ГОДИШНИК

НА

СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

Том 106
2019

ANNUAL

OF

SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”

FACULTY OF MATHEMATICS AND INFORMATICS

Volume 106
2019

СОФИЯ • 2019 • SOFIA
УНИВЕРСИТЕТСКО ИЗДАТЕЛСТВО „СВ. КЛИМЕНТ ОХРИДСКИ“
“ST. KLIMENT OHRIDSKI” UNIVERSITY PRESS

Annual of Sofia University "St. Kliment Ohridski"
Faculty of Mathematics and Informatics

Годишник на Софийския университет „Св. Климент Охридски"

Факултет по математика и информатика

Address for correspondence:

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5, J. Bourchier Blvd., P.O. Box 48
BG-1164 Sofia, Bulgaria

Fax xx(359 2) 8687 180
Electronic mail:
annuaire@fmi.uni-sofia.bg

Aims and Scope. The *Annual* is the oldest Bulgarian journal, founded in 1904, devoted to pure and applied mathematics, mechanics and computer science. It is reviewed by *Zentralblatt für Mathematik*, *Mathematical Reviews* and the Russian *Referativnii Jurnal*. The *Annual* publishes significant and original research papers of authors both from Bulgaria and abroad in some selected areas that comply with the traditional scientific interests of the Faculty of Mathematics and Informatics at the "St. Kliment Ohridski" University of Sofia, i.e., algebra, geometry and topology, analysis, probability and statistics, mathematical logic, theory of approximations, numerical methods, computer science, classical, fluid and solid mechanics, and their fundamental applications.

ISSN 1313–9215 (Print)
ISSN 2603–5529 (Online)

CONTENTS

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 106

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 106

NEURAL NETWORKS FOR FACILITY LOCATION PROBLEMS

VLADISLAV HARALAMPIEV

This paper presents a new self-organizing neural network approach for solving graph-based facility location problems. It is designed to have small amount of parameters and to not need much tuning. We test our approach on several groups of problems and show that it consistently finds good feasible solutions.

Keywords: Neural networks, facility location, combinatorial optimization.

2010 Math. Subject Classification: 62M45, 90C27.

1. INTRODUCTION

Facility location problems are a large class of optimization problems, modelling the search for optimal placement of facilities to minimize costs. Many of these problems are known to be NP-hard to solve exactly. In this paper, we investigate the possibility to use neural networks for solving two graph variants of facility location problems.

There are two main neural approaches for solving difficult combinatorial optimization problems — Hopfield networks [7] and variations of Kohonen's Self-Organizing Feature Map [8]. Unfortunately, both of these approaches have problems. Hopfield's method has a tendency to settle in poor local minima, often not representing a feasible solution, and it is difficult to select appropriate parameters leading to a good solution. The problems, associated with Hopfield's approach, are well documented [11]. Vast majority of Self-Organizing Feature Maps, on the other hand, are based on the Elastic Net method [4]. This method relies on the fact that the 'elastic band' moves in Euclidean space, and distances between vertices

are measured in the same space. This greatly limits the set of problems suitable for the method. In fact, most of the applications of Self-Organizing Feature Maps are to the Travelling Salesman Problem [5].

We propose a new neural network architecture for graph variants of facility location problems, inspired by the self-organizing approach to optimization. The network is designed to always find a feasible solution and to have small amount of parameters. It is often believed that involved mathematical instruments are more powerful than any heuristics, based on physical or biological analogies (see the preface in [1]). Our goal is not to outperform methods, designed for solving specific facility location problems, but to develop a robust neural network approach for facility location that needs minimal work with the specifics of the problem. This is important in practice, when we need to find acceptable solution with small investment. The proposed neural network is tested on several groups of problems and achieves good results, most of the time exactly solving the input instances.

2. GRAPH-BASED FACILITY LOCATION PROBLEMS

Let $G(V, E)$ be a connected, undirected and weighted graph with vertex set V and edge set E. We will denote the distance between two vertices $u, w \in V$ as $dist(u, w)$. The two facility location problems we are interested in are called $MiniSum$ and $MiniMax$. Good survey of the types of facility location problems is [3]. Intuitively, $MiniSum$ models the placement of several warehouse facilities, where the goal is to minimize the average travel distance. $MiniMax$ models the placement of fire stations, in which case we want to minimize the maximum time of travel to every vertex.

Definition 1 ($p - MiniSum$ problem)**.** Find a subset $u_1, u_2, \ldots, u_p$ of p vertices from V that minimizes $\sum_{w \in V} \min_{i \in \{1,\ldots,p\}} dist(w, u_i)$.

Definition 2 ($p - MiniMax$ problem)**.** Find a subset $u_1, u_2, \ldots, u_p$ of p vertices from V that minimizes $\max_{w \in V} \min_{i \in \{1,\ldots,p\}} dist(w, u_i)$.

The neural network we develop will only solve the $p - MiniSum$ problem. The following reduction is used for solving $p - MiniMax$.

Theorem 1 ($p - MiniMax$ to $p - MiniSum$ reduction)**.** *Solving a $p - MiniMax$ problem with any required positive precision ε can be reduced to solving a sequence of $p - MiniSum$ problems.*

Proof. Assume we need to solve a $p - MiniMax$ problem with input graph $G(V, E)$ and the optimal solution has value opt. For a given value c we can check if $opt \leq c$ by solving a $p - MiniSum$ problem in a modified version G' of G. G' is a fully connected graph with the same vertex set V. For every pair of vertices $u \neq w$ the weight in G' of the edge between u and w is one if $dist(u, w) \leq c$ in G,

otherwise it is ten. Now, $opt \leq c \iff$ the optimal solution to the $p - MiniSum$ problem in G' has value $n - p$. This is because if $opt \leq c$, there is a solution in G' that only uses edges of weight one (and vice versa).

To solve the original $p - MiniMax$ problem, we can do a binary search on the value c. This way, we reduced the problem to $O(lg(MAX) + |lg(\varepsilon)|)$ instances of $p - MiniSum$, where MAX is the maximum distance between two vertices in G.□

Note that the reduction assumes all instances of $p - MiniSum$ are solved correctly. Our neural approach provides only approximate solutions to $p - MiniSum$ problems, so the solution we get for $p - MiniMax$ is also approximate. Intuitively, errors early in the sequence of $p - MiniSum$ problems directly lead to a poor quality solution for the $p - MiniMax$ problem. But, early in the sequence, the value opt from the reduction is far from c, which makes much simpler the corresponding $p - MiniSum$ problem. Another difficulty arises from the way we define distances in G'. The distances between vertices in this graph do not change smoothly, making it harder for local search methods to find good solutions.

3. NETWORK ARCHITECTURE

The architecture of the proposed network is shown in Figure 1. There are three layers, which we call A, B and C.

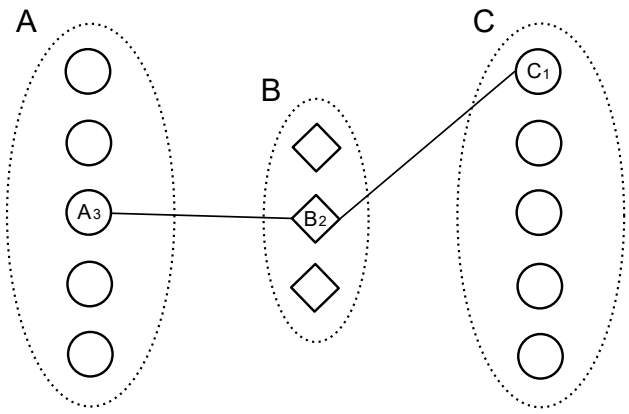


Figure 1. Neural network architecture for facility location problems

Layer A contains $|V|$ nodes, corresponding to the clients (vertices of the graph G). Layer B contains p nodes, corresponding to the facilities we need to locate. Layer C contains $|V|$ nodes, corresponding to the locations of the facilities (vertices of G). Layers $A - B$ and $B - C$ are fully connected. When we talk about outgoing edges, we assume edges are directed from A to B and from B to C. The weight of the edge between A_3 and B_2 (a number between zero and one) shows to what degree

client A_3 uses facility B_2. The weight of the $B_2 - C_1$ edge shows to what degree facility B_2 is located in C_1. The interpretation of the other edges is analogous. For each node in layers A and B, we require that the sum of the weights of the outgoing edges is one. These weights are initialized randomly. As the algorithm progresses, for each node one of the outgoing edges starts to dominate and its weight approaches one. To produce the final solution from the network, we assign clients to facilities and facilities to locations by following the dominating edges. As a side note, sometimes, because of symmetries in the graph, several edges start to dominate for a node (their weights become comparable and much larger than the weights of the other edges). We observed that in such situations choosing each one of these edges produces equally good solution. In our experiments, when multiple edges dominate for a node, we always pick the edge with the smallest index.

4. OPTIMIZATION

The neural network minimizes the function

$$\sum_{a_i \in A} \sum_{b_j \in B} \sum_{c_k \in C} weight(a_i, b_j) \cdot weight(b_j, c_k) \cdot dist(a_i, c_k) \tag{1}$$

Here $weight(a_i, b_j)$ represents to what degree client a_i uses facility b_j, $weight(b_j, c_k)$ represents to what degree facility b_j is located in c_k, and $dist(a_i, c_k)$ is the distance in G between the vertices, corresponding to a_i and c_k. As for each node one of the outgoing edges starts to dominate and its weight approaches one, this function becomes equivalent to the $MiniSum$ function.

The optimization starts from a randomly initialized state and consists of a series of iterations, until the weights converge. In each iteration we go through all the nodes in layers A and B in random order and update the weights of their outgoing edges. Assume we are processing node $a_1 \in A$. The update consists of three steps:

- *Evaluate.* For each facility $b_j \in B$ calculate the cost of assigning a_1 to it, $cost_j = \sum_{c_k \in C} weight(b_j, c_k) \cdot dist(a_1, c_k)$. After this compute the value $prefer_j = \frac{min_{b_s \in B} cost_s}{cost_j}$ which is between zero and one. Intuitively, values closer to one are more preferable for the client.

- *Strengthen.* Differences between the $prefer$ values are often small. We increase them by settings $prefer'_j = \frac{e^{mult \cdot prefer_j}}{max_{b_s \in B} e^{mult \cdot prefers}}$. Here $mult$ is a parameter.

- *Update.* First transform the weights of the outgoing edges to have the same meaning as the $prefer'$ values. This is achieved by setting $weight(a_1, b_j)$ to $\frac{weight(a_1, b_j)}{max_{b_s \in B} weight(a_1, b_s)}$. Then update each $weight(a_1, b_j)$ to be equal to

$(1-\alpha)\cdot weight(a_1,b_j)+\alpha\cdot prefer'_j$. Finally, normalize the weights so that they sum to one (by dividing each weight by the sum of all weights). α is a parameter, analogous to learning rate in the learning algorithms of classical neural networks.

The updates are done similarly for all other nodes in layers A and B. There are two parameters, the learning rate α and the *mult* parameter that scales the $prefer$ values. Exponential grid search is used to select the parameters. More specifically, the *mult* dimension of the grid consists of the values 1.2^x for $x \in 1,2,\ldots,50$. The α dimension consists of $0.2\cdot 0.8^y$ for $y \in 0,1,2,3,4$. For each cell of the grid we run the optimization with the corresponding parameters. We then pick the best solution found. To guarantee convergence in the allocated time, during each optimization run the learning rate decreases exponentially with the number of iterations. From our experience, the initial value of the learning rate affects the speed of convergence, but does not affect significantly the quality of the final solution (assuming the optimization runs long enough). On the other hand, *mult* affects the quality of the solution.

5. TEST PROBLEMS AND RESULTS

The proposed network is tested on four groups of problems:

- *Unweighted trees (TU).* Random trees with 50 to 100 nodes and p (number of facilities) between two and six. All edges are of length one. The random trees are generated using Prüfer's code, a mapping of trees to number sequences [10].
- *Weighted trees (TW).* Trees with the same parameters as the unweighted trees above, but with random floating point edge lengths between 1 and 100.
- *Chordal graphs (CH).* Chordal graphs are graphs without induced s-cycles for s more than three [2]. They have more complex structure than trees, but still are simple enough to allow efficient algorithms for many problems that are hard in general graphs. We generate chordal graphs with 50 to 100 vertices and set p (number of facilities) to a value between three and six. To generate them we use two methods — producing a perfect elimination order and using the equivalence to intersections of subtrees of a tree [6].
- *Bulgarian road network (BR).* For various geographic regions in Bulgaria we take the populated places and the roads connecting them. We choose regions with 60 to 400 populated places and set p (number of facilities) to a value between two and four. Data about populated places and roads is taken from OpenStreetMaps [9] (using Overpass queries).

For each of the first three groups, we randomly generate 200 graphs. For the last group, we generate 30 graphs.

For each of the instances above we compute the optimal answer by trying all possibilities of locating the facilities. Constraints in the instances are chosen small enough so that this computation runs in reasonable time. We also run a local search for each instance to compare its results with the results of the proposed neural network. Local search [1] is a classical general method for solving optimization problems that often gives very good results. It starts from a random solution and repeatedly tries to improve it by choosing a better solution from some neighbourhood of the current one. In our case, the neighbourhood is defined by changing the location of a facility, or by changing the facility that services a client. Since the results of local search depend on the initial solution, for each instance we run 1000 independent local searches and pick the best value they return.

The results for the $MiniSum$ problem are presented in Table 1. Both methods have excellent performance, most of the time finding an optimal solution. Local search has slightly better performance, probably because it runs 1000 independent searches.

Table 1. Results of local search and the proposed neural network for the $MiniSum$ problem. *Max* is the maximum error over all inputs (as percentage from the optimal answer), *Avg* is the average error, and *Exact* is the percentage of inputs, solved exactly.

	Local search			Neural network		
	Max, %	*Avg*, %	*Exact*, %	*Max*, %	*Avg*, %	*Exact*, %
TU	0.312	0.009	97	0.295	0.007	98
TW	1.130	0.030	92	2.000	0.090	90
CH	3.191	0.072	98	4.000	0.500	80
BR	0.021	0.012	96	0.000	0.000	100

The results for the $MiniMax$ problem are presented in Table 2. Our approach to $MiniMax$ requires solving a sequence of harder $MiniSum$ instances, so, as expected, the results are worse than the ones for $MiniSum$. The neural network approach performs significantly better than local search. It is often able to exactly solve the instance. As a side note, the excellent results on chordal graphs (CH) are probably because they, intuitively, are a sequence of attached cliques, which makes $MiniMax$ simpler to solve.

Table 2. Results of local search and the proposed neural network for the $MiniMax$ problem. *Max* is the maximum error over all inputs (as percentage from the optimal answer), *Avg* is the average error, and *Exact* is the percentage of inputs, solved exactly.

	Local search			Neural network		
	Max, %	*Avg*, %	*Exact*, %	*Max*, %	*Avg*, %	*Exact*, %
TU	32.900	9.770	60	5.080	0.827	87
TW	31.640	7.180	47	3.042	0.253	87
CH	0.000	0.000	100	0.000	0.000	100
BR	36.150	16.210	0	10.000	3.090	67

6. CONCLUSION

We presented a new neural network architecture for solving graph-based facility location problems and evaluated its performance on several groups of $MiniSum$ and $MiniMax$ problems. Our method is based on the self-organizing approach to optimization and shows good performance. On simpler instances its results are comparable to local search, and it significantly outperforms local search on harder instances.

ACKNOWLEDGEMENT. Map data used for generating the facility location instances based on the Bulgarian road network is copyrighted by OpenStreetMap contributors and can be found at https://www.openstreetmap.org

7. REFERENCES

[1] Aarts, E., Lenstra, J. K.: *Local Search in Combinatorial Optimization*. Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, 1997.

[2] Blair, J. R. S., Peyton, B.: *An Introduction to Chordal Graphs and Clique Trees. Graph Theory and Sparse Matrix Computation*. IMA Volumes in Mathematics and its Applications, vol. 56, 1993.

[3] Cappanera, P.: A Survey on Obnoxious Facility Location Problems. University of Pisa, technical report, 1999.

[4] Durbin, R., Willshaw, D.: An analogue approach to the travelling salesman problem using an elastic net method. *Nature*, **326**, 1987, 689–691.

[5] Favata, F., Walker, R.: A study of the application of Kohonen-type neural networks to the Travelling Salesman Problem. *Biological Cybernetics*, **64**, 1991, 463–468.

[6] Gavril, F.: The intersection graphs of subtrees in trees are exactly the chordal graphs. *J. Combin. Theory Ser. B*, **16**, 1974, 47–56.

[7] Hopfield, J. J., Tank, D. W.: "Neural" computation of decisions in optimization problems. *Biological Cybernetics*, **52**, 1985, 141–152.

[8] Kohonen, T.: Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, **43**, 1982, 59–69.

[9] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org. https://www.openstreetmap.org (2017)

[10] Prüfer, H.: Neuer Beweis eines Satzes über Permutationen. *Archiv für Mathematik und Physik*, **27**, 1918, 142–144.

[11] Wilson, G. V., Pawley, G. S.: On the stability of the Travelling Salesman Problem algorithm of Hopfield and Tank. *Biological Cybernetics*, **58**, 1988, 63–70.

Received on April 18, 2019
Received in a revised form on February 29, 2020

Vladislav Haralampiev

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 blvd. J. Bourchier, BG-1164 Sofia
BULGARIA

E-mail: vladislav.haralampiev@gmail.com

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 106

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 106

ON THE STRUCTURE OF SOME ARCS RELATED TO CAPS AND THE NONEXISTENCE OF SOME OPTIMAL CODES

ASSIA P. ROUSSEVA

In this paper we solve two instances of the main problem in coding theory for linear codes of dimension 5 over $\mathbb{F}_4$. We prove the nonexistence of $[395, 5, 295]_4$- and $[396, 5, 296]_4$-codes which implies the exact values $n_4(5, 295) = 396$ and $n_4(5, 296) = 397$. As a by-product, we characterize the arcs with parameters $(100, 26)$ in $\mathrm{PG}(3, 4)$.

Keywords: Linear codes, finite projecive geometries, arcs, extendable arcs, the Griesmer bound, Griesmer codes, Griesmer arcs.

2010 Math. Subject Classification: Primary: 51A20, 51A21, 94B65; Secondary: 51A22.

1. INTRODUCTION

In this paper we study two instances of the main problem in coding theory: the problem of determining the exact value of $n_q(k, d)$ defined as the minimal length of a k-dimensional linear code of minimum distance d over the field with q elements. This problem has been studied intensively in the past 30 years and has been solved completely for some small fields $\mathbb{F}_q$, and small dimensions k. The problem has a clear geometric relevance since every linear code of full length is known to be equivalent to an arc in the appropriate finite projective space and optimal codes correspond in the rule to nice geometric configurations.

A natural lower bound on $n_q(k, d)$ is the Griesmer bound [5]:

$$n_q(k, d) \geq g_q(k, d) \stackrel{def}{=} \sum_{i=1}^{k-1} \left\lceil \frac{d}{q^i} \right\rceil .$$

Linear codes meeting this bound are called *Griesmer codes*. Arcs associated with Griesmer codes are called *Griesmer arcs*. Given an integer k and a prime power q, Griesmer codes are known to exist for all sufficiently large values of d. A standard approach to the problem of finding the exact value of $n_q(k,d)$ is to solve the problem for fixed k and q for all d. In this setting the main problem in coding theory is a finite one. This paper deals with linear codes over the field with four elements. The exact value of $n_4(k,d)$ was found for $k \leq 4$ for all d [4,12]. For the next dimension $k = 5$ there exist 104 values of d for which $n_4(5,d)$ is unknown [12].

In this paper, we prove the nonexistence of the hypothetical quaternary Griesmer codes of dimension $k = 5$ with $d = 295, 296$, a fact which was hitherto unknown. The problem is studied purely geometrically due to the equivalence of linear $[n,k,d]_q$-codes and arcs with parameters $(n, n-d)$ in $\mathrm{PG}(k-1,q)$ [3,8,9,11]. Thus the existence of the codes in question that have parameters $[395,5,295]_4$ and $[396,5,296]_4$ is equivalent to the existence of $(395,100)$- and $(396,100)$-arcs in $\mathrm{PG}(4,4)$. The nonexistence proof relies on the classification of arcs with parameters $(100,26)$ in $\mathrm{PG}(3,4)$. These arcs are related to caps in $\mathrm{PG}(3,4)$ and can be obtained trivially from $(102,26)$-arcs by deleting two points. The latter are obtained as the sum of the maximal 17-cap in $\mathrm{PG}(3,4)$ and the whole space. Remarkably, there exists a $(100,26)$-arc which is not extendable to the unique $(102,26)$-arc.

This paper is organized as follows. In section 2 we present some basic facts on arcs in the geometries $\mathrm{PG}(r,q)$. We explain briefly the connections between linear codes over finite fields and arcs in finite projective geometries. Furthermore, we state without proof some results that are used in the paper. These include the so-called Hill–Lizak's Extension Theorem and H. N. Ward's Divisibility Theorem. Both theorems are formulated in their geometric form. Section 3 contains the geometric characterization of the arcs with parameters $(100,26)$ in $\mathrm{PG}(3,4)$. In section 4, we prove the nonexistence of arcs with parameters $(395,100)$, and $(396,100)$ in $\mathrm{PG}(4,4)$, which settles the problem of finding the exact value of $n_4(5,d)$ for $d = 295, 296$.

2. PRELIMINARIES

A *multiset in* $\mathrm{PG}(k-1,q)$ is a mapping $\mathcal{K}\colon \mathcal{P} \to \mathbb{N}_0$, where $\mathcal{P}$ denotes the pointset of $\mathrm{PG}(k-1,q)$. The integer $\mathcal{K}(\mathcal{P}) = \sum_{P\in\mathcal{P}} \mathcal{K}(P)$ is called the *cardinality* of the multiset $\mathcal{K}$. For a subset $\mathcal{Q}$ of $\mathcal{P}$, we set $\mathcal{K}(\mathcal{Q}) = \sum_{P\in\mathcal{Q}} \mathcal{K}(P)$. The integer $\mathcal{K}(\mathcal{Q})$ is called the *multiplicity* of the subset $\mathcal{Q}$. A point of multiplicity i is called an i-point; i-lines, i-planes, i-solids etc. are defined in a similar way. Given a set of points $S \subseteq \mathcal{P}$, we define the *characteristic function* χ_S of S by

$$\chi_S(P) = \begin{cases} 1 & \text{if } P \in S; \\ 0 & \text{if } P \notin S. \end{cases}$$

A multiset $\mathcal{K}$ in $\mathrm{PG}(k-1,q)$ is called an $(n,w,k-1,q)$-*arc*, or an (n,w)-arc for short, if

(a) $\mathcal{K}(\mathcal{P}) = n$;

(b) for each hyperplane H in $\mathrm{PG}(k-1,q)$, $\mathcal{K}(H) \leq w$, and

(c) there is a hyperplane H with $\mathcal{K}(H) = w$.

In a similar way, we define an $(n, w; k-1, q)$-*blocking set* (or just (n, w)-blocking set if the geometry is clear from the context) as a multiset $\mathcal{K}$ in $\mathrm{PG}(k-1,q)$ satisfying

(d) $\mathcal{K}(\mathcal{P}) = n$;

(e) for each hyperplane H in $\mathrm{PG}(k-1,q)$, $\mathcal{K}(H) \geq w$, and

(f) there is a hyperplane H with $\mathcal{K}(H) = w$.

Given a $(n, w; k-1, q)$-arc $\mathcal{K}$, we denote by $\gamma_i(\mathcal{K})$ the maximal multiplicity of an i-dimensional flat in $\mathrm{PG}(k-1,q)$, i.e. $\gamma_i(\mathcal{K}) = \max_\delta \mathcal{K}(\delta)$, $i = 0, \ldots, k-1$, where δ runs over all i-dimensional flats in $\mathrm{PG}(k-1,q)$. If $\mathcal{K}$ is clear from the context we shall write just γ_i. In what follows, we repeatedly use the following lemma which is proved by straightforward counting.

Lemma 1. *Let $\mathcal{K}$ be an $(n, w; k-1, q)$-arc, and let Π be an $(s-1)$-dimensional flat in $\mathrm{PG}(k-1,q)$, $2 \leq s < k$, with $\mathcal{K}(\Pi) = u$. Then, for any $(s-2)$-dimensional flat Δ contained in Π, we have*

$$\mathcal{K}(\Delta) \leq \gamma_{s-1}(\mathcal{K}) - \frac{n-u}{q^{k-s} + \ldots + q}.$$

For an $(n, w; k-1, q)$ arc $\mathcal{K}$, denote by a_i the number of hyperplanes H in $\mathrm{PG}(k-1,q)$ with $\mathcal{K}(H) = i$, $i \geq 0$. Let further λ_j be the number of points P from $\mathcal{P}$ with $\mathcal{K}(P) = j$. The sequence $(a_0, a_1, \ldots)$ is called *the spectrum* of $\mathcal{K}$. Simple counting arguments yield the following identities, which are equivalent to the first three MacWilliams identities for linear codes:

$$\sum_{i=0}^{n-d} a_i = \frac{q^k - 1}{q-1}, \tag{2.1}$$

$$\sum_{i=1}^{n-d} i a_i = n \cdot \frac{q^{k-1} - 1}{q-1}, \tag{2.2}$$

$$\sum_{i=2}^{n-d} \binom{i}{2} a_i = \binom{n}{2} \frac{q^{k-2} - 1}{q-1} + q^{k-2} \cdot \sum_{i=2}^{\gamma_0} \binom{i}{2} \lambda_i. \tag{2.3}$$

Set $w = n-d$ and $v_i = (q^i - 1)/(q-1)$. The following identity is easily obtained from (2.1)–(2.3):

$$\sum_{i=0}^{w} \binom{w-i}{2} a_i = \binom{w}{2} v_k - n(w-1) v_{k-1} + \binom{n}{2} v_{k-2} + q^{k-2} \cdot \sum_{i=2}^{\gamma_0} \binom{i}{2} \lambda_i. \tag{2.4}$$

Note that the sum on the left-hand side can be written as $\sum_H \binom{w-\mathcal{K}(H)}{2}$, where H runs over all hyperplanes of $\mathrm{PG}(k-1,q)$. Let us fix a hyperplane H_0. Given a subspace δ of codimension 2 contained in H_0, denote by $H_1, H_2, \ldots, H_q$ the remaining hyperplanes through δ. Set

$$\eta_i = \max_{\delta\colon \mathcal{K}(\delta)=i} \sum_{j=1}^{q} \binom{w-\mathcal{K}(H_j)}{2}. \tag{2.5}$$

Here the maximum is taken over all hyperlines δ of multiplicity i contained in H_0. Assume the spectrum (b_i) of the restriction of $\mathcal{K}$ to H_0, is known. We have

$$\sum_H \binom{w-\mathcal{K}(H)}{2} \le \sum_j b_j\eta_j + \binom{w-\mathcal{K}(H_0)}{2},$$

which by (2.4) implies

$$\sum_j b_j\eta_j + \binom{w-\mathcal{K}(H_0)}{2} \ge \binom{w}{2}v_k - n(w-1)v_{k-1} + \binom{n}{2}v_{k-2} + q^{k-2}\cdot\sum_{i=2}^{\gamma_0}\binom{i}{2}\lambda_i. \tag{2.6}$$

Clearly, (2.6) is a necessary condition for the existence of an (n,w)-arc in $\mathrm{PG}(k-1,q)$. It can also be used to rule out the existence of hyperplanes H for which $\mathcal{K}|_H$ has a given spectrum.

The following argument will be used throughout the paper. Let $\mathcal{K}$ be an $(n, n-d; k-1, q)$-arc, i.e. an arc associated with an $[n,k,d]_q$-code. Fix an i-dimensional flat δ in $\mathrm{PG}(k-1,q)$, with $\mathcal{K}(\delta)=t$. Let further π be a j-dimensional flat in $\mathrm{PG}(k-1,q)$ of complementary dimension, i.e. $i+j=k-2$ and $\delta\cap\pi=\emptyset$. Define the projection $\varphi=\varphi_{\delta,\pi}$ from δ onto π by

$$\varphi\colon \begin{cases} \mathcal{P}\setminus\delta & \to & \pi \\ Q & \to & \pi\cap\langle\delta,Q\rangle. \end{cases} \tag{2.7}$$

Here $\mathcal{P}$ is the set of points of $\mathrm{PG}(k-1,q)$. Note that φ maps $(i+s)$-flats containing δ into $(s-1)$-flats in π. Given a set of points $\mathcal{F}\subset\pi$, define the induced arc $\mathcal{K}^\varphi$ by

$$\mathcal{K}^\varphi(\mathcal{F}) = \sum_{\varphi_{\delta,\pi}(P)\in\mathcal{F}} \mathcal{K}(P).$$

If $\mathcal{F}$ is a k'-dimensional flat in π then $\mathcal{K}^\varphi(\mathcal{F}) \le \gamma_{k'+i+1} - t$.

In this paper, we consider arcs in $\mathrm{PG}(3,4)$ or $\mathrm{PG}(4,4)$ and always take δ to be a point (in the three-dimensional case) or a line (in the four-dimensional case); in both cases π will be a plane disjoint from δ. Every line L in π is then the image

of a hyperplane (a plane or a solid) containing δ. If $P_0, \ldots, P_q$ are the points on L we call the $(q+1)$-tuple $(\mathcal{K}^\varphi(P_0), \ldots, \mathcal{K}^\varphi(P_q))$ the *type of* L.

It was mentioned already, that the existence of linear $[n,k,d]_q$ codes of full length is equivalent to that of $(n, n-d; k-1, q)$-arcs. Two linear codes with the same parameters are semilinearly isomorphic if and only if the corresponding arcs are projectively equivalent. H.N. Ward proved in [13] a remarkable theorem on the divisibility of codes meeting the Griesmer bound. Below we give Ward's result restated for arcs in $\mathrm{PG}(k-1,q)$ (cf. [9]).

Theorem 1. *Let $\mathcal{K}$ be a Griesmer (n,w)-arc in $\mathrm{PG}(k-1,p)$, p prime, with $w \equiv n \bmod p^e$, $e \geq 1$. Then $\mathcal{K}(H) \equiv n \bmod p^e$ for every hyperplane H.*

For codes over $\mathbb{F}_4$ (resp. arcs in geometries over $\mathbb{F}_4$) we have the following weaker version of this result [13].

Theorem 2. *Let $\mathcal{K}$ be a Griesmer (n,w)-arc in $\mathrm{PG}(k-1,4)$ with $w \equiv n \bmod 2^e$. Then $\mathcal{K}(H) \equiv n \bmod 2^{e-1}$ for every hyperplane H.*

An (n,w)-arc $\mathcal{K}$ in $\mathrm{PG}(k-1,q)$ is called *extendable* if there exists an $(n+1,w)$-arc $\mathcal{K}'$ in $\mathrm{PG}(k-1,q)$ with $\mathcal{K}'(x) \geq \mathcal{K}(x)$ for every point of $\mathrm{PG}(k-1,q)$. The next extension result about arcs stated below follows directly from Hill-Lizak's extension theorem [6,7]:

Theorem 3. *Let $\mathcal{K}$ be an $(n,w;k-1,q)$-arc with $\gcd(n-w,q)=1$. Assume that the multiplicities of all hyperplanes are congruent to n or w modulo q. Then $\mathcal{K}$ can be extended to an $(n+1,w)$-arc.*

The following theorem from [10] follows from a result by Beutelspacher [1] and can be viewed as a generalization of Hill-Lizak's extension theorem.

Theorem 4. *Let $\mathcal{K}$ be a (n,w)-arc in $\mathrm{PG}(k-1,q)$, $q=p^s$, with spectrum $(a_i)_{i\geq 0}$. If $w \not\equiv n \bmod p$ and*

$$\sum_{i \not\equiv w \bmod q} a_i \leq q^{k-2} + q^{k-3} + \ldots + 1 + q^{k-3} \cdot r(q)$$

where $q + r(q) + 1$ is the minimal size of a non-trivial blocking set of $\mathrm{PG}(2,q)$, then there exists an $(n+1,w)$-arc.

As a corollary we can derive the following useful result [10]:

Corollary 1. *Let $\mathcal{K}$ be a nonextendable (n,w)-arc in $\mathrm{PG}(k-1,q)$, $q=p^s$, with $w \equiv n+1 \pmod q$ and with spectrum $(a_i)_{i\geq 0}$. Let θ denote the maximal number of hyperplanes of multiplicity $\not\equiv n+1 \pmod q$ incident with any subspace of codimension 2 of H, where H is a hyperplane of multiplicity $\mathcal{K}(H) \equiv w \pmod q$. Then $\sum_{i \not\equiv n, n+1 \pmod q} a_i > q^{k-3} \cdot r(q)/(\theta - 1)$, where $r(q)$ is as in Theorem 4. In particular, we have $\sum_{i \not\equiv n, n+1 \pmod q} a_i > q^{k-3} \cdot r(q)/(q-1)$.*

3. CLASSIFICATION OF THE $(100, 26)$-ARCS IN PG$(3, 4)$

In this section we classify the arcs with parameters $(100, 26)$ in PG$(3, 4)$. It is known that a $(102, 26)$-arc in PG$(3, 4)$ is the sum of a 17-cap plus the whole space, and hence is unique. By Hill-Lizak's extension theorem every $(101, 26)$-arc is extendable to a $(102, 26)$-arc. One obvious way to construct $(100, 26)$-arcs in PG$(3, 4)$ is to delete a point from a $(101, 26)$-arc, or equivalently, to delete two points from a $(102, 26)$-arc. It turns out however that there exist $(100, 26)$-arcs that cannot be obtained in this way.

Let $\mathcal{K}$ be a $(100, 26)$-arc. By Lemma 1

$$\gamma_0(\mathcal{K}) = 2, \gamma_1(\mathcal{K}) = 7, \gamma_2(\mathcal{K}) = 26.$$

From now on we assume that $\mathcal{K}$ is a non-extendable $(100, 26)$-arc in PG$(3, 4)$. The restriction of $\mathcal{K}$ to a maximal hyperplane is a $(26, 7)$-arc. The characterization of such arcs is given by the following lemma.

Lemma 2. *A* $(26, 7)$*-arc in* PG$(2, 4)$ *is one of the following:*

(1) two copies of the plane minus three non-concurrent lines minus a point (type (A));

(2) the sum of the plane plus a hyperoval minus a point (type (B));

(3) two 7-lines through a common 0-point; all points outside these two 7-lines are 1-points (type (C)).

The arcs of the first two types are extendable while an arc of the third type is not. This result is easily obtained from the known results on arcs and blocking sets in PG$(2, 4)$ and we omit the proof. Below we present the spectra of these arcs, as well as the possible line types after a projection from a 0-point. For the second spectrum of type (B) there are no 0-points.

Type	a_7	a_6	a_5	a_4	a_3	a_2	λ_2	λ_1	λ_0	Line types
(A)	14	4	0	0	2	1	9	8	4	77732 77633 66662
	13	5	0	0	3	0	8	10	3	77633
(B)	12	3	4	2	0	0	6	14	1	66644
	10	5	6	0	0	0	5	16	0	-
(C)	11	6	1	3	0	0	6	14	1	77444

It is important to note that a $(26, 7)$-arc does not have 0- or 1-lines, as well as 5-lines with a 0-point. It is also worth noting that a $(26, 7)$-arc cannot contain a 3- and a 4-line simultaneously.

Lemma 3. *Let $\mathcal{K}$ be a $(25,7)$-arc in* $\mathrm{PG}(2,4)$ *having a 6-line L with three 2- and two 0-points. Then $\mathcal{K}$ has a also a 7-line incident with a 0-point.*

Proof. Denote by λ_i, $i = 0,1,2$ the number of i-points in $\mathcal{K}$. Obviously $\lambda_2 - \lambda_0 = 4$, and since $\lambda_0 \geq 2$ and $\lambda_2 \leq 9$, we are left with four cases: $\lambda_2 = 6+i$, $\lambda_0 = 2+i$, where $i = 0,1,2,3$.

Assume for a contradiction that there is no 7-line with three 2-points and one 1-point. We consider the case $\lambda_0 = 2$. Let the three 2-points outside L be collinear. Then the line defined by them meets the 6-line L in a 0-point. This line should have another 0-point, because of our assumption, which gives $\lambda_0 \geq 3$, a contradiction. If the three 2-points outside L form a triangle, at least one of the lines defined by the vertices of this triangle meets L in a 2-point and hence there must be another 0-point, again a contradiction.

The cases $\lambda_0 = 3,4,5$ are dealt with in a similar way. □

Lemma 4. *Let $\mathcal{K}$ be a $(100,26)$-arc in* $\mathrm{PG}(3,4)$. *Then for every plane π in* $\mathrm{PG}(3,4)$ $\mathcal{K}(\pi) \geq 12$.

Proof. Let us note that by Lemma 1 $\mathcal{K}(\pi) \neq 7, 10, 11, 23$. Without loss of generality we consider the case when $\mathcal{K}$ is a non-extendable arc. If $\mathcal{K}$ is extendable the possible plane multiplicities are $26, 25, 24, 22, 21, 20$, and the lemma holds trivially.

Planes π of multiplicity ≤ 5 are ruled out by the nonexistence of 0- or 1-lines in $(26,7)$-arcs. It is easily seen that for planes of multiplicity at most 5, there is always a 0-point P in π which is incident only with 0- or 1-lines. Since P lies in at least one 26-plane π' ($\mathcal{K}$ was assumed to be non-extendable) the line $\pi \cap \pi'$ is a 0- or 1-line, a contradiction.

Assume there exists a 6-plane π_0. Consider a 2-line L in π_0. The line L is incident with at least two 26-planes, π_1 and π_2 say. Clearly π_1 and π_2 are of type (A). There exists a 0-point on L such that after a projection from that point the images of π_1 and π_2 are $(7,7,7,3,2)$. Now in the projection plane there is a line of type $(3,3,2/0,x,y)$ for some integers x, y. Now $x, y \leq 4$ since a 26-plane does not have a 5-line with a 0-point. This is a contradiction since $\mathcal{K}$ cannot have a 6-plane and a plane of multiplicity < 14 simultaneously.

Assume there exists a plane π_0 of multiplicity 8. Consider a projection from a 0-point in a 3-line L in π_0. The images of the other four planes through L are of type $(7,7,7-\epsilon,3,2+\epsilon)$ with $\epsilon \in \{0,1\}$. Now the projection plane necessarily contains a line of type $(7,7,6,6,0)$ which is an impossible type by Lemma 2. Planes of multiplicity 9 are ruled out similarly. In this case, we can even select the point on the 9-plane in such way that its image is of type $(3,3,3,0,0)$, which simplifies the proof. □

Lemma 5. *Let $\mathcal{K}$ be a $(100,26)$-arc in* $\mathrm{PG}(3,4)$. *Then there is no plane π in* $\mathrm{PG}(3,4)$ *with* $12 \leq \mathcal{K}(\pi) \leq 15$.

Proof. First, we shall rule out planes of multiplicity 15. Assume π_0 is a plane with $\mathcal{K} = 15$. The restriction of $\mathcal{K}$ to π_0 is a plane minus a line L and minus a point Q which lies off L. Assume there is a 0-point P outside π_0. A 26-plane π_1 through P (it exists since $\mathcal{K}$ is not extendable) has at least two 0-points (P and one on π_0). Hence this plane contains an arc of type (A) and therefore contains also Q. Now consider a 7-line L' in π_1 through P. It is incident with at least two further 26-planes that have at least two and hence at least three 0-points. On the other hand they meet π_0 in a 4-line which contradicts Lemma 4 (a $(26, 7)$-arc of type (A) does not have a 4-line). We have proved so far that there are no 0-points outside π_0. Now Q should be incident with a 26-plane that meets π_0 in a 3-line and hence has two 0-points, which is impossible.

In the same way we can rule out the existence of 14-planes (the complement of a line and two points or the complement of a Baer subplane), and of 16-planes in which the 0-points are collinear.

Now we are going to prove that planes of multiplicity 13 do not exist. The proof of the nonexistence of 12-planes is similar and more simple.

Assume there exists a 13-plane π_0. Fix a 4-line L in π_0 and denote the other four planes through L by π_i, $i = 1, \ldots, 4$. Without loss of generality π_1, π_2, π_3 are 26-planes and π_4 is a 25-plane. Consider a projection from P which we denote by φ and set $L_i = \varphi(\pi_i)$. Let us note that L_4 does not contain a 7-point. This follows from the fact that this point must be incident with three 26-lines and the types of L_1, L_2, L_3 are $(7, 7, 4, 4, 4)$ or $(6, 6, 6, 4, 4)$ and L_0 is forced to be of type $(4, 4, 4, 4, 0)$, a contradiction. Hence the type of L_4 is one of $(4, 6, 6, 6, 3)$, $(4, 6, 6, 5, 4)$ or $(4, 6, 5, 5, 5)$.

Now a 13-plane is the complement of a $(8, 1)$-blocking set, and hence one of the following: (a) the complement of a line and three points or (b) the complement of a Baer subplane and a point.

In case (a) there exists a point P such that the projection of π_0 from that point is of type $(4, 3, 3, 3, 0)$. Now none of the lines L_1, L_2, L_3 is of type $(7, 7, 4, 4, 4)$ since a 26-line through a 7-point should have two points of multiplicity at most 3. Consequently, the line L_4 should have two 3-points which is impossible. Therefore L_1, L_2, L_3 are of type $(6, 6, 6, 4, 4)$. Now with all three possibilities for L_4 we get a contradiction. For instance, if L_4 is of type $(4, 6, 6, 6, 3)$, the set of points

$$\mathcal{F} = \{X \in L_1 \cup L_2 \cup L_3 \cup L_4 \mid \mathcal{K}^{\varphi}(X) = 6\} \cup \{Y \in L_1 \cup L_2 \cup L_3 \cup L_4 \mid \mathcal{K}^{\varphi}(Y) = 3\}$$

must be a $(15, 4)$-arc and there is a line of type $(4, 4, 4, 3, 0)$. But we have already ruled out the existence of 15-planes. The other two possibilities for L_4 are dealt with in a similar fashion.

(b) As in the nonexistence proof for 15-planes we can show that there are no 0-points outside π_0. Now denote by P the extra 0-point on π_0 which is not from the removed Baer subplane. The lines in π_0 through P have multiplicities 3,3,3,3,1. hence a 26-plane through P (which necessarily exists) has two 0-points, a contradiction. □

Lemma 6. *There exists a unique* $(100, 26)$*-arc in* $\mathrm{PG}(3, 4)$ *with the following property:* $\mathrm{PG}(3, 4)$ *has a 24-plane with a 7-line consisting of three 2-points, one 1-point, and one 0-point.*

Proof. Denote by π_0 the 24-plane from the condition of the lemma. Let L be a 7-line in π_0 and let P be the unique 0-point on L. The remaining four planes through L, denoted by $\pi_1, \dots, \pi_4$, are 26-planes. We consider a projection φ from the point P. Set $Q = \varphi(L)$, and $L_i = \varphi(\pi_i)$, $i = 0, \dots, 4$. Clearly, $\mathcal{K}|_{\pi_i}$ $i = 1, \dots, 4$, are of type (A) or (C) (cf. Lemma 4). Hence the possible types of the lines $L_1, \dots, L_4$ are $(7, 7, 7 - \varepsilon, 3, 2 + \varepsilon)$, $\varepsilon = 0$ or 1, or $(7, 7, 4, 4, 4)$. Now we have the following possibilities for the four 26-planes through L:

(i) AAAA, (ii) AAAC, (iii) AACC, (iv) ACCC, (v) CCCC.

(i) The lines $L_1, \dots, L_4$ are all of type $(7, 7, 7-\varepsilon, 3, 2+\varepsilon)$. Assume the pointset $\mathcal{X} = \{X \mid \mathcal{K}^{\varphi}(X) \geq 6\}$ in the projection plane has four collinear points and denote by M the line incident with them. Let Z be the fifth point on M. It has multiplicity at most 2. Now every line through Z, different from L_0 or M has at least two points from $\mathcal{X}$, which is impossible. Hence $\mathcal{X}$ is a $(9, 3)$-arc. Moreover, there is no external line to $\mathcal{X}$ since it would be of multiplicity ≤ 15. Now for every point $R \neq Q$ on L_0 we have $\mathcal{K}^{\varphi}(R) \leq 4$. This implies $\mathcal{K}^{\varphi}(L_0) \leq 7 + 4 \cdot 4 = 23 < 24$, a contradiction.

(ii) Let L_4 be the line of type $(7, 7, 4, 4, 4)$. In this case there exists a 26-line through a 7-point on L_1 (different from P) which is of type $(7, 4, *, *, *)$ and hence is forced to be of type $(7, 7, 4, 4, 4)$. This is clearly impossible since only L_0 and L_4 can have points of multiplicity 4.

(iii) The proof is similar to that of (ii).

(iv) Let L_1 be of type $(7, 7, 7 - \varepsilon, 3, 2 + \varepsilon)$, and let L_2, L_3, L_4 be of type $(7, 7, 4, 4, 4)$. Now L_0 is forced to be of type $(7, 5, 4, 4, 4)$. Two of the 7-points on L_1 plus the 7-points on L_2, L_3, and L_4 form an oval which is extendable to a hyperoval by adding a point on L_0. Now through the point of multiplicity $7 - \varepsilon$ on L_1 we have a secant to the hyperoval (different from L_1) which is of type $(7, 7, 7-\varepsilon, 4, 4)$, which is impossible.

(v) The pointset $\{X \mid \mathcal{K}^{\varphi}(P) = 7\}$ is an oval. Denote the nucleus of the oval by N. Clearly, $\mathcal{K}^{\varphi}(N) \geq 5$ since there is a line of type $(7, \mathcal{K}^{\varphi}(N), 4, 4, 4)$. If L_0 is of type $(7, \mathcal{K}^{\varphi}(N), x_1, x_2, x_3)$ then $2 \leq x_i \leq 4$. Hence the structure of $\mathcal{K}$ can be represented as

$$\mathcal{K} = \mathcal{F} - \mathcal{B}.$$

Here $\mathcal{F}$ is the sum of the whole space plus a cone with an hyperoval as a base curve minus twice the vertex of the cone P; $\mathcal{B}$ is a set of 8 points blocking once every plane that does not contain P. Clearly $\mathcal{B} \cup \{P\}$ is a $(9, 1)$-blocking set with a 4-line. □

Lemma 7. *There exists no* $(100, 26)$*-arc in* $\mathrm{PG}(3, 4)$ *with the following property: every 7-line with a 0-point is contained in two 25-planes.*

Proof. With L and P as in the previous theorem, let π_0, π_1, and π_2 be the 26-planes and π_3, π_4 – the 25-planes through L. We have the following possibilities for the three 26-planes through L:

(i) AAA, (ii) AAC, (iii) ACC, (iv) CCC.

In all cases we consider a projection φ from P. We set $L_i := \varphi(\pi_i)$, $Q = \varphi(L)$.

(i) Here we shall deal with the case when L_0, L_1, L_2 are all of type $(7, 7, 7, 3, 2)$. The case when some (or all) of these lines are of type $(7, 7, 6, 3, 3)$ is ruled out in the same way. Assume three of the 7-points different from Q are collinear. Then the line defined by them meets one of L_3 or L_4 (L_3 say) in a point of multiplicity at most 2. Hence there is a line through this point which is of type $(7, 3, 3, \leq 2, x)$ or $(3, 3, 3, \leq 2, x)$. In the first case we get a line of multiplicity at most 22, a contradiction. In the second we get $x \leq 4$, which gives a line of multiplicity at most 15 and is again impossible.

Now the six points different from Q on L_0, L_1, L_2 that are of multiplicity 7 form a hyperoval. Through the 2-point on L_0, there exist two external lines to the hyperoval and one of them has to be of type $(2, 2, 3, *, *)$ (or $(2, 2, 2, *, *)$). Now it is an easy check that this line should be of multiplicity less than 16, which is impossible.

(ii) Let L_0 and L_1 be of type $(7, 7, 7 - \varepsilon, 3, 2 + \varepsilon)$, $\varepsilon \in \{0, 1\}$, and let L_2 be of type $(7, 7, 4, 4, 4)$. First observe that L_3 or L_4 do not have a point of multiplicity 7. In such case there is a line of type $(7, 4, 3/2, 3/2, *)$, which is impossible since a 25- and a 24-plane do not meet in a 7-line. Assume one of L_0 and L_1, L_0 say, is of type $(7, 7, 7, 3, 2)$. Now through the 2-point on L_0 there exist two lines of type $(2, 3, 4, 4, 4)$ or $(2, 2, 4, 4, 4)$, and hence each of L_3, L_4 has two points of multiplicity 4. Since type $(7, 5, 5, 4, 4)$ is impossible for L_3 or L_4 (by the nonexistence of 26-planes with a 5-line which contains a 0-point), both lines are of type $(7, 6, 4, 4, 4)$. This implies that L_1 is also of type $(7, 7, 7, 3, 2)$ and the set

$$\{X \mid \mathcal{K}^{\varphi}(X) = 7, X \neq P\} \cup \{Y \mid Y \in L_3, \mathcal{K}^{\varphi}(Y) = 6\}$$

is not a hyperoval (since it has tangents). Hence there is a line of type $(7, 7, 7, 4, 4)$ or $(7, 7, 6, 4, 2)$. The former is clearly impossible and the latter is ruled out by Lemma 2.

Now we are left with the case where the lines L_0 and L_1 are both of type $(7, 7, 6, 3, 3)$. The three 7-points different from P on L_0, L_1, L_2 are obviously not collinear. Now there exists a 26-line of type $(7, 4, 6/3, *, *)$ which is again ruled out by Lemma 2.

(iii) The proof is similar to that of (ii).

(iv) In this case the lines L_0, L_1, L_2 are all of type $(7, 7, 4, 4, 4)$ and the three 7-points different from Q are not collinear. Then L_3 and L_4 have three points of multiplicity at most 4, whence they are of type $(7, 6, 4, 4, 4)$. Now $\{X \mid \mathcal{K} \geq 6\}$ is a hyperoval. The arc $\mathcal{K}$ can be represented as $\mathcal{K} = \mathcal{F} - \mathcal{B}$, where $\mathcal{F}$ and $\mathcal{B}$ are as in Lemma 6(v). Again $\mathcal{B} \cup \{P\}$ is a $(9, 1)$-blocking set with two 3-lines meeting in

a point and four coplanar points in a general position. A blocking set with this structure does not exist. □

Summing up the results from Lemma 6 and Lemma 7 we get the following theorem.

Theorem 5. *Let $\mathcal{K}$ be a $(100, 26)$-arc in* PG$(3, 4)$. *Then $\mathcal{K}$ is one of the following:*

(1) the sum of a cap and the whole space minus two points;

(2) the arc from the cone construction (case (v) in Lemma 6).

4. THE NONEXISTENCE OF $(395, 100)$- AND $(396, 100)$-ARCS IN PG$(4, 4)$

In this section we prove the nonexistence of arcs with parameters $(395, 100)$ and $(396, 100)$ in PG$(4, 4)$. Equivalently, there exist no $[395, 5, 295]_4$- and $[396, 5, 296]_4$-codes. This resolves two open cases in Maruta's tables for optimal linear codes with $k = 5$, $q = 4$, namely $n_4(5, 295) = 396$ and $n_4(5, 296) = 397$.

As already noted, we will tackle the problem geometrically and will prove the nonexistence of arcs in PG$(4, 4)$ with parameters $(395, 100)$ and $(396, 100)$. The proof is based on the knowledge of the structure of the maximal planes which was completed in the previous section.

Theorem 6. *There exist no $(396, 100)$-arcs in* PG$(4, 4)$.

Proof. Assume that $\mathcal{K}$ is a $(396, 100)$-arc in PG$(4, 4)$. From the geometric version of Ward's divisibility theorem, as well as by easy counting we have that the admissible hyperplane multiplicities with respect to $\mathcal{K}$ are the following: $100, 96, 92, 86, 84, 82, 80, 78,$ and 76. Smaller hyperplanes are impossible since a $(100, 26)$-arc in PG$(3, 4)$ does not have planes of multiplicity less than 20.

Since the number of 2-points in $\mathcal{K}$ is at least 55 and the maximal size of a cap in PG$(4, 4)$ is 41 [2], there exist three collinear 2-points. A line incident with three 2-points is either a 6- or a 7-line.

First assume there is a 7-line L with three 2-points and consider a projection φ from L. This line is necessarily contained in a 26-plane, π say, and hence in a 100-solid. The five solids through π, denoted by Δ_i, $i = 0, \ldots, 4$, are 100-solids of type (2). Hence $\mathcal{K}^\varphi$ has five 17- and sixteen 19-points. Moreover, the 17-points should form a blocking set and hence are collinear. Therefore there is a 92-solid through L.

Now consider another projection, denoted by ψ from the 0-point P of L. The image of a 100-solid has five 7-points, one 5-point and fifteen 4-points with the 7- and 5-points forming a hyperoval. Clearly, $\mathcal{K}^\psi$ has seventeen 7-points that form a

cap. Each 7-point is incident with a unique tangent plane to the cap. This plane is forced to contain all the 5-points (since it is the image of the 92-solid above). This observation is true for every 7-point, which gives a contradiction.

Now consider a 6-line L consisting of three 2- and two 0-points, and a projection φ from L. Since a 100-solid does not have such a line, we have that L is contained in solids of multiplicity at most 96. Since every point is contained in five solids through L, counting the multiplicity of $\mathcal{K} - \chi_L$ we get

$$390 = |\mathcal{K} - \chi_L| \leq \frac{21 \cdot 90}{5} = 378,$$

a contradiction. □

Now we are going to prove the nonexistence of $(395, 100)$-arcs in $\mathrm{PG}(4, 4)$ by demonstrating that if such an arc exists it is extendable to the nonexisting $(396, 100)$-arc.

Theorem 7. *There exist no* $(395, 100)$*-arcs in* $\mathrm{PG}(4, 4)$.

Proof. Assume $\mathcal{K}$ is a $(395, 100)$-arc in $\mathrm{PG}(4, 4)$. As in the proof of theorem 6, there exist three collinear 2-points. We consider two cases: (a) the line L defined by these points is a 7-line, and (b) the line L defined by these points is a 6-line.

(a) The line L is necessarily contained in a 100-solid Δ_0, which is forced to be nonextedable. Since a $(100, 26)$-arc in $\mathrm{PG}(3, 4)$ has just planes of multiplicity 26, 24, 22, and 20, the possible multiplicities for solids with respect to $\mathcal{K}$ are: $100, 99, 92, 91, 86, \ldots, 83$, and $78, \ldots, 75$.

First we rule out the existence of 78- and 77-solids. By easy counting, solids of this multiplicity have to be projective. Hence such solids are either the complement of a line and two (resp. three points, or the complement of a Baer subplane (resp. Baer subplane and a point). Denote such a solid by Δ_1. Note that Δ_1 must meet Δ_0 in a 20-plane since the latter has no planes of smaller multiplicity. Consider a projection φ from a 4-line K in the plane $\Delta_0 \cap \Delta_1$. Now $\varphi(\Delta_0)$ is of type $(22, 22, 20, 16, 16)$. The possible types of the line $\varphi(\Delta_1)$ are the following:

if Δ_1 is a 77-plane: $(16, 16, 15, 15, 11)$, $(16, 16, 16, 14, 11)$, $(16, 16, 16, 15, 10)$, $(16, 16, 16, 16, 9)$;

if Δ_1 is a 78-plane: $(16, 16, 16, 14, 12)$, $(16, 16, 16, 15, 11)$, $(16, 16, 16, 16, 10)$.

We shall deal with the case when Δ_1 is a 78-solid (the case $\mathcal{K}(\Delta_1) = 77$ is treated analogously). The other three solids Δ_i, $i = 2, 3, 4$, through $\Delta_0 \cap \Delta_1$ are forced to be of multiplicity 99. Since a 26-plane in a 99-solid is contained in four 100-solids the image $\varphi(\Delta_i)$, $i = 2, 3, 4$, does not have a 22-point. Hence the lines $\varphi(\Delta_i)$, $i = 2, 3, 4$, are of type $(20, 20, 20, 19, 16)$. A 22-point in the projection plane is incident with four 96-lines (images of 100-solids) and one 95-line (the image of 99-solid). Therefore the 95-lines through each 22-point contain the 19-points on the lines $\varphi(\Delta_i)$, $i = 2, 3, 4$. This is obviously impossible.

Next we rule out the existence of 86-solids. A 22-plane in a 100-solid has one 2-point and twenty 1-points. Hence a 86-solid has one 2-point and eighty-four 1-points. Such a solid has just 21- and 22-planes and therefore the arc $\mathcal{K}$ has no further solids of multiplicity 85 and 86. This implies that $\mathcal{K}$ is extendable to the nonexistent $(396, 100)$-arc by Corollary 1.

Finally, an 85-solid Δ_1 must have a 2-point, otherwise all points are 1-points and all planes are 21-planes, which is impossible. If Δ_1 is an 85-solid then every 6-line is incident with one 21-plane, and four 22-planes; consequently, a 6-line has exactly one 2-point. This implies that Δ_1 consist either of one 2-point, one 0-point and all the rest 1-points, or of two 2-points, two 0-points (all these collinear) and all the rest 1-points. Now this is the only 85-solid since two such solids meet in a plane of multiplicity at most 18. Again $\mathcal{K}$ is extendable by Corollary 1 and we arrive at a contradiction.

(b) Now we assume that every three collinear 2-points determine a 6-line. Note that every 100-solid is an extendable $(100, 26)$-arc and consequently 26-planes cannot have three collinear 2-points. Consider such a 6-line, L say. Note that L is not contained in a 100-solid. Assume that L is contained in a 99-solid Δ. There exists a 25-plane π with $L \subset \pi \subset \Delta$. Denote by P, Q the two 0-points on L. If there exists a 7-line in π through P then counting the multiplicities of the planes through this line we get

$$99 \leq \mathcal{K}(\Delta) \leq 5 \cdot 25 - 4 \cdot 7 = 97,$$

a contradiction. This implies that $\mathcal{K}|_\pi$ is extendable to a $(26, 7)$-arc by turning P into 1-point. But this implies that Q is incident with a 7-line, again a contradiction.

We have proved that the multiplicities of the solids through L do not exceed 97. Consider a projection φ from L. We have $|\mathcal{K}^\varphi| = 389$ and by the above argument $\mathcal{K}^\varphi(M) \leq 91$ for every line M in the projection plane. Now counting the multiplicities of all lines in the plane of projection, we get

$$389 = |\mathcal{K}^\varphi| \leq \frac{21 \cdot 91}{5} = \frac{1911}{5} < 383,$$

a contradiction. □

Finally, we state Theorems 6 and 7 in coding-theoretic terms.

Corollary 2. *There exist no linear codes with parameters* $[395, 5, 295]_4$, *and* $[396, 5, 296]_4$. *Consequently,* $n_4(5, 295) = 396$, *and* $n_4(5, 296) = 397$.

ACKNOWLEDGEMENTS. This research has been supported by the Science Research Fund of Sofia University under Contract No. 80-10-81/15.04.2019.

5. REFERENCES

[1] Beutelspacher, A.: Blocking sets and partial spreads in finite projective spaces. *Geom. Dedicata*, **9**, 1980, 130–157.

[2] Edel Y., Bierbrauer, J.: 41 is the largest size of a cap in PG$(4, 4)$. *Des. Codes Cryptogr.*, **16**, 1999, 151–160.

[3] Dodunekov, S., Simonis, J.: Codes and projective multisets. *Electron. J. Combin.*, **5**, 1998, R37.

[4] Grassl, M.: Bounds on the minimum distance of linear codes and quantum codes. http://www.codetables.de

[5] Griesmer, J. H.: A bound for error-correcting codes. *IBM J. Res. Develop.*, **4**, 1960, 532–542.

[6] Hill, R.: An extension theorem for linear codes. *Des. Codes Cryptogr.*, **17**, 1999, 151–157.

[7] Hill, R., Lizak, P.: Extensions of linear codes. In: *Proc. Int. Symp. on Inf. Theory*, Whistler, BC, Canada, 1995.

[8] Landjev, I.: Linear codes over finite fields and finite projective geometries. *Discrete Math.*, **213**, 2000, 211–244.

[9] Landjev, I.: The geometric approach to linear codes. In: *Finite Geometries*, (A. Blokhuis et al., Eds.), Kluwer Acad. Publ., 2001, 247–256.

[10] Landjev, I., Rousseva, A.: An extension theorem for arcs and linear codes. *Probl. Inf. Transm.*, **42**, 2006, 65–76.

[11] Landjev, I., Storme, L.: Galois Geometries and Coding Theory. In: *Current Research Topics in Galois Geometry*, Chapter 8, (Jan De Beule and Leo Storme, Eds.), NOVA Science Publishers, 2012, 187–214.

[12] Maruta, T.: Giesmer bound for linear codes over finite fields. http://www.mi.s.osakafu-u.ac.jp/maruta/griesmer/

[13] Ward, H. N.: Divisibility of codes meeting the Griesmer bound. *J. Combin. Theory Ser. A*, **83**, 1998, 79–93.

Received on June 13, 2019
Received in a revised form on March 2, 2020

Assia P. Rousseva

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 blvd. J. Bourchier, BG-1164 Sofia
BULGARIA

E-mail: assia@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 106

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 106

A NONREALIZATION THEOREM IN THE CONTEXT OF DESCARTES' RULE OF SIGNS

HASSEN CHERIHA, YOUSTRA GATI AND VLADIMIR PETROV KOSTOV

For a real degree d polynomial P with all nonvanishing coefficients, with c sign changes and p sign preservations in the sequence of its coefficients ($c + p = d$), Descartes' rule of signs says that P has $pos \leq c$ positive and $neg \leq p$ negative roots, where $pos \equiv c(\mod 2)$ and $neg \equiv p(\mod 2)$. For $1 \leq d \leq 3$, for every possible choice of the sequence of signs of coefficients of P (called sign pattern) and for every pair (pos, neg) satisfying these conditions there exists a polynomial P with exactly pos positive and neg negative roots (all of them simple); that is, all these cases are realizable. This is not true for $d \geq 4$, yet for $4 \leq d \leq 8$ (for these degrees the exhaustive answer to the question of realizability is known) in all nonrealizable cases either $pos = 0$ or $neg = 0$. It was conjectured that this is the case for any $d \geq 4$. For $d = 9$, we show a counterexample to this conjecture: for the sign pattern $(+, -, -, -, -, +, +, +, +, -)$ and the pair $(1, 6)$ there exists no polynomial with 1 positive, 6 negative simple roots and a complex conjugate pair and, up to equivalence, this is the only case for $d = 9$.

Keywords: Real polynomials, Descartes' rule of signs, sign pattern.

