Contents

ROMAI J., **4**, 1(2008), 1–25

ALGORITHMIC TECHNIQUES FOR SEVERAL OPTIMIZATION PROBLEMS REGARDING DISTRIBUTED SYSTEMS WITH TREE TOPOLOGIES

Mugurel Ionuţ Andreica

Computer Science Department, Politehnica University of Bucharest

mugurel.andreica@cs.pub.ro

Abstract In this paper we present novel algorithmic techniques for solving several optimization problems regarding distributed systems with tree topologies. I address topics like: reliability improvement, partitioning, coloring, content delivery, optimal matchings, as well as some tree counting aspects. Some of the presented techniques are only of theoretical interest, while others can be used in practical settings.

Keywords: distributed systems, tree topologies, optimization.

2000 MSC: 05C05, 05C15, 05C35.

1. INTRODUCTION

Distributed systems are being increasingly developed and deployed all around the world, because they present efficient solutions to many practical problems. However, as their development progresses, many problems related to scalability, fault tolerance, stability, efficient resource usage and many other topics need to be solved. Developing efficient distributed systems is not an easy task, because many system parameters need to be fine tuned and optimized. Because of this, optimization techniques are required for designing efficient distributed systems or improving the performance of existing, already deployed ones. In this paper I present several novel algorithmic techniques for some optimization problems regarding distributed systems with a tree topology.

Trees are some of the simplest non-trivial topologies which appear in real-life situations. Many of the existing networks have a hierarchical structure (a tree or tree-like graph), with user devices at the edge of the network and router backbones at its core. Some peer-to-peer systems used for content retrieval and indexing have a tree structure. Multicast content is usually delivered using multicast trees. Furthermore, many graph topologies can be reduced to tree topologies, by choosing a spanning tree or by covering the graph's edges with edge disjoint spanning trees [1]. In a tree, there exists a unique path between any two nodes. Thus, the network is quite fragile. The fragility is compensated by the simplicity of the topology, which makes many decisions become easier.

This paper is structured as follows. Section 2 defines the main notations which are used in the rest of the paper. In Section 3 we consider the minimum weight cycle completion problem in trees. In Section 4 I discuss two tree partitioning problems and in Section 5 we consider two content delivery optimization problems. In Section 6 we solve several optimal matching problems in trees and powers of trees and in Section 7 we analyze the *first fit online coloring* heuristic, applied to trees. In Section 8 we consider three other optimization and tree counting problems. In Section 9 we discuss related work and in Section 10 we conclude and present future work.

2. NOTATIONS

A tree is an undirected, connected, acyclic graph. A tree may be rooted, in which case a special vertex r will be called its root. Even if the tree is unrooted, we may choose to root it at some vertex. In a rooted tree, we define $parent(i)$ as the parent of vertex i and $ns(i)$ as the number of sons of vertex i. For a leaf vertex i, $ns(i) = 0$ and for the root r, $parent(r)$ is undefined. The sons of a vertex i are denoted by $s(i, j)$ $(1 \leq j \leq ns(i))$. A vertex j is a *descendant* of vertex i if $(parent(j) = i)$ or $parent(j)$ is also a descendant of vertex i. We denote by $T(i)$ the subtree rooted at vertex i, i.e. the part of the tree composed of vertex i and all of its descendants (together with the edges connecting them). In the paper, the terms *node* and *vertex* will be used with the same meaning.

A matching M of a graph G is a set of edges of the graph, such that any two edges in the set have distinct endpoints (vertices). A maximum matching is a matching with maximum cardinality (maximum number of edges).

3. MINIMUM WEIGHT CYCLE COMPLETION OF A TREE

We consider a tree network with n vertices. For m pairs of vertices (i, j) which are not adjacent in the tree, we are given a weight $w(i, j)$ (we can consider $w(i, j) = +\infty$ for the other pairs of vertices). We want to connect some of these m pairs (i.e. add extra edges to the tree), such that, in the end, every vertex of the tree belongs to exactly one cycle. The objective consists of minimizing the total weight of the edges added to the tree. For the unweighted case ($w(i, j) = 1$) and when we can connect any pair of vertices which is not connected by a tree edge, there exists the following simple greedy algorithm [3]. We select an arbitrary root vertex r and then traverse the tree bottom-up (from the leaves towards the root). For each vertex i we compute a value $l(i)$, representing the largest number of vertices on a path $P(i)$ starting at i and continuing in $T(i)$, such that every vertex $j \in (T(i) \setminus P(i))$ belongs to exactly one cycle and the vertices in $P(i)$ are the only ones who do not belong to a cycle. We denote by $e(i)$ the second endpoint of the path (the first one being vertex i). For a leaf vertex i, we have $l(i) = 1$ and $e(i) = i$. For a non-leaf vertex i, we first remove from its list of sons the sons $s(i, j)$ with $l(s(i, j)) = 0$, update $ns(i)$ and renumber the other sons starting from 1. If i remains with only one son, we set $l(i) = l(s(i, 1)) + 1$ and $e(i) = e(s(i, 1))$. If i remains with $ns(i) > 1$ sons, we sort them according to the values $l(s(i, j))$, such that $l(s(i, 1)) \leq l(s(i, 2)) \leq \ldots \leq l(s(i, ns(i)))$. We connect by an edge the vertices $e(s(i, 1))$ and $e(s(i, 2))$. This way, every vertex on the paths $P(s(i, 1))$ and $P(s(i, 2))$, plus the vertex i, belong to exactly one cycle. For the other sons $s(i, j)$ $(3 \leq j \leq ns(i))$, we have to connect $s(i, j)$ to $e(s(i, j))$. This will only be possible if $l(s(i, j)) \geq 3$; otherwise, the tree admits no solution. Afterwards, we set $l(i) = 0$. If the root r has only one son, then we must have $l(r) \geq 3$, such that we can connect r to $e(r)$.

For the general case, we describe a dynamic programming algorithm (as the greedy algorithm cannot be extended to this case). We again root the tree at an arbitrary vertex r, thus defining parent-son relationships. For each vertex i, we compute two values: $wA(i)$=the minimum total weight of a subset of edges added to the tree such that every vertex in $T(i)$ belongs to exactly one cycle, and $wB(i)$=the minimum total weight of a subset of edges added to the tree such that every vertex in $(T(i) \setminus \{i\})$ belongs to exactly one cycle (and vertex i belongs to no cycle). We compute the values from the leaves towards the root. For a leaf vertex i, we have $wA(i) = +\infty$ and $wB(i) = 0$. For a non-leaf vertex i, we have: $wB(i) = \sum_{j=1}^{ns(i)} wA(s(i,j))$. In order to compute $wA(i)$ we first traverse $T(i)$ and for each vertex j, we compute $wAsum(i,j)$=the sum of all the $wA(p)$ values, where p is a son of a vertex q which is located on the path from i to j ($P(i \ldots j)$) and p does not belong to $P(i \ldots j)$. We have $wAsum(i,i) = wB(i)$ and for the other vertices j we have $wAsum(i,j) = wAsum(i, parent(j)) - wA(j) + wB(j)$. Now we try to add an edge, such that it closes a cycle in the tree which contains vertex i. We first try to add edges of the form (i,j), where j is a descendant of i (but not a son of i, of course) - these will be called *type 1* edges. Adding such an edge (i,j) provides a candidate value $wcand(i,i,j)$ for $wA(i)$: $wcand(i,i,j) = wAsum(i,j) + w(i,j)$. We then consider edges of the form (p,q) ($p \neq i$ and $q \neq i$), where the lowest common ancestor of p and q ($LCA(p,q)$) is vertex i - these will be called *type 2* edges (we consider every pair of distinct sons $s(i,a)$ and $s(i,b)$, and for each such pair we consider every pair of vertices $p \in T(s(i,a))$ and $q \in T(s(i,b))$ and verify if the edge (p,q) can be added to the tree). Adding such an edge (p,q) provides a candidate value $wcand(i,p,q)$ for $wA(i)$: $wcand(i,p,q)$=$wAsum(i,p)$+$wAsum(i,q)$-$wB(i)$+$w(p,q)$. $wA(i)$ will be equal to the minimum of the candidate values $wcand(i,*,*)$ (or to $+\infty$ if no candidate value exists). We can implement the algorithm in $O(n^2)$ time, which is optimal in a sense, because $m \leq (n \cdot (n-1)/2 - n + 1)$, which is $O(n^2)$. $wA(r)$ is the answer to our problem and we can find the actual edges to add to the tree by tracing back the way the $wA(*)$ and $wB(*)$ values were computed.

However, when the number m of edges which can be added to the tree is significantly smaller, we can improve the time complexity to $O((n+m)\cdot log(n))$. We compute for each of the m edges (i,j) the lowest common ancestor of the vertices i and j $(LCA(i,j))$ in the rooted tree. This can be achieved by preprocessing the tree in $O(n)$ time and then answering each LCA query in $O(1)$ time [2]. If $LCA(i,j)=k$, then we add the edge (i,j) to a list $Ledge(k)$. Then, for each non-leaf vertex i, we traverse the edges in $Ledge(k)$. For each edge (p,q) we can easily determine if it is of type 1 $(i=p$ or $i=q)$ or of type 2 and use the corresponding equation. However, we need the values $wAsum(i,p)$ and $wAsum(i,q)$. Instead of recomputing these values from scratch, we update them incrementally. It is obvious that $wAsum(parent(i),p)$=$wAsum(i,p)$+$wB(parent(i))$-$wA(i)$. We preprocess the tree, by assigning to each vertex i its DFS number $DFSnum(i)$ ($DFSnum(i)$=j if vertex i was the j^{th} distinct vertex visited during a DFS traversal of the tree which started at the root). Then, for each vertex i, we compute $DFSmax(i)$=the maximum DFS number of a vertex in its subtree. For a leaf node i, we have $DFSmax(i)=DFSnum(i)$. For a non-leaf vertex i, $DFSmax(i)$=$max\{DFSnum(i), DFSmax(s(i,1)),\dots,DFSmax(s(i,ns(i)))\}$. We maintain a segment tree, using the algorithmic framework from [15]. The operations we use are range addition update and point query. Initially, each leaf i $(1\leq i\leq n)$ has a value $v(i)=0$. Before computing $wA(i)$ for a vertex i, we set the value of leaf $DFSnum(i)$ in the segment tree to $wB(i)$. Then, for each son $s(i,j)$, we add the value $(wB(i)-wA(s(i,j))$ to the interval $[DFSnum(s(i,j)),DFSmax(s(i,j))]$ (range update). We can obtain $wAsum(i,p)$ for any vertex $p\in T(i)$ by querying the value of the cell $DFSnum(p)$ in the segment tree: we start from the (current) value of the leaf $DFSnum(p)$ and add the update aggregates *uagg* stored at every anestor node of the leaf in the segment tree. Queries and updates take $O(log(n))$ time each.

If the objective is to minimize the largest weight W_{max} of an edge added to the tree, we can binary search W_{max} and perform the following feasibility test on the values W_{cand} chosen by the binary search: we consider only the "extra" edges (i, j) with $w(i, j) \leq W_{cand}$ and run the algorithm described above for these edges; if $wA(r) \neq +\infty$, then W_{cand} is feasible.

4. TREE PARTITIONING TECHNIQUES

4.1. TREE PARTITIONING WITH LOWER AND UPPER SIZE BOUNDS

Given a tree with n vertices, we want to partition the tree into several parts, such that the number of vertices in each part is at least Q and at most $k \cdot Q$ $(k \geq 1)$. Each part P must have a representative vertex u, which does not necessarily belong to P. However, $(P \cup \{u\})$ must form a connected subtree. We present an algorithm which works for $k \geq 3$. We root the tree at any vertex r, traverse the tree bottom-up and compute the parts in a greedy manner. For each vertex i we compute $w(i)$=the size of a connected component $C(i)$ in $T(i)$, such that vertex $i \in C(i)$, $|C(i)| < Q$, and all the vertices in $(T(i) \setminus C(i))$ were split into parts satisfying the specified properties. For a leaf vertex i, $w(i) = 1$ and $C(i) = \{i\}$. For a non-leaf vertex i, we traverse its sons (in any order) and maintain a counter $ws(i)$=the sum of the $w(s(i, j))$ values of the sons traversed so far. If $ws(i)$ exceeds $Q - 1$ after considering the son $s(i, j)$, we form a new part from the connected components $C(s(i, last\_son + 1)), \ldots, C(s(i, j))$ and assign vertex i as its representative. Then, we reset $ws(i)$ to 0. $(last\_son < j)$ is the previous son where $ws(i)$ was reset to 0 (or 0, if $ws(i)$ was never reset to 0).

After considering every son of vertex i, we set $w(i) = ws(i) + 1$ and the component $C(i)$ is formed from the components $C(s(i, j))$ which were not used for forming a new part, plus vertex i. If $ws(i) + 1 = Q$, then we form a new part from the component $C(i)$ and set $w(i) = 0$ and $C(i) = \{\}$. During the algorithm, the maximum size of any part formed is $2 \cdot Q - 2$. At the end of the algorithm, we may have that $w(r) > 0$. In this case, the vertices in $C(r)$ were not assigned to any part. However, at least one vertex from $C(r)$ is adjacent to a vertex assigned to some part P. Then, we can extend that part P in order to contain the vertices in $C(r)$. This way, the maximum size of a part becomes $3 \cdot Q - 3$. The pseudocode of the first part of the algorithm is presented below. In order to compute the parts, we maintain for each vertex i a value $part(i)$, which is 0, initially (0 means that the vertex was not assigned to any part). In order to assign distinct part numbers, we maintain a global counter $part\_number$, whose initial value is 0. The first part of the algorithm has linear time complexity ($O(n)$). The second part (adding $C(r)$ to an already existing part) can also be performed in linear time, by searching for an edge (p, q), such that $part(p) = 0$ and $part(q) > 0$ (there are only $n - 1 = O(n)$ edges in a tree).

```
LowerUpperBoundTreePartitioning(Q, i):
if (ns(i)=0) then w(i)=1 else
 ws(i)=last_son=0
 for j=1 to ns(i) do // j=1,2,...,ns(i)
  LowerUpperBoundTreePartitioning(Q, s(i,j))
  ws(i)=ws(i)+w(s(i,j))
  if ($ws(i) \geq Q$) then
   part_number=part_number + 1; last_son=j; ws(i)=0
   for k=last_son+1 to j do AssignPartNumber(s(i,k), part_number)
 w(i)=ws(i)+1
 if ($w(i) \geq Q$) then
  part_number=part_number + 1; w(i)=0
  AssignPartNumber(i, part_number)

AssignPartNumber(i, part_number):
if ($part(i) \neq 0$) then return()
part(i)=part_number
for j=1 to ns(i) do AssignPartNumber(s(i,j), part_number)
```

4.2. CONNECTED TREE PARTITIONING

We now present an efficient algorithm for identifying k connected parts of given sizes in a tree (if possible), subject to minimizing the total cost. Thus, given a tree with n vertices, we want to find k vertex-disjoint components (called parts), such that the i^{th} part ($1 \leq i \leq k$) has $sz(i)$ vertices ($sz(1) + sz(2) + \ldots + sz(k) \leq n$ and $sz(i) \leq sz(i+1)$ for $1 \leq i \leq k-1$). Each tree edge

(i,j) has a cost $ce(i,j)$ and each tree vertex i has a cost $cv(i)$. We want to minimize the sum of the costs of the vertices and edges which do not belong to any part. An edge (i,j) belongs to a part p if both vertices i and j belong to part p.

In order to obtain k connected components of the given sizes we need to keep $Q-k$ edges of the tree and remove the others, where $Q = sz(1) + \ldots + sz(n)$. We could try all the *(*$(n-1)$ *choose* $(Q-k)$*)* possibilities of choosing $Q-k$ edges out of the $n-1$ edges of the tree. For each possibility, we obtain $k' = n-Q+k$ connected components with sizes $sz'(1) \leq sz'(2) \leq \ldots \leq sz'(k')$; in case of several components with equal sizes, we sort them in increasing order of the total cost of the vertices in them. Then, we must have $sz(j) = sz'(k'-k+j)$ and the total cost of the possibility is the sum of the costs of the removed edges plus the sum of the costs of the vertices in the components $1, 2, \ldots, k'-k$ (which should have only one vertex each, if the size conditions hold). However, this approach is quite inefficient in most cases. We present an algorithm with time complexity $O(n^3 \cdot 3^k)$. We root the tree at an arbitrary vertex r. Then, we compute a table $Cmin(i,j,S)$=the minimum cost of obtaining from $T(i)$ the parts with indices in the set S and, besides them, we are left with a connected component consisting of j vertices which includes vertex i and, possibly, several vertices which are ignored (if $j = 0$, then every vertex in $T(i)$ is assigned to one of the parts in S or is ignored). We compute this table bottom-up:

```
ConnectedTreePartitioning(i):
for each $S \subseteq \{1,2,\ldots,k\}$ do for $j$=0 to $n$ do $Cmin(i,j,S)=+\infty$
$Cmin(i, 1, \{\})=0$; $Cmin(i, 0, \{\})=cv(i)$
for $x$=1 to $ns(i)$ do
  ConnectedTreePartitioning($s(i,x)$)
  for each $S \subseteq \{1,2,\ldots,k\}$ do for $j$=0 to $n$ do
    $Caux(i,j,S)=Cmin(i,j,S)$; $Cmin(i,j,S)=+\infty$
  for each $S \subseteq \{1,2,\ldots,k\}$ do for $j$=0 to $n$ do
    for each $W \subseteq S$ do for $q$=0 to $qlimit(j)$ do
      $Cmin(i,j,S)=min\{Cmin(i,j,S),\ Caux(i,j-q,S \setminus W) + extra\_cost(i,s(i,x),q) + Cmin(s(i,x),q,W)\}$
```

```
for each $S \subseteq \{1, 2, \ldots, k\}$ do
 for $j$=0 to $n$ do if ($Cmin(i, j, S) < +\infty$) then
  for $q$=1 to $k$ do if (($j$=$sz(q)$) and ($q \notin S$)) then
   $Cmin(i,0,S \cup \{q\})=min\{Cmin(i,j,S), Cmin(i,0,S \cup \{q\})\}$
```

We define *extra_cost(i, son_x_i, q)=if* $(q > 0)$ *then return(0) else return(ce(i, son_x_i))* and *qlimit(j)=max{j-1,0}*. The algorithm computes *Cmin(i,*,*)* from the values of vertex i's sons, using the principles of tree knapsack. The total amount of computations for each vertex is $O(ns(i) \cdot 3^k \cdot n^2)$. Summing over all the vertices, we obtain $O(n^3 \cdot 3^k)$. The minimum total cost is $Cmin(r, 0, \{1, 2, \ldots, k\})$ (if this value is $+\infty$, then we cannot obtain k parts with the given sizes). In order to find the actual parts, we need to trace back the way the $Cmin(*, *, *)$ values were computed, which is a standard procedure. When the sum of the sizes of the k parts is n, then every vertex belongs to one part.

5. CONTENT DELIVERY OPTIMIZATION PROBLEMS

5.1. MINIMUM NUMBER OF UNICAST STREAMS

Consider a directed acyclic graph G with n vertices and m edges. Every directed edge (u, v) has a lower bound $lbe_G(u, v)$, an upper bound $ube_G(u, v)$ and a cost $ce_G(u, v)$. Every vertex u has a lower bound $lbv_G(u)$, an upper bound $ubv_G(u)$ and a cost $cv_G(u)$. We need to determine the minimum number of unicast communication streams p and a path for each of the p streams, such that the number of stream paths $npe(u, v)$ containing an edge (u, v) satisfies $lbe_G(u, v) \leq npe(u, v) \leq ube_G(u, v)$ and the number of paths $npv(u)$ containing a vertex u satisfies $lbv_G(u) \leq npv(u) \leq ubv_G(u)$. Each vertex u can be a *source* node, a *destination* node, both or none. A stream path may start at any source node and finish at any destination node. Moreover, for the number of streams p, we want to compute the paths such that the sum S over all the values $(npe(u, v) - lbe_G(u, v)) \cdot ce_G(u, v)$ and $(npv(u) - lbv_G(u)) \cdot cv_G(u)$ is minimum.

Particular cases of this problem have been studied previously. When $lbv_G(u)$ *=1* and $ubv_G(u) = 1$ for every vertex u, $lbe_G(u, v) = 0$ and $ube_G(u, v) = +\infty$ for every directed edge (u, v), all the costs are 0, and every vertex is a source and destination node, we obtain the *minimum path cover* problem in directed acyclic graphs, which is solved as follows [18]. Construct a bipartite graph B with n vertices $x_1, \ldots, x_n$ on the left side and n vertices $y_1, \ldots, y_n$ on the right side. We add an edge (x_i, y_j) in B if the directed edge (i, j) appears in G. Then, we compute a maximum matching in B. If the cardinality of this matching is C, then we need $p = n - C$ streams. The paths are computed as follows. Having an edge (x_i, y_j) in the maximum matching means that the edge (i, j) in G belongs to some stream path. If two edges (x_i, y_j) and (x_j, y_k) in B belong to the matching, then the edges (i, j) and (j, k) in G belong to the path of the same stream. For non-zero costs, we compute a minimum (total) weight matching in B (where every edge (x_i, y_j) has a weight equal to $ce(i, j)$).

In order to solve the problem we mentioned, we use a standard transformation and construct a new graph G', where every vertex u is represented by two vertices u_{in} and u_{out}. For every directed edge (u, v) in G, we add an edge (u_{out}, v_{in}) in G', with the same cost and lower and upper bounds. We also add a directed edge from u_{in} to u_{out} in G' (for every vertex u in G), with cost $cv_G(u)$, lower bound $lbv_G(u)$ and upper bound $ubv_G(u)$. Then we add two special vertices s (source) and t (sink) to G'. For every source node u in G, we add a directed edge (s, u_{in}) in G', with lower bound and cost 0 and upper bound $+\infty$. For every destination node v in G, we add a directed edge (v_{out}, t), with lower bound and cost 0 and upper bound $+\infty$. We also add the edges (s, t) and (t, s) with lower bound and cost 0 and upper bound $+\infty$. The resulting graph G' has costs, lower and upper bounds only on its edges and not on its vertices. In order to compute the minimum number of communication streams which satisfy the constraints imposed by G, it is enough to compute a (minimum cost) minimum feasible flow in G', from s to t. Decomposing the flow into unit-flow paths (in order to obtain the path of each communication stream) can then be done easily. We repeatedly perform a graph traversal (DFS or BFS) from s to t in G', considering only directed edges with positive flow on them. From the traversal tree, by following the "parent" pointers, we can find a path P from s to t, containing only edges with positive flow. We compute the minimum flow fP on any edge of P, transform P into fP unit paths and then decrease the flow on the edges in P by fP. If we remove the first and last vertices on any unit path (i.e. s and t), we obtain a path from a vertex u_{in} to a vertex v_{out}, where u is a source node in G and v is a destination node in G. We use the algorithm presented in [18] for determining a feasible flow (not necessarily minimum) in a flow network with lower and upper bounds on its edges. We denote this algorithm by $A(F, s, t)$ (F is the flow network given as argument, s is the source vertex and t is the sink vertex). I will describe $A(F, s, t)$ briefly. We construct a new graph F' from F, as follows. We maintain all the vertices and edges in F. For every directed edge (u, v) in F, the directed edge (u, v) in F' has the same cost, lower bound 0 and upper bound $(ube_F(u, v) - lbe_F(u, v))$. We add two extra vertices s' and t' and the following zero-cost directed edges: (s', u) and (u, t') for every vertex u in F (including s and t). The lower bound of every edge will be 0. The upper bound of a directed edge (s', u) in F' is equal to the sum of the lower bounds of the directed edges $(*, u)$ in F. The upper bound of every directed edge (u, t') in F' is equal to the sum of the lower bounds of the directed edges $(u, *)$ in F. The algorithm $A(F, s, t)$ computes a minimum cost maximum flow g in the graph F' (which, as stated, only has upper bounds); if all the costs are 0, only a maximum flow is computed. If g is equal to the sum of the upper bounds of the edges $(s', *)$ (or, equivalently, of the edges $(*, t')$), then a feasible flow from s to t exists in F: the flow on every directed edge (u, v) in F will be $lbe_F(u, v)$ plus the flow on the edge (u, v) in F'.

We first run the algorithm on G' (i.e. call $A(G', s, t)$) in order to verify if a feasible flow exists). If no feasible flow exists, then the constraints cannot be satisfied by any number of streams. Otherwise, we construct a graph G'' from G', by adding a new vertex $snew$ and a zero-cost directed edge $(snew, s)$ with lower bound 0 and upper bound x. $snew$ will be the new source vertex and x is a parameter which is used in order to limit the amount of flow entering the old source vertex s. We now perform a binary search on x, between 0 and $gmax$, where $gmax$ is the value of the feasible flow computed by calling $A(G', s, t)$. The feasibility test consists of verifying if there exists a feasible flow in the graph G'' (i.e. calling $A(G'', snew, t)$). The minimum value of x for which a feasible flow exists in G'' is the value of the minimum feasible flow in G', from s to t. Obtaining the feasible flow in G' from the feasible flow in G'' is trivial: for every directed edge (u, v) in G', we set its amount of flow to the flow of the same edge (u, v) in G''. The time complexity of the presented algorithm is $O(MF(n, m) \cdot log(gmax))$, where $gmax$ is a good upper bound on the value of a feasible flow and $MF(n, m)$ is the best time complexity of a (minimum cost) maximum flow algorithm in a directed graph with n vertices and m edges.

5.2. DEGREE-CONSTRAINED MINIMUM SPANNING TREE

In [13], the following problem was considered: given an undirected graph with n verices and m edges, where each edge (i, j) has a weight $w(i, j) > 0$, compute a spanning tree MST of minimum total weight, such that a special vertex r has degree exactly k in MST. A solution was proposed, based on using a parameter d and setting the cost of each edge (r, j) adjacent to r, $c(r, j) = d + w(r, j)$; the cost of the other edges is equal to their weight. Parameter d can range from $-\infty$ to $+\infty$. We denote by $MST(d)$=the minimum spanning tree using the cost functions defined previously. When $d = -\infty$, $MST(d)$ contains the maximum number of edges adjacent to r. For $d = +\infty$, $MST(d)$ contains the minimum number of edges adjacent to r. We define the function $ne(d)$=the number of edges adjacent to r in $MST(d)$. $ne(d)$ is non-increasing on the interval $[-\infty, +\infty]$. We will binary search the smallest value $dopt$ of the parameter d in the interval $[-\infty, +\infty]$, such that $ne(dopt) \leq k$. We finish the binary search when the length of the search interval is smaller than a small constant $\varepsilon > 0$.

If $ne(dopt) = k$, then the edges in $MST(dopt)$ form the required minimum spanning tree. If $ne(dopt) < k$, then $ne(dopt - \varepsilon) > k$. We define $S(d)$=the set of edges adjacent to vertex r in $MST(d)$. It is easy to prove that $S(dopt)$ is included in $S(dopt - \varepsilon)$. The required minimum spanning tree is constructed in the following manner. The edges adjacent to vertex r will be the edges in

$S(dopt)$, to which we add $(k - ne(dopt))$ arbitrary edges from the set $S(dopt - \varepsilon) \setminus S(dopt)$. Once these edges are fixed, we construct the following graph G: we set the cost of the chosen edges to 0 and the cost of the other edges (i, j) to $w(i, j)$. We now compute a minimum spanning tree MST_G in G. The edges in MST_G are the edges of the minimum spanning tree of the original graph, in which vertex r has degree exactly k. The time complexity of this approach is $O(m \cdot log(m) \cdot log(DMAX))$, where $DMAX$ denotes the range over which we search the parameter d. When m is not too large (i.e. m is not of the order $O(n^2)$), this represents an improvement over the $O(n^2)$ solution given in [13].

6. MATCHING PROBLEMS

6.1. MAXIMUM WEIGHT MATCHING IN AN EXTENDED TREE

Let us consider a rooted tree (with vertex r as the root). Each vertex i has a weight $w(i)$. We want to find a matching in the following graph G (extended tree), having the same vertices as T and an edge (x, y) between two vertices x and y, if: *(i)* x and y are adjacent in the tree; *(ii)* x and y have the same parent in the tree. The weight of an edge (x, y) in G is $|w(x) - w(y)|$. The weight of a matching is the sum of the weights of its edges. We are interested in a maximum weight matching in the graph G. For each vertex i, we sort its sons $s(i, 1), \ldots, s(i, ns(i))$ in non-decreasing order of their weights, i.e. $w(s(i, 1)) \leq \ldots \leq w(i, ns(i))$. We compute for each vertex i two values: $A(i)$=the maximum weight of a matching in $T(i)$ if vertex i is the endpoint of an edge in the matching and $B(i)$=the maximum weight of a matching in $T(i)$ if vertex i is not the endpoint of any edge in the matching. In order to compute these values, we will compute the following tables for every vertex i: $CA(i, j, k)$=the maximum weight of a matching in $T(i)$ if vertex i is the endpoint of an edge in the matching and we only consider its sons $s(i, j), s(i, j + 1), \ldots, s(i, k)$ (and their subtrees). Similarly, we have $CB(i, j, k)$, where vertex i does not belong to any edge in the matching. The maximum weight of a matching is $max\{A(r), B(r)\}$. The actual matching can be computed easily, by tracing back the way the $A(i)$, $B(i)$, $CA(i, *, *)$ and

$CB(i, *, *)$ values were computed. A recursive algorithm (called with r as its argument) is given below. The time complexity is $O(ns(i)^2)$ for a vertex i and, thus, $O(n^2)$ overall.

```
MaximumWeightMatching-ExtendedTree(i):
if (ns(i)=0) then A(i)=B(i)=0 else
 for j=1 to ns(i) do MaximumWeightMatching-ExtendedTree(s(i,j))
 for j=1 to ns(i) do
   CA(i, j, j - 1)= $-\infty$; $CA(i, j, j) = |w(i) - w(s(i, j))| + B(s(i, j))$
   CB(i, j, j - 1)= 0; CB(i, j, j)=max{A(s(i,j)), B(s(i,j))}
 for count=1 to (ns(i)-1) do for j=1 to (ns(i)-count) do
   k = j + count
   CA(i,j,k)=max{$|w(s(i, j)) - w(s(i, k))|$ + B(s(i,j)) + B(s(i,k)) + CA(i, j + 1, k - 1), $|w(i) - w(s(i, j))|$ + B(s(i,j)) + CB(i, j+1, k), $|w(i) - w(s(i, k))|$ + B(s(i,k)) + CB(i, j, k-1), max{A(s(i,j)), B(s(i,j))} + CA(i, j+1, k), max{ A(s(i,k)), B(s(i,k))} + CA(i, j, k-1)}
   CB(i,j,k)=max{$|w(s(i, j))-w(s(i, k))|$ + B(s(i,j)) + B(s(i,k)) + CB(i, j + 1, k - 1), max{A(s(i,j)), B(s(i,j))} + CB(i, j+1, k), max{A(s(i,k)), B(s(i,k))} + CB(i, j, k-1)}
 A(i)=CA(i,1,ns(i)); B(i)=CB(i,1,ns(i))
```

6.2. MAXIMUM MATCHING IN THE POWER OF A GRAPH

The k^{th} power G^k ($k \geq 2$) of a graph G is a graph with the same set of vertices as G, where there exists an edge (x, y) between two vertices x and y if the distance between x and y in G is at most k. The distance between two vertices (x, y) in a graph is the minimum number of edges which need to be traversed in order to reach vertex y, starting from vertex x. A maximum matching in G^k of a graph G can be found by restricting our attention to a spanning tree T of G. The following linear algorithm (called with $i = r$), using observations from [12], solves the problem (we consider that, initially, no vertex is matched):

```
MaximumMatchingGk(i):
 if (ns(i)=0) then return() else
  last_son=0
  for j=1 to ns(i) do // j=1,2,...,ns(i)
   MaximumMatchingGk(s(i,j))
   if (not matched(s(i,j))) then
    if (last_son = 0) then last_son = s(i,j) else
     add edge (last_son, s(i,j)) to the matching
     matched(last_son) = matched(s(i,j)) = true; last_son = 0
  if (last_son > 0) then
   add edge (i, last_son) to the matching
   matched(i) = matched(last_son) = true
```

7. FIRST FIT ONLINE TREE COLORING

A very intuitive algorithm for coloring a graph with n vertices is the *first-fit online coloring heuristic*. We traverse the vertices in some order $v(1), v(2), \ldots,$ $v(n)$. We assign color 1 to $v(1)$ and for $i = 2, \ldots, n$, we assign to $v(i)$ the minimum color $c(i) \geq 1$ which was not assigned to any of its neighbours $v(j)$ $(j < i)$.

A tree is *2-colorable*: we root the tree at any vertex r and then compute for each vertex i its level in the tree (distance from the root); we assign the color 1 to the vertices on even levels and the color 2 to those on odd levels. However, in some situations, we might be forced to process the vertices in a given order. In this case, it would be useful to compute the worst-case coloring that can be obtained by this heuristic, i.e. the largest number of colors that are used, under the worst-case ordering of the tree vertices (*Grundy number*). I will present an $O(n \cdot log(log(n)))$ algorithm for this problem, similar in nature to the linear algorithm presented in [4]. For each vertex i, we will compute $cmax(i)$=the largest color the can be assigned to vertex i in the worst-case, if vertex i is the last vertex to be colored. The value $max\{cmax(i)|1 \leq i \leq n\}$ is the largest number of colors that can be assigned by the first fit online coloring heuristic.

We root the tree at an arbitrary vertex r. The algorithm consists of two stages. In the first stage, the tree is traversed bottom-up and for each vertex i we compute $c(1,i)$=the largest color that can be assigned to vertex i, considering only the tree $T(i)$. For a leaf vertex i, we have $c(1,i) = 1$. For a non-leaf vertex i, we will sort its sons $s(i,1), \ldots, s(i, ns(i))$, such that $c(1, s(i,1)) \leq c(1, s(i,2)) \leq \ldots \leq c(1, s(i, ns(i)))$. We will initialize $c(1,i)$ to 1 and then consider the sons in the sorted order. When we reach son $s(i,j)$, we compare $c(1, s(i,j))$ with $c(1,i)$. If $c(1, s(i,j)) \geq c(1,i)$, then we increment $c(1,i)$ by 1 (otherwise, $c(1,i)$ stays the same). The justification of this algorithm is the following: if a vertex i can be assigned color $c(1,i)$ in some ordering of the vertices in $T(i)$, then there exists an ordering in which it can be assigned any other color c', such that $1 \leq c' \leq c(1,i)$. Then, when traversing the sons and reaching a son $s(i,j)$ with $c(1, s(i,j)) \geq c(1,i)$, we consider an ordering of the vertices in $T(s(i,j))$, where the color of vertex $s(i,j)$ is $c(1,i)$; thus, we can increase the maximum color that can be assigned to vertex i.

After the bottom-up tree traversal, we have $cmax(r) = c(1,r)$, but we still have to compute the values $cmax(i)$ for the other vertices of the tree. We could do that by rooting the tree at every vertex i and running the previously described algorithm, but this would take $O(n^2 \cdot log(log(n)))$ time. However, we can compute these values faster, by traversing the tree vertices in a top-down manner (considering the tree rooted at r). For each vertex i, we will compute $colmax(parent(i), i)$=the maximum color that can be assigned to $parent(i)$ if we remove $T(i)$ from the tree and afterwards we consider $parent(i)$ to be the (new) root of the tree. We will use the values $c(2,i)$ as temporary storage variables. $c(2,i)$ is initialized to $c(1,i)$, for every vertex i. When computing $cmax(i)$, we consider that vertex i is the root of the tree. Assume that we computed the value $cmax(i)$ of a vertex i and now we want to compute the value $cmax(j)$ of a vertex j which is a son of vertex i. We remove j from the list of sons of vertex i and add $parent(i)$ to this list ($parent(i)$=vertex i's parent in the tree rooted at the initial vertex r). We now need to lift vertex j above vertex i and make j the new root of the tree. In order to do this, we recompute the value $c(2,i)$, which is computed similarly to $c(1,i)$, except that we consider the new list of sons for vertex i (and their $c(2,*)$ values).

Afterwards, we add vertex i to the list of sons of vertex j. We compute the value $cmax(j)$ similarly to the value $c(1,j)$, using the values $c(2,*)$ of vertex j's sons (instead of the $c(1,*)$ values of the sons). After computing $cmax(j)$ we restore the lists of sons of vertices i and j to their original states (as if the tree were rooted at the initial vertex r). After computing the values $cmax(u)$ of all the descendants u of a vertex j, we reset the value $c(2,j)$ to $c(1,j)$.

Both traversals take $O(n \cdot log(n))$ time, if we sort the $ns(i)$ sons of every vertex i in $O(ns(i) \cdot log(ns(i)))$ time. However, it has been proved in [4] that the minimum number of vertices of a tree with the Grundy number q is 2^{q-1}, which is the binomial tree $B(q-1)$. The binomial tree $B(0)$ consists of only one vertex. The binomial tree $B(k \geq 1)$ has a root vertex with k neighbors; the i^{th} of these neighbors $(0 \leq i \leq k-1)$ is the root of a $B(i)$ binomial tree. Thus, every value $c(1,*)$, $c(2,*)$ and $cmax(*)$ can be represented using $O(log(log(n)))$ bits. We can use radix-sort and obtain an $O(n \cdot log(log(n)))$ time complexity. The pseudocode of the functions is given below. The main algorithm consists of calling *FirstFit-BottomUp(r)*, initializing the $c(2,*)$ values to the $c(1,*)$ values, setting $cmax(r) = c(1,r)$ and then calling *FirstFit-TopDown(r)*

```
Compute(i, idx):
sort the sons of vertex i, such that c(idx,s(i,1))≤ ...≤c(idx,s(i,ns(i)))
c(idx,i)=1
for j=1 to ns(i) do if (c(idx,s(i,j))≥c(idx,i)) then c(idx,i)=c(idx,i)+1

FirstFit-BottomUp(i):
for j=1 to ns(i) do FirstFit-BottomUp(s(i,j))
Compute(i, 1)

FirstFit-TopDown(i):
if (i ≠ r) then
  remove vertex i from the list of sons of parent(i)
  add parent(parent(i)) to the list of sons of parent(i) (if parent(i) ≠ r)
  Compute(parent(i),2); colmax(parent(i),i)=c(2,parent(i))
  add parent(i) to the list of sons of vertex i
  Compute(i,2); cmax(i)=c(2,i)
  restore the original lists of sons of the vertices parent(i) and i
```

```
for j=1 to ns(i) do FirstFit-TopDown(s(i,j))
c(2,i)=c(1,i)
```

8. OTHER OPTIMIZATION AND COUNTING PROBLEMS

8.1. BUILDING A (CONSTRAINED) TREE WITH MINIMUM HEIGHT

In this subsection I consider the following optimization problem: We are given a sequence of n leaves and each leaf i ($1 \leq i \leq n$) has a height $h(i)$. We want to construct a (strict) binary tree with $n-1$ internal nodes, such that, in an inorder traversal of the tree, we encounter the n leaves in the given order. The height of an internal node i is $h(i) = 1{+}max\{h(leftson(i), h(rightson(i))\}$ (the height of the leaves is given). We are interested in computing a tree whose root has minimum height. A straight-forward dynamic programming solution is the following: compute $Hmin(i,j)$=the minimum height of a tree containing the leaves i, $i+1$, …, j. We have: $Hmin(i,j)$=$1 + min_{i \leq k \leq j-1}$ $max\{Hmin(i,k)$, $Hmin(k+1,j)\}$. $Hmin(1,n)$ is the answer to our problem. However, the time complexity of this algorithm is $O(n^3)$, which is unsatisfactory. An optimal, linear-time algorithm was given in [14]. The main idea of this algorithm is the following. We traverse the leaves from left to right and maintain information about the rightmost path of the optimal tree for the first i leaves. Then, we can add the $(i+1)^{st}$ leaf by modifying the rightmost path of the optimal tree for the first i leaves. Let's assume that we processed the first i leaves and the optimal tree for these leaves contains, on its rightmost path, the vertices $v(1)$, $v(2)$, …, $v(nv(i))$, in order, from the root to the rightmost leaf ($v(1)$ is the root). Let's assume that the heights of the subtrees rooted at these vertices are $hv(1)$, …, $hv(nv(i))$. It is easy to build this tree for $i = 1$ and $i = 2$ (it is unique). When adding the $(i+1)^{st}$ leaf, we traverse the rightmost path from $nv(i)$ down to 2. Assume that we are considering the vertex $v(j)$. If $hv(j-1) < (2 + max\{hv(j), h(i+1)\})$, then we disconsider the vertex $v(j)$ from the rightmost path and move to the next vertex $(v(j-1))$. Let's assume that the path now contains the vertices $v(1)$, …, $v(nv'(i))$. We replace vertex $v(nv'(i))$ by a new vertex $vnew$, whose left son will be $v(nv'(i))$ (together with its subtree) and whose right son will be the $(i+1)^{st}$ leaf. The height of the new vertex will be $1 + max\{hv(nv'(i)), h(i+1)\}$. The rightmost path of the optimal tree behaves like a stack and, thus, the overall time complexity is linear.

We present a sub-optimal $O(n \cdot log(n))$ time algorithm which is interesting on its own. The algorithm is similar to Huffman's algorithm for computing optimal prefix-free codes, except that it maintains the order of the leaves. A suggestion that such an approach might work was given to me by C. Gheorghe. At step i $(1 \leq i \leq n-1)$ of the algorithm, we have $n-i+1$ subtrees of the optimal tree. Each subtree j contains an interval of leaves $[leftleaf(j), rightleaf(j)]$ and its height is $h(j)$. We combine the two adjacent subtrees j and $j+1$ whose combined height *(1+max{height(subtree j),height(subtree j+1)})* is minimum among all the $O(n)$ pairs of adjacent subtrees. At the first step, the n subtrees are represented by the n leaves, whose heights are given. A straight-forward implementation of this idea leads to an $O(n^2)$ algorithm. However, the processing time can be improved by using two segment trees [15], A and B, with n and $n-1$ leaves, respectively. Each node q of a segment tree corresponds to an interval of leaves $[left(q), right(q)]$ (leaves are numbered starting from 1). Each leaf node of the segment tree A can be in the *active* or *inactive* state. Each node q of A (whether leaf or internal node) maintains a value $nactive(q)$, denoting the number of active leaves in its subtree. Initially, each of the n leaves of A is active and the $nactive(*)$ values are initialized appropriately, in a bottom-up manner (1, for a leaf node, and $nactive(leftson(q)) + nactive(rightson(q))$, for an internal node q). Segment tree B has $n-1$ leaves and each node of B (leaf or internal node) stores a value *hc*. If leaf i $(1 \leq i \leq n-1)$ is *active* in A, then *hc(leaf i)=1+max{h(i), h(j)}*, where $j > i$ is the next active leaf. If leaf i is not *active* in A or is the last *active* leaf, then *hc(leaf i)=+∞*. The value *hc* of each internal node q of B is the minimum among all the *hc* values of the leaves in node q's subtree, i.e. *hc(node q)=min{hc(leftson(q)), hc(rightson(q))}*. Moreover, each node q of B maintains the number *lnum* of the leaf in its subtree which gives the value *hc(node* *q)*. We have *lnum(leaf i)=i* and *lnum(internal node q)=if (hc(leftson(q))* $\leq$ *hc(rightson(q))) then lnum(leftson(q)) else lnum(rightson(q))*.

At each step i ($1 \leq i \leq n-1$), each *active* leaf is the leftmost leaf of a subtree of the optimal tree. After every step, the number of active leaves decreases by 1. We can find in $O(log(n))$ time the pair of adjacent subtrees to combine. The height of the combination of these subtrees is *hc(root node of B)*, the leftmost leaf of the first subtree is *i=lnum(root node of B)* and that of the second subtree is $j = next\_active(i)$. We define the function *next_active* by using two other functions: $rank(i)$ and $unrank(r)$. $rank(i)$ returns the number of *active* leaves before leaf i ($0 \leq rank(i) \leq$ *nactive(root node of A)-1*). $unrank(r)$ returns the index of the leaf whose rank is r. The two functions are inverses of each other: $unrank(rank(i)) = i$ and $rank(unrank(r)) = r$. We have *rank(i)=rank'(i, root node of A)*, *unrank(r)=unrank'(r, root node of A)* and *next_active(i)=unrank(rank(i) + 1)*.

```
rank'(i, q):
if (q is a leaf node) then
  if (left(q)=right(q)=i) then return(0) else return(-1)
else if ($i > right(leftson(q))$) then
  return(nactive(leftson(q))+rank'(i, rightson(q)))
else return(rank'(i, leftson(q)))
```

```
unrank'(r, q):
if (q is a leaf node) then
  if ($r > 0$) then return(-1) else return(left(q))
else if ($nactive(leftson(q)) \leq r$) then
  return(unrank'(r-nactive(leftson(q)), rightson(q)))
else return(unrank'(r, leftson(q)))
```

The functions *rank*, *unrank* and *next_active* take $O(log(n))$ time each. After obtaining the indices of the two active leaves i and j whose corresponding subtrees are united (by adding a new internal node whose left son is the root of i's subtree and whose right son is the root of j's subtree), we mark leaf j as *inactive*. We do this by traversing the segment tree A from leaf j towards the root (from j to $parent(j)$, $parent(parent(j))$, ..., *root node of*

A) and decrement by 1 the *nactive* values of the visited nodes. Then, we change the h values of leaves i and j. We set *h(i)=hc(root node of B)* and $h(j) = +\infty$. After this, we also change the *hc* values associated to the leaves i and j in the segment tree B. The new *hc* value of leaf j will be $+\infty$. If i is now the last active leaf, then *hc(leaf i)* becomes $+\infty$, too. Otherwise, let $j' = next\_active(i)$, the next *active* leaf after i (at this point, leaf j is not *active* anymore). We change *hc(leaf node i)* to $(1 + max\{h(i), h(j')\})$. After changing the *hc* value of a leaf k, we traverse the tree from leaf k towards the root (visiting all of k's ancestors, in order, starting from $parent(k)$ and ending at the root of B). For each ancestor node q, we recompute *hc(node q)* as $min\{hc(leftson(q)), hc(rightson(q))\}$.

8.2. THE NUMBER OF TREES WITH A FIXED NUMBER OF LEAVES

In order to compute the number of labeled trees with n vertices and exactly p leaves, we compute a table $NT(i, j)$=the number of trees with i vertices and exactly j leaves ($1 \leq j \leq i \leq n$). Obviously, we have $NT(1, 1) = NT(2, 2) = 1$ and $NT(i, j) = 0$ for $i = 1, 2$ and $j \neq i$. For $i > 2$, we have $NT(i, i) = 0$ and for $1 \leq j \leq i - 1$, we proceed as follows. The j leaves can be chosen in $C(i, j)$ ways (i *choose* j). After choosing the identifiers of the j leaves, we conceptually remove the leaves from the tree, thus remaining with a tree having $i-j$ vertices and any number of leaves k ($1 \leq k \leq j$). Each of the j leaves that we conceptually removed is adjacent to one of these k vertices. Furthermore, each of these k vertices is adjacent to at least one of the j leaves from the larger tree. Thus, we need to compute the number of surjective functions f from a domain of size j to a domain of size k. We denote this value by $NF(j, k)$. This is a "classical" problem, but I will present a simple solution, nevertheless. We have $NF(0, 0) = 1$ and $NF(j, k) = 0$, if $j < k$. In order to compute the values for $k \geq 1$ and $j \geq k$, we consider every number g of values x from the set $\{1, \ldots, j\}$ for which $f(x) = k$. Once g is fixed, we have $C(j, g)$ ways of choosing the g values from the set $\{1, \ldots, j\}$. For each such possibility we have $NF(j-g, k-1)$ ways of extending it to a surjective function. Thus, $NF(j, k) =$

$\sum_{g=1}^{j} C(j,g) \cdot NF(j-g,k-1)$. We can tabulate all the $NF(*,*)$ values in $O(n^3)$ time (after tabulating the combinations $C(*,*)$ in $O(n^2)$ time, first). With the $NF(*,*)$ values computed, we have $NT(i,j) = C(i,j) \cdot \sum_{k=1}^{j} NF(j,k)$. We can easily compute each entry $NT(i,j)$ in $O(n)$ time. However, if we compute a table SNF of partial sums for the NF values, we can reduce the time complexity to $O(1)$ per entry. We have $SNF(j,0) = 0$ and $SNF(j,k) = SNF(j,k-1) + NF(j,k)$. With this table, the definition of $NT(i,j)$ becomes $C(i,j) \cdot SNF(j,j)$. Even with this improvement, however, the overall time complexity is $O(n^3)$, because of the step where the NF values are computed. The technique of performing dynamic programming on successive layers of leaves of a tree is also useful in several other counting problems.

8.3. THE NUMBER OF TREES WITH DEGREE CONSTRAINTS

We want to compute the number of unlabeled, rooted trees with $n \geq 2$ vertices, such that the (degree / number of sons) of each vertex belongs to a set S, which is a subset of $\{0, 1, 2, \ldots, n-1\}$. By *(a/b)* we mean that a refers to the degree-constrained problem and b refers to the number-of-sons-constrained problem (everything else being the same). Because every tree with $n \geq 2$ vertices must contain at least a leaf (a vertex of degree 1) and at least one vertex with at least 1 son, the set S will always contain the subset $(\{1\}/\{0,1\})$. We compute a table $NT(i,j,p)$=the number of trees with i vertices, such that the root has degree j (j sons) and the maximum number of vertices in the subtree of any son of the root is p; moreover, except perhaps the tree root, the (degrees/numbers of sons) of all the other vertices belong to the set S. Because the trees are unlabeled, we can sort the sons of each vertex in non-decreasing order of the numbers of vertices in their subtrees. Thus, we compute the table NT in increasing order of p. $NT(1,0,p) = 1$ and $NT(1, j > 0, p) = NT(i \geq 2, j, 0) = 0$. For $p \geq 1$ and $i \geq 2$, we have:

$$NT(i,j,p) = NT(i,j,p-1) + \sum_{k=1}^{\left\lfloor \frac{i-1}{p} \right\rfloor} NT(i-k \cdot p, j-k, p-1) \cdot CR(TT(p),k)$$

$TT(p)$ is the total number of trees with p vertices, for which the (degree / number of sons) of the root is equal to some $((x-1)/(x))$, $x \in S$, and the (degrees / numbers of sons) of the other vertices belong to the set S. By $CR(i, j)$ we denote combinations with repetitions of i elements, out of which we choose j. Because the argument i can be very large, we cannot tabulate $CR(i, j)$. Instead, we compute it on the fly. We know that $CR(i, j) = C(i + j - 1, j)$ and that $C(i, j) = \frac{i-j+1}{j} \cdot C(i, j - 1)$. Thus, $CR(i, j)$ can be computed in $O(j)$ time. Before computing any value $NT(*, *, p)$, we need to compute and store the values $TT(p)$, $TT(p) = \sum_{x \in S} NT(p, ((x-1)/(x)), p-1)$, and $CR(TT(p), k)$, for all the values of k $(1 \le k \le \left\lfloor \frac{n-1}{p} \right\rfloor)$. We can compute all of these values in $O(n^3 \cdot log(n))$ time. The desired number of trees is $\sum_{x \in S} NT(n, x, n - 1)$. The memory storage can be reduced from $O(n^3)$ to $O(n^2)$, by noticing that the values $NT(*, *, p)$ are computed based only on the values $NT(*, *, p - 1)$. Thus, we can maintain these values only for the most recent two values of p.

A less efficient method is to compute the numbers $Tok(i)$=the number of trees with i vertices, such that each vertex satisfies the (degree/number of sons) constraints. $Tok(1) = Tok(2) = 1$. We make use of the $TT(i)$ values defined previously, except that they will be computed differently. For every $i \ge 2$, we consider every possible number x of sons of the tree root and compute $NT2(i, x)$=the number of trees with i vertices, such that the tree root has x sons and all the other vertices satisfy the (degree/number of sons) constraints. We generate all the possibilities $(y(1), y(2), ..., y(i-1))$, with $0 \le y(j) \le \left\lfloor \frac{i-1}{j} \right\rfloor$ $(1 \le j \le i - 1)$ and $y(1) + \ldots + y(i - 1)$=x. $y(j)$ is the number of sons of the tree root which have j vertices in their subtrees. The number of trees "matching" such a partition is equal to $\prod_{j=1}^{i-1} CR(TT(j), y(j))$. $NT2(i, x)$ is computed by summing the numbers of trees "matching" every partition. Afterwards, if $x \in S$, we add $NT2(i, x)$ to $Tok(i)$. If $x = ((y - 1)/(y))$ and $y \in S$, then we add $NT2(i, x)$ to $TT(i)$. $NT2(i, x)$ may be added to both $Tok(i)$ and $TT(i)$.

9. RELATED WORK

Reliability analysis and improvement techniques for distributed systems were considered in [6,7]. Reliability analysis and optimization for tree networks in particular were considered in [3,5,8]. Different kinds of tree partitioning algorithms, based on optimizing several objectives, were proposed in [9,10,16]. Problems related to tree coloring were studied in [4]. Content delivery in distributed systems is a subject of high practical and theoretical interest and is studied from multiple perspectives. Communication scheduling in tree networks was considered in many papers (e.g. [17]) and the optimization of content delivery trees (multicast trees) was studied in [11].

10. CONCLUSIONS AND FUTURE WORK

In this paper I considered several optimization problems regarding distributed systems with tree topologies (e.g. peer-to-peer networks, wireless networks, Grids), which have many practical applications: minimum weight cycle completion (reliability improvement), constrained partitioning (distributed coordination and control), minimum number of streams and degree-constrained minimum spanning trees (efficient content delivery), optimal matchings (data replication and resource allocation), coloring (resource management and frequency allocation) and tree counting aspects. All these problems are variations or extensions of problems which have been previously posed in other research papers. The presented techniques are either better (faster or more general) than the previous solutions or easier to implement.

References

[1] J. Roskind, R. E. Tarjan, *A Note on Finding Minimum-Cost Edge-Disjoint Spanning Trees*, Mathematics and Operations Research **10 (4)** (1985), 701-708.

[2] M. A. Bender, M. Farach-Colton, *The LCA Problem revisited*, Lecture Notes in Computer Science **1776** (2000), 88-94.

[3] M. Scortaru, *National Olympiad in Informatics*, Gazeta de informatica (Informatics Gazzette) **12 (7)** (2002), 8-13.

[4] S. M. Hedetniemi, S. T. Hedetniemi, T. Beyer, *A Linear Algorithm for the Grundy (Coloring) Number of a Tree*, Congressus Numerantium **36** (1982), 351-362.

[5] M. I. Andreica, N. Tapus, *Reliability Analysis of Tree Networks Applied to Balanced Content Replication*, Proc. of the IEEE Intl. Conf. on Automation, Robotics, Quality and Testing (2008), 79-84.

[6] D. J. Chen, T. H. Huang, *Reliability Analysis of Distributed Systems Based on a Fast Reliability Algorithm*, IEEE Trans. on Par. and Dist. Syst. **3** (1992), 139-154.

[7] A. Kumar, A. S. Elmaghraby, S. P. Ahuja, *Performance and reliability optimization for distributed computing systems*, Proc. of the IEEE Symp. on Comp. and Comm. (1998), 611-615.

[8] H. Abachi, A.-J. Walker, *Reliability analysis of tree, torus and hypercube message passing architectures*, Proc. of the IEEE S.-E. Symp. on System Theory (1997), 44-48.

[9] G. N. Frederickson, *Optimal algorithms for tree partitioning*, Proc. of the ACM-SIAM Symposium on Discrete Algorithms (SODA) (1991), 168-177.

[10] R. Cordone, *A subexponential algorithm for the coloured tree partition problem*, Discrete Applied Mathematics **155 (10)** (2007), 1326-1335.

[11] Y. Cui, Y. Xue, K. Nahrstedt, *Maxmin overlay multicast: rate allocation and tree construction*, Proc. of the IEEE Workshop on QoS (IWQOS) (2004), 221-231.

[12] Y. Qinglin, *Factors and Factor Extensions*, M.Sc. Thesis, Shandong Univ., 1985.

[13] T. L. Magnanti, L. A. Wolsey, *Optimal Trees*, Handbooks in Operations Research and Management Science, **vol. 7, chap. 9** (1995), 513-616.

[14] S.-C. Mu, R. S. Bird, *On Building Trees with Minimum Height, Relationally*, Proc. of the Asian Workshop on Programming Languages and Systems (2000).

[15] M. I. Andreica, N. Tapus, *Optimal Offline TCP Sender Buffer Management Strategy*, Proc. of the Intl. Conf. on Comm. Theory, Reliab., and QoS (2008), 41-46.

[16] B. Y. Wu, H.-L. Wang, S. T. Kuan, K.-M. Chao, *On the Uniform Edge-Partition of a Tree*, Discrete Applied Mathematics **155 (10)** (2007), 1213-1223.

[17] M. R. Henzinger, S. Leonardi, *Scheduling multicasts on unit-capacity trees and meshes*, J. of Comp. and Syst. Sci. **66 (3)** (2003), 567-611.

[18] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, *Introduction to Algorithms*, MIT Press and McGraw-Hill (2001).

ROMAI J., **4**, 1(2008), 27–46

BIT REVERSAL THROUGH DIRECT FOURIER PERMUTATION METHOD AND VECTORIAL DIGIT REVERSAL GENERALIZATION

Nicolaie Popescu-Bodorin

“Spiru Haret” University, Bucharest

bodorin@iee.org

Abstract This paper describes the Direct Fourier Permuation Algorithm, an efficient method of computing Bit Reversal of natural indices $[1, 2, 3, \dots, 2^k]$ in a vectorial manner (k iterations). It also proposes the Vectorial Digit Reversal Algorithm, a natural generalization of Direct Fourier Permutation Algorithm that is enabled to compute the r-digit reversal of natural indices $[1, 2, 3, \dots, r^k]$ where r is an arbitrary radix. Matlab functions implementing these two algorithms and various test and comparative results are presented in this paper to support the idea of inclusion of these two algorithms in the next Matlab Signal Processing Toolbox official distribution package as much faster alternatives to current Matlab functions *bitrevorder* and *digitrevorder*.

Keywords: Bit Reversal, Direct Fourier Permutation Method, Additive Constants Method, Vectorial Digit Reversal, bitrevorder, digitrevorder, FFT, Fast / Discrete Fourier Transform, Danielson-Lanczos Lemma.

2000 MSC: 65T50, 65T99, 65K05, 68W01.

1. INTRODUCTION

Recurrence relations in the Danielson-Lanczos Lemma allow for the immediate implementation of an explicitly recursive function for FFT computation. Each time it calls itself, a 2^k point FFT computation is reduced to two 2^{k-1} FFT computations. This direct approach (divide et impera and explicit backward recursion) allows for the FFT computation algorithm to be expressed by decimation-in-time implementation (1). Subdividing the initial vector x up to a set of pairs on which the FFT computation is very simple is sometimes called *buterfly operation*, a name suggested by the graphical representation of the computation [1].

```
function $X = mkFFT(x)$
% $x$ - 1x$2^k$ line vector
% $X$ - Discrete Fourier Transform of $x$ calculated through 2-radix Decimation
% in Time Fast Fourier Transform with Explicit Backward Recursion.
$N = max(size(x));$
$if N == 1$
    $X = x;$
else                                                              (1)
    $oddind = 1 : 2 : N; xodd = x(oddind);$
    $evenind = 2 : 2 : N; xeven = x(evenind);$
    $O = mkFFT(xodd);$
    $E = mkFFT(xeven);$
    For $k = 1 : 1 : N/2$
        $X(k) = O(k) + (exp(-2 * pi * i * (k-1)/N)) * E(k);$
        $X(N/2 + k) = O(k) - (exp(-2 * pi * i * (k-1)/N)) * E(k);$
    end
end;
```

Implicitly, during this computation, a permutation of the argument vector x takes place whenever the function calls itself. Consequently, a composed permutation of the initial vector argument x will occur up to and within the innermost call. This permutation will be referred further in this paper as *Fourier Permutation* and is sometimes named the *Buterfly Permutation* [1], or *Bit Reversal Permutation* [2] due to the fact that the reversed bit representation of the permuted index of the initial 0-based index (see the following example) is implicitly sorted in ascending order.

Example 1.

permuted index	bit reprezentation	reversed reprezentation	0-based index
0	000	000	0
4	100	001	1
2	010	010	2
6	110	011	3
1	001	100	4
5	101	101	5
3	011	110	6
7	111	111	7

The Fourier Permutation of the 1-based index $[1, 2, 3, 4, 5, 6, 7, 8]$ is $[1, 5, 3, 7, 2, 6, 4, 8]$. The stages leading to this permutation through the explicit calls in (1) are the following: $[1, 2, 3, 4, 5, 6, 7, 8] \to [1, 3, 5, 7, 2, 4, 6, 8] \to [1, 5, 3, 7, 2, 6, 4, 8] \to [1, 5, 3, 7, 2, 6, 4, 8]$, where a left to right reading shows the permutations operated from the first to the last (the innermost) call.

2. ADDITIVE CONSTANTS METHOD

In what follows we aim to compute the Fourier permutation in an iterative manner, by other methods than the 'Bit Reversal' algorithms proposed in [1] - [9].

Our primary intention is to formulate an algorithm for computing the function $Y_N = mkFPerm(N)$, where Y_N is the Fourier permutation of the vector of indices $[1, 2, \dots, N]$ and $N = 2^k$, with acceptable efficiency. The formulation of such an algorithm requires that a recurrence relation of first order should be found between the consecutive components of the resulting vector Y_N

$$\forall p \in \overline{1, N-1} : Y_N(p+1) = Y_N(p) + C_N(p);\ Y_N(1) = 1, \qquad (2)$$

where $C_N(1:N)$ is an additive constants vector to be determined.

Example 2:

$$\begin{array}{lccccccccc} C_8(1:7) = [& & (+4) & (-2) & (+4) & (-5) & (+4) & (-2) & (+4) &] \\ Y_8(1:8) = [& 1 & 5 & 3 & 7 & 2 & 6 & 4 & 8 &] \end{array}$$

The Fourier permutation of indices $\overline{1,8}$ can be computed taking into account that each component of the resulting vector is the sum of the preceding component and a constant that depends on N, and that the first component always has the value 1.

The quantity to be found in all the odd rank positions of vector C_N (+4 in the example under discussion) will be further referred to in this paper as the *trivial additive constant of the Fourier permutation*, all the other being called nontrivial constants.

Definition 1. Let $N = 2^k, k \in N^*$, N*. We shall call the *aditive constants* of the Fourier permutation of indices $[1, 2, 3, ..., 2^k]$ the $(2^k - 1)$ components of vector C_N thus constructed:

i. $C_{2^1} = 1$;

ii. $\forall k \in N^*,\ k \geq 2: \ C_{2^k} = [2C_{2^{k-1}}, -2^k + 3, 2C_{2^{k-1}}]$.

The basic properties of the additive constants corresponding to the Fourier permutation of indices $[1, 2, 3, ..., 2^k]$ are given in the following

Proposition 1. *If $k \geq 2$ and $N = 2^k$, the vector of the additive constants of the Fourier permutation (C_N) has the following properties:*

i. *all odd-rank components store the value of the trivial additive constant:*
$\forall p \in \overline{1, N/2} : C_N(2p-1) = 2^{k-1}$;

ii. *all non-trivial additive constants are negative:*
$\forall p \in \overline{1, N/2} : C_N(2p) < 0$;

iii. *the lowest non-trivial additive constant splits the C_N vector into two equal vectors:*

$C_N(1 : 2^{k-1} - 1) = C_N(2^{k-1} + 1 : 2^k - 1)$;

iv. *the value of the minor nontrivial constant:* $C_N(2^{k-1}) = -2^k + 3$;

v. *all non-trivial additive constants, except the minor nontrivial constant, are even:*

$\forall p \in \overline{1, (N/2 - 1)} : |C_N(2p)|\, mod\, 2 = 0$;

vi. *the components that are symmetrically placed in relation to the minor nontrivial constant are equal:*

$\forall p \in \overline{1, (N/2 - 1)} : C_N(2^k - p) = C_N(p)$;

vii. *forward recursion: at step $(k + 1)$ each of the vectors formed with the components on the left and on the right of the minor non-trivial additive constant, respectively, is double the vector of constants computed at step k:*

if $N_1 = 2^k$ and $N_2 = 2^{k+1}$ *then*: $C_{N_2}(1 : N_1 - 1) = 2C_{N_1} = C_{N_2}(N_1 + 1 : N_2)$;

viii. *forward recursion between minor nontrivial constants:* if $N_1 = 2^k$ *and* $N_2 = 2^{k+1}$ *then*: $C_{N_2}(N_1) = 2C_{N_1}(2^{k-1})$.

Consequence. If $N = 2^k$ and $k \geq 2$, then the number of unique nontrivial additive constants of the Fourier permutation is $(k - 1)$, and the value of the trivial additive constant is 2^{k-1}. Consequently the computation of the Fourier permutation of indices $[1, 2, 3, ..., 2^k]$ is reduced to:

- the computation of these k additive constants and their distribution in a template vector of length $2^k - 1$.

- the computation of each component of the resulting vector as the sum of the preceeding component and the corresponding additive constant.

3. COMPUTING FOURIER PERMUTATION THROUGH ADDITIVE CONSTANTS ALGORITHM

Due to property P1.iii, in order to obtain the template vector of additive constants it suffices to determine its first 2^{k-1} components, i.e. its first 2^{k-2} non-trivial additive constants (as all the others, i.e. all odd rank components store the trivial additive constant). According to the above considerations, the matlab function for generating the additive constants of the Fourier permutation and the function for computing the Fourier permutation of indices $[1, 2, 3, ..., 2^k]$ can be written as:

```
function $V = mkCAPF(N)$
% $VI$ = intermediate vector of the first $2^{k-1}$ non-trivial additive constants
% $N = 2^k, k > 2$
% $ct$ = trivial constant
$ct = N/2$; % property (P1.i)
$k = log2(N)$;
$VI = [-2, -5]$;
% the two non-trivial additive constants corresponding to case k=3
for $p = 4 : k$
    $VI = 2 * VI$; % property (P1.vii)
    $c = max(size(VI))$;
    $VI = [VI, VI(1 : c - 1)]$; % property (P1.iii)
    $VI = [VI, VI(c) - 3]$; % properties (P1.vii, P1.viii)
end;
$c = max(size(VI))$;
$VI = [VI, VI(1 : c - 1)]$; % property (P1.iii)
$V = zeros(1, N - 1) + ct$; % property (P1.i)
$V(2 : 2 : N - 1) = VI$;

function $V = mkFPerm(N)$
$V = zeros(1, N); V(1) = [1]$;
```

```
$CN = mkCAPF(N);$
for $p = 1 : N - 1$
    $V(p + 1) = V(p) + CN(p);$
end;
```

The complexity of the computation of permuted indices depends on the multiplication operations in the mkCAPF function and on the addition iterated in the mkFPerm function. Consequently the complexity of computing the function mkFPerm is $O(Nlog_2(N))$ and may decrease if multiplications are excluded from the computational mechanism and the number of iterations decreases, possibly to k.

4. DIRECT FOURIER PERMUTATION METHOD

Proposition 2. *If $N = 2^k$ and $k \geq 2$, the vector of the additive constants of the Fourier permutation, $C_N = mkCAPF(N)$, has the following properties:*

i. . *the sum of all the additive constants of the Fourier permutation is*

$$\sum_{i=1}^{N-1} C_N(i) = 2^k - 1;$$

ii. the sum of all non-trivial constants of the Fourier permutation is

$$\sum_{i=1}^{N/2} C_N(2i-1) = -(2^{k-1} - 1)^2;$$

iii. the sum of the first 2^{k-1} constants of the Fourier permutation is 1

$$\sum_{i=1}^{N/2} C_N(i) = 1;$$

iv. the sum of any 2^{k-1}consecutive constants of the Fourier permutation is1

$$\forall p \in \overline{1, N/2} : \sum_{i=p}^{N/2+p-1} C_N(i) = 1.$$

Consequences.

By construction, for any $i \in \overline{1, 2^{k-1}}$ the difference between the rank $(i + 2^{k-1})$ component and the rank i component of vector Y is the sum of the 2^{k-1} consecutive additive constants starting with (and including) the rank i constant.

Consequently $Y_N(2^{k-1}+1 : 2^k) = Y_N(1 : 2^{k-1}) + 1$.

Moreover, using properties P1.vii and P2.iii,iv applied to case $N = 2^{k-1}$, it follows that the sum of any 2^{k-2} consecutive constants selected from the first $(2^{k-1} - 1)$ components of C_N is 2. However, by construction, for any $i \in \overline{1, 2^{k-2}}$, the difference between the rank $(i + 2^{k-2})$ component and the rank i component of the vector Y_N is the sum of the 2^{k-2} consecutive additive constants starting with the rank i constant $Y_N(2^{k-2} + 1 : 2^{k-1}) = Y_N(1 : 2^{k-2}) + 2$.

The sum of any 2^{k-3} consecutive constants selected from the first $(2^{k-2} - 1)$ components of C_N is 2^2 , and therefore $Y_N(2^{k-3}+1 : 2^{k-2}) = Y_N(1 : 2^{k-3})+2^2$.

And the procedure can go on until the first component of vector C_N is reached by successive truncations like those above: $C_N(1) = 2^{k-1}$; $Y_N(2) = Y_N(1) + 2^{k-1}$.

Proposition 3. *If $N = 2^k$ and $k \geq 2$, the Fourier permutation of the indices $[1, \ldots, N]$, has the property*

$$\forall p \in \overline{0, (k-1)} : Y_N(2^{k-p-1} + 1 : 2^{k-p}) = Y_N(1 : 2^{k-p-1}) + 2^p.$$

5. DIRECT FOURIER PERMUTATION ALGORITHM IMPLICIT BIT REVERSAL

By applying the properties mentioned in Proposition 2 and their consequences, the generating function of the the Fourier permutation of indices $\overline{1, 2^k}$ can be rewritten in k iterations, as follows:

```
function $V = dfp(b, N)$
% $N = 2^k, k > 0$;
% $b$ is a natural number;
% $V$ is the Fourier permutation of indices $[b, b+1, ..., b+2^k-1]$
$V = [b]; p2 = N/2;$
while $p2 \geq 1$
    $V = [V, V+p2]$; % one vectorial addition and one memory reallocation of $V$
    $p2 = p2/2$; % update by one division
end;
```

Let $V = dfp(0, N)$.

Taking into account that the decimal number obtained by reversing the k-bit representation of the decimal number 2^{k-p} is $2^{(k-1)-(k-p)} = 2^{p-1}$, the function that returns the decimal values corresponding to the binary representations obtained by reversing the k-bit representation of all of the components of V is the following:

```
function $W = mkB10RevKBit(N)$
% $N = 2^k, k > 0;$
$W = [0]; p2 = 1;$
while $p2 \leq N/2$
    $W = [W, W+p2];$
    $p2 = p2 * 2;$
end;
```

Proposition 4. *On each iteration within $mkB10RevKBit$ function, the intermediate result W is ascendently ordered.*

Consequence. The result $W = mkB10RevKBit(N)$ is ascendently ordered. As the length of W is $N = 2^k$ and the values in W are distinct natural numbers corresponding to binary k-bit representations, it follows that the maximal item in W is $2^k - 1$. Consequently $W = [0, 1, 2, 3, \ldots, 2^k - 1]$.

The above considerations allow for the formulation of the following theorem.

Theorem 1. (Correctness and Complexity of Direct Fourier Permutation Algorithm) *If $V = dfp(b, N)$, $N = 2^k$, then by implication the vector V meets the relation $V = b + R$, where R is a permutation of W, specifically the one that corresponds to the reversed k-bit representations of the components of W (i.e. $R = dfp(0, N)$, or in other words, reversed k-bit representation of R is sorted in ascending order, i.e. R=bitrevorder(W)). The arithmetical complexity of the computation of V is $O(N + k - 1)$.*

Remark. The above theorem proves that the Fourier permutation is uniquely determined by its first component and by its most important property formulated as Proposition 3. This is because neither Proposition 3 nor Proposition 2 really depends on the first component of the index to be permuted ($[0, 1, 2, \dots, 2^{k-1}]$ or $[1, 2, \dots, 2^k]$ or $[b, b + 1, b + 2, \dots, b + 2^{k-1}]$). In other words, all the properties mentioned in this paper regarding both the Fourier permutation and the set of additive constants of the Fourier permutation, including Theorem 1, are independent of the particular choice of the initial index (languages like C uses 0-based indexing while Matlab uses 1-based indexing). These are the reasons why the formalism in Theorem 1 has been chosen to unify the descriptions of both cases mentioned in Example 1.

6. VECTORIAL DIGIT REVERSAL GENERALIZATION

As a natural generalization of the Direct Fourier Permutation Algorithm we propose the following algorithm that is enabled to compute the r-digit reversal of natural index $[1, 2, 3, ..., r^k]$ for arbitrary radices r in a vectorial manner:

```
function $V = vdigitrevorder(N, r)$
$\%N = r^k$
$crN = N/r;$
$KV = crN * [0 : r - 1] + 1;$
$V = KV;$
$crLenV = r;$
$crStep = 1;$
```

```
while $crLenV < N$
    if $crLenV == r^{crStep}$
        $crStep = crStep + 1;$
        $crN = crN/r;$
        $KV = V + crN;$
        $crLenKV = crLenV;$
        $crLenV = 2 * crLenV;$
    else
        $KV = KV + crN;$
        $crLenV = crLenV + crLenKV;$
    end;
    $V = [V, KV];$
end;
```

The idea of the above algorithm is to compute (at each step of the iteration) the current kernel vector KV and to concatenate its value at the end of the currently calculated partial result V using the following rule:

While V is only a partial result (i.e. length of $V < N$):

- if the current length of V is equal to a natural power of the radix r^p then
  - update the current kernel vector *increasing the number of its components and increasing the components themselves* $KV = V + r^{k-p-1}$;
  - update the current partial result *doubling the number of its components* $V = [V, V + r^{k-p-1}]$;
- else (the current length of V is between r^p and r^{p+1})
  - update the current kernel vector *increasing all of its components* $KV = KV + r^{k-p-1}$
  - update the current partial result *increasing the number of its components with* r^p $V = [V, KV]$.

Another form of the same algorithm, in which the similarity to the Direct Fourier Permutation is obvious, is the following:

Vectorial Digit Reversal:

```
function $V = vdro(N, r)$
$V = [1]; pr = N/r;$
while $pr \geq 1$
    $KV = V;$
    for $cont = 1 : r - 1$
        $V = [V, KV + cont * pr];$
    end;
    $pr = pr/r;$
end;
```

Direct Fourier Permutation:

```
function $V = dfp(b, N)$
$V = [b]; p2 = N/2;$
while $p2 \geq 1$
    $V = [V, V + p2];$
    $p2 = p2/2;$
end;
```

We prefer the above formulation of the Vectorial Digit Reversal Algorithm for two reasons: it is less redundant and it enables us to note that these two functions, $vdro(N, r)$, $dfp(b, N)$, have the same arithmetic complexity when $r = 2$ and $N = 2^k$. Also, during various tests performed by the author, the above formulation of the Direct Fourier Permutation Algorithm has been proved to be the fastest Matlab script for computing bit reversal permutation. On the other hand, both the present mathematical formulation of the Bit Reversal computation and the theoretical arithmetic complexity obtained in Theorem 1 suggest that the Direct Fourier Permutation Algorithm defines the minimal computational effort (minimal arithmetic complexity) for computing bit reversal permutation of an arbitrary-based index $[b, b + 1, \ldots, b + 2^{k-1}]$ in natural arithmetic, unless a stronger property than Proposition 3 can be formulated.

Despite the Matlab formalism that has been chosen in the description of both of the above algorithms, up to this point of the present work, there was no particular hypothesis (concerning some particular implementation or some particular computational advantage that could have been gained by programming in one specific language, medium or platform) being pursued, and consequently, all the above results are purely arithmetical. In the following section we will see how the particular result of Theorem 1 that concerns complexity can be refined by making the most of the speed of the vectorized Matlab calculation.

7. BENCHMARK

7.1. GENERALITIES

In this section we obtain experimentally determined time-complexity results for the two Matlab script functions, *dfp* and *vdro*, coded above.

Test variables: the arithmetic complexity of the *dfp* function does not depend on the starting value (b) of the index $[b, b+1, \ldots, b+2^{k-1}]$ and consequently the *dfp* function will be tested only against increasing values of the length of the index ($N = 2^k$). According to the above considerations, almost identical time-complexity results are expected to be found for the $dfp(b, 2^k)$ and $vdro(2^k, 2)$ computations. The *vdro* function will be tested against the variable (k, r).

Computing the medium execution times: the medium execution time for each test variable (k and (k, r) respectively) will be the average of execution times cumulated over a great number of repetitions. Each medium execution time thus calculated will obviously depend on the performance of the computer running the tests, but the *nature* of variation of the medium execution time against the test variables is a characteristic of the Matlab implementations of the algorithms in themselves and does not depend on the performance of the computer. On the other hand, it is all the more necessary to establish with accuracy the medium execution time, as the execution time intended to be estimated is shorter (cases of immediate practical interest, where test variables are small). In these cases the number of repetitions will be the greatest and will decrease gradually as the value of the test variables increases, up to a value that guarantees that the resulting vector of medium execution times is a fair statistical reflection of reality. In establishing the number of repetitions corresponding to each individual test variable, as well as in establishing the minimal number of repetitions corresponding to high values of the test variables, we will consider to be a "fair reflection of reality" such a representation of the medium execution times in which the first symptom of convergence is present, i.e. the curve of medium execution times is nearly a smooth, ascendent one.

Limitations of the present 'digitrevorder' Matlab functions: there are three reasons for replacing the existing *digitrevorder* Matlab function within the Signal Processing Toolbox:

i) the arbitrary radix r is limited to the integer range from 2 to 36. In the proposed implementation of the Vectorial Digit Reversal Algorithm (*vdro* function) there is no such limitation;

ii) the computation of the *vdro* proposed function is much faster than that required to be done in the present *digitrevorder* function;

iii) unexpected behavior of the *digitrevorder* function could be sometimes obtained ($digitrevorder(1:81,3)$ - for example). This is because of a data validation issue: the power k of the radix r is computed in the digitrevorder.m (present file) as being: $radixpow = floor(log10(N)/log10(radixbase))$ and this instruction sometimes fails to return the correct result of the calculus (see $floor(log10(27)/log10(3))$, $floor(log10(81)/log10(3))$, $floor(log10(7^7)/log10(7))$ for example) and consequently, that instruction must be replaced anyway with: $radixpow = floor(log(N)/log(radixbase))$, and moreover, a refined, more accurate method of computing the $log10$ function may be needed.

7.2. TESTING THE PROPOSED FUNCTIONS AGAINST THE EXISTING DIGITREVORDER MATLAB FUNCTION

This section assumes that the tested functions are called with the following syntaxes: $dfp(1,2^k)$, $vdro(2^k,2)$, $digitrevorder(1:2^k,2)$, where k takes several increasing natural values. We preferred to test *digitrevorder* instead of *bitrevorder* function because, due to its implementation, the latter is calling the former.

