha)\Gamma(\alpha+1)}} \tag{25}$$

Such that

$$\| GN(t_2) - GN(t_1) \| \leq \varepsilon$$

Hencefort $G(M)$ is equi-continuous and according to the well-known Arzela-Ascoli theorem, $\bar{G(M)}$ is compact.

Theorem 3: V: $[a, b] \times [0, \infty) \rightarrow [0, \infty)$ be a continuous function and $V(t,.)$ increasing for each t in $[a, b]$. Let us assume that one can find v, w satisfying $K(D)v \leq V(t, v)$, $K(D)w \geq V(t, w)$, $0 \leq v(t) \leq w(t)$, $a \leq t \leq b$. Then our new equation has a positive solution.

Proof: The fixed-point of the operator G is needed to be considered. Nevertheless, within the framework of lemma 1, the considered operator G: $H \rightarrow H$ is completely continuous. Let us choose two arbitrary densities of population in the N_1 and N_2 in H satisfying $N_1 \leq N_2$ then,

by assuming that, V is a positive function, then

$$\begin{aligned} GN_1(t) &\leq \frac{1-\alpha}{AB(\alpha)} \parallel V(t, N_1(t)) \parallel \\ &\quad + \frac{\alpha}{AB(\alpha)\Gamma(\alpha)} \int_0^t (t-y)^{\alpha-1} \parallel V(y, N_1(y)) \parallel dy \\ &\leq GN_2(t) \end{aligned} \tag{26}$$

Henceforth the mapping G is increasing. By the conjecture, we get $Gm \geq m$, $Gn \leq n$. Henceforth the operator G: $\langle n, m\rangle \to \langle n, m\rangle$ is compact within the framework of lemma 2 and continuous in view of lemma 1. Since H is a normal cone of G.

Uniqueness of solution

In this section, we discuss with care the conditions under which the unicity of the solution are obtained. To establish these conditions, we consider evaluating the following quantity.

$$\begin{aligned} &\parallel GN(t) - GM(t) \parallel \\ &\leq \left\| \begin{array}{c} \frac{1-\alpha}{AB(\alpha)} (V(t, N(t)) - V(t, M(t))) \\ + \frac{\alpha}{AB(\alpha)\Gamma(\alpha)} \int_0^t (t-y)^{\alpha-1} (V(y, N(y)) - V(y, M(y)))dy \end{array} \right\| \\ &\leq \frac{1-\alpha}{AB(\alpha)} \parallel V(t, N(t)) - V(t, M(t)) \parallel + \\ &\quad \frac{\alpha}{AB(\alpha)\Gamma(\alpha)} \int_0^t (t-y)^{\alpha-1} \parallel V(y, N(y)) - V(y, M(y)) \parallel dy \\ &\leq \frac{1-\alpha}{AB(\alpha)} J \parallel N(t) - M(t) \parallel + \\ &\quad \frac{\alpha}{AB(\alpha)\Gamma(\alpha)} J \int_0^t (t-y)^{\alpha-1} \parallel N(t) - M(t) \parallel dy \\ \parallel GN(t) - GM(t) \parallel &\leq \left\{ \frac{1-\alpha}{AB(\alpha)} J + \frac{\alpha b^\alpha}{AB(\alpha)\Gamma(\alpha+1)} J \right\} \parallel N(t) - M(t) \parallel \end{aligned} \tag{27}$$

Therefore if the following condition holds $\frac{1-\alpha}{AB(\alpha)} J + \frac{\alpha b^\alpha}{AB(\alpha)\Gamma(\alpha+1)} J < 1$ then, the mapping G is a contraction, which implies it has a fixed-point, thus, the new model admits a unique positive solution.

Numerical solution via forward-corrector method

The recent development of fractional differentiation based on the Mittag-Leffler function has induced a new type of Volterra fractional differential equations. As presented earlier, the fractional integral calculus associated o the new fractional calculus is the set of functions for which the their fractional integral in Atangana and Baleanu sense is an average of the given function and the Riemann-Liouville fractional integral. This new design has therefore opened way to many new studies, for instance what can we do to solve the Volterra version of a given equation numerically. It is well known that the Corrector method is very accurate method to handle

Volterra equations, due to the Riemann-Liouville, however, with the new fractional integral one could possible apply also the corrector method in the integral part and apply another numerical method in the other part. In this section, we introduce the Forward-Corrector method to solve our new model. Nevertheless there are two ways to handle numerically fractional differential equation based on the new fractional differentiation. In our case, we could solve our problem directly in its present form or solve its Volterra version. We shall start with the Volterra version.

$$\begin{aligned} N(t)-N(0) &= \frac{1-\alpha}{AB(\alpha)}\{aN^2(t)-bN(t)-(1-p)v(t)N^2(t)\} \\ &\quad +\frac{\alpha}{AB(\alpha)\Gamma(\alpha)}\int_0^t (t-y)^{\alpha-1}\left\{\begin{array}{c} aN^2(y)-bN(y) \\ -(1-p)v(y)N^2(y) \end{array}\right\}dy \end{aligned}$$

The part within the integral could be handled with the Corrector method, which is provided as follow

$$\begin{aligned} &\int_0^{t_n}(t_{n+1}-y)^{\alpha-1}V(y,N(y))dy \\ &= \frac{h^\alpha}{\alpha(\alpha+1)}\sum_{j=0}^{n+1}b_{j,n+1}V(t_j,N(t_j)) \\ b_{j,n+1} &= \begin{cases} n^{\alpha+1}-(n-\alpha)(n+1)^\alpha, & \text{if } j=0 \\ (n-j+2)^{\alpha+1}-(n-j)^{\alpha+1}-2(n-j+1)^{\alpha+1} & \text{if } 1\le j\le n, \\ 1, & \text{if } j=n+1 \end{cases} \end{aligned} \tag{28}$$

Therefore according to [18] the fractional variant of the one step Adam-Moulton method for the second part of our equation is given by:

$$\frac{\alpha h^\alpha}{AB(\alpha)\Gamma(\alpha+2)}V(t_{n+1},N_h^p(t_{n+1}))+\frac{\alpha h^\alpha}{AB(\alpha)\Gamma(\alpha+2)}\sum_{j=0}^{n}b_{j,n+1}V(t_j,N_h(t_j)) \tag{29}$$

In the second part we use the forward approximation as follows

$$N(t_{n+1})=N(t_n)+\frac{1-\alpha}{AB(\alpha)}hV(t_{n+1},N(t_{n+1})) \tag{30}$$

Putting Eqs (29) and (28) into Eq (17), we obtain the following numerical approximation:

$$\begin{aligned} N(t_{n+1}) &= N(t_n)+\frac{1-\alpha}{AB(\alpha)}hV(t_{n+1},N(t_{n+1})) \\ &\quad \frac{\alpha h^\alpha}{AB(\alpha)\Gamma(\alpha+2)}V(t_{n+1},N_h^p(t_{n+1})) \\ &\quad +\frac{\alpha h^\alpha}{AB(\alpha)\Gamma(\alpha+2)}\sum_{j=0}^{n}b_{j,n+1}V(t_j,N_h(t_j)) \end{aligned} \tag{31}$$

This approach can be used to solve many other fractional differential equations based on the new fractional differentiation.

The second approach to solve our problem is to discretize the Atangana-Baleanu time fractional derivative. Koca and Atangana suggested the numerical approximation of the new derivative as follow [19]:

$$
\begin{aligned}
&{}_{0}^{ABC}D_{t}^{\alpha}(N(t_{n+1})) \\
&= \frac{AB(\alpha)}{1-\alpha}\sum_{k=1}^{n+1}\frac{N^{k+1}-N^{k}}{\Delta t}\left\{\begin{array}{l}(t_n - t_{k+1})E_{\alpha,2}\left(-\frac{\alpha}{1-\alpha}(t_n - t_{k+1})\right) \\ -(t_n - t_k)E_{\alpha,2}\left(-\frac{\alpha}{1-\alpha}(t_n - t_k)\right)\end{array}\right\}
\end{aligned} \tag{32}
$$

$$
E_{\alpha,2}(z) = \sum_{j=0}^{\infty}\frac{z^{j}}{j!\Gamma(\alpha j+2)}
$$

Replacing the above in Eq (8), using also the forward numerical scheme, then the numerical approximation solution of the new model is given as:

$$
\begin{aligned}
&\frac{AB(\alpha)}{1-\alpha}\sum_{k=1}^{n+1}\frac{N^{k+1}-N^{k}}{\Delta t}\left\{\begin{array}{l}(t_n - t_{k+1})E_{\alpha,2}\left(-\frac{\alpha}{1-\alpha}(t_n - t_{k+1})\right) \\ -(t_n - t_k)E_{\alpha,2}\left(-\frac{\alpha}{1-\alpha}(t_n - t_k)\right)\end{array}\right\} \\
&= aN^{2}(t_{n+1}) - bN(t_n) - (1-p)v(t_n)N^{2}(t_{n+1}) \\
&\frac{AB(\alpha)}{1-\alpha}\sum_{k=1}^{n+1}\frac{N^{k+1}-N^{k}}{\Delta t}\beta_{k,n} \\
&= aN^{2}(t_{n+1}) - bN(t_n) - (1-p)v(t_n)N^{2}(t_{n+1})
\end{aligned} \tag{33}
$$

Numerical simulations

In this section, we present the numerical replication of the model for different values of fractional order using the proposed numerical scheme. The numerical solutions are depicted in Fig 1 for $\alpha = 0.95$, Fig 2 for $\alpha = 0.75$, Fig 3 for $\alpha = 0.45$ and finally Fig 4 for $\alpha = 0.25$.

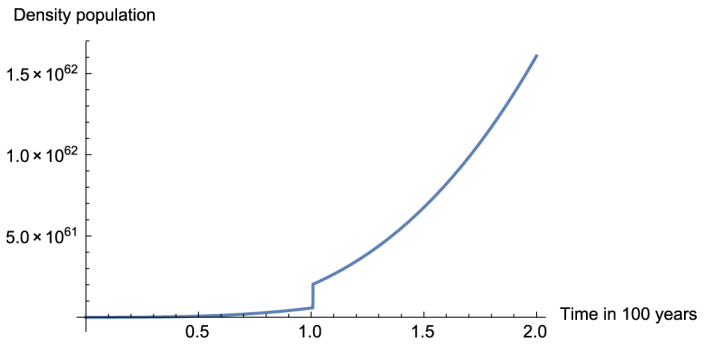

Fig 1. Numerical simulation for $\alpha = 0.95$ and $t = 100$.

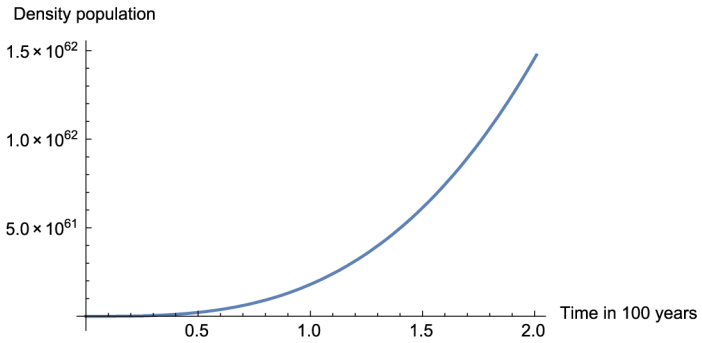

Fig 2. Numerical simulation for $\alpha = 0.75$ and $t = 100$.

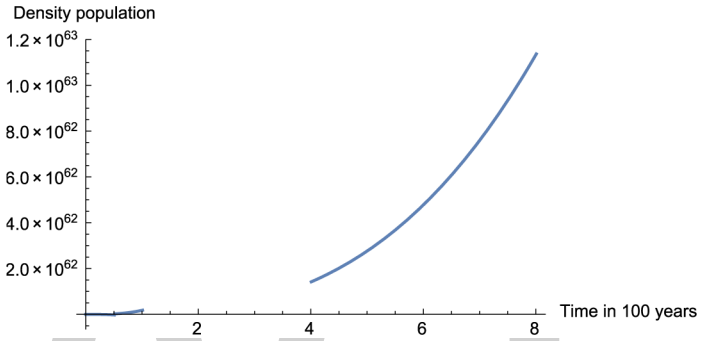

Fig 3. Numerical simulation for $\alpha = 0.45$ and $t = 100$.

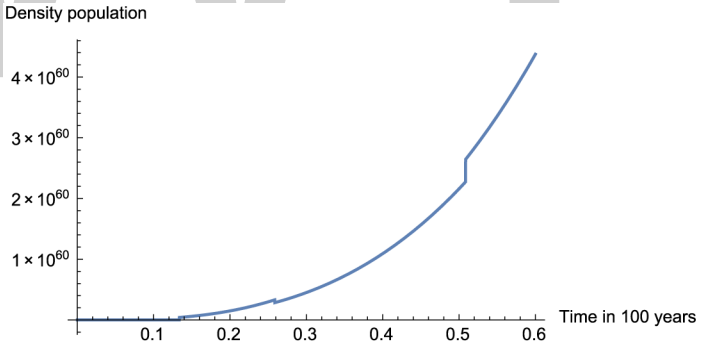

Fig 4. Numerical simulation for $\alpha = 0.25$ and $t = 100$.

Conclusion

The aim of this work was to suggest a nonlinear fractional differential equation that could be used to describe the density of population growth taking into account real world behaviors. To do this, we introduced a new component that considers the choice of partner. The analysis of existence of positive solution of the new model was examined via the fixed-point theorem. The new model was solved numerically using the modified approach that fit well the new fractional integral. Some numerical simulations were done as function of fractional order.

Author Contributions

Conceptualization: Badr Saad T. Alkahtani, Ilknur Koca.

Writing – original draft: Abdon Atangana.

References

1. Kingsland S, Modeling Nature: Episodes in the History of Population Ecology. University of Chicago Press. 1995: 127–146.
2. Uyenoyama M, Rama S, Ed. The Evolution of Population Biology. Cambridge University Press. 2004: 1–19.
3. Renshaw E, Modeling Biological Populations in Space and Time. Cambridge University Press. 1991: 6–9.
4. Worster D, Nature's Economy. Cambridge University Press. 1994: 398–401.
5. Verhulst PF, Notice sur la loi que la population poursuit dans son accroissement. Correspondance mathématique et physique. 1838; 10: 113–121.
6. Verhulst PF, Recherches mathématiques sur la loi d'accroissement de la population. Mathematical Researches into the Law of Population Growth Increase. Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles 1845; 18: 1–42.
7. McKendricka AG, Kesava Paia M,XLV. The Rate of Multiplication of Micro-organisms: A Mathematical Study. Proceedings of the Royal Society of Edinburgh. 1912; 31: 649–653. https://doi.org/10.1017/S0370164600025426
8. Jannedy S, Bod R, Hay J, Probabilistic Linguistics. Cambridge, Massachusetts: 2003, MIT Press.
9. Rach R, Wazwaz A, Two reliable methods for solving the Volterra integral equation with weakly singular kernel. Journal of computational and Applied Mathematics. 2016; 302: 71–80.
10. Atangana A, Baleanu D, New fractional derivatives with nonlocal and non-singular kernel: Theory and application to heat transfer model. Therm. Sci. 2016.
11. Atangana A, Koca I, Chaos in a simple nonlinear system with Atangana–Baleanu derivatives with fractionalorder. Chaos Solitons Fractals. 2016.
12. Algahtani OJJ, Comparing the Atangana–Baleanu and Caputo–Fabrizio derivative with fractional order: Allen Cahn model, Chaos, Solitons and Fractals. 2016.
13. Alkahtani BST, Chua's circuit model with Atangana–Baleanu derivative with fractional order, Chaos, Solitons and Fractals. 2016.
14. Atangana A, Baleanu D, Caputo-Fabrizio applied to groundwater flow within a confined aquifer. J Eng Mech. 2016 Press.
15. Wazwaz A, Analytical approximations and Pade' approximants for Volterra's population model. Applied Mathematics and Computation. 1999; 100: 13–25. https://doi.org/10.1016/S0096-3003(98)00018-6
16. Wazwaz A, The decomposition method applied to systems of partial differential equations and to the reaction-diffusion Brusselator model. Applied Mathematics and Computation. 2000; 110: 251–264. https://doi.org/10.1016/S0096-3003(99)00131-9
17. Wazwaz A, Eltantawy S, A new (3+1)-dimensional generalized Kadomtsev–Petviashvili equation. Nonlinear Dynamice. 2016; 1529–1534.
18. Benson D, Wheatcraft S, Meerschaert M. Application of a fractional advec- tion-dispersion equation. Water Resour Res 2000; 36:1403–12. Näsholm SP, Holm S. On a fractional zener elastic wave equation. Fract. Calc. Appl. Anal. (2013), 16 (1):26–50. https://doi.org/10.1029/2000WR900031
19. Koca I, Atangana A, Solutions of Cattaneo-Hristov model of elastic heat diffusion with Caputo-Fabrizio and Atangana Baleanu fractional derivatives. Thermal Science. 2016 OnLine-First (00):102–102.
20. Arafa AA, Rida SZ, Mohamed H, Homotopy analysis method for solving biological population model. Commun Theor Phys. 2011; 56: 797–800. https://doi.org/10.1088/0253-6102/56/5/01
21. Roul P, Application of homotopy perturbation method to biological population model. Appl Appl Math. 2010; 10: 1369–1378.
22. Macías-Díaz JE, Existence and Uniqueness of Positive and Bounded Solutions of a Discrete Population Model with Fractional Dynamics. Discrete Dynamics in Nature and Society. 2017; 1–7.
23. Syed TMD, Ayyaz A, Bandar BM, On biological population model of fractional order. Int. J. Biomath. 2016; 9: no. 5, Article ID 1650070, 13 pages.

Adaptive fractional fuzzy sliding mode control of microgyroscope based on backstepping design

Xiao Liang, Juntao Fei *

College of IoT Engineering, Hohai University, Changzhou, China

* jtfei@hhu.edu.cn

Abstract

In this paper, a robust sliding mode control (SMC) based on backstepping technique is studied for a microgyroscope in the presence of unknown model uncertainties and external disturbances using adaptive fuzzy compensator and fractional calculus. At first, the dynamic of microgyroscope is transformed into analogically cascade system to guarantee the application of backstepping design. Then a novel fractional differential sliding surface is proposed which integrates the capacities of the fractional calculus and SMC. In order to reduce the chattering in SMC, a fuzzy logical system is utilized to approximate the external disturbances. In addition, fractional order adaptive laws are derived to estimate the damping and stiffness coefficients and angular velocity online based on Lyapunov stability theory which also guarantees the stability of the closed loop system. Finally, simulation results signify the robustness and effectiveness of the proposed control schemes and the comparison of root mean square error under different fractional orders and integer order are given to demonstrate the better performance of proposed controller.

Editor: Chee Kong Chui, National University of Singapore

Introduction

Microgyroscope has many applications in military and civil fields such as navigation, automobile and traffic etc. due to their superior features in angular velocity measurement. However, constrained by manufacturing process and design principle, it is difficult to meet desired requirements and its performance is sensitive to time varying system parameters, external disturbances, ambient conditions including temperature and pressure and so on. In order to obtain better dynamic performance, lots of robust control methods have been applied to microgyroscope for many years. Park[1] proposed an adaptive control scheme with velocity estimation to compensate fabrication imperfects so as to operate insusceptibly in varying environments for a z-axis microgyroscope. In [2], two adaptive controllers were developed to tune the natural frequency of the drive axis for a vibrational microgyroscope. Adaptive neural sliding mode control algorithms were proposed for the unknown system dynamics and nonlinearities in the microgyroscope in [3–4]. A direct model reference adaptive control scheme with an estimating observer to modify disturbance was investigated which ensured the resonant

Funding: This work is partially supported by National Science Foundation of China under Grant No. 61873085, Natural Science Foundation of Jiangsu Province under Grant No. BK20171198; The University Graduate Research and Innovation Projects of Jiangsu Province under Grant No. KYCX17_0540; The Fundamental Research Funds for the Central Universities under Grant No. 2017B679X14, 2017B20014.

Competing interests: The authors have declared that no competing interests exist.

oscillations of the microgyroscope in [5]. The tuning algorithm for systems parameters is derived based on Lyapunov stability theorem which guarantees the stability of the closed–loop system. By constructing suitable Lyapunov functions and combing with matrix inequality technique, new simple sufficient conditions are presented for stochastic delayed cellular neural networks in [6–7] and global asymptotic stability of the cohen-grossberg neural network models in [8–9] respectively.

Fractional calculus which expends the order of differential and integral from integer to fraction has been studied for three centuries. In recent years, more and more attention has been paid on its application in controller design instead of a pure theoretical mathematical subject owing to its higher modeling accuracy and degree of freedom compared to integer order controllers. Some researches about fractional calculus have been studied in [10–13]. Fractional order controllers were employed for microgrid in [14]. A fractional model was established to solve some physical problem in [15–16]. A model reference adaptive control strategy with fractional operators was demonstrated to improve the plant dynamics in [17]. A local fractional differential equation of fractal dimensional order is applied to a non-differentiable model of the LC-electric circuit in [18]. Some researches about the solvability for nonlinear fractional differential equations have been studied in [19–21].

The sliding mode control (SMC) technique is considered to be an effective control scheme for robust control which has been applied to both linear and nonlinear systems. The main idea of SMC is to choose a linear manifold of the state variables such as deviations and their derivatives as sliding surface and then design a control law for driving and constraining the system state into the previous designed sliding surface. SMC has shown great superiorities in dealing with nonlinear systems with uncertainties which is benefit from its robustness and insensitivity to parameters variation and external disturbances. Sling mode control and observations were focused on in [22] for complex industrial systems because of the advantages above. An adaptive novel SMC using neural network and fuzzy system are designed for the uncertain nonlinear system in [23] [24]. An adaptive SMC with a new adaptive law, whose adaptive gains was inversely proportional to the sliding variables, offered the fast dynamic and reduced chattering for robot manipulators in [25]. Neural network and fuzzy system are also utilized to deal with uncertainties and suppress the harmonics for active power filter in [26] [27] [28] [29].

SMC applies not only to integer order systems, but also to fractional order systems. Thus fractional order calculus can also be incorporated in sliding mode control [30–31]. Chen et al. [30] proposed an adaptive sliding mode control scheme for a fractional order nonlinear system with uncertainties. A fractional order fuzzy sliding mode controller was designed for robotic manipulators which retained the advantages of SMC and reduced the chattering simultaneously in [31].

Backstepping method has been well known for its recursive and systematic design in nonlinear feedback systems [32] [33] [34]. The concept of backstepping control is to choose appropriate functions of the state variables as virtual controls for subsystems and then design control laws based on Lyapunov functions. A simplified adaptive backstepping scheme was proposed for a full-car active suspension system with external disturbances in [35]. An adaptive backstepping controller was proposed for vehicle active suspensions in [36] to guarantee the stability of the attitude of vehicle and the improvement of ride comfort. Unfortunately, backstepping control scheme does not work well for systems with discontinuous disturbance and parameter variations. So it is usually combined with other intelligent control methods such as sliding mode control, fuzzy control and so on. An adaptive sliding mode controller based on backstepping technique was proposed for robotic manipulator in [37] which estimates the system uncertainties and external disturbances by the adaptive laws. Park et al. [38] designed a backstepping integral sliding mode controller based on T-S fuzzy model for an Interior Permanent Magnet Synchronous Motor. A backstepping fractional order sliding

mode control was developed for power systems and microgyroscope system respectively which showed good dynamic performances and great robustness compared to traditional methods in [39–40]. Feng et. al [41] proposed a novel adaptive Super-Twisting sliding mode control for a microgyroscope.

In this paper, in order to incorporate the advantages of fractional control, sliding mode control, fuzzy control and backstepping control, an adaptive fractional fuzzy sliding mode controller based on backstepping design is proposed for a microgyroscope. The output trajectory of microgyroscope track the reference trajectory accurately and effectively and the estimation of system parameters have been verified to converge to their true values asymptotically. The main contributions of this paper are emphasized as follows:

1. The superior characteristic of this designed controller is that a fractional order term is adopted in the sliding manifold which generates an extra degree of freedom and makes the design of control law more flexible, consequently the performance of the closed loop system has been improved a lot compared to the traditional SMC whose sliding surface is based on integer order calculus of the state variables.

2. Based on backstepping fractional sliding mode control scheme, a fuzzy logical system is designed to deal with the unknown uncertainties and external disturbances which weakened the chattering phenomenon. Furthermore, adaptive algorithm for parameters of microgyroscope is derived based on Lyapunov stability theory, which guarantees the stability of the closed-loop system and the unknown parameters of microgyroscope system can be identified on line simultaneously. In general, the method proposed in this paper both improves the system performance and enhancing system robustness against model uncertainties and external disturbances as well.

This paper is organized as follows: In section 2, the dynamics of microgyroscope is described. The structure of backstepping fractional sliding mode control and adaptive fractional fuzzy sliding mode control based on backstepping technique are proposed in section 3 and section 4 respectively. Simulation results are shown in section 5 and finally for the conclusions.

Materials and methods

In this section, the mathematical model of z-axis microgyroscope is described, and the preliminary of fractional calculus is introduced, then for solving the trajectory tracking problem of microgyroscope system with unknown model uncertainties and external disturbances, an adaptive fractional fuzzy sliding mode controller based on backstepping design is proposed based on Lyapunov theory.

Dynamics of microgyroscope

The microgyroscope is composed of a proof mass, sensing mechanisms and electrostatic actuation used to force an oscillatory motion and velocity of the proof mass and to sense the position. In order to achieve the dynamics of the MEMS, some assumptions have been made: 1) the motion of the proof mass is limited to x and y axis as shown in Fig 1; 2) the microgyroscope rotates at a constant angular velocity; 3) the centrifugal forces is neglected. Under the above assumptions, the dynamics of the microgyroscope can be simplified as follows:

$$\begin{aligned} m\ddot{x} + d_x\dot{x} + [k_x - m(\Omega_y{}^2 + \Omega_z{}^2)]x + m\Omega_x\Omega_y y = u_x + 2m\Omega_z\dot{y} \\ m\ddot{y} + d_y\dot{y} + [k_y - m(\Omega_x{}^2 + \Omega_z{}^2)]y + m\Omega_x\Omega_y x = u_y - 2m\Omega_z\dot{x} \end{aligned} \quad (1)$$

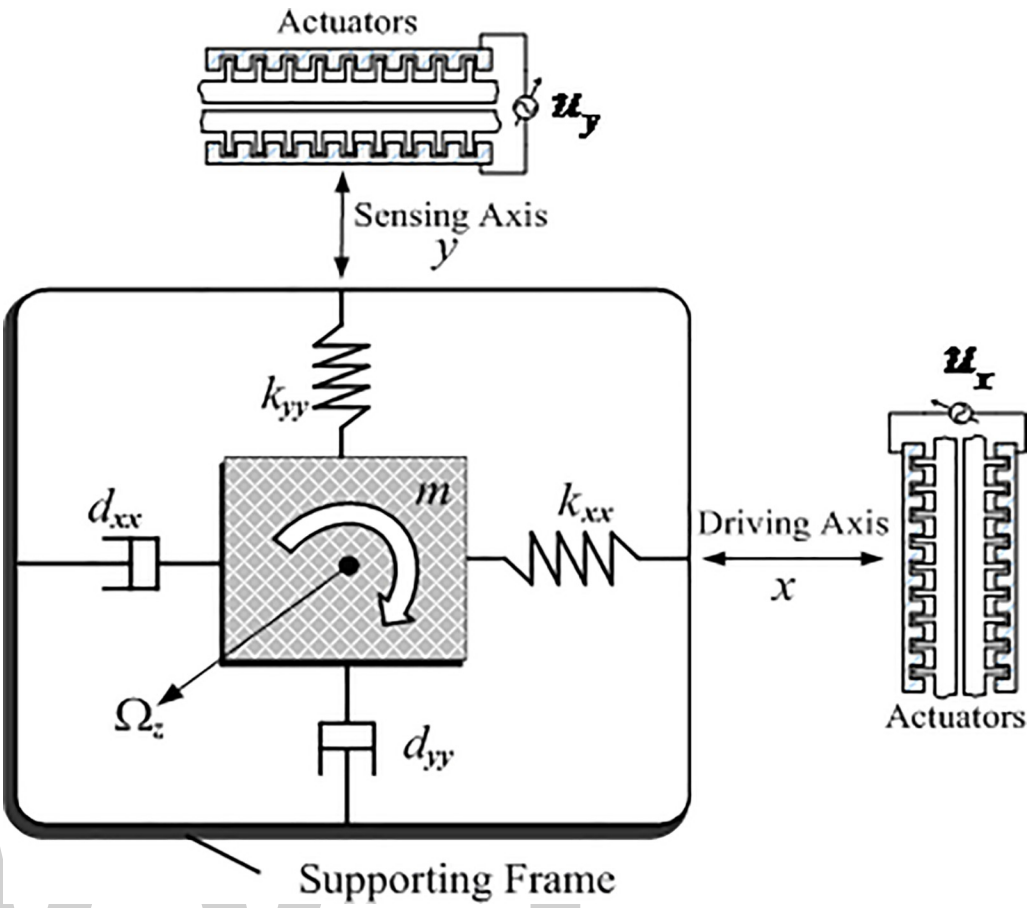

Fig 1. Schematic diagram of a z-axis microgyroscope.

where m is the mass of proof mass, $d_{x,y}$ and $k_{x,y}$ are damping and spring coefficients terms along x- and y-axis respectively. $\Omega_{x,y,z}$ are the angular velocity along each axis, and $u_{x,y}$ are the control forces in x and y directions.

Considering fabrication defects, which may cause extra coupling between x- and y- axis, the dynamics for a z-axis microgyroscope is revised as:

$$\begin{aligned} m\ddot{x} + d_{xx}\dot{x} + d_{xy}\dot{y} + k_{xx}x + k_{xy}y &= u_x + 2m\Omega_z\dot{y} \\ m\ddot{y} + d_{xy}\dot{x} + d_{yy}\dot{y} + k_{xy}x + k_{yy}y &= u_y - 2m\Omega_z\dot{x} \end{aligned} \tag{2}$$

In the above equations, d_{xx} and d_{yy} are damping terms; k_{xx} and k_{yy} are spring coefficients terms; d_{xy} and k_{xy} are coupled damping and spring terms, respectively.

Dividing both sides of Eq (2) by proof mass m, reference length q_0 and natural resonance frequency ω_0 simultaneously results:

$$\begin{aligned} \ddot{x} + d_{xx}\dot{x} + d_{xy}\dot{y} + \omega_x{}^2 x + \omega_{xy}y &= u_x + 2\Omega_z\dot{y} \\ \ddot{y} + d_{xy}\dot{x} + d_{yy}\dot{y} + \omega_{xy}x + \omega_y{}^2 y &= u_y - 2\Omega_z\dot{x} \end{aligned} \tag{3}$$

which is the nondimensional dynamics of microgyroscope.

In (3),

$$\frac{d_{xx}}{m\omega_0} \to d_{xx}, \frac{d_{xy}}{m\omega_0} \to d_{xy}, \frac{d_{yy}}{m\omega_0} \to d_{yy}$$
$$\frac{k_{xx}}{m{\omega_0}^2} \to \omega_x^2, \frac{k_{xy}}{m{\omega_0}^2} \to \omega_{xy}, \frac{k_{yy}}{m{\omega_0}^2} \to \omega_y^2, \frac{\Omega_z}{m\omega_0} \to \Omega_z \tag{4}$$

Through the equivalent transformation, the vector form of the model is described as:

$$\ddot{q} + D\dot{q} + Kq = u - 2\Omega\dot{q} \tag{5}$$

where

$$q = \begin{bmatrix} x \\ y \end{bmatrix}, D = \begin{bmatrix} d_{xx} & d_{xy} \\ d_{xy} & d_{yy} \end{bmatrix}, K = \begin{bmatrix} \omega_x^2 & \omega_{xy} \\ \omega_{xy} & \omega_y^2 \end{bmatrix}, u = \begin{bmatrix} u_x \\ u_y \end{bmatrix}, \Omega = \begin{bmatrix} 0 & -\Omega_z \\ \Omega_z & 0 \end{bmatrix} \tag{6}$$

Backstepping fractional sliding mode control

Preliminary introduction of fractional order. As the extended form of differentiation and integration, Caputo(C), Riemann-Liouville(RL), and Grunwald-Letnikov(GL) definitions are the three most commonly used definitions in engineering, science and economics fields, especially the Caputo fractional order calculus which happens to be adopted in this paper.

The Caputo fractional derivative of order α of function $f(x)$ is denoted as:

$${}_aD_t^{\alpha} f(t) = \frac{1}{\Gamma(n-\alpha)} \int_a^t \frac{f^{(n)}(\tau)}{(t-\tau)^{\alpha-n+1}} d\tau, \ n-1 < \alpha < n \tag{7}$$

where t and a are the upper and lower bounds of the operator respectively and Γ is the Gamma function which satisfies:

$$\Gamma(\gamma) = \int_0^{\infty} e^{-t} t^{\gamma-1} dt \tag{8}$$

For convenience, ${}_aD_t^{\alpha}$ is replaced by D^{α} in the following parts.

It is noted that if $\alpha = 0$, then the operation $D^0 f(x)$ satisfies $D^0 f(x) = f(x)$.

Fractional differential sliding mode surface is proposed in this part since its higher control precision compared to the integer order for the adjustable fractional order α. Backstepping control is usually applied to a class of special nonlinear dynamical systems which can be built from subsystems by choosing appropriate Lyapunov functions. Thanks to the recursive procedure, good tracking performance and global stability are guaranteed.

Design of backstepping fractional sliding mode control. Considering the system parameter variations and external disturbances, the dynamic of the MEMS gyroscope is described as follows:

$$\ddot{q} + (D + 2\Omega)\dot{q} + Kq = u + d \tag{9}$$

where d denotes the lumped bounded uncertainties and disturbances which satisfies $\|d\| \leq \rho$, and ρ is a positive constant, referring to the upper bound of the uncertainties and disturbances.