2010 Math. Subject Classification: Primary: 26C10; Secondary: 30C15.

1. INTRODUCTION

In his work *La Géométrie* published in 1637, René Descartes (1596–1650) announces his classical rule of signs which says that for the real polynomial $P(x, a) := x^d + a_{d-1}x^{d-1} + \cdots + a_0$, the number c of sign changes in the sequence of its coefficients serves as an upper bound for the number of its positive roots. When roots are counted with multiplicity, then the number of positive roots

has the same parity as c. One can apply these results to the polynomial $P(-x)$ to obtain an upper bound on the number of negative roots of P. For a given c, one can find polynomials P with c sign changes with exactly c, $c-2$, $c-4$, ... positive roots. One should observe that by doing so one does not impose any restrictions on the number of negative roots.

Remark 1. It is mentioned in [1] that 18th century authors used to count roots with multiplicity while omitting the parity conclusion; later this conclusion was attributed (see [3]) to a paper of Gauss of 1828 (see [7]), although it is absent there, but was published by Fourier in 1820 (see p. 294 in [6]).

In the present paper we consider polynomials P without zero coefficients. We denote by p the number of sign preservations in the sequence of coefficients of P, and by pos_P (resp. neg_P) the number of positive and negative roots of P. Thus the following condition must be fulfilled:

$$pos_P \le c\,, \quad pos_P \equiv c\,(\mathrm{mod}\ 2)\,, \quad neg_P \le p\,, \quad neg_P \equiv p\,(\mathrm{mod}\ 2)\,. \tag{1.1}$$

Definition 1. A *sign pattern* is a finite sequence σ of ($\pm$)-signs; we assume that the leading sign of σ is $+$. For a given sign pattern of length $d+1$ with c sign changes and p sign preservations, we call (c,p) its *Descartes pair*, $c+p=d$. For a given sign pattern σ with Descartes pair (c,p), we call (pos, neg) an *admissible pair* for σ if conditions (1.1), with $pos_P = pos$ and $neg_P = neg$, are satisfied.

It is natural to ask the following question: *Given a sign pattern σ of length $d+1$ and an admissible pair (pos, neg) can one find a degree d real monic polynomial the signs of whose coefficients define the sign pattern σ and which has exactly pos simple positive and exactly neg simple negative roots ?* When the answer to the question is positive we say that the couple $(\sigma, (pos, neg))$ is *realizable*.

For $d = 1$, 2 and 3, the answer to this question is positive, but for $d = 4$ D. J. Grabiner showed that this is not the case, see [8]. Namely, for the sign pattern $\sigma^* := (+,+,-,+,+)$ (with Descartes pair $(2,2)$), the pair $(2,0)$ is admissible, see (1.1), but the couple $(\sigma^*, (2,0))$ is not realizable. Indeed, for a monic polynomial $P_4 := x^4 + a_3x^3 + \cdots + a_0$ with signs of the coefficients defined by σ^* and having exactly two positive roots $u<v$ one has $a_j>0$ for $j \ne 2$, $a_2<0$ and $P_4((u+v)/2) < 0$. Hence $P_4(-(u+v)/2) < 0$ because $a_j((u+v)/2)^j = a_j(-(u+v)/2)^j$, $j = 0$, 2, 4 and $0 < a_j((u+v)/2)^j = -a_j(-(u+v)/2)^j$, $j = 1$, 3. As $P_4(0) = a_0 > 0$, there are two negative roots $\xi < -(u+v)/2 < \eta$ as well.

Definition 2. We define *the standard $\mathbb{Z}_2 \times \mathbb{Z}_2$-action* on couples of the form (sign pattern, admissible pair) by its two generators. Denote by $\sigma(j)$ the jth component of the sign pattern σ. The first of the generators replaces the sign pattern σ by σ^r, where σ^r stands for the *reverted* (i.e. read from the back) sign pattern multiplied by $\sigma(1)$, and keeps the same pair (pos, neg). This generator corresponds to the fact that the polynomials $P(x)$ and $x^dP(1/x)/P(0)$ are both

monic and have the same numbers of positive and negative roots. The second generator exchanges *pos* with *neg* and changes the signs of σ corresponding to the monomials of odd (resp. even) powers if d is even (resp. odd); the rest of the signs are preserved. We denote the new sign pattern by σ_m. This generator corresponds to the fact that the roots of the polynomials (both monic) $P(x)$ and $(-1)^d P(-x)$ are mutually opposite, and if σ is the sign pattern of P, then σ_m is the one of $(-1)^d P(-x)$.

Remark 2. For a given sign pattern σ and an admissible pair (pos, neg), the couples $(\sigma, (pos, neg))$, $(\sigma^r, (pos, neg))$, $(\sigma_m, (neg, pos))$ and $((\sigma_m)^r, (neg, pos))$ are simultaneously realizable or not. One has $(\sigma_m)^r = (\sigma^r)_m$.

Modulo the standard $\mathbb{Z}_2 \times \mathbb{Z}_2$-action Grabiner's example is the only nonrealisable couple (sign pattern, admissible pair) for $d = 4$. All cases of couples (sign pattern, admissible pair) for $d = 5$ and 6 which are not realizable are described in [1]. For $d = 7$, this is done in [5] and for $d = 8$ in [5] and [11]. For $d = 5$, there is a single nonrealizable case (up to the $\mathbb{Z}_2 \times \mathbb{Z}_2$-action). The sign pattern is $(+, +, -, +, -, -,)$ and the admissible pair is $(3, 0)$. For $n = 6$, 7 and 8 there are respectively 4, 6, and 19 nonrealizable cases. In all of them one of the numbers *pos* or *neg* is 0. In the present paper we show that for $d = 9$ this is not so.

Notation 1. For $d = 9$, we denote by σ^0 the following sign pattern (we give on the first and third lines below respectively the sign patterns σ^0 and σ^0_m while the line in the middle indicates the positions of the monomials of odd powers):

$$\begin{array}{lcccccccccccc} \sigma^0 & = & (& + & - & - & - & - & + & + & + & + & - &) \\ & & & 9 & & 7 & & 5 & & 3 & & 1 & & \\ \sigma^0_m & = & (& + & + & - & + & - & - & + & - & + & + &) \end{array}$$

In a sense σ^0 is centre-antisymmetric – it consists of one plus, four minuses, four pluses and one minus.

Theorem 1. (1) *The sign pattern σ^0 is not realizable with the admissible pair $(1, 6)$.*

(2) *Modulo the standard $\mathbb{Z}_2 \times \mathbb{Z}_2$-action, for $d \leq 9$, this is the only nonrealizable couple (sign pattern, admissible pair) in which both components of the admissible pair are nonzero.*

Remark 3. It is shown in [10] that for $d = 11$, the admissible pair $(1, 8)$ is not realizable with the sign pattern $(+ - - - - - + + + + + -)$. Hence Theorem 1 shows an example of a nonrealisable couple, with both components of the admissible pair different from zero, in the least possible degree (namely, 9).

Section 2 contains comments concerning the above result and realizability of sign patterns and admissible pairs in general. Section 3 contains some technical lemmas which allow to simplify the proof of Theorem 1. The plan of the proof

of part (1) of Theorem 1 is explained in Section 4. The proof results from several lemmas whose proofs can be found in Section 5. The proof of part (2) of Theorem 1 is given in Section 8.

2. COMMENTS

It seems that the problem to classify, for any degree d, all couples (sign pattern, admissible pair) which are not realizable, is quite difficult. This is confirmed by Theorem 1. For the moment, only certain sufficient conditions for realizability or nonrealizability have been formulated:

- in [5] and [13] series of nonrealizable cases were found, for $d \geq 4$, even and for $d \geq 5$, odd respectively;
- in [5] sufficient conditions are given for the nonrealizability of sign patterns with exactly two sign changes.
- in [4] sufficient conditions are given for the realizability and the nonrealizability of sign patterns with exactly two sign changes.

Remark 4. For $d \leq 8$, all couples (sign pattern, admissible pairs) with $pos \geq 1$, $neg \geq 1$, are realizable. That is, in the examples of nonrealizability given in [5] and [13] one has either $pos = 0$ or $neg = 0$, so the question to construct an example of nonrealizability with $pos \neq 0 \neq neg$ was a challenging one.

The result in [5] about sign patterns with exactly two sign changes, consisting of m pluses followed by n minuses followed by q pluses, with $m + n + q = d + 1$, is formulated in terms of the following quantity:

$$\kappa := \frac{d-m-1}{m} \cdot \frac{d-q-1}{q} \ .$$

Lemma 1. *For $\kappa \geq 4$, such a sign pattern is not realizable with the admissible pair $(0, d-2)$. The sign pattern is realizable with any admissible pair of the form $(2, v)$.*

Lemma 1 coincides with Proposition 6 of [5]. One can construct new realizable cases with the help of the following concatenation lemma (see its proof in [5]):

Lemma 2. *Suppose that the monic polynomials P_j of degrees d_j and with sign patterns of the form $(+, \sigma_j)$, $j = 1, 2$ (where σ_j contains the last d_j components of the corresponding sign pattern) realize the pairs (pos_j, neg_j). Then:*

(1) if the last position of σ_1 is $+$, then for any $\varepsilon > 0$ small enough, the polynomial $\varepsilon^{d_2} P_1(x) P_2(x/\varepsilon)$ realizes the sign pattern $(+, \sigma_1, \sigma_2)$ and the pair $(pos_1 + pos_2, neg_1 + neg_2)$;

(2) *if the last position of* σ_1 *is* $-$*, then for any* $\varepsilon > 0$ *small enough, the polynomial* $\varepsilon^{d_2} P_1(x) P_2(x/\varepsilon)$ *realizes the sign pattern* $(+, \sigma_1, -\sigma_2)$ *and the pair* $(pos_1 + pos_2, neg_1 + neg_2)$ *(here* $-\sigma_2$ *is obtained from* σ_2 *by changing each* $+$ *by* $-$ *and vice versa).*

Remark 5. If Lemma 2 were applicable to the case treated in Theorem 1, then this case would be realizable and Theorem 1 would be false. We show here that Lemma 2 is indeed inapplicable. It suffices to check the cases $\deg P_1 \geq 5$, $\deg P_2 \leq 4$ due to the centre-antisymmetry of σ^0 and the possibility to use the $\mathbb{Z}_2 \times \mathbb{Z}_2$-action. In all these cases the sign pattern of the polynomial P_1 has exactly two sign changes (including the first sign $+$, the four minuses that follow and the next between one and four pluses). With the notation from Lemma 1, these cases are $m = 1$, $n = 4$, $q = 1, \ldots, 4$. The respective values of κ are 9, 6, 5 and $9/2$. All of them are > 4. By Descartes' rule the polynomial P_1 can have either 0 or 2 positive roots. In the case of 2 positive roots, Lemma 2 implies that its concatenation with P_2 has at least 2 positive roots which is a contradiction. Hence P_1 has no positive roots. The polynomials P_1 and P_2 define sign patterns with $3+q-1$ and $4-q$ sign preservations respectively. The polynomial P_1 has $\leq 1+(q-1)$ negative roots (see Lemma 1) and P_2 has $\leq 4-q$ ones. Therefore the concatenation of P_1 and P_2 has ≤ 6 negative roots and a polynomial realizing the couple $(\sigma^0, (1,6))$ (if any) could not be represented as a concatenation of P_1 and P_2. This, of course, does not a priori mean that such a polynomial does not exist.

3. PRELIMINARIES

Notation 2. By S we denote the set of tuples $a \in \mathbb{R}^9$ for which the polynomial $P(x, a) = x^9 + a_8 x^8 + \cdots + a_0$ realizes the pair $(1, 6)$ and the signs of its coefficients define the sign pattern σ^0. We denote by T the subset of S for which $a_8 = -1$. The notation $\bar{S}$ and $\bar{T}$ stands for the closures of the sets S and T.

By writing $a \in S$ (resp. $a \in T$) we mean that the coefficient vector a of the polynomial $P(x, a)$ (excluding the coefficient of x^9) is in S (resp. in T).

For a polynomial $P \in S$, the conditions $a_9 = 1$, $a_8 = -1$ can be obtained by rescaling the variable x and by multiplying P by a nonzero constant (a_9 is the leading coefficient of P).

Lemma 3. *For* $a \in \bar{S}$*, one has* $a_j \neq 0$ *for* $j = 7, 6, 3, 2$*, and one does not have* $a_4 = 0$ *and* $a_5 = 0$ *simultaneously.*

Proof of Lemma 3. For $a_j = 0$ (where j is one of the indices $7, 6, 3, 2$) there are less than 6 sign changes in the sign pattern σ^0_m. Descartes' rule of signs implies that the polynomial $P(., a)$ has less than 6 negative roots counted with multiplicity. The same is true for $a_5 = a_4 = 0$. □

Lemma 4. *For $a \in \bar{S}$, one has $a_0 \neq 0$.*

Remark 6. A priori the set $\bar{S}$ can contain polynomials with all roots real and nonzero. The positive ones can be either a triple root or a double and a simple roots (but not three simple roots). If $a \in S$, then $P(x, a)$ has the maximal possible number of negative roots (equal to the number of sign preservations in the sign pattern). If $a^{'} \in \bar{S}$, then the polynomial $Q(x, a^{'})$ is the limit of polynomials $Q(x, a)$ with $a \in S$. In the limit as a $\to a^{'}$, the complex conjugate pair can become a double positive, but not a double negative root, because there are no 8 sign preservations in the sign pattern.

Proof of Lemma 4. In the proof we consider the two cases $a_0 = 0 \neq a_1$ and $a_0 = a_1 = 0$, and for each of them the three possibilities $a_4 \neq 0 \neq a_5$, $a_4 = 0 \neq a_5$ and $a_4 \neq 0 = a_5$, see Lemma 3.

Suppose that for $P \in \bar{S}$, one has $a_0 = 0$ and for $j \neq 0$, $a_j \neq 0$. Hence the polynomial $P_1 := P/x$ has 6 negative roots and either 0 or 2 positive roots. We show that 0 positive roots is impossible. Indeed, the polynomial P_1 defines a sign pattern with exactly 2 sign changes. Suppose that all negative roots are distinct. If P_1 has no positive roots, then one can apply Lemma 1, according to which, as one has $\kappa = 9/2 > 4$, such a polynomial does not exist. If P_1 has a negative root $-b$ of multiplicity $m > 1$, then its perturbation

$$P_{1,\epsilon} := (x + b + \epsilon)P_1/(x + b) \ , \ 0 < \epsilon \ll 1 \, ,$$

defines the same sign pattern and instead of the root $-b$ of multiplicity m has a root $-b$ of multiplicity $m - 1$ and a simple root $-b - \epsilon$. After finitely many such perturbations, one is in the case when all negative roots are distinct, which leads to a contradiction as above.

If P_1 has 2 positive roots, then this is a double positive root g, see Remark 6. In this case, we add to P_1 a linear term $\pm\epsilon x$ (with ϵ small enough in order not to change the sign pattern) to make the double root bifurcate into a complex conjugate pair. The sign is chosen depending on whether P_1 has a minimum or a maximum at g. After this, if there are multiple negative roots, we apply perturbations of the form $P_{1,\epsilon}$ to arrive again at a contradiction.

Suppose that $a_1 = a_0 = 0$, and that for $j \geq 2$, $a_j \neq 0$. Then one considers the polynomial $P_2 := P/x^2$. It defines a sign pattern with two sign changes and one has $\kappa = 5 > 4$. Hence it has 2 positive roots, otherwise one obtains a contradiction with Lemma 1.

Suppose now that exactly one of the coefficients a_4 or a_5 is 0. We assume this to be a_4, for a_5 the reasoning is similar. Suppose also that either $a_1 \neq 0$, $a_0 = 0$ or $a_1 = a_0 = 0$, and that for $j \geq 2$, $j \neq 4$, one has $a_j \neq 0$. We treat in detail the case $a_1 \neq 0$, $a_0 = 0$, the case $a_1 = a_0 = 0$ is treated by analogy. We first make the double positive root if any bifurcate into a complex conjugate pair as above. This does not change the coefficient a_4. After this instead of perturbations

$P_{1,\epsilon}$ we use perturbations preserving the condition $a_4 = 0$. Suppose that $P_1 = (x+b)^m Q_1 Q_2$, where Q_1 and Q_2 are monic polynomials, $\deg Q_2 = 2$, Q_2 having a complex conjugate pair of roots, Q_1 having $6-m$ negative roots counted with multiplicity. Then we set:

$$P_1 \mapsto P_1 + \epsilon(x+b)^{m-1}(x+h_1)(x+h_2)Q_1\,,$$

where the real numbers h_i are distinct, different from any of the roots of P and chosen in such a way that the coefficient δ of x^3 of P_1 is 0. Such a choice is possible, because all coefficients of the polynomial $(x+b)^{m-1}Q_1$ are positive, hence δ is of the form $A+(h_1+h_2)B+Ch_1h_2$, where $A>0$, $B>0$ and $C>0$. The result of the perturbation is a polynomial P_1 having six negative distinct roots and a complex conjugate pair; its coefficient of x^3 is 0. By adding a small positive number to this coefficient, one obtains a polynomial P_1 with roots as before and defining the sign pattern $(+----++++)$. For this polynomial one has $\kappa = 9/2 > 4$ which contradicts Lemma 1.

In the case $a_1 = a_0 = 0$, the polynomial P_1 thus obtained has five negative distinct roots, a complex conjugate pair of roots and a root at 0. One adds small positive numbers to its constant term and to its coefficient of x^3 and one proves in the same way that such a polynomial does not exist. □

Remark 7. One deduces from Lemmas 3 and 4 that for a polynomial in $\bar{T}$ exactly one of the following conditions holds true:

(1) all its coefficients are nonvanishing;

(2) exactly one of them is vanishing and this coefficient is either a_1 or a_4 or a_5;

(3) exactly two of them are vanishing, and these are either a_1 and a_4 or a_1 and a_5.

Lemma 5. *There exists no real degree 9 polynomial satisfying the following conditions:*

- *the signs of its coefficients define the sign pattern σ^0,*
- *it has a complex conjugate pair of roots with nonpositive real part,*
- *it has a single positive root,*
- *it has negative roots of total multiplicity 6.*

Proof. Suppose that such a monic polynomial P exists. We can write P in the form $P = P_1P_2P_3$, where $\deg P_1 = 6$.

All roots of P_1 are negative hence $P_1 = \sum_{j=0}^{6} \alpha_j x^j$, $\alpha_j > 0$, $\alpha_6 = 1$; $P_2 = x - w$, $w > 0$; $P_3 = x^2 + \beta_1 x + \beta_0$, $\beta_j \geq 0$, $\beta_1^2 - 4\beta_0 < 0$.

By Descartes' rule of signs, the polynomial $P_1P_2 = \sum_{j=0}^{7} \gamma_j x^j$, $\gamma_7 = 1$, has exactly one sign change in the sequence of its coefficients. It is clear that as

$0 > a_8 = \gamma_6 + \beta_1$, and as $\beta_1 \geq 0$, one must have $\gamma_6 < 0$. But then $\gamma_j < 0$ for $j = 0, \ldots, 6$. For $j = 2$, 3 and 4, one has $a_j = \gamma_{j-2} + \beta_1\gamma_{j-1} + \beta_0\gamma_j < 0$ which means that the signs of a_j do not define the sign pattern σ^0. □

Remark 8. It follows from Lemma 5 that polynomials of $\bar{T}$ can only have negative roots of total multiplicity 6 and positive roots of total multiplicity 1 or 3 (i.e., either one simple, or one simple and one double or one triple positive root); these polynomials have no root at 0 (Lemma 4). Indeed, when approaching the boundary of T, the complex conjugate pair can coalesce to form a double positive (but never nonpositive) root; the latter might eventually coincide with the simple positive root.

4. PLAN OF THE PROOF OF PART (1) OF THEOREM 1

Suppose that there exists a monic polynomial $P(x, a^*)|_{a_8^*=-1}$ with signs of its coefficients defined by the sign pattern σ^0, with 6 distinct negative, a simple positive and two complex conjugate roots.

Then for a close to $a^* \in \mathbb{R}^8$, all polynomials $P(x, a)$ share with $P(x, a^*)$ these properties. Therefore the interior of the set T is nonempty. In what follows we denote by Γ the connected component of T to which a^* belongs. Denote by $-\delta$ the value of a_7 for $a = a^*$ (recall that this value is negative).

Lemma 6. *There exists a compact set $K \subset \bar{\Gamma}$ containing all points of $\bar{\Gamma}$ with $a_7 \in [-\delta, 0)$. Hence there exists $\delta_0 > 0$ such that for every point of $\bar{\Gamma}$, one has $a_7 \leq -\delta_0$, and for at least one point of K and for no point of $\bar{\Gamma}\backslash K$, the equality $a_7 = -\delta_0$ holds.*

Proof. Suppose that there exists an unbounded sequence $\{a^n\}$ of values $a \in \bar{\Gamma}$ with $a_7^n \in [-\delta, 0)$. Hence one can perform rescalings $x \mapsto \beta_n x$, $\beta_n > 0$, such that the largest of the moduli of the coefficients of the monic polynomials $Q_n := (\beta_n)^{-9} P(\beta_n x, a^n)$ equals 1. These polynomials belong to $\bar{S}$, not necessarily to $\bar{T}$ because a_8 after the rescalings, in general, is not equal to -1. The coefficient of x^7 in Q_n equals $a_7^n/(\beta_n)^2$. The sequence $\{a^n\}$ is unbounded, so there exists a subsequence β_{n_k} tending to ∞. This means that the sequence of monic polynomials $Q_{n_k} \in \bar{S}$ with bounded coefficients has a polynomial in $\bar{S}$ with $a_7 = 0$ as one of its limit points which contradicts Lemma 3.

Hence the moduli of the roots and the tuple of coefficients a_j of $P(x, a) \in \bar{\Gamma}$ with $a_7 \in [-\delta, 0)$ remain bounded from which the existence of K and δ_0 follows. □

The above lemma implies the existence of a polynomial $P_0 \in \bar{\Gamma}$ with $a_7 = -\delta_0$. We say that P_0 is *a_7-maximal*. Our aim is to show that no polynomial of $\bar{\Gamma}$ is a_7-maximal which contradiction will be the proof of Theorem 1.

Definition 3. A real univariate polynomial is *hyperbolic* if it has only real (not necessarily simple) roots. We denote by $H \subset \bar{\Gamma}$ the set of hyperbolic polynomials in $\bar{\Gamma}$. Hence these are monic degree 9 polynomials having positive and negative roots of respective total multiplicities 3 and 6 (roots at the origin are impossible by Lemma 4). By $U \subset \bar{\Gamma}$ we denote the set of polynomials in $\bar{\Gamma}$ having a complex conjugate pair, a simple positive root and negative roots of total multiplicity 6. Thus $\bar{\Gamma} = H \cup U$ and $H \cap U = \emptyset$. We denote by U_0, U_2, $U_{2,2}$, U_3 and U_4 the subsets of U for which the polynomial $P \in U$ has respectively 6 simple negative roots, one double and 4 simple negative roots, at least two negative roots of multiplicity ≥ 2, one triple and 3 simple negative roots and a negative root of multiplicity ≥ 4.

The following lemma on hyperbolic polynomials is proved in [10]. It is used in the proofs of the other lemmas.

Lemma 7. *Suppose that V is a hyperbolic polynomial of degree $d \geq 2$ with no root at* 0*. Then:*

(1) *V does not have two or more consecutive vanishing coefficients.*

(2) *If V has a vanishing coefficient, then the signs of its surrounding two coefficients are opposite.*

(3) *The number of positive (of negative) roots of V is equal to the number of sign changes in the sequence of its coefficients (in the one of $V(-x)$).*

By a sequence of lemmas we consecutively decrease the set of possible a_7-*maximal* polynomials until in the end it turns out that this set must be empty. The proofs of the lemmas of this section except Lemma 6 are given in Sections 5 (Lemmas 7 – 12), 6 (Lemma 13) and 7 (Lemmas 14 –16).

Lemma 8. (1) *No polynomial of $U_{2,2} \cup U_4$ is a_7-maximal.*

(2) *For each polynomial of U_3, there exists a polynomial of U_0 with the same values of a_7, a_5, a_4 and a_1.*

(3) *For each polynomial of $U_0 \cup U_2$, there exists a polynomial of $H \cup U_{2,2}$ with the same values of a_7, a_5, a_4 and a_1.*

Lemma 8 implies that if there exists an a_7-maximal polynomial in $\bar{\Gamma}$, then there exists such a polynomial in H. So from now on, we aim at proving that H contains no such polynomial hence H and $\bar{\Gamma}$ are empty.

Lemma 9. *There exists no polynomial in H having exactly two distinct real roots.*

Lemma 10. *The set H contains no polynomial having one triple positive root and negative roots of total multiplicity* 6*.*

Lemma 10 and Remark 6 imply that a polynomial in H (if any) satisfies the following condition:

Condition A. Any polynomial $P \in H$ has a double and a simple positive roots and negative roots of total multiplicity 6.

Lemma 11. *There exists no polynomial $P \in H$ having exactly three distinct real roots and satisfying the conditions $\{a_1 = 0, a_4 = 0\}$ or $\{a_1 = 0, a_5 = 0\}$.*

It follows from Lemma 11 and Lemma 3 that a polynomial $P \in H$ having exactly three distinct real roots (hence a double and a simple positive and an 6-fold negative one) can satisfy at most one of the conditions $a_1 = 0$, $a_4 = 0$ and $a_5 = 0$.

Lemma 12. *No polynomial in H having exactly three distinct real roots is a_7-maximal.*

Thus an a_7-maximal polynomial in H (if any) must satisfy Condition A and have at least four distinct real roots.

Lemma 13. *The set H contains no polynomial having a double and a simple positive roots and exactly two distinct negative roots of total multiplicity 6, and which satisfies either the conditions $\{a_1 = a_4 = 0\}$ or $\{a_1 = a_5 = 0\}$.*

At this point we know that an a_7-maximal polynomial of H satisfies Condition A and one of the two following conditions:

Condition B. It has exactly four distinct real roots and satisfies exactly one or none of the equalities $a_1 = 0$, $a_4 = 0$ or $a_5 = 0$.

Condition C. It has at least five distinct real roots.

Lemma 14. *The set H contains no a_7-maximal polynomial satisfying Conditions A and B.*

Therefore an a_7-maximal polynomial in H (if any) must satisfy Conditions A and C.

Lemma 15. *The set H contains no a_7-maximal polynomial having exactly five distinct real roots.*

Lemma 16. *The set H contains no a_7-maximal polynomial having at least six distinct real roots.*

Hence the set H contains no a_7-maximal polynomial at all. It follows from Lemma 8 that there is no such polynomial in $\bar{\Gamma}$. Hence $\bar{\Gamma} = \emptyset$.

5. PROOFS OF LEMMAS 7, 8, 9, 10, 11 AND 12

Proof of Lemma 7. Part (1). Suppose that a hyperbolic polynomial V with two or more vanishing coefficients exists. If V is degree d hyperbolic, then $V^{(k)}$ is also hyperbolic for $1 \leq k < d$. Therefore we can assume that V is of the form $x^\ell L + c$, where $\deg L = d - \ell$, $\ell \geq 3$, $L(0) \neq 0$ and $c = V(0) \neq 0$. If V is hyperbolic and $V(0) \neq 0$, then such is also $W := x^d V(1/x) = cx^d + x^{d-\ell} L(1/x)$ and also $W^{(d-\ell)}$

which is of the form $ax^\ell + b$, $a \neq 0 \neq b$. However given that $\ell \geq 3$, this polynomial is not hyperbolic.

For the proof of part (2) we use exactly the same reasoning, but with $\ell = 2$. The polynomial $ax^2 + b$, $a \neq 0 \neq b$, is hyperbolic if and only if $ab < 0$.

To prove part (3) we consider the sequence of coefficients of $V := \sum_{j=0}^{d} v_j x^j$, $v_0 \neq 0 \neq v_d$. Set $\Phi := \sharp\{k | v_k \neq 0 \neq v_{k-1}, v_k v_{k-1} < 0\}$, $\Psi := \sharp\{k | v_k \neq 0 \neq v_{k-1}, v_k v_{k-1} > 0\}$ and $\Lambda := \sharp\{k | v_k = 0\}$. Then $\Phi + \Psi + 2\Lambda = d$. By Descartes' rule of signs the number of positive (of negative) roots of V is $pos_V \leq \Phi + \Lambda$ (resp. $neg_V \leq \Psi + \Lambda$). As $pos_V + neg_V = d$, one must have $pos_V = \Phi + \Lambda$ and $neg_V = \Psi + \Lambda$. It remains to notice that $\Phi + \Lambda$ is the number of sign changes in the sequence of coefficients of V (and $\Psi + \Lambda$ equals the number of sign changes in the sequence of coefficients of $V(-x)$), see part (2) of the lemma. □

Proof of Lemma 8. Part (1). A polynomial of $U_{2,2}$ or U_4 respectively is representable in the form:

$$P^\dagger := (x+u)^2(x+v)^2 S\Delta \quad \text{and} \quad P^* := (x+u)^4 S\Delta\,,$$

where $\Delta := (x^2 - \xi x + \eta)(x - w)$ and $S := x^2 + Ax + B$. All coefficients u, v, w, ξ, η, A, B are positive and $\xi^2 - 4\eta < 0$ (see Lemma 5); for A and B this follows from the fact that all roots of $P^\dagger/\Delta$ and P^*/Δ are negative. (The roots of $x^2 + Ax + B$ are not necessarily different from $-u$ and $-v$.) We consider the two Jacobian matrices

$$J_1 := (\partial(a_8, a_7, a_1, a_4)/\partial(\xi, \eta, w, u)) \quad \text{and} \quad J_2 := (\partial(a_8, a_7, a_1, a_5)/\partial(\xi, \eta, w, u))\,.$$

In the case of $P^\dagger$ their determinants equal

$$\begin{aligned} \det J_1 &= (A^2u^2v + 2A^2uv^2 + 2Au^2v^2 + Auv^3 + 2ABu^2 + 5ABuv \\ &\quad + 2ABv^2 + 3Bu^2v + 2Buv^2 + Bv^3 + 2B^2u + B^2v)\Pi\,, \\ \det J_2 &= (A^2uv + Au^2v + 2Auv^2 + 2ABu \\ &\quad + ABv + 2Bu^2 + 4Buv + 2Bv^2)\Pi\,, \end{aligned}$$

where $\Pi := -2v(w+u)(-\eta - w^2 + w\xi)(\xi u + \eta + u^2)$.

These determinants are nonzero. Indeed, each of the factors is either a sum of positive terms or equals $-\eta - w^2 + w\xi < -\xi^2/4 - w^2 + w\xi = -(\xi/2 - w)^2 \leq 0$. Thus one can choose values of (ξ, η, w, v) close to the initial one (u, A and B remain fixed) to obtain any values of (a_8, a_7, a_1, a_4) or (a_8, a_7, a_1, a_5) close to the initial one. In particular, with $a_8 = -1$, $a_1 = a_4 = 0$ or $a_8 = -1$, $a_1 = a_5 = 0$ while a_7 can have values larger than the initial one. Hence this is not an a_7-maximal polynomial. (If the change of the value of (ξ, η, w, v) is small enough, the values of the coefficients a_j, $j = 0$, 2, 3, 5 or 4 and 6 can change, but their signs remain the same.) The same reasoning is valid for P^* as well in which case one has

$$\begin{aligned} \det J_1 &= (3A^2u^2 + 3Au^3 + 9ABu + 6Bu^2 + 3B^2)M\,, \\ \det J_2 &= (A^2u + 3Au^2 + 3AB + 8Bu)M\,, \end{aligned}$$

with $M := -4u^2(w+u)(-\eta - w^2 + w\xi)(\xi u + \eta + u^2)$.

To prove part (2), we observe that if the triple root of $P \in U_3$ is at $-u < 0$, then in case when P is increasing (resp. decreasing) in a neighbourhood of $-u$ the polynomial $P - \varepsilon x^2(x+u)$ (resp. $P + \varepsilon x^2(x+u)$), where $\varepsilon > 0$ is small enough, has three simple roots close to $-u$; it belongs to $\bar{\Gamma}$, its coefficients a_j, $2 \neq j \neq 3$, are the same as the ones of P, the signs of a_2 and a_3 are also the same.

For the proof of part (3), we observe first that 1) for $x < 0$ the polynomial P has three maxima and three minima and 2) for $x > 0$ one of the following three things holds true: either $P' > 0$, or there is a double positive root γ of P', or P' has two positive roots $\gamma_1 < \gamma_2$ (they are both either smaller than or greater than the positive root of P). Suppose first that $P \in U_0$. Consider the family of polynomials $P - t$, $t \geq 0$. Denote by t_0 the smallest value of t for which one of the three things happens: either $P - t$ has a double negative root v (hence a local maximum), or $P - t$ has a triple positive root γ or $P - t$ has a double and a simple positive roots (the double one is at γ_1 or γ_2). In the second and third cases one has $P - t_0 \in H$. In the first case, if $P - t_0$ has another double negative root, then $P - t_0 \in U_{2,2}$ and we are done. If not, then consider the family of polynomials

$$P_s := P - t_0 - s(x^2 - v^2)^2(x^2 + v^2) = P - t_0 - s(x^6 - v^2x^4 - x^2v^4 + v^6)\,, \quad s \geq 0\,.$$

The polynomial $-(x^6 - v^2x^4 - x^2v^4 + v^6)$ has double real roots at $\pm v$ and a complex conjugate pair. It has the same signs of the coefficients of x^6, x^4 and 1 as $P - t_0$ and P. The rest of the coefficients of $P - t_0$ and P_s are the same. As s increases, the value of P_s for every $x \neq \pm v$ decreases. So for some $s = s_0 > 0$ for the first time one has either $P_s \in U_{2,2}$ (another local maximum of P_s becomes a double negative root) or $P_s \in H$ (P_s has positive roots of total multiplicity 3, but not three simple ones). This proves part (3) for $P \in U_0$.

If $P \in U_2$ and the double negative root is a local minimum, then the proof of part (3) is just the same. If this is a local maximum, then one skips the construction of the family $P - t$ and starts constructing the family P_s directly. □

Proof of Lemma 9. Suppose that such a polynomial exists. Then it must be of the form $P := (x+u)^6(x-w)^3$, $u > 0$, $w > 0$. The conditions $a_8 = -1$ and $a_1 > 0$ read:

$$6u - 3w = -1 \quad \text{and} \quad 3u^5w^2(u - 2w) > 0\,.$$

In the plane of the variables (u, w) the domain $\{u > 0, w > 0, u - 2w > 0\}$ does not intersect the line $6u - 3w = -1$ which proves the lemma. □

Proof of Lemma 10. Represent the polynomial in the form $P = (x+u_1)\cdots(x+u_6)(x-\xi)^3$, where $u_j > 0$ and $\xi > 0$. The numbers u_j are not necessarily distinct. The coefficient a_8 then equals $u_1 + \cdots + u_6 - 3\xi$. The condition $a_8 = -1$ implies $\xi = \xi_* := (u_1 + \cdots + u_6 + 1)/3$. Thus

$$P(x) = (x+u_1)\cdots(x+u_6)\Big(x - \frac{u_1 + \cdots + u_6 + 1}{3}\Big)^3$$

and for the coefficient a_1 we have

$$27a_1 = (u_1 + \cdots + u_6 + 1)^2\, u_1 u_2 \cdots u_6 \Big(3 - (u_1 + \cdots + u_6 + 1) \sum_{j=1}^{6} \frac{1}{u_j}\Big)\,.$$

The last factor in this representation is negative, hence $a_1 < 0$, a contradiction. □

Proof of Lemma 11. Suppose that such a polynomial exists. Then it must be of the form $(x+u)^6(x-w)^2(x-\xi)$, where $u > 0$, $w > 0$, $\xi > 0$, $w \neq \xi$. One checks numerically (say, using MAPLE), for each of the two systems of algebraic equations $a_8 = -1$, $a_1 = 0$, $a_4 = 0$ and $a_8 = -1$, $a_1 = 0$, $a_5 = 0$, that each real solution (u, w, ξ) or (u, v, w) contains a nonpositive component. □

Proof of Lemma 12. Making use of Condition A formulated after Lemma 10, we consider only polynomials of the form $(x+u)^6(x-w)^2(x-\xi)$, where u, w, ξ are positive and $w \neq \xi$. Consider the Jacobian matrix

$$J_1^* := (\partial(a_8, a_7, a_1)/\partial(u, w, \xi))\,.$$

Its determinant equals $-12u^4(u+w)(u-5w)(w-\xi)(u+\xi)$. All factors except $u - 5w$ are nonzero. Thus for $u \neq 5w$, one has $\det J_1 \neq 0$, so one can fix the values of a_8 and a_1 and vary the one of a_7 arbitrarily close to the initial one by choosing suitable values of u, w and ξ. Hence the polynomial is not a_7-maximal. For $u = 5w$, one has $a_3 = -2500w^5(\xi + 5w) < 0$ which is impossible. Hence there exist no a_7-maximal polynomials which satisfy only the condition $a_1 = 0$ or none of the conditions $a_1 = 0$, $a_4 = 0$ or $a_5 = 0$. To see that there exist no such polynomials satisfying only the condition $a_4 = 0$ or $a_5 = 0$ one can consider the matrices $J_4^* := (\partial(a_8, a_7, a_4)/\partial(u, w, \xi))$ and $J_5^* := (\partial(a_8, a_7, a_5)/\partial(u, w, \xi))$. Their determinants equal respectively

$$-60u(u+w)(2u-w)(\xi-w)(\xi+u) \quad \text{and} \quad -12u(u+w)(5u-w)(\xi-w)(\xi+u)\,.$$

They are nonzero respectively for $2u \neq w$ and $5u \neq w$, in which cases in the same way we conclude that the polynomial is not w_7-maximal. If $u = w/2$, then $a_1 = -(1/64)w^7(10\xi - w)$ and $a_8 = w - \xi$. As $a_1 > 0$ and $a_8 < 0$, one has $w > 10\xi$ and $\xi > w > 10\xi$ which is a contradiction. If $w = 5u$, then $a_6 = 20u^2(u+\xi) > 0$ which is again a contradiction. □

6. PROOF OF LEMMA 13

The multiplicities of the negative roots of P define the following a priori possible cases:

$$\text{A)} \quad (5,1)\,, \quad \text{B)} \quad (4,2) \quad \text{and} \quad \text{C)} \quad (3,3)\,.$$

In all of them the proof is carried out simultaneously for the two possibilities $\{a_1 = a_4 = 0\}$ and $\{a_1 = a_5 = 0\}$. In order to simplify the proof we fix one of the roots to be equal to -1 (this can be achieved by a change $x \mapsto \beta x$, $\beta > 0$, followed by $P \mapsto \beta^{-9} P$). This allows to deal with one less parameter. By doing so we may no longer require that $a_8 = -1$, but only that $a_8 < 0$.

Case A) We use the following parametrization:

$$P = (x+1)^5 (s\,x+1)(t\,x-1)^2 (w\,x-1)\,,\ s>0\,,\ t>0\,,\ w>0\,,\ t\neq w\,,$$

i.e. the negative roots of P are at -1 and $-1/s$ and the positive ones at $1/t$ and $1/w$.

The condition $a_1 = w + 2t - s - 5 = 0$ yields $s = w + 2t - 5$. With this s one has

$$a_3 = a_{32} w^2 + a_{31} w + a_{30}\,, \quad a_4 = a_{42} w^2 + a_{41} w + a_{40}\,, \quad \text{where}$$

$$a_{32} = -2t + 5\,, \quad a_{31} = -(2t-5)^2\,, \quad a_{30} = -2t^3 + 20t^2 - 50t + 40\,,$$

$$a_{42} = t^2 - 10t + 10\,, \quad a_{41} = 2t^3 - 25t^2 + 70t - 50\,, \quad a_{40} = -10t^3 + 55t^2 - 100t + 45\,.$$

The coefficient a_{30} has a single real root $6.7245\ldots$ hence $a_{30} < 0$ for $t > 6.7245\ldots$. On the other hand, for $t > 6.7245\ldots$,

$$a_{32} w^2 + a_{31} w = w(-2t+5)(w + 2t - 5) = w(-2t+5)s < 0\,.$$

Thus the inequality $a_3 > 0$ fails for $t > 6.7245\ldots$. Observing that $a_{41} = (2t-5)a_{42}$ one can write

$$a_4 = (w + 2t - 5) w\, a_{42} + a_{40} = s\, w\, a_{42} + a_{40}\,.$$

The real roots of a_{42} (resp. a_{40}) equal $1.127\ldots$ and $8.872\ldots$ (resp. $0.662\ldots$). Hence for $t \in [1.127\ldots, 8.872\ldots]$, the inequality $a_4 > 0$ fails. There remains to consider the possibility $t \in (0, 1.127\ldots)$.

It is to be checked directly that for $s = w + 2t - 5$, one has

$$a_8/t = 10t^2 w + 5t\, w^2 - 2t^2 - 29t\, w - 2w^2 + 5t + 10w = (5t-2) w\, s + t(5-2t)\,,$$

which is nonnegative (hence $a_8 < 0$ fails) for $t \in [2/5, 5/2]$. Similarly

$$a_6 = a_6^* w(w + 2t - 5) + a_6^\dagger = a_6^* w\, s + a_6^\dagger\,, \quad \text{where}$$

$$a_6^* = 10t^2 - 20t + 5\,, \quad a_6^\dagger = -5(t-1)(4t^2 - 9t + 1)\,.$$

The real roots of a_6^* (resp. $a_6^\dagger$) equal $1.707\ldots > 2/5 = 0.4$ and $0.293\ldots$ (resp. $1 > 2/5$, $0.117\ldots$ and $2.133\ldots$) hence for $t \in (0, 2/5)$ one has $a_6^* > 0$ and $a_6^\dagger > 0$, i.e. $a_6 > 0$ and the equality $a_6 = 0$ or the inequality $a_6 < 0$ is impossible. □

Case B) We parametrize P as follows:

$$P = (x+1)^4 (T x^2 + S x - 1)^2 (w\,x - 1)\,,\ T>0\,,\ w>0\,.$$

Here we presume S to be real, not necessarily positive. The factor $(Tx^2+Sx-1)^2$ contains the double positive and negative roots of P.

From $a_1 = w + 2S - 4 = 0$ one finds $S = (4-w)/2$. With this S one has

$$a_8/T = (4w-1)T + 4w - w^2\,, \quad a_5 = a_{52}T^2 + a_{51}T + a_{50}\,, \quad \text{where}$$

$$a_{52} = w - 4\,, \quad a_{51} = -4w^2 + 10w - 16\,, \quad a_{50} = (3/2)w^3 - 9w^2 + 16w - 12\,.$$

Suppose first that $w > 1/4$. The inequality $a_8 < 0$ is equivalent to

$$T < T_0 := (w^2 - 4w)/(4w-1)\,.$$

As $T > 0$, this implies $w > 4$.

For $T = T_0$, one obtains $a_5 = 3C/2(4w-1)^2$, where the numerator $C := 6w^5 - 40w^4 + 85w^3 - 54w^2 + 32w - 8$ has a single real root $0.368\ldots$. Hence for $w > 4$, one has $C > 0$ and $a_5|_{T=T_0} > 0$. On the other hand, $a_{50} = a_5|_{T=0}$ has a single real root $3.703\ldots$, so for $w > 4$ one has $a_5|_{T=0} > 0$. For $w > 4$ fixed, and for $T \in [0, T_0]$, the value of the derivative

$$\partial a_5/\partial T = (2w-8)T - 4w^2 + 10w - 16$$

is maximal for $T = T_0$; this value equals

$$-2(7w^3 - 14w^2 + 21w - 8)/(4w-1)\,,$$

which is negative because the only real root of the numerator is $0.510\ldots$. Thus $\partial a_5/\partial T < 0$ and a_5 is minimal for $T = T_0$. Hence the inequality $a_5 < 0$ fails for $w > 1/4$. For $w = 1/4$ one has $a_8 = 15/16 > 0$.

So suppose that $w \in (0, 1/4)$. In this case the condition $a_8 < 0$ implies $T > T_0$. For $T = T_0$ one gets

$$a_4 = 3D/2(4w-1)^2\,, \quad \text{where} \quad D := 8w^5 - 32w^4 + 54w^3 - 85w^2 + 40w - 6$$

has a single real root $2.719\ldots$. Therefore for $w \in (0, 1/4)$ one has $D < 0$ and $a_4|_{T=T_0} < 0$. The derivative $\partial a_4/\partial T = -w^2 - 2T - 4$ being negative one has $a_4 < 0$ for $w \in (0, 1/4)$, i.e. the inequality $a_4 > 0$ fails. □

Case C) We set

$$P := (x+1)^3(sx+1)^3(tx-1)^2(wx-1)\ ,\ s > 0\ ,\ t > 0\ ,\ w > 0\ ,\ t \neq w\,.$$

The condition $a_1 = w + 2t - 3s - 3 = 0$ implies $s = s_0 := (w + 2t - 3)/3$. For $s = s_0$, one has $27a_8 = t(w + 2t - 3)^2 H^*$, where

$$H^* := 6wt^2 - 2t^2 + 3w^2t - 5wt + 3t + 6w - 2w^2\,. \tag{6.1}$$

We show first that for $s = s_0$, the case $a_1 = a_5 = 0$ is impossible. To fix the ideas, we represent in Figure 1 the sets $\{H^* = 0\}$ (solid curve) and $\{a_5^* = 0\}$ (dashed

curve), where $a_5^* := a_5|_{s=s_0}$. Although we need only the nonnegative values of t and w, we show these curves also for the negative values of the variables to make things more clear. (The lines $t = 2/3$ and $w = 1/3$ are asymptotic lines for the set $\{H^* = 0\}$). For $t \geq 0$ and $w \geq 0$, the only point, where $H^* = a_5^* = 0$, is the point $(0; 3)$. However, at this point one has $a_8 = 0$, i.e. this does not correspond to the required sign pattern.

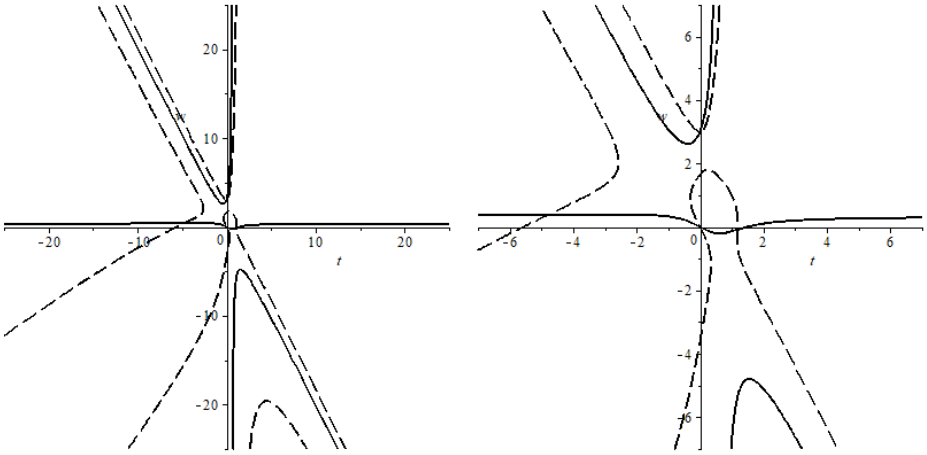


Figure 1: The sets $\{H^* = 0\}$ (solid curve) and $\{a_5^* = 0\}$ (dashed curve), with 3 and 4 connected components respectively.

Lemma 17. (1) *For $(t, w) \in \Omega_1 \cup \Omega_2$, where $\Omega_1 = [3/2, \infty) \times [1/3, \infty)$ and $\Omega_2 = [0, 3/2] \times [0, 3]$, one has $H^* \geq 0$.*

(2) *For $(t, w) \in \Omega_3 := [3/2, \infty) \times [0, 1/3]$, one has $a_5^* < 0$.*

(3) *For $(t, w) \in \Omega_4 := [0, 3/2] \times [3, \infty)$, the two conditions $H^* < 0$ and $a_5^* = 0$ do not hold simultaneously.*

Lemma 17 (which is proved after the proof of Lemma 12) implies that in each of the sets Ω_j, $1 \leq j \leq 4$, at least one of the two conditions $H^* < 0$ (i. e. $a_8 < 0$) and $a_5^* = 0$ fails. There remains to notice that $\Omega_1 \cup \Omega_2 \cup \Omega_3 \cup \Omega_4 = \{t \geq 0, w \geq 0\}$.

Now, we show that for $s = s_0$, the case $a_1 = a_4 = 0$ is impossible. In Figure 2 we show the sets $\{H^* = 0\}$ (solid curve) and $\{a_4^* = 0\}$ (dashed curve), where $a_4^* := a_4|_{s=s_0}$. We use the notation introduced in Lemma 17. By part (1) of Lemma 17 the case $a_1 = a_4 = 0$ is impossible for $(t, w) \in \Omega_1 \cup \Omega_2$.

Lemma 18. (1) *For $(t, w) \in \Omega_3$, one has $a_4^* > 0$.*

(2) *For $(t, w) \in \Omega_4$, the two conditions $H^* < 0$ and $a_4^* = 0$ do not hold simultaneously.*

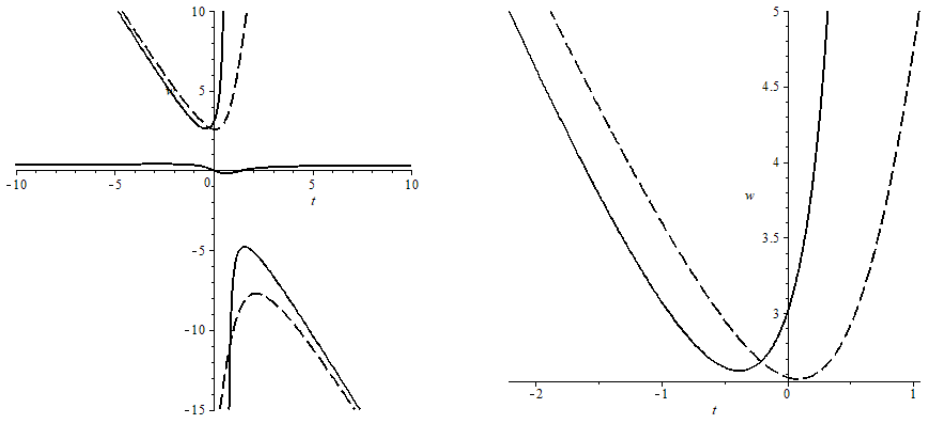


Figure 2: The sets $\{H^* = 0\}$ (solid curve) and $\{a_4^* = 0\}$ (dashed curve), with 3 and 2 connected components respectively.

Thus the couple of conditions $H^* < 0$, $a_4^* = 0$ fails for $t \geq 0$, $w \geq 0$. This proves Lemma 13. Lemma 18 is proved after Lemma 17 . □

Proof of Lemma 17. Part (1). Consider the quantity H^* as a polynomial in the variable w:

$$H^* = b_2 w^2 + b_1 w + b_0 \,,$$

where

$$b_2 = 3t - 2\,, \quad b_1 = 6t^2 - 5t + 6\,, \quad b_0 = -2t(t - 3/2)\,.$$

Its discriminant $\Delta_w := b_1^2 - 4b_0 b_2 = 9(2t^2 - 3t + 2)(2t^2 + t + 2)$ is positive for any real t. This is why for $t \neq 2/3$, the polynomial H^* has 2 real roots; for $t = 2/3$, it is a linear polynomial in w and has a single real root $-5/24$. When H^* is considered as a polynomial in the variable t, one sets

$$\begin{aligned} &H^* := c_2 t^2 + c_1 t + c_0\,, \quad \text{where} \\ &c_2 = 6w - 2\,, \quad c_1 = 3w^2 - 5w + 3\,, \quad c_0 = -2w(w - 3)\,. \end{aligned} \tag{6.2}$$

Its discriminant

$$\Delta_t := c_1^2 - 4c_0 c_2 = 9(w^2 + 5w + 1)(w^2 - 3w + 1)$$

is negative if and only if $w \in (-4.79\ldots, 0.20\ldots) \cup (-0.38\ldots, 2,61\ldots)$. One checks directly that $H^*|_{w=1/3} = (5/3)t + 16/9$ which is positive for $t \geq 0$. Next, one has $H^*|_{w=0} = b_0$ which is negative for $t > 3/2$. Finally, for $t > 3/2$, the ratio b_0/b_2 is negative which means that for $t > 3/2$ fixed, the polynomial H^* has one positive and one negative root, so the positive root belongs to the interval $(0, 1/3)$ (because

$H^*|_{w=1/3} > 0$). Hence $H^* \geq 0$ for $(t,w) \in \Omega_1$ and $H^* > 0$ for (t,w) in the interior of Ω_1.

Suppose now that $(t,w) \in [0,3/2] \times [0,3]$. For $t \in (2/3, 3/2]$ fixed, one has $b_2 > 0$, $b_1/b_2 > 0$ and $b_0/b_2 > 0$ which implies that H^* has two negative roots, and for $(t,w) \in (2/3,3/2] \times [0,3]$, one has $H^* > 0$. For $t \in [0,2/3)$ fixed, one has $b_2 < 0$, $b_1/b_2 < 0$, $b_0/b_2 < 0$ and H^* has a positive and a negative root; given that $b_2 < 0$, H^* is positive between them. For $w = 3$ and $t \geq 0$, one has $H^* = t(16t + 15) \geq 0$, with equality only for $t = 0$. Therefore $H^* > 0$ for $(t,w) \in [0,2/3) \times [0,3]$. And for $t = 2/3$, one obtains $H^* = (16/3)w + 10/9$ which is positive for $w \geq 0$.

Part (2). One has

$$\begin{aligned} a_5^* &= -8t^5 + 8t^4 w + 6t^3 w^2 - 4t^2 w^3 - 2tw^4 - 24t^4 \\ &\quad -66t^3 w - 63t^2 w^2 - 12tw^3 + 3w^4 + 84t^3 + 153t^2 w \\ &\quad +90tw^2 - 3w^3 - 144t^2 - 144tw - 36w^2 + 108t + 54w\,. \end{aligned}$$

Consider a_5^* as a polynomial in w. Set $R_w :=\mathrm{Res}(a_5^*, \partial a_5^*/\partial w, w)/2125764$. Then $R_w = (2t-3)R_w^1 R_w^2$, where

$$\begin{aligned} R_w^1 &= 32t^5 + 16t^4 - 80t^3 + 184t^2 - 142t - 63\,, \\ R_w^2 &= 10t^{10} - 80t^9 + 365t^8 - 928t^7 + 1564t^6 - 1788t^5 \\ &\quad +1345t^4 - 668t^3 + 208t^2 - 40t + 4\,. \end{aligned}$$

The real roots of R_w^1 (resp. R_w^2) equal $-2.56\ldots$, $-0.30\ldots$ and $1.18\ldots$ (resp. $0.34\ldots$ and $1.16\ldots$). That is, the largest real root of R_w is $3/2$. One has

$$a_5^*|_{w=0} = -4t(2t^4 + 6t^3 - 21t^2 + 36t - 27)\,,$$

with real roots equal to $-5.55\ldots$, 0 and $1.18\ldots$. This means that for $t > 3/2$, the signs of the real roots of a_5^* do not change and their number (counted with multiplicity) remains the same. For $t = 3/2$ and $t = 2$, one has

$$a_5^* = -30w^3 - (45/2)w^2 - (243/4) \quad \text{and} \quad a_5^* = -w^4 - 43w^3 - 60w^2 - 22w - 328$$

respectively, which quantities are negative. Hence $a_5^* < 0$ for $t \geq 3/2$ from which Part (2) follows.

Part (3). Consider the resultant

$$\begin{aligned} R^\flat &:= \mathrm{Res}(H^*, a_5^*, t) = -52488w(w-3)R^\sharp(w^2 - w + 1)^2\,, \\ R^\sharp &:= 5w^6 - 16w^5 + 40w^4 - 23w^3 + 61w^2 - 16w - 2\,. \end{aligned}$$

The real roots of $R^\sharp$ equal $-0.09\ldots$ and $0.37\ldots$; the factor $w^2 - w + 1$ has no real roots. Thus the largest real root of $R^\flat$ equals 3. For $w = 3$, one has

$$a_5^* = -4t^2(2t^3 + 15t + 90) \leq 0\,,$$

with equality if and only if $t = 0$. For $w > 3$ and $t \geq 0$, the sets $\{H^* = 0\}$ and $\{a_5^* = 0\}$ do not intersect (because $R^\flat < 0$). We showed in the proof of part (1) of the lemma that the discriminant Δ_t is positive for $w \geq 3$. Hence each horizontal line $w = w_0 > 3$ intersects the set $\{H^* = 0\}$ for two values of t; one of them is positive and one of them is negative (because $c_0/c_2 < 0$); we denote them by t_+ and t_-.

The discriminant $R_t :=\mathrm{Res}(a_5^*, \partial a_5^*/\partial t, t)$ equals $2176782336(w-3)R_t^1 R_t^2$, where

$$\begin{aligned} R_t^1 &:= 5w^{12} + 50w^{11} + 100w^{10} - 2513w^9 + 10781w^8 - 25932w^7 + 46604w^6 \\ &\quad -70411w^5 + 86678w^4 - 82706w^3 + 65264w^2 - 43104w + 16896\,, \end{aligned}$$

$$R_t^2 := 8w^4 + 154w^3 - 68w^2 - 239w - 352\,.$$

The factor R_t^1 is without real roots. The real roots of R_t^2 (both simple) equal $-19.61\ldots$ and $1.81\ldots$. Hence for each $w = w_0 > 3$, the polynomial a_5^* has one and the same number of real roots. Their signs do not change with t. Indeed, a_5^* is a degree 5 polynomial in t, with leading coefficient and constant term equal to -8 and $3w(w-3)(w^2+2w-6)$ respectively; the real roots of the quadratic factor equal $-3.64\ldots$ and $1.64\ldots$.

For $w_0 > 3$, the polynomial a_5^* has exactly 3 real roots $t_1 < t_2 < t_3$. For any $w_0 > 3$, the signs of these roots and of the roots $t_\pm$ of H^* and the order of these 5 numbers on the real line are the same. For $w = 4$, one has

$$t_1 = -3.3\ldots < t_- = -1.6\ldots < t_2 = -0.8\ldots < t_+ = 0.2\ldots < t_3 = 0.3\ldots$$

Hence the only positive root t_3 of a_5^* belongs to the domain where $H^* > 0$. Hence one cannot have $a_5^* = 0$ and $H^* < 0$ at the same time. Lemma 17 is proved. □

Proof of Lemma 18. Part (1). One has

$$\begin{aligned} a_4^* &:= -20t^4 - 22t^3w - 30t^2w^2 - 10tw^3 + w^4 + 66t^3 + 45t^2w + 36tw^2 \\ &\quad +15w^3 - 135t^2 - 54tw - 54w^2 + 108t + 54w - 81\,. \end{aligned}$$

Consider a_4^* as a polynomial in t. Its discriminant $\Delta_t^\bullet :=\mathrm{Res}(a_4^*, \partial a_4^*/\partial t, t)$ is of the form $170061120\,\Delta^\flat \Delta^\sharp (w^2 - w + 1)^2$, where

$$\begin{aligned} \Delta^\flat &:= 9w^4 + 48w^3 + 82w^2 + 56w + 205\,, \\ \Delta^\sharp &:= 3w^4 + 14w^3 - 63w^2 + 51w - 82\,. \end{aligned}$$

Only the factor $\Delta^\sharp$ has real roots, and these are $w_- := -7.72\ldots$ and $w_+ := 2.56\ldots$; they are simple. For $w \in (w_-, w_+)$, the quantity a_4^* is negative. Indeed, $a_4^*|_{w=0} = -20t^4 + 66t^3 - 135t^2 + 108t - 81$ which polynomial has no real roots; hence this is the case of $a_4^*|_{w=w_0}$ for any $w_0 \in (w_-, w_+)$. This proves Part (1), because the set Ω_3 belongs to the strip $\{w_- < w < w_+\}$.

Part (2). The discriminant $\mathrm{Res}(a_4^*, H^*, t)$ equals $-26244\, R^{\triangle}(w^2 - w + 1)^2$ whose factor

$$R^{\triangle} := 2w^6 + 16w^5 - 61w^4 + 23w^3 - 40w^2 + 16w - 5$$

has exactly two real (and simple) roots which equal $-10.90\ldots$ and $2.68\ldots$. Hence for $w \geq 3 > w_+$,

(1) the sets $\{H^* = 0\}$ and $\{a_4^* = 0\}$ do not intersect;

(2) the numbers of positive and negative roots of H^* and a_4^* do not change; for H^* this follows from formula (6.2); for a_4^* whose leading coefficient as a polynomial in t equals -20, this results from $a_4^*|_{t=0} = w^4 + 15w^3 - 54w^2 + 54w - 81$ whose real roots $-18.1\ldots$ and $2.5\ldots$ (both simple) are < 3.

Hence for $w = w_0 \geq 3$, one has $h_- < A_- < 0 \leq h_+ < A_+$, where h_- and h_+ (resp. A_- and A_+) are the two roots of $H^*|_{w=w_0}$ (resp. of $a_4^*|_{w=w_0}$), with equality only for $w_0 = 3$. It is sufficient to check this string of inequalities for one value of w_0, say, for $w_0 = 4$, in which case one obtains

$$h_- = -1.63\ldots < A_- = -1.26\ldots < h_+ = 0.22\ldots < A_+ = 0.85\ldots\,.$$

Hence for $w = w_0 \geq 3$, the only positive root of the polynomial $a_4^*|_{w=w_0}$ belongs to the domain $\{H^* > 0\}$. This proves Part (2) of the Lemma. □

7. PROOFS OF LEMMAS 14, 15 AND 16

Proof of Lemma 14. We are using the following:

Notation 3. If ζ_1, ζ_2, …, ζ_k are distinct roots of the polynomial P (not necessarily simple), then by P_{ζ_1}, P_{ζ_1,ζ_2}, …, $P_{\zeta_1,\zeta_2,\ldots,\zeta_k}$ we denote the polynomials

$$P/(x-\zeta_1)\,,\ P/(x-\zeta_1)(x-\zeta_2)\,,\ \ldots\,,\ P/(x-\zeta_1)(x-\zeta_2)\ldots(x-\zeta_k)\,.$$

Denote by u, v, w and t the four distinct roots of P (all nonzero). Hence

$$P = (x-u)^m(x-v)^n(x-w)^p(x-t)^q\ ,\ m+n+p+q = 9\,.$$

For $j = 1, 4$ or 5, we show that the Jacobian matrix $J := (\partial(a_8, a_7, a_j)/\partial(u, v, w, t))^{\top}$ (where a_8, a_7, a_j are the corresponding coefficients of P expressed as functions of (u, v, w, t)) is of rank 3. (The entry in position $(2, 3)$ of J is $\partial a_7/\partial w$.) Hence one can vary the values of (u, v, w, t) in such a way that a_8 and a_j remain fixed (the value of a_8 being -1) and a_7 takes all possible nearby values. Hence the polynomial is not a_7-maximal.