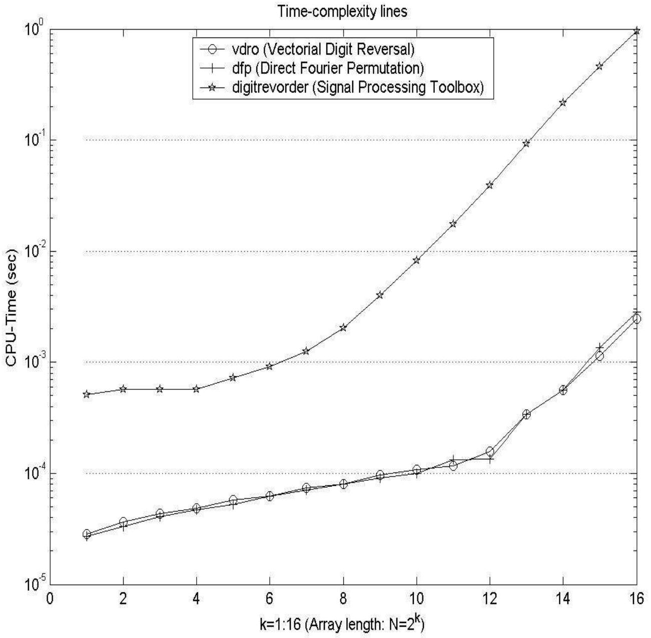


Fig. 1. Testing *dfp*, *vdro*, *digitrevorder* against increasing *k* values

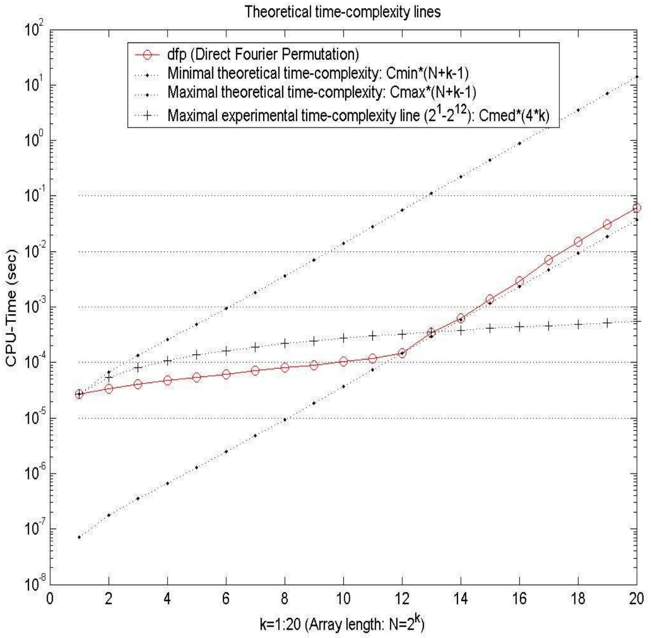


Fig. 2. Time-complexity lines of the dfp computation

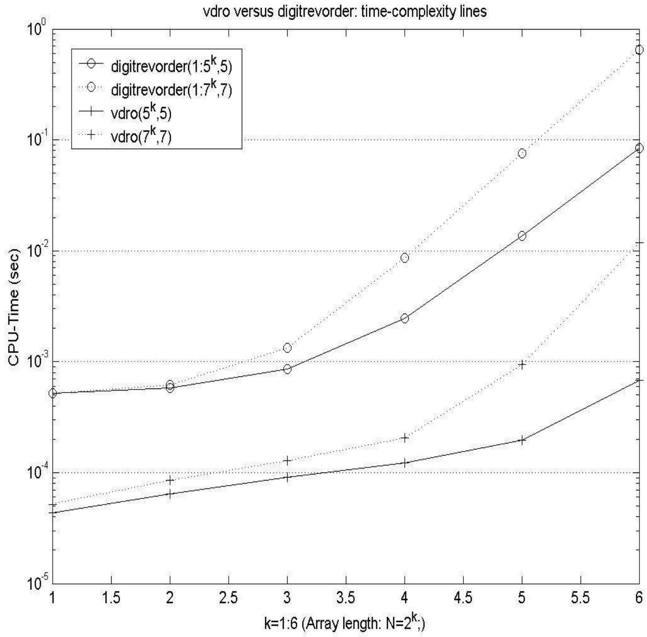


Fig. 3. Time-complexity lines: vdro versus digitrevorder (arbitrary radices)

As expected, the time-complexity lines (fig.1) of the two proposed algorithms (dfp and $vdro$) are almost identical when the radix is $r = 2$. In this case the graphical representation in fig. 1 enables us to distinguish two variation regimes along the time-complexity lines of the both algorithms. Fig. 2 reveals more clearly the point where the variation of the time-complexity line undergoes a significant change. During all the tests done by the author so far, the existence of this point has been proved to be an invariant of the algorithm. This is because the medium execution time calculated for each k depends on the following factors: k itself, the number of repetitions, the type of the cache and influences from other processes that are running on the same CPU.

The greater the increase of k and in the number of repetitions, the greater the contribution of the two other factors will be to the calculated medium execution time and the computation itself will becomes less cache-optimal. In fig. 2, Cmin, Cmed and Cmax are the minimal, medium, maximal duration, respectively, of the computational cycle of the dfp algorithm, all of them being experimentally determined.

Also, fig. 2 and the dfp function itself allow for the formulation of the following remarks:

i) due to its simplicity, the dfp function certainly has a lower complexity and a higher performance than the *bitrevorder* function within the Signal Processing Toolbox;

ii) the dfp function allows us to compute the bit reversal order of indices $\overline{1, 2^k}$ in just k iterations, each of these involving only three operations: one vectorial addition, one memory reallocation, and one division;

iii) for k ranging between 2 and 12, the medium execution time needed to compute $V = dfp(1, 2^k)$ is increasing linearly with $4k$ (reflecting the very few operations within each iteration). On this range of k values, the dfp function takes all computational advantage of its simplicity. Consequently, this is the range on which the dfp function reaches its maximal efficiency and its minimal computational complexity $O(4k)$;

iv) for k ranging between 13 and 20, due to the increasing size of the variable V, the computational complexity of the dfp function suddenly turns into its own worst case scenario - predicted by its theoretical complexity $O(N+k-1)$ - on which the medium execution time needed to compute $V = dfp(1, 2^k)$ is increasing exponentially with k (i.e. linearily with $N = 2^k$). But even on this range of k values, the multiplicative constant characterising the time complexity of the computation is relatively small. Consequently, on this range of k values the dfp function is still operative, even though it reaches its minimal efficiency and its maximal computational complexity $O(2^k + k - 1)$.

Fig. 3 reveals the great improvement in the performance of the proposed *vdro* function compared to the present Matlab function digitrevorder. Even for arbitrary radices r the computation of the *vdro* function is faster and more reliable.

7.3. REPLICABILITY OF RESULTS

The essential qualitative results illustrated by the tests above are the following:

i) higher performance of the dfp function compared to the Matlab $bitrevorder$ function;

ii) linear variation proportional to $4k$ of the medium execution time needed to compute $V = dfp(1, 2^k)$ for k ranging between 2 and 12;

iii) exponential variation proportional to (2^k+k-1) of the medium execution time needed to compute $V = dfp(1, 2^k)$ for k ranging between 13 and 20;

iv) higher performance of the $vdro$ function compared to the Matlab $digitrevorder$ function for arbitrary radices r.

The replication of the qualitative results mentioned above does not depend on the PC hardware architecture used. Nonetheless, the replication of any quantitative results requires a hardware architecture which must be very similar to the one used for the purposes of this paper. All the tests mentioned above were run on the following configuration: *CPU:* Prescott, Intel Pentium 4E, 2800Mhz; *Memory bus features:* Type - Dual DDR SDRAM, Bus width - 64 bit, Real clock - 200MHz (DDR), Effective clock - 400MHz, Bandwidth - 6400MB/s; *Memory module features:* Size - 512 MB, Module type - Unbuffered, Speed - PC3200 (200MHz);

As far as the variation of experimental medium execution times along the range of k is concerned, all the tests that have been run have high statistical relevance, being based on a number of repetitions that is high enough for statistic phenomena to become observable. All the above tests have been validated by the author by performing more tests on other hardware configurations, with similar results.

8. APPENDIX

1. The Discrete Fourier Transform of the signal of finite length $x(0 : N-1)$ is a vector $X(0 : N-1)$, possessing the components

$$\forall k \in \overline{0, (N-1)} : X(k) = \sum_{n=0}^{N-1} x(n) W_N^{kn},\ W_N = exp(-2\pi i/N);$$

2. Fast Fourier Transform is the algorithm that allows for the computation of the Discrete Fourier Transform with a complexity of order $O(Nlog_2 N)$. The FFT algorithm for computing the Discrete Fourier Transform of signals of length 2^k is based on the Danielson-Lanczos Lemma.

Danielson-Lanczos Lemma. *Let signal* $x(0 : N-1)$*,* $N = 2^k$ *and* $f(0 : N/2-1)$*,* $g(0 : N/2-1)$ *be defined by*

$$f(n) = x(2n),\ g(n) = x(2n+1),\ n \in \overline{0, (N/2-1)};$$

Let $X(0 : N-1)$*,* $F(0 : N/2-1)$*,* $G(0 : N/2-1)$ *be the Fourier transforms of signals* x*,* f*, and* g*, respectively.*
Then X *has the following components:* $X(k) = F(k) + W_N^k G(k),\ k \in \overline{0, (N/2-1)}$; $X(N/2+k) = F(k) - W_N^k G(k),\ k \in \overline{0, (N/2-1)}$;

Acknowledgements The author wishes to thank Luminiţa State for carefully reading this text and also for all discussions and advice on Fourier Analysis and Signal Processing subjects.

References

[1] D. Sundararajan, M. Omair Ahmad, M. N. S. Swamy, *Fast computation of the discrete Fourier transform of real data*, IEEE Trans. on Signal Processing, **45**, 8(1997), 2010-2022.

[2] L. Carter, K. Su Gatlin, *Towards an optimal bit-reversal permutation program*, Proc. of IEEE-FOCS'98, November 8-11 in Palo Alto, CA.

[3] A. A. Yong, *A Better FFT bit-reversal algorithm without tables*, IEEE Trans. on Signal Processing, **39**, 10(1991), 2365-2367.

[4] A. Biswas, *Bit reversal in FFT from matrix viewpoint*, IEEE Trans. on Signal Processing, **39**, 6(1991), 1415-1418.

[5] J. Jeong, W. J. Williams, *A unified fast recursive algorithm for data shuffling in various orders*, IEEE Transactions on Signal Processing, **40**, 5(1992), 1091-1095.

[6] M. Orchard, *Fast bit-reversal algorithms based on index representations in* $GF(2^b)$, IEEE Transactions on Signal Processing, **40**, 4(1992), 1004-1007.

[7] J. M. Rius, R. De Porrata-Doria, *New bit-reversal algorithm*, IEEE Trans. on Signal Processing, **43**, 4(1995), 991-994.

[8] K. Drouiche, *A new efficient computational algorithm for bit reversal mapping*, IEEE Transactions on Signal Processing, 49, 1(2001), 251-254.

[9] S.-Ch. Pei, K.-W. Chang, *Efficient bit and digital reversal algorithm using vector calculation*, IEEE Trans. on Signal Processing, **55**, 3(2007), 1173-1175.

ROMAI J., **4**, 1(2008), 47–60

THEORY OF OLIGOPOLIES: DYNAMICS AND STABILITY OF EQUILIBRIA

Konstantinos Andriopoulos, Tassos Bountis, Nikos Papadopoulos

Centre for Research and Applications of Nonlinear Systems and Department of Mathematics, University of Patras, Greece

kand@aegean.gr; bountis@math.upatras.gr

Abstract The theory of oligopolies is a particularly active area of research using applied mathematics to answer questions that arise in microeconomics. It basically studies the occurrence of equilibria and their stability in market models involving few firms and has a history that goes back to the work of Cournot in the 19th century. More recently, interest in this approach has been revived, owing to important advances in analogous studies of Nash equilibria in game theory. In this paper, we first attempt to highlight the basic ingredients of this theory for a concrete model involving two firms. Then, after reviewing earlier work on this model, we describe our modifications and improvements, presenting results that demonstrate the robustness of the approach of nonlinear dynamics in studying equilibria and their stability properties. On the other hand, plotting the profit functions resulting from our modified model we show that its behavior is more realistic than that of other models reported in the literature.

Keywords: economic dynamics, nonlinear dynamics, stability of equilibrium points.

2000 MSC: 91B62, 93A30, 70K20.

1. INTRODUCTION

People constantly want to buy products, choose amongst those which they find more appropriate to their needs and finally pay the price requested. How many of us, however, really understand how the market works? Do persons play a vital role in the determination of the prices for each product, or are they simply recipients of the flow? How can one attack these problems mathematically and how close are the mathematical findings to reality?

Depending on the product a market is born. The number of firms activated in that market depends on the easiness to achieve profitability, the measures followed by Government, the strength of the market itself and many other possibly minor factors. An immediate question is: how do these firms decide how much to produce and at what price? When this question is answered, production takes place and depending upon the evolution of the market new firms enter, while others collapse and exit the market.

In economic theory two extremes are usually studied first: Monopoly, where there is just one firm (most of the times due to government intervention) and the so-called perfect (complete) competition. Somewhere in between lies the case of *oligopoly*, that is few firms (oligo) selling (polo) products, which may be either identical or differentiated.

The theory of oligopoly is an active field of research and has attracted much attention through the last decades. This paper first presents an introduction to oligopoly theory with emphasis on two firms (duopoly) for reasons of simplicity and then focuses on the resulting dynamics, which is of particular interest. Currently, one can find many research papers, in both mathematical and economical journals, approaching the problem from the perspective of nonlinear dynamics based on some 'reasonable' assumptions.

One such approach was followed recently by Matsumoto and Szidarovszky (MS) in their paper [6]. They postulated a particular dependence of prices on production levels and proposed a dynamical system based solely on deviations from equilibrium configurations. In the present paper, we begin from a more fundamental set of differential equations describing the dynamics and also introduce more realistic price functions. What we find is that the resulting equilibria are distributed in much the same way as in the MS approach and they are all stable. However, the corresponding profit values at these equilibria are distributed very differently on the parameter plane and suggest that our approach is more 'natural' than the one followed by MS.

2. FORMULATION OF A REASONABLE MODEL

We start with a number of assumptions: Consider the case of two firms with the restriction that no other firm can enter the market in the future. Both firms produce the same product, which cannot be stored and sold later. The process takes place in finite time and the firms have no information on the other's actions/choices and do not cooperate.

Cournot [4] assumed that firms choose their output level (productivity) first and then the market sets the price straightforwardly, based on a demand curve and the total quantity offered. The quantity of production chosen by one firm affects the profit of all others including itself and assumptions are made by each firm, regarding the output of all the others.

A set of quantities sold for which, holding the quantities of all other firms constant, no firm can obtain a higher profit by choosing a different quantity is called a Cournot (Nash) equilibrium. At such an equilibrium, no firm wants to change its behaviour. Each firm is on its best-response curve and attains its maximal profit, given that it has the correct information about its rivals' output.

Stackelberg [12] and Bertrand [3] formulated other models originating from different beliefs concerning the behaviour of the market. The former studied the case of a firm setting its output level first, followed by the other firms, assuming Cournot's ideas for the determination of the equilibrium point, while the latter postulated that firms firstly choose prices and then let the quantities of the product be designated by the market.

In this paper, we study Cournot behaviour in a market where two firms are active: firm 1 and firm 2, which sell the identical product at quantities x and y, respectively, at the same price. The objective functions, which represent the overall profit (or loss) of each firm, are

$$u_1 = \{p(x,y) - c_1(x,y)\}x \quad \text{and} \quad u_2 = \{p(x,y) - c_2(x,y)\}y, \tag{1}$$

where $p(x,y)$ is the price requested per item x or y sold and $c_i, i = 1,2$, are the cost functions for each firm. The first-order conditions for the computation of the Cournot equilibrium point are

$$\left\{\frac{\partial p}{\partial x} - \frac{\partial c_1}{\partial x}\right\} x + \{p(x,y) - c_1(x,y)\} = 0 \ \text{ and}$$

$$\left\{\frac{\partial p}{\partial y} - \frac{\partial c_2}{\partial y}\right\} y + \{p(x,y) - c_2(x,y)\} = 0 \Leftrightarrow \qquad x = X(y) \ \text{ and } y = Y(x). \tag{2}$$

The solution of the simultaneous system, (2), yields the Cournot (Nash) equilibrium point.

2.1. INVERSE DEMAND FUNCTION

The choice of an inverse demand function is of obvious fundamental importance. Driven by economic reasoning, one expects the price to vary as follows: for small demand, the price remains approximately the same, while, as demand increases, the slope decreases steeply and then levels off and falls slowly to zero, as shown in Figure 1.

Puu [7, 8, 9] considered a price function of the form

$$p(x,y) = \frac{1}{x+y}, \tag{3}$$

according to an assumption first made in [5]. Of course, this assumption has a serious drawback due to its singularity as production levels go to zero [1, 2].

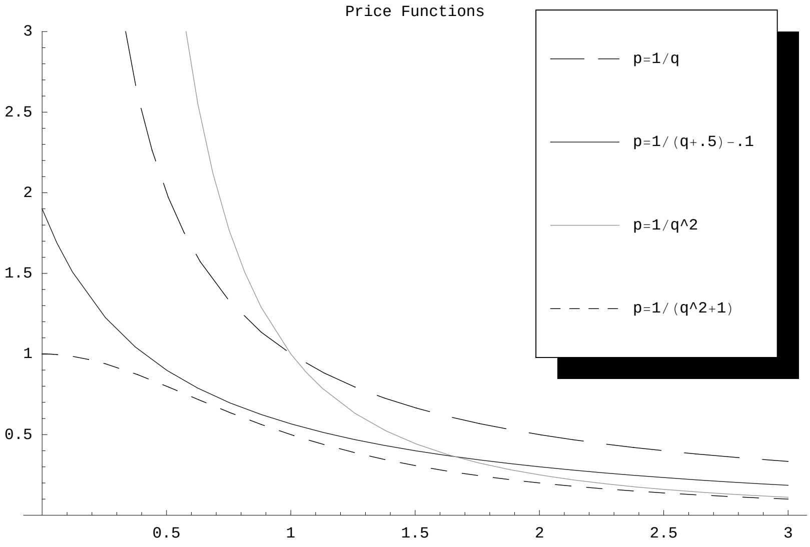


Fig. 1. Four representative choices for the price function vs. output level.

This is 'amended' by considering a price function of the form

$$p + p_0 = \frac{1}{q+q_0}, \quad \text{where} \quad q = x + y. \tag{4}$$

Still, this choice suffers from the fact that a steep price decrease occurs already at very small output levels.

We propose a more realistic price function (see Figure 1) given by the expression

$$p = \frac{1}{q^2+1}, \tag{5}$$

which overcomes the above difficulties.

Note that in the case of differentiated goods, the way this differentiation enters in the corresponding expressions is very important. A simple and realistic way to achieve this is by introducing two parameters, θ_1 and θ_2, as in [6] and by defining two inverse demand functions as follows

$$p_1 = \frac{1}{x+\theta_1 y} \quad \text{and} \quad p_2 = \frac{1}{\theta_2 x + y}, \tag{6}$$

where $0 < \theta_i < 1$ keeping $\mathrm{d}p_1/\mathrm{d}x < 0$ and $\mathrm{d}p_2/\mathrm{d}y < 0$, an assumption first made in [7] and later extended to differentiated goods [6].

We suggest that (6) may be appropriate for the analysis followed in [6], but it is not optimal. For reasons mentioned above, a better choice might be

$$p_1 = \frac{1}{(x+\theta_1 y)^2+1} \quad \text{and} \quad p_2 = \frac{1}{(\theta_2 x + y)^2+1}. \tag{7}$$

2.2. FORMULATION OF A CONTINUOUS DYNAMICAL SYSTEM

We consider the case of constant marginal costs to simplify the calculations and best illustrate the idea. The objective functions for the two firms are

$$u_1(x,y) = p_1(x,y)x - c_1 x \quad \text{and} \quad u_2(x,y) = p_2(x,y)y - c_2 y. \tag{8}$$

Let us formulate a continuous dynamical system modeling the situation of a Cournot duopoly. Matsumoto and Szidarovszky [6] introduced the so-called 'reaction functions', $R_1(y)$ and $R_2(x)$, by solving the first-order conditions for the system (8) with inverse demand functions as given in (6), i.e. $\theta_1 y = c_1(x+\theta_1 y)^2$ and $\theta_2 x = c_2(y+\theta_2 x)^2$ and, solving for x and y, respectively, obtained

$$R_1(y) = \sqrt{\frac{\theta_1 y}{c_1}} - \theta_1 y \quad \text{and} \quad R_2(x) = \sqrt{\frac{\theta_2 x}{c_2}} - \theta_2 x. \tag{9}$$

Hence, they postulated that the differential equations generate the continuous dynamics of the system are

$$\dot{x}(t) = k_1(R_1(y(t)) - x(t)) \quad \text{and} \quad \dot{y}(t) = k_2(R_2(x(t)) - y(t)), \tag{10}$$

where the dot denotes differentiation with respect to time t and k_i, $i = 1, 2$, are some positive constants (adjustment coefficients).

This choice seems rather *ad hoc*, in the sense that there are several ways to construct a dynamical system. In [6] the authors try to express the change of output level depending on the deviation from the Cournot equilibrium point. An alternative approach, which we investigate here, is to assume that the change of output is proportional to the *rate of change of profits* with respect to the production level, according to the equations

$$\dot{x}(t) = k_1 \frac{\partial u_1}{\partial x} \quad \text{and} \quad \dot{y}(t) = k_2 \frac{\partial u_2}{\partial y} \tag{11}$$

or, explicitly for the case of (6) and (8),

$$\dot{x}(t) = k_1 \left(\frac{\theta_1 y}{(x + \theta_1 y)^2} - c_1 \right) \quad \text{and} \quad \dot{y}(t) = k_2 \left(\frac{\theta_2 x}{(y + \theta_2 x)^2} - c_2 \right). \tag{12}$$

2.3. CHOICE OF A COST FUNCTION

Marginal costs determine the dependence of cost functions on amounts of productivity and, in most research papers, are assumed to be constant. This makes the analysis much easier, but lacks sufficient economical justification. Other choices include the introduction of capacity constraints for each firm as proposed, for example, by Puu and Norin [9] and Puu [10, 11]. More specifically, Puu and Norin [9] introduce as *total production cost functions* the expressions $-\ln(1 - x/v_1)$ for the first firm and $-\ln(1 - y/v_2)$ for the second, where v_1 and v_2 are the capacity limits for firm 1 and 2, respectively. This is a reasonable assumption since zero production levels result in zero cost, increasing output levels increase the total cost and finally, as production reaches capacity limits, costs go to infinity. These total costs lead to the profit functions

$$\begin{aligned} u_1(x, y) &= x p_1(x, y) + \ln(1 - x/v_1), \\ u_2(x, y) &= y p_2(x, y) + \ln(1 - y/v_2), \end{aligned}$$

see (1). However, this choice turns out to give results similar to those we obtain assuming constant marginal costs (see Sections 3 and 4) and will not be further pursued here.

3. PUTTING THESE IDEAS TO WORK

3.1. THE FORMULATION DUE TO MATSUMOTO AND SZIDAROVSZKY [6]

Before we use the modified price functions (7) let us consider the problem as expressed by our equations (12), involving the price functions introduced in [6]. Thus, the objective functions for the two firms are given by

$$u_1(x,y) = \frac{x}{x+\theta_1 y} - c_1 x, \quad u_2(x,y) = \frac{y}{\theta_2 x + y} - c_2 y. \tag{13}$$

The vanishing of the first-order partial derivatives

$$\frac{\partial u_1}{\partial x} = 0 \quad \text{and} \quad \frac{\partial u_2}{\partial y} = 0 \tag{14}$$

determines the Cournot equilibrium point by the relations

$$\begin{aligned} \frac{\theta_1 y}{(x+\theta_1 y)^2} - c_1 &= 0, \\ \frac{\theta_2 x}{(\theta_2 x + y)^2} - c_2 &= 0. \end{aligned} \tag{15}$$

In order to study the stability of these equilibrium points, we need to consider the Jacobian matrix associated with the dynamical system (12)

$$J = \begin{bmatrix} k_1 \dfrac{\partial^2 u_1}{\partial x^2} & k_1 \dfrac{\partial^2 u_1}{\partial x \partial y} \\ k_2 \dfrac{\partial^2 u_2}{\partial x \partial y} & k_2 \dfrac{\partial^2 u_2}{\partial y^2} \end{bmatrix}$$

and obtain its eigenvalues from the characteristic equation

$$\lambda^2 + 2A\lambda + B = 0, \tag{16}$$

where

$$\begin{aligned} A &= -\tfrac{1}{2} k_1 \frac{\partial^2 u_1}{\partial x^2} - \tfrac{1}{2} k_2 \frac{\partial^2 u_2}{\partial y^2}, \\ B &= \left(\frac{\partial^2 u_1}{\partial x^2} \frac{\partial^2 u_2}{\partial y^2} - \frac{\partial^2 u_1}{\partial x \partial y} \frac{\partial^2 u_2}{\partial x \partial y} \right) k_1 k_2. \end{aligned} \tag{17}$$

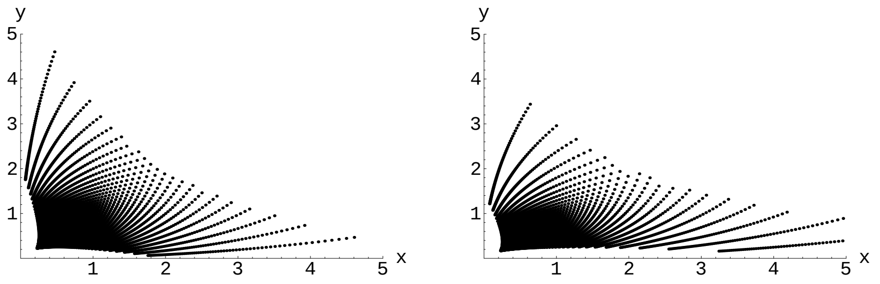


Fig. 2. Stable equilibrium points of the MS model for $\theta_1 = \theta_2 = 0.5$, $k_1 = 1$, $k_2 = 1.1$ (left) and $\theta_1 = 0.7$, $\theta_2 = 0.2$, $k_1 = 0.4$, $k_2 = 1.3$ (right).

The eigenvalues are

$$\lambda_{1,2} = -A \pm \sqrt{A^2 - B} \tag{18}$$

and stability is achieved if and only if both $A > 0$ and $B > 0$ for every equilibrium point.

Following this approach – which we call the MS formulation – Matsumoto and Szidarovszky found that the equilibrium point of (10) is unique and always locally asymptotically stable [6]. Solving (12) numerically, we also find a unique equilibrium point, which is always asymptotically stable, as follows: The eigenvalues (18) evaluated at the fixed points for every choice of parameters in the set $(\theta_1, \theta_2, c_1, c_2)$ have negative real part. We then allow the marginal costs c_1 and c_2 to vary in the interval $(0, 1)$, compute the corresponding equilibrium point and plot it in the (x, y) plane of Figure 2 for two choices of proportionality constants k_1 and k_2.

3.2. USING THE PRICE FUNCTIONS (7)

Consider the following objective functions for the two firms

$$\begin{aligned} u_1(x,y) &= \frac{x}{(x+\theta_1 y)^2+1} - c_1 x, \\ u_2(x,y) &= \frac{y}{(\theta_2 x+y)^2+1} - c_2 y. \end{aligned} \tag{19}$$

The first-order conditions (14) give

$$\frac{1}{(x+\theta_1 y)^2+1} - \frac{2x(x+\theta_1 y)}{[(x+\theta_1 y)^2+1]^2} = c_1, \tag{20}$$

$$\frac{1}{(\theta_2 x + y)^2 + 1} - \frac{2y(\theta_2 x + y)}{[(\theta_2 x + y)^2 + 1]^2} = c_2. \tag{21}$$

In order to specify the reaction functions explicitly, one must solve (19) and (21) for $x = R_1(y)$ and $y = R_2(x)$ analytically, which appears impossible. Assuming that $x = R_1(y)$ and $y = R_2(x)$ we determine the first-order derivatives,

$$\frac{\partial x}{\partial y} = \frac{\partial R_1}{\partial y} \quad \text{and} \quad \frac{\partial y}{\partial x} = \frac{\partial R_2}{\partial x}, \tag{22}$$

by differentiating (19) with respect to y keeping $x = x(y)$ and (21) with respect to x. Thus, we obtain

$$\frac{\partial x}{\partial y} = \frac{\theta_1^2 y(3x^2 - \theta_1^2 y^2 - 1) + 2\theta_1 x(x^2 - 1)}{\theta_1 y(3\theta_1 xy + 2\theta_1^2 y^2 + 2) - x(x^2 - 3)} = F(x, y) = \frac{\partial R_1}{\partial y}, \tag{23}$$

$$\frac{\partial y}{\partial x} = \frac{\theta_2^2 x(3y^2 - \theta_2^2 x^2 - 1) + 2\theta_2 y(y^2 - 1)}{\theta_2 x(3\theta_2 xy + 2\theta_2^2 x^2 + 2) - y(y^2 - 3)} = G(x, y) = \frac{\partial R_2}{\partial x}. \tag{24}$$

Consider now the continuous dynamical system of the form (10), with $k_1 = k_2 = 1$ for simplicity. Its Jacobian matrix is

$$J = \begin{bmatrix} -1 & \dfrac{\partial R_1(y)}{\partial y} \\ \dfrac{\partial R_2(x)}{\partial x} & -1 \end{bmatrix},$$

whose eigenvalues satisfy the characteristic equation

$$\lambda^2 + 2\lambda + 1 - \frac{\partial R_2}{\partial x}\frac{\partial R_1}{\partial y} = 0. \tag{25}$$

Again, as in the previous subsection, we locate all equilibrium points, (x_e, y_e), of the system, for various choices of the parameter values, solving (20) and (21) numerically. Note that, combining (22), (23) and (24), the characteristic equation (25) takes the form

$$\lambda^2 + 2\lambda + 1 - G(x_e, y_e)F(x_e, y_e) = 0 \tag{26}$$

whose roots are given by

$$\lambda_\pm = -1 \pm \sqrt{G(x_e, y_e)F(x_e, y_e)}. \tag{27}$$

Figure 3 shows all the equilibrium points for c_1, c_2 choices ranging from 0 to 1. Clearly all (c_1, c_2) pairs leading to $x_e < 0$ or $y_e < 0$ are rejected as being 'unphysical'.

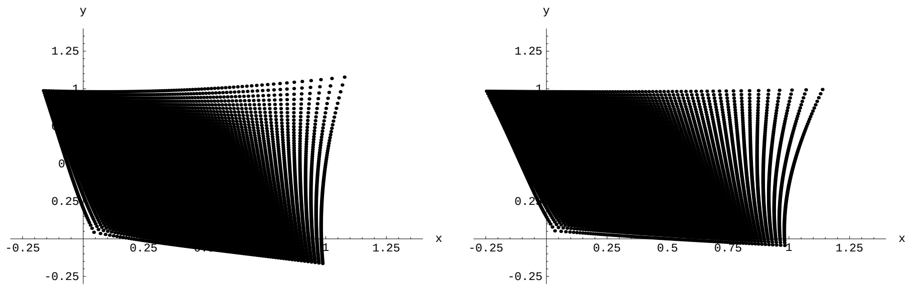


Fig. 3. Stable equilibrium points of our system for $\theta_1 = \theta_2 = 0.5$, $k_1 = 1$, $k_2 = 1.1$ (left) and $\theta_1 = 0.7$, $\theta_2 = 0.2$, $k_1 = 1$, $k_2 = 1.1$ (right). Compare with Figure 2.

The important result is that our approach, utilising (7) and (11), also yields that the equilibrium points, (x_e, y_e), are *all stable* and have an analogous distribution in the x, y plane as in the case of the formulation described in subsection 3.1.

4. A STUDY OF THE PROFIT FUNCTIONS

Having determined the Cournot equilibrium points and their stability, firms naturally worry about their profits. Provided that in the short term a given firm is in a position to change its parameters and/or initial conditions, the two firms try to maximise their profits, u_1 and u_2, as in (13) or (19). Let us now examine how these profits evaluated at the various equilibrium points vary as functions of the parameters c_1 and c_2:

Following the MS formulation, we plot in Figure 4 for each equilibrium point shown in Figure 2 the corresponding u_1 and u_2 values for the profits of firms 1 and 2, respectively. Observe that these profit pairs lie on two nonintersecting curves, indicating that the MS approach exhibits a rather restrictive and unrealistic distribution of profit values. It seems unnatural to expect that such an extensive variation of the c_1 and c_2 parameters would yield profits belonging to so limited a (1-dimensional) subset of the (u_1, u_2) plane.

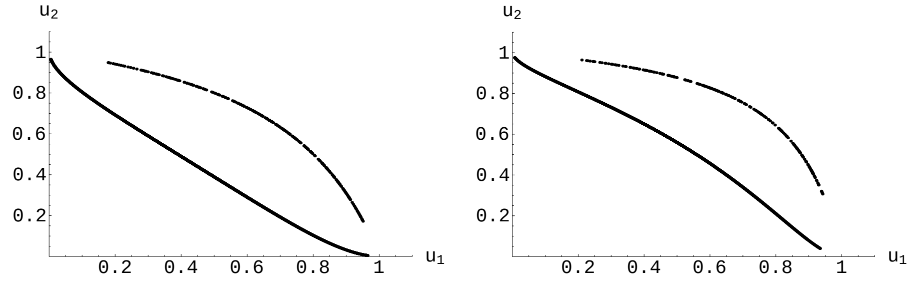


Fig. 4. Profit distributions for the MS model, (13), at $\theta_1 = \theta_2 = 0.5$ (left) and $\theta_1 = 0.7$, $\theta_2 = 0.2$ (right).

By contrast, our approach for the corresponding c_1, c_2 values leads to a distibution of profits that occupies a significant (2-dimensional) subset of the (u_1, u_2) plane, as shown in Figure 5.

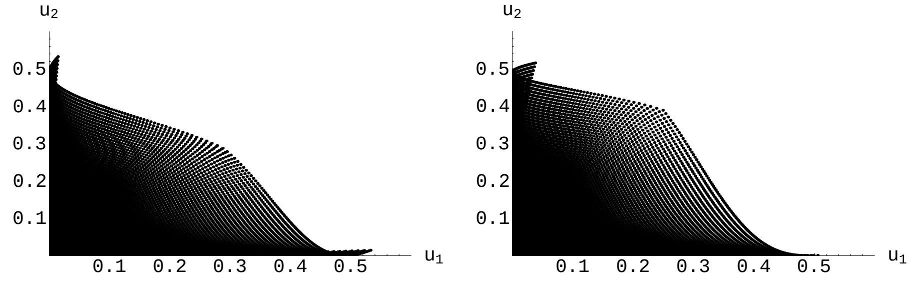


Fig. 5. Profit distributions for our formulation (19) at $\theta_1 = \theta_2 = 0.5$ (left) and $\theta_1 = 0.7$, $\theta_2 = 0.2$ (right).

We also plot profits vs. prices for each firm separately in Figure 6. We find that, at $\theta_1 > \theta_2$, the profits of firm 2 are considerably *higher* than those of firm 1 for a large set of c_1 and c_2 parameters corresponding to prices of firm 1 that are *lower* than those of firm 2, i.e. $p_1 < p_2$.

We believe that this is an interesting observation. It implies, at least for the parameter values studied, that a firm (here firm 2) which pays less attention to the level of production of its rival (here firm 1) and consequently sets its prices with $\theta_2 < \theta_1$, achieves *higher profits* when its products are more expensive than those of its rival.

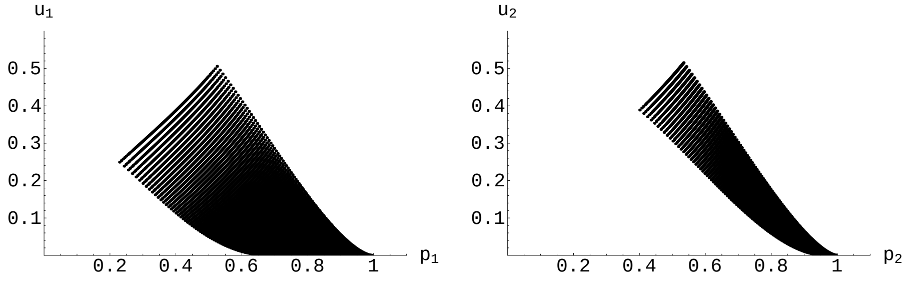


Fig. 6. Profits vs. prices resulting from our approach (19) with $\theta_1 = 0.7$, $\theta_2 = 0.2$ for firm 1 (left) and firm 2 (right).

5. CONCLUDING REMARKS

The operation of a free market is a complex problem whose time evolution depends on many factors and generally relies on choices made by humans. Still, in a market that is well-developed, it appears that competing firms operate on a basis of well-defined principles whose effectiveness has been validated by experience. It follows, therefore, that it would be useful to analyse free market operation using deterministic models of nonlinear dynamics (like systems of nonlinear ordinary differential equations) to describe firm competition.

In this paper, we have adopted such a dynamical approach for a market consisting of two firms producing a single, but differentiated, product. Following a formulation suggested by Matsumoto and Szidarovszky (MS) we have performed a detailed study of a model that uses more realistic price functions and a dynamical system derived from more fundamental principles. Our first conclusion is that in terms of existence and stability of equilibria our approach yields results very similar to MS, demonstrating the robustness of these duopoly models.

However, extending our study to the evaluation of the associated price functions we have found that the results of our model are richer and, therefore, more realistic than those of MS. Furthermore, an important observation can be made concerning the choice of the two parameters, θ_1 and θ_2, that reflect the degree to which each firm takes into account the output level of the other firm, when setting the price of its own product. We have found that if firm

2 sets its prices with less regard for the output level of firm 1, i.e. $\theta_2 < \theta_1$, its profits are higher over the full regime where the prices of firm 1 are lower than those of firm 2.

These results may lead to the conclusion that a two-firm market has simple dynamics which in all cases we examined is attracted by a stable equilibrium point. Clearly, to observe more complicated behavior, one has to go beyond duopoly models and investigate systems of $n \geq 3$ competing firms.

Acknowledgements

We thank Stelios Dimas for his assistance with the computations and the State (Hellenic) Scholarships Foundation (IKY) for partial financial support for KA throughout this work.

References

[1] Agiza, H. N., *Explicit stability zones for Cournot game with* 3 *and* 4 *competitors*, Chaos, Solitons and Fractals, **9**(1998), 1955-1966.

[2] Ahmed, E., Agiza, H. N., *Dynamics of a Cournot game with n-competitors*, Chaos, Solitons and Fractals, **9**(1998), 1513-1517.

[3] Bertrand, J., *Journal des Savants*, (1883, September).

[4] Cournot, A. A., *Recherches sur les principes mathématiques de la théorie des richesses*, Hachette, Paris, 1838.

[5] Dana, R., Montrucchio, L., *Dynamic complexity in duopoly games*, Journal of Economic Theory, **40**(1986), 40-56.

[6] Matsumoto, A., Szidarovszky, F., *Delayed nonlinear Cournot and Bertrand dynamics with product differentiation*, Nonlinear Dynamics, Psychology and Life Sciences, **11**(2007), 367-395.

[7] Puu, T., *Chaos in duopoly pricing*, Chaos, Solitons and Fractals, **1**(1991), 573-581.

[8] Puu, T., *The chaotic duopolists revisited*, Journal of Economic Behavior and Organization, **33**(1998), 385-394.

[9] Puu, T., Norin, A., *Cournot duopoly when the competitors operate under capacity constraints*, Chaos, Solitons and fractals, **18**(2003), 577-592.

[10] Puu, T., *The layout of a new industry: from oligopoly to competition*, PU.M.A., **4**(2005), 1-18.

[11] Puu, T., *On the stability of the Cournot equilibrium when the number of competitors increases*, Journal of Economic Behavior and Organization, **66**(2008), 445-456.

[12] Stackelberg, H. von, *Marktform und Gleichgewicht*, Springer, Vienna, 1934.

ROMAI J., **4**, 1(2008), 61–67

ALGEBRAS OF CONTINUOUS FUNCTIONS AND COMPACTIFICATIONS OF SPACES

Mitrofan M.Choban, Laurenţiu I.Calmuţchi

Tiraspol State University, Chişinău, Republic of Moldova

mmchoban@email.ro

Keywords: g-compactification, compactification, algebra of functions.

2000 MSC: 54D35, 54D40.

1. INTRODUCTION

The notion of compactness is one of the most important notions.

A generalized compactification or a g-compacification of a space X is a pair (Y, e_Y), where Y is a compact Hausdorff space and $e_Y : X \to Y$ is a continuous mapping such that the set $e_Y(X)$ is dense in Y. If e_Y is an embedding, then Y is called a compactification of X, we identify $x = e_Y(x)$ for any $x \in X$ and consider that $X \subseteq Y$.

Let (Y, e_Y) and (Z, e_Z) be two g-compactifications of a space X. We consider that $(Y, e_Y) \leq (Z, e_Z)$ if there exists a continuous mapping $f : Z \to Y$ such that $e_Y = f \circ e_z$, i.e. $e_Y(x) = f(e_Z(x))$ for any $x \in X$. If $(Y, e_Y) \leq (Z, e_Z)$ and $(Z, e_Z) \leq (Y, e_Y)$, then f is a homeomorphism and we say that the g-compactifications (Y, e_Y), (Z, e_Z) are equivalent. We identify the equivalent g-compactifications. In this case the set $GC(X)$ of all g-compactifications of the space X is a complete lattice with the maximal element $(\beta X, \beta_X)$. The minimal element of the lattice $GC(X)$ is the one-point space. The compactification βX is the Stone-Čech compactification of the space X.

Let L be a non-empty subset of the lattice $GC(X)$. Then the maximal element $\vee L$ and the minimal element $\wedge L$ are determined in $GC(X)$. Problems connected with the g-compactifications $\vee L$, $\wedge L$ are among the most interesting problems of the topology.

We use the terminology from [3]. Denote by $|A|$ the cardinality of the set A, by $cl_X A$ or clA the closure of the set A in the space X.

2. PARTIAL ALGEBRAS

Denote by K the ring of complex numbers.

A set A is called a partial algebra if for some pairs $x, y \in X$ there are determined the sum $x + y$ and the multiplication $x \cdot y$ such that:

1. the multiplication and sum are associative, commutative and distributive.

2. for every $x \in A$ and $\alpha \in K$ it is defined $\alpha X \in A$ such that:

2.1. $\alpha(x + y) = \alpha x + \alpha y$ for all $x, y \in A$ and $\alpha \in K$;

2.2. $(\alpha + \beta)x = \alpha x + \beta x$ for all $x \in A$ and $\alpha, \beta \in K$;

2.3. $1 \cdot x = x$ for any $x \in A$;

3. $\alpha(\beta x) = (\alpha\beta)x$ for all $x \in A$ and $\alpha, \beta \in K$;

4. there exist two distinct elements $0, 1 \in A$ such that $0 \cdot x = 0$ and $0 + x = 1 \cdot x = x$ for any $x \in X$.

Let A be a partial algebra. A subset $I \subseteq A$ is called an ideal of A if: $I \neq \emptyset$; $x + y \in I$ provided $x, y \in I$ and $x + y$ is defined in A; $x \cdot y \in I$ provided $x \in I$ and $x \cdot y$ is defined in A; if $\alpha \in K$ and $x \in I$, then $\alpha x \in I$. A maximal ideal of A is called a proper ideal of A if it is not contained in other proper ideal of A. Let $M(A)$ be the set of maximal ideals of A. For every $x \in A$ we put $M(x, A) = \{I \in M(A) : x \in I\}$. Then $M(x_1, ..., x_n, A) = \cap\{M(x_i, A) : i \leq n, n \in \mathcal{N} = \{1, 2, ...\}\}$. If $L \subseteq A$, then $M(L, A) = \cap\{M(x, A) : x \in L\}$.

The family $\{M(L, A) : L \subseteq A\}$ is a closed basis of the topology of the space $M(A)$.

Theorem 2.1. *The ideal space $M(A)$ is a compact T_1-space.*

Proof. Let $\{M(L_\lambda, A) : \lambda \in \Gamma\}$ be a given family of closed sets and $\cap\{M(L_\lambda, A) : \lambda \in P\} \neq \emptyset$ for any non-empty finite subset $P \subseteq \Gamma$. Assume that: for every two elements $\alpha, \beta \in \Gamma$ there exists $\gamma \in \Gamma$ such that $M(L_\gamma, A) \subseteq M(L_\alpha, A) \cap M(L_\beta, A)$; if the set $P \subseteq A$ is finite and $M(P, A) \cap M(L_\alpha, A) \neq \emptyset$ for every $\alpha \in \Gamma$, then $M(P, A) = M(L_\beta, A)$ for some $\beta \in \Gamma$.

We put $I = \cup\{L_\lambda : \lambda \in \Gamma\}$. Then $I \in M(A)$ and $I \in \cap\{M(L_\alpha, A) : \alpha \in \Gamma\}$. It is obvious that $M(A)$ is a T_1-space. The proof is complete.

3. SPECIAL COMPACTIFICATION OF THE FIELD *K*

Let $C = \{z \in K :\mid z \mid = 1\}$. For every $z \in C$ we fix an improper number ∞_z such that:

1. $\infty_z \neq \infty_x$ provided $x \neq z$;

2. if $x, y \in C,\ x \neq -y$ and $z = \alpha(x+y)$ for some positive number α, then we consider that $\infty_z = \infty_x + \infty_y$;

3. if $x, y \in C$ and $z = xy$, then $\infty_z = \infty_x \cdot \infty_y$;

4. $x + \infty_y = \infty_y + x$ for any $x \in K$ and $y \in C$;

5. if $x \in K, y \in C, z \in C$ and $z = \lambda xy$ for some positive λ, then $\infty_z = x \cdot \infty_y = \infty_y \cdot x$;

6. $0 \cdot \infty_z = \infty_z \cdot 0 = 0$ for any $z \in C$.

We put $\bar{K} = K \cup \{\infty_z : z \in C\}$ and $\Omega = \{\infty_z : z \in C\}$. There exists an one-to-one mapping $\varphi : \bar{K} \rightarrow B = \{x \in K :\mid x \mid \leq 1\}$ such that $\varphi(0) = 0, \varphi(\infty_z) = z$ for any $z \in C, \varphi(x) = x \cdot (1+ \mid x \mid)$ for any $x \in K$. On $\bar{K}$ we consider the topology with respect to which φ is a homeomorphism. Then $\bar{K}$ is a compactification of the space K and $\bar{K}$ is a partial algebra. If $x, y \in C$ and $x = -y$, then $\infty_x + \infty_y$ is not determined. We consider that $-\infty_x = \infty_{(-x)}$. If R is the field of reals, then $+\infty = \infty_1$, $-\infty = \infty_{(-1)}$ and $R \cup \{-\infty, +\infty\} \subseteq K$. We put $\bar{R} = R \cup \{-\infty, +\infty\} = [-\infty, +\infty]$. Thus $\bar{R}$ is a compactification of the reals and $\bar{R}$ is homeomorphic to the closed interval.

4. PARTIAL ALGEBRAS OF FUNCTIONS

Fix a topological space X. Denote by $C(X, \bar{K})$ the set of all continuous functions of X into $\bar{K}$ and $C^0(X, K) = \{f \in C(X, K) : f(X)$ is a bounded subset of $K\}$. If $f \in C^0(X, K)$, then $cl_K f(X)$ is a compact subset.

A function $f \in C^0(X, K)$ possesses a compact support if there exists a compact subset $F \subseteq X$ such that $f^{-1}(K \backslash \{0\}) \subseteq F$. In this case $supp(f) = cl_X f^{-1}(K \backslash \{0\})$ is a compact subset. Let $C_0(X, K) = \{f \in C(X, K) : supp(f)$ is a compact set$\}$.

Definition 4.1. *A subset $A \subseteq C(X, \bar{K})$ is an algebra of functions on X if*:

- $0 \in A$, *where* $0(x) = 0$ *for any* $x \in X$;

- *if f is a constant function, then $f \in A$;*
- *if $f, g \in A$ and $f + g \in C(X, \bar{K})$, then $f + g \in A$;*
- *if $f, g \in A$ and $f \cdot g \in C(X, \bar{K})$, then $f \cdot g \in A$;*
- *if $f \in A$, then $-f \in A$ and $\bar{f} \in A$;*
- *if $f \in A$, then $\lambda f \in A$ for any $\lambda \in K$.*

For every subset $L \subset C(X, \bar{K})$ there exists a minimal algebra $a(L)$ generated by the set L.

Definition 4.2. *Let (Y, e_Y) be a g-compactification of a space X. Then $C(X, K, Y, e_Y) = \{f \circ e_y : f \in C(X, \bar{K})\}$ is called the algebra of functions on X continuously extendable on Y.*

Theorem 4.3. *Let X be a topological space $X_0 = \cup\{U : U$ is open in X and $cl_X U$ is a compact Hausdorff subspace$\}$* and $L \subseteq C(X, \bar{K})$. *Then there exists a unique g-compactification (Y, e_Y) of the space X with the following properties:*

1. *every function $f \in L$ is continuously extendable on Y, i.e. there exists a unique continuous function $ef \in C(Y, K)$ such that $f = ef \circ e_Y$.*

2. *if $y_1, y_2 \in Y \backslash e_y(X_0)$ and $y_1 \neq y_2$, then there exists $f \in L$ such that $ef(y_1) \neq ef(y_2)$.*

3. *if the set U is open in X and $cl_X U$ is a Hausdorff compact subset of X, then $e_y(U)$ is an open subset of Y and $e_Y \mid U : U \to e_Y(U)$ is a homeomorphism.*

Proof. If X is a compact Hausdorff space, then $Y = X$. Suppose that the space X is not a compact Hausdorff space. Let $\{(x_\alpha, F_\alpha) : \alpha \in A\}$ be the set of all pairs (x, F), where $x \in X_0, F$ is a closed subset of X and $x \notin F$. For every $\alpha \in A$ fix a continuous function $\varphi_\alpha : X \to [0, 1] \subseteq K$ such that $\varphi_\alpha(x_\alpha) = 0$ and $F_\alpha \cup (X \backslash X_0) \subseteq \varphi_\alpha^{-1}(1)$. Let X_α be the closure of the set $\varphi_\alpha(X)$ in K. Then $(X_\alpha, \varphi_\alpha) \in GC(X)$.

For every $f \in L$ denote by Y_f the closure of the set $f(X)$ in $\bar{K}$. Then $(Y_f, f) \in GC(X)$.

If $X_0 \cup L = \emptyset$, then (Y, e_Y) is the one-point g-compactification.

Suppose that $X_0 \cup L \neq \emptyset$.

Denote by (Y, e_Y) the minimal g-compactification with the following properties:

- $(Y, e_Y) \geq (X_\alpha, \varphi_\alpha)$ for any $\alpha \in A$;

- $(Y, e_Y) \geq (Y_f, f)$ for any $f \in L$.

By construction, the functions from $L \cup \{\varphi_\alpha : \alpha \in A\}$ are continuously extendable on (Y, e_Y). Hence, $e_Y \mid Z : Z \to e_Y(Z)$ is a homeomorphism. If $\alpha \in A$ and $e\varphi_\alpha$ is the extension of φ_α on Y, then $x_\alpha \alpha \in e\varphi_\alpha^{-1}[0, 1)$ and $e\varphi_\alpha^{-1}[0, 1)$ is an open subset of Y. Thus $e_Y(X_0)$ is an open subset of Y.

Thus (Y, e_Y) satisfies all conditions of Theorem 4.3.

Fix $(S, e_S) \in GC(X)$ with the properties of the Theorem 4.3. Then the functions from $L \cup \{\varphi_\alpha : \alpha \in A\}$ are continuously extendable on S and $(S, e_S) \geq$ $sup(\{(Y_f, f) : f \in L\} \cup \{(X_\alpha, \varphi_\alpha) : \alpha \in A\}) = (Y, e_Y)$ and there exists a continuous mapping $g : S \to Y$ such that $e_Y = g \circ e_S$.

We affirm that g is a homeomorphism and the g-compactifications (Y, e_Y), (S, e_S) are equivalent. Suppose that g is not a homeomorphism. Then there exists two distinct points $s, s_2 \in S$ such that $g(s_1) = g(s_2)$.

If $x_1 \in X \setminus X_0$ and $x_2 \in X\ X_0$, then $\varphi_\alpha(x_1) \neq \varphi_\alpha(x_2)$ provided $x_1 = x_2$. In this case $e_Y(x_1) = e_Y(x_2)$. Therefore $s_1, s_2 \in S \backslash e_S(X_0)$. Let ef be the continuous extension of the function $f \in L$ on (S, e_S). Thus $ef(s_1) = ef(s_2)$ for every $f \in L$, a contradiction with the condition 2. The proof is complete.

Corollary 4.4 *Let X be a locally compact Hausdorff space and $L \subseteq C(X, K)$. Then there exists a unique compactification Y of the space X such that:*

1. *every function $f \in L$ is continuously extendable on Y;*

2. *if $y_1, y_2 \in Y \setminus X$ and $y_1 \neq y_2$, then there exists $f \in L$ such that $ef(y_1) \neq ef(y_2)$;*

3. *X is an open and dense subspace of the space Y.*

Theorem 4.5. *Let X be a topological space, $X_0 = \cup\{U : U$ is open in X and $cl_X U$ is a compact Hausdorff subspace$\}$, $L \subseteq C(X, K)$ and $L \cup \{f \in C(X, K) : supp(f) \subseteq X_0\}$. Then the maximal ideal space $M(\bar{L})$ of the algebra L is the g-compactification with the properties from Theorem 4.3.*

Proof. Let (Y, e_f) be the g-compactification from Theorem 4.3. For every

$f \in I$ denote by ef the continuous extension of f on (Y, e_Y). We consider that $e_Y(Z) = Z \subseteq Y$. From the Stone-Weierstrass theorem ([3], Theorem 3.2.21) the algebra $\{ef : f \in \bar{L}\}$ is dense in the Banach algebra $C(Y, K)$ of all continuous functions of Y into K. The maximal ideal spaces $C(\bar{L})$ and $M(C(Y, K))$ are homeomorphic to the space Y [4]. The proof is complete.

Remark 4.6 For the Riemanian surfaces X and functions $L \subseteq C(X, \bar{R})$ the Corollary 4.4 was proved by C. Constantinescu and A. Cornea in [2], while for any locally compact X Hausdorff space, by M. Brelot [1].

Remark 4.7. The set $L \subseteq C^0(X, K)$ and the subalgebra $I \subseteq C^0(X, K)$ generate the same g-compactification of the Brelot-Constantinescu-Cornea type.

Let $A(X, K)$ be the set of all closed subalgebras of the Banach algebra $C^0(X, K)$ with the sup-norm $||f|| = sup\{|f(x)| : x \in X\}$. Then there exist an one-to-one correspondence $k : A(X, K) \rightarrow GC(X)$ and a mapping $c : A(X, K) \rightarrow GC(X)$ such that:

1. if $A \in A(X, K)$ and $k(A) = (Y, e_Y)$, then $A = \{f \circ e_Y : f \in C(Y, K)\}$ and Y is the maximal ideal space of the algebra A;

2. $A, B \in A(X, K)$, then $A \subseteq B$ iff $k(A) \leq k(B)$;

3. if $A \in A(X, K)$ and $c(A) = (Y, e_Y)$, then (Y, e_Y) is the g-compactification of the Brelot-Constantinescu-Cornea type generated by the algebra A;

4. by construction, $k(A) \leq c(A)$ for any algebra $A \in A(X, K)$;

5. $c(A(X, K))$ is the set of all g-compactifications of the Brelot-Constantinescu-Cornea type;

6. if X is a locally compact Hausdorff space, then $c(A(X, K))$ is the set of all compactifications of the space X;

7. if $|X| \leq 1$, then $c = k$.

Example 4.8. Let X be a locally compact non-compact Hausdorff space and A be the algebra of constant functions. Then $c(A)$ is the one-point Alexandroff compactification of X and $k(A)$ is the one-point minimal g-compactification of X.

Example 4.9. Let X be the space of reals and $A = \{f \in C^0(X, K) : [-3, 3] \subseteq f^{-1}(0)$. Then $c(A)$ is the Stone-Čech compactification βX of X and

$k(A) = (Y, e_Y)$ is a g-compactification. In this case $e_Y([-3, 3])$ is an one-point subset of the space Y.

Example 4.10. Let $X = [0, 1)$ be endowed with the topology generated by the open base $\{X \cap [x, x+\epsilon) : x \in X, \epsilon > 0\}$, Y be the space $[0, 1]$ in the natural topology, $e_Y(x) = x$ for any $x \in X$ and $A = C^0(Y, K) \subseteq C^0(X, K)$. Then (Y, e_Y) is a g-compactification of the space X and $c(A) = k(A) = (Y, e_Y)$. In this case $c = k$.

References

[1] M. Berlot, *On topologies and boundaries in potential theory*, Springer, Berlin, 1971.

[2] C. Constantinescu, A. Cornea, *Ideale Raunder Riemannscher Flachen*, Ergeb. Math., **32**, Springer, Berlin, 1963.

[3] R. Engelking, *General topology*, PWN, Warszava, 1977.

[4] I. M. Gelfand, A. N. Kolmogoroff, *On rings of continuous functions on topological spaces*, Doklady Akad. Nauk SSSR, **22**(1939), 11-15.

ROMAI J., **4**, 1(2008), 69–84

ABOUT SOME PARTICULAR CLASSES OF BOUNDED OPERATORS ON PSEUDO-HILBERT SPACES

Loredana Ciurdariu

"Politehnica" University of Timişoara

cloredana43@yahoo.com

Abstract In this paper we define n-quasicontractions, n-quasi-isometries, n-quasi-hyponormal and power bounded operators on pseudo-Hilbert spaces, give some properties and show that a $(T^*T)^\alpha$-contraction with $0 < \alpha < 1$ is a power bounded operator. Some basic properties of a quasi-isometry on pseudo-Hilbert spaces are investigated too.

Keywords: Loynes spaces, gramian p-hyponormal operators.

2000 MSC: 47A20, 44A66.

1. Introduction

We recall first that R. M. Loynes defined in [9] and [10] the notions of VE-spaces and VH-spaces (or LVH-spaces) respectively, as generalizations of prehilbertian and Hilbert spaces, respectively. The name ”pseudo-Hilbert spaces” for LVH-spaces appears first in [14] and later these spaces was called ”Loynes spaces” in [3] and [15]. These spaces are important in the study of the stochastic processes.

Thus the pseudo-Hilbert spaces (Loynes spaces) are characterized by the property of possessing an ”inner product” (gramian) which takes its values in a suitable ordered topological space, instead of a scalar valued inner product [9], [10]. This property is sufficient to ensure the existence of many results known from the Hilbert spaces. But, a main difference is given to the failure of the Riesz representation theorem. Another difference is the absence of the classical Cauchy-Schwarz inequality, because the product $[x,x]\cdot[y,y]$, $x, y \in Z$

does not exist, and the existence of the gramian adjoint and of the gramian projections is not always checked.

Some immediately examples of Loynes Z-spaces are the complex Hilbert spaces with the scalar product and Hilbert C^*-modulus (which are Hilbert modulus on a C^*-algebra).

Herein we define these classes of operators, mentioned in the abstract, on spaces of operators more general than Hilbert spaces, on pseudo-Hilbert spaces and to generalize many properties of them known from the Hilbert spaces.

$\mathcal{B}^*(\mathcal{H})$ is a particular C^*-algebra (see Corollary 1) and this fact allows us to recover many properties from Hilbert spaces in our cases.

Corollary 1, Corollary 2 and Theorem 1 were proved in [6].

A few results concerning the quasi-isometries and quasi-contractions, given in Hilbert spaces in [12] and generalized in [2] as Remark1, Remark 2, Proposition 1, Proposition 2 and Proposition 3 are presented for pseudo-Hilbert spaces. Then a characterization of hyponormal operators from the book of P.Halmos [8] was reverified and given also for n-hyponormal operators in Proposition 6 and Proposition 7.

Gramian hyponormal operators can be obtained from the polar decomposition of a gramian hyponormal invertible operator in Proposition 4 and Proposition 5, using the model from [11].

Theorem 4, Consequence 1 and Theorem 6 are another extensions in Loynes spaces of two results of Aluthge [1] concerning the hyponormality of $\tilde{T}$ defined by using the polar decomposition of a gramian hyponormal and invertible operator.

Definition 1[6] *A locally convex space Z is called admissible in the Loynes sense if the following conditions are satisfied:*

(A.1) *Z is complete;*

(A.2) *there is a closed convex cone in Z, denoted by Z_+, that defines an order relation on Z (that is $z_1 \leq z_2$ if $z_2 - z_1 \in Z_+$);*

(A.3) *there is an involution in Z, $Z \ni z \to z^* \in Z$ (that is $z^{**} = z$, $(\alpha z)^* = \overline{\alpha} z^*$, $(z_1 + z_2)^* = z_1^* + z_2^*$), such that $z \in Z_+$ implies $z^* = z$;*

(A.4) *the topology of Z is compatible with the order (that is there exists a basis of convex solid neighbourhoods of the origin);*

(A.5) *any monotonously decreasing sequence in Z_+ is convergent.*

Remark 1 [6] *A set $C \subset Z$ is called solid if $0 \leq z' \leq z''$ and $z'' \in C$ implies $z' \in C$.*

Example 1[6]*$Z = C$, a C^*–algebra with topology and natural involution.*

Definition 2[6] *Let Z be an admissible space in the Loynes sense. A linear topological space $\mathcal{H}$ is called pre-Loynes Z–space if it satisfies the following properties:*

(L1) *$\mathcal{H}$ is endowed with a Z–valued inner product (gramian), i.e. there exists an application $\mathcal{H} \times \mathcal{H} \ni (h, k) \to [h, k] \in Z$ having the properties:*

(G$_1$) *$[h, h] \geq 0$; $[h, h] = 0$ implies $h = 0$;*

(G$_2$) *$[h_1 + h_2, h] = [h_1, h] + [h_2, h]$;*

(G$_3$) *$[\lambda h, k] = \lambda[h, k]$;*

(G$_4$) *$[h, k]^* = [k, h]$;*

for all $h, k, h_1, h_2 \in \mathcal{H}$ and $\lambda \in \mathbb{C}$;

(L2) *the topology of $\mathcal{H}$ is the weakest locally convex topology on $\mathcal{H}$ for which the application $\mathcal{H} \ni h \to [h, h] \in Z$ is continuous.*

Moreover, if $\mathcal{H}$ is a complete space with this topology, then $\mathcal{H}$ is called a Loynes Z–space.

Example 2 [6] *Let $Z = C$ in Example 1, then Z with $[z_1, z_2] = z_2^* z_1$ is a Loynes-Z space.*

An important result which can be used below is given in the next statement, and was proved in [6].

Let $\mathcal{H}$ and $\mathcal{K}$ be two Loynes Z-spaces.

We recall that [6], [9], [10] an operator $T \in \mathcal{L}(\mathcal{H}, \mathcal{K})$ is called gramian bounded, if there exists a constant $\mu > 0$ such that in the sense of order of Z

holds

$$[Th, Th]_{\mathcal{K}} \leq \mu[h, h]_{\mathcal{H}}, \quad h \in \mathcal{H}. \tag{1.3.1}$$

Denote the class of such operators by $\mathcal{B}(\mathcal{H}, \mathcal{K})$ and $\mathcal{B}^*(\mathcal{H}, \mathcal{K}) = \mathcal{B}(\mathcal{H}, \mathcal{K}) \cap \mathcal{L}^*(\mathcal{H}, \mathcal{K})$.

We also denote the introduced norm by

(1.3.2)

$$\|T\| = \inf \left\{ \sqrt{\mu},\ \mu > 0 \text{ and satisfies } (1.3.1) \right\}.$$

Corollary 1 [6] *The space $\mathcal{B}^*(\mathcal{H}, \mathcal{K})$ is a Banach space, and its involution $\mathcal{B}^*(\mathcal{H}, \mathcal{K})$ in $\mathcal{B}^*(\mathcal{K},\ \mathcal{H})$ satisfies*

$$\|T^*T\| = \|T\|^2, \quad T \in \mathcal{B}^*(\mathcal{H}, \mathcal{K}).$$

In particular $\mathcal{B}^*(\mathcal{H})$ is a C^*–algebra.

Corollary 2 [6] *Let $T : \mathcal{H} \to \mathcal{K}$ be an operator between Loynes Z–spaces $\mathcal{H}$ and $\mathcal{K}$ for which there is T^*, i.e. $T \in \mathcal{L}^*(\mathcal{H}, \mathcal{K})$.*

(a) *If T is surjective, then T^* is injective;*

(b) *If T^* is surjective, then T is injective;*

(c) *If $T \in \mathcal{C}^*(\mathcal{H}, \mathcal{K})$, then $\mathcal{N}(T) = \mathcal{R}(T^*)^{\perp} = \{h \in \mathcal{H} \mid [h, T^*k] = 0,\ k \in \mathcal{K}\}$.*

Theorem 1 [6] *Let $\mathcal{H}$, $\mathcal{K}$ be two Loynes Z–spaces and $T \in \mathcal{L}(\mathcal{H}, \mathcal{K})$. The following assertions are equivalent:*

(a) *the operator $T \in \mathcal{B}(\mathcal{H}, \mathcal{K})$, $\mathcal{R}(T)$ is closed and there exists $T^{-1} \in \mathcal{B}(\mathcal{R}(T), \mathcal{H})$;*

(b) *there exists $\mu, \nu > 0$ such that:*

$$\nu[h, h] \leq [Th, Th] \leq \mu[h, h] \tag{1.3.3}$$

for any $h \in \mathcal{H}$.

If T admits an adjoint, then each condition (a) *or* (b) *ensures the existence of a bounded extension $S \in \mathcal{B}^*(\mathcal{K}, \mathcal{H})$ of T^{-1} with the property $S\big|_{\mathcal{N}(T^*)} = 0$.*

2. The main results

Let Z be an admissible space and $\mathcal{H}$ a Loynes Z- space.

Consider $A \in \mathcal{B}^*(\mathcal{H})$ a fixed positive operator. In the sense of [2], we define also in pseudo-Hilbert spaces a *gramian A-contraction* on $\mathcal{H}$, as being an operator $T \in \mathcal{B}^*(\mathcal{H})$ which satisfies the inequality

$$T^*AT \leq A.$$

As a particular case of the *gramian A-contraction* we obtain

Definition 1 *An operator $T \in \mathcal{B}^*(\mathcal{H})$ is called $(T^*T)^\alpha$- gramian contraction, $0 < \alpha < 1$, if $T^*(T^*T)^\alpha T \leq (T^*T)^\alpha$.*

Definition 2 *An operator $T \in \mathcal{B}^*(\mathcal{H})$ is gramian power bounded if*

$$\sup_{n \in \mathbb{N}} \|T^n\| < \infty.$$

Definition 3 *An operator $T \in \mathcal{B}^*(\mathcal{H})$ is called a gramian n-quasicontraction if $T$is gramian $T^{*n}T^n$- contraction, i.e. $T^*(T^{*n}T^n)T \leq T^{*n}T^n$.*

Definition 4 *An operator $T \in \mathcal{B}^*(\mathcal{H})$ is called a gramian n-quasi-isometry if T is gramian $T^{*n}T^n$- isometry, that is $T^*(T^{*n}T^n)T = T^{*n}T^n$.*

Remark 1 *If $T \in \mathcal{B}^*(\mathcal{H})$ is a gramian n-quasi-isometry, then $\| T \| \geq 1$.*

$T \in \mathcal{B}^*(\mathcal{H})$ and $n \in \mathbb{N}$ fixed implies $T^n \in \mathcal{B}^*(\mathcal{H})$. The conclusion is obvious from

$\| T^{n+1} \|^2 = \| T^{*n+1}T^{n+1} \| = \| T^{*n}T^n \| = \| T^n \|^2$ or $\| T^{n+1} \| = \| T^n \| \leq \| T^n \| \| T \|$.

Remark 2 (i) *The operator $T \in \mathcal{B}^*(\mathcal{H})$ is a gramian n-quasicontraction iff the sequence of operators $\{T^{*m}T^m\}_{m \geq n}$ is decreasing.*

(ii) *If T is a gramian n-quasicontraction then, for all $m > n$, T is a gramian m-quasicontraction too.*

(iii) *If T is a gramian n-quasi-isometry then, for all $m > n$, T is a gramian m-quasi-isometry too.*

Proof: (i) By Definition 3, we can write, $T^{*n+1}T^{n+1} \leq T^{*n}T^n$ and $T^{*n+2}T^{n+2} = T^*(T^{*n+1}T^{n+1})T \leq T^*(T^{*n}T^n)T = T^{*n+1}T^{n+1}$ that is the sequence $\{T^{*m}T^m\}_{m\geq n}$ is decreasing. We used the known inequality, $0 \leq A \leq B \Rightarrow T^*AT \leq T^*BT$, $A, B, T \in \mathcal{B}^*(\mathcal{H})$, see 2.6 pp. 44 [13], for example. Conversely, if $\{T^{*m}T^m\}$ is decreasing for all $m \geq n$ then $T^*(T^{*n}T^n)T = T^{*n+1}T^{n+1} \leq T^{*n}T^n$.

(ii) It is immediate from (i).

(iii) $T^{*n+2}T^{n+2} = T^*(T^{*n+1}T^{n+1})T = T^*(T^{*n}T^n)T = T^{*n+1}T^{n+1}$.

□

Definition 5 *An operator $T \in \mathcal{B}^*(\mathcal{H})$ is gramian quasinormal if $(T^*T)T = TT^*T$.*

This definition, was considered in the case of Hilbert spaces in [2] and in [7] for this notion.

It is easy to see that if T is a gramian quasinormal contraction, then T and T^* are T^*T- contractions and T^*T commutes with T and T^*.

Remark 3 *If T is a gramian n-quasicontraction on $\mathcal{H}$ satisfying $T \mid T^n \mid = \mid T^n \mid T$, then it easily follows that T^n is gramian quasinormal and when T is a gramian n-quasi-isometry one has $T^{*n}T^n = T^{*2n}T^{2n} = (T^{*n}T^n)^2$. Hence T^n is a partial gramian isometry, being a gramian projection.*

Proof: The condition $T \mid T^n \mid = \mid T^n \mid T$ implies $T \mid T^n \mid^2 = \mid T^n \mid^2 T$ and $T^n \mid T^n \mid^2 = \mid T^n \mid^2 T^n$, or exactly $T^n(T^{n*}T^n) = (T^{n*}T^n)T^n$.

By Remark 1 (iii), we have $T^{*n}T^n = T^{*n+1}T^{n+1} = T^*(T^{*n+1}T^{n+1})T = ... = T^{*2n}T^{2n}$ and $(T^{*n}T^n)^2 = (T^{*n}T^n)T^{*n}T^n = \mid T^n \mid^2 T^{*n}T^n = T^{*n} \mid T^n \mid^2 T^n = T^{*2n}T^{2n}$, by $T^{*n} \mid T^n \mid^2 = \mid T^n \mid^2 T^{*n}$.

□

We recall that the modulus of an operator $T \in \mathcal{B}^*(\mathcal{H})$, is defined by $\mid T \mid = (T^*T)^{\frac{1}{2}}$. The existence and the positivity is assured in $\mathcal{B}^*(\mathcal{H})$ by the functional calculus with continuous functions on spectrum [6]. Also $\mid T \mid^{2\alpha} \geq 0$.

Proposition 1 *If* T *is a* $(T^*T)^\alpha$*- gramian contraction with* $0 < \alpha < 1$, *then* T *is a gramian power bounded operator.*

Proof: Writing the condition from the Definition 1, we have

$$0 \leq T^* \mid T \mid^{2\alpha} T \leq \mid T \mid^{2\alpha} .$$

But from,

$$[\mid T \mid^\alpha T^n h, \mid T \mid^\alpha T^n h] = [T^{*n} \mid T \mid^{2\alpha} T^n h, h] \leq [\mid T \mid^{2\alpha} h, h] =$$

$$= [\mid T \mid^\alpha h, \mid T \mid^\alpha h],\ (\forall)\ n \in \mathbb{N},\ \mathrm{h} \in \mathcal{H}$$

we obtain

$$[\mid T \mid T^n h, \mid T \mid T^n h] \leq [\mid T \mid^{1-\alpha} \mid T \mid^\alpha T^n h, \mid T \mid^{1-\alpha} \mid T \mid^\alpha T^n h] \leq$$

$$\leq \| \mid T \mid^{1-\alpha} \|^2\ [\mid T \mid^\alpha T^n h, \mid T \mid^\alpha T^n h] \leq \| \mid T \mid^{1-\alpha} \|^2 \cdot \| \mid T \mid^\alpha \|^2\ [h,h],\ (\forall)\ n \in \mathbb{N},\ \mathrm{h} \in \mathcal{H}.$$

In these inequalities we used that $\mid T \mid^\alpha, \mid T \mid^{1-\alpha} \in \mathcal{B}^*(\mathcal{H})$, which follows from $T \in \mathcal{B}^*(\mathcal{H})$ a C^*-algebra implies $T^*T \in \mathcal{B}^*(\mathcal{H})$ or $\mid T \mid \in \mathcal{B}^*_+(\mathcal{H})$, i.e. $\| \mid T \mid \| \leq M$ or $\mid T \mid \leq MI, 0 < \alpha < 1$ and because x^α is a monotone operator, we have $\mid T \mid^\alpha \leq M^\alpha I$.

Also, by induction, we found, $T^{*n} \mid T \mid^{2\alpha} T^n = T^*(T^{*n-1} \mid T \mid^{2\alpha} T^{n-1})T \leq$ $\leq T^* \mid T \mid^{2\alpha} T \leq \mid T \mid^{2\alpha}$.

Because

$$[T^{n+1}h, T^{n+1}h] = [TT^n h, TT^n h] = [T^{*n}T^*TT^n h, h] = [T^{*n} \mid T \mid^2 T^n h, h] =$$

$= [\mid T \mid T^n h, \mid T \mid T^n h]$, we deduce that

$$[T^{n+1}h, T^{n+1}h] \leq \| \mid T \mid^{1-\alpha} \|^2 \cdot \| \mid T \mid^\alpha \|^2\ [h,h],\ h \in \mathcal{H},$$

i.e. $\| T^n \| \leq \| \mid T \mid^{1-\alpha} \| \cdot \| \mid T \mid^\alpha \|,\ n \geq 1$ that is T is gramian power bounded.

□

Definition 6 *We say that an operator* $T \in \mathcal{B}^*(\mathcal{H})$ *is gramian n-quasi hyponormal, if* $[T^{n+1}h, T^{n+1}h] \geq [T^*T^n h, T^*T^n h]$ *for all* $h \in \mathcal{H}$.

Remark 4 *Using the previous definition, it easily follows that* $\|T^{n+1}\| \geq \|T^*T^n\|$.

Remark 5 *Every gramian n-quasihyponormal operator is gramian m-quasi hyponormal, for all* $m > n,\ m \in \mathbb{N}$.

Proof: It is obvious from,

$$[T^{m+1}h, T^{m+1}h] = [T^{n+1}T^{m-n}h, T^{n+1}T^{m-n}h] \geq$$

$$\geq [T^*T^nT^{m-n}h, T^*T^nT^{m-n}h] = [T^*T^m h, T^*T^m h].$$

□

Proposition 2 *Let* T *be a gramian n-quasi-isometry on Loynes* Z*-space* $\mathcal{H}$ *with* $T^n \neq 0$. *Then the following implications are equivalent:*
(i) $\| T \| = 1$*;*
(ii) T *is gramian n-quasihyponormal.*

Corollary 3 *If* T *is a gramian n-quasi-isometry with* $\|T\| = 1$, *then* $\mathcal{N}(T) \subset \mathcal{N}(T^{*n})$.

If in Definition 6 we take, $n = 0$, then we obtain the definition of gramian hyponormal operators.

Proposition 3 *Let* T *be a gramian n-quasi-isometry with* $n \geq 1$ *on Loynes* Z*-space* $\mathcal{H}$ *and* $T^n \neq 0$. *Then the following statements are equivalent:*
(i) $\| T^n \| = 1$*;*
(ii) T^n *is gramian hyponormal.*

Proof: (ii)⇒(i) T^n being gramian hyponormal, it follows that $\|T^n\|^m = \|T^{nm}\|,\ m \in \mathbb{N}$ by Proposition 2, [5].

T is a gramian n-quasi-isometry so $T^{*n+1}T^{n+1} = T^{*n}T^n$ and from Remark 1, (iii) $T^{*m+1}T^{m+1} = T^{*m}T^m$, $(\forall)\ m > n, m \in \mathbb{N}$. Thus

$$\|T^n\|^2 = \|T^{*n}T^n\| = \|T^{*n+1}T^{n+1}\| = ... = \|T^{*2n}T^{2n}\| = \|T^{2n}\|^2,$$

or $\|T^n\| = \|T^n\|^2$. Using now the hypothesis $T^n \neq 0$, we obtain $\|T^n\| = 1$.

(i)⇒(ii) We calculate

$$[T^nh - T^{n*}T^{2n}h, T^n - T^{n*}T^{2n}h] = [T^nh, T^nh] - [T^{n*}T^{2n}h, T^nh] -$$

$$-[T^nh, T^{n*}T^{2n}h] + [T^{n*}T^{2n}h, T^{n*}T^{2n}h] = [T^nh, T^nh] - 2[T^{2n}h, T^{2n}h] +$$

$$+[T^{n*}T^{2n}h, T^{n*}T^{2n}h] \leq [T^nh, T^nh] - 2[T^{2n}h, T^{2n}h] + \|T^{n*}\|^2\|T^n\|^2[T^nh, T^nh] \leq$$

$$\leq [T^nh, T^nh] - 2[T^{2n}h, T^{2n}h] + [T^nh, T^nh] = 0,$$

by hypothesis (i) and Remark 1.

This implies $T^nh - T^{n*}T^{2n}h = 0$ or $T^n = T^{n*}T^{2n}$.

Therefore, using that $\|T^{2n*}\| = 1$ from Remark 1, we have

$$[T^{n*}h, T^{n*}h] = [T^{2n*}T^nh, T^{2n*}T^nh] \leq \|T^{2n*}\|^2[T^nh, T^nh] = [T^nh, T^nh],$$

which means that T^n is gramian hyponormal.

□

Corollary 4 (i) *If $T \in \mathcal{B}^*(\mathcal{H})$ is a gramian hyponormal operator, then $\mathcal{N}(T) \subseteq \mathcal{N}(T^*)$.*

(ii) *For every gramian hyponormal operator T we have $\mathcal{R}(T^* - \overline{\lambda}I)^\perp \subseteq \mathcal{R}(T - \lambda I)^\perp$.*

(iii) *If $T \in \mathcal{B}^*(\mathcal{H})$ is gramian hyponormal and surjective, then T is invertible and T^{-1} is gramian hyponormal.*

(iv) *Every gramian hyponormal operator T for which $T - \lambda I$ is surjective or $T^* - \overline{\lambda}I$ is injective satisfies $\mathcal{R}(T - \lambda I) \subseteq \mathcal{R}(T^* - \overline{\lambda}I)$.*

Proof: (i) Let $h \in \mathcal{N}(T)$. Now, the gramian hyponormality of T implies that

$$0 \leq [T^*h, T^*h] \leq [Th, Th] = 0$$

that is $T^*h = 0$.

(ii) T gramian hyponormal implies $T - \lambda I$ is gramian hyponormal, so that $\mathcal{N}(T - \lambda I) \subseteq \mathcal{N}(T^* - \overline{\lambda} I)$ or $\mathcal{R}(T^* - \overline{\lambda} I)^{\perp} \subseteq \mathcal{R}(T - \lambda I)^{\perp}$, using that $\mathcal{N}(A) = \mathcal{R}(A^*)^{\perp}$ for every $A \in \mathcal{C}^*(\mathcal{H})$.

(iii) $T \in \mathcal{B}^*(\mathcal{H})$ being surjective, it follows that T^* is injective, see Corollary 1.2.1, [5]. Then $\mathcal{N}(T^*) = \{0\}$ or $\mathcal{N}(T) = \{0\}$ by (i). Thus T is invertible in $\mathcal{B}^*(\mathcal{H})$ and the inverse, T^{-1} is also gramian hyponormal.

(iv) By (ii), we have $\mathcal{R}(T^* - \overline{\lambda} I)^{\perp} \subseteq \mathcal{R}(T - \lambda I)^{\perp}$ or $\mathcal{R}(T - \lambda I) \subseteq \mathcal{R}(T - \lambda I)^{\perp\perp} \subseteq \mathcal{R}(T^* - \overline{\lambda} I)^{\perp\perp} = \mathcal{R}(T^* - \overline{\lambda} I)$, because $\mathcal{R}(T^* - \overline{\lambda} I)$ will be closed (by Theorem 1.3.1, see [6], $T^* - \overline{\lambda} I$ being injective), $\mathcal{N}(T - \lambda I) \oplus \overline{\mathcal{R}(T^* - \overline{\lambda} I)} = \mathcal{H}$ and $\mathcal{R}(T^* - \overline{\lambda} I)^{\perp} \oplus \mathcal{R}(T^* - \overline{\lambda} I)^{\perp\perp} = \mathcal{H}$.

□

As a particular case of the Proposition 3, when $n = 1$, we obtain a generalization of Theorem 2.2 from [12].

Theorem 2 *It T is a gramian quasi-isometry and $\|T\| = 1$, then T is a gramian hyponormal operator.*

The following result generalizes a part of Theorem 2.9 from [12] on Loynes spaces.

Theorem 3 *Let T be a gramian quasi-isometry. If T^* is gramian n-quasihyponormal, then T is gramian normal.*

Proof: Assume that T^* is gramian n-quasihyponormal. Then

$$[T^{*n+1}T^n h, T^{*n+1}T^n h] \geq [TT^{*n}T^n h, TT^{*n}T^n h]$$

for all $h \in \mathcal{H}$.

Since T is a gramian quasi-isometry, we have

$$T^{*n}T^n = T^*T \tag{1}$$

and then

$$[T^{*2}Th, T^{*2}Th] = [T^*T^{*n}T^n h, T^*T^{*n}T^n h] = [T^{*n+1}T^n h, T^{*n+1}T^n h] \geq$$

$$\geq [TT^{*n}T^n h, TT^{*n}T^n h] = [TT^*Th, TT^*Th]$$

, or $[T^2T^{*2}Th, Th] \geq [(TT^*)^2Th, Th]$.
This inequality shows that

$$[T^2T^{*2}h, h] \geq [(TT^*)^2h, h]$$

for all $h \in \overline{\mathcal{R}(T)}$.

Thus from $[T^2T^{*2}h, h] = 0 = [(TT^*)^2h, h]$ for all $h \in \mathcal{N}(T^*)$, we find

$$[T^2T^{*2}h, h] \geq [(TT^*)^2h, h] \tag{2}$$

for all $h \in \mathcal{H}$. By inequality (2), we deduce that $\|T^2\| = \|T\|^2$. Now, combining the relation (1) with (2), we have $\|T\|^2 = \|T^2\| = \|T\|$ or $\|T\| = 1$.