For the application of backstepping technique, coordinate transformation of the dynamic is necessary.

Define two variables x_1 and x_2. Let

$$x_1 = q,\ x_2 = \dot{q} \tag{10}$$

Then a mathematical model of MEMS gyroscope can be expressed as follows:

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -(D + 2\Omega)x_2 - Kx_1 + u + d \end{cases} 0(11)$$

Making the position vector q follow its desired trajectory strictly is the main object of the controller design. The specific controller design is divided into two steps.

Step 1: Assume that q_r is the ideal tracking value, then the tracking error vector can be defined as:

$$e_1 = x_1 - q_r \tag{12}$$

Then its time derivative is

$$\dot{e}_1 = \dot{x}_1 - \dot{q}_r = x_2 - \dot{q}_r \tag{13}$$

Define the virtual control variable as:

$$\alpha_1 = -c_1 e_1 + \dot{q}_r \tag{14}$$

where c_1 is a constant and $c_1>0$.

Define the new error variable as:

$$e_2 = x_2 - \alpha_1 \tag{15}$$

Select a Lyapunov function as Eq (16):

$$V_1 = \frac{1}{2} {e_1}^T e_1 \tag{16}$$

By deriving both sides of (16) one can obtain:

$$\begin{aligned} \dot{V}_1 &= {e_1}^T \dot{e}_1 = {e_1}^T (x_2 - \dot{q}_r) \\ &= {e_1}^T (e_2 - c_1 e_1) \\ &= {e_1}^T e_2 - c_1 {e_1}^T e_1 \end{aligned} \tag{17}$$

If $e_2 = 0$, then

$$\dot{V}_1 = -c_1 {e_1}^T e_1 \le 0 \tag{18}$$

Step 2: The time derivative of (15) is

$$\begin{aligned} \dot{e}_2 &= \dot{x}_2 - \dot{\alpha}_1 \\ &= -(D + 2\Omega)x_2 - Kx_1 + u + d - \dot{\alpha}_1 \end{aligned} \tag{19}$$

A fractional order sliding mode surface is defined as:

$$s = \lambda_1 e_1 + \lambda_2 D^{\alpha-1} e_1 + \lambda_3 e_2 \tag{20}$$

where λ_1,λ_2,λ_3 refer to the positive sliding surface parameters and α–1 is the fractional order of fractional derivate operation.

Taking the time derivative of s, we get:

$$\dot{s} = \lambda_1 \dot{e}_1 + \lambda_2 D^{\alpha} e_1 + \lambda_3 \dot{e}_2 \tag{21}$$

A new Lyapunov function is described as:

$$V_2 = V_1 + \frac{1}{2} s^T s \tag{22}$$

By making derivative of (22), we have:

$$\begin{aligned} \dot{V}_2 &= \dot{V}_1 + s^T \dot{s} \\ &= {e_1}^T e_2 - c_1 {e_1}^T e_1 + s^T (\lambda_1 \dot{e}_1 + \lambda_2 D^{\alpha} e_1 + \lambda_3 \dot{e}_2) \end{aligned} \tag{23}$$

where

$$e_2 = \frac{s - \lambda_1 e_1 - \lambda_2 D^{\alpha-1} e_1}{\lambda_3} \tag{24}$$

Then, Eq (24) is added into Eq (23), which yields

$$\begin{aligned} \dot{V}_2 &= {e_1}^T e_2 - c_1 {e_1}^T e_1 + s^T (\lambda_1 \dot{e}_1 + \lambda_2 D^{\alpha} e_1 + \lambda_3 \dot{e}_2) \\ &= -c_1 {e_1}^T e_1 + {e_1}^T \frac{s - \lambda_1 e_1 - \lambda_2 D^{\alpha-1} e_1}{\lambda_3} + s^T [\lambda_1 \dot{e}_1 + \lambda_2 D^{\alpha} e_1 + \lambda_3 (f + u + d - \dot{\alpha}_1)] \end{aligned} \tag{25}$$

where $f = -(D + 2\Omega)\dot{q} - Kq = -(D + 2\Omega)x_2 - Kx_1$.

In order to keep $\dot{V}_2 \leq 0$, the corresponding control law is designed as:

$$\begin{aligned} u &= -f - \rho \frac{s}{\|s\|} + \dot{\alpha}_1 + \frac{1}{\lambda_3} \left(-\lambda_1 \dot{e}_1 - \lambda_2 D^{\alpha} e_1 - \frac{e_1}{\lambda_3} + \frac{\lambda_2 s {e_1}^T}{\|s\|^2 \lambda_3} D^{\alpha-1} e_1 \right) \\ &= (D + 2\Omega)(e_2 + \alpha_1) + K(e_1 + q_r) - \rho \frac{s}{\|s\|} + \dot{\alpha}_1 \\ &\quad + \frac{1}{\lambda_3} \left(-\lambda_1 \dot{e}_1 - \lambda_2 D^{\alpha} e_1 - \frac{e_1}{\lambda_3} + \frac{\lambda_2 s {e_1}^T}{\|s\|^2 \lambda_3} D^{\alpha-1} e_1 \right) \end{aligned} \tag{26}$$

Since $s^T e_1 = {e_1}^T s$, substituting Eq (26) into Eq (25) results:

$$\begin{aligned} \dot{V}_2 &= -c_1 {e_1}^T e_1 - \frac{\lambda_1}{\lambda_3} {e_1}^T e_1 + s^T \lambda_3 \left(d - \rho \frac{s}{\|s\|} \right) \\ &\leq -c_1 {e_1}^T e_1 - \frac{\lambda_1}{\lambda_3} {e_1}^T e_1 + \lambda_3 (\|s\| \|d\| - \rho \|s\|) \\ &\leq -c_1 {e_1}^T e_1 - \frac{\lambda_1}{\lambda_3} {e_1}^T e_1 \leq 0 \end{aligned} \tag{27}$$

The derivative of V_2 keeps negative semi- definite. According to Barbalart lemma, it can be proved $\lim_{t \to \infty} e_1(t) = 0, \lim_{t \to \infty} s(t) = 0$, that is to say the proposed control strategy can ensure the asymptotical stability of the closed loop system.

Adaptive fractional fuzzy sliding mode control based on backstepping technique. In previous controller design, the control law (26) is derived under the condition of the available parameter variations D,K,Ω and external disturbances ρ. On the contrary, these uncertainty bounds are unknown in actual systems. So for the better conduction of the backstepping fractional SMC system in practice, a good estimate of the unknown parameters with $\hat{D}, \hat{K}, \hat{\Omega}$ is

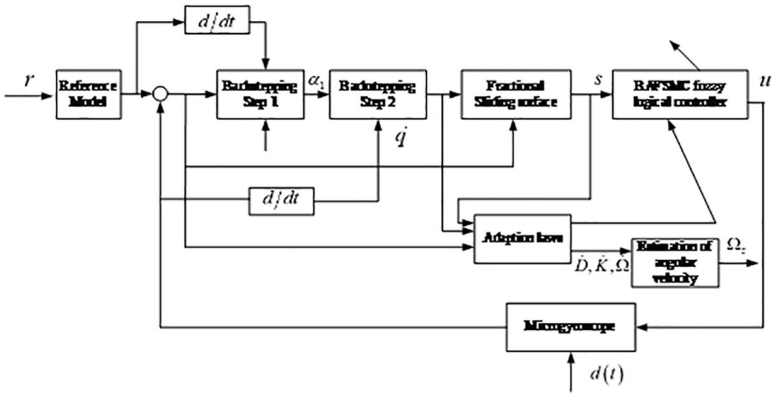

Fig 2. The architecture of the adaptive backstepping fractional fuzzy sliding mode controller.

necessary. Adaptive schemes combined are used online to collect data and adjust the parameters automatically. In addition, an adaptive fuzzy compensator $\hat{h}(s)$ is proposed to handle the chattering caused by the sliding mode surface. The architecture of the proposed adaptive fractional fuzzy sliding mode controller based on backstepping technique is shown in Fig 2.

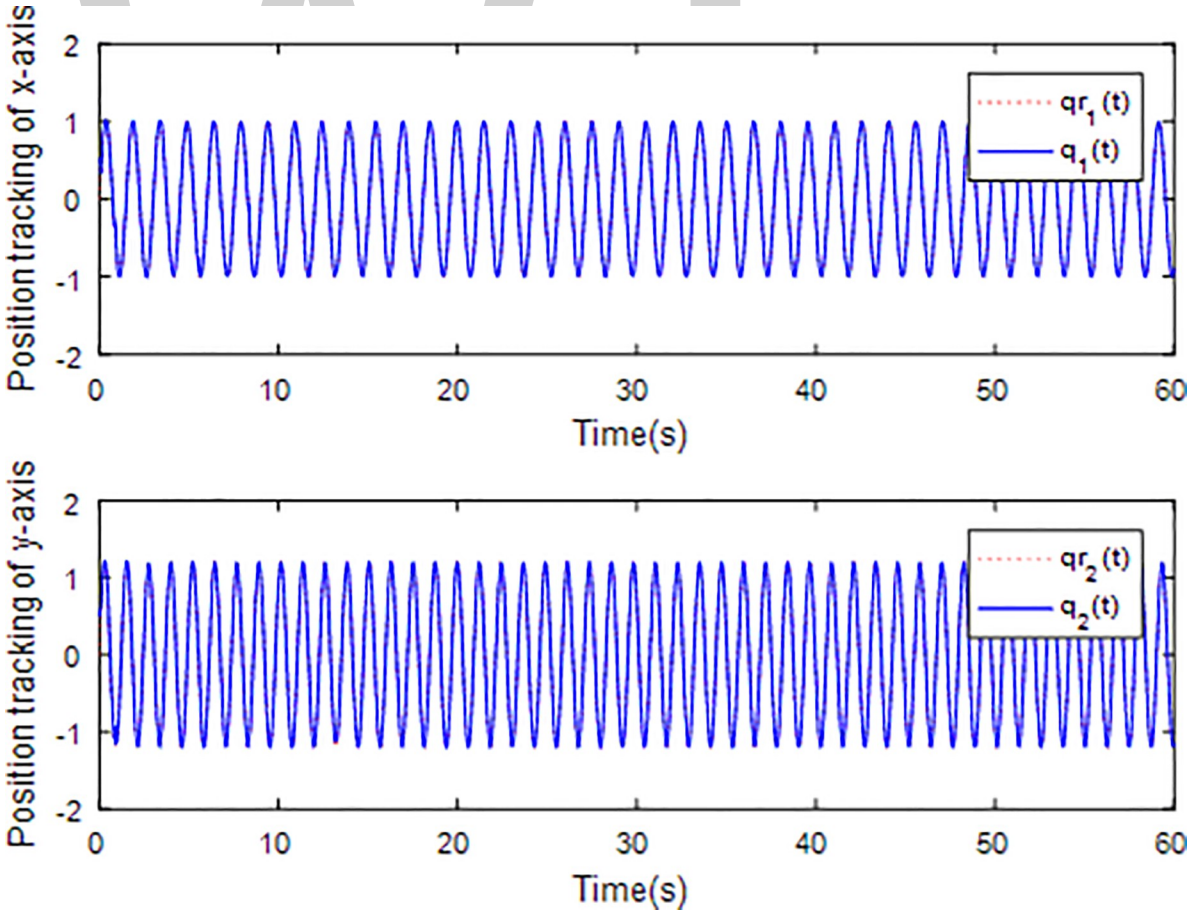

Fig 3. Tracking trajectory using fractional order sliding surface.

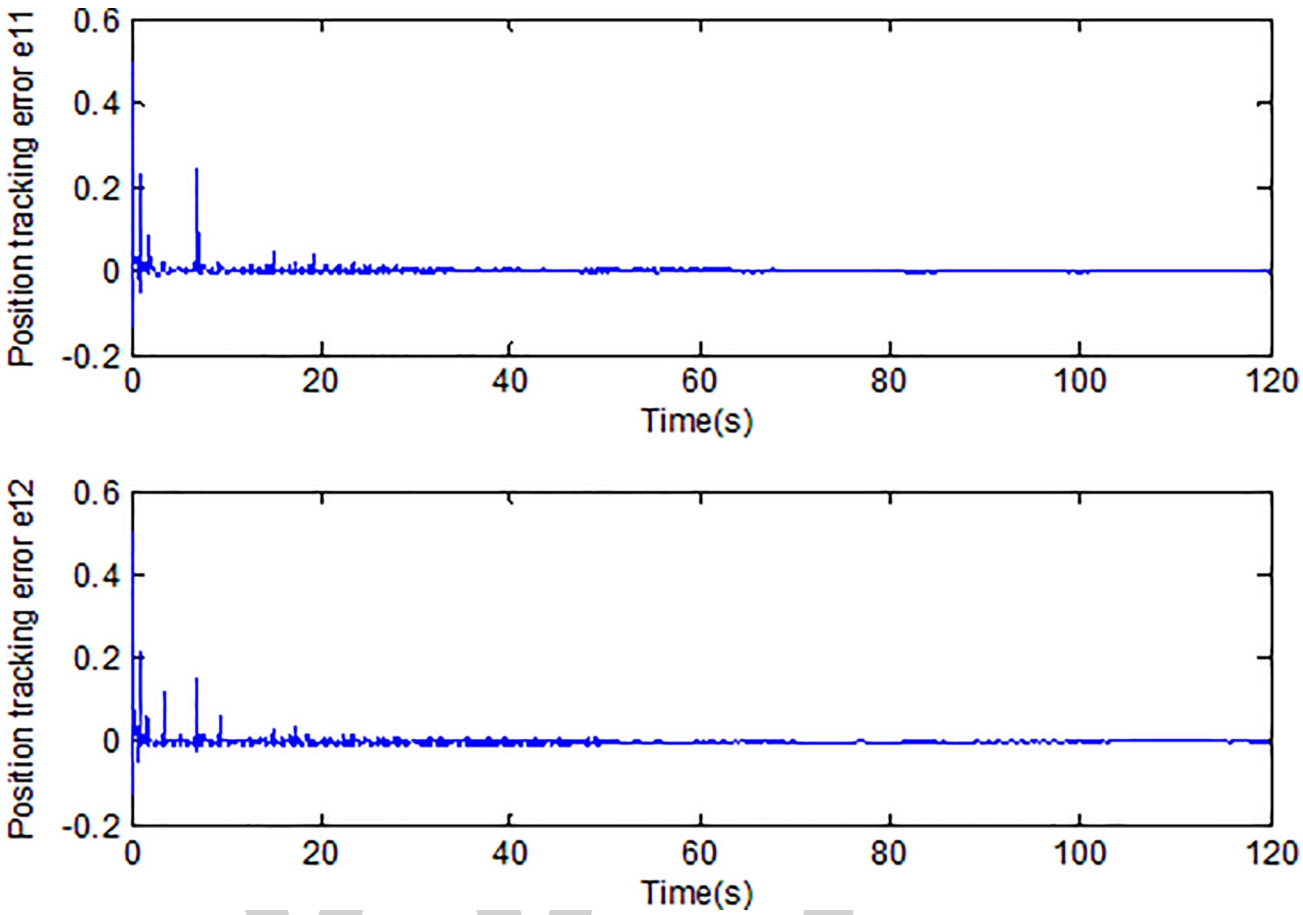

Fig 4. Tracking error using fractional order sliding surface.

Define the parameter estimation error as:

$$\begin{aligned} \tilde{D} &= \hat{D} - D \\ \tilde{K} &= \hat{K} - K \\ \tilde{\Omega} &= \hat{\Omega} - \Omega \\ \tilde{\theta}_h &= \theta_h{}^* - \theta_h \end{aligned} \tag{28}$$

The adaptive control law u' can be derived as:

$$u' = (\hat{D} + 2\hat{\Omega})(e_2 + \alpha_1) + \hat{K}(e_1 + q_r) + \dot{\alpha}_1 - \hat{h}(s) + \frac{1}{\lambda_3}\left(-\lambda_1 \dot{e}_1 - \lambda_2 D^{\alpha} e_1 - \frac{e_1}{\lambda_3} + \frac{\lambda_2 s e_1{}^T}{\|s\|^2 \lambda_3} D^{\alpha-1} e_1\right) \tag{29}$$

where

$$\hat{h}(s|\theta) = [\hat{h}_1 \;\; \hat{h}_2]^T = [\theta_{h1}{}^T \Phi(s_1) \;\; \theta_{h2}{}^T \Phi(s_2)]^T \tag{30}$$

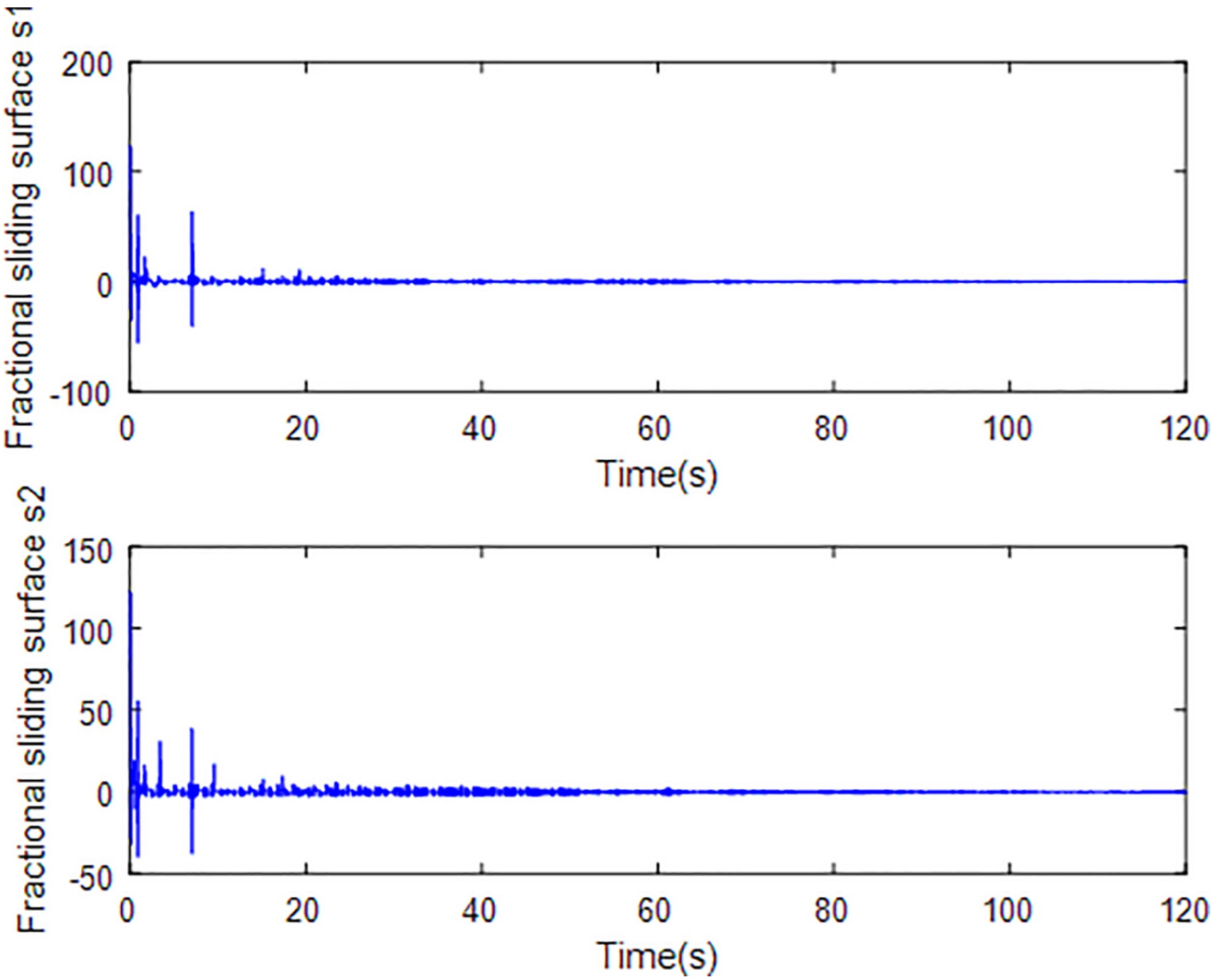

Fig 5. Fractional sliding surface.

Assuming that $\hat{h}(s|\theta_h^*) = \rho\frac{s}{\|s\|}$, then the optimal parameters of fuzzy system is defined as:

$$\theta_h^* = \arg\min_{\theta_h\in\Omega_h}\left[\sup_{x\in R^n}|\hat{h}(s|\theta_h) - \hat{h}(s|\theta_h^*)|\right] \tag{31}$$

where Ω_h are the collections of parameter and θ_h.

Substituting the control law (29) into $\dot{s}$ as in (21) results in:

$$\begin{aligned}\dot{s} &= \lambda_1\dot{e}_1 + \lambda_2 D^\alpha e_1 + \lambda_3(f + u + d - \dot{\alpha}_1)\\ &= \lambda_3\left((\tilde{D} + 2\tilde{\Omega})(e_2 + \alpha_1) + \tilde{K}(e_1 + q_r) + d - \hat{h}(s) + \frac{1}{\lambda_3}\left(-\frac{e_1}{\lambda_3} + \frac{\lambda_2 s e_1^T}{\lambda_3\|s\|^2}D^{\alpha-1}e_1\right)\right)\end{aligned} \tag{32}$$

Define the Lyapunov function candidate as:

$$V = \frac{1}{2}e_1^T e_1 + \frac{1}{2}s^T s + \frac{1}{2}tr\{\tilde{D}M^{-1}\tilde{D}^T\} + \frac{1}{2}tr\{\tilde{K}N^{-1}\tilde{K}^T\} + \frac{1}{2}tr\{\tilde{\Omega}P^{-1}\tilde{\Omega}^T\} + \frac{1}{2r}\sum_{i=1}^{2}\tilde{\theta}_{hi}^T\tilde{\theta}_{hi} \tag{33}$$

where $M = M^T > 0$, $N = N^T > 0$, $P = P^T > 0$ are positive definite matrices and $tr\{\bullet\}$ denotes the matrix trace operator.

Taking the time derivation on both sides of V yields

$$\begin{aligned}
\dot{V} &= -c_1 {e_1}^T e_1 + {e_1}^T \frac{s - \lambda_1 e_1 - \lambda_2 D^{\alpha-1} e_1}{\lambda_3} \\
&+ s^T \lambda_3 \left((\tilde{D} + 2\tilde{\Omega})(e_2 + \alpha_1) + \tilde{K}(e_1 + q_r) + d - \hat{h}(s) + \frac{1}{\lambda_3} \left(-\frac{e_1}{\lambda_3} + \frac{\lambda_2 s {e_1}^T}{\|s\|^2 \lambda_3} D^{\alpha-1} e_1 \right) \right) + tr\{\tilde{D} M^{-1} \dot{\tilde{D}}^T\} \\
&+ tr\{\tilde{\Omega} P^{-1} \dot{\tilde{\Omega}}^T\} + tr\{\tilde{K} N^{-1} \dot{\tilde{K}}^T\} + \frac{1}{r} \sum_{i=1}^{2} \tilde{\theta}_{hi}{}^T \dot{\tilde{\theta}}_{hi} \\
&= -c_1 {e_1}^T e_1 - \frac{\lambda_1 {e_1}^T e_1}{\lambda_3} + s^T \lambda_3 \tilde{D}(e_2 + \alpha_1) + tr\{\tilde{D} M^{-1} \dot{\tilde{D}}^T\} + s^T \lambda_3 \tilde{K}(e_1 + q_r) + tr\{\tilde{K} N^{-1} \dot{\tilde{K}}^T\} \\
&+ 2 s^T \lambda_3 \tilde{\Omega}(e_2 + \alpha_1) + tr\{\tilde{\Omega} P^{-1} \dot{\tilde{\Omega}}^T\} + s^T \lambda_3 (\hat{h}(s|\theta^*) - \hat{h}(s) + d - \hat{h}(s|\theta^*)) + \frac{1}{r} \sum_{i=1}^{2} \tilde{\theta}_{hi}{}^T \dot{\tilde{\theta}}_{hi}
\end{aligned} \tag{34}$$

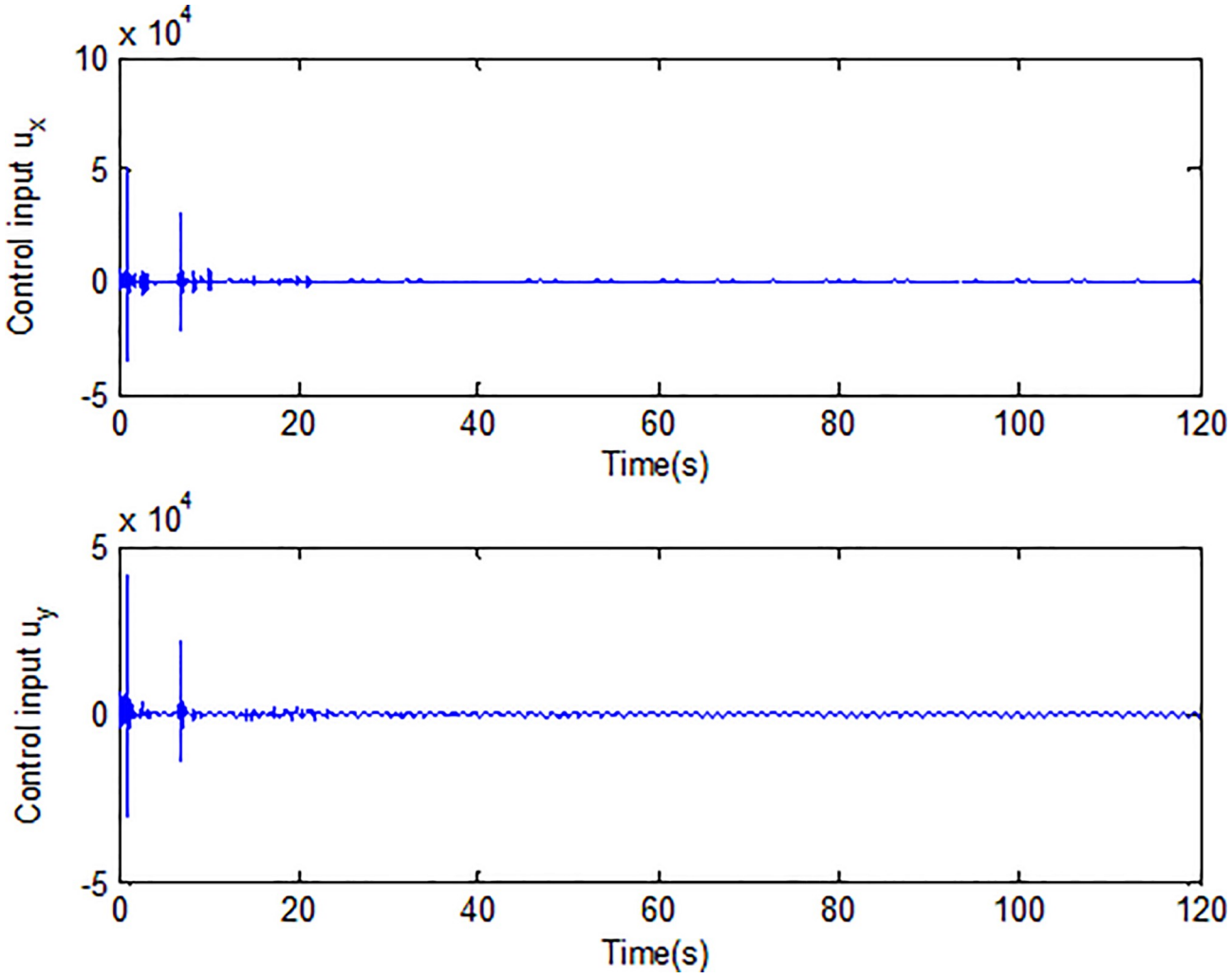

Fig 6. Control input signals.

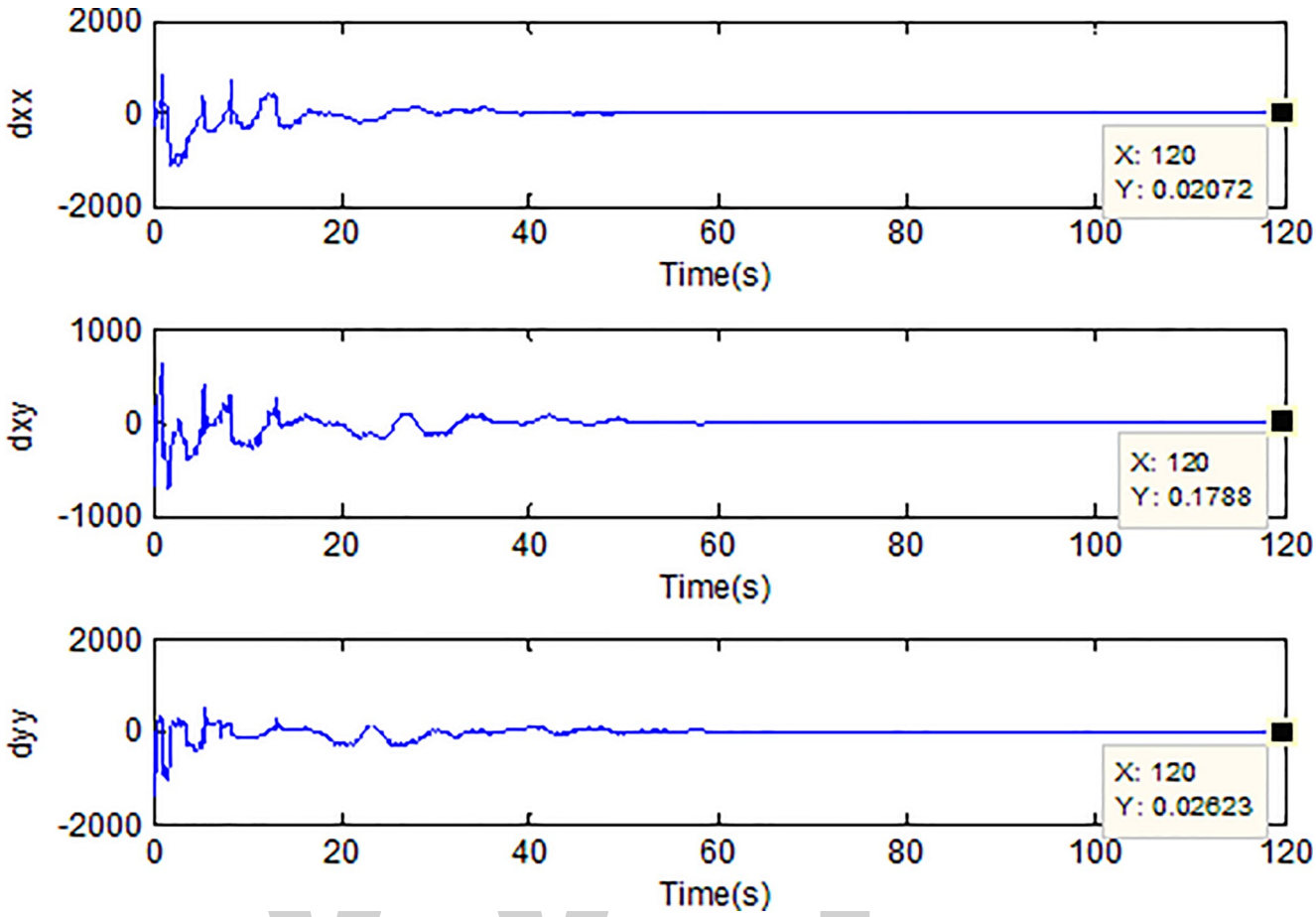

Fig 7. Adaption of damping coefficients of microgyroscope.