The entries of the four columns of J are the coefficients of x^8, x^7 and x^j of the polynomials $-mP_u = \partial P/\partial u$, $-nP_v$, $-pP_w$ and $-qP_t$. By abuse of language

we say that the linear space $\mathcal{F}$ spanned by the columns of J is generated by the polynomials P_u, P_v, P_w and P_t. As

$$P_{u,v} = \frac{P_u - P_v}{v-u}\,, \quad P_{u,w} = \frac{P_u - P_w}{w-u} \quad \text{and} \quad P_{u,t} = \frac{P_u - P_t}{t-u}\,,$$

one can choose as generators of $\mathcal{F}$ the quadruple $(P_u,\, P_{u,v},\, P_{u,w},\, P_{u,t})$; in the same way one can choose $(P_u,\, P_{u,v},\, P_{u,v,w},\, P_{u,v,t})$ or $(P_u,\, P_{u,v},\, P_{u,v,w},\, P_{u,v,w,t})$ (the latter polynomials are of respective degrees 8, 7, 6 and 5). As $(x-t)P_{u,v,w,t} = P_{u,v,w}$, $(x-w)P_{u,v} = P_{u,v,w}$ etc. one can choose as generators the quadruple

$$\psi := (x^3 P_{u,v,w,t}\ ,\ x^2 P_{u,v,w,t}\ ,\ x P_{u,v,w,t}\ ,\ P_{u,v,w,t})\,.$$

Set $P_{u,v,w,t} := x^5 + Ax^4 + \cdots + G$. The coefficients of x^8, x^7 and x^5 of the quadruple ψ define the matrix

$$J^* := \begin{pmatrix} 1 & 0 & 0 & 0 \\ A & 1 & 0 & 0 \\ D & C & B & A \end{pmatrix}\,.$$

Its columns span the space $\mathcal{F}$ hence $\operatorname{rank} J^* = \operatorname{rank} J$. As at least one of the coefficients B and A is nonzero (Lemma 7) one has $\operatorname{rank} J^* = 3$ and the lemma follows (for the case $j = 6$). In the cases $j = 5$ and $j = 1$ the last row of J^* equals respectively $(\,E\ D\ C\ B\,)$ and $(\,0\ 0\ G\ F\,)$ and in the same way $\operatorname{rank} J^* = 3$. □

Proof of Lemma 15. We are using Notation 3 and the method of proof of Lemma 14. Denote by u, v, w, t, h the five distinct real roots of P (not necessarily simple). Thus using Lemma 10 one can assume that

$$\begin{gathered} P = (x+u)^\ell (x+v)^m (x+w)^n (x-t)^2 (x-h)\,, \\ u,\, v,\, w,\, t,\, h > 0\,, \quad \ell + m + n = 6\,. \end{gathered} \tag{7.1}$$

Set $J := (\partial(a_8, a_7, a_j, a_1)/\partial(u, v, w, t, h))^\top$, $j = 4$ or 5. The columns of J span a linear space $\mathcal{L}$ defined by analogy with the space $\mathcal{F}$ of the proof of Lemma 14, but spanned by 4-vector-columns.

Set $P_{u,v,w,t,h} := x^4 + ax^3 + bx^2 + cx + d$. Consider the vector-column

$$(0, 0, 0, 0, 1, a, b, c, d)^\top\,.$$

The similar vector-columns defined when using the polynomials $x^s P_{u,v,w,t,h}$, $1 \le s \le 4$, instead of $P_{u,v,w,t,h}$ are obtained from this one by successive shifts by one position upward. To obtain generators of $\mathcal{L}$ one has to restrict these vector-columns to the rows corresponding to x^8 (first), x^7 (second), x^j ($(9-j)$th) and x (eighth row).

Further we assume that $a_1 = 0$. If this is not the case, then at most one of the conditions $a_4 = 0$ and $a_5 = 0$ is fulfilled and the proof of the lemma can be finished by analogy with the proof of Lemma 14.

Consider the case $j = 5$. The rank of J is the same as the rank of the matrix

$$M := \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ a & 1 & 0 & 0 & 0 \\ c & b & a & 1 & 0 \\ 0 & 0 & 0 & d & c \end{pmatrix} \quad \begin{matrix} x^8 \\ x^7 \\ x^5 \\ x \end{matrix} .$$

One has $\operatorname{rank} M = 2 + \operatorname{rank} N$, where $N = \begin{pmatrix} a & 1 & 0 \\ 0 & d & c \end{pmatrix}$. Given that $d \neq 0$, see Lemma 4, one can have $\operatorname{rank} N < 2$ only if $a = c = 0$. We show that the condition $a = c = 0$ leads to the contradiction that one must have $a_8 > 0$. We set $u = 1$ to reduce the number of parameters, so we require only the inequality $a_8 < 0$, but not the equality $a_8 = -1$, to hold true. We have to consider the following cases for the values of the triple (ℓ, m, n) (see (7.1)): 1) $(4, 1, 1)$, 2) $(3, 2, 1)$ and 3) $(2, 2, 2)$. Notice that

$$P_{u,v,w,t,h}|_{u=1} = (x + 1)^{\ell-1}(x + v)^{m-1}(x + w)^{n-1}(x - t) .$$

In case 1) one has

$$a = 3 - t , \quad b = 3 - 3t , \quad c = 1 - 3t \quad \text{and} \quad d = -t , \tag{7.2}$$

so the condition $a = c = 0$ leads to the contradiction $3 = t = 1/3$.

In case 2) one obtains

$$a = 2 + v - t , \quad b = 1 + 2v - (2 + v)t , \quad c = v - (1 + 2v)t \quad \text{and} \quad d = -vt . \tag{7.3}$$

Thus, the condition $a = c = 0$ yields $v = -1$, $t = 1$. This is also a contradiction because v must be positive.

In case 3) one gets

$$\begin{aligned} &a = 1 + v + w - t , \quad b = v + (1 + v)w - (1 + v + w)t , \\ &c = vw - (v + (1 + v)w)t , \quad d = -vwt . \end{aligned} \tag{7.4}$$

Expressing v and w as functions of t from the system of equations $a = c = 0$, one obtains two possible solutions: $v = t$, $w = -1$ and $v = -1$, $w = t$. In both cases one of the variables (v, w) is negative which is a contradiction.

Now consider the case $j = 4$. The matrices M and N equal respectively

$$M := \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ a & 1 & 0 & 0 & 0 \\ d & c & b & a & 1 \\ 0 & 0 & 0 & d & c \end{pmatrix} , \quad N = \begin{pmatrix} b & a & 1 \\ 0 & d & c \end{pmatrix} .$$

One has $\operatorname{rank} N < 2$ only for $b = 0$, $d = ac$ (because $d \neq 0$).

In case 1) these conditions lead to the contradiction $1 = t = (3 \pm \sqrt{5})/2$, see (7.2).

In case 2) one expresses the variable t from the condition $b = 0$: $t = t^\bullet := (1 + 2v)/(2 + v)$. Set $a^\bullet := a|_{t=t^\bullet}$, $c^\bullet := c|_{t=t^\bullet}$ and $d^\bullet := d|_{t=t^\bullet}$. The quantity $d^\bullet - a^\bullet c^\bullet$ equals $3(v^2 + v + 1)^2/(2 + v)^2$ which vanishes for no $v \geq 0$. So case 2) is also impossible.

In case 3) the condition $b = 0$ implies $t = t^\triangle := (vw + v + w)/(1 + v + w)$. Set $a^\triangle := a|_{t=t^\triangle}$, $c^\triangle := c|_{t=t^\triangle}$ and $d^\triangle := d|_{t=t^\triangle}$. The quantity $d^\triangle - a^\triangle c^\triangle$ equals $(w^2+w+1)(v^2+v+1)(v^2+vw+w^2)/(1+v+w)^2$ which is positive for any $v \geq 0$, $w \geq 0$. Hence case 3) is impossible. The lemma is proved. □

Proof of Lemma 16. We use the same ideas and notation as in the proof of Lemma 15. Six of the six or more real roots of P are denoted by (u, v, w, t, h, q). The space $\mathcal{L}$ is defined by analogy with the one of the proof of Lemma 15. The Jacobian matrix J is of the form

$$J := (\partial(a_8, a_7, a_j, a_1)/\partial(u, v, w, t, h, q))^\top .$$

Set $P_{u,v,w,t,h,q} := x^3 + ax^2 + bx + c$ and consider the vector-column

$$(0, 0, 0, 0, 0, 1, a, b, c)^\top .$$

Its successive shifts by one position upward correspond to the polynomials $x^s P_{u,v,w,t,h,q}$, $s \leq 5$. In the case $j = 5$ the matrices M and N look like this:

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ a & 1 & 0 & 0 & 0 & 0 \\ c & b & a & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & c & b \end{pmatrix}, \quad N = \begin{pmatrix} a & 1 & 0 & 0 \\ 0 & 0 & c & b \end{pmatrix} .$$

One has $\operatorname{rank} M = 2 + \operatorname{rank} N$ and $\operatorname{rank} N = 2$, because at least one of the two coefficients b and c is nonzero (Lemma 7). Hence $\operatorname{rank} M = 4$ and the lemma is proved by analogy with Lemmas 14 and 15. In the case $j = 4$ the matrices M and N look like this:

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ a & 1 & 0 & 0 & 0 & 0 \\ 0 & c & b & a & 1 & 0 \\ 0 & 0 & 0 & 0 & c & b \end{pmatrix}, \quad N = \begin{pmatrix} b & a & 1 & 0 \\ 0 & 0 & c & b \end{pmatrix} .$$

The matrix N is of rank 4, because either $b \neq 0$ or $b = 0$ and both a and c are nonzero (Lemma 7). Hence $\operatorname{rank} M = 4$. □

8. PROOF OF PART (2) OF THEOREM 1

We remind that we consider polynomials with positive leading coefficients. For $d = 9$, we denote by σ a sign pattern and by σ^* the shortened sign pattern (obtained from σ by deleting its last component).

Lemma 19. *For $d = 9$, if $pos \geq 2$ and $neg \geq 2$, then such a couple (sign pattern, admissible pair) is realizable.*

Proof. Suppose that the last two components of σ are equal (resp. different). Then the pair $(pos, neg-1)$ (resp. $(pos-1, neg)$) is admissible for the sign pattern σ^* and the couple $(\sigma^*,\ (pos\ ,\ neg-1))$ (resp. $(\sigma^*,\ (pos-1\ ,\ neg))$) is realizable by some degree 8 polynomial P, see Remark 4. Hence the couple $(\sigma,\ (pos\ ,\ neg))$ is realizable by the concatenation of the polynomials P and $x+1$ (resp. P and $x-1$). □

Lemma 19 implies that in any nonrealizable couple with $pos > 0$ and $neg > 0$, one of the numbers pos, neg equals 1. Using the the standard $\mathbb{Z}_2 \times \mathbb{Z}_2$-action (i.e changing if necessary $P(x)$ to $-P(-x)$) one can assume that $pos = 1$. This implies that the last component of the sign pattern is $-$.

Lemma 20. *For $d = 9$, if $pos = 1$, $neg \geq 2$ and the last two components of σ are $(-\ ,\ -)$, then such a couple $(\sigma,\ (pos\ ,\ neg))$ is realizable.*

Proof. The couple $(\sigma^*,\ (pos\ ,\ neg-1))$ is realizable by some polynomial P, see Remark 4. Hence the concatenation of P and $x+1$ realizes the couple $(\sigma,\ (pos\ ,\ neg))$. □

Hence for any nonrealizable couple $(\sigma,\ (pos\ ,\ neg))$, one has $pos = 1$, $neg \geq 2$ and the last two components of σ are $(+\ ,\ -)$. Thus, the couple $(\sigma^*,\ (0\ ,\ neg))$ is nonrealizable. The first and the last components of σ^* are $+$. There are 19 such couples modulo the $\mathbb{Z}_2 \times \mathbb{Z}_2$-action, see [11]:

Case	Sign pattern	Admissible pair(s)
A	$(++------++)$	$(0,6)$
B	$(+-------++)$	$(0,6)$
C	$(++++----+)$	$(0,6)$
D	$(+++-----+)$	$(0,6)$
E	$(+-+---+-+)$	$(0,2)$
F	$(+-+-+---+)$	$(0,2)$
$G1-G2$	$(+-+-------+)$	$(0,2)\,,\ (0,4)$
$H1-H2$	$(+----+---+)$	$(0,2)\,,\ (0,4)$
$I1-I3$	$(+---------+)$	$(0,2)\,,\ (0,4)\,,\ (0,6)$

$$
\begin{array}{lll}
J & (+++----++) & (0,6) \\
K & (+-----+--+) & (0,4) \\
L & (+-------++) & (0,4) \\
M & (+-+++----+) & (0,4) \\
N & (+-+-----++) & (0,4) \\
Q & (+-----+-++) & (0,4)
\end{array}
$$

To obtain all couples $(\sigma^*, (0,\ neg))$ giving rise to nonrealizable couples $(\sigma, (1,\ neg))$ by concatenation with $x-1$, one has to add to the above list of cases $(A-Q)$ the cases obtained from them by acting with the first generator of the $\mathbb{Z}_2 \times \mathbb{Z}_2$-action, i.e. the one replacing σ by σ^r, see Definition 2. The second generator (the one replacing σ by σ^m) has to be ignored, because it exchanges the two components of the admissible pair and the condition $pos = 1$ could not be maintained. The cases that are to be added are denoted by $(A^r - Q^r)$. E.g.

$$
N^r \qquad (++-----+-+) \qquad (0,4)\,.
$$

One can observe that, due to the center-symmetry of certain sign patterns, one has $A = A^r$, $E = E^r$, $Hj = Hj^r$, $j = 1,\ 2$ and $Ij = Ij^r$, $j = 1,\ 2,\ 3$.

With the only exception of case C^r, we show that all cases $(A-Q)$ and $(A^r - Q^r)$, are realizable which proves part (2) of the theorem. We do this by means of Lemma 2. We explain this first for the following cases:

$$
\begin{array}{llllllllllll}
B^r, & C, & D, & E, & F, & F^r, & G1, & G1^r, & G2, & G2^r, & H1, & H2, \\
I1, & I2, & I3, & K, & K^r, & L^r, & M, & M^r, & N^r & \text{and} & Q^r\,.
\end{array}
$$

In all of them the last three components of σ are $(-+-)$, and we set $P_2^\dagger := x^2-x+1$ (see part (2) of Lemma 2). The polynomial $P_2^\dagger$ has no real roots and defines the sign pattern $\sigma^\dagger := (+-+)$. Denote by $\tilde{\sigma}$ the sign pattern obtained from σ by deleting its two last components. Hence $(1, neg)$ is an admissible pair for the sign pattern $\tilde{\sigma}$, and the couple $(\tilde{\sigma}, (1, neg))$ is realizable by some degree 7 monic polynomial $\tilde{P}_1$, see Remark 4. By Lemma 2 the concatenation of $\tilde{P}_1$ and $P_2^\dagger$ realizes the couple $(\sigma, (1, neg))$.

In cases A, B, J, L, N and Q, the last four components of the sign pattern σ are $(-++-)$. We set $P_2^\triangle := (x+2)((x^2-2)+1) = x^3-2x^2-3x+10$. Hence $P_2^\triangle$ realizes the couple $((+--+), (0,1))$. Denote by $\sigma^\triangle$ the sign pattern obtained from σ by deleting its three last components. Hence $(1, neg-1)$ is an admissible pair for the sign pattern $\sigma^\triangle$, and the couple $(\sigma^\triangle, (1, neg-1))$ is realizable by some

degree 6 monic polynomial $P_1^{\triangle}$, see Remark 4. By Lemma 2 the concatenation of $P_1^{\triangle}$ and $P_2^{\triangle}$ realizes the couple $(\sigma, (1, neg))$.

In the two remaining cases D^r and J^r, the last six components of σ are $(- - + + + -)$. The sign pattern $\sigma^{\ddagger} := (+ + - - - +)$ is realizable by some degree 5 polynomial $P_2^{\ddagger}$, see [1]. Denote by $\sigma^{\diamond}$ the sign pattern obtained from σ by deleting its five last components. Hence in cases D^r and J^r one has $\sigma^{\diamond} = (+ - - - -)$ and $\sigma^{\diamond} = (+ + - - -)$ respectively. Thus the couple $(\sigma^{\diamond}, (1, 3))$ is realizable by some monic degree 4 polynomial $P_1^{\diamond}$ (see Remark 4), and the concatenation of $P_1^{\diamond}$ and $P_2^{\ddagger}$ realizes the couple $(\sigma, (1, neg))$. Part (2) of Theorem 1 is proved.

9. REFERENCES

[1] Albouy, A., Fu, Y.: Some remarks about Descartes' rule of signs. *Elem. Math.*, **69**, 2014, 186–194. Zbl 1342.12002, MR3272179

[2] Anderson, B., Jackson, J., Sitharam, M.: Descartes' rule of signs revisited. *Amer. Math. Monthly*, **105**, 1998, 447–451. Zbl 0913.12001, MR1622513

[3] Cajori, F.: A history of the arithmetical methods of approximation to the roots of numerical equations of one unknown quantity. *Colorado College Publication, Science Series* ,**12–7** (1910), 171–215.

[4] Cheriha, H., Gati, Y., Kostov, V. P.: Descartes' rule of signs, Rolle's theorem and sequences of admissible pairs. (submitted). arXiv:1805.04261.

[5] Forsgård, J., Shapiro, B., Kostov, V. P.: Could René Descartes have known this? *Exp. Math.*, **24(1)**, 2015, 438–448. Zbl 1326.26027, MR3383475

[6] Fourier, J.: Sur l'usage du théorème de Descartes dans la recherche des limites des racines. *Bulletin des sciences par la Société philomatique de Paris (1820)* 156–165, 181–187; œuvres 2, 291–309, Gauthier- Villars, 1890.

[7] Gauss, C. F.: Beweis eines algebraischen Lehrsatzes. *J. Reine Angew. Math.*, **3**, 1828, 1–4; Werke 3, 67–70, Göttingen, 1866. ERAM 003.0089cj, MR1577673

[8] Grabiner, D. J.: Descartes' rule of signs: another construction. *Amer. Math. Monthly*, **106**, 1999, 854–856. Zbl 0980.12001, MR1732666

[9] Kostov, V. P.: *Topics on hyperbolic polynomials in one variable*. Panoramas et Synthèses **33**, 2011, vi + 141 p., SMF. Zbl 1259.12001, MR2952044

[10] Kostov, V. P.: Polynomials, sign patterns and Descartes' rule of signs. *Math. Bohem.*, **144(1)**, 2019, 39–67.

[11] Kostov, V. P.: On realizability of sign patterns by real polynomials. *Czechoslovak Math. J.* (to appear) arXiv:1703.03313.

[12] Kostov, V. P.: On a stratification defined by real roots of polynomials. *Serdica Math. J.*, **29(2)**, 2003, 177–186. Electronic preprint math.AG/0208219. Zbl 1049.12002, MR1992553

[13] Kostov, V. P., Shapiro, B.: Something you always wanted to know about real polynomials (but were afraid to ask). (submitted). arXiv:1703.04436.

Received on November 25, 2019
Received in a revised form on March 20, 2020

Hassen Chericha
Université Côte d'Azur, LJAD
FRANCE
and
University of Carthage, EPT - LIM
TUNISIA
E-mail: hassan.cheriha@unice.fr

Youstra Gati
University of Carthage, EPT - LIM
TUNISIA
E-mail: yousra.gati@gmail.com

Vladimir Petrov Kostov
Université Côte d'Azur, LJAD
FRANCE
E-mail: vladimir.kostov@unice.fr

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 106

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“

FACULTY OF MATHEMATICS AND INFORMATICS
Volume 106

SATURATED AND PRIMITIVE SMOOTH COMPACTIFICATIONS OF BALL QUOTIENTS

P. G. BESHKOV, A. K. KASPARYAN, G. K. SANKARAN

Let $X = (\mathbb{B}/\Gamma)'$ be a smooth toroidal compactification of a quotient of the complex 2-ball $\mathbb{B} = \mathrm{PSU}_{2,1}/\mathrm{PS}(U_2 \times U_1)$ by a lattice $\Gamma < \mathrm{PSU}_{2,1}$, $D := X \setminus (\mathbb{B}/\Gamma)$ be the toroidal compactifying divisor of X, $\rho : X \to Y$ be a finite composition of blow downs to a minimal surface Y and $E(\rho)$ be the exceptional divisor of ρ. The present article establishes a bijective correspondence between the finite unramified coverings of ordered triples (X, D, E) and the finite unramified coverings of $(\rho(X), \rho(D), \rho(E))$. We say that $(X, D, E(\rho))$ is saturated if all the unramified coverings $f : (X', D', E'(\rho')) \to (X, D, E)$ are isomorphisms, while $(X, D, E(\rho))$ is primitive exactly when any unramified covering $f : (X, D, E(\rho)) \to (f(X), f(D), f(E(\rho)))$ is an isomorphism. The covering relations among the smooth toroidal compactifications $(\mathbb{B}/\Gamma)'$ are studied by Uludag's [7], Stover's [6], Di Cerbo and Stover's [2] and other articles.

In the case of a single blow up $\rho = \beta : X = (\mathbb{B}/\Gamma)' \to Y$ of finitely many points of Y, we show that there is an isomorphism $\Phi : \mathrm{Aut}(Y, \beta(D)) \to \mathrm{Aut}(X, D)$ of the relative automorphism groups and $\mathrm{Aut}(X, D)$ is a finite group. Moreover, when Y is an abelian surface then any finite unramified covering $f : (X, D, E(\beta)) \to (f(X), f(D), f(E(\beta)))$ factors through an $\mathrm{Aut}(X, D)$-Galois covering. We discuss the saturation and the primitiveness of X with Kodaira dimension $\kappa(X) = -\infty$, as well as of X with $K3$ or Enriques minimal model Y.

Keywords: Smooth toroidal compactifications of quotients of the complex 2-ball, unramified coverings.

2010 Math. Subject Classification: Primary: 14M27; Secondary: 14J25, 51H30.

1. UNRAMIFIED PULL BACK OF A SMOOTH COMPACTIFICATION

Lemma 1. *Let M be a complex manifold and N be a complex analytic subvariety of M or an open subset of M.*

(i) If $f : M \to f(M)$ is an unramified covering of degree d then $f : N \to f(N)$ is an unramified covering of degree d exactly when $f : M \setminus N \to f(M) \setminus f(N)$ is an unramified covering of degree d.

(ii) Let us suppose that $f : M \to f(M)$ is a holomorphic map onto a complex manifold, $f(N) \cap f(M \setminus N) = \emptyset$ and $f : N \to f(N)$, $f : M \setminus N \to f(M \setminus N)$ are unramified coverings of degree d. Then $f : M \to f(M)$ is an unramified covering of degree d.

Proof. (i) Let $X := N$ or $X := M \setminus N$. Then $f : X \to f(X)$ is an unramified covering of degree $\deg(f|_X) = \deg(f|_M) = d$ exactly when $f^{-1}(f(X)) = X$. If so, then the intersection $f^{-1}(f(M \setminus X)) \cap X = \emptyset$ is empty, whereas $f^{-1}(f(M \setminus X)) = M \setminus X$, the union $f(M) = f(X) \coprod f(M \setminus X)$ is disjoint and $f : M \setminus X \to f(M \setminus X) = f(M) \setminus f(X)$ is an unramified covering of degree d.

(ii) The union $f(M) = f(N) \coprod f(M \setminus N)$ is disjoint, so that $f^{-1}(f(M \setminus N)) = M \setminus N$, $f^{-1}(f(N)) = N$ and $f : M \to f(M)$ is an unramified covering of degree d.□

Lemma 2. *Let $f : X \to X'$ be an unramified covering of degree d of smooth projective surfaces.*

(i) Suppose that $D = \coprod_{j=1}^{k} D_j$ is a divisor on X with disjoint smooth irreducible components D_j and f restricts to an unramified covering $f : D \to f(D)$ of degree d. Then $f(D) = \cup_{j=1}^{k} f(D_j)$ has smooth irreducible components $f(D_j)$, f restricts to unramified coverings $f : D_j \to f(D_j)$ for all $1 \leq j \leq k$ and $f(D_i) \cap f(D_j) = \emptyset$ for $f(D_i) \not\equiv f(D_j)$.

In particular, D_j are smooth elliptic curves if and only if $f(D_j)$ are smooth elliptic curves.

(ii) If C' is a smooth irreducible rational curve on X' then the complete preimage $f^{-1}(C') = \coprod_{i=1}^{d} C_i$ consists of d disjoint smooth irreducible rational curves C_i and f restricts to isomorphisms $f : C_i \to C'$ for all $1 \leq i \leq d$.

Proof. (i) The unramified covering $f : D \to f(D)$ is a local biholomorphism, so that $f(D)$ is a smooth divisor on X'. Thus, all the irreducible components $f(D_j)$ of $f(D)$ are smooth curves and $f(D_i) \cap f(D_j) \neq \emptyset$ requires $f(D_i) \equiv f(D_j)$. For any $1 \leq i \leq k$ let $J(i)$ be the set of those $1 \leq j \leq k$, for which $f(D_j) \equiv f(D_i)$. Then there exists a subset $I \subseteq \{1, \ldots, k\}$ with $\coprod_{i \in I} J(i) = \{1, \ldots, k\}$ and $f(D) = \coprod_{i \in I} f(D_i)$. By the very definition of $J(i)$, there holds the inclusion $\coprod_{j \in J(i)} D_j \subseteq f^{-1}(f(D_i))$.

Since f restricts to an unramified covering $f : D \to f(D)$ of degree d, any $p \in f^{-1}(f(D_i))$ belongs to D_s for some $1 \leq s \leq k$. Then $f(p) \in f(D_i)$ specified that $s \in J(i)$, whereas $f^{-1}(f(D_i)) \subseteq \coprod\limits_{j\in J(i)} D_j$ and $f^{-1}(f(D_i)) = \coprod\limits_{j\in J(i)} D_j$. Thus, for any $i \in I$ the morphism f restricts to an unramified covering $f : \coprod\limits_{j\in J(i)} D_j \to f(D_i)$ of degree d. By definition, any $f(p) \in f(D_i)$ with $p \in \coprod\limits_{j\in J(i)} D_j$ has a trivializing neighborhood U on $f(D_i)$, whose pull back $f^{-1}(U) = \coprod\limits_{q\in f^{-1}(p)} V_q$ is a disjoint union of neighborhoods V_q of $q \in f^{-1}(p)$ on $\coprod\limits_{j\in J(i)} D_j$ with biholomorphic restrictions $f : V_q \to U$. For a sufficiently small U one can assume that $V_q \subset D_j$ for $q \in D_j$. That is why f restricts to unramified coverings $f : D_j \to f(D_j) = f(D_i)$. In particular, D_j are smooth elliptic curves exactly when $f(D_j)$ are smooth elliptic curves.

(ii) Let $f^{-1}(C') = \sum\limits_{i=1}^{k} C_i$ be a union of k irreducible curves C_i, $d_i := \deg[f|_{C_i} : C_i \to C']$ and $\mathrm{Br}(f|_{C_i}) := \{q \in C' \mid \left|f^{-1}(q) \cap C_i\right| < d_i\}$ be the branch locus of $f|_{C_i}$ for $1 \leq i \leq k$. Any $\mathrm{Br}(f|_{C_i})$ is a finite set, as well as the intersection $\cup_{1\leq i<j\leq k} C_i \cap C_j$ of different irreducible components, so that

$$\Sigma := \left[\cup_{i=1}^{k} \mathrm{Br}(f|_{C_i})\right] \cup \left[\cup_{1\leq i<j\leq k} f(C_i \cap C_j)\right]$$

is a finite subset of C'. For any $q \in C' \setminus \Sigma$ one has $f^{-1}(q) = \coprod\limits_{i=1}^{k} f^{-1}(q) \cap C_i$, whereas

$$d = \left|f^{-1}(q)\right| = \sum_{i=1}^{k} \left|f^{-1}(q) \cap C_i\right| = \sum_{i=1}^{k} d_i.$$

If $q_j \in \mathrm{Br}(f|_{C_j})$ then $f^{-1}(q_j) = \cup_{i=1}^{k} f^{-1}(q_j) \cap C_i$ with $\left|f^{-1}(q_j) \cap C_j\right| < d_j$, so that

$$d = \left|f^{-1}(q_j)\right| \leq \sum_{i=1}^{k} \left|f^{-1}(q_j) \cap C_i\right| < \sum_{i=1}^{k} d_i = d.$$

This is absurd, justifying $\mathrm{Br}(f|_{C_j}) = \emptyset$ for all $1 \leq j \leq k$. Similarly, for any $p \in C_i \cap C_j$ there holds

$$d = \left|f^{-1}(p)\right| < \sum_{i=1}^{k} \left|f^{-1}(p) \cap C_i\right| = \sum_{i=1}^{k} d_i = d.$$

The contradiction shows that the irreducible components C_i of $f^{-1}(C')$ are disjoint. The unramified coverings $f|_{C_i} : C_i \to C'$ of the smooth irreducible rational curve C' are of degree $d_i = 1$, due to $\pi_1(C') = \{1\}$. Therefore $d = \sum\limits_{i=1}^{k} d_i = k$ and

$f^{-1}(C') = \coprod_{i=1}^{d} C_i$ consists of d disjoint smooth irreducible rational curves with biholomorphic restrictions $f|_{C_i} : C_i \to C'$ for all $1 \leq i \leq d$. $\square$

A (-1)-curve L_i on a smooth projective surface Y is a smooth irreducible rational curve with self-intersection $L_i^2 = -1$. Throughout, we say that a smooth projective surface Y is minimal if it does not contain a (-1)-curve. This is slightly different from the contemporary viewpoint of the Minimal Model Program, which considers a smooth projective surface Y to be minimal if its canonical divisor K_Y is nef (i.e., $K_Y.C \geq 0$ for all effective curves $C \subset Y$). The numerical effectiveness of K_Y excludes the existence of (-1)-curves on Y. If Y is of Kodaira dimension $\kappa(Y) = -\infty$ then K_Y is not nef, regardless of the presence of (-1)-curves on Y. That is the reason for exploiting the older, out of date notion of minimality of a smooth projective surface, which requires the non-existence of (-1)-curves on Y. By a theorem of Castelnuovo (Theorem V.5.7 [5]), for any smooth irreducible projective surface X there is a birational morphism $\rho : X \to Y$ onto a minimal smooth projective surface Y, which is a composition of blow downs of (-1)-curves. If X is of Kodaira dimension $\kappa(X) \geq 0$ then the minimal model Y of X is unique (up to an isomorphism). This is not true when X is birational to a rational or a ruled surface.

Lemma 3. *(i) Let* $\mathrm{Bl} : X_1 \to Y_1$ *be a blow down of a* (-1)*-curve* $L_1 \subset X_1$ *and* $\varphi : Y_2 \to Y_1$ *be an unramified covering of degree* d*. Then the fibered product commutative diagram*

$$\begin{array}{ccc} X_2 := X_1 \times_{Y_1} Y_2 & \xrightarrow{\beta} & Y_2 \\ \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle \varphi} \\ X_1 & \xrightarrow{\mathrm{Bl}} & Y_1 \end{array} \tag{1}$$

consists of an unramified covering $f : X_2 \to X_1$ *of degree* d *and the blow down* $\beta : X_2 \to Y_2$ *of the disjoint union* $f^{-1}(L_1) = \coprod_{j=1}^{d} L_{1,j}$ *of the* (-1)*-curves* $L_{1,j}$*.*

(ii) Let $\rho_1 : \mathrm{Bl}_1 \dots \mathrm{Bl}_{r-1}\mathrm{Bl}_r : T_r := X_1 \to Y_1 =: T_0$ *be a composition of blow downs* $\mathrm{Bl}_i : T_i \to T_{i-1}$ *of* (-1)*-curves* $L_i \subset T_i$ *and* $\varphi : Y_2 \to Y_1$ *be an unramified covering of degree* d*. Then the fibered product commutative diagrams*

$$\begin{array}{ccc} S_i := T_i \times_{T_{i-1}} S_{i-1} & \xrightarrow{\beta_i} & S_{i-1} \\ \downarrow{\scriptstyle \varphi_i} & & \downarrow{\scriptstyle \varphi_{i-1}} \\ T_i & \xrightarrow{\mathrm{Bl}_i} & T_{i-1} \end{array} \tag{2}$$

fit into a commutative diagram

$$\begin{array}{ccccc} S_r & \ldots S_i := T_i \times_{T_{i-1}} S_{i-1} & \xrightarrow{\beta_i} & S_{i-1} & \ldots S_0 := Y_2 \\ \downarrow{\scriptstyle f} & \downarrow{\scriptstyle \varphi_i} & & \downarrow{\scriptstyle \varphi_{i-1}} & \downarrow{\scriptstyle \varphi=\varphi_0} \\ T_r := X & \ldots T_i & \xrightarrow{\mathrm{Bl}_i} & T_{i-1} & \ldots T_0 := Y_1 \end{array} \tag{3}$$

and induce a fibered product commutative diagram

$$\begin{array}{ccc} X_2 = X_1 \times_{Y_1} Y_2 & \xrightarrow{\rho_2} & Y_2 \\ \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle \varphi} \\ X_1 & \xrightarrow{\rho_1} & Y_1 \end{array} \tag{4}$$

with an unramified covering $f : X_2 \to X_1$ of degree d and a composition $\rho_2 = \beta_1 \ldots \beta_{r-1}\beta_r : X_2 \to Y_2$ of blow downs of $\varphi_i^{-1}(L_i) = \coprod_{j=1}^{d} L_{i,j}$ for all $1 \leq i \leq r$.

Proof. (i) By the very definition of a blow down $\mathrm{Bl} : X_1 \to Y_1$ of L_1 to $\mathrm{Bl}(L_1) = q_1 \in Y_1$, one has $X_1 \setminus L_1 = Y_1 \setminus \{q_1\}$. Then

$$X_2 := X_1 \times_{Y_1} Y_2 = [(X_1 \setminus L_1) \times_{Y_1} Y_2] \coprod [L_1 \times_{Y_1} Y_2]$$

decomposes into the disjoint union of

$$(X_1 \setminus L_1) \times_{Y_1} Y_2 = \{(x_1, y_2) \,|\, x_1 = \mathrm{Bl}(x_1) = \varphi(y_2)\} \simeq Y_2 \setminus \varphi^{-1}(q_1) \quad \text{and}$$

$$L_1 \times_{Y_1} Y_2 = \{(x_1, y_2) \,|\, q_1 = \mathrm{Bl}(x_1) = \varphi(y_2)\} = L_1 \times \varphi^{-1}(q_1).$$

If $\varphi^{-1}(q_1) = \{p_{1,j} \,|\, 1 \leq j \leq d\}$ then X_2 is the blow up of Y_2 at $\{p_{1,j} \,|\, 1 \leq j \leq d\}$. Due to $\mathrm{Bl}f = \varphi\beta$, the exceptional divisor of β is $\beta^{-1}(\{p_{1,j} \,|\, 1 \leq j \leq d\}) = \beta^{-1}\varphi^{-1}(q_1) = (\varphi\beta)^{-1}(q_1) = (\mathrm{Bl}f)^{-1}(q_1) = f^{-1}\mathrm{Bl}^{-1}(q_1) = f^{-1}(L_1) = \coprod_{j=1}^{d} L_{1,j}$. According to Corollary 17.7.3 (i) from Grothendieck's [4], $f : X_2 \to X_1$ is an unramified covering, since $\varphi : Y_2 \to Y_1$ is an unramified covering.

(ii) By an increasing induction on $1 \leq i \leq r$, one applies (i) to the fibered product commutative diagrams (2) and justifies (ii). □

Lemma 4. *(i) In the notations from Lemma 3 (i) and the fibered product commutative diagram (1), let $D^{(2)}$ be a (possibly reducible) divisor on X_2, which does not contain an irreducible component of the exceptional divisor of β and $D^{(1)}$ be a (possibly reducible) divisor on X_1, which does not contain the exceptional divisor L_1 of* Bl. *Then the restriction $f : D^{(2)} \to D^{(1)}$ is an unramified covering of degree*

$d = \deg[f : X_2 \to X_1]$ *if and only if* $\varphi : \beta(D^{(2)}) \to \mathrm{Bl}(D^{(1)})$ *is an unramified covering of degree* d.

(ii) In the notations from Lemma 3 (ii) and the fibered product commutative diagram (4), let $D^{(2)}$ *be a (possibly reducible) divisor on* X_2*, which does not contain an irreducible component of the exceptional divisor of* ρ_2 *and* $D^{(1)}$ *be a (possibly reducible) divisor on* X_1*, which does not contain an irreducible component of the exceptional divisor of* ρ_1*. Then the restriction* $f : D^{(2)} \to D^{(1)}$ *is an unramified covering of degree* d *if and only if the restriction* $\varphi : \rho_2(D^{(2)}) \to \rho_1(D^{(1)})$ *is an unramified covering of degree* d.

Proof. (i) If $f : D^{(2)} \to D^{(1)}$ is an unramified covering of degree d then $f^{-1}(D^{(1)} \cap L_1) = f^{-1}(D^{(1)}) \cap f^{-1}(L_1) = D^{(2)} \cap f^{-1}(L_1)$ and the restriction $f : D^{(1)} \cap f^{-1}(L_1) \to D^{(1)} \cap L_1$ is an unramified covering of degree d. After denoting $f^{-1}(L_1) = \coprod\limits_{j=1}^{d} L_{1,j}$, $\beta(L_{1,j}) = p_{1,j}$ and $\mathrm{Bl}(L_1) = q_1$, one applies Lemma 1 (i), in order to conclude that

$$\varphi \equiv f : \beta(D^{(2)}) \setminus \{p_{1,j} \,|\, 1 \leq j \leq d\} \equiv D^{(2)} \setminus f^{-1}(L_1) \longrightarrow D^{(1)} \setminus L_1 \equiv \mathrm{Bl}(D^{(1)}) \setminus \{q_1\}$$

is an unramified covering of degree d. As a result, the morphism φ restricts to $\varphi : \{p_{1,j} \,|\, 1 \leq j \leq d\} \to \{q_1\}$, so that

$$\varphi : \beta(D^{(2)}) = \beta(D^{(2)}) \setminus \{p_{1,j} \,|\, 1 \leq j \leq d\} \coprod \{p_{1,j} \,|\, 1 \leq j \leq d\} \longrightarrow$$
$$\longrightarrow \Big[\mathrm{Bl}(D^{(1)}) \setminus \{q_1\}\Big] \coprod \{q_1\} = \mathrm{Bl}(D^{(1)})$$

is an unramified covering of degree d by Lemma 1 (ii).

Conversely, assume that $\varphi : \beta(D^{(2)}) \to \mathrm{Bl}(D^{(1)})$ is an unramified covering of degree d. Choose a sufficiently small neighborhood V of $q_1 = \mathrm{Bl}(L_1)$ on Y_1, such that $\varphi^{-1}(V) = \coprod\limits_{j=1}^{d} U_j$ is a disjoint union of neighborhoods U_j of $p_{1,j}$, $1 \leq j \leq d$ on Y_2 with biholomorphic restrictions $\varphi : U_j \to V$ of φ. Bearing in mind that $\mathrm{Bl}_1 : X_1 \to Y_1$ is the blow up of Y_1 at q_1, one decomposes

$$\mathrm{Bl}(D^{(1)}) = \Big[\mathrm{Bl}(D^{(1)}) \setminus V\Big] \coprod \Big[\mathrm{Bl}(D^{(1)}) \cap V\Big] \quad \text{and}$$

$$D^{(1)} = \Big[\mathrm{Bl}(D^{(1)}) \setminus V\Big] \coprod \mathrm{Bl}^{-1}(\mathrm{Bl}(D^{(1)}) \cap V).$$

Similarly, $\beta : X_2 \to Y_2$ is the blow up of Y_2 at $\varphi^{-1}(q_1) = \{p_{1,j} \,|\, 1 \leq j \leq d\}$, so that there are decompositions

$$\beta(D^{(2)}) = \Big[\beta(D^{(2)}) \setminus \varphi^{-1}(V)\Big] \coprod \Big[\beta(D^{(2)}) \cap \varphi^{-1}(V)\Big] \quad \text{and}$$

$$D^{(2)} = \Big[\beta(D^{(2)}) \setminus \varphi^{-1}(V)\Big] \coprod \beta^{-1}(\beta(D^{(2)}) \cap \varphi^{-1}(V)).$$

According to $\varphi^{-1}(\mathrm{Bl}(D^{(1)})\cap V)=\varphi^{-1}(\mathrm{Bl}(D^{(1)}))\cap\varphi^{-1}(V)=\beta(D^{(2)})\cap\varphi^{-1}(V)$, the restriction $\varphi:\beta(D^{(2)})\cap\varphi^{-1}(V)\to\mathrm{Bl}(D^{(1)})\cap V$ is an unramified covering of degree d. Now, Lemma 1 (ii) applies to provide that

$$f\equiv\varphi:\beta(D^{(2)})\setminus\varphi^{-1}(V)\longrightarrow\mathrm{Bl}(D^{(1)})\setminus V$$

is an unramified covering of degree d. According to Lemma 1 (ii), it sufficed to show that

$$f:\beta^{-1}(\beta(D^{(2)})\cap\varphi^{-1}(V))\longrightarrow\mathrm{Bl}^{-1}(\mathrm{Bl}(D^{(1)})\cap V)$$

is an unramified covering of degree d, in order to conclude that $f:D^{(2)}\to D^{(1)}$ is an unramified covering of degree d. To this end, note that

$$\varphi^{-1}(\mathrm{Bl}(D^{(1)})\cap V)=\beta(D^{(2)})\cap\varphi^{-1}(V)=\beta(D^{(2)})\cap\left(\coprod_{j=1}^{d}U_j\right)=\coprod_{j=1}^{d}\Big[\beta(D^{(2)})\cap U_j\Big],$$

so that

$$\varphi:\coprod_{j=1}^{d}\Big[\beta(D^{(2)})\cap U_j\Big]\longrightarrow\mathrm{Bl}(D^{(1)})\cap V$$

is an unramified covering of degree d. Thus, the biholomorphisms $\varphi:U_j\to V$ restrict to biholomorphisms $\varphi:\beta(D^{(2)})\cap U_j\to\mathrm{Bl}(D^{(1)})\cap V$. According to $\varphi(p_{1,j})=q_1$, there arise biholomorphisms

$$\varphi:(\beta(D^{(2)})\cap U_j)\setminus\{p_{1,j}\}\longrightarrow(\mathrm{Bl}(D^{(1)})\cap V)\setminus\{q_1\}.$$

By the very definition of a blow up at a point, these induce biholomorphisms

$$f:\Big[(\beta(D^{(2)})\cap U_j)\setminus\{p_{1,j}\}\Big]\coprod L_{1,j}\longrightarrow\Big[(\mathrm{Bl}(D^{(1)})\cap V)\setminus\{q_1\}\Big]\coprod L_1$$

for all $1\leq j\leq d$. Bearing in mind that

$$\coprod_{j=1}^{d}\Big\{\Big[(\beta(D^{(2)})\cap U_j)\setminus\{p_{1,j}\}\Big]\coprod L_{1,j}\Big\}=\beta^{-1}(\beta(D^{(2)})\cap\varphi^{-1}(V)),$$

one concludes that φ induces an unramified covering

$$f:\beta^{-1}(\beta(D^{(2)})\cap\varphi^{-1}(V))\longrightarrow\mathrm{Bl}^{-1}(\mathrm{Bl}(D^{(1)})\cap V)$$

of degree d.

(ii) Along the commutative diagram (3), if $f:D^{(2)}\to D^{(1)}$ is an unramified covering of degree d then by a decreasing induction on $r\geq i\geq 1$ and making use of (i), one observes that $\varphi_i:\beta_{i+1}\ldots\beta_r(D^{(2)})\to\mathrm{Bl}_{i+1}\ldots\mathrm{Bl}_r(D^{(1)})$ is an unramified covering of degree d, whereas $\varphi:\rho_2(D^{(2)})\to\rho_1(D^{(1)})$ is an unramified covering of degree d. Conversely, suppose that $\varphi:\rho_2(D^{(2)})\to\rho_1(D^{(1)})$ is an unramified

covering of degree d. Then by an increasing induction on $1 \leq i \leq r$ and making use of (i), one concludes that

$$\varphi_i : \beta_{i+1} \dots \beta_r(D^{(2)}) \to \mathrm{Bl}_{i+1} \dots \mathrm{Bl}_r(D^{(1)})$$

is an unramified covering of degree d. As a result, $f : D^{(2)} \to D^{(1)}$ is an unramified covering of degree d. □

Corollary 5. *Let $X_1 = (\mathbb{B}/\Gamma_1)$ be a smooth toroidal compactification, $\rho_1 : X_1 \to Y_1$ be a composition of blow downs onto a minimal surface Y_1, $\varphi : Y_2 \to Y_1$ be an unramified covering of degree d and (4) be the defining commutative diagram of the fibered product $X_2 = X_1 \times_{Y_1} Y_2$. Then:*

(i) there is a subgroup Γ_2 of Γ_1 of index $[\Gamma_1 : \Gamma_2] = d$, such that $X_2 = (\mathbb{B}/\Gamma_2)'$ is the toroidal compactification of $\mathbb{B}/\Gamma_2$;

(ii) $f : X_2 \to X_1$ restricts to unramified coverings $f : \mathbb{B}/\Gamma_2 \to \mathbb{B}/\Gamma_1$, respectively, $f : D^{(2)} := X_2 \setminus (\mathbb{B}/\Gamma_2) \to X_1 \setminus (\mathbb{B}/\Gamma_1) =: D^{(1)}$ of degree d;

(iii) the composition $\rho_2 : X_2 \to Y_2$ of blow downs maps onto a minimal surface Y_2;

(iv) φ restricts to an unramified covering $\varphi : \rho_2(D^{(2)}) \to \rho_1(D^{(1)})$ of degree d.

Proof. By Lemma 3 (ii), the fibered product diagram (4) consists of an unramified covering $f : X_2 \to X_1$ of degree d and a composition $\rho_2 : X_2 \to Y_2$ of blow downs. The surface Y_2 is minimal. Otherwise any (-1)-curve L'_i on Y_2 maps isomorphically onto a (-1)-curve $\varphi(L'_i) \subset Y_1$, according to Lemma 2 (ii). That contradicts the minimality of Y_1 and shows the minimality of Y_2.

The unramified covering $f : X_2 \to X_1 = (\mathbb{B}/\Gamma_1)'$ of degree d restricts to an unramified covering $f : f^{-1}(\mathbb{B}/\Gamma_1) \to \mathbb{B}/\Gamma_1$ of degree d. The smoothness of $\mathbb{B}/\Gamma_1$ excludes the existence of isolated branch points of the Γ_1-Galois covering $\zeta_1 : \mathbb{B} \to \mathbb{B}/\Gamma_1$. However, ζ_1 can ramify along divisors and $\mathbb{B}$ is not the usual universal cover of the complex manifold $\mathbb{B}/\Gamma_1$. Nevertheless, $\mathbb{B}$ is the orbifold universal cover of $\mathbb{B}/\Gamma_1$ and the orbifold universal covering map $\zeta_1 : \mathbb{B} \to \mathbb{B}/\Gamma_1$ factors through a (possibly ramified) covering $\zeta_2 : \mathbb{B} \to f^{-1}(\mathbb{B}/\Gamma_1)$ and the covering $f : f^{-1}(\mathbb{B}/\Gamma_1) \to \mathbb{B}/\Gamma_1$, i.e., $\zeta_1 = f\zeta_2$. Since $\pi_1^{\mathrm{orb}}(\mathbb{B}) = \{1\}$ is a normal subgroup of $\Gamma_2 := \pi_1^{\mathrm{orb}}(f^{-1}(\mathbb{B}/\Gamma))$, the covering ζ_2 is Galois and its Galois group Γ_2 is a subgroup of $\Gamma_1 = \pi_1^{\mathrm{orb}}(\mathbb{B}/\Gamma_1)$ of index $[\Gamma_1 : \Gamma_2] = d$. In particular, $f^{-1}(\mathbb{B}/\Gamma_1) = \mathbb{B}/\Gamma_2$. By Lemma 1 (i), f restricts to an unramified covering $f : D^{(2)} := X_2 \setminus (\mathbb{B}/\Gamma_2) \to X_1 \setminus (\mathbb{B}/\Gamma_1) =: D^{(1)}$ of degree d of the toroidal compactifying divisor $D^{(1)} = \coprod_{j=1}^{k} D_j^{(1)}$ of $\mathbb{B}/\Gamma_1$. Note that for any $1 \leq j \leq k$ the restriction $f : f^{-1}(D_j^{(1)}) \to D_j^{(1)}$ is an unramified covering of degree d, whereas a local biholomorphism. Therefore $f^{-1}(D_j^{(1)}) = \cup_{i=1}^{r_j} D_{j,i}^{(2)}$ is smooth and has disjoint smooth irreducible components $D_{j,i}^{(2)}$. As a result,

$$D^{(2)} = f^{-1}(D^{(1)}) = \coprod_{j=1}^{k} f^{-1}(D_j^{(1)}) = \coprod_{j=1}^{k} \coprod_{i=1}^{r_j} D_{j,i}^{(2)}$$

has disjoint smooth irreducible components $D^{(2)}_{j,i}$. By assumption, $D^{(1)}_j$ are smooth elliptic curves, so that all $D^{(2)}_{j,i}$ are smooth elliptic curves by Lemma 2 (i). That is why $X_2 = (\mathbb{B}/\Gamma_2)'$ is the toroidal compactification of $\mathbb{B}/\Gamma_2$. According to Lemma 4 (ii), $\varphi : Y_2 \to Y_1$ restricts to an unramified covering $\varphi : \rho_2(D^{(2)}) \to \rho_1(D^{(1)})$ of degree d. □

Lemma 6. *(i) Let $f : X_2 \to X_1$ be an unramified covering of degree d of smooth projective surfaces and $\mathrm{Bl} : X_1 \to Y_1$ be a blow down of a (-1)-curve $L_1 \subset X_1$. Then the Stein factorization $\varphi\beta$ of $\mathrm{Bl}f$ consists of the blow down $\beta : X_2 \to Y_2$ of $f^{-1}(L_1) = \coprod_{j=1}^{d} L_{1,j}$ and an unramified covering $\varphi : Y_2 \to Y_1$ of degree d, so that $X_2 = X_1 \times_{Y_1} Y_2$ is the fibered product of X_1 and Y_2 over Y_1.*

(ii) Let $\rho_1 = \mathrm{Bl}_1 \dots \mathrm{Bl}_r : T_r := X_1 \to Y_1 =: T_0$ be a composition of blow downs of (-1)-curves $L_i \subset T_i$ and $f : X_2 \to X_1$ be an unramified covering of degree d. Then the Stein factorization $\varphi\rho_2$ of $\rho_1 f : X_2 \to Y_1$ closes the fibered product commutative diagram (4) with the composition $\rho_2 = \beta_1 \dots \beta_r : S_r := X_2 \to Y_2 := S_0$ of the blow downs $\beta_i : S_i \to S_{i-1}$ of $\varphi_i^{-1}(L_i) = \coprod_{j=1}^{d} L_{i,j}$ for all $1 \le i \le r$ and an unramified covering $\varphi : Y_2 \to Y_1$ of degree d.

Proof. (i) If $\mathrm{Bl}f = \varphi\beta : X_2 \to Y_1$ is the Stein factorization of $\mathrm{Bl}f$ and $q_1 := \mathrm{Bl}(L_1)$ then $(\mathrm{Bl}f)^{-1}(q_1) = f^{-1}\mathrm{Bl}^{-1}(q_1) = f^{-1}(L_1) = \coprod_{j=1}^{d} L_{1,j}$ has irreducible components $L_{1,j}$ by Lemma 4. For any $q \in Y_1 \setminus \{q_1\}$ one has $(\mathrm{Bl}f)^{-1}(q) = f^{-1}\mathrm{Bl}^{-1}(q) = f^{-1}(q)$ of cardinality $\left|f^{-1}(q)\right| = d$. Therefore, the surjective morphism $\beta : X_2 \to Y_2$ with connected fibres is the blow down of $L_{1,j}$, $\forall 1 \le j \le d$. According to Lemma 1 (i), the restriction $f : X_2 \setminus f^{-1}(L_1) \to X_1 \setminus L_1$ is an unramified covering of degree d, since $f : f^{-1}(L_1) \to L_1$ is an unramified covering of degree d. In such a way, there arises a commutative diagram

$$\begin{array}{ccc} X_2 \setminus f^{-1}(L_1) & \xrightarrow{\beta=\mathrm{Id}} & Y_2 \setminus \beta f^{-1}(L_1) \\ \Big\downarrow{\scriptstyle f} & & \Big\downarrow{\scriptstyle \varphi} \\ X_1 \setminus L_1 & \xrightarrow{\mathrm{Bl}=\mathrm{Id}} & Y_1 \setminus \{q_1\} \end{array}$$

and $\varphi : Y_2 \setminus \beta f^{-1}(L_1) \to Y_1 \setminus \{q_1\}$ is an unramified covering of degree d. If $p_{1,j} := \beta(L_{1,j})$ then $\beta^{-1}\varphi^{-1}(q_1) = (\varphi\beta)^{-1}(q_1) = (\mathrm{Bl}f)^{-1}(q_1) = \coprod_{j=1}^{d} L_{1,j}$ reveals that $\varphi^{-1}(q_1) = \{p_{1,j} \,|\, 1 \le j \le d\}$ consists of d points and $\varphi : Y_2 \to Y_1$ is an unramified covering of degree d. By Lemma 3 (i), the fibered product $X_2' := X_1 \times_{Y_1} Y_2$ is the blow up of Y_2 at $\varphi^{-1}(q_1) = \{p_{1,j} \,|\, 1 \le j \le d\}$, so that $X_2' = X_2$.

According to Grothendieck's Corollary 17.7.3 (i) from [4], it suffices to show that $X_2' = X_2$, in order to conclude that $\varphi : Y_2 \to Y_1$ is an unramified covering of

degree d. We have justified straightforwardly that $\varphi : Y_2 \to Y_1$ is an unramified covering of degree d, in order to use it towards the coincidence of X_2 with the fibered product $X_2' := X_1 \times_{Y_1} Y_2$.

(ii) is an immediate consequence of the fact that the composition of morphisms with connected fibres has connected fibres. □

Corollary 7. *Let $f : X_2 \to X_1 = (\mathbb{B}/\Gamma_1)'$ be an unramified covering of degree d of a smooth toroidal compactification $X_1 = (\mathbb{B}/\Gamma_1)'$, $\rho_1 : X_1 \to Y_1$ be a composition of blow downs onto a minimal surface Y_1 and $D^{(1)} := X_1 \setminus (\mathbb{B}/\Gamma_1)$ be the toroidal compactifying divisor of $\mathbb{B}/\Gamma_1$. Then:*

(i) there exist a composition $\rho_2 : X_2 \to Y_2$ of blow downs onto a minimal surface Y_2 and an unramified covering $\varphi : Y_2 \to Y_1$ of degree d, which exhibits $X_2 = X_1 \times_{Y_1} Y_2$ as a fibered product of X_1 and Y_2 over Y_1;

(ii) there is a subgroup $\Gamma_2 < \Gamma_1$ of index $[\Gamma_1 : \Gamma_2] = d$, such that $X_2 = (\mathbb{B}/\Gamma_2)'$ is the toroidal compactification of $\mathbb{B}/\Gamma_2$ and f restricts to unramified coverings $f : \mathbb{B}/\Gamma_2 \to \mathbb{B}/\Gamma_1$, $f : D^{(2)} := X_2 \setminus (\mathbb{B}/\Gamma_2) \to X_1 \setminus (\mathbb{B}/\Gamma_2) =: D^{(1)}$ of degree d;

(iii) φ restricts to an unramified covering $\varphi : \rho_2(D^{(2)}) \to \rho_1(D^{(1)})$ of degree d.

Proof. (i) is an immediate consequence of Lemma 6 (ii) and the fact that any inramified cover Y_2 of a minimal surface Y_1 is minimal.

(ii) The unramified covering $f : X_2 \to X_1 = (\mathbb{B}/\Gamma_1)'$ of degree d restricts to an unramified covering $f : f^{-1}(\mathbb{B}/\Gamma_1) \to \mathbb{B}/\Gamma_1$ of degree d. As in the proof of Corollary 5, there is a subgroup $\Gamma_2 < \Gamma_1$ of index $[\Gamma_1 : \Gamma_2] = d$, such that $X_2 = (\mathbb{B}/\Gamma_2)'$ is the toroidal compactification of $\mathbb{B}/\Gamma_2$ and f restricts to unramified coverings $f : \mathbb{B}/\Gamma_2 \to \mathbb{B}/\Gamma_1$, $f : D^{(2)} := X_2 \setminus (\mathbb{B}/\Gamma_2) \to X_1 \setminus (\mathbb{B}/\Gamma_1) =: D^{(1)}$ of degree d.

(iii) is an immediate consequence of Lemma 4 (ii). □

Definition 8. A smooth toroidal compactification $X_1 = (\mathbb{B}/\Gamma_1)'$ is saturated if there is no unramified covering $f : X_2 = (\mathbb{B}/\Gamma_2)' \to (\mathbb{B}/\Gamma_1)' = X_1$ of degree d, which restricts to an unramified covering $f : \mathbb{B}/\Gamma_2 \to \mathbb{B}/\Gamma_1$ of degree d.

Bearing in mind that the fundamental group of a smooth projective variety is a birational invariant, one combines Corollary 5 with Corollary 7 and obtains the following

Corollary 9. *A smooth toroidal compactification $X_1 = (\mathbb{B}/\Gamma_1)'$ is saturated if and only if one and, therefore, any minimal model Y_1 of X_1 is simply connected.*

2. UNRAMIFIED PUSH FORWARD OF A SMOOTH COMPACTIFICATION

Let X_2 be a smooth projective surface, $\beta : X_2 \to Y_2$ be a blow down with exceptional divisor $E(\beta) = \coprod_{s=1}^{d} L_{1,s}$ and $f : X_2 \to X_1$ be an unramified covering of degree d, which restricts to an unramified covering $f : E(\beta) \to f(E(\beta))$ of degree d. According to Lemma 2 (ii), $L_1 := f(E(\beta))$ is a (-1)-curve on X_1. Then Lemma 6 (i) implies that there is a fibered product commutative diagram (1) with the blow down $\mathrm{Bl} : X_1 \to Y_1$ of L_1 and an unramified covering $\varphi : Y_2 \to Y_1$ of degree d, which shrinks $\beta(E(\beta)) = \{p_{1,j} := \beta(L_{1,j}) \,|\, 1 \le j \le d\}$ to a point $q_1 \in Y_1$. We say that φ is induced by f.

Suppose that $\rho_2 = \beta_1 \dots \beta_r : S_r := X_2 \to Y_2 =: S_0$ is a composition of blow downs

$$\beta_i : S_i := \beta_{i+1} \dots \beta_r(S_r) \longrightarrow S_{i-1} := \beta_i \dots \beta_r(S_r) \tag{5}$$

with exceptional divisors $E(\beta_i) = \coprod_{s=1}^{d} L_{i,s}$ for all $1 \le i \le r$. By a decreasing induction on $r \ge i \ge 1$, let us assume that there is a fibered product commutative diagram

$$\begin{array}{ccc} S_r & \xrightarrow{\beta_r} & S_{r-1} \\ {\scriptstyle f=\varphi_r}\downarrow & & \downarrow{\scriptstyle \varphi_{r-1}} \\ f(S_r) & \xrightarrow{\mathrm{Bl}_r} & \varphi_{r-1}(S_{r-1}) \end{array} \qquad \begin{array}{ccc} \dots S_{i+1} & \xrightarrow{\beta_{i+1}} & S_i \\ {\scriptstyle \varphi_{i+1}}\downarrow & & \downarrow{\scriptstyle \varphi_i} \\ \dots \varphi_{i+1}(S_{i+1}) & \xrightarrow{\mathrm{Bl}_{i+1}} & \varphi_i(S_i) \end{array}$$

with fibered product squares $\mathrm{Bl}_j\varphi_j = \varphi_{j-1}\beta_j$, such that φ_j restricts to an unramified covering $\varphi_j : E(\beta_j) \to L_j := \varphi_j(E(\beta_j))$ of degree d and φ_{j-1} shrinks the set $\beta_j(E(\beta_j)) = \{p_{j,s} := \beta_j(L_{j,s}) \,|\, 1 \le s \le d\}$ to a point $q_j \in \varphi_{j-1}(S_{j-1})$ for all $r \ge j \ge i+1$. If $\varphi_i : S_i \to \varphi_i(S_i)$ restricts to an unramified covering $\varphi_i : E(\beta_i) \to L_i := \varphi_i(E(\beta_i))$ of degree d then there is an unramified covering $\varphi_{i-1} : S_{i-1} \to \varphi_{i-1}(S_{i-1})$ of degree d, which shrinks $\beta_i(E_{\beta_i}) = \{p_{i,s} = \beta_i(L_{i,s}) \,|\, 1 \le s \le d\}$ to a point $q_i \in S_{i-1}$ and closes the fibered product commutative diagram $\varphi_{i-1}\beta_i = \mathrm{Bl}_i\varphi_i$. Thus, if an unramified covering $f : X_2 \to X_1$ of degree d induces unramified coverings $E(\beta_i) = \coprod_{s=1}^{d} L_{i,s} \to L_i$ of degree d for all $1 \le i \le r$ then there is an unramified covering $\varphi := \varphi_0 : Y_2 = S_0 \to \varphi_0(S_0) =: Y_1$ of degree d, which induces unramified coverings $\beta_i(E(\beta_i)) = \{p_{i,s} := \beta_i(L_{i,s}) \,|\, 1 \le s \le d\} \to \{q_i\} \subset \varphi_{i-1}(S_{i-1})$ of degree d for all $1 \le i \le r$.