Now, using Theorem 2, the operator T will be gramian hyponormal and then $\mathcal{N}(T) \subset \mathcal{N}(T^*)$. Using (2) we obtain $[T^*T^*h, T^*T^*h] \geq [TT^*h, TT^*h]$ or $[T^*h, T^*h] \geq [Th, Th]$ for all $h \in \overline{\mathcal{R}(T^*)}$. If $h \in \mathcal{N}(T) \subset \mathcal{N}(T^*)$, then $[T^*h, T^*h] = 0 = [Th, Th]$ i.e. $[T^*h, T^*h] \geq [Th, Th]$, for all $h \in \mathcal{H}$.

This argument shows that T^* is gramian hyponormal and T being gramian hyponormal we find that T is gramian normal.

□

Proposition 4 *Let $T = U \mid T \mid$ be an invertible gramian hyponormal operator. Then $U^n \mid T \mid$ is gramian hyponormal for any $n \geq 1$. Moreover, $U^n \mid T \mid$ is invertible in $\mathcal{B}^*(\mathcal{H})$ and also gramian hyponormal.*

Proof: T being gramian hyponormal, we have

$$T^*T = \mid T \mid U^*U \mid T \mid \geq TT^* = U \mid T \mid^2 U^*$$

or

$$\mid T \mid^2 \geq U \mid T \mid^2 U^*, \tag{3}$$

using the fact that the operator U from the polar decomposition of an invertible operator is gramian unitary from $\mathcal{B}^*(\mathcal{H})$, see [6].

Because $A \geq B \geq 0 \Rightarrow X^*AX \geq X^*BX$, $(\forall)$ $A, B, X \in \mathcal{B}^*(\mathcal{H})$ applying the inequality (3) n times, we find, $\mid T \mid^2 \geq U^n \mid T \mid^2 U^{*n}$.

The last inequality shows that

$$(U^n \mid T \mid)^*(U^n \mid T \mid) = \mid T \mid^2 \geq U^n \mid T \mid^2 U^{*n} = (U^n \mid T \mid)(U^n \mid T \mid)^*.$$

The operators T and U being invertible in $\mathcal{B}^*(\mathcal{H})$, so are $\mid T \mid$ and U^n.

Then $U^n \mid T \mid$ will be invertible in $\mathcal{B}^*(\mathcal{H})$ and its inverse will be gramian hyponormal.

□

Proposition 5 (i) *Let $T = U \mid T \mid$ be an invertible gramian hyponormal operator, with $T = U \mid T \mid$ the polar decomposition of T. Then $\mid T \mid \geq U \mid T \mid U^*$ and the operator $U \mid T \mid^{\frac{1}{2}}$ is also gramian hyponormal.*

(ii) *Let $T \in \mathcal{B}^*(\mathcal{H})$ be invertible and the polar decomposition of T, $T = U \mid T \mid$. Then T is gramian n-hyponormal iff $S = U \mid T \mid^n$, $n \in \mathbb{N}$ is gramian hyponormal.*

Proof: (i) T being a gramian hyponormal, yields the inequalities

$$T^*T \geq TT^* \Rightarrow \mid T \mid^2 \geq \mid T^* \mid^2 .$$

But, polar decomposition of T, which exists T being invertible, $T = U \mid T \mid$ with U gramian unitary, implies, $T^* = \mid T \mid U^*$ and

$$TT^* = \mid T^* \mid^2 = U \mid T \mid^2 U^*.$$

This relations lead to $\mid T \mid^2 \geq U \mid T \mid^2 U^*$, which becomes

$$\mid T \mid^2 \geq (U \mid T \mid U^*)(U \mid T \mid U^*) = (U \mid T \mid U^*)^2,$$

where $\mid T \mid$, $(U \mid T \mid U^*)$ $\mid T \mid^2$, $U \mid T \mid^2 U^* \geq 0$, and $\mid T \mid$, $\mid T \mid^2$, $(U \mid T \mid U^*)$, $U \mid T \mid^2 U^* \in \mathcal{B}_h^*(\mathcal{H})$. Since $f(x) = \sqrt{x}$ is a monotone operator on $[0, \infty)$, see pp. 45, [13], we deduce that

$$\mid T \mid \geq U \mid T \mid U^*.$$

The last inequality will be $(U \mid T \mid^{\frac{1}{2}})^*(U \mid T \mid^{\frac{1}{2}}) \geq (U \mid T \mid^{\frac{1}{2}})(U \mid T \mid^{\frac{1}{2}})^*$.

(ii) $(T^*T)^n \geq (TT^*)^n \Leftrightarrow \mid T \mid^{2n} \geq (U \mid T \mid^2 U^*)^n = U \mid T \mid^{2n} U^* \Leftrightarrow$
$\Leftrightarrow (U \mid T \mid^n)^*(U \mid T \mid^n) \geq U \mid T \mid^{2n} U^* = (U \mid T \mid^n)(U \mid T \mid^n)^*$
$\Leftrightarrow U \mid T \mid^n$ is gramian hyponormal.

□

The following result was given in Hilbert spaces by A. Aluthge in [1]. The existence of the polar decomposition with U gramian unitary is provided for the invertible operator in $\mathcal{B}^*(\mathcal{H})$, where $\mathcal{H}$ is a Loynes Z-space.

Theorem 4 *Let* $T = U \mid T \mid$ *be a gramian p-hyponormal operator* $\frac{1}{2} \leq p < 1$, *and* U *be gramian unitary. Then the operator* $\tilde{T} =\mid T \mid^{\frac{1}{2}} U \mid T \mid^{\frac{1}{2}}$ *is gramian hyponormal.*

Proof: The check will be as in [1]. Taking into account that any gramian p-hyponormal operator, $\frac{1}{2} \leq p < 1$, is gramian $(\frac{1}{2})$-hyponormal , we have

$$\mid T \mid= (\mid T \mid U^*U \mid T \mid)^{\frac{1}{2}} = (T^*T)^{\frac{1}{2}} \geq (TT^*)^{\frac{1}{2}} =$$

$$= (U \mid T \mid\mid T \mid U^*)^{\frac{1}{2}} = (U \mid T \mid U^*U \mid T \mid U^*)^{\frac{1}{2}} = [(U \mid T \mid U^*)^2]^{\frac{1}{2}} = U \mid T \mid U^*.$$

Using again the inequality, $A \geq B \geq 0 \Rightarrow X^*AX \geq X^*BX$, $(\forall)$ $A, B, X \in \mathcal{B}^*(\mathcal{H})$, we obtain,

$$U^* \mid T \mid U \geq U^*U \mid T \mid U^*U =\mid T \mid\geq U \mid T \mid U^*.$$

But, $\tilde{T}$ is gramian hyponormal because

$$\tilde{T}^*\tilde{T} - \tilde{T}\tilde{T}^* =\mid T \mid^{\frac{1}{2}} U^* \mid T \mid^{\frac{1}{2}}\mid T \mid^{\frac{1}{2}} U \mid T \mid^{\frac{1}{2}} - \mid T \mid^{\frac{1}{2}} U \mid T \mid^{\frac{1}{2}}\mid T \mid^{\frac{1}{2}} U^* \mid T \mid^{\frac{1}{2}}=$$

$$=\mid T \mid^{\frac{1}{2}} U^* \mid T \mid U \mid T \mid^{\frac{1}{2}} - \mid T \mid^{\frac{1}{2}} U \mid T \mid U^* \mid T \mid^{\frac{1}{2}}=$$

$$=\mid T \mid^{\frac{1}{2}} (U^* \mid T \mid U - U \mid T \mid U^*) \mid T \mid^{\frac{1}{2}}\geq 0.$$

□

Consequence 1 *Every gramian hyponormal operator,* $T = U \mid T \mid$ *with* U *gramian unitary has* $\tilde{T} =\mid T \mid^{\frac{1}{2}} U \mid T \mid^{\frac{1}{2}}$ *gramian hyponormal.*

The Problem 165, from the book of P. Halmos can be extended also in pseudo-Hilbert spaces as follows.

Is a gramian contraction, similar to a gramian unitary operator, necessarily gramian unitary?

Proposition 6 *If U is gramian unitary, P is positive and invertible in $\mathcal{B}^*(\mathcal{H})$, $C = P^{-1}UP$ and $A = UP$, then a necessary and sufficient condition that C be a gramian contraction is that A be gramian hyponormal operator.*

Proof: We use the demonstration given by P. Halmos in Hilbert spaces for linear and bounded operators.

Considering the above assumptions we obtain the following assertions:

C is gramian contraction is equivalent to

$$CC^* \le I \Leftrightarrow P^{-1}UP^2U^*P^{-1} \le I \Leftrightarrow$$

$$\Leftrightarrow UP^2U^* \le P^2 \Leftrightarrow (UP)(UP)^* \le (UP)^*(UP) \Leftrightarrow$$

$$\Leftrightarrow AA^* \le A^*A.$$

□

Answer: If W is gramian unitary, S is invertible in $\mathcal{B}^*(\mathcal{H})$ and $C = S^{-1}WS$, then using the polar decomposition of the invertible operator S, $S = VP$ we notice that $C = P^{-1}UP$, where $U = V^{-1}WV$ is gramian unitary and P is positive.

It is sufficient to consider the transformations of gramian unitary operators by the positive operators. Then the statement just proved is applicable to them.

In the case of gramian n-hyponormal operators, Proposition 5 has the following variant below.

Proposition 7 *If U is gramian unitary, $P \ge 0$ and invertible in $\mathcal{B}^*(\mathcal{H})$, $C = P^{-n}UP^n$, $n \in \mathbb{N}$ and $A = UP$, then a necessary and sufficient condi-*

tion that C be a gramian contraction is that A be a gramian n-hyponormal operator.

Proof: Using the hypothesis we obtain the following assertions:
C is gramian contraction is equivalent to

$$CC^* \leq I \Leftrightarrow P^{-n}UP^{2n}U^*P^{-n} \leq I \Leftrightarrow$$

$$\Leftrightarrow P^{-n}(UP^2U^*)(UP^2U^*)...(UP^2U^*)P^{-n} \leq I \Leftrightarrow$$

$$\Leftrightarrow P^{-n}(UP^2U^*)^nP^{-n} \leq I \Leftrightarrow$$

$$\Leftrightarrow (UP^2U^*)^n \leq (P^2)^n \Leftrightarrow ((UP)(UP)^*)^n \leq ((UP)^*(UP))^n \Leftrightarrow$$

$$\Leftrightarrow (AA^*)^n \leq (A^*A)^n.$$

□

A similar result with Theorem 5 can be stated.

Theorem 6 *Let $T = U \mid T \mid$ be a gramian $(\frac{p}{2})$-hyponormal operator* $0 < p < 1$*, and U be gramian unitary. Then the operator $\tilde{T_1} =\mid T \mid^{\frac{p}{2}} U \mid T \mid^{\frac{p}{2}}$ is gramian hyponormal.*

Proof: The proof of this theorem is similar to the proof of Theorem 4.

□

References

[1] A. Aluthge, *On p-hyponormal operators for* $0 < p < 1$, Integral Equations Operator Theory, **13** (1990), 307-315.

[2] G. Cassier, L. Suciu, *Mapping theorems and similarity to contractions*, Hot Topics in Operator Theory, Theta, 2008.

[3] S. A. Chobanyan, A. Weron, *Banach-space-valued stationary processes and their linear prediction*, Dissertations Math., **125**(1975), 1–45.

[4] L. Ciurdariu, A. Crăciunescu, *On spectral representation of gramian normal operators on pseudo-Hilbert spaces*, Anal. Univ. de Vest Timişoara, **XLV**, 1 (2007), 131-149.

[5] L. Ciurdariu, *Classes of linear operators on pseudo-Hilbert spaces and applications*, PhD Thesis, West Univ., Timişoara, 2005. (Romanian)

[6] L. Ciurdariu, *Hyponormal and p-hyponormal operators on pseudo-Hilbert spaces*, to appear.

[7] J. B. Conway, *Subnormal operators*, Pitman, Boston, 1981.

[8] P. R. Halmos, *A Hilbert space problem book*, 2nd ed., Springer, New York, 1982.

[9] R. M. Loynes, *Linear operators in VH-spaces*, Trans. American Math. Soc., **116** (1965), 167-180.

[10] R. M. Loynes, *On generalized positive definite functions*, Proc. London Math. Soc., **3** (1965), 373-384.

[11] M. Martin, M. Putinar, *Lectures on hyponormal operators*, Operator Theory, **39**, Birkhauser, Basel, 1989.

[12] S. M. Patel, *A note on quasi-isometries*, Glasnik Matematicki, **35**, 55 (2000), 307-312.

[13] Ş. Strătilă, L. Zsido, *Operator algebras*, Part I, II, T.U.T., Timişoara, 1995.

[14] A. Weron, S. A. Chobanyan, *Stochastic processes on pseudo-Hilbert spaces* Bull. Acad. Polon., Ser. Math. Astr. Phys., **XXI**, 9 (1973), 847-854.

[15] A. Weron, *Prediction theory in Banach spaces*, Proc. of Winter School on Probability, Karpacz, Springer, London, 1975, 207-228.

ROMAI J., **4**, 1(2008), 85–92

SYSTEMS OF EQUATIONS INVOLVING ALMOST PERIODIC FUNCTIONS

Silvia - Otilia Corduneanu

Dept. of Math., "Gheorghe Asachi" Technical University of Iaşi

scorduneanu@yahoo.com

Abstract The theory of Fourier series for almost periodic functions for solving systems of equations which are linear with respect to convolution by functions belonging to $L^1(G)$ is used.

Keywords: almost periodic function, functional equation, Fourier series.

2000 MSC: 43A60.

1. INTRODUCTION

The theory of almost periodic functions on groups was elaborated by John Von Neumann. The main results which were developed by Harald Bohr for almost periodic functions on the real line, as the existence of mean value, the theory of the Fourier series, were extended to this class of functions. The set $AP(G)$ of all almost periodic functions on a Hausdorff locally compact abelian group G is a Banach algebra with respect to the supremum norm. If $f \in AP(G)$ then the Fourier series of f is

$$\sum_{n=1}^{\infty} c_{\gamma_n}(f)\gamma_n,$$

where $\hat{G}$ is the dual of the group G and $\{\gamma_n \in \hat{G} |\ n \in \mathbb{N}\}$ is the set of those characters with the property that for every $n \in \mathbb{N}$ the Fourier coefficient $c_{\gamma_n}(f)$ is different by the value 0. Based on the property that two almost periodic functions coincide if they have the same Fourier coefficients, we solve the following system of equations

$$f_i = \sum_{j=1}^{p} g_{ij} * f_j + h_i, \quad i = 1, 2, ..., p. \tag{1}$$

In this context the functions $g_{ij},\ (i,j) \in \{1,2,...,p\}^2$ belong to $L^1(G)$ and the functions $h_i,\ i \in \{1,2,...,p\}$ are in $AP(G)$. A solution of the system (1) is an element $F = (f_1, f_2, ..., f_p) \in (AP(G))^p$ such that the almost periodic functions $f_1, f_2, ..., f_p$ satisfy the relations (1). In the same manner we can discuss the system of functional equations

$$f_i = \sum_{j=1}^{p} \nu_{ij} * f_j + h_i, \quad i = 1,2,...,p, \tag{2}$$

where $\nu_{ij},\ (i,j) \in \{1,2,...,p\}^2$ are bounded measures and $h_i,\ i \in \{1,2,...,p\}$, are almost periodic functions.

2. PRELIMINARIES

Consider a Hausdorff locally compact Abelian group G and let λ be the Haar measure on G. Let us denote by $\mathcal{C}(G)$ the set of all bounded continuous complex-valued functions on G. The space of Haar measurable functions f on G, with $\int_G |f(x)|d\lambda(x) < \infty$, will be denoted by $L^1(G)$. As usually the norm $\|\cdot\|_1$ is defined by

$$\|f\|_1 = \int_G |f(x)|d\lambda(x), \quad f \in L^1(G).$$

We use $m_F(G)$ to denote the space of all bounded measures on G. We denote by $|\nu|$ the variation measure which corresponds to a measure ν on G. For $f \in \mathcal{C}(G)$ and $a \in G$, the translate of f by a is the function $f_a(x) = f(xa)$ for all $x \in G$. In [2], [5], [6], [7], there are defined the almost periodic functions.

Definition 2.1. *A function $f \in \mathcal{C}(G)$ is called an almost periodic function on G, if the family of translates of f, $\{f_a : a \in G\}$ is relatively compact in the sense of uniform convergence on G.*

The set $AP(G)$ of all almost periodic functions on G is a Banach algebra with respect to the supremum norm, closed to conjugation. There exists a unique positive linear functional $M : AP(G) \to \mathbb{C}$ such that $M(f_a) = M(f)$, for all $a \in G$, $f \in AP(G)$ and $M(\mathbf{1}) = 1$. We denote by $\mathbf{1}$ the constant function which is 1 for all $x \in G$. If $f \in AP(G)$ we define the mean of f as being the above complex number $M(f)$. Denote by $\hat{G}$ the dual of G. It is easy to see that $\hat{G} \subset AP(G)$. We put $c_\gamma(f) = M(\overline{\gamma}f)$ for all $\gamma \in \hat{G}$, $f \in AP(G)$ and we call $c_\gamma(f)$, the Fourier coefficient of f corresponding to $\gamma \in \hat{G}$. Next, we recall the definition of the Fourier series of an almost periodic function. First, let us say that if $f \in AP(G)$, then there exists only a countably subset of $\hat{G}$, denoted by $\{\gamma_n \in \hat{G} | \ n \in \mathbb{N}\}$, such that $M(\overline{\gamma}_n f) \neq 0, \ n \in \mathbb{N}$.

Definition 2.2. *Let $f \in AP(G)$ and $\{\gamma_n \in \hat{G} | \ n \in \mathbb{N}\}$ be the subset of $\hat{G}$ such that $M(\overline{\gamma}_n f) \neq 0, \ n \in \mathbb{N}$. We define the Fourier series of f by*

$$\sum_{n=1}^{\infty} c_{\gamma_n}(f)\gamma_n.$$

If $g \in L^1(G)$ and $f \in AP(G)$, their convolution, $g * f$, belongs to $AP(G)$. We recall that

$$g * f(x) = \int_G f(xy^{-1})g(y)d\lambda(y), \quad x \in G,$$

and that the Fourier transform of $g \in L^1(G)$, denoted by $\hat{g}$ is given by

$$\hat{g}(\gamma) = \int_G g(x)\overline{\gamma}(x)d\lambda(x), \quad \gamma \in \hat{G}.$$

We also recall that for $\nu \in m_F(G)$ and $f \in AP(G)$ we have that their convolution belongs to $AP(G)$; we use the notation $\nu * f$ for the convolution between f and ν and the meaning of that is

$$\nu * f(x) = \int_G f(xy^{-1})d\nu(y), \quad x \in G.$$

The Fourier - Stieltjes transform of $\nu \in m_F(G)$, denoted by $\hat{\nu}$, is given by

$$\hat{\nu}(\gamma) = \int_G \overline{\gamma}(x)d\nu(x), \quad \gamma \in \hat{G}.$$

3. SYSTEMS OF EQUATIONS

Consider $p \in \mathbb{N}, \ p \geq 2$, the functions $g_{ij} \in L^1(G)$, $(i,j) \in \{1,2,...,p\}^2$ and the almost periodic functions l_i, $i = 1,2,...,p$. Denote by $[g]$ the matrix having the elements g_{ij}, $(i,j) \in \{1,2,...,p\}^2$. Consider $i \in \{1,2,...,p\}$. The subset of $\hat{G}$ which contains the characters with the property that the corresponding Fourier coefficient of l_i is different by the value 0, is

$$\mathcal{S}_{l_i} = \{\gamma_n^i | \ n \in \mathbb{N}\}.$$

This means $M\left(\overline{\gamma_n^i} l_i\right) \neq 0,\ n \in \mathbb{N}$. The Fourier series of l_i is

$$\sum_{n=1}^{\infty} c_{\gamma_n^i}(l_i)\gamma_n^i.$$

Consider

$$\mathcal{S} = \bigcup_{i=1}^{p} \mathcal{S}_{l_i},$$

and denote $\mathcal{S} = \{\gamma_n \in \hat{G} |\ n \in \mathbb{N}\}$. Let $(a_n)_n$ be a sequence of complex numbers such that the series $\sum_{n=1}^{\infty} |a_n|^2$ is convergent.

Lemma 3.1. *For every $i = 1, 2, ..., p$ the Fourier series*

$$\sum_{n=1}^{\infty} c_{\gamma_n^i}(l_i)a_n\gamma_n^i$$

is uniformly convergent in the space $AP(G)$.

Proof. For every $i \in \{1, 2, ..., p\}$ we have that the function l_i satisfies the Parseval equality

$$\sum_{n=1}^{\infty} |c_{\gamma_n^i}(l_i)|^2 = M(|l_i|^2). \tag{3}$$

The conclusion follows from (3) and from the inequalities

$$\left(\sum_{k=n}^{n+p} |c_{\gamma_k^i}(l_i)a_k\gamma_k^i(x)|\right)^2 = \left(\sum_{k=n}^{n+p} |c_{\gamma_k^i}(l_i)a_k|\right)^2 \leq$$

$$\leq \sum_{k=n}^{n+p} |c_{\gamma_k^i}(l_i)|^2 \sum_{k=n}^{n+p} |a_k|^2, \quad x \in G,\ n \in \mathbb{N},\ p \in \mathbb{N}. \quad \square$$

Using Lemma 3.1 we can easily obtain the following result.

Corollary 3.1. *There exist the functions $h_1, h_2, ..., h_p$ such that for every $i = 1, 2, ..., p$ we have*

$$h_i = \sum_{n=1}^{\infty} c_{\gamma_n^i}(l_i)a_n\gamma_n^i. \tag{4}$$

in the space $AP(G)$.

Notation 3.1. *For every $n \in \mathbb{N}$ we make the following notation*

$$C_{(h_1,h_2,...,h_p,\gamma_n)} = \begin{bmatrix} c_{\gamma_n}(h_1) \\ \\ c_{\gamma_n}(h_2) \\ \\ \vdots \\ \\ c_{\gamma_n}(h_p) \end{bmatrix}.$$

For $n \in \mathbb{N}$ we consider the matrix

$$M_{([g],\gamma_n)} = \begin{bmatrix} \hat{g}_{11}(\gamma_n) - 1 & \hat{g}_{12}(\gamma_n) & \dots & \hat{g}_{1p}(\gamma_n) \\ \\ \hat{g}_{21}(\gamma_n) & \hat{g}_{22}(\gamma_n) - 1 & \dots & \hat{g}_{2p}(\gamma_n) \\ \\ \dots & \dots & \dots & \dots \\ \\ \hat{g}_{p1}(\gamma_n) & \hat{g}_{p2}(\gamma_n) & \dots & \hat{g}_{pp}(\gamma_n) - 1 \end{bmatrix} \tag{5}$$

and we denote the determinant of the matrix $M_{([g],\gamma_n)}$ by Δ_n. We also consider

$$\max_{(i,j)\in\{1,2,...,p\}^2} \|g_{ij}\|_1 = V_{\max}.$$

Theorem 3.1. *Consider the functions $g_{ij} \in L^1(G)$, $(i,j) \in \{1,2,...,p\}^2$ and the functions h_i, $i \in \{1,2,...,p\}$ defined in* (4).
If there exists $\delta > 0$ such that $\inf_{n\in\mathbb{N}} |\Delta_n| > \delta$, then the system (1) *has a solution $(f_1, f_2, ..., f_p) \in (AP(G))^p$.*

Proof. Consider $n \in \mathbb{N}$. The Fourier coefficients of the convolutions $g_{ij} * f_j$, $(i,j) \in \{1,2,...,p\}^2$ are given by

$$c_{\gamma_n}(g_{ij} * f_j) = \hat{g}_{ij}(\gamma_n) c_{\gamma_n}(f_j).$$

Taking into account the relations (1), we calculate the Fourier coefficients of the functions $f_1, f_2, ..., f_p$, so, we obtain the linear algebric system

$$c_{\gamma_n}(f_i) = \sum_{j=1}^{p} \hat{g}_{ij}(\gamma_n) c_{\gamma_n}(f_j) + c_{\gamma_n}(h_i), \quad i = 1,2,...,p, \tag{6}$$

which has the matrix (5). From (6) it follows that for every $n \in \mathbb{N}$ we have

$$C_{(f_1,f_2,...,f_p,\gamma_n)} = -M_{([g],\gamma_n)}{}^{-1} C_{(h_1,h_2,...,h_p,\gamma_n)}, \tag{7}$$

where

$$C_{(f_1,f_2,...,f_p,\gamma_n)} = \begin{bmatrix} c_{\gamma_n}(f_1) \\ \\ c_{\gamma_n}(f_2) \\ \\ \vdots \\ \\ c_{\gamma_n}(f_p) \end{bmatrix}.$$

We prove that for every $i \in \{1, 2, ..., p\}$, the Fourier series

$$\sum_{n=1}^{\infty} c_{\gamma_n}(f_i)\gamma_n \tag{8}$$

is uniformly convergent in the space $AP(G)$. For every $(i,j) \in \{1, 2, ..., p\}^2$ we have

$$|\hat{g}_{ij}(\gamma_n)| = |\int_G g_{ij}(x)\overline{\gamma}_n(x)d\lambda(x)| \leq \|g_{ij}\|_1 \leq V_{\max}$$

and

$$|\hat{g}_{ij}(\gamma_n) - 1| \ \leq 1 + V_{\max}.$$

If β^n_{ij} is the element of the matrix $M_{([g],\gamma_n)}{}^{-1}$ situated on the line i and the column j then

$$|\beta^n_{ij}| \leq \frac{(p-1)!}{\delta}\,(1 + V_{\max})^{p-1}\,.$$

Taking into account (7) it follows that for every $i \in \{1, 2, ..., p\}$

$$\begin{aligned} &|c_{\gamma_n}(f_i)\gamma_n(x)| \leq \\ &\leq \frac{(p-1)!}{\delta}\,(1 + V_{\max})^{p-1}\sum_{k=1}^{p}|c_{\gamma_n}(h_k)|, \quad n \in \mathbb{N},\ x \in G. \end{aligned} \tag{9}$$

On the other hand, for every $k \in \{1, 2, ..., p\}$ the series

$$\sum_{n=1}^{\infty}|c_{\gamma_n}(h_k)| \tag{10}$$

is convergent. Therefore, from (9) we see that the Fourier series (8) are uniformly convergent. Based on the property that two almost periodic functions coincide if they have the same Fourier coefficients, we conclude that the sums of these series satisfy the equations of the system (1). □

In the same manner we can treat the system (2). Consider the measures $\nu_{ij} \in m_F(G),\ (i,j) \in \{1,2,...,p\}^2$. We denote by $[\nu]$ the matrix having the elements $\nu_{ij},\ (i,j) \in \{1,2,...,p\}^2$. For $n \in \mathbb{N}$ we consider the matrix

$$M_{([\nu],\gamma_n)} = \begin{bmatrix} \hat{\nu}_{11}(\gamma_n) - 1 & \hat{\nu}_{12}(\gamma_n) & \dots & \hat{\nu}_{1p}(\gamma_n) \\ \hat{\nu}_{21}(\gamma_n) & \hat{\nu}_{22}(\gamma_n) - 1 & \dots & \hat{\nu}_{2p}(\gamma_n) \\ \dots & \dots & \dots & \dots \\ \hat{\nu}_{p1}(\gamma_n) & \hat{\nu}_{p2}(\gamma_n) & \dots & \hat{\nu}_{pp}(\gamma_n) - 1 \end{bmatrix} \tag{11}$$

and we denote the determinant of the matrix $M_{([\nu],\gamma_n)}$ by Ω_n. We have the inequalities

$$|\hat{\nu}_{ij}(\gamma)| \leq |\nu|(G), \quad \gamma \in \hat{G}, \quad (i,j) \in \{1,2,...,p\}^2.$$

Using formulas for calculating the Fourier coefficients, as

$$c_\gamma(\nu_{ij} * f_j) = \hat{\nu}_{ij}(\gamma)c_\gamma(f_j), \quad \gamma \in \hat{G}, \quad (i,j) \in \{1,2,...,p\}^2,$$

and performing similar steps as in the proof of Theorem 3.1, we obtain the following result.

Theorem 3.2. *Consider the measures $\nu_{ij} \in m_F(G),\ (i,j) \in \{1,2,...,p\}^2$ and the functions $h_i,\ i \in \{1,2,...,p\}$ defined in* (4).
If there exists $\omega > 0$ such that $\inf_{n\in\mathbb{N}} |\Omega_n| > \omega$, then the system (2) *has a solution $(f_1, f_2, ..., f_p) \in (AP(G))^p$.*

References

[1] L. N. Argabright, J. Gil de Lamadrid, *Fourier analysis of unbounded measures on locally compact Abelian groups*, Mem. Amer. Math. Soc. **145** (1974).

[2] L. N. Argabright, J. Gil de Lamadrid, *Almost Periodic Measures*, Mem. Amer. Math. Soc. **428** (1990).

[3] C. Corduneanu, *Almost periodic functions*, New York, 1968.

[4] N. Dinculeanu, *Integration on locally compact spaces*, Ed. Academiei R.P.R., Bucureşti, 1965 (Romanian).

[5] W. F. Eberlein, *Abstract ergodic theorems and weak almost periodic Functions*, Trans Amer. Math. Soc. **67** (1949), 217-240.

[6] E. Hewitt, K. A. Ross, *Abstract harmonic analysis* **I**, Springer, Berlin, 1963.

[7] J. Gil de Lamadrid, *Sur les mesures presque périodiques*, Astérisque **4** (1973), 61-89.

[8] W. Rudin, *Fourier analysis on groups*, Interscience Tracts in Pure and Applied Mathematics, **12**, Interscience Publishers – John Wiley and Sons, New York, 1962.

ROMAI J., **4**, 1(2008), 93–100

THE IMPACT OF THE STOCHASTIC LIMITER CONTROL ON ECONOMIC DYNAMICS

Adriana-Daniela Gurgui Costea

Theoretical High School "Ovidius", Constanţa

adrianagurgui@yahoo.com

Abstract The impact of the stochastic limiter on the dynamics of the economy is analyzed. The conventional economic model is modified by adding with or multiplying the characteristic parameters by a random quantity. It is shown that if the limiter contains a stochastic factor, then it is possible for the governing invariant sets of the dynamics generated by the model to change into an equilibrium or a cyclic regime or, even more important, a cyclic regime can turn into a chaotic one. For economists, it is highly important to be able to predict the economic asymptotic behaviors in order to prevent the unwanted situations. That's why MathCAD, which is a very useful and powerful program, is going to be used. The numerical examples and diagrams in the paper are realized by us and they use MathCAD.

Keywords: economic dynamics, logistic function, Feigenbaum diagram, limiter control.

2000 MSC: 37N40.

1. RANDOM REGIME OF THE MODEL

The economic model described by the logistic function was intensively studied in the framework of the topological dynamical systems. However, in reality, the factors determining the dynamics generated by the model are random. In the theory of discrete dynamic systems these random variations are known as *noise*, in analogy to electronic circuits.

In what follows, the limiter factor, the parameter 'h' is random. We investigate the effects of the stochastic limiter on the dynamics of the economical system. Thus, the model will be modified up to a version that implies the existence of a *noise* in the limiter factor $f_{h,\omega}(x_t) = \min\{ax_t(1-x_t), h(1+\omega\xi_t)\}$, or $f_{h,\omega}(x_t) = \min\{ax_t(1-x_t), h(1+rnd(d))\}$ where $\omega > 0$, is *the intensity* *of noise,* ξ_t is the Gaussian distribution of the variable with the average 0 and the variation 1, and 'rnd' is the random function, $0 < d < 1$. Of course, dynamics are determined by circumstances and can be random. But, as we are mainly interested in the consequences of the *noise* limiter, we shall refer to the last cases.

2. RANDOM VARIATIONS

Previously it has been studied the effect of the introduction of a fixed limiter threshold in order to control the dependence of the logistic function. In what follows, we take into consideration the fact that in reality, the values situated under this threshold have a variation which will be stimulated by the introduction of a random element. Two kinds of variations will be analyzed: one, which is stimulated by random numbers characterized by the 'rnd' function and the other which is of a Gaussian type, 'dnorm'. Moreover, there will also be taken into consideration the following cases: the variations will influence X_t function; the variations will influence the parameter a. The graphical representations of the two variations are illustrated in fig. 1.

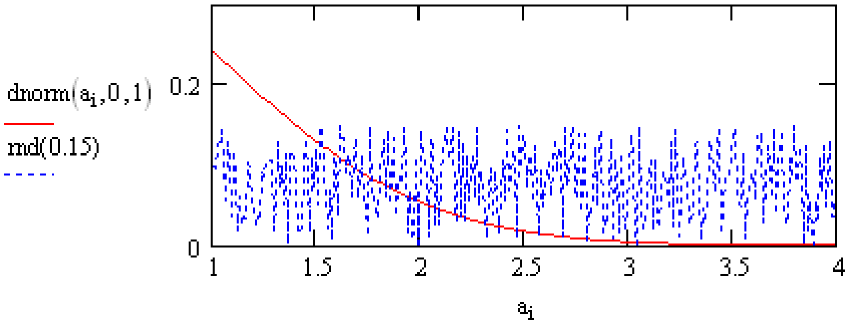


Fig. 1. Variations of the limiter threshold.

If **hl** stands for the random form of the threshold, then its expression in terms of the initial determinist threshold **h** will have one of the forms: $h_1 = h(1 + rnd(d))$, or $h_1 = h(1 + \omega dnorm(h, 0, 1))$, where "rnd" is the random function, $0 < d < 1$ is its domain and **dnorm** is the function of normal distribution (of 1 and 0 parameters in the present case).

We are going to illustrate a few representations of the logistic function with both determinist and random threshold in order to establish their influence. In our analysis we also use the numerical series. We start with the logistic function in a chaotic regime corresponding to $a = 3.9$ applying first a deterministic threshold with a superior limiter of $h = 0.9$.

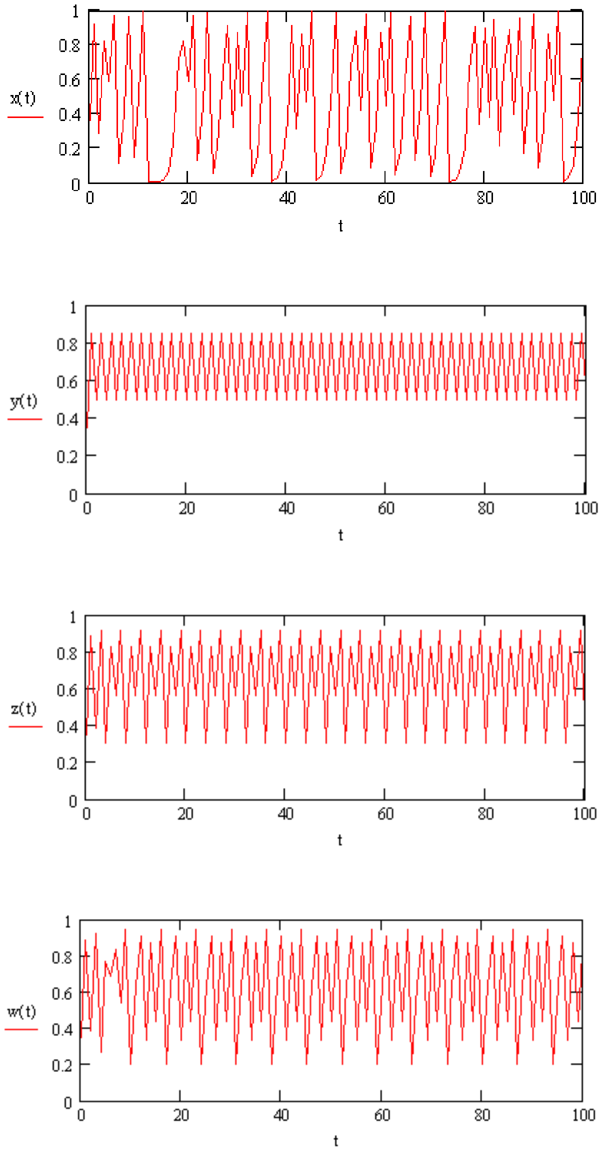
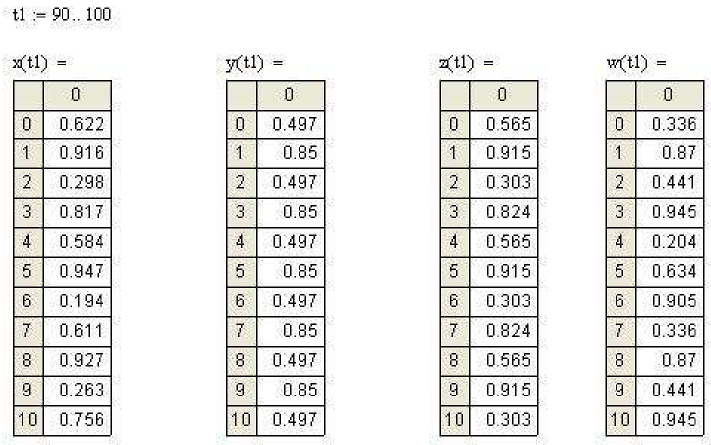


Fig. 2. The change of logistic function in: a) chaotic regime; b) with a determinist threshold, h = 0,85; c) with a random threshold h1=h(1+0,4rnd(0,2)); d) with a Gaussian threshold, h2=h(1+0,4dnorm(h,0,1)).

The result is given in fig 2 a). Thus $\mathbf{x(t)}$ stands for the initial logistic function, $\mathbf{y(t)}$ for the logistic function limited by a logistic threshold, $\mathbf{z(t)}$ for the logistic function limited by a random threshold and $\mathbf{w(t)}$ for the logistic function limited by a Gaussian threshold. Here and in the following by logistic function we mean the motion generated by it.

We notice that, when the determinist threshold is applied, the function changed from its chaotic regime into a cyclic regime of period 2. But, as we can see in the fig 2 c) and in the following table, when a random threshold, whose values vary around the values of the determinist threshold, is applied, the state of the model will change into a cycle of period 4. Also, if a Gaussian threshold is applied, the system changes into a cyclic regime of period 7, (fig 2 d)).



t1 := 90.. 100

x(t1) =

	0
0	0.622
1	0.916
2	0.298
3	0.817
4	0.584
5	0.947
6	0.194
7	0.611
8	0.927
9	0.263
10	0.756

y(t1) =

	0
0	0.497
1	0.85
2	0.497
3	0.85
4	0.497
5	0.85
6	0.497
7	0.85
8	0.497
9	0.85
10	0.497

z(t1) =

	0
0	0.565
1	0.915
2	0.303
3	0.824
4	0.565
5	0.915
6	0.303
7	0.824
8	0.565
9	0.915
10	0.303

w(t1) =

	0
0	0.336
1	0.87
2	0.441
3	0.945
4	0.204
5	0.634
6	0.905
7	0.336
8	0.87
9	0.441
10	0.945

3. FEIGENBAUM DIAGRAMS

In order to study the influence of variations over the limited logistic function (HLC), we shall analyze the influence of variations in the bifurcation Feigenbaum diagram. In this representation we are interested in the bifurcation variable a_i, which determines the various regimes of the logistic function (equilibrium, cycles, chaos). For these we are going to define the following variables:

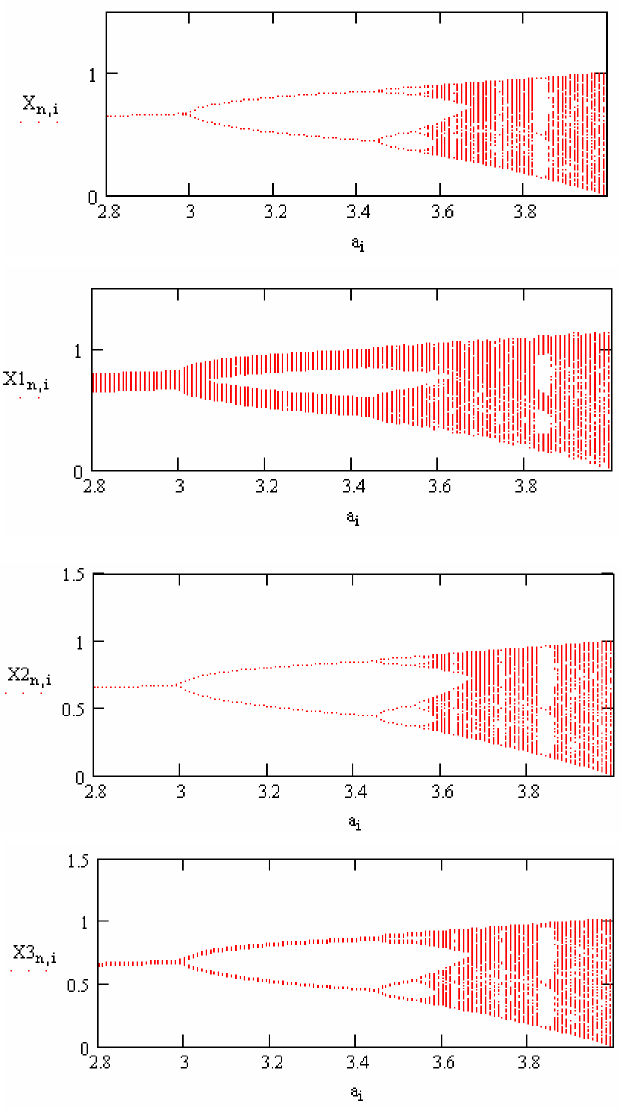
$X_{n,i}$–the basic logistic function (n is the number of iterations and i determines the step of increase of variable a_i) (fig 3. a));

$X1_{n,i}$–the logistic function to which there has been added a random variation (rnd) over $X_{n,i}$ (fig 3. b));

$X2_{n,i}$–the logistic function to which there has been added a Gaussian variation (dnorm) over $X_{n,i}$ (fig 3. c));

$X3_{n,i}$–the logistic function to which there has been added a random variation (rnd) over a_i (fig 3. d));

$X4_{n,i}$–the logistic function to which there has been added a Gaussian variation (dnorm) over a_i (fig 3. e)).



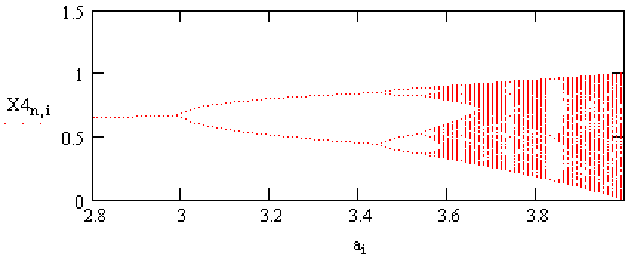


Fig. 3. Feigenbaum diagram for the logistic function a) $X_{n,i}$-basic; b) $X1_{n,i}$- to which there has been added a random variation (rnd) over $X_{n,i}$; c) $X2_{n,i}$- to which there has been added a Gaussian variation (dnorm) over $X_{n,i}$; d) $X3_{n,i}$- to which there has been added a random variation (rnd) over a_i; e) $X4_{n,i}$- to which there has been added a Gaussian variation (dnorm) over a_i

That is: $X_{n+1,i} := a_i X_{n,i}(1 - X_{n,i})$; $X1_{n+1,i} := a_i X_{n,i}(1 - X_{n,i}) + rnd(0.15)$; $X2_{n+1,i} := a_i X_{n,i}(1 - X_{n,i}) + dnorm(a_i, 0, 1)$; $X3_{n+1,i} := (a_i + rnd(0.1))X_{n,i}(1 - X_{n,i})$; $X4_{n+1,i} := (a_i + dnorm(a_i, 0, 1))X_{n,i}(1 - X_{n,i})$.

In the case of a random variation, this variation is represented by thick lines. When we take into consideration the Gaussian variation, the dependence of the logistic function is reflected by the change in form and sizes of it. The differences will be illustrated in what follows by calculations. The changes that take place in the case of variations are very important not only for the functions themselves, but also for their averages, calculated for a random number of iterations. We choose a number of 50 iterations between 100 and 150. Define the following averages:

$m\,(n,i)$ the average for 50 iterations of the basic logistic function $X_{n,i}$;

$m1\,(n,i)$ the average for 50 iterations of the basic logistic function $X1_{n,i}$;

$m2\,(n,i)$ the average for 50 iterations of the basic logistic function $X2_{n,i}$;

$m3\,(n,i)$ the average for 50 iterations of the basic logistic function $X3_{n,i}$;

$m4\,(n,i)$ the average for 50 iterations of the basic logistic function $X4_{n,i}$.

where $m\,(n,i) = \frac{1}{50}\sum\limits_{n=100}^{150} X_{n+1,i}$, $m1\,(n,i) = \frac{1}{50}\sum\limits_{n=100}^{150} X1_{n+1,i}$, $m2\,(n,i) = \frac{1}{50}\sum\limits_{n=100}^{150} X2_{n+1,i}$, $m3\,(n,i) = \frac{1}{50}\sum\limits_{n=100}^{150} X3_{n+1,i}$, $m4\,(n,i) = \frac{1}{50}\sum\limits_{n=100}^{150} X4_{n+1,i}$. In order to avoid the effect of variations over these averages, we calculate the differences (in absolute value) that follows between these and the basic function average $m\,(n,i)$. The relations of definition are

$$d1(n,i) := |m(n,i) - m1(n,i)|, d2(n,i) := |m(n,i) - m2(n,i)|,$$

$$d3(n,i) := |m(n,i) - m3(n,i)|, d4(n,i) := |m(n,i) - m4(n,i)|.$$

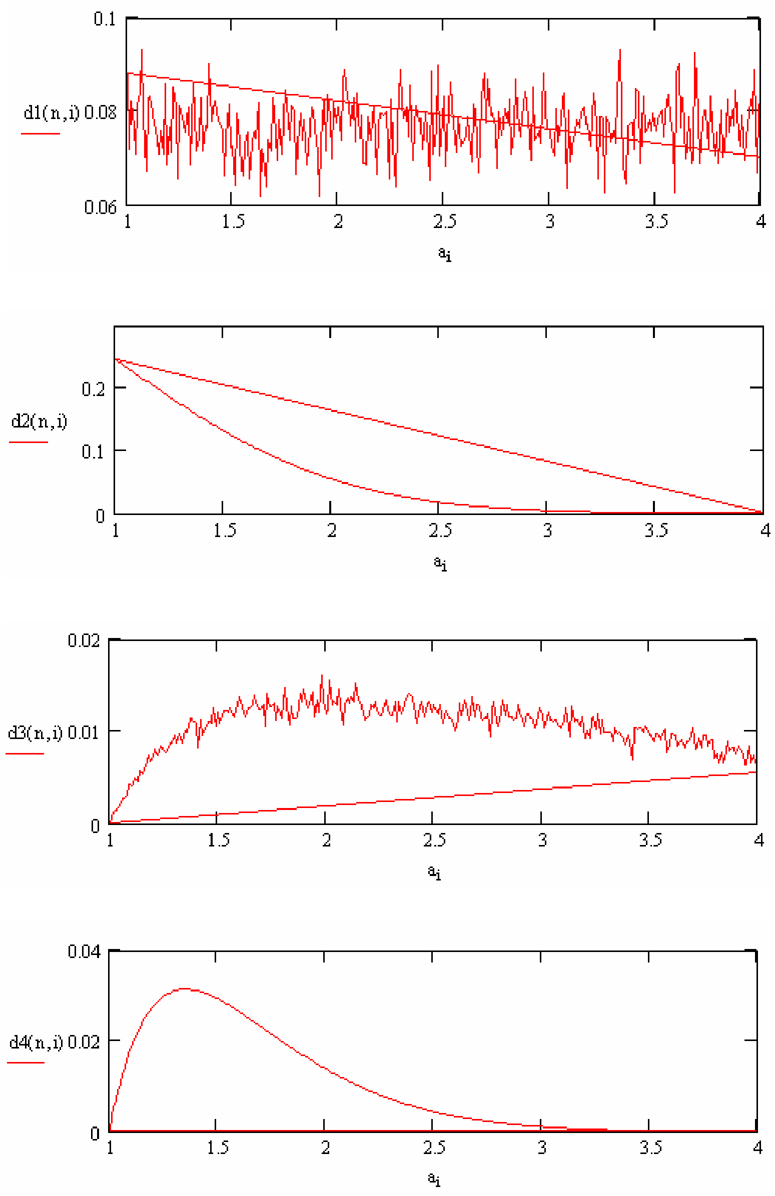


Fig. 4. The effects of variations over averages.

4. CONCLUSIONS

The mechanisms of hard limiter control are often met in economy. The control generates a stable behavior of the system, even if it is chaotic or periodic. A frequent change in the position of the limiter can appear as a proper strategy which can compensate the new cycles that appear. This strategy will lead to irregular behavior of the system.

Our analysis shows that it is to keep the limiter fixed, changing it only in those moments when the value of the parameter is modified substantially. In this case there will appear adjustable cycles of a smaller period. Out of these the cycle of period one seems to be the best from several points of view in relation to economy. A powerful and permanent control is necessary in order to obtain this regime. Otherwise its behavior will not be at the right standards. When economy is being discussed, these effects could be used as arguments against the control politics. In order to counterbalance such arguments, a simple control politics should be formulated in order to reach the peak of economy.

References

[1] Cristi, M., *Periodic economic cycles, the effect of evolution towards criticality and crisis*, Journal of Statistical Mechanics, 2005.

[2] Zang, W.-B., *Discrete dynamical systems, bifurcations and chaos in economics*, Amsterdam, 2006.

[3] Hilker, F. M., Westerhoff, F. H. (eds), *Control of chaotic population dynamics: Ecological and economic considerations*, Beiträge des Instituts für Umweltsystemforschung der Universität Osnabrück, **32**, November 2005.

[4] Mathsoft Mathcad User Guide.

[5] Matsoft Engineering & Education.

ROMAI J., **4**, 1(2008), 101–106

NUMERICAL APPROXIMATION OF POINCARÉ MAPS

Raluca Efrem

Faculty of Mathematics-Informatics, University of Craiova

ra_efrem@yahoo.com

Abstract A classical technique for analyzing dynamical systems is due to Poincaré [4]. It replaces the flow of a n^{th}-order continuous-time system by a $(n-1)^{th}$-order discrete-time system called the Poincaré map. A Poincaré map essentially describes how points on a plane Σ (the Poincaré section), which is transverse to an orbit Γ and which are sufficiently close to Γ, get mapped back onto Σ by the flow [5]. Unfortunately, except under the most trivial circumstances, the Poincaré map cannot be expressed by explicit equations. Here the numerical analysis interpose. In this paper we present an algorithm for constructing the Poincaré map, we build its Maple code, and implement it on an example.

Keywords: systems of differential equations, Poincaré map.

2000 MSC: 37C05.

1. INTRODUCTION

Consider a system of autonomous differential equations

$$\dot{x} = f(x), x \in \mathbb{R}^n. \tag{1}$$

Throughout the text, we assume that the vector field f is sufficiently smooth (C^1 will do nicely). Let $\varphi(x,t)$ denote the solution of (1) satisfying $\varphi(x,0) = x$. The curve $O(x) = \{\varphi(x,t) : t \in \mathbb{R}\}$ is called the *orbit* or *trajectory* passing through the point x.

In many physical applications and particulary in the theory of dynamical systems, one is often interested in computing a *Poincarè map* P (also known as the *first return map*) of a system such as (1). This map is produced by considering successive intersections of a trajectory with a codimension-one surface Σ - called the *Poincaré section*- of the phase space $\mathbb{R}^n$.

Given a system (1), the existence of a Poincarè map is far from being obvious; in many cases it simply does not exist. However, in the case when the system admits periodic solutions Γ, the Poincarè map is well-defined. Indeed, let x^* be a point on such a solution ($x^* \in \Gamma$). There exists a positive number T, called the *period* of the orbit, such that $\varphi(x^*, T+t) = \varphi(x^*, t)$ for all $t \in \mathbb{R}$. In particular, $\varphi(x^*, T) = \varphi(x^*, 0) = x^*$, so the point x^* returns to itself after having flowed T time units. Now consider a surface Σ that is transversal to the flow, i.e. the surface normal at x^* satisfies $< n_\Sigma(x^*), f(x^*) > \neq 0$, where $< \cdot >$ denotes the inner product in $\mathbb{R}^n$. By the implicit function theorem we can find an open neighborhood $U \subset \Sigma$ of x^* such that for all $x \in U$, there exists a positive number $\tau(x)$ such that if $z = \varphi(x, \tau(x))$ then: (a) $z \in \Sigma$ (x returns to the plane Σ after time $\tau(z)$); (b) $sign < n_\Sigma(x), f(x) >= sign < n_\Sigma(z), f(z) >$ (Σ is crossed from the same direction). The function $\tau : \mathbb{R}^n \to \mathbb{R}_+$ is continuous, and represents the time taken by the point x to return to Σ according to the condition (b). The point $z = \varphi(x, \tau(x))$ is called the *fist return* of x, and the Poincaré map $P : U \to \Sigma$ is defined by $P(x) = \varphi(x, \tau(x))$. Note, by definition, we have $\tau(x^*) = T$ and $P(x^*) = x^*$.

The advantages of a Poincaré map consists in the facts that it reduces the study of a flow to the study of maps and it also reduce the dimension of the problem by 1. Except for some cases, the Poincaré map can not be expressed by explicit equations, but it is implicitly defined by the vector field f and the section Σ.

2. LOCATING HYPERPLANE CROSSINGS Σ

In practical implementations the (n-1)-dimensional hyperplane Σ can be chosen in several ways, but if the program already knows the position of a limit cycle, it can choose a point in the hyperplane $x_\Sigma = x^*$ and the normal vector $h = f(x^*)$ where x^* is any point on the limit cycle. So the hyperplane is represented by the equation $H(x) =< h, x - x_\Sigma >= 0$. In practice the hyperplane Σ divides $\mathbb{R}^n$ in two regions: $\Sigma_- = \{x \in \mathbb{R}^n | < h, x - x_\Sigma >< 0\}$ and $\Sigma_+ = \{x \in \mathbb{R}^n | < h, x - x_\Sigma >> 0\}$ and the trajectory will repeatedly cross Σ from Σ_- to Σ_+ to Σ_- etc.

In order to locate the first hyperplane crossing of a trajectory $\varphi(x,t)$, integrate the trajectory and calculate $H(\varphi(x,t))$ at every time-step. We keep integrating until two consecutive points, $x_1 = \varphi(x,t_1)$ and $x_2 = \varphi(x,t_2)$, lie on different sides of Σ, that is $H(x_1)$ and $H(x_2)$ have opposite signs. Once x_1 and x_2 are found, the exact crossing is some point $\alpha = \varphi(x,\tau)$ with $t_1 < \tau < t_2$.

If we want to calculate α and τ, assuming x_1, x_2, t_1 and t_2 are known, we use a simple linear interpolation scheme. Let $\alpha_1 = H(x_1)$ and $\alpha_2 = H(x_2)$. Then $\alpha \approx \frac{\alpha_2}{\alpha_2-\alpha_1}x_1 + \frac{\alpha_1}{\alpha_1-\alpha_2}x_2$, $\tau \approx \frac{\alpha_2}{\alpha_2-\alpha_1}t_1 + \frac{\alpha_1}{\alpha_1-\alpha_2}t_2$.

3. A NUMERICAL EXAMPLE FOR CONSTRUCTING THE POINCARÉ MAP

Consider the dynamical system

$$\dot{x_1} = -x_2 + x_1(1 - x_1^2 - x_2^2),$$

$$\dot{x_2} = x_1 + x_2(1 - x_1^2 - x_2^2).$$

If we pass to the polar system of coordinates $x_1 = r\cos\theta$ and $x_2 = r\sin\theta$ this system becomes the product system

$$\dot{r} = r(1-r^2),\ \dot{\theta} = 1, \tag{2}$$

the general solution of which is

$$\Phi(r,\theta,t) = ([1 + (\frac{1}{r^2} - 1)e^{-2t}]^{-1/2}, t+\theta). \tag{3}$$

We note that the periodic solution is $r(t) = 1$ (with the period $T = 2\pi$) and the transversal section is $\Sigma_p = \{(r,\theta) \in \mathbb{R}^+ \times S^1 | r > 0, \theta = 0\}$.

Consider now the diffeomorphism who help us to pass from polar coordinates to Cartesian $h : \mathbb{R}^2 \to \mathbb{R}^2, h(r,\theta) = (r\cos\theta, r\sin\theta) = (x_1, x_2)$. So, the two systems are topological conjugate, and the solution of the first system will be $\varphi(x,t) = h \circ \Phi(r,\theta,t)$. Doing the calculations we obtain

$$\varphi(x,t) = [1+(\frac{1}{x_1^2+x_2^2}-1)e^{-2t}]^{\frac{-1}{2}} \frac{1}{\sqrt{x_1^2+x_2^2}}(x_1\cos t - x_2\sin t, x_1\sin t + x_2\cos t). \tag{4}$$

The periodic solution $r(t) = 1$ of the system in polar coordinates will be transform in the unit circle $x_1^2 + x_2^2 = 1$ in Cartesian coordinates. The principal period is also $T = 2\pi$ and a point on this solution is $x^* = (1,0)$. So we can choose the point in the hyperplane Σ, $x_\Sigma = x^*$ and by using the equation of H we conclude that the Poincaré section is $\Sigma = \{(x_1, x_2) \in \mathbb{R}^2 | x_2 = 0\}$. We denote by $\Sigma_+ = \{(x_1, x_2) \in \mathbb{R}^2 | x_2 > 0\}$ and $\Sigma_- = \{(x_1, x_2) \in \mathbb{R}^2 | x_2 < 0\}$ the two regions in which Σ divide $\mathbb{R}^2$. We also choose the point $x = (0.5, 0)$ in the neighborhood of x^*, and construct the image $P(x)$ of this point under the Poincaré map. We note that in an Euclidian space the point $P(x)$ is not the first point at which $\varphi(x,t)$ intersects Σ; $\varphi(x,t)$ will pass at least once after returning in the neighborhood U. If Σ is properly chosen, then the trajectory in study will successively cross Σ coming from Σ_+ to Σ_- to Σ_+ etc. The point $x = (0.5, 0)$ situated in the neighborhood of $x^* = (1, 0)$ first returns in the same neighborhood after $\tau = 6.283185307 \approx 2\pi$.

The Maple code of the programm which constructs the Poincaré map and the numerical results are presented below.

```
f := proc(x :: Vector)
return < −x[2]+x[1]∗(1−x[1]∧2−x[2]∧2), x[1]+x[2]∗(1−x[1]∧2−x[2]∧2) >:
end:
H := proc(x :: Vector)
local y, h :: Vector :
y := x − xs : h := f(xs) :
return DotProduct(h, y) :
end:
phi := proc(x :: Vector, t :: float)
local r :: float :
r := 1/(x[1] ∧ 2 + x[2] ∧ 2) :
return < 1/sqrt(1 + (r − 1) ∗ exp(−2 ∗ t)) ∗ (x[1] ∗ cos(t) ∗ sqrt(r) − x[2]∗
sin(t)∗sqrt(r)), 1/sqrt(1+(r−1)∗exp(−2∗t))∗(x[1]∗sin(t)∗sqrt(r)+x[2]∗
cos(t) ∗ sqrt(r)) >:
end:
crossings := proc(x0 :: Vector, t0, tf :: float)
local t1, t2, tau, alfa1, alfa2 :: float :
x1, x2, alfa :: Vector :
```

```
$x2 := x0 :$
$t2 := t0 :$
$alfa2 := H(x0) :$
while $(t2 < tf)$ do
$x1 := x2 :$
$t1 := t2 :$
$alfa1 := alfa2 :$
$t2 := t1 + 0.001 :$
if $t2 > tf$ then
$t2 := tf : fi;$
$x2 := phi(x, t2) :$
$alfa2 := H(x2) :$
if $(alfa1 * alfa2 < 0.0)$ then
print(‘There is a crossing between the points:‘):
$print(x1) : print(x2) :$
$alfa := alfa2/(alfa2 - alfa1) * x1 + alfa1/(alfa1 - alfa2) * x2 :$
$tau := alfa2/(alfa2 - alfa1) * t1 + alfa1/(alfa1 - alfa2) * t2 :$
print(‘The point of crossing is‘): print($alfa$):
print(‘The time when the crossing took place is‘):print($tau$):
fi:
od:
end:
```

It	Crossing between points		Crossing point	Time of crossing
1	−0.9972070812	−0.9972127303	−0.997210429151753176	3.141592652
	0.0005909984257	−0.0004062110483	$-2.31352480575819364 * 10^{-14}$	
2	0.9999947498	0.9999944460	0.999994693503674648	6.283185307
	−0.0001853062088	0.0008146884761	$-2.62652584143846513 * 10^{-14}$	
3	−0.9999996874	−0.9999999652	−0.99999990351749724	9.424777961
	0.0007779606830	−0.0002220392266	$2.76520855182651720 * 10^{-14}$	
4	0.9999999312	0.9999998020	0.99999988331662348	12.56637061
	−0.0003706143508	0.0006293855992	$-3.07184291024886758 * 10^{-14}$	
5	−0.9999995360	−0.9999999992	−0.99999998219571140	15.70796327
	0.0009632678000	−0.00003673205104	$-1.97446090510408624 * 10^{-15}$	
6	0.9999998456	0.9999999014	0.99999987622042166	18.84955592
	−0.0005559215100	0.0004440784466	$-2.6994571107613378 * 10^{-14}$	

References

[1] S. H. M. J. Houben, J. M. L. Maubach, R. M. M. Mattheij, *An accelerated Poincaré-map method for autonomous oscillators*, Applied Mathematics and Computation, **140**, 2-3(2003).

[2] W. Just, H. Kantz, *Some considerations on Poincaré maps for chaotic flows*, Journal of Physics A: Mathematical and General, **33**, 1(2000).

[3] T. Parker, L. Chua, *Practical numerical algorithms for chaotic systems*, Springer, New York, 1989.

[4] H. Poincaré, *Les méthodes nouvelles de la méchanique céleste*, Gauthier-Villars, Paris, 1899 .

[5] H. G. Schuster, *Deterministic chaos*, VCH, Weinheim, 1989.

ROMAI J., **4**, 1(2008), 107–111

GOOD AND SPECIAL WPO PROPERTIES FOR BERNSTEIN OPERATORS IN P VARIABLES

Loredana-Florentina Galea

Faculty of Law and Economics, The Agora University of Oradea

loredana.galea@univagora.ro

Abstract In 1969, D. D. Stancu have introduced the Bernstein operators with arguments real functions in p variables. Using weakly Picard operators and contraction principle, C. Bacotiu studied the convergence of these operators' iterates. In the present paper good and special convergence for Bernstein operators in p variables are investigated.

Keywords: weakly Picard operators, Bernstein operators, good and special weakly Picard operators.

2000 MSC: 47H10, 41A36.

1. INTRODUCTION

In [8], D.D. Stancu has introduced the Bernstein operators in p variables, like as:

Definition 1.1. *[8]. Consider the set*

$$\overline{\Delta} = \{(x_1, x_2, ..., x_{p+}) \in \mathbb{R}^p : x_1 \geqslant 0,\ x_2 \geqslant 0, ..., x_p \geqslant 0;\ x_1 + x_2 + ... + x_p \leqslant 1\}$$

The operators $B_n : C\left(\overline{\Delta}\right) \to C\left(\overline{\Delta}\right)$ *defined by*

$$B_n(f)(x_1, ..., x_p) = \sum_{0 \leqslant i_1 + ... + i_p \leqslant n} p_{n;i_1,...,i_p}(x_1, ..., x_p)\, f\left(\tfrac{i_1}{n}, ..., \tfrac{i_p}{n}\right)$$

for any $f \in C\left(\overline{\Delta}\right)$ *and* $(x_1, ..., x_p) \in \overline{\Delta}$ *are called Bernstein operators with arguments real functions in p variables.*

The polynomials $p_{n;i_1,...,i_p}$ are generalizations of Bernstein fundamental polynomials and are defined by

$$p_{n;i_1,...,i_p}(x_1, ..., x_p) = \tfrac{n!}{i_1!i_2!...i_p!} x_1^{i_1} ... x_p^{i_p} (1 - x_1 - x_2 - ... - x_p)^{n - i_1 - i_2 - ... - i_p}$$

for any $(x_1, ..., x_p) \in \overline{\Delta}$.

The set $\nu_{\overline{\Delta}} = \{(0,0,...,0), (0,1,...,0), ..., (0,0,...,1)\} \subset \overline{\Delta}$ represents the set of knots and if $\alpha_0 = (0,0,...,0), \alpha_1 = (0,1,...,0), ..., \alpha_p = (0,0,...,1)$, then $\nu_{\overline{\Delta}} = \left\{\alpha_k : k = \overline{0,p}\right\}$.

In the following, we present some properties of Bernstein operators in p variables, which were studied in [8]:

(P1) B_n are linear and positive;

(P2) for any $k = \overline{0,p}$, $e_{\alpha k}$ is a fixed point for B_n, so

$$B_n(e_{\alpha k})(x_1, ..., x_p) = e_{\alpha k}(x_1, ..., x_p),\ \forall\ (x_1, ..., x_p) \in \overline{\Delta};$$

(P3) B_n is an interpolation for any $f \in C\left(\overline{\Delta}\right)$ and for the knots of $\nu_{\overline{\Delta}}$, so

$$B_n(f)(\alpha_k) = f(\alpha_k), \forall k = \overline{0,p}.$$

Definition 1.2. *[5], [6], [7]. Let (X, d) be a metric space.*

1) An operator $A : X \to X$ is weakly Picard operator (WPO) if the sequence of successive approximations $(A^m(x_0))_{m \in \mathbb{N}}$ converges for all $x_0 \in X$ and the limit (which may depend on x_0) is a fixed point of A.

2) If the operator $A : X \to X$ is an WPO and $F_A = \{x^*\}$, then the operator A is called a Picard operator (PO).

3) If the operator $A : X \to X$ is an WPO, then the operator A^∞ is defined by $A^\infty : X \to X,\ A^\infty(x) := \lim\limits_{m \to \infty} A^m(x)$.

The basic result in the WPO's theory is the following characterization theorem

Theorem 1.1. *[5], [6], [7]. An operator $A : X \to X$ is WPO if and only if there exists a partition of X, $X = \bigcup\limits_{\lambda \in \Lambda} X_\lambda$, such that:*

(a) $X_\lambda \in I(A),\ \forall \lambda \in \Lambda$;

(b) $A|_{X_\lambda} : X_\lambda \to X_\lambda$ is PO, $\forall \lambda \in \Lambda$.

In [7], I.A. Rus has introduced the notions of good and special WPO

Definition 1.3. . *Let (X, d) be a metric space and $A : X \to X$ an WPO.*

1) $A : X \to X$ is a good WPO, if the series $\sum\limits_{m=1}^{\infty} d\left(A^{m-1}(x), A^m(x)\right)$ converges, for all $x \in X$ [7]. If the sequence $\left(d\left(A^{m-1}(x), A^m(x)\right)\right)_{m \in \mathbb{N}^*}$ is *strictly decreasing for all $x \in X$, the operator A is a good WPO of type M [2].*

2) $A : X \to X$ is a special WPO, if the series $\sum\limits_{m=1}^{\infty} d\left(A^m\left(x\right), A^{\infty}\left(x\right)\right)$ converges, for all $x \in X$ [7]. If the sequence $\left(d\left(A^m\left(x\right), A^{\infty}\left(x\right)\right)\right)_{m\in\mathbb{N}^*}$ is strictly decreasing for all $x \in X$, A is a special WPO of type M [2].

Theorem 1.2. *[4]. Let (X, d) be a metric space and $A : X \to X$ a WPO. If A is a special WPO then A is a good WPO.*

Theorem 1.3. *[3]. The Bernstein operators in p variables $B_n : C\left(\overline{\Delta}\right) \to C\left(\overline{\Delta}\right)$ are weakly Pixard and $B_n^{\infty}\left(f\right) = \varphi_f^*, \forall f \in C\left(\overline{\Delta}\right)$, where the function $f \in C\left(\overline{\Delta}\right)$ are given by*

$$\varphi_f^*\left(x_1, ..., x_p\right) = f\left(\alpha_0\right) + \left[f\left(\alpha_1\right) - f\left(\alpha_0\right)\right] x_1 + \left[f\left(\alpha_2\right) - f\left(\alpha_0\right)\right] x_2 + ...+$$

$$+\left[f\left(\alpha_p\right) - f\left(\alpha_0\right)\right] x_p,\ \forall\ f \in C\left(\overline{\Delta}\right),\ \forall\ \left(x_1, ..., x_p\right) \in \overline{\Delta}.$$

The convergence is in the space $C\left(\overline{\Delta}\right)$.

In order to apply the characterization theorem of weakly Picard operators, the partition of space $C\left(\overline{\Delta}\right)$

$$C\left(\overline{\Delta}\right) := \bigcup_{\Lambda\in\mathbb{R}^{p+1}} X_{\Lambda}$$

was considered, where $X_{\Lambda} := \left\{f \in C\left(\overline{\Delta}\right) : f\left(\alpha_k\right) = \lambda_k, \text{ for } k = \overline{0, p}\right\}$, for any $\Lambda = \left(\lambda_0, \lambda_1, ..., \lambda_p\right) \in \mathbb{R}^{p+1}$.

Proposition 1.1. *[3]. The Bernstein operators in p variables satisfy the following contraction property*

$$\left\|B_n\left(f\right) - B_n\left(g\right)\right\|_C \leqslant \left(1 - \frac{1}{p^{n-1}}\right) \left\|f - g\right\|_C, \forall f, g \in X_{\Lambda}. \qquad (1)$$

2. MAIN RESULTS

In this section, we study the good and special weakly Picard operators convergence for Bernstein operators in p variables.

Using the inequality (1), we obtain the estimation

$$\left|B_n^1\left(f\right)\left(x_1, ..., x_p\right) - B_n^{\infty}\left(f\right)\left(x_1, ..., x_p\right)\right| =$$

$$= \left|B_n^1\left(f\right)\left(x_1, ..., x_p\right) - B_n^1\left(B_n^{\infty}\left(f\right)\right)\left(x_1, ..., x_p\right)\right| \leqslant$$

$$\leqslant \left(1-\tfrac{1}{p^{n-1}}\right)\left|f\left(x_1,...,x_p\right)-B_n^{\infty}\left(f\right)\left(x_1,...,x_p\right)\right| =$$

$$=\left(1-\tfrac{1}{p^{n-1}}\right)\left|f\left(x_1,...,x_p\right)-\left\{f\left(\alpha_0\right)+\left[f\left(\alpha_1\right)-f\left(\alpha_0\right)\right]x_1+...+\left[f\left(\alpha_p\right)-f\left(\alpha_0\right)\right]x_p\right\}\right| \leqslant$$

$$\leqslant \left(1-\tfrac{1}{p^{n-1}}\right)\left(p+1\right)C,\ \forall\ \left(x_1,...,x_p\right)\in\overline{\Delta},$$

where $C = diam\left(\operatorname{Im} f\right) = diam\left(f\left(\overline{\Delta}\right)\right) =$

$$=\max\left\{\left|f\left(x_1,...,x_p\right)-f\left(y_1,...,y_p\right)\right| : \left(x_1,...,x_p\right),\left(y_1,...,y_p\right)\in\overline{\Delta}\right\}.$$

$$\left|B_n^2\left(f\right)\left(x_1,...,x_p\right)-B_n^{\infty}\left(f\right)\left(x_1,...,x_p\right)\right| =$$

$$=\left|B_n^1\left(B_n^1\left(f\right)\right)\left(x_1,...,x_p\right)-B_n^1\left(B_n^{\infty}\left(f\right)\right)\left(x_1,...,x_p\right)\right| \leqslant$$

$\leqslant \left(1-\tfrac{1}{p^{n-1}}\right)\left\|B_n^1\left(f\right)-B_n^{\infty}\left(f\right)\right\| = \left(1-\tfrac{1}{p^{n-1}}\right)^2\left(p+1\right)C,\ \forall\ \left(x_1,...,x_p\right)\in$ $\overline{\Delta}$.

By induction, for $m\in\mathbb{N}^*$, we have

$$\left|B_n^m\left(f\right)\left(x_1,...,x_p\right)-B_n^{\infty}\left(f\right)\left(x_1,...,x_p\right)\right| =$$

$$=\left|B_n^1\left(B_n^{m-1}\left(f\right)\right)\left(x_1,...,x_p\right)-B_n^1\left(B_n^{\infty}\left(f\right)\right)\left(x_1,...,x_p\right)\right| \leqslant$$

$$\leqslant \left(1-\tfrac{1}{p^{n-1}}\right)^m\left(p+1\right)C,\ \forall\ \left(x_1,...,x_p\right)\in\overline{\Delta}.$$

So, $\sum\limits_{m=1}^{\infty}\left|B_n^m\left(f\right)\left(x_1,...,x_p\right)-B_n^{\infty}\left(f\right)\left(x_1,...,x_p\right)\right| \leqslant C\left(p+1\right)\left(p^{n-1}-1\right)$, $\forall\ \left(x_1,...,x_p\right)\in\overline{\Delta}$.

On the other hand, we have

$$\left|B_n^1\left(f\right)\left(x_1,...,x_p\right)-B_n^0\left(f\right)\left(x_1,...,x_p\right)\right| =$$

$$=\left|\sum_{0\leqslant i_1+...+i_p\leqslant n} p_{n;i_1,i_2,...,i_p}\left(x_1,...,x_p\right)f\left(\tfrac{i_1}{n},...,\tfrac{i_p}{n}\right)-f\left(x_1,...,x_p\right)\right| =$$

$$=C\cdot\sum_{0\leqslant i_1+...+i_p\leqslant n} p_{n;i_1,i_2,...,i_p}\left(x_1,...,x_p\right)=C,\ \forall\ \left(x_1,...,x_p\right)\in\overline{\Delta}\ .$$

By induction, we have

$$\left|B_n^m\left(f\right)\left(x_1,...,x_p\right)-B_n^{m-1}\left(f\right)\left(x_1,...,x_p\right)\right| =$$

$$=\left|B_n^1\left(B_n^{m-1}\left(f\right)\right)\left(x_1,...,x_p\right)-B_n^1\left(B_n^{m-2}\left(f\right)\right)\left(x_1,...,x_p\right)\right| \leqslant$$

$$\leqslant \left(1 - \tfrac{1}{p^{n-1}}\right)^{m-1} \cdot C,\ \forall\ (x_1, ..., x_p) \in \overline{\Delta}.$$

So, $\sum\limits_{m=1}^{\infty} \left|B_n^m(f)(x_1, ..., x_p) - B_n^{m-1}(f)(x_1, ..., x_p)\right| \leqslant C \cdot p^{n-1}$, $\forall\ (x_1, ..., x_p) \in \overline{\Delta},\ \forall\ f \in C\left(\overline{\Delta}\right)$.

From above results, we have the following

Proposition 2.1. *The Bernstein operators in p variables are special weakly Picard and good weakly Picard on* $C\left(\overline{\Delta}\right)$.

References

[1] L. D'Apuzzo, *On the convergence of the method of succesive approximations in metric spaces*, Ann. Istit. Univ. Navale Napoli, **45/46**(1976/1977). (in Italian).

[2] L. D'Apuzzo, *On the notion of good and special convergence of the method of successive approximations*, Ann. Istit. Univ. Navale Napoli, **45/46**(1976/1977), 123-138. (in Italian).

[3] C. Bacotiu, *Picard operators and application*, PhD Thesis, Univ. Babes-Bolyai, Cluj-Napoca, 2004. (in Romanian).

[4] S. Mureşan, L. Galea, *On good and special weakly Picard operators* (submitted).

[5] I. A. Rus, *Generalized contractions and applications*, Cluj University Press, 2001. (in Romanian).

[6] I. A. Rus, *Weakly Picard operators and applications*, Seminar on Fixed Point Theory, Cluj-Napoca, **2**(2000), 41-58.

[7] I. A. Rus, *Picard operators and applications*, Sci. Math.Japon., **58**(1)(2003), 191-219.

[8] D. D. Stancu, *A new class of uniform approximating polynomial operators in two and several variables*, Proc. Conf. on Constructive Theory of functions, Budapest, (1969), 443-455.

ROMAI J., **4**, 1(2008), 113–119

EXAMPLES OF DIFFERENTIAL GAMES WITH STOCHASTIC PERTURBATION ASSOCIATED WITH NASH EQUILIBRIUM SOLUTIONS AND OPEN LOOP STRATEGY

Daniela Ijacu

Academy of Economic Studies, Bucharest

Abstract We assimilate the stochastic systems, which are describing the stochastic dynamics of a differential game, as a characteristic system associated with Hamilton Iacobi equation.We present in detail an example of stochastic differential game, where the calculus can be explained as an algorithm if the hypothesis (H) on the diffusion coefficients is personalized as linear fields with constant coefficients.

Keywords: differential games, Nash equilibrium, open loop strategy.

2000 MSC: 49N70, 60H10.

A differential game with stochastic perturbation is determined by a dynamics of the state variable $x \in X \subseteq R^n$ defined by a system of stochastic differential equations

$$1) \begin{cases} d_t x = f\left(t, x, u_1, .., u_N\right) dt + \sum\limits_{j=1}^{d} g_j(t,x) \otimes dw_j(t), \\ x(0) = x_0, (t, x, u_1, .., u_N) \in [0, t_f] \times X \times U_1 \times .. \times U_N, \end{cases}$$

where $U_i \subseteq R^{m_i}$ are some fixed closed sets and $w\left(t\right) = \left(w_1\left(t\right)..w_d\left(t\right)\right)$ is a standard Wiener d-dimensional process over a complete filtered probability space $\{\Omega, \mathcal{F}, P, \{\mathcal{F}_t\} \uparrow \mathcal{F}\}$ and $"\otimes"$ is a special type of stochastic integral $[3]$.

Remark 0.1. *Choosing an admissible command (open loop strategy) is a key point in the analysis of some differential stochastic systems as much in define the sense of associated solution with* (1) *as in formulating some required "reasonable" conditions associated with Nash equilibrium solutions.*

The simplest case is when the fields $g_j \in R^n, j \in \{1, .., d\}$, from stochastic perturbation, does not depend on the state $x \in X$ and associate the continuous process

$$\text{I) } \eta(t,\omega)=\sum_{j=1}^{d}\int_{0}^{t} g_j(s)\,dw_j(s)\,, t\in[0,t_f]\,,$$

where we used the standard Ito integral.

Then by a translation $x(t)=y(t)+\eta(t,\omega)\,, t\in[0,t_f]$ the solution of stochastic system (1) is replaced by $y(t)$ satisfying the system of differential equations with random parameter $\omega\in\Omega$

$$\begin{cases} \dfrac{dy}{dt}=f(t,y+\eta(t,\omega)\,,u)=f^{\omega}(t,y,u)\,,\\ y(0,\omega)=x_0,(t,y,u)\in[0,t_f]\times X\times U. \end{cases}$$

A functional $J(u)=F(x(t_f,\omega,u))+\int\limits_0^{t_f} f_0(t,x(t,\omega,u)\,,u(t,\omega))\,dt$ associated with (1) is rewritten

$$J^{\omega}(u)=F(y(t_f,\omega,u))+\eta(t_f,\omega)+\int\limits_0^{t_f} f_0(t,y(t,\omega,u)))\,dt$$
$$+\eta(t,\omega)\,,u(t,\omega)=$$
$$F^{\omega}(y(t_f,\omega,u))+\int\limits_0^{t_f} f_0^{\omega}(t,y(t,\omega,u)\,,u(t,\omega))\,dt,\omega\in\Omega.$$

The necessary condition of optimality (Pontriaghin principle) will be associated with an optimal solution and with the Hamilton function $H^{\omega}(t,y,u,\psi)\overset{\Delta}{=}\psi^t f^{\omega}(t,y,u)+f_0^{\omega}(t,y,u)$. Let us explain this conclusion. We have

$$\begin{aligned} c_1^{\omega})\frac{dy^*}{dt}(t,\omega) &= \nabla_{\psi}H^{\omega}(t,y^*(t,\omega),u^*(t,\omega),\psi(t,\omega))\\ &= f^{\omega}(t,y^*(t,\omega),u^*(t,\omega))\,,y^*(0,\omega)=x_0\in X, \end{aligned}$$

$$\begin{aligned} c_2^{\omega})\frac{d\psi^i}{dt}(t,\omega) &= -\nabla_y H^{\omega}(t,y^*(t,\omega),u^*(t,\omega),\psi(t,\omega))=\\ &= -\left[\frac{\partial f^{\omega}}{\partial y}(t,y^*(t,\omega),u^*(t,\omega))\right]^T\psi(t,\omega)-\frac{\partial f_0^{\omega}}{\partial y}(t,y^*(t,\omega),\\ u^*(t,\omega)),t &\in [0,t_f]\,, \end{aligned}$$

$$c_3^{\omega})\psi(t_f,\omega)=\nabla_y F^{\omega}(y^*(t_f,\omega))=\ \nabla_x F(y^*(t_f,\omega)+\eta(t_f,\omega))$$

$$\begin{aligned} c_4^{\omega})H^{\omega}(t,y^*(t,\omega),u^*(t,\omega),\psi(t,\omega)) &= \min_{u\in U}H^{\omega}(t,y^*(t,\omega),u,\psi(t,\omega))\,,\\ \text{a.p.t in } t &\in [0,t_f] \text{ and for each } \omega\in\Omega. \end{aligned}$$

This shows that the differentiable function $\psi(t,\omega)$, $t\in[0,t_f]$ (covector) is just $\mathfrak{F}$- measurable for each $t\in[0,t_f]$ (is not $\mathfrak{F}_t$ adapted) relying $\omega\in\Omega$

we obtain (see c_4^{ω})) $u^* (t, \cdot)$ just as $\mathcal{F}$-measurable function for each $t \in [0, t_f]$ (non $\mathcal{F}_t$ adapted). Therefore $x^* (t, \omega) = y^* (t, \omega) + \eta (t, \omega), t \in [0, t_f]$ will be a non $\mathcal{F}_t$ adapted solution associated with stochastic system (1), where the stochastic integral ⊗" is taken in the Ito sense, if $g_i (t, x) = g_i (t)$, $i \in \{1, .., d\}$. Using a class of commands as $u (t, \cdot)$ to be $\mathcal{F}_t-$ measurable $t \in [0, t_f]$ (command $\mathcal{F}_t-$adapted) forces us to define the adjunct function $\psi (t, \omega), t \in [0, t_f]$, as a continuous process (nondifferentiable) and $\mathcal{F}_t-$adapted and with final condition $\psi (t_f, \omega) = \nabla_x F (x^* (t_f, \omega))$ fixed.