Since $D = D^T, K = K^T, \Omega = -\Omega^T$ and $s^T\tilde{D}(e_2 + \alpha_1) = (e_2 + \alpha_1)^T\tilde{D}s$ are scalar, we have

$$\begin{aligned}\lambda_3 s^T\tilde{D}(e_2+\alpha_1) &= \frac{1}{2}(\lambda_3 s^T\tilde{D}(e_2+\alpha_1) + \lambda_3(e_2+\alpha_1)^T\tilde{D}s) \\ &= tr\left\{\frac{1}{2}\lambda_3(\tilde{D}(e_2+\alpha_1)s^T + \tilde{D}s(e_2+\alpha_1)^T)\right\}\end{aligned} \tag{35}$$

Simultaneously, we obtained

$$\begin{aligned}\lambda_3 s^T\tilde{K}(e_1+q_r) &= \frac{1}{2}\left(\lambda_3 s^T\tilde{K}(e_1+q_r) + \lambda_3(e_1+q_r)^T\tilde{K}s\right) \\ &= tr\left\{\frac{1}{2}\lambda_3\left(\tilde{K}(e_1+q_r)s^T + \tilde{K}s(e_1+q_r)^T\right)\right\} \\ 2\lambda_3 s^T\tilde{\Omega}(e_2+\alpha_1) &= \lambda_3 s^T\tilde{\Omega}(e_2+\alpha_1) - \lambda_3(e_2+\alpha_1)^T\tilde{\Omega}s \\ &= tr\left\{\lambda_3\left(\tilde{\Omega}(e_2+\alpha_1)s^T - \tilde{\Omega}s(e_2+\alpha_1)^T\right)\right\}\end{aligned} \tag{36}$$

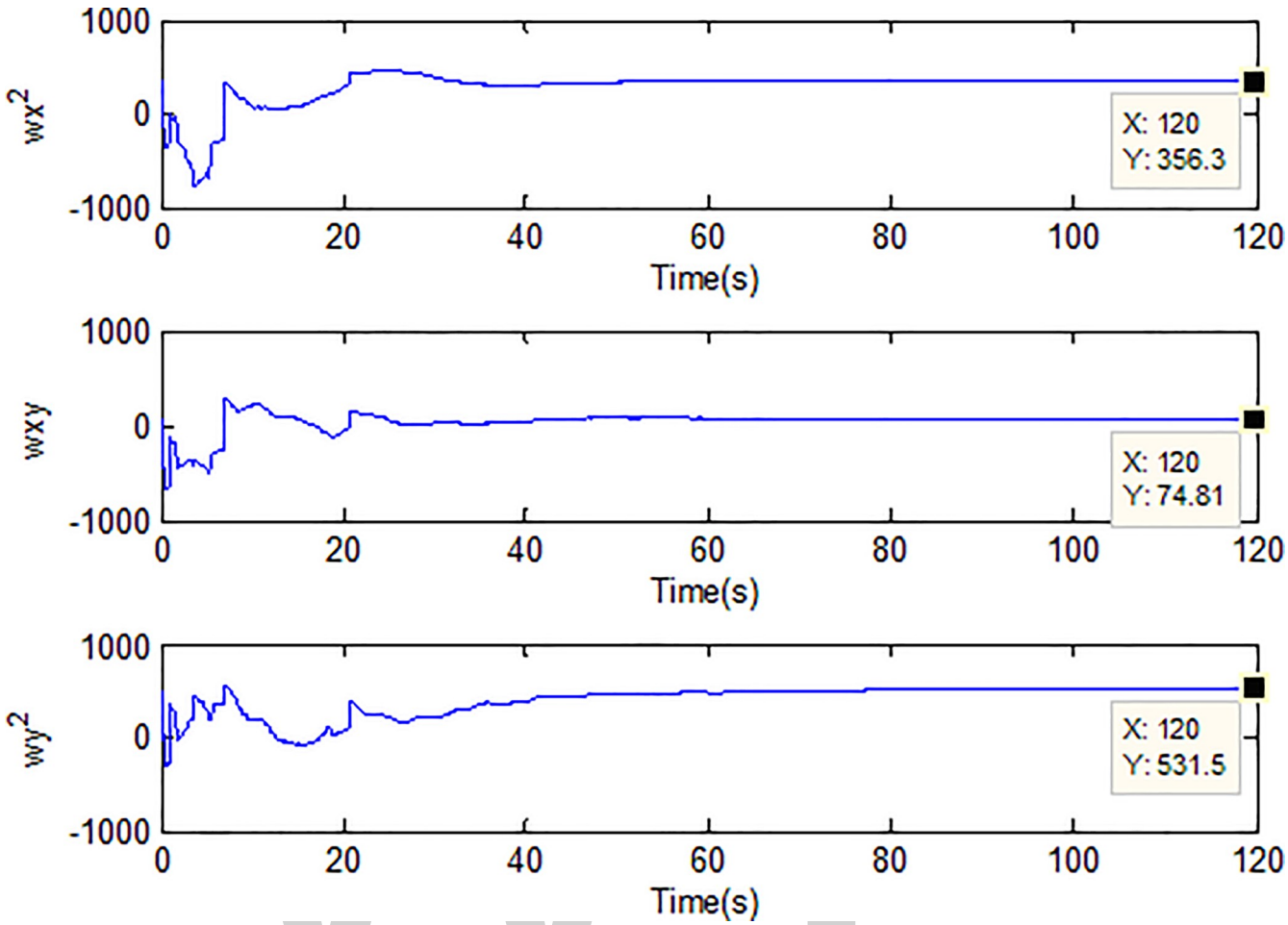

Fig 8. Adaption of spring constants of microgyroscope.

Substituting (35) and (36) into (34) results:

$$\dot{V} = -c_1 {e_1}^T e_1 - \frac{\lambda_1 {e_1}^T e_1}{\lambda_3} + tr\left\{\tilde{D}\left(M^{-1}\dot{\tilde{D}}^T + \frac{1}{2}\lambda_3\left((e_2+\alpha_1)s^T + s(e_2+\alpha_1)^T\right)\right)\right\}$$
$$+tr\left\{\tilde{K}\left(N^{-1}\dot{\tilde{K}}^T + \frac{1}{2}\lambda_3\left((e_1+q_r)s^T + s(e_1+q_r)^T\right)\right)\right\}$$
$$+tr\left\{\tilde{\Omega}\left(P^{-1}\dot{\tilde{\Omega}}^T + \lambda_3\left((e_2+\alpha_1)s^T - s(e_2+\alpha_1)^T\right)\right)\right\}$$
$$+\frac{1}{r}\sum_{i=1}^{2}{\tilde{\theta}_{hi}}^T\left(r\lambda_3 s_i \Phi(s_i) + \dot{\tilde{\theta}}_{hi}\right) + \lambda_3 s^T\left(d - \hat{h}(s|\theta^*)\right) \quad (37)$$

where $i = 1,2$ represents the two-axis vector.

In order to guarantee $\dot{V} \leq 0$, the online adapting laws for parameters are as follows:

$$\begin{aligned}
\dot{\hat{D}}^T &= \dot{\tilde{D}}^T = -\frac{1}{2}\lambda_3 M\left((e_2+\alpha_1)s^T + s(e_2+\alpha_1)^T\right) \\
\dot{\hat{K}}^T &= \dot{\tilde{K}}^T = -\frac{1}{2}\lambda_3 N\left((e_1+q_r)s^T + s(e_1+q_r)^T\right) \\
\dot{\hat{\Omega}}^T &= \dot{\tilde{\Omega}}^T = -\lambda_3 P\left((e_2+\alpha_1)s^T - s(e_2+\alpha_1)^T\right) \\
\dot{\theta}_{hi} &= -\dot{\tilde{\theta}}_{hi} = r\lambda_3 s_i \Phi(s_i), i = 1,2
\end{aligned} \tag{38}$$

Substituting (38) into (37), it is obvious that

$$\begin{aligned}
\dot{V} &= -c_1 {e_1}^T e_1 - \frac{\lambda_1 {e_1}^T e_1}{\lambda_3} + \lambda_3 s^T (d - \hat{h}(s|\theta^*)) \\
&\leq -c_1 {e_1}^T e_1 - \frac{\lambda_1 {e_1}^T e_1}{\lambda_3} + \lambda_3 \left(\|s\|\|d\| - \rho \frac{s^T s}{\|s\|} \right) \\
&\leq -c_1 {e_1}^T e_1 - \frac{\lambda_1 {e_1}^T e_1}{\lambda_3} + \lambda_3 (\|s\|\|d\| - \rho\|s\|) \\
&\leq -c_1 {e_1}^T e_1 - \frac{\lambda_1 {e_1}^T e_1}{\lambda_3} \leq 0
\end{aligned} \tag{39}$$

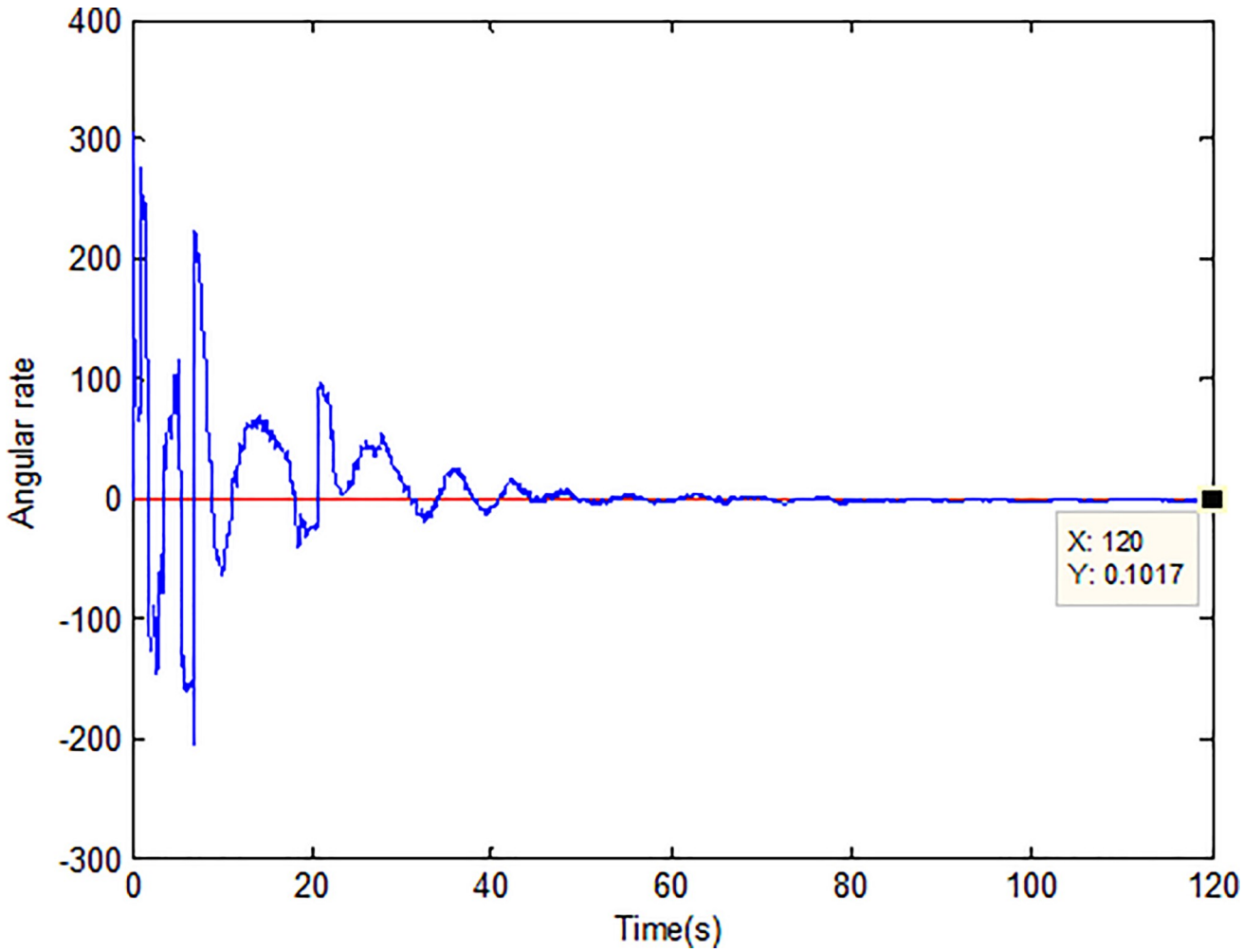

Fig 9. Adaption of angular velocity.

$\dot{V}$ is proved to be negative semi-definite which means $V, s, \tilde{D}, \tilde{K}, \tilde{\Omega}$ are all bounded. According to (32), $\dot{s}$ is also bounded. Integrating $\dot{V}$ with respect to time, we have $\int_0^t c_1 {e_1}^T e_1 + \frac{\lambda_1 {e_1}^T e_1}{\lambda_3} + \lambda_3(\|s\|\|d\| - \rho\|s\|) \leq V(0) - V(t)dt$. Since $V(0)$ is bounded and $V(t)$ is bounded and non-increasing, it can be concluded that $\int_0^t c_1 {e_1}^T e_1 + \frac{\lambda_1 {e_1}^T e_1}{\lambda_3} dt$ and $\int_0^t \lambda_3(\|s\|\|d\| - \rho\|s\|)dt$ are all bounded. According to Barbalart lemma, $\lim_{t\to\infty} e_1(t) = 0$, $\lim_{t\to\infty} s(t) = 0$, that is to say the tracking error and fractional sliding mode surface will asymptotically converge to zero which guarantees the stability of the gyroscope system.

Results and discussions

A z-axis MEMS gyroscope dynamical model is chosen as a simulation example to validate the effectiveness of the proposed control strategy. The parameters of the microgyroscope are chosen as follows:

$$m = 1.8 \times 10^{-7}\text{kg},\ d_{xx} = 1.8 \times 10^{-6}\,\text{N s/m}, d_{yy} = 1.8 \times 10^{-6}\,\text{N s/m},$$
$$d_{xy} = 3.6 \times 10^{-7}\,\text{N s/m},\ k_{xx} = 63.955\,\text{N/m}, k_{yy} = 95.92\,\text{N/m}, k_{xy} = 12.779\,\text{N/m}.$$

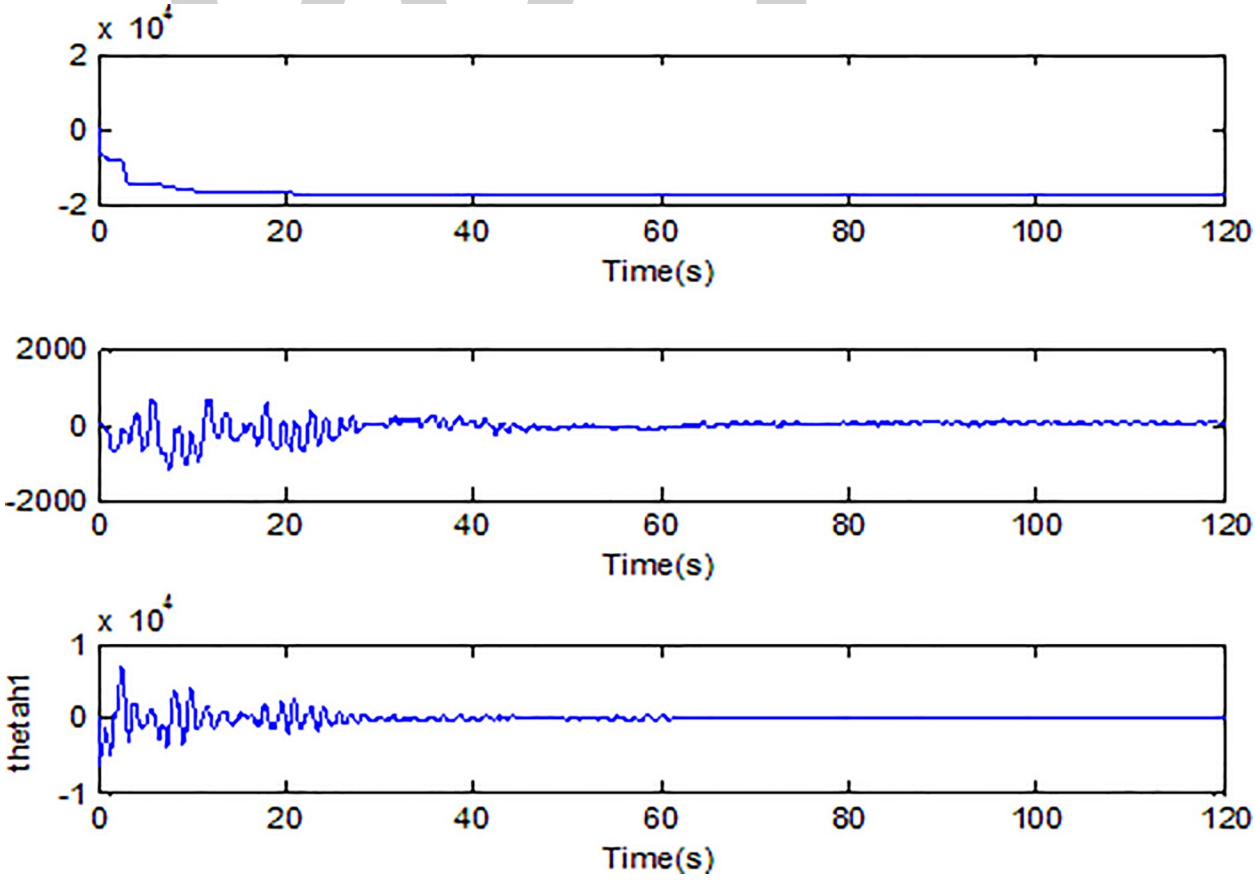

Fig 10. Adaption of θ_h along X axis.

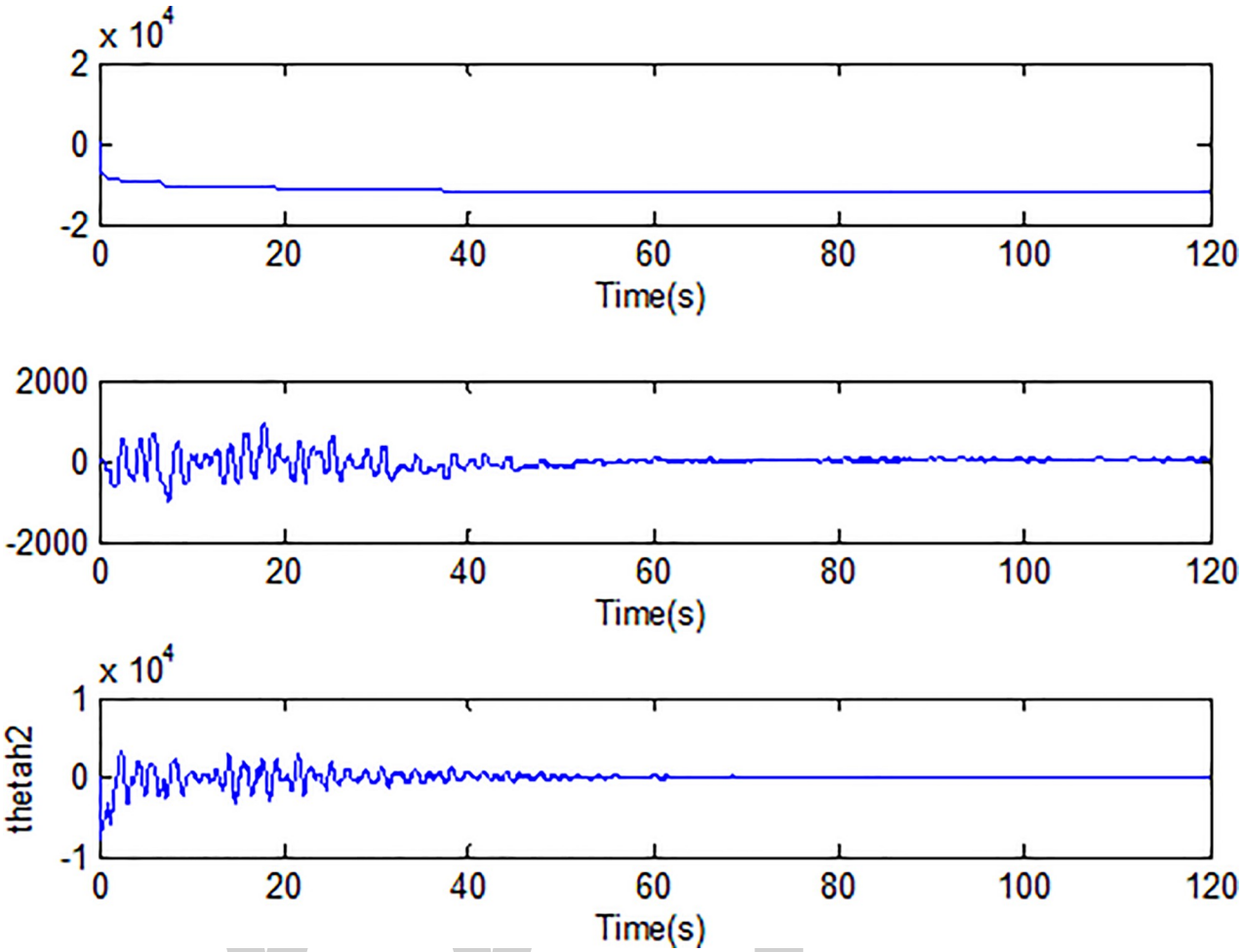

Fig 11. Adaption of θ_h along Y axis.

Assume that the unknown angular velocity is Ω_z = 100 rad/s. Then the non-dimensional gyroscope parameter matrices can be derived as follows:

$$D = \begin{bmatrix} 0.01 & 0.002 \\ 0.002 & 0.01 \end{bmatrix}, K = \begin{bmatrix} 355.3 & 70.99 \\ 70.99 & 532.9 \end{bmatrix}, \Omega = \begin{bmatrix} 0 & -0.1 \\ 0.1 & 0 \end{bmatrix} \tag{40}$$

The membership functions of the fuzzy variable s are defined as:

$$\begin{aligned} &\mu_{NM}(s) = 1/(1+\exp(5(s+3))), \mu_{ZO}(s) = \exp(-s^2) \\ &\mu_{PM}(s) = 1/(1+\exp(5(s-3))) \end{aligned} \tag{41}$$

In this simulation example, reference trajectory is selected as $q_{r1} = \sin(4.17t)$, $q_{r2} = 1.2\sin(5.11t)$ and the initial states of the system are set as $q_1(0) = 0.5, \dot{q}_1(0) = 0, q_2(0) = 0.5, \dot{q}_2(0) = 0$. Choose the initial conditions of $\hat{D}, \hat{K}, \hat{\Omega}$ as $\hat{D}_0 = 0.95{*}D$, $\hat{K}_0 = 0.95{*}K$, $\hat{\Omega}_0 = 0.95{*}\Omega$. Select the sliding surface parameters λ_1 = 55,λ_2=10,λ_3 = 1, the control parameters c_1 = 200, r = 10000 and the adaptive gains $M = N = diag(150,150), P = diag(20,20)$.

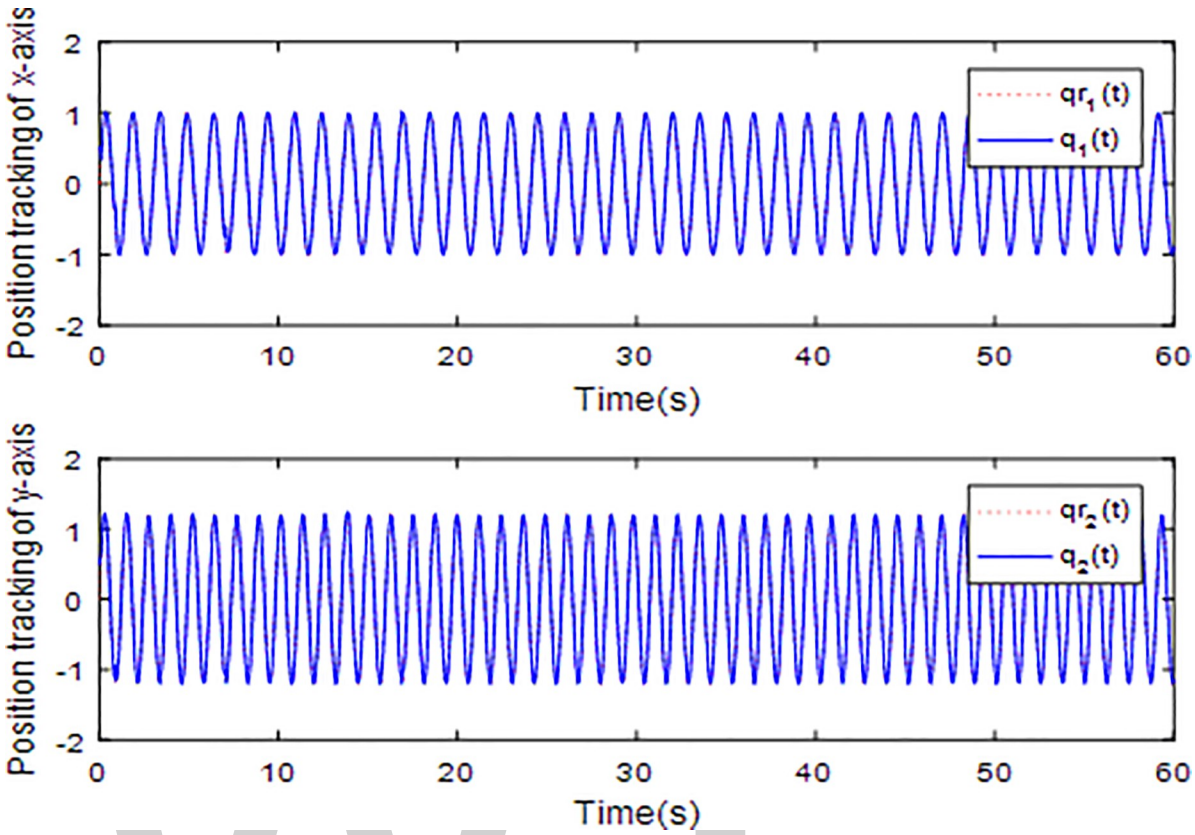

Fig 12. Tracking trajectory using integer order sliding surface.

When the fractional order is set as $\alpha = 0.9$ and the disturbance is applied as random signal $d = [0.5^{*}randn(1,1);0.5^{*}randn(1,1)]$, the corresponding simulation results are shown in Figs 3–10.

Fig 3 describes the trajectories of the system states. It is obvious that the tracking performance is well achieved with the existence of external disturbance by the proposed adaptive fractional fuzzy sliding mode control based on backstepping technique. Fig 4 plots the tracking error of the microgyroscope system which converges to zero in a short time and guarantees the asymptotical stability of the system. In addition, the tracking error under the condition of $\alpha = 0.9$ is demonstrated to be the lowest that will be introduced in detail below.

Fig 5 depicts the convergence of the fractional sliding surface s. It is intuitive that the sliding surface converges to zero within a short time which ensures that the trajectory of the system attains to sliding surface. In Fig 6, the time evolution of the input control signal is brought. The chattering is effectively reduced as a result of the approximation for switching function of fuzzy system. Fig 7 and Fig 8 draw the adaption of the system parameter matrix D and K respectively. With persistent sinusoidal signals, the estimation of D and K have been verified to converge to their true values which allows the existence of small range of errors. Fig 9 describes the adaption of angular velocity whose estimate also converges to its actual value. Fig 10 and Fig 11 depict the adaption of fuzzy parameter θ_h along X and Y axis respectively. It is obvious that the parameter reaches a steady state after 40 seconds.

Fig 12 and Fig 13 plot the tracking trajectories and tracing errors of microgyroscope along x-axis and y-axis respectively using integer sliding mode controller. It can be seen that the

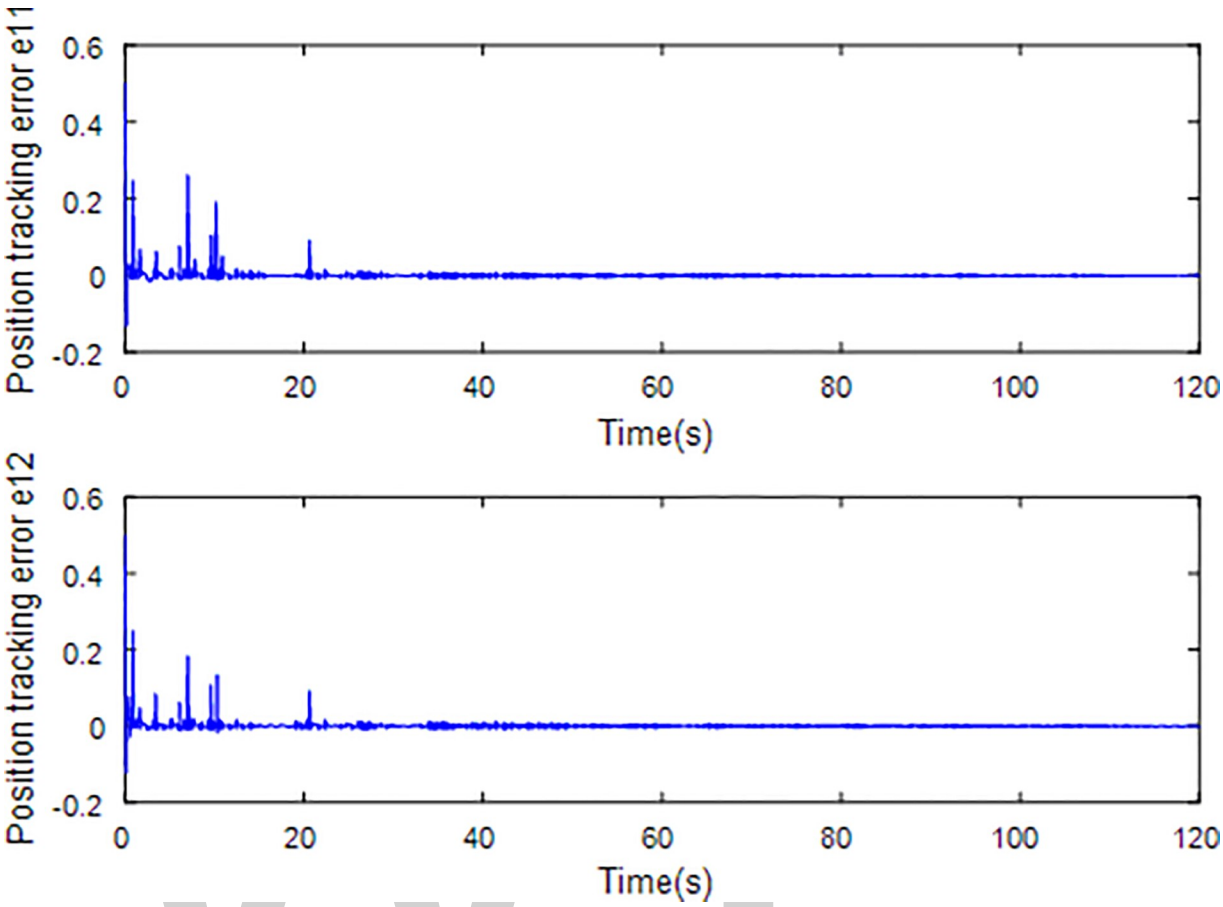

Fig 13. Tracking error using integer order sliding surface.

tracking performance also meet the expected requirements and the tracking error converges to zero asymptotically. However, compared to previous fractional sliding mode controller, the tracking performance seems to be a little inferior. In order to see the tracking performance under different fractional orders and integer order visually, a universal standard is used to quantify tracking error by calculating root mean square error (rms error). The rms error reflects how much the measured value deviates from the true value. The smaller the rms error is, the higher the measurement accuracy is. So, it can be a criterion to assess the tracking performance of the control scheme under different orders. Table 1 shows the rms errors along x-axis and y-axis under different fractional orders.

Table 1. RMS errors of x and y axis under different fractional orders.

RMS ERROR α	*X*	*Y*
0.1	0.0134	0.0121
0.3	0.0156	0.0125
0.5	0.0157	0.0120
0.7	0.0149	0.0133
0.9	0.0129	0.0118

For fairness, the fractional order α is added in a ladder-type increase and the integer order is set $\alpha = 1$. It is intuitive to see that the fractional order has impact on tracking errors. When fractional order $\alpha = 0.9$, the rms error seems to be minimal that is why we choose $\alpha = 0.9$ in previous design procedure. In the case of $\alpha = 1$, the rms errors along x-axis and y-axis are 0.0154 and 0.0131 which are slightly larger than the case of fractional order $\alpha = 0.9$. This effectively verified that the adaptive fractional fuzzy sliding mode control based on backstepping technique is superior to the conventional integer order ones.

Remark: The computational cost of the proposed fractional order sliding mode control and traditional integer order sliding mode control is about 2 minutes and 70 seconds.

Conclusion

An adaptive fractional fuzzy sliding mode controller for microgyroscope system based on backstepping design is presented in this paper. The object of the controller design is to make the output trajectory of microgyroscope track the reference trajectory accurately and effectively. Compared to the earlier control methods such as AGC technique and PLL technique, the proposed technique has advantages in terms of control accuracy and adaptability in engineering applications. Unlike traditional SMC with integer order, a fractional differential sliding surface is proposed which has more design freedom. Then a fuzzy system is incorporated into fractional sliding mode control to attenuate the chattering in the sliding phase. Furthermore, adaptive estimators are used to identify the angular velocity and other unknown system parameters. In order to find the best fractional order α for the system, simulations under different fractional orders are carried out, verifying the efficacy of the proposed control schemes. In the future research, we will focus on the design of hardware circuits and control method, build a test platform and complete the test of the microgyroscope system based on FPGA.

Author Contributions

Conceptualization: Juntao Fei.

Data curation: Xiao Liang.

Formal analysis: Xiao Liang.

Investigation: Xiao Liang.

Project administration: Juntao Fei.

Writing – original draft: Juntao Fei.

Writing – review & editing: Juntao Fei.

References

1. Park S, Horowitz R, Tan C W. Adaptive Control for MEMS Gyroscopes. 2002.
2. Leland R P. Adaptive control of a MEMS gyroscope using Lyapunov methods. IEEE Trans. on Control Systems Technology, 2006, 14(2):278–283.
3. Fei J, Lu C. Adaptive sliding mode control of dynamic systems using double loop recurrent neural network structure, IEEE Trans. on Neural Network and Learning System, 2018, 29(4): 1275–1286.
4. Fei J, Lu C. Adaptive fractional order sliding mode controller with neural estimator, Journal of the Franklin Institute, 2018, 355(5): 2369–2391.
5. Sue C Y. Integrated model reference adaptive control and time-varying angular rate estimation for micro-machined gyroscopes. International Journal of Control, 2010, 83(2):246–256.
6. Guo Y. Mean square exponential stability of stochastic delay cellular neural networks[J]. Electronic Journal of Qualitative Theory of Differential Equations, 2013, 16(34):1–10.