Conversely, assume that Y_2 is a smooth projective surface, $\beta : X_2 \to Y_2$ is a blow down with exceptional divisor $E(\beta) = \coprod_{s=1}^{d} L_{1,s}$ and $\varphi : Y_2 \to Y_1$ is an unramified covering of degree d, which shrinks $\beta(E(\beta)) = \{p_{1,s} = \beta(L_{1,s}) \,|\, 1 \le s \le d\}$ to a point $q_1 \in Y_1$. According to Lemma 3 (i), there is a fibered product

commutative diagram (1), where $\mathrm{Bl} : X_1 \to Y_1$ is the blow up of Y_1 at $q_1 \in Y_1$ and $f : X_2 \to X_1$ is an unramified covering of degree d, which restricts to an unramified covering $f : E(\beta) = \coprod_{s=1}^{d} L_{1,s} \to L_1 := \mathrm{Bl}^{-1}(q_1)$ of degree d. Let $\rho_2 = \beta_1 \dots \beta_r : S_r := X_2 \to Y_2 =: S_0$ be a composition of blow downs (5) with exceptional divisors $E(\beta_i) = \coprod_{s=1}^{d} L_{i,s}$. By an increasing induction on $1 \leq i \leq r$, suppose that

$$\begin{array}{ccc} S_i & \xrightarrow{\beta_i} & S_{i-1} \\ {\scriptstyle\varphi_i}\downarrow & & \downarrow{\scriptstyle\varphi_{i-1}} \\ \varphi_i(S_i) & \xrightarrow{\mathrm{Bl}_i} & \varphi_{i-1}(S_{i-1}) \end{array} \qquad \begin{array}{ccc} \dots S_1 & \xrightarrow{\beta_1} & S_0 = Y_2 \\ {\scriptstyle\varphi_1}\downarrow & & \downarrow{\scriptstyle\varphi=\varphi_0} \\ \dots \varphi_1(S_1) & \xrightarrow{\mathrm{Bl}_1} & \varphi(Y_2) \end{array}$$

is a fibered product commutative diagram with fibered product squares $\varphi_{j-1}\beta_j = \mathrm{Bl}_j\varphi_j$, such that φ_{j-1} restricts to an unramified covering

$$\varphi_{j-1} : \beta_j(E(\beta_j)) = \{p_{j,s} := \beta_j(L_{j,s}) \,|\, 1 \leq s \leq d\} \longrightarrow \{q_j\} \subset \varphi_{j-1}(S_{j-1})$$

of degree d and φ_j restricts to an unramified covering

$$\varphi_j : E(\beta_j) = \coprod_{s=1}^{d} L_{j,s} \longrightarrow \varphi_j(E(\beta_j)) =: L_j$$

of degree d for all $1 \leq j \leq i$. If φ_i restricts to an unramified covering

$$\varphi_i : \beta_{i+1}(E(\beta_{i+1})) = \{p_{i+1,s} = \beta_{i+1}(L_{i+1,s}) \,|\, 1 \leq s \leq d\} \longrightarrow \{q_{i+1}\} \subset \varphi_i(S_i)$$

of degree d then there is an unramified covering

$$\varphi_{i+1} : S_{i+1} \longrightarrow \varphi_{i+1}(S_{i+1})$$

of degree d, which restricts to an unramified covering

$$\varphi_{i+1} : E(\beta_{i+1}) = \coprod_{s=1}^{d} L_{i+1,s} \longrightarrow L_{i+1} := \varphi_{i+1}(E(\beta_{i+1}))$$

of degree d and closes the fibered product commutative diagram $\varphi_i\beta_{i+1} = \mathrm{Bl}_{i+1}\varphi_{i+1}$ with the blow down $\mathrm{Bl}_{i+1} : \varphi_{i+1}(S_{i+1}) \to \varphi_i(S_i)$ of L_{i+1}. In such a way, if $\varphi : Y_2 \to Y_1$ is an unramified covering of degree d, which induces unramified coverings

$$\beta_i(E(\beta_i)) = \{p_{i,s} := \beta_i(L_{i,s}) \,|\, 1 \leq s \leq d\} \longrightarrow \{q_i\} \subset \varphi_{i-1}(S_{i-1})$$

of degree d for all $1 \leq i \leq r$ then $f := \varphi_r : X_2 \to f(X_2)$ is an unramified covering of degree d, which induces unramified coverings $E(\beta_i) = \coprod_{s=1}^{d} L_{i,s} \to L_i$ of degree d for all $1 \leq i \leq r$. The above considerations justify the following

Lemma-Definition 10. *Let X_2, Y_2 be smooth projective surfaces and*

$$\rho_2 = \beta_1 \dots \beta_r : S_r := X_2 \longrightarrow Y_2 =: S_0$$

be a composition of blow downs (5) with exceptional divisors $E(\beta_i)$ for all $1 \leq i \leq r$. Then the following are equivalent:

(i) there is an unramified covering $f : X_2 \to f(X_2)$ of degree d, which induces unramified coverings $E(\beta_i) = \coprod_{s=1}^{d} L_{i,s} \to L_i$ of degree d for all $1 \leq i \leq r$;

(ii) there is an unramified covering $\varphi : Y_2 \to \varphi(Y_2)$ of degree d, which induces unramified coverings $\beta_i(E(\beta_i)) = \{p_{i,s} = \beta_i(L_{i,s}) \,|\, 1 \leq s \leq d\} \to \{q_i\} \subset \varphi_{i-1}(S_{i-1})$ of degree d for all $1 \leq i \leq r$.

If there holds one and, therefore, any one of the aforementioned conditions then there is a fibered product commutative diagram (4), where

$$\rho_1 = \mathrm{Bl}_1 \dots \mathrm{Bl}_r : X_1 := \varphi(X_2) \to \varphi(Y_2) =: Y_1$$

is the composition of blow downs Bl_i of L_i for all $1 \leq i \leq r$ and we say that $f : X_2 \to f(X_2)$ and $\varphi : Y_2 \to \varphi(Y_2)$ are compatible with ρ.

Corollary 11. *Let $X_2 = (\mathbb{B}/\Gamma_2)'$ be a smooth toroidal compactification and $\rho_2 : X_2 \to Y_2$ be a composition of blow downs onto a minimal surface Y_2. If there is an unramified covering $f : X_2 = (\mathbb{B}/\Gamma_2)' \to f(X_2) =: X_1$ of degree d, which is compatible with ρ_2 and restricts to an unramified covering $f : \mathbb{B}/\Gamma_2 \to f(\mathbb{B}/\Gamma_2)$ of degree d then:*

(i) there is a fibered product commutative diagram (4) with an unramified covering $\varphi : Y_2 \to \varphi(Y_2) =: Y_1$ of degree d and a composition of blow downs $\rho_1 : X_1 \to Y_1$ onto a minimal surface Y_1;

(ii) there is a lattice Γ_1 of $\mathrm{Aut}(\mathbb{B}) = PU(2,1)$, containing Γ_2 as a subgroup of index $[\Gamma_1 : \Gamma_2] = d$ and such that $X_1 = (\mathbb{B}/\Gamma_1)'$ is the toroidal compactification of $\mathbb{B}/\Gamma_1$;

(iii) φ restricts to an unramified covering $\varphi : \rho_2(D^{(2)}) \to \rho_1(D^{(1)})$ of degree d, where $D^{(j)} := X_j \setminus (\mathbb{B}/\Gamma_j)$ are the compactifying divisors of $\mathbb{B}/\Gamma_j$, $1 \leq j \leq 2$.

Proof. (i) is an immediate consequence of Lemma 10.

Towards (ii), let us note that the composition $f\zeta_2 : \mathbb{B} \to f(\mathbb{B}/\Gamma_2)$ of the orbifold universal covering $\zeta_2 : \mathbb{B} \to \mathbb{B}/\Gamma_2$ with the unramified covering $f : \mathbb{B}/\Gamma_2 \to f(\mathbb{B}/\Gamma_2)$ is Galois, since $\pi_1^{\mathrm{orb}}(\mathbb{B}) = \{1\}$ is a normal subgroup of $\Gamma_1 := \pi_1^{\mathrm{orb}}(f(\mathbb{B}/\Gamma_2))$. Moreover, $\pi_1^{\mathrm{orb}}(\mathbb{B}/\Gamma_2) = \Gamma_2$ is a subgroup of Γ_1 of index $[\Gamma_1 : \Gamma_2] = d$ and $f(\mathbb{B}/\Gamma_2) = \mathbb{B}/\Gamma_1$. By Lemma 1 (i), $f : X_2 \to X_1$ restricts to an unramified covering $f : D^{(2)} = X_2 \setminus (\mathbb{B}/\Gamma_2) \to D^{(1)} := X_1 \setminus (\mathbb{B}/\Gamma_1)$ of degree d. The toroidal compactifying divisor $D^{(2)}$ of $\mathbb{B}/\Gamma_2$ has disjoint smooth elliptic irreducible components, so that Lemma 2 (i) applies to provide that $D^{(1)}$ consists of disjoint smooth elliptic irreducible components and $X_1 = (\mathbb{B}/\Gamma_1)'$ is the toroidal compactification

of $\mathbb{B}/\Gamma_1$. According to Lemma 4 (ii), that suffices for $\varphi : Y_2 \to Y_1$ to restrict to an unramified covering $\varphi : \rho_2(D^{(2)}) \to \rho_1(D^{(1)})$. □

Corollary 12. *Let $X_2 = (\mathbb{B}/\Gamma_2)'$ be a smooth toroidal compactification, $D^{(2)} := X_2 \setminus (\mathbb{B}/\Gamma_2)$ be the compactifying divisor of $\mathbb{B}/\Gamma_2$ and $\rho_2 : X_2 \to Y_2$ be a composition of blow downs onto a minimal surface Y_2. If $\varphi : Y_2 \to \varphi(Y_2)$ is an unramified covering of degree d, which is compatible with ρ_2 and restricts to an unramified covering $\varphi : \rho_2(D^{(2)}) \to \varphi\rho_2(D^{(2)})$ of degree d then:*

(i) there is a fibered product commutative diagram (4) with an unramified covering $f : X_2 \to f(X_2) =: X_1$ of degree d and a composition of blow downs $\rho_1 : X_1 \to Y_1$ onto a minimal surface Y_1;

(ii) there is a lattice Γ_1 of $\mathrm{Aut}(\mathbb{B}) = PU(2,1)$, containing Γ_2 as a subgroup of index $[\Gamma_1 : \Gamma_2] = d$ and such that $X_1 = (\mathbb{B}/\Gamma_1)'$ is the toroidal compactification of $\mathbb{B}/\Gamma_1$;

(iii) f restricts to an unramified covering $f : \mathbb{B}/\Gamma_2 \to \mathbb{B}/\Gamma_1$ of degree d.

Proof. Lemma 10 justifies (i). According to Lemma 4 (ii), f restricts to an unramified covering $f : D^{(2)} \to f(D^{(2)})$ of degree d. Then Lemma 1 (i) applies to provide that $f : X_2 \setminus D^{(2)} = \mathbb{B}/\Gamma_2 \to X_1 \setminus f(D^{(2)})$ is an unramified covering of degree d. The proof of Corollary 11 (ii) has established that this is sufficient for the existence of a lattice Γ_1 of $\mathrm{Aut}(\mathbb{B}) = PU(2,1)$, containing Γ_2 as a subgroup of index $[\Gamma_1 : \Gamma_2] = d$ and such that $X_1 \setminus f(D^{(2)}) = \mathbb{B}/\Gamma_1$. That justifies (iii). By assumption, $D^{(2)}$ consists of smooth elliptic irreducible components. Therefore $f(D^{(2)})$ has smooth elliptic irreducible components and $X_1 = (\mathbb{B}/\Gamma_1) \coprod f(D^{(2)})$ is the toroidal compactification of $\mathbb{B}/\Gamma_1$. □

Definition 13. Let $X = (\mathbb{B}/\Gamma)'$ be a smooth toroidal compactification. If there is no unramified covering $f : X \to f(X)$ of degree d, which restricts to an unramified covering $f : \mathbb{B}/\Gamma \to f(\mathbb{B}/\Gamma)$ of degree d and is compatible with some composition of blow downs $\rho : X \to Y$ onto a minimal surface Y, we say that $X = (\mathbb{B}/\Gamma)'$ is primitive.

The Euler characteristic of a smooth toroidal compactification $X = (\mathbb{B}/\Gamma)'$ is a natural number $e(X) = e(\mathbb{B}/\Gamma)$. That is why there exists a primitive smooth toroidal compactification $X_0 = \overline{\mathbb{B}/\Gamma_0}$ and a finite sequence

$$X_n := X \xrightarrow{f_n} X_{n-1} \qquad \dots X_i \xrightarrow{f_i} X_{i-1} \dots \qquad X_1 \xrightarrow{f_1} X_0$$

of unramified coverings $f_i : X_i = (\mathbb{B}/\Gamma_i)' \to (\mathbb{B}/\Gamma_{i-1})' = X_{i-1}$ of degree d_i of smooth toroidal compactifications $X_j = (\mathbb{B}/\Gamma_j)'$, which restrict to unramified coverings $f_i : \mathbb{B}/\Gamma_i \to \mathbb{B}/\Gamma_{i-1}$ of degree d_i and are compatible with some compositions of blow downs $\rho_i : X_i \to Y_i$ onto minimal surfaces Y_i. Combining Corollary 11 with Corollary 12, one obtains the following

Corollary 14. *Let $X = (\mathbb{B}/\Gamma)'$ be a smooth toroidal compactification with toroidal compactifying divisor $D := X \setminus (\mathbb{B}/\Gamma)$. Then X is primitive if and only if no minimal model Y of X with a composition of blow downs $\rho : X \to Y$ admits an unramified covering $\varphi : Y \to \varphi(Y)$ of degree $d > 1$, which restricts to an unramified covering $\varphi : \rho(D) \to \varphi\rho(D)$ of degree d and is compatible with ρ.*

Let us suppose that a smooth toroidal compactification $X = (\mathbb{B}/\Gamma)'$ with toroidal compactifying divisor $D := X \setminus (\mathbb{B}/\Gamma)$ admits a blow down $\beta : X \to Y$ of $n \in \mathbb{N}$ smooth irreducible rational (-1)-curves onto a minimal surface Y and there is an unramified covering $\varphi : Y \to \varphi(Y)$ of degree d, which restricts to unramified coverings $\varphi : \beta(D) \to \varphi\beta(D)$ and $\varphi : \beta(E(\beta)) \to \varphi\beta(E(\beta))$ of degree d. Then the Euler number of the smooth surface $\varphi(Y)$ is $e(\varphi(Y)) = \frac{e(Y)}{d} \in \mathbb{Z}$ and the cardinality of $\varphi\beta(E(\beta))$ if $|\varphi\beta(E(\beta))| = \frac{|\beta(E(\beta))|}{d} = \frac{n}{d} \in \mathbb{N}$, so that $d \in \mathbb{N}$ divides $e(Y)$ and $n = |\beta(E(\beta))|$. As a result, d divides the greatest common divisor $\mathrm{GCD}(|\beta(E(\beta))|, e(Y))$.

Note that the compatibility of an unramified covering $\varphi : Y \to \varphi(Y)$ with $\beta : X \to Y$ reduces to $\varphi^{-1}(\varphi\beta(E(\beta)) = \beta(E(\beta))$ and is detected on Y. When $\rho = \beta_1 \dots \beta_r : X \to Y$ is a composition of $r \geq 2$ blow downs, the compatibility of an unramified covering $\varphi : Y \to \varphi(Y)$ of degree d with ρ cannot be traced out on the minimal model Y of X alone. Namely, if $S_0 := Y$, $T_0 := \varphi(Y)$ then in the notations from the commutative diagram (3), the unramified covering $\varphi_1 : S_1 \to T_1$ of degree d may restrict to an unramified covering $\varphi_1 : \beta_2(E(\beta_2)) \to \varphi_1\beta_2(E(\beta_2))$ of degree d, but $\varphi_0 := \varphi$ is not supposed to restrict to an unramified covering $\varphi : \beta_1\beta_2(E(\beta_2)) \to \varphi\beta_1\beta_2(E(\beta_2))$ of degree d. More precisely, if an irreducible component $L_{1,j}$ of $E(\beta_1)$ intersects $\beta_2(E(\beta_2))$ in at least two points then $|\beta_1\beta_2(E(\beta_2))| < d$ and $\varphi : \beta_1\beta_2(E(\beta_2)) \to \varphi\beta_1\beta_2(E(\beta_2))$ is of degree $< d$.

3. SATURATED AND PRIMITIVE SMOOTH COMPACTIFICATIONS OF NON-POSITIVE KODAIRA DIMENSION

Definition 15. Let $X = (\mathbb{B}/\Gamma)'$ and $X_0 = (\mathbb{B}/\Gamma_0)'$ be smooth toroidal compactification. We say that X dominates X_0 and write $X \succeq X_0$ or $X_0 \preceq X$ if there exist a finite sequence of ball lattices

$$\Gamma_n := \Gamma < \Gamma_{n-1} < \ldots < \Gamma_i < \Gamma_{i-1} < \ldots < \Gamma_1 < \Gamma_0,$$

with smooth toroidal compactifications $X_i = (\mathbb{B}/\Gamma_i)'$ of the corresponding ball quotients $\mathbb{B}/\Gamma_i$ and a finite sequence of unramified coverings

$$X_n := X \xrightarrow{f_n} X_{n-1} \qquad \ldots X_i \xrightarrow{f_i} X_{i-1} \ldots \qquad X_1 \xrightarrow{f_1} X_0$$

of degree $\deg[f_i : X_i \to X_{i-1}] = [\Gamma_{i-1} : \Gamma_i] = d_i \in \mathbb{N}$, which restrict to unramified coverings $f_i : \mathbb{B}/\Gamma_i \to \mathbb{B}/\Gamma_{i-1}$ of degree d_i and are compatible with some compositions $\rho_i = \beta_{i,1} \dots \beta_{i,r_i} : X_i \to Y_i$ of blow downs $\beta_{i,j}$ onto minimal surfaces Y_i.

It is clear that a smooth toroidal compactification $X = \overline{\mathbb{B}/\Gamma}$ is saturated if and only if it is maximal with respect to the partial order $\succeq$. Similarly, X is primitive exactly when it is minimal with respect to $\succeq$. Note that the partial order $\succeq$ on the set $\mathcal{S}$ of the smooth toroidal compactifications $X = (\mathbb{B}/\Gamma)'$ is artinian, i.e., any subset $\mathcal{S}_o \subseteq \mathcal{S}$ has a minimal element $X_o = (\mathbb{B}/\Gamma_o)' \in \mathcal{S}_o$. The minimal $X \in \mathcal{S}$ are exactly the primitive ones, but the minimal $X_o \in \mathcal{S}_o$ are not necessarily primitive, since such X_o is not supposed to be a minimal element of $\mathcal{S}$.

The present section discusses the saturated and the primitive smooth toroidal compactifications $X = (\mathbb{B}/\Gamma)'$ of Kodaira dimension $\kappa(X) \leq 0$.

Proposition 16. *If $X = (\mathbb{B}/\Gamma)'$ is a smooth toroidal compactification of Kodaira dimension $\kappa(X) = -\infty$ then X is a rational surface or X has a ruled minimal model $\pi : Y \to E$ with an elliptic base E.*

Any smooth rational $X = (\mathbb{B}/\Gamma)'$ is both saturated and primitive.

There is no smooth saturated $X = (\mathbb{B}/\Gamma)'$, whose minimal model is a ruled surface $\pi : Y \to E$ with an elliptic base E.

Proof. (i) Let $\rho : X = (\mathbb{B}/\Gamma)' \to Y$ be a composition of blow downs onto a minimal surface Y of $\kappa(Y) = -\infty$, Then $Y = \mathbb{P}^2(\mathbb{C})$ is the complex projective plane or $\pi : Y \to E$ is a ruled surface with a base E of genus $g \in \mathbb{Z}^{\geq 0}$. The toroidal compactifying divisor $D := X \setminus (\mathbb{B}/\Gamma) = \coprod\limits_{j=1}^{k} D_j$ has disjoint smooth irreducible elliptic components D_j. If $g \geq 2$ then the morphisms $\pi\rho : D_j \to E$ map to points $p_j := \pi\rho(D_j) \in E$, so that $\rho(D_j) \subseteq \pi^{-1}(p_j)$ for all $1 \leq j \leq k$. The exceptional divisor L of $\rho : X \to Y$ has finite image $\rho(L) = \{q_1, \ldots, q_m\}$ on Y and $\rho(L) \subseteq \coprod\limits_{i=1}^{m} \pi^{-1}(\pi(q_i))$. Therefore

$$Y' := Y \setminus \left[\coprod_{i=1}^{m} \pi^{-1}(\pi(q_i))\right] \subseteq Y \setminus \rho(L) \equiv X \setminus L$$

and ρ acts identically on Y'. Moreover,

$$Y'' := Y' \setminus \left[\coprod_{j=1}^{k} \pi^{-1}(p_j)\right] = Y \setminus \left[\left(\coprod_{i=1}^{m} \pi^{-1}(\pi(q_i))\right) \coprod \left(\coprod_{j=1}^{k} \pi^{-1}(p_j)\right)\right] \subseteq \mathbb{B}/\Gamma.$$

However, Y'' contains (infinitely many) fibres $\pi^{-1}(e) \simeq \mathbb{P}^1(\mathbb{C})$, $e \in E$ of $\pi : Y \to E$ and that contradicts the Kobayashi hyperbolicity of $\mathbb{B}/\Gamma$. In such a way, we have shown that any minimal model Y of a smooth toroidal compactification $X = (\mathbb{B}/\Gamma)'$ of $\kappa(X) = -\infty$ is birational to $\mathbb{P}^2(\mathbb{C})$ or to a minimal ruled surface $\pi : Y \to E$ with an elliptic base E.

Any rational $X = (\mathbb{B}/\Gamma)'$ is simply connected and does not admit finite unramified coverings $X_1 \to X$ of degree $d > 1$. That is why X is saturated. Let us

suppose that $f : X = (\mathbb{B}/\Gamma)' \to X_0 = (\mathbb{B}/\Gamma_0)'$ is an unramified covering of degree $d > 1$, which is compatible with some composition of blow downs $\rho : X \to Y$ onto a minimal rational surface Y and restricts to an unramified covering $f : \mathbb{B}/\Gamma \to \mathbb{B}/\Gamma_0$ of degree d. The Kodaira dimension is preserved under finite unramified coverings, so that $\kappa(X_0) = \kappa(X) = -\infty$. The surface X_0 is not simply connected, whereas non-rational. Therefore, there is a composition $\rho_0 : X_0 \to Y_0$ of blow downs onto a ruled surface $\pi_0 : Y_0 \to E_0$ with base E_0 of genus $g_0 \in \mathbb{N}$. The surjective morphism $\rho_0 f : X = (\mathbb{B}/\Gamma)' \to Y_0$ induces an embedding $(\rho_0 f)^* : H^{0,1}(Y_0) \to H^{0,1}(X)$. On one hand, the irregularity of Y_0 is $h^{0,1}(Y_0) := \dim_{\mathbb{C}} H^{0,1}(Y_0) = g_o \in \mathbb{N}$. On the other hand, the rational surface X has vanishing irregularity $h^{0,1}(X) = 0$. That contradicts the presence of a finite unramified covering $f : X \to X_0$ of degree $d > 1$ and shows that any smooth rational toroidal compactification $X = (\mathbb{B}/\Gamma)'$ is primite.

Let $X = (\mathbb{B}/\Gamma)'$ be a smooth toroidal compactification, whose minimal model Y is a ruled surface $\pi : Y \to E$ with an elliptic base E. Since Y is birational to $\mathbb{P}^1(\mathbb{C}) \times E$ and the fundamental group is a birational invariant, one has $\pi_1(X) \simeq \pi_1(Y) \simeq \pi_1(E) \simeq (\mathbb{Z}^2, +)$. In particular, Y is not simply connected. According to Corollary 9, X cannot be saturated. □

According to the Enriques-Kodaira classification, there are four types of minimal smooth projective surfaces Y of Kodaira dimension $\kappa(Y) = 0$. These are the abelian and the bi-elliptic surfaces with universal cover $\mathbb{C}^2$, as well as the $K3$ and the Enriques surfaces with $K3$ universal cover. If $\varphi : Y_2 \to Y_1$ is a finite unramified covering of smooth projective surfaces then the Kodaira dimension $\kappa(Y_1) = \kappa(Y_2)$ and the universal covers $\widetilde{Y_1} = \widetilde{Y_2}$ coincide. Let Y_2 be a smooth projective surface with a fixed point free involution $g_o : Y_2 \to Y_2$ and $\beta : X_2 \to Y_2$ be the blow up of Y_2 at a $\langle g_o \rangle$-orbit $\{p_{1,1}, p_{1,2} = g_o(p_{1,1})\} \subset Y_2$. Then by the very definition of a blow up, g_o induces a fixed point free involution $g_1 : X_2 \to X_2$, which leaves invariant the exceptional divisor $E(\beta) = L_{1,1} \coprod L_{1,2}$, $L_{1,i} := \beta^{-1}(p_{1,i})$ of β and there is a fibered product commutative diagram (4) with a $\langle g_o \rangle$-Galois covering $\varphi : Y_2 \to Y_1$, a $\langle g_1 \rangle$-Galois covering $f : X_2 \to X_1$ and the blow up $\mathrm{Bl} : X_1 \to Y_1$ of Y_1 at $\{q_1\} = \varphi(\{p_{1,1}, p_{1,2}\})$. Now, suppose that $\rho_2 = \beta_1 \dots \beta_r : S_r := X_2 \to Y_2 =: S_0$ is a composition of blow downs with exceptional divisors $E(\beta_i) = L_{i,1} \coprod L_{i,2}$ and $g_o : S_0 \to S_0$ is a fixed point free involution. By an increasing induction on $1 \le i \le r$, if $g_{i-1} : S_{i-1} \to S_{i-1}$ is a fixed point free involution, which leaves invariant $\beta_i(E(\beta_i)) = \{p_{i,1}, p_{i,2}\}$ then there is a fixed point free involution $g_i : S_i \to S_i$, which leaves invariant $E(\beta_i) = L_{i,1} \coprod L_{i,2}$. In such a way, if a fixed point free involution $g_0 : S_0 \to S_0$ induces isomorphisms $L_{i,1} \to L_{i,2}$ for all $1 \le i \le r$ then there is a fixed point free involution $g_r : S_r \to S_r$ and a fibered product commutative diagram (4) with a $\langle g_o \rangle$-Galois covering $\varphi : Y_2 \to Y_1$, a $\langle g_r \rangle$-Galois covering $f : X_2 \to X_1$ and the composition $\rho_1 = \mathrm{Bl}_1 \dots \mathrm{Bl}_r : X_1 \to Y_1$ of the blow downs of $E(\beta_i)/\langle g_i \rangle = L_i \simeq \mathbb{P}^1(\mathbb{C})$. If $g_o : S_0 \to S_0$ induces isomorphisms $L_{i,1} \to L_{i,2}$ of the irreducible components of $E(\beta_i) = L_{i,1} \coprod L_{i,2}$ for all $1 \le i \le r$, we say that g_o is compatible with $\rho_2 = \beta_1 \dots \beta_r$.

Proposition 17. *Let $X = (\mathbb{B}/\Gamma)'$ be a smooth toroidal compactification, $D := X \setminus (\mathbb{B}/\Gamma)$ be the toroidal compactifying divisor of $\mathbb{B}/\Gamma$ and $\rho = \beta_1 \dots \beta_r : X \to Y$ be a composition of blow downs onto a $K3$ surface Y. Then:*

(i) X is a saturated compactification;

(ii) X is non-primitive exactly when there is a fixed point free involution $g_o : Y \to Y$, which is compatible with ρ and leaves invariant $\rho(D)$;

(iii) if X is non-primitive then there is a fibered product commutative diagram

$$\begin{array}{ccc} X & \xrightarrow{\rho} & Y \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle \varphi} \\ X_0 & \xrightarrow{\rho_0} & Y_0 \end{array}$$

with a primitive smooth toroidal compactification $X_0 = (\mathbb{B}/\Gamma_0)'$, a composition of blow downs $\rho_0 : X_0 \to Y_0$ onto a minimal Enriques surface Y_0 and unramified double covers $f : X \to X_0$, $\varphi : Y \to Y_0$.

Proof. (i) is an immediate consequence of $\pi_1(Y) = \{1\}$, according to Corollary 9.

(ii) and (iii) follow from Corollary 14 and the fact that a minimal projective surface Y_0 admits an unramified covering $\varphi : Y \to Y_0$ by a $K3$ surface Y if and only if Y_0 is the quotient of Y by a fixed point free involution $g_o : Y \to Y$. Such $Y_0 = Y/\langle g_o \rangle$ are called minimal Enriques surfaces and do not admit unramified coverings $\varphi_0 : Y_0 \to \varphi_0(Y_0)$ of degree > 1. □

Proposition 18. *Let $X = (\mathbb{B}/\Gamma)'$ be a smooth toroidal compactification and $\rho : \beta_1 \dots \beta_r : X \to Y$ be a composition of blow downs onto a minimal Enriques surface Y. Then:*

(i) X is a primitive compactification;

(ii) X is not saturated;

(iii) there is an unramified double cover $f : X_1 = \overline{\mathbb{B}/\Gamma_1} \to \overline{\mathbb{B}/\Gamma} = X$ by a saturated smooth toroidal compactification $X_1 = (\mathbb{B}/\Gamma_1)'$ with $K3$ minimal model Y_1.

Proof. (i) is due to the lack of an unramified covering $\varphi : Y \to \varphi(Y)$ of degree $d > 1$.

(ii) follows from $\pi_1(Y) = (\mathbb{Z}_2, +) \neq \{1\}$.

(iii) is an immediate consequence of the Enriques-Kodaira classification of the smooth projective surfaces. □

Let $X = (\mathbb{B}/\Gamma)'$ be a smooth toroidal compactification with abelian or bi-elliptic minimal model Y. According to Theorem 1.3 from Di Cerbo and Stover's article [3], X can be obtained from Y by blow up $\beta : X \to Y$ of $n \in \mathbb{N}$ points $p_1, \dots, p_n \in Y$.

Proposition 19. *Let $X = (\mathbb{B}/\Gamma)'$ be a smooth toroidal compactification with a blow down $\beta : X \to Y$ onto a minimal surface Y with exceptional divisor $E(\beta) = \coprod_{i=1}^{n} L_i$ and $D := X \setminus (\mathbb{B}/\Gamma)$ be the toroidal compactifying divisor of $\mathbb{B}/\Gamma$. Then:*

(i) β transforms $E(\beta)$ onto the singular locus $\beta(E(\beta)) = \beta(D)^{\rm sing}$ of $\beta(D) \subset Y$;

(ii) X is non-primitive if and only if there is an unramified covering $\varphi : Y \to \varphi(Y)$ of degree $d > 1$, which restricts to an unramified covering $\varphi : \beta(D) \to \varphi\beta(D)$ of degree d;

(iii) the relative automorphism group $\mathrm{Aut}(Y, \beta(D)) = \mathrm{Aut}(Y, \beta(D), \beta(D)^{\rm sing})$ admits an isomorphism

$$\Phi : \mathrm{Aut}(Y, \beta(D)) \longrightarrow \mathrm{Aut}(X, D)$$

with the relative automorphism group $\mathrm{Aut}(X, D) = \mathrm{Aut}(X, D, E(\beta))$;

(iv) $g_o \in \mathrm{Aut}(Y, \beta(D)$ is fixed point free if and only if it corresponds to a fixed point free $g = \Phi(g_o) \in \mathrm{Aut}(X, D)$.

Proof. (i) If $D = \coprod_{j=1}^{k} D_j$ has irreducible components D_j then the singular locus of $\beta(D)$ is

$$\beta(D)^{\rm sing} = \left[\cup_{j=1}^{k} \beta(D_j)^{\rm sing}\right] \cup \left[\cup_{1 \le i < j \le k} \beta(D_i) \cap \beta(D_j)\right].$$

Since D_j are smooth irreducible elliptic curves, $\beta(D)^{\rm sing} \subseteq \beta(E(\beta))$. Conversely, any (-1)-curve L_i on $X = (\mathbb{B}/\Gamma)'$ intersects $D = \coprod_{j=1}^{k} D_j$ in at least three points, due to the Kobayashi hyperbolicity of $\mathbb{B}/\Gamma$. In fact, $|L_i \cap F| \ge 4$, according to Theorem 1.1 (2) from Di Cerbo and Stover's article [3]. Therefore, the multiplicity of $\beta(L_i) = p_i$ with respect to $\beta(D)$ is ≥ 4 and $p_i \in \beta(D)^{\rm sing}$. That justifies $\beta(E(\beta)) \subseteq \beta(D)^{\rm sing}$ and $\beta(E(\beta)) = \beta(D)^{\rm sing}$.

(ii) By Corollary 14 and (i), $X = (\mathbb{B}/\Gamma)'$ is non-primitive if and only if there is an unramified covering $\varphi : Y \to \varphi(Y)$ of degree $d > 1$, which restricts to unramified coverings $\varphi : \beta(D) \to \varphi\beta(D)$ and $\varphi : \beta(D)^{\rm sing} \to \beta(D)^{\rm sing}$ of degree d. Let us observe that any unramified covering $\varphi : \beta(D) \to \varphi\beta(D)$ of degree d restricts to an unramified covering $\varphi : \beta(D)^{\rm sing} \to \beta(D)^{\rm sing}$ of degree d, as far as the local biholomorphism $\varphi : \beta(D) \to \varphi\beta(D)$ preserves the multiplicities of the points with respect to $\beta(D)$ and $\beta(D)^{\rm sing}$ consists of the points of $\beta(D)$ of multiplicity ≥ 2.

(iii) If a holomorphic automorphism $g_o : Y \to Y$ restricts to a holomorphic automorphism $g_o : \beta(D) \to \beta(D)$ then g_o preserves the multiplicities of the points with respect to $\beta(D)$ and $\beta(D)^{\rm sing}$ is $\langle g_o \rangle$-invariant. That justifies $\mathrm{Aut}(Y, \beta(D)) \le \mathrm{Aut}(Y, \beta(D), \beta(D)^{\rm sing})$ and $\mathrm{Aut}(Y, \beta(D)) = \mathrm{Aut}(Y, \beta(D), \beta(D)^{\rm sing})$.

In order to show the existence of a group isomorphism

$$\Phi : \mathrm{Aut}(Y, \beta(D), \beta(D)^{\rm sing}) \longrightarrow \mathrm{Aut}(X, D, E(\beta)),$$

let us pick a $g_o \in \mathrm{Aut}(Y, \beta(D), \beta(D)^{\mathrm{sing}})$. Then $X \setminus E(\beta) = Y \setminus \beta(E(\beta)) = Y \setminus \beta(D)^{\mathrm{sing}}$ is acted by $\Phi(g_o)|_{X\setminus E(\beta)} := g_o|_{Y\setminus\beta(D)^{\mathrm{sing}}}$. By the definition of a blow up at a point, the bijection $g_o : \beta(D)^{\mathrm{sing}} \to \beta(D)^{\mathrm{sing}}$ with $g_o(\beta(L_{1,i})) = \beta(L_{1,j})$ induces isomorphisms $\Phi(g_o) : L_{1,i} \to L_{1,j}$ and provides an element $\Phi(g_o) \in \mathrm{Aut}(X, E(\beta))$. After observing that $\Phi(g_o)(D \setminus E(\beta)) = g_o(\beta(D) \setminus \beta(D)^{\mathrm{sing}}) = \beta(D) \setminus \beta(D)^{\mathrm{sing}} = D \setminus E(\beta)$, one concludes that $\Phi(g_o)$ transforms the Zariski closure D of $D \setminus E(\beta)$ onto itself and $\Phi(g_o) \in \mathrm{Aut}(D)$.

The correspondence Φ is a group homomorphism since g_o and $\Phi(g_o)$ coincide on Zariski open subsets of Y, respectively, X. Towards the bijectiveness of Φ, let $g \in \mathrm{Aut}(X, D, E(\beta))$ and note that $Y \setminus \beta(D)^{\mathrm{sing}} = X \setminus E(\beta)$. That allows us to define $\phi^{-1}(g)|_{Y\setminus\beta(D)^{\mathrm{sing}}} := g|_{X\setminus E(\beta)}$. The isomorphism $g : E(\beta) \to E(\beta)$ of the exceptional divisor $E(\beta)$ of β induces a permutation $\Phi^{-1}(g) : \beta(D)^{\mathrm{sing}} \to \beta(D)^{\mathrm{sing}}$ of the finite set $\beta(D)^{\mathrm{sing}}$ and provides an automorphism $\Phi^{-1}(g) \in \mathrm{Aut}(Y, \beta(D)^{\mathrm{sing}})$. Bearing in mind that $\Phi^{-1}(g)(\beta(D) \setminus \beta(D)^{\mathrm{sing}}) = g(D \setminus E(\beta)) = D \setminus E(\beta) = \beta(D) \setminus \beta(D)^{\mathrm{sing}}$, one concludes that $\Phi^{-1}(g) \in \mathrm{Aut}(\beta(D))$ is an automorphism of the Zariski closure $\beta(D)$ of $\beta(D) \setminus \beta(D)^{\mathrm{sing}} = \beta(D)^{\mathrm{smooth}}$.

Note that any automorphism $g \in \mathrm{Aut}(X, D)$ acts on the set of the smooth irreducible rational curves on X. Moreover, g preserves the self-intersection number of such a curve and $\langle g \rangle$ acts on the set $E(\beta) = \coprod_{i=1}^{n} L_i$ of the (-1)-curves on X. Thus, $g \in \mathrm{Aut}(X, D, E(\beta))$ and $\mathrm{Aut}(X, D) \subseteq \mathrm{Aut}(X, D, E(\beta))$, whereas $\mathrm{Aut}(X, D, E(\beta)) = \mathrm{Aut}(X, D)$.

(iv) If $g \in \mathrm{Aut}(X, D)$ has no fixed point on X then $g_o := \Phi^{-1}(g) \in \mathrm{Aut}(Y, \beta(D))$ restricts to $g_o|_{Y\setminus\beta(E(\beta))} = g|_{X\setminus E(\beta)}$ without fixed points. The assumption $g_o(p_i) = p_i = \mathrm{Bl}(L_i)$ for some $1 \leq i \leq n$ implies that g restricts to an automorphism $g : L_i \to L_i$. Any biholomorphism $g \in \mathrm{Aut}(L_i) = \mathrm{Aut}(\mathbb{P}^1(\mathbb{C})) = PGL(2, \mathbb{C})$ of the projective line $L_i = \mathbb{P}^1(\mathbb{C})$ is a fractional linear transformation and has two fixed points, counted with their multiplicities. That contradicts the lack of fixed points of g on X and implies that the associated automorphism $g_o = \Phi^{-1}(g) \in \mathrm{Aut}(Y, \beta(D))$ has no fixed points on Y.

Conversely, if $g_o \in \mathrm{Aut}(Y, \beta(D))$ has no fixed points on Y and $g := \Phi(g_o)$ then the restriction $g|_{X\setminus E(\beta)} = g_o|_{Y\setminus\beta(\beta)}$ has no fixed points. If $g(x) = x$ for some $x \in E(\beta) = \coprod_{i=1}^{n} L_i$ then $x \in L_i$ for some $1 \leq i \leq n$ and $g(L_i) = L_i$. As a result, g_o fixes $p_i = \beta(L_i) \in Y$, which is absurd. In such a way, any fixed point free $g_o \in \mathrm{Aut}(Y, \beta(D))$ corresponds to a fixed point free $g = \Phi(g_o) \in \mathrm{Aut}(X, D)$. □

Proposition 20. *Let $X = (\mathbb{B}/\Gamma)'$ be a smooth toroidal compactification with toroidal compactifying divisor $D := X \setminus (\mathbb{B}/\Gamma)$ and a blow down $\beta : X \to Y$ of $n \in \mathbb{N}$ smooth irreducible rational (-1)-curves. Then $\mathrm{Aut}(X, D)$ is a finite group.*

Proof. By Proposition 19 (iii), $\mathrm{Aut}(X, D) = \mathrm{Aut}(X, D, E(\beta))$. Any $g \in \mathrm{Aut}(X, D)$ acts on $D = \coprod_{j=1}^{k} D_j$ and induces a permutation of the smooth elliptic

irreducible components $D_1, \ldots, D_k$ of D. In such a way, there arises a representation

$$\Sigma_1 : \mathrm{Aut}(X, D) \longrightarrow \mathrm{Sym}(D_1, \ldots, D_k) = \mathrm{Sym}(k).$$

The image of Σ_1 in the finite group $\mathrm{Sym}(k)$ is a finite group, so that it suffices to show the finiteness of $\ker(\Sigma_1)$, in order to conclude that $\mathrm{Aut}(X, D)$ is a finite group. Similarly, $\mathrm{Aut}(X, D) = \mathrm{Aut}(X, D, E(\beta))$ acts on the exceptional divisor $E(\beta) = \coprod_{i=1}^{n} L_i$ of $\beta : X \to Y$ and defines a representation

$$\Sigma_2 : \mathrm{Aut}(X, D) \longrightarrow \mathrm{Sym}(L_1, \ldots, L_n) = \mathrm{Sym}(n).$$

Since $\Sigma_2(\ker(\Sigma_1))$ is a finite group, it suffices to show that $G := \ker(\Sigma_2) \cap \ker(\Sigma_1)$ is a finite group. For any $1 \leq i \leq n$, $1 \leq j \leq k$ and $g \in G$, the finite set $L_i \cap D_j$ is transformed into itself, according to $g(L_i \cap D_j) \subseteq g(L_i) \cap g(D_j) = L_i \cap D_j$. Therefore, there is a representation

$$\Sigma_{i,j} : G \longrightarrow \mathrm{Sym}(L_i \cap D_j).$$

The image $\Sigma_{i,j}(G)$ is a finite group, while the kernel $K_{i,j} := \ker(\Sigma_{i,j})$ fixes any point $p \in L_i \cap D_j$ and acts on D_j. It is well known that the holomorphic automorphisms $\mathrm{Aut}_p(D_j)$ of an elliptic curves D_j, which fix a point $p \in D_j$, form a cyclic group of order 2, 4 or 6. Therefore, $K_{i,j} \leq \mathrm{Aut}_p(D)$, G, $\ker(\Sigma_1)$ and $\mathrm{Aut}(X, D)$ are finite groups. □

Definition 21. A smooth toroidal compactification $X = (\mathbb{B}/\Gamma)'$ with a blow down $\beta : X \to Y$ of $n \in \mathbb{N}$ smooth irreducible rational (-1)-curves onto a minimal surface Y is Galois non-primitive if there is a fixed point free automorphism $g \in \mathrm{Aut}(X, D) \setminus \{\mathrm{Id}_X\}$.

Any Galois non-primitive $X = (\mathbb{B}/\Gamma)'$ is non-primitive, because the $\langle g \rangle$-Galois covering $\zeta : X \to \zeta(X) = X/\langle g \rangle$ is unramified and restricts to unramified coverings $\zeta : \mathbb{B}/\Gamma \to \zeta(\mathbb{B}/\Gamma)$ and $\zeta : E(\beta) = \coprod_{i=1}^{n} L_i \to \zeta(E(\beta))$ of degree $|\langle g \rangle| = \mathrm{ord}(g)$.

Note that the presence of an unramified covering $\varphi : Y \to \varphi(Y)$ implies the coincidence $\widetilde{Y} = \widetilde{\varphi(Y)}$ of the universal cover $\widetilde{Y}$ of Y with the universal cover $\widetilde{\varphi(Y)}$ of $\varphi(Y)$. The fundamental group $\pi_1(\varphi(Y))$ of $\varphi(Y)$ acts on $\widetilde{Y}$ by biholomorphic automorphisms without fixed points and contains the fundamental group $\pi_1(Y)$ of Y as a subgroup of index $[\pi_1(\varphi(Y)) : \pi_1(Y)] = d$.

Proposition 22. *Let $X = (\mathbb{B}/\Gamma)'$ be a smooth toroidal compactification with toroidal compactifying divisor $D := X \setminus (\mathbb{B}/\Gamma)$, $\beta : X \to Y$ be a blow down of $n \in \mathbb{N}$ smooth irreducible rational (-1)-curves to a minimal surface Y and $N(\pi_1(Y))$ be the normalizer of the fundamental group $\pi_1(Y)$ of Y in the biholomorphism group $\mathrm{Aut}(\widetilde{Y})$ of the universal cover $\widetilde{Y}$ of Y. Then X is Galois non-primitive if and only if there exist a natural divisor $d > 1$ of $\mathrm{GCD}(\left|\beta(D)^{\mathrm{sing}}\right|, e(Y)) \in \mathbb{N}$ and an unramified covering $\varphi : Y \to \varphi(Y)$ of degree d, such that $\pi_1(\varphi(Y)) \cap N(\pi_1(Y)) \gneq \pi_1(Y)$ and $\varphi : \beta(D) \to \varphi\beta(D)$ is an unramified covering of degree d.*

Proof. If $X = (\mathbb{B}/\Gamma)'$ is Galois non-primitive then there exists a fixed point free biholomorphism $g \in \mathrm{Aut}(X, D) \setminus \{\mathrm{Id}_X\}$ of X. By Proposition 19(iv), g induces a fixed point free biholomorphism $g_o = \Phi^{-1}(g) \in \mathrm{Aut}(Y, \beta(D)) \setminus \{\mathrm{Id}_Y\}$ of Y. The element g_o of the finite group $\mathrm{Aut}(Y, \beta(D))$ is of finite order $d \in \mathbb{N} \setminus \{1\}$ and the $\langle g_o \rangle$-Galois coverings $\zeta : Y \to Y/\langle g_o \rangle$, $\zeta : \beta(D) \to \zeta\beta(D)$ are unramified and of degree d. The automorphism g_o of Y lifts to an automorphism $\sigma \in \mathrm{Aut}(\widetilde{Y})$ of the universal cover $\widetilde{Y}$ of Y, which normalizes $\pi_1(Y)$ and belongs to

$$\begin{aligned}\pi_1(\zeta(Y)) = \pi_1(Y/\langle g_o \rangle) &= \pi_1\left((\widetilde{Y}/\pi_1(Y))/\langle \sigma\pi_1(Y) \rangle\right) \\ &= \pi_1\left(\widetilde{Y}/\langle \sigma, \pi_1(Y) \rangle\right) = \langle \sigma, \pi_1(Y) \rangle.\end{aligned}$$

Conversely, suppose that $\varphi : Y \to \varphi(Y)$ is an unramified covering of degree $d > 1$, which restricts to an unramified covering $\varphi : \beta(D) \to \varphi\beta(D)$ of degree d and there exists $\sigma \in [\pi_1(\varphi(Y)) \cap N(\pi_1(Y))] \setminus \pi_1(Y)$. Then $g_o := \sigma\pi_1(Y) \in \mathrm{Aut}(Y) = N(\pi_1(Y))/\pi_1(Y)$ is a non-identical biholomorphism $g_o : Y \to Y$. Since $\langle \sigma, \pi_1(Y) \rangle$ is a subgroup of $\pi_1(\varphi(Y))$, the unramified covering $\varphi : Y \to \varphi(Y)$ factors through the $\langle g_o \rangle$-Galois covering $\zeta : Y \to Y/\langle g_o \rangle$ and a covering $\varphi_o : Y/\langle g_o \rangle \to \varphi(Y)$ along the commutative diagram

$$\begin{array}{ccc} Y & \xrightarrow{\ \zeta\ } & Y/\langle g_o \rangle \\ & {\scriptstyle\varphi}\searrow & \downarrow{\scriptstyle\varphi_o} \\ & & \varphi(Y) \end{array} \tag{6}$$

The finite coverings $\zeta : Y \to Y/\langle g_o \rangle$ and $\varphi_o : Y/\langle g_o \rangle \to \varphi(Y)$ are unramified, because their composition $\varphi = \varphi_o\zeta : Y \to \varphi(Y)$ is unramified. That is why g_o has no fixed points on Y. If $\beta(D) \subset Y$ is not $\langle g_o \rangle$-invariant then there is an orbit $\mathrm{Orb}_{\langle g_o \rangle}(y_o) \subset Y$ of some $y_o \in \beta(D)$ which intersects both $\beta(D)$ and $Y \setminus \beta(D)$. Therefore, $\zeta : \beta(D) \to \zeta\beta(D)$ has a fibre $\zeta^{-1}(\zeta(y_o))$ of cardinality $\left|\zeta^{-1}(\zeta(y_o))\right| < \deg(\zeta) = |\langle g_o \rangle| = \mathrm{ord}(g_o)$ and $\zeta : \beta(D) \to \zeta\beta(D)$ is ramified. As a result, the composition $\varphi = \varphi_o\zeta : \beta(D) \to \varphi\beta(D)$ is ramified. The contradiction shows the $\langle g_o \rangle$-invariance of $\beta(D)$. According to Proposition 19 (iv), the fixed point free $g_o \in \mathrm{Aut}(Y, \beta(D)) \setminus \{\mathrm{Id}_Y\}$ corresponds to a fixed point free $g = \Phi(g_o) \in \mathrm{Aut}(X, D) \setminus \{\mathrm{Id}_X\}$ and X is Galois non-primitive. □

Definition 23. A covering $\varphi : Y \to \varphi(Y)$ by a smooth projective surface Y has Galois factorization if there exist $g_o \in \mathrm{Aut}(Y) \setminus \{\mathrm{Id}_Y\}$ and a covering $\varphi_o : Y/\langle g_o \rangle \to \varphi(Y)$, such that $\varphi = \varphi_o\zeta$ factors through the $\langle g_o \rangle$-Galois covering $\zeta : Y \to Y/\langle g_o \rangle$ and a covering φ_o along the commutative diagram (6).

Now, Proposition 22 can be reformulated in the form of the following

Corollary 24. *Let $X = (\mathbb{B}/\Gamma)'$ be a non-primitive smooth toroidal compactification with toroidal compactifying divisor $D := X \setminus (\mathbb{B}/\Gamma)$, $\beta : X \to Y$ be a blow down of $n \in \mathbb{N}$ smooth irreducible rational (-1)-curves onto a minimal surface Y and $\varphi : Y \to \varphi(Y)$ be an unramified covering of degree d, which restricts to an unramified covering $\varphi : \beta(D) \to \varphi\beta(D)$ of degree d. Then X is Galois non-primitive if and only if φ admits a Galois factorization.*

Corollary 25. *(i) Let $X = (\mathbb{B}/\Gamma)'$ be a smooth toroidal compactification with abelian minimal model Y. Then X is not saturated and X is non-primitive if and only if it is Galois non-primitive.*

(ii) If $X = (\mathbb{B}/\Gamma)'$ is a smooth toroidal compactification with bi-elliptic minimal model Y then X is not saturated.

Proof. (i) Any abelian surface Y has non-trivial fundamental group $\pi_1(Y) \simeq (\mathbb{Z}^4, +)$. According to Corollary 9, that suffices for a smooth toroidal compactification $X = (\mathbb{B}/\Gamma)'$ with abelian minimal model Y to be non-saturated.

By Theorem 1.3 from Di Cerbo and Stover's article [3], if a smooth toroidal compactification $X = (\mathbb{B}/\Gamma)'$ has abelian minimal model Y then there is a blow down $\beta : X \to Y$ of $n \in \mathbb{N}$ smooth irreducible rational (-1)-curves on X onto Y. Such X is non-primitive if and only if there exists an unramified covering $\varphi : Y \to \varphi(Y)$ of degree $d > 1$, which restricts to an unramified covering $\varphi : \beta(D) \to \varphi\beta(D)$ of degree d. Since Y and $\varphi(Y)$ have one and the same universal cover $\widetilde{\varphi(Y)} = \widetilde{Y} = \mathbb{C}^2$ and one and the same Kodaira dimension $\kappa(\varphi(Y)) = \kappa(Y) = 0$, the minimal smooth irreducible projective surface $\varphi(Y)$ is abelian or bi-elliptic.

If $\varphi(Y)$ is an abelian surface then its fundamental group $\pi_1(\varphi(Y)) \simeq (\mathbb{Z}^4, +)$ is abelian and $\pi_1(Y) \simeq (\mathbb{Z}^4, +)$ is a normal subgroup of $\pi_1(\varphi(Y))$. As a result, $\varphi : Y \to \varphi(Y)$ is a $\pi_1(\varphi(Y))/\pi_1(Y)$-Galois covering and Y is Galois non-primitive.

Let us suppose that $\varphi(Y)$ is a bi-elliptic surface. According to Bagnera-de Franchis classification of the bi-elliptic surfaces from [1], there is an abelian surface A and a cyclic subgroup $\langle g \rangle \leq \mathrm{Aut}(A)$ of order $d \in \{2, 3, 4, 6\}$ with a non-translation generator $g \in \mathrm{Aut}(A)$, such that $\varphi(Y) = A/\langle g \rangle$. Let $\mathrm{AffLin}(\mathbb{C}) := \mathcal{T}(\mathbb{C}^2) \rtimes \mathrm{GL}(2, \mathbb{C})$ be the group of the affine linear transformations of $\mathbb{C}^2 = \widetilde{Y} = \widetilde{\varphi(Y)} = \widetilde{A}$ and

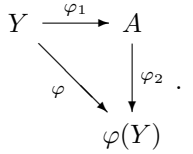
$$\mathcal{L} : \mathrm{AffLin}(\mathbb{C}^2) \longrightarrow \mathrm{GL}(2, \mathbb{C})$$

be the group homomorphism, associating to $\sigma \in \mathrm{AffLin}(\mathbb{C}^2)$ its linear part $\mathcal{L}(\sigma) \in \mathrm{GL}(2, \mathbb{C})$. Then the fundamental group of A is the maximal translation subgroup

$$\pi_1(A) = \pi_1(\varphi(Y)) \cap \ker(\mathcal{L})$$

of $\pi_1(\varphi(Y))$. The translation subgroup $\pi_1(Y) \leq \pi_1(\varphi(Y)) \cap \ker(\mathcal{L})$ of $\pi_1(\varphi(Y))$ is contained in $\pi_1(A)$ and the unramified covering $\varphi : Y \to \varphi(Y)$ factors through unramified coverings $\varphi_1 : Y \to A$ and $\varphi_2 : A \to \varphi(Y)$, along the commutative

diagram



The covering $\varphi_1 : Y \to A$ is $\pi_1(A)/\pi_1(Y)$-Galois, so that $\varphi = \varphi_2\varphi_1$ is a Galois factorization of φ for $\pi_1(Y) \lneq \pi_1(A)$. In the case of $\pi_1(Y) = \pi_1(A)$, there is an isomorphism $Y \simeq \mathbb{C}^2/\pi_1(Y) \simeq \mathbb{C}^2/\pi_1(A) = A$ and the covering $\varphi : Y \simeq A \to \varphi(Y) = A/\langle g\rangle$ is $\langle g\rangle$-Galois. Thus, X is Galois non-primitive and a co-abelian smooth toroidal compactification $X = (\mathbb{B}/\Gamma)'$ is non-primitive if and only if it is Galois non-primitive.

(ii) The fundamental group $\pi_1(Y)$ of a bi-elliptic surface Y is subject to an exact sequence

$$1 \longrightarrow \pi_1(Y) \cap \ker(\mathcal{L}) \longrightarrow \pi_1(Y) \longrightarrow \langle g\rangle \longrightarrow 1$$

with a non-translation cyclic subgroup $\langle g\rangle$ of $\mathrm{Aut}\left(\mathbb{C}^2/\pi_1(Y) \cap \ker(\mathcal{L})\right) = \mathrm{Aut}(A_o)$ of order 2, 3, 4 or 6. In particular, Y is not simply connected and a smooth toroidal compactification $X = (\mathbb{B}/\Gamma)'$ with bi-elliptic minimal model Y is not saturated. □

ACKNOWLEDGEMENT. This research was partially supported by Contract 80-10-209/17.04.2019 with the the Scientific Foundation of Kliment Ohridski University of Sofia.

4. REFERENCES

[1] Bagnera, G., de Franchis, M.: Sur les surfaces hyperelliptiques. *C. R. Acad. Sci.*, **145**, 1907, 747–749.

[2] Di Cerbo, L. F., Stover, M.: Multiple realizations of varieties as ball quotient compactifications. *Michigan Math. J.*, **65(2)**, 2016, 441–447.

[3] Di Cerbo, L. F., Stover, M.: Punctured spheres in complex hyperbolic surfaces and bi-elliptic ball quotient compactifications. *Trans. Amer. Math. Society*, **372**, 2019, 4627–4646.

[4] Grothendieck A.: *Éléments de géométrie algébrique: IV. Étude locale des schémas et des morphismes de schémas*, Quatrième partie. Publications mathématiques de l'I.H.E.S., tome 32, 1967, pp. 5–361.

[5] Hartshorne R.: *Algebraic Geometry*. Graduate Texts in Mathematics **52**, Springer, 1977.

[6] Stover, M.: Volumes of Picard modular surfaces. *Proc. Amer. Math. Soc.*, **139**, 2011, 3045–3056.

[7] Uludağ, M.: Covering relations between ball quotient orbifolds. *Math. Ann.*, **328(2)**, 2004, 503–523.

Received on January 3, 2020

Pancho G. Beshkov

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 blvd. J. Bourchier, BG-1164 Sofia
BULGARIA

E-mail: panchob@fmi.uni.sofia.bg

Azniv K. Kasparian

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 blvd. J. Bourchier, BG-1164 Sofia
BULGARIA

E-mail: kasparia@fmi.uni-sofia.bg

Gregory K. Sankaran

Department of Mathematical Sciences
University of Bath
Bath BA2 7AY
UNITED KINGDOM

E-mail: masgks@bath.ac.uk

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 106

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 106

SHAPE PRESERVING PROPERTIES OF THE BERNSTEIN POLYNOMIALS WITH INTEGER COEFFICIENTS

BORISLAV R. DRAGANOV

The Bernstein polynomials with integer coefficients do not generally preserve monotonicity and convexity. We establish sufficient conditions under which they do. We also observe that they are asymptotically shape preserving.

Keywords: Bernstein polynomials, integer coefficients, integral coefficients, shape preserving, monotone, convex.

2010 Math. Subject Classification: Primary: 41A10; Secondary: 41A29, 41A35, 41A36.

1. MAIN RESULTS

The Bernstein polynomial is defined for $n \in \mathbb{N}_+$, $f \in C[0,1]$ and $x \in [0,1]$ by

$$B_n f(x) := \sum_{k=0}^{n} f\left(\frac{k}{n}\right) p_{n,k}(x), \quad p_{n,k}(x) := \binom{n}{k} x^k (1-x)^{n-k}.$$

It is known that if $f \in C[0,1]$, then (see e.g. [1, Chapter 1, Theorem 2.3])

$$\lim_{n\to\infty} B_n f(x) = f(x) \quad \text{uniformly on } [0,1].$$

In order to show that any continuous function on $[0,1]$, which has integer values at the ends of the interval, can be approximated with algebraic polynomials with

integer coefficients, Kantorovich [5] introduced the operator

$$\widetilde{B}_n(f)(x) := \sum_{k=0}^{n} \left[f\left(\frac{k}{n}\right) \binom{n}{k} \right] x^k (1-x)^{n-k},$$

where $[\alpha]$ denotes the largest integer that is less than or equal to the real α. L. Kantorovich showed that if $f \in C[0,1]$ is such that $f(0), f(1) \in \mathbb{Z}$, then (see [5], or e.g. [4, pp. 3–4], or [6, Chapter 2, Theorem 4.1])

$$\lim_{n\to\infty} \widetilde{B}_n(f)(x) = f(x) \quad \text{uniformly on } [0,1].$$

Instead of the integer part we can take the nearest integer. More precisely, if $\alpha \in \mathbb{R}$ is not a half-integer, we set $\langle \alpha \rangle$ to be the integer at which the minimum $\min_{m\in\mathbb{Z}} |\alpha - m|$ is attained. When α is a half-integer, we can define $\langle \alpha \rangle$ to be either of the two neighbouring integers even without following a given rule. The results we will prove are valid regardless of our choice in the latter case. The integer modification of the Bernstein polynomial based on the nearest integer function is given by

$$\widehat{B}_n(f)(x) := \sum_{k=0}^{n} \left\langle f\left(\frac{k}{n}\right) \binom{n}{k} \right\rangle x^k (1-x)^{n-k}.$$

Similarly to [5], it is shown that

$$\lim_{n\to\infty} \widehat{B}_n(f)(x) = f(x) \quad \text{uniformly on } [0,1]$$

provided that $f \in C[0,1]$ and $f(0), f(1) \in \mathbb{Z}$.

Let us note that the operators $\widetilde{B}_n$ and $\widehat{B}_n$ are not linear for $n \geq 2$.

As is known, the Bernstein polynomials possess good shape preserving properties. In particular, if f is monotone, then $B_n f$ is monotone of the same type, or, if $f(x)$ is convex or concave then so is, respectively, $B_n f$ (see e.g. [1, Chapter 10, Theorem 3.3, (i) and (ii)]). Our main goal is to extend these assertions to the integer forms of the Bernstein polynomials.

The operators $\widetilde{B}_n$ and $\widehat{B}_n$ possess the property of simultaneous approximation, that is, the derivatives of $\widetilde{B}_n(f)$ and $\widehat{B}_n(f)$ approximate the corresponding derivatives of f in the uniform norm on $[0,1]$. This was established in [2, 3] under certain necessary and sufficient conditions, as estimates of the rate the convergence were proved. Hence, trivially, under these conditions, if $f^{(r)}(x)$ is strictly positive or negative, then so are $(\widetilde{B}_n(f))^{(r)}(x)$ and $(\widehat{B}_n(f))^{(r)}(x)$ at least for n large enough, depending on f. We will establish sufficient conditions on the shape of f that imply the corresponding monotonicity or convexity of $\widetilde{B}_n(f)$ and $\widehat{B}_n(f)$ for all n regardless of the smoothness of f.

The properties we will present below are not hard to prove. However, they seem interesting and might be useful in the applications of the approximation of

functions by polynomials with integer coefficients and in CAGD. Let us note that computer manipulation of polynomials with integer coefficients is faster.

The operators $\widetilde{B}_n$ and $\widehat{B}_n$ do not generally preserve monotonicity or convexity. We include counter examples in Section 4. It is quite straightforward to show that the monotonicity of $f(x)$ implies the monotonicity of the same type of $\widetilde{B}_n(f)(x)$ and $\widehat{B}_n(f)(x)$ for $n = 1$ and $n = 2$ (see Remark 2.1 below). However, both operators almost preserve monotonicity or convexity. In order to make this precise, we will introduce the notions of asymptotic monotonicity and convexity preservation.