This is an interesting, but difficult to use in applications, mathematical problem.

From now on we use the admissible command in measurable and bounded functions class on product space $\{[0, t_f] \times \Omega, \mathcal{B} \otimes \mathcal{F}, dt \otimes P\}$, where $\mathcal{B}-\sigma$ algebra of measurable Lebesgue set from $[0, t_f]$ and $\mathcal{F} \supseteq \{\mathcal{F}_t\} \uparrow$ is a $\sigma-$algebra of measurable set given with filtered probability space.

With the dynamics (1) we associate the functionals

2) $J_i^{\omega} (u) = F^i (x (t_f, \omega, u)) + \eta (t_f, \omega) + \int\limits_0^{t_f} f_0^i (t, x (t, \omega, u), u (t, \omega)) \, dt$ for each $\omega \in \Omega, i \in \{1, .., N\}$, where $x (t, \omega, u), t \in [0, t_f]$ represents the solution of system (1) corresponding to the admissible command $u (\cdot) \triangleq (u_1 (\cdot), .., u_N (\cdot)) \in \mathcal{A} = \otimes \mathcal{A}_i$, which, for each $t \in [0, t_f]$ fixed, is just $\mathcal{F}-$measurable (is non $\mathcal{F}_t-$adapted).

Remark 0.2. *The sense of solution for (1) will be make precise minutely later accepting that $u_i (\cdot) \in \mathcal{A}_i$ ($u_i (\cdot)$ admitted) if $u_i (t, \omega) : [0, t_f] \times \Omega \longrightarrow U_i$ is measurable and bounded and possesses the trajectories $u_i (\cdot, \omega)$ piecewise continuous for each $\omega \in \Omega, i \in \{1, .., N\}$.*

The definition of an equilibrium Nash solution for differential game with stochastic perturbation represented by system (1) and functionals (2) will include the interpretation that this time we have a set of differential deterministic game with N players index after parameter $\omega \in \Omega$.

In this sense a solution of stochastic system (1) is searched as a continuous process and $\mathcal{F}_t-$adapted with value in mappings set $x = G (p (t, \omega), y)$ relaying to $y \in B (x_0, \rho (\omega)) \subseteq R^n$. Therefore, if $y (t, \omega, u) \in B (x_0, \rho (\omega)) \subseteq R^n$ is the solution of differential equations system

$$3)\begin{cases} \dfrac{dy}{dt}(t,\omega,u)=\left[\dfrac{dG}{dy}(p(t,\omega);y)\right]^{-1} f(t,G(p(t,\omega);y),u(t,\omega)) \\ \qquad\qquad \triangleq \tilde{f}(\omega,t,u(t,\omega)),t\in[0,t_f], \\ y(0,\omega,u)=x_0, \end{cases}$$

then $\chi(t,\omega,u)=G(p(t,\omega);y(t,\omega,u))$ is a solution non $\mathcal{F}_t-$adapted of differential stochastic system (1) corresponding to admissible command $u(\cdot)\in\mathcal{A}$ (non $\mathcal{F}_t-$adapted). Correspondingly, the functionals $J_i^\omega(u)$ are rewritten as

$$4)\ J_i^\omega(u)=F^i(G(p(t_f,\omega),y(t_f,\omega,u))) \\ +\int_0^{t_f} f_0^i(t,G(p(t,\omega),y(t,\omega,u)),u(t,\omega))\,dt \\ =\tilde{F}^i(\omega,y(t_f,\omega,u))+\int_0^{t_f}\tilde{f}_0^i(\omega,t,y(t,\omega,u),u(t,\omega))\,dt,$$

where $\tilde{F}^i(\omega,y)=F^i(G(p(t,\omega),y)),$

$\tilde{f}_0^i(\omega,t,y,u)=f_0^i(t,G(p(t,\omega);y),u),i\in\{1,..,N\}.$

Remark 0.3. *The functions $\tilde{f}(\omega,t,y,u)\in R^n$ in (3) and $\tilde{F}^i(\omega,y),\tilde{f}_0^i(\omega,t,y,u)$ in (4) establish a differential game with N players and open loop strategy,*

5) $\Gamma_N(\omega)=\left\{[0,t_f],Y=R^n,U_i,\mathcal{A}_i,\tilde{f}(\omega,\cdot),x_0,J_i^\omega\right\}_{i\in\{1,..,N\}}.$

Here we accept that the vectorial functions $g_j(t,x)$ are continuous in $(t,x)\in[0,t_f]\times R^n$, having the structure given in: II) $g_j(t,x)=A_jx+b_j(t)$, where A_j is a constant $(n\times n)$ matrix and $b_j(t)$ is continuous, $j\in\{1,..,d\}.$

Remark 0.4. *In hypothesis (II), the solution of the system (1) becomes 6)* $x(t,\omega,u)=G(t,\omega)(y(t,\omega,u))+\eta(t,\omega),t\in[0,t_f],$

where the $n\times n$ matrix G and the vector $\eta\in R^n$ are defined as a continuous process satisfying

$$\begin{aligned} 6.1)\ d_tG &= \sum_{j=1}^d A_jG\circ dw_j(t),G(0,\omega)=I_n, \\ \eta(t,\omega) &= \sum_{j=1}^d\int_0^t b_j(s)\cdot dw_j(s),t\in[0,t_f], \end{aligned}$$

where "$\circ$" is the Fisk-Stratonovich stochastic integral, and "$\cdot$" is Ito integral.

The vector $y\left(t,\omega,u\right)\in R^{n}$ is defined as a differential process in $t\in\left[0,t_{f}\right]$ and non $\mathcal{F}_{t}-$adapted satisfying

$$6.2)\left\{\begin{array}{c}\dfrac{dy}{dt}=\left[G\left(t,\omega\right)\right]^{-1}f\left(t,G\left(t,\omega\right)\left(y\right)+\eta\left(t,\omega\right);u\left(t,\omega\right)\right)\\ =\tilde{f}\left(\omega,t,y,u\left(t,\omega\right)\right),t\in\left[0,t_{f}\right],\\ y\left(0,\omega,u\right)=x_{0}\in R^{n}.\end{array}\right.$$

The matrix $G\left(t,\omega\right)$ is invertible and $K\left(t,\omega\right)=\left[G\left(t,\omega\right)\right]^{-1}$ is the solution of the system

$$6.3)\ d_{t}K=-\sum_{j=1}^{d}KA_{j}\circ dw_{j}\left(t\right),K\left(0,\omega\right)=I_{n},t\in\left[0,t_{f}\right].$$

A Nash equilibrium point $\left(x^{*}\left(t,\omega\right),u^{*}\left(t,\omega\right)\right),t\in\left[0,t_{f}\right],\ \omega\in\Omega$, convert in (6), leads us to a Nash equilibrium point $\left(y^{*}\left(\cdot,\omega\right),u^{*}\left(\cdot,\omega\right)\right)$ for the differential game $\Gamma_{N}\left(\omega\right)$ defined in (5) for $\omega\in\Omega$.

Lemma 0.1. *Let $g_{j}\left(t,x\right)\in R^{n},j\in\{1,..,d\}$ are fulfilling the hypothesis $\left(i_{0}\right)$. Let $f\left(t,x,u\right)\in R^{n},F_{i}\left(x\right)\in R,f_{0}^{i}\left(t,x,u\right)\in R,i\in\{1,..,N\}$ continuous on $\left[0,t_{f}\right]\times R^{n}\times U$ i.e*

$i_{1})\left|f\left(t,x,u\right)\right|\leq k_{R}\left(1+\left|x\right|\right)$,for any $x\in R^{n},u\in B\left(0,R\right)\subseteq U,\ t\in\left[0,t_{f}\right]$

$i_{2})\left|f\left(t,x\prime\prime,u\right)-f\left(t,x\prime,u\right)\right|\leq L_{R}^{\rho}\left|x\prime\prime-x\prime\right|$, for any $t\in\left[0,t_{f}\right]$,

$u\in B\left(0,R\right),x\prime,x\prime\prime\in B\left(x_{0},\rho\right)\subseteq R^{n}$, where $k_{R}\geq0,L_{R}^{\rho}\geq0$ are constant.

Let $\left(x^{*}\left(\cdot,\omega\right),u^{*}\left(\cdot,\omega\right)\right)$ a Nash equilibrium point associated with system (1) and functionals (2) for any $\omega\in\Omega$. Then

$y^{*}\left(t,\omega\right)=\left[G\left(t,\omega\right)\right]^{-1}\left(x^{*}\left(t,\omega\right)-\eta\left(t,\omega\right)\right)$ and $u^{*}\left(t,\omega\right),t\in\left[0,t_{f}\right]$ is a Nash equilibrium point for differential game $\Gamma_{N}\left(\omega\right)$ defined in (5) for each $\omega\in\Omega$.

The proof of this lemma proceeds easily if we use the solution representation $x^{*}\left(t,\omega\right)=G\left(t,\omega\right)\left(y^{*}\left(t,\omega\right)\right)+\eta\left(t,\omega\right)$ as in (6). Thus, supposing the conditions of differentiability for initial date $f\left(t,x,u\right)\in R^{n},F^{i}\left(x\right)\in R,f_{0}^{i}\left(t,x,u\right)\in R,i\in\{1,..,N\}$, relying on variable $x\in R^{n}$ we have the necessary conditions for Nash equilibrium for $\left(y^{*}\left(\cdot,\omega\right),u^{*}\left(\cdot,\omega\right)\right)$ which will represent the necessary conditions "reasonable" associated with the Nash equilibrium point $\left(x^{*}\left(\cdot,\omega\right),u^{*}\left(\cdot,\omega\right)\right)$ for each $\omega\in\Omega$.

Theorem 0.1. *Let $g_{j}\left(t,x\right)\in R^{n},j\in\{1,..,d\}$ fulfil the hypothesis $\left(i_{0}\right)$. Let $f\left(t,x,u\right)\in R^{n},F_{i}^{i}\left(x\right)\in R,i\in\{1,..,N\}$ fulfil the hypotheses $\left(i_{1}\right)$ and $\left(i_{2}\right)$*

from Lemma 5 and $f(t,\cdot,u) \in C^1(R^n,R^n), F^i(\cdot), f_0^i(t,\cdot,u) \in C^1(R^n,R)$ *for any* $(t,u) \in [0,t_f] \times U, i \in \{1,..,N\}$. *Let* $(x^*(\cdot,\omega), u^*(\cdot,\omega))$ *be a Nash equilibrium solution associated with dynamics (1) and functionals (2). Then the differentiable functions* $\psi^i(t,\omega), t \in [0,t_f], i \in \{1,..,N\}$ *exist such that with*

$$H^i\left(\omega,t,x,u_{(-i)},u_i,\psi^i\right) \triangleq \left(\psi^i\right)^T K(t,\omega) f(t,x,u_1,..,u_N)$$

the following equations

$$c_1)\ d_t x^*(t,\omega) = f(t,x^*(t,\omega),u^*(t,\omega))\,dt + \sum_{j=1}^{d} g_j(t,x^*(t,\omega)) \otimes dw_j(t),$$

$x^*(0,\omega) = x_0,$

$$c_2)\ \left(\frac{d\psi^i}{dt}(t,\omega)\right)^T = -\nabla_x H^i\left(\omega,t,x^*(t,\omega),u^*(t,\omega),\psi^i(t,\omega)\right)^T G(t,\omega),$$

$$c_3)\ \left(\psi^i(t_f,\omega)\right)^T = \left(\nabla_x F^i(x^*(t_f,\omega))\right)^T G(t_f,\omega),$$

$$c_4)\ H^i\omega,t,x^*(t,\omega),u^*(t,\omega),\psi^i(t,\omega) =$$
$$= \min_{u_i\in U_i} H^i\left(\omega,t,x^*(t,\omega),u^*_{(-i)}(t,\omega),u_i,\psi^i(t,\omega)\right),$$

hold for any $i \in \{1,..,N\}, t \in [0,t_f]$, *for* $u^*(\cdot,\omega)$ *and any* $\omega \in \Omega$, *where* $u_{(-i)} = (u_1,..,u_{i-1},u_{i+1},..,u_N)$ *and the* $(n \times n)$ *matrix satisfies (6.3).*

The proof of this theorem is a direct consequence of Lemma 5 if we write the necessary conditions for $(y^*(\cdot,\omega), u^*(\cdot,\omega))$.

References

[1] W. M. McEneancy, G. G. Yin, Q. Zang (eds.), *Stochastic analysis, control, optimization and applications*, A volume in Honor of W.H.Fleming, Birkhäuser, Basel, 1999.

[2] A. Friedman, *Stochastic differential equations and applications*, I, Academic, New York, 1975.

[3] D. Ijacu, C. Vârsan, *Smooth mappings and non $\mathcal{F}_t$ adapted solutions associated with Hamilton Iacobi stochastic equations*, IFIP Proceedings "Analysis and optimization of differential systems", Kluwer, Dordrecht, 2003.

[4] D. Ijacu, I. Molnar, C. Vârsan, *On some parabolic SPDE involving gradient representation of stochastic flows*, Rev. Roum. Pures Appl., **LI**, 4(2006).

[5] D. Ijacu, C. Vârsan, *Stochastic differential equations associated with smooth mappings and non $\mathcal{F}_t$ adapted solutions*, Preprint nr. 6/2003.

ROMAI J., **4**, 1(2008), 119–129

IMPROVEMENT OF AN INEQUALITY FOR THE SOLUTION OF THE TWO-DIMENSIONAL NAVIER-STOKES EQUATIONS

Anca-Veronica Ion

"Gh. Mihoc - Caius Iacob" Institute of Mathematical Statistic and Applied Mathematics, Bucharest

averionro@yahoo.com

Abstract An improvement of a classical estimate of the norm of a certain projection of the solution of the two-dimensional Navier-Stokes equations, for periodic boundary conditions, is formulated and proved.

Keywords: Navier-Stokes equations, large time behaviour of solutions.

2000 MSC: 35B05, 35B40.

1. INTRODUCTION

We consider the Navier-Stokes equations for a two-dimensional flow, with periodic boundary conditions.

As in the usual Galerkin method for this problem, the Hilbert space of functions used as phase space (defined in Section 2) is split into a direct sum of two subspaces: one is the finite dimensional space spanned by the eigenfunctions of the linear operator $\mathbf{A}$ corresponding to a finite set, Γ_m (defined in Section 3), of eigenvalues of $\mathbf{A}$, and the other is the orthogonal complement of the first.

The solution $\mathbf{u}$ of the Navier-Stokes equations and these equations themselves are projected on these subspaces. Let $\mathbf{P}$ be the projector of $\mathcal{H}$ on the finite dimensional space, and define $\mathbf{Q} = \mathbf{I} - \mathbf{P}$, $\mathbf{p} = \mathbf{Pu}$, and $\mathbf{q} = \mathbf{Qu}$, hence the decomposition $\mathbf{u} = \mathbf{p} + \mathbf{q}$.

In [3] an estimate for $\mathbf{q}$ is proved, namely the $\left[L^2(\Omega)\right]^2$ norm of this function is found to be less than $K_0 L^{1/2}\delta$, with K_0 depending on the data of the problem (kinematic viscosity coefficient, area of periodicity cell, body forces), $\delta = \lambda/\Lambda$, where λ is the least eigenvalue of $\mathbf{A}$, Λ is the least eigenvalue of $\mathbf{A}$ not belonging to Γ_m, and L is an increasing function of the number of eigenvalues in Γ_m (it is of the order of $\ln \Lambda$).

In the literature, this inequality is used (e.g. [6], [2]) to estimate the distance between an approximate inertial manifold of the Navier-Stokes problem (belonging to the family of approximate inertial manifolds defined in [3], [6], [7]) and the exact solution. For the j^{th} approximate inertial manifold defined in [6], [7], this distance is found to be less than $\kappa_j L^{(j+1)/2} \delta^{(j+3)/2}$, with κ_j depending on the data, therefore, it is clear that the presence of L affects the accuracy of the estimate.

In this paper we obtain independent of L estimates for the norm of $\mathbf{q}$.

2. THE EQUATIONS, THE FUNCTIONAL FRAMEWORK

Let $\Omega = (0,l) \times (0,l)$ be a domain in $\mathbb{R}^2$, representing the periodicity cell for the flow of a viscous incompressible fluid filling the plane. The plane flow in Ω of this fluid is modeled by the Navier-Stokes equations, with periodic boundary conditions on the boundary $\partial\Omega$ of Ω

$$\frac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \boldsymbol{\nabla})\, \mathbf{u} + \nabla p = \mathbf{f}, \tag{1}$$

$$\operatorname{div} \mathbf{u} = 0, \tag{2}$$

$$\mathbf{u}(t, \mathbf{x}) = \mathbf{u}(t, \mathbf{x} + l\mathbf{e}_j),\ t \geq 0,\ \mathbf{x} \in \Omega,\ j = 1, 2, \tag{3}$$

$$\mathbf{u}(0, \cdot) = \mathbf{u}_0(\cdot), \tag{4}$$

where $\mathbf{u} = \mathbf{u}(t, \mathbf{x}) \in \mathbb{R}^2$ is the fluid velocity, $t \in \mathbb{R}^+$ is the time, $\mathbf{x} \in \Omega$, $p = p(t, \mathbf{x}) \in \mathbb{R}$ is the fluid pressure, ν is the coefficient of kinematic viscosity, and $\mathbf{f}$ is the body force. The vectors $\mathbf{e}_j$, $j = 1, 2$ in condition (3) are the versors of the axes in $\mathbb{R}^2$.

Assume that $\mathbf{f}$ is independent of time and is an element of $\left[L^2_{per}(\Omega)\right]^2$, that is the space of $\left[L^2(\mathbb{R}^2)\right]^2$ functions, which are periodic, with Ω as periodicity cell. Due to the periodicity of the boundary conditions, it is usually assumed that [8], [5]

$$\bar{\mathbf{f}} = \frac{1}{l^2}\int_\Omega \mathbf{f}(\mathbf{x})\,d\mathbf{x} = \mathbf{0}, \tag{5}$$

and that the pressure is a periodic function. For simplicity we assume also that the average $\overline{\mathbf{u}}$ of the velocity over the periodicity cell is equal to zero, too.

The velocity $\mathbf{u}$ is thus looked for in the space $\mathcal{H} = \left\{\mathbf{v};\ \mathbf{v}\in\left[L^2_{per}(\Omega)\right]^2,\ \operatorname{div}\mathbf{v}=\mathbf{0},\ \bar{\mathbf{v}}=0\right\}$. The scalar product in $\mathcal{H}$ is $(\mathbf{u},\mathbf{v}) = \int_\Omega (u_1v_1+u_2v_2)\,dx$, (where $\mathbf{u}=(u_1,u_2)$, $\mathbf{v}=(v_1,v_2)$). The induced norm is denoted by $|\quad|$.

We also define the space $\mathcal{V} = \left\{\mathbf{v}\in\left[H^1_{per}(\Omega)\right]^2,\ div\,\mathbf{v}=\mathbf{0},\bar{\mathbf{v}}=0\right\}$, with the scalar product $((\mathbf{u},\mathbf{v})) = \sum_{i,j=1}^2\left(\frac{\partial u_i}{\partial x_j},\frac{\partial v_i}{\partial x_j}\right)$, and the induced norm, denoted by $\|\ \|$.

Then the variational formulation of the Navier-Stokes equations, obtained by projecting the generalized equations on the space of solenoidal vectors [8], reads as the Cauchy problem

$$\frac{d\mathbf{u}}{dt} - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\boldsymbol{\nabla})\,\mathbf{u} = \mathbf{f} \quad \text{in } \mathcal{V}', \tag{6}$$

$$\mathbf{u}(0) = \mathbf{u}_0,\ \ \mathbf{u}_0\in\mathcal{H}. \tag{7}$$

In the sequel the notations $\mathbf{B}(\mathbf{u},\mathbf{v}) = (\mathbf{u}\cdot\boldsymbol{\nabla})\mathbf{v}$, $\mathbf{B}(\mathbf{u}) = \mathbf{B}(\mathbf{u},\mathbf{u})$, $\mathbf{b}(\mathbf{u},\mathbf{v},\mathbf{w}) = (\mathbf{B}(\mathbf{u},\mathbf{v}),\mathbf{w})$ will be used.

For the bilinear mapping $\mathbf{B}(\mathbf{u},\mathbf{v})$ the following inequalities

$$|\mathbf{B}(\mathbf{u},\mathbf{v})| \le c_1\,|\mathbf{u}|^{\frac12}\,|\boldsymbol{\Delta}\mathbf{u}|^{\frac12}\,\|\mathbf{v}\|, \qquad (\forall)\ \ \mathbf{u}\in D(\mathbf{A}),\ \mathbf{v}\in\mathcal{V}, \tag{8}$$

$$|\mathbf{B}(\mathbf{u},\mathbf{v})| \le c_2\,\|\mathbf{u}\|\,\|\mathbf{v}\|\left[1+\ln\left(\frac{|\boldsymbol{\Delta}\mathbf{u}|^2}{\lambda_1\|\mathbf{u}\|^2}\right)\right]^{\frac12},\ (\forall)\ \ \mathbf{u}\in D(\mathbf{A}),\ \mathbf{v}\in\mathcal{V}. \tag{9}$$

hold [4], [8], [6]. We remind the following properties of the trilinear form $\mathbf{b}(\mathbf{u},\mathbf{v},\mathbf{w})$ (valid for periodic boundary conditions [5]):

$$\mathbf{b}(\mathbf{u},\mathbf{v},\mathbf{w}) = -\mathbf{b}(\mathbf{u},\mathbf{w},\mathbf{v}), \tag{10}$$

$$\mathbf{b}(\mathbf{u},\mathbf{v},\mathbf{v}) = 0, \tag{11}$$

as well as the following inequalities [5]

$$|\mathbf{b}(\mathbf{u},\mathbf{v},\mathbf{w})| \le c_3\,|\mathbf{u}|^{\frac12}\,\|\mathbf{u}\|^{\frac12}\,\|\mathbf{v}\|\,|\mathbf{w}|^{\frac12}\,\|\mathbf{w}\|^{\frac12},\ (\forall)\,\mathbf{u},\mathbf{v},\mathbf{w}\in\mathcal{V}, \tag{12}$$

$$|\mathbf{b}(\mathbf{u},\mathbf{v},\mathbf{w})| \leq c_4 |\mathbf{u}|^{\frac{1}{2}} \|\mathbf{u}\|^{\frac{1}{2}} \|\mathbf{v}\|^{\frac{1}{2}} |\Delta\mathbf{v}|^{\frac{1}{2}} |\mathbf{w}|, \ (\forall)\, \mathbf{u} \in \mathcal{V},\ \mathbf{v} \in \mathbf{D}(\mathbf{A}), \mathbf{w} \in \mathcal{H}. \tag{13}$$

We denote $\mathbf{A} = -\mathbf{\Delta}$ and note that $\mathbf{A}$ is defined on $D(\mathbf{A}) = \mathcal{V} \cap H^2(\Omega)$. For the problem (6), (7) we have the classical existence and uniqueness results:

Theorem 1 [8]. *a) If* $\mathbf{u}_0 \in \mathcal{H}$, $\mathbf{f} \in \mathcal{H}$, *then the problem* (6), (7) *has an unique solution* $\mathbf{u} \in C^0([0,T];\mathcal{H}) \cap L^2(0,T;\mathcal{V})$, *for any* $T > 0$.

b) If, in addition, $\mathbf{u}_0 \in \mathcal{V}$, *then* $\mathbf{u} \in C^0([0,T];\mathcal{V}) \cap L^2(0,T;D(\mathbf{A}))$, *for any* $T > 0$. *Moreover, the solution is analytic in time for* $t > 0$.

The semi-dynamical system $\{S(t)\}_{t\geq 0}$ generated by problem (6) is dissipative [9], [5]. More precisely, there is a $\rho_0 > 0$ such that for every $R > 0$, there is a $t_0(R) > 0$ with the property that for every $\mathbf{u}_0 \in \mathcal{H}$ with $|\mathbf{u}_0| \leq R$, we have $|S(t)\mathbf{u}_0| \leq \rho_0$ for $t > t_0(R)$. In addition, there are absorbing balls in $\mathcal{V}$ and $D(\mathbf{A})$ for $\{S(t)\}_{t\geq 0}$, i.e. there are $\rho_1 > 0$, $\rho_2 > 0$ and $t_1(R)$, $t_2(R)$ with $t_2(R) \geq t_1(R) \geq t_0(R)$ such that for every $R > 0$, $|\mathbf{u}_0| \leq R$ implies $\|S(t)\mathbf{u}_0\| \leq \rho_1$ for $t > t_1(R)$ and $|\mathbf{A}S(t)\mathbf{u}_0| \leq \rho_2$ for $t > t_2(R)$.

3. THE DECOMPOSITION OF THE SPACE, THE PROJECTED EQUATIONS

The eigenvalues of $\mathbf{A}$ are $\lambda_{j_1,j_2} = \frac{4\pi^2}{l^2}\left(j_1^2 + j_2^2\right)$, $(j_1,\ j_2) \in \mathbb{N}^2 \setminus \{(0,0)\}$, and the corresponding eigenfunctions are [6]

$$\begin{aligned}
\mathbf{w}^{s\pm}_{j_1,j_2} &= \frac{\sqrt{2}}{l}\frac{(j_2, \mp j_1)}{|\mathbf{j}|} \sin\left(2\pi\frac{j_1x_1 \pm j_2x_2}{l}\right),\\
\mathbf{w}^{c\pm}_{j_1,j_2} &= \frac{\sqrt{2}}{l}\frac{(j_2, \mp j_1)}{|\mathbf{j}|} \cos\left(2\pi\frac{j_1x_1 \pm j_2x_2}{l}\right),
\end{aligned}$$

where $|\mathbf{j}| = \left(j_1^2 + j_2^2\right)^{\frac{1}{2}}$. These eigenfunctions form a total system for $\mathcal{H}$. In order to be able to write easily sums involving the four eigenfunctions above, we denote them as follows

$$\mathbf{w}^{s+}_{j_1,j_2} = \mathbf{w}^1_{j_1,j_2},\ \ \mathbf{w}^{s-}_{j_1,j_2} = \mathbf{w}^2_{j_1,j_2},\ \ \mathbf{w}^{c+}_{j_1,j_2} = \mathbf{w}^3_{j_1,j_2},\ \ \mathbf{w}^{c-}_{j_1,j_2} = \mathbf{w}^4_{j_1,j_2}.$$

For a fixed $m \in \mathbb{N}$ we consider the set Γ_m of eigenvalues λ_{j_1,j_2} having $0 \leq j_1, j_2 \leq m$ and define

$$\lambda := \lambda_{1,0} = \lambda_{0,1} = \frac{4\pi^2}{l^2}, \tag{14}$$

$$\Lambda := \lambda_{m+1,0} = \lambda_{0,m+1} = \frac{4\pi^2}{l^2}(m+1)^2, \tag{15}$$

$$\delta = \delta(m) := \frac{\lambda}{\Lambda} = \frac{1}{(m+1)^2}. \tag{16}$$

Λ is the least eigenvalue not belonging to Γ_m. The eigenfunctions corresponding to the eigenvalues of Γ_m span a finite-dimensional subspace, $\mathcal{H}_m$, of $\mathcal{H}$. We denote by $\mathbf{P}$ the orthogonal projection operator on this subspace and by $\mathbf{Q}$ the orthogonal projection operator on the complementary subspace, i.e. $\mathbf{P} : \mathcal{H} \to \mathcal{H}_m$, $\mathbf{Q} = \mathbf{I} - \mathbf{P}$, $\mathbf{I}$ being the identity operator. Then the solution $\mathbf{u}$ of (6)-(7) reads $\mathbf{u} = \mathbf{p} + \mathbf{q}$, where

$$\mathbf{p} = \mathbf{P}\mathbf{u},\ \mathbf{q} = \mathbf{Q}\mathbf{u}.$$

By projecting equation (6) on the above constructed spaces, we obtain

$$\frac{d\mathbf{p}}{dt} - \nu\Delta\mathbf{p} + \mathbf{PB}(\mathbf{p}+\mathbf{q}) = \mathbf{Pf}, \tag{17}$$

$$\frac{d\mathbf{q}}{dt} - \nu\Delta\mathbf{q} + \mathbf{QB}(\mathbf{p}+\mathbf{q}) = \mathbf{Qf}. \tag{18}$$

4. OUR NEW ESTIMATES FOR THE “SMALL” COMPONENT OF THE SOLUTION

In [3] it is proved that for every $R > 0$, there is an instant $t_3(R) \geq t_2(R)$, such that, for every $|\mathbf{u}_0| \leq R$,

$$|\mathbf{q}(t)| \leq K_0 L^{\frac{1}{2}} \delta, \quad \|\mathbf{q}(t)\| \leq K_1 L^{\frac{1}{2}} \delta^{\frac{1}{2}}, \tag{19}$$
$$|\mathbf{q}'(t)| \leq K_0' L^{\frac{1}{2}} \delta, \quad |\Delta\mathbf{q}(t)| \leq K_2 L^{\frac{1}{2}}, \qquad t \geq t_3(R),$$

where K_0, K_0', K_1, K_2 depend of ν, $|f|$, λ and, for our choice of the projection subspaces, $L = L(m) = 1 + \ln 2m^2$ (see also [6]). The constant L occurs when using inequality (9) to prove (19). More specific,

$$L = \sup_{\mathbf{p}\in\mathbf{P}\mathcal{H}} \left(1 + \ln \frac{|\Delta\mathbf{p}|^2}{\lambda_1 \|\mathbf{p}\|^2}\right) = \max_{\lambda\in\Gamma_m} \left(1 + \ln\frac{\lambda}{\lambda_1}\right) = 1 + \ln 2m^2. \tag{20}$$

We note that L tends to infinity as $m \to \infty$. In the sequel we improve the estimates (19), such that L will not occur anymore in our new inequalities. The idea is that of refining the contribution of the term $\mathbf{QB}(\mathbf{p})$ from $\mathbf{QB}(\mathbf{p}+\mathbf{q})$ in (18).

We start from the trigonometric relation

$$\sin\left(2\pi\frac{j_1x_1 \pm j_2x_2}{l}\right)\sin\left(2\pi\frac{k_1x_1 \pm k_2x_2}{l}\right) = \frac{1}{2}\left[\cos 2\pi\frac{(j_1-k_1)\,x_1 \pm (j_2-k_2)\,x_2}{l} - \right.$$

$$\left. - \cos 2\pi\frac{(j_1+k_1)\,x_1 \pm (j_2+k_2)\,x_2}{l}\right],$$

and the similar ones for all other combinations of sine and cosine that might occur in the scalar product of two eigenfunctions. Since $\mathbf{p} = \sum\limits_{0\le j_1,j_2\le m}\sum\limits_{i=1}^{4} p^i_{j_1,j_2}\mathbf{w}^i_{j_1,j_2}$, from the term $(\mathbf{p}\boldsymbol{\nabla})\,\mathbf{p} = \left(\sum\limits_{0\le j_1,j_2\le m}\sum\limits_{i=1}^{4} p^i_{j_1,j_2}\mathbf{w}^i_{j_1,j_2}\nabla\right)\left(\sum\limits_{0\le k_1,k_2\le m}\sum\limits_{l=1}^{4} p^l_{k_1,k_2}\mathbf{w}^l_{k_1,k_2}\right)$, only those products of terms that have $j_1+k_1 \ge m+1$ or $j_2+k_2 \ge m+1$ will belong to $\mathbf{Q}\mathcal{H}$.

From this point on we assume that m is even, and we put $m = 2n$.

If for $\mathbf{w}^i_{j_1,j_2}$ and $\mathbf{w}^l_{k_1,k_2}$ we have $j_1, j_2 \le n$ and $k_1, k_2 \le n$, then $\left(\mathbf{w}^i_{j_1,j_2}\nabla\right)\mathbf{w}^l_{k_1,k_2}$ belongs to $\mathbf{P}\mathcal{H}$. Whence the idea of defining the subspace of $\mathcal{H}$ spanned by the eigenfunctions $\mathbf{w}^i_{j_1,j_2}$ with $0 \le j_1,\ j_2 \le n$, $1 \le i \le 4$. We denote this space by $\mathcal{H}_n$, the projection operator on this subspace by $\mathbf{P_p}$ and we set: $\mathbf{P_q} = \mathbf{P} - \mathbf{P_p}$, $\mathbf{p}_p = \mathbf{P_p}\mathbf{p}$, $\mathbf{p}_q = \mathbf{P_q}\mathbf{p}$. Obviously

$$\mathbf{Q}\,(\mathbf{p}_p\boldsymbol{\nabla})\,\mathbf{p}_p\!= \mathbf{0}. \tag{21}$$

Then, by setting $\delta_1 = \delta\,(n) = \frac{1}{(n+1)^2}$, $L_1 = L(n) = 1 + \ln 2n^2$, the estimates (19) imply

$$|(\mathbf{I}-\mathbf{P_p})\mathbf{u}| \le K_0L_1^{\frac{1}{2}}\delta_1, \quad \|(\mathbf{I}-\mathbf{P_p})\mathbf{u}\| \le K_1L_1^{\frac{1}{2}}\delta_1^{\frac{1}{2}}, \tag{22}$$

$$\left|(\mathbf{I}-\mathbf{P_p})\mathbf{u}'\right| \le K_0'L_1^{\frac{1}{2}}\delta_1, \quad |\Delta(\mathbf{I}-\mathbf{P_p})\mathbf{u}| \le K_2L_1^{\frac{1}{2}}. \tag{23}$$

On the other hand, we note that $(\mathbf{I}-\mathbf{P_p})\mathbf{u} = \mathbf{p_q} + \mathbf{q}$, hence $|\mathbf{p}_q| \le |(\mathbf{I}-\mathbf{P_p})\mathbf{u}|$, $\|\mathbf{p}_q\| \le \|(\mathbf{I}-\mathbf{P_p})\mathbf{u}\|$.

We use these inequalities to refine the estimates (19). In the rest of the paper we assume that for a fixed $R \ge 0$, the function $\mathbf{u}_0$ is such that $|\mathbf{u}_0| \le R$. We assert and prove

Theorem 1. *There are some constants* $\widetilde{C}_0$, $\widetilde{C}_1$, $\widetilde{C}\,'_0$, $\widetilde{C}_2$, *depending only on* ν, λ, $|\mathbf{Qf}|$ *such that, for* t *large enough, the inequalities*

$$|\mathbf{q}(t)| \le \widetilde{C}_0\delta, \tag{24}$$

$$\|\mathbf{q}(t)\| \leq \widetilde{C}_1 \delta^{\frac{1}{2}}, \tag{25}$$

$$\left|\mathbf{q}'(t)\right| \leq \widetilde{C}'_0 \delta, \tag{26}$$

$$|\Delta \mathbf{q}(t)| \leq \widetilde{C}_2, \tag{27}$$

hold.

Proof. With the notation introduced before the Proposition, we have

$$\begin{aligned}\mathbf{QB}(\mathbf{p}) = \mathbf{QB}(\mathbf{p}_p + \mathbf{p}_q) = \mathbf{QB}(\mathbf{p}_p) + \mathbf{QB}(\mathbf{p}_p, \mathbf{p}_q) + \mathbf{QB}(\mathbf{p}_q, \mathbf{p}_p) + \mathbf{QB}(\mathbf{p}_q) \quad (28)\\ = \mathbf{QB}(\mathbf{p}_p, \mathbf{p}_q) + \mathbf{QB}(\mathbf{p}_q, \mathbf{p}_p) + \mathbf{QB}(\mathbf{p}_q),\end{aligned}$$

since $\mathbf{QB}(\mathbf{p}_p) = 0$. Now we take the scalar product of (18) by $\mathbf{q}$, and by using (11) and (28), we obtain

$$\begin{aligned}\frac{1}{2}\frac{d\,|\mathbf{q}|^2}{dt} + \nu\,\|\mathbf{q}\|^2 \leq \left|\left(\mathbf{B}(\mathbf{p}_p, \mathbf{p}_q), \mathbf{q}\right)\right| + \left|\left(\mathbf{B}(\mathbf{p}_q, \mathbf{p}_p), \mathbf{q}\right)\right| + \quad (29)\\ + \left|(\mathbf{B}(\mathbf{p}_q), \mathbf{q})\right| + \left|(\mathbf{B}(\mathbf{q}, \mathbf{p}), \mathbf{q})\right| + \left|(\mathbf{Qf}, \mathbf{q})\right|.\end{aligned}$$

For the first term of the right-hand side, the following estimates, obtained by using (8), the dissipativity of the problem and (22), hold

$$\begin{aligned}\left|\left(\mathbf{B}(\mathbf{p}_p, \mathbf{p}_q), \mathbf{q}\right)\right| \leq c_1 \left|\mathbf{p}_p\right|^{1/2} \left|\Delta \mathbf{p}_p\right|^{1/2} \|\mathbf{p}_q\| \, |\mathbf{q}| \leq c_1 \rho_0^{1/2} \rho_2^{1/2} K_1 L_1^{\frac{1}{2}} \delta_1^{\frac{1}{2}} \frac{1}{\Lambda^{\frac{1}{2}}} \, \|\mathbf{q}\|\\ \leq c_1^2 \rho_0 \rho_2 K_1^2 L_1 \delta_1 \frac{2}{\nu\Lambda} + \frac{\nu}{8} \|\mathbf{q}\|^2.\end{aligned}$$

For the second term we obtain, with the same arguments and (23),

$$\begin{aligned}\left|\left(\mathbf{B}(\mathbf{p}_q, \mathbf{p}_p), \mathbf{q}\right)\right| &\leq c_1 \left|\mathbf{p}_q\right|^{1/2} \left|\Delta \mathbf{p}_q\right|^{1/2} \|\mathbf{p}_p\| \, |\mathbf{q}|\\ &\leq c_1 K_0^{1/2} L_1^{\frac{1}{4}} \delta_1^{\frac{1}{2}} K_2^{1/2} L_1^{\frac{1}{4}} \rho_1 \frac{1}{\Lambda^{\frac{1}{2}}} \, \|\mathbf{q}\|\\ &\leq c_1^2 K_0 K_2 \rho_1^2 L_1 \delta_1 \frac{2}{\nu\Lambda} + \frac{\nu}{8} \|\mathbf{q}\|^2.\end{aligned}$$

For the third term we have, by using (9) and the above arguments,

$$\begin{aligned}\left|(\mathbf{B}(\mathbf{p}_q), \mathbf{q})\right| &\leq c_2 L^{\frac{1}{2}} \|\mathbf{p}_q\|^2 \, |\mathbf{q}| \leq c_2 L^{\frac{1}{2}} K_1^2 L_1 \delta_1 \frac{1}{\Lambda^{\frac{1}{2}}} \, \|\mathbf{q}\|\\ &\leq c_2^2 K_1^4 L L_1^2 \delta_1^2 \frac{2}{\nu\Lambda} + \frac{\nu}{8} \|\mathbf{q}\|^2,\end{aligned}$$

and for the fourth, by using (13) and the above arguments,

$$\begin{aligned}|(\mathbf{B}(\mathbf{q},\mathbf{p}),\mathbf{q})| &\leq c_4 |\mathbf{q}|^{\frac{1}{2}} \|\mathbf{q}\|^{\frac{1}{2}} \|\mathbf{p}\|^{\frac{1}{2}} |\Delta \mathbf{p}|^{\frac{1}{2}} |\mathbf{q}| \leq \\ &\leq c_4 K_0^{1/2} L^{\frac{1}{4}} \delta^{\frac{1}{2}} K_1^{1/2} L^{\frac{1}{4}} \delta^{\frac{1}{4}} \rho_1^{\frac{1}{2}} \rho_2^{\frac{1}{2}} \frac{1}{\Lambda^{\frac{1}{2}}} \|\mathbf{q}\| \\ &\leq c_4^2 K_0 K_1 \rho_1 \rho_2 L \delta^{\frac{3}{2}} \frac{2}{\nu\Lambda} + \frac{\nu}{8} \|\mathbf{q}\|^2 .\end{aligned}$$

At last

$$|(\mathbf{Qf},\mathbf{q})| \leq |\mathbf{Qf}|\, |\mathbf{q}| \leq \frac{2\,|\mathbf{Qf}|^2}{\nu\Lambda} + \frac{\nu}{8} \|\mathbf{q}\|^2 .$$

The above inequalities and (29) imply

$$\frac{1}{2}\frac{d}{dt}|\mathbf{q}|^2 + 3\frac{\nu\Lambda}{8}|\mathbf{q}|^2 \leq C_0^2 \delta,$$

where

$$\begin{aligned}C_0^2 &= \frac{2}{\nu\lambda}\left[c_1^2 \rho_0 \rho_2 K_1^2 L_1 \delta_1 + c_1^2 K_0 K_2 \rho_1^2 L_1 \delta_1 + c_2^2 K_1^4 L L_1^2 \delta_1^2 + \right. \\ &\left. + c_4^2 K_0 K_1 \rho_1 \rho_2 L \delta^{\frac{3}{2}} + |\mathbf{Qf}|^2 \right].\end{aligned}$$

By using the Gronwall Lemma, it follows,

$$|\mathbf{q}(t)|^2 \leq |\mathbf{q}(0)|^2 e^{-\frac{3}{4}\nu\Lambda t} + \frac{8C_0^2}{3\nu\lambda}\delta^2,$$

hence, for $t_4(R) \geq t_3(R)$, taken as to have $|\mathbf{q}(0)|^2 e^{-\frac{3}{4}\nu\Lambda t} \leq \frac{4C_0^2}{3\nu\lambda}\delta^2$ for $t \geq t_4(R)$, we obtain (24), with $\widetilde{C}_0 = \frac{2C_0}{\sqrt{\nu\lambda}}$.

The functions of $n:$ $L_1\delta_1 = L(n)\,\delta(n) = \left(1+\ln 2n^2\right)/(n+1)^2$, $L L_1^2 \delta_1^2 = L(2n)\, L(n)^2 \,\delta(n)^2 = \left(1+\ln 8n^2\right)\left(1+\ln 2n^2\right)^2/(n+1)^4$ and $L\delta^{\frac{3}{2}} = \left(1+\ln 8n^2\right)/(2n+1)^3$, that occur in the structure of C_0^2, have at $n=2$ values less than 1 and are decreasing as n increases (for $n \geq 2$). Then, for $n \geq 2$, we have

$$C_0^2 \leq \frac{2}{\nu\lambda}\left(c_1^2 \rho_0 \rho_2 K_1^2 + c_1^2 K_0 K_2 \rho_1^2 + c_2^2 K_1^4 + c_4^2 K_0 K_1 \rho_1 \rho_2 + |\mathbf{Qf}|^2\right)$$

and the right-hand side of this inequality depends only on $\nu,\ \lambda,\ |\mathbf{Qf}|$. More than that, since all functions defined above tend to zero as $n \to \infty$, we can choose n large enough so that $\frac{2|\mathbf{Qf}|^2}{\nu\lambda}$ becomes the leading term in C_0^2.

For that n, $\widetilde{C}_0$ will be of the order of $\frac{|\mathbf{Qf}|}{\nu\lambda}$. However, the structure of $K_0,\ K_1,\ \rho_0,\ \rho_1,\ \rho_2$ shows that if ν is very small, then n with the above property must be very large.

Now, we attempt to estimate $\|\mathbf{q}\|$. Multiplying equation (18) by $\mathbf{\Delta q}$, and using the equalities $\mathbf{u} = \mathbf{p} + \mathbf{q} = \mathbf{p_p} + \mathbf{p_q} + \mathbf{q}$ and (21), we obtain

$$\begin{aligned} \frac{1}{2}\frac{d\,\|\mathbf{q}\|^2}{dt} + \nu\,|\Delta\mathbf{q}|^2 \quad \leq \quad & \big|\big(\mathbf{B}(\mathbf{p}_p, \mathbf{p}_q + \mathbf{q}), \Delta\mathbf{q}\big)\big| + \big|\big(\mathbf{B}(\mathbf{p}_q{+}\mathbf{q}, \mathbf{p}_p), \Delta\mathbf{q}\big)\big| + \\ & + \big|\big(\mathbf{B}(\mathbf{p}_q{+}\mathbf{q}, \mathbf{p}_q{+}\mathbf{q}), \Delta\mathbf{q}\big)\big| + \\ & + |(\mathbf{Qf}, \mathbf{\Delta q})|\,. \end{aligned}$$

For the first term in the right-hand side with (8), by using the dissipativity of the problem and (22), we have

$$\begin{aligned} \big|\big(\mathbf{B}(\mathbf{p}_p, \mathbf{p}_q + \mathbf{q}), \Delta\mathbf{q}\big)\big| \quad & \leq \quad c_1\,|\mathbf{p}_p|^{\frac{1}{2}}\,|\mathbf{\Delta p}_p|^{\frac{1}{2}}\,\|\mathbf{p}_q + \mathbf{q}\|\,|\Delta\mathbf{q}| \\ & \leq \quad c_1\rho_0^{1/2}\rho_2^{1/2}L_1^{\frac{1}{2}}K_1\delta_1^{\frac{1}{2}}\,|\Delta\mathbf{q}| \\ & \leq \quad c_1^2\rho_0\rho_2K_1^2L_1\delta_1\frac{2}{\nu} + \frac{\nu}{8}\,|\Delta\mathbf{q}|^2\,, \end{aligned}$$

for the second, with (13), (24), (22) and the dissipativity of the problem,

$$\begin{aligned} \big|\big(\mathbf{B}(\mathbf{p}_q{+}\mathbf{q}, \mathbf{p}_p), \Delta\mathbf{q}\big)\big| \quad & \leq \quad c_4\,|\mathbf{p}_q{+}\mathbf{q}|^{\frac{1}{2}}\,\|\mathbf{p}_q{+}\mathbf{q}\|^{\frac{1}{2}}\,\|\mathbf{p}_p\|^{\frac{1}{2}}\,|\Delta\mathbf{p}_p|^{\frac{1}{2}}\,|\Delta\mathbf{q}| \\ & \leq \quad c_4\widetilde{C}_0^{\frac{1}{2}}\delta_1^{\frac{1}{2}}L_1^{\frac{1}{4}}K_1^{\frac{1}{2}}\delta_1^{\frac{1}{4}}\rho_1^{\frac{1}{2}}\rho_2^{\frac{1}{2}}\,|\Delta\mathbf{q}| \\ & \leq \quad c_4^2\widetilde{C}_0\rho_1\rho_2K_1L_1^{\frac{1}{2}}\delta_1^{\frac{3}{2}}\frac{2}{\nu} + \frac{\nu}{8}\,|\Delta\mathbf{q}|^2 \end{aligned}$$

for the third, with the above quoted results and (23)

$$\begin{aligned} \big|\big(\mathbf{B}(\mathbf{p}_q{+}\mathbf{q}, \mathbf{p}_q{+}\mathbf{q}), \Delta\mathbf{q}\big)\big| \quad & \leq \quad c_4\,|\mathbf{p}_q{+}\mathbf{q}|^{\frac{1}{2}}\,\|\mathbf{p}_q{+}\mathbf{q}\|\,|\Delta\,(\mathbf{p}_q{+}\mathbf{q})|^{\frac{1}{2}}\,|\Delta\mathbf{q}| \\ & \leq \quad c_4\widetilde{C}_0^{\frac{1}{2}}\delta_1^{\frac{1}{2}}L_1^{\frac{1}{2}}K_1\delta_1^{\frac{1}{2}}L_1^{\frac{1}{4}}K_2^{\frac{1}{2}}\,|\Delta\mathbf{q}| \\ & \leq \quad c_4^2\widetilde{C}_0K_1^2K_2L_1^{\frac{3}{2}}\delta_1^2\frac{2}{\nu} + \frac{\nu}{8}\,|\Delta\mathbf{q}|^2\,, \end{aligned}$$

and for the fourth

$$|(\mathbf{Qf}, \mathbf{\Delta q})| \leq \frac{2\,|\mathbf{Qf}|^2}{\nu} + \frac{\nu}{8}\,|\Delta\mathbf{q}|^2\,.$$

Denote

$$\begin{aligned} \frac{1}{2}C_1^2 \quad = \quad & \frac{2c_1^2}{\nu}\rho_0\rho_2K_1^2L_1\delta_1 + \frac{2c_4^2}{\nu}\widetilde{C}_0\rho_1\rho_2K_1L_1^{\frac{1}{2}}\delta_1^{\frac{3}{2}} + \\ & + \frac{2c_4^2}{\nu}\widetilde{C}_0K_1^2K_2L_1^{\frac{3}{2}}\delta_1^2 + \frac{2\,|\mathbf{Qf}|^2}{\nu}. \end{aligned}$$

Then the differential inequality for $\|\mathbf{q}\|$ becomes

$$\frac{d\|\mathbf{q}\|^2}{dt}+\nu\Lambda\|\mathbf{q}\|^2\leq C_1^2,$$

yielding

$$\|\mathbf{q}(t)\|^2\leq\|\mathbf{q}(0)\|^2e^{-\nu\Lambda t}+\frac{1}{\nu\lambda}C_1^2\delta.$$

Let $t_5(R)\geq t_4(R)$ be such that for $t\geq t_5(R)$ the inequality

$$\|\mathbf{q}(0)\|^2e^{-\nu\Lambda t}\leq\frac{1}{\nu\lambda}C_1^2\delta$$

is satisfied. For $t\geq t_5(R)$,

$$\|\mathbf{q}(t)\|^2\leq\frac{2}{\nu\lambda}C_1^2\delta \tag{30}$$

holds.

Remark that $L_1\delta_1=L(n)\delta(n)$, $L_1^{\frac{1}{2}}\delta_1^{\frac{3}{2}}=L(n)^{\frac{1}{2}}\delta(n)^{\frac{3}{2}}$, $L_1^{\frac{3}{2}}\delta_1^2=L(n)^{\frac{3}{2}}\delta(n)^2$ have values less than 1 for $n=2$, decrease as n increases beyond $n=2$ and tend to zero as $n\to\infty$. Hence,

$$\frac{2}{\nu\lambda}C_1^2\leq\frac{8}{\nu^2\lambda}[c_1^2\rho_0\rho_2K_1^2+c_4^2\widetilde{C}_0\rho_1\rho_2K_1+c_4^2\widetilde{C}_0K_1^2K_2+|\mathbf{Qf}|^2]. \tag{31}$$

We denote by $\widetilde{C}_1^2$ the right hand side of the above inequality. The inequalities (30) and (31) imply (25), with the coefficient $\widetilde{C}_1$ depending only on ν, λ, $|\mathbf{Qf}|$ and not on n.

Moreover, n can be chosen large enough such that each of the first three terms of C_1^2 becomes smaller than $\frac{|\mathbf{Qf}|^2}{\nu}$, hence for this n, $\widetilde{C}_1$ maybe taken of the order of $\frac{|\mathbf{Qf}|}{\nu\lambda^{\frac{1}{2}}}$.

With the same method as that used for the solution $\mathbf{u}(t)$ in [1], [8], it can be proved that $\mathbf{q}(t)$ is analytic in time for $t>0$ and is the restriction to the real axis of an analytic function of a complex variable defined on a neighborhood of the real axis. By using the Cauchy formula, we obtain (26).

Finally, from (18) we have

$$\Delta\mathbf{q}=\frac{1}{\nu}\left[\frac{d\mathbf{q}}{dt}+\mathbf{QB}(\mathbf{p}+\mathbf{q})-\mathbf{Qf}\right]$$

and with the above estimates we obtain (27).□

Acknowledgements

Work supported by the Romanian Ministry of Education and Research (Grant CEEX-05-D11-25/5.11.2005).

References

[1] P. Constantin, C. Foiaş, *Navier-Stokes equations,* Chicago Lectures in Math., Univ. of Chicago, 1988.

[2] C. Devulder, M. Marion, *A class of numerical algorithms for large time integration: the nonlinear Galerkin methods*, SIAM J. Numer. Anal., **29**(1992), 462-483.

[3] C. Foiaş, O. Manley, R. Temam, *Modelling of the interactions of the small and large eddies in two dimensional turbulent flows*, Math. Modelling and Num. Anal., **22**(1988), 93-114.

[4] A. Georgescu, *Hydrodynamic stability theory*, Kluwer, Dordrecht, 1985.

[5] J. C.Robinson, *Infinite-dimensional dynamical systems; An introduction to dissipative parabolic PDEs and the theory of global attractors,* Cambridge University Press, 2001.

[6] R. Temam, *Induced trajectories and approximate inertial manifolds*, Math. Mod. Num. Anal., **23(**1989**),** 541-561.

[7] R. Temam, *Attractors for the Navier-Stokes equations, localization and approximation,* J. Fac. Sci. Univ. Tokyo, Soc. IA, Math., **36**(1989), 629-647.

[8] R. Temam, *Navier-Stokes equations and nonlinear functional analysis,* CBMS-NSF Reg. Conf. Ser. in Appl. Math., SIAM, Philadelphia, 1995.

[9] R. Temam, *Infinite-dimensional dynamical systems in mechanics and physics,* Appl. Math. Sci., **68**, Springer, New York, 1997.

ROMAI J., **4**, 1(2008), 131–137

ON A DIFFUSION PROCESS INTERMEDIATE BETWEEN STANDARD BROWNIAN MOTION AND THE ORNSTEIN-UHLENBECK PROCESS

Mario Lefebvre

École Polytechnique de Montréal, Canada

mlefebvre@polymtl.ca

Abstract We first consider a time-inhomogeneous diffusion process that is a generalization of the standard Brownian motion. We find that it has a Gaussian probability density function with the same mean as an Ornstein-Uhlenbeck process, and variance that generalizes that of the standard Brownian motion. Next, the problem of finding diffusion processes having a Gaussian $\mathrm{N}(0, t)$ probability density function is treated.

Keywords: Wiener process, Kolmogorov forward equation, infinitesimal parameters.

2000 MSC: 60H15.

1. INTRODUCTION AND THEORETICAL RESULTS

Arguably the two most important diffusion processes are the Wiener process $\{W(t), t \geq 0\}$ and the Ornstein-Uhlenbeck process $\{U(t), t \geq 0\}$ defined respectively by the stochastic differential equations

$$dW(t) = \mu\, dt + \sigma\, dB_1(t)$$

and

$$dU(t) = -\alpha U(t)\, dt + \sigma\, dB_2(t),$$

where $\{B_i(t), t \geq 0\}$, $i = 1, 2$, is a standard Brownian motion, and $\mu \in \mathbb{R}$ and $\sigma > 0$ are constants. As is well known, conditional on $W(t_0) = w_0$ and $U(t_0) = u_0$, we may write that $W(t) \sim \mathrm{N}\left(w_0 + \mu(t - t_0), \sigma^2(t - t_0)\right)$ and $U(t) \sim \mathrm{N}\left(u_0 e^{-\alpha(t-t_0)}, \frac{\sigma^2}{2\alpha}\left[1 - e^{-2\alpha(t-t_0)}\right]\right)$ see Lefebvre (2007), pp. 184 and 203, for instance).

In this note, we first consider the time-inhomogeneous diffusion process $\{X(t), t \geq 0\}$ that satisfies the stochastic differential equation

$$dX(t) = -\frac{k}{2} X(t) \, dt + (1 + kt)^{1/2} \, dB(t), \tag{1}$$

where k is a non-negative constant and $\{B(t), t \geq 0\}$ is a standard Brownian motion. Notice that it generalizes the standard Brownian motion, which corresponds to the case when $k = 0$. That is, if $k = 0$, then $\{X(t), t \geq 0\}$ is a Wiener process with infinitesimal parameters $\mu = 0$ and $\sigma = 1$.

In Section 2, we find the probability density function of the random variable $X(t)$. We see that if $t_0 = 0$ and $x_0 = 0$, then $X(t)$ has the same probability density function as a standard Brownian motion, namely a Gaussian $\mathrm{N}(0, t)$ distribution. Then, in Section 3, we consider the problem of finding other diffusion processes having a Gaussian $\mathrm{N}(0, t)$ distribution. Finally, a few remarks conclude this work in Section 4.

2. PROBABILITY DENSITY FUNCTION OF $\boldsymbol{X(T)}$

Let $\Phi(t)$ be the function that satisfies the ordinary differential equation

$$\frac{d}{dt}\Phi(t) = -\frac{k}{2}\Phi(t), \quad \text{for } t > t_0,$$

subject to the initial condition $\Phi(t_0) = 1$. Its solution is $\Phi(t) = \exp\left\{-\frac{k}{2}(t - t_0)\right\}$, for $t \geq t_0$. We can state the following proposition.

Proposition 2.1. *Conditional on $X(t_0) = x_0$, the distribution of the random variable $X(t)$ is given by*

$$X(t) \sim N\left(x_0 \, e^{-(k/2)(t-t_0)}, t - t_0 \, e^{-k(t-t_0)}\right) \quad \text{for } t \geq t_0.$$

Proof. This result is an application of Proposition 4.3.1, p. 211, in Lefebvre (2007) [see Remark iii), p. 212]. Indeed, we deduce from this proposition that $X(t) \mid \{X(t_0) = x_0\}$ has a Gaussian distribution with mean

$$m(t) = \Phi(t)\left(x_0 + \int_{t_0}^{t} \Phi^{-1}(u) \cdot 0 \, du\right) = x_0 \, e^{-(k/2)(t-t_0)}$$

and variance $\sigma^2(t) = \Phi^2(t) \int_{t_0}^t \Phi^{-2}(u)\,(1+ku)\,du$. That is,

$$\sigma^2(t) = e^{-k(t-t_0)} \int_{t_0}^t e^{k(u-t_0)}\,(1+ku)\,du = t - t_0\,e^{-k(t-t_0)} \quad \text{for } t \geq t_0. \quad \blacksquare$$

Remarks. i) We can also obtain the probability density function of $X(t)$ by proceeding as follows: the function

$$f(x,t)\; (= f(x,t;x_0,t_0)) \;:= \frac{P\left[X(t) \in (x, x+dx) \mid X(t_0) = x_0\right]}{dx} \tag{2}$$

satisfies the Kolmogorov forward equation (also called Fokker-Planck equation; see Cox and Miller (1965), for instance)

$$\begin{aligned} & \frac{1}{2}\frac{\partial^2}{\partial x^2}\left\{(1+kt)\,f(x,t)\right\} - \frac{\partial}{\partial x}\left\{-\frac{k}{2}\,x\,f(x,t)\right\} = \frac{\partial}{\partial t} f(x,t) \\ \iff \quad & \frac{1+kt}{2}\,f_{xx} + \frac{k}{2}\,(f + x\,f_x) = f_t. \end{aligned}$$

Taking the Fourier transform on both sides of this partial differential equation, we obtain that $F(\omega,t) := \int_{-\infty}^{\infty} e^{i\omega x}\, f(x,t)\,dt$, where $\omega \in \mathbb{R}$, is a solution of

$$F_t + \frac{k}{2}\,\omega\,F_\omega + \frac{\omega^2}{2}\,(1+kt)F = 0. \tag{3}$$

Moreover, the function F is such that

$$F(0,t) = 1 \tag{4}$$

and

$$\lim_{t \downarrow t_0} F(\omega,t) = e^{i\omega x_0}. \tag{5}$$

This last condition follows from the fact that

$$\lim_{t \downarrow t_0} f(x,t) = \delta(x - x_0).$$

Next, the general solution of equation (3) can be written as $F(\omega,t) = e^{-\omega^2 t/2}\, G\left(\omega\, e^{-kt/2}\right)$, where G is an arbitrary function. We then infer from (4) and (5) that $F(\omega,t)$ is of the form

$$F(\omega,t) = \exp\left\{ i\omega\, x_0\, e^{-k(t-t_0)/2} - \frac{\omega^2}{2}\left[t - t_0\, e^{-k(t-t_0)}\right]\right\}.$$

Finally, the result in the proposition is obtained by remembering that if $X \sim \mathrm{N}(\mu, \sigma^2)$, then the characteristic function of X is given by $C_X(\omega) := E\left[e^{i\omega X}\right] = \exp\left\{i\omega\mu - \frac{\sigma^2}{2}\omega^2\right\}$.
ii) When the initial time is $t_0 = 0$, the distribution of $X(t)$ reduces to

$$X(t) \mid \{X(0) = x_0\} \sim \mathrm{N}\left(x_0 e^{-kt/2}, t\right).$$

Notice that the mean of $X(t)$ is the same as that of an Ornstein-Uhlenbeck process with $\alpha = k/2$, while its variance corresponds to that of a standard Brownian motion. Furthermore, if the process starts at $x_0 = 0$, then $X(t) \sim \mathrm{N}(0, t)$. In the next section, the problem of finding diffusion processes having the same probability density function as a standard Brownian motion starting at the origin will be treated.

iii) From the previous remark, we can state that the diffusion process $\{X(t), t \geq 0\}$ defined by (1) is intermediate between the Wiener process with $\mu = 0$ and $\sigma = 1$, and the Ornstein-Uhlenbeck process with $\alpha = k/2$. In applications where the mean of $X(t)$ tends to 0 with increasing t, rather than remaining constant, and the variance of $X(t)$ is a linear function of t, this process would be a model better than either the Wiener or the Ornstein-Uhlenbeck process. In the case of the Ornstein-Uhlenbeck process, its variance is bounded (from above) by $\sigma^2/(2\alpha)$ $(= \sigma^2/k)$.

3. DIFFUSION PROCESSES HAVING A GAUSSIAN PROBABILITY DENSITY FUNCTION

In the preceding section, we found that the diffusion process $\{X(t), t \geq 0\}$ defined by (1) has the same probability density function as a standard Brownian motion starting from the origin, if $x_0 = 0$ and $t_0 = 0$. Now, we try to find other diffusion processes having a Gaussian $\mathrm{N}(0, t)$ probability density function.

Let $m(x, t)$ and $v(x, t) \geq 0$ be the infinitesimal parameters of $\{X(t), t \geq 0\}$. These functions must be such that (see Lamberton and Lapeyre (1997, p. 58), in particular), for all $s \geq 0$,

$$\int_0^s |m(x,t)|\,dt < \infty \quad \text{and} \quad \int_0^s v(x,t)\,dt < \infty. \tag{6}$$

Then, the function $f(x,t)$ defined in (2) satisfies the Kolmogorov forward equation

$$\frac{1}{2}\frac{\partial^2}{\partial x^2}\{v(x,t)f(x,t)\} - \frac{\partial}{\partial x}\{m(x,t)f(x,t)\} = \frac{\partial}{\partial t}f(x,t). \tag{7}$$

When $v(x,t) \equiv 1$ and $m(x,t) \equiv 0$, we know that (if $x_0 = 0$ and $t_0 = 0$)

$$f(x,t) = \frac{1}{\sqrt{2\pi t}} \exp\left\{-\frac{x^2}{2t}\right\}$$

for $x \in \mathbb{R}$ and $t > 0$. Substituting the function $f(x,t)$ into (7), we obtain that

$$\frac{1}{2}\left\{v\left(\frac{x^2}{t^2} - \frac{1}{t}\right) + 2v_x\left(-\frac{x}{t}\right) + v_{xx}\right\} - \left\{m_x + m\left(-\frac{x}{t}\right)\right\} = -\frac{1}{2t} + \frac{x^2}{2t^2}.$$

Since we have only one differential equation and two unknown functions, there are many possible solutions for which $X(t)$ is a diffusion process.

First, notice that we cannot have $m(x,t) = m(t)$ and $v(x,t) = v(t)$ at the same time, except when $m(x,t) \equiv 0$ and $v(x,t) \equiv 1$. That is, when $X(t)$ is a standard Brownian motion. Assume that $v(x,t) = v(t)$, but that $m(x,t)$ depends on x. We find that m satisfies the ordinary differential equation

$$m_x - \left(\frac{x}{t}\right)m + \frac{v(t)-1}{2t}\left(1 - \frac{x^2}{t}\right) = 0,$$

whose general solution is

$$m(x,t) = c_1 \exp\left\{\frac{x^2}{2t}\right\} - x\left(\frac{v(t)-1}{2t}\right).$$

Let us choose the constant $c_1 = 0$. We see that the process considered in the previous section corresponds to the infinitesimal variance $v(t) = 1 + kt$, with k a non-negative constant. Indeed, we then have $m(x,t) = -\frac{k}{2}x$. Note that the conditions in (6) are satisfied with this choice of infinitesimal parameters. There are however other interesting possibilities. For example, we could take $v(x,t) = v(t) = 1+kt^2$ and $m(x,t) = -\frac{k}{2}tx$. Furthermore, we can of course consider the case when $m(x,t) = m(t)$, but $v(x,t)$ depends on x, as well as the general case when both $m(x,t)$ and $v(x,t)$ depend on x (and t).

4. CONCLUSION

In this note, we first considered the diffusion process $\{X(t), t \geq 0\}$ whose infinitesimal parameters are $m(x,t) = -kx/2$ and $v(x,t) = 1 + kt$, where the constant k is non-negative. Although this process is time-inhomogeneous, we were able to calculate explicitly the probability density function of the random variable $X(t)$. We saw that $X(t)$ is normally distributed and that its parameters are related to those that correspond to the standard Brownian motion and the Ornstein-Uhlenbeck process.

The diffusion process $\{X(t), t \geq 0\}$ is a good compromise between the Wiener and Ornstein-Uhlenbeck processes, in that it behaves partly like these two very important diffusion processes. Moreover, if $\{X(t), t \geq 0\}$ starts from the origin at time $t_0 = 0$, then $X(t) \sim \mathrm{N}(0, t)$, exactly like a standard Brownian motion.

In Section 3, we saw that there are other time-inhomogeneous diffusion processes $\{X(t), t \geq 0\}$ for which $X(t)$ has a Gaussian $\mathrm{N}(0, t)$ distribution. This is true when $t_0 = 0$ and $X(0) = 0$. Making use of the proposition in Lefebvre (2007) mentioned above, we could calculate their probability density function in the general case when the initial time is $t_0 \geq 0$ and $X(t_0) = x_0 \in \mathbb{R}$.

As a sequel to this work, we could, in particular, try to find diffusion processes having a lognormal probability density function, like the geometric Brownian motion. This diffusion process is used extensively in financial mathematics. Moreover, it would be nice to have some real-life data for which the diffusion process $\{X(t), t \geq 0\}$ introduced in Section 1 would be a good model. Finally, we could also study first passage time problems involving $\{X(t), t \geq 0\}$.

Acknowledgment. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] D. R. Cox, H. D. Miller, *The theory of stochastic processes*, Methuen, London, 1965.

[2] D. Lamberton, B. Lapeyre, *Introduction au calcul stochastique appliqué à la finance*, 2nd ed., Ellipses, Paris, 1997.

[3] M. Lefebvre, *Applied stochastic processes*, Springer, New York, 2007.

ROMAI J., **4**, 1(2008), 137–168

ON CRYSTALLIZATION PROBLEM IN STATISTICAL CRYSTAL THEORY WITH SYMMETRIES OF COMPOSITE LATTICE

Boris V. Loginov, Oleg V. Makeev

Ulyanovsk State Technical University, Russia

loginov@ulstu.ru, o.makeev@ulstu.ru

Abstract Bifurcation theory methods under group symmetry conditions [6, 7] are applied to crystallization problem with composite lattices in statistical crystal theory. The obtained results are supported by RFBR-RA grant No. 07-01-91680.

Keywords: statistical crystal theory, bifurcation theory, group symmetry.

2000 MSC: 58E09.

Crystallization of liquid phase state in the case of composite lattice is described by the system of nonlinear integral equations with kernels depending on modulus of arguments difference [1], obtained by the uncoupling of N.N. Bogolyubov equations hierarchy on second distribution function. Suppose that forming crystal molecules belong to M different classes and take Bogolyubov equations hierarchy, that is the system of equations connecting simple $F_i(q)$ and binary $F_{ij}(q, q')$ densities of particles distribution

$$\frac{\partial F_i}{\partial q^\alpha} + \frac{1}{\theta v}\sum_{j=1}^{M} n_j \int \frac{\partial \Phi_{ij}(|q-q'|)}{\partial q^\alpha} F_{ij}(q,q')dq' = 0, \quad q = (q^1, q^2, q^3). \tag{1}$$

Here $\theta = kT$, k – Boltzman constant, T – temperature, $n_i = \frac{N_i}{N}$, $v = \frac{V}{N} \Rightarrow \frac{n_j}{v} = \frac{N_j}{V} = \frac{1}{v_j}$, $\Phi_{ij}(|q-q'|) = \Phi_{ji}(|q-q'|)$ – the potential energy of i-th and j-th molecule classes interaction which are disposed at the points q and q'.

Carrying out the approximation $F_{ij}(q,q') = F_i(q)F_j(q')G_{ij}(|q-q'|)$, $\lim\limits_{|q-q'|\to\infty} G_{ij}(|q-q'|) = 1$; $G_{ij}(|q-q'|) = 0$ at $|q-q'| \le a$, where $G_{ij}(|q-q'|)$ is radial density distribution of two types particles, transform (1) to the form

$$\frac{\partial F_i(q)}{\partial q^\alpha} + \frac{1}{\theta v}\sum_{j=1}^{M} n_j \frac{\partial U_{ij}(q)}{\partial q^\alpha} F_i(q) = 0,$$

$$U_{ij}(q) = \int\limits_{(q')} \Bigg\{ \int\limits_{\infty}^{|q-q'|} \frac{d\Phi_{ij}(r)}{dr} G_{ij}(r)dr \Bigg\} F_j(q')dq'.$$

Setting $F_i(q) = \frac{1}{\lambda_i}\exp\big[-\frac{1}{\theta v}\sum\limits_{j=1}^{M} n_j U_{ij}(q)\big]$, where $\frac{1}{\lambda_i}$ is a constant not depending on coordinates and defining from the condition of density normalization $\lim\limits_{V\to\infty}\frac{1}{V}\int\sum\limits_{i=1}^{M}F_i(q)dq = 1$, and $\rho_i(q) = \frac{1}{v_i}F_i(q) = \frac{1}{\lambda_i v_i}e^{u_i(q)}$ we obtain the system of nonlinear integral equations

$$\ln\{\lambda F_i(q)\} = u_i(q) = -\frac{1}{\theta}\sum_{j=1}^{M}\int\frac{1}{\lambda v_j}K_{ij}(|q-q'|)e^{u_j(q')}dq'.$$

$K_{ij}(|q-q'|) = K_{ji}(|q-q'|) = \int\limits_{\infty}^{|q-q'|}\frac{d\Phi_{ij}(r)}{dr}G_{ij}(r)dr$, $q = (q^1, q^2, q^3)$ – Cartesian coordinates. As far as $\frac{1}{v_i} = \frac{n_i}{v} = \frac{1}{Mv}$, the system of integral equations takes the form

$$u_i(q) + \frac{1}{Mv\theta\lambda}\sum_{j=1}^{M}\int K_{ij}(|q-q'|)e^{u_j(q')}dq' = 0. \tag{2}$$

As far as the composite lattice consists of M identical sublattices, here, like in [2], the common constant of normalization λ is introduced.

Further the general case of composite lattice will be illustrated by an example of crystallization with translation lattice consisting of four primitive sublattices Γ_m of monoclinic syngony with nonsymmorphic group C_{2h}^5 [3], which have nontrivial screw rotation and glide reflection.

First, give a brief introduction to crystallographic groups. It is know [3] that all symmetry groups of 3-dimensional homogeneous discrete space – spatial crystallographic groups – are three times periodic. Translations group $T = \{\mathbf{a} = m_1\mathbf{a}_1 + m_2\mathbf{a}_2 + m_3\mathbf{a}_3\}$, $m_i \in \mathbb{Z}$ propagates any point into $3D$-periodic system of points, that is spatial lattice. Crystalline lattices are divided into 7 crystalline systems, that are called syngonies. Bravais mathematically showed that for the crystals of 7 syngonies 14 types of lattices are possible. Besides of translational symmetry crystallographic groups are characterized by point symmetry K – the aggregate of rotation and reflection operations being the symmetry of elementary cell. There are 32 point groups called crystalline classes which are compatible with the translation group. However, space crystallographic groups have new elements of symmetry that are absent in translation and point groups: screw displacements and glide reflections. Screw displacement is a translation with subsequent rotation on some angle around translation axis. Glide reflection is a reflection in some plane with subsequent translation along this plane. Both indicated symmetry elements are formed by commuting elements, these elements themselves can be absent in crystallographic group.

To the considered crystallographic group C_{2h}^5 it corresponds the rotation-reflection point group symmetry – the eldest crystalline class of monoclinic syngony $C_{2h} = \{e, r, \sigma_h, \sigma_h r\}$, where r is the rotation of angle π around Oz, σ_h is the reflection in the xOy plane.

Here the basic translation vectors should be chosen in the form $\mathbf{a}_1 = \alpha\mathbf{i}$, $\mathbf{a}_2 = \beta\mathbf{i} + \gamma\mathbf{j}$, $\mathbf{a}_3 = \delta\mathbf{k}$, and the inverse lattice vectors take the form

$$\mathbf{l}^{(1)} = \frac{[\mathbf{a}_2, \mathbf{a}_3]}{\Omega} = \frac{1}{\alpha}\mathbf{i} - \frac{\beta}{\alpha\gamma}\mathbf{j}; \quad \mathbf{l}^{(2)} = \frac{[\mathbf{a}_3, \mathbf{a}_1]}{\Omega} = \frac{1}{\gamma}\mathbf{j}; \quad \mathbf{l}^{(3)} = \frac{[\mathbf{a}_1, \mathbf{a}_2]}{\Omega} = \frac{1}{\delta}\mathbf{k},$$

where $\Omega = \langle \mathbf{a}_1, [\mathbf{a}_2, \mathbf{a}_3] \rangle = \alpha\gamma\delta$.

Elements of the nonsymmorphic group $G = C_{2h}^5$ take the form (ξ, O), $\xi \in T$, $O \in C_{2h}$, the product is defined by the formula

$$(\xi_1, O_1) \cdot (\xi_2, O_2) = (\xi_1 + O_1\xi_2, O_1 O_2).$$

Moreover, the elements of point group C_{2h} are entered into G only accompanied by the translations $(\frac{1}{2}t_z, r)$, $(\frac{1}{2}t_x, \sigma_h)$, $(\frac{1}{2}t_x + \frac{1}{2}t_z, \sigma_h r)$.

Remark 1. Give the geometric interpretation of composite lattice. There are identical particles in the lattice nodes possessing the colour symmetry of the point group K. For nonsymmorphic group C_{2h}^5 of monoclinic syngony such particles may be interpreted like a ball, divided into 4 parts by two mutual perpendicular planes passing through the ball center, each part of the ball is colored into white or black. The translation group T propagates such particles into space-periodic systems that are sublattices of the considered crystal. The transformations of nonsymmorphic group transfers one sublattice into another, which is shifted by some translation $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, $\alpha_i \in (0; 1)$ (see the table of nonsymmorphic crystallographic groups in the appendix to the monograph [3]). Moreover, every particle in the new sublattice is turned by corresponding transformation of the point group K .

At the crystallization phenomenon investigation, the problem of periodic solutions construction, $u_i(q) = u_{0i} + w_i(q, \varepsilon)$, $w_j(q, \varepsilon) = \sum_{\mathbf{l}} w_{\mathbf{l}} e^{2\pi i \langle \mathbf{l}_j, q \rangle}$ ($\mathbf{l}_j = m_j^{(1)}\mathbf{l}^{(1)} + m_j^{(2)}\mathbf{l}^{(2)} + m_j^{(3)}\mathbf{l}^{(3)}$ is the inverse lattice vector), in the form of

Fourier series on inverse lattice vectors, bifurcating from homogeneous densities distribution $\rho_i(q) = \rho_{0i} = \frac{1}{v_{0i}}$, $v_{0j} = Mv_0$, naturally arises. Since sublattices consist of identical but having different orientation particles and have the same translation group we can take $\rho_{i0} = \rho_0$ and $u_{i0} = u_0$, and the small parameter ε should be determined by the relation $\frac{\exp u_0}{Mv\theta\lambda} = \frac{\exp u_0}{Mv_0\theta_0\lambda_0} + \varepsilon = \mu_0 + \varepsilon$.

The considered system of nonlinear integral equations (2) with respect to vector-functions $w = \{w_i(q,\varepsilon)\}_1^M$ takes the form

$$B_s w_s \equiv w_s(q) + \mu_0 \sum_{j=1}^{M} \int K_{sj}(|q-q'|) w_j(q') dq' =$$
$$- \varepsilon \sum_{j=1}^{M} \int K_{sj}(|q-q'|) e^{w_j(q')} dq' - \mu_0 \sum_{j=1}^{M} \int K_{sj}(|q-q'|)[e^{w_j(q')} -$$
$$- w_j(q') - 1] dq' \equiv R_s(w,\varepsilon). \quad (3)$$

Remark 2. By virtue of Remark 1, system (2) possesses the group symmetry of nonsymmorphic crystallographic group corresponding to the composite lattice consisting of M sublattices of one type molecules oriented by point group $K = C_{2h}$ ($|K| = M$, $|C_{2h}| = 4$) transformations. Nonsymmorphic group transformations transfer equations of the system (2) into each other, leaving invariant the entire system. The connection between sublattices of composite lattice and equations are realized by screw rotation and glide reflection.

The system of integral equations (3) is considered in the space of vector-functions continuously differentiable on the elementary periodicity cell, and kernels $K_{sj}(|q-q'|)$ are sufficiently smooth, so $\int K_{sj}(|q-q'|)u_j(q')dq'$, $q \in \Pi_0$ can be differentiated with respect to q.