7. Guo Y. Exponential stability analysis of traveling waves solutions for nonlinear delayed cellular neural networks[J]. Dynamical Systems, 2017, 32(4):1–14.
8. Guo Y. Global stability analysis for a class of Cohen-Grossberg neural network models[J]. Bulletin of the Korean Mathematical Society, 2012, 49(6):1193–1198.
9. Guo Y. Globally Robust Stability Analysis for Stochastic Cohen–Grossberg Neural Networks with Impulse Control and Time-Varying Delays[J]. Ukrainian Mathematical Journal, 2018, 69(8):1220–1233.
10. Coronel-Escamilla A, Gómez-Aguilar J.F, Torres L, Escobar-Jimenez R. F, Valtierra-Rodriguez M. Synchronization of chaotic systems involving fractional operators of Liouville-Caputo type with variable-order[J]. Physica A: Statistical Mechanics and its Applications, 2017, 487, 1–21.
11. Atangana A, Gómez-Aguilar J. F. Decolonisation of fractional calculus rules: Breaking commutativity and associativity to capture more natural phenomena[J]. The European Physical Journal Plus, 2018, 133(4):1–23.
12. Atangana Abdon. Non validity of index law in fractional calculus: A fractional differential operator with Markovian and non-Markovian properties[J]. Physica A: Statistical Mechanics and its Applications, 2018, 505: 688–706.
13. Atangana A, Gómez-Aguilar J. F. Numerical approximation of Riemann- Liouville definition of fractional derivative: From Riemann-Liouville to Atangana-Baleanu[J]. Numerical Methods for Partial Differential Equations, 2017.
14. Indranil P.; Saptarshi D. Kriging based surrogate modeling for fractional order control of microgrids. IEEE Trans. on Smart Grid, 2015, 1, 36–44.
15. Atangana A, Gómez-Aguilar J.F. Hyperchaotic behaviour obtained via a nonlocal operator with exponential decay and Mittag-Leffler laws[J]. Chaos, Solitons & Fractals, 2017.
16. Coronel-Escamilla A, Torres F, Gómez-Aguilar J. F, Escobar-Jiménez R.F, Guerrero-Ramirez G.V. On the trajectory tracking control for an SCARA robot manipulator in a fractional model driven by induction motors with PSO tuning[J]. Multibody System Dynamics, 2018, 43(3):257–277.
17. Ladaci S, Charef A. On Fractional Adaptive Control [J]. Nonlinear Dynamics, 2006, 43(4):365–378.
18. Yang X J, Machado J A T, Cattani C, Gao F. On a fractal LC-electric circuit modeled by local fractional calculus [J]. Communications in Nonlinear Science & Numerical Simulation, 2017, 47:200–206.
19. Guo Y. Solvability for a nonlinear fractional differential equation [J]. Bulletin of the Australian Mathematical Society, 2009, 80(1):125–138.
20. Ghanbari K, Gholami Y, Mirzaei H. Nontrivial solutions for boundary-value problems of nonlinear fractional differential equations[J]. Bulletin of the Korean Mathematical Society, 2010, 47(1):81–87.
21. Guo Y. Solvability of boundary value problems for a nonlinear fractional differential equations[J]. Ukrainian Mathematical Journal, 2010.
22. Wu L, Mazumder S K, Kaynak O. Sliding Mode Control and Observation for Complex Industrial Systems—Part I. IEEE Trans. on Industrial Electronics, 2017, 64(8):6680–6683.
23. Li H, Wang J, Lam H K, Zhou Q. Adaptive Sliding Mode Control for Interval Type-2 Fuzzy Systems. IEEE Trans. on Systems Man & Cybernetics Systems, 2016, (99):1–10.
24. Fei J., Ding H., Adaptive Sliding Mode Control of Dynamic System Using RBF Neural Network, Nonlinear Dynamics. 70(2): 1563–1573, 2012.
25. Baek J, Jin M, Han S. A New Adaptive Sliding-Mode Control Scheme for Application to Robot Manipulators. IEEE Trans. on Industrial Electronics, 2016, 63(6):3628–3637.
26. Fei J, Wang T. Adaptive Fuzzy-Neural-Network Based on RBFNN Control for Active Power Filter, International Journal of Machines Learning and Cybernetics, 2019(10):1139–1150.
27. Zhu Y., Fei J., Disturbance Observer Based Fuzzy Sliding Mode Control of PV Grid Connected Inverter, IEEE Access, 6: 21202–21211, 2018.
28. Hou S., Fei J., Chen C., Finite-Time Adaptive Fuzzy-Neural-Network Control of Active Power Filter, IEEE Trans. on Power Electronics, https://doi.org/10.1109/TPEL.2019.2893618, 2019.
29. Chu Y, Fei J. Dynamic Global PID Sliding Mode Control Using RBF Neural Compensator for Three-Phase Active Power Filter, Transactions of the Institute of Measurement and Control. 2018, 40 (12): 3549–3559.
30. Chen L, Wu R, He Y, Chai Y. Adaptive sliding-mode control for fractional-order uncertain linear systems with nonlinear disturbances. Nonlinear Dynamics, 2015, 80(1–2):51–58.
31. Fang Y., Fei J., Cao Di, Adaptive Fuzzy-Neural Fractional-Order Current Control of Active Power Filter with Finite-Time Sliding Controller, International Journal of Fuzzy System, https://doi.org/10.1007/s40815-019-00648-4, 2019.

32. Fang Y., Fei J., Yang Y., Adaptive Backstepping Design of a Microgyroscope, Micromachines, 9 (7):338, 2018. https://doi.org/10.3390/mi907033
33. Tong SC, Li YM. Adaptive fuzzy output feedback tracking backstepping control of strict-feedback nonlinear systems with unknown dead zones. IEEE Trans. on Fuzzy Systems 2012; 20(1):168–180.
34. Fang Y, Fei J. Hu T., Adaptive Backstepping Fuzzy Sliding Mode Vibration Control of Flexible Structure, Journal of Low Frequency Noise Vibration and Active Control, 2018, 37(4): 1079–2096.
35. Yoshimura T. Adaptive backstepping discrete-time control for a full-car active suspension. International Journal of Vehicle Autonomous Systems, 2017, 13(3):221.
36. Sun W, Gao H, Kaynak O. Adaptive Backstepping Control for Active Suspension Systems With Hard Constraints. IEEE/ASME Trans. on Mechatronics, 2013, 18(3):1072–1079.
37. Mustafa A, Dhar N, Agrawal P, Yerma N K. Adaptive backstepping sliding mode control based on nonlinear disturbance observer for trajectory tracking of robotic manipulator, International Conference on Control and Robotics Engineering. IEEE, 2017:29–34.
38. Park S, Taesung Y, Kwak G, Ahn H. T-S fuzzy control of IPMSM using backstepping integral sliding mode, International Conference on Control, Automation and Systems. IEEE, 2015:1113–1118.
39. Liu C., Cai G., Gao J. and Yang D. Design of Nonlinear Robust Damping Controller for Power Oscillations Suppressing Based on Backstepping-Fractional Order Sliding Mode. Energies, 2017, 10(5), 676.
40. Fei J., Liang X., Adaptive Backstepping Fuzzy-Neural-Network Fractional Order Control of Microgyroscope Using Nonsingular Terminal Sliding Mode Controller, Complexity, 2018. https://doi.org/10.1155/2018/5094179
41. Feng Z, Fei J. Design and Analysis of Adaptive Super-Twisting Sliding Mode Control for a Microgyroscope, PLOS ONE, 13(1), https://doi.org/10.1371/journal.pone.0189457, 2018. PMID: 29298297

Controlling and synchronizing a fractional-order chaotic system using stability theory of a time-varying fractional-order system

Yu Huang[1], Dongfeng Wang[2], Jinying Zhang[3], Feng Guo[4]*

1 Department of Automation, North China Electric Power University, Baoding, China, **2** Department of Automation, North China Electric Power University, Baoding, China, **3** Shenhua Guohua Electric Power Research Institute Corporation, Beijing, China, **4** Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, Xiamen, China

* betop@xmu.edu.cn

Abstract

Control and synchronization of fractional-order chaotic systems have attracted wide attention due to their numerous potential applications. To get suitable control method and parameters for fractional-order chaotic systems, the stability analysis of time-varying fractional-order systems should be discussed in the first place. Therefore, this paper analyzes the stability of the time-varying fractional-order systems and presents a stability theorem for the system with the order $0<\alpha<1$. This theorem is a sufficient condition which can discriminate the stability of time-varying systems conveniently. Feedback controllers are designed for control and synchronization of the fractional-order Lü chaotic system. The simulation results demonstrate the effectiveness of the proposed theorem.

Editor: Jun Ma, Lanzhou University of Technology, CHINA

1. Introduction

Fractional-order calculus which extends the descriptive abilities of integer-order calculus can be traced to the work of Leibniz and Hospital in 1695. The integer-order calculus depends only on the local characteristics of a function's, but fractional-order calculus accumulates all information of the function in a certain time, which is also called memory property. Mathematical models based on fractional-order calculus can describe the dynamic behavior of an actual system accurately in many areas, thereby it is necessary to facilitate the improvement of its design and control stability for fractional-order dynamic systems [1]. Recently, fractional-order chaotic control and synchronization have attracted increasing attention. In [2], Razminia A *et al.* synchronized a unidirectional coupling structure for the two fractional order chaotic systems using a sliding mode control methodology. In [3], Wu GC *et al.* presented a nonlinear coupling method to study the master-slave synchronization for the fractional differential equation. In [4], Golmankhaneh AK *et al.* have presented the chaos synchronization of two identical and non-identical fractional orders of a new chaotic system by using active control. In [5], Jajarmi A *et al.* used a linear state feedback controller together with an active control technique in order to control a hyperchaotic financial system. In [6], a Lyapunov approach is adopted for deriving the

Funding: This work was supported by the National Key R&D Program of China (2016YFB0600701) and the Fundamental Fund for the Central Universities of China under Grant 2015MS66. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

parameter adaptation laws and proving the stability of the generalized projective synchronization (GPS) of two incommensurate fractional-order chaotic closed-loop systems. A linear feedback controller is proposed to achieve synchronisation of a fractional-order system with uncertainties and disturbance and guarantees the bounded state error for any bounded interference infinite time [7]. In [8], a simple but practical method to synchronize almost all familiar fractional-order chaotic systems which are including the commensurate system and incommensurate case, autonomous system, and the nonautonomous case has been put forward, and sufficient conditions are derived to guarantee synchronization of these systems. In [9], Shao SY *et al.* studies the fractional-order disturbance observer (FODO)-based adaptive sliding mode synchronization control for a class of fractional-order chaotic systems with unknown bounded disturbances. In [10], Soukkou A *et al.* proposed a fractional-order prediction-based feedback control scheme (Fo-PbFC) to stabilize the unstable equilibrium points and to synchronize the fractional-order chaotic systems (FoCS). In [11], Nourian *et al.* estimated the unknown coefficients of the system and demonstrated the stabilization of the synchronizer system by using the adaptive rule and a proper Lyapunov candidate function. In [12], Maheri *et al.* put forward a robust adaptive nonlinear feedback controller scheme to realize the synchronization of two different fractional-order chaotic systems in the condition of fully unknown parameters, external disturbance and uncertainties. In [13], Zhou *et al.* designed an adaptive controller to synchronize two entirely different fractional-order chaotic systems with uncertain parameters. Combining with appropriate parameter estimation laws. In [14], Yang proposed a single-state proportional feedback method to synchronize two identical generalized Lorenz systems. Used Lyapunov stability theory and a fractional-order differential inequality. In [15], Zhang *et al.* developed a modified adaptive control scheme and adaptive parameter laws to robustly synchronize coupled with fractional-order chaotic systems without certain parameters and perturbations. In [16], Xiang *et al.* investigated a robust synchronization for a class of systems with external disturbances.

In addition, many scholars have made great contributions in the field of the control and stability of time-varying fractional order systems. In [17], Aguila-Camacho N *et al.* put forward a new lemma for the Caputo fractional derivatives which has been proved to be useful in order to find the fractional-order extension of Lyapunov functions and can be used to demonstrate the stability of many fractional order systems including nonlinear and time-varying. In [18], Bao HB *et al.* put forward sufficient conditions which ensure the drive–response systems to achieve adaptive synchronization of fractional-order memristor-based neural networks with time-varying delay. In [19], the authors dealt with the fractional-order neural networks with impulsive effects and time-varying delay, and established several sufficient conditions guaranteeing the global Mittag–Leffler stability of the equilibrium point of the neural networks.

However, the most basic control and synchronization problem of chaotic systems are that of stability. Stability is a precondition for normal operation of systems and the main factor of system designs. A Lyapunov direct method is a core issue in integer-order stability theory, which is also a basic stability theorem for control systems.

It has been proven that the Lyapunov direct method is a relatively complete theoretical for integer-order systems both in theoretical study and practical application. As the transfer function of fractional-order systems is usually not a rational function of complex variable s, the stability analysis of fractional-order systems is far more complicated than that of integer-order systems. Many scholars have carried out extensive research on time-invariant fractional-order systems and made considerable achievements. For fractional-order LTI systems, in [20], Semary *et al.* discussed their physical and non-physical transfer functions, stability, poles, time domain, frequency domain, their relationships for different fractional-order differential equations and other basic concepts. In [21], Wang *et al.* used the argument principle of complex analysis to deduce two stability criteria for linear time-invariant fractional-order systems,

which can determine system stability without utilizing characteristic roots. They also used Laplace transform and residue theorem to discuss the internal and external stability conditions of a linear time-invariant fractional-order system [22]. Pakzad put forward a practical analytical model to discuss the stability robustness of a class of linear time-invariant fractional-order systems with single and multiple commensurate delays of retarded type, against delay uncertainties [23].

All the above stability analyses are for time-invariant fractional-order systems. However, the above results are not widely used due to various reasons. For example, the eigenvalue criterion cannot be applied in time-varying fractional-order systems [24]. Therefore, this paper analyzes the stability of the time-varying fractional-order systems and presents a stability theorem for the system with the order $0<\alpha<1$. This theorem is a sufficient condition which can discriminate the stability of time-varying systems conveniently. Feedback controllers are designed for control and synchronization of the fractional-order Lü chaotic system.

The rest of the paper is organized as follows. Section 2 analyzes the development status and the stability of fractional-order systems. Section 3 presents a stability theorem for these systems with the order $0<\alpha<1$. Feedback controllers for fractional-order Lü chaotic system's control and synchronization are designed on the basis of previous stability theorem in Section 4. Finally, the conclusion is drawn according to the present study in Section 5.

2. Development status of fractional-order system and stability

2.1 Definition of fractional-order calculus

Nowadays, many different definitions of fractional-order calculation were presented, in [25]. The most common definition, with $\alpha\epsilon(0,1)$, is shown as Eq 1 and was proposed by M. Caputo in 1967. Eq 1 is important for integral transformation because the initial value expressions generated in integral transformation are all in the form of integer-order derivatives, which can be effectively applied in practice.

$$ {}_{t_0}I_t^{\alpha}x(t) = \frac{1}{\Gamma(\alpha)}\int_{t_0}^{t}\frac{x(\tau)}{(t-\tau)^{1-\alpha}}d\tau, \tag{1}$$

Where $x(t)$ is a function with an arbitrary integer order; the fractional order meets $0<\alpha<1$; ${}_{t_0}I_t^{\alpha}$ is a fractional-order integral with order α of function $x(t)$ between $[t_0,t]$; $\Gamma(\cdot)$; denotes the gamma function.

Definition 1 For any real number q, $\lfloor q \rfloor$ denotes the integer part of q, that is to say, $\lfloor q \rfloor$ is the largest integer no more than q. ${}_{t_0}D_t^{\alpha}$ is a Caputo fractional differential operator. Thus, the differential of $x(t)$ with fractional-order q is

$$ {}_{t_0}D_t^{q}x(t) = \frac{1}{\Gamma(\lfloor q \rfloor + 1 - q)}\int_{t_0}^{t}\frac{x^{(\lfloor q \rfloor+1)}(\tau)}{(t-\tau)^{q-\lfloor q \rfloor}}d\tau \tag{2}$$

2.2 Development of stability analysis of fractional-order system

Theorem 1 When $0<\alpha<1$, $x\in R^n$, $A\in R^{nxn}$, the fractional-order system ${}_{t_0}D^{\alpha}{}_t x(t) = Ax(t)$, $t\geq t_0$ is asymptotically stable if and only if all the characteristic values of matrix A satisfy $|\arg(\mathrm{eig}(A))|>\alpha\pi/2$. Furthermore, the system is stable if and only if all the characteristic values of matrix A satisfy $|\arg(\mathrm{eig}(A))|\geq\alpha\pi/2$, which can be found in [21].

Theorem 1 is the existing stability criterion of a linear time-invariant fractional-order system with $0<\alpha<1$. This theorem is suitable only for a linear time-invariant fractional-order system, but it is often misused [26]. For time-invariant fractional-order nonlinear systems, if all the eigenvalues of the Jacobi matrix at equilibrium are stable, then the equilibrium is called

stable equilibrium point. However, Theorem 1 is not suitable for time-varying fractional-order systems.

Considering the time-varying fractional-order system with order $0<\alpha<1$ and initial value $x(t_0)$, the following is obtained:

$$ {}_{t_0}D_t^{\alpha}x(t) = f(t,x) \tag{3}$$

Where $\alpha\in(0,1)$, $f:[t_0,\infty]\times\Omega\rightarrow R^n$ is piecewise continuous and meets the local Lipschitz condition ($\Omega\in R^n$is a domain that contains $x = 0$).

Definition 2 A continuous function β: $[0,t)\rightarrow[0,\infty)$ is said to belong to class-k if it is strictly increasing and $\beta(0) = 0$.

Definition 3 If and only if $f(t,x_e) = {}_{t_0}D_t^{\alpha}x_e$, then constant x_e is the equilibrium point of the Caputo-defined fractional-order dynamic system (3). Without loss of generality, we assume $x_e = 0$.

Theorem 2 [25] Let $x_e = 0$ be an equilibrium point of the fractional-order system (3). Assume that Lyapunov function $V(t,x(t))$ and class-k functions $\beta_i(i = 1, 2, 3)$ exist, which satisfy

$$\beta_1(\|x\|) \le V(t,x) \le \beta_2(\|x\|), \tag{4}$$

$$ {}_0D_t^{\gamma}V(t,x(t)) \le -\beta_3(\|x\|), \tag{5}$$

Where $\gamma\in(0,1)$, then the equilibrium point of the system (3) is asymptotically stable.

3. Stability of time-varying fractional-order systems

3.1. Fractional-order system stability analysis

For linear time-varying fractional-order systems, the system (3) can be generally described in the following form:

$$ {}_{t_0}D_t^{\alpha}x(t) = \boldsymbol{A}(\boldsymbol{t})x(t), t \ge t_0 \tag{6}$$

For the system (6), we present stability Theorem 3 after introducing Lemma 1 as follows:

Lemma 1 For a continuous function $f(x) = x^T Ax$, $x \in R^{nx1}$, if $A\in R^{nxn}$ is a positive definite matrix, then

$$\lambda_{\min}(\boldsymbol{A})\|x\|^2 \le f(x) \le \lambda_{\max}(\boldsymbol{A})\|x\|^2 \tag{7}$$

in which $\lambda_{max}(\cdot)$ and $\lambda_{min}(\cdot)$ are the maximum and minimum eigenvalues, respectively, of the corresponding matrix.

Theorem 2 provides a guiding stability determination framework for general fractional-order systems, but its complexity is inconvenient when analyzing specific problems. Furthermore, Theorem 1 is not suitable for time-varying systems (6) [21]. Hence, for the time-varying fractional-order system (6), a stability analysis method will be given, we define a real symmetric matrix $H(t)$as follows:

$$\boldsymbol{H}(\boldsymbol{t}) = \boldsymbol{A}(\boldsymbol{t}) + \boldsymbol{A}^T(\boldsymbol{t}) \tag{8}$$

As ***H(t)*** is a real symmetric matrix, whose eigenvalues are real numbers. Let λ_{min} and λ_{max} be the respective minimum and maximum eigenvalues of $H(t)$. We thus obtain the following asymptotic stability sufficiency with Theorem 3.

Theorem 3 The sufficient condition for asymptotic stability of the fractional-order system (6) with equilibrium point x_e=0 is that the maximum eigenvalue of $H(t)$ satisfies $\lambda_{max} < 0$.

Proof: Let $\boldsymbol{V}(x,t) = x^T(t)x(t) = \|x(t)\|^2$,
Taking the α-order derivative of $\mathbf{V}(x,t)$with respect to time t, we have

$$\begin{aligned} V^{\alpha}(x,t) &= \frac{d^{\alpha}}{dt^{\alpha}}x^T(t)\cdot x(t) + x^T(t)\cdot\frac{d^{\alpha}}{dt^{\alpha}}x(t) \\ &= x^T(t)[\boldsymbol{A^T}(\boldsymbol{t}) + \boldsymbol{A}(\boldsymbol{t})]x(t) \\ &= x^T(t)\boldsymbol{H}(\boldsymbol{t})x(t) \end{aligned}$$

As $\lambda_{\min}$ and $\lambda_{\max}$are the minimum and maximum eigenvalues of the real symmetric matrix ***H(t)***, respectively, according to **Lemma 1**, we have

$$\lambda_{\min}\|x(t)\|^2 \le x^T(t)\boldsymbol{H}(\boldsymbol{t})x(t) \le \lambda_{\max}\|x(t)\|^2$$

Therefore $V^{\alpha}(x,t) \le \lambda_{\max}\|x(t)\|^2$

Considering the theorem's condition $\lambda_{\max} < 0$, we can easily obtain the conclusion because *V(x,t)*satisfies the condition (1) of Theorem 2, and V^{α}(x,t) satisfies the condition (2) of Theorem 2. Hence, according to Theorem 2, the time-varying fractional-order system (6) with equilibrium point x_e is asymptotically stable.

3.2 Examples of fractional-order system stability analysis

The stability analysis of two typical systems is given to demonstrate the effectiveness of the proposed stability theory.

Example 1: Consider the linear time-varying fractional-order system ($\alpha = 0.95$)

$${}_{t_0}D_t^{\alpha}x(t) = \begin{bmatrix} ((-b+a)+a\cos(\omega t))x_1 + (b - a\sin(\omega t))x_2 \\ (-b - a\sin(\omega t))x_1 + ((-b+a) - a\cos(\omega t))x_2 \end{bmatrix}. \tag{9}$$

The system matrix of (9) is

$$\boldsymbol{A}(\boldsymbol{t}) = \begin{bmatrix} (-b+a)+a\cos(\omega t) & b - a\sin(\omega t) \\ -b - a\sin(\omega t) & (-b+a) - a\cos(\omega t) \end{bmatrix}.$$

We assume that a = 0.25, b = 1, ω = 2, the real symmetric matrix *H(t)* is:

$$\boldsymbol{H}(\boldsymbol{t}) = \boldsymbol{A}(\boldsymbol{t}) + \boldsymbol{A^T}(\boldsymbol{t}) = \begin{bmatrix} -1.5 + 0.5\cos(2t) & -0.5\sin(2t) \\ -0.5\sin(2t) & -1.5 - 0.5\cos(2t) \end{bmatrix}.$$

The eigenvalues of *H(t)*can be obtained:

$$\lambda_1 = -1, \lambda_2 = -2.$$

All the eigenvalues are negative, we can conclude that system (9) is stable from Theorem 3.

Example 2: Consider the following linear time-varying fractional-order system

$${}_{t_0}D_t^{\alpha}\dot{x}(t) = \begin{bmatrix} (-b + a\sin\omega t)x_1 + a(\cos\omega t)x_2 \\ a(\cos\omega t)x_1 + (-b - a\sin\omega t)x_2 \end{bmatrix}. \tag{10}$$

The system matrix $A(t)$ and the corresponding real symmetric matrix $H(t)$ of (10) are acquired as follows:

$$\boldsymbol{A}(\boldsymbol{t}) = \begin{bmatrix} -b + a\sin\omega t & a\cos\omega t \\ a\cos\omega t & -b - a\sin\omega t \end{bmatrix}, \boldsymbol{H}(\boldsymbol{t}) = \begin{bmatrix} -2b + 2a\sin\omega t & 2a\cos\omega t \\ 2a\cos\omega t & -2b - 2a\sin\omega t \end{bmatrix}$$

The eigenvalues of $H(t)$ can be calculated: $\lambda_{1,2} = -2b \pm 2a$ Taking a = 0.25, b = 0.5 ($\lambda_1 = -0.5$, $\lambda_2 = -1.5$),this system is stable according to Theorem 3. And if we take a = 1, b = 0.5($\lambda_1 = 1$,$\lambda_2 = -3$), it is unable to determine the stability of this system from Theorem 3.

These examples show that Theorem 3 can discriminate the stability of time-varying fractional-order systems accurately. However, it is worth noting that this theorem is only a sufficient condition rather than a sufficient and necessary condition.

4. Control and synchronization of fractional-order Lü chaotic system

With the global boom of complex network research [27], chaotic systems as a part of complex networks are being widely applied [28]. The robust control and synchronization of chaotic systems have been gaining increasing attention. However, because of the lack of a stability analysis method for fractional-order systems, no systematic solution exists for the control and synchronization of a fractional-order chaotic system. With the use of the time-varying fractional-order stability theorem proposed in this paper, two controllers are designed for the fractional-order Lü chaotic system's tracking control and synchronization.

4.1 Tracking control of fractional-order Lü chaotic system

The mathematic model of fractional-order Lü chaotic system is described as follows [26]:

$${}_{t_0}D_t^{\alpha}x(t) = g(x) = \begin{bmatrix} 30(x_2 - x_1) \\ -x_1x_3 + 22.2x_2 \\ x_1x_2 - 8.8x_3/3 \end{bmatrix}. \tag{11}$$

Evidently, the above system is a typical nonlinear fractional-order system. To make the fractional-order Lü chaotic system (11) stable, the controller is designed as follows:

$$u(t) = Kx(t). \tag{12}$$

Where $u = [u_1,u_2,u_3]^T$, $x = [x_1,x_2,x_3]^T$, $k = [k_1,k_2,k_3]^T$, and the real number k_1,k_2,k_3 must be selected properly. By exerting the control action (12) into the system (11), we obtain

$${}_{t_0}D_t^{\alpha}x(t) = \begin{bmatrix} 30(x_2 - x_1) + u_1 \\ -x_1x_3 + 22.2x_2 + u_2 \\ x_1x_2 - 8.8x_3/3 + u_3 \end{bmatrix}. \tag{13}$$

We then simplify the controller as $u_1 = 0$,$u_2 = kx_2$,$u_3 = 0$ Single-variable linear feedback needs to be used to control the system. The controlled fractional-order Lü system can thus be

written in the form of the following linear time-varying fractional-order system:

$$ {}_{t_0}D_t^{\alpha}x(t) = \begin{bmatrix} -30 & 30 & 0 \\ 0 & k+22.2 & -x_1 \\ 0 & x_1 & -8.8/3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \tag{14} $$

The matrix ***A(t)*** of the controlled system (14) is

$$ \boldsymbol{A}(t) = \begin{bmatrix} -30 & 30 & 0 \\ 0 & k+22.2 & -x_1 \\ 0 & x_1 & -8.8/3 \end{bmatrix} \tag{15} $$

The matrix ***A(t)*** the control parameter k and is a function of the state variable x_1. Thus, it is a time-varying matrix even if the control parameter k is fixed. According to Eq 8, ***H(t)***is

$$ \boldsymbol{H}(t) = \begin{bmatrix} -60 & 30 & 0 \\ 30 & 2(k+22.2) & 0 \\ 0 & 0 & -2*8.8/3 \end{bmatrix} \tag{16} $$

By solving $det(\lambda I – H(t)) = 0$, we obtain the eigenvalues of ***H(t)***.Combining the root locus analysis, we determine that all the eigenvalues of ***H(t)*** are less than 0 if $k<-29.7$. According to Theorem 3, the controlled Lü system (13) is uniformly asymptotically stable.

Fig 1. shows the fractional-order Lü system (13) controlled to the zero equilibrium point with k = -35.

The solid lines in Fig 1. shows the motion curves of each state of the fractional-order Lü chaotic system when control action is added at t = 10s. Clearly, the system gradually converges to equilibrium point S_0 = (0,0,0) after the control action is added. The above design shows that we can easily obtain a feedback control parameter k to make the system stable using ***H(t)***-based Theorem 3. Given the time-varying state x_1 contained in ***A(t)***, obtaining a feedback control parameter k using ***A(t)***-based Theorem 1 is difficult.

4.2 Synchronization of fractional-order Lü chaotic system

System (11) is selected as the driving system

$$ {}_{t_0}D_t^{\alpha}x(t) = \begin{bmatrix} 30(x_2 - x_1) \\ -x_1x_3 + 22.2x_2 \\ x_1x_2 - 8.8x_3/3 \end{bmatrix}. \tag{17} $$

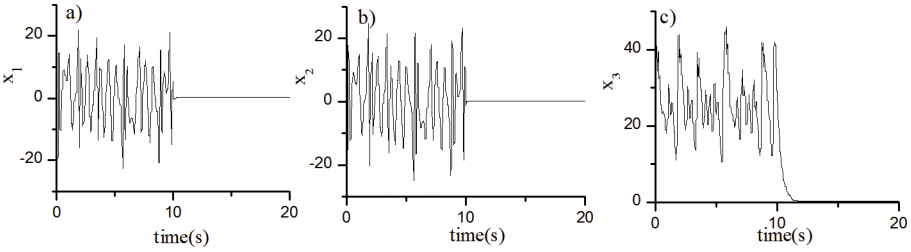

Fig 1. The state diagram of fractional-order Lü chaotic system with robust controller. The motion curves of each state of the fractional-order Lü chaotic system when control action is added at t = 10s.

The response system is

$$ {}_{t_0}D_t^{\alpha} y(t) = \begin{bmatrix} 30(y_2 - y_1) + u_1 \\ -y_1 y_3 + 22.2 y_2 + u_2 \\ y_1 y_2 - 8.8 y_3/3 + u_3 \end{bmatrix}. \tag{18} $$

The synchronization error is defined as follows:

$$ e_1 = y_1 - x_1, \quad e_2 = y_2 - x_2, \quad e_3 = y_3 - x_3. $$

Our purpose is to design $u(t) = [u_1, u_2, u_3]^T$ to obtain $\lim_{t\to\infty} \|e\| = \lim_{t\to\infty} \|y - x\| = 0$. Then, the error system is

$$ {}_{t_0}D_t^{\alpha} e(t) = \begin{bmatrix} 30(e_2 - e_1) + u_1 \\ x_1 x_3 - y_1 y_3 + 22.2 e_2 + u_2 \\ y_1 y_2 - x_1 x_2 - 8.8 e_3/3 + u_3 \end{bmatrix} = \begin{bmatrix} 30(e_2 - e_1) + u_1 \\ -y_3 e_1 - x_1 e_3 + 22.2 e_2 + u_2 \\ y_2 e_1 + x_1 e_2 - 8.8 e_3/3 + u_3 \end{bmatrix}. \tag{19} $$

The controller is designed as $u_1 = 0$, $u_2 = y_3 e_1 + k e_2$, $u_3 = -y_2 e_1$. The objective is to use a simple signal feedback control to synchronize the systems. The controlled fractional-order Lü system can be written as

$$ {}_{t_0}D_t^{\alpha} e(t) = \begin{bmatrix} -30 & 30 & 0 \\ 0 & k + 22.2 & -x_1 \\ 0 & x_1 & -8.8/3 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}. \tag{20} $$

In accordance with the design process of the tracking controller in Section 4.1, the same feedback coefficient k can guarantee the stability of the synchronization systems. The synchronization results of the fractional-order Lü system when k = -35 are shown in Fig 2.

The solid lines in Fig 2 show the motion curves of each state of the fractional-order Lü chaotic driving system and response system when the control action is added at t = 10 s. Clearly, the response curves tend to the driving curves after the control action is added. The error curves in sub-figures d)–e) of Fig 2 show the quickness and effectiveness of the method.

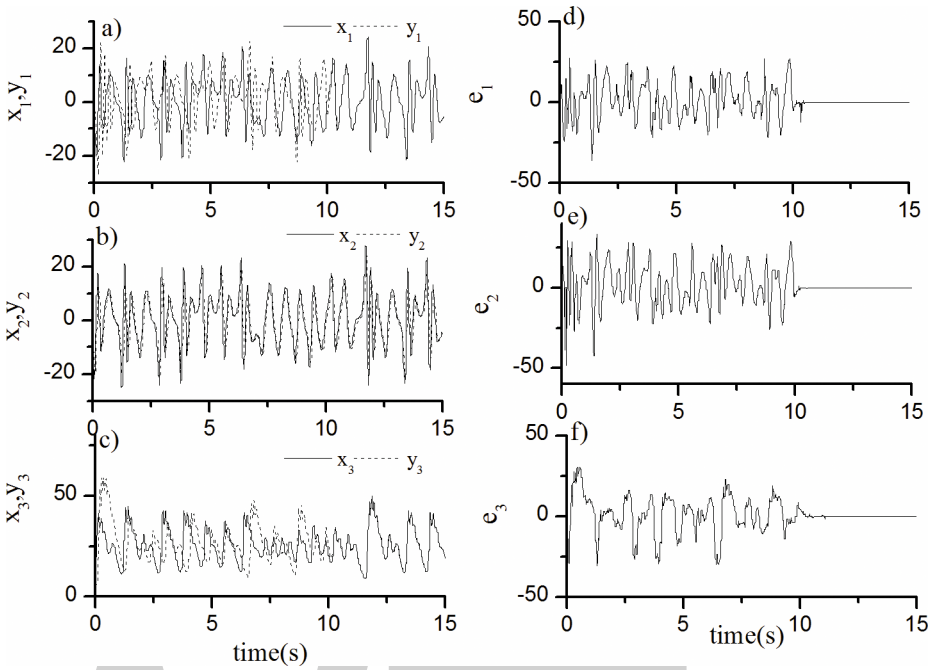

Fig 2. Synchronization results of fractional-order Lü chaotic system. The motion curves of each state of the fractional-order Lü chaotic driving system and response system when the control action is added at $t = 10$ s.

The above design shows that we can easily obtain a feedback control parameter k to make the system stable according to the proposed method. However, the time-varying state x_1 is contained in $\boldsymbol{A(t)}$, so it is difficult to obtain a feedback control parameter k based on Theorem 1.

5. Conclusion

A sufficient stability theorem for time-varying fractional-order systems is proposed because the existing stability determination methods for fractional-order systems are complicated and difficult to apply. On the basis of the proposed theorem, a feedback controller for the fractional-order Lü chaotic system is designed for tracking control and synchronization. Simulation results demonstrate the effectiveness of the method.

Supporting information

S1 Data.
(ZIP)

Acknowledgments

This work was supported by the National Key R&D Program of China (2016YFB0600701) and the Fundamental Fund for the Central Universities of China under Grant 2015MS66. The authors would like to give our heartfelt thanks to anonymous reviewers for the constructive suggestions in improving the paper.

Author Contributions

Conceptualization: Yu Huang.

Formal analysis: Feng Guo.

Investigation: Dongfeng Wang, Jinying Zhang.

Methodology: Yu Huang.

Software: Dongfeng Wang, Jinying Zhang.