Definition 1.1. Let X be a class of functions defined on $I \subseteq \mathbb{R}$ and $L_n : X \to X$, $n \in \mathbb{N}_+$, be a family of operators. We say that L_n *uniformly asymptotically preserves monotonicity* on X if there exist $n_0 \in \mathbb{N}_+$ and functions $\varepsilon_n, \eta_n : I \to \mathbb{R}$, $n \geq n_0$, with the properties:

(i) $\lim_{n\to\infty} \varepsilon_n(x) = \lim_{n\to\infty} \eta_n(x) = 0$ uniformly on I;

(ii) If $f(x)$ is monotone increasing on I, then so is $L_n(f)(x)+\varepsilon_n(x)$ for all $n \geq n_0$;

(iii) If $f(x)$ is monotone decreasing on I, then so is $L_n(f)(x)+\eta_n(x)$ for all $n \geq n_0$.

Remark 1.2. Let us note that conditions (ii) and (iii) are equivalent if the operators L_n are linear.

We will show that the following result holds.

Theorem 1.3. *The operators $\widetilde{B}_n$ and $\widehat{B}_n$ uniformly asymptotically preserve monotonicity on the class of continuous functions on $[0, 1]$ with integer values at 0 and 1.*

Similarly, we introduce the following notion.

Definition 1.4. Let X be a class of functions defined on $I \subseteq \mathbb{R}$ and $L_n : X \to X$, $n \in \mathbb{N}_+$, be a family of operators. We say that L_n *uniformly asymptotically preserves convexity* on X if there exist $n_0 \in \mathbb{N}_+$ and functions $\varepsilon_n, \eta_n : I \to \mathbb{R}$, $n \geq n_0$, with the properties:

(i) $\lim_{n\to\infty} \varepsilon_n(x) = \lim_{n\to\infty} \eta_n(x) = 0$ uniformly on I;

(ii) If $f(x)$ is convex on I, then so is $L_n(f)(x) + \varepsilon_n(x)$ for all $n \geq n_0$;

(iii) If $f(x)$ is concave on I, then so is $L_n(f)(x) + \eta_n(x)$ for all $n \geq n_0$.

Remark 1.5. As above, if the operators L_n are linear, then (ii) and (iii) are equivalent.

We will show that $\widetilde{B}_n$ and $\widehat{B}_n$ possess the property described in the definition.

Theorem 1.6. *The operators $\widetilde{B}_n$ and $\widehat{B}_n$ uniformly asymptotically preserve convexity on the set of continuous functions on* [0,1] *with integer values at* 0 *and* 1.

On the other hand, it will be useful to establish sufficient conditions on the function f under which we have that $\widetilde{B}_n(f)$ and $\widehat{B}_n(f)$ are monotone, or, respectively, convex or concave. A straightforward corollary of some of our main results is the following assertion.

Theorem 1.7. *Let $f : [0,1] \to \mathbb{R}$ and $f(0), f(1) \in \mathbb{Z}$.*

(a) If $f(x)-x$ is monotone increasing on $[0,1]$, then so are $\widetilde{B}_n(f)(x)$ and $\widehat{B}_n(f)(x)$ for all n.

(b) If $f(x)+x$ is monotone decreasing on $[0,1]$, then so are $\widetilde{B}_n(f)(x)$ and $\widehat{B}_n(f)(x)$ for all n.

Let us explicitly note that if $f(x) - x$ is monotone increasing on $[0,1]$, then so is $f(x)$, and similarly, if $f(x) + x$ is monotone decreasing on $[0,1]$, then so is $f(x)$.

Also, we will establish the following stronger result.

Theorem 1.8. *Let $f : [0,1] \to \mathbb{R}$ and $f(0), f(1) \in \mathbb{Z}$. Set for $n \in \mathbb{N}_+$ and $x \in [0,1]$*

$$\varphi_n(x) := (n+1) \int_0^1 t(1-t)^{n(1-x)} \frac{(1-t)^{nx-1} - t^{nx-1}}{1-2t}\, dt.$$

(a) If $f(x) - \varphi_n(x)$ is monotone increasing on $[0,1]$, then so are $\widetilde{B}_n(f)(x)$ and $\widehat{B}_n(f)(x)$.

(b) If $f(x) + \varphi_n(x)$ is monotone decreasing on $[0,1]$, then so are $\widetilde{B}_n(f)(x)$ and $\widehat{B}_n(f)(x)$.

As it follows from Remark 2.10, the function $\varphi_n(x)$ is monotone increasing on $[0,1]$ for each $n \in \mathbb{N}_+$ and it is of small magnitude – it satisfies the estimates

$$0 \le \varphi_n(x) \le \frac{6}{n}, \quad x \in \left[\frac{1}{n}, 1 - \frac{1}{n}\right].$$

In Section 2 we will establish even less restrictive conditions on f that imply the monotonicity of $\widetilde{B}_n(f)$ and $\widehat{B}_n(f)$. They show how to construct functions φ_n, which beside the property given in the theorem above, are also such that $|\varphi_n(x)| \le c/n$ for all $x \in [0,1]$ and all $n \in \mathbb{N}_+$, where c as an absolute positive constant; moreover, the functions φ_n can be constructed in such a way that if $f(x) \mp \varphi_{n_0}(x)$ is monotone increasing, respectively, decreasing on $[0,1]$ with some n_0, then so are $\widetilde{B}_n(f)(x)$ and $\widehat{B}_n(f)(x)$ for all $n \ge n_0$.

Concerning the preservation of convexity and concavity, we will establish

Theorem 1.9. *Let $f : [0,1] \to \mathbb{R}$ and $f(0), f(1) \in \mathbb{Z}$. Set*

$$\Phi(x) := 6\big(x \ln x + (1-x)\ln(1-x)\big).$$

(a) *If* $f(x)-\Phi(x)$ *is convex on* $[0,1]$*, then so are* $\widetilde{B}_n(f)(x)$ *and* $\widehat{B}_n(f)(x)$ *for all* $n\in\mathbb{N}_+$.

(b) *If* $f(x)+\Phi(x)$ *is concave on* $[0,1]$*, then so are* $\widetilde{B}_n(f)(x)$ *and* $\widehat{B}_n(f)(x)$ *for all* $n\in\mathbb{N}_+$.

Note that $\Phi(x)$ is convex and the assumption $f(x)\mp\Phi(x)$ is convex/concave implies that $f(x)$ is convex/concave, respectively.

A less restrictive sufficient condition is given in the following assertion.

Theorem 1.10. *Let* $f:[0,1]\to\mathbb{R}$ *and* $f(0),f(1)\in\mathbb{Z}$*. Set for* $n\in\mathbb{N}_+$*,* $n\geq 3$*, and* $x\in[0,1]$

$$\Phi_n(x):=(n+1)\int_0^1\left(t^2+(1-t)^2\right)\times\frac{(nx-3)t^2(1-t)^{n-2}-(nx-2)t^3(1-t)^{n-3}+t^{nx}(1-t)^{n(1-x)}}{(1-2t)^2}\,dt.$$

(a) *If* $f(x)-\Phi_n(x)$ *is convex on* $[0,1]$*, then so are* $\widetilde{B}_n(f)(x)$ *and* $\widehat{B}_n(f)(x)$.

(b) *If* $f(x)+\Phi_n(x)$ *is concave on* $[0,1]$*, then so are* $\widetilde{B}_n(f)(x)$ *and* $\widehat{B}_n(f)(x)$.

As we will establish in Proposition 3.4, the function $\Phi_n(x)$ is convex on $[0,1]$ and it is of small magnitude – it satisfies the estimates

$$-\frac{4}{n}\leq\Phi_n(x)\leq\frac{16}{n},\quad x\in\left[\frac{1}{n},1-\frac{1}{n}\right].$$

In Section 3 we will establish even less restrictive conditions on f that imply the convexity or concavity of $\widetilde{B}_n(f)$ and $\widehat{B}_n(f)$. They show how to construct functions Φ_n, which beside the property given in the theorem above, are also such that $|\Phi_n(x)|\leq c/n$ for all $x\in[0,1]$ and all $n\in\mathbb{N}_+$ with some absolute positive constant c; moreover, the functions Φ_n can be constructed in such a way that if $f(x)\mp\Phi_{n_0}(x)$ is convex, respectively, concave on $[0,1]$ with some n_0, then so are $\widetilde{B}_n(f)(x)$ and $\widehat{B}_n(f)(x)$ for all $n\geq n_0$.

We proceed to the proof of the results stated above. In the next section we will establish Theorem 1.3 as well as sufficient conditions that imply the monotonicity of $\widetilde{B}_n(f)(x)$ and $\widehat{B}_n(f)(x)$. In particular, we will get Theorems 1.7 and 1.8. In Section 3 we derive analogues of these results concerning convexity. We present several examples that illustrate the notion of the asymptotic shape preservation and some of the sufficient conditions stated above in Section 4.

2. PRESERVING MONOTONICITY

We set

$$\tilde{b}_n(k) := \left[f\left(\frac{k}{n}\right)\binom{n}{k}\right]\binom{n}{k}^{-1}$$

and

$$\hat{b}_n(k) := \left\langle f\left(\frac{k}{n}\right)\binom{n}{k}\right\rangle\binom{n}{k}^{-1},$$

where $k = 0, \ldots, n$. Then the operators $\widetilde{B}_n$ and $\widehat{B}_n$ can be written respectively in the form

$$\widetilde{B}_n(f)(x) = \sum_{k=0}^{n} \tilde{b}_n(k)\, p_{n,k}(x)$$

and

$$\widehat{B}_n(f)(x) = \sum_{k=0}^{n} \hat{b}_n(k)\, p_{n,k}(x).$$

For their first derivatives we have (by direct computation, or see [7] or [1, Chapter 10, (2.3)])

$$(\widetilde{B}_n(f))'(x) = n\sum_{k=0}^{n-1}\left(\tilde{b}_n(k+1) - \tilde{b}_n(k)\right) p_{n-1,k}(x) \tag{2.1}$$

and

$$(\widehat{B}_n(f))'(x) = n\sum_{k=0}^{n-1}\left(\hat{b}_n(k+1) - \hat{b}_n(k)\right) p_{n-1,k}(x). \tag{2.2}$$

Proof of Theorem 1.3. First, let $f(x)$ be monotone increasing on $[0,1]$. We will estimate from below $(\widetilde{B}_n(f))'(x)$.

Since $f(x)$ is increasing on $[0, 1/n]$, then

$$nf\left(\frac{1}{n}\right) \geq nf(0).$$

We have that $nf(0) \in \mathbb{Z}$; consequently

$$\left[nf\left(\frac{1}{n}\right)\right] \geq nf(0).$$

Hence

$$\tilde{b}_n(1)-\tilde{b}_n(0)=\frac{1}{n}\left(\left[nf\left(\frac{1}{n}\right)\right]-nf(0)\right)\geq 0. \tag{2.3}$$

Next, using $[\alpha]\leq\alpha$, we arrive at

$$\begin{aligned}\tilde{b}_n(n)-\tilde{b}_n(n-1)&=f(1)-\frac{1}{n}\left[nf\left(\frac{n-1}{n}\right)\right]\\&\geq f(1)-f\left(\frac{n-1}{n}\right)\geq 0.\end{aligned} \tag{2.4}$$

Now, let $1\leq k\leq n-2$, $n\geq 3$. Using the trivial inequalities $\alpha-1\leq[\alpha]\leq\alpha$, we get

$$\begin{aligned}\tilde{b}_n(k+1)-\tilde{b}_n(k)&=\left[f\left(\frac{k+1}{n}\right)\binom{n}{k+1}\right]\binom{n}{k+1}^{-1}-\left[f\left(\frac{k}{n}\right)\binom{n}{k}\right]\binom{n}{k}^{-1}\\&\geq\left(f\left(\frac{k+1}{n}\right)\binom{n}{k+1}-1\right)\binom{n}{k+1}^{-1}-f\left(\frac{k}{n}\right)\\&=f\left(\frac{k+1}{n}\right)-f\left(\frac{k}{n}\right)-\binom{n}{k+1}^{-1}.\end{aligned} \tag{2.5}$$

Therefore

$$\tilde{b}_n(k+1)-\tilde{b}_n(k)\geq-\binom{n}{k+1}^{-1},\quad 1\leq k\leq n-2,\ n\geq 3. \tag{2.6}$$

Below we follow the convention that a sum, whose lower index bound is larger than the upper one, is identically 0.

We combine (2.3), (2.4) and (2.6) with (2.1), and use the inequality $\binom{n}{k+1}\geq\binom{n}{2}$ for $k=1,\dots,n-3$, and the identity $\sum_{k=0}^{n-1}p_{n-1,k}(x)\equiv 1$ to arrive at

$$\begin{aligned}(\widetilde{B}_n(f))'(x)&\geq n\sum_{k=1}^{n-2}\left(\tilde{b}_n(k+1)-\tilde{b}_n(k)\right)p_{n-1,k}(x)\\&\geq-\frac{2}{n-1}\sum_{k=1}^{n-3}p_{n-1,k}(x)-(n-1)x^{n-2}(1-x)\\&\geq-\frac{2}{n-1}-(n-1)x^{n-2}(1-x).\end{aligned}$$

We set $\varepsilon_n(x):=2x/(n-1)+x^n/n+x^{n-1}(1-x)$, $n\geq 2$. It satisfies condition (i). Its derivative is $\varepsilon_n'(x)=2/(n-1)+(n-1)x^{n-2}(1-x)$; hence $\varepsilon_n(x)$ satisfies (ii) in Definition 1.1 with $n_0=2$.

The case of monotone decreasing functions is reduced to the one of monotone increasing by applying the latter to the function $\bar{f}(x):=f(1-x)$ and using that $\widetilde{B}_n(f)(x)=\widetilde{B}_n(\bar{f})(1-x)$.

The considerations for the operator $\widehat{B}_n$ are quite similar as we use that $|\alpha - \langle\alpha\rangle| \leq 1/2$. □

Remark 2.1. Formula (2.1) and estimates (2.3) and (2.4) show that if $f(x)$ is monotone increasing on $[0,1]$, then so are $\widetilde{B}_1(f)(x)$ and $\widetilde{B}_2(f)(x)$. Similarly, we have that $\widehat{B}_1(f)(x)$ and $\widehat{B}_2(f)(x)$ are monotone increasing if $f(x)$ is such.

We proceed to establishing sufficient conditions on f that imply the monotonicity of $\widetilde{B}_n(f)$ and $\widehat{B}_n(f)$. We first consider the operator $\widetilde{B}_n$ and the case of monotone increasing functions.

Proposition 2.2. *Let $f : [0,1] \to \mathbb{R}$, $f(0), f(1) \in \mathbb{Z}$ and $n \in \mathbb{N}_+$. If $n \geq 3$, let also $\phi_n : [0,1] \to \mathbb{R}$ be such that*

$$\phi_n\left(\frac{k+1}{n}\right) - \phi_n\left(\frac{k}{n}\right) \geq \binom{n}{k+1}^{-1}, \quad k = 1, \ldots, n-2. \tag{2.7}$$

If $f(x)$ is monotone increasing on $[0, 1/n]$ and on $[1-1/n, 1]$ and if $f(x) - \phi_n(x)$ is monotone increasing on $[1/n, 1-1/n]$, then $\widetilde{B}_n(f)(x)$ is monotone increasing on $[0,1]$.

Proof. We will show that $\tilde{b}_n(k+1) - \tilde{b}_n(k) \geq 0$, $k = 0, \ldots, n-1$. Then, by virtue of (2.1) we will have $(\widetilde{B}_n(f))'(x) \geq 0$ on $[0,1]$.

As we have already established in (2.3) and (2.4), $\tilde{b}_n(k+1) - \tilde{b}_n(k) \geq 0$ for $k = 0$ and $k = n-1$.

Let $1 \leq k \leq n-2$, $n \geq 3$. Since $f(x) - \phi_n(x)$ is monotone increasing on $[1/n, 1-1/n]$ and $\phi_n(x)$ satisfies (2.7), we have

$$\begin{aligned} f\left(\frac{k+1}{n}\right) - f\left(\frac{k}{n}\right) &\geq \phi_n\left(\frac{k+1}{n}\right) - \phi_n\left(\frac{k}{n}\right) \\ &\geq \binom{n}{k+1}^{-1}. \end{aligned}$$

Then (2.5) implies that $\tilde{b}_n(k+1) - \tilde{b}_n(k) \geq 0$, $k = 1, \ldots, n-2$. The proof is completed. □

Clearly, the function $\phi_n(x) := x$ satisfies (2.7) for all $n \in \mathbb{N}_+$; hence Theorem 1.7(a) follows for the operator $\widetilde{B}_n$. The next corollary contains a less restrictive choice of $\phi_n(x)$. Actually, the function, defined in it, satisfies (2.7) as an equality.

Corollary 2.3. *Let $f : [0,1] \to \mathbb{R}$ and $f(0), f(1) \in \mathbb{Z}$. Let $n \in \mathbb{N}_+$, $n \geq 3$, be fixed. Set*

$$\phi_n(x) := (n+1) \int_0^1 t(1-t)^{n(1-x)} \frac{(1-t)^{nx} - t^{nx}}{1-2t}\, dt, \quad x \in [0,1]. \tag{2.8}$$

If $f(x) - \phi_n(x)$ is monotone increasing on $[0,1]$, then so is $\widetilde{B}_n(f)(x)$.

Proof. The motivation for the definition of $\phi_n(x)$ comes from the following formula, which is derived by the relationship between the beta and gamma functions (see e.g. [9] or [[1, Chapter 10, (1.8)]). We have

$$\begin{aligned}\int_0^1 t^k(1-t)^{n-k}dt &= B(k+1,n-k+1)\\ &= \frac{\Gamma(k+1)\,\Gamma(n-k+1)}{\Gamma(n+2)}\\ &= \frac{1}{n+1}\binom{n}{k}^{-1}.\end{aligned} \tag{2.9}$$

Consequently, for $k=1,\dots,n-2$ we have

$$\begin{aligned}\phi_n\left(\frac{k+1}{n}\right)-\phi_n\left(\frac{k}{n}\right) &= (n+1)\int_0^1 t^{k+1}(1-t)^{n-(k+1)}dt\\ &= \binom{n}{k+1}^{-1}.\end{aligned} \tag{2.10}$$

Thus $\phi_n(x)$ satisfies (2.7).

It remains to observe that $\phi_n(x)$ is differentiable and

$$\phi_n'(x) = n(n+1)\int_0^1 t^{nx+1}(1-t)^{n(1-x)}\frac{\ln(1-t)-\ln t}{1-2t}\,dt > 0, \quad x\in[0,1].$$

Therefore $\phi_n(x)$ is monotone increasing on $[0,1]$; hence so is $f(x)$.

Now, the assertion of the corollary follows from Proposition 2.2. □

Remark 2.4. The function $\phi_n(x)$, defined in (2.8), can be represented in the following symmetric form

$$\phi_n(x) = \frac{n+1}{2}\int_0^1 \frac{t(1-t)\left((1-t)^{n(1-x)-1}+t^{n(1-x)-1}\right)\left((1-t)^{nx}-t^{nx}\right)}{1-2t}\,dt.$$

Next, we will note the following elementary estimates for the function $\phi_n(x)$, defined in (2.8).

Lemma 2.5. *The function $\phi_n(x)$, defined in* (2.8)*, satisfies the estimates*

$$0\le\phi_n(x)\le\frac{4}{n}, \quad x\in\left[0,1-\frac{1}{n}\right], \quad n\ge 3.$$

Proof. As we noted in the proof of Corollary 2.3, $\phi_n(x)$ is monotone increasing; hence

$$\phi_n(0)\le\phi_n(x)\le\phi_n\left(1-\frac{1}{n}\right), \quad x\in\left[0,1-\frac{1}{n}\right].$$

Clearly, $\phi_n(0) = 0$.

Next, summing the equalities in (2.10) on $k = 1, \dots, n-2$, we arrive at

$$\phi_n\left(1 - \frac{1}{n}\right) - \phi_n\left(\frac{1}{n}\right) = \sum_{k=2}^{n-1} \binom{n}{k}^{-1}.$$

In view of (2.8) and (2.9) with $k = 1$, we have

$$\phi_n\left(\frac{1}{n}\right) = (n+1)\int_0^1 t(1-t)^{n-1}\,dt = \binom{n}{1}^{-1}.$$

It remains to take into account that $\binom{n}{k} \geq \binom{n}{2}$ for $k = 2, \dots, n-2$, to deduce that

$$\begin{aligned}\phi_n\left(1 - \frac{1}{n}\right) &\leq \binom{n}{1}^{-1} + (n-3)\binom{n}{2}^{-1} + \binom{n}{n-1}^{-1} \\ &\leq \frac{4}{n}.\end{aligned}$$

□

Rockett [8, Theorem 1] established a neat formula for the sum of the reciprocals of the binomial coefficients.

Since generally $[-\alpha] \neq -[\alpha]$ and $\langle -\alpha \rangle \neq -\langle \alpha \rangle$ (however, $\langle \alpha \rangle$ is an odd function for some definitions of the nearest integer), the cases of monotone decreasing or concave functions cannot be reduced, respectively, to the cases of increasing or convex functions by considering $-f$ in place of f. However, we can swap between increasing and decreasing functions using the transformation $\bar{f}(x) := f(1-x)$. Thus we derive the following sufficient condition concerning the preservation of the monotone decreasing behaviour from Proposition 2.2.

Proposition 2.6. *Let $f : [0,1] \to \mathbb{R}$, $f(0), f(1) \in \mathbb{Z}$ and $n \in \mathbb{N}_+$. If $n \geq 3$, let also $\psi_n : [0,1] \to \mathbb{R}$ be such that*

$$\psi_n\left(\frac{k+1}{n}\right) - \psi_n\left(\frac{k}{n}\right) \geq \binom{n}{k}^{-1}, \quad k = 1, \dots, n-2. \tag{2.11}$$

If $f(x)$ is monotone decreasing on $[0, 1/n]$ and on $[1-1/n, 1]$ and if $f(x)+\psi_n(x)$ is monotone decreasing on $[1/n, 1-1/n]$, then $\widetilde{B}_n(f)(x)$ is monotone decreasing on $[0,1]$.

The second assertion of Theorem 1.7 concerning the operator $\widetilde{B}_n$ follows from the last proposition with $\psi_n(x) := x$. A less restrictive ψ_n is defined in the following corollary of Proposition 2.6.

Corollary 2.7. *Let $f : [0,1] \to \mathbb{R}$ and $f(0), f(1) \in \mathbb{Z}$. Let $n \in \mathbb{N}_+$, $n \geq 3$, be fixed. Set*

$$\psi_n(x) := (n+1) \int_0^1 t(1-t)^{n(1-x)+1} \frac{(1-t)^{nx-1} - t^{nx-1}}{1-2t}\, dt, \quad t \in [0,1].$$

If $f(x) + \psi_n(x)$ is monotone decreasing on $[0,1]$, then so is $\widetilde{B}_n(f)(x)$.

Proof. The assertion is established similarly to Corollary 2.3 as instead of (2.10) we show that

$$\psi_n\left(\frac{k+1}{n}\right) - \psi_n\left(\frac{k}{n}\right) = \binom{n}{k}^{-1}$$

for $k = 1, \ldots, n-2$. The function $\psi_n(x)$ is monotone increasing. □

Remark 2.8. Similarly to Lemma 2.5, it is shown that $\psi_n(x)$, defined in Corollary 2.7, satisfies

$$0 \leq \psi_n(x) \leq \frac{4}{n}, \quad x \in \left[\frac{1}{n}, 1\right].$$

Analogous results hold for the operator $\widehat{B}_n$. They are verified similarly to Proposition 2.2, as we use $|\alpha - \langle \alpha \rangle| \leq 1/2$. Let us note that now the assumptions concerning the two types of monotonicity are symmetric unlike those for the operator $\widetilde{B}_n$.

Proposition 2.9. *Let $f : [0,1] \to \mathbb{R}$, $f(0), f(1) \in \mathbb{Z}$ and $n \in \mathbb{N}_+$. If $n \geq 3$, let also $\tilde{\varphi}_n : [0,1] \to \mathbb{R}$ be such that*

$$\tilde{\varphi}_n\left(\frac{k+1}{n}\right) - \tilde{\varphi}_n\left(\frac{k}{n}\right) \geq \frac{1}{2}\left(\binom{n}{k}^{-1} + \binom{n}{k+1}^{-1}\right), \quad k = 1, \ldots, n-2. \quad (2.12)$$

(a) If $f(x)$ is monotone increasing on $[0, 1/n]$ and on $[1 - 1/n, 1]$ and if $f(x) - \tilde{\varphi}_n(x)$ is monotone increasing on $[1/n, 1 - 1/n]$, then $\widehat{B}_n(f)(x)$ is monotone increasing on $[0,1]$.

(b) If $f(x)$ is monotone decreasing on $[0, 1/n]$ and on $[1 - 1/n, 1]$ and if $f(x) + \tilde{\varphi}_n(x)$ is monotone decreasing on $[1/n, 1 - 1/n]$, then $\widehat{B}_n(f)(x)$ is monotone decreasing on $[0,1]$.

The assertions of Theorem 1.7 for $\widehat{B}_n(f)(x)$ follow from the last proposition with $\tilde{\varphi}_n(x) := x$.

Remark 2.10. Another function satisfying (2.12) is

$$\tilde{\varphi}_n(x) := \frac{n+1}{2} \int_0^1 t(1-t)^{n(1-x)} \frac{(1-t)^{nx-1} - t^{nx-1}}{1-2t}\, dt, \quad x \in [0,1]. \tag{2.13}$$

As in the previous cases, it is shown that it is differentiable, as

$$\tilde{\varphi}_n'(x) = \frac{n(n+1)}{2} \int_0^1 t^{nx}(1-t)^{n(1-x)} \frac{\ln(1-t) - \ln t}{1-2t}\, dt, \quad x \in [0,1]. \tag{2.14}$$

Consequently, $\tilde{\varphi}_n(x)$ is monotone increasing on $[0,1]$ and satisfies the estimates

$$0 \le \tilde{\varphi}_n(x) \le \frac{3}{n}, \quad x \in \left[\frac{1}{n}, 1 - \frac{1}{n}\right].$$

Proof of Theorem 1.8. The function $\tilde{\varphi}_n(x)$, defined in (2.13), satisfies (2.12). Then $\varphi_n(x) := 2\tilde{\varphi}_n(x)$ satisfies the conditions (2.7), (2.11) and (2.12). Since $f(x) - \varphi_n(x)$ is monotone increasing on $[0,1]$, then so is $f(x)$. Now, Propositions 2.2 and 2.9(a) yield that $\widetilde{B}_n(f)(x)$ and $\widehat{B}_n(f)(x)$ are monotone increasing on $[0,1]$. The proof of assertion (b) of the theorem is similar. □

3. PRESERVING CONVEXITY

For the second derivatives of $\widetilde{B}_n(f)$ and $\widehat{B}_n(f)$ we have (by direct computation, or see [7] or [1, Chapter 10, (2.3)])

$$(\widetilde{B}_n(f))''(x) = n(n-1) \sum_{k=0}^{n-2} \left(\tilde{b}_n(k+2) - 2\tilde{b}_n(k+1) + \tilde{b}_n(k)\right) p_{n-2,k}(x) \tag{3.1}$$

and

$$(\widehat{B}_n(f))''(x) = n(n-1) \sum_{k=0}^{n-2} \left(\hat{b}_n(k+2) - 2\hat{b}_n(k+1) + \hat{b}_n(k)\right) p_{n-2,k}(x). \tag{3.2}$$

Proof of Theorem 1.6. Similarly to the proof of the corresponding result in the monotone case, we estimate the second derivative of $\widetilde{B}_n(f)(x)$ and $\widehat{B}_n(f)(x)$. We will consider in detail only the former operator in the case of convex functions; the arguments for the latter operator are quite alike. The case of concave functions is analogous too.

Let $f(x)$ be convex on the interval $[0,1]$. Then

$$f\left(\frac{k+2}{n}\right) - 2f\left(\frac{k+1}{n}\right) + f\left(\frac{k}{n}\right) \ge 0, \quad k = 0, \ldots, n-2,\ n \ge 2.$$

Using $\alpha - 1 \le [\alpha] \le \alpha$ and $f(0) \in \mathbb{Z}$, we get

$$\begin{aligned}\tilde{b}_n(2)-2\tilde{b}_n(1)+\tilde{b}_n(0) &= \left[f\left(\frac{2}{n}\right)\binom{n}{2}\right]\binom{n}{2}^{-1} - 2\left[f\left(\frac{1}{n}\right)\binom{n}{1}\right]\binom{n}{1}^{-1} + f(0)\\ &\ge \left(f\left(\frac{2}{n}\right)\binom{n}{2} - 1\right)\binom{n}{2}^{-1} - 2f\left(\frac{1}{n}\right) + f(0)\\ &= f\left(\frac{2}{n}\right) - 2f\left(\frac{1}{n}\right) + f(0) - \binom{n}{2}^{-1}.\end{aligned} \tag{3.3}$$

Similarly, we get for $n \ge 2$

$$\tilde{b}_n(1) - 2\tilde{b}_n(n-1) + \tilde{b}_n(n-2) \ge f(1) - 2f\left(\frac{n-1}{n}\right) + f\left(\frac{n-2}{n}\right) - \binom{n}{n-2}^{-1}. \tag{3.4}$$

Let $k = 1, \dots, n-3$, $n \ge 4$. Just analogously, we arrive at the estimates

$$\begin{aligned}&\tilde{b}_n(k+2) - 2\tilde{b}_n(k+1) + \tilde{b}_n(k)\\ &\quad= \left[f\left(\frac{k+2}{n}\right)\binom{n}{k+2}\right]\binom{n}{k+2}^{-1} - 2\left[f\left(\frac{k+1}{n}\right)\binom{n}{k+1}\right]\binom{n}{k+1}^{-1}\\ &\qquad + \left[f\left(\frac{k}{n}\right)\binom{n}{k}\right]\binom{n}{k}^{-1}\\ &\quad\ge \left(f\left(\frac{k+2}{n}\right)\binom{n}{k+2} - 1\right)\binom{n}{k+2}^{-1} - 2f\left(\frac{k+1}{n}\right)\\ &\qquad + \left(f\left(\frac{k}{n}\right)\binom{n}{k} - 1\right)\binom{n}{k}^{-1}\\ &\quad\ge f\left(\frac{k+2}{n}\right) - 2f\left(\frac{k+1}{n}\right) + f\left(\frac{k}{n}\right) - \left(\binom{n}{k}^{-1} + \binom{n}{k+2}^{-1}\right).\end{aligned} \tag{3.5}$$

Thus we have shown that

$$\tilde{b}_n(2) - 2\tilde{b}_n(1) + \tilde{b}_n(0) \ge -\binom{n}{2}^{-1}, \quad \tilde{b}_n(1) - 2\tilde{b}_n(n-1) + \tilde{b}_n(n-2) \ge -\binom{n}{n-2}^{-1},$$

$$\tilde{b}_n(k+2) - 2\tilde{b}_n(k+1) + \tilde{b}_n(k) \ge -\left(\binom{n}{k}^{-1} + \binom{n}{k+2}^{-1}\right), \quad k = 1, \dots, n-3\,.$$

Consequently, from (3.1) we obtain

$$\begin{aligned}(\widetilde{B}_n(f))''(x) \ge& -n(n-1)\binom{n}{2}^{-1}(1-x)^{n-2} - n(n-1)\binom{n}{n-2}^{-1}x^{n-2}\\ &- n(n-1)\sum_{k=1}^{n-3}\left(\binom{n}{k}^{-1} + \binom{n}{k+2}^{-1}\right)p_{n-2,k}(x)\,.\end{aligned}$$

Next, by virtue of the inequality $\binom{n}{k} \geq \binom{n}{3}$ for $k = 3, \ldots, n-3$, and the identity $\sum_{k=0}^{n-2} p_{n-2,k}(x) \equiv 1$, we have for $n \geq 6$

$$\begin{aligned}
(\widetilde{B}_n(f))''(x) \geq & -2(1-x)^{n-2} - (n-1)(n-2)x(1-x)^{n-3} \\
& - (n-2)(n-3)x^2(1-x)^{n-4} - n(n-1)\binom{n}{3}^{-1} \sum_{k=3}^{n-3} p_{n-2,k}(x) \\
& - n(n-1)\binom{n}{3}^{-1} \sum_{k=1}^{n-5} p_{n-2,k}(x) - (n-2)(n-3)x^{n-4}(1-x)^2 \\
& - (n-1)(n-2)x^{n-3}(1-x) - 2x^{n-2} \\
\geq & -2(1-x)^{n-2} - (n-1)(n-2)x(1-x)^{n-3} \\
& - (n-2)(n-3)x^2(1-x)^{n-4} - \frac{12}{n-2} \\
& - (n-2)(n-3)x^{n-4}(1-x)^2 - (n-1)(n-2)x^{n-3}(1-x) \\
& - 2x^{n-2} =: -b_n(x).
\end{aligned}$$

We set

$$\begin{aligned}
\varepsilon_n(x) := & \frac{6x^2}{n-2} - \frac{2(n-3)}{n(n-1)}\left(x^n + (1-x)^n\right) \\
& + \frac{4}{n-1}\left(x^{n-1} + (1-x)^{n-1}\right) + x^{n-2}(1-x) + x(1-x)^{n-2}.
\end{aligned}$$

We have $\varepsilon_n''(x) = b_n(x)$; hence $\varepsilon_n(x)$ satisfies condition (ii) in Definition 1.4 with $n_0 = 6$. Clearly, it satisfies condition (i) too. □

Further, we will derive sufficient conditions on the function f that imply the convexity and concavity of $\widetilde{B}_n(f)(x)$ and $\widehat{B}_n(f)(x)$.

Proposition 3.1. *Let $f : [0,1] \to \mathbb{R}$ and $f(0), f(1) \in \mathbb{Z}$. Let $n \in \mathbb{N}_+$, $n \geq 2$, be fixed and $\Phi_n : [0,1] \to \mathbb{R}$ be such that*

$$\begin{aligned}
& \Phi_n\left(\frac{2}{n}\right) - 2\Phi_n\left(\frac{1}{n}\right) + \Phi_n(0) \geq \binom{n}{2}^{-1} \\
& \Phi_n\left(\frac{k+2}{n}\right) - 2\Phi_n\left(\frac{k+1}{n}\right) + \Phi_n\left(\frac{k}{n}\right) \\
& \qquad \geq \binom{n}{k}^{-1} + \binom{n}{k+2}^{-1}, \quad k = 1, \ldots, n-3,\ n \geq 4,
\end{aligned}$$

and

$$\Phi_n(1) - 2\Phi_n\left(\frac{n-1}{n}\right) + \Phi_n\left(\frac{n-2}{n}\right) \geq \binom{n}{n-2}^{-1}.$$

If $f(x) - \Phi_n(x)$ is convex on $[0,1]$, then so is $\widetilde{B}_n(f)(x)$.

Proof. Since $f(x)-\Phi_n(x)$ is convex on $[0,1]$, then

$$f\left(\frac{k+2}{n}\right)-2f\left(\frac{k+1}{n}\right)+f\left(\frac{k}{n}\right)$$
$$\geq \Phi_n\left(\frac{k+2}{n}\right)-2\Phi_n\left(\frac{k+1}{n}\right)+\Phi_n\left(\frac{k}{n}\right), \quad k=0,\dots,n-2.$$

Then (3.3)-(3.5) and the assumptions on $\Phi_n(x)$ imply

$$\tilde{b}_n(k+2)-2\tilde{b}_n(k+1)+\tilde{b}_n(k)\geq 0, \quad k=0,\dots,n-2,$$

which, by virtue of (3.1), completes the proof of the proposition. □

Similarly to Proposition 3.1 we prove the following sufficient condition for preserving concavity.

Proposition 3.2. *Let $f:[0,1]\to\mathbb{R}$ and $f(0),f(1)\in\mathbb{Z}$. Let $n\in\mathbb{N}_+$, $n\geq 2$, be fixed and $\Phi_n:[0,1]\to\mathbb{R}$ be such that*

$$\Phi_n\left(\frac{k+2}{n}\right)-2\Phi_n\left(\frac{k+1}{n}\right)+\Phi_n\left(\frac{k}{n}\right)\geq 2\binom{n}{k+1}^{-1}, \quad k=0,\dots,n-2.$$

If $f(x)+\Phi_n(x)$ is concave on $[0,1]$, then so is $\widetilde{B}_n(f)(x)$.

Similarly to Propositions 3.1 and 3.2 , we have the following result for the other integer modification of the Bernstein polynomials, the operator $\widehat{B}_n$.

Proposition 3.3. *Let $f:[0,1]\to\mathbb{R}$ and $f(0),f(1)\in\mathbb{Z}$. Let $n\in\mathbb{N}_+$, $n\geq 2$, be fixed and $\Phi_n:[0,1]\to\mathbb{R}$ be such that*

$$\Phi_n\left(\frac{2}{n}\right)-2\Phi_n\left(\frac{1}{n}\right)+\Phi_n(0)\geq\frac{1}{2}\left(2\binom{n}{1}^{-1}+\binom{n}{2}^{-1}\right),$$
$$\Phi_n\left(\frac{k+2}{n}\right)-2\Phi_n\left(\frac{k+1}{n}\right)+\Phi_n\left(\frac{k}{n}\right)$$
$$\geq\frac{1}{2}\left(\binom{n}{k}^{-1}+2\binom{n}{k+1}^{-1}+\binom{n}{k+2}^{-1}\right), \quad k=1,\dots,n-3,\ n\geq 4,$$

and

$$\Phi_n(1)-2\Phi_n\left(\frac{n-1}{n}\right)+\Phi_n\left(\frac{n-2}{n}\right)\geq\frac{1}{2}\left(\binom{n}{n-2}^{-1}+2\binom{n}{n-1}^{-1}\right).$$

(a) If $f(x)-\Phi_n(x)$ is convex on $[0,1]$, then so is $\widehat{B}_n(f)(x)$.

(b) If $f(x)+\Phi_n(x)$ is concave on $[0,1]$, then so is $\widehat{B}_n(f)(x)$.

We proceed to the proof of Theorem 1.9.

Proof of Theorem 1.9. For $n = 1$ the assertion is trivial since $\widetilde{B}_1(f)(x)$ and $\widehat{B}_1(f)(x)$ are linear functions. Let $n \geq 2$. We will verify that the function $\Phi(x)$ defined in the theorem satisfies the conditions in the propositions stated so far in this section. We set

$$\Delta(k) := \Phi\left(\frac{k+2}{n}\right) - 2\Phi\left(\frac{k+1}{n}\right) + \Phi\left(\frac{k}{n}\right), \quad k = 0, \ldots, n-2.$$

First, we observe that

$$2\binom{n}{1}^{-1} \geq \frac{1}{2}\left(2\binom{n}{1}^{-1} + \binom{n}{2}^{-1}\right) \geq \binom{n}{2}^{-1}, \tag{3.6}$$

$$2\binom{n}{n-1}^{-1} \geq \frac{1}{2}\left(\binom{n}{n-2}^{-1} + 2\binom{n}{n-1}^{-1}\right) \geq \binom{n}{n-2}^{-1}, \tag{3.7}$$

$$\binom{n}{k}^{-1} + \binom{n}{k+2}^{-1} \geq \frac{1}{2}\left(\binom{n}{k}^{-1} + 2\binom{n}{k+1}^{-1} + \binom{n}{k+2}^{-1}\right) \tag{3.8}$$

$$k = 1, \ldots, n-3,\ n \geq 4,$$

and

$$\binom{n}{k}^{-1} + \binom{n}{k+2}^{-1} \geq 2\binom{n}{k+1}^{-1}, \quad k = 1, \ldots, n-3,\ n \geq 4. \tag{3.9}$$

Relations (3.6) and (3.7) are identical and trivial. It is straightforward to see that (3.8) and (3.9) are equivalent too. Let us verify the last one. It reduces to

$$(n-k-1)(n-k) + (k+1)(k+2) \geq 2(k+1)(n-k-1).$$

We divide both sides of the inequality above by $(k+1)(n-k-1)$, to arrive at

$$\frac{n-k}{k+1} + \frac{k+2}{n-k-1} \geq 2.$$

It remains to observe that the second term on the left hand-side is larger than the reciprocal of the first one and then to take into account that the sum of a positive real and its reciprocal is always at least 2.

Thus to show that $\Phi(x)$ satisfies the assumptions in Propositions 3.1, 3.2 and 3.3, it is sufficient to prove that

$$\Delta(k) \geq 2\binom{n}{1}^{-1} = \frac{2}{n}, \quad k = 0, n-2, \tag{3.10}$$

and

$$\Delta(k) \geq \binom{n}{k}^{-1} + \binom{n}{k+2}^{-1}, \quad k = 1, \ldots, n-3,\ n \geq 4. \tag{3.11}$$

The function $\Phi(x)$ is twice continuously differentiable in $(0,1)$ and

$$\Phi''(x) = \frac{6}{x(1-x)}.$$

By Taylor's formula we get for $k = 0, \ldots, n-2$

$$\Delta(k) = \int_{k/n}^{(k+2)/n} M_{n,k}(t)\, \Phi''(t)\, dt, \tag{3.12}$$

where

$$M_{n,k}(t) := \begin{cases} t - \frac{k}{n}, & t \in \left[\frac{k}{n}, \frac{k+1}{n}\right], \\ \frac{k+2}{n} - t, & t \in \left(\frac{k+1}{n}, \frac{k+2}{n}\right]. \end{cases}$$

For $k = 0$ formula (3.12) implies

$$\begin{aligned} \Delta(0) &= 6 \int_0^{1/n} \frac{dt}{1-t} + 6 \int_{1/n}^{2/n} \left(\frac{2}{n} - t\right) \frac{dt}{t(1-t)} \\ &\geq 6 \int_0^{1/n} \frac{dt}{1-t} \\ &\geq \frac{6}{n}. \end{aligned}$$

Thus (3.10) is verified for $k = 0$. The case $k = n-2$ is symmetric to $k = 0$.

For $k = 1, \ldots, n-3$, (3.12) yields

$$\begin{aligned} \Delta(k) &\geq \frac{6}{\max_{x \in [k/n, (k+2)/n]} x(1-x)} \int_{k/n}^{(k+2)/n} M_{n,k}(t)\, dt \\ &= \frac{6}{n^2 \max_{x \in [k/n, (k+2)/n]} x(1-x)}. \end{aligned}$$

If $(k+2)/n \leq 1/2$, then $\max_{x \in [k/n, (k+2)/n]} x(1-x) = (k+2)(n-k-2)/n^2$ and $\binom{n}{k} \leq \binom{n}{k+2}$; hence (3.11) will follow from

$$\frac{6}{(k+2)(n-k-2)} \geq 2\binom{n}{k}^{-1}.$$

This inequality follows from

$$\frac{3}{n(k+2)} \geq \binom{n}{k}^{-1},$$

which is trivial for $k = 1$, and otherwise follows from

$$\frac{3}{n(k+2)} \geq \binom{n}{2}^{-1} = \frac{2}{n(n-1)}.$$

The case $k/n \geq 1/2$ is symmetric to the case just considered.

It remains to verify (3.11) for k such that $1/2 \in (k/n, (k+2)/n)$. Then $\max_{x\in[k/n,(k+2)/n]} x(1-x) = 1/4$. The condition $1/2 \in (k/n, (k+2)/n)$ is equivalent to $n/2 - 2 < k < n/2$.

If n is even, then $k = n/2 - 1$ and $\binom{n}{k} = \binom{n}{k+2}$. In this case (3.11) will follow from

$$\frac{12}{n^2} \geq \binom{n}{n/2-1}^{-1}.$$

This is verified directly for $n = 4$; otherwise, it follows from

$$\frac{12}{n^2} \geq \binom{n}{2}^{-1} = \frac{2}{n(n-1)}, \tag{3.13}$$

which is trivial.

Finally, if n is odd, then $k = (n-3)/2$ or $k = (n-1)/2$. These two cases are symmetric and it suffices to consider $k = (n-3)/2$. Then $\binom{n}{k} < \binom{n}{k+2}$. Therefore (3.11) will follow from

$$\frac{12}{n^2} \geq \binom{n}{(n-3)/2}^{-1}.$$

This is checked directly for $n = 5$; otherwise, it follows from (3.13). □

We proceed to the proof of Theorem 1.10.

Proof of Theorem 1.10. Direct computations and (2.9) yield for $n \geq 3$ and $k = 0, \ldots, n-2$ the relation

$$\Phi_n\left(\frac{k+2}{n}\right) - 2\Phi_n\left(\frac{k+1}{n}\right) + \Phi_n\left(\frac{k}{n}\right) = \binom{n}{k}^{-1} + \binom{n}{k+2}^{-1}. \tag{3.14}$$

Therefore, by virtue of (3.8) and (3.9), the function $\Phi_n(x)$ satisfies the conditions in Propositions 3.1-3.3; hence the assertions of the theorem follow. □

Proposition 3.4. *The function $\Phi_n(x)$, defined in Theorem 1.10, is convex on $[0,1]$ and satisfies the estimates*

$$-\frac{4}{n} \leq \Phi_n(x) \leq \frac{16}{n}, \quad x \in \left[\frac{1}{n}, 1 - \frac{1}{n}\right]. \tag{3.15}$$

Proof. As we assumed in the statement of Theorem 1.10, $n \geq 3$. The function $\Phi_n(x)$ is twice continuously differentiable on $[0,1]$, as

$$\Phi_n'(x) = n(n+1)\int_0^1 \left(t^2+(1-t)^2\right) \times \frac{t^2(1-t)^{n-2} - t^3(1-t)^{n-3} + t^{nx}(1-t)^{n(1-x)}(\ln t - \ln(1-t))}{(1-2t)^2}\, dt$$

and

$$\Phi_n''(x) = n^2(n+1)\int_0^1 \left(t^2+(1-t)^2\right) t^{nx}(1-t)^{n(1-x)} \left(\frac{\ln t - \ln(1-t)}{1-2t}\right)^2 dt.$$

We have that $\Phi_n''(x) > 0$ on $[0,1]$; hence $\Phi_n(x)$ is convex on $[0,1]$.

Further, since $\Phi_n(x)$ is convex, then

$$\Phi_n(x) \leq \max\left\{\Phi_n\left(\frac{1}{n}\right), \Phi_n\left(1-\frac{1}{n}\right)\right\}, \quad x \in \left[\frac{1}{n}, 1-\frac{1}{n}\right]. \tag{3.16}$$

Straightforward computations and (2.9) show that

$$\Phi_n\left(\frac{1}{n}\right) = \binom{n}{1}^{-1} + \binom{n}{3}^{-1} \leq \frac{4}{n}. \tag{3.17}$$

The definition of $\Phi_n(x)$ readily yields that $\Phi_n(2/n) = \Phi_n(3/n) = 0$. This, combined with (3.16) and (3.17), implies (3.15) for $n = 3, 4$.

To estimate $\Phi_n(1-1/n)$ for $n \geq 5$ we sum relations (3.14) on $k = 2, \ldots, j$ and then on $j = 2, \ldots, n-3$. As we take into account $\Phi_n(2/n) = \Phi_n(3/n) = 0$, we arrive at

$$\Phi_n\left(\frac{n-1}{n}\right) = \sum_{k=2}^{n-3}(n-k-2)\binom{n}{k}^{-1} + \sum_{k=2}^{n-3}(n-k-2)\binom{n}{k+2}^{-1}. \tag{3.18}$$

Next, we estimate the right-hand-side of (3.18):

$$\begin{aligned}
\sum_{k=2}^{n-3}(n-k-2)\binom{n}{k}^{-1} &\leq \sum_{k=2}^{n-2}(n-k+1)\frac{k!\,(n-k)!}{n!} \\
&= (n+1)\sum_{k=2}^{n-2}\binom{n+1}{k}^{-1} \\
&\leq (n+1)\binom{n+1}{2}^{-1} + (n+1)(n-4)\binom{n+1}{3}^{-1} \\
&\leq \frac{8}{n}.
\end{aligned}$$

Similarly, we get

$$\sum_{k=2}^{n-3}(n-k-2)\binom{n}{k+2}^{-1} \leq \frac{8}{n}.$$

By virtue of the last two estimates, the fact that $\Phi_n(2/n) = \Phi_n(3/n) = 0$ and (3.18), we arrive at

$$\Phi_n\left(\frac{n-1}{n}\right) \leq \frac{16}{n}.$$

This along with (3.16) and (3.17) imply the upper estimate in (3.15) for $n \geq 5$.

In order to verify the lower estimate, we use that $\Phi_n(x)$ is convex and $\Phi_n(2/n) = \Phi_n(3/n) = 0$ to deduce that $\Phi_n(x)$ attains its global minimum on the interval $(2/n, 3/n)$. Since $\Phi_n(x)$ is convex, its graph on the interval $[2/n, 3/n]$ lies above the secant line through the points $(1/n, \Phi_n(1/n))$ and $(2/n, \Phi_n(2/n))$. Thus we arrive at

$$\Phi_n(x) \geq \Phi_n\left(\frac{1}{n}\right)(2-nx) \geq -\Phi_n\left(\frac{1}{n}\right), \quad x \in \left[\frac{2}{n}, \frac{3}{n}\right].$$

Hence, taking into account (3.17), we get the left inequality in (3.15). □

4. EXAMPLES

We will give several examples to illustrate some of the results obtained above.

We begin with an example, which shows that the operator $\widetilde{B}_n$ does not preserve monotonicity for all n. It can be shown that if f is monotone increasing, then so is $\widetilde{B}_n(f)$ for $n \leq 5$. Here is a counterexample for $n = 6$.

Example 4.1. Let

$$f(0) = 0; \qquad f\left(\frac{1}{6}\right) = \frac{50}{60}; \quad f\left(\frac{2}{6}\right) = \frac{56}{60}; \qquad f\left(\frac{3}{6}\right) = \frac{57}{60};$$
$$f\left(\frac{4}{6}\right) = \frac{58}{60}; \qquad f\left(\frac{5}{6}\right) = \frac{59}{60}; \quad f(1) = 1.$$

Then

$$\widetilde{B}_n(f)(x) = 5x(1-x)^5 + 14x^2(1-x)^4 + 19x^3(1-x)^3 + 14x^4(1-x)^2 + 5x^5(1-x) + x^6.$$

Its derivative is

$$(\widetilde{B}_n(f))'(x) = 5(1-x)^5 + 3x(1-x)^4 + x^2(1-x)^3 - x^3(1-x)^2 - 3x^4(1-x) + x^5$$

and $(\widetilde{B}_n(f))'(7/10) = -73/2000$.

It seems that it is quite difficult to construct a monotone function f, for which $\widetilde{B}_n(f)$ or $\widehat{B}_n(f)$ are not monotone, by means of elementary functions.

In the next example we consider the sufficient condition stated in Theorem 1.7.

Example 4.2. The function $f(x) = (x+1)^5$ satisfies the assumptions in Theorem 1.7. Thus the polynomials $\widetilde{B}_n(f)$ are monotone increasing for all n. Figure 1 contains the plot of $f(x)$ and $\widetilde{B}_n(f)$ for $n = 5$ and $n = 10$.

Finally, let us demonstrate that $\widetilde{B}_n$ preserves asymptotically convexity.

Example 4.3. Consider the concave function $f(x) = \sqrt{x}$. Figure 2 shows the plots of $f(x)$ and $\widetilde{B}_n(f)$ for $n = 5$ and $n = 10$. We can see that the graphs of $\widetilde{B}_5(f)$ and $\widetilde{B}_{10}(f)$ have an inflection point. It moves to 1 as n increases. This example shows that generally $\widetilde{B}_n$, and similarly $\widehat{B}_n$, does not preserve convexity.

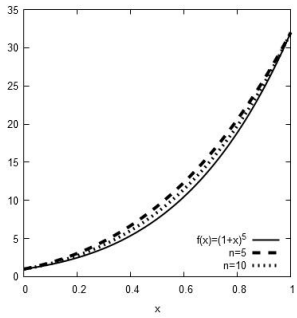


Figure 1. $\widetilde{B}_n$ and monotonicity

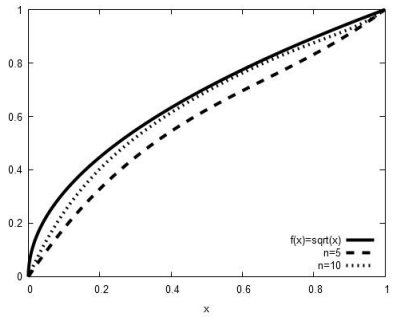


Figure 2. $\widetilde{B}_n$ and convexity

The computations and the plots were made with wmMaxima 16.04.2.

ACKNOWLEDGEMENTS. This work was supported by grant DN 02/14 of the Fund for Scientific Research of the Bulgarian Ministry of Education and Science. I am thankful to the Referee for the corrections and suggestions—they improved the presentation. Especially, I owe the Referee the elegant idea how to reduce the case of decreasing functions to the case of increasing.

5. REFERENCES

[1] DeVore, R. A., Lorentz, G. G.: *Constructive Approximation*. Springer-Verlag, Berlin, 1993.

[2] Draganov, B. R.: Simultaneous approximation by Bernstein polynomials with integer coefficients. *J. Approx. Theory*, **237**, 2019, 1–16.

[3] Draganov, B. R.: Converse estimates for the simultaneous approximation by Bernstein polynomials with integer coefficients. arXiv:1904.09417, 2019.

[4] Ferguson, Le Baron O.: *Approximation by Polynomials with Integral Coefficients*. Mathematical Surveys Vol. **17**, American Mathematical Society, 1980.

[5] Kantorovich, L. V.: Some remarks on the approximation of functions by means of polynomials with integer coefficients. *Izv. Akad. Nauk SSSR, Ser. Mat.*, **9**, 1931, 1163–1168 (in Russian).

[6] Lorentz, G. G., v.Golitschek, M., Makovoz, Y.: *Constructive Approximation, Advanced Problems*. Springer-Verlag, Berlin, 1996.

[7] Martini, R.: On the approximation of functions together with their derivatives by certain linear positive operators. *Indag. Math.*, **31**, 1969, 473–481.

[8] Rockett, A. M.: Sums of the inverses of binomial coefficients. *Fibonacci Quart.*, **19**, 1981, 433–437.

[9] Sury, B.: Sum of the reciprocals of the binomial coefficients. *European J. Combin.*, **14**, 1993, 351–535.

Received on January 15, 2020
Received in a revised form on May 25, 2020

Borislav R. Draganov
Department of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

Institute of Mathematics and Informatics
Bulgarian Academy of Science
bl. 8 Acad. G. Bonchev Str.
1113 Sofia
BULGARIA

E-mail: bdraganov@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 106

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 106

DEFINITE QUADRATURE FORMULAE OF 5-TH ORDER WITH EQUIDISTANT NODES

ANA AVDZHIEVA, GENO NIKOLOV

We construct sequences of definite quadrature formulae of order five which use equidistant nodes. The error constants of these quadratures are evaluated and simple a posteriori error estimates derived under the assumption that the integrand's fifth derivative does not change its sign in the integration interval.

Keywords: Definite quadrature formulae, Peano kernels, Euler-MacLaurin summation formulae, a posteriori error estimates.

2000 Math. Subject Classification: 41A55, 65D30, 65D32.

1. INTRODUCTION AND STATEMENT OF THE RESULTS

The definite integral

$$I[f] := \int_0^1 f(x)\,dx$$

is evaluated approximately by a quadrature formula, which is a linear functional of the form

$$Q[f] = \sum_{i=0}^{n} a_i\, f(x_i), \qquad 0 \le x_0 < x_1 < \cdots < x_n \le 1\,. \tag{1.1}$$

Two reasonable though rather demanding requirements for a quadrature formula (1.1) are : 1) to have the smallest possible maximal error for integrands f belonging

to a given class of functions and 2) to provide the exact value of the integral for integrands from a linear space of the highest possible dimension. Quadrature formulae satisfying these two requirements are called *optimal* and *Gauss type* quadratures, respectively. Regardless which of these criteria is applied for the design of a quadrature formula, typically its nodes and weights are evaluated numerically, thus they are only approximately known. For this reason often in practice a preference is given to quadrature formulae having other useful properties, e.g. quadrature formulae whose knots and weights are *explicitly* known, or which allow easy error estimation. For instance, using quadrature formulae with equispaced nodes, we save half of the integrand evaluations when doubling the number of nodes; using quadrature formulae of (almost) Chebyshev type (i.e., with almost all weights equal to each other) we reduce the error induced by rounding. In automated routines for numerical integration, definite quadrature formulae are widely used for derivation of criteria for termination of calculations (the so-called stopping rules), see e.g. [5].

This paper is a continuation of our previous study on definite quadrature formulae of low order which use equidistant nodes and are of almost Chebyshev type. Before formulating our results, let us recall some definitions.

Quadrature formula (1.1) is said to have *algebraic degree of precision* m (in short, $ADP(Q) = m$), if its remainder

$$R[Q; f] := I[f] - Q[f]$$

vanishes whenever $f \in \pi_m$, and $R[Q; f] \neq 0$ when f is a polynomial of degree $m+1$. Here and henceforth, π_k stands for the set of real algebraic polynomials of degree at most k.

Definition 1. Quadrature formula (1.1) is said to be *definite of order* r, $r \in \mathbb{N}$, if there exists a real non-zero constant $c_r(Q)$ such that its remainder functional admits the representation

$$R[Q; f] = I[f] - Q[f] = c_r(Q)\, f^{(r)}(\xi)$$

for every real-valued function $f \in C^r[0, 1]$, with some $\xi \in [0, 1]$ depending on f.

Furthermore, Q is called *positive definite* (resp., *negative definite*) of order r, if $c_r(Q) > 0$ ($c_r(Q) < 0$).

Definition 2. A real-valued function $f \in C^r[0, 1]$ is called $r-$*positive* (resp., $r-$*negative*) if $f^{(r)}(x) \geq 0$ (resp. $f^{(r)}(x) \leq 0$) for every $x \in [0, 1]$.

A definite quadrature formula of order r provides one-sided approximation to $I[f]$ whenever f is $r-$positive or $r-$negative. If $\{Q^+,\ Q^-\}$ is a pair of a positive and a negative definite quadrature formula of order r and f is an r-positive function, then $Q^+[f] \leq I[f] \leq Q^-[f]$. Most of quadrature formulae used in practice (e.g., quadrature formulae of Gauss, Radau, Lobatto, Newton-Cotes) are definite of certain order.

In [1] we constructed several sequences of asymptotically optimal definite quadrature formulae of order four with all but a few boundary nodes being equidistant. For some pairs of these definite quadrature formulae we derived a posteriori error estimates. In [2, 3] definite quadrature formulae of order three based on the nodes of compound trapesium and midpoint quadratures were constructed and a posteriori error estimates derived. It turns out that definite quadrature formulae of odd order offer some additional advantages, see Proposition 1 below.

Definition 3. Quadrature formula (1.1) is called:

- *symmetrical*, if

$$a_k = a_{n-k}\,, \qquad k = 0, \dots, n\,, \tag{1.2}$$

$$x_k = 1 - x_{n-k}\,, \qquad k = 0, \dots, n\,; \tag{1.3}$$

- *nodes-symmetrical*, if only condition (1.3) is satisfied;

- The quadrature formula

$$\widetilde{Q}[f] = \widetilde{Q}[Q; f] := \sum_{k=0}^{n} a_k\, f(x_{n-k}) \tag{1.4}$$

is called *reflected quadrature formula* to (1.1).

Proposition 1 ([**2**])**.** *(i) If Q is a positive definite quadrature formula of order r (r – odd), then its reflected quadrature formula $\widetilde{Q}$ is negative definite of order r and vice versa. Moreover, $c_r(\widetilde{Q}) = -c_r(Q)$.*

(ii) If quadrature formula Q in (1.1) *is nodes-symmetrical and definite of order r (r – odd), and f is an r-positive or r-negative function, then, with Q^* standing for either Q or $\widetilde{Q}$ we have*

$$\begin{aligned} \big|R[Q^*; f]\big| \le B[Q; f] &:= \big|\widetilde{Q}[f] - Q[f]\big| \\ &= \Big|\sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} \big(a_k - a_{n-k}\big)\big(f(x_{n-k}) - f(x_k)\big)\Big|. \end{aligned} \tag{1.5}$$

(iii) Under the same assumptions for Q and f as in (ii), for $\hat{Q} = (Q + \widetilde{Q})/2$ we have

$$\big|R[\hat{Q}; f]\big| \le \frac{1}{2} B[Q; f]\,.$$

Proposition 1(i) implies that definite quadrature formulae of odd order are never symmetrical. Let us point out that the error estimate (1.5) becomes especially simple when almost all coefficients of Q are equal to each other.

For $n \in \mathbb{N}$ and a function f defined on the interval $[0,1]$, we denote

$$x_i = x_{i,n} = \frac{i}{n}, \quad f_i = f(x_{i,n}), \qquad i = 0 \ldots, n.$$

Recall that the finite differences $\Delta^k f_i$ are defined recursively by

$$\Delta^1 f_i = \Delta f_i := f_{i+1} - f_i \quad \text{and} \quad \Delta^{k+1} f_i = \Delta\big(\Delta^k f_i\big), \quad k \geq 1.$$

Set

$$c := \frac{3+\sqrt{30}}{21600}\sqrt{1-2\sqrt{\frac{2}{15}}}, \quad c \approx 0.000203818.$$

Our main result reads as follows:

Theorem 1. *(i) For every $n \geq 11$, the quadrature formula*

$$Q_n[f] = \frac{1}{n}\sum_{k=0}^{n-1} A_k\, f_k + \frac{c}{n}\big(\Delta^4 f_0 - \Delta^4 f_{n-5}\big),$$

where $A_k = 1$ for $5 \leq k \leq n-6$ and

$$A_0 = \frac{95}{288}, \quad A_1 = \frac{317}{240}, \quad A_2 = \frac{23}{30}, \quad A_3 = \frac{793}{720}, \quad A_4 = \frac{157}{160},$$
$$A_{n-5} = \frac{383}{288}, \quad A_{n-4} = -\frac{481}{720}, \quad A_{n-3} = \frac{22}{5}, \quad A_{n-2} = -\frac{1823}{720}, \quad A_{n-1} = \frac{4277}{1440},$$

is positive definite of order 5 with the error constant

$$c_5(Q_n) = \frac{c}{n^5} + \frac{5(19-288c)}{288n^6}. \tag{1.6}$$

(ii) If f is a 5-positive or 5-negative function, then

$$\big|R[Q_n; f]\big| \leq \frac{1}{n}\left|\Big(\frac{95}{288} - c\Big)\big(\Delta^5 f_0 + \Delta^5 f_{n-5}\big) + 2c\,\big(\Delta^4 f_{n-4} - \Delta^4 f_0\big)\right|.$$

As an immediate consequence of Theorem 1 and Proposition 1 we have:

Corollary 1. *The reflected to Q_n from Theorem 1 quadrature formula $\widetilde{Q}_n$ is negative definite of order 5 with the error constant $c_5(\widetilde{Q}_n) = -c_5(Q_n)$.*

If f is a 5-positive or 5-negative function and $\hat{Q}_n = \frac{1}{2}\left(Q_n + \widetilde{Q}_n\right)$, then

$$\big|R[\widetilde{Q}_n; f]\big| \leq \frac{1}{n}\left|\Big(\frac{95}{288} - c\Big)\big(\Delta^5 f_0 + \Delta^5 f_{n-5}\big) + 2c\,\big(\Delta^4 f_{n-4} - \Delta^4 f_0\big)\right|.$$

$$\big|R[\hat{Q}_n; f]\big| \leq \frac{1}{2n}\left|\Big(\frac{95}{288} - c\Big)\big(\Delta^5 f_0 + \Delta^5 f_{n-5}\big) + 2c\,\big(\Delta^4 f_{n-4} - \Delta^4 f_0\big)\right|.$$

Remark 1. It is worth noting that while the implied by (1.6) error estimate

$$|R[Q_n; f]| \le c_5(Q_n)\|f^{(5)}\|_{C[0,1]}$$

requires knowledge of the magnitude of the $C[0,1]$–norm of the integrand's derivative, the error bounds in Theorem 1(ii) and Corollary 1 in terms of finite diferences are easy to evaluate and may serve as a simple criteria for the number of nodes n needed to guarantee the evaluation of $I[f]$ with a prescribed tolerance. (Note however that these error bounds apply only for 5-positive or 5-negative integrands.) Let us also mention that, according to Corollary 1, the symmetrical (and hence not definite) quadrature formula $\hat{Q}_n$ has smaller error bound than the definite quadrature formulae Q_n and $\widetilde{Q}_n$.