Describe the zero-subspace of linearized system (3)

$$B_s w_s \equiv w_s(q) + \mu_0 \sum_{j=1}^{M} \int K_{sj}(|q-q'|) w_j(q') dq' = 0, \quad s = 1,\dots,M. \quad (4)$$

By presenting the components of vector-function w by Fourier series on inverse lattice vectors (index k is numbering the three-tuples of integers $(m_j^{(1)}, m_j^{(2)}, m_j^{(3)})$)

$$w_j(q) = \sum_k w_{kj} e^{2\pi i \langle \mathbf{l}_{kj}, q\rangle}, \quad \mathbf{l}_{kj} = m_{kj}^{(1)} \mathbf{l}^{(1)} + m_{kj}^{(2)} \mathbf{l}^{(2)} + m_{kj}^{(3)} \mathbf{l}^{(3)}, \quad m_{kj}^{(p)} \in \mathbb{Z}$$

one gets the equation

$$B_s w_s = \sum_k w_{ks} e^{2\pi i \langle \mathbf{l}_{ks}, q\rangle} + \mu_0 \sum_{j=1}^{M} \int K_{sj}(|q-q'|) \sum_k w_{kj} e^{2\pi i \langle \mathbf{l}_{kj}, q'\rangle} dq' =$$

$$\sum_k w_{ks} e^{2\pi i \langle \mathbf{l}_{ks}, q\rangle} + \mu_0 \sum_{j=1}^{M} \int K_{sj}(|q-q'|) \sum_k w_{kj} e^{2\pi i \langle \mathbf{l}_{kj}, q\rangle} e^{-2\pi i \langle \mathbf{l}_{kj}, q-q'\rangle} dq' =$$

$$\sum_k w_{ks} e^{2\pi i \langle \mathbf{l}_{ks}, q\rangle} + \mu_0 \sum_k \sum_{j=1}^{M} w_{kj} e^{2\pi i \langle \mathbf{l}_{kj}, q\rangle} \int K_{sj}(|q-q'|) e^{-2\pi i \langle \mathbf{l}_{kj}, q-q'\rangle} dq'. \quad (5)$$

Compute the integrals $I_{ts} = \int K_{ts}(|q-q'|) e^{-2\pi i \langle \mathbf{l}_{ks}, q-q'\rangle} dq'$. By setting $\tilde{q} = q - q' = x\mathbf{e}_1 + y\mathbf{e}_2 + z\mathbf{e}_3 = \tilde{x}\mathbf{a}_1 + \tilde{y}\mathbf{a}_2 + \tilde{z}\mathbf{a}_3$; $\mathbf{a}_j = a_{1j}\mathbf{e}_1 + a_{2j}\mathbf{e}_2 + a_{3j}\mathbf{e}_3$

$$\begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \mathbf{a}_3 \end{pmatrix} = A^T \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{pmatrix} \Rightarrow \begin{pmatrix} x \\ y \\ z \end{pmatrix} = A \begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix},$$

using the change of variables $x = \tilde{x}a_{11} + \tilde{y}a_{12} + \tilde{z}a_{13}$, $y = \tilde{x}a_{21} + \tilde{y}a_{22} + \tilde{z}a_{23}$, $z = \tilde{x}a_{31} + \tilde{y}a_{32} + \tilde{z}a_{33}$, and then carrying out the transition to spherical coordinates one gets

$$I_{ts} = \int_0^\infty \rho^2 K(\rho) 2\pi \int_{-1}^{1} \exp\left[-\frac{2\pi i \rho}{det\, A} R_{ks} t\right] dt = \frac{2 det\, A}{R_{ks}} \int_0^\infty \rho K(\rho) \sin\left(\frac{2\pi\rho R_{ks}}{det\, A}\right) d\rho.$$

Omitting tedious computations, write the expression for R_{ks}

$$\begin{aligned}
R_{ks} = R(m_{ks}^{(1)}, m_{ks}^{(2)}, m_{ks}^{(3)}) = \\
\Big\{ m_{ks}^{(1)^2} [(a_{21}^2 + a_{22}^2 + a_{23}^2)(a_{31}^2 + a_{32}^2 + a_{33}^2) - (a_{21}a_{31} + a_{22}a_{32} + a_{23}a_{33})^2] + \\
m_{ks}^{(2)^2} \left[(a_{11}^2 + a_{12}^2 + a_{13}^2)(a_{31}^2 + a_{32}^2 + a_{33}^2) - (a_{11}a_{31} + a_{12}a_{32} + a_{13}a_{33})^2\right] + \\
m_{ks}^{(3)^2} \left[(a_{21}^2 + a_{22}^2 + a_{23}^2)(a_{11}^2 + a_{12}^2 + a_{13}^2) - (a_{11}a_{21} + a_{12}a_{22} + a_{13}a_{23})^2\right] + \\
+ 2m_{ks}^{(1)} m_{ks}^{(2)} [(a_{11}a_{31} + a_{12}a_{32} + a_{13}a_{33})(a_{21}a_{31} + a_{22}a_{32} + a_{23}a_{33}) - \\
- (a_{11}a_{21} + a_{12}a_{22} + a_{13}a_{23})(a_{31}^2 + a_{32}^2 + a_{33}^2)] + \\
+ 2m_{ks}^{(1)} m_{ks}^{(3)} [(a_{11}a_{21} + a_{12}a_{22} + a_{13}a_{23})(a_{21}a_{31} + a_{22}a_{32} + a_{23}a_{33}) - \\
- (a_{11}a_{31} + a_{12}a_{32} + a_{13}a_{33})(a_{21}^2 + a_{22}^2 + a_{23}^2)] + \\
+ 2m_{ks}^{(2)} m_{ks}^{(3)} [(a_{11}a_{21} + a_{12}a_{22} + a_{13}a_{23})(a_{11}a_{31} + a_{12}a_{32} + a_{13}a_{33}) - \\
(a_{21}a_{31} + a_{22}a_{32} + a_{23}a_{33})(a_{11}^2 + a_{12}^2 + a_{13}^2)] \Big\}^{\frac{1}{2}}. \quad (6)
\end{aligned}$$

The linearized system (5) takes the form

$$\sum_k w_{ks} e^{2\pi i (m_{ks}^{(1)} x + m_{ks}^{(2)} y + m_{ks}^{(3)} z)} -$$

$$\mu_0 \sum_k \sum_{j=1}^{M} w_{kj} e^{2\pi i (m_{kj}^{(1)} x + m_{kj}^{(2)} y + m_{kj}^{(3)} z)} \frac{2 det\, A}{R_{kj}} \int_0^\infty \rho K_{sj}(\rho) \sin\left(\frac{2\pi\rho R_{kj}}{det\, A}\right) d\rho = 0 \quad (7)$$

with determinant

$$\Delta_k = \left[I\delta_{sj} - \mu_0 \frac{2det\, A}{R_{kj}(m_{kj})} \int_0^\infty \rho K_{sj}(\rho) \sin\left(\frac{2\pi\rho R_{kj}(m_{kj})}{det\, A} \right) d\rho \right], \; s, j = \overline{1, M} \tag{8}$$

Thus, conversion to zero of the determinant (8) defines eigenvalues μ_0 and presents the crystallization criterion [4] with corresponding composite lattice.

Kernels $K_{sj}(|q-q'|) = K_{js}(|q-q'|)$ are invariant with respect to the group of Euclidean space motion $\mathbb{R}^3$ including also transformations g of nonsymmorphic crystallographic groups G $K_{sj}(|gq - gq'|) = K_{sj}(|q - q'|)$. The corresponding operators $K_{sj} f(q) = \int K_{sj}(|q - q'|) f(q') dq'$ are invariant with respect to shift operators by virtue of kernels K_{sj} invariance relative to simultaneous motions in $\mathbb{R}^3$ of arguments q and q'. Indeed, since dq' is invariant relative to G measure in $\mathbb{R}^3$, one has

$$(K_{sj} f)(gq) = \int K_{sj}(|gq - q'|) f(q') dq' \overset{q' \equiv g\bar{q}}{=} \int K_{sj}(|q - \bar{q}|) f(g\bar{q}) d\bar{q} = K_{sj} T(g) f(q).$$

Therefore applying the nonsymmorphic transformation $T(g)$ to the s-th equation (4) one gets

$$T(g) w_s(q) + \mu_0 \sum_{j=1}^{M} \int K_{sj}(|q - \bar{q}|) w_j(g\bar{q}) d\bar{q} =$$
$$= T(g) w_s(q) + \mu_0 \sum_{j=1}^{M} \int K_{sj}(|q - \bar{q}|) T(g) w_j(\bar{q}) d\bar{q} = 0.$$

Hence, together with some solution $w = (w_1, \dots, w_M)^T$ of the linearized system (4) it has the solutions $T(g)w = (T(g)w_1, \dots, T(g)w_M)^T$, so system (4) is invariant relative to the transformations $T(g)$.

The connection between sublattices of composite lattice and equations is realized by transformations of nonsymmorphic group

$$(\frac{1}{2} t_z, r) \cong (1,4)(2,3), \quad (\frac{1}{2} t_x, \sigma_h) \cong (1,3)(2,4), \quad (\frac{1}{2} t_x + \frac{1}{2} t_z, \sigma_h r) \cong (1,2)(3,4). \tag{9}$$

For the equations changing into each other, it is necessary the fulfilment of the following symmetry relations between the kernels of the integral operators

$$\begin{aligned} &K_{11} = \ldots = K_{44}; && K_{14} = K_{23} = K_{32} = K_{41}; \\ &K_{12} = K_{21} = K_{34} = K_{43}; && K_{13} = K_{31} = K_{24} = K_{42}. \end{aligned} \tag{10}$$

Choose the basis vectors of the null subspace $N(\mathbf{B})$ in the form

$$\Phi_1 = \begin{pmatrix} \varphi_1 & = & e^{2\pi i(x+y+z)} \\ \varphi_2 & = & e^{2\pi i(-x-y-z)} \\ \varphi_3 & = & -e^{2\pi i(x+y-z)} \\ \varphi_4 & = & -e^{2\pi i(-x-y+z)} \end{pmatrix}, \quad \Phi_2 = (\tfrac{1}{2}t_x + \tfrac{1}{2}t_z, \sigma_h r)\Phi_1 = \begin{pmatrix} \varphi_2 \\ \varphi_1 \\ \varphi_4 \\ \varphi_3 \end{pmatrix},$$

$$\Phi_3 = (\tfrac{1}{2}t_x, \sigma_h)\Phi_1 = \begin{pmatrix} \varphi_3 \\ \varphi_4 \\ \varphi_1 \\ \varphi_2 \end{pmatrix}, \qquad \Phi_4 = (\tfrac{1}{2}t_z, r)\Phi_1 = \begin{pmatrix} \varphi_4 \\ \varphi_3 \\ \varphi_2 \\ \varphi_1 \end{pmatrix}.$$

For the simplification of designations hereinafter we omit the symbol “tilde”, i.e. x, y, z are considered as coordinates along the axes $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ respectively.

For the computation of bifurcating solutions in neighborhoods of parameter critical value, bifurcation theory methods [5] are applied.

Let E_1 and E_2 be Banach spaces. The nonlinear equation

$$Bx = R(x,\lambda), \quad R(0,0) = 0,\ R_x(0,0) = 0 \tag{11}$$

is considered. Here $B : E_1 \to E_2$ is a closed linear Fredholm operator ($R(B) = \overline{R(B)}$, $R(B)$ is the range of the operator B) with dense in E_1 domain $D(B)$, $N(B) = \operatorname{span}\{\Phi_1, \dots, \Phi_n\}$ is its null-subspace, $N^*(B) = \operatorname{span}\{\Psi_1, \dots, \Psi_n\} \subset E_2^*$ is its defect-subspace. The nonlinear operator $R(x,\lambda)$ is supposed to be defined and sufficiently smooth in x and λ in a neighborhood of $(0,0) \in E_1 \dot{+} \Lambda$, Λ is the parameter space. According to Hahn-Banach theorem there exist biorthogonal systems $\{\Gamma_j\}_1^n \in E_1$, $\langle \Phi_i, \Gamma_j \rangle = \delta_{ij}$ and $\{Z_k\}_1^n \in E_2$, $\langle Z_k, \Psi_l \rangle = \delta_{kl}$, generating the projectors $P = \sum\limits_{j=1}^{n} \langle \cdot, \Gamma_j \rangle \varphi_j : E_1 \to N(B)$, $Q = \sum\limits_{j=1}^{n} \langle \cdot, \Psi_j \rangle z_j : E_2 \to E_{2,n} = \operatorname{span}\{z_1, \dots, z_n\}$ and the following direct sum expansions $E_1 = E_1^n \dot{+} E_1^{\infty-n}$, $E_1^n = N(B)$, $E_2 = E_{2,n} \dot{+} E_{2,\infty-n}$, $E_{2,\infty-n} = R(B)$. Then the Lyapounov-Schmidt method allows to reduce the problem (11) of construction of small norm solutions to nonlinear finite-dimensional equations system, that is the bifurcation equation.

Here $Z_s = \Phi_s$,

$$\Gamma_1 = \Psi_1 = \frac{1}{4|\Pi_0|}\Phi_2,\ \Gamma_2 = \Psi_2 = \frac{1}{4|\Pi_0|}\Phi_1,\ \Gamma_3 = \Psi_3 = \frac{1}{4|\Pi_0|}\Phi_4,\ \Gamma_4 = \Psi_4 = \frac{1}{4|\Pi_0|}\Phi_3.$$

Write system (3) in the power series expansion introducing parameters ξ_k, $k=1,\ldots,4$ and Schmidt correction operator

$$\begin{aligned}
\tilde{B}W &= \begin{pmatrix} w_1(q) \\ \cdots \\ w_4(q) \end{pmatrix} + \mu_0 \mathcal{K} W + \sum_{j=1}^{4} \langle W, \Gamma_j \rangle Z_j = -\varepsilon \begin{pmatrix} \sum_{j=1}^4 K_{1j}(|q-q'|)dq' \\ \cdots \\ \sum_{j=1}^4 K_{4j}(|q-q'|)dq' \end{pmatrix} \qquad (11) \\
&-\varepsilon \mathcal{K} \begin{pmatrix} w_1(q') + \frac{w_1(q')^2}{2!} + \ldots \\ \cdots \\ w_4(q') + \frac{w_4(q')^2}{2!} + \ldots \end{pmatrix} - \mu_0 \mathcal{K} \begin{pmatrix} \frac{w_1(q')^2}{2!} + \frac{w_1(q')^3}{3!} + \ldots \\ \cdots \\ \frac{w_4(q')^2}{2!} + \frac{w_4(q')^3}{3!} + \ldots \end{pmatrix} + \sum_{j=1}^{4} \xi_j Z_j, \\
\xi_j &= \langle W, \Gamma_j \rangle,
\end{aligned}$$

$$\mathcal{K} = \begin{pmatrix} \int K_{11}(q')dq' & \ldots & \int K_{14}(q')dq' \\ \cdots & \cdots & \cdots \\ \int K_{41}(q')dq' & \ldots & \int K_{44}(q')dq' \end{pmatrix}.$$

By the implicit operators theorem the first equation (12) has a unique solution $W = W(\xi, \varepsilon)$.

Branching system takes the form $L^{(i)}(\xi,\varepsilon) \equiv \xi_i - \langle W(\xi,\varepsilon), \Gamma_i \rangle = 0, \quad i=1,\ldots,4.$

We find the solutions of the first equation (12) in the form of the series $W(q,\varepsilon) = \sum\limits_{|\alpha|+k\geq 1} W_{\alpha;k} \xi^\alpha \varepsilon^k$.

Omitting tedious computations we write out the main part of the branching system

$$\begin{aligned}
f_1(\xi,\varepsilon) &= A\xi_1\varepsilon + B{\xi_2}^3 + C{\xi_1}^2\xi_2 + D\xi_1\xi_3\xi_4 + E\xi_2{\xi_3}^2 + F\xi_2{\xi_4}^2 + \ldots = 0, \qquad (12) \\
f_2(\xi,\varepsilon) &= A\xi_2\varepsilon + B{\xi_1}^3 + C{\xi_2}^2\xi_1 + D\xi_2\xi_3\xi_4 + E\xi_1{\xi_4}^2 + F\xi_1{\xi_3}^2 + \ldots = 0, \\
f_3(\xi,\varepsilon) &= A\xi_3\varepsilon + B{\xi_4}^3 + C{\xi_3}^2\xi_4 + D\xi_1\xi_2\xi_3 + E\xi_4{\xi_1}^2 + F\xi_4{\xi_2}^2 + \ldots = 0, \\
f_4(\xi,\varepsilon) &= A\xi_4\varepsilon + B{\xi_3}^3 + C{\xi_4}^2\xi_3 + D\xi_1\xi_2\xi_4 + E\xi_3{\xi_2}^2 + F\xi_3{\xi_1}^2 + \ldots = 0.
\end{aligned}$$

The obtained system admits the group (9) of substitutions $p_1 = (12)(34)$, $p_2 = (13)(24)$, $p_3 = (13)(23)$.

Passing to real variables $\xi_1 = \tau_1 + i\tau_2$, $\xi_2 = \tau_1 - i\tau_2$, $\xi_3 = \tau_3 + i\tau_4$, $\xi_4 = \tau_3 - i\tau_4$ one gets branching system in the new basis

$$\widehat{\Phi}_1 = \begin{pmatrix} \cos 2\pi(x+y+z) \\ \cos 2\pi(x+y+z) \\ -\cos 2\pi(x+y-z) \\ -\cos 2\pi(x+y-z) \end{pmatrix}, \quad \widehat{\Phi}_2 = \begin{pmatrix} \sin 2\pi(x+y+z) \\ -\sin 2\pi(x+y+z) \\ -\sin 2\pi(x+y-z) \\ \sin 2\pi(x+y-z) \end{pmatrix},$$

$$\widehat{\Phi}_3 = \begin{pmatrix} -\cos 2\pi(x+y-z) \\ -\cos 2\pi(x+y-z) \\ \cos 2\pi(x+y+z) \\ \cos 2\pi(x+y+z) \end{pmatrix}, \quad \widehat{\Phi}_4 = \begin{pmatrix} -\sin 2\pi(x+y-z) \\ \sin 2\pi(x+y-z) \\ \sin 2\pi(x+y+z) \\ -\sin 2\pi(x+y+z) \end{pmatrix},$$

$$\begin{aligned}
t_1(\tau,\varepsilon) &= A\tau_1\varepsilon + B\tau_1(\tau_1^2-3\tau_2^2) + C\tau_1(\tau_1^2+\tau_2^2) + D\tau_1(\tau_3^2+\tau_4^2) + \\
&\quad + E[\tau_1(\tau_3^2-\tau_4^2) + 2\tau_2\tau_3\tau_4] + F[\tau_1(\tau_3^2-\tau_4^2) - 2\tau_2\tau_3\tau_4] + \ldots = 0, \\
t_2(\tau,\varepsilon) &= A\tau_2\varepsilon + B\tau_2(\tau_2^2-3\tau_1^2) + C\tau_2(\tau_1^2+\tau_2^2) + D\tau_2(\tau_3^2+\tau_4^2) + \\
&\quad + E[-\tau_2(\tau_3^2-\tau_4^2) + 2\tau_1\tau_3\tau_4] - F[\tau_2(\tau_3^2-\tau_4^2) + 2\tau_1\tau_2\tau_4] + \ldots = 0, \\
t_3(\tau,\varepsilon) &= A\tau_3\varepsilon + B\tau_3(\tau_3^2-3\tau_2^2) + C\tau_3(\tau_3^2+\tau_4^2) + D\tau_3(\tau_1^2+\tau_2^2) + \\
&\quad + E[\tau_3(\tau_1^2-\tau_2^2) + 2\tau_1\tau_2\tau_4] + F[\tau_3(\tau_1^2-\tau_2^2) - 2\tau_1\tau_2\tau_4] + \ldots = 0, \\
t_4(\tau,\varepsilon) &= A\tau_4\varepsilon + B\tau_4(\tau_4^2-3\tau_3^2) + C\tau_4(\tau_3^2+\tau_4^2) + D\tau_4(\tau_1^2+\tau_2^2) + \\
&\quad + E[-\tau_4(\tau_1^2-\tau_2^2) + 2\tau_1\tau_2\tau_3] - F[\tau_4(\tau_1^2-\tau_2^2) + 2\tau_1\tau_2\tau_3] + \ldots = 0.
\end{aligned} \tag{13}$$

By applying the transformations

$$\begin{aligned}
t_1\tau_2 - t_2\tau_1 &= 2B\tau_1\tau_2(\tau_1^2-\tau_2^2) + E[\tau_1\tau_2(\tau_3^2-\tau_4^2) - \tau_3\tau_4(\tau_1^2-\tau_2^2)] + \\
&\quad + F[\tau_1\tau_2(\tau_3^2-\tau_4^2) + \tau_3\tau_4(\tau_1^2-\tau_2^2)] + \ldots = 0, \\
t_1\tau_2 + t_2\tau_1 &= A\tau_1\tau_2\varepsilon + (C-2B)\tau_1\tau_2(\tau_1^2+\tau_2^2) + D\tau_1\tau_2(\tau_3^2+\tau_4^2) + \\
&\quad + (E-F)\tau_3\tau_4(\tau_1^2+\tau_2^2) + \ldots = 0, \\
t_3\tau_4 - t_4\tau_3 &= 2B\tau_3\tau_4(\tau_3^2-\tau_4^2) + E[\tau_3\tau_4(\tau_1^2-\tau_2^2) - \tau_1\tau_2(\tau_3^2-\tau_4^2)] + \\
&\quad + F[\tau_3\tau_4(\tau_1^2-\tau_2^2) + \tau_1\tau_2(\tau_3^2-\tau_4^2)] + \ldots = 0, \\
t_3\tau_4 + t_4\tau_3 &= A\tau_3\tau_4\varepsilon + (C-2B)\tau_3\tau_4(\tau_3^2+\tau_4^2) + D\tau_3\tau_4(\tau_1^2+\tau_2^2) + \\
&\quad + (E-F)\tau_1\tau_2(\tau_3^2+\tau_4^2) + \ldots = 0,
\end{aligned} \tag{14}$$

by adding and subtracting first and third, second and fourth equation of the system (14), we bring branching system to the form

$$\begin{aligned}
\tilde{t}_1(\tau,\varepsilon) &= B[\tau_1\tau_2(\tau_1^2-\tau_2^2) + \tau_3\tau_4(\tau_3^2-\tau_4^2)] + \\
&\quad + F[\tau_1\tau_2(\tau_3^2-\tau_4^2) + \tau_3\tau_4(\tau_1^2-\tau_2^2)] + \ldots = 0, \\
\tilde{t}_2(\tau,\varepsilon) &= B[\tau_1\tau_2(\tau_1^2-\tau_2^2) - \tau_3\tau_4(\tau_3^2-\tau_4^2)] + \\
&\quad + E[\tau_1\tau_2(\tau_3^2-\tau_4^2) - \tau_3\tau_4(\tau_1^2-\tau_2^2)] + \ldots = 0, \\
\tilde{t}_3(\tau,\varepsilon) &= A\varepsilon(\tau_1\tau_2+\tau_3\tau_4) + (C-B)[\tau_1\tau_2(\tau_1^2+\tau_2^2) + \tau_3\tau_4(\tau_3^2+\tau_4^2)] + \\
&\quad + (D+E-F)[\tau_1\tau_2(\tau_3^2+\tau_4^2) + \tau_3\tau_4(\tau_1^2) + \tau_2^2] + \ldots = 0, \\
\tilde{t}_4(\tau,\varepsilon) &= A\varepsilon(\tau_1\tau_2-\tau_3\tau_4) + (C-B)[\tau_1\tau_2(\tau_1^2+\tau_2^2) - \tau_3\tau_4(\tau_3^2+\tau_4^2)] + \\
&\quad + (D-E+F)[\tau_1\tau_2(\tau_3^2+\tau_4^2) - \tau_3\tau_4(\tau_1^2) + \tau_2^2] + \ldots = 0.
\end{aligned} \tag{15}$$

Note, that the written system in real basis allows substitutions

$$\begin{gathered}
p_1: \tau_1 \to \tau_1,\ \tau_2 \to -\tau_2,\ \tau_3 \to \tau_3,\ \tau_4 \to -\tau_4; \\
p_2: \tau_1 \leftrightarrow \tau_3,\ \tau_2 \leftrightarrow \tau_4; \quad p_3: \tau_1 \leftrightarrow \tau_3,\ \tau_2 \leftrightarrow -\tau_4.
\end{gathered} \tag{17}$$

The first two equations (16) are considered as the system relative to $(\tau_1^2-\tau_2^2)$, $(\tau_3^2-\tau_4^2)$ with determinant $\Delta = 2[\tau_1\tau_2\tau_3\tau_4(EF-B^2)+B(E-F)(\tau_1^2\tau_2^2+\tau_3^2\tau_4^2)]$.

I. If $\Delta \neq 0$, it is possible for $(B^2 - EF)^2 - 4B^2(E-F)^2 < 0$ $(|B^2 - EF| < 2|B(E-F)|)$, then $\tau_1^2 = \tau_2^2 \neq 0$ and $\tau_3^2 = \tau_4^2 \neq 0$ (inequality to zero must take place at least in one case). Then the last two equations of the system (16) are brought to one of the following forms

A. $\tau_1 = \tau_2$, $\tau_3 = \tau_4$. For $\tau_1 \neq 0$, $\tau_3 \neq 0$, one has

$$A\varepsilon + 2(C-B)\tau_1^2 + 2(D+E-F)\tau_3^2 = 0,$$
$$A\varepsilon + 2(C-B)\tau_3^2 + 2(D+E-F)\tau_1^2 = 0,$$

$\tau_1 = \pm\tau_3 = \pm\sqrt{\frac{-A\varepsilon}{2(-B+C-D-E+F)}} + o(|\varepsilon|^{1/2})$, $sign\,\varepsilon = -sign\,A(-B+C-D-E+F)$, any sign combinations are possible.

For $\tau_1 \neq 0$, $\tau_3 = 0$, one gets the equation $A\tau_1^2\varepsilon - 2B\tau_1^4 = 0$,

$$\tau_1 = \pm\sqrt{\frac{A\varepsilon}{2B}} + o(|\varepsilon|^{1/2}), \quad sign\,\varepsilon = sign\,(AB).$$

For $\tau_1 = 0$, $\tau_3 \neq 0$, one gets $A\tau_3^2\varepsilon - 2B\tau_3^4 = 0$,

$$\tau_3 = \pm\sqrt{\frac{A\varepsilon}{2B}} + o(|\varepsilon|^{1/2}), \quad sign\,\varepsilon = sign\,(AB).$$

In the case **B.** $\tau_1 = -\tau_2$, $\tau_3 = -\tau_4$, we obtain to the same equations.

C. $\tau_1 = -\tau_2$, $\tau_3 = \tau_4$. For $\tau_1 \neq 0$, $\tau_3 \neq 0$, one has

$$A\varepsilon + 2(-B+C)\tau_1^2 + 2(D-E+F)\tau_3^2 = 0,$$
$$-A\varepsilon + 2(B-C)\tau_3^2 + 2(-D+E-F)\tau_1^2 = 0,$$

$$\tau_1 = \pm\sqrt{\frac{-A\varepsilon}{2(-B+C+D-E+F)}} + o(|\varepsilon|^{1/2}),$$
$$sign\,\varepsilon = -sign\,A(-B+C+D-E+F);$$
$$\tau_3 = \pm\sqrt{\frac{A\varepsilon}{2(B-C+D-E+F)}} + o(|\varepsilon|^{1/2}),$$
$$sign\,\varepsilon = sign\,A(B-C+D-E+F).$$

For $\tau_1 \neq 0$, $\tau_3 = 0$, one gets the equation $A\varepsilon + 2(-B+C)\tau_1^2 = 0$,

$$\tau_1 = \pm\sqrt{\frac{A\varepsilon}{2(B-C)}} + o(|\varepsilon|^{1/2}), \quad sign\,\varepsilon = sign\,A(B-C).$$

For $\tau_1 = 0$, $\tau_3 \neq 0$, one gets $-A\varepsilon + 2(B-C)\tau_3^2 = 0$,

$$\tau_3 = \pm\sqrt{\frac{A\varepsilon}{2(B-C)}} + o(|\varepsilon|^{1/2}), \quad sign\,\varepsilon = sign\,A(B-C).$$

In the case **D.** $\tau_1 = -\tau_2$, $\tau_3 = \tau_4$ we obtain the same equations.

II. Let $\Delta = 0$, it is possible for $(B^2 - EF)^2 \geq 4B^2(E-F)^2$ $(|B^2 - EF| \geq 2|B(E-F)|)$, then if $E \neq F$

$$\tau_3\tau_4 = \frac{B^2 - EF \pm \sqrt{(B^2-EF)^2 - 4B^2(E-F)^2}}{2B(E-F)}\tau_1\tau_2 = k\tau_1\tau_2.$$

Then the last two equations of the system (16) can be written in the form

$$(\tau_1^2 + \tau_2^2)[(C-B) + k(E-F)] + (\tau_3^2 + \tau_4^2)D = -A\varepsilon,$$
$$(\tau_1^2 + \tau_2^2)k + (\tau_3^2 + \tau_4^2)[(C-B) + k(E-F)] = -Ak\varepsilon,$$

whence

$$\tau_1^2 + \tau_2^2 = \frac{-A\varepsilon[(C-B) + k(E-F) - D]}{[(C-B) + k(E-F)]^2 - kD}, \tau_3^2 + \tau_4^2 = \frac{-A\varepsilon[(C-B) + k(E-F) - 1]}{[(C-B) + k(E-F)]^2 - kD}.$$

Thus, all obtained solutions are presented in the form of converging in the small neigbourhood of $\varepsilon = 0$ series of the form $W = \sum \tau_k^0(\varepsilon^{1/2})\widehat{\Phi}_k + O(|\varepsilon|)$, where $\tau_k^0(\varepsilon^{1/2})$ are the leading terms of asymptotics of obtained solutions. Taking into account group transformations their number can be decreased.

References

[1] N. N. Bogolyubov, *Dynamic theory problems in statistical physics*, Gostekhizdat, Moscow, 1946. (Russian)

[2] B. N. Rolov, A. V. Ivin, V. N. Kuzovkov, *Statistics and kinetics of phase transitions*, Riga, 1979.

[3] G. Ya Lyubarskii, *Group theory and its applications in physics,* GITTL., Moscow, 1958. (Russian)

[4] A. A. Vlasov, *Many particles theory*, Gostekhizdat, Moscow, 1950. (Russian)

[5] M. M. Vainberg, V. A. Trenogin, *Branching theory of solutions of nonlinear equations,* Nauka, Moscow, 1969; Engl. transl. Wolters Noordorf, Leyden, 1974.

[6] B. V. Loginov, *Branching theory of solutions of nonlinear equations under group invariance conditions*, Fan, Tashkent, 1985.

[7] N. Sidorov, B. Loginov, A. Sinitsyn, M. Falaleev, *Lyapounov-Schmidt methods in nonlinear analysis and applications*, Kluwer, MIA, Dordrecht, **550, 2002.**

ROMAI J., **4**, 1(2008), 149–168

PSEUDOPERTURBITERATION METHODS IN GENERALIZED EIGENVALUE PROBLEMS

Boris V. Loginov, Olga V. Makeeva

Ulyanovsk State Technical University, Russia,

Technological Inst. - branch of Ulyanovsk State Agricultural Academy, Dimitrovgrad, Russia

loginov@ulstu.ru, omakeeva@hotbox.ru

Abstract **Abstract** The review of the authors' results is given on pseudoperturbiteration methods allowing to refine the spectral characteristics, eigenvalues, eigen- and Jordan elements of the operator-functions depending on spectral parameter and adjoint to them. The main attention is payed to the development of iteration processes and their various applications.

Keywords: generalized eigenvalue problems; adjoint problem; refining of spectral characteristics; perturbation and bifurcation theory; iteration processes; applications in mathematical physics.

2000 MSC: 47A55, 47A75, 58E07, 65J99.

1. INTRODUCTION

The idea of pseudoperturbation method (PPM), i.e. perturbation operator construction such that the known approximations to spectral characteristics would became exact for the perturbed operator, is due to M.K. Gavurin [1] in his application to the refining of simple eigenvalues and eigenelements of self-adjoint operators in Hilbert spaces. Later on his PhD-student F. Kuhnert [2] has solved this problem for non-self-adjoint operators. The further development of PP-method was given in the articles by B. V. Loginov, D. G. Rakhimov and N. A. Sidorov (see review article [3]), where the refining problem was solved for multiple eigenvalues, eigenvectors and generalized Jordan chains (GJCh) of linear by spectral parameter operator-function in Banach spaces. The suggested there PP-operator did not make possible to use GJChs of adjoint operator-function, and in the general case, the problem remained unsolved. In 2003 [4, 5] two forms of PP-operators were suggested which allow to refine multiple eigenvalues and GJChs of the eigenvalue problem linear by spectral parameter and adjoint to it in Banach spaces. They symmetrically used the known approximations to enumerating spectral characteristics with subsequent application of Newton-Kantorovich method to their refining. In the articles [6, 7] the development of PP-method to nonlinear with respect to spectral parameter operator-function was given based on its linearization possibilities.

Here four iteration processes for the determination of exact spectral characteristics are suggested together with the investigation of their convergence rates and stability with respect to small computation errors. The main attention is payed to various illustrations of PP-iteration method; some of them are reviewed in [8]. The PP-iteration (PPI) methods are considered in a group symmetry conditions. Also their connections with parameter continuation methods [9] are investigated. The theory of PPI-methods and the proof of their computational stability is based on bifurcation and perturbation theories [10].

The obtained results are supported by grant RFBR-Romanian Academy, N 07-01-91680-PA_a.

2. PSEUDOPERTURBATION OPERATOR CONSTRUCTION

In Banach spaces E_1 and E_2 the linear by spectral parameter eigenvalue problem with bounded for simplicity linear operators $B, A \in L(E_1, E_2)$

$$(B - tA)x = 0 \tag{1}$$

is considered. Let the unknown eigenvalue λ be the Fredholm point of the operator-function $B - tA$ with eigenelements $N(B - \lambda A) = span\{\varphi_1^{(1)}, \dots, \varphi_n^{(1)}\}$, $N(B^* - \lambda A^*) = span\{\psi_1^{(1)}, \dots, \psi_n^{(1)}\}$ and corresponding $A-$ and A^*- Jordan chains [10] of lengths $p_1 \leq p_2 \leq \dots \leq p_n$

$$\begin{gathered}(B - \lambda A)\varphi_i^{(s)} = A\varphi_i^{(s-1)},\ (B^* - \lambda A^*)\psi_i^{(s)} = A^*\psi_i^{(s-1)},\ s = \overline{2, p_i},\ i = \overline{1, n},\\ k = \det[\langle A\varphi_i^{(p_i)}, \psi_j^{(1)}\rangle] \neq 0,\ L_{ij} = [\langle A\varphi_i^{(p_i+1-s)}, \psi_j^{(l)}\rangle]_{l=\overline{1,p_j},\, s=\overline{1,p_i}} \neq 0,\\ L = \det[L_{ij}] \neq 0,\end{gathered} \tag{2}$$

which can be chosen [10, 11] to satisfy the biorthogonality relations

$$\begin{aligned}&\langle\varphi_i^{(j)},\gamma_k^{(l)}\rangle=\delta_{ik}\delta_{jl},\quad \langle z_i^{(j)},\psi_k^{(l)}\rangle=\delta_{ik}\delta_{jl},\quad j(l)=1,\ldots,p_i(p_k)\\ &\gamma_k^{(l)}=A^*\psi_k^{(p_k+1-l)},\quad z_i^{(j)}=A\varphi_i^{(p_i+1-j)},\quad i,k=1,\ldots,n.\end{aligned}\tag{3}$$

Let some sufficient good approximations λ_0, $\varphi_{i0}^{(s)}$, $\psi_{i0}^{(s)}$ to unknown eigenvalue λ and GJChs be given, $|\lambda-\lambda_0|\le\varepsilon$, $\|\varphi_i^{(s)}-\varphi_{i0}^{(s)}\|\le\varepsilon$, $\|\psi_i^{(s)}-\psi_{i0}^{(s)}\|\le\varepsilon$, with close to unit relevant magnitudes k_0 and L_0 (2). The problem is posed: to construct pseudoperturbation operator D_0, for which the given approximations would be exact for the perturbed operator.

Lemma 1. *Passing to linear combinations the systems* $\{\gamma_{k0}^{(l)}\}_{k=\overline{1,n}}^{l=\overline{1,p_k}}$, $\gamma_{k0}^{(l)}=A^*\psi_{k0}^{(p_k+1-l)}$, $\{z_{i0}^{(j)}\}_{i=\overline{1,n}}^{j=\overline{1,p_i}}$, $z_{i0}^{(j)}=A\varphi_{i0}^{(p_i+1-j)}$ *satisfying the biorthogonality relations* (3) $\langle\varphi_{i0}^{(j)},\gamma_{k0}^{(l)}\rangle=\delta_{ik}\delta_{jl}$, $\langle z_{i0}^{(j)},\psi_{k0}^{(l)}\rangle=\delta_{ik}\delta_{jl}$ *can be determined.*

Proof. In fact, let some sufficiently good approximations $\{\varphi_{i0}^{(j)}\}_{i=\overline{1,n}}^{j=\overline{1,p_i}}$, $\{\widetilde{\psi}_{\mu0}^{(\nu)}\}_{\mu=\overline{1,n}}^{\nu=\overline{1,p_\mu}}$ and $z_{i0}^{(j)}=A\varphi_{i0}^{(p_i+1-j)}$ be given. Setting $\widetilde{\gamma}_{\mu0}^{(\nu)}=A^*\widetilde{\psi}_{\mu0}^{(p_\mu+1-\nu)}$ form the linear combinations $\gamma_{k0}^{(l)}=\sum\limits_{\mu=1}^{n}\sum\limits_{\nu=1}^{p_\mu}K_{k\mu}^{l\nu}\widetilde{\gamma}_{\mu0}^{(\nu)}$, subject to the conditions $\langle\varphi_{i0}^{(j)},\gamma_{k0}^{(l)}\rangle=\delta_{ik}\delta_{jl}$, $i,k=1,\ldots,n$, $j(l)=1,\ldots,p_i(p_k)$. Then according to (2) the coefficients $K_{k\mu}^{l\nu}$ can be uniquely determined by the system of equations $\sum\limits_{\mu=1}^{n}\sum\limits_{\nu=1}^{p_\mu}K_{k\mu}^{l\nu}\langle A\varphi_{i0}^{(j)},\widetilde{\psi}_{\mu0}^{(p_\mu+1-\nu)}\rangle=\delta_{ik}\delta_{jl}$, $i,k=1,\ldots,n$, $j(l)=1,\ldots,p_i(p_k)$. Setting now $\psi_{k0}^{(p_k+1-l)}=\sum\limits_{\mu=0}^{n}\sum\limits_{\nu=1}^{p_\mu}K_{k\mu}^{l\nu}\widetilde{\psi}_{\mu0}^{(p_\mu+1-\nu)}$ from this system it follows $\langle z_{i0}^{(p_i+1-j)},\psi_{k0}^{(p_k+1-l)}\rangle=\delta_{ik}\delta_{jl}$, $z_{i0}^{(p_i+1-j)}=A\varphi_{i0}^{(j)}$.

Computing the discrepancies

$$\begin{aligned}&\sigma_{i0}^{(1)}=(B-\lambda_0A)\varphi_{i0}^{(1)},\quad \sigma_{i0}^{(j)}=(B-\lambda_0A)\varphi_{i0}^{(j)}-A\varphi_{i0}^{(j-1)},\\ &\tau_{i0}^{(1)}=(B^*-\lambda_0A^*)\psi_{i0}^{(1)},\quad \tau_{i0}^{(j)}=(B^*-\lambda_0A^*)\psi_{i0}^{(j)}-A^*\psi_{i0}^{(j-1)},\end{aligned}\qquad j=2,\ldots,p_i$$

determine [4] two forms of pseudoperturbation operator

$$D_0x=\sum_{i=1}^{n}\sum_{j=1}^{p_i}\langle x,\gamma_{i0}^{(j)}\rangle\Big[\sigma_{i0}^{(j)}-\sum_{k=1}^{n}\sum_{s=1}^{p_k}\langle\sigma_{i0}^{(j)},\psi_{k0}^{(s)}\rangle z_{k0}^{(s)}\Big]+\sum_{i=1}^{n}\sum_{j=1}^{p_i}\langle x,\tau_{i0}^{(j)}\rangle z_{i0}^{(j)}\tag{4}$$

$$D_0x=\sum_{i=1}^{n}\sum_{j=1}^{p_i}\langle x,\gamma_{i0}^{(j)}\rangle\sigma_{i0}^{(j)}+\sum_{i=1}^{n}\sum_{j=1}^{p_i}\Big[\langle x,\tau_{i0}^{(j)}-\sum_{k=1}^{n}\sum_{s=1}^{p_k}\langle\varphi_{k0}^{(s)},\tau_{i0}^{(j)}\rangle\gamma_{k0}^{(s)}\rangle\Big]z_{i0}^{(j)}\tag{5}$$

Theorem 1. [4] *Pseudoperturbation operators* (4) *and* (5) *have the following properties*

$$\begin{array}{l} D_0\varphi_{i0}^{(s)} = \sigma_{i0}^{(s)} \\ D_0^*\psi_{i0}^{(s)} = \tau_{i0}^{(s)} \end{array} \Rightarrow \begin{array}{ll} (B-\lambda_0 A - D_0)\varphi_{i0}^{(1)} = 0, & (B-\lambda_0 A - D_0)\varphi_{i0}^{(s)} = A\varphi_{i0}^{(s-1)} \\ (B^*-\lambda_0 A^* - D_0^*)\psi_{i0}^{(1)} = 0, & (B^*-\lambda_0 A^* - D_0^*)\psi_{i0}^{(s)} = A^*\psi_{i0}^{(s-1)} \end{array} \tag{6}$$

3. ITERATIVE PROCESSES AND THEIR CONVERGENCE RATES

A. Newton-Kantorovich method

According to the generalized E. Schmidt lemma [10] and general perturbation theory at the sufficiently exact initial approximations ($\|D_0\|$ is small) there exists the bounded operator $\Gamma = \widetilde{B}^{-1} = \left[(A_0 - \lambda_0 A_1 - D_0) + \sum\limits_{i=1}^{n} \left\langle \cdot, \gamma_{i0}^{(1)} \right\rangle z_{i0}^{(1)}\right]^{-1}$. Consequently the Jordan elements for the exact eigenvalue are determined by formulae [4]

$$\varphi_i^{(s)} = (I - (\lambda-\lambda_0)\Gamma A + \Gamma D_0)^{-1}\left[\Gamma A(I - (\lambda-\lambda_0)\Gamma A + \Gamma D_0)^{-1}\right]^{s-1}\varphi_{i0}^{(1)},$$
$$s = 1, 2, \ldots \tag{7}$$

and it is not difficult to see [4] that the exact eigenvalue $t = \lambda$ is $K = \sum\limits_{i=1}^{n} p_i$-multiple root of the equation

$$f(t) = \det[l_{ij}(t-\lambda_0)] = 0,$$
$$l_{ij}(t-\lambda_0) = \left\langle ((t-\lambda_0)\,A - D_0)\,[I - \Gamma(t-\lambda_0)A + \Gamma D_0]^{-1}\,\varphi_{i0}^{(1)}, \psi_{k0}^{(1)}\right\rangle$$
$$= -\left\langle\left\{I - [I - \Gamma((t-\lambda_0)A_1 - D_0)]^{-1}\right\}\varphi_{i0}^{(1)}, \gamma_{k0}^{(1)}\right\rangle,$$
$$l_{ij,s}(t-\lambda_0) = \frac{1}{s!}\frac{d^s l_{ij}(t-\lambda_0)}{dt^s}$$
$$= \left\langle [I - (t-\lambda_0)\Gamma A + \Gamma D_0]^{-1}\left(\Gamma A\left[I - (t-\lambda_0^{)}\Gamma A + \Gamma D_0\right]^{-1}\right)^s \varphi_{i0}^{(1)}, \gamma_{j0}^{(1)}\right\rangle,$$

i.e. $\left.\dfrac{d^s f(t)}{dt^s}\right|_{t=\lambda} = 0$ for $s < K$ and $\left.\dfrac{d^K f(t)}{dt^K}\right|_{t=\lambda} = \det\left[\dfrac{d^{p_i} l_{ij}(t-\lambda_0)}{dt^{p_i}}\right]_{t=\lambda} \neq 0$ by virtue of the relation

$$\frac{1}{p_i!}\frac{d^{p_i} l_{ik}(0)}{dt^{p_i}} = \left\langle (I+\Gamma D_0)^{-1}\left(\Gamma A(I+\Gamma D_0)^{-1}\right)^{p_i}\varphi_{i0}^{(1)}, \gamma_{k0}^{(1)}\right\rangle$$
$$= \left\langle (\Gamma A)^{p_i}\varphi_{i0}^{(1)}, \gamma_{k0}^{(1)}\right\rangle -$$
$$- \left\langle \left[\Gamma D_0(\Gamma A)^{p_i} + (\Gamma A)(\Gamma D_0)(\Gamma A)^{p_i-1} + \ldots + (\Gamma A)^{p_i}(\Gamma D_0)\right]\varphi_{i0}^{(1)}, \gamma_{k0}^{(1)}\right\rangle +$$
$$+ o(\|D_0\|) = \left\langle \varphi_{i0}^{(1)}, \gamma_{k0}^{(1)}\right\rangle + O(\|D_0\|) = \delta_{ik} + O(\|D_0\|). \tag{8}$$

In order to determine the exact λ of the equation $f^{(K-1)}(t) = 0$, the modified or basic Newton-Kantorovich (N.-K.) method can be applied taking $t = \lambda_0$ as the initial approximation

$$\lambda_\nu = \lambda_{\nu-1} - \left[f^{(K)}(\lambda_0)\right]^{-1} f^{(K-1)}(\lambda_{\nu-1}), \nu = 1, 2, \ldots \tag{9}$$

$$\lambda_\nu = \lambda_{\nu-1} - \left[f^{(K)}(\lambda_{\nu-1})\right]^{-1} f^{(K-1)}(\lambda_{\nu-1}), \nu = 1, 2, \ldots \tag{10}$$

Theorem 2. *Newton-Kantorovich method applied to the equation $f^{(K-1)}(t) = 0$, for sufficiently small $\|D_0\|$, determines its unique solution λ and has the square convergence rate.*

Proof. Indeed, by virtue of formulae [10,§31]

$$\varphi_{j0}^{(s)} = (\Gamma_0 A)^{s-1}\varphi_{j0}^{(1)} = \varphi_{j0}^{(s-\left[\frac{s}{p_j}\right]p_j)} \text{ and } \left\langle \varphi_{j0}^{(s)}, \gamma_{k0}^{(1)} \right\rangle = \delta_{jk}\delta_{(s-\left[\frac{s}{p_j}\right]p_j),1} \tag{11}$$

and according to Hadamard inequality for determinant, there exists a constant C such that $\left|f^{(s)}(\lambda_0)\right| \leq C\|D_0\|$, $s \neq K$. Also the relation (8) gives $f^{(K)}(\lambda_0) = p_1! \ldots p_n! \det\left[\delta_{ik} + O(\|D_0\|)\right]$, whence by virtue of the continuity $f^{(K)}(t)$ there exists a constant such that $\left|f^{(K)}(\lambda_v)\right|^{-1} \leq m(\rho)$ in some ρ-neighborhood $S_\rho(\lambda_0)$ of the point λ_0. The continuity of $f^{(K+1)}(\lambda)$ in $S_\rho(\lambda_0)$ gives the Lipschitz condition with some constant l $\left|f^{(K)}(\lambda_1) - f^{(K)}(\lambda_2)\right| \leq l\left|\lambda_1 - \lambda_2\right|, \quad \lambda_1, \lambda_2 \in S_\rho(\lambda_0)$. Consequently [12, §34.2] if $q = \frac{1}{2}m^2 lC\|D_0\| < 1$ and $\rho' = mC\|D_0\| \sum\limits_{k=0}^{\infty} q^{2^K-1} < \rho$, the equation $f^{(K-1)}(t) = 0$ has in the ball $S'_\rho(\lambda_0)$the unique solution λ to which the iterations (10) are square convergent.

Corollary. *According to theorem [12, §34.3] , if $2m^2 lC\|D_0\| < 1$ and $r' = \frac{1}{ml}(1 - \sqrt{1 - 2m^2 lC\|D_0\|}) < \rho$ the iterations* (9) *of the modified N.-K. method converge in the ball $\overline{S'_r(\lambda_0)}$ to the unique solution λ. The convergence rate of $\{\lambda_\nu\}$ to λ is given by the inequality $|\lambda - \lambda_\nu| \leq \frac{1 - \sqrt{1 - 2m^2 lC\|D_0\|}}{\sqrt{1 - 2m^2 lC\|D_0\|}} mC\|D_0\|$.*

After the determination of the exact eigenvalue λ the elements of GJChs can be determined directly from the following equations [4]

$$\begin{aligned} &(B - \lambda A)\varphi_j^{(1)} + \sum_{i=1}^{n} \left\langle \varphi_j^{(1)}, \gamma_{i0}^{(1)} \right\rangle z_{i0}^{(1)} = z_{j0}^{(1)}, \\ &(B - \lambda A)\varphi_j^{(s)} + \sum_{i=1}^{n} \left\langle \varphi_j^{(s)}, \gamma_{i0}^{(1)} \right\rangle z_{i0}^{(1)} = A\varphi_j^{(s-1)}, \end{aligned} \quad s = 2, \ldots, p_j, j = 1, \ldots, n \tag{12}$$

$$\begin{array}{l}(B^*-\lambda A^*)\psi_j^{(1)}+\sum\limits_{i=1}^{n}\left\langle z_{i0}^{(1)},\psi_j^{(1)}\right\rangle\gamma_{i0}^{(1)}=\gamma_{j0}^{(1)},\\ (B^*-\lambda A^*)\psi_j^{(s)}+\sum\limits_{i=1}^{n}\left\langle z_{i0}^{(1)},\psi_j^{(s)}\right\rangle\gamma_{i0}^{(1)}=A^*\psi_j^{(s-1)}.\end{array}\quad s=\overline{2,p_j},j=\overline{1,n} \qquad (13)$$

B. Newton-Kantorovich method with cubic convergence

In the articles [13, 14] one modification of the basic N.-K. method was suggested that has the cubic convergence. It is based on the change of the original nonlinear equation $g(t) = 0$ by the equation $F(t) = g(t)e^{-kt} = 0$, which has the same roots. This change gives the modification of the iteration process by the following one

$$t_{n+1}=t_n-g(t_n)\left[g'(t_n)-kg(t_n)\right]^{-1}, \qquad (14)$$

where the arbitrary parameter k is chosed in the special form $k=\dfrac{g''(t_n)}{2g'(t_n)}$. In fact, the setting $\varepsilon_n = t^*-t_n$ gives for the iterations (14) the following relation $\varepsilon_{n+1}=\varepsilon_n+g(t_n)\left[g(t_n)-kg(t_n)\right]^{-1}$, whence by using of the Taylor expansions of $g(t_n)$ and $g'(t_n)$ in a neighbourhood of the exact solution t^*, the equality

$$\varepsilon_{n+1}=-\varepsilon_n^2\frac{\frac{1}{2}g''(t^*)-kg'(t^*)-\varepsilon_n\left(\frac{1}{3}g'''(t^*)-\frac{k}{2}g''(t^*)\right)+o\left(|\varepsilon_n|\right)}{g'(t^*)-\varepsilon_n\left(g''(t^*)-kg'(t^*)\right)+o\left(|\varepsilon_n|\right)}$$

$$=\varepsilon_n^3\frac{\frac{1}{3}g'''(t^*)-\frac{k}{2}g''(t^*)+o\left(|\varepsilon_n|\right)}{g'(t^*)-\varepsilon_n\frac{1}{2}g''(t^*)+o\left(|\varepsilon_n|\right)}.$$

Under the theorem of [12, §34.2] assumptions about convergence rate of the basic Newton method in [14] its analog about cubic convergence of the suggested modification is proved.

Consequently the relevant iteration scheme for our problem has the form

$$t_n=t_{n-1}-\frac{2g(t_{n-1})g'(t_{n-1})}{2g'^2(t_{n-1})-g''(t_{n-1})g(t_{n-1})}= \qquad (15)$$

$$g=f^{(K-1)}(t)t_{n-1}-\frac{2f^{(K-1)}(t_{n-1})f^{(K)}(t_{n-1})}{2\left[f^{(K)}(t_{n-1})\right]^2-f^{(K+1)}(t_{n-1})f^{(K-1)}(t_{n-1})},$$

which has the cubic convergence at sufficiently small $\|D_0\|$.

C. Eitken-Steffensen iteration process

In the monographs [15, 16, 17] it was investigated the Eitken-Steffensen iteration process which does not require the derivatives computation, however possessing the square convergence rate

$$\lambda_n = \lambda_{n-1} - \frac{g(\lambda_{n-1})}{g(\lambda_{n-1}) - g\left(\psi(\lambda_{n-1})\right)} g(\lambda_{n-1}), \psi(t) = t - f(t). \tag{16}$$

For our problem we must take $g(t) = f^{(K-1)}(t)$, and (16) takes the form

$$\lambda_n = \lambda_{n-1} - \frac{f^{(K-1)}(\lambda_{n-1})}{f^{(K-1)}(\lambda_{n-1}) - f^{(K-1)}\left(\psi(\lambda_{n-1})\right)} f^{(K-1)}(\lambda_{n-1}).$$

On every step of this process it is required to solve $4n^2(K+1)^2$ linear equations (12), (13) for $\lambda = \lambda_{n-1}$ and $\psi(\lambda_{n-1}) = \lambda_{n-1} - f^{(K-1)}(\lambda_{n-1})$, respectively.

Remark 1. The iteration processes **A** and **C** can be changed on the relevant processes for the computation of multiple roots [15], for our problem of K-multiple roots of the equation $f(t) = \det[l_{ij}(t - \lambda_0)] = 0$,

A. $t_n = t_{n-1} - K\dfrac{f(t_{n-1})}{f'(t_{n-1})}$, **C.** $\lambda_n = \lambda_{n-1} - K\dfrac{f(\lambda_{n-1})}{f(\lambda_{n-1}) - f(\psi(\lambda_{n-1}))} f(\lambda_{n-1})$, which have superlinear convergence rate.

D. Gavurin iteration process

By means of pseudoperturbation operator D_0 (4), (5) the eigenvalue problem (1) is rewritten in the form of the perturbation problem for the eigenvalue λ_0 with the small parameter D_0 [10, §32]

$$B_0\varphi \equiv (B - \lambda_0 A - D_0)\varphi = (\Delta\lambda)A\varphi - D_0\varphi, \quad \mu = \Delta\lambda = \lambda - \lambda_0, \tag{17}$$

which can be reduced [10] to the branching equation (BEq)

$$L(\mu, \|D_0\|) = \det\left[\langle(((\Delta\lambda)A - D_0)\left[I - \Gamma_0\left((\Delta\lambda)A - D_0\right)\right]^{-1} \varphi_{i0}^{(1)}, \psi_{k0}^{(1)}\rangle\right] = 0. \tag{18}$$

The length of the decreasing part of the Newton diagram for (13) is equal to $K = \sum\limits_{i=1}^{n} p_i$, since $L(\mu, 0) = \sum\limits_{s=k}^{\infty} L_{s0}\mu^s$, $L_{s0} = 0$, $s < k$,

$L_{K0} = \det\left[\langle \mu A\left(I - \Gamma_0 A\right)^{-1} \varphi_{i0}^{(1)}, \psi_{k0}^{(1)}\rangle\right] = \det\left[\langle A\varphi_{i0}^{(p_i)}, \psi_{k0}^{(1)}\rangle\right] = 1.$

However, some of BEq coefficients $L_{s,K-s}$ may be nonzero. Let ν be the first number for which $L_{\nu,K-\nu} \neq 0$.

From formulae (7) it follows that as first approximations to the exact Jordan chain elements we must to take the following

$$\begin{aligned}
\varphi_{j1}^{(1)} &= (I + \Gamma D_0)^{-1}\varphi_{j0}^{(1)}, \\
\varphi_{j1}^{(2)} &= (I + \Gamma D_0)^{-1}\Gamma A(I + \Gamma D_0)^{-1}\varphi_{j0}^{(1)} = (I + \Gamma D_0)^{-1}\Gamma A\varphi_{j1}^{(1)}, \ldots, \\
\varphi_{j1}^{(p_i)} &= (I + \Gamma D_0)^{-1}\left[\Gamma A(I + \Gamma D_0)^{-1}\right]^{p_j - 1}\varphi_{j0}^{(1)} = (I + \Gamma D_0)^{-1}\Gamma A\varphi_{j1}^{(p_i - 1)},
\end{aligned} \tag{19}$$

$$\begin{aligned}
\psi_{j1}^{(1)} &= (I+\Gamma^* D_0^*)^{-1}\psi_{j0}^{(1)},\\
\psi_{j1}^{(2)} &= (I+\Gamma^* D_0^*)^{-1}\Gamma^* A^*(I+\Gamma^* D_0^*)^{-1}\psi_{j0}^{(1)} = (I+\Gamma^* D_0^*)^{-1}\Gamma^* A^*\psi_{j1}^{(1)},\dots,\\
\psi_{j1}^{(p_i)} &= (I+\Gamma^* D_0^*)^{-1}\left[\Gamma^* A^*(I+\Gamma^* D_0^*)^{-1}\right]^{p_j-1}\psi_{j0}^{(1)}\\
&= (I+\Gamma^* D_0^*)^{-1}\Gamma^* A^*\psi_{j1}^{(p_i-1)}.
\end{aligned} \tag{20}$$

These elements are the solutions of the equations

$$\left.\begin{aligned}
&(B-\lambda_0 A)\varphi_{j1}^{(1)} + \sum_{k=1}^{n}\left\langle \varphi_{j1}^{(1)},\gamma_{k0}^{(1)}\right\rangle z_{k0}^{(1)} = z_{j0}^{(1)},\\
&(B-\lambda_0 A)\varphi_{j1}^{(s)} + \sum_{k=1}^{n}\left\langle \varphi_{j1}^{(s)},\gamma_{k0}^{(1)}\right\rangle z_{k0}^{(1)} = A\varphi_{j1}^{(s-1)},
\end{aligned}\right. \quad s=2,\dots,p_j, j=1,\dots,n$$

$$\left.\begin{aligned}
&(B^*-\lambda_0 A^*)\psi_{j1}^{(1)} + \sum_{k=1}^{n}\left\langle z_{k0}^{(1)},\psi_{j1}^{(1)}\right\rangle \gamma_{k0}^{(1)} = \gamma_{j0}^{(1)},\\
&(B^*-\lambda_0 A^*)\psi_{j1}^{(s)} + \sum_{k=1}^{n}\left\langle z_{k0}^{(1)},\psi_{j1}^{(s)}\right\rangle \gamma_{k0}^{(1)} = A^*\psi_{j1}^{(s-1)},
\end{aligned}\right. \quad s=2,\dots p_j, j=1,\dots,n$$

respectively. The relevant first approximation λ_1 to the exact eigenvalue λ can be found by the Newton diagram method from the BEq (18). Now after Lemma 1 application, i.e. after the biorthogonalization of the first approximations, we repeat the iteration process: the pseudoperturbation operator D_1 is determinating and the second approximations to Jordan elements $\varphi_{j2}^{(s)}$, $\psi_{j2}^{(s)}$, $s=1,\dots,p_j$, $j=1,\dots,n$ with the relevant λ_2 determination and so on.

Theorem 3. *The M.K. Gavurin iteration processes is converging with the rate of the order* $1+\dfrac{1}{\nu}$, *i.e. it is superlinearly convergent.*

The technically difficult proof coincides with the proof of relevant theorems in [18, 19].

4. STABILITY OF ITERATION PROCESSES

In [20, 21] regularization by A.N. Tychonov questions in perturbation and bifurcation theories are investigated. The symbol $\sim$ denotes δ-approximations of the corresponding magnitudes, $\|\widetilde{A}x - Ax\| \le \delta(\|x\| + a\|Ax\|)$, $\forall x \in D(A)$. Here we give the result for linear equations, on the base of which the computational stability of suggested iteration processes can be satisfied by Theorem 4 [20, 21]. Let in the equation $Ax=f$, $A: E_1 \supset D(A) \to E_2$, $\overline{D(A)} = E_1$ the operator A be Fredholm; $N(A) = span\{\varphi_1,\dots,\varphi_n\}$, $N^*(A) = span\{\psi_1,\dots,\psi_n\}$ be its zero and defect subspaces; $\{\gamma_i\}_1^n \in E_1^*$ and $\{z_i\}_1^n \in E_2^*$ be corresponding biorthogonal systems. Let $x^* = \Gamma f$, $\Gamma = \big(A + \sum\limits_{i=1}^{n} \langle\cdot, \gamma_i\rangle z_i\big)^{-1}$ be the normal solution of the equation $Ax = f$. If $\delta < q[a + (1+na)\|\Gamma\|]^{-1}$, $0 < q < 1$, then the unique solution of the regularized equation $\widetilde{A}x + \sum\limits_{i=1}^{n} \langle x, \gamma_i\rangle z_i = \widetilde{f}$ is determined by the formula $\widetilde{x} = \big(\widetilde{A}x + \sum\limits_{i=1}^{n} \langle x, \gamma_i\rangle z_i\big)^{-1}\widetilde{f}$ and satisfies the following estimate $\|\widetilde{x} - \Gamma f\| \le (1-q)^{-1}\|\Gamma\|(1 + a\|f\| + \|\Gamma f\|)\delta$.

5. APPLICATIONS OF PSEUDOPERTURBITERATION METHODS

A. Pseudoperturbiteration method for differential equations with displacements in boundary conditions

Here we consider three problems with displacements: 1. one-dimensional Bitsadze-Samarskii problem [22]; 2. four eigenvalue problems with two displacements [23]; 3. E. Schmidt eigenvalue problems with displacement [24, 25, 26].

1. In the space $C^2[(0, x_0) \cup (x_0, 1)] \cap C^1[0, 1]$ the linear eigenvalue problem $u'' + \lambda u = 0$, $u(0) = 0$, $u(x_0) = u(1)$ is considered. It has two sets of eigenvalues $\mu_m = \dfrac{2m\pi}{1 - x_0}$, $\mu_s = \dfrac{(2s+1)\pi}{1 + x_0}$, to which the eigenfunctions $\varphi_m = \sin\mu_m x$, $\varphi_s = \sin\mu_s x$ correspond. The adjoint problem $v'' + \lambda v = 0$, $v(0) = 0$, $v(1) = 0$, $v'(x_0 + 0) - v'(x_0 - 0) = v'(1)$ in the space $C^2[(0, x_0) \cup (x_0, 1)] \cap C[0, 1]$ has the same eigenvalues with relevant eigenfunctions

$$\psi_m = \begin{cases} 0, & 0 \le x \le x_0, \\ \sin\mu_m(1 - x), & x_0 \le x \le 1; \end{cases}$$

$$\psi_s = \begin{cases} \dfrac{\sin\mu_s(1 - x_0) \cdot \sin\mu_s x}{\sin\mu_s x_0}, & 0 \le x \le x_0, \\ \sin\mu_s(1 - x), & x_0 \le x \le 1. \end{cases}$$

The Jordan chains of the length two exist in the condition $\mu_m = \mu_s = \mu_0 \Rightarrow x_0 = \dfrac{2s - 2m + 1}{2s + 2m + 1}$. They are computed in the form satisfying the biorthogonality conditions (3).

The realization of PPI-method is fulfilled for $m = s = 1$, $x_0 = \frac{1}{5}$, $\lambda = \mu_0^2 = \frac{25\pi^2}{4}$

$$\begin{aligned} \varphi^{(1)} &= \sin\frac{5\pi}{2}x, & \psi^{(1)} &= \begin{cases} 0, & 0 \le x \le \frac{1}{5}, \\ \frac{125\pi}{6}\cos\frac{5\pi}{2}x, & \frac{1}{5} \le x \le 1; \end{cases} \\ \varphi^{(2)} &= -\frac{1}{5\pi}x\cos\frac{5\pi}{2}x, & \psi^{(2)} &= \begin{cases} -\frac{10}{3}\sin\frac{5\pi}{2}x, & 0 \le x \le \frac{1}{5}, \\ -\frac{25}{6}(1-x)\sin\frac{5\pi}{2}x, & \frac{1}{5} \le x \le 1. \end{cases} \end{aligned} \tag{21}$$

As approximations $\widehat{\varphi}_0^{(1)}$, $\widehat{\psi}_0^{(1)}$, $\widehat{\varphi}_0^{(2)}$, $\widehat{\psi}_0^{(2)}$ to eigenfunctions $\varphi_0^{(1)}$, $\widetilde{\psi}_0^{(1)}$ and Jordan elements $\varphi_0^{(2)}$, $\widetilde{\psi}_0^{(2)}$ were selected relevant parts of Taylor series for (7), and the approximation to eigenvalue λ_0 was determined to within 1 from the equation $\left\langle \frac{d^2}{dx^2}\widehat{\varphi}_0^{(2)} + t\widehat{\varphi}_0^{(2)} + \widehat{\varphi}_0^{(1)}, \widehat{\psi}_0^{(1)} \right\rangle = 0$, by stopping the iteration process at the achievement of a sufficient accuracy of approximation to λ_0. Note that in the considered example the exact spectral characteristics were known.

The problems 2. are investigated in [23], the problems 3. will be described in part **C** 4.

B. The refining of critical spectral parameter at Poincaré-Andronov-Hopf bifurcation

This problem in the general form is considered in [25], separately for the DE of the first order $Ax' = Bx - R(x,\varepsilon)$, $R(0,0) = 0$, $R_x(0,0) = 0$ and DE of higher order of the form $A_s x^{(s)} + A_{s-1}x^{(s-1)} + \ldots + A_1 x' = Bx - R(x, x^{(1)}, \ldots, x^{(s-1)}, \varepsilon)$, $R(0,\ldots,0,0) = 0$, $R_{x^k}(0,\ldots,0,0) = 0$ with bounded, for simplicity, linear operators. Therefore we give here only concrete example in part **C**.3° as spatially one-dimensional dynamical problem with a displacement illustrating the results [25].

C. Generalized pseudoperturbiteration methods for computation of E. Schmidt eigenvalue problems

At the beginning of the past century, E. Schmidt has introduced [26] systems of eigenvalues $\{\lambda_k\}$ (counted with their multiplicity) and eigenelements $\{\varphi_k\}_1^\infty$, $\{\psi_k\}_1^\infty$ satisfying the relations $B\varphi_k = \lambda_k\psi_k$, $B^*\psi_k = \lambda_k\varphi_k$ for integral Fredholm operators and allowing to generalize Hilbert-Schmidt theory on non-self-adjoint completely continuous operators in abstract separable Hilbert space [27]. As s-numbers these systems have found many applications in computational mathematics and ill-posed problems theory.

1. For a pair of linear, bounded for simplicity, operators $B, A \in L(H)$ the generalized E. Schmidt eigenvalues and eigenelements are introduced [28] by the following equalities: $B\varphi = \lambda A\psi$, $B^*\psi = \lambda A^*\varphi$. Here H is a Hilbert space, the eigenvalue λ can be chosen real. The above system can be rewritten in matrix form

$$(\mathcal{B} - \lambda\mathcal{A})\Phi = 0, \quad (\mathcal{B}^* - \lambda\mathcal{A}^*)\Psi = 0, \tag{22}$$

where $\mathcal{A} = \text{diag}\{A^*, A\}$, $\mathcal{B} = \text{secondarydiag}\{B^*, B\}$, which is non-selfadjoint. Therefore Schmidt eigenelements can have generalized Jordan chains $\Phi_k^{(j)} = \left(\varphi_k^{(j)}, \psi_k^{(j)}\right)$ and $\Psi_k^{(j)} = \left(\hat{\varphi}_k^{(j)}, \hat{\psi}_k^{(j)}\right)$, $j = \overline{1, p_k}$, $k = \overline{1, n}$. They can be chosen so that the following orthogonality relations be satisfied

$$\left(\Phi_i^{(j)}, \Gamma_k^{(l)}\right) = \left(\varphi_i^{(j)}, A\hat{\varphi}_k^{(p_k+1-l)}\right) + \left(A\psi_i^{(j)}, \hat{\psi}_k^{(p_k+1-l)}\right) = \delta_{ik}\delta_{jl},$$

$$\Gamma_k^{(l)} = \mathcal{A}^*\Psi_k^{(p_k+1-l)},$$

$$\left(Z_i^{(j)}, \Psi_k^{(l)}\right) = \left(\varphi_i^{(p_i+1-j)}, A\hat{\varphi}_k^{(l)}\right) + \left(A\psi_i^{(p_i+1-j)}, \hat{\psi}_k^{(l)}\right) = \delta_{ik}\delta_{jl},$$

$$Z_i^{(j)} = \mathcal{A}\Phi_i^{(p_i+1-j)}, \ i, k = \overline{1, n}, j(l) = \overline{1, p_i(p_k)}.$$

Let for n-multiple Schmidt eigenvalue λ and generalized Jordan chains (GJCh) some sufficiently good approximations λ_0, $\Phi_{k0}^{(j)}$, $\Psi_{k0}^{(j)}$, $j = \overline{1, p_k}$, $k = \overline{1, n}$, be given. The problem arises to construct the pseudoperturbation operator such that these approximations would be exact magnitudes for the perturbed problem and to give iteration procedure for their refining (as an eigenvalue and GJCh of the given operators) on the base of perturbation theory [28].

2. The second problem connected with the E. Schmidt spectrum arises in abstract Dirac type systems [29] at dynamical bifurcation

$$\mathcal{A}\frac{dX}{dt} \equiv \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix} \begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = \begin{pmatrix} B & 0 \\ 0 & -B^* \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - R(X, \varepsilon) \equiv \mathcal{B}X - R(X, \varepsilon),$$

from pure imaginary n-multiple $\mathcal{A}$-eigenvalue $\pm i\alpha$ of the operator $\mathcal{B}$. Then according to [30]

$$\mathbf{B}(\alpha)\Phi_k^{(1)} \equiv \begin{pmatrix} \mathcal{B} & \alpha\mathcal{A} \\ -\alpha\mathcal{A} & \mathcal{B} \end{pmatrix} \begin{pmatrix} U_{1k}^{(1)} \\ U_{2k}^{(1)} \end{pmatrix} = 0,$$

$$\mathbf{B}^*(\alpha)\Psi_k^{(1)} \equiv \begin{pmatrix} \mathcal{B}^* & -\alpha\mathcal{A}^* \\ \alpha\mathcal{A}^* & \mathcal{B}^* \end{pmatrix} \begin{pmatrix} V_{1k}^{(1)} \\ V_{2k}^{(1)} \end{pmatrix} = 0,$$

$U_{sk}^{(1)} = \left(u_{s1k}^{(1)}, u_{s2k}^{(1)}\right)^T$, $V_{sk}^{(1)} = \left(v_{s1k}^{(1)}, v_{s2k}^{(1)}\right)^T \in H\dot{+}H$, $s = 1,2$, $k = \overline{1,n}$, which means that the numbers $\pm\alpha$ are n-multiple Schmidt A-eigenvalues of the operator B, resp. Schmidt A^*-eigenvalues of the operator B^*

$$Bu_{11k}^{(1)} = -\alpha Au_{22k}^{(1)},\ Bu_{21k}^{(1)} = \alpha Au_{12k}^{(1)};\ Bv_{22k}^{(1)} = \alpha A^*v_{11k}^{(1)},\ Bv_{12k}^{(1)} = -\alpha A^*v_{21k}^{(1)};$$

$$B^*u_{22k}^{(1)} = -\alpha A^*u_{11k}^{(1)},\ B^*u_{12k}^{(1)} = \alpha A^*u_{21k}^{(1)};\ B^*v_{11k}^{(1)} = \alpha Av_{22k}^{(1)},\ B^*v_{21k}^{(1)} = -\alpha Av_{12k}^{(1)}.$$

Again the biorthogonality properties for the conveniently chosen $\mathbf{A}$-Jordan chains of the operator $\mathbf{B}(\alpha) = \mathbf{B} - \alpha\mathbf{A}$, $\mathbf{B} = \operatorname{diag}(B, -B^*, B, -B^*)$, $\mathbf{A} =$ secondary $\operatorname{diag}(-A, -A^*, A, A^*)$, take place

$$\left(\mathbf{A}\Phi_{\mu k}^{(l_k+1-j)}, \Psi_{\nu s}^{(l)}\right) = \delta_{\mu\nu}\delta_{ks}\delta_{jl},\ \left(\Phi_{\mu k}^{(j)}, \mathbf{A}^*\Psi_{\nu s}^{(p_s+1-l)}\right) = \delta_{\mu\nu}\delta_{ks}\delta_{jl},\ \mu,\nu = 1,2,$$

$$j = \overline{1,p_k},\ l = \overline{1,p_s},\ k,s = \overline{1,n}, \Phi_{1k}^{(s)} = \left(U_{1k}^{(s)}, U_{2k}^{(s)}\right)^T,$$

$$\Phi_{2k}^{(s)} = \left(-U_{2k}^{(s)}, U_{1k}^{(s)}\right)^T, \Psi_{1k}^{(s)} = \left(V_{2k}^{(s)}, -V_{1k}^{(s)}\right)^T,\ \Psi_{2k}^{(s)} = \left(V_{1k}^{(s)}, V_{2k}^{(s)}\right)^T.$$

Thus the problem of refining the approximately given E. Schmidt n-multiple eigenvalue α and relevant eigenelements with Jordan chains $u_{ijk0}^{(s)}, v_{ijk0}^{(s)}$ arises as a generalization of the previous one.

The problem 2. can be presented in the same form (22) as the problem 1. by replacing $\mathcal{A}$ and $\mathcal{B}$ by $\mathbf{A}$ and $\mathbf{B}$, λ by α. In the following arguments the forms (22) will be used both for the problem 1. and 2.

For the approximations to n-multiple Schmidt eigenvalue of the problem 1. (analogously to the problem 2.) with relevant GJChs we introduce the following notations: λ_0, $\Phi_{i0}^{(j)}$, $\Psi_{k0}^{(j)}$, $\Gamma_{k0}^{(j)}$, $Z_{i0}^{(j)}$. Passing, if necessary, to linear combinations (lemma 1) they can be considered satisfying the biorthogonality relations. Computing the discrepancies

$$\sigma_{i0}^{(j)} = (\mathcal{B} - \lambda_0\mathcal{A})\Phi_{i0}^{(j)} + (1 - \delta_{1j})\mathcal{A}\Phi_{i0}^{(j-1)},$$

$$\tau_{i0}^{(j)} = (\mathcal{B}^* - \lambda_0 \mathcal{A}^*)\Psi_{i0}^{(j)} + (1 - \delta_{1j})\mathcal{A}^* \Psi_{i0}^{(j-1)},$$

introduce two following forms of pseudoperturbation operators D_0 (see (4), (5)), such that $D_0 \Phi_{i0}^{(s)} = \sigma_{i0}^{(s)}$, $D_0^* \Psi_{i0}^{(s)} = \tau_{i0}^{(s)}$, so the approximations to eigenvalue and GJChs are exact for the perturbed operator $\mathcal{B} - \lambda_0 \mathcal{A} - D_0$.

Now one of iteration procedure of 3. can be applied.

Remark 2. To generalized E. Schmidt eigenvalue problem with s operators

$$\begin{aligned} &B_1 \varphi_1^{(1)} = \lambda A \psi_s^{(1)}, \quad B_2 \varphi_2^{(1)} = \lambda A \psi_{s-1}^{(1)}, \quad \ldots, \quad B_s \varphi_s^{(1)} = \lambda A \psi_1^{(1)}, \\ &B_1^* \psi_s^{(1)} = \lambda A^* \varphi_1^{(1)}, \quad B_2^* \psi_{s-1}^{(1)} = \lambda A^* \varphi_2^{(1)}, \quad \ldots, \quad B_s^* \psi_1^{(1)} = \lambda A^* \varphi_s^{(1)}, \end{aligned}$$

and problems with polynomial or analytic operator-function of spectral parameter

$$B\varphi = A(\lambda)\psi \equiv \sum_{k=1}^{s} \lambda^k A_k \psi, \quad B^* \psi = A^*(\lambda)\varphi \equiv \sum_{k=1}^{s} \lambda^k A_k^* \varphi,$$

as generally in generalized eigenvalue problems with polynomial (analytic) dependence on spectral parameter their linearization by means of matrix operators [6, 7] is applied [28].

3. Consider here the example of one dimensional dynamical problem, arising at the determination of spectral parameter critical value at Poicaré-Andronov-Hopf bifurcation [25] for the equation

$$\mathcal{A}\frac{dX}{dt} = \mathcal{B}X + R(X, \varepsilon),\ R(0, \varepsilon) = 0, \quad \mathcal{A} = \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix}, \mathcal{B} = \begin{pmatrix} B & 0 \\ 0 & -B^* \end{pmatrix},$$

$$A = A^* = I,\ B = B^* = \frac{d^2 x}{dt^2} + I$$

which can be written in the form of the periodical solutions determination for the boundary value problem

$$\begin{aligned} \frac{\partial v}{\partial t} &= \frac{d^2 u}{dx^2} + u + R_1(u, v, \varepsilon), \quad R_1(0,0,\varepsilon) = 0, \quad u(0,t) = 0, \quad u(x_0, t) = u(1,t), \\ \frac{\partial u}{\partial t} &= -\frac{d^2 v}{dx^2} - v + R_2(u, v, \varepsilon), \quad R_2(0,0,\varepsilon) = 0, \quad v(1,t) = 0, \quad v(x_0, t) = v(0,t). \end{aligned}$$

The linearized problem has the pure imaginary critical eigenvalues
$u, v \in C^2\left([0, x_0) \cup (x_0, 1]\right) \cap C^1 [0,1]$, $0 < x_0 < 1$,

$$u'' + u = i\alpha v, \quad v'' + v = -i\alpha u, \tag{23}$$

$$u(0) = 0, \quad u(x_0) = u(1), \quad v(1) = 0, \quad v(x_0) = v(0),$$

where it is sufficient to find positive values α and relevant eigenfunctions $u(x)$, $v(x)$. It is established that nontrivial solutions exist only for $\alpha > 1$. The adjoint problem in the space $C^2([0, x_0) \cup (x_0, 1]) \cap C[0, 1]$, has the form

$$\begin{aligned} &\tilde{u}'' + \tilde{u} = -i\alpha\tilde{v}, \quad \tilde{v}'' + \tilde{v} = i\alpha\tilde{u}, \quad 0 < x_0 < 1, \\ &\quad \tilde{u}(0) = 0, \ \tilde{u}(1) = 0, \ \tilde{u}(x_0 + 0) = \tilde{u}(x_0 - 0), \\ &\quad \tilde{u}'(x_0 + 0) - \tilde{u}'(x_0 - 0) = \tilde{u}'(1), \\ &\quad \tilde{v}(0) = 0, \tilde{v}(1) = 0, \tilde{v}(x_0 + 0) = \tilde{v}(x_0 - 0), \\ &\quad \tilde{v}'(x_0 + 0) - \tilde{v}'(x_0 - 0) = -\tilde{v}'(0). \end{aligned} \tag{24}$$

The equation

$$\begin{aligned} &\Delta \equiv [\sin\nu - \sin\nu(1 - x_0)]\,(\sin h\mu - \sin h\mu x_0) + \\ &+ (\sin\nu - \sin\nu x_0)\,[\sinh\mu - \sinh\mu(1 - x_0)] = 0, \end{aligned}$$

$\mu = \sqrt{\alpha - 1}$, $\nu = \sqrt{\alpha + 1}$, determines, for every $0 < x_0 < 1$, the eigenvalue $\alpha = \alpha(x_0)$ of the problem (23). The system $\Delta = 0$, $\Delta'_\alpha = 0$ is inconsistent and Jordan elements are absent.

The PPI-method is illustrated here for the case $x_0 = 0,5$. Here $\Delta = 32\sinh\dfrac{\mu}{4}\cosh\dfrac{3\mu}{4}\sin\dfrac{\nu}{4}\cos\dfrac{\nu}{4}(1 - 4\sin^2\dfrac{\nu}{4}) \Rightarrow \alpha = 16\pi^2 n^2 - 1$, $n = 1, 2, \ldots$; $\alpha = 4\pi^2(1 + 2m)^2 - 1$, $m = 0, 1, 2, \ldots$; $\alpha = 16(\pi s \pm \dfrac{\pi}{6})^2 - 1$, $s = 0, 1, 2, \ldots$

The smallest eigenvalue is $\alpha = \dfrac{4\pi^2}{9} - 1$ at $s = 0$, for which the explicit formulae for eigenfunctions of the problems are obtained. For this eigenvalue the computational experiment is made with approximation $\widetilde{\alpha} = 3.4375$, convergent on the fifth step up to 10^{-16}. Schmidt's eigenfunctions are determined then after refining $\widetilde{\alpha}$ on the relevant formulae.

4. In [24] in the functional class $C^2([0, x_0) \cup (x_0, 1]) \cap C^1[0; 1]$, $0 < x_0 < 1$ the E. Schmidt boundary eigenvalue problem

$$u'' + \lambda v = 0, \ v'' + \lambda u = 0, \quad u(0) = 0, \ u(x_0) = u(1), \quad v(1) = 0, \ v(x_0) = v(0)$$

is considered. The adjoint problem is stated. The determinant of the boundary conditions is calculated

$$\Delta = \big(\sinh\mu - \sinh\mu x_0\big)\big[\sin\mu - \sin\mu(1-x_0)\big]+$$
$$+\big(\sin\mu - \sin\mu x_0\big)\big[\sinh\mu - \sinh\mu(1-x_0)\big],$$
$$\lambda = \mu^2.$$

Analysis of the system $\Delta = 0$, $\Delta'_\mu = 0$ shows that Jordan elements are absent. The eigenfunctions for direct and adjoint problems are determined. The computational experiment is made again for symmetric particular case $x_0 = 0,5$, where $\Delta = -8\sin\frac{\mu}{2}\big[\sinh\mu - \sinh\frac{\mu}{2}\big]\big[2\cos\frac{\mu}{2} - 1\big] \Rightarrow \mu_1 = 2\pi n$ or $\mu_2 = \pm\frac{2\pi}{3} + 4\pi s$; $n, s = 0, 1, 2, \ldots$; $\mu > 0$. For μ_1 the eigenfunctions are absent, for μ_2 $\sin\mu = \sin\mu x_0 = \pm\frac{\sqrt{3}}{2} \neq 0$ and for the direct problem

$$u(x) = \pm 2\sin(\mu x \pm \frac{\pi}{6}) + (\sinh\frac{\mu}{2})^{-1}\sinh(\mu x - \frac{\mu}{2}),$$
$$v(x) = \pm 2\sin(\mu x \pm \frac{\pi}{6}) - (\sinh\frac{\mu}{2})^{-1}\sinh(\mu x - \frac{\mu}{2}),$$

for the adjoint one

$$u(x) = v(x) = \begin{cases} \sin\mu x, & 0 \le x \le \frac{1}{2}, \\ \sin\mu(1-x), & \frac{1}{2} \le x \le 1. \end{cases}$$

The computational experiment is fulfilled for the smallest eigenvalue μ_0^2, $\mu_0 = \frac{2\pi}{3}$, with approximation $\widetilde{\mu}_0 = 2,0938$ to within of the order 10^{-21}.

One case $x_0 \neq 0,5$ is considered, where the approximation $\mu_0 = 2.0938$ was determined graphically. This is the *unique* example substantive of PPI-method application, when we can't indicate the exact solution.

5.° Model problems of electromagnetic oscillations in lossless resonators.