References

1. Bagley RL, Torvik PJ. On the appearance of the fractional derivative in the behavior of real materials. J Appl Mech. 1984; 51(2):294–8.
2. Razminia A, Baleanu D. Complete synchronization of commensurate fractional order chaotic systems using sliding mode control. Mechatronics. 2013; 23 (7): 873–879.
3. Wu GC, Baleanu D. Chaos synchronization of the discrete fractional logistic map. Signal processing. 2014; 102: 96–99.
4. Golmankhaneh AK, Arefi R, Baleanu D. Synchronization in a nonidentical fractional order of a proposed modified system. Journal of vibration and control. 2015; 21(6): 1154–1161.
5. Jajarmi A, Hajipour Mojtaba, Baleanu D. New aspects of the adaptive synchronization and hyperchaos suppression of a financial model. Chaos solutions & fractals. 2017; 99: 285–296.
6. Boulkroune A, Bouzeriba A, Bouden T. Fuzzy generalized projective synchronization of incommensurate fractional-order chaotic systems. Neurocomputing. 2016; 173:606–14.
7. Li CL, Zhang J. Synchronisation of a fractional-order chaotic system using finite-time input-to-state stability. International Journal of Systems Science. 2016; 47(10):2440–8.
8. Li RH, Chen WS. Lyapunov-based fractional-order controller design to synchronize a class of fractional-order chaotic systems. Nonlinear Dynamics. 2014; 76(1):785–95.
9. Shao SY, Chen M, Yan XH. Adaptive sliding mode synchronization for a class of fractional-order chaotic systems with disturbance. Nonlinear Dynamics. 2016; 83(4):1855–66.
10. Soukkou A, Boukabou A, Leulmi S. Prediction-based feedback control and synchronization algorithm of fractional-order chaotic systems. Nonlinear Dynamics. 2016; 85(4):2183–206.
11. Nourian A, Balochian S. The adaptive synchronization of fractional-order Liu chaotic system with unknown parameters. Pramana. 2016; 86(6):1401–7.
12. Maheri M, Arifin NM. Synchronization of two different fractional-order chaotic systems with unknown parameters using a robust adaptive nonlinear controller. Nonlinear Dynamics. 2016; 85(2):825–38.
13. Zhou P, Ding R. Adaptive function projective synchronization between different fractional-order chaotic systems. Indian Journal of Physics. 2012; 86(6):497–501.
14. Yang CC. One input control for exponential synchronization in generalized Lorenz systems with uncertain parameters. Journal of the Franklin Institute. 2012; 349(1):349–65.
15. Zhang RX, Yang SP. Robust chaos synchronization of fractional-order chaotic systems with unknown parameters and uncertain perturbations. Nonlinear Dynamics. 2012; 69(3):983–92.
16. Xiang W, Chen FQ. Robust synchronization of a class of chaotic systems with disturbance estimation. Communications in Nonlinear Science and Numerical Simulation. 2011; 16(8):2970–7.
17. Aguila-Camacho N, Manuel A. Duarte-Mermoud, Gallegos JA. Lyapunov functions for fractional order systems. Communications in Nonlinear Science and Numerical Simulation. 2014; 19(9): 2951–2957.
18. Bao HB, Park JH, Cao JD. Adaptive synchronization of fractional-order memristor-based neural networks with time delay. Nonlinear Dynamics. 2015; 82(3):1343–1354
19. Global S. Mittag-Leffler stability and synchronization of impulsive fractional-order neural networks with time-varying delays. Nonlinear Dynamics. 2014; 77(4):1251–1260.
20. Semary MS, Radwan AG, Hassan HN. Fundamentals of fractional-order LTI circuits and systems: number of poles, stability, time and frequency responses. International Journal of Circuit Theory and Applications. 2016; 44(12):2114–33.
21. Wang ZB, Cao GY, Zhu XJ. Stability conditions and criteria for fractional order linear time-invariant systems. Control Theory & Applications(China). 2004; 21(6):922–6.
22. Wang Z, Cao G, Zhu X. Research on the internal and external stability of fractional order linear systems. Control and Decision. 2004; 19(10):1171–4.

23. Pakzad MA, Pakzad S, Nekoui MA. Exact method for the stability analysis of time-delayed linear-time invariant fractional-order systems. IET Control Theory & Applications. 2015; 9(16):2357–68.

24. Li LX, Peng HP, Luo Q, Yang YX, Liu Z. Problem and analysis of stability decidable theory for a class of fractional order nonlinear system. 2013.

25. Oustaloup A, Mathieu B, Lanusse P, editors. Second generation CRONE control. Systems, Man and Cybernetics, 1993'Systems Engineering in the Service of Humans', Conference Proceedings, International Conference on; 1993: IEEE. pp. 136–142.

26. Li Y, Chen Y, Podlubny I. Stability of fractional-order nonlinear dynamic systems: Lyapunov direct method and generalized Mittag–Leffler stability. Computers & Mathematics with Applications. 2010; 59 (5):1810–21.

27. Lu JH, Chen GR. A time-varying complex dynamical network model and its controlled synchronization criteria. IEEE Transactions on Automatic Control. 2005; 50(6):841–6.

28. Zhu ZW, Leung H. Adaptive identification of nonlinear systems with application to chaotic communications. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications. 2000; 47 (7):1072–80.

Parameter identification for gompertz and logistic dynamic equations

Elvan Akın[1☯]*, **Neslihan Nesliye Pelen**[2☯], **Ismail Uğur Tiryaki**[3☯], **Fusun Yalcin**[4☯]

1 Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, Missouri, United States of America, **2** Department of Mathematics, Ondokuz Mayıs University, Arts and Science Faculty, Samsun, Turkey, **3** Department of Mathematics, Bolu Abant Izzet Baysal University, Faculty of Arts and Science, Bolu, Turkey, **4** Department of Mathematics, Faculty of Science, Akdeniz University, Antalya, Turkey

☯ These authors contributed equally to this work.
* akine@mst.edu

Abstract

In this paper, we generalize and compare Gompertz and Logistic dynamic equations in order to describe the growth patterns of bacteria and tumor. First of all, we introduce two types of Gompertz equations, where the first type 4-paramater and 3-parameter Gompertz curves do not include the logarithm of the number of individuals, and then we derive 4-parameter and 3-parameter Logistic equations. We notice that Logistic curves are better in modeling bacteria whereas the growth pattern of tumor is described better by Gompertz curves. Increasing the number of parameters of Logistic curves give favorable results for bacteria while decreasing the number of parameters of Gompertz curves for tumor improves the curve fitting. Moreover, our results overshadow some of the existing results in the literature.

Editor: J. Alberto Conejero, IUMPA - Universitat Politecnica de Valencia, SPAIN

Introduction

Most of the growth curves are described by linear, power, parabolic, power-exponential, logistic, log-logistic, von Bertalanffy, Gompertz, and Richards curves; see [1], [2], [3], [4], [5], and [6] for the tumor, [7] for the human fetus, [8] for the human life. A recent research article [8] related with a human life modeled by Gompertz and Mirror Gompertz differential equations are

$$(\ln x)' = -\beta \ln x, \tag{1}$$

and

$$x' = -\beta(1-x)\ln(1-x), \tag{2}$$

respectively, where β is a positive parameter. The well-known logistic equation in the

Funding: The authors received no specific funding.

Competing interests: The authors have declared that no competing interests exist.

literature, see [9] is given by

$$x' = kx\left(1 - \frac{x}{K}\right), \tag{3}$$

where k is the proportionality constant and K is the carrying capacity.

In this study, mathematical modeling is applied to the Pseudomonas putida and mammary tumor datas given in [10, 11], respectively. Note that Pseudomonas putida is a bacterium found in most soil and water habitats, and is significant to the environment due to its complex metabolism and ability to control pollution, [12] and [13]. We model their growth patterns by continuous and discrete Gompertz and Logistic curves. To achieve our goal, we derive 4-parameter and 3-parameter Gompertz and Logistic dynamic equations. We first propose two types of Gompertz dynamic equations: The first type Gompertz dynamic equations are motivated by [14]. We contribute two first type continuous Gompertz curves to the literature. All of the discrete Gompertz curves in this type are new. 4-parameter second type Gompertz dynamic equations are motivated by [2] in which only 3-parameter discrete Gompertz curves are considered. 3-parameter second type continuous Gompertz are investigated earlier in [10]. Inspired by [15], we come up with 4-parameter Logistic dynamic equations while 3-parameter Logistic dynamic equations are constructed earlier in [16]. 4-parameter Logistic discrete curves are new. To establish both dynamic equations, we use the variation of constant formulas together with the circle dot multiplication and the circle minus substraction on time scales. We refer readers to [17] and [18] by Bohner and Peterson for the theory of time scales calculus.

The parameters of these models are estimated by NonlinearModelFit function of Wolfram Mathematica 11.0 applying Monte Carlo simulation and our comparison is based on outputs following from the p-values of parameters, adjusted R-squared, and RMSE (root mean square error), RRMSE (Relative Root Mean Square Error), MAPE (Mean Absolute Percent Error), MAE (mean absolute error), U1 (Theil inequality coefficient, Theil's U1), and U2 (Theil inequality coefficient, Theil's U2). We use the Mathematica 11 program for the goodness of fit test of the models. Having at least three small values of each determined statistical criterion, the p value less than 0.05 for each parameter, and adjusted R-squared value close to 1 show better performance in terms of goodness of fit.

Outline of this paper is as follows: In Section 2, we introduce the time scales calculus together with some preliminary results. Sections 3 and 4 are related with first and second type Gompertz dynamic equations. In each section we obtain 4-parameter and 3-parameter continuous and discrete Gompertz curves. In Section 6, Logistic dynamic equations are introduced and we explicitly calculate 4-parameter and 3-parameter continuous and discrete Logistic curves. In the last section, we discuss how Gompertz and Logistic curves fit the growth of Pseudomonas putida and mammary tumor and include our conclusion.

Preliminary results

A *time scale*, $\mathbb{T}$, is an arbitrary nonempty closed subset of the real numbers $\mathbb{R}$. The theory of time scales is to introduce a new calculus so that we can unify the continuous and discrete analysis. Here, we give basic definitions and some essential results without proofs. Nevertheless, we mainly refer readers two books [17] and [18] by Bohner and Peterson and the manuscript [16] by Akin-Bohner and Bohner.

The *forward jump operator* σ on $\mathbb{T}$ is defined as $\sigma(t) := \inf\{s > t : s \in \mathbb{T}\} \in \mathbb{T}$, for all $t \in \mathbb{T}$. For this definition we also have $\sigma(\emptyset) = \sup \mathbb{T}$. The *backward jump operator* ρ on $\mathbb{T}$ is defined by $\rho(t) := \sup\{s < t : s \in \mathbb{T}\} \in \mathbb{T}$, for all $t \in \mathbb{T}$. Here, we have $\rho(\emptyset) = \inf \mathbb{T}$. If $\sigma(t) > t$, we say

t is *right-scattered*, while if $\rho(t) < t$ we say t is *left-scattered*. If $\sigma(t) = t$, we say t is *right-dense*, while if $\rho(t) = t$ we say t is *left-dense*. The *graininess* function $\mu : \mathbb{T} \mapsto [0, \infty)$ is defined by $\mu(t)$: $= \sigma(t) - t$. It is apparent that for $\mathbb{T} = \mathbb{Z}$, $\sigma(t) = t + 1$, $\rho(t) = t - 1$ and for $\mathbb{T} = \mathbb{R}$, $\sigma(t) = t$, $\rho(t) = t$. The set $\mathbb{T}^\kappa$ is derived from $\mathbb{T}$. If $\mathbb{T}$ has left-scattered maximum m, then $\mathbb{T}^\kappa = \mathbb{T} - \{m\}$. Otherwise, $\mathbb{T}^\kappa = \mathbb{T}$. The following notations are also useful: $f^\sigma(t) = f(\sigma(t))$. Note that $t \in [t_0, \infty)_{\mathbb{T}} = [t_0, \infty) \cap \mathbb{T}$.

Assume $f : \mathbb{T} \mapsto \mathbb{R}$ and let $t \in \mathbb{T}^\kappa$, then we define $f^\Delta(t)$ to be the number (provided it exists) with the property that given any $\epsilon > 0$, there is a neighborhood U of t such that

$$| [f(\sigma(t)) - f(s)] - f^\Delta(t)[\sigma(t) - s] | \leq \epsilon | \sigma(t) - s |,$$

for all $s \in U$. $f^\Delta(t)$ is called the *delta derivative* of $f(t)$ at t. Note that the delta-derivative turns out to be the usual derivative when $\mathbb{T} = \mathbb{R}$ while it is the forward difference operator when $\mathbb{T} = \mathbb{Z}$. If f is differentiable at t, then f is continuous at t. If f is continuous at t and t is right-scattered, then f is differentiable at t with

$$f^\Delta(t) = \frac{f(\sigma(t)) - f(t)}{\mu(t)}.$$

If f is differentiable and t is right-dense, then

$$f^\Delta(t) = \lim_{s \to t} \frac{f(t) - f(s)}{t - s}.$$

If f is differentiable at t, then

$$f^\sigma(t) = f(t) + \mu(t) f^\Delta(t). \tag{4}$$

If $f, g : \mathbb{T} \mapsto \mathbb{R}$ are differentiable at $t \in \mathbb{T}^\kappa$, then the product $fg : \mathbb{T} \mapsto \mathbb{R}$ is also differentiable at t with

$$(fg)^\Delta(t) = f^\Delta(t) g(t) + f(\sigma(t)) g^\Delta(t).$$

If f is continuous at each right-dense point $t \in \mathbb{T}$ and whenever $t \in \mathbb{T}$ is left-dense $\lim_{s \to t^-} f(s)$ exists as a finite number, then we say that $f : \mathbb{T} \mapsto \mathbb{R}$ is *rd-continuous*. A function $F : \mathbb{T}^\kappa \mapsto \mathbb{R}$ is called an *antiderivative* of $f : \mathbb{T} \mapsto \mathbb{R}$ provided $F^\Delta(t) = f(t)$ holds for all $t \in \mathbb{T}^\kappa$. In this case, we define the integral of f by

$$\int_a^t f(s) \Delta s = F(t) - F(a) \text{ for } t \in \mathbb{T}. \tag{5}$$

If $1 + \mu(t) p(t) \neq 0$ for all $t \in \mathbb{T}^\kappa$, $p : \mathbb{T}^\kappa \mapsto \mathbb{R}$ is called *regressive*. The set of all regressive and rd-continuous functions is denoted by R. If $1 + \mu(t) p(t) > 0$ for all $t \in \mathbb{T}^\kappa$, $p : \mathbb{T}^\kappa \mapsto \mathbb{R}$ is called *positively regressive*. The set of all positively regressive and rd-continuous functions is denoted by R^+.

If $p, q \in R$ and α is a constant, then we define

$$\ominus p(t) = -\frac{p(t)}{1 + \mu(t) p(t)}, \qquad p(t) \ominus q(t) = \frac{p(t) - q(t)}{1 + \mu(t) q(t)}, \tag{6}$$

and

$$(p \oplus q)(t) = p(t) + q(t) + \mu(t) p(t) q(t)$$

for all $t \in \mathbb{T}^\kappa$. Finding a simple formula of the derivative of any power of a function yields to

the introduction of a circle dot multiplication. A circle dot multiplication $\odot$ is defined in [16] as

$$(\alpha \odot p)(t) = \alpha p(t) \int_0^1 (1 + h\mu(t)p(t))^{\alpha-1} dh.$$

Note that $\ominus p = -p$, $p \oplus q = p + q$ and $\alpha \odot p = \alpha p$ for the continuous case. If p is regressive, then we define the exponential function by

$$e_p(t, s) = \exp\left(\int_s^t \xi_\mu(p(\tau))\Delta\tau\right) \quad for\ s, t \in \mathbb{T}, \tag{7}$$

where $\xi_h(z) = \frac{1}{h} Log(1 + hz), h > 0$ is the cylinder transformation such that $\xi_0(z) = z$. If $p : \mathbb{T}^\kappa \mapsto \mathbb{R}$ is rd-continuous and regressive, then the *exponential function* $e_p(t, t_0)$ is the unique solution of the IVP

$$x^\Delta = p(t)x, \quad x(t_0) = 1$$

on $\mathbb{T}$ for each fixed $t_0 \in \mathbb{T}^\kappa$. For data analysis we need to calculate exponential functions

$$e_\beta(t, t_0) = e^{\beta(t-t_0)}, \quad e_{\ominus\beta}(t, t_0) = e^{-\beta(t-t_0)} \text{ when } \mathbb{T} = \mathbb{R} \tag{8}$$

$$e_\beta(t, t_0) = (1 + \beta)^{t-t_0}, \quad e_{\ominus\beta}(t, t_0) = (1 + \beta)^{-(t-t_0)} \text{ when } \mathbb{T} = \mathbb{Z} \tag{9}$$

for a regressive constant β, see Table 2.4 in [17].

We use the following properties of the exponential function $e_p(t, s)$, $t, s \in \mathbb{T}$.

Theorem 0.1. *If p, q are regressive and $t_0 \in \mathbb{T}$, then*

1. $e_p(t, t) \equiv$ *and* $e_0(t, s) \equiv 1$;
2. $e_p(\sigma(t), s) = (1 + \mu(t)p(t))e_p(t, s)$;
3. $\dfrac{1}{e_p(t, s)} = e_{\ominus p}(t, s) = e_p(s, t)$;
4. $\dfrac{e_p(t, s)}{e_q(t, s)} = e_{p\ominus q}(t, s)$;
5. $e_p(t, s)e_q(t, s) = e_{p\oplus q}(t, s)$;
6. *if $p > 0$ for all $t \in \mathbb{T}$, then $e_p(t, t_0) > 0$ for all $t \in \mathbb{T}$*;
7. *if $p \in R^+$, then $e_p(t, t_0) > 0$ for all $t \in \mathbb{T}$.*

In addition, two of the useful formulas for a circle dot are

$$e_{\alpha\odot p}(t, t_0) = (e_p(t, t_0))^\alpha, \tag{10}$$

and

$$1 + \mu(\alpha \odot p) = (1 + \mu p)^\alpha, \tag{11}$$

where p is a regressive function and α is a constant, see [16].

The followings are the variation of constants formulas, see Theorems 2.74 and 2.77 in [17]. The equation

$$x^\Delta = p(t)x + f(t) \tag{12}$$

is called *regressive* if $x^{\Delta} = p(t)x$ is regressive (i.e., p is regressive) and $f : \mathbb{T} \to \mathbb{R}$ is rd-continuous.

Theorem 0.2. *Suppose* (12) *is regressive. Let* $t_0 \in \mathbb{T}$ *and* $x_0 \in \mathbb{R}$. *The unique solution of the IVP*

$$x^{\Delta} = p(t)x + f(t), \;\; x(t_0) = x_0$$

is given by

$$x(t) = e_p(t, t_0)x_0 + \int_{t_0}^{t} e_p(t, \sigma(\tau))f(\tau)\Delta\tau.$$

Theorem 0.3. *Suppose* (12) *is regressive. Let* $t_0 \in \mathbb{T}$ *and* $x_0 \in \mathbb{R}$. *The unique solution of the IVP*

$$x^{\Delta} = -p(t)x^{\sigma} + f(t), \;\; x(t_0) = x_0$$

is given by

$$x(t) = e_{\ominus p}(t, t_0)x_0 + \int_{t_0}^{t} e_{\ominus p}(t, \tau)f(\tau)\Delta\tau.$$

First type gompertz dynamic equations

In this section, we will introduce Gompertz dynamic curves motivated by the 4-parameter Gompertz curve

$$\omega(t) = B + A \exp\left(-\exp(-K(t - t_0)\right), t \in \mathbb{R} \tag{13}$$

given in [19] for the growth curve analyses of bacterial counts. Here, K can be found as the growth rate coefficient, t_0 is the initial time, $A + B$ is the carrying capacity of the environment for the population. To explain the carrying capacity notion we can say that every environment has its own limits, therefore it is impossible for species to grow up infinitely. Thus, the number of the population should be finite.

In order to obtain the Gompertz model in the continuous case, we differentiate Eq (13) and obtain

$$\begin{aligned} \omega' &= AK \exp\{-\exp\{-K(t - t_0)\}\} \exp\{-K(t - t_0)\} \\ &= [\omega(t) - B]K \exp\{-K(t - t_0)\}. \end{aligned}$$

In addition, note that we have

$$\begin{aligned} e_{\ominus(K\odot\ominus e_{\ominus K})}(t, t_0) &= \frac{1}{e_{K\odot \ \ominus e_{\ominus K}}(t, t_0)} = \left(\frac{1}{e_{\ominus e_{\ominus K}}(t, t_0)}\right)^{K} \\ &= (e_{e_{\ominus K}}(t, t_0))^{K} = e_{K\odot e_{\ominus K}}(t, t_0) \end{aligned} \tag{14}$$

on $[t_0, \infty)_{\mathbb{T}}$, where we use Theorem 0.1 and (10). Since $e_{\ominus K}(t, t_0) = e^{-K(t-t_0)}$ for $t \in \mathbb{R}$, and (14) holds, then we obtain

$$\begin{aligned} e_{\ominus(K\odot\ominus e_{\ominus K})}(t, t_0) &= \exp\left\{\frac{\exp\left(-K(t - t_0)\right)}{K} - \frac{1}{K}\right\}^{-K} \\ &= e \exp\{-\exp\{-K(t - t_0)\}\}, t \in \mathbb{R}. \end{aligned} \tag{15}$$

Motivated by the calculation above, we have the following initial value problem modeling 4-parameter Gompertz curve on time scales.

Theorem 0.4. *The initial value problem*

$$\begin{aligned} \omega^{\Delta} &= -(K \odot \ominus e_{\ominus K}(t, t_0))\omega^{\sigma} + B(K \odot \ominus e_{\ominus K}(t, t_0)) \\ \omega(t_0) &= \omega_0 \end{aligned} \tag{16}$$

has the solution of the form

$$\omega = B + (\omega_0 - B)e_{K \odot e_{\ominus K}}(t, t_0)$$

$t \in [t_0, \infty)_{\mathbb{T}}$, *where K is the growth rate and* t_0 *is the initial time,* ω_0 *is the value of the function at the initial time and B is the coefficient that has an impact on carrying capacity.*

Proof. We notice that the positivity of K implies the positivity of $e_{\ominus K}$ by Theorem 0.1. Since $1 + \mu(\ominus e_{\ominus K}) = \frac{1}{1+\mu e_{\ominus K}} > 0$, we have the positively regressivity of $\ominus e_{\ominus K}$. Since $1 + \mu(K \odot \ominus e_{\ominus K}) = (1 + \mu(\ominus e_{\ominus K}))^K > 0$ by (11), the dynamic equation in the IVP (16) is regressive. Therefore, we apply Theorem 0.3 and obtain the unique solution for $t \in [t_0, \infty)_{\mathbb{T}}$

$$\begin{aligned} \omega &= e_{\ominus(K \odot \ominus e_{\ominus K})}(t, t_0)\omega_0 + B\int_{t_0}^{t} e_{\ominus(K \odot \ominus e_{\ominus K})}(t, \tau)(K \odot \ominus e_{\ominus K}(\tau, t_0))\Delta\tau \\ &= e_{K \odot e_{\ominus K}}(t, t_0)\omega_0 \\ &\quad + Be_{K \odot \ominus e_{\ominus K}}(t_0, t)\int_{t_0}^{t} e_{K \odot \ominus e_{\ominus K}}(\tau, t_0)(K \odot \ominus e_{\ominus K}(\tau, t_0))\Delta\tau \\ &= e_{K \odot e_{\ominus K}}(t, t_0)\omega_0 + Be_{K \odot \ominus e_{\ominus K}}(t_0, t)\int_{t_0}^{t} (e_{K \odot \ominus e_{\ominus K}}(\tau, t_0))^{\Delta}\Delta\tau \\ &= e_{K \odot e_{\ominus K}}(t, t_0)\omega_0 + Be_{K \odot \ominus e_{\ominus K}}(t, t_0)[e_{K \odot \ominus e_{\ominus K}}(t, t_0) - 1] \\ &= B + (\omega_0 - B)e_{K \odot e_{\ominus K}}(t, t_0), \end{aligned} \tag{17}$$

where we use (14) and Theorem 0.1.

Example 0.5. Let $\mathbb{T} = \mathbb{R}$. Then the continuous Gompertz curve

$$\omega = B + e(\omega_0 - B)\exp\{-\exp\{-K(t - t_0)\}\} \tag{18}$$

is obtained from (17) for $t \in [t_0, \infty)_{\mathbb{R}}$ by using Eqs (8) and (15). This is compatible with the continuous Gompertz growth curve (13) by taking $A = e(\omega_0 - B)$ in (13).

Example 0.6. Let $\mathbb{T} = \mathbb{Z}$. Since $e_{\ominus K}(t, t_0) = (1 + K)^{-(t-t_0)}$ for $t \in [t_0, \infty)_{\mathbb{Z}}$ by (9), (14) yields

$$\begin{aligned} e_{K \odot e_{\ominus K}}(t, t_0) &= [e_{e_{\ominus K}}(t, t_0)]^K \\ &= \left[\exp\left(\sum_{s=t_0}^{t-1} \ln\left(1 + \frac{1}{(1+K)^{s-t_0}}\right)\right)\right]^K \\ &= \left[\exp\left(\ln\left(\prod_{s=t_0}^{t-1} 1 + \frac{1}{(1+K)^{s-t_0}}\right)\right)\right]^K \\ &= \left[\prod_{s=t_0}^{t-1} \frac{1 + (1+K)^{s-t_0}}{(1+K)^{s-t_0}}\right]^K, \quad t \in [t_0, \infty)_{\mathbb{Z}} \end{aligned}$$

and so

$$e_{\ominus(K\odot\ominus e_{\ominus K})}(t,t_0)=\left[\prod_{\tau=t_0}^{t-1}\frac{(1+K)^{\tau-t_0}}{1+(1+K)^{\tau-t_0}}\right]^{-K} \tag{19}$$

for $t\in[t_0,\infty)_{\mathbb{Z}}$. Thus, the discrete Gompertz growth curve

$$\omega=B+(\omega_0-B)\left[\prod_{\tau=t_0}^{t-1}\frac{1+(1+K)^{\tau-t_0}}{(1+K)^{\tau-t_0}}\right]^{K} \tag{20}$$

again follows from (17) for $t\in[t_0,\infty)_{\mathbb{Z}}$.

Motivated by the first variation of constant formula, Theorem 0.2, we derive another Gompertz curve on time scales.

Theorem 0.7. *The initial value problem*

$$\begin{aligned}&\omega^{\Delta}=\ominus(K\odot e_{\ominus K}(t,t_0))\omega+B(\ominus(K\odot e_{\ominus K}(t,t_0)))\\&\omega(t_0)=\omega_0\end{aligned} \tag{21}$$

has the solution of the form

$$\omega=(\omega_0+B)e_{\ominus(K\odot e_{\ominus K})}(t,t_0)-B \tag{22}$$

for $t\in[t_0,\infty)_{\mathbb{T}}$, where K is the decay rate coefficient and regeressive, t_0 is the initial time, ω_0 is the value of the function at the initial time and B is the coefficient that has an impact on carrying capacity.

Proof. Since K is regressive, $e_{\ominus K}$ is also regressive by (7). The dynamic equation in the IVP (21) is regressive. Then, in order to obtain the unique solution (22) we apply Theorem 0.3 for $t\in[t_0,\infty)_{\mathbb{T}}$

$$\begin{aligned}\omega &= e_{\ominus(K\odot e_{\ominus K})}(t,t_0)\omega_0+B\int_{t_0}^{t}e_{\ominus(K\odot e_{\ominus K})}(t,\sigma(\tau))(\ominus(K\odot e_{\ominus K}(\tau,t_0)))\Delta\tau\\&= e_{\ominus(K\odot e_{\ominus K})}(t,t_0)\omega_0\\&\quad -Be_{\ominus(K\odot e_{\ominus K})}(t,t_0)\int_{t_0}^{t}e_{K\odot e_{\ominus K}}(\tau,t_0)(K\odot e_{\ominus K}(\tau,t_0))\Delta\tau\\&= e_{\ominus(K\odot e_{\ominus K})}(t,t_0)\omega_0-Be_{\ominus(K\odot e_{\ominus K})}(t,t_0)\int_{t_0}^{t}(e_{K\odot e_{\ominus K}}(\tau,t_0))^{\Delta}\Delta\tau\\&= e_{\ominus(K\odot e_{\ominus K})}(t,t_0)\omega_0-Be_{\ominus(K\odot e_{\ominus K})}(t,t_0)[e_{K\odot e_{\ominus K}}(t,t_0)-1]\\&= (\omega_0+B)e_{\ominus(K\odot e_{\ominus K})}(t,t_0)-B,\end{aligned}$$

where we use (14) and Theorem 0.1.

Example 0.8. Let $\mathbb{T}=\mathbb{R}$. Then the alternative continuous Gompertz curve

$$\omega=\frac{1}{e}(\omega_0+B)\exp\{\exp\{-K(t-t_0)\}\}-B \tag{23}$$

is obtained from (22) for $t\in\mathbb{R}$. It is worth to mention that Eq (23) is a new Gompertz curve in the continuous case.

Example 0.9. Let $\mathbb{T} = \mathbb{Z}$. Then using (19), we have

$$e_{\ominus(K\odot e_{\ominus K})}(t,t_0) = \left[\prod_{\tau=t_0}^{t-1}\frac{(1+K)^{\tau-t_0}}{1+(1+K)^{\tau-t_0}}\right]^K$$

for $t \in [t_0,\infty)_{\mathbb{Z}}$. Since

$$e_{\ominus(K\odot e_{\ominus K})} = \frac{1}{e_{K\odot e_{\ominus K}}} = \left[\frac{1}{e_{e_{\ominus K}}}\right]^K,$$

the alternative discrete Gompertz growth curve

$$\omega = (\omega_0 + B)\left[\prod_{\tau=t_0}^{t-1}\frac{(1+K)^{\tau-t_0}}{1+(1+K)^{\tau-t_0}}\right]^K - B \tag{24}$$

again follows from (22) for $t \in [t_0,\infty)_{\mathbb{Z}}$.

The Gompertz growth curve (23) is given as 4-parameter Gompertz growth curve in [19]. From this point of view, the 3-parameter Gompertz growth curve on time scales

$$\omega = \omega_0 e_{K\odot e_{\ominus K}}(t,t_0) \tag{25}$$

is obtained from (17) when $B = 0$, and so the 3-parameter continuous and discrete Gompertz curves are

$$\omega = e\omega_0\exp\{-\exp\{-K(t-t_0)\}\} \tag{26}$$

for $t \in [t_0,\infty)_{\mathbb{R}}$ and for $t \in [t_0,\infty)_{\mathbb{Z}}$

$$\omega = \omega_0\left[\prod_{\tau=t_0}^{t-1}\frac{1+(1+K)^{\tau-t_0}}{(1+K)^{\tau-t_0}}\right]^K, \tag{27}$$

respectively. From (22) when $B = 0$, the alternative 3-parameter continuous and discrete Gompertz curves

$$\omega = \frac{1}{e}\omega_0\exp\{\exp\{-K(t-t_0)\}\} \tag{28}$$

and

$$\omega = \omega_0\left[\prod_{\tau=t_0}^{t-1}\frac{1+(1+K)^{\tau-t_0}}{(1+K)^{\tau-t_0}}\right]^{-K}, \tag{29}$$

are gained to the literature, respectively.

The zwietering modification of gompertz growth curve

The Gompertz growth curve is reparameterized in order to model the bacteria growth population in food and is stated as

$$w = A\exp\{-\exp\{\frac{eK_z}{A}(T-t)+1\}\} \tag{30}$$

in [14] for $t \in \mathbb{R}$, where K_z is the absolute growth rate at time T, so called lag time, which is

interpreted as the time between when a microbial population is transferred to a new habitat recovers and when a considerable cell division occurs.

In order to find a corresponding dynamic model, we rewrite (30) as

$$w = A(\exp\{-\exp\{-\frac{eK_z}{A}t\}\})^{e^{\frac{eK_z}{A}T+1}} \tag{31}$$

so that we can use the property of circle dot (10) for the unification the continuous and discrete cases. Therefore, using (15) and (14) yield the dynamic Zwietering Modification Gompertz curve

$$\begin{aligned} w &= Ae^{-\left(e^{\frac{eK_z}{A}T+1}\right)}\left(e_{\ominus\left(\frac{eK_z}{A}\odot\ominus e_{\ominus\frac{eK_z}{A}}\right)}(t,0)\right)^{e^{\frac{eK_z}{A}T+1}} \\ &= Ae^{-\left(e^{\frac{eK_z}{A}T+1}\right)}e_{e^{\left(\frac{eK_z}{A}T+1\right)}\odot\ominus\left(\frac{eK_z}{A}\odot\ominus e_{\ominus\frac{eK_z}{A}}\right)}(t,0) \\ &= Ae^{-\left(e^{\frac{eK_z}{A}T+1}\right)}e_{e^{\left(\frac{eK_z}{A}T+1\right)}\odot\left(\frac{eK_z}{A}\odot e_{\ominus\frac{eK_z}{A}}\right)}(t,0). \end{aligned} \tag{32}$$

Therefore, we claim that (32) is the solution of the IVP

$$\begin{aligned} \omega^{\Delta} &= \left(e^{\left(e^{\frac{eK_z}{A}T+1}\right)}\odot\left(\frac{eK_z}{A}\odot e_{\ominus\frac{eK_z}{A}}\right)\right)w \\ \omega(0) &= Ae^{-\left(e^{\frac{eK_z}{A}T+1}\right)}. \end{aligned} \tag{33}$$

Since (30) is the continuous modified Gompertz growth curve, the discrete modified Gompertz growth curve follows from (32).