The rest of the paper is organised as follows. Section 2 contains some preliminaries. In Section 2.1 we give some known facts about the Peano kernel representation of linear functionals, and prove a simple necessary condition for a quadrature formula to be positive definite. Some facts about Bernoulli polynomials and numbers and the Euler-MacLaurin summation formula are given in Section 2.2. In Sections 3 we present some formulae for numerical differentiation to be used for replacement of the derivatives occurring in the Euler-MacLaurin formula. The proof of Theorem 1 and Corollary 1 is given in Section 4.

2. PRELIMINARIES

2.1. PEANO KERNEL REPRESENTATION OF LINEAR FUNCTIONALS

For $r \in \mathbb{N}$, the Sobolev class of functions $W_1^r = W_1^r[0,1]$ is defined by

$$W_1^r[0,1] := \{f \in C^{r-1}[0,1] \,:\, f^{(r-1)} \text{ loc. abs. continuous}, \int_0^1 |f^{(r)}(t)|\,dt < \infty\}$$

and contains, in particular, the class $C^r[0,1]$.

If $\mathcal{L}$ is a linear functional defined in $W_1^r[0,1]$ which vanishes on π_{r-1}, then, by a classical result of Peano [11], $\mathcal{L}$ is represented in the form

$$\mathcal{L}[f] = \int_0^1 K_r(t) f^{(r)}(t)\,dt\,,$$

where $K_r(t) = K_r(\mathcal{L}; t)$ is given by

$$K_r(t) = \mathcal{L}\Big[\frac{(\cdot - t)_+^{r-1}}{(r-1)!}\Big], \quad t \in [0,1]\,, \qquad u_+(t) = \max\{t, 0\}\,, \quad t \in \mathbb{R}\,.$$

When $\mathcal{L}$ is the remainder $R[Q;\cdot]$ of a quadrature formula Q with $ADP(Q) \geq r-1$, with some notational and language abuse, $K_r(t) = K_r(Q;t)$ is referred to as the r-th Peano kernel of Q. For Q as in (1.1), explicit representations for $K_r(Q;t)$, $t \in [0,1]$, are

$$K_r(Q;t) = \frac{(1-t)^r}{r!} - \frac{1}{(r-1)!}\sum_{i=0}^{n} a_i (x_i - t)_+^{r-1}, \tag{2.1}$$

and

$$K_r(Q;t) = (-1)^r \Big[\frac{t^r}{r!} - \frac{1}{(r-1)!}\sum_{i=0}^{n} a_i (t - x_i)_+^{r-1}\Big]. \tag{2.2}$$

Thus, for $f \in C^r[0,1]$ and a quadrature formula Q with $ADP(Q) = r-1$,

$$R[Q;f] = \int_0^1 K_r(Q;t)\, f^{(r)}(t)\, dt\,.$$

It is clear now that Q is a positive (negative) definite quadrature formula of order r if and only if $ADP(Q) = r-1$ and $K_r(Q;t) \geq 0$ (resp. $K_r(Q;t) \leq 0$) for all $t \in [0,1]$, and if this is the case, then

$$c_r(Q) = \int_0^1 K_r(Q;t)\, dt\,.$$

From (2.1) and (2.2) one easily derives the following necessary condition for positive (negative) definiteness of a quadrature formula.

Lemma 1. *Let*

$$Q[f] = \sum_{k=0}^{n} a_k\, f(x_k)\,, \quad 0 = x_0 < x_1 < \cdots < x_n = 1,$$

be a quadrature formula for $I[f] = \int_0^1 f(x)\, dx$. A necessary condition for Q to be positive (resp., negative) definite of order r is

$$(-1)^r\, a_0 \leq 0 \quad \textit{and} \quad a_n \leq 0 \quad (\textit{resp.,}\ (-1)^r\, a_0 \geq 0 \quad \textit{and} \quad a_n \geq 0)\,.$$

Proof. If Q is positive or negative definite of order r, then $ADP(Q) = r-1$, and therefore $K_r(Q;t) \geq 0$ (resp., $K_r(Q;t) \leq 0$) for every $t \in (0,1)$. From (2.1) and (2.2) we find that for sufficiently small $\varepsilon > 0$

$$\operatorname{sign}\, K_r(Q; x_n - \varepsilon) = -\operatorname{sign} a_n\,, \quad \operatorname{sign}\, K_r(Q; x_0 + \varepsilon) = (-1)^{r+1} \operatorname{sign} a_0\,,$$

whence the conclusion follows. □

Assuming f is smooth enough, the remainder of the n-th compound trapezium quadrature formula

$$Q_n^{Tr}[f] = \frac{1}{2n}\big(f_0 + f_n\big) + \frac{1}{n}\sum_{k=1}^{n-1} f_k$$

(with $f_i = f(x_i)$ and $x_i = i/n$, $i = 0, \dots, n$) admits an expansion of the form

$$R[Q_n^{Tr}; f] = -\sum_{\nu=1}^{[\frac{s}{2}]} \frac{B_{2\nu}(0)}{n^{2\nu}}\big[f^{(2\nu-1)}(1) - f^{(2\nu-1)}(0)\big] + \frac{(-1)^s}{n^s}\int_0^1 \widetilde{B}_s(nx) f^{(s)}(x) dx\,.$$

This is the so-called Euler-Maclaurin summation formula (see, e.g., [4, Satz 98]). Here, $\{B_\nu\}$ are the Bernoulli polynomials, which are defined recursively by

$$B_0(x) = 1, \quad B_\nu'(x) = B_{\nu-1}(x), \qquad \int_0^1 B_\nu(t)\,dt = 0\,, \qquad \nu \in \mathbb{N},$$

and $\widetilde{B}_\nu$ is the one-periodic extension of B_ν, i.e., $\widetilde{B}_\nu(x) = B_\nu(\{x\})$, where $\{x\}$ is the fractional part of $x \in \mathbb{R}$.

In the case $s = 5$ the Euler–Maclaurin summation formula reads as

$$\begin{aligned} I[f] =& Q_n^{Tr}[f] - \frac{1}{12n^2}\big[f'(1) - f'(0)\big] + \frac{1}{720n^4}\big[f'''(1) - f'''(0)\big] \\ & - \frac{1}{n^5}\int_0^1 \widetilde{B}_5(nx) f^{(5)}(x)\,dx\,, \end{aligned} \tag{2.3}$$

with the explicit form of B_5

$$B_5(x) = \frac{x^5}{120} - \frac{x^4}{48} + \frac{x^3}{72} - \frac{x}{720}\,.$$

Let us note that for $x \in \mathbb{R}$

$$-c \le \widetilde{B}_5(x) \le c\,, \quad \text{where} \quad c := \frac{3+\sqrt{30}}{21600}\sqrt{1 - 2\sqrt{\frac{2}{15}}} \approx 0.000203818\,.$$

Rewriting (2.3) in the form

$$\begin{aligned} I[f] =& Q_n^{Tr}[f] - \frac{1}{12\,n^2}\big[f'(1) - f'(0)\big] + \frac{1}{720\,n^4}\big[f'''(1) - f'''(0)\big] \\ & - \frac{c}{n^5}\big[f^{(4)}(1) - f^{(4)}(0)\big] + \frac{1}{n^5}\int_0^1 \Big(c - \widetilde{B}_5(nx)\Big) f^{(5)}(x)\,dx\,, \end{aligned}$$

we observe that the quadrature formula

$$\begin{aligned} Q_n^*[f] =& Q_n^{Tr}[f] - \frac{1}{12\,n^2}\big[f'(x_n) - f'(x_0)\big] + \frac{1}{720\,n^4}\big[f'''(x_n) - f'''(x_0)\big] \\ & - \frac{c}{n^5}\big[f^{(4)}(x_n) - f^{(4)}(x_0)\big] \end{aligned} \tag{2.4}$$

is positive definite of order 5, since

$$K_5(Q_n^*;t) = n^{-5}\big(c - \widetilde{B}_5(nt)\big) \geq 0, \qquad t \in \mathbb{R}\,. \tag{2.5}$$

However, Q_n^* is not of desired form, as it involves evaluations of both the integrand and of its derivatives. In order to obtain a quadrature formula using only integrand's evaluation, we need some formulae for numerical differentiation.

3. FORMULAE FOR NUMERICAL DIFFERENTIATION

The following formulae for numerical differentiation will be used to replace the derivatives occurring in quadrature formula Q_n^* (recall that $x_i = i/n$ and $f_i = f(x_i)$ for $i = 0, \ldots, n$):

$$\begin{aligned} f'(x_0) &\approx D_1[f] := \frac{n}{12}\big[-25\,f_0 + 48\,f_1 - 36\,f_2 + 16\,f_3 - 3\,f_4\big]\,, \\ f'''(x_0) &\approx D_3[f] := \frac{n^3}{2}\big[-5\,f_0 + 18\,f_1 - 24\,f_2 + 14\,f_3 - 3\,f_4\big]\,, \\ f^{(4)}(x_0) &\approx D_4[f] := n^4\big[f_0 - 4\,f_1 + 6\,f_2 - 4\,f_3 + f_4\big] = n^4\Delta^4 f_0\,, \end{aligned}$$

$$\begin{aligned} f'(x_n) &\approx \widetilde{D}_1[f] := \frac{n}{12}\big[25\,f_n - 48\,f_{n-1} + 36\,f_{n-2} - 16\,f_{n-3} + 3\,f_{n-4}\big]\,, \\ f'''(x_n) &\approx \widetilde{D}_3[f] := \frac{n^3}{2}\big[5\,f_n - 18\,f_{n-1} + 24\,f_{n-2} - 14\,f_{n-3} + 3\,f_{n-4}\big]\,, \\ f^{(4)}(x_n) &\approx \widetilde{D}_4[f] := n^4\big[f_n - 4\,f_{n-1} + 6\,f_{n-2} - 4\,f_{n-3} + f_{n-4}\big] = n^4\Delta^4 f_{n-4}\,. \end{aligned}$$

These formulae are sharp for $f \in \pi_4$, i.e., the linear functionals

$$L_j[f] := f^{(j)}(x_0) - D_j[f]\,, \quad \widetilde{L}_j[f] := f^{(j)}(x_n) - \widetilde{D}_j[f]\,, \quad j = 1,\,3,\,4$$

vanish on π_4. According to Peano's theorem, for $f \in C^5[0,1]$ they admit integral representations, in particular,

$$L_j[f] := \int_0^1 K_5(L_j;t) f^{(5)}(t)\,dt\,, \quad K_5(L_j;t) = L_j\Big[\frac{(\cdot - t)_+^4}{4!}\Big], \quad j = 1,\,3,\,4\,. \tag{3.1}$$

Proposition 2. *The Peano kernels $K_5(L_j;\cdot)$, $j=1,3,4$, vanish identically on the interval $[x_4,x_n]$. Moreover,*

$$\int_0^1 K_5(L_1;t)\,dt=\frac{1}{5n^4}\,, \tag{3.2}$$

$$\int_0^1 K_5(L_3;t)\,dt=\frac{7}{4n^2}\,, \tag{3.3}$$

$$\int_0^1 K_5(L_4;t)\,dt=-\frac{2}{n}\,. \tag{3.4}$$

Proof. The first claim follows from (3.1): for $t\geq x_4$ and $x\leq x_4$ we have, by definition, $(x-t)_+^4\equiv 0$, hence $K_5(L_j;t)=L_j[(\cdot-t)_+^4]/4!\equiv 0$ for $t\in[x_4,x_n]$.

Equality (3.2) is verified as follows:

$$\begin{aligned}\int_0^1 K_5(L_1;t)\,dt&=-\frac{n}{288}\int_0^1\left[48(x_1-t)_+^4-36(x_2-t)_+^4+16(x_3-t)_+^4-3(x_4-t)_+^4\right]dt\\&=\frac{n}{1440}\Big[48(x_1-t)^5\Big|_0^{x_1}-36(x_2-t)^5\Big|_0^{x_2}+16(x_3-t)^5\Big|_0^{x_3}-3(x_4-t)^5\Big|_0^{x_4}\Big]\\&=\frac{1}{5n^4}\,.\end{aligned}$$

Equalities (3.3) and (3.4) are verified in the same way. □

In order to deduce an analogous statement for the linear functionals $\widetilde{L}_j$, we need a more convenient formula for their Peano kernels. Since

$$(x-t)_+^4+(t-x)_+^4=(x-t)^4\quad\text{ for every }\ x,t\in\mathbb{R}\,,$$

and $\widetilde{L}_j$ vanish on π_4, it follows that $\widetilde{L}_j[(\cdot-t)_+^4]=-\widetilde{L}_j[(t-\cdot)_+^4]$, hence

$$\widetilde{L}_j[f]:=\int_0^1 K_5(\widetilde{L}_j;t)f^{(5)}(t)\,dt\,,\quad K_5(\widetilde{L}_j;t)=-\widetilde{L}_j\Big[\frac{(t-\cdot)_+^4}{4!}\Big]\,,\quad j=1,\,3,\,4\,. \tag{3.5}$$

By using (3.5), we establish in the same manner the following:

Proposition 3. *The Peano kernels $K_5(\widetilde{L}_j;\cdot)$, $j=1,3,4$, vanish identically on the interval $[x_0,x_{n-4}]$. Moreover,*

$$\int_0^1 K_5(\widetilde{L}_1;t)\,dt=\frac{1}{5n^4}\,,$$

$$\int_0^1 K_5(\widetilde{L}_3;t)\,dt=\frac{7}{4n^2}\,,$$

$$\int_0^1 K_5(\widetilde{L}_4;t)\,dt=\frac{2}{n}\,.$$

4. PROOF OF THEOREM 1

Replacement of the derivatives in (2.4) with the formulae for numerical differentiation from Section 3 yields

$$\begin{aligned} Q_n^*[f] =& Q_n^{Tr}[f] + \frac{D_1[f]}{12n^2} - \frac{D_3[f]}{720n^4} + c\,\frac{D_4[f]}{n^5} - \frac{\widetilde{D}_1[f]}{12n^2} + \frac{\widetilde{D}_3[f]}{720n^4} - c\,\frac{\widetilde{D}_4[f]}{n^5} \\ &+ \frac{1}{12n^2}\big(L_1[f]-\widetilde{L}_1[f]\big) - \frac{1}{720n^4}\big(L_3[f]-\widetilde{L}_3[f]\big) + \frac{c}{n^5}\big(L_4[f]-\widetilde{L}_4[f]\big) \\ =:& \widehat{Q}_n[f] + L[f]\,, \end{aligned} \tag{4.1}$$

where the linear functional L is given by

$$L = \frac{1}{12n^2}\left(L_1 - \widetilde{L}_1\right) - \frac{1}{720n^4}\left(L_3 - \widetilde{L}_3\right) + \frac{c}{n^5}\big(L_4 - \widetilde{L}_4\big) \tag{4.2}$$

and $\widehat{Q}_n$ is the quadrature formula

$$\widehat{Q}_n[f] = \frac{1}{n}\sum_{k=0}^{n} a_k\, f_k + \frac{c}{n}\left(\Delta^4 f_0 - \Delta^4 f_{n-4}\right) \tag{4.3}$$

with coefficients

$$\begin{aligned} &a_0 = a_n = \frac{95}{288}, \qquad a_1 = a_{n-1} = \frac{317}{240}, \quad a_2 = a_{n-2} = \frac{23}{30}, \\ &a_3 = a_{n-3} = \frac{793}{720}, \quad a_4 = a_{n-4} = \frac{157}{160}\,, \quad a_k = 1, \quad 5 \le k \le n-5\,. \end{aligned} \tag{4.4}$$

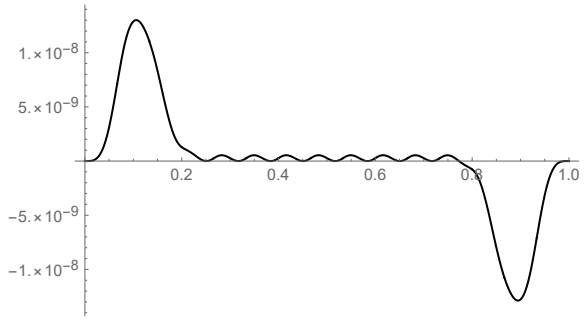


Figure 1. The graph of $K_5(\widehat{Q}_n;t)$, $n = 15$.

Clearly, $ADP(\widehat{Q}_n) \ge 4$. Unfortunately, $\widehat{Q}_n$ is not positive definite of order 5, as $K_5(\widehat{Q}_n;t)$ is negative in a neighborhood of $x_n = 1$, see Figure 1. In fact, $\widehat{Q}_n$ fails to satisfy the criteria for positive definiteness of Lemma 1, as the coefficient of $f_n = f(x_n)$ in $\widehat{Q}_n$ is

$$\kappa = \frac{1}{n}\Big(\frac{95}{288} - c\Big) > 0\,.$$

In order to fulfill the necessary condition for positive definiteness of Lemma 1, we modify $\widehat{Q}_n$ so that the coefficient of $f(x_n)$ equals zero:

$$\begin{aligned} Q_n[f] &= \widehat{Q}_n[f] - \kappa\, L_5[f]\,, && (4.5)\\ L_5[f] &= -f_{n-5} + 5\, f_{n-4} - 10\, f_{n-3} + 10\, f_{n-2} - 5\, f_{n-1} + f_n\,. && (4.6) \end{aligned}$$

Since the finite difference functional $L_5[f] = \Delta^5 f_{n-5}$ vanishes on π_4, the newly built quadrature formula Q_n uses the equispaced nodes and $ADP(Q_n) = 4$. Assuming $n \geq 11$ and using (4.3), (4.4), (4.5) and (4.6), we find that

$$Q_n[f] = \frac{1}{n} \sum_{k=0}^{n-1} A_k\, f_k + \frac{c}{n} \left(\Delta^4 f_0 - \Delta^4 f_{n-5}\right),$$

where $A_k = 1$ for $5 \leq k \leq n-6$ and

$$\begin{aligned} &A_0 = \frac{95}{288}, \quad A_1 = \frac{317}{240}, \quad A_2 = \frac{23}{30}, \quad A_3 = \frac{793}{720}, \quad A_4 = \frac{157}{160},\\ &A_{n-5} = \frac{383}{288}, \quad A_{n-4} = -\frac{481}{720}, \quad A_{n-3} = \frac{22}{5}, \quad A_{n-2} = -\frac{1823}{720}, \quad A_{n-1} = \frac{4277}{1440}\,. \end{aligned}$$

Hence, Q_n is the quadrature formula from Theorem 1.

We need to show that Q_n is positive definite of order 5, i.e. that $K_5(Q_n; t) \geq 0$ for $t \in (0, 1)$. To this end, we observe that, by virtue of (4.1) and (4.2),

$$Q_n = Q_n^* - L - \kappa\, L_5$$

with L and L_5 given by (4.2) and (4.6), respectively. Consequently,

$$K_5(Q_n; t) = K_5(Q_n^*; t) + K_5(L; t) + \kappa K_5(L_5; t)\,. \tag{4.7}$$

According to (2.5), $K_5(Q_n^*; t) \geq 0$ for $t \in (0, 1)$. From (4.2) and Propositions 2 and 3 we infer that

$$K_5(L; t) \equiv 0 \quad \text{for} \quad t \in [x_4, x_{n-4}]\,. \tag{4.8}$$

A similar conclusion is true for $K_5(L_5; t)$, as it is a B-spline of degree 4 with knots x_i, $n-5 \leq i \leq n$, and therefore

$$K_5(L_5; t) \equiv 0 \quad \text{for} \quad t \in [x_0, x_{n-5}]\,. \tag{4.9}$$

It follows from (4.7), (2.5), (4.8) and (4.9) that $K_5(Q_n; t) \equiv K_5(Q_n^*; t) \geq 0$ on the interval $[x_4, x_{n-5}]$, therefore we only need to verify that $K_5(Q_n; t) \geq 0$ in the cases $t \in (x_0, x_4)$ and $t \in (x_{n-5}, x_n)$.

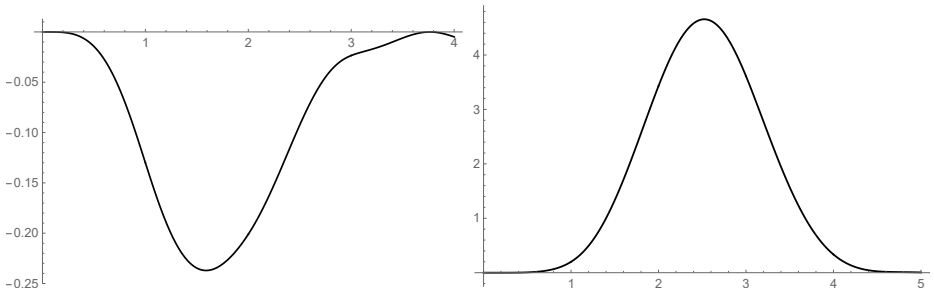
Case 1: $t \in (x_0, x_4)$. By the change of variable $t = u/n$, $u \in (0, 4)$, we obtain

$$K_5(Q_n; t) = -\frac{1}{4!\, n^5}\, \varphi_1(u)\,, \quad u \in (0, 4)\,,$$

where the function φ_1 does not depend on n, namely,

$$\varphi_1(u) = \frac{u^5}{5} - (A_0+c)u^4 - (A_1-4c)(u-1)_+^4 - (A_2+6c)(u-2)_+^4 - (A_3-4c)(u-3)_+^4 .$$

The graph of φ_1, depicted in Figure 2(a), shows that $\varphi_1(u) < 0$ in the interval $(0,4)$ (φ_1 has a local maximum at $u = 3.76475$, equal to -0.000059). Therefore, $K_5(Q_n;t) > 0$ for $t \in (x_0, x_4)$.



(a) The graph of $\varphi_1(u)$, $0 \le u \le 4$. (b) The graph of $\varphi_2(u)$, $0 \le u \le 5$.

Figure 2

Case 2: $t \in (x_{n-5}, x_n)$. By the change of variable $t = 1 - u/n$ we obtain

$$K_5(Q_n;t) = \frac{1}{4!\,n^5}\,\varphi_2(u), \quad u \in (0,5),$$

where

$$\varphi_2(u) = \frac{u^5}{5} - B_1(u-1)_+^4 - B_2(u-2)_+^4 - B_3(u-3)_+^4 - B_4(u-4)_+^4 ,$$

with $B_i = A_{n-i} + (-1)^i \binom{4}{i-1} c$, $i = 1, \dots, 5$. Again, φ_2 does not depend on n and is positive for $u \in (0,5)$, as shown in Figure 2(b). Consequently, $K_5(Q_n;t) > 0$ for $t \in (x_{n-5}, x_n)$, and the proof that Q_n is a positive definite quadrature formula of order 5 is completed.

Having established the positive definiteness of Q_n, we proceed with evaluating its error constant $c_5(Q_n) = I[K_5(Q_n;\cdot)]$. From (4.7) we have

$$c_5(Q_n) = \int_0^1 K_5(Q_n^*;t)\,dt + \int_0^1 K_5(L;t)\,dt + \kappa \int_0^1 K_5(L_5;t)\,dt . \tag{4.10}$$

We evaluate the three integrals on the right-hand side of (4.10). For the first one, we find from (2.5)

$$\int_0^1 K_5(Q_n^*;t)\,dt = \frac{1}{n^5}\int_0^1 \big(c - \widetilde{B}_5(nt)\big)\,dt = \frac{c}{n^5} .$$

According to (4.2),

$$K_5(L;t) = \frac{1}{12n^2}\left(K_5(L_1;t) - K_5(\widetilde{L}_1;t)\right) - \frac{1}{720n^4}\left(K_5(L_3;t) - K_5(\widetilde{L}_3;t)\right) + \frac{c}{n^5}\left(K_5(L_4;t) - K_5(\widetilde{L}_4;t)\right)$$

and using Propositions 2 and 3, we obtain

$$\int_0^1 K_5(L;t)\,dt = -\frac{4c}{n^6}\,.$$

Recall that $L_5[f] = \Delta^5 f_{n-5}$, and from Peano's representation theorem,

$$K_5(L_5;t) = \frac{1}{4!}\big[(x_n - t)_+^4 - 5(x_{n-1} - t)_+^4 + 10(x_{n-2} - t)_+^4 - 10(x_{n-3} - t)_+^4 + 5(x_{n-4} - t)_+^4 - (x_{n-5} - t)_+^4\big]\,.$$

Hence,

$$\int_0^1 K_5(L_5;t)\,dt = \frac{1}{5!n^5}\left[n^5 - 5(n-1)^5 + 10(n-2)^5 - 10(n-3)^5 + 5(n-4)^5 - (n-5)^5\right] = \frac{1}{n^5}\,.$$

Substituting the found values of the three integrals in (4.10), we obtain

$$c_5(Q_n) = \frac{c}{n^5} - \frac{4c}{n^6} + \frac{1}{n^6}\Big(\frac{95}{288} - c\Big) = \frac{c}{n^5} + \frac{5(19 - 288c)}{288n^6}\,,$$

which was to be proved. This accomplishes the proof of Theorem 1(i).

For the proof of Theorem 1(ii) we apply Proposition 1(ii). We set $A_n = 0$, hence Q_n becomes a nodes-symmetrical quadrature formula. Now, according to (1.5),

$$B(Q_n;f) = \left|\widetilde{Q}_n[f] - Q_n[f]\right| = \left|Q_n[\widetilde{f}] - Q_n[f]\right|, \quad \text{where } \widetilde{f}(t) = f(1-t)\,.$$

In view of (4.3) and (4.5),

$$Q_n[f] = \frac{1}{n}\sum_{k=0}^{n} a_k\, f_k + \frac{c}{n}\left(\Delta^4 f_0 - \Delta^4 f_{n-4}\right) - \kappa\,\Delta^5 f_{n-5}\,.$$

Making use of relations $\Delta^4 \widetilde{f}_0 = \Delta^4 f_{n-4}$, $\Delta^4 \widetilde{f}_{n-4} = \Delta^4 f_0$, $\Delta^5 \widetilde{f}_{n-5} = -\Delta^5 f_0$ and $a_k = a_{n-k}$, $k = 0, \ldots, n$ (see (4.2)), we obtain

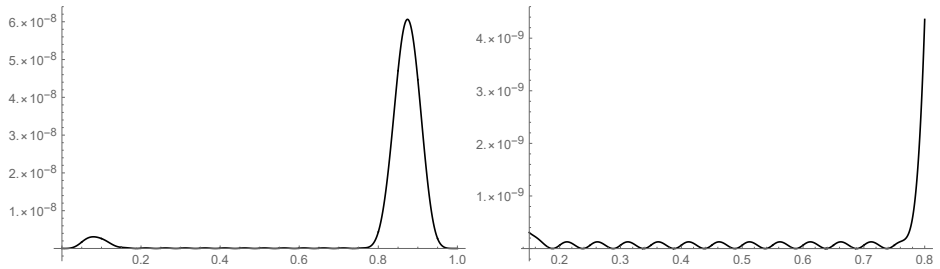
$$Q_n[\widetilde{f}] = \frac{1}{n}\sum_{k=0}^{n} a_k\, f_k + \frac{c}{n}\left(\Delta^4 f_{n-4} - \Delta^4 f_0\right) + \kappa\,\Delta^5 f_0\,.$$

Hence,

$$B(Q_n;f)=\left|Q_n[\widetilde{f}]-Q_n[f]\right|=\left|\kappa\left(\Delta^5 f_0+\Delta^5 f_{n-5}\right)+\frac{2c}{n}\left(\Delta^4 f_{n-4}-\Delta^4 f_0\right)\right|$$
$$=\frac{1}{n}\left|\left(\frac{95}{288}-c\right)\left(\Delta^5 f_0+\Delta^5 f_{n-5}\right)+2c\left(\Delta^4 f_{n-4}-\Delta^4 f_0\right)\right|.$$

Claim (ii) of Theorem 1 now follows from Proposition 1(ii). The proof of Theorem 1 is completed. Corollary 1 is a consequence of Theorem 1 and Proposition 1.

Remark 2. The magnitude of the Peano kernel $K_5(Q_n;t)$ in the interval $[x_4,x_{n-5}]$ is much smaller compared to its magnitude near the endpoints of $(0,1)$, see Figure 3. A further perturbation of Q_n of the form $Q_n'[f]=Q_n[f]+\kappa_1\,\Delta^5 f_0$ is possible, with $\kappa_1>0$ small enough so that $0\le K_5(Q_n';t)<K_5(Q_n;t)$ in (x_0,x_5). Eventually, this leads to a quadrature formula Q_n' which is positive definite of order 5 and has a slightly smaller error constant, $c_5(Q_n')<c_5(Q_n)$. The improvement however is negligible, so we decided not to perform this step.



(a) Graph of $K_5(Q_n;t)$, $t\in[0,1]$. (b) Graph of $K_5(Q_n;t)$, $t\in[x_3,x_{n-4}]$.

Figure 3: Graphs of $K_5(Q_n;t)$, $n=20$.

ACKNOWLEDGEMENT. The authors were supported by the Sofia University Research Fund through Contract No. 80-10-17/09.04.2019. The first author was supported of the Bulgarian Ministry of Education and Science under the National Research Programme "Young scientists and postdoctoral students" approved by DCM No. 577/17.08.2018.

5. REFERENCES

[1] Avdzhieva, A., Nikolov, G.: Asymptotically optimal definite quadrature formulae of 4th order. *J. Comput. Appl. Math.*, **311**, 2017, 565–582.

[2] Avdzhieva, A., Gushev, V., Nikolov, G.: Definite quadrature formulae of order three with equidistant nodes. Ann. Univ. Sofia, Fac. Math. Inf. **104**, 155–170 (2017).

[3] Avdzhieva, A., Gushev, V., Nikolov, G.: Definite quadrature formulae of order three based on the compound midpoint rule. In: *Numerical Methods and Applications. 9th International Conference, NMA 2018* (G. Nikolov, N. Kolkovska and K. Georgiev, Eds.), Lecture Notes in Computer Science **11189**, Springer Nature, 227–234, 2019.

[4] Braß, H.: *Quadraturverfahren.* Vandenhoeck & Ruprecht, Göttingen, 1977.

[5] Förster, K.-J.: Survey on stopping rules in quadrature based on Peano kernel methods. *Suppl. Rend. Circ. Math. Palermo*, Ser. II **33**, 1993, 311–330.

[6] Jetter, K.: Optimale Quadraturformeln mit semidefiniten Peano-Kernen. *Numer. Math.* **25**, 1976, 239–249.

[7] Köhler, P., Nikolov, G.: Error bounds for optimal definite quadrature formulae. *J. Approx. Theory*, **81**, 1995, 397–405.

[8] Lange, G.: *Beste und optimale definite Quadraturformel.* Ph.D. Thesis, Technical University Clausthal, Germany, 1977.

[9] Lange, G.: Optimale definite Quadraturformel. In: *Numerische Integration* (G. Hämmerlin, ed.), ISNM vol. 45, Birkhäuser, Basel, Boston, Stuttgart, 1979, pp. 187–197.

[10] Nikolov, G.: On certain definite quadrature formulae. *J. Comput. Appl. Math.*, **75**, 1996, 329–343.

[11] Peano, G.: Resto nelle formule di quadratura espresso con un integrale definito. *Atti della Reale Accademia dei Lincei: Rendiconti* (Ser. 5), **22**, 1913, 562–569.

[12] Schmeisser, G.: Optimale Quadraturformeln mit semidefiniten Kernen. *Numer. Math.* **20**, 1972, 32–53.

Received on December 17, 2019

Ana Avdzhieva, Geno Nikolov

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5, J. Bourchier blvd., BG-1164 Sofia
BULGARIA

E-mails: aavdzhieva@fmi.uni-sofia.bg
geno@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 106

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 106

NONEXISTENCE OF (17, 108, 3) TERNARY ORTHOGONAL ARRAY

SILVIA BOUMOVA, TANYA MARINOVA, TEDIS RAMAJ, MAYA STOYANOVA

We develop a combinatorial method for computing and reducing of the possibilities of distance distributions of ternary orthogonal array (TOA) of given parameters (n, M, τ). Using relations between distance distributions of arrays under consideration and their relatives we prove certain constraints on the distance distributions of TOAs. This allows us to collect rules for removing distance distributions as infeasible. The main result is nonexistence of $(17, 108, 34)$ TOA. Our approach allows substantial reduction of the number of feasible distance distributions for known arrays. This could be helpful for other investigations over the classification of the ternary orthogonal arrays.

Keywords: Hamming space, orthogonal arrays, Krawtchouk polynomials, distance distributions.

2010 Math. Subject Classification: Primary: 05B15; Secondary: 94B25.

1. INTRODUCTION

Let $H(n, 3)$ be the Hamming space of dimension n over the alphabet $\{0, 1, 2\}$. The Hamming distance $d(x, y)$ between two points $x, y \in H(n, 3)$ is equal to the number of coordinates where they differ.

Definition 1.1. An orthogonal array (OA) of strength τ and index λ in $H(n, 3)$ (also called ternary orthogonal array or TOA), consists of the rows of an $M \times n$ matrix C with the property that every $M \times \tau$ submatrix of C contains all ordered τ-tuples of $H(\tau, 3)$, each one exactly $\lambda = M/3^\tau$ times as rows. We denote such orthogonal array as (n, M, τ) TOA.

Let $C \subset H(n,3)$ be an (n, M, τ) TOA and $c \in H(n,3)$ is a fixed point of the space.

Definition 1.2. The distance distribution of C with respect to the point c is the $(n+1)$-tuple $W = W(c) = (w_0, w_1, \ldots, w_n)$, where

$$w_i = |\{x \in C \mid d(x,c) = i\}|, \quad i = 0, \ldots, n.$$

If $w_0 \geq 1$ then the point c is a word in the array C and such points we denote as internal points. The case $w_0 = 0$ denote an external point for the orthogonal array C. For simplicity and differentiation the distance distributions of internal and external points will be denoted as $P = P(c) = (p_0, p_1, \ldots, p_n)$ and $Q = Q(c) = (q_0, q_1, \ldots, q_n)$, respectively.

Let n, M and $\tau \leq n$ be fixed. The sets of all possibilities for distance distributions of a given (n, M, τ) TOA with respect to internal points and external points are denoted by $P(n, M, \tau)$ and $Q(n, M, \tau)$, respectively. Their union is the set $W(n, M, \tau) = P(n, M, \tau) \cup Q(n, M, \tau)$.

There is a method [6, 2] for computation of the sets $P(n, M, \tau)$, $Q(n, M, \tau)$ and $W(n, M, \tau)$. This method is based on the fact that each orthogonal array is a design in $H(n,3)$.

We consider the Hamming space $H(n,3)$ as polynomial metric space where zonal orthogonal polynomials are the Krawtchouk polynomials. For fixed n and $q = 3$, the (normalized) Krawtchouk polynomials are defined by

$$Q_i^{(n)}(t) := \frac{1}{r_i} K_i^{(n,3)}(z),$$

where $z = n(1-t)/2$, $r_i := 2^i \binom{n}{i}$, and

$$K_i^{(n,3)}(z) := \sum_{j=0}^{i} (-1)^j 2^{i-j} \binom{z}{j} \binom{n-z}{i-j},$$

$i = 0, 1, \ldots, n$, are the (usual) Krawtchouk polynomials [1, 14].

Definition 1.3. [10] A code $C \subset H(n,3)$ is a τ-design if and only if for every real polynomial $f(t)$ of degree at most τ and for every point $c \in H(n,3)$ the equality

$$\sum_{x \in C} f(\langle c, x \rangle) = f_0 |C|$$

holds, where f_0 is the first coefficient in the expansion $f(t) = \sum_{i=0}^{n} f_i Q_i^{(n)}(t)$ and $\langle c, x \rangle = 1 - 2d(c,x)/n$.

Since every (n, M, τ) TOA is a τ-design, the following theorem holds.

Theorem 1.4 ([6, 2]). *Let $C \subset H(n,3)$ is an (n, M, τ) TOA and $c \in H(n,q)$ is a fixed point. The following propositions are valid*

(a) If $c \in C$, for the distance distribution of C with respect of c the following system holds:

$$\sum_{i=0}^{n} p_i \left(1 - \frac{2i}{n}\right)^k = b_k|C|, k = 0, 1, \ldots, \tau, \tag{1.1}$$

(b) If $c \notin C$, for the distance distribution of C with respect of c the following system holds:

$$\sum_{i=1}^{n} q_i \left(1 - \frac{2i}{n}\right)^k = b_k|C|, k = 0, 1, \ldots, \tau, \tag{1.2}$$

where $b_k = f_0$ is the first coefficient in the expansion of the polynomial t^k by the normalized Krawchouk polynomials.

Through this theorem all initially feasible distance distributions of TOA of parameters (n, M, τ) can be computed effectively for relatively small n and τ.

Boyvalenkov and two of authors [4] have presented and implemented an algorithm for investigation binary orthogonal arrays. In this paper we develop a similar algorithm that reduces the possible elements of the set $P(n, M, \tau)$. This algorithm uses some connections between a given TOA and its related TOAs. During the implementation of the algorithm this set $P(n, M, \tau)$ is changed by ruling out some distance distributions.

In Section 2 we prove several relations between distance distributions of arrays under consideration and their relatives. This imposes significant constraints on the targeted TOAs and allows us to collect rules for removing distance distributions from the set $P(n, M, \tau)$. The algorithm and one nonexistence result are described in Section 3.

2. RELATIONS BETWEEN DISTANCE DISTRIBUTIONS OF (N, M, τ) TOA AND ITS DERIVED

Let n, M and $2 \leq \tau < n$ be fixed. Let $C \subset H(n,3)$ be an (n, M, τ) TOA with sets of distance distributions $P(n, M, \tau)$, $Q(n, M, \tau)$ and $W(n, M, \tau)$ after calculating the results of the systems (1.1) and (1.2). We proceed with the removing a column from C. Using well-known properties of the orthogonal arrays [8] we obtain another orthogonal array C' with the same strength and cardinality and length $n-1$. Without loss of generality (see [8]) let $P \in P(n, M, \tau)$ be the distance distribution of C with respect to $c = \mathbf{0} \in C$. Then the point $c' = \mathbf{0} \in C'$ and the distance distribution of C' with respect to c' we denote by $P' \in P(n-1, M, \tau)$. The scheme of this construction is shown in the Figure 1 bellow.

Definition 2.1. For every $i \in \{0, 1, \ldots, n\}$ the submatrix which consists of the rows of C with i nonzero coordinates is called an i-block.

It follows from the definition of distance distribution that the i-block is a $w_i \times n$ matrix. Next we denote by y_i the number of the zeros in the intersection of the fixed column of C and the rows of the i-block. The number of the nonzero elements in this intersection is denoted by $\overline{y}_i$.

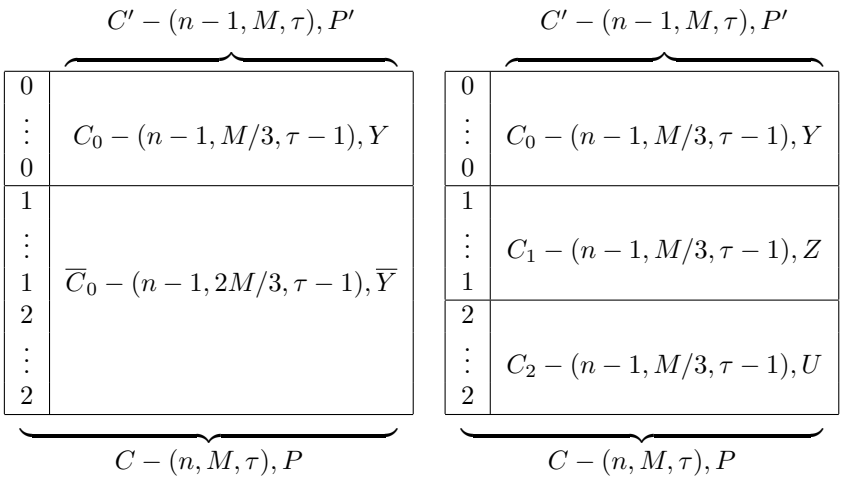


Figure 1

Theorem 2.2. *The nonnegative integer numbers y_i and $\overline{y}_i$, for $i = 0, 1, \ldots, n$, satisfy the following system of linear equations*

$$\left| \begin{array}{l} y_i + \overline{y}_i = p_i,\ i = 1, 2, \ldots, n-1 \\ y_i + \overline{y}_{i+1} = p'_i,\ i = 0, 1, \ldots, n-1 \\ y_0 = p_0,\ \overline{y}_n = p_n \\ \overline{y}_i, y_i \in \mathbb{Z},\ x_i \geq 0,\ y_i \geq 0,\ i = 0, 1, \ldots, n \end{array} \right. . \tag{2.1}$$

Proof. From the definition of the numbers y_i and $\overline{y}_i$ directly we obtain the equalities:

$$y_i + \overline{y}_i = p_i, \quad i = 1, \ldots, n-1, \quad \overline{y}_n = p_n, \quad y_0 = p_0.$$

Let us have a look at the i-th element p'_i in the distance distribution P' of C' with respect to $c' = \mathbf{0} \in C'$. It denotes the number of points in C' that have exactly i nonzero coordinates. Such points can be obtained from C by removing the first column in two possible ways. The first one is from a point which first coordinate is zero and has i nonzero entries. The number of these words of C is exactly y_i. Second is from a point of C with $i+1$ nonzero entries such that one of them is in the first column. These are the points in the $(i+1)$-block and their number is $\overline{y}_{i+1}$. Therefore

$$y_i + \overline{y}_{i+1} = p'_i$$

for every $i = 0, 1, \ldots, n-1$. □

Remark 2.3. There is a generalization of Theorem 2.2, i.e. the assertion is valid not only for internal points but also for every distance distribution $W \in W(n, M, \tau)$. However, for the purposes of the algorithm in the next section we can limit to the distance distributions in $P(n, M, \tau)$.

Corollary 2.4. *The distance distribution $P \in P(n, M, \tau)$ is not feasible if no system (2.1) obtained when P' runs $P(n-1, M, \tau)$ has a solution.*

Corollary 2.4 rules out some distance distributions P but its main purpose is to define feasible pairs (P, P') which we will investigate further.

If we order the elements in some column (for example the first column) of (n, M, τ) TOA C and remove this column (as shown in the Figure 1) we obtain three different $(n-1, M/3, \tau-1)$ TOAs. One of them is C_0 – the TOA obtained from C by ordering the zeros in the first column and taking the corresponding points of C'.

Theorem 2.5. *The vector Y is the distance distribution of C_0 with respect to the internal point $c' = \mathbf{0} \in C'$, i.e. $Y = (y_1, y_2, \ldots, y_{n-1}) \in P(n-1, M/3, \tau-1)$.*

Proof. We know from the definition of i-block that y_i is exactly the numbers of points in C with distance i to the fixed point $c = \mathbf{0} \in C$. Therefore, the number of points in C_0 with distance i to the point $c' = \mathbf{0} \in C'$ is exactly y_i. □

Corollary 2.6. *If $Y \notin P(n-1, M/3, \tau-1)$ then the pair (P, P') is not feasible.*

Let us return to the construction in Figure 1. We denote by C_1 and C_2 the orthogonal arrays corresponding to the sorted and removed elements one and two in the first column of C, respectively. Another property of the orthogonal arrays says that an union of C_1 and C_2, two ternary orthogonal arrays with parameters $(n-1, M/3, \tau-1)$ and $(n-1, M/3, \tau-1)$ is also a TOA with parameters $(n-1, 2M/3, \tau-1)$. This union will be denoted by $\overline{C}_0$. Note that there may be repeating points in the considered orthogonal arrays.

Theorem 2.7. *If $\overline{y}_0 \geq 1$, then $\overline{Y}$ is the distance distribution of $\overline{C}_0$ with respect to the fixed point $c' = \mathbf{0}$, i.e. $\overline{Y} \in P(n-1, 2M/3, \tau-1)$.*

Proof. The nonzero entries of the first column of C are selected and removed and this way the orthogonal array $\overline{C}_0$ is obtained. We have from the definition of i-block that $\overline{y_i}$ is exactly the numbers of points in C with distance i to the point $c = \mathbf{0}$. Therefore, the numbers of points in $\overline{C}_0$ with distance i to the point $c' = \mathbf{0}$ is exactly $\overline{y}_i$. The condition $\overline{y}_0 \geq 1$ determines that we check if $\overline{C}_0$ contains the point $c' = \mathbf{0}$, i.e. $c' \in \overline{C}_0$ is the internal point and $\overline{Y} \in P(n-1, 2M/3, \tau-1)$. □

Corollary 2.8. *If* $\overline{y}_0 \geq 1$ *and* $\overline{Y} \notin P(n-1, 2M/3, \tau-1)$*, then the pair* (P, P') *is not feasible.*

After applying Corollary 2.8 for fixed distance distribution $P \in P(n, M, \tau)$ we continue with the remaining feasible pairs (P, P'). Let

$$(\overline{y}_0^{(r)} = 0, \overline{y}_1^{(r)}, \ldots, \overline{y}_n^{(r)}; y_0^{(r)}, y_1^{(r)}, \ldots, y_{n-1}^{(r)}, y_n^{(r)} = 0), \quad r = 1, \ldots, s,$$

be all solutions of system (2.1) when P' runs the set $P(n-1, M, \tau)$ such that the corresponding pair (P, P') is not ruled out by Corollaries 2.6 and 2.8. Denote by k_r the numbers of columns corresponding to the r-th solution of the system (2.1) for $r = 1, \ldots, s$.

After the sieve from Corollaries 2.6 and 2.8, we formulate another necessary condition for the existence of C.

Theorem 2.9. *The system*

$$\left|\begin{array}{lllll} k_1 & +k_2 & +\ldots & +k_s & = n \\ k_1\overline{y}_1^{(1)} & +k_2\overline{y}_1^{(2)} & +\ldots & +k_s\overline{y}_1^{(s)} & = p_1 \\ k_1\overline{y}_2^{(1)} & +k_2\overline{y}_2^{(2)} & +\ldots & +k_s\overline{y}_2^{(s)} & = 2p_2 \\ \vdots & & & & \\ k_1\overline{y}_n^{(1)} & +k_2\overline{y}_n^{(2)} & +\ldots & +k_s\overline{y}_n^{(s)} & = np_n \\ k_j \in \mathbb{Z}, & k_j \geq 0, & j = 1, & \ldots, s & \end{array}\right. \tag{2.2}$$

with respect to $k_1, k_2, \ldots, k_s$ *has a solution, i.e. the ternary orthogonal array* C *of parameters* (n, M, τ) *exists if the system (2.2) has a solution.*

Proof. For every cutting of a column of C we solve the system (2.1) for every possible $P' \in P(n-1, M, \tau)$. Let i be fixed. In the i-block the numbers of nonzero entries is exactly ip_i. On the other hand we know that $\overline{y}_i$ is the number of points in i-block with entries 1 or 2 in the first column. So the count of nonzero entries in i-block is equal to $k_1\overline{y}_i^{(1)} + k_2\overline{y}_i^{(2)} + \ldots + k_s\overline{y}_i^{(s)}$. Therefore for every $i = 0, \ldots, n$ the equalities in the system (2.2) hold. □

3. OUR ALGORITHM AND ONE NONEXISTENCE RESULT

Based on the observations and conclusions in the previous section an algorithm for reducing the feasible distance distributions in the set $P(n, M, \tau)$ for fixed n, M and τ can be developed. If the result from the algorithm is an empty set we can conclude that ternary orthogonal arrays with parameters (n, M, τ) do not exist.

By calculating the sets $P(n-1, 2M/3, \tau-1)$ we observe that these sets become very large so the Theorem 2.7 and the Corollary 2.8 are not easy to be applied for the computations. Even more, when $\tau > 3$ the set $P(n-1, 2M/3, \tau-1)$ itself should

be reduced through generation and reduction of the set $P(n-2, 4M/9, \tau-2)$ which cardinality is even bigger. That is the reason why the algorithm is based only on Theorems 2.2, 2.5 and 2.9 and their corollaries.

First, we generate with Theorem 1.4 the following rows of distance distribution sets when the length varies from τ to n

$$P(\tau, M, \tau), P(\tau+1, M, \tau), \ldots, P(n, M, \tau)$$

$$P(\tau-1, M/3, \tau-1), P(\tau, M/3, \tau-1), \ldots, P(n-1, M/3, \tau-1)$$

$$\ldots$$

For fixed j, $j = \tau, \tau+1, \ldots, n$, the algorithm is applied over the set $P(j, M, \tau)$ and its derived as the algorithm ends either if $j = n$ or if an empty set is obtained for some j.

From the set $P(j, M, \tau)$ a distance distribution P is selected. For this fixed distance distribution and for every distance distribution in $P' \in P(j-1, M, \tau)$ the system (2.1) is resolved. If for every P' this system does not have a solution, the distance distribution P is ruled out from $P(j, M, \tau, 3)$, (see Corollary 2.4).

Otherwise, for the solution $(Y, \overline{Y})$ we check the condition in Theorem 2.5. If it is not fulfilled the pair (P, P') is not feasible (see Corollary 2.6).

For every feasible (P, P') we collect the solution $(Y, \overline{Y})$. After all solutions are collected when P' runs over $P(j-1, M, \tau)$ the system (2.2) is solved. If there is no solution, the distance distribution P is ruled out from the set $P(j, M, \tau)$, (see Theorem 2.9).

This is the step for fixed j. After reducing the elements of $P(j, M, \tau)$ we increase j by 1 and proceed with the next set of investigation of the distance distributions $P(j+1, M, \tau)$. We continue until $j < n$. If the set $P(j, M, \tau)$ is empty for some $j_0 < n$ the algorithm ends with the conclusion that (j, M, τ) TOAs do not exist for $j = j_0, j_0+1, \ldots, n$.

For the sake of clarity a pseudocode of the algorithm is provided bellow.

In what follows, our investigation is focused on the set $P(17, 108, 3)$, one of the cases in [11] where the existence was marked as undecided. Moreover, for $n = 12, \ldots, 17$ there are no evidence whether orthogonal arrays with parameters $(n, 108, 3)$ exist. Several teams of authors ([5, 12, 13]) have investigated these among many others cases, but the issue of the existence of a ternary orthogonal array with parameters $(17, 108, 3)$ has not been clarified so far.

We calculate all possible distance distributions for internal point, i.e. we generate the sets $P(n, 108, 3)$ for $n = 3, \ldots, 17$. Along with this we need the sets $P(n, 36, 2)$ for $n = 2, \ldots, 16$. The Algorithm 1 is applied on these sets. First the sets $P(j, 36, 2)$ are reduced, starting from $j = 2$. After that the sets $P(n, 108, 3)$ are reduced. Then last reduced set is $P(17, 108, 3)$. In the tables below the cardinalities of all these sets are provided. In the first table the results for $|P(n, 108, 3)|$ are presented for $n = 3, 4, \ldots, 17$. The entry $A \to B$ in the first table means that

Algorithm 1 Algorithm over TOAs

```
1: procedure NDDA(P(n, M, τ), P(n − 1, M, τ), P(n − 1, M/3, τ − 1))
2:     Input: n, M, τ, P(n, M, τ), P(n − 1, M, τ), P(n − 1, M/3, τ − 1)
3:     filteredP ← empty set
4:     for P ∈ P(n, M, τ) do
5:         allȲ ← empty set
6:         for P′ ∈ P(n − 1, M, τ) do
7:             Y, Ȳ ← solve system (2.1) for integer nonnegative solutions
8:             if no integer solutions then
9:                 next;
10:            if Y ∈ P(n − 1, M/3, τ − 1) then
11:                add Ȳ to allȲ
12:        if allȲ is empty then
13:            add P to filteredP
14:        else
15:            if system (2.2) has no integer nonnegative solutions then
16:                add P to filteredP
17:    Output: P(n, M, τ) \ filteredP
```

in the beginning there is A initially feasible distance distributions of $(n, 108, 3)$ TOA, i.e. the set $P(n, 108, 3)$ has A elements, starting from $n = 3$ in the first row and the first column and ending to $n = 17$ in the third row and the fifth column, successively. The number B after the arrow (in corresponding entry) represents the reduced value B of elements in the set $P(n, 108, 3)$ in the end of the algorithm, $n = 3, 4, \ldots, 17$. Analogously, the results for $|P(n, 36, 2)|$ are presented in the second table for $n = 2, 3, \ldots, 16$ and $\tau = 2$.

$\|P(n,108,3)\|$:	$1 \to 1$	$4 \to 4$	$18 \to 16$	$48 \to 43$	$113 \to 89$
	$271 \to 208$	$440 \to 368$	$701 \to 540$	$1002 \to 702$	$879 \to 699$
	$901 \to 660$	$631 \to 337$	$119 \to 29$	$49 \to 6$	$10 \to 0$

$\|P(n,36,2)\|$:	$1 \to 1$	$4 \to 4$	$16 \to 14$	$31 \to 30$	$52 \to 49$
	$85 \to 79$	$109 \to 105$	$121 \to 109$	$127 \to 111$	$108 \to 100$
	$85 \to 79$	$62 \to 50$	$28 \to 26$	$12 \to 11$	$6 \to 4$

The zero in the last element $10 \to 0$ of the first table corresponds to the the number of elements in the set $P(17, 108, 3)$, i.e. our algorithm ends with the empty set $P(17, 108, 3) = \emptyset$. Therefore, we obtain the following nonexistence result.

Theorem 3.1. *There exist no ternary orthogonal array of parameters* $(17, 108, 3)$.

All calculations in this paper were performed by programs in Maple. All results (in particular all possible distance distributions in the beginning) can be seen at [15]. All programs are available upon request.

ACKNOWLEDGEMENTS. The research of the first author was supported, in part, by the Bulgarian National Science Fund (NSF) under contract KP-06-N32/1-2019. The research of the third author was supported, in part, by the Bulgarian NSF under contract DN2-02/2016. The research of the fourth author was supported, in part, by the Bulgarian NSF under contract KP-06-N32/2-2019.

4. REFERENCES

[1] Abramowitz, M., Stegun, I. A.: *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*. New York, Dover, 1964.

[2] Boyvalenkov, P., Kulina, H.: Investigation of binary orthogonal arrays via their distance distributions. *Probl. Inf. Transm.*, **49(4)**, 2013, 320–330.

[3] Boyvalenkov, P., Kulina, H., Marinova, T., Stoyanova, M.: Nonexistence of binary orthogonal arrays via their distance distributions. *Probl. Inf. Transm.*, **51(4)**, 2015, 326-334.

[4] Boyvalenkov, P., Marinova, T., Stoyanova, M.: Nonexistence of a few binary orthogonal arrays. *Discrete Appl. Math.*, **217(2)**, 2017, 144–150.

[5] D. A. Bulutoglu, D. A., Margot, F.: Classification of orthogonal arrays by integer programming. *J. Statist. Plann. Inference*, **138**, 2008, 654–666.

[6] Delsarte, P.: An Algebraic Approach to the Association Schemes in Coding Theory. *Philips Res. Rep. Suppl.*, **10**, 1973.

[7] Delsarte, P., Levenshtein, V. I.: Association schemes and coding theory. *IEEE Trans. Inform. Theory*, **44**, 1998, 2477–2504.

[8] Hedayat, A., Sloane, N. J. A., Stufken, J.: *Orthogonal Arrays: Theory and Applications*. Springer-Verlag, New York, 1999.

[9] Levenshtein, V. I.: Krawtchouk polynomials and universal bounds for codes and designs in Hamming spaces. *IEEE Trans. Infor. Theory*, **41**, 1995, 1303–1321.

[10] Levenshtein, V. I.: Universal bounds for codes and designs. In: *Handbook of Coding Theory*. (V. S. Pless and W. C. Huffman, Eds.), **Ch. 6**, Elsevier, Amsterdam, 1998, 499–648.

[11] http://neilsloane.com/oadir/index.html

[12] Schoen, E.D., Eendebak, P. T., Nguyen, M. V. M.: Complete enumeration of pure-level and mixed-level orthogonal arrays. *J. Combin. Des.*, **18**, 2009, 123–140.

[13] Seiden, E., Zemach, R.: On orthogonal arrays. *Ann. Math. Statist.*, **37**, 1996, 1355–1370.

[14] Szegő, G.: *Orthogonal Polynomials*. American Mathematical Society Colloquium Publications 23, AMS, Providence, RI, 1939.

[15] https://store.fmi.uni-sofia.bg/fmi/algebra/stoyanova/toa.html

Received on February 6, 2020
Received in a revised form on March 17, 2020

Silvia Boumova

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 blvd. J. Bourchier, BG-1164 Sofia
BULGARIA

E-mail: boumova@fmi.uni-sofia.bg

Tanya Marinova

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 blvd. J. Bourchier, BG-1164 Sofia
BULGARIA

E-mail: tanya.marinova@fmi.uni-sofia.bg

Tedis Ramaj

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 blvd. J. Bourchier, BG-1164 Sofia
BULGARIA

E-mail: tramaj@fmi.uni-sofia.bg

Maya Stoyanova

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 blvd. J. Bourchier, BG-1164 Sofia
BULGARIA

E-mail: stoyanova@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 106

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 106

QUALITATIVE ANALYSIS OF A MATHEMATICAL MODEL OF CALCIUM DYNAMICS INSIDE THE MUSCLE CELL

ZDRAVKA D. NEDYALKOVA, TIHOMIR B. IVANOV

In this paper, we consider a mathematical model of calcium dynamics inside the muscle cell, proposed by Williams. We make a qualitative study of the model solutions. In particular, we study the existence and stability of equilibrium points of the model with respect to the model parameters in two limiting cases—when a constant stimulus is present and when there is no stimulus that triggers muscle activity. Numerical examples are given for each case, in order to illustrate the analytic results.

Keywords: Calcium dynamics, muscle cell, stability analysis, equilibrium points, dynamical system.

2010 Math. Subject Classification: 37N25.

1. INTRODUCTION

A general cross-section of a skeletal muscle can be seen in Figure 1.

The hierarchical structure in the skeletal muscle is described as follows [3]:

- A skeletal muscle is surrounded by fibrous tissue, called *epimysium*. It serves as a protection shield and protects the muscle from friction against other muscles and bones;
- Within the muscle, there is another connective tissue, the *perimysium*, which connects muscle fibers into bundles, called fascicles. A large muscle contains more fibers in each bundle, while a small one contains less;

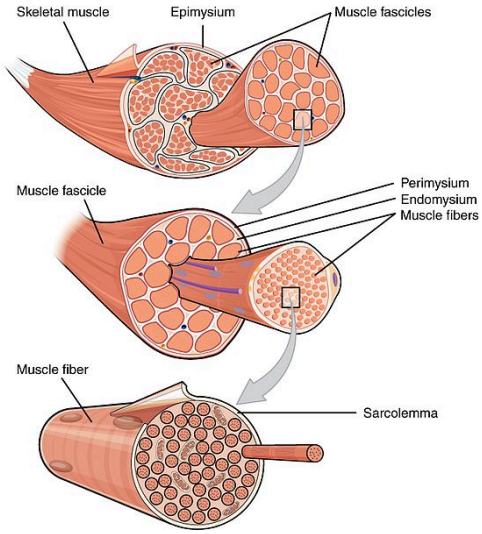


Figure 1: Skeletal muscle morphology [1].

- Inside the fascicles there is another connective tissue, which isolates each fiber, called *endomysium*;
- The endomysium contains the muscle cells/fibers or *myofibers*, formed in the process of myogenesis. Every myofiber can have a different length up to several centimeters, which is the reason that the muscle cells have multiple nuclei.

In Figure 2, the structure of a muscle fiber is shown. The membrane of the muscle cell, called sarcolemma, contains a bunch of tubes called myofibrils—the contractile units of the cell. Each muscle fiber contains hundreds or thousands of myofibrils, which are divided into segments called sarcomeres. The sarcomeres are the basis for muscle contraction theory, known as the sliding filament theory.

Each sarcomere is separated by a border, called a *Z-line* or a *Z-disc*. As in Figure 2, the sarcomere is composed of long fibrous proteins. It contains two main types of long protein chains, called **filaments**[1]—thin, made of actin protein strands, and thick—composed of myosin protein strands. Muscle contraction happens, because of thin and thick filaments sliding past each other through complex biochemical processes, triggered by calcium dynamics inside the muscle cell.

Each muscle cell has the so-called **sarcoplasmic reticulum** (SR), which is a membrane-bound network of tubules that wraps the myofibrils. **The main func-**

[1]We have marked in bold the crucial terms related to the muscle structure that will be used throughout the paper.

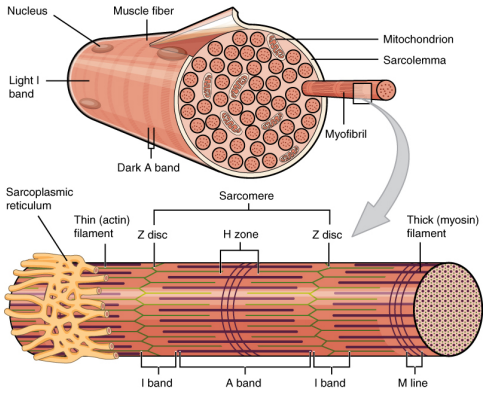


Figure 2: A muscle fiber structure [1].

tion of the SR is to store calcium ions.