Let be given the closed domain $V \subset R^3$ with piecewise smooth boundary $S = \partial V$. The eigenvalue oscillations of lossless resonator are called solutions of the boundary eigenvalue problems for the homogeneous Maxwell system

$$\begin{array}{ll} \mathrm{rotE} = i\omega\mu\mathrm{H}, & \mathrm{rotH} = -i\omega\varepsilon\mathrm{E}; \\ \mathrm{divE} = 0, & \mathrm{divH} = 0; \end{array} \qquad \big[\mathrm{n},\mathrm{E}\big]\big|_S = 0, \quad \big(\mathrm{n},\mathrm{H}\big)\big|_S = 0 \qquad (25)$$

n is the unit outer normal to S, ε and μ are dielectric and magnetic permeabilities of the medium. (25) is the typical E. Schmidt eigenvalue problem, ω

is the Schmidt's eigenvalue parameter. In the article [31] the adjoint to (25) system

$$\begin{array}{ll} \mathrm{rot}\mathcal{E} = i\omega\varepsilon\mathcal{H}, & \mathrm{rot}\mathcal{H} = -i\omega\mu\mathcal{E}; \\ \mathrm{div}\mathcal{E} = 0, & \mathrm{div}\mathcal{H} = 0; \end{array} \quad [\mathrm{n}, \mathcal{H}]\big|_S = 0, \quad (\mathrm{n}, \mathcal{E})\big|_S = 0$$

is constructed and on the base of I.S. Arzânykh basic formulae of the field theory for problems related to the Helmholtz operator [32, 33] and potential theory the system of integral equations equivalent to eigenvalue problem (25) is derived. For this the Fredholm property of the system (25) is proved.

In the monographs [34]-[38] and articles [39, 40] exact solutions of (25) for rectangular, cylindrical and spherical resonators are found obtained by using variables separation method and symmetry properties of the domain V. The definite symmetries of the domain generates the development of PPI-methods in conditions of group symmetry.

D. Solution of Algebraic Equations.

The method of replacing an algebraic equation by the characteristic equation of the corresponding Frobenius matrix, which, as known, is not rational, makes possible to apply the pseudoperturbation construction to refine roots of polynomials [41]. Although here we can not give the explicit formula for the E. Schmidt operator, this approach leads to interesting relations and summation formulae.

6. PSEUDOPERTURBITERATION METHOD UNDER GROUP SYMMETRY CONDITIONS

Here we give only the general scheme of pseudoperturbiteration methods. For the case of discrete group symmetry the eigenelements and generalized Jordan chains for direct and adjoint eigenvalue problems (1) and considered generalizations can be restored by the group action on relevant basis magnitudes in generating trajectories spaces. Analogously in the case of continuous symmetry these magnitudes can be restored by the action of infinitesimal operators of Lie algebra on the basis of magnitudes in generating trajectories subspace [42]. The simplest example is given by the rotation operator $s\frac{\partial}{\partial t} - t\frac{\partial}{\partial s}$ on the eigenspace $N(B) = \{s^2, st, t^2\}$, where the generating subspaces is formed by the basis elements s^2 and t^2. In the monograph [42] another examples in capillary-gravity surface waves or phase transitions in statistical crystal theory can be found. Such examples for critical imaginary eigenvalues at Poincarè-Andronov-Hopf bifurcation are contained in [43].

Therefore in order to apply PPI-methods it is sufficient to know approximations to basic elements of generating trajectories subspaces. Reproducing the other approximations by the group action in the case of discrete symmetry and the relevant infinitesimal operators in the case of continuous symmetry, we can obtain the complete approximate basis of eigenelements and generalized Jordan chains corresponding to multiple eigenvalue. Further we can construct the PP-operator having the relevant group symmetry with subsequent application of one iterational process of n. 3.

However we can not indicate substantial examples for illustration of PPI-method. The considered in [34, 35, 37, 40, 42, 43] examples of eigenvalue problems are solved exactly. We hope to find such examples on the base of the article [39].

Remark 3. Connections of PPI-methods with parameter continuation methods are remained for now on the level of the article [3].

References

[1] M. K. Gavurin, *On pseudoperturbation method for eigenvalues determination*, J. Comput. Math. Math. Phys. **1**, 5 (1961), 751-770, Russian, Engl. transl.

[2] F. Kuhnert, *Die Pseudoperturbation method*, Math. Forschungsberichte, **26** (1971), 1-119.

[3] B. V. Loginov, D. G. Rakhimov, N. A. Sidorov, *Development of M.K. Gavurin's pseudoperturbation method*, Fields Inst. Comm., **25** (2000), 367-381.

[4] B. V. Loginov, O. V. Makeeva, A. V. Tsyganov, *Refining of approximately given Jordan chains of linear operator-function of spectral parameter on the base of perturbation theory*, Interuniv. Proc. "Functional Analysis", Ulyanovsk Pedagogical University, **38** (2003), 53-62. (Russian)

[5] B. Karasözen, B. V. Loginov, V. A. Trenogin, *Pseudoperturbation method for sharpening of approximately given generalized Jordan chains*, Bul. Şt. Ser. Mat.-Inform., Univ. din Piteşti, Romania, **9** (2003), 183-190.

[6] I. V. Konopleva, B. V. Loginov, O. V. Makeeva, *Linearization on spectral parameter in branching theory*, Proc. Middle-Volga Math. Soc. **7**, 1 (2005), 105-113. (Russian)

[7] O. V. Makeeva, *On Jordan chains of polynomial operator-function of spectral parameter and its linearization*, Interuniv. Proc. "Functional Analysis", Ulyanovsk Pedagogical University, **39** (2005), 31-38. (Russian)

[8] B. V. Loginov, O. V. Makeeva, *The pseudoperturbation method in generalized eigenvalue problems*, Doklady Mathematics, **77**, 2 (2009), 194-197. Pleiades Publ.,Ltd.; Dokl. Akad. Nauk **419**, 2 (2009), 160-163.

[9] V. I. Shalashilin, E. B. Kuznetsov, *Parameter continuation method and the best parametrization* (in applied mathematics and mechanics), Editorial URSS, Moscow, Russian, Engl. transl.

[10] M. M. Vainberg, V. A. Trenogin, *Branching theory of solutions of nonlinear equations*, Nauka, Moscow, 1969; Engl. transl. Volters-Noordorf Int. Publ., Leyden, 1974.

[11] B. V. Loginov, Yu. B. Rousak, *Generalized Jordan structure in branching theory*, Direct and Inverse Problems for Partial Differential Equations, M.S. Salakhitdinov (ed), Akad. Nauk UzbekSSR, Fan, Tashkent, (1978), 133-148. (Russian)

[12] V. A. Trenogin, *Functional analysis*, Nauka, Moskow; Novosibirsk; Fizmatlit, Moscow (1980, 1999, 2002). Russian; French. transl. as Analyse Functionnelle, Mir, Moscow, 1985.

[13] A. Varhelyi, *On the improved Newton method for the solving nonlinear real equation*, Ann. Univ. Sci. Budapest, Sec. Computator, **3** (1982), 85-91.

[14] A. O. Kuznetsov, *On modification of Newton method and its usage for the sharpening of eigenvalues and eigenvectors of linear operators*, Mixed Type Equations and Free Boundary Value Problems, M.S. Salakhitdinov, T.D. Džhuraev (eds), Fan, Tashkent, 1987, 196-201. (Russian)

[15] A. M. Ostrovski, *Solution of equations and systems of equations*, Academic Press, New York; Russian transl., IL, Moscow, 1963.

[16] V. L. Zaguskin, *Handbook on numerical methods for the solving of equations*, Fizmatgiz, Moscow, 1960.

[17] J. M. Ortega, W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York; Russian transl., Mir, Moscow, 1975.

[18] B. V. Loginov, N. A. Sidorov, *Computation of eigenvalues and eigenelements of bounded operators by pseudoperturbation method*, Matem. Zametki, **19**, 1 (1976), 105-108; Engl. transl.

[19] B. V. Loginov, D. G. Rakhimov, *On the refining of eigenvalues and eigenvectors of analytic operator-function by pseudoperturbation method*, Izvestiya Akad. Nauk UzbekSSR, Fiz.-Mat., **1** (1977), 12-20. (Russian)

[20] N. A. Sidorov, V. A. Trenogin, *On the approach to regularization problem on the base of linear operators perturbation*, Matem. Zametki, **20**, 5 (1976), 747-752, Russian; Engl. transl.

[21] N. A. Sidorov, *General questions of regularization in branching theory problems*, Irkutsk University, 1982.

[22] O. V. Makeeva, *Pseudoperturbation method application to one-dimensional Bitsadze-Samarskii eigenvalue problem*, Vestnik of Ulyanovsk State Pedagogical University, **2** (2006), 58-61. (Russian)

[23] V. R. Kim-Tyan, B. V. Loginov, O. V. Makeeva, *One-dimensional boundary value problem with two displacements and pseudoperturbation method*, ROMAI Journal, **3**, 2 (2007), 199-212.

[24] B. V. Loginov, O. V. Makeeva, *On one spectral problem of E. Schmidt with displacements in boundary conditions*, Vestnik of Samara State University, **9** (2006), 14-18. (Russian)

[25] B. V. Loginov, O. V. Makeeva, E. V. Foliadova, *Sharpening of critical spectral parameter at dynamic bifurcation by pseudoperturbation method*, ROMAI Journal, **2**, 1 (2006), 119-126.

[26] E. Schmidt, *Zur Theorie linearen und nichtlinearen Integralgleichungen*, Teils 1-3, Mathematische Annallen, **65** (1908), 370-399.

[27] M. Sh. Mogilevsky, *On the representation of completely continuous operators in abstract separable Hilbert space*, Izvestiya VUZ, Mathematics, **3**, 4 (1958), 183-186. (Russian)

[28] B. V. Loginov, O. V. Makeeva, E. V. Foliadova, *On the sharpening of approximately given generalized Schmidt's eigenvalues of linear operators by pseudoperturbation method*, Interuniv. Proc. "Functional Analysis", Ulyanovsk Pedagogical University, **39** (2005), 21-30. (Russian)

[29] B. M. Levitan, I. S. Sargsyan, *Introduction to spectral Theory*, Nauka, Moscow, 1970.

[30] O. V. Makeeva, *Pseudoperturbation method for the determination of spectral parameter critical value in abstract Dirac type systems at dynamic bifurcation*, Interuniv. Proc. "Functional Analysis", Ulyanovsk Pedagogical University, **39** (2005), 44-52. (Russian)

[31] B. V. Loginov, O. V. Makeeva, *E. Schmidt spectral problem on eigenoscillations of lossless resonator*, Proc. Middle-Volga Mathem. Soc., **9**, 1 (2007), 31-38.

[32] I. S. Aržanykh, *Integral equations of basic problems of field theory and elasticity theory*, Akad. Nauk UzbekSSR, Tashkent, 1954.

[33] I. S. Aržanykh, *Invertibility of wave operators*, Fan, Akad. Nauk UzbekSSR, Tashkent, 1962.

[34] L. A. Vainstein, *Electromagnetic waves*, Radio and Communication, Moscow, 1989.

[35] V. I. Vol'man, Yu. V. Pimenov, *Technical electrodynamics*, Communication, Moscow, 1971.

[36] A. S. Il'ynskii, G. Ya. Slepyan, *Oscillations and waves in electrodynamical systems with losses*, Moscow State University, 1983.

[37] V. V. Nikol'skii, *Electrodynamics and radiowaves propagation*, Moscow, Nauka, 1974.

[38] A. A. Vlasov, *Macroscopic electrodynamics*, GTI, Moscow, 1955.

[39] G. Ja. Slepyan, *To design of electrodynamic oscillations for bodies of revolution*, J. Comput. Math. Math. Phys. **17**, 3 (1977), 776-790.

[40] N. A. Sapogova, V. M. Kantorovich, *Group theory application to investigation of degeneration removal in spherical resonators*, Izvestya VUZ, Radiophysics, **14**, 12 (1971), 1869-1877.

[41] B. V. Loginov, O. V. Makeeva, *On the application of pseudoperturbation method to the solving of algebraic equations*, Proc. Middle-Volga Math. Soc. **10**, 1 (2008), 315-320.

[42] B. V. Loginov, *Branching theory of solutions of nonlinear equation under group symmetry conditions.* Tashkent. Fan, Acad. Sci. Uzbek SSR, 1985.

[43] B. V. Loginov, I. V. Konopleva, Yu. B. Rousak, *Symmetry and potentiality in general problem of branching theory*, Izvestiya VUZ, Mathematics, 4, **527** (2006), 30-40.

ROMAI J., 4, 1(2008), 169–183

ON THE RECONSTRUCTION OF THE PHASE SPACE OF A DYNAMICAL SYSTEM

Maria Mădălina Mazilu, Stephane Azou, Alexandru Şerbănescu

Military Technical Academy, Bucharest; Université de Bretagne Occidentale, France; Military Technical Academy, Bucharest

madalinamazilu@yahoo.com; azou@univ-brest.fr; serbal@mta.ro

Abstract The study of time series with the aim of trying to characterize the underlying dynamic system (or data generating process) is a field of study of great interest in chaos based communications, namely spread spectrum communications with chaotic spreading code. In order to perform any kind of analysis upon a data series first there should be correctly reconstructed the phase space with the key parameters time delay and embedding dimension. In this paper several methods to determine these parameters are presented and their applicability in communications is discussed. Also there is taken into account the influence of noise as the spread spectrum communication have the power spectra lower than the spectra of noise. Also the modulation of informational symbols influence the dynamics of the data, therefore the parameters of interest are much more difficult to determine.

Keywords: time series analysis, chaos based communications.

2000 MSC: 34C20, 68P30.

1. INTRODUCTION

The concept of chaos is one of the most exciting and rapidly expanding research topics of recent decades. Ordinarily, chaos is disorder or confusion. In the scientific sense, chaos does involve some disarray, but there is much more to it than that. The states of natural or technical systems typically change in time, sometimes in a rather intricate manner. The study of such complex dynamics is an important task in numerous scientific disciplines and their applications as well in communications. Usually the aim is to find mathematical models which can be adapted to the real processes. In the last two decades there were developed nonlinear methods for data analysis that are based on a metric or topological analysis of the phase space of the underlying dynamics or an appropriate reconstruction of it. The subject of discussion in this paper focuses on phase space reconstruction.

Because chaos is very sensitive to initial conditions and chaotic sequences possess a noise-like wide-spectrum characteristic, yet they can be reproduced there very suitable as spreading code in Direct Sequence Spread Spectrum (CD3S) technique. Also, when using chaotic systems, there is no need to assume the randomness, since when observed in a coarse-grained state-space they do not behave randomly, but they do in a long run. Two chaotic trajectories diverge quickly, therefore chaotic systems provide by their nature a large sequence of uncorrelated sequences [13].

Due to several of such special features, including the important diffusion, confusion and mixing (ergodicity) properties [2], chaos-based communication systems can demonstrate some superiority over the conventional DSSS codes.

The proposed structure of a CD3S transmitter [3] :

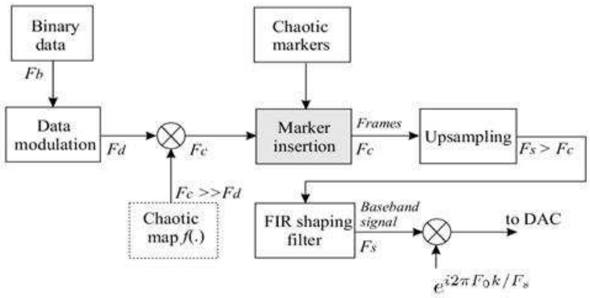


Fig. 1. Structure of a DSSS transmitter

Direct-sequence spectra vary somewhat in spectral shape, depending on the actual carrier and data modulation used. In fig. 1 is employed a binary phase shift keyed (BPSK) signal, which is the most common modulation type used in direct sequence systems. Afterward the spectra is spread along a much larger bandwidth and upsampled in order to obtain a sampled signal with a "continuous" dynamics.

In this paper the possibility to reconstruct the phase space of CD3S received signal is investigated. In the next section the methods employed with examples for a classical dynamical system, Lorenz, are presented and, afterward, the methods are applied to a modulated signal.

2. METHOD OF DELAYS

Systems of very different kinds, from very large to very small time-space scales can be modeled mathematically by (deterministic) differential equations.

Let the possible states of a system be represented by points in a finite-dimensional phase space, $\Re^m$. The transition form a system state from t_1 at time t_2 is a deterministic rule $T_{t_2-t_1}$[11]. For **continuous-time dynamics** with variables $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_m(t)]$, there can be defined a set of m ordinary differential equations

$$\dot{\mathbf{x}}(t) = \frac{\partial \mathbf{x}(t)}{\partial t} = \mathbf{G}(x) \tag{1}$$

where the vector field $\mathbf{G}(x)$ is always taken to be continuous in its variables and also taken to be differentiable as often as needed. The family of transition rules T_t, or its realization in the forms (1), are referred to as a *dynamical system.* Formally, a dynamical system is given by $\mathbf{x}(t)$ in a m dimensional space, which is called the *phase space*, for continuous time and a time evolution law (1).

Usually the observation of a real world process does not allow all possible state variables. Either not all state variables are known or not all of them can be measured. Most often only one observation is available

$$u(t) = h(x(t)) \tag{2}$$

where $h(\cdot)$ is the *measurement function.* Since the measurement results in discrete time series, the observations will be referred to as $u(k) = x(t_0 + kT)$, T being the sampling rate of the measurement.

As stated before, trying to reconstruct the state space of a system from information on scalar measurements $u(k) = x(t_0 + kT)$ is equivalent to finding a connection between the derivatives and the states variables, namely the differential equations which produced the observations. This is named the method of *derivatives coordinates.* Since this methods employs high order coordinates, the presence of noise can affect the reconstruction and therefore this method is not very useful for experimental data, so the mostly approached method is the *method of delays* (MOD) which is the point of the further discussion.

This method was first introduced into dynamical systems theory independently by Packard *et al* [8],and by David Ruelle and Takens [12] and it consists in the fact there really is no need of derivatives to form the system coordinates in which to capture the structure of orbits in phase space, but that there can be used directly the lagged samples.

The more rigorous formulation of this theorem is given as: consider K a smooth $\left(C^2\right)$ m-dimensional manifold that constitutes the original state space of the dynamical system under investigation and let $\phi^t : K \to K$ be the corresponding flow. Suppose that one can measure some scalar quantity $u\left(t\right) = h\left(x\left(t\right)\right)$ that a given by the measurement function $h : K \to \Re$, where $u\left(t\right) = \phi^t\left(u\left(0\right)\right).$ Then one may reconstruct a *delay coordinates map*:

$$F : K \to \Re^m$$

$$u \to \mathbf{y}(k) = [u(k),\, u(k+T),\, u(k+2T), ..., u(k+(m-1)J] \tag{3}$$

$$k = \overline{1, M} M = N-(m-1)J$$

that maps a state u from the original state space K to a point $\mathbf{y}$ in the *reconstructed state space* $\Re^m$, where m is the embedding dimension, J gives the time lag to be used [9], and N gives the number of measurements.

The MOD reconstructs the attractor dynamics by using delay coordinates to form multiple state-space vectors, $\mathbf{y}\left(\mathbf{i}\right)$. That is the reconstructed trajectory, $\mathbf{Y}$, is given by

$$\mathbf{Y} = \left[\mathbf{y}\left(\mathbf{1}\right)\, \mathbf{y}\left(\mathbf{2}\right) \ldots \mathbf{y}\left(\mathbf{M}\right)\right]^{\mathbf{T}}. \tag{4}$$

If the time series represents a continuous flow with samples taken every Δt seconds, then the delay time τ is the time period between successive of each of the embedding space vector. τ is considered an integer multiple of the sampling period and it can be expressed as $\tau = J \cdot T$. Here there should be stressed the difference between time delay τ and time lag J.

This is possible because in nonlinear systems the multiple dynamical variables interact with one another [1]. This was demonstrated numerically by Packard *et al* [8] and was proved by Takens [12]. The method of delays is the most widespread approach because it is the most straightforward and the noise level is constant for each delay component. Still this methods has a major drawback: the quality of reconstruction depends upon choosing the appropriate embedding parameters, the dimension m and the time delay J.

2.1. ESTIMATING TIME DELAY

As suggested, when employing the MOD, the only parameter that affects the quality of reconstruction is the embedding dimension. Also there should be stressed that it assumes the availability of an infinite amount of noiseless data. Instead, in real experiments, finite noisy data sets are used, therefore proper care should be taken when choosing the time delay (lag).

In estimating the value of τ many methods were considered but none is universal and when selecting it, two major problems should be considered: *redundancy* and *irrelevance*, as called by Casdagli *et al.* in *[5]*. In order to better exemplify the presented phenomena the Lorenz attractor in fig. 2 is considered.

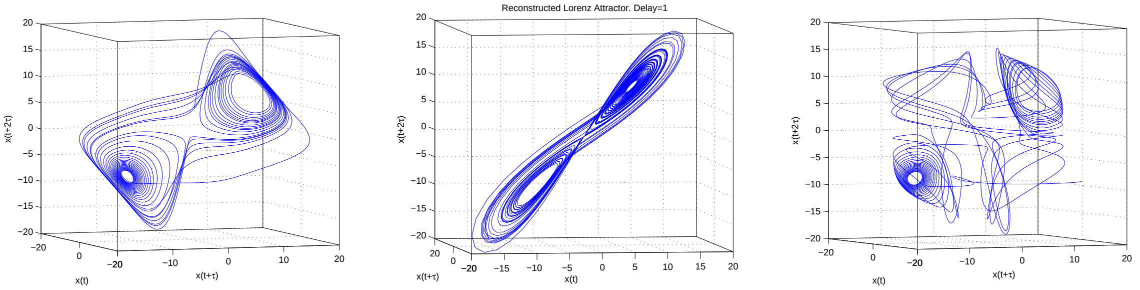


Fig. 2. Reconstructed Lorenz attractor with different time delays. Noiseless data set. $\sigma = 10$, $R = 28$, $b = \frac{8}{3}$, sampling period $T = 0.01$, number of samples $N = 3500$, $(a)\, \tau = 17T$, $(b)\, \tau = T$, $(c)\, \tau = 51T$.

The first one is related to the choice of τ as small as possible. In this case the consecutive measurements of the reconstructed vectors will give nearly the same results. Hence, the topological vectors constructed via the method of delays, will be stretched along the main diagonal, or line of identity, in the m-dimensional embedding space, leading to a resemblance for the reconstructed attractor with dimension close to one and thus the analysis of the picture of the attractor will be very difficult.

The second problem, irrelevance, appears with the choice of τ too large. In this case the reconstructed vectors become totally uncorrelated and the extraction of any information from this phase space picture becomes impossible, therefore resulting a random distribution of points in the embedding space.

Considering all mentioned above, τ should make independent each component in the reconstructed vector.

Ideally a method to estimate the delay should be computationally efficient, work well with noisy data and lead to consistent, accurate estimate of key descriptors of the original attractor.

Average displacement method. There are several criteria for the selection of the time delay τ, the main of them being the autocorrelation method. The main disadvantage of this methods is that it gives reasonable time delay only for two dimensional systems and in the literature are presented several criteria that were proved to be very sensitive to the employed dynamical system. Another popular method is the first local minimum of the delayed mutual information. Like the autocorrelation method, this approach works well for low dimensional dynamical systems and, in addition, it is sensitive to the number of bins employed for the segmentation of coordinates. Therefore there should be considered a method that offers different results for different number of dimensions. The investigated method, named average displacement, was introduced by Rosenstein and his colleagues [10] and quantifies the expansion of the attractor from the main diagonal. It is intended to overcome the drawbacks of autocorrelation.

For various embedding dimensions the *Average Displacement Method* (AD) can be computed as a function of τ such that

$$\langle S_m(J)\rangle = \frac{1}{M}\sum_{i=1}^{M}\|Y_i^{\tau} - Y_i\|, \tag{5}$$

where the upper scripts denote the delay between successive embedding components[10]. If they are used, the scalar time series (5) can be rewritten as

$$\langle S_m(J)\rangle = \frac{1}{M}\sum_{i=1}^{M}\sqrt{\sum_{j=1}^{m-1}\left[x_{i+jJ} - x_i\right]^2}. \tag{6}$$

The average displacement is useful for quantifying the decrease in redundancy error with increasing τ. As the time delay increases from zero the average displacement increases accordingly until it reaches a plateau. When increasing the value of m, the time until $\langle S_m\rangle$ reaches a plateau is shorter, therefore a constant embedding window can be maintained.

The two dimensional phase portrait for the Lorenz time series with the time delay generated by the average displacement method can be seen in fig.3.

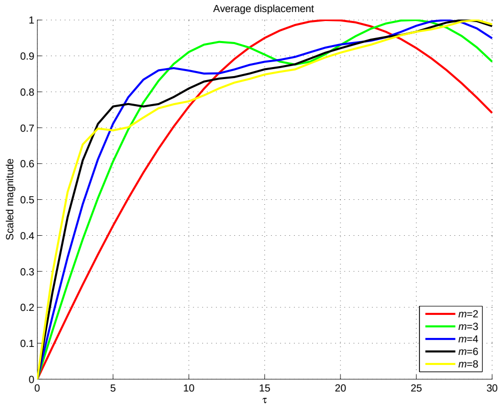


Fig. 3. Average displacement for Lorenz dynamical system and 2D attractor reconstruction with the provided time delay. $\sigma = 10$, $R = 28$, $b = \frac{8}{3}$. Initial conditions $x_0 = \frac{1}{10}$, $y_0 = -\frac{1}{5}$, $z_0 = \frac{3}{10}$, sampling period $T = 0.02$, number of samples $N = 3500$. (b) $m = 3$, $\tau = 9T$.

Based on empirical results, the authors suggested choosing time delay, "*as the point where the slope first decreased to less than 40% of its initial value*" [10]. Still when the wave shapes reach a certain saturation there are some waviness, and using the changing slope to determine the time delay may introduce some errors.

There can be seen an anomalous jump from $J = 0$ to $J = 1$, due to the fact that when using a zero time delay the phase-space vector is aligned with the line of identity.

Multiple correlation method. Multiple autocorrelation approach is derived from autocorrelation and *Average Displacement Method.* Considering (5), the square average displacement for a chaotic time series, $\{x_i\}$ in an m-dimensional space can be written as

$$\left\langle S_m^2 \left(J\right)\right\rangle = \frac{1}{M} \sum_{i=1}^{M} \|y_i^\tau - y_i\|^2, \tag{7}$$

such that (6) becomes

$$\left\langle S_m^2 \left(J\right)\right\rangle = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{m-1} \left[x_{i+jJ} - x_i\right]^2. \tag{8}$$

Before deriving a relationship between $\left\langle S_m^2 \left(J\right)\right\rangle$and the autocorrelation, we note that the expression for a finite data set for autocorrelation is given by

$$R_{xx}\left(J\right) \approx \frac{1}{N-J} \sum_{i=1}^{N-J} x_i \cdot x_{i+J}. \tag{9}$$

Extending the left-hand side of (8) and ignoring the errors caused by the border data we obtain

$$\left\langle S_m^2 \left(J\right)\right\rangle = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{m-1} \left[x_{i+jJ}^2 - 2x_i \cdot x_{i+jJ} + x_i^2\right] = 2\left(m-1\right)E{-}2\sum_{j=1}^{m-1} R_{xx}\left(jJ\right), \tag{10}$$

where $E = \frac{1}{M}\sum_{i=1}^{M} x_i^2 = \frac{1}{M}\sum_{i=1}^{M} x_{i+jJ}^2$, for $1 \leq j \leq m-1$ and $R_{xx}^m\left(J\right) = \sum_{j=1}^{m-1} R_{xx}\left(jJ\right).$

As it was considered by the authors of [7], the multiple autocorrelation delay can be described as: select the corresponding time as the time delay τ when the value of $R_{xx}^m\left(J\right)$ decreased to the $1/e^{-1}$ times of its initial value.

Moreover, when replacing the autocorrelation method with the non bias multiple autocorrelation we have

$$C_{xx}^{m}\left(J\right)=\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{m-1}\left(x_{i}-\overline{x}\right)\left(x_{i+jJ}-\overline{x}\right)=R_{xx}^{m}\left(J\right)-\left(m-1\right)\left(\overline{x}\right)^{2}. \quad (11)$$

In fig. 4 the results for x variable of Lorenz system perturbed with noise are illustrated.

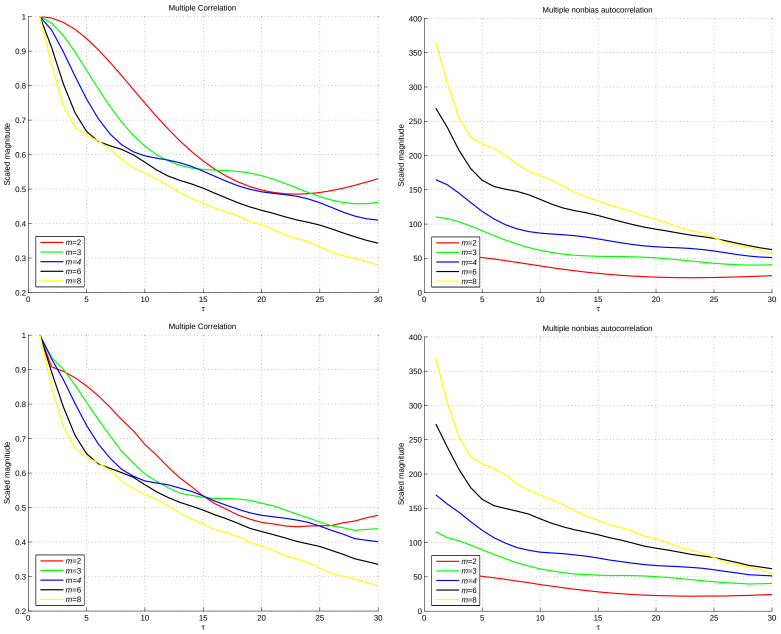


Fig. 4. Multiple autocorrelation for Lorenz dynamical system. $\sigma = 10$, $R = 28$, $b = \frac{8}{3}$. Initial conditions $x_0 = \frac{1}{10}$, $y_0 = -\frac{1}{5}$, $z_0 = \frac{3}{10}$, sampling period $T = 0.02$, number of samples $N = 3500$.

This algorithm can be regarded as an extension of the autocorrelation approach in the high order dimension.

2.2. ESTIMATING EMBEDDING DIMENSION

Mańé and Takens theorem guarantees that the reconstructed dynamics, if properly embedded, is equivalent to the dynamics of the true, underlying system in the sense that dynamical invariants, such as generalized dimensions and the Lyapunov spectrum, are identical.

There were different methods proposed for determining the minimum embedding dimension, but the most commonly used is the geometrical approach, false nearest neighbors (FNN). The idea of this method is quite intuitive. Suppose that the minimal embedding dimension for a given time series $\{x_i\}$ is m_0. Suppose that one embeds the time series in an m-dimensional space with $m < m_0$. Due to this projection the topological structure is no longer preserved. Points are projected into neighborhoods of other points to which they would not belong in higher dimensions. These points are called false neighbors. If now the dynamics is applied, these false neighbors are not typically mapped into the image of the neighborhood, but somewhere else, so that the average "diameter" becomes quite large.

The idea of the algorithm false nearest neighbors is the following: since each point y_i from the reconstructed vector as in (2) has a nearest neighbor y_j with the nearness in the sense of some distance function (in the simulations the Euclidean distance is used), then

$$\begin{aligned} R_m\left(i\right)^2 = [&\left(x_i - x_{n(i,m)}\right)^2 + \left(x_{i+J} - x_{n(i,m)+J}\right)^2 + \dots \\ &+ \left(x_{i+(m-1)J} - x_{n(i,m)+(m-1)J}\right)^2], \end{aligned} \tag{12}$$

where $R_m\left(i\right)$ is rather small when there is a large amount of data and can be approximated by $1/N^{\frac{1}{m}}$, where N represents the number of samples.

In the dimension $m+1$ the nearest neighbor distance is changed due to the $(m+1)^{st}$ coordinates x_{i+mJ} and $x_{n(i,m)+mJ}$ and R_{m+1} can be computed as

$$R_{m+1}^2\left(i\right) = R_m^2\left(i\right) + \left[x_{i+mJ} - x_{n(i,m)+mJ}\right]^2 . \tag{13}$$

If $R_{m+1}\left(i,j\right)$ is too large, there can be presumed that this happens because the near neighborliness of the two points being compared is due to the projection from some higher dimensional attractor down to dimension m. Some threshold is required to decide when neighbors are false. Then, if

$$\frac{\left|x_{i+mJ} - x_{n(i,m)+mJ}\right|}{R_m\left(i\right)} > R_T, \tag{14}$$

the nearest neighbor at time point j are declared false. In practice for values of R_T in the range $10 \le R_T \le 50$ the number of false neighbors identified by this criterion is constant.

For the Lorenz attractor, the embedding dimension, where the percentage of false neighbors drops to zeros, is 3, while the sufficient condition from the embedding Takens theorem is 5.

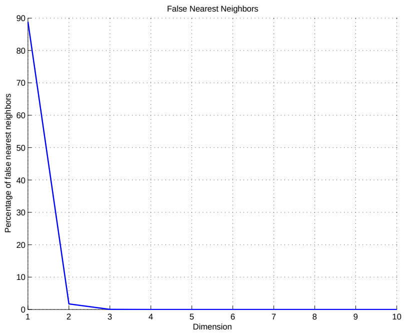


Fig. 5. Percentage of false nearest neighbors as a function of embedding dimension for clean Lorenz data. $\sigma = 10$, $R = 28$, $b = \frac{8}{3}$. Initial conditions $x_0 = \frac{1}{10}$, $y_0 = -\frac{1}{5}$, $z_0 = \frac{3}{10}$, sampling period $T = 0.02$, time lag $J = 9T$, number of samples $N = 3500$.

The so called flaw of this method is that it involves the time lag, and without a proper time delay J, the embedding dimension can not be estimated accurately.

2.3. EMBEDDING WINDOW

The embedding window is defined as the length of the interval spanned by the first and last delay coordinate

$$\tau_w = \tau \cdot (m-1)\,. \tag{15}$$

The parameter is thoroughly discussed in [6] and it is shown to be more useful, since it determines the amount of information passed form the time series to the embedding vectors, than the dependent parameters: m the embedding dimension and τ time delay.

There are several methods employed in order to establish the time window and the most of them are related to the saturation of the correlation dimension.

As shown in [4, 6] the determination of minimal necessary embedding dimension is strongly dependent upon the choice of the time delay. As one expects the embedding dimension decreases when increasing the time delay, as the delayed coordinates are more and more independent. As a consequence, a wise solution would be to apply the false nearest neighbor method, but considering proper time delays for each dimension. Therefore, the time delay should be established for each $d = 2, 3, ...$ with one of the prior learned approaches: average displacement, multiple autocorrelation or multiple non-bias autocorrelation.

3. EXPERIMENTAL RESULTS

A particularly interesting candidate for **discrete time** chaotic sequences generators is the family of Chebyshev polynomials, whose chaoticity can be verified easily and many other properties of which are accessible to a rigorous mathematical analysis. The independent binary sequences generated by a chaotic Chebyshev map were shown to be not significantly different from random binary sequences. For this reason, a k^{th}-order Chebyshev map is employed in the following discussion for masking the information. This map is defined by

$$\begin{aligned} x_{k+1} = F\left(x_k\right) &= T_p\left(x_k\right), \\ T_0 &= 1, \\ T_1 &= x, \\ T_{p+1}\left(x\right) &= 2xT_p\left(x\right) - T_p\left(x\right), \end{aligned} \tag{16}$$

p being the order of the polynomial.

Consider a data signal

$$D = [1, -1, 1, 1, -1, 1, 1, 1, -1, 1, 1, -1, -1, 1, 1, 1, 1, 1, -1, -1, -1, 1, 1]. \tag{17}$$

After the spreading process, a raised cosine shaping filter with the roll off factor $\alpha = 0.3$, that introduces a delay of 3 samples, is applied. Considering the sampling frequency $F_e = 5F_c$, after this shaping the signal will have a more smooth dynamics suitable to methods that require a continuous time systems (smooth dynamics). This is equivalent to interpolating with a rate $R = 5$. These algorithms can be applied to the data as they offer a quite good estimation for the time lag when discussing continuous data. However, when considering discrete chaotic time series, they cannot process it accurately, because the sampling spacing of the discrete time series is too large, therefore the maps behave like random series.

The delay introduced by the filter was discarded.

Trying to reconstruct the phase space of the signal the following results for the time delay were obtained

Table 1: Estimated time lag for the filtered spread spectrum sequence with spreading sequence 2^{nd} order Chebyshev polynomial

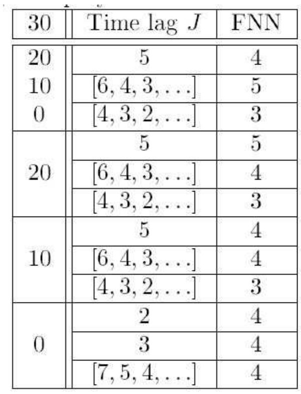
	Processing gain $G = 31$			Processing gain $G = 63$			Processing gain $G = 1024$		
Method	$m = 2$	$m = 3$	$m \geq 4$	$m = 2$	$m = 3$	$m \geq 4$	$m = 2$	$m = 3$	$m \geq 4$
Average Displacement	6	4	3	6	4	3	6	4	3
Multiple Autocorrelation	4	3	2	4	3	2	4	3	2
Multiple non bias Autocorrelation	4	3	2	4	3	2	4	3	2

For the estimation of the embedding dimension we did not consider the whole data set as both implemented methods require a great computational cost; instead an amount of $N = 1500$ was considered, and the segment of data was chosen randomly. Taking into account the time lags estimated before and the previous remarks, the minimum embedding dimension was determined to be $m = 3$.

From the previous work we saw that trying to reconstruct the phase space in the presence of noise can be troubling especially in the case of a low signal to noise ratio which is the case in the most situations that deal with CD3S. Modulation, as proved before, makes even more difficult. Anyway, knowing that the expected signal should be low dimensional, one could guide oneself when choosing the appropriate m. Also there should be kept in mind that a too high embedding dimension can cause spurious correlation. Again for a low Signal to Noise Ratio (SNR) the dimension will be inflated. Therefore a compromise should be made.

Table 2: Estimated time lag for noisy CD3S with spreading sequence 2nd order Chebyshev polynomial

Processing gain $G = 31$	SNR=30 dB			SNR=20 dB			SNR=10 dB			SNR=0 dB			
Method	m=2	m=3	m≥4	m=2	m=3	m≥4	m=2	m=3	m≥4	m=2	m=3	m≥4	m≥5
Average Displacement	6	4	3	6	4	3	6	4	3	7	5	4	3
Multiple Autocorrelation	4	3	2	4	3	2	4	3	2	2			
Multiple non bias Autocorrelation	4	3	2	4	3	2	4	3	2	2			



30	Time lag J	FNN
20	5	4
10	$[6,4,3,\ldots]$	5
0	$[4,3,2,\ldots]$	3
20	5	5
	$[6,4,3,\ldots]$	4
	$[4,3,2,\ldots]$	3
10	5	4
	$[6,4,3,\ldots]$	4
	$[4,3,2,\ldots]$	3
0	2	4
	3	4
	$[7,5,4,\ldots]$	4

The table above presents the minimum embedding dimension estimated by False Nearest Neighbor method for noisy CD3S spread with spreading sequence 2^{nd} order Chebyshev polynomial

From the previous tables one can draw the conclusion that the *Average Displacement Method* yields the most reliable results when estimating the time delay, even in noisy environment for all investigated systems. We allege that as until a $SNR = 0\,dB$ these two methods are quite robust to noise (white Gaussian noise). Still there cannot be designated one method superior to another, because in order to draw such a conclusion one should investigate more replications of systems of interest with different data length.

As far as the embedding dimension is concerned, the results coincide with expectations, meaning that as the signal to noise ratio decreases the embedding dimension increases. Also both methods show dependence upon the choice of time lag, and only in case of the an appropriate J the embedding dimension will be as expected.

4. CONCLUSIONS

In this paper several methods for reconstructing the phase space of a dynamical system from the measurements of one variable were presented. Even though these methods were largely used in other search domains as finances (critical regime in financial indices) and medical engineering (cardiology, neurology etc.), these methods find their applicability in spread spectrum communications, were the modulation alters the natural dynamics of data. The presence of noise was considered too. There can be seen that the results conveyed are different from noiseless data: the time delay decreases as the data becomes more and more uncorrelated and the embedding dimension increases as the noise tends to fill itself the phase space.

References

[1] H. D. I. Abarbanel, R. Brown, J. J. Sidorovich, L. Sh. Tsimring, *The analysis of observed chaotic data in physical systems*, Rev. Mod. Phys., **65**, 4(1993), 1331-1392.

[2] G. Alvarez, L. I. Shujun, *Some basic cryptographic requirements for chaos-based cryptosystems*, Int. J. of Bifurcation and Chaos, **16** (2006), 2129-2151.

[3] S. Azou, G. Burel, C. Pistre, *A chaotic direct-sequence spread spectrum system for underwater communication*, IEEE-Oceans'02, Biloxi, MS, USA, 2002.

[4] L. Cao, *Practical method for determining the minimum embedding dimension of a scalar time series*, Phys. D, **110**, 1-2(1997), 43-50.

[5] M. Casdagli, S. Eubank, J. Doyne Farmer, J. Gibson, *State space reconstruction in the presence of noise*, Phys. D, **51**, 1-3(1991), 52-98, 1991.

[6] D. Kugiumtzis, *State space reconstruction parameters in the analysis of chaotic time series - the role of the time window length*, Physica D, **95**(1996), 13-28.

[7] H. - G. Ma, C. Z. Han, *Selection of embedding dimension and delay time in phase space reconstruction*, Frontiers of Electrical and Electronic Engineering in China, **1**(2006), 111-114.

[8] N. H. Packard, J. P. Crutchfield, J. D. Farmer, R. S. Shaw, *Geometry from a time series*, Phys. Rev. Lett., **45**, 9(1980), 712-716.

[9] U. Parlitz, *Nonlinear time-series analysis*, in Nonlinear Modeling - Advanced Black-Box Techniques, J.A.K. Suykens, J. Vandewalle (eds.), Kluwer, 1998.

[10] M. T. Rosenstein, J. J. Collins, C. J. De Luca, *Reconstruction expansion as a geometry-based framework for choosing proper delay times*, Physica D, **73**(1994), 82-98.

[11] T. Schreiber, *Interdisciplinary application of nonlinear time series methods*, Phys. Rep., **308**, 2(1999), 1-64.

[12] F. Takens, *Detecting strange attractors in turbulence*, Dynamical Systems and Turbulence, **898**(1981), 366-381.

[13] J. Tou, P. Yip, H. Leung, *Spread-spectrum signals and the chaotic logistic map*, Circuits, Systems, and Signal Processing, **18**(1999), 59-73.

ROMAI J., **4**, 1(2008), 185–192

MACROECONOMIC MODEL WITH RATIONAL EXPECTATIONS FOR REPUBLIC OF MOLDOVA

Elvira Naval

Institute of Mathematics and Computer Science of The Academy of Sciences of Moldova, Chişinău, Republic of Moldova

nvelvira@math.md

Abstract A small macroeconomic model [1] was adapted for Republic of Moldova and estimated using annual data. Three goods model with one domestically produced good consumed both at home and abroad and one imported good are examined. The aggregate demand, the aggregate supply, the money market and the government sector are considered. The consumption is the sum of real rate interest, disposable income, lagged consumption term and lagged disposable income term. Consumer disposable income is defined as *GDP* plus the earnings on net assets held abroad, minus interest paid on domestic debt and taxes. Investment is a linear function of the real interest rate, real output and the beginning-of-period capital stock. Export is a function of the real exchange rate, level of real output abroad, and a lagged export term. Real import is the function of the real exchange rate and real domestic output, a lagged import and a lagged reserve-import ratio.

Keywords: macroeconomic models.

2000 MSC: 91B60, 91B64.

A small macroeconomic model based on familiar theoretical considerations [1] was adapted to the national economy reality for Republic of Moldova and estimated using annual data. The goal of the researches consists in the estimation of a set of macroeconomic indicators supposed as a behavioral variable and in solving the obtained system of nonlinear equations and conduct some simulations. For these purposes, the macroeconomic model with widely-accepted developing-country specifications for the key behavioral relationships [1-2] is used .

Three goods model with one domestically produced good consumed both at home and abroad and one imported good is considered. The model is divided into aggregate demand, aggregate supply, the money market, foreign sector and the government sector.

Aggregate demand. Real aggregate demand for the internal output is considered to be equal to the sum of private consumption, investment, public consumption and net export

$$Y_t = Cp_t + Inv_t + Cg_t + X_t - \frac{e_t P_t^*}{P_t}. \tag{1}$$

The variables from equation (1) are defined as follows: Y_t is the real GDP; Cp_t are real expenditures on private consumption; Inv_t are real gross domestic investments; Cg_t is the real governmental consumption for domestic goods; X_t denotes real exports; e_t is the nominal exchange rate (the price of foreign currency in comparison with national currency); Z_t is real imports, measured in the units of the foreign goods; P_t^* is the import price in foreign currency; P_t is the price of internal output in national currency.

Private consumption, one of the demand components, is specified as behavioral equation in the following matter

$$lnC_t = \alpha_0 + \alpha_1 r_{t-1} + \alpha_2 lnC_{t-1} + \alpha_3 Y_t^d + \epsilon_1, \tag{2}$$

where r_t is real internal interest rate, Y_t^d is real disposable income and α_i are the coefficients that must be estimated.

Consumers disposable income is supposed to be GDP minus internal taxes

$$Y_t^d = Y_t - T_t, \tag{3}$$

where T_t are the real taxes (taxes incomes). On the other hand, the disposable income and consumption expenditures are correlated with the net modifications in the consumers' wealth through the budget restrictions to the private sector

$$Y_t^d = Cp_t + Inv_t + \{(M_t - M_{t-1}) + e_t \Delta Fp_t - (DCp_t - DCp_{t-1})\}/P_t, \tag{4}$$

where M_t denotes the money supply and ΔFp_t are changes in foreign private assets, measured in foreign currency. So the disposable income is allocated in consumption, investments and net changes in financial assets.

Investments, other demand component, are represented as behavioral function depending of the real output and one period lag investments

$$lnInv_t = k_0 + k_1 lnY_t + k_2 Inv_{t-1} + \epsilon_2. \tag{5}$$

The following demand component export is assumed to be a behavioral function of the real world output level (Y^*) and of the one period lag export term, both variables have positive coefficients,

$$lnX_t = \tau_0 + \tau_1 lnY_t^* + \tau_2 X_{t-1} + \epsilon_3. \tag{6}$$

At last, real imports are positively depending on the domestic real output. The term of one period lag import is included in estimated equation in order to obtain partial adjustment behavior. Moreover, because disposability of the foreign exchange often presents a restriction for states in transition, one period lag of reserves is frequently included in the regression for imports behavioral function. So, the imports equation may be written as

$$lnZ_t = \delta_0 + \delta_1 ln\frac{Y_t P_t}{P_t^* e_t} + \delta_2 ln\frac{Rez_{t-1}}{P_{t-1}^* Z_{t-1}} + \delta_3 lnZ_{t-1} + \epsilon_4. \tag{7}$$

Aggregate supply represents the GDP and is considered to be a function of Cobb-Douglas type depending on two production factors - labor and capital

$$Y_t = \theta_0 K_t^{\theta_1} L_t^{\theta_2}, \tag{8}$$

where K_t and L_t are the stocks of aggregate capital and labor, and θ_i, $i = 0, 1, 2$ are the coefficients which must be estimated. Since aggregate capital stock data possess some developing countries, estimation of supply implies a serious problem. Thus, in order to obtain aggregate capital stock data, similar to [1], the following procedure will be applied. The solution of the differential equation $K_t = (1-\rho)K_{t-1} + Inv_t$, where ρ is the depreciation rate, may be written as

$$lnK_t \approx ln2 + \frac{1}{2} ln \sum_{i=0}^{t-1} (1-\rho)^i Inv_{t-1} + \frac{t}{2} ln(1-\rho) + \frac{1}{2} K_0, \tag{9}$$

where K_0 is the initial capital stock. So,

$$lnY_t = ln\theta_0 + \theta_1 lnK_t + \theta_2 lnL_t = \theta_0^{'} + \theta_1 K_t^{'} + \theta_2 lnL_t, \tag{10}$$

where $\theta_0^{`} = \ln \theta_0 + \frac{\theta_1}{2} K_0$,

$$K_t^{`} = \ln 2 + \frac{1}{2} \ln \sum_{i=0}^{t-1} (1-\rho)^i Inv_{i-1} + \frac{t}{2} \ln(1-\rho). \tag{11}$$

Establishing constant returns to the scale ($\theta_1 + \theta_2 = 1$), dividing (9) by lnL_t we obtain

$$ln(\frac{Y_t}{L_t}) = \theta_0^{`} + \theta_1 (K_t^{`} - lnL_t). \tag{12}$$

Finally the specifications lagged variable $\ln(\frac{Y_{t-1}}{L_{t-1}})$ and a time trend t were introduced as additional variables. So, empirical production function takes the form

$$ln(\frac{Y_t}{L_t}) = \theta_0^{`} + \theta_1 (K_t^{`} - lnL_t) + g \cdot t + \theta_3 ln(\frac{Y_{t-1}}{L_{\ t-1}}). \tag{13}$$

In the assumption of the complete wage-price flexibility equation (13) represents the aggregate supply function of the economy.

Money market (M_t) in economy consists of reserves (Rez_t), internal credits (DC_t) and other components (Oth_t)

$$M_t = e_t Rez_t + DC_t + Oth_t. \tag{14}$$

Reserves are endogenously determined by the balance of payments, while internal credits accorded to private sector (DCp_t), and to public sector (DCg_t) are determined by the government policy:

$$DC_t = DCp_t + DCg_t. \tag{15}$$

Money demand is usually supposed to be positively correlated with the income level, while negatively related to nominal interest rate with the introduction of partial adjustment mechanism for obtaining lagged reactions

$$ln\frac{M_t}{P_t} = \beta_0 + \beta_1 lnY_t + \beta_2 i_t + \beta_3 lnY_{t-1} + \beta_4 ln\frac{M_{t-1}}{P_{t-1}} + \epsilon_6. \tag{16}$$

Foreign sector. Balance of payment identity is

$$e_t \Delta R_t = CA_t - e_t(\Delta Fg_t + \Delta Fp_t) - \Delta Oth_t, \tag{17}$$

where

$$CA_t = P_t X_t - e_t P_t^* Z_t + i_t^* e_t (Fp_{t-1} + Fg_{t-1} + R_{t-1}), \tag{18}$$

the authorities may obtain the accepted level of private capital flows, ΔFp_t conditioned by current account CA_t and public capital flows ΔFg_t.

Real interest rate is given by the equation

$$r_t = i_t - \frac{E_t P_{t+1} - P_t}{P_t}, \tag{19}$$

meaning that the real interest rate is the nominal rate minus expected inflation rate, where $E_t P_{t+1}$ is the expectation in year t of the $t+1$ year price.

Public sector. Dynamic specification of the model is completed with the description of the non-financial public sector behavior. The public sector borrows from external markets (ΔFg_t), as well as from the internal banking sector (ΔDCg_t). Its revenues consist of tax receipts and interest on foreign asset. Expenditures (Cg_t) consist of purchases of domestic goods for consumption purposes and interest payments on domestic debt. Combining these elements, the governmental budget restriction may be written as

$$e_t \Delta Fg_t - \Delta DCg_t = P_t(T_t - Cg_t) + i_t^* e_t Fg_{t-1} - i_t DCg_{t-1}. \tag{20}$$

Model equations estimation

Further the estimations of some model equations will be presented.

Private consumption

The equations were estimated with T.S.L.S., using as instruments exogenous variables and lagged values of exogenous and endogenous variables

$Cp_t = exp(-4.1197 - 0.5955 * r_{t-1} + 2.1792 * lnY^dt - 0.6904 * lnCp_{t-1})$

σ	(0.86)	(0.13)	(0.35)	0.28)
t	(-4.79)	(-4.48)	(6.18)	(-2.49)

$R^2 = 0.99$; $F = 112.22$; $DW = 2.21$.

All coefficients have the anticipated signs, conforming well to theory and estimations available in the literature. The coefficient of r_{t-1} is negative, showing an inverse dependence between private consumption and one lag term real interest rate. The short term semi elasticity is -0.59. Real interest rate influences consumption in the long-run with the elasticity of -0.35. Private consumption positively depends on disposable income, with a short-run elasticity of 2.18 and long-run elasticity of 1.29. The coefficient of the lagged consumption is -0.69, negative, the response to the change in the lagged interest rate and in the disposable income is not prolonged over time. All coefficients are significant and the variables connection is very strong. The values of the t-Student test point out the lack of multi collinearity. The h-D$urbin$ test is equal to $-0.41 > -1.96$, thus there is no residuals autocorrelation.

Exports

$$X_t = exp(-1.0639 + 0.1707 * lnY^* + 0.5993 * lnX_{t-1} - 0.1892 * DUM)$$

σ	(1.35)	(0.06)	(0.12)	(0.04)
t	(-0.79)	(3.05)	(5.20)	(-5.12)

$R^2 = 0.96$; $F = 41.21$; $DW = 2.07$.

The DUMMY variable takes value 1 in years 1998 and 2000. The estimated equation has all the coefficients significant and with anticipated signs. The export positively depends on the world GDP, with the short-run elasticity of 0.17. The coefficient of the lagged export is equal to 0.59, thus the effect of relative prices and world GDP modification will not be prolonged over time. All coefficients are significant for an $\alpha = 0.01$, the variables are strongly correlated. The test h-D$urbin$ has the value $0.013 < 1.96$ the hypothesis of residuals autocorrelation is rejected.

Imports

$$Z_t = exp(1.1775 + 0.5791 * ln\frac{Y_t P_t}{P_t^* e_t} + 0.0726 * ln\frac{Rez_{t-1}}{P_{t-1}^*} + 0.2661 * lnZ_{t-1})$$

σ	(0.43)	(0.07)	(0.06)	(0.01)
t	(2.76)	(8.10)	(4.42)	(9.12)

$R^2 = 0.99$; $F = 147.89$; $DW = 2.23$.

The import positively depends on real domestic GDP measured in foreign currency (USD), with a short-run elasticity of 0.58 and that of the long-run of 0.79. We may say that the GDP impact on the import is not significant. The reserve-import ratio influences the import positively and significantly, with an elasticity of 0.07 in the short-run and 0.09 in the long-run. The lagged import coefficient is 0.27, this caused the reduced effects of the variables on the import in the long-run. All the coefficients are significant for $\alpha = 0.01$, the variables are correlated very strong, confirmed by a big F-statistic value. The test h-D$urbin$ $= -0.33 < -1.96$; there is no residuals autocorrelation.

One scenario of macroeconomic development for Republic of Moldova (years 2005-2010)

This scenario is based on the forecasting calculations taking into account the following assumptions:

- Increasing of the world economy growth indexes.
- Maintaining of the slow dollar depreciation and euro appreciation on the world markets.
- Macroeconomic stability oriented state policy.
- Improvement of the enterprisers management.
- Relatively favorable climate conditions.

In the forecast the following values for exogenous and policy variables were selected:

- World GDP will increase annually with the average modification, calculated for years 1999-2005.
- World price values are extrapolated using O.S.L.S..
- Domestic nominal interest rate is extrapolated using average growth index and will represent 17%; 15%; 14%;12% and 11% in 2006-2010.
- Domestic currency exchange rate at the end of year 2010 is supposed to be 12 lei for 1 USA dollar, thus for years 2006-2009 it will register the following values 13.13; 12.85; 12.57 and 12.28 lei for 1 USA dollar.
- Governmental consumption is constant, equal to the year 2005 level: 1406.12 mil.lei in constant prices.
- Public credits increase with 3% per year.
- Private credits increase with 10% per year.
- Other monetary components were extrapolated using average modification for years 1995-2005.

The main goal of the forecast effectuated on the basis of the examined model consists in obtaining of the macroeconomic dynamics starting with specified exogenous variables and giving policy variables that assures model solution.

References

[1] Haque, N. U., Lahiri, K., Montiel P., *An econometric rational-expectations macroeconomic model for developing countries with capital controls*, Policy Research Department, Washington, DC, 1990.

[2] Khan, M. S., Montiel P., Haque N., *Adjustment with growth*, Journal of Development Economics 32, 1990.

[3] Pecican, E., Tanasoiu, O., Iacob, A. I., *Economic models*, Biblioteca Digitală, Bucureşti, 2000. (Romanian)

[4] Ţigănescu, E., Dobre, I., Roman, M., *Macroeconomics. Strategic decisions*, Editura ASE, 2000. (Romanian)

ROMAI J., 4, 1(2008), 193–201

CONDITIONS OF SINGLE NASH EQUILIBRIUM EXISTENCE IN THE INFORMATIONAL EXTENDED TWO-MATRIX GAMES

Ludmila Novac

Moldova State University, Chişinău, Republic of Moldova

novacliuda@yahoo.com

Abstract The informational aspect in game theory is manifested by: the devise of possession of information about strategy's choice, the payoff functions, the order of moves, and optimal principles of players; the using methods of possessed information in the strategy's choice by players. The player's possession of supplementary information about unfolding of the game can influence appreciably the player's payoffs.

Keywords: two-matrix games, Nash equilibrium.

2000 MSC: 91A40.

An important element for the players is the possession of information about the behaviour of his opponents. Thus for the same sets of strategies and the same payoff functions it is possible to obtain different results, if the players have supplementary information. So the information for the players about the strategy choice by the others players have a significant role for the unfolding of the game.

Consider the two-matrix game in the normal form $\Gamma = \langle N, X_1, X_2, A, B \rangle$, where , $A = \{a_{ij}\}$, $B = \{b_{ij}\}$, $i = \overline{1,m}$, $j = \overline{1,n}$ (A and B are the payoff matrices for the first and the second player respectively. Each player can choose one of his strategies and his purpose is to maximize his payoff. The player can choose his strategy independently of his opponent and the player does not know the chosen strategy of his opponent.

According to [1] we define the Nash equilibrium.

Definition 1. *The pair* (i^*,j^*), $i^* \in X_1, j^* \in X_2$ *is called the Nash equilibrium (NE) for the game* Γ, *if the relations*

$$\begin{cases} a_{i^*j^*} \geqslant a_{ij^*}, \forall i \in X_1, \\ b_{i^*j^*} \geqslant a_{i^*j}, \forall j \in X_2, \end{cases} \tag{1}$$

hold. Notation: $(i^*,j^*) \in NE\,(\Gamma)$.

There are two-matrix games for which the set of the Nash equilibria is empty: $NE\,(\Gamma) = \emptyset$ (solutions do not exist in pure strategies).

For every two-matrix game we can construct some informational extended games. If one of the players knows the strategy chosen by the other, we consider that it is one form of the informational extended two-matrix game for the initial game. Even if the initial two-matrix game has no solutions in pure strategies, for the informational extended games at least one solution in pure strategies always exists (Nash equilibria). The proof of this assertion can be found in [2], [3]. In the case of informational extended games the player which knows the chosen strategy of his opponent has one advantage and he will obtain one of his greater payoff.

According to [1], let us define two forms of informational extended games ${}_1\Gamma$ and ${}_2\Gamma$. We consider that for the game ${}_1\Gamma$ the first player knows the chosen strategy of the second player, and for the game ${}_2\Gamma$ the second player knows the chosen strategy of the first player.

If one of the players knows the chosen strategy of the other, then the set of the strategies for this player can be represented by a set of mappings defined on the set of strategies of his opponent.

Definition 2. *(The game* ${}_1\Gamma$ according to [1]*) The informational extended two-matrix game* ${}_1\Gamma$ *can be defined in the normal form by:* ${}_1\Gamma = \left\langle N, \overline{X_1}, X_2, \overline{A}, \overline{B} \right\rangle$, *where* $N = \{1,2\}$, $\overline{X_1} = \{\varphi_1 : X_2 \longrightarrow X_1\}$, $\overline{A} = \{\overline{a}_{ij}\}$, $\overline{B} = \left\{\overline{b}_{ij}\right\}$, $i = \overline{1,m^n}$, $j = \overline{1,n}$.

For the game ${}_1\Gamma$ we have $\overline{X_1} = \{1,2,\ldots,m^n\}$, $X_2 = \{1,2,\ldots,n\}$, $\left|\overline{X_1}\right| = m^n$, the matrices $\overline{A}$ and $\overline{B}$ have dimension $[m^n \times n]$ and they are formed from elements of initial matrices A and B respectively.

Choosing one element from each of the rows from the matrix A we will build one column in the matrix $\overline{A}$. The columns from the matrix $\overline{B}$ are built in the same manner, choosing one element from each of the rows from the matrix B. Thus, the matrices $\overline{A}$ and $\overline{B}$ have the dimension $[m^n \times n]$.

Definition 3. *(The game* ${}_2\Gamma$ according to [1]*) The informational extended two-matrix game* ${}_2\Gamma$ *can be defined in the normal form by:* ${}_2\Gamma = \left\langle N, X_1, \overline{X_2}, \widetilde{A}, \widetilde{B} \right\rangle$, *where* $\overline{X_2} = \{\varphi_2 : X_1 \longrightarrow X_2\}$, $\left|\overline{X_2}\right| = n^m$, $\widetilde{A} = \{\widetilde{a}_{ij}\}$, $\widetilde{B} = \left\{\widetilde{b}_{ij}\right\}$, $i = \overline{1,m}$, $j = \overline{1,n^m}$.

For the game ${}_2\Gamma$ we have $X_1 = \{1, 2, \ldots, m\}$, $\overline{X_2} = \{1, 2, \ldots, n^m\}$, the matrices $\widetilde{A}$ and $\widetilde{B}$ have dimension $[m \times n^m]$ and they are formed from elements of initial matrices A and B respectively.

The extended matrices $\widetilde{A}$ and $\widetilde{B}$ will be built in a similar way as in the case of the game ${}_1\Gamma$. That is each of rows in the matrices $\widetilde{A}$ (or in the matrix $\widetilde{B}$, respectively) will be built choosing one element from each of the columns from the matrix A (or B, respectively).

The following algorithm can be used for the determination of Nash equilibria in the informational extended two-matrix games ${}_1\Gamma$ and ${}_2\Gamma$.

Algorithm.

For the game ${}_1\Gamma$, we determine the maximum element in each column from the matrix A, i. e. $a_{i_j j} = \max\limits_i \{a_{1j}, a_{2j}, \ldots, a_{mj}\}$, for $\forall j = \overline{1,n}$.

For each element $a_{i_j j}$, $j = \overline{1,n}$ thus obtained, we determine the corresponding elements with the same indices from the matrix $B : b_{i_j j}$, $j = \overline{1,n}$.

For each of these pairs $a_{i_j j}, b_{i_j j}$, $\left(j = \overline{1,n}\right)$ we determine if these values can be the payoffs for players for some Nash equilibria.

Thus if $\forall k \in X_2 \backslash \{j\}$ $\exists b_{ik} : b_{ik} \leqslant b_{i_j j}$, then the pair $a_{i_j j}, b_{i_j j}$ can be the payoffs for players for some Nash equilibria in the game ${}_1\Gamma$; consider this pair $a_{i^* j^*}, b_{i^* j^*}$.

It is possible that for the pair $a_{i^* j^*}, b_{i^* j^*}$ there are several Nash equilibria.

If we wish to determine how many Nash equilibria there are in the game ${}_1\Gamma$ for the pair $a_{i^* j^*}, b_{i^* j^*}$ we determine the number of elements which are in each column $k \in X_2 \backslash \{j\}$ from the matrix B for which $b_{ik} \leqslant b_{i^* j^*}$. Denote by n_j, $j = \overline{1,n}$ the number of elements b_{ij} from the column j for which $b_{ij} \leqslant b_{i^* j^*}$. For j^* we have $n_{(j^*)} = 1$.

Then the number of Nash equilibria for which the players have the payoff $a_{i^*j^*}$ and $b_{i^*j^*}$, respectively, can be determined by

$$N_{j^*} = n_1 \cdot n_2 \cdot \ldots \cdot n_{(j^*-1)} \cdot 1 \cdot n_{(j^*+1)} \cdot \ldots n_n. \tag{2}$$

The number of all Nash equilibria in the game ${}_1\Gamma$ can be determined by: $N = \sum\limits_j N_j$.

If the pair of elements $a_{i_j j}, b_{i_j j}$ can be the payoffs of the players for some Nash equilibrium in the informational extended game ${}_1\Gamma$, then j will be the strategy for the second player. And because $\overline{X_1} \neq X_1$, we have to determine the strategy for the first player, for which the elements $a_{i^*j^*}, b_{i^*j^*}$ will correspond to one Nash equilibrium.

In this way we determine the elements $b_{i_1 1}, b_{i_2 2}, \ldots, b_{i_j j}, \ldots, b_{i_n n}$, for which $b_{i_k k} \leqslant b_{i_j j}$, $\forall k \in X_2 \backslash \{j\}$.

Then using the indices of the rows of these elements, we can determine the strategy for the first player by

$$i' = (i_1 - 1)\, m^{n-1} + (i_2 - 1)\, m^{n-2} + \ldots + (i_j - 1)\, m^{n-j} + \ldots + (i_n - 1)\, m^0 + 1. \tag{3}$$

So, the pair i', j is the Nash equilibrium for the informational extended game ${}_1\Gamma : (i', j) \in NE\left({}_1\Gamma\right)$.

Similarly, for the game ${}_2\Gamma$, we can determine the strategy for the second player by

$$j' = (j_1 - 1)\, n^{m-1} + (j_2 - 1)\, n^{m-2} + \ldots + (j_i - 1)\, n^{m-i} + \ldots + (j_m - 1)\, n^0 + 1, \tag{4}$$

where the indices j_i $\left(i = \overline{1, m}\right)$ are determined by the indices of columns of the elements $b_{ij_i} = \max\limits_j \{b_{i1}, b_{i2}, \ldots, b_{in}\}$, $\forall i = \overline{1, m}$.

Theorem 1. *Assume that for the initial game Γ the next conditions hold:*

1) for each column $j \in X_2$ there exists a single maximum element in the matrix $A : a_{i^*j} = \max\limits_{i \in X_1} a_{ij}$ and the corresponding element from the matrix B is the minimum element from the matrix $b_{i^*j} = \min\limits_{i \in X_1, j \in X_2} b_{ij}$;

2) there exists a single column $j^* \in X_2$ in the matrix A so that for the maximum element from this column $\max\limits_{i \in X_1} a_{ij^*} = a_{i'j^*}$ the corresponding element from the matrix B is greater than the minimum element from this matrix: $b_{i'j^*} > \min\limits_{i \in X_1, j \in X_2} b_{ij}$;

3) for each column from the matrix B the other elements are greater than $b_{i'j^*}$;

Then the informational extended game ${}_1\Gamma$ has a single Nash equilibrium, so $|NE\left({}_1\Gamma\right)| = 1$.

Theorem 2. *Assume that for the initial game Γ the next conditions hold:*

1) for each row $i \in X_1$ there exists a single maximum element in the matrix $B : b_{ij^*} = \max\limits_{j \in X_2} b_{ij}$ and the corresponding element from the matrix A is the minimum element from the matrix $a_{ij^*} = \min\limits_{i \in X_1, j \in X_2} a_{ij}$;

2) there exists a single row $i^* \in X_1$ in the matrix B so that for the maximum element from this row $\max\limits_{j \in X_2} b_{i^*j} = b_{i^*j'}$ the corresponding element from the matrix A is greater than the minimum element from the matrix $A : a_{i^*j'} > \min\limits_{i \in X_1, j \in X_2} a_{ij}$;

3) for each row from the matrix A the other elements are greater than $a_{i^*j'}$;

Then the informational extended game ${}_2\Gamma$ has a single Nash equilibrium, so $|NE\left({}_2\Gamma\right)| = 1$.

Proof. We prove the theorem for the informational extended game ${}_1\Gamma$.

Consider that all conditions hold.

For the proof we use the algorithm for the determination of Nash equilibria and the relation (formula) for the number of Nash equilibria in the informational extended game.

According to this algorithm, in the informational extended game ${}_1\Gamma$ Nash equilibria will exist for the elements $a_{i^*j} = \max\limits_{i \in X_1} a_{ij}$, $j \in X_2$, if for the corresponding element b_{i^*j} in each column from the matrix B there are elements less than b_{i^*j}.

There are two cases:

1) the element a_{i^*j} is from column j^* (according to the second condition from the theorem);

2) the element a_{i^*j} is not from the column j^*.

Consider the first case. So for $j' \in X_2$, $j' \neq j^*$ we consider the element $a_{i^*j'} = \max\limits_{i \in X_1} a_{ij'}$. Then from first condition of theorem it follows that $b_{i^*j'} = \min\limits_{i \in X_1, j \in X_2} b_{ij} = \min\limits_{i \in X_1} b_{ij'}$, so $\min\limits_{i \in X_1} b_{ij^*} > \min\limits_{i \in X_1, j \in X_2} b_{ij}$. Then in the relation (formula) for the number of Nash equilibria in the informational extended game ${}_1\Gamma$ we have the component with index j^* equal to zero. Thus for the pair of elements $a_{i^*j'}$ and $b_{i^*j'}$ we have the number of Nash equilibria in the game ${}_1\Gamma$: $N_{j'} = 1 \cdot 1 \cdot \ldots \cdot 0 \cdot \ldots \cdot 1 = 0$. So for all pairs of elements for which the element a_{i^*j} is not from the column j^* Nash equilibria do not exist in the informational extended game ${}_1\Gamma$.

Consider now the second case. For the column $j' \in X_2$, $j' = j^*$ we consider the element $a_{i^*j'} = \max\limits_{i \in X_1} a_{ij^*} = a_{i^*j^*}$ and the corresponding element $b_{i^*j^*} = \min\limits_{i \in X_1} b_{ij^*} > \min\limits_{i \in X_1, j \in X_2} b_{ij}$ (according to the second condition from the theorem). Then for each column $j \in X_2$ in the matrix B a single minimum element exists, so $n_1 = \ldots = n_{j^*} = \ldots = n_n = 1$ and $N_{j^*} = 1 \cdot 1 \cdot \ldots \cdot 1 \cdot \ldots \cdot 1 = 1$.

Thus for all pairs of elements $\left(a_{i^*j'}, b_{i^*j'}\right)$ we have a single Nash equilibrium: $\sum\limits_{j \in X_2} N_j = 1$ therefore $\left|NE\left({}_1\Gamma\right)\right| = 1$. The theorem is proved.

Remark 1. *From the conditions of this theorem it follows:*

1) for the game ${}_1\Gamma$ there is a single minimum element in each column in the matrix B;

2) for the game ${}_2\Gamma$ there is a single minimum element in each row in the matrix A.

We can do some specifications for the third condition from this theorem.

Remark 2. *For the informational extended game ${}_1\Gamma$ the column j^* in the matrix B can contain some elements $b_{i'j^*}$ less than $b_{i^*j^*}$, and greater than $\min\limits_{i \in X_1, j \in X_2} b_{ij}$, only if $a_{i'j^*} \neq \max\limits_{i \in X_1} a_{ij^*}$.*

For the informational extended game ${}_2\Gamma$ the row i^* in the matrix A can contain some elements $a_{i^*j'}$ less than $a_{i^*j^*}$, and greater than $\min\limits_{i \in X_1, j \in X_2} b_{ij}$, only if $b_{i^*j'} \neq \max\limits_{j \in X_2} b_{i^*j}$.

Example 1. *Consider the initial game Γ with the matrices*

$$A = \begin{pmatrix} 0 & \underline{4} & 2 & 3 \\ \underline{5} & 0 & 3 & 2 \\ 3 & 3 & \underline{6} & 0 \\ 1 & 2 & 1 & \underline{\underline{7}} \end{pmatrix}; \qquad B = \begin{pmatrix} \underline{4} & 0 & 2 & 2 \\ 0 & \underline{5} & 3 & 3 \\ 3 & 3 & 0 & \underline{6} \\ 2 & 2 & \underline{\underline{5}} & 1 \end{pmatrix}.$$

For the initial game Γ, Nash equilibria do not exist.

For the elements of these matrices all conditions of theorem hold and for each of the informational extended games ${}_1\Gamma$ and ${}_2\Gamma$ there exists a single Nash equilibrium.

Consider firstly the informational extended game ${}_1\Gamma$. For this game the matrices have the dimension $[256 \times 4]$.

We determine the maximum elements from each column in the matrix A : $a_{21} = 5$, $a_{12} = 4$, $a_{33} = 6$, $a_{44} = 7$.

For each of these elements we examine the pairs of elements from the matrices A and B. Thus for the pairs of elements $(a_{21}, b_{21}) = (5, 0)$, $(a_{12}, b_{12}) = (4, 0)$, $(a_{33}, b_{33}) = (6, 0)$ Nash equilibria do not exist in the informational extended game ${}_1\Gamma$.

For the pair of elements $(a_{44}, b_{44}) = (7, 1)$ in the extended matrix $\widetilde{A}$ there exists a row formed by elements $(a_{21}, a_{12}, a_{33}, a_{44}) = (5, 4, 6, 7)$ while the extended matrix $\widetilde{B}$ will contain the corresponding row formed by the corresponding elements $(0, 0, 0, 1)$; in the extended matrices these rows will have the index $i^* = 76$. For the determination of this number we use the relation (3). For the determination of Nash equilibrium we determine the maximum element from the row i^* in the extended matrix $\widetilde{B}$. Thus the single Nash equilibrium in the informational extended game ${}_1\Gamma$ is $(i^*, j^*) = (76, 4)$, and the payoffs for the players are the corresponding elements $(a_{44}, b_{44}) = (7, 1)$.

Consider now the informational extended game ${}_2\Gamma$. For this game the matrices have the dimension $[4 \times 256]$.

The maximum elements in each row in the matrix B are: $b_{11} = 4$, $b_{22} = 5$, $b_{34} = 6$, $b_{43} = 5$.

Thus for the pairs of elements $(a_{11}, b_{11}) = (0, 4)$, $(a_{22}, b_{22}) = (0, 5)$, $(a_{34}, b_{34}) = (0, 6)$ Nash equilibria do not exist in the informational extended game ${}_2\Gamma$.

For the pair of elements $(a_{43}, b_{43}) = (1, 5)$ in the extended matrix $\overline{B}$ there exists a column formed by the elements $(b_{11}, b_{22}, b_{34}, b_{43}) = (4, 5, 6, 5)$ and the corresponding column in the extended matrix $\overline{A}$ will be formed by the corresponding elements $(0, 0, 0, 1)$; in the extended matrices these columns will have the index $j^* = 31$ (using the relation (4)). Further we determine the maximum elements in the column j^* from the extended matrix $\overline{A}$. So the single Nash equilibrium in the informational extended game ${}_2\Gamma$ is $(i^*, j^*) = (4, 31)$ and the payoffs for the players are the corresponding elements $(a_{43}, b_{43}) = (1, 5)$.□

Example 2. *We examine a game for which Nash equilibria do not exist and the conditions of theorem hold only for the informational extend game* ${}_2\Gamma$,

$$A = \begin{pmatrix} 0 & \underline{4} & \underline{6} \\ \underline{5} & 0 & \underline{6} \\ \underline{5} & 2 & 3 \end{pmatrix} \qquad B = \begin{pmatrix} \underline{7} & 2 & 1 \\ 1 & \underline{9} & 1 \\ 1 & 4 & \underline{\underline{6}} \end{pmatrix}.$$

For the informational extended game ${}_2\Gamma$ there is a single Nash equilibrium which will be determined in the columns from the extended matrices formed by the elements $(b_{11}, b_{22}, b_{33}) = (7, 9, 6)$ and $(a_{11}, a_{22}, a_{33}) = (0, 0, 3)$; and the payoffs for the players are the elements $(a_{33}, b_{33}) = (3, 6)$ respectively.

For the game ${}_1\Gamma$ there will exist four Nash equilibria, because in the matrix A in each of the first and the third columns there are two maximum elements. So in the extended matrix $\widetilde{A}$ will be four rows formed by the maximum elements from the columns of the matrix $A : (5, 4, 6)$ and in the matrix $\widetilde{B}$ there will exist four rows formed by the corresponding elements $(1, 2, 1)$. Thus the Nash equilibria will be determined by the elements $(a_{12}, b_{12}) = (4, 2)$, because $b_{12} = 2$ is the maximum element in the row $(1, 2, 1)$.

Example 3. *Consider a game in which Nash equilibria do not exist and for which all conditions of theorem hold for both informational extended games* ${}_1\Gamma$ *and* ${}_2\Gamma$, *so for each of these informational extended games exists a single Nash equilibrium*

$$A = \begin{pmatrix} 1 & \underline{\underline{4}} & 2 \\ \underline{5} & 0 & 3 \\ 0 & 3 & \underline{6} \end{pmatrix} \qquad B = \begin{pmatrix} 4 & 2 & \underline{\underline{5}} \\ 0 & \underline{5} & 3 \\ \underline{3} & 1 & 0 \end{pmatrix}.$$

For informational extended game ${}_1\Gamma$ the Nash equilibrium will be determined by the pair of elements $(a_{12}, b_{12}) = (4, 2)$; and for the game ${}_2\Gamma$ the Nash equilibrium will be determined by the pair of elements $(a_{13}, b_{13}) = (2, 5)$.

References

[1] Kukushkin, N. S., Morozov, V. V., *Teoria neantagonisticeskih igr*, Moscova, 1984, 46-51.

[2] Novac L., *Existence of Nash equilibrium situations in extended bimatriceal informational games*, Analele Ştiintifice, Fac. Mat. si Inf., USM, Chişinău, **4**(2002), 66-71.(Romanian)

[3] Hancu B., Novac L., *Informational aspects in the game theory*, Annals of the Tiberiu Popoviciu Seminar of Functional Equations, Approximation and Convexity, Cluj-Napoca, **3**(2005), 25-34.

[4] Novac L., *Informational extended games*, Second Conference of the Mathematical Society of the Republic of Moldova (dedicated to the 40th Anniversary of the foundation of the Institute of Mathematics and Computer Science of ASM), Communications, Chishinau, 2004, 232-234.

[5] Novac L., *Existence of Nash equilibria in extended bimatriceal informational games*, Analele ATIC, USM, Chişinău, **4**(2004), 32-35.(Romanian)

ROMAI J., **4**, 1(2008), 203–206

ABOUT SOME FUNCTIONAL INTEGRAL EQUATIONS IN SPACES WITH PERTURBATED METRIC

Ion-Marian Olaru, Vasilica Olaru

University "Lucian Blaga" of Sibiu, High School "Gustav Gundish", Sibiu

Abstract In spaces with perturbated metric the following functional integral equation

$$u(x) = h(x, u(0)) + \int_0^{x_1} \cdots \int_0^{x_m} K(x, s, u(\theta_1 s, \cdots, \theta_m s))ds, \tag{1}$$

where

$$x, s \in D = \prod_{i=1}^{m} [0, b_i],\ m(D) \leq 1,\ \theta_i \in (0, 1), (\forall) i = \overline{1, m}$$

is studied.

Keywords: fixed point, weakly Picard operators.

2000 MSC: 47K10.

1. INTRODUCTION

Let (X, d) be a metric space and $A : X \to X$ an operator. We shall use the following notations
$F_A := \{x \in X \mid A(x) = x\}$ the fixed points set of A;
$I(A) := \{Y \in P(X) \mid A(Y) \subset Y\}$ the family of the nonempty invariant subsets of A;
$A^{n+1} = A \circ A^n, A^0 = 1_X, A^1 = A, n \in N.$

Definition 1.1. *[4] An operator A is an weakly Picard operator (WPO) if the sequence*

$$(A^n(x))_{n \in N}$$

converges for all $x \in X$, and the limit (which depends on x) is a fixed point of A.

Definition 1.2. *[4] If the operator A is WPO and $F_A = \{x^*\}$ then A is called a Picard operator.*

Definition 1.3. *[4] If A is an WPO, then we define the operator*

$$A^\infty : X \to X,\ A^\infty(x) = \lim_{n\to\infty} A^n(x).$$

We remark that $A^\infty(X) = F_A$.

Definition 1.4. *[4] Let be A an WPO and $c > 0$. The operator A is called a c-WPO if*

$$d(x, A^\infty(x)) \leq c \cdot d(x, A(x)).$$

We have the following characterization of the WPOs

Theorem 1.1. *[4] Let (X,d) be a metric space and $A : X \to X$ an operator. The operator A is an WPO (c-WPO) if and only if there exists a partition of X,*

$$X = \bigcup_{\lambda\in\Lambda} X_\lambda$$

such that

(a) $X_\lambda \in I(A)$,

(b) $A \mid X_\lambda : X_\lambda \to X_\lambda$ is a Picard (c-Picard) operator, for all $\lambda \in \Lambda$.

2. MAIN RESULTS

Let (X,d) be a complete metric space. We denote by P the set of functions $g : \mathbb{R}_+ \to \mathbb{R}_+$ which are strictly increasing, continuous and surjective. By Φ we denote the set of functions introduced by

Definition 2.1. *We say that the function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ belongs to the Φ class if the following conditions are met:*

(1) φ is increasing;

(2) $\varphi(t) < t$, for all $t \in \mathbb{R}_+$;

(3) φ is right continuous.

Example 2.1. *The function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$, $\varphi(t) = at$, $a < 1$ belongs to the set Φ.*

Example 2.2. *The function $g : \mathbb{R}_+ \to \mathbb{R}_+$, $g(t) = t^2$, belongs to the set P.*

Proposition 2.1. *[3] Let $f : X \to X$ be an operator and $\varphi \in \Phi$, $g \in P$ such that:*

(i) $g(d(f(x), f(y))) \le \varphi(g(d(x, y)))$, for all $x, y \in X$.