Example 0.10. Let $\mathbb{T} = \mathbb{Z}$. Then the discrete Zwietering modification of Gompertz growth curve is

$$\omega = Ae^{-\left(e^{\frac{eK_z}{A}T+1}\right)}\left[\prod_{\tau=0}^{t-1}\frac{\left(1+\frac{eK_z}{A}\right)^{\tau}}{1+\left(1+\frac{eK_z}{A}\right)^{\tau}}\right]^{-e^{\frac{eK_z}{A}}\left(e^{\frac{eK_z}{A}T+1}\right)} \tag{34}$$

obtained from (19) for $t \in [0,\infty)_{\mathbb{Z}}$.

Gompertz-laird growth curve

This model is mainly used for the modeling of tumor growth. The Laird re-parameterization prevails even today as the most frequently fitted Gompertz version in cancer research, and is now also commonly fitted to growth data in other fields such as those of domestic (e.g. poultry and livestock, marine (e.g. molluscs, fish, and dolphins) animals.

The continuous Gompertz-Laird growth curve is given by

$$\omega = \omega_0 e^{-\frac{L}{K}(e^{-Kt}-1)}$$

for $t \in [0,\infty)_{\mathbb{R}}$ in [14], which is equivalent to

$$w = w_0 e^{\frac{L}{K}}(e^{-e^{-Kt}})^{\frac{L}{K}} \tag{35}$$

for $t \in [0,\infty)_{\mathbb{R}}$, where the parameter L describes the initial specific growth rate that is not a notion that measures the relative growth or absolute growth. More precisely, we can say that

the absolute growth rate at $t = 0$ is $\omega_0.L$ Thus, the term L can be described as division of the initial absolute growth rate with the initial value.

Similarly, by using (15) we obtain the Gompertz-Laird growth curve on time scales as

$$\begin{aligned} w &= w_0(e_{\ominus(K\odot\ominus e_{\ominus K})}(t,0))^{\frac{L}{K}} \\ &= w_0 e_{\frac{L}{K}\odot(\ominus(K\odot\ominus e_{\ominus K}))}(t,0) \\ &= w_0 e_{\frac{L}{K}\odot(K\odot e_{\ominus K})}(t,0) \end{aligned} \tag{36}$$

for $t \in [0,\infty)_{\mathbb{T}}$ and so (36) is the solution of the IVP

$$\begin{aligned} w^{\Delta} &= \frac{L}{K}\odot(K\odot e_{\ominus K})w \\ \omega(0) &= \omega_0. \end{aligned} \tag{37}$$

Since (35) is the continuous Gompertz-Laird growth curve, the following example gives the discrete Gompertz-Laird growth curve.

Example 0.11. Let $\mathbb{T} = \mathbb{Z}$. Then we obtain

$$\omega = \omega_0\left[\prod_{\tau=0}^{t-1}\frac{(1+K)^{\tau}}{1+(1+K)^{\tau}}\right]^{-L} \tag{38}$$

as the discrete Gompertz-Laird growth curve for $t \in [t_0,\infty)_{\mathbb{Z}}$, where we use again (19).

If $L = Km$ in (36), the **Zweifel and Lasker** re-parametrization dynamic equation is obtained for studying fish growth. Moreover, the continuous and discrete curves are given in (35), (38), respectively. Similarly, if $L = \ln\left(\frac{A}{\omega_0}\right)^K$ in (36), we derive the dynamic form of **Simpler W_0 form** of Gompertz Laird growth curve which prevails on. Moreover, the continuous and discrete curves are given in (35) and (38), respectively. Note that all of first type Gompertz curves in the discrete case are new.

Second type gompertz dynamic equations

It is clear that (13) is not a Gompertz model when the dependent variable is log-transformed. In this subsection, we will derive Gompertz dynamic equations involving logarithmic functions. This idea of the derivation of Gompertz dynamic equations is inspired from the Gompertz difference equation

$$\ln G(t+1) = a + b\ln G(t), \quad t \in \mathbb{Z} \tag{39}$$

where a is taken as the growth rate and b is taken as the exponential rate of growth deceleration, which was firstly given by Bassukas et. al. [3]. The equivalent form of (39) is given by

$$\Delta\ln G(t) = a + (b-1)\ln G(t), \quad t \in \mathbb{Z} \tag{40}$$

and so the continuous version becomes

$$[\ln G(t)]' = a + (b-1)\ln G(t), \quad t \in \mathbb{R}. \tag{41}$$

Unifying (40) and (41), we end up by

$$[\ln G(t)]^{\Delta} = a + (b-1)\ln G(t), \quad t \in \mathbb{T}. \tag{42}$$

By the second variation of constants formula, Theorem 0.3, we get the alternative dynamic

equation

$$[\ln G(t)]^{\Delta} = a - (b-1)\ln G(\sigma(t)), \quad t \in \mathbb{T}. \tag{43}$$

Notice that Eq (42) turns out to be (39) when $\mathbb{T} = \mathbb{Z}$ while (43) turns out to be

$$\ln G(t+1) = \frac{a}{b} + \frac{\ln G(t)}{b}, \tag{44}$$

for a nonzero constant b and when $\mathbb{T} = \mathbb{R}$, we obtain the following Gompertz differential equation from the Gompertz dynamic Eq (43) as

$$[\ln G(t)]' = a - (b-1)\ln G(t), \tag{45}$$

which is equivalent to Gompertz differential Eq (1) when $a = 0$, $b - 1 = \beta$, and $G = x$ in (45). In [8], Gompertz differential Eq (41) with $a = 0$ is called the Mirror Gompertz differential equation, and is equivalent to (2) when $a = 0$, $b - 1 = \beta$, and $G = 1 - x$ in (41).

From now on, we take $a = \alpha$ and $b - 1 = \beta$ in (42) and (43) and assume

$$\ln G(t_0) = g_0, \tag{46}$$

where g_0 is a real number. The following theorems yield the second type Gompertz dynamic curves.

Theorem 0.12. *Suppose that β is regressive and $\alpha > 0$. Then the solution of the IVP (42)–(46) is given by*

$$G(t) = \exp\left(e_{\beta}(t,t_0)\left[g_0 + \alpha\int_{t_0}^{t} e_{\beta}(t_0,\sigma(\tau))\Delta\tau\right]\right) \tag{47}$$

for $t \in [t_0,\infty)_{\mathbb{T}}$.

Proof. If $\ln G(t)$ is taken as $u(t)$, then the IVP (42)–(46) becomes $u^{\Delta}(t) = \alpha + \beta u(t)$, $u(t_0) = g_0$, $t \in [t_0,\infty)_{\mathbb{T}}$. Then by Theorems 0.1 and 0.2, we obtain

$$\begin{aligned} u(t) &= e_{\beta}(t,t_0)g_0 + \alpha\textstyle\int_{t_0}^{t} e_{\beta}(t,\sigma(\tau))\Delta\tau \\ &= e_{\beta}(t,t_0)[g_0 + \alpha\textstyle\int_{t_0}^{t} e_{\ominus\beta}(t,t_0)e_{\beta}(t,\sigma(\tau))\Delta\tau] \\ &= e_{\beta}(t,t_0)[g_0 + \alpha\textstyle\int_{t_0}^{t} e_{\beta}(t_0,\sigma(\tau))\Delta\tau], \quad t \in [t_0,\infty)_{\mathbb{T}}. \end{aligned}$$

Since $G = e^{u}$, (47) is obtained as the solution of the IVP (42)–(46).

Theorem 0.13. *Suppose β is regressive and $\alpha > 0$. Then the solution of the IVP (43)–(46) is given by*

$$G(t) = \exp\left(e_{\ominus\beta}(t,t_0)\left[g_0 + \alpha\int_{t_0}^{t} e_{\ominus\beta}(t_0,\tau)\Delta\tau\right]\right) \tag{48}$$

for $t \in [t_0,\infty)_{\mathbb{T}}$.

Proof. If ln $G(t)$ is taken as $u(t)$, then the IVP (43)–(46) turns out to be $u^{\Delta}(t) = \alpha - \beta u^{\sigma}(t)$, $u(t_0) = g_0$, $t \in [t_0, \infty)_{\mathbb{T}}$. Again, Theorems 0.1 and 0.3 yield

$$\begin{aligned} u(t) &= e_{\ominus\beta}(t, t_0)g_0 + \alpha \int_{t_0}^{t} e_{\ominus\beta}(t, \tau)\Delta\tau \\ &= e_{\ominus\beta}(t, t_0)[g_0 + \alpha \int_{t_0}^{t} e_{\beta}(t, t_0)e_{\ominus\beta}(t, \tau)\Delta\tau] \\ &= e_{\ominus\beta}(t, t_0)[g_0 + \alpha \int_{t_0}^{t} e_{\ominus\beta}(t_0, \tau)\Delta\tau] \end{aligned}$$

for $t \in [t_0, \infty)_{\mathbb{T}}$. Since $G = u$, (48) is obtained as the solution of the IVP (43)–(46).

Example 0.14. Let $\mathbb{T} = \mathbb{R}$. Then the solutions of the IVPs (42)–(46) and (43)–(46) are

$$\begin{aligned} G(t) &= \exp\left(e^{\beta(t-t_0)}[g_0 + \alpha \int_{t_0}^{t} e^{-\beta(\tau-t_0)}d\tau]\right) \\ &= \exp\left(e^{\beta(t-t_0)}\left[g_0 + \frac{\alpha}{\beta}(1 - e^{-\beta(t-t_0)})\right]\right) \\ &= \exp\left(\left(g_0 + \frac{\alpha}{\beta}\right)e^{\beta(t-t_0)} - \frac{\alpha}{\beta}\right), \end{aligned} \tag{49}$$

and

$$\begin{aligned} G(t) &= \exp\left(e^{-\beta(t-t_0)}[g_0 + \alpha \int_{t_0}^{t} e^{\beta(\tau-t_0)}d\tau]\right) \\ &= \exp\left(e^{-\beta(t-t_0)}\left[g_0 + \frac{\alpha}{\beta}(e^{\beta(t-t_0)} - 1)\right]\right) \\ &= \exp\left(\left(g_0 - \frac{\alpha}{\beta}\right)e^{-\beta(t-t_0)} + \frac{\alpha}{\beta}\right), \end{aligned} \tag{50}$$

respectively for $t \in [t_0, \infty)$ and here we use (8).

Example 0.15. Let $\mathbb{T} = \mathbb{Z}$. Then the solutions of the IVPs (42)–(46) and (43)–(46) are

$$\begin{aligned} G(t) &= \exp\left(e_{\beta}(t, t_0)[g_0 + \alpha \int_{t_0}^{t} e_{\beta}(t_0, \sigma(\tau))\Delta\tau]\right) \\ &= \exp\left((\beta + 1)^{(t-t_0)}[g_0 + \alpha\sum_{\tau=t_0}^{t-1}(1 + \beta)^{-(\tau+1-t_0)}]\right) \\ &= \exp\left((\beta + 1)^{(t-t_0)}[g_0 + \alpha(1 + \beta)^{-(t-t_0)}\sum_{\tau=t_0}^{t-1}(1 + \beta)^{\tau-t_0}]\right) \\ &= \exp\left((\beta + 1)^{(t-t_0)}\left[g_0 + \alpha(1 + \beta)^{-(t-t_0)}\frac{(1 + \beta)^{t-t_0} - 1}{\beta}\right]\right) \\ &= \exp\left(\left(g_0 + \frac{\alpha}{\beta}\right)(\beta + 1)^{t-t_0} - \frac{\alpha}{\beta}\right) \end{aligned} \tag{51}$$

and

$$\begin{aligned} G(t) &= \exp\left(e_{\ominus\beta}(t,t_0)[g_0+\alpha\int_{t_0}^{t} e_{\ominus\beta}(t_0,\tau)\Delta\tau]\right) \\ &= \exp\left((\beta+1)^{-(t-t_0)}[g_0+\alpha\textstyle\sum_{\tau=t_0}^{t-1}(1+\beta)^{\tau-t_0}]\right) \\ &= \exp\left((\beta+1)^{-(t-t_0)}\left[g_0+\alpha\frac{(1+\beta)^{t-t_0}-1}{\beta}\right]\right) \\ &= \exp\left(\left(g_0-\frac{\alpha}{\beta}\right)(\beta+1)^{-(t-t_0)}+\frac{\alpha}{\beta}\right), \end{aligned} \tag{52}$$

respectively for $t \in [t_0,\infty)_{\mathbb{Z}}$ and here we use (9).

In (51) by taking $t_0 = 0, \beta = b-1$ and $\alpha = a$ Equation 3.1 is obtained in [11]. Both continuous Gompertz growth curves (18) and (50) are obtained from the IVPs (13) and (43)–(46). If we let $B = 0$, $e\omega_0 = e^{\frac{\alpha}{\beta}}$, $K = \beta$ in (18), and $g_0 - \frac{\alpha}{\beta} = -1$ in (50), we observe that

$$e.e^{g_0}e^{-\frac{\alpha}{\beta}} = e.e^{-1} = 1$$

which implies $\omega_0 = e^{g_0}$. Similarly, the continuous Gompertz curves (23) and (49) are the solutions of the IVPs (21) and (42)–(46). If we let $B = 0$, $\frac{1}{e}\omega_0 = e^{-\frac{\alpha}{\beta}}$, $\beta = -K$ in (23), and $g_0 + \frac{\alpha}{\beta} = 1$ in (49), we observe that

$$e^{-1}.e^{g_0}e^{\frac{\alpha}{\beta}} = e^{-1}e = 1$$

which implies $\omega_0 = e^{g_0}$. From these observations, we conclude the 3-parameter first type continuous Gompertz curve and the second type continuous Gompertz curve are identical. However, since such an intimate relation among the discrete curves cannot be observed, considering first and second type discrete Gompertz curves contributes to the literature.

Logistic dynamic equations

In this section, we derive 4-parameter and 3-parameter Logistic continuous and discrete curves from Logistic dynamic equations.

3-Parameter logistic dynamic curves

Since there are two versions of linear equations

$$u^{\Delta} = p(t)u + f(t), \qquad u^{\Delta} = -p(t)u^{\sigma} + f(t),$$

there are two Logistic dynamic equations

$$L^{\Delta} = [\ominus(p(t) + f(t)L)]L, \tag{53}$$

and

$$L^{\Delta} = [p(t) \ominus (f(t)L)]L, \tag{54}$$

respectively, see [16]. By using the definition of circle minus (6), Logistic dynamic Eqs (53) and (54) turn out to be the typical Logistic differential Eq (3) under certain conditions on p and f when $\mathbb{T} = \mathbb{R}$. In [16], it is shown that the solutions of (53) and (54) subject to

$$L(t_0) = l_0 \neq 0 \tag{55}$$

are given by

$$L(t) = \frac{e_{\ominus p}(t, t_0)}{\frac{1}{l_0} + \int_{t_0}^{t} e_{\ominus p}(\sigma(\tau), t_0) f(\tau) \Delta\tau}, \tag{56}$$

and

$$L(t) = \frac{e_p(t, t_0)}{\frac{1}{l_0} + \int_{t_0}^{t} e_p(\tau, t_0) f(\tau) \Delta\tau}, \tag{57}$$

see Theorem 4.2 in [16]. Here, we assume that p is regressive, f is a rd-continuous function. We now calculate continuous and discrete Logistic solutions in order to compare their data fitting. In population dynamics, one often assumes that there exists a constant $N \neq 0$ such that $p(t) = Nf(t)$ for all $t \in \mathbb{T}$.

Example 0.16. Let $\mathbb{T} = \mathbb{R}$, $t_0 = 1, f = \alpha$ and $p = \beta$, where α and β are constants. Then (56) and (57) turn out to be

$$L = \frac{1}{-\frac{\alpha}{\beta} + \left(\frac{1}{l_0} + \frac{\alpha}{\beta}\right) e^{\beta(t-1)}} \tag{58}$$

and

$$L = \frac{1}{\frac{\alpha}{\beta} + \left(\frac{1}{l_0} - \frac{\alpha}{\beta}\right) e^{-\beta(t-1)}}, \tag{59}$$

respectively.

Example 0.17. Let $\mathbb{T} = \mathbb{Z}$, $t_0 = 1, f = \alpha$ and $p = \beta$, where α and β are constants. (56) and (57) turn out to be

$$L = \frac{1}{-\frac{\alpha}{\beta} + \left(\frac{1}{l_0} + \frac{\alpha}{\beta}\right) (1 + \beta)^{t-1}} \tag{60}$$

and

$$L = \frac{1}{\frac{\alpha}{\beta} + \left(\frac{1}{l_0} - \frac{\alpha}{\beta}\right) (1 + \beta)^{-t+1}}, \tag{61}$$

respectively.

4-Parameter logistic dynamic curves

The 4-parameter Logistic curve

$$\omega(t) = f - \frac{1}{\frac{b}{f} + \left(\frac{1}{f-s} - \frac{b}{f}\right) e^{kt}}, \quad t \in \mathbb{R} \tag{62}$$

is introduced and discussed in [15], where k, f, and s are positive constants and b is a real

number. If we let $b = 1$ in (62), then (62) and (3) are equivalent. Equivalently, we have

$$\omega(t) = f - \frac{f}{b + \left(\frac{f}{f-s} - b\right) e^{kt}}, \quad t \in \mathbb{R}. \tag{63}$$

Motivated by (63), we purpose the 4-parameter logistic dynamic curve as

$$\omega(t) = f - \frac{f}{b + \left(\frac{f}{f-s} - b\right) e_k(t,0)}, \quad t \in \mathbb{T}. \tag{64}$$

To obtain the 4-parameter logistic dynamic equation, we differentiate Eq (64) and derive

$$\begin{aligned}
\omega^{\Delta} &= \frac{f\left(\frac{f-s}{b} - b\right) k e_k(t,0) + kfb - kfb}{\left(b + \left(\frac{f}{f-s} - b\right) e_k(t,0)\right)\left(b + \left(\frac{f}{f-s} - b\right) e_k^{\sigma}(t,0)\right)} \\
&= \frac{kf\left[b + \left(\frac{f-s}{b} - b\right) e_k(t,0)\right] - kfb}{\left(b + \left(\frac{f}{f-s} - b\right) e_k(t,0)\right)\left(b + \left(\frac{f}{f-s} - b\right) e_k^{\sigma}(t,0)\right)} \\
&= \frac{kf}{\left(b + \left(\frac{f}{f-s} - b\right) e_k^{\sigma}(t,0)\right)} - \frac{kfb}{\left(b + \left(\frac{f}{f-s} - b\right) e_k(t,0)\right)\left(b + \left(\frac{f}{f-s} - b\right) e_k^{\sigma}(t,0)\right)} \\
&= -k\omega^{\sigma} + kf - \frac{kfb}{\left(b + \left(\frac{f}{f-s} - b\right) e_k(t,0)\right)\left(b + \left(\frac{f}{f-s} - b\right) e_k^{\sigma}(t,0)\right)} \\
&+ \frac{kfb}{b + \left(\frac{f}{f-s} - b\right) e_k(t,0)} - \frac{kfb}{b + \left(\frac{f}{f-s} - b\right) e_k(t,0)} \\
&= -k\omega^{\sigma} + kf + \frac{kfb}{b + \left(\frac{f}{f-s} - b\right) e_k(t,0)} \omega^{\sigma} - \frac{kfb}{b + \left(\frac{f}{f-s} - b\right) e_k(t,0)} \\
&= -k\left[1 - \frac{b}{b + \left(\frac{f}{f-s} - b\right) e_k(t,0)}\right] \omega^{\sigma} + kf\left[1 - \frac{b}{b + \left(\frac{f}{f-s} - b\right) e_k(t,0)}\right].
\end{aligned} \tag{65}$$

Hence, we obtain the 4-parameter logistic dynamic equation as:

$$\omega^{\Delta}(t) = -p(t)\omega^{\sigma}(t) + fp(t), \quad t \in \mathbb{T} \tag{66}$$

where

$$p(t) = k\left[1 - \frac{b}{b + \left(\frac{f}{f-s} - b\right) e_k(t,0)}\right], \quad t \in \mathbb{T}. \tag{67}$$

Theorem 0.18. *Let k, f, s be positive constants. Consider the 4-parameter logistic dynamic Eq (66) with the initial condition*

$$\omega(0) = s. \tag{68}$$

Then,

$$\omega(t) = f + (s - f) e_{\ominus p}(t, 0), \quad t \in \mathbb{T} \tag{69}$$

is the unique solution of the IVP (66)–(68) where p is defined as in (67).

Proof. To get the desired result, we use Theorem 0.3. Therefore, we have

$$\begin{aligned}\omega(t) &= se_{\ominus p}(t,0)+\int_0^t fpe_{\ominus p}(t,\tau)\Delta\tau\\ &= se_{\ominus p}(t,0)+\int_0^t fp(\tau)e_p(\tau,t)\Delta\tau\\ &= se_{\ominus p}(t,0)+f(1-e_{\ominus p}(t,0))\\ &= f+(s-f)e_{\ominus p}(t,0),\ \ t\in\mathbb{T},\end{aligned}$$

which completes the proof.

Example 0.19. If $\mathbb{T}=\mathbb{Z}$, then the solution in (69) turns out to be

$$w(t)=f+(s-f)\frac{1}{\prod_{\tau=0}^{t-1}\left[1+k\left(-\frac{b}{b+\left(\frac{f}{f-s}-b\right)(1+k)^{\tau}}+1\right)\right]}, \tag{70}$$

where we use (9).

Let $b=1$ in (67). Note that $p=\frac{k}{f}\omega$ and this means that Eq (66) turns out to be

$$\begin{aligned}\omega^{\Delta} &= -\frac{k}{f}\omega\omega^{\sigma}+k\omega\\ &= k\omega\left(1-\frac{\omega^{\sigma}}{f}\right).\end{aligned} \tag{71}$$

If $\mathbb{T}=\mathbb{R}$, then we obtain (3) from (71). One of the logistic dynamic equations is (54). By taking (54) into account and using the definition of minus circle, we get

$$L^{\Delta}=\left(\frac{p-f_1L}{1+\mu f_1L}\right)x.$$

This implies that

$$L^{\Delta}+\mu L^{\Delta}f_1L=(p-f_1L)L.$$

By the simple useful formula, we have

$$L^{\Delta}+(L^{\sigma}-L)f_1L=(p-f_1L)L.$$

Solving the above equation for L^{Δ}, we get

$$\begin{aligned}L^{\Delta} &= (p-f_1L^{\sigma})L\\ &= pL\left(1-\frac{f_1}{p}L^{\sigma}\right).\end{aligned} \tag{72}$$

If we take $L=\omega$, $p=k$ and $\frac{f_1}{p}=\frac{1}{f}$ in (72), then we obtain Eq (71).

At this point the following question arises: Is it possible to find an alternative 4-parameter logistic dynamic equation which turns out to be the equivalent form of the 3-parameter logistic dynamic equation? To find the answer of this question consider Eq (65) where e_k is replaced

by e_k^σ in the last two terms of Eq (65). Then we obtain that

$$\begin{aligned}\omega^\Delta &= -k(\omega+\mu\omega^\Delta)+kf-\frac{kfb}{\left(b+\left(\frac{f}{f-s}-b\right)e_k(t,0)\right)\left(b+\left(\frac{f}{f-s}-b\right)e_k^\sigma(t,0)\right)}\\ &+ \frac{kfb}{b+\left(\frac{f}{f-s}-b\right)e_k^\sigma(t,0)}-\frac{kfb}{b+\left(\frac{f}{f-s}-b\right)e_k^\sigma(t,0)}.\end{aligned} \tag{73}$$

Solving the above equation for ω^Δ yields

$$\omega^\Delta(1+k\mu)=-k\omega+kf+\frac{kb}{b+\left(\frac{f}{f-s}-b\right)e_k^\sigma(t,0)}\omega-\frac{kfb}{b+\left(\frac{f}{f-s}-b\right)e_k^\sigma(t,0)},$$

or

$$\begin{aligned}\omega^\Delta &= -k\omega\left[\frac{1-\frac{b}{b+\left(\frac{f}{f-s}-b\right)e_k^\sigma(t,0)}}{1+k\mu}\right]+\frac{kf-\frac{kfb}{b+\left(\frac{f}{f-s}-b\right)e_k^\sigma(t,0)}}{1+k\mu}\\ &= (\ominus k)\omega\left[1-\frac{b}{b+\left(\frac{f}{f-s}-b\right)e_k^\sigma(t,0)}\right]-f(\ominus k)\left[1-\frac{b}{b+\left(\frac{f}{f-s}-b\right)e_k^\sigma(t,0)}\right].\end{aligned}$$

Hence, we get the following logistic dynamic equation

$$\omega^\Delta=(\ominus k)q\omega-f(\ominus k)q,\quad t\in\mathbb{T}, \tag{74}$$

where

$$q(t)=1-\frac{b}{b+\left(\frac{f}{f-s}-b\right)e_k^\sigma(t,0)},\quad t\in\mathbb{T}. \tag{75}$$

Theorem 0.20. *Let k, f, s be positive constants and q be taken as in (75). Then, Eq (74) with the initial condition (68) has the solution*

$$\omega=(s-f)e_{(\ominus k)q}(t,0)+f,\quad t\in\mathbb{T}. \tag{76}$$

Proof. By using the definition of minus circle (6), we have

$$\begin{aligned}1+\mu[\ominus((\ominus k)q)] &= 1+\mu\left[\ominus\frac{-kq}{1+\mu k}\right]\\ &= 1+\mu\frac{kq}{1+\mu k-\mu kq}\\ &= \frac{1+\mu k}{1+\mu k-\mu kq}.\end{aligned} \tag{77}$$

By using Theorem 0.2 and the properties of exponential functions given Theorem 0.1, we

arrive at the unique solution as follows:

$$\begin{aligned}
\omega(t) &= se_{(\ominus k)q}(t,0) - f\int_0^t e_{(\ominus k)q}(t,\sigma(\tau))(\ominus k)q(\tau)\Delta\tau \\
&= se_{(\ominus k)q}(t,0) - fe_{(\ominus k)q}(t,0)\int_0^t e_{\ominus((\ominus k)q)}(\sigma(\tau),0)(\ominus k)q(\tau)\Delta\tau \\
&= se_{(\ominus k)q}(t,0) \\
&\quad -fe_{(\ominus k)q}(t,0)\int_0^t e_{\ominus((\ominus k)q)}(\tau,0)[1+\mu(\tau)\ominus((\ominus k)q(\tau))]\frac{-kq(\tau)}{1+\mu(\tau)k}\Delta\tau \\
&= se_{(\ominus k)q}(t,0) \\
&\quad +fe_{(\ominus k)q}(t,0)\int_0^t e_{\ominus((\ominus k)q)}(\tau,0)\frac{1+\mu(\tau)k}{1+\mu(\tau)k-\mu(\tau)kq(\tau)}\frac{kq(\tau)}{1+\mu(\tau)k}\Delta\tau \\
&= se_{(\ominus k)q}(t,0) + fe_{(\ominus k)q}(t,0)\int_0^t e_{\ominus((\ominus k)q)}(\tau,0)\ominus((\ominus k)q(\tau))\Delta\tau \\
&= se_{(\ominus k)q}(t,0) + f(1-e_{(\ominus k)q}(t,0)) \\
&= (s-f)e_{(\ominus k)q}(t,0) + f, \quad t\in\mathbb{T},
\end{aligned}$$

which completes the proof.

Example 0.21. If $\mathbb{T}=\mathbb{Z}$, then the solution in (76) turns out to be

$$w(t) = f + (s-f)\prod_{\tau=0}^{t-1}\left[1+\frac{k}{k+1}\left(\frac{b}{b+\left(\frac{f}{f-s}-b\right)(1+k)^{\tau+1}}-1\right)\right]. \tag{78}$$

If we take $b=1$ in (75), then we get $q=\frac{\omega^\sigma}{f}$. Then Eq (74) turns out to be

$$\omega^\Delta = \frac{\ominus k}{f}\omega^\sigma(\omega-f). \tag{79}$$

Furthermore, if $\mathbb{T}=\mathbb{R}$, (79) turns out to be the logistic differential Eq (3). Note that logistic dynamic Eq (79) is equal neither (53) nor (54). Therefore, (78) with $b=1$ is different than (60) and (61). It means that obtaining 3-parameter logistic curves from 4-parameter logistic curves yields new 3-parameter logistic discrete curves. The following example consists of two new 3-parameter logistic discrete curves.

Example 0.22. If we let $\mathbb{T}=\mathbb{Z}$ and $b=1$ in (70) and (78), then we obtain

$$w(t) = f + (s-f)\frac{1}{\prod_{\tau=0}^{t-1}\left[1+k\left(-\frac{1}{1+\left(\frac{f}{f-s}-1\right)(1+k)^{\tau}}+1\right)\right]}, \tag{80}$$

and

$$w(t) = f + (s-f)\prod_{\tau=0}^{t-1}\left[1+\frac{k}{k+1}\left(\frac{1}{1+\left(\frac{f}{f-s}-1\right)(1+k)^{\tau+1}}-1\right)\right], \tag{81}$$

respectively.

Goodness-of-fit test for gompertz and logistic curves and conclusion

The main aim of this study for the statistical analysis of Gompertz and Logistic curves is to determine whether their equations are able to model Pseudomonas Putita and tumor data sets given in [10] and [11]. In order to achieve our goal, p-values of the parameters, adjusted R-squared values and six types of errors, namely, RMSE (root mean square error), RRMSE (Relative Root Mean Square Error), MAPE (Mean Absolute Percent Error), MAE (mean absolute error), U1 (Theil inequality coefficient, Theil's U1), U2 (Theil inequality coefficient, Theil's U2) for each data set calculated, where

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} e_t^2}{T}}, \quad RRMSE = \sqrt{\frac{\sum_{t=1}^{T} |\frac{e_t}{y_t}|^2}{T}}, \quad MAE = \frac{\sum_{t=1}^{T} |e_t|}{T}$$

$$MAPE = \frac{\sum_{t=1}^{T} \frac{|e_t|}{y_t}}{T}, \quad U1 = \frac{\sqrt{\frac{\sum_{t=1}^{T} e_t^2}{T}}}{\sqrt{\frac{\sum_{t=1}^{T} y_t^2}{T}} + \sqrt{\frac{\sum_{t=1}^{T} \hat{y}_t^2}{T}}}, \quad U2 = \frac{\sqrt{\frac{\sum_{t=1}^{T} e_t^2}{T}}}{\sqrt{\frac{\sum_{t=1}^{T} y_t^2}{T}}}$$

Here, e_t shows the error component, y_t the original time series values, $\hat{y}_t$ the estimated values of the time series, and T the number of observations of the series. The criteria to determine an equation showing better performance in terms of goodness of fit is to have statistically significant coefficients; in other words, meaningful p-values for each coefficient, the larger adjusted R-squared value and smaller errors (see S1 and S2 Figs). The coefficient estimates of these models are obtained by the Mathematica 11.0 and the Wolfram Language uses "ConjugateGradient", "Gradient", "LevenbergMarquardt", "Newton", "NMinimize", and "QuasiNewton" methods.

The curves which successfully model Pseudomonas putita data are 4-parameter first type continuous Gompertz curve (18), 3-parameter first type Gompertz curves (26), (27), (29), Zwietering modification of continuous Gompertz curve (31), Gompertz-Liard curves (35), (38), and 2-parameter second type Gompertz curves 49α0, 50α0, 51α0 and 52α0 that are obtained from (49), (50), (51), (52) with $t_0 = 1$ and $\alpha = 0$. According to the results in S1 Fig Eq (18) has the best fit among Gompertz curves. Therefore, we observe that the performance of 4-parameter Gompertz curves for bacteria is better than 3 and 2-parameter Gompertz growth curves. In addition to these, Eqs (27), (29) and (38) are the 3- parameter discrete Gompertz growth curves are new, thus, contribute to the literature.

When we concentrate attention on the Logistic type growth curves, from S1 Fig, it is apparent that all of the Logistic type equations are successful in modeling the bacteria data set. In addition, growth curves (70), (78), (80), (81) are the new 4-parameter discrete Logistic curves that are obtained in this study. According the results in S1 Fig, among the Logistic type growth curves, 4-parameter Logistic growth curves are better in modeling when it is compared with 3-parameter ones.

Moreover, we can infer that the 4-parameter continuous Logistic curve (63) and (70), (78) model better than Eq (18). Thus, Pseudomonas Putita data is modeled by Logistic type equations better than Gompertz type equations. In addition, Eq (63) highlighted with orange, (70) and (78) highlighted with green in S1 Fig, have the smallest errors among the other Logistic equations. Eqs (70) and (78) have smaller errors after the eighth decimal when they are compared with Eq (63) highlighted yellow in S1 Fig. Therefore, new discrete growth curves (70) and (78) are the best in modeling bacteria data.

Eq (29) highlighted with yellow, (35) highlighted with green, 49α0, 50α0, 51α0 and 52α0 highlighted with orange in S2 Fig are the curves that successfully model the tumor data among the Gompertz curves. Eq (35) has the best fit in modeling based on S2 Fig. This equation is the Gompertz Liard continuous equation that was developed for tumor modeling in the literature, so our result is compatible with the one in [14]. On the other hand, 3-parameter discrete Gompertz growth curve (29) also models the tumor data set and this equation is developed in this study as well. In addition, 4-parameter continuous and discrete second type Gompertz curves (51) and (52) are also new. 2-parameter version 51a0 of (51) was studied in [8] as Mirror Gompertz curve. Nevertheless, its discrete version 52α0 is a new contribution to the literature. At this point, we declare that 3-parameter Gompertz curves are more successful in modeling tumor data than 4-parameter Gompertz curves. Although the errors of the Logistic type curves are smaller than the errors of Gompertz type curves, all of their parameters are not statistically significant. Therefore, the Gompertz-Liard curve (35) is the best curve in modeling when it is compared with all the other curves and so one can say that Gompertz type curves have better fitting than Logistic type curves for tumor data.

As a result, Logistic curves are better in modeling bacteria data whereas tumor data is modeled better by Gompertz curves. Increasing the number of parameters of Logistic curves give favorable results for bacteria data while decreasing the number of the parameters of Gompertz curves for tumor data turns out to be reasonable. S3 Fig gives us the curve fittings of 4-parameter discrete Logistic curve (70) and 3-parameter discrete Gompertz curve (29) for bacteria data set and also shows the importance of the number of parameters.