It has been shown that calcium plays a central role in the process of activation of a muscle cell. In general, the process that leads to a contraction of a muscle fiber can be described in the following steps [9]:

1. An impulse travels through the axon of the motor neuron to the axon terminal;
2. At the axon terminal there are voltage-gated calcium channels, which open due to the action potential and calcium ions diffuse into the terminal;
3. The calcium presence in the axon terminal opens the so-called synaptic vesicles to release a neurotransmitter, called acetylcholine (ACh);
4. The released ACh diffuses, crosses the synaptic cleft and binds to ACh receptors on the motor end plate of the muscle, which contains cation channels. The cation channels open and sodium ions enter the muscle fiber, causing potassium ions to exit the muscle fiber;
5. The input flux of the sodium ions changes the membrane potential, causing depolarization or the so-called end plate potential (EPP). Once the membrane potential reaches a threshold value, an axon potential propagates along the sarcolemma;
6. Inside the muscle cell, the sarcoplasmic reticulum (SR), which is a network of tubules that regulates calcium concentration, then releases calcium so that it can bind to contractile filaments (actin and myosin filaments) in the muscle fiber. The binding of calcium to the contractile filaments (CFs) causes a shift in the filaments and allows them to bind to each other and contract. The latter is the so-called contractile filament theory, developed independently by two research teams in the 20th century [8].

Various authors have worked on the mathematical description of calcium dynamics inside the muscle cell, see e.g. [5, 6, 7] and the references therein. In the present work, we consider a mathematical model proposed by Williams in [7]. Here, we study the local asymptotic behaviour of the model solutions, depending on the parameter values in the two limiting cases—when a constant stimulus is present and when there is no stimulus to trigger muscle activity.

The paper is structured as follows. In Section 2, we derive the mathematical model. The general properties like existence and uniqueness, positivity, and boundedness of the solutions are shown in Section 3. An analytic study of model's dynamics is carried out in Section 4. In particular, existence and local stability study of the equilibria is derived. Numerical experiments are given in Section 5 to illustrate the analytic results and to further discuss their biological meaning in Section 6.

2. MATHEMATICAL MODEL

As discussed earlier, when a nerve impulse comes to the muscle, the action potential results in the release of Ca^{2+} ions from the SR. Ca^{2+} ions then flow into the sarcomere where the CFs are situated. Then, Ca^{2+} ions start binding to the receptors in the CFs and as a result, the filaments start sliding, causing the sarcomere to shorten. When the stimulus is turned off, the Ca^{2+} ions are transported back into the SR and the sarcomere relaxes. Having in mind the aforementioned, one needs to model the dynamics of calcium ions, SR, and CFs, in order to understand the process of muscle contraction.

For this purpose, we consider a mass action kinetics model, proposed by Williams [7], further considered by McMillen [6] and used by Meredith in [5]. The model is based on the principle of mass action kinetics, which assumes that the rate of a chemical reaction is proportional to the concentration of the reactants. Let us denote the following:

- c—concentration of free calcium ions;
- r_u—concentration of unbound sarcoplasmic reticulum sites;
- r_b—concentration of bound sarcoplasmic reticulum sites;
- f_u—concentration of unbound CF sites;
- f_b—concentration of bound CF sites;
- k_1—rate of release of calcium ions from the SR;
- k_2—rate of binding of calcium ions to the SR;
- k_3—rate of binding of calcium ions to the CFs;

- k_4—rate of release of calcium ions from the CFs.

The flow of calcium is illustrated in Figure 3:

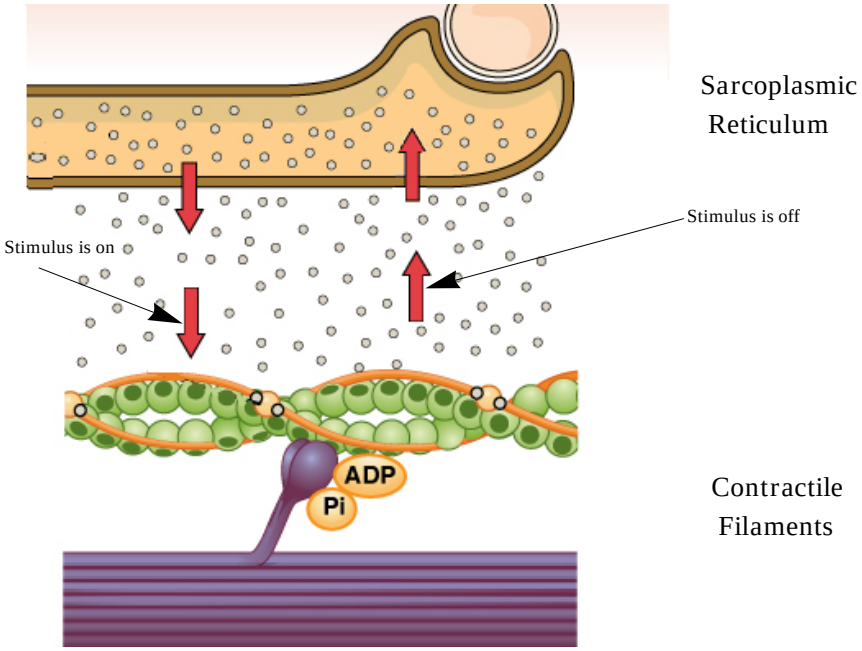


Figure 3: Flow of calcium in the muscle cell.

Based on the principle of mass action, the following statements are valid:

1. When the stimulus is on, i.e., when there is an action potential in the muscle cell, the rate of unbinding of calcium ions from the SR is proportional to the concentration of calcium-bound SR sites with a rate constant k_1;
2. When the stimulus is off, the rate of binding of calcium ions to the SR is proportional to the product of the concentrations of free calcium ions and unbound SR calcium-binding sites with a rate constant k_2;
3. The rate of binding of calcium ions to the CF sites is proportional to the product of the concentrations of free calcium ions and unbound filament sites with a rate constant k_3.

Further, because of empirical evidence, the rate of release of calcium ions from the CFs is chosen to be proportional to the product of concentration of bound and unbound filament sites with a rate constant k_4. This is meant to account for some

cooperativity between the bound and unbound CF sites in the process of calcium release.

In mathematical terms, the above assumptions result in the following system of five ODEs:

$$\begin{aligned}
\frac{dc}{dt} &= k_1 r_b - k_2 r_u c - k_3 f_u c + k_4 f_b f_u, \\
\frac{dr_b}{dt} &= -k_1 r_b + k_2 r_u c, \\
\frac{dr_u}{dt} &= k_1 r_b - k_2 r_u c, \\
\frac{df_b}{dt} &= k_3 c f_u - k_4 f_b f_u, \\
\frac{df_u}{dt} &= -k_3 c f_u + k_4 f_b f_u,
\end{aligned} \tag{2.1}$$

where k_1 and k_2 are non-negative coefficients and k_3, k_4 are positive constants. Further, the following assumptions are made by Williams [7]:

1. when the stimulus is on, $k_1 > 0,\ k_2 = 0$;
2. when the stimulus is off, $k_1 = 0, k_2 > 0$.

Adding together the first, second, and fourth equations, it follows that the total amount of calcium is constant:

$$c + f_b + r_b = C. \tag{2.2}$$

Analogously, one can show that the total numbers of bound and unbound SR and CF sites are also constant, i.e,

$$\begin{aligned}
r_u + r_b &= S, \\
f_b + f_u &= F,
\end{aligned} \tag{2.3}$$

where S and F are the total numbers of SR and CF sites.

By using (2.2)–(2.3), one reduces the ODE system (2.1) to the following two-dimensional model for the concentrations of free calcium ions and calcium-bound sites:

$$\begin{aligned}
\frac{dc}{dt} &= (k_4 f_b - k_3 c)(F - f_b) + k_1 (C - c - f_b) + k_2 c (C - S - c - f_b), \\
\frac{df_b}{dt} &= -(k_4 f_b - k_3 c)(F - f_b).
\end{aligned} \tag{2.4}$$

Further, we scale the model by the total amount of the CF sites F:

$$\begin{aligned}
&\hat{f}_b = f_b/F, \quad \hat{c} = c/F, \quad \hat{C} = C/F, \quad \hat{S} = S/F, \\
&\hat{k}_2 = F k_2, \quad \hat{k}_3 = F k_3, \quad \hat{k}_4 = F k_4.
\end{aligned} \tag{2.5}$$

Substituting (2.5) in (2.4) and skipping the hats for notational simplicity, one obtains

$$\begin{aligned} \frac{dc}{dt} &= (k_4 f_b - k_3 c)(1 - f_b) + k_1 (C - c - f_b) + k_2 c (C - S - c - f_b), \\ \frac{df_b}{dt} &= -(k_4 f_b - k_3 c)(1 - f_b). \end{aligned} \tag{2.6}$$

Remark 1. The above scaling leads to certain restrictions for f_b and c, which we shall use later in the qualitative analysis of the system (2.6). Dividing both sides of (2.2) and (2.3) by F, it follows that:

$$\hat{c} + \hat{f}_b + \hat{r}_b = \hat{C},$$
$$\hat{f}_b + \hat{f}_u = 1.$$

From the latter equations and the restrictions $\hat{c} \geq 0,\ \hat{f}_b \geq 0,\ \hat{r}_b \geq 0,\ \hat{f}_u \geq 0$, we obtain

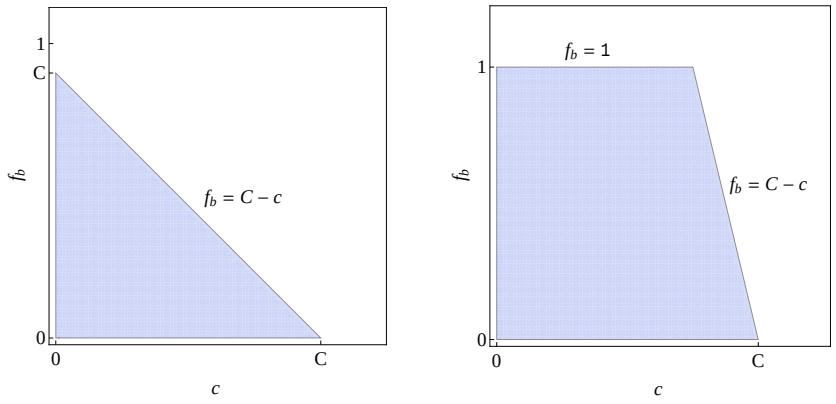
$$0 \leq \hat{c} + \hat{f}_b \leq \hat{C},$$
$$0 \leq \hat{f}_b \leq 1.$$

Therefore, system (2.6) is considered in the phase space

$$\{(c, f_b) \in \mathbb{R}^2 :\ 0 \leq c + f_b \leq C,\ 0 \leq f_b \leq 1,\ c \geq 0\}. \tag{2.7}$$

3. GENERAL PROPERTIES OF MODEL'S SOLUTIONS

Proposition 1. *The solutions of the model* (2.6) *are bounded for each choice of the model parameters.*



(a) Case $C \leq 1$. (b) Case $C > 1$.

Figure 4: Geometry of the phase space

Proof. For the proof, we shall consider the following two cases, which determine different geometry of the phase space: $C \leq 1$ and $C > 1$, see Figure 4a and Figure 4b.

Case $C \leq 1$. We shall prove that the vector field at the boundary points to the inside of the phase space. At $f_b = C - c$, it holds that

$$\frac{df_b}{dc} = -\frac{(k_4 f_b - k_3 c)(1 - f_b)}{(k_4 f_b - k_3 c)(1 - f_b) - k_2 c S} \leq -1.$$

The latter means that at this part of the boundary the slope of the vectors in the vector field is less than the slope of the line $f_b = C - c$, thus, the vector field points to the inside of the phase space.

If $c = 0$, then

$$\frac{dc}{dt} = k_4 f_b (1 - f_b) + k_1 (C - f_b) > 0$$

is valid.

Finally, when $f_b = 0$,

$$\frac{df_b}{dt} = k_3 c > 0$$

holds true.

Case $C > 1$. Let us again consider the boundary of the phase space. If $f_b = 1$, it follows that

$$\frac{df_b}{dt} = 0.$$

Thus, the solution stays on the boundary.

The results for the behaviour of the vector field on the rest boundary of the considered phase space coincide with the results in the case $C \leq 1$.

Since the vector field points to the inside of the phase space at all of its boundary, it follows that the solutions of the model (2.6) are bounded for every choice of the model parameters. □

Now, following a standard result (see, e.g., [4, pp, 17–18]), the following proposition holds true.

Proposition 2. *For the model* (2.6)*, there exists a unique trajectory through every point* $(x_0, y_0) \in \mathbb{R}^2_+$ *and it is defined for every* $t \in [0, +\infty)$.

4. LOCAL QUALITATIVE ANALYSIS OF MODEL'S DYNAMICS IN THE LIMITING CASES $k_1 = 0$, or $k_2 = 0$

In this section, we shall study qualitatively the system of differential equations (2.6). We shall consider the two limiting cases—when the stimulus is on, i.e., when $k_2 = 0,\ k_1 = const > 0$, and when the stimulus is off, i.e., $k_1 = 0,\ k_2 = const > 0$.

4.1. CASE $k_1 = const > 0,\ k_2 = 0$.

Let us first consider the case when the rate constant for binding of calcium to the SR, k_2, is equal to zero. Thus, the system we consider is:

$$\begin{aligned} \frac{dc}{dt} &= (k_4 f_b - k_3 c)(1 - f_b) + k_1 (C - c - f_b), \\ \frac{df_b}{dt} &= -(k_4 f_b - k_3 c)(1 - f_b). \end{aligned} \tag{4.1}$$

Existence of equilibrium points

The equilibria of the system (4.1) are the solutions of the system of algebraic equations

$$\begin{aligned} (k_4 f_b - k_3 c)(1 - f_b) + k_1 (C - c - f_b) &= 0, \\ -(k_4 f_b - k_3 c)(1 - f_b) &= 0. \end{aligned}$$

Solving the latter system, we find two possible equilibrium points:

$$E_1 = (C - 1, 1) \text{ and } E_2 = \left(\frac{Ck_4}{k_3 + k_4}, \frac{Ck_3}{k_3 + k_4} \right).$$

First, let us consider the conditions for the existence of the equilibrium points.

Proposition 3. *The equilibrium point E_1 exists iff $C \geq 1$. The equilibrium point E_2 exists exactly when $0 \leq C \leq \dfrac{k_3 + k_4}{k_3}$.*

Proof. In order for the equilibrium points to exist (i.e., to be in the phase space), they must satisfy the restrictions (2.7).

- Equilibrium $E_1 = (C - 1, 1)$.
  We substitute $c = C - 1$ and $f_b = 1$ in (2.7) and derive the existence condition $C \geq 1$.
- Equilibrium $E_2 = \left(\dfrac{Ck_4}{k_3 + k_4}, \dfrac{Ck_3}{k_3 + k_4} \right)$.
  We substitute the latter in (2.7) and derive:

  $$0 \leq \frac{Ck_4}{k_3 + k_4} + \frac{Ck_3}{k_3 + k_4} \leq C, \qquad 0 \leq \frac{Ck_3}{k_3 + k_4} \leq 1.$$

  The first condition is trivially fulfilled, while the latter one is satisfied for $0 \leq C \leq \dfrac{k_3 + k_4}{k_3}$. □

Local stability of equilibrium points

To analyze the local stability of the equilibrium points we use the Hartman–Grobman theorem [4]. The Jacobi matrix of (4.1) as a function of the phase variables c and f_b is:

$$J(c, f_b) = \begin{pmatrix} -k_3(1-f_b) - k_1 & k_4(1-f_b) - k_4 f_b + k_3 c - k_1 \\ k_3(1-f_b) & -k_4(1-f_b) + k_4 f_b - k_3 c \end{pmatrix}.$$

Proposition 4. *The conditions for the stability of the equilibrium points E_1 and E_2 in terms of C are given in Table 1.*

C	$0 < C < 1$	$1 < C < \dfrac{k_3+k_4}{k_3}$	$C > \dfrac{k_3+k_4}{k_3}$
E_1	$\nexists$	saddle	stable
E_2	stable	stable	$\nexists$

Table 1: Classification of equilibria for the case $k_2 = 0$ in terms of C.

Proof. We shall analyze the stability of the equilibrium points separately.

1. Local stability of $E_1 = (C-1, 1)$.
   As derived in Proposition 3, the condition for the existence of the equilibirum point is $C \geq 1$. Substituting E_1 in the Jacobi matrix, we derive:

   $$J(E_1) = \begin{pmatrix} -k_1 & -k_4 + k_3(C-1) - k_1 \\ 0 & k_4 - k_3\,(C-1) \end{pmatrix}.$$

   For the eigenvalues $\lambda_1,\ \lambda_2$ of $J(E_1)$, we have

   $$\lambda_1 = -k_1 < 0,\ \lambda_2 = k_4 - k_3(C-1).$$

   Using the latter, we consider two cases for determining the stability of E_1:

   - $k_4 - k_3\,(C-1) > 0 \iff C < \dfrac{k_3+k_4}{k_3}$.
     In this case, the eigenvalues are with opposite signs. That is, the equilibrium is a saddle point.
   - $k_4 - k_3\,(C-1) < 0 \iff C > \dfrac{k_3+k_4}{k_3}$
     In this case, both eigenvalues are negative and E_1 is asymptotically stable.

2. Local stability of $E_2 = \left(\dfrac{Ck_4}{k_3+k_4}, \dfrac{Ck_3}{k_3+k_4}\right)$.
   We compute the Jacobi matrix at E_2:

$$J(E_2) = \begin{pmatrix} -k_3\left(1-\dfrac{Ck_3}{k_3+k_4}\right)-k_1 & k_4\left(1-\dfrac{Ck_3}{k_3+k_4}\right)-k_1 \\ k_3\left(1-\dfrac{Ck_3}{k_3+k_4}\right) & -k_4\left(1-\dfrac{Ck_3}{k_3+k_4}\right) \end{pmatrix}$$

and obtain

$$\begin{aligned} \lambda_1\lambda_2 &= det\ J(E_2) = k_1\left(k_4 - k_3(C-1)\right), \\ \lambda_1+\lambda_2 &= trace\ J(E_2) = -k_1 - k_4 + k_3\left(C-1\right). \end{aligned}$$

By the existence condition for E_2, derived in Proposition 3, we conclude that the determinant is always positive, with $\lambda_1 + \lambda_2 < 0$ and, therefore, the equilibrium is asymptotically stable, whenever it exists. □

4.2. CASE $k_1 = 0,\ k_2 = const > 0$.

Let us now consider the case, when the rate constant for release of calcium from the SR, k_1, is equal to zero. Thus, we consider the following system:

$$\begin{aligned} \frac{dc}{dt} &= \left(k_4 f_b - k_3 c\right)\left(1-f_b\right) + k_2 c\left(C - S - c - f_b\right), \\ \frac{df_b}{dt} &= -\left(k_4 f_b - k_3 c\right)\left(1-f_b\right). \end{aligned} \tag{4.2}$$

Existence of equilibrium points

To find the equilibrium points of the latter system of ODEs, we solve the system of algebraic equations

$$\left(k_4 f_b - k_3 c\right)\left(1-f_b\right) + k_2 c\left(C - S - c - f_b\right) = 0, \tag{4.3}$$

$$-\left(k_4 f_b - k_3 c\right)\left(1-f_b\right) = 0. \tag{4.4}$$

The solutions of (4.4) are $f_b = 1$ and $f_b = \dfrac{k_3}{k_4}\, c$. Therefore, the four possible equilibrium points to the system (4.2) are:

$$E_1 = (0,1),\ E_2 = (C-S-1,1),\ E_3 = (0,0),\ \text{and}\ E_4 = \left(\frac{k_4(C-S)}{k_3+k_4}, \frac{k_3(C-S)}{k_3+k_4}\right).$$

We shall derive conditions for the existence of each of the equilibrium points E_1–E_4 in terms of the total amount of calcium C.

Proposition 5. *The following statements are valid:*

- *Equilibrium point E_1 exists exactly when $C \geq 1$;*
- *Equilibrium points E_2 exists if and only if $C \geq S + 1$;*
- *Equilibrium point E_3 exists for every choice of the parameters in the model (4.2);*
- *Equilibrium point E_4 exists iff $S \leq C \leq S + \dfrac{k_3 + k_4}{k_3}$.*

Proof. We shall derive the conditions for the existence of the equilibrium points separately.

1. Existence of $E_1 = (0, 1)$.
   Taking into consideration the inequalities in (2.7) and substituting $c = 0$ and $f_b = 1$, we obtain the condition $C \geq 1$.
2. Existence of $E_2 = (C - S - 1, 1)$.
   We substitute the values for c and f_b in (2.7) and derive $C \geq S + 1$.
3. Existence of $E_3 = (0, 0)$.
   The existence of this equilibrium is trivial since the point (0,0) satisfies the conditions in (2.7) and, therefore, exists for every choice of the parameters in the model (4.2).
4. Existence of $E_4 = \left(\dfrac{k_4(C-S)}{k_3+k_4}, \dfrac{k_3(C-S)}{k_3+k_4} \right)$.
   Substituting the latter in the inequalities in (2.7), we derive

$$0 \leq \frac{k_4(C-S)}{k_3+k_4} + \frac{k_3(C-S)}{k_3+k_4} \leq C,$$
$$0 \leq \frac{k_3(C-S)}{k_3+k_4} \leq 1.$$

   Taking into consideration the positivity of the constants k_3, k_4, we derive the condition $S \leq C \leq S + \dfrac{k_3 + k_4}{k_3}$. □

Local stability of equilibrium points

Proposition 6. *The conditions for the stability of the equilibrium points $E_1 = (0, 1)$, $E_2 = (C - S - 1, 1)$, $E_3 = (0, 0)$, and $E_4 = \left(\dfrac{k_4(C-S)}{k_3+k_4}, \dfrac{k_3(C-S)}{k_3+k_4} \right)$ of the system (4.2) in terms of C, given in Table 2 for the case $S < 1$ and in Table 3 for the case $S > 1$, are valid.*

C	$0<C<S$	$S<C<1$	$1<C<S+1$	$S+1<C<S+\frac{k_3+k_4}{k_3}$	$C>S+\frac{k_3+k_4}{k_3}$
E_1	$\nexists$	$\nexists$	saddle	unstable	unstable
E_2	$\nexists$	$\nexists$	$\nexists$	saddle	stable
E_3	stable	saddle	saddle	saddle	saddle
E_4	$\nexists$	stable	stable	stable	$\nexists$

Table 2: Classification of equilibria for the case $k_1 = 0$ in terms of the total amount of calcium ions C, when $S < 1$.

C	$0<C<1$	$1<C<S$	$S<C<S+1$	$S+1<C<S+\frac{k_3+k_4}{k_3}$	$C>S+\frac{k_3+k_4}{k_3}$
E_1	$\nexists$	saddle	saddle	unstable	unstable
E_2	$\nexists$	$\nexists$	$\nexists$	saddle	stable
E_3	stable	stable	saddle	saddle	saddle
E_4	$\nexists$	$\nexists$	stable	stable	$\nexists$

Table 3: Classification of equilibria for the case $k_1 = 0$ in terms of the total amount of calcium ions C, when $S > 1$ is valid.

Proof. Let us consider the four possible equilibrium points:

$$E_1 = (0,1),\ E_2 = (C-S-1,1),\ E_3 = (0,0),\ \text{and}\ E_4 = \left(\frac{k_4\,(C-S)}{k_3+k_4}, \frac{k_3\,(C-S)}{k_3+k_4}\right).$$

We linearize the system of equations (4.2) to analyze the stability of the equilibria, by using the Hartman–Grobman theorem. The Jacobi matrix of the system is

$$J(c,f_b) = \begin{pmatrix} -k_3\,(1-f_b) + k_2\,(C-S-2c-f_b) & k_4 + k_3 c - 2k_4 f_b - k_2 c \\ k_3\,(1-f_b) & -k_4 + 2k_4 f_b - k_3 c \end{pmatrix}. \tag{4.5}$$

We shall evaluate the Jacobi matrix at the four equilibrium points and determine the type of the equilibria by the signs of the eigenvalues of the matrix.

1. Equilibrium point $E_1 = (0,1)$.
   Let us first note that the point E_1 exists only for $C \geq 1$, see Proposition 5. Substituting the latter equilibrium point in (4.5), we derive:

   $$J(E_1) = \begin{pmatrix} k_2\,(C-S-1) & -k_4 \\ 0 & k_4 \end{pmatrix}.$$

   The eigenvalues of $J(E_1)$ are $\lambda_1 = k_2(C-S-1)$ and $\lambda_2 = k_4$. Then, obviously, E_1 is a saddle point if $C < S+1$ holds and an unstable node if $C > S+1$ is valid.

2. Equilibrium point $E_2 = (C - S - 1, 1)$.
We substitute E_2 in (4.5) and obtain

$$J(E_2) = \begin{pmatrix} -k_2\,(C - S - 1) & (C - S - 1)\,(k_3 - k_2) - k_4 \\ 0 & k_4 - k_3\,(C - S - 1) \end{pmatrix}.$$

The eigenvalues of the triangular matrix are $\lambda_1 = -k_2(C - S - 1) < 0$ (from the existence condition) and $\lambda_2 = k_4 - k_3(C - S - 1)$. Thus, the equilibrium point is a stable node when $k_4 < k_3(C - S - 1) \iff C > \dfrac{k_4}{k_3} + S + 1$ and is a saddle point when $S + 1 < C < \dfrac{k_4}{k_3} + S + 1$.

3. Equilibrium point $E_3 = (0, 0)$. We compute the determinant and trace of the Jacobi matrix:

$$J(E_3) = \begin{pmatrix} -k_3 + k_2\,(C - S) & k_4 \\ k_3 & -k_4 \end{pmatrix}$$

and obtain

$$det\ J(E_3) = -k_2 k_4\,(C - S)\,, \quad trace\ J(E_3) = -k_3 - k_4 + k_2(C - S).$$

The sign of the determinant in this case depends on the factor $C - S$, therefore, we shall consider the following two cases:

- $C - S > 0$.
  In this case, the determinant is negative and, therefore, E_3 is a saddle point.
- $C - S < 0$.
  In this case, the determinant is positive and the trace is negative. The equilibrium is, thus, asymptotically stable.

4. Equilibrium point $E_4 = \left(\dfrac{k_4(C - S)}{k_3 + k_4}, \dfrac{k_3(C - S)}{k_3 + k_4}\right)$.

$$J(E_4) = \begin{pmatrix} -k_3\left(1 - \dfrac{k_3(C - S)}{k_3 + k_4}\right) - \dfrac{k_2 k_4(C - S)}{k_3 + k_4} & k_4\left(1 - \dfrac{k_3(C - S)}{k_3 + k_4}\right) - \dfrac{k_2 k_4(C - S)}{k_3 + k_4} \\ k_3\left(1 - \dfrac{k_3(C - S)}{k_3 + k_4}\right) & -k_4\left(1 - \dfrac{k_3(C - S)}{k_3 + k_4}\right) \end{pmatrix}.$$

For the eigenvalues, after some computations, we obtain

$$\begin{aligned}
\lambda_1\lambda_2 &= det\ J(E_4)\\
&= \frac{k_2k_4\,(C-S)\,(k_4-k_3(C-S-1))}{k_3+k_4},\\
\lambda_1+\lambda_2 &= trace\ J(E_4)\\
&= \frac{-\,k_3(k_3+k_4-k_3(C-S))-k_2k_4(C-S)-k_4(k_3+k_4-k_3(C-S))}{k_3+k_4}\\
&= \frac{k_3k_4(-1-1+C-S)+k_3^2(-1+C-S)-k_4(k_2(C-S)+k_4)}{k_3+k_4}\\
&= \frac{k_3k_4\,(C-S-2)+k_3^2(C-S-1)-k_4\,(k_2(C-S)+k_4)}{k_3+k_4}.
\end{aligned}$$

In order for the equilibrium point to exist, using Proposition 5, we consider the case when $S < C < S + \dfrac{k_3+k_4}{k_3}$. In this case, the determinant is always positive, therefore, we have to determine the sign of the trace. Further, we shall give an upper bound for the expression of the trace:

$$\begin{aligned}
trace\ J(E_4) &= \frac{k_3k_4\,(C-S-2)+k_3^2(C-S-1)-k_4\,(k_2(C-S)+k_4)}{k_3+k_4}\\
&= \frac{k_3k_4(C-S-1)}{k_3+k_4}-\frac{k_3k_4}{k_3+k_4}+\frac{k_3^2(C-S-1)}{k_3+k_4}-\frac{k_2k_4(C-S)}{k_3+k_4}-\frac{k_4^2}{k_3+k_4}\\
&< \frac{k_3k_4^2}{k_3(k_3+k_4)}-\frac{k_3k_4}{k_3+k_4}+\frac{k_3^2k_4}{k_3(k_3+k_4)}-\frac{k_2k_4(C-S)}{k_3+k_4}-\frac{k_4^2}{k_3+k_4}\\
&= -\frac{k_2k_4(C-S)}{k_3+k_4}.
\end{aligned}$$

The latter expression is always negative for $C > S$—the case, which we are interested in. Therefore, the equilibrium is asymptotically stable. □

5. NUMERICAL EXAMPLES

5.1. LIMITING CASE $k_1 = const > 0, k_2 = 0$

In this section, we give example phase portraits for the three different cases, considered in the classification of the equilibria in Proposition 4. For the numerical experiments, we consider the model parameters, taken from Table 4:

$$k_1 = 9.6,\ k_3 = 65,\ k_4 = 45,$$

and $S = 2$. Let us note that the initial conditions for the system (4.1) must satisfy conditions (2.7).

Experiment 1. We consider the following parameter value: $C = 0.8$, which corresponds to the case $0 < C < 1$. Thus, as concluded in Proposition 4, in this case the point $E_1 = (C-1, 1)$ does not exist, while $E_2 = \left(\dfrac{Ck_4}{k_3+k_4}, \dfrac{Ck_3}{k_3+k_4}\right)$ is asymptotically stable. The numerical results are shown in Figure 5 and are in agreement with the analytical conclusions.

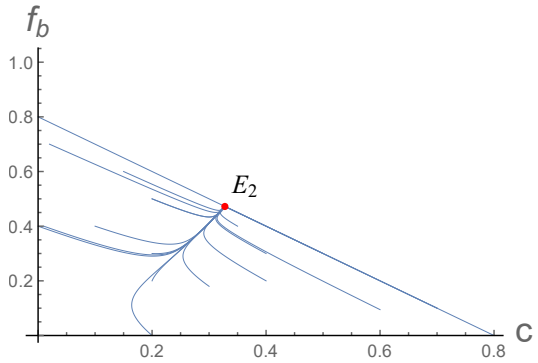


Figure 5: Phase portrait for the case $k_2 = 0$ with parameter value $C = 0.8$. E_1 does not exist, while E_2 is a stable equilibrium.

Experiment 2. We consider the parameter $C = 1.6$, which corresponds to the case $1 < C < \dfrac{k_3+k_4}{k_3}$. By Proposition 4, in this case the equilibrium point $E_1 = (C-1, 1)$ is a saddle point, while $E_2 = \left(\dfrac{Ck_4}{k_3+k_4}, \dfrac{Ck_3}{k_3+k_4}\right)$ is again asymptotically stable. The numerical results are in agreement with the conclusions in Proposition 4 and are depicted in Figure 6.

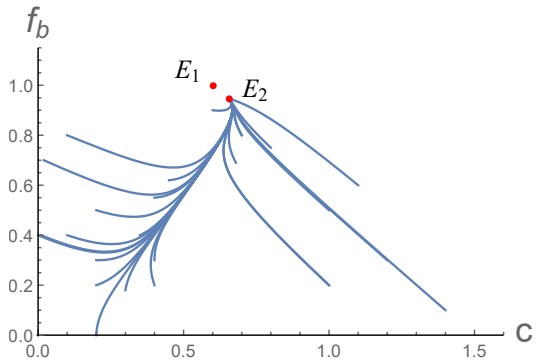


Figure 6: Phase portrait for the case $k_2 = 0$ with parameter value $C = 1.6$. E_1 is a saddle point, E_2 is a stable equilibrium.

Experiment 3. In this experiment, we consider the parameter $C = 2$, which corresponds to the case $C > \dfrac{k_3 + k_4}{k_3}$. Following Proposition 4, $E_1 = (C - 1, 1)$ is to be asymptotically stable, while $E_2 = \left(\dfrac{Ck_4}{k_3 + k_4}, \dfrac{Ck_3}{k_3 + k_4} \right)$ does not exist. The numerical results are shown in Figure 7. Again, the numerical experiments are in agreement with the analytic results.

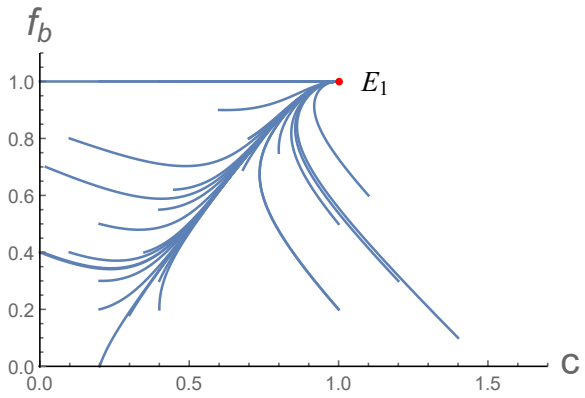


Figure 7: Phase portrait for the case $k_2 = 0$ with parameter value $C = 2$. E_1 is a stable equilibrium, E_2 does not exist.

Remark 2. By the correspoding results in Fig 5, 6, and 7, we can further suppose that the locally stable equilibrium points in each of the considered experiments are also globally asymptotically stable.

5.2. LIMITING CASE $k_1 = 0, k_2 = const > 0$

Here, we shall present several phase portraits, illustrating Proposition 6. For the numerical experiments, we consider the following values for the parameters, taken from Table 4:

$$k_2 = 5.9,\ k_3 = 65,\ k_4 = 45.$$

Let us note that the initial conditions for the system (4.2) must satisfy conditions (2.7).

Experiment 1. In this experiment, we consider the model parameters $C = 0.8$ and $S = 0.5$. Thus, we consider the case $0 < S < C < 1$. By Proposition 6, in this case $E_1 = (0, 1)$ and $E_2 = (C - S - 1, 1)$ do not exist, $E_3 = (0, 0)$ is a saddle point, and $E_4 = \left(\dfrac{k_4 (C - S)}{k_3 + k_4}, \dfrac{k_3 (C - S)}{k_3 + k_4} \right)$ is a stable equilibrium. The following is illustrated by the numerical results, depicted in Fig 8.

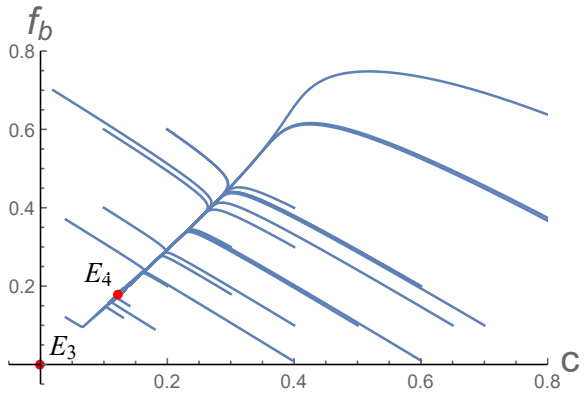


Figure 8: Phase portrait for the case $k_1 = 0$ with parameters $C = 0.8,\ S = 0.5$. E_1 and E_2 do not exist, while E_3 is a saddle and E_4 is a stable equilibrium.

Experiment 2. We consider the case $0 < C < 1 < S$, thus, we choose the model parameters $C = 0.8$ and $S = 4$. By Proposition 6, $E_1 = (0, 1)$, $E_2 = (C - S - 1, 1)$, $E_4 = \left(\dfrac{k_4 (C - S)}{k_3 + k_4}, \dfrac{k_3 (C - S)}{k_3 + k_4} \right)$ do not exist, while $E_3 = (0, 0)$ is a stable equilibrium. The obtained results, shown in Figure 9, illustrate the latter.

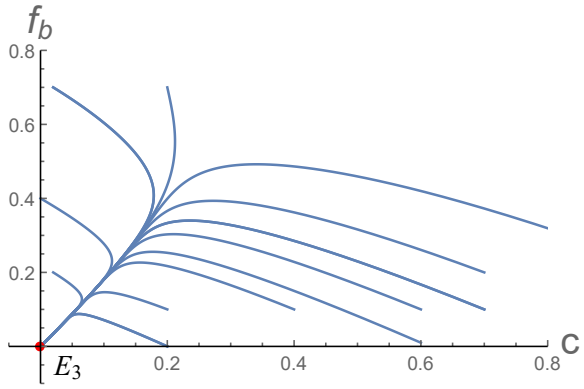


Figure 9: Phase portrait for the case $k_1 = 0$ with parameters $C = 0.8,\ S = 4$. In this case, E_1, E_2, and E_4 do not exist, while E_3 is a stable equilibrium.

Experiment 3. We shall consider model parameters $C = 4$, $S = 6$, thus, the case $1 < C < S$ holds. Following the statement of Proposition 6, equilibrium points $E_2 = (C - S - 1, 1)$ and $E_4 = \left(\dfrac{k_4 (C - S)}{k_3 + k_4}, \dfrac{k_3 (C - S)}{k_3 + k_4} \right)$ do not exist, while $E_1 = (0, 1)$ is a saddle point, and $E_3 = (0, 0)$ is an asymptotically stable equilibrium

point. The numerical results, which illustrate the statement of Proposition 6, are shown in Figure 10.

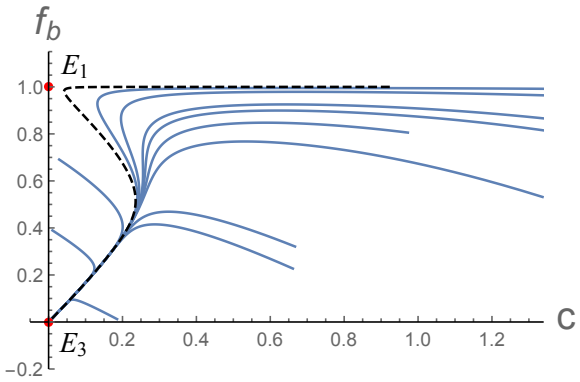


Figure 10: Phase portrait for the case $k_1 = 0$ with parameters $C = 4,\ S = 6$. In this case, E_2 and E_4 do not exist, while E_1 is a saddle and E_3 is an asymptotically stable equilibrium. Note: The dashed trajectory will be discussed further in the next section.

Experiment 4. In the following experiment, we consider the conditions $S < C < S + 1$ and choose the model parameters $C = 5.2$ and $S = 5$. Taking into account Proposition 6, in this case, E_1 and E_3 are saddle points, E_2 does not exist and E_4 is a stable equilibrium. The numerical results, illustrate the statement of the latter proposition, see Figure 11.

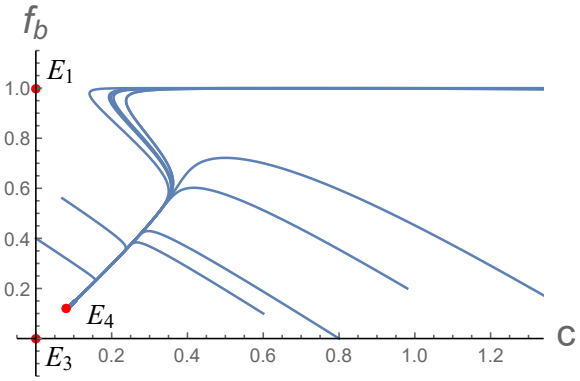


Figure 11: Phase portrait for the case $k_1 = 0$ with parameters $C = 5.2,\ S = 5$. In this case E_1 and E_3 are saddle points, E_2 does not exist, and E_4 is a stable equilibrium.

Experiment 5. For Experiment 5, we consider the case $S + 1 < C < S + \dfrac{k_3 + k_4}{k_3}$

and choose model parameters $C = 5.2$ and $S = 4$. Using Proposition 6, $E_1 = (0, 1)$ is an unstable equilibrium, $E_2 = (C - S - 1, 1)$ and $E_3 = (0, 0)$ are saddle points, while $E_4 = \left(\frac{k_4 (C - S)}{k_3 + k_4}, \frac{k_3 (C - S)}{k_3 + k_4} \right)$ is asymptotically stable. The numerical results in Figure 12 are in agreement with the analytic results.

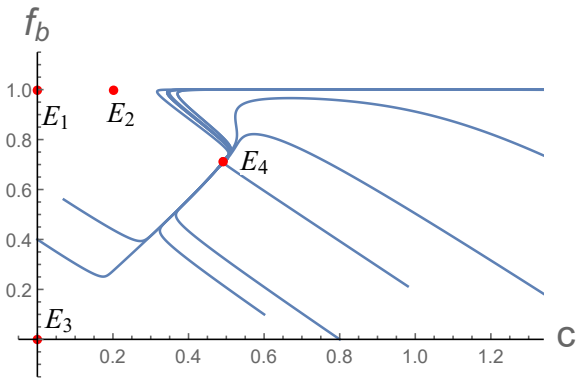


Figure 12: Phase portrait for the case $k_1 = 0$ with parameters $C = 5.2,\ S = 4$. In this case, E_1 is an unstable equilibrium, E_2 and E_3 are saddle points, and E_4 is an asymptotically stable equilibrium.

Experiment 6. Here, we shall consider the case $C > S + \frac{k_3 + k_4}{k_3}$ and choose model parameters $C = 7$, $S = 4$. By Proposition 6, E_1 is an unstable equilibrium, E_2 is asymptotically stable, E_3 is a saddle point, and E_4 does not exist. The numerical results in Figure 13 are in agreement with the analytic results.

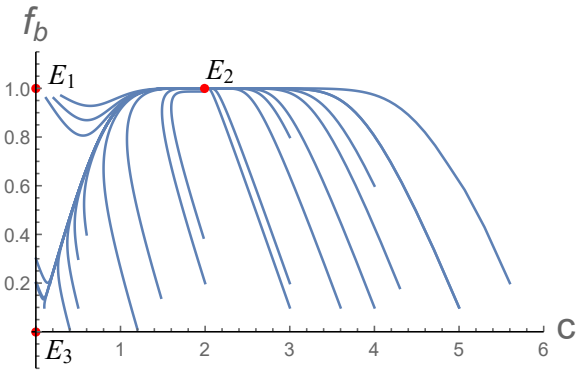


Figure 13: Phase portrait for the case $k_1 = 0$ with parameters $C = 7,\ S = 4$. In this case, E_1 is an unstable equilibrium, E_2 is a stable equilibrium, E_3 is a saddle point, and E_4 does not exist.

6. BIOLOGICAL IMPLICATIONS OF THE QUALITATIVE ANALYSIS

Based on the qualitative analysis of the model for the calcium dynamics in a muscle cell, we make the following observations:

- Case $k_1 = const > 0,\ k_2 = 0$.

  Let us first discuss the case when there is a stimulus, i.e., when $k_2 = 0$. For each choice of the parameters, depending on the ratio C between the total concentrations of calcium ions and CF sites, the biological system tends to a certain equilibrium.

  - Following Proposition 4, when $C < 1$ holds, i.e., when the total concentration of CF sites is more than the total concentration of calcium (or, stated otherwise, there is not enough calcium to fill the CF sites), the system always reaches the equilibrium point $E_2 = \left(\dfrac{Ck_4}{k_3+k_4}, \dfrac{Ck_3}{k_3+k_4}\right)$.
  - However, even in the case when there are sufficient calcium ions, depending on the ratio $\dfrac{k_4}{k_3}$ between the rates of binding and release from the CF sites, the system might also stabilize at this point. This is the case, when $C < 1 + \dfrac{k_4}{k_3}$, or equivalently $\dfrac{k_4}{k_3} > C - 1$, thus, the rate of binding of calcium ions to the CF is relatively small, compared to the rate of release;
  - Vice versa, if $\dfrac{k_4}{k_3} < C - 1$, then calcium ions eventually bind to all CF sites, which corresponds to the stable equilibrium $E_1 = (C-1, 1)$, where $f_b = 1$.

  Let us further note that the equilibrium state of the system does not depend on the rate of release from the SR sites k_1. Therefore, the asymptotic behaviour of the system does not depend on the strength of the incoming signal. However, it determines the rate at which the biological system tends to the equilibrium point. For the sake of example, numerical results for the concentration of free calcium ions, obtained for two different values of k_1, are shown in Figure 14.

- Case $k_1 = 0,\ k_2 = const > 0$.

  Here, we shall discuss from a biological point of view the qualitative results for the case, when there is no stimulus present in the muscle cell, i.e., when $k_1 = 0$.

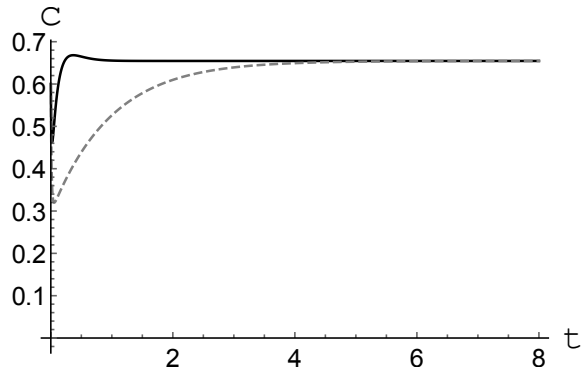


Figure 14: Concentration of free calcium ions c in time. Results for $k_1 = 1$ are depicted with dashed line, for $k_1 = 9.6$—with solid line.

- Following Proposition 6, if $0 < C < S$ holds true, which biologically means that the total concentration of calcium ions is less than the total concentration of SR sites, then the system reaches the equilibrium state $c = 0$, $f_b = 0$. The latter means that all calcium ions get bound to the SR, thus, the muscle cell is relaxed. Let us emphasize that the case $0 < C < S$ is the natural one for the process, since the free calcium ions were originally released from the SR.
- If, however, the total concentration C is higher than S, then different equilibrium points are reached.

We have discussed in this section the two limiting cases when k_1 and k_2 are held constant, one of them 0. Of course, in reality the process is characterized with consecutive changes in their values. Therefore, the results, presented here, will give us information for the two separate parts of the process—when the stimulus is on and off.

Let us further consider one numerical result to illustrate the process of calcium dynamics, described by model equations (2.6). Here, for model parameters we shall use values from [6], systematized in Table 4.

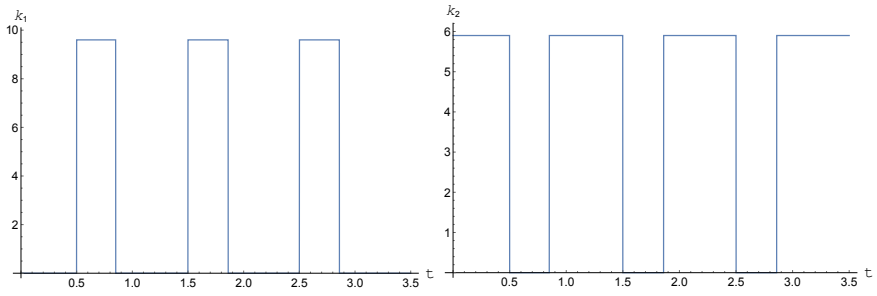
Further, we define a square wave stimulus by introducing the piecewise constant functions k_1 and k_2 in the following way:

$$k_1 = \begin{cases} k_{10}, & \text{stimulus is on,} \\ 0, & \text{stimulus is off,} \end{cases} \qquad k_2 = \begin{cases} 0, & \text{stimulus is on,} \\ k_{20}, & \text{stimulus is off.} \end{cases}$$

For our numerical experiment, we consider the particular choice of k_1 and k_2, depicted in Figure 15.

Parameter	Value	Parameter	Value
C	2	μ_s	600 mN/mm
S	6	l_{s0}	0.234 mm
k_{10}	9.6 s^{-1}	l_{c0}	2.6 mm
k_{20}	5.9 s^{-1}	a	-2.23 mm^{-2}
k_3	65 s^{-1}	α_{max}	1.8
k_4	45 s^{-1}	α_m	0.4 s/mm
L	2.7 mm	α_p	1.33 s/mm
P_0	60.86 mN/mm^2		

Table 4: Model parameters for (2.6), taken from [6].



(a) $k_1(t)$ (b) $k_2(t)$

Figure 15: Graphs of coefficients k_1 and k_2.

The numerical solutions for the concentrations c and f_b, using fourth-order Runge–Kutta method [2] with time discretization step 10^{-3} are shown in Figure 16.

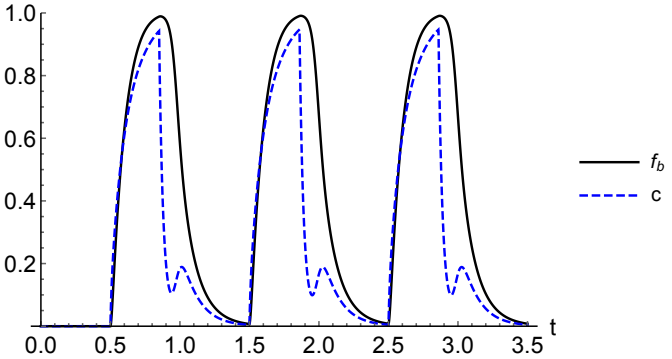


Figure 16: Modelling of calcium dynamics—concentration of free calcium ions (dashed line), concentration of filament-bound calcium sites (solid line).

To explain the numerical results, let us consider the two distinct situations in the process—when the stimulus is on and off.

- Presence of stimulus

  Let us first note that in the case, when $k_2 = 0$, this choice of parameters corresponds to the case of an asymptotically stable point $E_1 = (C - 1, 1)$ in Proposition 4. Thus, for $C = 2$ and $S = 6$, the solution would "try to reach" the corresponding equilibrium point $E_1 = (1, 1)$. The latter is clearly seen from the numerical experiments in Figure 16.

- Absence of stimulus

  In the case, when the stimulus is off, or equivalently, when $k_1 = 0$, by the qualitative analysis, summarized in Proposition 6, there exist the saddle equilibrium point $E_1 = (0, 1)$ and the asymptotically stable $E_3 = (0, 0)$. The latter explains the peculiar behaviour of the solution for c, that is observed, e.g., around $t = 1$. In particular, let us consider the dashed trajectory in Figure 10, which is obtained for an initial condition corresponding to the peak of the graphs in Figure 6. When close to the saddle point, the trajectory is repelled with a change in the sign of the derivative for the concentration c, which results in a rise of the solution for c, followed by a decrease to the equilibrium $c = 0$.

7. CONCLUSION

In this paper, we have considered a mathematical model, described in terms of ordinary differential equations, for the process of calcium dynamics inside the muscle cell. We have obtained results for the qualitative behaviour of the model solutions in the two limiting cases $k_1 = 0$ and $k_2 = 0$ that to the best of our knowledge are not known in the scientific literature. On one hand, such kind of qualitative information is useful in the mathematical modelling of biological processes and it helps to better understand the dynamical properties of the mathematical model. On the other hand, it gives valuable information about the influence of the different model parameters. The latter is particularly interesting, when considering the process in different conditions, e.g., when there are certain deceases present, which affect the normal calcium activity inside the muscle cell.

ACKNOWLEDGEMENTS. The work of the authors was partially funded under contract 80-10-17/09.04.2019 with Science Fund, Sofia University.

8. REFERENCES

[1] Betts, J., Desaix, P., et.al.: Anatomy and physiology, OpenStax, 2013. https://opentextbc.ca/anatomyandphysiology

[2] Butcher, J.: *Numerical methods for ordinary differential equations*. Wiley, 2009.

[3] Frontera, W., Ochala, J.: Skeletal muscle: a brief review of structure and function. *Calcif. Tissue Int.*, **96**, 2015, 183–195. https://doi.org/10.1007/s00223-014-9915-y

[4] Hale, J.: *Ordinary differential equations*. Dover Books, 2009.

[5] Meredith, T.: A mathematical model of the neuromuscular junction and muscle force generation in the pathological condition myasthenie gravis. https://pdfs.semanticscholar.org/a97b/79416876e52dd357ca5f59f52bbfab23b1e7.pdf

[6] McMillen, T., Williams, T., Holmes, P.: Nonlinear muscles, passive viscoelasticity and body taper conspire to create neuromuscular phase legs in anguilliform swimmers. *PLoS. Comput. Biol.*, **4**, 2008, e10000157. https://doi.org/10.1371/journal.pcbi.1000157

[7] Williams, T., Bowtell, G., Curtin, N.: Predicting force generation by lamprey muscle during applied sinusoidal movement using a simple dynamic model. *J. Exp. Biol.*, **201**, 1998, 869–875. https://jeb.biologists.org/content/201/6/869.article-info

[8] Sliding filament theory, Wikipedia, Retrieved on December 17, 2019. https://en.wikipedia.org/wiki/Sliding_filament_theory

[9] Muscle contraction, Wikipedia, Retrieved on December 17, 2019. https://en.wikipedia.org/wiki/Muscle_contraction

Received on February 6, 2020
Received in a revised form on March 23, 2020

Zdravka Nedyalkova, Tihomir Ivanov

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 James Bourchier Blvd.
BG-1164 Sofia
BULGARIA

E-mails: znedyalkova@fmi.uni-sofia.bg
tbivanov@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 106

ANNUAL OF SOFIA UNIVERSITY „ST. KLIMENT OHRIDSKI“
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 106

SMOOTHEST INTERPOLATION WITH BOUNDARY CONDITIONS IN $W_2^3[a,b]$

VELINA IVANOVA, RUMEN ULUCHEV

We study the problem on the smoothest interpolant with boundary conditions in the Sobolev space $W_2^3[a,b]$. Characterization and uniqueness of the best interpolant with free knots of interpolation, satisfying boundary conditions, are proved. Based on our proofs we present an algorithm for finding the unique oscillating spline interpolant. Numerical results are given.

Keywords: Smoothest interpolation, splines, Birkhoff interpolation.
2010 Math. Subject Classification: 65D05, 65D07.

1. INTRODUCTION

Let $[a,b]$ be a closed finite subinterval of the real line, r be a natural number, and $1 < p < \infty$. As usual, by $W_p^r[a,b]$ we denote the Sobolev space

$$W_p^r[a,b] = \big\{ f \,:\, f^{(r-1)} \text{ is abs. continuous in } [a,b],\ f^{(r)} \in L_p[a,b] \big\},$$

and by $\|\cdot\|_p$ the norm in $L_p[a,b]$,

$$\|g\|_p = \left(\int_a^b |g(t)|^p \, dt \right)^{1/p}, \qquad g \in L_p[a,b].$$

Suppose that we are given real numbers $\mathbf{y} = (y_0, y_1, \ldots, y_{N+1})$. We shall use the notation $\mathbf{x} = (x_0, x_1, \ldots, x_{N+1})$ for the elements of the set

$$X_N := \big\{(x_0, x_1, \ldots, x_{N+1}) \in \mathbb{R}^{N+2} \,:\, a = x_0 < x_1 < \cdots < x_{N+1} = b\big\}.$$

In 1988, Pinkus [12] considered the problem on existence, characterization, and uniqueness of knots $\mathbf{x}^* \in X_N$ and a function $f^* \in W_p^r[a,b]$ for which the following quantity is attained:

$$\inf_{\mathbf{x}\in X_N} \inf_{f\in W_p^r[a,b]} \big\{\|f^{(r)}\|_p : f(x_i) = y_i,\ i = 0, \ldots, N+1\big\}. \tag{1.1}$$

That is, we seek for the *smoothest* interpolant in $W_p^r[a,b]$ with free interpolation knots in $[a,b]$. The paper of Pinkus [12] may be regarded as a further development of de Boor's study [6] on the "best" interpolant with fixed interpolation knots.

Henceforth we assume that the data $\mathbf{y} = (y_0, y_1, \ldots, y_{N+1})$ satisfy the inequalities

$$(y_i - y_{i-1})(y_{i+1} - y_i) < 0, \qquad i = 1, \ldots, N. \tag{1.2}$$

Note that conditions (1.2) are not essential restrictions. Indeed, if $y_{i-1} < y_i < y_{i+1}$ or $y_{i-1} > y_i > y_{i+1}$ for some i and f takes values y_{i-1}, y_{i+1} at knots $x_{i-1} < x_{i+1}$, respectively, then by the continuity of the functions from $W_p^r[a,b]$, f takes the intermediate value y_i at some point $x_i \in (x_{i-1}, x_{i+1})$. It means that if there exists a solution to (1.1) in the case of oscillating data, we easily obtain a solution when the data $\mathbf{y}$ do not oscillate by taking the maximal subsequence of values in $\mathbf{y}$ satisfying (1.2). We also assume that

$$N + 2 > r, \tag{1.3}$$

for otherwise a trivial solution to (1.1) is given by the Lagrange interpolation polynomial of degree $r-1$ with arbitrary knots from the set X_N.

Taking into account the above remarks we henceforth assume that r, N, and the data $\mathbf{y}$ satisfy (1.2) and (1.3).

We give below a brief account on the results on problem (1.1).

The case $r = 1$ is elementary (see [12]).

In 1984 Marin [10] completely solved (1.1) for $r = p = 2$. He first characterized the solution $(\mathbf{x}^*, f^*)$ as follows:

$$f^* \text{ is strictly monotone in } [x_i^*, x_{i+1}^*], \quad i = 0, \ldots, N, \tag{1.4}$$

and explicitly found the optimal knots $\mathbf{x}^*$ and the interpolant f^*. The extremal function is actually the unique interpolating natural cubic spline with knots $\mathbf{x}^*$ satisfying (1.4).

For $p \in (1, \infty)$, Pinkus [12] proved the existence and characterization of the solution to (1.1) for all r, but the uniqueness for $p = 1$ and $r = 2$ only. The following

result is a keystone in the survey on the smoothest interpolation, where as usual $f[x_i, \dots, x_{i+r}]$ is the divided difference of the function f at knots $x_i, \dots, x_{i+r}$.

Theorem A (Pinkus [12]) *Let $1 < p < \infty$, $\mathbf{y} = (y_0, y_1, \dots, y_{N+1})$ be real numbers satisfying* (1.2) *and* (1.3)*, and let f^* be a solution of* (1.1)*. There exist $a = x_0^* < \cdots < x_{N+1}^* = b$, such that $f^*(x_i^*) = y_i$, $i = 0, \dots, N+1$. Furthermore,*

(a)

$$f^{*(r)}(t) = \left| \sum_{i=0}^{N+1-r} \eta_i B_i(t) \right|^{q-1} \operatorname{sign} \left(\sum_{i=0}^{N+1-r} \eta_i B_i(t) \right),$$

where $\dfrac{1}{p} + \dfrac{1}{q} = 1$, $B_i(t)$ is the B-spline of degree $r-1$ with knots $x_i^*, \dots, x_{i+r}^*$, and the coefficients $\{\eta_i\}_{i=0}^{N+1-r}$ satisfy

$$\int_a^b B_i(t) f^{*(r)}(t)\, dt = f[x_i^*, \dots, x_{i+r}^*], \qquad i = 0, \dots, N+1-r;$$

(b) f^* is strictly monotone in $[x_i^*, x_{i+1}^*]$, $i = 0, \dots, N$.

The uniqueness of the smoothest interpolant in general was conjectured but it is still an open problem. There are a few particular cases where it was proved, e.g., for $p = 2$ and $r = 2$ by Marin [10], for $p = 2$ and $r = 3$ by Uluchev [20], for $p \in (1, \infty)$ and $r = 2$ by Rademacher and Scherer [14] and independently by Uluchev [20]. Based on key results of Bojanov [1] concerning interpolation by perfect splines, Pinkus [12] proved the uniqueness of the smoothest interpolant, which is actually a perfect spline, for the case $p = \infty$ and $r \in \mathbb{N}$. In 1995, Naidenov [11] proposed an algorithm for construction of the unique smoothest perfect spline. The case $p = 1$ was studied by Pinkus [12].

Various modifications of the problem have been studied by Bojanov [4], Draganova in [7] for the periodic case and on interpolation in mean values in [8]. Multidimensional aspects of the problem (1.1) have been considered by Marin [10], Rademacher and Scherer [14], Scherer and Smith [16], Scherer [15].

A short summary of the results on the topic was presented by Pinkus in [13].

Here we study a problem on the smoothest interpolant with free knots in the space $W_2^3[a, b]$ with additional boundary conditions imposed on the interpolant. The paper is organized as follows. We state our main results in Section 2. Section 3 consists of preliminaries on Birkhoff interpolation and B-splines with Bitkhoff type of knots. In Section 4 we study an extremal problem for interpolation at fixed knots with functions from $W_2^3[a, b]$ satisfying boundary conditions. Then we give characterization of the smoothest interpolant for our problem with free interpolation knots. Applying a constructive approach used in [20] by the second author we prove that there exists a unique fifth degree oscillating spline interpolant in Section 5. A direct consequence is the uniqueness of the smoothest interpolant with boundary conditions. In the final Section 6 we suggest a numerical algorithm for

finding the oscillating spline interpolant. We conclude this section with results of numerical experiment for a given data.

2. MAIN RESULTS

Suppose that $[a,b] \subset \mathbb{R}$, $r \in \mathbb{N}$, and $\mathbf{y} = (y_0, \dots, y_{N+1})$ are arbitrary real numbers. For a fixed $\mathbf{x} = (x_0, \dots, x_{N+1}) \in X_N$, we denote by $F(\mathbf{x}, \mathbf{y})$ the set of all functions $f \in W_2^3[a,b]$, such that

$$f(x_i) = y_i, \qquad i = 0, \dots, N+1, \tag{2.1}$$

$$f'(x_0) = 0, \quad f'''(x_0) = 0, \quad f'(x_{N+1}) = 0, \quad f'''(x_{N+1}) = 0. \tag{2.2}$$

In addition to the usual interpolation conditions we impose boundary conditions for the first and third derivative of the function at the endpoints $a = x_0$ and $b = x_{N+1}$. At first glance it seems that conditions (2.2) are very restrictive. Note that in the case of smoothest interpolation in $W_2^3[a,b]$ satisfying (2.1) only, the extremal function is a natural fifth degree spline whose third and fourth derivatives a priori vanish at the endpoints of the interval $[a,b]$, see [20]. Henceforth, $S_m(x_1, \dots, x_N)$ will stand for the space of spline functions of degree m with knots $x_1, \dots, x_N$.

Here we study the problem

$$\inf_{\mathbf{x} \in X_N} \inf_{f \in F(\mathbf{x}, \mathbf{y})} \|f'''\|_2. \tag{2.3}$$

The following result answers some questions concerning (2.3), including a characterization of the smoothest interpolant.