Then f has a unique fixed point, which is the limit of successively approximations sequence.

Proposition 2.2. *We suppose that:*

(i) $h \in C(D \times \mathbb{R}^n)$ and $K \in C(D \times D \times \mathbb{R}^n)$;

(ii) $h(0, \alpha) = \alpha$, $(\forall)\alpha \in \mathbb{R}^n$;

(iii) there exists $g \in P$, $\varphi \in \Phi$ such that

$$g(\|K(x, s, u_1) - K(x, s, u_2)\|_{\mathbb{R}^n}) \le \varphi(g(\|u_1 - u_2\|_{\mathbb{R}^n})),$$

for all $x, s \in D$ and $u_1, u_2 \in \mathbb{R}^n$.

In these conditions the equation(1) has in $C(D, \mathbb{R})$ an infinity of solutions.

Proof: Consider the operator

$$A : (C(D, \mathbb{R}^n), |\cdot|) \to (C(D, \mathbb{R}^n), |\cdot|),$$

$$A(u)(x) := h(x, u(0)) + \int_0^{x_1} \cdots \int_0^{x_m} K(x, s, u(\theta_1 s, \cdots, \theta_m s))ds.$$

Here $|u| = \max\limits_{x \in D} |u(x)|$.

Let $\lambda \in \mathbb{R}^n$ and $X_\lambda = \{u \in C(D, \mathbb{R}^n) \mid u(0) = \lambda\}$. Then

$$C(D, \mathbb{R}^n) = \bigcup_{\lambda \in \mathbb{R}^n} X_\lambda.$$

is a partition of $C(D, \mathbb{R}^n)$ and $X_\lambda \in I(A)$, for all $\lambda \in \mathbb{R}^n$.

For all $u, v \in X_\lambda$, we have

$$g(\|A(u)(x) - A(v)(x)\|_{\mathbb{R}^n}) \le$$

$$g(\int_0^{x_1}\cdots\int_0^{x_m} g^{-1}(\varphi(g(\|K(x,s,u(\theta_1 s,\cdots,\theta_m s) - K(x,s,v(\theta_1 s,\cdots,\theta_m s))\|))))) ds$$

$$\leq g(m(D)g^{-1}\varphi(g(|u-v|))) \leq \varphi(g(|u-v|))$$

Then, via Proposition 2.1, $A \mid X_\lambda$ is a Picard, while A is an weakly Picard operator.

References

[1] P. Corazza, *Introduction to metric-preserving function, Amer. Math. Monthly*, **104**, 4 (1999), 309-323.

[2] M. S. Khan, M. Swaleh, S. Sessa, *Fixed point theorems by altering distances between the points*, Bull. Austral. Math. Soc., **30**(1984), 1.

[3] I.-M. Olaru, *A fixed point result in spaces with perturbated metric*, (to appear) .

[4] I. A. Rus, *Weakly Picard operators and applications*, Seminar on fixed point theory Cluj-Napoca, 2 (2001), 41-57.

[5] I. A. Rus, *Generalized contractions*, Seminar on fixed point theory, **3** (1983), 1-130.

[6] M.-A. Şerban, *Spaces with perturbated metrics and fixed point theorems*, the Twelfth International Conference on Applied Mathematics, Computer Science and Mechanics, Băişoara, Semptember 10-13, 2008.

ROMAI J., **4**, 1(2008), 207–214

SOLVING FUZZY LINEAR SYSTEMS OF EQUATIONS

A. Panahi, T. Allahviranloo, H. Rouhparvar

Depart. of Math., Science and Research Branch, Islamic Azad University, Tehran, Iran

panahi53@gmail.com

Abstract Systems of linear equations, with uncertainty on the parameters, play a major role in various problems in economics and finance. Fuzzy system of linear equations has been discussed in [1] using LU decomposition when the matrix A in $Ax = b$ is a crisp matrix. Also the Adomian decomposition method and iterative methods have been studied in [2, 6] for fuzzy system of linear equations. In this paper we study such a system with fuzzy coefficients, i.e. the matrix A is a fuzzy matrix. We find two fuzzy matrices, the lower triangular L and the upper triangular U such that $A = LU$ and give a procedure to solve the fuzzy system of linear equations.

Keywords: fuzzy matrix, fuzzy linear system, decomposition.

2000 MSC: 15A33.

1. INTRODUCTION

In this paper we intend to find a new solution for x in the matrix equation

$$Ax = b, \tag{1}$$

where A is a fuzzy square matrix and x and b are fuzzy vectors. Buckly and Qu [4] argued that the classical solution, based on extension principle and regular fuzzy arithmetic, should rejected since it too often fails to exist. They defined six other solutions and showed that five of them are identical. Basically, in their work the solutions of all systems of linear crisp equations formed by the α-levels are calculated. In [8] another method for solving system of linear fuzzy equations based on using parametric functions in which the variables are given by the fuzzy coefficients of the system, was proposed.

2. NOTATIONS AND BASIC DEFINITIONS

First we recall some definitions concerning fuzzy numbers. We denote by E^1 the set of all fuzzy numbers.

Definition 2.1 *A fuzzy subset u of the real line* $\mathbb{R}$ *with membership function* $u(t) : \mathbb{R} \to [0,1]$ *is called a fuzzy number if:*
(a) u is normal, i.e., there exist an element t_0 *such that* $u(t_0) = 1$*;*
(b) u is fuzzy convex, i.e., $u(\lambda t_1 + (1-\lambda)t_2) \geq \min\{u(t_1), u(t_2)\}$*,* $\forall t_1, t_2 \in \mathbb{R}, \forall \lambda \in [0,1]$*;*
(c) $u(t)$ *is upper semicontinuous;*
(d) supp u is bounded, where supp u =cl $(\{t \in \mathbb{R} : u(t) > 0\})$*, and cl is the closure operator.*

Definition 2.2 *We represent an arbitrary fuzzy number by an ordered pair of functions* $[x](\alpha) = [x^1(\alpha), x^2(\alpha)]$, $0 \leq \alpha \leq 1$*, which satisfy the following requirements [6]:*
(a) $x^1(\alpha)$ *is a bounded left continuous nondecreasing function over* $[0,1]$*;*
(b) $x^2(\alpha)$ *is a bounded left continuous nonincreasing function over* $[0,1]$*;*
(c) $x^1(\alpha) \leq x^2(\alpha)$, $0 \leq \alpha \leq 1$.

For arbitrary $[x]^\alpha = [x^1(\alpha), x^2(\alpha)]$ and $[y]^\alpha = [y^1(\alpha), y^2(\alpha)]$ and $k > 0$ we define addition $[x \oplus y](\alpha)$ and scalar multiplication by k as

$$[x \oplus y]^\alpha = [x](\alpha) + [y](\alpha) = [x^1(\alpha) + y^1(\alpha), x^2(\alpha) + y^2(\alpha)],$$

and

$$[kx]^\alpha = \begin{cases} [kx^1(\alpha), kx^2(\alpha)], & k \geq 0 \\ [kx^2(\alpha), kx^1(\alpha)], & k < 0 \end{cases}$$

respectively, for every $\alpha \in [0,1]$. We denote by $-x = (-1)x \in E^1$ the symmetric of E^1. The product $x \odot y$ of two fuzzy numbers x and y, based on Zadeh's extension principle, is defined by

$$(x \odot y)^2(\alpha) = \max\{x^1(\alpha)y^1(\alpha), x^1(\alpha)y^2(\alpha), x^2(\alpha)y^1(\alpha), x^2(\alpha)y^2(\alpha)\},$$

$$(x \odot y)^1(\alpha) = \min\{x^1(\alpha)y^1(\alpha), x^1(\alpha)y^2(\alpha), x^2(\alpha)y^1(\alpha), x^2(\alpha)y^2(\alpha)\}.$$

Definition 2.3 *A fuzzy number* $x \in E^1$ *is said to be positive if* $x^1(1)$ *≥0, strictly positive if* $x^1(1) > 0$*, negative if* $x^2(1) \leq 0$ *and strictly negative if*

$x^2(1) < 0$. *We say that x and y have the same sign if they are either both positive or both negative [3].*

Definition 2.4 *A matrix $A = [a_{ij}]$ is called a fuzzy matrix, if each element of A is a fuzzy number. A is positive (negative) and denoted by $A > 0$ ($A < 0$) if each element of A is positive (negative). Similarly, nonnegative and nonpositive fuzzy matrices can be defined [5].*

The product of fuzzy numbers defined based on Zadeh's extension principle is not very practical from the computational point of view but the cross product is a computational method.

Now we study summary from the theoretical properties of the cross product of fuzzy numbers, for more details see [3]. Firstly we begin with a theorem which was obtained by using the stacking theorem [7].

Definition 2.5 *The binary operation $\otimes$ on E^1 that will be introduced by Theorem 2.1 and Corollary 2.1 is called the cross product of fuzzy numbers.*

Theorem 2.1. *If x and y are positive fuzzy numbers, then $w = x \otimes y$, defined by $[w]^\alpha = [w^1(\alpha), w^2(\alpha)]$, where*

$$\begin{cases} w^1(\alpha) = x^1(\alpha)y^1(1) + x^1(1)y^1(\alpha) - x^1(1)y^1(1), \\ w^2(\alpha) = x^2(\alpha)y^2(1) + x^2(1)y^2(\alpha) - x^2(1)y^2(1), \end{cases} \tag{2}$$

for every $\alpha \in [0,1]$, is a positive fuzzy number.

Corollary 2.1. *Let x and y be two fuzzy numbers.*

(a) If x is positive and y is negative then $x \otimes y = -(x \otimes (-y))$ is a negative fuzzy number.

(b) If x is negative and y is positive then $x \otimes y = -((-x) \otimes y)$ is a negative fuzzy number.

(c) If x and y are negative then $x \otimes y = (-x) \otimes (-y)$ is a positive fuzzy number.

Remark 2.1. *The below formulas of calculus can be easily proved ($\alpha \in [0,1]$):*

$$\begin{cases} (x \otimes y)^1(\alpha) = x^2(\alpha)y^1(1) + x^2(1)y^1(\alpha) - x^2(1)y^1(1), \\ (x \otimes y)^2(\alpha) = x^1(\alpha)y^2(1) + x^1(1)y^2(\alpha) - x^1(1)y^2(1), \end{cases} \tag{3}$$

if x is positive and y is negative,

$$\begin{cases} (x \otimes y)^1(\alpha) = x^1(\alpha)y^2(1) + x^1(1)y^2(\alpha) - x^1(1)y^2(1), \\ (x \otimes y)^2(\alpha) = x^2(\alpha)y^1(1) + x^2(1)y^1(\alpha) - x^2(1)y^1(1), \end{cases} \tag{4}$$

if x is negative and y is positive. In the last possibility, if x and y are negative, then

$$\begin{cases} (x \otimes y)^1(\alpha) = x^2(\alpha)y^2(1) + x^2(1)y^2(\alpha) - x^2(1)y^2(1), \\ (x \otimes y)^2(\alpha) = x^1(\alpha)y^1(1) + x^1(1)y^1(\alpha) - x^1(1)y^1(1). \end{cases} \tag{5}$$

Remark 2.2. *The cross product extends the scalar multiplication of fuzzy numbers. Indeed, if one of operands is the real number k identified with its characteristic function, then $k^1(\alpha) = k^2(\alpha), \forall \alpha \in [0,1]$ and using the above formulas of calculus we get the result.*

3. NEW METHOD FOR FINDING THE SOLUTION OF A FUZZY SYSTEM OF LINEAR EQUATIONS

In the previous section we have analyzed the properties and main features of the cross product for multiplying fuzzy numbers. In this section we are going to show some ideas about the use of this operation to solve fuzzy system of linear equations of the form (1).

We consider that the fuzzy coefficients and the elements of fuzzy right-hand side vector are all triangular fuzzy numbers. To find our new solution, first we find two fuzzy matrices L and U such that $A = L \otimes U$ and L is lower triangular with diagonals $l_{ii} = 1$ and U is an upper triangular fuzzy matrix. Consider $A = [a_{ij}]$, $L = [l_{ij}]$ and $U = [u_{ij}]$ be three fuzzy matrices such that

$$[a_{ij}]^\alpha = [a^1_{ij}(\alpha), a^2_{ij}(\alpha)], \quad [l_{ii}]^\alpha = [1,1], \quad 1 \le i \le n, \, 1 \le j \le n.$$

For $\alpha = 1$ we have

$$u_{1i}(1) = a_{1i}(1) \quad , \quad l_{i1}(1) = \frac{a_{i1}(1)}{u_{11}(1)}, \quad i = 1, \ldots, n\,, \tag{6}$$

and

$$u_{ri}(1)=a_{ri}(1)-\sum_{k=1}^{r-1}l_{rk}(1)u_{ki}(1),\quad l_{ir}(1)=\frac{a_{ir}(1)-\sum_{k=1}^{r-1}l_{ik}(1)u_{kr}(1)}{u_{rr}(1)},\qquad(7)$$

for $r=2,3,\ldots,n$ and $i=r,r+1,\ldots,n$. In case $\alpha\in[0,1)$ we obtain

$$\begin{cases}1\otimes[u^1_{1j}(\alpha),u^2_{1j}(\alpha)]=[a^1_{1j}(\alpha),a^2_{1j}(\alpha)], & j=1,2,\ldots,n\\[l^1_{i1}(\alpha),l^2_{i1}(\alpha)]\otimes[u^1_{11}(\alpha),u^2_{11}(\alpha)]=[a^1_{i1}(\alpha),a^2_{i1}(\alpha)], & i=1,2,\ldots,n\end{cases}$$

and

$$\begin{cases}[u^1_{ri}(\alpha),u^2_{ri}(\alpha)]\oplus\sum_{k=1}^{r-1}[l^1_{rk}(\alpha),l^2_{rk}(\alpha)]\otimes[u^1_{ki}(\alpha),u^2_{ki}(\alpha)]=[a^1_{ri}(\alpha),a^2_{ri}(\alpha)],\\[l^1_{ir}(\alpha),l^2_{ir}(\alpha)]\otimes[u^1_{rr}(\alpha),u^2_{rr}(\alpha)]\oplus\sum_{k=1}^{r-1}[l^1_{ik}(\alpha),l^2_{ik}(\alpha)]\otimes[u^1_{kr}(\alpha),u^2_{kr}(\alpha)]=[a^1_{ir}(\alpha),a^2_{ir}(\alpha)],\\r=2,3,\ldots,n;\quad i=r,r+1,\ldots,n.\end{cases}$$

Now suppose the elements of matrices L and U are positive, therefore according to (2) we get

$$u^1_{1j}(\alpha)=a^1_{1j}(\alpha),\quad u^2_{1j}(\alpha)=a^2_{1j}(\alpha),\quad j=1,\ldots,n\qquad(8)$$

$$l^1_{i1}(\alpha)=\frac{a^1_{i1}(\alpha)-l^1_{i1}(1)u^1_{11}(\alpha)+l^1_{i1}(1)u^1_{11}(1)}{u^1_{11}(1)},\qquad(9)$$

$$l^2_{i1}(\alpha)=\frac{a^2_{i1}(\alpha)-l^2_{i1}(1)u^2_{11}(\alpha)+l^2_{i1}(1)u^2_{11}(1)}{u^2_{11}(1)},$$

for $i=1,\ldots,n$ and

$$\begin{cases}u^1_{rj}(\alpha)=a^1_{rj}(\alpha)-\sum_{k=1}^{r-1}(l^1_{rk}(\alpha)u^1_{kj}(1)+l^1_{rk}(1)u^1_{kj}(\alpha)-l^1_{rk}(1)u^1_{kj}(1)),\\u^2_{rj}(\alpha)=a^2_{rj}(\alpha)-\sum_{k=1}^{r-1}(l^2_{rk}(\alpha)u^2_{kj}(1)+l^2_{rk}(1)u^2_{kj}(\alpha)-l^2_{rk}(1)u^2_{kj}(1)),\\j=r,r+1,\ldots,n,\quad r=2,3,\ldots,n\,,\end{cases}\qquad(10)$$

also

$$l^1_{ir}(\alpha) = \frac{a^1_{ir}(\alpha) - \sum_{k=1}^{r-1} \psi - l^1_{ir}(1)u^1_{rr}(\alpha) + l^1_{ir}(1)u^1_{rr}(1)}{u^1_{rr}(1)}, \tag{11}$$

$$l^2_{ir}(\alpha) = \frac{a^2_{ir}(\alpha) - \sum_{k=1}^{r-1} \omega - l^2_{ir}(1)u^2_{rr}(\alpha) + l^2_{ir}(1)u^2_{rr}(1)}{u^2_{rr}(1)},$$

where $j = r, r+1, \ldots, n$, $r = 2, 3, \ldots, n$, $\psi = l(1)_{ik}(\alpha)u(1)_{kr}(1) + l(1)_{ik}(1)u(1)_{kr}(\alpha) - l(1)_{ik}(1)u(1)_{kr}(1)$ and $\omega = l^2_{ik}(\alpha)u^2_{kr}(1) + l^2_{ik}(1)u^2_{kr}(\alpha) - l^2_{ik}(1)u^2_{kr}(1)$.

Now we solve the system $Ly = b$, and after finding y we solve the system $Ux = y$ to find the solution x for the fuzzy system $Ax = b$.

4. AN EXAMPLE

Example 4.1. *Consider the* 3×3 *fuzzy system of linear equations* $Ax = b$ *and let* A *be a fuzzy matrix and* b *be a fuzzy vector in the* α*-cut representation as follows*

$$b = \begin{bmatrix} [1+\alpha, 3-\alpha] \\ [\alpha, \quad 2-\alpha] \\ [-3, -2-\alpha] \end{bmatrix}, A = \begin{bmatrix} [5+\alpha, 10-4\alpha] & [3+2\alpha, 7-2\alpha] & [1+2\alpha, 4-\alpha] \\ [4+8\alpha, 32-20\alpha] & [2+12\alpha, 29-15\alpha] & [8\alpha, 18-10\alpha] \\ [14+10\alpha, 58-34\alpha] & [2+30\alpha, 62-30\alpha] & [1+19\alpha, 44-24\alpha] \end{bmatrix}.$$

For $\alpha = 1$ we have

$$\begin{bmatrix} 6 & 5 & 3 \\ 12 & 14 & 8 \\ 24 & 32 & 20 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \times \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix},$$

where u_{ij} and l_{ij} are $u_{ij}(1)$ and $l_{ij}(1)$ respectively. In view of relations (6) and (7) we have

$$L(1) = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 4 & 3 & 1 \end{bmatrix}, \quad U(1) = \begin{bmatrix} 6 & 5 & 3 \\ 0 & 4 & 2 \\ 0 & 0 & 2 \end{bmatrix}.$$

Therefore, according to the sign of elements of matrixes L and U, also relations (8-11) for $\alpha \in [0, 1)$, we obtain L and U as follows

$$L = \begin{bmatrix} 1 & 0 & 0 \\ [1+\alpha, 4-2\alpha] & 1 & 0 \\ [3+\alpha, 7-3\alpha] & [1+2\alpha, 4-\alpha] & 1 \end{bmatrix},$$

$$U = \begin{bmatrix} [5+\alpha, 10-4\alpha] & [3+2\alpha, 7-2\alpha] & [1+2\alpha, 4-\alpha] \\ 0 & [1+3\alpha, 5-\alpha] & [1+\alpha, 4-2\alpha] \\ 0 & 0 & [1+\alpha, 5-3\alpha] \end{bmatrix}.$$

Now we solve the system $Ly = b$, considering the cross product in each needed multiplication

$$\begin{bmatrix} 1 & 0 & 0 \\ [1+\alpha, 4-2\alpha] & 1 & 0 \\ [3+\alpha, 7-3\alpha] & [1+2\alpha, 4-\alpha] & 1 \end{bmatrix} \cdot \begin{bmatrix} [y_1^1, y_1^2] \\ [y_2^1, y_2^2] \\ [y_3^1, y_3^2] \end{bmatrix} = \begin{bmatrix} [1+\alpha, 3-\alpha] \\ [\alpha, \quad 2-\alpha] \\ [-3, -2-\alpha] \end{bmatrix}.$$

After finding y we solve in the same way the system $Ux = y$ to find the solution x for the fuzzy system $Ax = b$.

$$\begin{bmatrix} [5+\alpha, 10-4\alpha] & [3+2\alpha, 7-2\alpha] & [1+2\alpha, 4-\alpha] \\ 0 & [1+3\alpha, 5-\alpha] & [1+\alpha, 4-2\alpha] \\ 0 & 0 & [1+\alpha, 5-3\alpha] \end{bmatrix} \cdot \begin{bmatrix} [x_1^1, x_1^2] \\ [x_2^1, x_2^2] \\ [x_3^1, x_3^2] \end{bmatrix} = \begin{bmatrix} [y_1^1, y_1^2] \\ [y_2^1, y_2^2] \\ [y_3^1, y_3^2] \end{bmatrix}.$$

5. CONCLUSION

In this paper we studied fuzzy linear system of the form $Ax = b$ with A square matrix of fuzzy coefficients and b fuzzy number vector. We introduced two fuzzy matrices, the lower triangular L and the upper triangular U such that $A = LU$ and solved the fuzzy system of linear equations $Ly = b$ and $Ux = y$ respectively.

References

[1] S. Abbasbandy, R. Ezzati, A. Jafarian, *LU decomposition method for solving fuzzy system of linear equations*, Appl. Math. and Comput., **172** (2006), 633-643.

[2] T. Allahviranloo, *Numerical methods for fuzzy system of linear equations*, Appl. Math. and Comput., **155** (2004), 493-502.

[3] B. Bede, J. Fodor, *Product type operations between fuzzy numbers and their applications in geology*, Acta Polytechnica Hungarica, **3**, 1 (2006), 123-139.

[4] J. J. Buckly, Y. Qu, *Solving systems of linear fuzzy equations*, Fuzzy Sets and Systems, **43** (1991) 33-43.

[5] M. Dehghan, B. Hashemi, M. Ghatee, *Computational methods for solving fully fuzzy linear systems*, Appl. Math. and Comput., **179** (2006), 328-343.

[6] M. Ma, M. Friedman, A. Kandel, *A new fuzzy arithmetic*, Fuzzy Sets and Systems, **108** (1999), 83-90.

[7] M. L. Puri, D. A. Ralescu, *Differentials of fuzzy functions*, J. Math. Anal. Appl., **91** (1983), 552-558.

[8] A. Vroman, G. Deschrijver, E. E. Kerre, *Solving systems of equations by parametric functions-An improved algorithm*, Fuzzy Sets and Systems.

ROMAI J., 4, 1(2008), 215–224

DESIGN SENSITIVITY ANALYSIS AND SHIP STERN HYDRODYNAMIC FLOW FIELD IMPROVEMENT

Horaţiu Tănăsescu, Nasta Tănăsescu

ICEPRONAV Galaţi, Univ. "Dunărea de Jos" Galaţi

Abstract The paper draws attention and briefly focuses on ship hulls stern flows in the light of two original ideas (concepts) in ship hydrodynamics, belonging to the author: 1. a new stern hydrodynamic concept (NSHC), with radial crenelated-corrugated sections (Tănăsescu's stern shape); 2. using of an inverse piezo-electric effect [(electric current→high-frequency power generator→piezoelectric driver made of certain ceramic material, which induces an elliptical vibratory movement (high frequency over 20 kHz), into the elastic side plates (15 mm thickness) in the streamlines direction (of the external flowing water)], able to reduce the total forward resistance.

Keywords: ship hydrodynamics.

2000 MSC: 76D55.

Introduction. A naval architect, a ship owner, or a simple passenger, in looking over the stern of a ship and the turbulent tossing above the propeller race, instinctively realizes that most of this upheaval is an wasted effort. The irregular nature of the currents which flow into the propeller disc imposes to this propulsion device to take such a complex water flow and make so much out of it in the way of useful thrust, inducing the idea that the turmoil surging out of the propeller disc can be converted into useful power that increases the ship velocity. To this aim it is necessary a profound knowledge of the basic phenomena and fundamental principles governing the motion of water and the causes for the particular behavior of a ship and its propulsion devices.

Theoretical aspects. In real conditions, a propeller is fitted behind the ships (models) hull stern, working in a non-uniform water stream, which has been disturbed by the ship hull during its forward motion. The ship moving hull carries with it a volume of surrounding water forming a region, known under the name of boundary layer, across which there is a steep change in velocity, the propeller being placed behind the ship hull stern, in the ship body trail. As a consequence, the velocity (even considering its average value) of water particles relative to the propeller disk is no longer (neither in magnitude nor in direction) equal to the velocity of advance of the propeller relative to still water. This trail, where there is a difference between the ship speed and the speed of the water particles relative to the ship is also termed wake. The wake is a zone poorly investigated theoretically (analytically), due to very complex, aleatory flow character within it. In ship propeller theory, a distinctive importance is given only to the incipient part of the trail (wake), located immediately in the front of the propeller disk plane. The movement in this zone is called wake movement or simply wake. The wake movement can be investigated either in the presence or in the absence of the propeller, bearing the name of the effective wake or the nominal wake, respectively. However, the wake movement of interest is only that from the plane where the propeller follows to be situated. The flow average velocity from that plane is termed wake speed V_W, and, in general, it is smaller than the ship speed V_S, relative to the infinite upstream water. If the water is moving in the same direction as the ship, the wake is said to be positive. Then

$$Wake = V_S - V_W.$$

In order to non-dimensionalize this relation, we can use as characteristic speed either V_W or V_S, leading to two wake factors

-Froude wake factor= $w_F = (V_s - V_W)/V_W$;
-Taylor wake factor= $w = (V_s - V_W)/V_S$.

Besides, this general effect of the ship hull, there exist local perturbations due to the shaft, shaft bossings or shaft brackets and other appendages. These effects combined lead to the so-called relative rotative efficiency (RRE), defined by

$RRE = \eta_R =$ efficiency of propeller behind the ship hull/ efficiency of propeller in open water (at speed V_W).

Always, but especially in present circumstances, for a better functioning it is necessary a propeller cavitation reducing, overall propulsive efficiency and stability improving. As already mentioned, the dynamics of a cavitating propeller depends on the system environment in which it is operating, the flow field within a propeller mounting behind a ship hull is very different from that one in an open water test or in a section of a cavitation tunnel. Thus, a propeller that is very efficient in open water can not be suited for a certain kind of stern shape architecture. For this reason, *the wake distribution in the propeller disk plane represents a key element for designing a ship hull stern form.* A uniform wake distribution from an immediately upstream propeller parallel plane disk can diminish the propeller cavitation (having as an indirect consequence on the noise and vibration level induced on board and in the hull stern structure, lowering) and increases the propulsive efficiency (and so, obtaining the minimum energy consumption). Therefore, to obtain a good nominal wake distribution is an important objective of naval architects. In addition, the global - directional hydrodynamic stability improving by using a special kind of stern having more appropriate architecture, can not be but favorable.

The state of the art in the field. The present-day tendency in maritime transportation industry is represented by designing and building of bigger, faster, more energy-efficient and stable ships but simultaneously having stricter noise and vibration levels for stern hull structure. A modern ships hull lines are designed to minimize the forward resistance, to reduce the propeller cavitation, to improve the propulsion performance and to increase the global hydrodynamic stability. Since the apparition of the first ships, the naval architects tried to improve the existing hull forms. As a general recently accepted opinion, the ships of the future will be designed and built only on the basis of some *new devised concepts.* It is well known that the stern flow problem is very complex. However, the ship hull stern flows have received much attention these last years, in particular with respect to their modeling and design principles. The most recently known industrial achievements focused on flow improvement in the stern region, which consist in symmetrically flattening of the stern lateral surfaces towards the central plane. This concept has resulted in a huge amount of inconveniences almost in all practical applications to real ships (unsuitable placing of equipments, lack of necessary spaces for inspections, repairs etc.). For a long time, the authors thought how to redesign the two systems - hydroframe system and propulsion system - very important (critical) for a ship, so that the hydroframe may meet the propulsion and the propulsion may meet the hydroframe in an optimal way.

Scientific research objectives: - total forward resistance reduction, propulsive efficiency increasing;
- propeller cavitation reduction (for level of noise and vibration induced on board and in the stern structure decreasing);
- development of a numerical parameterized model;
- design sensitivity analysis of fields generated;
- optimizations;
- original concepts and ideas validation;
- new methodologies establishing.

Project description, results obtained, future prospects. Having in view the presented ideas, within the contract no. 208, CEEX, I have proposed (intuitively, based on experience), a new stern hydrodynamic concept of streamline tube type, (having quasi-cylindrical increasing sections), which starts from front propeller disk and stretches until hull cylindrical region (fig. 1). In devising this new design concept, the author referred (as a supplementary basic background) to two existing theories:

- *the streamline tube theory* (the water particles axial velocities distribution at entrance in the propeller disk can be configured favorably - homogenized - by comprising the radial crenelated - corrugated stern sections in a stream tube that also comprises the propeller disk);

- *the Bernoulli effect* (increasing of water particles axial velocities in the regions within which the water pressure is decreased).

Taking into account the streamline tube theory and the Bernoulli effect, we can estimate that *the 3D spectrum* of flow generated around and outside a classical stern hull having practiced transversal crenellated-corrugated stern sections can be substantially improved by an architectural optimization in the direction of axial velocities from a propulsion propeller immediate front plane uniformization (fig.2).

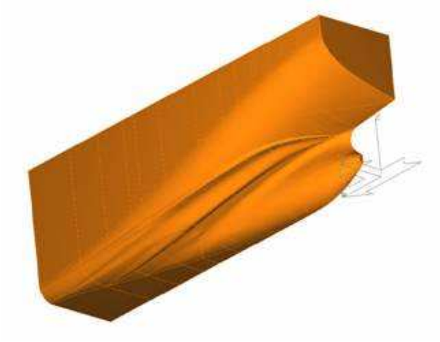


Fig. 1. The new stern hydrodynamic concept.

The directions of the crenelated-corrugated sections teeth crests and troughs longitudinal curved lines, will be those of the stern natural streamlines (which can be established experimentally in a flow visualization test) for vortices turning up avoiding and for a minimum forward resistance obtaining.

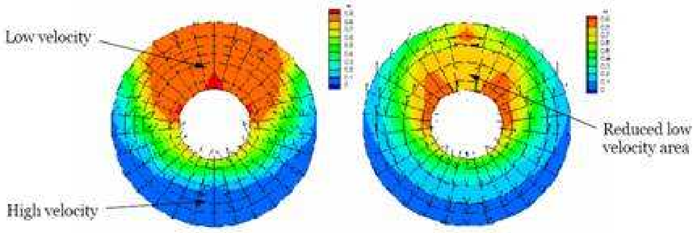


Fig. 2. Comparison between experimental wake obtained for the model with initial stern shape design (left) and for the model with modified stern shape according to our design (right).

Finally, the most important, until now, proved result, is the reducing of propeller cavitation (working in the simulated nominal wake of the hull using the new stern hydrodynamic concept with radial crenellated - corrugated sections, Tănăsescu's stern shape), practically to zero (fig.3).

Unfortunately, this cavitation decreasing (lack of cavitation) is associated with a total forward resistance (of the ship) increasing (approximately 4-5%)

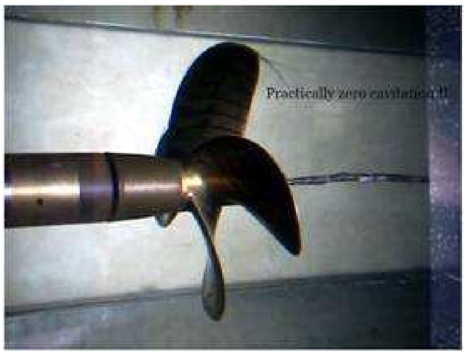


Fig. 3. Simulated nominal wake testing, in 850x850 mm section of the cavitation tunnel at 25 rps rotative speed (it can be remarked lack of cavitation).

due to initiation and movement of some multiple increased vortices (fig. 4), resulted from the separation (although a low one - fig. 5) of the boundary layer (destruction of an important part of fluid mechanical energy, pressure decreasing downwards the ship stern body etc).

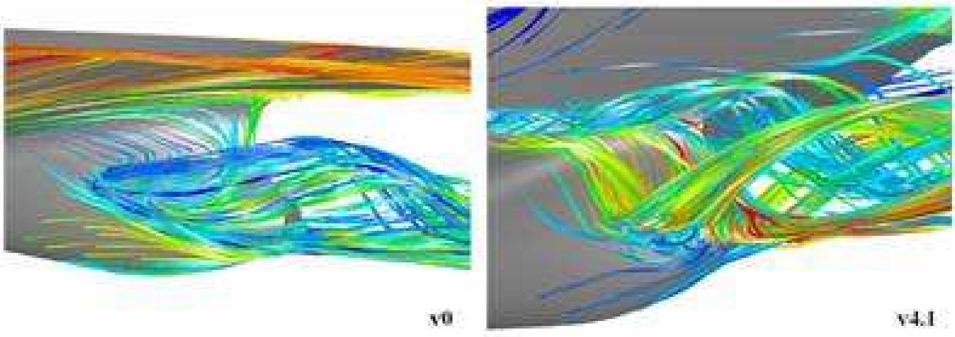


Fig. 4. Vortex initiation and separation - FLUENT 6.3. (left - the model with initial stern shape design - simple vortex; right - the model with modified stern shape according to our new concept design - multiple vortices).

Therefore, it would be necessary a much more reduction or even complete separation and multiple vortices phenomena (within the turbulent boundary layer) avoiding. *In this respect I thought that I should try to use the inverse piezoelectric effect (electric current → high-frequency generator → piezoelectric driver made of certain ceramic material - fig. 6), which induces an elliptical*

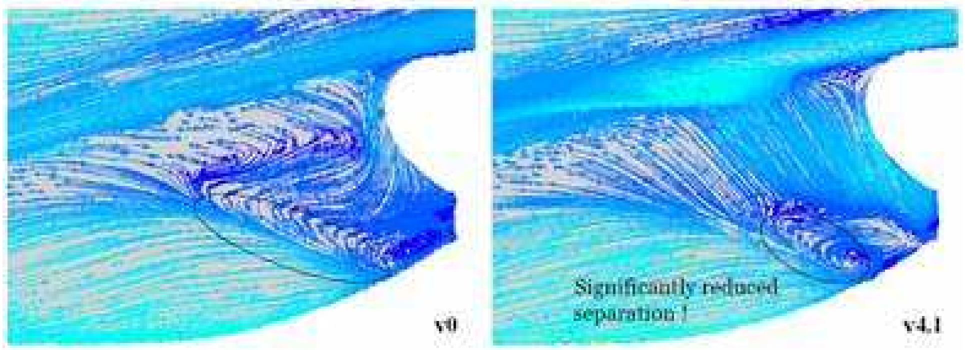


Fig. 5. Limit streamlines on stern surface - FLUENT 6.3. (left -the model with initial stern shape design; right - the model with modified stern shape according to our new concept design).

vibratory movement (high frequency over 20 kHz), into the elastic side plates (15 mm thickness) in the streamlines direction (of the external flowing water).

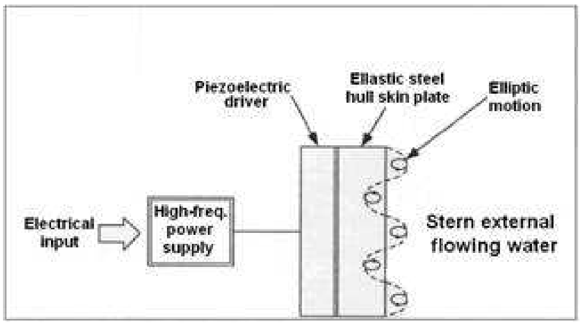


Fig. 6. Principle scheme of an ultrasonic vibrator.

Piezoelectricity is the ability of crystals and certain ceramic materials to generate a voltage in response to an applied mechanical stress. Piezoelectric effect was discovered by Pierre and Jacques Curie in 1880.

The basic principle would be the following: certain piezoelectric ceramic materials can be used to convert electrical energy into mechanical energy in the form of vibrations of an elastic body (ship hull stern plates), whose surface points perform an linear elliptic motion (in the streamlines direction of the external flowing water), with an ultrasonic frequency over 20 kHz.

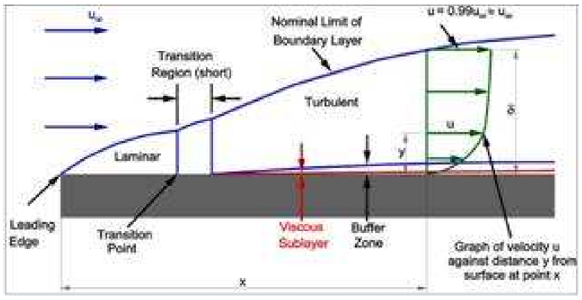

u_∞
Nominal Limit of Boundary Layer
$u = 0.99u_\infty \approx u_\infty$
Transition Region (short)
Turbulent
Laminar
u
y
δ
Leading Edge
Transition Point
Viscous Sublayer
Buffer Zone
Graph of velocity u against distance y from surface at point x
x

Fig. 7. Boundary-layer flow regions.

Water particles (from within the ship hull stern turbulent boundary layer - fig. 7), are pressed against vibrating steel plates reducing the interface (hull - water), skin friction drag. It is hoped that such a combination of devices can reduce ship forward resistance due to hull skin-water friction reduction by controlling the inside turbulent boundary layer flow characteristics.

In addition, as a continuation and a completion of the researches accomplished in this CEEX contract, author considers of interest the realization of:

- *a parameterized geometrical model* streamline tube type, (including the effects of new stern design having quasi-cylindrical increasing radial crenelated-corrugated sections on inside propeller flow);

- *a design sensitivity analysis* of the new stern fields generated.

In these cases different geometries (as necessary form, width and depth, along hull distances, for flow separation avoiding), should be studied theoretically, numerically and experimentally.

Design sensitivity analysis consists in determining derivatives of a system response with respect to its design parameters x_i. In the context of design optimization (of the new hydrodynamic stern concept proposed), the response is expressed in terms of objective and constraint functions and, accordingly, the overall aim of design sensitivity analysis is to find the gradients of these functions. However, since any such problem function depends explicitly on the dependent variables Φ of the considered problem, sensitivity formulations in essence aim at the calculation of the derivatives $\delta\Phi/\delta x_i$. In other words, the changes in the flowfield Φ following from a given change in design must be predicted. After the determination of these flowfield sensitivities, it is a matter of a straightforward calculus to compute the design sensitivities of any problem function

$$\nabla f = \frac{df}{dx_i} = \underbrace{\frac{\partial f}{\partial x_i} + \frac{\partial f}{\partial \gamma}\left[\frac{\partial \gamma}{\partial x_i}\right]}_{grid\,sensitivity} + \underbrace{\frac{\partial f}{\partial \phi}\left[\frac{\partial \phi}{\partial x_i}\right]}_{flow\,sensitivity} ,$$

$i = 1, ..., n_{dv}$, where: dv are the design variables;
$f(x_i)$ - problem function (typically identical to objective and constraints function);
x_i - design parameters;
$\gamma(x_i)$ - geometrical quantities;
$\Phi(x_i)$ - vector containing unknown flow variables (velocity, static pressure, possibly turbulence modeling quantities), determined by the governing equations. Obviously, both geometry and flow are implicitly controlled by the design parameters through hull stern surface parametrization, mesh generation and flow analysis.

Interdisciplinary research. The main disciplines involved into the present project are: mathematical physics (partial differential and integral equations); applied physics; technical physics; materials physics; modeling and simulation; hydraulics and fluid mechanics; naval hydrodynamics.

References

[1] G. K. Batchelor, *An introduction to fluid dynamics*, Cambridge University Press, 1991.

[2] C. A. J. Fletcher, *Computational techniques for fluid dynamics*, Springer, Berlin, 1991.

[3] FLUENT 6.3, *User's Manual*, Fluent Incorporated, Lebanon, NH, 2006.

[4] J. P. Ghose, R. P. Gokarn, *Basic ship propulsion*, Applied Publishers, Pvt. Ltd., New Delhi, 2004.

[5] S. K. Godunov, V. S. Reabenki, *Finite difference computing schemes*, Bucharest, 1977.

[6] H. Lamb, *Hydrodynamics*, Cambridge University Press, 1952.

[7] J. N. Newman, *Marine hydrodynamics*, Massachusetts Institute of Technology, 1982.

[8] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannerly, *Numerical recipes in Fortran - The art of scientific computing*, Cambridge University Press, 1992.

[9] K. J. Rawson, E. C. Tupper, *Basic ship theory*, Longman, London, 1995.

[10] M. G. Salvadori, M. L. Baron, *Numerical methods in engineering*, Bucharest, 1972.

[11] H. E. Saunders, *Hydrodynamics in ship design*, The Society of Naval Architects and Marine Engineers, New York, 1957.

[12] N. Tănăsescu, *Numerical methods in MATLAB*, MatrixRom, Bucharest, 2002.

[13] H. C. Tănăsescu, Al. A. Vasilescu, *Fluid mechanics, Course for engineers*, University of Galaţi, 2000.

[14] H. C. Tănăsescu, *Numerical analysis, Course for master of science engineers*, University of Galaţi, 2001.

[15] H. C. Tănăsescu, *Contributions concerning surface ships forward resistance studied by numerical simulation*, PhD thesis, University of Galati, 1998.

[16] Al. A. Vasilescu, *Hydromechanics course*, University of Galaţi, 1962.

ROMAI J., **4**, 1(2008), 225–242

SOLUTION PRINCIPLES FOR SIMULTANEOUS AND SEQUENTIAL GAMES MIXTURE

Valeriu Ungureanu

State University of Moldova, Chişinău, Republic of Moldova

valungureanu@usm.md

Abstract Equilibrium principles for the hierarchical and the mixture of hierarchical and simultaneous games are defined and examined by means of the graphs of the best response mappings, Pareto optimal response mappings and graphs intersection.

Keywords: Noncooperative game, strategic form game, multiobjective game, Nash equilibrium (NE), Stackelberg equilibrium (SE), Nash-Stackelberg equilibrium (NSE), Pareto-Nash-Stackelberg equilibrium (PNSE), graph of the best response mapping, graph of the Pareto optimal response mapping, graphs intersection.

2000 MSC: 91A05, 91A06, 91A10, 91A20, 91A40, 91A44, 91A65, 90C29, 90C39, 90C90.

1. INTRODUCTION

Interactive decisions situations, which involve both sequential decisions and simultaneous decisions made by independent and interdependent players with one or more objectives, can be modelled by means of strategic games (Stackelberg game [21, 3], Nash game [4], Pareto-Nash game [1, 7, 2, 5], Pareto-Nash-Stackelberg game). At every stage (level) of the Nash-Stackelberg game a Nash game is played. The stage profiles (joint decisions) are executed sequentially throughout the hierarchy as a Stackelberg game. At every stage of the multiobjective Pareto-Nash-Stackelberg game a multiobjective Pareto-Nash game is played. Stage profiles are executed sequentially throughout the hierarchy. Via notion of best response mapping graph we define unsafe and safe Stackelberg equilibria for Stackelberg games, pseudo and multi-stage Nash-Stackelberg equilibria for Nash-Stackelberg games, and Pareto-Nash-Stackelberg equilibria for multiobjective Pareto-Nash-Stackelberg games.

The carried out investigation continues and extends the Nash game research via Nash equilibrium as an element of the best response mapping graphs intersection [6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20].

Consider the noncooperative strategic form game

$$\Gamma = \langle N, \{X_p\}_{p\in N}, \{f_p(x)\}_{p\in N}\rangle,$$

where

- $N = \{1, 2, ..., n\}$ is a set of players,
- $X_p \subseteq R^{k_p}$ is a set of strategies of player $p \in N$,
- $k_p < +\infty, p \in N$,
- and $f_p(x)$ is a p^{th} player cost function defined on the Cartesian product $X = \underset{p\in N}{\times} X_p$.

Elements $x = (x_1, x_2, ..., x_n) \in X$ are named profiles of the game.

Suppose that the players make their moves hierarchically:

first player chooses his strategy $x_1 \in X_1$ and communicates it to the second player,

the second player chooses his strategy $x_2 \in X_2$ and communicates x_1, x_2 to the third player,

...

n^{th} player selects his strategy $x_n \in X_n$ after observing the moves $x_1, ..., x_{n-1}$ of the preceding players.

On the resulting profile $x = (x_1, ..., x_n)$ every player computes the value of his cost function.

When player $p \in N$ moves, players $1, 2, ..., p-1$ are leaders or predecessors of player p and players $p+1, ..., n$ are followers or successors of the player p. Players have all the information about the predecessors choices and doesn't have information about the choices of the successors, but the p^{th} player ($p < n$) has all the information about the all strategy sets and the cost functions of *the players $p, p+1, ..., n$. Without loss of generality suppose that all players minimize the values of their cost functions.*

By backward induction, every player $n, n-1, ..., 2$ computes his best move mapping and the first player computes the set of his best moves:

$$\begin{aligned}
&B_n(x_1, ..., x_{n-1}) = \operatorname*{Arg\,min}_{y_n \in X_n} f_n(x_1, ..., x_{n-1}, y_n),\\
&B_{n-1}(x_1, ..., x_{n-2}) = \operatorname*{Arg\,min}_{y_{n-1},\, y_n :\, (x_1, ..., x_{n-2},\, y_{n-1},\, y_n) \in Gr_n} f_{n-1}(x_1, ..., x_{n-2},\, y_{n-1}, y_n),\\
&\ldots\\
&B_2(x_1) = \operatorname*{Arg\,min}_{y_2, ..., \, y_n :\, (x_1, y_2, ..., \, y_n) \in Gr_3} f_2(x_1, y_2, ..., y_n),\\
&\hat{X} = \operatorname*{Arg\,min}_{(y_1, ..., y_n) \in Gr_2} f_1(y_1, ..., y_n)
\end{aligned}$$

where

$$\begin{aligned}
&Gr_n = \{x \in X : x_1 \in X_1, ..., x_{n-1} \in X_{n-1}, x_n \in B_n(x_1, ..., x_{n-1})\},\\
&Gr_{n-1} = \{x \in Gr_n : x_1 \in X_1, ..., x_{n-2} \in X_{n-2}, (x_{n-1}, x_n) \in B_{n-1}(x_1, ..., x_{n-2})\},\\
&\ldots\\
&Gr_2 = \{x \in Gr_3 : x_1 \in X_1, (x_2, ..., x_n) \in B_2(x_1)\}.
\end{aligned}$$

Evidently, $Gr_2 \subseteq Gr_3 \subseteq \cdots \subseteq Gr_n$ and form a family of nested sets.

Definition. *Any profile $\hat{x} \in \hat{X}$ of the game is called an unsafe (optimistic, strong) Stackelberg equilibrium.*

This definition is equivalent to the respective [3] definition. For $n = 2$ the unsafe Stackelberg equilibrium notion and original Stackelberg equilibrium [21] notion are equivalent.

2. UNSAFE STACKELBERG EQUILIBRIUM. EXISTENCE AND PROPERTIES

Theorem 2.1. *For every finite hierarchical game the set $\hat{X}$ of unsafe Stackelberg equilibria is non empty.*

The proof is evident.

Corollary. *The unsafe Stackelberg equilibrium notion and the Nash equilibrium notion are not equivalent.*

For the proof it is sufficient to mention (in the Theorem 1 context) that the Nash equilibrium does not exist for every pure strategy finite game.

Example 1. Consider a three-player $2 \times 2 \times 2$ game with the cost matrices

$$a_{1**} = \begin{bmatrix} 5 & 2 \\ 1 & 3 \end{bmatrix}, \quad a_{2**} = \begin{bmatrix} 1 & 3 \\ 3 & -1 \end{bmatrix},$$

$$b_{1**} = \begin{bmatrix} 5 & 7 \\ 4 & 6 \end{bmatrix}, \quad b_{2**} = \begin{bmatrix} 3 & 4 \\ 8 & 5 \end{bmatrix},$$

$$c_{1**} = \begin{bmatrix} 2 & 10 \\ 7 & 3 \end{bmatrix}, \quad c_{2**} = \begin{bmatrix} 8 & 6 \\ 4 & 5 \end{bmatrix}.$$

The first player moves the first, the second player moves the second and the third player moves the last.

First of all we must determine the graph Gr_3 of the third player. The graph elements are emphasized by boxes

$$c_{1**} = \begin{bmatrix} \boxed{2} & 10 \\ 7 & \boxed{3} \end{bmatrix}, \quad c_{2**} = \begin{bmatrix} 8 & \boxed{6} \\ \boxed{4} & 5 \end{bmatrix}.$$

The second player graph Gr_2 is determined on the base of the Gr_3. Its elements are emphasized by two frames boxes

$$b_{1**} = \begin{bmatrix} \boxed{\boxed{5}} & 7 \\ 4 & \boxed{6} \end{bmatrix}, \quad b_{2**} = \begin{bmatrix} 3 & \boxed{\boxed{4}} \\ \boxed{8} & 5 \end{bmatrix},$$

Last, the set of the unsafe Stackelberg equilibria is determined on Gr_2

$$a_{1**} = \begin{bmatrix} \boxed{\boxed{5}} & 2 \\ 1 & \boxed{3} \end{bmatrix}, \quad a_{2**} = \begin{bmatrix} 1 & \boxed{\boxed{\boxed{3}}} \\ \boxed{3} & -1 \end{bmatrix}.$$

The game consists of one unsafe Stackelberg equilibrium $(2,1,2)$ with the players' cost functions values $(3,4,6)$. Remark, the profile $(2,1,2)$ is not a Nash equilibrium. Moreover, the corresponding three matrix game does not have pure strategies Nash equilibria. □

In Example 1 the player best move mappings are mono-valued and the realization of the unique unsafe Stackelberg equilibrium is natural. The situation is more complicated when the mappings are multi-valued. The realization of the unsafe Stackelberg equilibrium is problematic. The modification of Example 1 illustrates this fact.

Example 1 (continuation). We modify only the elements a_{211}, b_{211} and c_{211} of the cost matrices in Example 1

$$a_{1**} = \begin{bmatrix} 5 & 2 \\ 1 & 3 \end{bmatrix}, \quad a_{2**} = \begin{bmatrix} +\infty & 3 \\ 3 & -1 \end{bmatrix},$$

$$b_{1**} = \begin{bmatrix} 5 & 7 \\ 4 & 6 \end{bmatrix}, \quad b_{2**} = \begin{bmatrix} 4 & 4 \\ 8 & 5 \end{bmatrix},$$

$$c_{1**} = \begin{bmatrix} 2 & 10 \\ 7 & 3 \end{bmatrix}, \quad c_{2**} = \begin{bmatrix} 6 & 6 \\ 4 & 5 \end{bmatrix}.$$

The elements of the Gr_3 are emphasized by boxes

$$c_{1**} = \begin{bmatrix} \boxed{2} & 10 \\ 7 & \boxed{3} \end{bmatrix}, \quad c_{2**} = \begin{bmatrix} \boxed{6} & \boxed{6} \\ \boxed{4} & 5 \end{bmatrix}.$$

Elements of the Gr_2 are emphasized by two frames boxes

$$b_{1**} = \begin{bmatrix} \boxed{\boxed{5}} & 7 \\ 4 & \boxed{6} \end{bmatrix}, \quad b_{2**} = \begin{bmatrix} \boxed{\boxed{4}} & \boxed{\boxed{4}} \\ \boxed{8} & 5 \end{bmatrix},$$

Last, the set of the unsafe Stackelberg equilibria is determined on Gr_2

$$a_{1**} = \begin{bmatrix} \boxed{\boxed{5}} & 2 \\ 1 & \boxed{3} \end{bmatrix}, \quad a_{2**} = \begin{bmatrix} \boxed{\boxed{+\infty}} & \boxed{\boxed{\boxed{3}}} \\ \boxed{3} & -1 \end{bmatrix}.$$

The game consists of the same unique unsafe Stackelberg equilibrium $(2,1,2)$ with the players' cost functions values $(3,4,6)$. Unfortunately, the realization of this equilibrium is problematic because for the third player the first strategy (the profile $(2,1,1)$) give the same value for the cost function as the second strategy (the profile $(2,1,2)$). If, at the last stage, the third player will choose the first strategy, then at the profile $(2,1,1)$ the values of the cost functions are $(+\infty,4,6)$. It's a disaster result for the first player. $\Box$

Theorem 2.2. *If every strategy set $X_p \subset R^{k_p}, p = \overline{1,n}$ is compact and every cost function $f_p(x_1,...,x_p,...,x_n), p = \overline{1,n}$ is continuous by $(x_p,...,x_n)$ on $X_p \times \cdots \times X_n$ for every fixed $x_1 \in X_1,...,x_{p-1} \in X_{p-1}$, then the unsafe Stackelberg equilibria set $\hat{X}$ is non empty.*

The proof follows from the well known Calculus' Weierstrass theorem.

Theorem 2.3. *If every strategy set $X_p \subseteq R^{k_p}, p = \overline{1,n}$ is convex and every cost function $f_p(x_1,...,x_p,...,x_n), p = \overline{1,n}$ is strict convex by $(x_p,...,x_n)$ on $X_p \times \cdots \times X_n$ for every fixed $x_1 \in X_1,...,x_{p-1} \in X_{p-1}$, then the game has a unique unsafe Stackelberg equilibrium with the "guaranteed" realization property.*

"Guaranteed" realization means that the unsafe Stackelberg equilibrium is strictly preferred for the all players. The proof follows from the properties of the strict convex functions.

Remark that the strict convex requirements in the theorem 3 may be substituted by unimodal or other requirements that guaranteed the mono-valued characteristics of the involved best-move mappings.

3. SAFE STACKELBERG EQUILIBRIUM

In order to exclude the case illustrated in the continuation of Example 1 the notion of the safe Stackelberg equilibrium is introduced, which is equivalent with respective notion in [3].

By backward induction, every player $2, ..., n$ computes his best move mapping and the first player computes the set of his best moves:

$$B_n(x_1, ..., x_{n-1}) = \operatorname*{Arg\,min}_{y_n \in X_n} f_n\left(x_1, ..., x_{n-1}, y_n\right),$$

$$\tilde{B}_{n-1}(x_1, ..., x_{n-2}) = \operatorname*{Arg\,\underset{y_{n-1}}{\min}\,\underset{y_n}{\max}}_{(x_1,...,x_{n-2},y_{n-1},y_n)\in Gr_n} f_{n-1}\left(x_1, ..., x_{n-2}, y_{n-1}, y_n\right),$$

$$\tilde{B}_{n-2}(x_1, ..., x_{n-3}) = \operatorname*{Arg\,\underset{y_{n-2}}{\min}\,\underset{y_{n-1},y_n}{\max}}_{(x_1,...,x_{n-3},y_{n-2},...,y_n)\in \tilde{G}r_{n-1}} f_{n-1}\left(x_1, ..., x_{n-3}, y_{n-2}, ..., y_n\right),$$

$$\dots$$

$$\tilde{B}_2(x_1) = \operatorname*{Arg\,\underset{y_2}{\min}\,\underset{y_3,...,y_n}{\max}}_{(x_1,y_2,...,y_n)\in \tilde{G}r_3} f_2\left(x_1, y_2, ..., y_n\right),$$

$$\tilde{X} = \operatorname*{Arg\,\underset{y_1}{\min}\,\underset{y_2,...,y_n}{\max}}_{(y_1,...,y_n)\in \tilde{G}r_2} f_1\left(y_1, ..., y_n\right)$$

where

$$Gr_n = \{x \in X : x_1 \in X_1, ..., x_{n-1} \in X_{n-1}, x_n \in B_n(x_1, ..., x_{n-1})\},$$
$$\tilde{G}r_{n-1} = \left\{x \in Gr_n : x_1 \in X_1, ..., x_{n-2} \in X_{n-2}, (x_{n-1}, x_n) \in \tilde{B}_{n-1}(x_1, ..., x_{n-2})\right\},$$
$$\dots$$
$$\tilde{G}r_2 = \left\{x \in Gr_3 : x_1 \in X_1, (x_2, ..., x_n) \in \tilde{B}_2(x_1)\right\}.$$

Obviously, $\tilde{G}r_2 \subseteq \tilde{G}r_3 \subseteq \cdots \subseteq \tilde{G}r_{n-1} \subseteq Gr_n$, too.

Definition. *The profile $\tilde{x} \in \tilde{X}$ of the game is named a safe (pessimistic, weak) Stackelberg equilibrium.*

In general, the unsafe Stackelberg equilibria set is not equivalent to a safe Stackelberg equilibria set, *i.e.* $\hat{X} \neq \tilde{X}$. Remark that in Example 1 modified game the profile $(1, 1, 1)$ (with the costs $(5, 5, 2)$) is a safe Stackelberg equilibrium. The security's *"payment"* for the first player is *"supported"* also by the second player because the value of his cost function increases, too.

In addition, for $n = 2$ the safe Stackelberg equilibrium is not always an unsafe Stackelberg equilibrium.

Theorems 1−3 analogs for safe Stackelberg equilibrium may be formulated and proved. In the conditions of Theorem 3 the unsafe and safe Stackelberg equilibria are identical.

4. PSEUDO-EQUILIBRIUM. NASH-STACKELBERG EQUILIBRIUM

Consider the noncooperative strategic form game

$$\Gamma = \langle N, \{X_p^l\}_{l \in S, p \in N_l}, \{f_p^l(x)\}_{l \in S, p \in N_l} \rangle,$$

where

- $S = \{1, 2, ..., s\}$ is a set of stages,
- $N_l = \{1, 2, ..., n_l\}$ is a set of players at stage (level) $l \in S$,
- $X_p^l \subseteq R^{k_p^l}$ is a set of strategies of player $p \in N_l$ at stage $l \in S$,
- $s < +\infty, \quad n_l < +\infty, l \in S$,
- and $f_p^l(x)$ is a l^{th} stage p^{th} player cost function defined on the Cartesian product $X = \underset{p \in N_l, l \in S}{\times} X_p^l$.

Elements $x = (x_1^1, x_2^1, ..., x_{n_1}^1, x_1^2, x_2^2, ..., x_{n_2}^2, ..., x_1^s, x_2^s, ..., x_{n_s}^s) \in X$ are named profiles of the game.

Suppose that the players make their moves hierarchically:

> at the first stage players $1, 2, ..., n_1$ selects their strategies $x_1^1 \in X_1^1$, $x_2^1 \in X_2^1$, ..., $x_{n_1}^1 \in X_{n_1}^1$ simultaneously and communicate it to the second stage players $1, 2, ..., n_2$,
>
> the second stage players $1, 2, ..., n_2$ select simultaneously their strategies $x_1^2 \in X_1^2$, $x_2^2 \in X_2^2, ..., x_{n_2}^2 \in X_{n_2}^2$ and communicate the result to the third stage players,
>
> ...
>
> the s^{th} stage players $1, 2, ..., n_s$ select simultaneously their strategies $x_1^s \in X_1^s, x_2^s \in X_2^s, ..., x_{n_s}^s \in X_{n_s}^s$ at the last.

On the resulting profile $x = (x_1^1, x_2^1, ..., x_{n_1}^1, x_1^2, x_2^2, ..., x_{n_2}^2, ..., x_1^s, x_2^s, ..., x_{n_s}^s)$ every player computes the value of his cost function.

Suppose that the l^{th} stage p^{th} player has all information about all strategy sets and the cost functions of players of stages $l, l+1, ..., s$. Without loss of generality **suppose that all players minimize the values of their cost functions.**

Definition. *The profile $\hat{x} \in X$ of the game is a pseudo-equilibrium if for every $l \in S$ there exist $y^{l+1} \in X^{l+1}, ..., y^n \in X^n$ such that*

$$f_p^l(\hat{x}^1, ..., \hat{x}^{l-1}, x_p^l \| \hat{x}_{-p}^l, y^{l+1}, ..., y^n) \geq f_p^l(\hat{x}^1, ..., \hat{x}^l, y^{l+1}, ..., y^n), \forall x_p^l \in X_p^l, \ \forall p \in N_l,$$

where $\hat{x}_{-p}^l = (\hat{x}_1^l, ..., \hat{x}_{p-1}^l, \hat{x}_{p+1}^l, ..., \hat{x}_{n_l}^l)$.

According to the definition, players $1, 2, ..., n_l$, $l = 1, 2, ..., s-1, s$ select their pseudo-equilibrium strategies:

$$B_p^1(\chi_{-p}^1) = \operatorname*{Arg\,min}_{x_p^1 \in X_p^1} f_p^1\left(x_p^1 \| \chi_{-p}^1\right), p \in N_1,$$

$$(\hat{x}^1, x^2, ..., x^s) \in PE^1 = \bigcap_{p \in N_1} Gr_p^1,$$

$$B_p^2(\hat{x}^1, \chi_{-p}^2) = \operatorname*{Arg\,min}_{x_p^2 \in X_p^2} f_p^2\left(\hat{x}^1, x_p^2 \| \chi_{-p}^2\right), p \in N_2,$$

$$(\hat{x}^1, \hat{x}^2, x^3, ..., x^s) \in PE^2 = \bigcap_{p \in N_2} Gr_p^2,$$

$$\cdots$$

$$B_p^s(\hat{x}^1, \hat{x}^2, ..., \hat{x}^{s-1}, \chi_{-p}^s) = \operatorname*{Arg\,min}_{x_p^s \in X_p^s} f_p^s\left(\hat{x}^1, ..., \hat{x}^{s-1}, x_p^s \| \chi_{-p}^s\right), p \in N_s,$$

$$(\hat{x}^1, \hat{x}^2, ..., \hat{x}^s) \in PE^s = \bigcap_{p \in N_s} Gr_p^s,$$

where

$$Gr_p^1 = \left\{(x^1, ..., x^s) : x_p^1 \in B_p^1(\chi_{-p}^1)\right\}, p \in N_1,$$

$$Gr_p^2 = \left\{(\hat{x}^1, x^2, ..., x^s) : x_p^2 \in B_p^2(\hat{x}^1, \chi_{-p}^2)\right\}, p \in N_2,$$

$$\cdots$$

$$Gr_p^s = \left\{(\hat{x}^1, ..., \hat{x}^{s-1}, x^s) : x_p^s \in B_p^s(\hat{x}^1, ..., \hat{x}^{s-1}, \chi_{-p}^s)\right\}, p \in N_s,$$

$$\chi_{-p}^l = (x_{-p}^l, x^{l+1}, ..., x^s) \in X_{-p}^l \times X^{l+1} \times \cdots \times X^s,$$

$$X_{-p}^l = X_1^l \times ... \times X_{p-1}^l \times X_{p+1}^l \times ... \times X_{n_l}^l.$$

Surely, **the set of all pseudo-equilibria is** $PE = PE^s$**.**

Example 2. Consider a four-player $2 \times 2 \times 2 \times 2$ two-stage game with the cost matrices:

$$a_{**11} = \begin{bmatrix} \mathbf{5} & 8 \\ 6 & \mathbf{4} \end{bmatrix}, \quad a_{**12} = \begin{bmatrix} 6 & \mathbf{4} \\ \mathbf{5} & 9 \end{bmatrix}, \quad a_{**21} = \begin{bmatrix} 3 & 2 \\ \mathbf{1} & \mathbf{1} \end{bmatrix}, \quad a_{**22} = \begin{bmatrix} \mathbf{2} & 3 \\ 3 & \mathbf{2} \end{bmatrix},$$

$$b_{**11} = \begin{bmatrix} \mathbf{4} & 7 \\ \mathbf{6} & 8 \end{bmatrix}, \quad b_{**12} = \begin{bmatrix} \mathbf{3} & 7 \\ 6 & \mathbf{4} \end{bmatrix}, \quad b_{**21} = \begin{bmatrix} 7 & \mathbf{3} \\ 5 & \mathbf{3} \end{bmatrix}, \quad b_{**22} = \begin{bmatrix} 8 & \mathbf{3} \\ \mathbf{1} & 9 \end{bmatrix},$$

$$c_{**11} = \begin{bmatrix} 5 & 2 \\ 4 & 5 \end{bmatrix}, \quad c_{**12} = \begin{bmatrix} -1 & 3 \\ 3 & 2 \end{bmatrix}, \quad c_{**21} = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}, \quad c_{**22} = \begin{bmatrix} 3 & 5 \\ 4 & -2 \end{bmatrix},$$

$$d_{**11} = \begin{bmatrix} 6 & 2 \\ 1 & 3 \end{bmatrix}, \quad d_{**12} = \begin{bmatrix} -2 & 3 \\ 3 & 1 \end{bmatrix}, \quad d_{**21} = \begin{bmatrix} 7 & 2 \\ 1 & 3 \end{bmatrix}, \quad d_{**22} = \begin{bmatrix} 6 & 3 \\ 3 & -1 \end{bmatrix}.$$

The first and the second players move at the first stage, the third and the fourth players move at the second stage.

Elements of Gr_1^1 and Gr_2^1 are emphasized in matrices by bold fonts.

Obviously, PE^1 consists of two elements: $(1,1,1,1)$ and $(2,2,2,1)$.

Suppose, the first and the second players choose the strategies 1 and 1.

At the second stage, the third and the fourth players have to play a matrix games with matrices

$$c_{11**} = \begin{bmatrix} 5 & \mathbf{-1} \\ \mathbf{1} & 3 \end{bmatrix}, \quad d_{11**} = \begin{bmatrix} 6 & \mathbf{-2} \\ 7 & \mathbf{6} \end{bmatrix}.$$

PE^2 contains a profile $(1,1,1,2)$. Thus, the profile $(1,1,1,2)$ with costs $(6,3,-1,-2)$ is a two-stage pseudo-equilibrium.

Suppose, the first and the second players choose the strategies 2 and 2.

Then, at the second stage, the third and the fourth players have to play a matrix game with matrices

$$c_{22**} = \begin{bmatrix} 5 & 2 \\ \mathbf{3} & \mathbf{-2} \end{bmatrix}, \quad d_{22**} = \begin{bmatrix} 3 & \mathbf{1} \\ 3 & \mathbf{-1} \end{bmatrix}.$$

The set PE^2 contains profile $(2,2,2,2)$. Thus, the profile $(2,2,2,2)$ with costs $(2,9,-2,-1)$ is a two-stage pseudo-equilibrium.

Obviously, both pseudo-equilibria $(1, 1, 1, 2)$ and $(2, 2, 2, 2)$ aren't Nash equilibrium. □

The pseudo-equilibrium definition does not used the information that at the following stage the stage players will choose the strategy accordingly the pseudo-equilibrium statement. As the result, the profiles do not safe the required statement at all stages.

For excluding this inconvenient, it's reasonable to choose strategies by backward induction procedure and thus we obtain a new equilibrium notion.

By stage backward induction, players $1, 2, ..., n_l$, $l = s, s-1, ..., 2, 1$ select their equilibrium strategies

$$B_p^s(x^1, ..., x^{s-1}, x^s_{-p}) = \operatorname*{Arg\,min}_{y_p^s \in X_p^s} f_p^s\left(x^1, ..., x^{s-1},\, y_p^s \| x^s_{-p}\right),\ p \in N_s,$$

$$NSE^s = \bigcap_{p \in N_s} Gr_p^s,$$

$$B_p^{s-1}(x^1, ..., x^{s-2}, x^{s-1}_{-p}) =$$

$$= \operatorname*{Arg\,min}_{\substack{y_p^{s-1},\, y^s: \\ (x^1,...,x^{s-2},\, y_p^{s-1} \| x^{s-1}_{-p},\, y^s) \in NSE^s}} f_p^{s-1}\left(x^1, ..., x^{s-2}, y_p^{s-1} \| x^{s-1}_{-p}, y^s\right), p \in N_{s-1},$$

$$NSE^{s-1} = \bigcap_{p \in N_{s-1}} Gr_p^{s-1},$$

$$B_p^{s-2}(x^1, ..., x^{s-2}, x^{s-3}_{-p}) =$$

$$= \operatorname*{Arg\,min}_{\substack{y_p^{s-2},\, y^{s-1},\, y^s: \\ (x^1,...,x^{s-3},\, y_p^{s-2} \| x^{s-2}_{-p},\, y^{s-1},\, y^s) \in NSE^{s-1}}} f_p^{s-2}\left(x^1, ..., x^{s-3}, y_p^{s-2} \| x^{s-2}_{-p}, y^{s-1}, y^s\right),$$

$p \in N_{s-2}$,

$$NSE^{s-2} = \bigcap_{p \in N_{s-2}} Gr_p^{s-2},$$

$$\cdots$$

$$B_p^1(x^1_{-p}) = \operatorname*{Arg\,min}_{y_p^1,\, y^2,...,y^s : (y_p^1 \| x^1_{-p},\, y^2,...,y^s) \in NSE^2} f_p^1\left(y_p^1 \| x^1_{-p},\, y^2, ..., y^s\right), p \in N_1,$$

$$NSE^1 = \bigcap_{p \in N_1} Gr_p^1,$$

where

$$Gr_p^s = \left\{ x \in X : \begin{array}{c} x^l \in X^l,\ l = \overline{1, s-1}, \\ x_{-p}^s \in X_{-p}^s, \\ x_p^s \in B_p^s(x^1, ..., x^{s-1}, x_{-p}^s) \end{array} \right\},\ p \in N_s,$$

$$Gr_p^{s-1} = \left\{ x \in NSE^s : \begin{array}{c} x^l \in X^l, l = \overline{1, s-2}, \\ x_{-p}^{s-1} \in X_{-p}^{s-1}, \\ x_p^{s-1} \in B_p^{s-1}(x^1, ..., x^{s-2}, x_{-p}^{s-1}) \end{array} \right\},\ p \in N_{s-1},$$

$$\dots$$

$$Gr_p^1 = \left\{ x \in NSE^2 : \begin{array}{c} x_{-p}^1 \in X_{-p}^1, \\ x_p^1 \in B_p^1(x_{-p}^1) \end{array} \right\},\ p \in N_1.$$

Of course, $NSE^1 \subseteq NSE^2 \subseteq \cdots \subseteq NSE^s$.

Definition. *Every element of the NSE^1 is called a Nash-Stackelberg equilibrium.*

The set of all Nash-Stackeberg equilibria NSE^1 is denoted by NSE also.

If $s = 1$ and $n_1 > 1$, then every Nash-Stackelberg equilibrium is the Nash equilibrium. If $s > 1$ and $n_1 = n_2 = ... = n_s = 1$, then every equilibrium is an unsafe Stackelberg equilibrium. Thus, the Nash-Stackelberg equilibrium notion generalizes the both Stackelberg and Nash equilibria notions.

Example 2 (the first continuation). For the NSE illustration consider the same four-player 2 × 2 × 2 × 2 two-stage game with the cost matrices as in the example 2. It's opportune to represent the cost matrices as follows

$$a_{11**} = \begin{bmatrix} 5 & 6 \\ 3 & 2 \end{bmatrix},\ a_{12**} = \begin{bmatrix} 8 & 4 \\ 2 & 3 \end{bmatrix},\ a_{21**} = \begin{bmatrix} 6 & 5 \\ 1 & 3 \end{bmatrix},\ a_{22**} = \begin{bmatrix} 4 & 9 \\ 1 & 2 \end{bmatrix},$$

$$b_{11**} = \begin{bmatrix} 4 & 3 \\ 7 & 8 \end{bmatrix},\ b_{12**} = \begin{bmatrix} 7 & 7 \\ 3 & 3 \end{bmatrix},\ b_{21**} = \begin{bmatrix} 6 & 6 \\ 5 & 1 \end{bmatrix},\ b_{22**} = \begin{bmatrix} 8 & 4 \\ 3 & 9 \end{bmatrix},$$

$$c_{11**} = \begin{bmatrix} 5 & \mathbf{-1} \\ \mathbf{1} & 3 \end{bmatrix},\ c_{12**} = \begin{bmatrix} \mathbf{2} & 3 \\ \mathbf{2} & 5 \end{bmatrix},\ c_{21**} = \begin{bmatrix} 4 & \mathbf{3} \\ \mathbf{4} & \mathbf{4} \end{bmatrix},\ c_{22**} = \begin{bmatrix} 5 & \mathbf{2} \\ 3 & \mathbf{-2} \end{bmatrix},$$

$$d_{11**} = \begin{bmatrix} 6 & \mathbf{-2} \\ 7 & \mathbf{6} \end{bmatrix},\ d_{12**} = \begin{bmatrix} \mathbf{2} & 3 \\ \mathbf{2} & 3 \end{bmatrix},\ d_{21**} = \begin{bmatrix} \mathbf{1} & 3 \\ \mathbf{1} & 3 \end{bmatrix},\ d_{22**} = \begin{bmatrix} 3 & \mathbf{1} \\ 3 & \mathbf{-1} \end{bmatrix}.$$

Elements of Gr_3^2, Gr_4^2 are emphasized in matrices by bold fonts.

NSE^2 consists of profiles: $(1,1,1,2), (1,2,1,1), (1,2,2,1), (2,1,2,1), (2,2,1,2),$ $(2,2,2,2)$.

At the first stage, the first and the second players have to play a specific matrix game with incomplete matrices

$$a_{11**} = \begin{bmatrix} \cdot & 6 \\ \cdot & \cdot \end{bmatrix}, \quad a_{12**} = \begin{bmatrix} 8 & \cdot \\ \mathbf{2} & \cdot \end{bmatrix}, \quad a_{21**} = \begin{bmatrix} \cdot & \cdot \\ \mathbf{1} & \cdot \end{bmatrix}, \quad a_{22**} = \begin{bmatrix} \cdot & 9 \\ \cdot & \mathbf{2} \end{bmatrix},$$

$$b_{11**} = \begin{bmatrix} \cdot & \mathbf{3} \\ \cdot & \cdot \end{bmatrix}, \quad b_{12**} = \begin{bmatrix} 7 & \cdot \\ \mathbf{3} & \cdot \end{bmatrix}, \quad b_{21**} = \begin{bmatrix} \cdot & \cdot \\ 5 & \cdot \end{bmatrix}, \quad b_{22**} = \begin{bmatrix} \cdot & \mathbf{4} \\ \cdot & 9 \end{bmatrix},$$

Elements of Gr_1^1, Gr_2^1 are emphasized in matrices by bold fonts.

NSE^1 consists only from one profile $(1,2,2,1)$ with the costs $(2,3,2,2)$.

Obviously, $PE = PE^2 \neq NSE = NSE^1$.

As can be seen, if the first and the second players choose their strategies 1 and 2 at the first stage, then the third and the fourth players can move also 1 and 2 at the second stage, because the profile $(1,2,1,2)$ will have for them the same costs 2 and 2. Unfortunately, for the first and the second players the cost will be 8 and 7, respectively. For the first and second players the profile $(1,2,1,2)$ is worse than $(1,2,2,1)$.

Thus, as in Stackelberg game, it's necessary and reasonable to introduce a notion of safe Nash-Stackelberg equilibrium. □

By stage backward induction, players $1,2,...,n_l$, $l = s, s-1, ..., 2, 1$ select their equilibrium strategies

$$B_p^s(x^1,...,x^{s-1},x_{-p}^s) = \operatorname*{Arg\,min}_{y_p^s \in X_p^s} f_p^s\left(x^1,...,x^{s-1},\, y_p^s \| x_{-p}^s\right), \, p \in N_s,$$

$$SNSE^s = NSE^s = \bigcap_{p \in N_s} Gr_p^s,$$

$$\tilde{B}_p^{s-1}(x^1,...,x^{s-2},x_{-p}^{s-1}) =$$

$$= \operatorname*{Arg\,min\,max}_{\substack{y_p^{s-1} \quad y^s \\ (x^1,...,x^{s-2},y_p^{s-1}\|x_{-p}^{s-1},y^s) \in SNSE^s}} f_p^{s-1}\left(x^1,...,x^{s-2},y_p^{s-1}\|x_{-p}^{s-1},y^s\right), p \in N_{s-1},$$

$$SNSE^{s-1} = \bigcap_{p \in N_{s-1}} \tilde{G}r_p^{s-1},$$

$$\tilde{B}_p^{s-2}(x^1,...,x^{s-2},x_{-p}^{s-3}) =$$
$$= \operatorname{Arg} \min_{\substack{y_p^{s-2} \\ (x^1,...,x^{s-3},y_p^{s-2}\|x_{-p}^{s-2},y^{s-1},y^s)\in NSE^{s-1}}} \max_{y^{s-1},y^s} f_p^{s-2}\left(x^1,...,x^{s-3},y_p^{s-2}\|x_{-p}^{s-2},y^{s-1},y^s\right),$$
$$p \in N_{s-2},\ SNSE^{s-2} = \bigcap_{p\in N_{s-2}} \tilde{G}r_p^{s-2},$$
$$\dots$$
$$\tilde{B}_p^1(x_{-p}^1) = \operatorname{Arg} \min_{\substack{y_p^1 \\ (y_p^1\|x_{-p}^1,y^2,...,y^s)\in NSE^2}} \max_{y^2,...,y^s} f_p^1\left(y_p^1\|x_{-p}^1,\ y^2,...,y^s\right), p \in N_1,$$
$$SNSE^1 = \bigcap_{p\in N_1} \tilde{G}r_p^1,$$

where

$$Gr_p^s = \left\{ x \in X : \begin{array}{c} x^l \in X^l,\ l = \overline{1,s-1}, \\ x_{-p}^s \in X_{-p}^s, \\ x_p^s \in B_p^s(x^1,...,x^{s-1},x_{-p}^s) \end{array} \right\}, p \in N_s,$$
$$\tilde{G}r_p^{s-1} = \left\{ x \in NSE^s : \begin{array}{c} x^l \in X^l, l = \overline{1,s-2}, \\ x_{-p}^{s-1} \in X_{-p}^{s-1}, \\ x_p^{s-1} \in \tilde{B}_p^{s-1}(x^1,...,x^{s-2},x_{-p}^{s-1}) \end{array} \right\}, p \in N_{s-1},$$
$$\dots$$
$$\tilde{G}r_p^1 = \left\{ x \in NSE^2 : \begin{array}{c} x_{-p}^1 \in X_{-p}^1, \\ x_p^1 \in \tilde{B}_p^1(x_{-p}^1) \end{array} \right\}, p \in N_1.$$

Surely, $SNSE^1 \subseteq SNSE^2 \subseteq \cdots \subseteq SNSE^s$.

Definition. *Elements of $SNSE^1$ are called safe Nash-Stackelberg equilibria.*

The set of all safe Nash-Stackeberg equilibria $SNSE^1$ is denoted by $SNSE$ also.

Example 2 (the second continuation). Remark, that at the second stage $SNSE^2 = NSE^2$.

At the first stage, the first and the second players have to play a specific matrix game with the same incomplete matrices

$$a_{11**} = \begin{bmatrix} \cdot & 6 \\ \cdot & \cdot \end{bmatrix}, \quad a_{12**} = \begin{bmatrix} \mathbf{8} & \cdot \\ 2 & \cdot \end{bmatrix}, \quad a_{21**} = \begin{bmatrix} \cdot & \cdot \\ \mathbf{1} & \cdot \end{bmatrix}, \quad a_{22**} = \begin{bmatrix} \cdot & 9 \\ \cdot & 2 \end{bmatrix},$$

$$b_{11**} = \begin{bmatrix} \cdot & \mathbf{3} \\ \cdot & \cdot \end{bmatrix}, \quad b_{12**} = \begin{bmatrix} 7 & \cdot \\ 3 & \cdot \end{bmatrix}, \quad b_{21**} = \begin{bmatrix} \cdot & \cdot \\ \mathbf{5} & \cdot \end{bmatrix}, \quad b_{22**} = \begin{bmatrix} \cdot & 4 \\ \cdot & 9 \end{bmatrix},$$

The elements of $\tilde{Gr}_1^1$, $\tilde{Gr}_2^1$ are emphasized in matrices by bold fonts. (In every 2 × 2 matrix the maximal element is selected. After that, the first player with the fixed strategy of the second player, chooses the matrix with the minimal among the maximal selected elements. In the same manner, the second player constructs his graph. The graphs intersection consists of a single element.)

$SNSE^1$ consists only of one profile $(2,1,2,1)$ with costs $(1,5,4,1)$.

Consequently, for the considered game there are two PE $(1,1,1,2)$ and $(2,2,2,2)$ with costs $(6,3,-1,-2)$ and $(2,9,-2,-1)$, respectively, one NSE $(1,2,2,1)$ with costs $(2,3,2,2)$ and one $SNSE$ $(2,1,2,1)$ with costs $(1,5,4,1)$. None of these profiles are better for all players. The problem of the equilibrium principle selection is opened. □

5. MULTI-OBJECTIVE PSEUDO-EQUILIBRIUM. PARETO-NASH-STACKELBERG EQUILIBRIUM

Consider the multiobjective strategic form game

$$\Gamma = \langle N, \{X_p^l\}_{l\in S, p\in N_l}, \{\mathbf{f}_p^l(x)\}_{l\in S, p\in N_l}\rangle,$$

with vector cost functions $\{\mathbf{f}_p^l(x)\}_{l\in S, p\in N_l}$.

In the same manner as for the Nash-Stackelberg game the equilibrium principles can be introduced. An essential difference in corresponding definitions is the strong requirement that every minimization or maximization operator must be interpreted as Pareto maximization or minimization operator. Evidently, the Pareto optimal response mapping and the graph of the Pareto optimal response mapping are considered for every player. An intersection of the graphs of Pareto optimal response mappings is considered in every definition as the stage profile.