Supporting information

S1 Fig. Bacteria. Fitted parameters and statistical error analysis for bacteria. *: significant at.10 level, **: significant at.05 level and ***: significant at.01 level.

S2 Fig. Tumor. Fitted parameters and statistical error analysis for tumor data. *: significant at.10 level, **: significant at.05 level and ***: significant at.01 level.

S3 Fig. Compare. Compare with 4-parameter discrete Logistic curve and 3-parameter discrete Gompertz curve for bacteria data set.

S1 File. Bacteria. Data set for bacteria.

S2 File. Tumor. Data set for tumor.

Author Contributions

Formal analysis: Elvan Akın, Neslihan Nesliye Pelen.

Methodology: Elvan Akın, Neslihan Nesliye Pelen, Ismail Uğur Tiryaki, Fusun Yalcin.

Software: Ismail Uğur Tiryaki, Fusun Yalcin.

Supervision: Elvan Akın.

Validation: Elvan Akın, Neslihan Nesliye Pelen, Ismail Uğur Tiryaki, Fusun Yalcin.

Writing – original draft: Elvan Akın, Neslihan Nesliye Pelen.

Writing – review & editing: Elvan Akın.

References

1. Alverez-Arenas Arturo, Belmonte-Beitia J., and Calvo Gabriel Nonlinear waves in a simple model of hig-grade glioma. Applied Mathemaicts and Nonlinear Sciences. Volume 1: Issue 2, p.405–422 (2016).
2. Atıcı F. M., Atıcı M., Hrushesky W.J.M., Nguyen N. Modeling Tumor Volume with Basic Functions of Fractional Calculus. Progr. Fract. Differ. Appl. 1, No. 4, 229–241 (2015).
3. Bassukas I. D., Schultze B. M. The recursion formula of the Gompertz function: A simple method for the estimation and comparison of tumor growth curves. Growth Dev. Aging, Vol.52, (1988), pp.113–122.
4. Domingues Jose Sergio. Gompertz Model: Resolution and Analysis for Tumors. Journal of Mathematical Modelling and Application 1, No. 7, 70–77 (2012).
5. Durbin P. W., Jeung N., Williams M. H., and Arnold J. S. Construction of a Growth Curve for Mammary Tumors of the Rat. Cancer Research. Volume 27, p.1341–1347 (1967).
6. Rojas Clara, and Belmonte-Beitia J. Optimal control problems for differential equations applied to tumor growth: state of the art. Applied Mathemaicts and Nonlinear Sciences. Volume 3: Issue 2, p.375–402 (2018).
7. Dudek Krzysztof, Kedzia Wojciech, Kedzia Emilia, and Kedzia Alicja. Mathematical modelling of the growth of human fetus anatomical structures. Anat Sci Int.
8. Skiadas C.H.Comparing the Gompertz-Type Models with a First Passage Time density Model. Advances in Data Analysis. Springer/Birkhauser (2010), pp. 203–209.
9. Kelley W., and Peterson A. The Theory of Differential Equations: Classical and Qualitative. Springer, Second Edition.
10. Annadurai G., Rajesh Babu S., Srinivasamoorthy V. R. Development of mathematical models (Logistic, Gompertz and Richards models) describing the growth pattern of Pseudomonas putida(NICM 2174). Bioprocess Engineering, Vol.23, (2000), pp.607–612.
11. Şengül, S. Discrete fractional calculus and its applications to tumor growth. Master thesis, Paper 161. http://digitalcommons.wku.edu/theses/161, 2010.
12. Espinosa-Urgel M., Salido A., Ramos J. Genetic Analysis of Functions Involved in Adhesion of Pseudomonas putida to Seeds. Journal of Bacteriology. Volume 182, p.2363–2369 (2000).
13. Perz J., Craig A. S., Stratton C. W., Bodner S. J., Philipps W. E. Jr., and Schaffner W. Pseudomonas putida Septicemia in a Special Care Nursery Due to Contaminated Flush Solutions Prepared in a Hospital Pharmacy. Journal of Clinical Microbiology, volume 43, p.5316–5318 (2005).
14. Tjorve K.M.C., Tjorve E. The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family. PLOS ONE https://doi.org/10.1371/journal.pone.0178691. (2017). PMID: 28582419
15. Wan X., Wang M., Wang G., Zhong W. A new 4 parameter generalized logistic equation and its applications to mammalian somatic growth. Acta Theriologica 45(2):145–153,2000.
16. Akın-Bohner E. and Bohner M. Miscellaneous Dynamic Equations. Methods and Applications of Analysis. Vol 10, No. 1, pp. 011–030, (2003).
17. Bohner M. and Peterson A. C. Dynamic Equations on Time Scales: An Introduction with Applications. Birkhauser (2001).
18. Bohner M. and Peterson A. C. Advances in Dynamic Equations on Time Scales. Birkhauser (2003).
19. Gibson A.M, Bratchell N., Roberts T.A. Predicting microbial growth: growth responses of salmonellae in a laboratory medium as affected by pH, sodium chloride and storage temperature. Int. J. Food Microbiol. 1988; 6:155–78. https://doi.org/10.1016/0168-1605(88)90051-7 PMID: 3275296

Calculating the Malliavin derivative of some stochastic mechanics problems

Paul Hauseux[1], **Jack S. Hale**[1], **Stéphane P. A. Bordas**[1,2]*

1 Institute of Computational Engineering, University of Luxembourg, 6 Avenue de la Fonte, 4362 Esch-sur-Alzette, Luxembourg, **2** Cardiff School of Engineering, Cardiff University, The Queen's Building, The Parade, Cardiff, Wales, CF24 3AA, United Kingdom

* stephane.bordas@uni.lu

Abstract

The Malliavin calculus is an extension of the classical calculus of variations from deterministic functions to stochastic processes. In this paper we aim to show in a practical and didactic way how to calculate the Malliavin derivative, the derivative of the expectation of a quantity of interest of a model with respect to its underlying stochastic parameters, for four problems found in mechanics. The non-intrusive approach uses the Malliavin Weight Sampling (MWS) method in conjunction with a standard Monte Carlo method. The models are expressed as ODEs or PDEs and discretised using the finite difference or finite element methods. Specifically, we consider stochastic extensions of; a 1D Kelvin-Voigt viscoelastic model discretised with finite differences, a 1D linear elastic bar, a hyperelastic bar undergoing buckling, and incompressible Navier-Stokes flow around a cylinder, all discretised with finite elements. A further contribution of this paper is an extension of the MWS method to the more difficult case of non-Gaussian random variables and the calculation of second-order derivatives. We provide open-source code for the numerical examples in this paper.

Editor: Xiao-Jun Yang, China University of Mining and Technology, CHINA

Introduction

The classical derivative is a fundamental tool of calculus that is widely used across every field of mathematics, science and engineering. Various generalisations and extensions of the classical derivative, e.g. local and/or partial Frechét and Gâteaux derivatives [1], are now common tools in the repertoire of practitioners working in many fields. Modern extensions such as fractional and non-local derivatives are finding increasing use in several fields of science and technology, see e.g. [2–6]. The semi-inverse method of [7] is a powerful tool for the establishment of variational principles (Euler-Lagrange) from governing equations for physical problems.

By contrast, the Malliavin calculus [8], an extension of the notions of classical calculus of variations to stochastic processes, is certainly less widely known. In our view, this is probably because the vast majority of papers written on the subject require study of mathematics and stochastics to an advanced level. However, we think that Malliavin calculus deserves a wider audience. The objective of this paper then is introduce the Malliavin derivative as a useful numerical tool for practitioners to understand the behaviour of stochastic PDEs in mechanics,

Funding: We thank the financial support of the European Research Council Starting Independent Research Grant (ERC Stg grant agreement No. 279578) entitled "Towards real time multiscale simulation of cutting in non-linear materials with applications to surgical simulation and computer guided surgery." Paul Hauseux is supported by the internal MOMENTUM project at the University of

Luxembourg. Jack S. Hale is supported by the National Research Fund, Luxembourg, and cofunded under the Marie Curie Actions of the European Commission (FP7-COFUND) Grant No. 6693582. We also thank the funding from the Luxembourg National Research Fund (INTER/MOBILITY/14/8813215/CBM/Bordas).

Competing interests: The authors have declared that no competing interests exist.

rather than to fully explain the technicalities of Malliavin calculus. Interested readers are referred to e.g. [8–10] for a full mathematical treatment.

We are not the first to apply Malliavin calculus as a useful tool for practical computation. The Malliavin calculus can be used to efficiently calculate the Greeks, the sensitivity of financial instruments to their underlying parameters e.g. [11–14]. In the physical sciences we are aware of only a handful of recent papers that use techniques inspired by the Malliavin calculus to understand the behaviour of systems with stochastic behaviour. We are not aware of any papers in the engineering mechanics community on the topic. [15] introduced the methodology of Malliavin Weight Sampling (MWS), the method we adopt in this paper, and applied it to the simulation of particles undergoing Brownian motion. [16] presented a more general framework for deriving the MWS update rules and its practical implementation. [17] used the MWS to evaluated linear response functions of particle systems forced by coloured noise. When the coefficients of the models are assumed to follow known statistical distributions, then the likelihood ratio method can be seen as a Malliavin weighting function [11, 12, 18]. The Malliavin theory is however more general and allows the determination of the optimal weight with minimum variance even if the specification of the stochastic parameters involved in the model are not known explicitly.

The contribution of this paper is as follows; we show the application of the Malliavin Weight Sampling method [15] to four archetypal problems in mechanics. Unlike the examples in [15], we consider some models defined by partial differential equations (PDEs) that are discretised using the finite element method. We make a new extension of the MWS method to parameters defined by non-Gaussian distributions. This has important practical value because it is often important to model parameters with distributions that preclude realisations with non-physical values, e.g. positive viscosity in a fluid mechanics problem. Finally we extend the MWS method in [15] to the calculation of second-order derivatives.

An outline of this paper is as follows; we give an outline of the MWS method and use the MWS method to study the behaviour of a simple Kelvin-Voigt visco-elastic system with Gaussian and non-Gaussian stochastic variables respectively. We extend the analysis of the Kelvin-Voigt system to the second derivative. We then study; a 1D elastic bar, a hyperelastic bar prone to buckling, and Navier-Stokes flow around a cylinder, all discretised in space using the finite element method. Finally we summarise and suggest some interesting avenues for future research.

The Malliavin Weight Sampling (MWS) method

Problem setting

Consider a non-linear, possibly time-dependent stochastic partial differential equation $F(u, m) = 0$ with random parameter m. For each possible value of m, u is the solution of the PDE and therefore u depends explicitly on m ($m \mapsto u(m)$). To simplify the notation, the spatial position x and time t are omitted but it is understood that u can also depend on $x \in \Omega \subset \mathbb{R}^d$ where $d = \{1, 2, 3\}$ is the spatial dimension of the domain and/or $t \in \mathbb{R}^+$.

Let $(\Omega_p, \mathcal{F}, P)$ a probability space where Ω_p is the sample space, $\mathcal{F}$ is a σ-algebra of subsets of Ω_p and P is a probability measure. We are interested to evaluate the expected value of a quantity of interest $J(m) = J(u(m))$ denoted by $\mathbb{E}[J]$ [19]:

$$\mathbb{E}[J] \coloneqq \int_{\Omega_p} J(u(\omega)) \cdot dP(\omega). \tag{1}$$

In a practical way, if m is a random variable with probability density function f_m, Eq (1) writes:

$$\mathbb{E}[J] \coloneqq \int_{\mathbb{R}} J(u(x)) \cdot f_m(x) \mathrm{d}x. \tag{2}$$

As we will see, the Malliavin Weight Sampling method (MWS) [16] allows the evaluation of the sensitivity of the expected value of the quantity of interest with respect to the mean value of the stochastic parameter m as [9, 11, 16, 18]:

$$\frac{\partial \mathbb{E}[J]}{\partial \bar{m}} = \mathbb{E}[Jq_m], \tag{3}$$

where q_m is the Malliavin weight for the parameter m and $\bar{m}$ is the mean of m. Under certain condition of regularity [11, 20, 21] when the probability density function (PDF) of the parameter m is known, the Malliavin weight q_m associated can be computed directly from the PDF of m. This approach can be viewed as an integration by parts, and is a direct result of Malliavin calculus where we take the derivative of random functions rather than the classical derivative. We emphasise again the quite different nature of the above derivative Eq (3) to the classical notion of a derivative from elementary calculus.

In Eq (3), we suppose that the quantity of interest J does not depend explicitly on the parameter m. Later we introduce a more general equation Eq (35) that must be considered if in fact J does depend on m.

The simplest approach to calculate $\mathbb{E}[Jq_m]$, and the one we use exclusively in this paper, is to use the standard Monte Carlo estimator; that is, take Z independent and identically distributed (iid) realisations m_z of m, solve for $J_z \coloneqq J(u(m_z))$ before taking the sample mean of the set of realisations $\{J_1, \ldots, J_Z\}$:

$$\frac{\partial \mathbb{E}[J]}{\partial \bar{m}} = \mathbb{E}[Jq_m] \approx \frac{1}{Z} \sum_{z=1}^{Z} J(m_z) \cdot q_m(m_z). \tag{4}$$

From the central limit theorem, the error in Eq (4) is normally distributed with variance $Z^{-1} V$ where V is the variance of Jq_m.

What will not be clear to the reader at this stage is how to determine the Malliavin weights. Through a simple practical examples in the next section, we will explain how to use the MWS method, determine the specification of the weights for both Gaussian and non-Gaussian distributions on the parameter m, and thus calculate the Malliavin derivative Eq (3).

Kelvin-Voigt model

The Kelvin-Voigt constitutive model with Young's modulus E, viscosity η and loading stress σ can be written as the following linear ordinary differential equation:

$$E\epsilon(t) + \eta \frac{\mathrm{d}\epsilon(t)}{\mathrm{d}t} = \sigma. \tag{5}$$

A schematic of this model is shown in Fig 1.

The initial condition on the strain is $\epsilon(t = 0) = 0$ and we study the response of the system for time $t \in [0, T]$. Our quantity of interest functional is the value of the strain at time t, i.e.:

$$J \coloneqq \epsilon(t) \tag{6}$$

and we are interested in its expected value (mean) $\mathbb{E}[\epsilon(t)]$.

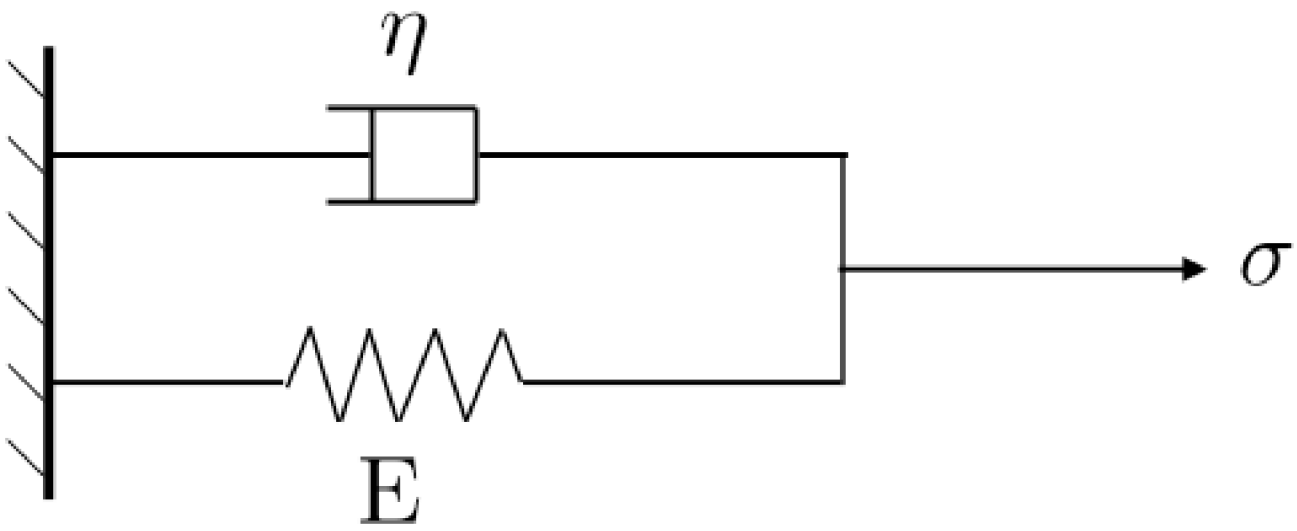

Fig 1. Schematic of a Kelvin-Voigt model with Young's modulus *E*, viscosity *η* and loading stress *σ*. We model the loading stress σ as a random noise (random variable), inducing a random strain ϵ as the output of the model.

Gaussian case

We first consider the case that the randomness can be modelled as a Gaussian random variable. A similar model is shown in [16].

We choose choose to model the stress as a random noise:

$$\sigma(t) = \sigma_0 + \alpha\xi, \tag{7}$$

where σ_0 and α are constant and ξ is a a Gaussian random variable with zero mean and unit variance. ξ represents the uncertainty related to the value of the stress σ.

From Eq (7), the mean value of σ is σ_0 and the variance of σ is equal to α^2. We assume throughout that the Young's modulus E and the viscosity η are perfectly known. Given that the forcing stress σ for the system is random, the strain ϵ is also random. The goal then is to evaluate the derivative of the expected value of the strain with respect to the mean value σ_0:

$$\frac{\partial \mathbb{E}[\epsilon(t)]}{\partial \sigma_0} \tag{8}$$

using the method of MWS.

We choose to solve the ODE Eq (5) using an Euler explicit finite difference method with time step δt:

$$\epsilon(t+\delta t) = \epsilon(t) + \frac{\delta t}{\eta}\left[\sigma_0 - E\epsilon(t) + \frac{\alpha\xi}{\sqrt{\delta t}}\right]. \tag{9}$$

Remark. Note that the multiplying term before ξ contains $\sqrt{\delta t}$ and not δt. This is a 'conforming' discretisation of the stochastic noise term, resulting in a dependence of the variance of the random parameter on the discretisation size. Informally, taking the limit, we can recover the original ODE Eq (5) as:

$$\mathbb{E}\left[\frac{\mathrm{d}\epsilon}{\mathrm{dt}}\right] = (\sigma_0 - E\epsilon(t))/\eta, \tag{10}$$

and:

$$\mathbb{V}\left[\eta\frac{\mathrm{d}\epsilon}{\mathrm{dt}}-(\sigma_0-E\epsilon(t))\right]=\alpha^2. \tag{11}$$

Where $\mathbb{V}[\cdot]$ is the the variance. Given that $\mathbb{E}[\sqrt{\delta t}\xi]=0$ and $\mathbb{E}[(\sqrt{\delta t}\alpha\xi)^2]=\alpha^2\delta t$, the numerical method in Eq (9) is consistent in the following sense:

$$\lim_{\delta t\to 0}\frac{1}{\delta t}\mathbb{E}[\epsilon(t+\delta t)-\epsilon(t)]=(\sigma_0-E\epsilon(t))/\eta, \tag{12}$$

$$\lim_{\delta t\to 0}\frac{1}{\delta t}\mathbb{E}[[\eta(\epsilon(t+\delta t)-\epsilon(t))-\delta t(\sigma_0-E\epsilon(t))]^2]=\alpha^2. \tag{13}$$

For this next part, we adopt the same notation as [16]. We denote ϵ the strain of the the system at time t and we denote ϵ' the strain of the the system at time $t+\delta t$. Furthermore, we let $P(\epsilon)$ and $P(\epsilon')$ be the probability that the strain of the system is ϵ and ϵ' respectively. The propagator $W(\epsilon\to\epsilon')$ must satisfy:

$$P(\epsilon')=\int_{\epsilon}W(\epsilon\to\epsilon')P(\epsilon)\mathrm{d}\epsilon, \tag{14}$$

$$\int_{\epsilon'}W(\epsilon\to\epsilon')\mathrm{d}\epsilon'=1. \tag{15}$$

Eq (14) means that the probability that the strain of the system is ϵ' is the sum (integral) of all the probabilities to be at ϵ multiplied by the probability that the system passes from state ϵ to ϵ' during δt. Condition Eq (15) comes from the integration of the first condition over ϵ'.

To derive the analytical expression of the propagator in Eq (18) we start with the fact that $\xi^\star=\alpha\sqrt{\delta t}\xi$ is Gaussian $N\sim(0,\delta t\alpha^2)$, hence the probability density function is known and must satisfy the following condition:

$$\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\delta t\alpha^2}}\exp\left(-\frac{\xi^{\star 2}}{2\delta t\alpha^2}\right)\mathrm{d}\xi^\star=1. \tag{16}$$

With an integration by substitution from Eq (9), with:

$$\xi^\star=((\epsilon'-\epsilon)\eta-\delta t\sigma_0+\delta tE\epsilon), \tag{17}$$

we can then show the expression of the propagator, the probability that the system passes from state ϵ to ϵ' during δt is given by [16]:

$$W(\epsilon\to\epsilon')=\frac{\eta}{\sqrt{2\pi\delta t\alpha^2}}\exp\left(-\frac{((\epsilon'-\epsilon)\eta-\delta t\sigma_0+\delta tE\epsilon)^2}{2\delta t\alpha^2}\right). \tag{18}$$

With the propagator in hand we will now see how it is possible to evaluate the Malliavin derivative with the MWS method. To recap, we denote by $J(\epsilon)$ a quantity of interest of our system and we want to compute the derivative of the mean value of this quantity of interest $\mathbb{E}[J]$ with respect to a parameter $\bar{m}$, in this case σ_0 when $\sigma=\sigma_0+\alpha\xi$.

The form of the Malliavin weights q_m can be obtained using the following procedure. First, we know that with $dP(\epsilon)=P(\epsilon)d\epsilon$, Eq (1) we can write:

$$\mathbb{E}[J]=\int_{\epsilon}JP(\epsilon)\mathrm{d}\epsilon, \tag{19}$$

and by taking the derivative of Eq (19) [11, 16, 18]:

$$\frac{\partial \mathbb{E}[J]}{\partial \bar{m}} = \int_{\epsilon} JP(\epsilon) \frac{\partial \ln P}{\partial m} \mathrm{d}\epsilon. \tag{20}$$

To define a set of rules for updating q_m, we differentiate Eq (14) with respect to m:

$$P(\epsilon') \frac{\partial \ln P'}{\partial \bar{m}} = \int_{\epsilon} W(\epsilon \to \epsilon') P(\epsilon) \frac{\partial \ln P + \partial \ln W}{\partial m} \mathrm{d}\epsilon, \tag{21}$$

and we obtain the following rule for updating the Malliavin weight:

$$q_m(t + \delta t) = q_m(t) + \frac{\partial \ln W}{\partial m}. \tag{22}$$

In the example of random stress with $\sigma = \sigma_0 + \xi$, we obtain:

$$\frac{\partial \ln W}{\partial \sigma} = \sqrt{\delta t} \xi / \alpha. \tag{23}$$

For random Young's modulus $E = E_0 + \xi$ we would have the same expression. In the case of random viscosity $\eta = \eta_0 + \xi$ we would obtain following the same logic:

$$\frac{\partial \ln W}{\partial \eta} = \xi / \alpha. \tag{24}$$

With the expression for the Malliavin weight Eq (23) in hand we can now implement an algorithm to calculate the derivative. The procedure is very simple; we take Z samples of the evolutions of the stochastic ODE using the explicit Euler scheme whilst simultaneously evolving the Malliavin weight q_m. At teach time step Algorithm 1 describes this procedure in more detail.

The deterministic constants are given to be $\eta = 1$, $E = 1$ and we take a time step of $\delta t = 0.01$ for the finite difference scheme. We evaluate by the MWS method the derivative of the expected value of ϵ with respect to σ_0 for a loading time $t \in [0, T]$ with $T = 30s$. In this example the number of realisations is fixed at $Z = 20000$. We compare the results with the analytical solution which is:

$$\frac{\partial \mathbb{E}[\epsilon]}{\partial \sigma_0} = \frac{1}{E} \left(1 - \exp\left\{ -\frac{Et}{\eta} \right\} \right). \tag{25}$$

We briefly remark that for all numerical results presented in this paper there are two sources of errors committed with respect to the undiscretised problem. The first error is due to the deterministic approximation of the PDE (finite difference or finite element method), and the second due to the stochastic approximation (Monte Carlo estimator). In all cases we drive the error in the deterministic approximation of the PDE far lower than that in the stochastic approximation, such that the error is dominated by the number of realisations Z used in the Monte Carlo estimator.

In Fig 2 we can see that the MWS method gives a good estimation of the Malliavin derivative, particularly in the non-steady state regime $t \leq 5s$. The relative error and so the global statistical error can become very high when the system reaches a steady state because the value of the sensitivity derivative is constant but the statistical error is compounded after each time step. To address this issue a technique can be employed based on the correlation function [16]. The reader is referred to [16] for further details. We have implemented this correlation correction, which we denote MWS-steady-state, and we can see in Fig 2 that the error is greatly

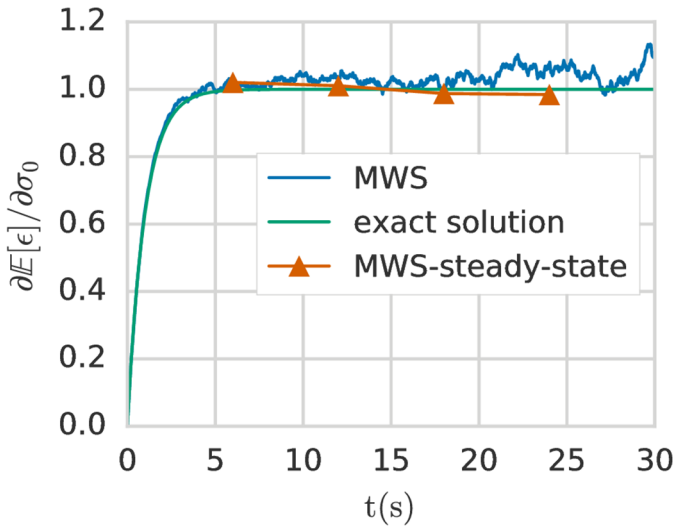

Fig 2. Malliavin derivative of the expected value of the strain with respect to the loading σ_0 for the Kelvin-Voigt model with uncertain stress modelled as a Gaussian random variable. Comparison between the exact solution, the MWS method and the the MWS-steady-state method with a correction using the correlation function to improve the convergence of the MWS method when the system transitions into the steady state. For the MWS method we use $Z = 20000$ realisations at each time step.

reduced in the steady-state regime. For the numerical examples presented in the following sections, we will consider only the systems undergoing transition or purely steady state systems. Therefore we will not use the MWS-steady-state method again in this paper.

Non-Gaussian case

In this section we explain how to calculate the derivative for non-Gaussian stochastic parameters using the MWS method. The procedure is similar to that shown in the previous part but the rule for updating the Malliavin weight must be modified.

We begin as before by considering uncertainty in the stress σ:

$$\begin{aligned} \sigma &= \sigma_1 + c\xi \\ &= \underbrace{\sigma_1 + c\mathbb{E}[\xi]}_{\sigma_0} + c(\xi - \mathbb{E}[\xi]). \end{aligned} \tag{26}$$

where ξ a random variable with probability density function $f(x)$ and c and σ_1 are two constants. We have written Eq (26) in the form of a constant $\sigma_0 = \sigma_1 + c\mathbb{E}[\zeta]$ plus a random variable $c(\zeta - \mathbb{E}[\zeta])$ with zero mean. Then it follows that σ_0 is the mean of the uncertain stress σ. We will use the MWS method to evaluate the derivative of the expected value of the quantity of interest with respect to σ_0.

Algorithm 1: Malliavin Weight Sampling algorithm for time dependent problem. The notation used is that of the Kelvin-Voigt example in but the same basic algorithm is used throughout the paper. Note that a correction term is needed for systems in steady state, see [16].

Data: σ_0, E, η and the random variable $\xi \in N(0, 1)$.
Result: $\partial\mathbb{E}[\epsilon(t)]/\partial\sigma_0$, the derivative of the mean of ϵ with respect to the mean stress σ_0 at time t.

```
∂𝔼[ϵ(t)]/∂σ_0 = 0 for all t.                          ▷initialisation
for z = 0 to Z - 1 do
  t = 0;                                                  ▷time
  ϵ(t) = 0;                                  ▷initial condition
  q_σ = 0;                                          ▷MWS weight
  for i = 1 to n do
    Draw realisation of random variable ξ_i;
    ϵ(t + δt) = ϵ(t) + σ_0 δt/η − E ϵ(t) δt/η + √δt ξ_i/η;
    q_σ(t + δt) = q_σ(t) + ∂lnW/∂σ = q_σ + √δt ξ_i;
    ∂𝔼[ϵ(t + δt)]/∂σ_0 += ϵ(t + δt) q_σ/Z;
    t += δt;
  end
end
```

To be able to use the MWS method the probability density function on the parameter must satisfy some regularity conditions, see [11, 20, 21]. Intuitively, the probability density function must be sufficiently "regular" on $\mathbb{R}$ which holds for the Gaussian, log-normal, Beta($\alpha > 1, \beta > 1$) and Gamma($k > 1, \theta$) distributions. However, a uniform distribution between two values a and b can not be considered "regular" because the probability density function is not differentiable at a and b. Instead, we choose to regularised approximation of a uniform distribution using a Beta($1 + e, 1 + e$) random variable with $e \ll 1$.

Continuing, we again discretise Eq (5) using an explicit Euler method with time step δt:

$$\epsilon(t + \delta t) = \epsilon(t) + \frac{\delta t}{\eta}\left(\sigma_0 - E_0\epsilon(t) + \frac{c}{\sqrt{\delta t}}(\xi - \mathbb{E}[\xi])\right). \tag{27}$$

Alternatively, in the case of uncertainty in the Young's modulus, the discretisation can be written:

$$\epsilon(t + \delta t) = \epsilon(t) + \frac{\delta t}{\eta}\left(\sigma_0 - E_0\epsilon(t) + \frac{c\ \epsilon(t)}{\sqrt{\delta t}}(\mathbb{E}[\xi] - \xi)\right), \tag{28}$$

or, for uncertainty related to the viscosity:

$$\epsilon(t + \delta t) = \epsilon(t) + \frac{\sigma_0 - E\epsilon(t)}{\xi + \eta}\delta t. \tag{29}$$

The probability density function of the beta distribution, for $0 \leq x \leq 1$, and shape parameters $\alpha, \beta > 0$, is:

$$f(x; \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)}x^{\alpha-1}(1 - x)^{\beta-1}. \tag{30}$$

The beta function B is a normalisation constant to ensure that the total probability integrates to 1. In general we will evaluate and update the Malliavin weight for the parameter m as:

$$q_m(t + \delta t) = q_m(t) + \frac{\partial \ln W}{\partial m} = q_m(t) + \frac{\partial \ln f(\xi)}{\partial \xi}\frac{\partial \xi}{\partial m}. \tag{31}$$

Note that in Eq (31), it is important to check that the condition $\mathbb{E}[q_m(t)] = 0$ is verified. If $\mathbb{E}[q_m(t)] \neq 0$, the updated rule must be corrected. An example of performing this correction is given in the next section entitled extension to second derivative. Finally, we note that for the initial condition we always impose $q_m(t = 0) = 0$.

For the uncertain Young's modulus modelled with a beta distribution, we have:

$$\frac{\partial \ln W}{\partial m} = \frac{(\beta - 1)\sqrt{\delta t}}{c(1-m)} - \frac{(\alpha - 1)\sqrt{\delta t}}{c\ m}. \tag{32}$$

For the uncertain stress with beta distribution, we have:

$$\frac{\partial \ln W}{\partial m} = \frac{(\beta - 1)\sqrt{\delta t}}{c(1-m)} - \frac{(\alpha - 1)\sqrt{\delta t}}{c\ m}. \tag{33}$$

For the uncertain viscosity with beta distribution, we have:

$$\frac{\partial \ln W}{\partial m} = \frac{\beta - 1}{c(1-m)} - \frac{\alpha - 1}{c\ m}. \tag{34}$$

These results and further calculations are summarised in Table 1.

The Malliavin derivatives of the Kelvin-Voigt model with respect to the mean of the three parameters $\{\sigma_0, \eta_0, E_0\}$ modelled as beta(2, 2) distributions are shown in Fig 3. The exact solution is computed semi-analytically using standard integration rules. Good agreement between the MWS and semi-analytical solution is observed for E_0 and σ_0. For the viscosity η_0 the number of Monte Carlo samples is not sufficient to achieve negligible error, but the overall trend is followed.