Theorem 1. *Let $\mathbf{y} = (y_0, \dots, y_{N+1})$, $N > 1$, be real numbers satisfying conditions* (1.2) *and let f^* be a solution to problem* (2.3)*. Then, there exist knots $\mathbf{x}^* = (x_0^*, \dots, x_{N+1}^*) \in X_N$ such that $f^* \in F(\mathbf{x}^*, \mathbf{y})$. Furthermore,*

(a) $f^* \in S_5(x_1^*, \dots, x_N^*)$;

(b) f^* is strictly monotone in $[x_i^*, x_{i+1}^*]$, for all $i = 0, \dots, N$.

Therefore, the smoothest interpolant with free knots is strictly monotone in each interval between two consecutive knots, thus its first derivative must vanish at the interior knots.

In Section 5 we show that there exists a unique fifth degree spline interpolant with knots in X_N satisfying the above characterization conditions for the smoothest interpolant to the problem (2.3). More precisely, we prove:

Theorem 2. *Let $N > 1$ and the real numbers $\mathbf{y} = (y_0, y_1, \dots, y_{N+1})$ oscillate in the sense that $(y_i - y_{i-1})(y_{i+1} - y_i) < 0$, $i = 1, \dots, N$. Then, there exists unique spline $s^* \in S_5(x_1^*, \dots, x_N^*)$ and knots $\mathbf{x}^* = (x_0^*, \dots, x_{N+1}^*) \in X_N$, such that*

$$\begin{aligned} &s^*(x_i^*) = y_i, && i = 0, \dots, N+1, \\ &s^{*\prime}(x_i^*) = 0, && i = 0, \dots, N+1, \\ &s^{*\prime\prime\prime}(x_0^*) = 0, && s^{*\prime\prime\prime}(x_{N+1}^*) = 0. \end{aligned} \tag{2.4}$$

A direct consequence of Theorem 1 and Theorem 2 is the next statement.

Theorem 3. *Let $N > 1$ and $\mathbf{y} = (y_0, \dots, y_{N+1})$ be real numbers satisfying inequalities* (1.2)*. If $(f^*, \mathbf{x}^*)$, $\mathbf{x}^* = (x_0^*, \dots, x_{N+1}^*) \in X_N$ is a solution to problem* (2.3)*, then f^* is the unique spline interpolant from the set $S_5(x_1^*, \dots, x_N^*) \cap F(\mathbf{x}^*, \mathbf{y})$ strictly oscillating at the knots $\mathbf{x}^*$.*

3. PRELIMINARIES ON BIRKHOFF INTERPOLATION

We need some basic definitions and results concerning Birkhoff interpolation and B-splines with Birkhoff type of knots, see for details [3, 5, 9]. Let $\mathbf{t} = (t_1, \dots, t_m)$, $t_1 < \dots < t_m$,

$$E = \begin{pmatrix} e_{10} & \dots & e_{1,r-1} \\ \dots & \dots & \dots \\ e_{m0} & \dots & e_{m,r-1} \end{pmatrix}$$

be an *incidence matrix* (E consists of 0's and 1's only), and $|E|$ be the total number of 1-entries in E. By π_r we denote the set of algebraic polynomials with real coefficients of degree at most r.

The incidence matrix E satisfies *Strong Pólya condition*, if $\sum_{j \le k} \sum_i e_{ij} > k+1$ for all $k = 0, \dots, r-2$.

A sequence of 1-entries $e_{ij}, \dots, e_{i,j+\ell-1}$ in i-th row of the matrix E is said to be *supported odd block* if ℓ is an odd number and there exist i_1, i_2, j_1, j_2, such that

$$e_{i_1 j_1} = e_{i_2 j_2} = 1, \qquad i_1 < i < i_2, \qquad j_1 < j, \qquad j_2 < j.$$

The matrix E is *conservative* if it does not contain supported odd blocks of 1's. The pair $(\mathbf{t}, E)$ is *s-regular*, if E is conservative and satisfies Strong Pólya condition.

Based on Birkhoff interpolation by polynomials, B-splines with Birkhoff type of knots were introduced preserving most important properties of the usual B-splines with simple (or multiple) knots (see [3]). Namely, for points $\mathbf{t} = (t_1, \dots, t_m)$, $t_1 < \dots < t_m$, and an s-regular incidence matrix E with $|E| = r + 1$, the B-spline of degree $r - 1$ with Birkhoff knots $(\mathbf{t}, E)$ is defined by

$$B\big((\mathbf{t}, E); t\big) = \frac{1}{(r-1)!} D\big[(\mathbf{t}, E); (\cdot - t)_+^{r-1}\big],$$

where $D[(\mathbf{t}, E); f]$ is the divided difference of the function f at $(\mathbf{t}, E)$, i.e. the coefficient of t^r in the polynomial $p(t) \in \pi_r$ which interpolates f at $(\mathbf{t}, E)$ in the sense

$$p^{(j)}(t_i) = f^{(j)}(t_i), \qquad e_{ij} = 1.$$

Given $r, N \in \mathbb{N}$ and a pair $(\mathbf{t}, E)$, $t_1 < \cdots < t_m$, $E = \{e_{ij}\}_{i=1,j=0}^{m\ r-1}$, $|E| = r+N$, we define a $(r+1)$-*partition of* $(\mathbf{t}, E)$ as a sequence of pairs $\{(\mathbf{t}_i, E_i)\}_{i=1}^{N}$, obtained in the following way. Let us order the elements of E in the manner

$$e_{10}, \ldots, e_{1,r-1}, e_{20}, \ldots, e_{2,r-1}, \ldots, e_{m0}, \ldots, e_{m,r-1}$$

and enumerate the 1's in the latter sequence from 1 to $r+N$. Let $\mathbf{e}_p, \mathbf{e}_{p+1}, \ldots, \mathbf{e}_q$ be the rows of E containing $r+1$ consecutive 1's starting with the i-th one. Suppose that the number of 1's of this $(r+1)$-sample in the row $\mathbf{e}_p$ is μ and the number of 1's in the sample in the row $\mathbf{e}_q$ is ν. We denote by $\mathbf{t}_i$ the set of knots $t_p < \cdots < t_q$ and by E_i the matrix composed from $\mathbf{e}_p, \ldots, \mathbf{e}_q$ in which all 1's in the first (resp., last) row of E_i except the first μ (resp., ν) ones are replaced by 0's.

It is said that the $(r+1)$-partition $\{(\mathbf{t}_i, E_i)\}_{i=1}^{N}$ of $(\mathbf{t}, E)$ is *s-regular* if all pairs $(\mathbf{t}_i, E_i)$, $i = 1, \ldots, N$ are s-regular.

In our study we need conditions for solvability of the Birkhoff interpolation problem by splines. The following general necessary and sufficient condition is due to Borislav Bojanov.

Theorem B (Bojanov [3], [5, Theorem 4.20]) *Let* $\mathbf{x} = (x_0, \ldots, x_{m+1})$, $a = x_0 < \cdots < x_{m+1} = b$, $E = \{e_{ij}\}_{i=0,\,j=0}^{m+1\ r-1}$ *and integers* $\{\nu_i\}_{i=1}^{n}$ *be given such that* $N = \nu_1 + \cdots + \nu_n$, $1 \le \nu_i \le r$, $i = 1, \ldots, n$, *and* $|E| = N + r$. *Assume that* $(\mathbf{x}, E)$ *has an s-regular* $(r+1)$*-partition* $\{(\mathbf{x}_i, E_i)\}_{i=1}^{N}$. *Then the interpolation problem*

$$s^{(j)}(x_i) = f_{ij}, \qquad e_{ij} = 1$$

by splines s *of degree* $r-1$ *with knots* $\xi_1, \ldots, \xi_n$ *of multiplicities* $\nu_1, \ldots, \nu_n$, *respectively, has a unique solution for each given data* $\{f_{ij}\}$ *if and only if*

$$B\big((\mathbf{x}_i, E_i); \theta_i\big) \neq 0, \qquad i = 1, \ldots, N,$$

where $(\theta_1, \ldots, \theta_N) = \big((\xi_1, \nu_1), \ldots, (\xi_n, \nu_n)\big)$.

4. PROOF OF THE CHARACTERIZATION THEOREM

Using the notations from Section 2, let $\mathbf{y} = (y_0, \ldots, y_{N+1})$ be arbitrary real numbers, $\mathbf{x} = (x_0, \ldots, x_{N+1}) \in X_N$ and $F(\mathbf{x}, \mathbf{y})$ be the set of all functions $f \in W_2^3[a, b]$ satisfying (2.1) and (2.2).

Lemma 1. *There exists a unique spline function* $s \in S_5(x_1, \ldots, x_N)$ *satisfying the interpolation conditions* (2.1) *and* (2.2).

Proof. The assertion follows immediately from Theorem B setting $r = 6$, $m = N$, $\nu_1 = \cdots = \nu_n = 1$, $n = N$, $\theta_i = \xi_i = x_i$, $i = 1, \ldots, N$. Indeed, the

$(N+2)\times r$ incidence matrix

$$E=\begin{pmatrix}1&1&0&1&0&0\\1&0&0&0&0&0\\\dots&\dots&\dots&\dots&\dots&\dots\\1&0&0&0&0&0\\1&1&0&1&0&0\end{pmatrix},\qquad |E|=N+6,$$

of the Birkhoff interpolation problem (2.1)–(2.2) has an s-regular $(r+1)$-partition $\big\{(\mathbf{t}_i,E_i)\big\}_{i=1}^{N}$. Obviously, $\theta_i=x_i\in\operatorname{supp}B\big((\mathbf{t}_i,E_i);t\big)$, $i=1,\dots,N$. Then Theorem B yields that there exists a unique spline s of degree $r-1=5$ with knots $(\xi_1\dots,\xi_N)=(x_1\dots,x_N)$ satisfying (2.1) and (2.2), i.e. $s\in S_5(x_1,\dots,x_N)$. □

Remark 1. Note that the spline s in Lemma 1 is a function from the class $F(\mathbf{x},\mathbf{y})$.

The following is a modified version of the classical Holladay's theorem.

Lemma 2. *Let s be the unique spline in the space $S_5(x_1,\dots,x_N)$ satisfying the interpolation conditions* (2.1) *and* (2.2)*. Then, for each function $f\in F(\mathbf{x},\mathbf{y})$,*

$$\|s'''\|_2\le\|f'''\|_2.$$

The equality holds if and only if $f=s$ in $[a,b]$.

Proof. We follow the standard line taking into account that both f and s satisfy the interpolation conditions (2.1) and (2.2), $x_0=a$, $x_{N+1}=b$, and $s^{V}(t)\big|_{(x_i,x_{i+1})}=c_i=const.$, $i=0,\dots,N$:

$$\begin{aligned}\int_a^b s'''(t)\big(f'''(t)-s'''(t)\big)\,dt&=\int_a^b s'''(t)\,d\big(f''(t)-s''(t)\big)\\&=s'''(t)\big(f''(t)-s''(t)\big)\Big|_a^b-\int_a^b s^{IV}(t)\big(f''(t)-s''(t)\big)\,dt\\&=-\int_a^b s^{IV}(t)\,d\big(f'(t)-s'(t)\big)\\&=-s^{IV}(t)\big(f'(t)-s'(t)\big)\Big|_a^b+\int_a^b s^{V}(t)\big(f'(t)-s'(t)\big)\,dt\\&=\int_a^b s^{V}(t)\,d\big(f(t)-s(t)\big)=\sum_{i=0}^{N}\int_{x_i}^{x_{i+1}} s^{V}(t)\,d\big(f(t)-s(t)\big)\\&=\sum_{i=0}^{N}\int_{x_i}^{x_{i+1}} c_i\,d\big(f(t)-s(t)\big)=\sum_{i=0}^{N}c_i\big(f(t)-s(t)\big)\Big|_{x_i}^{x_{i+1}}\\&=0.\end{aligned}$$

Then

$$\begin{aligned}\int_a^b \big(s'''(t)\big)^2\,dt &\le \int_a^b \big[\big(f'''(t)-s'''(t)\big)^2+\big(s'''(t)\big)^2\big]\,dt\\ &= \int_a^b \big[\big(f'''(t)-s'''(t)\big)+s'''(t)\big]^2\,dt\\ &= \int_a^b \big(f'''(t)\big)^2\,dt,\end{aligned}$$

i.e.

$$\|s'''\|_2 \le \|f'''\|_2,$$

where the equality holds if and only if $f'''(t)-s'''(t)=0$ in $[a,b]$. The last identity yields $f-s\in\pi_2$. Since $f,s\in F(\mathbf{x},\mathbf{y})$, the quadratic polynomial $f-s$ vanishes at the endpoins of $[a,b]$ and at least in one interior knot, hence $f-s=0$ in $[a,b]$. The proof of the lemma is complete. □

Remark 2. Lemma 2 claims that the only function for which

$$\inf_{f\in F(\mathbf{x},\mathbf{y})} \|f'''\|_2$$

is attained is the unique spline interpolant s from Lemma 1.

Remark 3. A general result on the existence, characterization and uniqueness of a function $f\in W_p^r[a,b]$ satisfying Birkhoff type interpolation conditions with minimal L_p norm of $f^{(r)}$ for fixed knots was proved by Bojanov [2]. However our case does not fall in the scope of Theorem 1 in [2].

Lemma 3. *Let $s\in S_5(x_1,\dots,x_N)$ be the unique spline satisfying* (2.1) *and* (2.2). *If the data $\mathbf{y}=(y_0,\dots,y_{N+1})$ satisfies condition* (1.2) *and $N>1$, then*

(a) s''' has exactly $N+2$ simple zeros in $[a,b]$;

(b) s' has exactly N simple zeros in (a,b).

Proof. (a) Since the data oscillates, s has at least N local extrema in (a,b). Then, the derivative s' has at least N zeros in (a,b). The interpolation conditions (2.2) give two additional zeros at the endpoints of the interval $[a,b]$ which means that s' has totally at least $N+2$ non-coinciding zeros in $[a,b]$. Applying Rolle's theorem for s', it follows that the second derivative s'' has at least $N+1$ non-coinciding zeros in (a,b), which give N zeros of s''' in (a,b). Because of (2.2), s''' has two more zeros at the endpoints of $[a,b]$. Therefore, s''' has at least $N+2$ zeros in $[a,b]$.

Observe that s''' is a spline function from the space $S_2(x_1,\dots,x_N)$. A well-known result (see [17, Theorem 4.53]) says that any spline from $S_2(x_1,\dots,x_N)$ has no more than $N+2$ zeros counting multiplicities, i.e. s''' has at most $N+2$ zeros in $[a,b]$.

So, we conclude that s''' has exactly $N+2$ simple zeros in $[a,b]$.

(b) From the proof of (a) it follows that s' has exactly N simple zeros in (a,b). Otherwise Rolle's theorem would give more than $N+2$ zeros for s''' in $[a,b]$, a contradiction. □

Proof of Theorem 1. (a) Let f^* solve the extremal problem (2.3). Therefore there exist $\mathbf{x}^* \in X_N$, such that $f^* \in F(\mathbf{x}^*, \mathbf{y})$. Since f^* solves (2.3), then f^* must solve the extremal problem for fixed knots at $\mathbf{x}^*$, namely

$$\inf_{f \in F(\mathbf{x}^*, \mathbf{y})} \|f'''\|_2.$$

By Lemma 1 and Lemma 2 it follows that f^* is the unique spline in $S_5(x_1^*, \dots, x_N^*)$, satisfying the interpolation conditions (2.1) and (2.2).

(b) From Lemma 3, we obtain that $f^{*\prime}$ has exactly N simple zeros in (a,b) which are the extremal points of f^* as well. Denote by $a < \eta_1 < \cdots < \eta_N < b$ all the extremal points of f^* in (a,b) and set $\eta_0 = a$, $\eta_{N+1} = b$. It is clear that the function f^* is strictly monotone in each interval $[\eta_i, \eta_{i+1}]$, $i = 0, \dots, N$. We remark that due to the oscillation of the data $\mathbf{y}$ we have $\eta_i \in (x_{i-1}^*, x_{i+1}^*)$, $i = 1, \dots, N$.

We will show that $\eta_i = x_i^*$ for all $i = 0, \dots, N+1$. Let us assume to the contrary that $\eta_j \neq x_j^*$ for some j. We set $z_i = f^*(\eta_i)$, $i = 0, \dots, N+1$ and consider the extremal problem

$$\inf_{f \in F(\boldsymbol{\eta}, \mathbf{z})} \|f'''\|_2$$

for fixed interpolation knots $\boldsymbol{\eta} = (\eta_0, \dots, \eta_{N+1}) \in X_N$ and $\mathbf{z} = (z_0, \dots, z_{N+1})$.

From Lemma 2 it follows that there exists a unique function $\hat{f} \in F(\boldsymbol{\eta}, \mathbf{z})$ for which the infimum is attained. Since by Lemma 2, $\hat{f} \in S_5(\eta_1, \dots, \eta_N)$ and $f^* \in S_5(x_1^*, \dots, x_N^*)$, and by assumption $\boldsymbol{\eta} \neq \mathbf{x}^*$, then it follows that $\hat{f} \neq f^*$. Note that $f^* \in F(\boldsymbol{\eta}, \mathbf{z})$ but the extremal interpolant in $F(\boldsymbol{\eta}, \mathbf{z})$ is the function $\hat{f}$. Therefore

$$\|\hat{f}'''\|_2 < \|f^{*\prime\prime\prime}\|_2.$$

Now, observe that $|z_i| = |f^*(\eta_i)| \geq |y_i|$, $i = 1, \dots, N$. Then for the continious function $\hat{f}$ there exist points $a = \zeta_0 < \zeta_1 < \cdots < \zeta_N < \zeta_{N+1} = b$ such that $\hat{f}(\zeta_i) = y_i$, $i = 0, \dots, N+1$. This means that $\hat{f} \in F(\boldsymbol{\zeta}, \mathbf{y})$, $\boldsymbol{\zeta} = (\zeta_0, \dots, \zeta_{N+1}) \in X_N$, and $\|\hat{f}'''\|_2 < \|f^{*\prime\prime\prime}\|_2$ which contradicts the minimality property of the function f^* for the extremal problem (2.3).

Thus, we proved that the extremal points of f^* coincide with interpolation knots, i.e. $\eta_i = x_i^*$ for all $i = 0, \dots, N+1$. Therefore, f^* is strictly monotone in $[x_i^*, x_{i+1}^*]$, $i = 1, \dots, N$. □

5. PROOF OF THEOREM 2

Let $x_i < x_{i+1}$ and arbitrary real numbers y_i, y_{i+1}, s_i'', s_{i+1}'' be given. We set $\Delta_i = x_{i+1} - x_i$, $\Delta y_i = y_{i+1} - y_i$ and denote by $P_i(t) \in \pi_5$ the polynomial satisfying

$$\begin{array}{lll} P_i(x_i) = y_i, & P_i'(x_i) = 0, & P_i''(x_i) = s_i'', \\ P_i(x_{i+1}) = y_{i+1}, & P_i'(x_{i+1}) = 0, & P_i''(x_{i+1}) = s_{i+1}''. \end{array} \tag{5.1}$$

We can find explicitly the polynomial P_i solving Hermite iterpolation problem (5.1). Standard calculations show that the following relations hold true:

$$\begin{aligned} P_{i-1}'''(x_i) &= \frac{6}{\Delta_{i-1}^3}\big(10\Delta y_{i-1} - \frac{1}{2}\Delta_{i-1}^2 s_{i-1}'' + \frac{3}{2}\Delta_{i-1}^2 s_i''\big), \\ P_i'''(x_i) &= \frac{6}{\Delta_i^3}\big(10\Delta y_i - \frac{3}{2}\Delta_i^2 s_i'' + \frac{1}{2}\Delta_i^2 s_{i+1}''\big), \\ P_{i-1}^{IV}(x_i) &= \frac{24}{\Delta_{i-1}^4}\big(15\Delta y_{i-1} - \Delta_{i-1}^2 s_{i-1}'' + \frac{3}{2}\Delta_{i-1}^2 s_i''\big), \\ P_i^{IV}(x_i) &= \frac{24}{\Delta_i^4}\big(-15\Delta y_i + \frac{3}{2}\Delta_i^2 s_i'' - \Delta_i^2 s_{i+1}''\big). \end{aligned} \tag{5.2}$$

We seek for a spline $s \in S_5(x_1, \ldots, x_N)$ with

$$s \in C^4[a,b], \qquad P_i = s\big|_{(x_i, x_{i+1})} \in \pi_5, \quad i = 0, \ldots, N, \tag{5.3}$$

satisfying the interpolation conditions (2.4).

Let us set

$$\begin{array}{lll} & s_i'' = s''(x_i), & i = 0, \ldots, N+1, \\ \Delta_i = x_{i+1} - x_i, & \Delta y_i = y_{i+1} - y_i, & i = 0, \ldots, N, \\ \alpha_i = \dfrac{\Delta_i}{\Delta_{i-1}}, & \delta_{i-1} = \dfrac{\Delta y_i}{\Delta y_{i-1}}, & i = 1, \ldots, N, \\ \beta_i = \dfrac{\Delta_i^2 s_i''}{2\Delta y_i}, & \gamma_i = \dfrac{\Delta_i^2 s_{i+1}''}{2\Delta y_i}, & i = 0, \ldots, N. \end{array} \tag{5.4}$$

The boundary conditions for $s'''(t)$ at the endpoints and the continuity conditions for $s'''(t)$ and $s^{IV}(t)$ at the knots $\{x_i\}_{i=1}^N$ can be written for the polynomial pieces P_i as follows:

$$\begin{array}{lll} P_0'''(x_0) = 0, & P_{i-1}'''(x_i) = P_i'''(x_i), \quad i = 1, \ldots, N, & P_N'''(x_{N+1}) = 0, \\ & P_{i-1}^{IV}(x_i) = P_i^{IV}(x_i), \quad i = 1, \ldots, N, & \end{array} \tag{5.5}$$

From (5.2)–(5.4) we obtain for $P_0'''(x_0) = 0$ in (5.5):

$$10 - 3\beta_0 + \gamma_0 = 0,$$

i.e.

$$\gamma_0 = 3\beta_0 - 10. \tag{5.6}$$

Using (5.2)–(5.4) we have for the continuity conditions (5.5) at x_1:

$$\begin{aligned}(20 - 2\beta_0 + 6\gamma_0)\alpha_1^3 &= (20 - 6\beta_1 + 2\gamma_1)\delta_0,\\ (60 - 8\beta_0 + 12\gamma_0)\alpha_1^4 &= (-60 + 12\beta_1 - 8\gamma_1)\delta_0.\end{aligned} \tag{5.7}$$

Now, taking into account (5.6) we rewrite (5.7) in the form

$$\begin{aligned}(8\beta_0 - 20)\alpha_1^3 &= (10 - 3\beta_1 + \gamma_1)\delta_0,\\ (7\beta_0 - 15)\alpha_1^4 &= (-15 + 3\beta_1 - 2\gamma_1)\delta_0,\end{aligned} \tag{5.8}$$

Note that $\beta_1\delta_0 = \gamma_0\,\alpha_1^2$ from (5.4). Then by elimination of γ_1 in (5.8) we get

$$(7\beta_0 - 15)\alpha_1^4 + (16\beta_0 - 40)\alpha_1^3 + (9\beta_0 - 30)\alpha_1^2 - 5\delta_0 = 0. \tag{5.9}$$

On the other hand, from equalities (5.8) it follows that

$$\begin{aligned}15 - 2\beta_1 + 3\gamma_1 &= \frac{-1}{12\delta_0}\big[28(7\beta_0 - 15)\alpha_1^4 + 10(16\beta_0 - 40)\alpha_1^3 + 40\delta_0\big],\\ 20 - 2\beta_1 + 6\gamma_1 &= \frac{-1}{3\delta_0}\big[16(7\beta_0 - 15)\alpha_1^4 + 7(16\beta_0 - 40)\alpha_1^3 + 40\delta_0\big],\\ 3\gamma_1 &= \frac{-3}{8\delta_0}\big[8(7\beta_0 - 15)\alpha_1^4 + 4(16\beta_0 - 40)\alpha_1^3 + 40\delta_0\big].\end{aligned} \tag{5.10}$$

Similarly, for $i = 2, \ldots, N$ we obtain from (5.2)–(5.5):

$$\begin{aligned}(10 - \beta_{i-1} + 3\gamma_{i-1})\alpha_i^3 &= (10 - 3\beta_i + \gamma_i)\delta_{i-1},\\ (15 - 2\beta_{i-1} + 3\gamma_{i-1})\alpha_i^4 &= (-15 + 3\beta_i - 2\gamma_i)\delta_{i-1}.\end{aligned} \tag{5.11}$$

Since $\delta_{i-1}\beta_i = \gamma_{i-1}\alpha_i^2$ by (5.4), equalities (5.11) give

$$(15 - 2\beta_{i-1} + 3\gamma_{i-1})\alpha_i^4 + (20 - 2\beta_{i-1} + 6\gamma_{i-1})\alpha_i^3 + 3\gamma_{i-1}\alpha_i^2 - 5\delta_{i-1} = 0, \tag{5.12}$$

and

$$\begin{aligned}&15 - 2\beta_i + 3\gamma_i\\ &\quad= \frac{-1}{12\delta_{i-1}}\big[28(15 - 2\beta_{i-1} + 3\gamma_{i-1})\alpha_i^4 + 10(20 - 2\beta_{i-1} + 6\gamma_{i-1})\alpha_i^3 + 40\delta_{i-1}\big],\\ &20 - 2\beta_i + 6\gamma_i\\ &\quad= \frac{-1}{3\delta_{i-1}}\big[16(15 - 2\beta_{i-1} + 3\gamma_{i-1})\alpha_i^4 + 7(20 - 2\beta_{i-1} + 6\gamma_{i-1})\alpha_i^3 + 40\delta_{i-1}\big],\\ &3\gamma_i = \frac{-3}{8\delta_{i-1}}\big[8(15 - 2\beta_{i-1} + 3\gamma_{i-1})\alpha_i^4 + 4(20 - 2\beta_{i-1} + 6\gamma_{i-1})\alpha_i^3 + 40\delta_{i-1}\big].\end{aligned} \tag{5.13}$$

Finally, from (5.2), (5.4), and (5.13) we obtain

$$
\begin{aligned}
s'''(x_{N+1}) = P_N'''(x_{N+1}) &= \frac{3\Delta y_N}{\Delta_N^3}\big(20 - 2\beta_N + 6\gamma_N\big) \\
&= \frac{-\Delta y_N}{\Delta_N^3\,\delta_{N-1}}\big[16(15 - 2\beta_{N-1} + 3\gamma_{N-1})\alpha_N^4 \\
&\qquad + 7(20 - 2\beta_{N-1} + 6\gamma_{N-1})\alpha_N^3 + 40\delta_{N-1}\big].
\end{aligned} \tag{5.14}
$$

Remark 4. We will make use of the equalities (5.10) and (5.13) as recurrence relations for the coefficients in the algebraic equations (5.9) and (5.12), and for the quantity (5.14).

Now we consider useful monotonicity properties of polynomial zeros under recurrence relations of the polynomial coefficients. The following two lemmas can be found in [20]; proofs in full details are given in the PhD Thesis of the second author [19].

By the classical *Descartes' rule* a polynomial $a_0x^n + a_1x^{n-1} + \cdots + a_n$ has no more positive zeros counting multiplicities than the number of strict sign changes in the sequence $a_0, \ldots, a_n$. In particular, if there is exactly one strict sign change in the sequence of coefficients then the polynomial has exactly one simple positive zero.

Lemma 4 (Uluchev [20, Lemma 3.3.1])**.** *Suppose that the coefficients $a_0(\tau)$, $a_1(\tau)$, $a_2(\tau)$, a_4 of the function*

$$Q(\tau, z) = a_0(\tau)z^4 + a_1(\tau)z^3 + a_2(\tau)z^2 + a_4$$

satisfy the conditions:

(i) $a_4 = const.,\ \ a_4 > 0$;

(ii) $a_i(\tau) \in C^1_{[t,T]},\ \ i = 0, 1, 2,\ \ t < T$;

(iii) $a_0(t) \le 0,\ \ a_1(t) < 0,\ \ a_2(t) < 0$;

(iv) $a_i'(\tau) > 0,\ \ i = 0, 1, 2,\ \ \tau \in (t, T)$;

(v) there exist $\{\tau_i\}_{i=0}^2,\ \ t \le \tau_0 < \tau_1 < \tau_2 < T$ *with* $a_i(\tau_i) = 0,\ \ i = 0, 1, 2.$

Then, there exist unique points t_1 and t_2, such that $t < t_1 < t_2 < T$, and the equation with respect to z,

$$Q(\tau, z) = 0,$$

(a) has exactly one positive simple root $z(\tau)$ if $\tau \in [t, t_1]$ is fixed;

(b) has exactly two positive simple roots $z(\tau) < \hat{z}(\tau)$ if $\tau \in (t_1, t_2)$ is fixed;

(c) has exactly one positive root $z(\tau) = \hat{z}(\tau)$ of multiplicity two if $\tau = t_2$;

(d) has no positive root if $\tau \in (t_2, T]$ is fixed;

(e) $z(\tau) \in C^1_{(t,t_2)}$ and $z'(\tau) > 0$ for $\tau \in (t, t_2)$.

Remark 5. More precisely, in Lemma 4, $t_1 = \tau_0$ and the larger positive zero $\hat{z}(\tau)$ of $Q(\tau, z)$ comes from $+\infty$ as τ runs to the right of t_1. For $\tau \in (t_1, t_2)$, $z(\tau)$ increases, $\hat{z}(\tau)$ decreases, and both positive zeros of $Q(\tau, z)$ coincide for $\tau = t_2$.

Let us set

$$\begin{aligned}
&Q(\tau, z) = a_0(\tau)z^4 + a_1(\tau)z^3 + a_2(\tau)z^2 + a_4,\\
&b_0(\tau, z) = A_0\big(28a_0(\tau)z^4 + 10a_1(\tau)z^3 - 8a_4\big), \quad A_0 = const., \quad A_0 > 0,\\
&b_1(\tau, z) = A_1\big(16a_0(\tau)z^4 + 7a_1(\tau)z^3 - 8a_4\big), \quad A_1 = const., \quad A_1 > 0,\\
&b_2(\tau, z) = A_2\big(8a_0(\tau)z^4 + 4a_1(\tau)z^3 - 8a_4\big), \quad A_2 = const., \quad A_2 > 0,\\
&\qquad b_4 = const., \quad b_4 > 0.
\end{aligned} \tag{5.15}$$

Lemma 5 (Uluchev [20, Lemma 3.3.2])**.** *Suppose that the coefficients $a_0(\tau)$, $a_1(\tau)$, $a_2(\tau)$, a_4 of function*

$$Q(\tau, z) = a_0(\tau)z^4 + a_1(\tau)z^3 + a_2(\tau)z^2 + a_4$$

satisfy the conditions:

(i) $a_4 = const.$, $a_4 > 0$;

(ii) $a_i(\tau) \in C^1_{[\tau_0,\tau_2]}$;

(iii) $a_i(\tau_i) = 0$, $i = 0, 1, 2$, $\tau_0 < \tau_1 < \tau_2$;

(iv) $a_i'(\tau) > 0$, $i = 0, 1, 2$, $\tau \in (\tau_0, \tau_2)$.

Now, Lemma 4 applies and let t_1, t_2, $z(\tau)$, $\hat{z}(\tau)$, $\xi(\tau)$ be as in Lemma 4. Then, for b_0, b_1, b_2, b_4 defined in (5.15)*,*

(a) the algebraic equation with respect to z,

$$b_0(\tau, \hat{z}(\tau))z^4 + b_1(\tau, \hat{z}(\tau))z^3 + b_2(\tau, \hat{z}(\tau))z^2 + b_4 = 0,$$

has no positive root if $\tau \in (t_1, t_2)$ is fixed;

(b) there exist unique points $\{\theta_i\}_{i=0}^2$ such that $t_1 < \theta_0 < \theta_1 < \theta_2 = t_2$ and

$$\begin{aligned}
b_i(\tau, z(\tau)) < 0, \quad \tau \in (t_1, \theta_i), \quad i = 0, 1, 2,\\
b_i(\theta_i, z(\theta_i)) = 0, \quad i = 0, 1, 2;
\end{aligned}$$

(c) the functions $b_j(\tau) = b_j(\tau, z(\tau))$, $j = 0, 1, 2$, and b_4 satisfy conditions (i)–(iv) of Lemma 4 for the interval $[\theta_0, \theta_2]$.

Proof of Theorem 2. Let us set

$$\begin{aligned}
\tau &= 7\beta_0 - 15,\\
a_{0,1}(\tau) &= \tau, \qquad a_{1,1}(\tau) = \frac{16}{7}\Big(\tau - \frac{5}{2}\Big), \qquad a_{2,1}(\tau) = \frac{9}{7}\Big(\tau - \frac{25}{3}\Big), \qquad a_{4,1} = -5\delta_0,\\
a_{0,i}(\tau) &= \frac{-1}{12\delta_{i-2}}\big(28a_{0,i-1}(\tau)\alpha_{i-1}^4 + 10a_{1,i-1}(\tau)\alpha_{i-1}^3 - 8a_{4,i-1}(\tau)\big), \quad i = 2,\dots,N,\\
a_{1,i}(\tau) &= \frac{-1}{3\delta_{i-2}}\big(16a_{0,i-1}(\tau)\alpha_{i-1}^4 + 7a_{1,i-1}(\tau)\alpha_{i-1}^3 - 8a_{4,i-1}(\tau)\big), \qquad i = 2,\dots,N,\\
a_{2,i}(\tau) &= \frac{-3}{8\delta_{i-2}}\big(8a_{0,i-1}(\tau)\alpha_{i-1}^4 + 4a_{1,i-1}(\tau)\alpha_{i-1}^3 - 8a_{4,i-1}(\tau)\big), \qquad i = 2,\dots,N,\\
a_{4,i} &= -5\delta_{i-1}, \qquad i = 2,\dots,N.
\end{aligned} \tag{5.16}$$

Using the recurrence relations (5.10) and (5.13) in view of the notations (5.16), we rewrite equations (5.9), (5.12) in the form

$$a_{0,i}(\tau)\alpha_i^4 + a_{1,i}(\tau)\alpha_i^3 + a_{2,i}(\tau)\alpha_i^2 + a_{4,i} = 0, \qquad i = 1,\dots,N, \tag{5.17}$$

and we seek for a solution $\tau, \alpha_1, \dots, \alpha_N$ of the nonlinear system (5.17) such that

$$\alpha_i > 0, \qquad i = 1,\dots,N. \tag{5.18}$$

In addition, by the interpolation conditions (2.4) the spline $s \in S_5(x_1,\dots,x_N)$ defined in (5.3) has to satisfy (5.14), which in view of notations (5.16) takes the form

$$s'''(x_{N+1}) = \frac{-\Delta y_N}{\Delta_N^3\,\delta_{N-1}}\big(16a_{0,N}(\tau)\alpha_N^4 + 7a_{1,N}(\tau)\alpha_N^3 + 40\delta_{N-1}\big) = 0. \tag{5.19}$$

Observe that $\delta_{i-1} < 0$, $i = 1,\dots,N$ and then

$$\frac{-1}{12\delta_{i-1}} > 0, \qquad \frac{-1}{3\delta_{i-1}} > 0, \qquad \frac{-3}{8\delta_{i-1}} > 0, \qquad i = 1,\dots,N.$$

We briefly sketch the idea of our proof. Let us denote the i-th equation of the system (5.17) by (5.17.i). We will bound τ for which the system (5.17) has a solution, satisfying (5.18), to a finite interval J. Moreover, for each fixed $\tau \in J$ we can uniquely determine $\alpha_i > 0$ satisfying (5.17.i), $i = 1,\dots,N-1$, and (5.17.N) would have positive roots $\alpha_N(\tau) < \hat{\alpha}_N(\tau)$. We will show that the function

$$\hat{\sigma}(\tau) = 16a_{0,N}(\tau)\hat{\alpha}_N^4(\tau) + 7a_{1,N}(\tau)\hat{\alpha}_N^3(\tau) + 40\delta_{N-1}$$

does not vanish in J, i.e. (5.19) cannot be satisfied if we choose the larger positive zero of (5.17.N). Using the smaller positive zero $\alpha_N(\tau)$ of (5.17.N) we will prove that

$$\sigma(\tau) = 16a_{0,N}(\tau)\alpha_N^4(\tau) + 7a_{1,N}(\tau)\alpha_N^3(\tau) + 40\delta_{N-1} \tag{5.20}$$

is monotone and has a unique zero in J. Hence, we will obtain a procedure and numerical algorithm for solving the system (5.17)–(5.19).

First, for any solution of (5.17)–(5.19), the relation $\tau \in [0, \frac{25}{3}]$ must hold. Otherwise we have two cases.

If $\tau < 0$, then $a_{j,1}(\tau) < 0$, $j = 0, 1, 2$ and $a_{4,1} > 0$. Hence equation (5.17.1) has only one positive root $\alpha_1(\tau)$. Recurrence formulae (5.16) yield that $a_{j,k}(\tau) < 0$, $j = 0, 1, 2$ and $a_{4,k} > 0$, hence (5.17.k) has a unique positive root $\alpha_k(\tau)$ for all $k = 2, \ldots, N$. Then $\sigma(\tau) < 0$ which means that $s'''(x_{N+1}) \neq 0$, i.e. (5.19) is not satisfied.

In case of $\tau > \frac{25}{3}$, we have $a_{j,1}(\tau) > 0$, $j = 0, 1, 2$ and $a_{4,1} > 0$. Then (5.17.1) has no positive root, hence the system (5.17) has no solution satisfying (5.18).

Let us set $\tau_0^{(1)} = 0$, $\tau_1^{(1)} = \frac{5}{2}$ and $\tau_2^{(1)} = \frac{25}{3}$. Since $\tau_0^{(1)} < \tau_1^{(1)} < \tau_2^{(1)}$, the coefficients $a_{j,1}(\tau)$, $j = 0, 1, 2$ and $a_{4,1}$ satisfy conditions (i)–(iv) of Lemma 5.

Suppose that for a fixed $k \in \{1, \ldots, N-1\}$ we have proved that any solution $\tau, \alpha_1, \ldots, \alpha_N$ of (5.17)–(5.19) is such that $\tau \in \left[\tau_0^{(k)}, \tau_2^{(k)}\right]$ and the coefficients $a_{j,k}(\tau)$, $j = 0, 1, 2$ and $a_{4,k}$ satisfy conditions (i)–(iv) of Lemma 5 for all $\tau \in \left[\tau_0^{(k)}, \tau_2^{(k)}\right]$. By Lemma 4 there exist points $t_1^{(k)}$ and $t_2^{(k)}$ such that $\tau_0^{(k)} = t_1^{(k)} < t_2^{(k)} < \tau_2^{(k)}$ and the equation (5.17.k) has:

- exactly one simple positive root $\alpha_k(\tau)$ if $\tau \leq t_1^{(k)}$;
- exactly two positive roots $\alpha_k(\tau) < \hat{\alpha}_k(\tau)$ if $\tau \in \left(t_1^{(k)}, t_2^{(k)}\right)$;
- exactly one positive root of multiplicity two $\alpha_k(\tau) = \hat{\alpha}_k(\tau)$ if $\tau = t_2^{(k)}$;
- no positive root if $\tau \in \left(t_2^{(k)}, \tau_2^{(k)}\right)$.

Assume that there exist a solution $\tau, \alpha_1, \ldots, \alpha_N$ of (5.17)–(5.19), such that $\alpha_k = \hat{\alpha}_k(\tau)$ for some $\tau \in \left(t_1^{(k)}, t_2^{(k)}\right)$. That is, α_k is the larger positive zero $\hat{\alpha}_k(\tau)$ of (5.17.k). From Lemma 5 (a) it follows that (5.17.$k+1$) has no positive root with respect to α_{k+1}, hence (5.17) has no solution satisfying (5.18).

Therefore for any solution $\tau, \alpha_1, \ldots, \alpha_N$ of (5.17)–(5.19) with $\tau \in \left(t_1^{(k)}, t_2^{(k)}\right)$, there holds $\alpha_k = \alpha_k(\tau)$ which is the smaller positive zero of (5.17.k). Application of Lemma 5 (b) gives that there exist unique points $\tau_j^{(k+1)}$, $j = 0, 1, 2$, with $t_1^{(k)} < \tau_0^{(k+1)} < \tau_1^{(k+1)} < \tau_2^{(k+1)} < t_2^{(k)}$ and $a_{j,k+1}(\tau_j^{(k+1)}) = 0$, $j = 0, 1, 2$. Now Lemma 5 (c) yields that the coefficients $a_{j,k+1}(\tau)$, $j = 0, 1, 2$ and $a_{4,k+1}$ satisfy conditions (i)–(iv) of Lemma 4 for $\tau \in \left[\tau_0^{(k+1)}, \tau_2^{(k+1)}\right]$.

Similar arguments as for $k = 1$ above show that for any solution $\tau, \alpha_1, \ldots, \alpha_N$ of (5.17)–(5.19) there holds $\tau \in \left[\tau_0^{(k+1)}, \tau_2^{(k+1)}\right]$. The arguments are the same as for $k = 1$ above.

For $\tau \in \left[t_1^{(k)}, \tau_0^{(k+1)}\right]$ we have $a_{j,k+1}(\tau) < 0$, $j = 0, 1, 2$ and $a_{4,k+1} > 0$. Then equation (5.17.$k+1$) has only one positive root $\alpha_{k+1} = \alpha_{k+1}(\tau)$. Recurrence formulae (5.16) yield that $a_{j,\ell}(\tau) < 0$, $j = 0, 1, 2$ and $a_{4,\ell} > 0$, hence (5.17.ℓ) has a

unique positive root $\alpha_\ell(\tau)$ for all $\ell = k+1, \dots, N$. But then $\sigma(\tau) < 0$ which means that $s'''(x_{N+1}) \neq 0$, i.e. (5.19) is not satisfied.

In the case $\tau \in \left(\tau_2^{(k+1)}, t_2^{(k)}\right]$ we have $a_{j,k+1}(\tau) > 0$, $j = 0, 1, 2$ and $a_{4,k+1} > 0$. Then $(5.17.k+1)$ has no positive root and the system (5.17) has no solution satisfying (5.18).

By Lemma 4 there exist unique points $t_1^{(N)}$ and $t_2^{(N)}$ such that $\tau_0^{(N)} = t_1^{(N)} < t_2^{(N)} < \tau_2^{(N)}$ and the equation $(5.17.N)$ has:

- exactly one simple positive root $\alpha_N(\tau)$ if $\tau \leq t_1^{(N)}$;
- exactly two positive roots $\alpha_N(\tau) < \hat{\alpha}_N(\tau)$ if $\tau \in \left(t_1^{(N)}, t_2^{(N)}\right)$;
- exactly one positive root of multiplicity two $\alpha_N(\tau) = \hat{\alpha}_N(\tau)$ if $\tau = t_2^{(N)}$;
- no positive root if $\tau \in \left(t_2^{(N)}, \tau_2^{(N)}\right)$.

So, we obtain a sequence of nested intervals

$$\left[t_1^{(N)}, t_2^{(N)}\right] \subset \left[t_1^{(N-1)}, t_2^{(N-1)}\right] \subset \cdots \subset \left[t_1^{(1)}, t_2^{(1)}\right] \subset \left[0, \tfrac{25}{3}\right],$$

and for any solution $\tau, \alpha_1, \dots, \alpha_N$ of (5.17)–(5.19) there holds $\tau \in \left[t_1^{(N)}, t_2^{(N)}\right]$.

Now we study functions $\hat{\sigma}(\tau)$ and $\sigma(\tau)$, $\tau \in \left[t_1^{(N)}, t_2^{(N)}\right]$. Observe that in view of notations (5.15),

$$\hat{\sigma}(\tau) = b_1(\tau, \hat{\alpha}_N(\tau)) \qquad \text{with} \quad A_1 = 1.$$

Also, the proof of Lemma 5 (a) relies on the inequalities $b_j(\tau, \hat{z}(\tau)) > 0$, $\tau \in (t_1, t_2)$ for each $j = 0, 1, 2$ (see [20, Eq. (3.3.6)]). If for some $\tau \in (t_1^{(N)}, t_2^{(N)})$ there is a solution $\alpha_1, \dots, \alpha_N$ of (5.17)–(5.18) with $\alpha_N = \hat{\alpha}_N(\tau)$, being the larger positive zero of the equation $(5.17.N)$, then $\hat{\sigma}(\tau) > 0$. Hence, $s'''(x_{N+1}) \neq 0$ and condition (5.19) is not satisfied.

It follows that for any solution $\tau, \alpha_1, \dots, \alpha_N$ of (5.17)–(5.19) there holds $\alpha_k = \alpha_k(\tau)$, being the smaller positive zero of the equation $(5.17.k)$ for all $k = 1, \dots, N$, and $\tau \in (t_1^{(N)}, t_2^{(N)})$. By the notations in (5.15) we have

$$\sigma(\tau) = b_1(\tau, \alpha_N(\tau)) \qquad \text{with} \quad A_1 = 1.$$

According to Lemma 5 (c) the function $\sigma(\tau)$ satisfies condition (iv) of Lemma 4, i.e. $\sigma'(\tau) > 0$, $\tau \in (t_1^{(N)}, t_2^{(N)})$. Then $\sigma(\tau)$ is strictly monotone for $\tau \in (t_1^{(N)}, t_2^{(N)})$. By Lemma 5 (b) there exists a unique point $\tau^* \in (t_1^{(N)}, t_2^{(N)})$ such that $\sigma(\tau^*) = 0$, which implies $s'''(x_{N+1}) = 0$, i.e. (5.19).

So, we have proved that there exists a unique spline function s^* and knots $\mathbf{x}^* = (x_1^*, \dots, x_N^*) \in X_N$, such that $s^* \in S_5(x_1^*, \dots, x_N^*) \cap F(\mathbf{x}^*, \mathbf{y})$ and s^* satisfies the characterization of the smoothest interpolant to the problem (2.3) given in Theorem 1. This completes the proof of the theorem. □

6. NUMERICAL ALGORITHM AND RESULTS

Here we discuss computational aspects of finding the unique oscillating spline interpolant from Theorem 2. We follow the procedure described in the proof of Theorem 2.

Let us fix τ as a point from an equidistant mesh for $\left[0, \frac{25}{3}\right]$. If the first equation (5.17.1) of the system (5.17) has not two simple positive roots we skip this value of τ and go to the next point of the mesh. If we do not succeed for that mesh, we decrease the mesh step and repeat. Thus, we find interval J_1 such that for each $\tau \in J_1$, (5.17.1) has two simple positive roots and we set α_1 to be the smaller of them. We represent the coefficients of the next algebraic equation (5.17.2) by α_1. If for a fixed τ from an equidistant mesh of J_1 the equation (5.17.2) of the system (5.17) has not two simple positive roots we skip this value of τ and go to the next point of the mesh in J_1. If we do not succeed for that mesh we decrease the mesh step and repeat. In this way we find interval $J_2 \subset J_1$ such that for each $\tau \in J_2$, (5.17.2) has two simple positive roots and we set α_2 to be the smaller of them. Repeating this process for each $i = 1, \ldots, N$ we find an interval J_i such that for $\tau \in J_i$ all the equations (5.17.1)–(5.17.i) have two positive roots. Moreover, $J_N \subset J_{N-1} \subset \cdots \subset J_1 \subset \left[0, \frac{25}{3}\right]$. Here intervals J_i are related to the intervals $\left[t_1^{(i)}, t_2^{(i)}\right]$, $i = 1, \ldots, N$ in the proof of Theorem 2.

Observe that (5.19)–(5.20) yield that $s'''(x_{N+1}) = s'''(\tau, x_{N+1}) = 0$ if $\sigma(\tau) = 0$. But the function $\sigma(\tau)$ defined in (5.20) is a monotone function of $\tau \in J_N$ and it changes sign in J_N. Then we find approximately τ^* by an equidistant mesh of the interval J_N, for which $\sigma(\tau)$ is minimal by absolute value.

Then we solve (5.17) and find $\alpha_i^* = \alpha_i(\tau^*)$, $i = 1, \ldots, N$.

In the next step we find Δ_i, $i = 0, \ldots, N$ using the formulae

$$\Delta_0 = \frac{b-a}{1 + \sum_{i=1}^{N} \prod_{j=1}^{i} \alpha_j^*}, \qquad \Delta_{i+1} = \alpha_i^* \Delta_i, \qquad i = 0, \ldots, N. \tag{6.1}$$

Hence, the optimal knots for the extremal problem (2.3) are

$$x_0^* = a, \qquad x_{i+1}^* = x_i^* + \Delta_i, \quad i = 0, \ldots, N-1, \qquad x_{N+1}^* = b. \tag{6.2}$$

From (5.6), (5.10), (5.13), and (5.16) we find recurrently β_i, $i = 0, \ldots, N$ and γ_i, $i = 0, \ldots, N$. Now, applying (5.4) we find

$$s''(x_i) = s_i'' = \frac{2\beta_i \Delta y_i}{\Delta_i^2}, \quad i = 0, \ldots, N, \quad s''(x_{N+1}) = s_{N+1}'' = \frac{2\gamma_N \Delta y_N}{\Delta_N^2}. \tag{6.3}$$

Next, we find the polynomial pieces $P_i \in \pi_5$ for $[x_i^*, x_{i+1}^*]$ by solving the Hermite interpolation problem (5.1), $i = 0, \ldots, N$. So, based on (5.3) we get the oscillating spline interpolant $s^*(t)$ satisfying (2.4).

Algorithm 1 Finding the Oscillating Spline Interpolant

Input:	Data $\mathbf{y} = (y_0, y_1, \ldots, y_{N+1})$ with (1.2) and (1.3)
Step 1.	Find τ^* such that the system (5.17) has a solution satisfying (5.18) and $s'''(x_{N+1}) \approx 0$
Step 2.	For τ^* obtained in Step 1 find $\{\alpha_i\}_{i=1}^{N}$ solving (5.17) with (5.18)
Step 3.	For τ^* as in Step 1 find the knots $\mathbf{x}^* = (x_0^*, x_1^*, \ldots, x_{N+1}^*)$ from (6.1)–(6.2)
Step 4.	For τ^* obtained in Step 1 find the quantities $\{s_i''\}_{i=0}^{N+1}$ from (6.3)
Step 5.	Construct the polynomial pieces $P_i \in \pi_5$ in $[x_i^*, x_{i+1}^*]$ by solving (5.1), $i = 0, \ldots, N$
Step 6.	Construct the oscillating spline interpolant $s^*(t)$ based on (5.3)
Output:	The knots $\mathbf{x}^*$ of $s^*(t)$, polynomial pieces $\{P_i\}_{i=0}^{N}$ of the spline interpolant $s^*(t)$, graphs of $s^*(t)$ and its derivatives

We summarize in an algorithm the basic steps we pass to find the fifth degree oscillating spline interpolant with boundary conditions.

With the assistance of *Mathematica* (by Wolfram Research Inc.) computer algebra system, we implement the above algorithm to a numerical example.

Example. We show results of numerical experiments for the data

$$N = 9, \qquad \mathbf{y} = (1,\ -2,\ 3,\ -1,\ 5,\ 2,\ 4,\ 0,\ 1,\ -3,\ 2),$$

satisfying conditions (1.2) and (1.3).

According to the algorithm described in the previous section, $J_i = [\ell_i, r_i]$ is an interval such that for $\tau \in J_i$ all the equations (5.17.1)–(5.17.i) have two positive roots, $i = 1, \ldots, 9$. Moreover, $J_9 \subset J_8 \subset \cdots \subset J_1 \subset \left[0, \frac{25}{3}\right]$. These nested intervals are given in Table 1.

J_i	ℓ_i	r_i
J_1	0.1	2.2
J_2	2.012	2.177
J_3	2.17	2.17495
J_4	2.1746	2.174867
J_5	2.174856	2.174864
J_6	2.1748639	2.174864
J_7	2.174864057	2.174864071
J_8	2.1748640706	2.17486407086
J_9	2.174864070844	2.1748640708585

Table 1: Nested intervals $J_9 \subset J_8 \subset \cdots \subset J_1 \subset \left[0, \frac{25}{3}\right]$

Our numerical results confirm monotonicity of the function $\sigma(\tau)$ for $\tau \in J_N$. Table 2 shows values of the function $\sigma(\tau)$ from (5.20), evaluated at equidistant points τ in a small interval $J \subset J_9$, where $\sigma(\tau) \approx 0$ and $\sigma(\tau)$ changes sign.

τ	$\sigma(\tau)$
2.17486407085837258	-0.00327161
2.17486407085837259	-0.00256242
2.1748640708583726	-0.00202345
2.17486407085837261	-0.00132029
2.17486407085837262	-0.00065045
2.17486407085837263	0.00001273
2.17486407085837264	0.00062011
2.17486407085837265	0.00132710
2.17486407085837266	0.00203236
2.17486407085837267	0.00274238

Table 2: $\sigma(\tau)$ for $\tau \in J = [2.17486407085837258, 2.17486407085837267]$

Now, we choose $\tau^* = 2.17486407085837263$ for which $\sigma(\tau) = 0.00001273$ is minimal in absolute value in Table 2, whence $s'''(x_{N+1}) \approx 0$. Solving the system (5.17) with (5.18) for $\tau = \tau^*$ we obtain the ratios $\alpha_i = \Delta_i/\Delta_{i-1}$, $i = 1, \ldots, 9$. Hence, using (6.1) and (6.2) we find the interpolation knots $\{x_i^*\}_{i=0}^{10}$, being also knots of the oscillating spline interpolant, for the interval $[a, b] = [0, 1]$. The knots are listed in Table 3.

x_0^*	0
x_1^*	0.093572609937859
x_2^*	0.207960413155138
x_3^*	0.310783910315965
x_4^*	0.43351596313152
x_5^*	0.52765838534873
x_6^*	0.60829886372617
x_7^*	0.71795857706244
x_8^*	0.77954122488487
x_9^*	0.88685751448236
x_{10}^*	1

Table 3: Interpolation and spline knots $(x_0^*, x_1^*, \ldots, x_{10}^*)$ for $[0, 1]$

Plot of the oscillating spline interpolant $s^*(t)$ satisfying (2.4), its first derivative, and its third derivative are shown in Figure 1, Figure 2, and Figure 3, respectively.

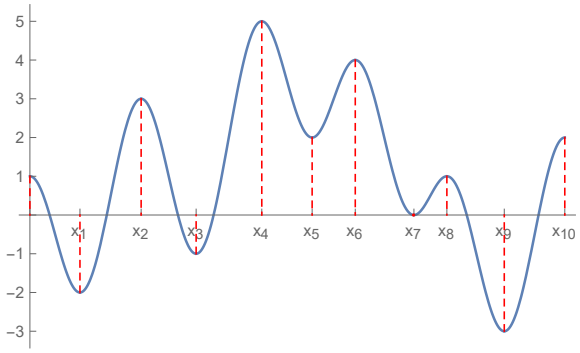


Figure 1: The smoothest interpolant $s^*(t)$

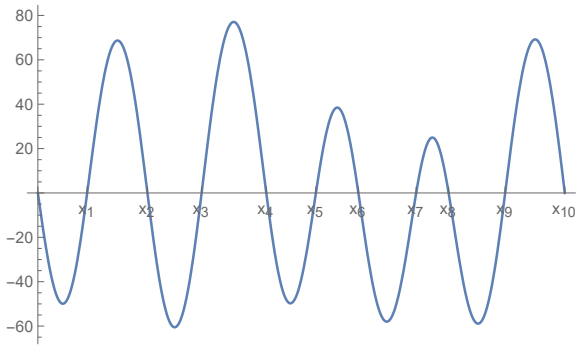


Figure 2: First derivative of the smoothest interpolant, $s^{*\prime}(t)$

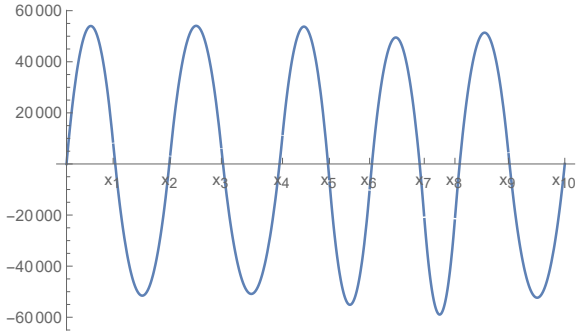


Figure 3: Third derivative of the smoothest interpolant, $s^{*\prime\prime\prime}(t)$

Acknowledgement. The authors are grateful to the referees for careful reading of the paper and their valuable suggestions and comments. This work was supported by the bilateral project DNTS-Austria KP-06/2019 (WTZ BG 03/2019), funded by Bulgarian National Science Fund and OeAD (Austria).

7. REFERENCES

[1] Bojanov, B.: Perfect splines of least deviation. *Anal. Math.* **16**, 1980, 185–197.

[2] Bojanov, B. σ-perfect splines and their applications to optimal recovery problems. *J. Complexity* **3**, 1987, 429–450.

[3] Bojanov, B.: B-splines with Birkhoff knots. *Constr. Approx.* **4**, 1988, 147–156.

[4] Bojanov, B.: Characterization of the smoothest interpolant. *SIAM J. Math. Anal.* **25**, 1994, No 6, 1642–1655.

[5] Bojanov, B., Hakopian, H., Sahakian, A.: *Spline Functions and Multivariate Interpolations*. Mathematics and Its Applications Vol. 248, Kluwer Academic Publishers, Dordrecht, 1993,

[6] de Boor, C.: On "Best" interpolation. *J. Approx. Theory* **16**, 1976, 28–42.

[7] Draganova, C.: Characterization of the smoothest periodic interpolant. *Math. Balkanica (N.S.)* **8**, 1994, 297–310.

[8] Draganova, C.: Smoothest interpolation in the mean. *J. Approx. Theory* **98** (1999), 223–247.

[9] Lorentz, G., Jetter, K., Riemenschneider, S.: *Birkhoff Interpolation*. Encyclopedia of Mathmatics and Its Applications, vol. 19, Addison & Wesley, 1983.

[10] Marin, S.: An approach to data parametrization in parametric cubic spline interpolation problems. *J. Approx. Theory* **41**, 1984, 64–86.

[11] Naidenov, N.: Algorithm for the construction of the smoothest interpolant. *East J. Approx.* **1**, 1995, 83–97.

[12] Pinkus, A.: On smoothest interpolants. *SIAM J. Math. Anal.* **19**, 1988, No 6, 1431–1441.

[13] Pinkus, A.: Uniqueness of the smoothest interpolants. *East J. Approx.* **3**, 1997, 377–380.

[14] Rademacher, C., Scherer, K.: Best parameter interpolation in L_p-norms. In: *Contributions to the Computation of Curves and Surfaces*, Monograf. Acad. Ciencia de Zaragoza, Puerto de la Cruz, 1989, pp. 67–80,

[15] Scherer, K.: Uniqueness of best parametric interpolation by cubic spline curves. *Constr. Approx.* **13**, 1997, 393–419.

[16] Scherer, K., Smith, P.: Existence of best parametric interpolation by curves. *SIAM J. Math. Anal.* **20**, 1989, 160–168.

[17] Schumaker, L. L.: *Spline Functios: Basic Theory*. Wiley Interscience, New York, 1981.

[18] Uluchev, R.: B-splines with Birkhoff knots. Applications in the approximations and shape-preserving interpolation. *Math. Balkanica (N.S.)* **3**, 1989, 225–239.

[19] Uluchev, R.: *Problems on Optimal Interpolation*, PhD Thesis, Sofia University, Sofia, 1990. (In Bulgarian)

[20] Uluchev, R.: Smoothest interpolation with free nodes in W_p^r, In: *Progress in Approximation Theory* (P. Nevai and A. Pinkus, Eds.), Academic Press, San Diego, 1991, pp. 787–806 (Special volume of *J. Approx. Theory*).

Received on April 18, 2020

Velina Ivanova, Rumen Uluchev

Faculty of Mathematics and Informatics
"St. Kliment Ohridski" University of Sofia
5 James Bourchier Blvd.
BG 1164 Sofia
BULGARIA

E-mails: vetotiv@abv.bg , rumenu@fmi.uni-sofia.bg

Submission of manuscripts. The *Annual* is published once a year. No deadline exists. Once received by the editors, the manuscript will be subjected to rapid, but thorough review process. If accepted, it is immediately scheduled for the nearest forthcoming issue. No page charge is made. The author(s) will be provided with a free of charge printable pdf file of their published paper.

The submission of a paper implies that it has not been published, or is not under consideration for publication elsewhere. In case it is accepted, it implies as well that the author(s) transfers the copyright to the Faculty of Mathematics and Informatics at the "St. Kliment Ohridski" University of Sofia, including the right to adapt the article for use in conjunction with computer systems and programs and also reproduction or publication in machine-readable form and incorporation in retrieval systems.

Instructions to Contributors. Preferences will be given to papers, not longer than 25 pages, written in English and typeset by means of a TEX system. A simple specimen file, exposing in detail the instruction for preparation of the manuscripts, is available upon request from the electronic address of the Editorial Board.

Manuscripts should be submitted for editorial consideration in pdf-format by e-mail to `annuaire@fmi.uni-sofia.bg`. Upon acceptance of the paper, the authors will be asked to send the text of the papers in `.tex` format and the appropriate graphic files (preferably in `.eps` format).

The manuscripts should be prepared in accordance with the instructions, given below.

The first page of manuscripts must contain a title, name(s) of the author(s), a short abstract, a list of keywords and the appropriate 2010 MSC codes (primary and secondary, if necessary). The affiliation(s), including the electronic address, should be given at the end of the manuscripts.

Figures have to be inserted in the text near their first reference. If the author cannot supply and/or incorporate the graphic files, drawings (in black ink and on a good quality paper) should be enclosed separately. If photographs are to be used, only black and white ones are acceptable.

Tables should be inserted in the text as close to the point of reference as possible. Some space should be left above and below the table.

Footnotes, which should be kept to a minimum and should be brief, must be numbered consecutively.

References must be cited in the text in square brackets, like [3], or [5, 7], or [11, p. 123], or [16, Ch. 2.12]. They have to be numbered either in the order they appear in the text or alphabetically. Examples (please note order, style and punctuation):

For books: Obreshkoff, N.: *Higher Algebra.* Nauka i Izkustvo, Second edition, Sofia, 1963 (in Bulgarian).

For journal articles: Frisch, H. L.: Statistics of random media. *Trans. Soc. Rheology*, **9**, 1965, 293–312.

For articles in edited volumes or proceedings: Friedman, H. Axiomatic recursive function theory. In: *Logic Colloquim 95*,(R. Gandy and F. Yates, eds.), North-Holland, 1971, 188–195.