6. CONCLUSIONS

The examined processes of decision making are very often phenomena of real life. Their mathematical moddeling as Pareto-Nash-Stackelberg games gives an powerful tool for investigation, analysis and solving hierarchical decision problems. Nevertheless, the problem of equilibrium principle choosing in real situations is a task for both a decision maker and a game theorist.

It is interesting to mention that the proposed models based on graphs of players optimal response mappings, give us the recurrent relations, similarly as that used traditionally on dynamic programming. It is obvious that a concrete n stage dynamic problem solving may be considered, for example, as an n players Stackelberg game, such that at every stage the same player moves, using the same objective function.

References

[1] Blackwell, D., *An analog of the minimax theorem for vector payoffs,* Pacific Journal of Mathematics, **6**(1956), 1-8.

[2] Born, P., Tijs, S., van den Aarssen, J. *Pareto equilibria in multiobjective games,* Methods of Operations Research, **60**(1988), 302-312.

[3] Leitmann, G., *On Generalized Stackelberg Strategies,* Journal of Optimization Theory and Applications, **26**(1978), 637-648.

[4] Nash, J. F., *Noncooperative game,* Annals of Mathematics, **54**(1951), 280-295.

[5] Podinovskii, V. V., Nogin V. D., *Pareto-optimal solutions of the multi-criteria problems,* Moscow, Nauka, 1982. (Russian)

[6] Sagaidac, M., Ungureanu, V., *Operational research,* Chişinău, CEP USM, 2004. (Romanian).

[7] Shapley, L. S., *Equilibrium Points in Games with Vector Payoffs,* Naval Research Logistics Quarterly, **6**(1959), 57-61.

[8] Ungureanu, V., *Mathematical programming*, Chişinău, USM, 2001. (Romanian)

[9] Ungureanu, V., Botnari, A., *Nash equilibria sets in mixed extension of* $2 \times 2 \times 2$ *games,* Computer Science Journal of Moldova, **13**, 1 (37)(2005), 13-28.

[10] Ungureanu, V., Botnari, A., *Nash equilibria sets in mixed extended* 2×3 *games,* Computer Science Journal of Moldova, **13**, 2 (38)(2005), 136-150.

[11] Ungureanu, V., *A method for Nash equilibria set computing in multi-matrix extended games,* Spanish-Italian-Netherlands Game Theory Meeting (SING2) and XVI Italian Meeting on Game Theory (XVI IMGTA), Foggia, Italy, June 14-17, 2006, Conference Book, 2006, p. 156.

[12] Ungureanu, V., *Nash equilibria set computing in finite extended games* Computer Science Journal of Moldova, **14**, 3 (42)(2006), 345-365.

[13] Ungureanu, V., *Nash Equilibrium Conditions for Normal Form Strategic Games,* The XIV-th Conference on Applied and Industrial Mathematics dedicated to the 60-th anniversary of the foundation of the Faculty of Mathematics and Computer Science of Moldova State University (CAIM XIV), Satellite Conference of ICM 2006, Chisinău, 17-19 august, 2006, 328-330.

[14] Ungureanu, V., *Nash Equilibrium Conditions — Extensions of Some Classical Theorems,* CODE-2007: Conference de la SMAI sur l'Optimisation et la Decision, Colloque satellite de la conference SMAI-2007, Institut Henri Poincaré, Paris, 18-20 avril 2007, http://www.ann.jussieu.fr/∼plc/code2007/ungureanu.pdf.

[15] Ungureanu, V., *Equilibrium principles for Pareto-Nash-Stackelberg games,* 6th Congress of Romanian Mathematicians, June 28 - July 4, 2007, University of Bucharest, Romania, Abstracts, 164-165.

[16] Ungureanu, V., *Equilibrium principles for Pareto-Nash-Stackelberg games,* Spain Italy Netherlands Meeting on Game Theory (SING3) and the 7th Spanish Meeting on Game Theory, 4-5-6 July, 2007, Facultad de Ciencias Matemáticas, Universidad Complutense, Madrid, Spain, Abstract Book, 81.

[17] Ungureanu, V., *Nash equilibrium conditions for strategic form games,* 6th International Congress on Industrial and Applied Mathematics, ETH and University of Zürich, Switzerland, July 16-20, 2007, Abstract Book, 436.

[18] Ungureanu, V., *Set of Nash equilibria in 2x2 mixed extended games,* The Wolfram Demonstrations Project
http://demonstrations.wolfram.com/SetOfNashEquilibriaIn2x2MixedExtendedGames/
Posted November 17, 2007.

[19] Ungureanu, V., *Nash Equilibrium Conditions for Strategic Form Games,* Libertas Mathematica, **XXVII**(2007), 131-140.

[20] Ungureanu, V., *Solution principles for generalized Stackelberg games,* Modelare Matematică, Optimizare şi Tehnologii Informaţionale, Materialele Conferinţei Internaţionale, ATIC, 19-21 martie 2008, Chişinău, Evrica, 2008, 181-189.

[21] Von Stackelberg, H., *Marktform und Gleichgewicht,* Springer Verlag, Vienna, 1934.

ROMAI J., **4**, 1(2008), 243–260

SOME CLASSES OF SOLUTIONS OF GAS DYNAMIC EQUATIONS

Petr A. Velmisov, Julia A. Kazakova

Ulyanovsk State Technical University, Russia

velmisov@ulstu.ru

Abstract In this work the solutions of gas dynamic equations are constructed by the parametric method, their classification is carried out, and the applications are showed for solutions of specific physical problems. In particular, the solutions of simple and dual waves are derived. The solutions, describing gas flows with local supersonic zones in Laval nozzles, are constructed too.

Keywords: gas dynamics, partial differential equations.

2000 MSC: 76N15.

1. METHOD DESCRIPTION

Consider a system of partial differential equations

$$F_k(x_1,...,x_m,u_1,...,u_n,u_{1x_1},...,u_{nx_1},...,u_{1x_m},...,u_{nx_m})=0,\ k=1\div n, \quad (1)$$

where $u_k(x_1,...,x_m)$ are functions of m variables x_1,x_2,...,x_m. After passing to new variables $\xi_1,\xi_2,...,\xi_m$, which are functions of the variables x_1,x_2,...,x_m, the solution of system (1) is found in the following form

$$u_k=U_k(\xi_1,\xi_2,...,\xi_m),\ k=1\div n; \qquad x_l=X_l(\xi_1,\xi_2,...,\xi_m),\ l=1\div m. \quad (2)$$

The partial derivatives $\dfrac{\partial u_k}{\partial x_l}$ are found by the formulas

$$\frac{\partial u_k}{\partial x_l}=\sum_{j=1}^{m}\frac{\partial U_k}{\partial \xi_j}\frac{\partial \xi_j}{\partial x_l}, \qquad \frac{\partial \xi_j}{\partial x_l}=\frac{\Delta_{jl}}{\Delta}, \qquad \Delta=\begin{vmatrix}\dfrac{\partial X_1}{\partial \xi_1} & \cdots\dfrac{\partial X_1}{\partial \xi_m}\\ \cdots & \cdots\\ \dfrac{\partial X_m}{\partial \xi_1} & \cdots\dfrac{\partial X_m}{\partial \xi_m}\end{vmatrix}\neq 0. \quad (3)$$

In (3) Δ_{jl} is a determinant derived from the determinant Δ by replacement of the j^{th} column by a column with zero elements except the element with number l which is equal one. Then system (1) is rearranged in the form

$$F_k(\xi_1,...,\xi_m,X_1,...,X_m,X_{1\xi_1},...X_{m\xi_m},U_1,...,U_n,U_{1\xi_1},...U_{n\xi_m})=0. \tag{4}$$

In this system $U_k(k=1\div n)$, $X_l(l=1\div m)$ are functions of variables $\xi_1,\xi_2,...,\xi_m$. The solution of the given system of equations can be found in polynomial form

$$\begin{aligned} U_k&=\sum_{i=0}^{\alpha_k}u_{ki}(\xi_1,...,\xi_{s-1},\xi_{s+1},...\xi_m)\xi_s^i,\ k=1\div n,\ s=1\div m,\\ X_l&=\sum_{j=0}^{\gamma_l}x_{lj}(\xi_1,...,\xi_{s-1},\xi_{s+1},...\xi_m)\xi_s^j,\ l=1\div m,\ s=1\div m,\end{aligned} \tag{5}$$

where $\alpha_k,\gamma_l\in N$ (N is the set of natural numbers). Particularly, the problem is the determination of parameters $\alpha_k,\gamma_l\in N$, for which the system of differential equations for $u_{ki}(\xi_1,...\xi_{m-1})$, $x_{lj}(\xi_1,...\xi_{m-1})$ is determined or underdetermined.

Consider the particular case, when the required functions depend on the coordinates x, y and the time t

$$F_k(x,y,t,u_1,...,u_n,u_{1x},...,u_{nx},u_{1y},...,u_{ny},u_{1t},...,u_{nt})=0,\ k=1\div n. \tag{6}$$

In this case the solution of the system is

$$u_k=u_k(\xi,\eta,t),\ k=1\div n,\quad x=x(\xi,\eta,t),\quad y=y(\xi,\eta,t). \tag{7}$$

The formulae of conversion to new variables are

$$\begin{aligned} u_{kx}&=\frac{u_{k\xi}y_\eta-u_{k\eta}y_\xi}{\Delta},\ u_{ky}=\frac{u_{k\eta}x_\xi-u_{k\xi}x_\eta}{\Delta},\\ u_{kt}&=u_{kt}+\frac{u_{k\xi}(y_tx_\eta-y_\eta x_t)+u_{k\eta}(y_\xi x_t-y_tx_\xi)}{\Delta},\end{aligned} \tag{8}$$

where $\Delta=x_\xi y_\eta-x_\eta y_\xi\neq 0$. The system (6) is rearranged in the form

$$F_k(\xi,\eta,x,y,t,x_\xi,x_\eta,x_t,y_\xi,y_\eta,y_t,u_1,...,u_n,u_{1\xi},...,u_{nt})=0. \tag{9}$$

In this system $x,y,u_k(k=1\div n)$ are functions of variables ξ,η,t. The solution of system (9) can be found in polynomial form

$$u_k=\sum_{i=0}^{\alpha_k}u_{ki}(\xi,t)\eta^i,\ x=\sum_{k=0}^{\gamma}x_k(\xi,t)\eta^k,\ y=\sum_{k=0}^{\omega}y_k(\xi,t)\eta^k, \tag{10}$$

where α_k,γ,$\omega\in N$. For some types of equations (for example, for quasi-linear equations of first order, in which coefficients are the polynomials relative to dependent and independent variables) the relations for parameters α_k,γ,$\omega$$\in N$ are obtained. These relations allow us to get all values of parameters α_k,γ,$\omega$$\in N$, for which the differential equation system in $x_k(\xi,t)$, $y_k(\xi,t)$, $u_{ki}(\xi,t)$ is determined or underdetermined, that is $s \geq r$, $s = r + j$, where s is the number of unknown functions dependent on ξ, t; r is the number of equations; j is a degree of sub-definite. If r_k is the maximum degree of variable η (when we substitute (10) in the equation $F_k = 0(k = 1 \div n)$), parameters s and r are

$$r = \sum_{k=1}^{n} r_k + n, \qquad s = \gamma + \omega + \sum_{k=1}^{n} \alpha_k + n + 2. \tag{11}$$

The program has been developed, by means of which we can determine the allowed values of parameters α_k, γ, $\omega \in N$. Having substituted (10) in (9), we put in set all maximum degrees of variables η for each k-th equation of system (9)

$$(J_{1,1}, J_{2,1}, ..., J_{i_1,1}, ..., J_{1,k}, J_{2,k}, ..., J_{i_k,k}, ..., J_{1,n}, ..., J_{i_n,n}),$$

where i_k the number of various degrees η in the k-th equation, $k = 1 \div n$. Each of values $J_{1,1}, J_{2,1}, ..., J_{i_1,1}, ..., J_{1,k}, J_{2,k}, ..., J_{i_k,k}, ..., J_{1,n}, ..., J_{i_n,n}$ is a linear combination of parameters α_k, γ, ω. Then the following system is formed and solved

$$\begin{cases} I_1 - J_{1,1} = x_{1,1}, \\ \dots\dots\dots\dots\dots\dots, \\ I_1 - J_{i_1,1} = x_{i_1,1}, \\ \dots\dots\dots\dots\dots\dots, \\ I_n - J_{1,n} = x_{1,n}, \\ \dots\dots\dots\dots\dots\dots, \\ I_n - J_{i_n,n} = x_{i_n,n}, \\ \sum\limits_{k=1}^{n} \alpha_k + \gamma + \omega + 2 + n = \sum\limits_{k=1}^{n} I_k + j, \end{cases} \tag{12}$$

where $I_k = max(J_{1,k}, ..., J_{i_k,k})$ is the maximum degree of η in the k-th equation; $x_{i_k,k} \in N$ are some natural numbers; j is a degree of sub-definite.

There exist $d = \prod\limits_{k=1}^{n} i_k$ variants of expression choice I_k.

In this program we formed n matrices $(S_1, S_2, ..., S_n)$. Each row of the k-th matrix consists of the coefficients α_k, γ, ω in degrees of variable η k-th equation (for example, $S_k = (J_{1,k}, ..., J_{i_k,k})^T$). The row number k-th matrix is i_k. There is realized a search cycle of maximum I_k degrees, and there are chosen parameters α_k, γ, ω for each set $\{I_k\}_{k=1\div n}$. From $\alpha_k, \gamma, \omega \in N, x_{i_k,k} \in N$ and the system (12), the parameters α_k, γ, ω are bounded. At every step of search the matrix (A) is formed, elements of which are coefficients of α_k, γ, ω in equations of system (12)

$$\begin{gathered}
A[m_1, q] = S_1[l_1, q] - S_1[m_1, q], m_1 = 1 \div i_1, \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
A[m_n, q] = S_n[l_n, q] - S_n[m_n, q], m_n = 1 \div i_n, \\
A[\sum_{k=1}^{n} i_k, q] = 1 - S_1[l_1, q] - S_2[l_2, q] - ... - S_n[l_n, q],
\end{gathered} \tag{13}$$

where $q = 1 \div (n + 2)$; the set $(l_1, l_2, ..., l_n)$ determines the numbers of rows in matrices $S_1, S_2, ..., S_n$, corresponding to maximum degrees $(I_1, I_2, ..., I_n)$. Values $x_{i_k,k}$ are put in vector b form

$$b[i] = \sum_{k=1}^{n+2} A[i, k]\varphi_k, \varphi_k \in \{\alpha_k, \gamma, \omega\}, i = 1 \div \left(1 + \sum_{k=1}^{n} i_k\right).$$

The last element of vector b is

$$b\left[1 + \sum_{k=1}^{n} i_k\right] = b\left[1 + \sum_{k=1}^{n} i_k\right] + n + 2.$$

Then, for $\gamma + \omega > 0$ and $b[i] \in N$, $\forall i$, the given values of α_k, γ, ω, are admissible.

2. APPLICATION OF METHOD TO GAS DYNAMIC EQUATIONS

The system of gas dynamics for an ideal gas in case of adiabatic process is

$$\begin{cases}
\rho(u_t + uu_x + vu_y + wu_z) = -p_x, \\
\rho(v_t + uv_x + vv_y + wv_z) = -p_y, \\
\rho(w_t + uw_x + vw_y + ww_z) = -p_z, \\
\rho_t + (\rho u)_x + (\rho v)_y + (\rho w)_z = 0, \\
\left(\frac{p}{\rho^\chi}\right)_t + u\left(\frac{p}{\rho^\chi}\right)_x + v\left(\frac{p}{\rho^\chi}\right)_y + w\left(\frac{p}{\rho^\chi}\right)_z = 0,
\end{cases} \tag{14}$$

where $u(t, x, y, z)$, $v(t, x, y, z)$, $w(t, x, y, z)$ are projections of the velocity vector, $\rho(t, x, y, z)$ is the density, $p(t, x, y, z)$ is the pressure, $\chi = \frac{c_p}{c_\nu} = const$, c_ν,

c_p are thermal coefficients at constant volume and constant pressure, respectively. The first three equations of system (14) are Euler equations of motion, the fourth equation is the continuity equation, the fifth equation is the energy equation.

For plane flows we have $\frac{\partial}{\partial z} = 0, w = 0$. Suppose that the flow is isentropic: $\frac{p}{\rho^{\chi}} = c = const$, $p = c\rho^{\chi}$, $c = \frac{p_0}{\rho_0^{\chi}}$, where p_0, ρ_0 are some constant values of pressure and density. For these flows we get

$$\begin{cases} u_t + uu_x + vu_y = -c\chi\rho^{\chi-2}\rho_x, \\ v_t + uv_x + vv_y = -c\chi\rho^{\chi-2}\rho_y, \\ \rho_t + \rho(u_x + v_y) + u\rho_x + v\rho_y = 0. \end{cases} \tag{15}$$

This is the system of equations of plane isentropic motion of gas for functions u, v, ρ.

Example1. Consider the system

$$\begin{cases} u_t + uu_x + vu_y = -\zeta w_x, \\ v_t + uv_x + vv_y = -\zeta w_y, \\ w(u_x + v_y) + \mu(w_t + uw_x + vw_y) = 0. \end{cases} \tag{16}$$

The system (16) is obtained from (15), when in (15) we consider $w(t, x, y) = \rho^{\chi-1}$, $\zeta = \frac{c\chi}{\chi - 1}$, $\mu = \frac{1}{\chi - 1}$. Putting the new variables ξ, η, t, the system (16) becomes

$$\begin{cases} u_t(x_\xi y_\eta - x_\eta y_\xi) + u_\xi(x_\eta y_t - y_\eta x_t) + u_\eta(y_\xi x_t - y_t x_\xi) + \\ u(u_\xi y_\eta - u_\eta y_\xi) + v(u_\eta x_\xi - u_\xi x_\eta) + \zeta(w_\xi y_\eta - w_\eta y_\xi) = 0, \\ v_t(x_\xi y_\eta - x_\eta y_\xi) + v_\xi(x_\eta y_t - y_\eta x_t) + v_\eta(y_\xi x_t - y_t x_\xi) + \\ u(v_\xi y_\eta - v_\eta y_\xi) + v(v_\eta x_\xi - v_\xi x_\eta) + \zeta(w_\eta x_\xi - w_\xi x_\eta) = 0, \\ w(u_\xi y_\eta - u_\eta y_\xi + v_\eta x_\xi - v_\xi x_\eta) + \mu(w_t(x_\xi y_\eta - x_\eta y_\xi) + w_\xi(x_\eta y_t - \\ y_\eta x_t) + w_\eta(y_\xi x_t - y_t x_\xi) + u(w_\xi y_\eta - w_\eta y_\xi) + v(w_\eta x_\xi - w_\xi x_\eta)) = 0. \end{cases} \tag{17}$$

The solution of (17) is found in the form

$$\begin{aligned} & u = \sum_{k=0}^{\alpha} u_k(\xi, t)\eta^k, \ v = \sum_{k=0}^{\beta} v_k(\xi, t)\eta^k, \ w = \sum_{k=0}^{\theta} w_k(\xi, t)\eta^k, \\ & x = \sum_{k=0}^{\gamma} x_k(\xi, t)\eta^k, \ y = \sum_{k=0}^{\omega} y_k(\xi, t)\eta^k. \end{aligned} \tag{18}$$

Here $\alpha, \beta, \theta, \gamma, \omega$ are natural numbers. Substituting expressions (18) in (17), we obtain the following maximum degrees of variable η

$$\begin{cases} J_1 = \alpha + \gamma + \omega - 1, J_2 = 2\alpha + \omega - 1, J_3 = \beta + \alpha + \gamma - 1, \\ J_4 = \theta + \omega - 1, J_5 = \beta + \gamma + \omega - 1, J_6 = \alpha + \beta + \omega - 1, \\ J_7 = 2\beta + \gamma - 1, J_8 = \theta + \gamma - 1, J_9 = \alpha + \omega + \theta - 1, \\ J_{10} = \beta + \gamma + \theta - 1, J_{11} = \theta + \gamma + \omega - 1. \end{cases} \qquad (19)$$

Parameters J_1, J_2, J_3, J_4 correspond to the first equation of system (17), J_5, J_6, J_7, J_8 correspond to the second equation of system (17), J_9, J_{10}, J_{11} correspond to the third equation of system (17). The number of coefficients in (18) is equal $r = \alpha + \beta + \theta + \gamma + \omega + 5$, and the number of equations in system (17) is determined by relation: $s = I_1 + I_2 + I_3 + 3$, where $I_1 = max(J_1, J_2, J_3, J_4)$, $I_2 = max(J_5, J_6, J_7, J_8)$, $I_3 = max(J_9, J_{10}, J_{11})$. The maximum values I_1, I_2, I_3 can be selected in 48 manners.

We use the program described above for the determination of variables $\alpha, \beta, \theta, \gamma, \omega \in N$, for which the system of equations is determined or underdetermined. The results of classification are written in the following summary Table

Table 1. Acceptable values for system (16)

Num.	α	β	θ	γ	ω	j	*Num.*	α	β	θ	γ	ω	j
1	0	0	0	0	1	3	30	1	1	0	0	1	0
2	0	0	0	0	2	1	31	1	1	1	0	1	0
3	0	1	0	0	1	3	32	1	1	2	0	1	0
4	0	1	0	0	2	1	33	1	2	0	0	1	0
5	0	2	0	0	2	1	34	1	2	1	0	1	0
6	1	0	0	1	0	3	35	1	2	2	0	1	0
7	1	0	0	1	1	1	36	2	0	0	1	0	0
8	1	0	1	1	1	1	37	2	0	1	1	0	0
9	1	1	0	1	1	1	38	2	1	0	1	0	0
10	1	1	1	1	1	1	39	2	1	1	1	0	0
11	1	1	2	1	1	1	40	2	1	2	1	0	0
12	2	0	0	2	0	1	41	0	1	0	1	0	0
13	0	0	0	1	0	3	42	0	1	1	1	0	0
14	0	0	0	2	0	1	43	0	1	2	1	0	0
15	0	1	0	1	1	1	44	0	2	0	0	1	0
16	0	1	1	1	1	1	45	0	2	1	0	1	0
17	1	0	0	2	0	1	46	1	1	0	1	0	0
18	0	0	0	1	1	1	47	1	1	1	1	0	0
19	0	0	1	1	1	1	48	1	1	2	1	0	0
20	1	0	1	1	0	2	49	0	0	1	0	1	2
21	1	0	2	1	0	1	50	0	0	1	0	2	0
22	1	0	2	1	1	0	51	0	1	1	0	1	2
23	2	0	1	2	0	0	52	0	1	1	0	2	0
24	0	0	1	1	0	2	53	0	1	2	0	1	1
25	0	0	1	2	0	0	54	0	2	1	0	2	0
26	1	0	1	2	0	0	55	0	1	2	1	1	0
27	1	0	0	0	1	0	56	0	0	2	0	1	0
28	1	0	1	0	1	0	57	0	0	2	1	0	0
29	1	0	2	0	1	0							

Example2. Let us consider the motion of an ideal liquid governed by the following system of equations

$$\begin{cases} \rho(u_t + uu_x + vu_y) = -p_x, \\ \rho(v_t + uv_x + vv_y) = -p_y, \\ \rho_t + \rho_x u + \rho_y v + \rho(u_x + v_y) = 0, \\ \rho(p_t + up_x + vp_y) - \chi p(\rho_t + u\rho_x + v\rho_y) = 0. \end{cases} \tag{20}$$

The system (20) is obtained from system (14) when flows are considered plane $\left(\frac{\partial}{\partial z}, w = 0\right)$ and entropy is not constant.

Putting the new variables ξ, η, t in system (20), we get the following system of equations

$$\begin{cases} \rho(u_t(x_\xi y_\eta - x_\eta y_\xi) + u_\xi(x_\eta y_t - y_\eta x_t) + u_\eta(y_\xi x_t - y_t x_\xi) + \\ u(u_\xi y_\eta - u_\eta y_\xi) + v(u_\eta x_\xi - u_\xi x_\eta)) = p_\eta y_\xi - p_\xi y_\eta, \\ \rho(v_t(x_\xi y_\eta - x_\eta y_\xi) + v_\xi(x_\eta y_t - y_\eta x_t) + v_\eta(y_\xi x_t - y_t x_\xi) + \\ u(v_\xi y_\eta - v_\eta y_\xi) + v(v_\eta x_\xi - v_\xi x_\eta)) = p_\xi x_\eta - p_\eta x_\xi, \\ \rho_t(x_\xi y_\eta - x_\eta y_\xi) + \rho_\xi(x_\eta y_t - x_t y_\eta) + \rho_\eta(x_t y_\xi - x_\xi y_t) + \\ u(\rho_\xi y_\eta - \rho_\eta y_\xi) + v(x_\xi \rho_\eta - x_\eta \rho_\xi) + \rho(u_\xi y\eta - \\ u_\eta y_\xi + x_\xi v_\eta - x_\eta v_\xi) = 0, \\ \rho(p_t(x_\xi y_\eta - x_\eta y_\xi) + p_\xi(x_\eta y_t - x_t y_\eta) + p_\eta(x_t y_\xi - x_\xi y_t) + \\ u(p_\xi y_\eta - p_\eta y_\xi) + v(x_\xi p_\eta - x_\eta p_\xi)) - \chi p(\rho_t(x_\xi y_\eta - x_\eta y_\xi) + \\ \rho_\xi(x_\eta y_t - x_t y_\eta) + \rho_\eta(x_t y_\xi - x_\xi y_t) + u(\rho_\xi y_\eta - \rho_\eta y_\xi) + \\ v(x_\xi \rho_\eta - x_\eta \rho_\xi)) = 0. \end{cases} \tag{21}$$

The solution of system (21) is found in form

$$\begin{gathered} x(\xi,\eta,t) = \sum_{k=0}^{\gamma} x_k(\xi,t)\eta^k, \ y(\xi,\eta,t) = \sum_{k=0}^{\omega} y_k(\xi,t)\eta^k, \\ u(\xi,\eta,t) = \sum_{k=0}^{\alpha} u_k(\xi,t)\eta^k, \ v(\xi,\eta,t) = \sum_{k=0}^{\beta} v_k(\xi,t)\eta^k, \\ p(\xi,\eta,t) = \sum_{k=0}^{\theta} p_k(\xi,t)\eta^k, \ \rho(\xi,\eta,t) = \sum_{k=0}^{\lambda} \rho_k(\xi,t)\eta^k. \end{gathered} \tag{22}$$

Having substituted (22) in system (21), we find the degrees of polynomials, occurring in each equation. The first equation contains polynomials of degrees

$$J_1 = \gamma + \omega + \alpha + \lambda - 1, J_2 = \omega + 2\alpha + \lambda - 1, J_3 = \lambda + \gamma + \alpha + \beta - 1, J_4 = \omega + \theta - 1.$$

The second equation contains polynomials of degrees

$$J_5 = \gamma+\omega+\beta+\lambda-1, J_6 = \omega+\alpha+\beta+\lambda-1, J_7 = \lambda+\gamma+2\beta-1, J_8 = \gamma+\theta-1.$$

The third equation contains polynomials of degrees

$$J_9 = \gamma + \omega + \lambda - 1, J_{10} = \omega + \alpha + \lambda - 1, J_{11} = \lambda + \gamma + \beta - 1.$$

The forth equation contains polynomials of degrees

$$J_{12} = \gamma + \omega + \theta + \lambda - 1, J_{13} = \omega + \alpha + \theta + \lambda - 1, J_{14} = \lambda + \gamma + \beta + \theta - 1.$$

Let us use the maximum degrees of polynomials, occurring in equations

$$I_1 = max(J_1, J_2, J_3, J_4),\ I_2 = max(J_5, J_6, J_7, J_8),$$
$$I_3 = max(J_9, J_{10}, J_{11}),\ I_4 = max(J_{12}, J_{13}, J_{14}).$$

The number of unknown functions is $s = \gamma+\omega+\alpha+\beta+\theta+\lambda+6$, the number of equations is $r = I_1 + I_2 + I_3 + I_4 + 4$, a degree of sub-definite in $j = s - 1$.

Below we give the table of acceptable values of parameters $\gamma, \omega, \alpha, \beta, \theta, \lambda$, obtained as a result of calculations by means of the program described above (Table2).

Table2. Acceptable values for system (20)

Num.	γ	ω	α	β	θ	λ	j
1	0	1	0	0	0	0	3
2	0	1	0	0	0	1	0
3	0	1	0	0	1	0	2
4	0	1	0	0	1	1	0
5	0	1	0	0	2	0	0
6	0	1	0	1	0	0	3
7	0	1	0	1	0	1	0
8	0	1	0	1	1	0	2
9	0	1	0	1	1	1	0
10	0	1	0	1	2	0	1
11	0	2	0	0	0	0	0
12	0	2	0	1	0	0	0
13	0	2	0	2	0	0	0
14	1	0	0	0	0	0	3
15	1	0	0	0	0	1	0
16	1	0	0	0	1	0	2
17	1	0	0	0	1	1	0
18	1	0	0	0	2	0	0
19	1	0	1	0	0	0	3
20	1	0	1	0	0	1	0
21	1	0	1	0	1	0	2
22	1	0	1	0	1	1	0
23	1	0	1	0	2	0	1
24	1	1	0	0	0	0	0
25	1	1	0	0	1	0	0
26	1	1	0	1	0	0	0
27	1	1	0	1	1	0	0
28	1	1	1	0	0	0	0
29	1	1	1	0	1	0	0
30	1	1	1	1	0	0	0
31	1	1	1	1	1	0	0
32	1	1	1	1	2	0	0
33	2	0	0	0	0	0	0
34	2	0	1	0	0	0	0
35	2	0	2	0	0	0	0

As an example we consider the particular case: $\lambda = \alpha = 1, \gamma = \beta = \omega = \theta = 0$, degree of sub-definite $j = 3$. Then the solution is

$$\begin{cases} x = x_0(\xi, t) + x_1(\xi, t)\eta, \;\; y = y(\xi, t), \\ u = u_0(\xi, t) + u_1(\xi, t)\eta, \;\; v = v(\xi, t), \\ p = p(\xi, t), \;\; \rho = \rho(\xi, t), \end{cases}$$

from which we get

$$\begin{cases} u = u_0(y, t) + u_1(y, t)x, \;\; v = v(y, t), \\ p = p(y, t), \;\; \rho = \rho(y, t). \end{cases}$$

Substituting in (20), grouping summands at degrees of x and equating total coefficients to 0, we get system (consider that $\rho \neq 0$)

$$\begin{cases} u_{1t} + u_1^2 + vu_{1y} = 0,\ u_{0t} + u_0 u_1 + vu_{0y} = 0, \\ \rho(v_t + vv_y) = -p_y,\ \rho_t + \rho_y v + \rho(u_1 + v_y) = 0, \\ \rho(p_t + vp_y) - \chi(\rho_t + v\rho_y) = 0. \end{cases} \tag{23}$$

Let us notice that the solution of system (23) can be found in form

$$u_0 = u_0(\delta),\ u_1 = u_1(\delta),\ v = v(\delta),\ \rho = \rho(\delta),\ p = p(\delta),$$

where $\delta = y + ct$, c is an arbitrary constant. Then we get the system of five ordinary differential equations for five functions, dependent on δ.

Notation1. Similarly we obtained full classifications of solutions in form (10) for the following systems.

1. Suppose that the fluid is incompressible ($\rho = const$). Then in (14) the energy equation is canceled. Then for plane flows we have system of equations

$$u_t + uu_x + vu_y = -\frac{1}{\rho}p_x,\ v_t + uv_x + vv_y = -\frac{1}{\rho}p_y,\ u_x + v_y = 0. \tag{24}$$

2. Suppose that there is $\chi = 2$ in system (15), then we obtain

$$\begin{cases} u_t + uu_x + vu_y = -2c\rho_x,\ v_t + uv_x + vv_y = -2c\rho_y, \\ \rho_t + u\rho_x + \rho u_x + v\rho_y + \rho v_y = 0. \end{cases} \tag{25}$$

3. Suppose that there is $\chi = 1$ and $z = ln\rho$ in system (15), then we get system of equations

$$\begin{cases} u_t + uu_x + vu_y = -cz_x,\ v_t + uv_x + vv_y = -cz_y, \\ z_t + uz_x + vz_y + u_x + v_y = 0. \end{cases} \tag{26}$$

4. Non-stationary transonic flows of gas are described in the first approximation by the asymptotic system of equations

$$\Big\{\ u_t + uu_x - v_y = 0,\ u_y - v_x = 0. \tag{27}$$

5. For stationary transonic flows of gas we consider the system

$$\Big\{\ uu_x - v_y = 0,\ u_y - v_x = 0. \tag{28}$$

Notation 2. Two-parameter method can be used to obtain the exact linearization of equation systems. For example, in the parametric form, the system (28) is [8]

$$\begin{cases} u(u_\xi y_\eta - u_\eta y_\xi) - (v_\eta x_\xi - v_\xi x_\eta) = 0, \\ u_\eta x_\xi - u_\xi x_\eta - (v_\xi y_\eta - v_\eta y_\xi) = 0. \end{cases} \tag{29}$$

Substituting $u = \xi,\ v = \eta$ in system (29), we get

$$\begin{cases} \xi y_\eta - x_\xi = 0, \\ -x_\eta + y_\xi = 0, \end{cases}$$

that is in a plane $(u, v) = (\xi, \eta)$ the nonlinear system (28) becomes a linear system in the functions $x(u, v),\ y(u, v)$. Let us notice, that system (28) is a particular case of the quasi-linear system

$$\begin{cases} f_1(u, v)u_x + f_2(u, v)u_y + f_3(u, v)v_x + f_4(u, v)v_y = 0, \\ g_1(u, v)u_x + g_2(u, v)u_y + g_3(u, v)v_x + g_4(u, v)v_y = 0, \end{cases}$$

which becomes linear after passing to the variables ξ, η and choosing $u = \xi$, $v = \eta$

$$\begin{cases} f_1(\xi, \eta)y_\eta - f_2(\xi, \eta)x_\eta - f_3(\xi, \eta)y_\xi + f_4(\xi, \eta)x_\xi = 0, \\ g_1(\xi, \eta)y_\eta - g_2(\xi, \eta)x_\eta - g_3(\xi, \eta)y_\xi + g_4(\xi, \eta)x_\xi = 0. \end{cases}$$

Notation 3. Parametric method is useful for construction of solutions such as “simple waves”.

As an example consider the solution of system (16) in the form of a simple wave

$$\begin{cases} u = u(\xi),\ v = v(\xi),\ w = w(\xi), \\ x = x_0(\xi) + x_1(\xi)y + x_2(\xi)t,\ y = \eta. \end{cases}$$

Substituting it in (17), then assuming that $u(\xi) = \xi$ and, considering that $w(\xi) \neq const$, we get

$$\begin{cases} x_1 = -v'(\xi), \\ x_2(\xi) = \xi + v(\xi)v'(\xi) - \zeta w'(\xi), \\ w(\xi)(1 + (v'(\xi))^2) - \mu\zeta(w'(\xi))^2. \end{cases}$$

Two functions $w(\xi)$ and $v(\xi)$ are found from one equation, therefore any of these functions is arbitrary. The function $x_0(\xi)$ is arbitrary too. Thus, the solution is obtained, and it depends on two arbitrary functions.

Let us find the solution in form of a simple wave for system (28)

$$u=\xi, v=v(\xi), y=\eta, x=x_0(\xi)+x_1(\xi)\eta.$$

Substituting this solution in (29), we get the system

$$\begin{cases} \xi+v'(\xi)x_1(\xi)=0, \\ -x_1(\xi)-v'(\xi)=0, \end{cases}$$

from which we obtain

$$\begin{cases} v(\xi)=\pm\frac{2}{3}\xi^{\frac{3}{2}}+c, \\ x_1(\xi)=\mp\xi^{\frac{1}{2}}. \end{cases} \tag{30}$$

Then we get the following solution of (29)

$$u=\xi,\ v=\pm\frac{2}{3}\xi^{\frac{3}{2}}+c,\ y=\eta,\ x=x_0(\xi)\mp\xi^{\frac{1}{2}}y, \tag{31}$$

where $x_0(\xi)$ is an arbitrary function.

Consider parametric solutions of equations of transonic gas flows and their application.

The unsteady transonic flows of an ideal gas in the asymptotic approximation are described by system of equations

$$2u_\tau+uu_x-v_y-w_z=0,\ u_y=v_x,\ u_z=w_x,\ v_z=w_y. \tag{32}$$

Here u, v, w are projections of velocity vector on the axes of a rectangular coordinate system x, y, z ; τ is the time. For velocity potential we have equation

$$2\varphi_{x\tau}+\varphi_x\varphi_{xx}-\varphi_{yy}-\varphi_{zz}=0. \tag{33}$$

For plane flows the equation (33) is

$$2\varphi_{x\tau}+\varphi_x\varphi_{xx}-\varphi_{yy}=0,$$

and for axisymmetric flows it reads

$$2\varphi_{x\tau}+\varphi_x\varphi_{xx}-\varphi_{yy}-\frac{1}{y}\varphi_y=0.$$

By differentiating the equation (33) by x, we get an equation for $u=\varphi_x$. Some solutions of system (32) and equation (33) have been considered in [1-4]. The solution of equation (33), describing flow in Laval nozzles with local

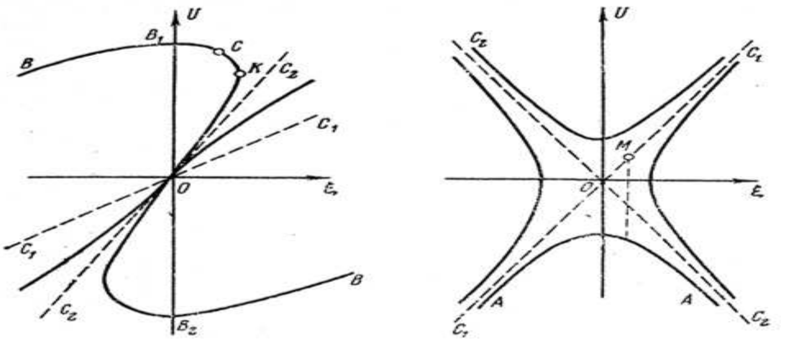


Fig. 1. The integral curves. *a*) for λ_1, λ_2 different but of the same sign; *b*) for λ_1, λ_2 of opposite signs

supersonic zones [5-7] (here and further on we consider a nozzle with two transversely-spaced planes of symmetry), is

$$u = U(\xi,\tau) + a_1(\tau)y^2 + a_2(\tau)z^2,\ x = m(\tau)\xi + n(\tau) + c_1(\tau)y^2 + c_2(\tau)z^2. \quad (34)$$

The equations for $U(\xi,\tau)$ are easily written. The primary purpose is to study the local supersonic zones (LSZ) during time. Therefore we consider just plane and axisymmetric flows. Generalization of results, obtained below, on a three-dimensional case (34) is an easy task.

System (32) admits the following solution

$$\begin{array}{l} u = \tau^{n-1}u_*(x_*,y_*,t) + 2\lambda'(\tau),\ v = \tau^{\frac{3}{2}(n-1)}v_*(x_*,y_*,t) + \dfrac{4\lambda''(\tau)}{\omega+1}y, \\ x_* = \dfrac{x-\lambda(\tau)}{\tau^n},\ y_* = y\tau^{-\frac{1}{2}(n+1)},\ t = ln(\tau). \end{array} \quad (35)$$

The solutions of u_*, v_* are found in the form (34)

$$\begin{array}{l} u_* = mU(\xi,t) + 2c(2c-1)y_*^2,\ x_* = m\xi + cy_*^2, \\ v_* = 2cm[2(2c-1)\xi - U(\xi,t)]y_* + \dfrac{8c(c-1)(2c-1)}{\omega+3}y_*^3. \end{array} \quad (36)$$

Here m, c, n are arbitrary constants, $\lambda(\tau)$ is an arbitrary function, $\omega = 0$ for plane flows and $\omega = 1$ for axisymmetric flows. For function $U(\xi,t)$ we get the equation

$$2U_t + (U - 2n\xi)U_\xi + 2[n-1+(\omega+1)c]U - 4c(2c-1)(\omega+1)\xi = 0. \quad (37)$$

First consider self-similar solutions (that is $U_t = 0$). Then U satisfies an ordinary differential equation, which contains two arbitrary parameters c and

n. In this case the behavior of integral curves depends on values of quantities λ_1, λ_2

$$\lambda_{1,2} = q_{1,2} - 2n,\ q_{1,2} = 1 - (\omega+1)c \mp [1 - (\omega+1)c(6-8c) + (\omega+1)^2 c^2]^{\frac{1}{2}}. \quad (38)$$

If λ_1, λ_2 are different but have the same sign, at the origin of coordinates of plane (U, ξ) we have a critical point, namely a node, (fig.1, *a*), if λ_1, λ_2 have opposite signs we have a saddle (fig.1, *b*); if $\lambda_1 = \lambda_2 \neq 0$, we have a degenerate node; if one of quantities λ_k (or both) is equal to zero, the solution on plane (U, ξ) is shown by parallel lines. Notice that curves $U = U(\xi, t)$ give distribution of velocity (pressure) $u = u(x,t)$ on axis $y = 0$. The solutions, shown by straight lines through a critical point (dotted straight lines correspond to them), is

$$U = q_1\xi,\ U = q_2\xi. \quad (39)$$

In the plane case $(\omega = 0)$ $q_1 = 2(1-2c), q_2 = 2c$.

In the case of a node the curves concern in a critical point of a straight line $U = q_1\xi$, if $|\lambda_1| < |\lambda_2|$, and a straight line $U = q_2\xi$, if $|\lambda_1| > |\lambda_2|$. In case of when $\lambda_1 \neq \lambda_2$, the solution of the equation (37) is written in form

$$(U - q_1\xi)^{-\lambda_1}(U - q_2\xi)^{-\lambda_2} = A = const, \quad (40)$$

or in parameter form

$$U = \frac{q_2}{q_2 - q_1}\eta + q_1 B\eta^{\chi},\ \xi = \frac{1}{q_2 - q_1}\eta + B\eta^{\chi},\ \chi = \frac{\lambda_1}{\lambda_2}. \quad (41)$$

For $\lambda_1 = \lambda_2 = \lambda_0$ $(q_1 = q_2 = q)$ the solution is easily written as

$$U = \frac{q}{\lambda_0}\eta \cdot ln(\eta) + \eta(1 + qB),\ \xi = \frac{1}{\lambda_0}\eta \cdot ln(\eta) + B\eta. \quad (42)$$

In formulae (40)-(42) A and B are arbitrary constants. The equation of sonic line for (36) in parametric form $y = y(\xi, \tau)$, $x = x(\xi, \tau)$ at $c \neq \frac{1}{2}$, $c \neq 0$ is

$$y^2 = \frac{m\tau^{(n+1)}}{2c(1-2c)}U(\xi, t) - \frac{\tau^2\lambda'(\tau)}{c(2c-1)},\ x = m\xi\tau^n + \frac{c}{\tau}y^2 + \lambda(\tau). \quad (43)$$

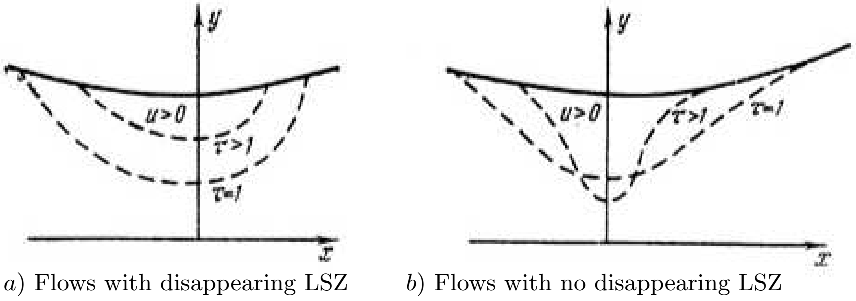
In case of $c = 0$ or $c = \frac{1}{2}$ a sonic line is $\xi = \xi_0(\tau)$. In what follows we consider a plane case $\omega = 0$ (in axisymmetric case the analysis is carried out similarly) and suppose that $\lambda'(\tau) = 0$. Then for (36) the equation of sonic line is

$$\begin{aligned} y^2 &= \frac{m\tau^{(n+1)}}{2c(1-2c)}\left[\frac{c}{3c-1}\eta + 2(1-2c)B\eta^{\chi}\right] = \frac{m\tau^{(n+1)}}{2c(1-2c)}U(\eta), \\ x &= m\tau^n + \left[\frac{1-c}{2(3c-1)(1-2c)}\eta + 2B\eta^{\lambda}\right], \ \chi = \frac{1-2c-n}{c-n}. \end{aligned} \tag{44}$$

Consider $m = 1 > 0$ (at $m < 0$ the reasoning is carried out similarly). From the first formula (44) it is obvious, that the sonic line can be constructed for $0 < c < \frac{1}{2}$, when $U > 0$, and for $c < 0$ or $c > \frac{1}{2}$, when $U < 0$. Analyzing the behavior of integral curves on fig.1 and considering the first formula (36), we conclude that the curves AA and C_1OC_2, represented on fig.1,*b* for $c < 0$ or $c > \frac{1}{2}$, can describe the flows with LSZ in Laval nozzle. The formulae (44) show us how LSZ changes in time. Since in transonic approximation the equations of nuzzle walls (they are easily constructed in parametric form) is

$$y = y_0 + \varepsilon f(x,\tau), \ \frac{\partial f}{\partial x} = v(y_0, x, \tau), \ y_0 = const, \ \varepsilon << 1, \tag{45}$$

we see from (44), that for $n > -1$ LSZ, borrowing a part of nozzle throat at the initial moment, during the time disappear, and the flow becomes subsonic everywhere, that is the solutions for $n > -1$ describe flows with disappearing LSZ. As an example, on fig.2,*a* the qualitative picture of this flow is shown (for $n = 2$, $c = -3$, $B > 0$). On the contrary, for $n < -1$, the development of LSZ is observed (the example is on fig.2,*b*, $n = -\frac{3}{2}$, $c = 4$, $B < 0$). Similarly, from the second formula (44) it is obvious that for $n > 0$, LSZ expands, for $n < 0$ it is narrowed as time is increased.



a) Flows with disappearing LSZ *b*) Flows with no disappearing LSZ

Fig. 2. The development of LSZ.

Let us notice, that in all formulae it is possible to replace τ by $(\tau + \tau_0)$, $0 \le \tau < \infty$, $\tau_0 > 0$. For $\lambda'(\tau) = 0$, according to (44), flows with LSZ, closing on nozzle axis for $\tau \to \infty$ for $n < -1$ take place. If $\lambda'(\tau) \neq 0$, at the same values of c $\left(c < 0, c > \frac{1}{2}\right)$ LSZ is closed on the nozzle axis for $\tau = \tau_1$, and then the supersonic zone occupies the whole segment of an axis (or with the increase of time there is a return process). Such LSZ are easy to construct under formulae (43),(44). Let us notice that solutions, which are given by curves, having closed parts $U > 0$ on axis $y = 0$, can describe flows with LSZ about profile (fig.3). The solutions, having such parts, are given, for example, by curves BOB (fig.1, a), and also by curves AA, C_1OC_2 (fig.1, b), for $\lambda'(\tau) > 0$.

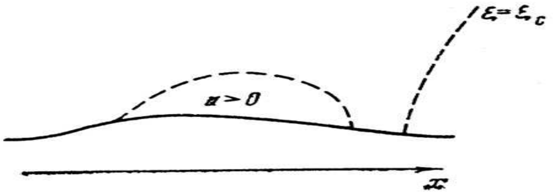


Fig. 3. The flow about profile

However, from the first formula (44) it follows that for such flows $0 < c < \frac{1}{2}$. But for such values c for the flows this type the solution is faulty (zones of ambiguity and zones of non-existence of solution). In fig. 3 we show the qualitative picture of flow about profile for $c = \frac{1}{4}$, $n = \frac{5}{8}$, $B > 0$, $y_0 = 3.35$. The integral curve BB_1C on fig.1,a corresponds to this solution ($\xi_C \leq \xi_K, U'(\xi_K) = \infty$). Flow behind $\xi = \xi_C$ can not be constructed in a class of solutions (36). Apparently, this solution can be continued behind $\xi = \xi_C$, using more general than (36), (40) solution, admissible for system (32), in the following form

$$u_* = U_0(\xi) + U_2(\xi)y_*^2, \; x_* = X_0(\xi) + X_2(\xi)y_*^2, \; v_* = V_1(\xi)y_* + V_3(\xi)y_*^3. \quad (46)$$

Consider now the solution (36) in the general case $U = U(\xi, t)$. For equation (37), using (41), the first two integrals are easily obtained. Then the general solution for $\lambda_1 \neq \lambda_2$ is written in form

$$F\left[(U - q_1\xi)^{-\lambda_1}(U - q_2\xi)^{\lambda_2}, (U - q_1\xi)exp\left(-\frac{1}{2}\lambda_2 t\right)\right] = 0. \quad (47)$$

Here F is an arbitrary function of two arguments. Having written the solution (47) in form solved relative to the second argument, we get the solution in the form $t = t(U, \xi)$. Let us notice, that in the function $t(U, \xi)$ the equation (37) is linear. In parametric form the solution (47) is written in the form (41), where B is an arbitrary function of variable $\eta \cdot exp\left(-\frac{1}{2}\lambda_2 t\right)$.

If $\lambda_1 = \lambda_2 = \lambda_0 (q_1 = q_2 = q)$, the general solution of equation (37) is

$$F\left[ln(U - q\xi) - \frac{\lambda_0 \xi}{U - q\xi}, (U - q\xi)exp\left(-\frac{1}{2}\lambda_0 t\right)\right] = 0. \tag{48}$$

In parametric form the solution is given by formulae (42), where $B\left(\eta \cdot exp\left(-\frac{\lambda_0 t}{2}\right)\right)$ is an arbitrary function.

References

[1] P. A. Vel'misov, *Asymptotic equations of gas dynamics*, Saratov University, 1986.

[2] P. A. Vel'misov, *Unsteady motion of gas in Laval nozzles*, in Aerodynamics, **2**(1994), Saratov University.

[3] O. S. Rizhov, *Study of transonic flows in Laval nozzles*, Moscow, 1965.

[4] O. S. Rizhov, *About work of Laval nozzles in off-nominal behavior*, J. Calculus Math. and Math. Phys., **7**, 4 (1967).

[5] S. Tomotika, K. Tamada, *Studies on two-dimensional transonic flows of compressible fluid (p.1)*, Quart. Appl. Math., **7**, 4(1950).

[6] S. Tomotika, Z. Hasimoto, *On the transonic flow of a compressible fluid through an axially symmetrical nozzle*, J. Mayh. Phys., **29**, 2(1950).

[7] T. C. Adamson,*Unsteady transonic flows in two-dimensional channels*, J. Fluid Mech., **52**, 3(1972), 437-449.

[8] P. A. Velmisov, S. V. Falkovich, *Some classes of solutions of transonic equations and equations of short waves*, in Selected Problems of Applied Mechanics, Moscow, 1974, 215-223.

ROMAI J., **4**, 1(2008), 261–270

PSEUDO QUATERNION REPRESENTATIONS IN THE THEORY OF MAPPINGS AND THEIR APPLICATIONS

Mukhamadi Zakhirov, Hurschidbek Yusupov

"M. Ulugbek" National University of Uzbekistan, Tashkent, Uzbekistan

mukhamadi@yahoo.com, hurschidbek@yahoo.de

Abstract The work deals with setting a correspondence between the main directions, main curvatures and theorema egregium of a surface with imaginary units, coefficients and discriminants of quadratic form, Beltrami equations system. Their applications in describing birth and uptake operators of bosons and fermions are investigated. In transonic gas dynamics it is established the connection between the Mach number and the imaginary units. Pseudoquaternions' matrix representation is investigated. New representation of Beltrami equations system solutions enable us to show that the Riemann theorem on mappings is valid on hyperbolical planes. The issue on the Poincaré's problem solution variant is studied too.

Keywords: systems of equations, fermion, commutative algebra, imaginary units, quaternion.

2000 MSC: 53Z05.

1. INTRODUCTION

The traditional representation of complex numbers is not adequate for the following reasons:

1) according to the Lagrange theorem, the direct sum of a quadratic functional in one-dimensional spaces reads $|z|^2 = x^2 - y^2$ instead of $|z|^2 = x^2 + y^2$. Consequently, there is at least one representation of complex numbers satisfying the Lagrange theorem [1];

2) the ellipticity of a point is connected with the theorema egregium, main surface curvatures. In the representation $z = x + iy$, x, y are the projections of z on the corresponding coordinate axis. The imaginary unit i is

connected with the theorema egregium, demonstrated further. An essential question arises: if the complex numbers represent the points of the surface of elliptic type, then what kind of numbers express the points of the surfaces of hyperbolic and parabolic types? Thus, an essential problem arises of deducing these numbers from equations system of elliptic type, i.e. from the Cauchy-Riemann and Beltrami equations system.

In order to do this, one must represent these equations in the operator form [2]. Factorization of the quadratic form is an example of the necessity of these numbers. These circumstances make us research the new representation of these numbers. Besides, in the following it is demonstrated that by means of these numbers the commutation relations for the birth and uptake operators of bosons can be deduced. In this respect, a fragment from the article of Dirac "Quantum electrodynamics" [3] is illuminating. "We discovered the mathematical description of electromagnetic field in the terms of decreasing and increasing of field components' excitation level by one quant. These operators might also be described as emission and uptake operators of a boson. All η are the birth operators which increase the excitation level by a unit, and all $\overline{\eta}$ represent the uptake operators which decrease the excitation level by a unit: η, $\overline{\eta}$ are the birth and uptake operators correspondingly. There are a pair of variables η^a and $\overline{\eta}^a$ for all boson states. Commutation relations for them include the following: 1) variables, corresponding to different boson states commute each other, i.e. commute all birth operators

$$\eta^a\eta^b - \eta^b\eta^a = 0, \tag{1}$$

and all uptake operators

$$\overline{\eta}^a\overline{\eta}^b - \overline{\eta}^b\overline{\eta}^a = 0; \tag{2}$$

2) the expression $\overline{\eta}^a\eta^b - \eta^b\overline{\eta}^a$ converts to zero when a and b are different and equal to zero when a is equal to b

$$\overline{\eta}^a\eta^b - \eta^b\overline{\eta}^a = \delta^{ab}, \tag{3}$$

where δ^{ab} – Kronecker's symbol.

If we deal with the electromagnetic field or any boson ensemble, we need the variables η and $\overline{\eta}$ for the description of quantum mechanical system which satisfy the abovestated commutation relations.

It is also possible to introduce the operators η and $\overline{\eta}$ for fermions which describe birth and uptake of the fermions as it was with bosons. Now any separate operator η and $\overline{\eta}$ correspond to the fermion state. Consequently, the operators η and $\overline{\eta}$ satisfy commutation relations which are different from the bosons' commutation relations

$$\eta^a\eta^b + \eta^b\eta^a = 0, \tag{4}$$

$$\overline{\eta}^a\overline{\eta}^b + \overline{\eta}^b\overline{\eta}^a = 0, \tag{5}$$

$$\overline{\eta}^a\eta^b + \eta^b\overline{\eta}^a = \delta^{ab}. \tag{6}$$

The equations (4)–(6) are the same to (1)–(3) correspondingly, only the sign minus is replaced with the sign plus. I consider it to be a very amazing mathematical fact. It is not certain what is hidden behind in fact, because we deal with two completely different physical situations. The equations (1)–(3) belong to particles, any number of which may be of any state. The equations (4)–(6) belong to the particles two of which can never be of the same state. Physically these two situations strongly differ, however, despite it, there is a close parallel between the equations corresponding to them".

2. BELTRAMI EQUATIONS

The following lemma is famous from algebra [1]: any commutative algebra with dividing possesses dimensionality of no more than 2. Thus, this lemma is one of the main reasons for the inapplicability of commutation relations in quantum mechanics while algebra with dividing is used. Consequently, the birth and uptake of bosons take place in a two-dimensional space, while fermions in a space of higher dimension.

If on the two-dimensional Riemann manifold V_2 the element of arc length is defined as

$$ds^2 = \gamma dx^2 - 2\beta dxdy + ady^2, \quad \alpha\gamma - \beta^2 = 1, \tag{7}$$

where x, y are curvilinear coordinates and $\gamma,\ \beta,\ \alpha$ are functions of variables, then there are the functions $u = u(x, y),\ \ v = v(x, y)$ which map any neighborhood of the point onto the domain of the Euclidean plane in such a way that

$$\gamma dx^2 - 2\beta dxdy + \alpha dy^2 = \sigma(u, v)(du^2 + dv^2),$$

where $\sigma(u, v) = \frac{\partial(x,y)}{\partial(u,v)} > 0$ is the Jacobian mapping. In this case, the functions satisfy Beltrami equations system taking into consideration the surface curvature

$$u_x = \beta v_x + \gamma v_y,\ -u_y = \alpha v_x + \gamma v_y,\ \alpha\gamma - \beta^2 = 1.$$

Further we consider smooth surfaces of more general form with the quadratic form (7) but without the ellipticity condition $\alpha\gamma - \beta^2 = 1$. Consequently, the quadratic form of the surface (7) is indefinite.

Remind the Euler theorem. *In each point of smooth surface there exist two perpendicular tangents $l_1,\ l_2$ in the direction of which a normal surface curvature assumes its maximal and minimal values $k_1,\ k_2$. If l is an arbitrary tangent forming the angle θ with the line l_1, then the normal curvature in the direction of l is $k_n = k_1 \cos^2\theta + k_2 \sin^2\theta$.*

Let $K = k_1 k_2$ be the theorem egregium of the surface. Let's consider the Beltrami first order partial differential equations system of mixed type on the surface [4]

$$\beta u_x + \gamma u_y = -(k_1 k_2) v_x = -K v_x,\ \beta v_x + \gamma v_y = u_y. \tag{8}$$

This system is equivalent to the following system of operator equations

$$\begin{pmatrix} \beta & \gamma \\ -\frac{\beta^2-K}{\gamma} & -\beta \end{pmatrix} \begin{pmatrix} u_x + \sqrt{-K} v_x \\ u_y + \sqrt{-K} v_y \end{pmatrix} = \sqrt{-K} \begin{pmatrix} u_x + \sqrt{-K} v_x \\ u_y + \sqrt{-K} v_y \end{pmatrix}. \tag{9}$$

Let us introduce the following imaginary units $\overrightarrow{i_e}, \overrightarrow{i_h}$ and $\overrightarrow{i_p}$.

i) Obviously, $\alpha\gamma - \beta^2 = K = -|K| i_e^2$, where $\overrightarrow{i_e}$ $(i_e^2 = -1)$ is the elliptic imaginary unit, where $i_e \neq \sqrt{-1}$ if $K > 0$, i.e. if the surface is of elliptic type;

ii) $\alpha\gamma - \beta^2 = K = -Ki_h^2$, where $\overrightarrow{i_h}$, $(i_h^2 = 1)$ is the hyperbolic imaginary unit, where $i_h = \pm 1$ if $K < 0$, i.e. if the surface is of hyperbolic type;

iii) $\alpha\gamma - \beta^2 = K = -Ki_p^2$ where $\overrightarrow{i_p}$, $(i_p^2 = 0)$ is the hyperbolic imaginary unit, where $i_p \neq 0$ if $K = 0$, i.e. if the surface is of parabolic type.

Generalizing the Lavrentev characteristic method for the mixed Beltrami equations system (equations system on smooth surfaces) we obtain

$$\gamma = k_1 \cos^2\theta + k_2 \sin^2\theta,\ \beta = (k_1 - k_2)\cos\theta\sin\theta,\ \alpha = k_1 \sin^2\theta\theta + k_2\cos^2\theta,$$

whence $\alpha\gamma - \beta^2 = k_1 k_2 = K$.

This allows one to make a geometrical interpretation of the coefficients of the system (7) and directions of coordinate axes (x, y). Thus, it is easy to establish the dependence of k_1, k_2 and θ on the coefficients of the system (8)

$$\tan\theta = \frac{2\beta}{\gamma - \alpha},\ k_{1,2} = \frac{\gamma + \alpha \pm \sqrt{(\gamma - \alpha)^2 + 4\beta^2}}{2}. \tag{10}$$

As the main curvatures as well as the theorem egregium of the surface are geometric objects, they do not depend on the introduced by us coordinates. Let us introduce the hyperbolic system of coordinates. Supposing

$$\gamma = k_1\cosh^2\vartheta + k_2\sinh^2\vartheta,\ \beta = (k_1 + k_2)\cosh\vartheta\sinh\vartheta,\ \alpha = k_1\sinh^2\vartheta + k_2\cosh^2\vartheta, \tag{11}$$

we shall receive $\alpha\gamma - \beta^2 = k_1 k_2$. The expressions of k_1, k_2 and ϑ in terms of α, β, γ read

$$\tanh 2\vartheta = \frac{2\beta}{\alpha + \gamma},\ k_{1,2} = \frac{\gamma - \alpha \pm \sqrt{(\gamma + \alpha)^2 - 4\beta^2}}{2}. \tag{12}$$

Consequently the Beltrami equations system of elliptical type arises when we reduce first quadratic form of the surface to the canonical form while the mixed Beltrami equations system arises when the equation of Dupin's indicatrixes are reduced to the canonical form.

The analogue of the Euler's formula, valid for hyperbolic trigonometric functions reads $e^{i_h\vartheta} = \cosh\vartheta + i_h\sinh\vartheta$. All famous interrelations of hyperbolical trigonometry can be studied by using this formula. Thus, we obtain the result

Theorem. *Riemannian metrics of a surface which satisfies the mixed Beltrami equations system defines main directions and main curvatures of the surface uniquely.*

Thus, the homeomorphism of Beltrami equations system forms a network of curves on the surface, i.e. the system of two families of curves, called surface curvature lines, each of which covers the surface pieces completely.

3. APPLICATION TO GAS DYNAMICS

Let us apply of the representations (2) and (10).

Let $u - i_e v = qe^{i\vartheta}$ be the complex potential of a flow. The equation of transonic gas dynamics reads

$$\begin{cases} \theta_x = -\frac{\mu^2}{q}\sin\theta\cos\theta q_x + \frac{1}{q}(1-\mu^2\sin^2\theta)q_y \equiv \beta q_x + \gamma q_y, \\ \theta_y = \frac{1}{q}(1-\mu^2\cos^2\theta)q_x - \frac{\mu^2}{q}\sin\theta\cos\theta q_y \equiv \alpha q_x + \beta q_y, \end{cases} \tag{13}$$

where μ is the Mach number. Obviously, $\alpha\gamma - \beta^2 = \frac{1-\mu^2}{q^2}$. Putting in (10) the coefficients from (13) we obtain: i) θ, which act in (13), are similar; ii) $\alpha + \gamma = \frac{1}{q}(2-\mu^2)$, $\alpha - \gamma = \frac{\mu^2}{q}\cos 2\theta$; iii) $k_1 = \frac{1}{q}$, $k_2 = \frac{1-\mu^2}{q}$;
iiii) $k_1 k_2 = \frac{1-\mu^2}{q^2} = K$, whence it is clear that:

i) if the flow is subsonic, i.e. $\mu < 1$, then $K > 0$ therefore the stream surface is of elliptical type;

ii) if the flow is transonic, i.e. $\mu = 1$, then $K = 0$ therefore the stream surface is of parabolic type;

iii) if the flow is supersonic, i.e. $\mu > 1$, then $K < 0$, therefore the stream surface is of hyperbolical type.

It is essential that mixed Beltrami equations system allows us to compute uniquely: i) values of main curvatures; ii) theorema egregium; iii) to define uniquely main directions of the surface flow in terms of the characteristics of the complex potential flow.

4. PSEUDOQUATERNIONS MATRIX REPRESENTATION

The following matrices will be put in correspondence to the quaternions

$$\begin{cases} \vec{i_e} \to \begin{pmatrix} \beta & \gamma \\ -\frac{\beta^2+1}{\gamma} & -\beta \end{pmatrix} = A_e, \\ \vec{j_e} \to i_e \begin{pmatrix} \beta & \gamma \\ -\frac{\beta^2-1}{\gamma} & -\beta \end{pmatrix} = B_e, \\ \vec{k_e} \to i_e \begin{pmatrix} 1 & 0 \\ -\frac{2\beta}{\gamma} & -1 \end{pmatrix} = C_e, \end{cases} \tag{14}$$

The following matrices will be put in correspondence to the pseudoquaternions

$$\begin{cases} \vec{i_k} \to i_e \begin{pmatrix} \beta & \gamma \\ \frac{\beta^2+1}{\gamma} & -\beta \end{pmatrix} = A_h, \\ \vec{j_k} \to i_h \begin{pmatrix} \beta & \gamma \\ -\frac{\beta^2-1}{\gamma} & -\beta \end{pmatrix} = B_h, \\ \vec{k_h} \to i_e i_h \begin{pmatrix} 1 & 0 \\ -\frac{2\beta}{\gamma} & -1 \end{pmatrix} = C_h, \end{cases} \tag{15}$$

Direct computations lead to the following properties of the matrices (14)

$$A_e^2 = -E,\ B_e^2 = -E,\ C_e^2 = -E,\ A_eB_e + B_eA_e = 0,$$

$$B_eC_e + C_eB_e = 0, \quad C_eA_e + A_eC_e = 0.$$

Correspondingly, we have

$$\vec{i_e}\vec{i_e} = \vec{i_e}^2 = -1, \quad \vec{j_e}\vec{j_e} = \vec{j_e}^2 = 1, \quad \vec{k_e}\vec{k_e} = \vec{k_e}^2 = -1,$$

$$\vec{i_e}\vec{j_e} + \vec{j_e}\vec{i_e} = 0, \quad \vec{j_e}\vec{k_e} + \vec{j_e}\vec{k_e} = 0, \quad \vec{k_e}\vec{i_e} + \vec{i_e}\vec{k_e} = 0.$$

The commutation properties of the matrices are defined exactly (15)

$$A_h^2 = E,\ B_h^2 = E,\ C_h^2 = E,\ A_hB_h - B_hA_h = 0,$$

$$B_hC_h - C_hB_h = 0, \quad C_hA_h - A_hC_h = 0.$$

Correspondingly, we have

$$\vec{i_h}\vec{i_h} = \vec{i_h}^2 = 1, \quad \vec{j_h}\vec{j_h} = \vec{j_h}^2 = 1, \quad \vec{k_h}\vec{k_h} = \vec{k_h}^2 = 1,$$

$$\vec{i_h}\vec{j_h} - \vec{j_h}\vec{i_h} = 0, \quad \vec{j_h}\vec{k_h} - \vec{j_h}\vec{k_h} = 0, \quad \vec{k_h}\vec{i_h} - \vec{i_h}\vec{k_h} = 0. \qquad (16)$$

Now we introduce the operators conjugate to the mentioned operators as the operators corresponding to the conjugate Beltrami equations system.

$$\bar{A}_e = -A_e, \quad \bar{B}_e = -B_e, \quad \bar{C}_e = -C_e, \bar{A}_h = -A_h, \quad \bar{B}_h = -B_h, \quad \bar{C}_h = -C_h.$$

Obviously, all operators with the index e satisfy (4), (5) and are considered as the birth and uptake operators of fermions, while all operators with the index h satisfy the relations (1), (2) and are considered as the birth and uptake operators of bosons. At the same time, these operators are the permutation operators of fermions and bosons respectively. Therefore, the relations (3) and (6) are replaced by squaring the operators with the indexes e, h.

5. THE POINCARÉ PROBLEM AND THE RIEMANN THEOREM

These results allow us to solve the Poincaré problem (conjecture) by using the Seifert fibration [6]. The domain D serves as the fibration base on the manifold, while all bounded domains which are the images of D through quasi-conformal mappings corresponding to the elliptical type of Beltrami equations system serve as the fibration space. Jacobians (consequently, theorema egregium of M) of these mappings define the fibration layers. The group of the conformal mappings serve as the structural group of layer transformations. Torus-torus in the fibration of Seifert arise as a sequence of the action of the operator B, while the Klein full bottle is the sequence of the action of the operator A. Torus-torus and the Klein full bottle arise because the elements of the considered by us operators-matrices are functions, namely the coefficients of the Beltrami equations system defined in D. The direct product $S^1 \times D$ where S^1 is the one-dimensional sphere, serves as the trivial puff torus while the puff Klein bottle is the direct sum in the representation of the Beltrami equations system [4]

$$w = u + i_e v + i_h(p - i_e q) = W_+ \oplus W_-, \tag{17}$$

which is covered twice by the trivial torus-torus. With the renormalization of (17) one can obtain the direct sum for the representation of the puff torus-torus $w = u + i_h p + i_e(v + i_h q)$, which is finite-to-one covered by the trivial torus-torus.

By the Dupin theorem [7], the surfaces of triple orthogonal system are always pairwise intersected by the curvature lines. The curvature lines are considered as integral curves of the equations

$$dx = 0, \quad dy = 0, \quad \gamma dx - \beta dy = 0.$$

Then, $u = M(\gamma dx - \beta dy)$, $v = Ndy$, $d\varphi = Ndx$, where $M = M(y)$ is the integration multiplier [8] of the forms $\psi_1 = \gamma dx - \beta dy$, $\psi_2 = dy$, $\psi_3 = dx$. Then, $M^2 \left[(\psi_1)^2 + (\psi_2)^2 + (\psi_3)^2\right] = du^2 + dv^2 + d\psi^2$, that is, the conformal mapping by the Gauss in a three-dimensional Euclidean space. Consequently, two arbitrary surfaces which possess the representation (17) must have common curvature lines. The following surface family serves as the mentioned triple of surface families

$$\omega_1 = Xi_e + Yj_h + Zi_e j_h, \quad \omega_2 = Xi_h + Yi_e i_h + Zi_e, \quad \omega_3 = Xi_e i_h + Yi_e + Zi_h.$$

Obviously, $\omega_1^2 + \omega_2^2 + \omega_3^2 = X^2 + Y^2 + Z^2$ and $X = X(x, y)$, $Y = Y(x, y)$, $Z = Z(x, y)$, where $(x, y) \in D$.

From (2) - (11) and (17) on may conclude that the substitution of the imaginary unit i_e by the unit i_h in the Beltrami equations system gives us the system of a hyperbolic type. Here, the bounded domains D and Δ, between which the mappings corresponding to the Beltrami equations system of elliptic type are considered as the geometrical objects in affine spaces, remain unchanged. In the same way, it can be proved the Riemann theorem on mappings for the hyperbolic Beltrami equations system.

References

[1] Postnikov, M. M., *Lie groups and algebras,* Nauka, Moskow, 1982.

[2] Zakhirov, M., *MES automorphisms of vector fields and their applications,* Third International Conference “Modern problems of mathematical physics and informational technologies”, Tashkent, **1**(2005), 275-280.

[3] Dirac, P. M. A., *Recollections of an exciting era,* Nauka, Moscow, 1990.

[4] Zakhirov, M., *Imaginary units: use in analysis and applications.* ROMAI J., **2**, 2(2006), 227-259.

[5] Bers, L., *Mathematical aspects of subsonic and transonic gas dynamics,* New York, 1959.

[6] Scott, P., *The geometries of 3-manifolds,* London Mathematical Society, 1983.

[7] Hilbert, D., *Anschauliche Geometrie*, Cohn-Vossen, Berlin, 1932.

[8] Zakirov, M., *Obobshennie kvaternioni sootvetsvuyushie uravneniya MES i prilojeniya*, Int. Conf. “Aktualniye voprosi kompleksnogo analiza”, Urgench, Xiva, Uzbekistan, 2008, 16-17.

MATHEMATICAL DESCRIPTION OF THE PROCESS OF REMOVING SOME IONS FROM AQUEOUS-ALCOHOL SOLUTIONS BY MEANS OF CATION-EXCHANGE RESINS

V. I. Zelentsov, A. M. Romanov

The Institute of Applied Physics of The Academy of Sciences of Moldova, Chişinău, Republic of Moldova

vzelen@yandex.ru

Abstract The constant of ion exchange (IE) equilibrium and the total exchange capacity of an ionit are important characteristics of ion exchange process [1-3].

For their determination the (IE) equations, which correspond to one or another of their adsorption isotherm, are used. Knowing these values one describes the ion exchange equilibrium in double or triple ions system, i.e. calculates the equilibrium relation of ions quantity in solid to this one in solution.

The most widely used equation of (IE) isotherm was obtained by P. Nikolsky

$$\frac{X_1^{1/z_1}}{X_2^{1/z_2}} = k_{1,2} \cdot \frac{a_1^{1/z_1}}{a_2^{1/z_2}}, \tag{1}$$

where

$X_{1,2}$ – the quantity of absorbed ions;

$z_{1,2}$ – the charge of the same ions;

$k_{1,2}$ – the exchange ions constant of these;

$a_{1,2}$ – the activities of ions in equilibrium solution.

In this work the (IE) equilibrium of ions of Ca and Fe on the cation-exchange U – 2.8 and C 100 in H – form has been studied.

Keywords: mathematical models for chemical reactions.

2000 MSC: 80A50 .

At the ionit introduction into the solution at some time the equilibrium settles that characterizes by the certain distribution of cations between the cationit and the solution.

The quantitative dependence of the exchange degree on the ions concentration in the solution is expressed by the law of acting masses.

For the exchange of bivalent Ca and Fe in the solution (Me^{2+}) to univalent hydrogen ions (H^{+}) of the cationit we can write

$$Me^{2+} + 2HR \leftrightarrow MeR_2 + 2H^{+}, \tag{2}$$

where Me^{2+} and H^{+} are exchanging ions; R is the insoluble anionic matrix of the cationit.

Then the thermodynamic constant of equilibrium of this equation is

$$k_T = \frac{C^2_{H^+} \cdot C_{MeR_2}}{C_{Me^{2+}} \cdot C^2_{HR}} \cdot \frac{f^2_{H^+} \cdot f_{MeR_2}}{f_{Me^{2+}} \cdot f^2_{HR}}, \tag{3}$$

where C$^{+}_{H}$, C$^{2+}_{Me}$ are the concentrations of ions in the solution; C$_{MeR}$, C$_{HR}$ are the concentrations of these ions in the cationit; f$_{i}$ are the coefficients of activity of corresponding ions in the cationit and in the solution.

The main difficulty when using equation (3) is the determination of the activity coefficients in the ionit phase. Often in order to solve this problem it is assumed the ionit phase as an ideal solid solution of exchanging ions and that ions activities are proportional to their concentration in the ionit phase. In this case the activity coefficients in equation (3) is omitted.

On the other hand, the relation of average activity coefficients of ions in the solution, if their concentrations are not so high (¡ 0,01 N), is close to 1. Then, under these hypotheses, equation (3) is

$$k_C = \frac{C^2_{H^+} \cdot C_{MeR_2}}{C_{Me^{2+}} \cdot C^2_{HR}}. \tag{4}$$

Thus, at the exchange of ions equal to their charge at rather low concentrations of initial salts the value of concentration constant (IE) k$_{C}$ is close to the value of thermodynamic constant (k$_{T}$).

In general, in order to determine experimentally the equilibrium constant (IE) one must know equilibrium concentrations of ions both in ionit phase and in solution that is difficult to do sometimes. If the cationit in H form is used, such a calculation becomes possible.

Consider from this point of view the equation (4). In this equation denote the full exchange capacity by X$_m$ (in mg-equiv) for the corresponding mass of the ionit (m), the initial quantity of ions in the solution by C$_o$ (in mg–equiv); the quantity of exchanging ions (Me^{2+} →2H$^+$) by X (in mg-equiv).

Then the equilibrium concentrations of ions Me^{2+} and H$^+$ in the ionit phase and in the solution (according to equation (2)) is equal to

$$\mathrm{C}_{Me}^{2+} = \mathrm{C}_o - \tfrac{X}{2}; \qquad \mathrm{C}_{HR} = \mathrm{X}_m - \mathrm{X}; \qquad \mathrm{C}MeR_2 = \tfrac{X}{2};\ \mathrm{C}_H^+ = \mathrm{X}.$$

Substituting these values into equation (4) we obtain the expression for constant k$_C$

$$k_C = \frac{X^2 . \frac{X}{2}}{(C_o - \frac{X}{2}) \cdot (X_m - X)^2}, \tag{5}$$

whence

$$(X_m - X)^2 = \frac{X^3}{k_C \cdot (2C_0 - X)}, \tag{6}$$

or

$$X_m - X = \frac{X \cdot \sqrt{X}}{\sqrt{k_C \cdot (2C_0 - X)}}. \tag{7}$$

By dividing both parts of equation (7) by X$_m$·X we obtain

$$\tfrac{1}{X} - \tfrac{1}{X_m} = \tfrac{1}{X_m \sqrt{k_c}} \sqrt{\tfrac{X}{2C_o - X}} \tag{8}$$

and, finally,

$$\tfrac{1}{X} = \tfrac{1}{Xm} + \tfrac{1}{X_m \sqrt{k_c}} \sqrt{\tfrac{X}{2C_o - X}}. \tag{9}$$

This is an equation of isotherm of ion exchange. If the initial concentrations of ions in the solution (C$_o$) are known and when determining experimentally the quantity of exchanging ions (Me^{2+}), that is equal to the quantity of evolving hydrogen ions (X), we can calculate the total exchangeable capacity of ionit (X$_m$) and the (IE) equilibrium constant (k$_c$).

Equation (9) represents an equation with two independent variables: X$_m$ - the total exchangeable capacity of ionit, k$_c$- equilibrium constant of (IE). As

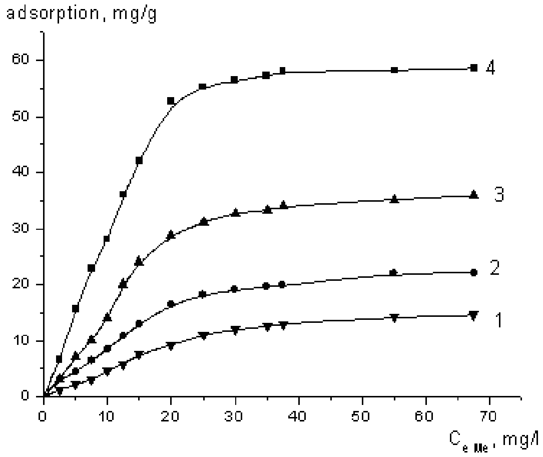


Fig. 1. Isotherms of adsorption of Ca (3, 4) and Fe (1, 2) on cationits KU-2*8H (1, 4) and C100H (2, 3). Weight of ionit - 0,25 g, volume of solution - 200 ml, time of contact of ionit with solution - 24 h.

any equation, it requires the determination of two parameters – X and C_0. Parameter X can be obtained from data of experimental adsorption isotherm of Me^{2+} (fig. 1), and C_0 - the quantity of Me^{2+} ions in initial solution is assigned in advance.

In order to verify the applicability of equation (9) at the description of (IE) equilibrium in the triple ion system Ca, Fe, H we studied the dependencies of quantities of ions of bivalent metals Ca and Fe adsorbed by the ionits KU-2·8H – Ca and C 100H – Fe on their equilibrium concentration in the solution. The experiments were carried out in static conditions. The procedure of preparing of alcohol –aqueous solutions of Ca and Fe and the carrying out of experiments, and the calculation of values of ions adsorption are described in [4]. The maximal concentration of Ca and Fe in mixture was about 200mg/l, and the minimal one – 10 mg/l for each ion.

Fig. 1 shows the obtained experimental data.

Apparently the entire experimental points well coincide with the straight line, that testifies that the studied process of ion exchange of H ions from cationits with Ca and Fe ions is satisfactory described by the given equation. Table 1 is under consideration.

Table 1 Total exchangeable capacity (X_m, mg/g) of the ionits.

X_m, mg/g						
Ionits	KU 2*8H			C100H		
\ Ions	Theoretic	Experimental	Deviation, %	Theoretic	Experimental	Deviation, %
Ca^{2+}	53,2	58,6	9,2	35,0	35,8	2,2
Fe^{2+}	21,5	22,9	6,3	16,1	14,5	9,3

In Table 1 the values of total exchange capacity X_m of cationits against Ca and Fe ions, calculated according to equation (9) and determined experimentally are shown. As can be seen the difference between values X_m calculated theoretically and those obtained experimentally does not exceed 10% what means that the proposed equation can be used for the determination of the total exchangeable capacity of ionits.

So, knowing the constants X_m and k_C one can calculate the value of adsorption of Me^{2+}ions by the ionit KU-2*8H and C100H at any of their concentrations in the initial solution and find the distribution of these ions in ionit phase and in the solution.

References

[1] Soldatov, V., Bychcova, V., *Ionoobmennoe ravnovesie v mnogocomponentnyh schemah*, Nauka i Tehnica, Minsk, 1988. (Russian)

[2] Kokotov, Yu., Zolotorev, P., El'kin, G., *Teoreticheskie osnovy ionnogo obmena*, Himiia, Leningrad, 1986. (Russian)

[3] Al'tschuller, G., Al'tschuller, O., *Raschot sostava ionita v ravnovesie s rastvorom electrolitov*, Jurnal Fiz. Himii, **75**, 12(2001), 2237-2241. (Russian)

[4] Zelentsov, V., Romanov, A., *Studiul procesului de înlăturare a ionilor metalelor grele din soluţii apoase de alcool prin metoda schimbului ionic*, Proc. XXX Annual ARA Congress, Chişinău, 2005, 242–245. (Romanian)