Extension to second derivative

The MWS method can also be used to compute the second Malliavin derivative of the expected value of a quantity of interest J. If the quantity of interest does not depend explicitly of the random parameter, the expression given in Eq (3) is valid, but the more general form is the following:

$$\frac{\partial \mathbb{E}[J]}{\partial \bar{m}} = \mathbb{E}\left[\frac{\partial J}{\partial \bar{m}}\right] + \mathbb{E}[J q_m]. \tag{35}$$

In Eq (35), when we want to compute the second derivative the term $\left[\frac{\partial J}{\partial \bar{m}}\right]$ does not vanish because in this case J is the first derivative with respect to $\bar{m}$ and therefore depends on the parameter $\bar{m}$ in general. By applying Eq (35), we can show for example in the case of uncertain Young's modulus that:

$$\frac{\partial^2 \mathbb{E}[\epsilon(t)]}{\partial E_0^2} = \mathbb{E}[\epsilon(t)(q_{EE}(t) + q_E(t)^2 - C_{EE} - C_E^2)], \tag{36}$$

Table 1. Summary of main results for Kelvin-Voigt model with three distributions on three different model parameters.

distribution	$\frac{\partial \ln W}{\partial \sigma}$ or $\frac{\partial \ln W}{\partial E}$	$\frac{\partial \ln W}{\partial \eta}$
Beta(α, β)	$\frac{(\beta-1)\sqrt{\delta t}}{c(1-m)} - \frac{(\alpha-1)\sqrt{\delta t}}{mc}$	$\frac{(\beta-1)}{c(1-m)} - \frac{(\alpha-1)}{mc}$
Gamma(κ, θ)	$\frac{-(\kappa-1)\sqrt{\delta t}}{mc} + \frac{\sqrt{\delta t}}{\theta c}$	$\frac{-(\kappa-1)}{mc} + \frac{1}{\theta c}$
log-Normal(μ, σ)	$\frac{\sqrt{\delta t}}{mc}\left(1 + \frac{\ln m - \mu}{\sigma^2}\right)$	$\frac{1}{mc}\left(1 + \frac{\ln m - \mu}{\sigma^2}\right)$

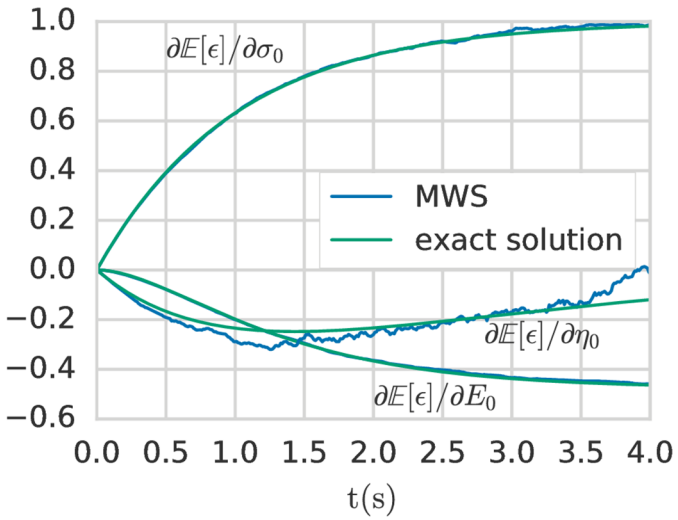

Fig 3. Malliavin derivatives of the expected value of the strain with respect to the mean of the stochastic parameters (Young's modulus E_0, viscosity η_0 and stress loading σ_0). Comparison between the exact solution and the MWS method. All uncertain parameters are modeled with a beta(2, 2) distribution. $Z = 10^5$ realisations are performed for each estimator and the mean value of 10 estimators is plotted for each parameter. Note that the value of Z is not large enough for the viscosity to converge with an negligible error compared to the two other parameters. By increasing Z, this error could be reduced.

with the following updating rule:

$$q_{EE}(t+\delta t) = q_{EE}(t) + \frac{\partial^2 \ln W}{\partial E^2}, \tag{37}$$

and:

$$q_E(t+\delta t) = q_E(t) + \frac{\partial \ln W}{\partial E}. \tag{38}$$

The constant C_E^2 and C_{EE} allow to ensure that the expected value of the global Malliavin weight $(q_{EE}(t) + q_E(t)^2 - C_{EE} - C_E^2)$ has an expected value equal to zero. In this specific case we have:

$$C_E^2 = \mathbb{E}\left[\left(\frac{\partial \ln W}{\partial E}\right)^2\right], \tag{39}$$

$$C_{EE} = \mathbb{E}\left[\frac{\partial^2 \ln W}{\partial E^2}\right]. \tag{40}$$

The precise specification of the constants depends on the distribution. We compute them analytically or by using standard numerical integration techniques found in e.g. Scipy or Maple.

In Fig 4, a comparison between the analytical solution and the MWS method is given for the value of the second derivative depending on time of the expected value of ϵ with respect to the Young's modulus. For the sake of example, the problem specification is the same as in previous sections, with the exception that the random variable follows a log-normal(μ, σ) distribution with mean equal to 0.5 and standard deviation equal to 0.25 which corresponds to μ =

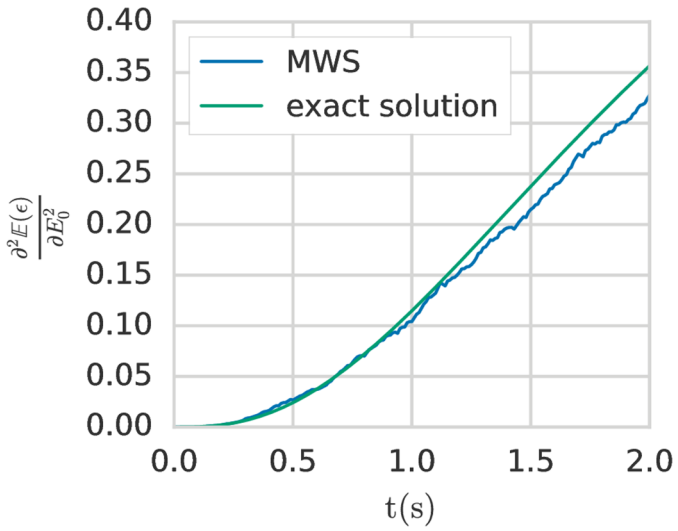

Fig 4. Second sensitivity derivative of the expected value of the strain with respect to the Young's modulus for the Kelvin-Voigt model. Comparison between the exact solution solution and the MWS method. The Young's modulus is modelled with a log-normal distribution. For the MWS method, $Z = 10^7$ realisations are performed. Note that the value of Z for the same order of magnitude for the error is higher for the second derivative compared to the first derivative because the variance V in the Malliavin estimator is bigger and we know from the central limit theorem that the error is in $\mathcal{O}(V^{1/2}Z^{-1/2})$.

-0.804 and $\sigma = 0.473$. The analytical solutions for the two constants C_{EE} and C_E^2 are in this case:

$$C_{EE} = C_E^2 = \left(1 + \frac{1}{\sigma^2}\right) \exp\left(2\sigma^2 - 2\mu\right). \tag{41}$$

As we can see in Fig 4, the MWS method gives a good approximation for the evaluation of the second derivative with $Z = 10^7$ realisations.

Extension to random process

In this paper we deal with random noise and in the next section we show numerical results of stochastic mechanics problems where models are defined as PDEs. The probability density function of the random variables used in these examples does not depend on time. Similarly to the Kelvin-Voigt model presented before, we study a time dependent problem in a finite dimensional space by splitting the time interval $[0, T]$ into a finite number of increments. Note that it is also possible to take into account the random noise only at the initial time instead of generating random variables at each time step. It would be possible to extend this work to random process, e.g. by using a Wiener process $W(t)$ which verifies in particular $(W(t + \delta t) - W(t)) \sim N(0, \delta t)$. In this case, even for simple ODEs, it is very difficult to obtain analytical solutions because the probability density function of a random process evolves in time. The Malliavin calculus is very well adapted to address these stochastic problems but requires much more advanced mathematical tools as those presented in this paper. In addition, the Malliavin calculus has the advantage and the specificity that it is possible to directly work in the continuum (infinite dimensional space) to evaluate the sensitivity derivatives. We hope that the first and simple approach restricted to random variables presented in this paper may be of interest to the engineering community and encourage them to investigate the benefits that the Malliavin calculus could provide in the context of stochastic PDEs.

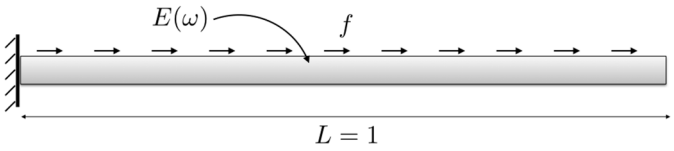

Fig 5. Elastic bar with stochastic Young's modulus.

PDE examples

We now turn our attention to models that are defined as PDEs. To solve the deterministic evaluations of the PDEs we use the finite element method. We have chosen to use DOLFIN, part of the FEniCS Project to implement the finite element method solvers [22].

Elastic bar with stochastic Young's modulus

The strong form PDE and boundary conditions of the behaviour of a 1-dimensional elastic bar (see Fig 5) are:

$$E\ \frac{\mathrm{d}^2 u(x)}{\mathrm{d}x^2} + f = 0;\ u(0) = 0 \text{ and } \frac{\mathrm{d}u(L)}{\mathrm{d}x} = 0. \tag{42}$$

We take $f = 1$, $L = 1$ and a stochastic Young's modulus:

$$E = 2(1 + \xi), \tag{43}$$

with ξ a random variable with beta(2, 2) distribution.

The forward model is described by the following weak residual formulation, find $u \in H_D^1(\Omega_s)$ such that:

$$F(u; \tilde{u}) \coloneqq -\int_{\Omega_s} E\ \nabla u \cdot \nabla \tilde{u}\ \mathrm{d}x + \int_{\Omega_s} f\tilde{u}\ \mathrm{d}x = 0 \quad \forall \tilde{u} \in H_0^1(\Omega_s), \tag{44}$$

where the space $H_D^1(\Omega_s)$ is the usual Sobolev space of square-integrable functions with square-integrable weak derivatives on the domain $\Omega_s \coloneqq [0, 1]$ that satisfy the Dirichlet boundary condition $u(0) = 0$ and $H_0^1(\Omega_s)$ vanish on the whole boundary. We solve the forward model using a piecewise linear finite element method with 1024 cells in the mesh.

The quantity of interest is:

$$J = \int_0^1 u(x)\mathrm{d}x. \tag{45}$$

The derivative of the expected value of J with respect to the mean value of the Young's modulus $\bar{E}$ can be computed analytically in this case:

$$\frac{\partial \mathbb{E}[J]}{\partial \bar{E}} = -\int_0^1 \frac{1}{3(2+2x)^2} \frac{x(1-x)}{\mathrm{B}(2,2)}\ \mathrm{d}x = 1 - \tfrac{3}{2}\ln(2). \tag{46}$$

This problem is a stationary (not time-dependent), in contrast to the Kelvin-Voigt model considered previously. However, this stationary problem can be solved using the same techniques. We introduce the concept of pseudo-time, where the system evolves from its initial state at $t = 0$ to the final solution at time $t = T$ through the a single solution of the PDE Eq (44).

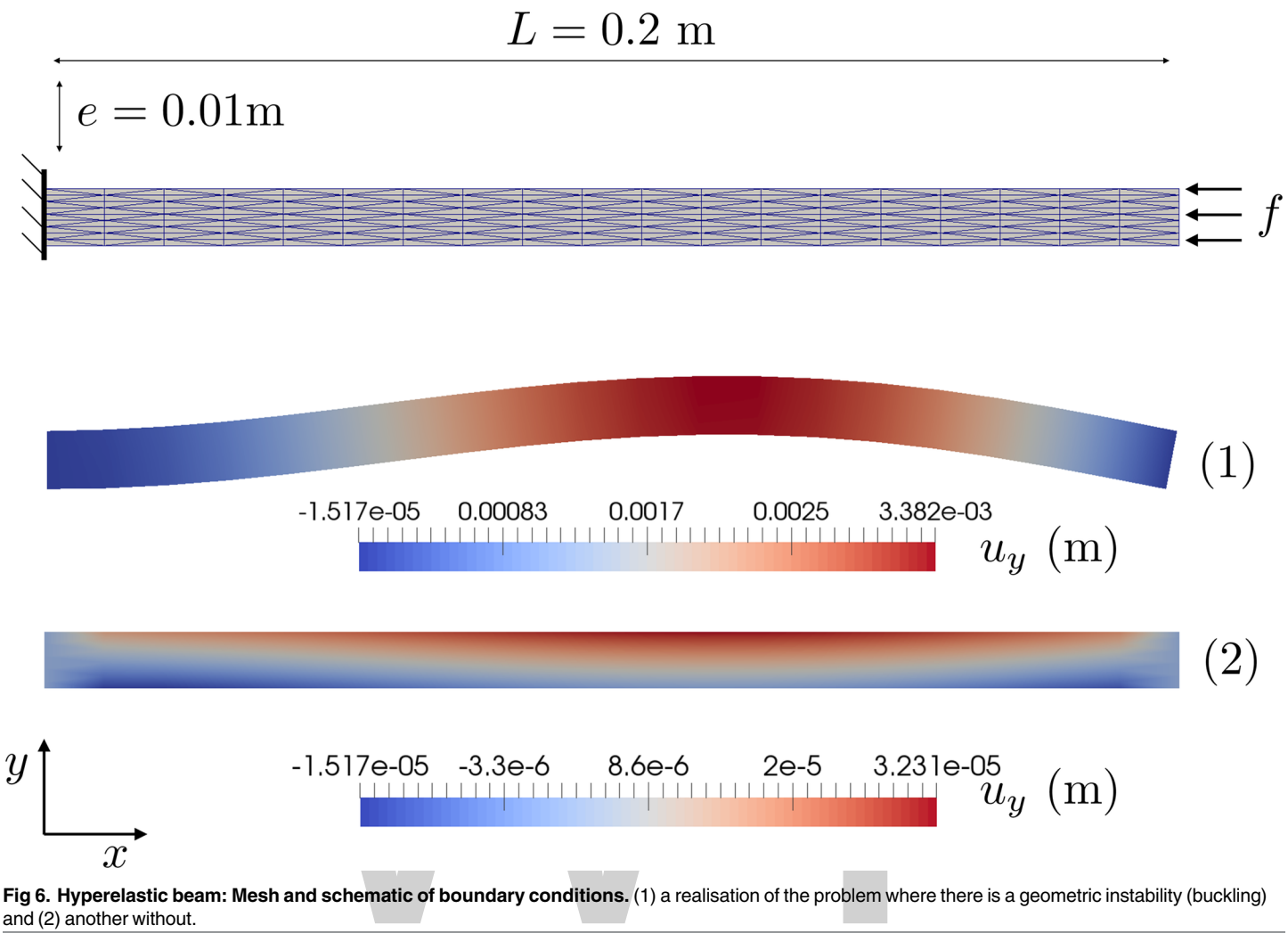

Fig 6. Hyperelastic beam: Mesh and schematic of boundary conditions. (1) a realisation of the problem where there is a geometric instability (buckling) and (2) another without.

Therefore in algorithm 1 we take the pseudo-time step as $\delta t = T$ and hence $n = 1$. As before, the Malliavin weight still has initial condition $q_m(0) = 0$.

Finally, the relative error between the MWS method with $Z = 5 \times 10^5$ realisations and the analytical solution is 3.0×10^{-3}.

Buckling of a hyperelastic beam with stochastic Young's modulus

We study the deformation of a 2D geometrically non-linear hyperelastic beam with stochastic Young's modulus E. We have deliberately designed this problem so that for some values of E the beam undergoes buckling, and for others not.

Consider a hyperelastic body that in its undeformed state occupies the domain $\Omega_0 = [0, L] \times [0, e] \subset \mathbb{R}^2$ with $L = 0.2m$ and $e = 0.01m$ (see Fig 6), and in its deformed state occupies some (unknown) domain $\Omega \subset \mathbb{R}^2$. $\boldsymbol{\varphi}$ is the map between the material points $\mathbf{X}$ in the undeformed domain Ω_0 and points $\mathbf{x}$ in the deformed domain Ω:

$$\boldsymbol{\varphi} : \Omega_0 \ni \mathbf{X} \to \mathbf{x} \in \Omega, \tag{47}$$

The deformation gradient can be written $\mathsf{F}(\mathbf{X}) \coloneqq \frac{\partial \boldsymbol{\varphi}}{\partial \mathbf{X}}$. The right Cauchy-Green tensor is then defined as $\mathsf{C} \coloneqq \mathsf{F}^T\mathsf{F}$.

The Neo-Hookean stored energy density of the body is then:

$$\mathcal{W}(\mathsf{F}) \coloneqq \mu(I_c - 2)/2 - \mu \log I_3 + \lambda(\log I_3)^2/2. \tag{48}$$

where $I_3 \coloneqq \det(\mathsf{F})$ and and $I_C = \operatorname{tr}(\mathsf{C})$. λ and μ are the Lame parameters and can be expressed as a function of the Young's modulus E and Poisson's ratio v as:

$$\lambda = \frac{Ev}{(1+v)(1-2v)} \text{ and } \mu = \frac{E}{2(1+v)}. \tag{49}$$

We choose to model the Young's modulus as a log-normal random variable with mean value 11 *MPa* and standard deviation 2 *MPa*. We take Poisson's ratio as a fixed constant $v = 0.3$.

Defining the displacement field as $\mathbf{u} \coloneqq \boldsymbol{\varphi} - \mathbf{X}$ and a linear functional $\mathbf{f}$ that encodes the external tractions we can characterise the elastic equilibrium displacement field $\mathbf{u}^*$ as the solution to the following minimisation problem:

$$\begin{aligned} \mathbf{u}^* &= \underset{\mathbf{u}\in[H^1_D(\Omega_0)]^2}{\arg\min}\ L(\mathbf{u}) \\ &= \underset{\mathbf{u}\in[H^1_D(\Omega_0)]^2}{\arg\min} \left\{ \int_{\Omega_0} \mathcal{W}(\mathsf{F})\ \mathrm{d}x_0 - \langle \mathbf{f}, \mathbf{u} \rangle \right\}, \end{aligned} \tag{50}$$

where $[H^1_D(\Omega_0)]^2$ is the usual vector-valued Sobolev space of square integrable functions with square integrable derivatives that satisfies the given Dirichlet boundary conditions and dx_0 is a measure on Ω_0. We fix the left hand of the beam, $u(0, y) = 0$ and apply a surface traction in the negative x direction on the right hand of the beam of magnitude f.

For one Monte Carlo realisation we solve the non-linear problem using a Newton method from SNES [23] with continuation in the loading parameter f and a third-order backtracking line search. We let the symbolic differentiation capabilities of UFL derive the residual and Jacobian of the forward model for use in the Newton method. We solve the linear systems arising from the Newton iterations using a conjugate gradient method preconditioned using algebraic multigrid (Hypre BoomerAMG [24]) interfaced from PETSc [23].

The quantity of interest is defined as:

$$J = \int_\Omega |u_y|\ \mathrm{d}x. \tag{51}$$

The Malliavin derivative of $\mathbb{E}[J]$ with respect to the mean Young's modulus obtained with the MWS method with $Z = 3 \times 10^3$ realisations is:

$$\frac{\partial \mathbb{E}[J]}{\partial E_0} \approx -3.1 \times 10^{-6}\ \mathrm{m}^3/\mathrm{MPa}. \tag{52}$$

No analytical solution exists for comparison. If we use dolfin-adjoint [25], we can compute the classical derivative of J with respect to the Young's modulus around the mean parameter:

$$\left.\frac{\partial J}{\partial E}\right|_{E=E_0} \approx -3.5 \times 10^{-8}\ \mathrm{m}^3/\mathrm{MPa}. \tag{53}$$

In this example the difference between the classical derivative and the Malliavin derivative is quite pronounced. This difference is caused by the presence of an instability (buckling). This instability is not activated when $E = E_0$, hence, the classical derivative tells us that J is relatively

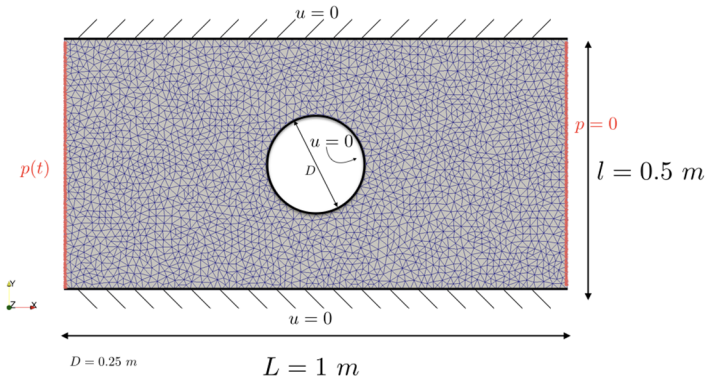

Fig 7. Mesh, geometry and boundary conditions for the incompressible Navier-Stokes problem.

insensitive to perturbations in the Young's modulus about E_0. However, the Malliavin derivative tells us that $\mathbb{E}[J]$ *is* in fact quite sensitive to changes in the mean of the Young's modulus E_0. The Malliavin derivative gives us quite a different perspective on the sensitivity of this problem than the classical one.

Incompressible Navier-Stokes equations with stochastic viscosity

We consider the incompressible Navier-Stokes equations on a domain Ω in $\mathbb{R}^2$ consisting of a pair of momentum and continuity equations:

$$\begin{aligned} \dot{\mathbf{u}} + \nabla \mathbf{u} \cdot \mathbf{u} - \nu \Delta \mathbf{u} + \nabla p &= \mathbf{f}, \\ \nabla \cdot \mathbf{u} &= 0. \end{aligned} \tag{54}$$

In Eq (54), $\mathbf{u}$ refers to the unknown velocity of the fluid, ν is the viscosity of the fluid, p the unknown pressure and $\mathbf{f}$ is a given source. The mesh, geometry and boundary conditions for the incompressible Navier-Stokes problem are shown in Fig 7. The viscosity is modelled as a random variable:

$$\nu = 0.015 + 0.01(\xi - 0.005), \tag{55}$$

with ξ a log-normal distribution with mean equal to 0.5 and standard deviation equal to 0.25.

We solve the PDE for a given parameter ν with FEniCS [26] (FE approximation) using Chorin's method with time step $\delta t = 0.01$ for $t \in [0, 1]$, see [27] for more details on the implementation. For one realisation of the viscosity the velocity at time $t = 1$ s is show in Fig 8.

The quantity of interest is the total volume of fluid that exits the right end of the domain:

$$J = \int_{t=0}^{t=1} \int_{S_{p=0}} \mathbf{u} \cdot \mathbf{n} \; \mathrm{d}s\mathrm{d}t, \tag{56}$$

where $S_{p=0}$ is the surface with normal vector $\mathbf{n}$ on the right side where the pressure is imposed to zero.

The derivative of $\mathbb{E}[J]$ with respect to η_0 obtained with the MWS method for $Z = 4 \times 10^5$ realisations is:

$$\frac{\partial \mathbb{E}[J]}{\partial \eta_0} \approx -7.9 \text{ s/m}^2. \tag{57}$$

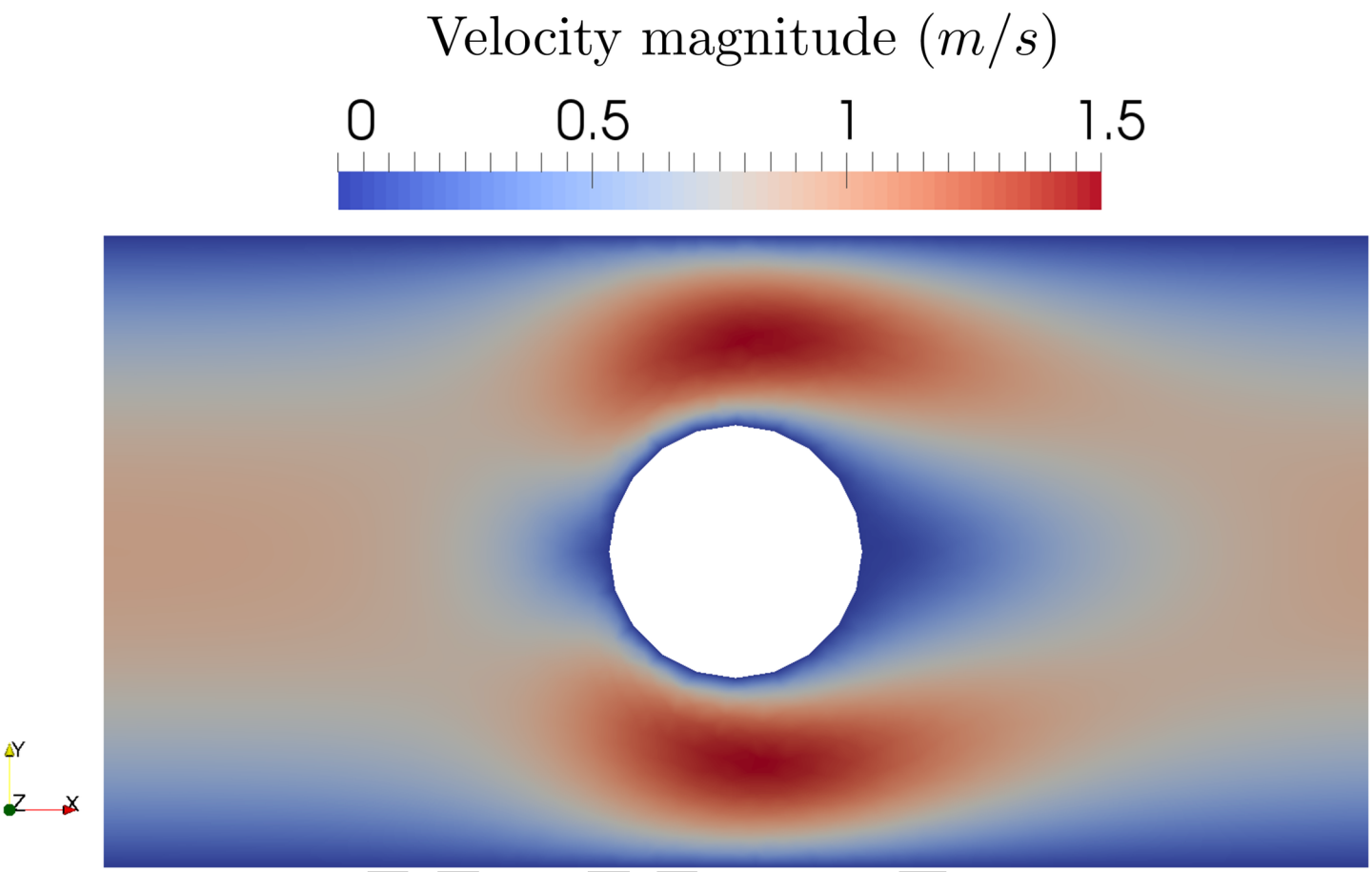

Fig 8. Velocity magnitude at time $t = 1$ s for one realisation of the viscosity.

No analytical solution exists for comparison. If we use dolfin-adjoint [25], we can compute the derivative of J with respect to the viscosity around the mean parameter:

$$\left.\frac{\partial J}{\partial \eta}\right|_{\eta=\eta_0} \approx -7.85 \text{ s/m}^2. \tag{58}$$

The two sensitivity derivatives are close. In this example, contrary to the hyperelastic example, the Malliavin approach does not give us a particularly different interpretation of the sensitivity.

Conclusion

In this paper we have shown how to calculate the Malliavin derivative using the method of Malliavin Weight Sampling. We have applied the method to some typical mechanics models that can be described by ODEs and PDEs, and solved those models using finite difference and finite element methods. In addition, we have extended the existing practical literature on MWS to non-Gaussian random variables and the calculation of second-order derivatives. We are currently investigating the extension of this work from random parameters to problems with variables modelled as random fields. We are also exploring the use of the Malliavin

derivative in derivative-driven variance reduction methods e.g. [28]. Code showing the calculation of the Malliavin derivative for the examples in this paper.

Acknowledgments

We would like to thank Prof. Ivan Nourdin, University of Luxembourg, for his helpful discussions. We thank the financial support of the European Research Council Starting Independent Research Grant (ERC Stg grant agreement No. 279578) entitled 'Towards real time multiscale simulation of cutting in non-linear materials with applications to surgical simulation and computer guided surgery'. Paul Hauseux is supported by the internal MOMENTUM project at the University of Luxembourg. Jack S. Hale is supported by the National Research Fund, Luxembourg, and cofunded under the Marie Curie Actions of the European Commission (FP7-COFUND) Grant No. 6693582. We also thank the funding from the Luxembourg National Research Fund (INTER/MOBILITY/14/8813215/CBM/Bordas).

Author Contributions

Conceptualization: Paul Hauseux.

Formal analysis: Paul Hauseux, Jack S. Hale.

Funding acquisition: Stéphane P. A. Bordas.

Methodology: Paul Hauseux, Jack S. Hale.

Project administration: Stéphane P. A. Bordas.

Software: Paul Hauseux, Jack S. Hale.

Supervision: Jack S. Hale, Stéphane P. A. Bordas.

Validation: Paul Hauseux, Jack S. Hale.

Visualization: Paul Hauseux, Jack S. Hale.

Writing – original draft: Paul Hauseux.

Writing – review & editing: Jack S. Hale, Stéphane P. A. Bordas.

References

1. Ambrosetti A, Prodi G. A Primer of Nonlinear Analysis. Cambridge University Press; 1995.
2. Yang XJ, Machado JAT. A new fractional operator of variable order: Application in the description of anomalous diffusion. Physica A: Statistical Mechanics and its Applications. 2017; 481(Supplement C):276–283. https://doi.org/10.1016/j.physa.2017.04.054
3. Atangana A, Koca I. In: Ruzhansky M, Cho YJ, Agarwal P, Area I, editors. On Uncertain-Fractional Modeling: The Future Way of Modeling Real-World Problems. Singapore: Springer Singapore; 2017. p. 121–143.
4. Atangana A, GÃmez-Aguilar JF. A new derivative with normal distribution kernel: Theory, methods and applications. Physica A: Statistical Mechanics and its Applications. 2017; 476(Supplement C):1–14. https://doi.org/10.1016/j.physa.2017.02.016
5. Atangana A, Baleanu D. New Fractional Derivatives with Nonlocal and Non-Singular Kernel: Theory and Application to Heat Transfer Model. Thermal Science. 2016; 20:763–769. https://doi.org/10.2298/TSCI160111018A
6. Yang XJ, Gao F, Machado JAT, Baleanu D. A new fractional derivative involving the normalized sinc function without singular kernel;. Available from: https://arxiv.org/abs/1701.05590
7. He JH. Variational principles for some nonlinear partial differential equations with variable coefficients. Chaos, Solitons & Fractals. 2004; 19(4):847–851. https://doi.org/10.1016/S0960-0779(03)00265-0

8. Stochastic Calculus of Variations in Mathematical Finance | Paul Malliavin | Springer;. Available from: http://www.springer.com/gp/book/9783540434313
9. Malliavin P. Stochastic analysis. Grundlehren der mathematischen Wissenschaften. Berlin, New York: Springer; 1997.
10. Nourdin I, Peccati G. Normal Approximations with Malliavin Calculus: From Stein's Method to Universality. Cambridge Tracts in Mathematics. Cambridge University Press; 2012.
11. Broadie M, Glasserman P. Estimating security price derivatives using simulation. Management science. 1996; 42(2):269–285. https://doi.org/10.1287/mnsc.42.2.269
12. Benhamou E. Optimal Malliavin Weighting Function for the Computation of the Greeks. Mathematical Finance. 2003; 13(1):37–53. https://doi.org/10.1111/1467-9965.t01-1-00004
13. Fournié E, Lasry JM, Lebuchoux J, Lions PL, Touzi N. Applications of Malliavin calculus to Monte Carlo methods in finance. Finance and Stochastics. 1999; 3(4):391–412. https://doi.org/10.1007/s007800050068
14. Chen N, Glasserman P. Malliavin Greeks without Malliavin calculus. Stochastic Processes and their Applications. 2007; 117(11):1689–1723. https://doi.org/10.1016/j.spa.2007.03.012
15. Warren PB, Allen RJ. Malliavin Weight Sampling for Computing Sensitivity Coefficients in Brownian Dynamics Simulations. Physical Review Letters. 2012; 109(25):250601. https://doi.org/10.1103/PhysRevLett.109.250601 PMID: 23368441
16. Warren PB, Allen RJ. Malliavin Weight Sampling: A Practical Guide. Entropy. 2014; 16(1):221–232. https://doi.org/10.3390/e16010221
17. Szamel G. Evaluating linear response in active systems with no perturbing field. EPL (Europhysics Letters). 2017; 117(5):50010. https://doi.org/10.1209/0295-5075/117/50010
18. Nualart D. The Malliavin calculus and related topics. vol. 1995. Springer; 2006. Available from: http://dx.doi.org/10.1007/3-540-28329-3
19. Matthies GH. Stochastic finite elements: Computational approaches to stochastic partial differential equations. Journal of Applied Mathematics and Mechanics. 2008; 88:849–873.
20. L'Ecuyer P. A Unified View of the IPA, SF, and LR Gradient Estimation Techniques. Manage Sci. 1990; 36(11):1364–1383. https://doi.org/10.1287/mnsc.36.11.1364
21. Capriotti L. Reducing the variance of likelihood ratio greeks in Monte Carlo. In: 2008 Winter Simulation Conference; 2008. p. 587–593.
22. Alnaes M, Blechta J, Hake J, A J, Kehlet B, Logg A, et al. The FEniCS Project Version 1.5. Archive of Numerical Software. 2015; 3(100).
23. Balay S, Abhyankar S, Adams MF, Brown J, Brune P, Buschelman K, et al. PETSc Users Manual. Argonne National Laboratory; 2016. ANL-95/11—Revision 3.7. Available from: http://www.mcs.anl.gov/petsc
24. Falgout RD, Yang UM. hypre: A Library of High Performance Preconditioners. In: Sloot PMA, Hoekstra AG, Tan CJK, Dongarra JJ, editors. Computational Science—ICCS 2002. No. 2331 in Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2002. p. 632–641.
25. Farrell PE, Ham DA, Funke SW, Rognes ME. Automated Derivation of the Adjoint of High-Level Transient Finite Element Programs. SIAM Journal on Scientific Computing. 2013; 35(4):C369–C393. https://doi.org/10.1137/120873558
26. Logg A, Wells GN. DOLFIN: Automated Finite Element Computing. ACM Trans Math Softw. 2010; 37(2):20:1–20:28. https://doi.org/10.1145/1731022.1731030
27. Langtangen HP, Logg A. Solving PDEs in Python—The FEniCS Tutorial I. No. 3 in Simula SpringerBriefs on Computing. Springer International Publishing; 2016. Available from: http://dx.doi.org/10.1007/978-3-319-52462-7
28. Hauseux P, Hale JS, Bordas SPA. Accelerating Monte Carlo estimation with derivatives of high-level finite element models. Computer Methods in Applied Mechanics and Engineering. 2017; 318(Supplement C):917–936. https://doi.org/10.1016/j.cma.2017.01.041
29. Hauseux P, Hale JS, Bordas SPA. Calculating the Malliavin Derivative of some numerical models using the Malliavin Weight Sampling method, 2017. https://dx.doi.org/10.6084/m9.figshare.5432722