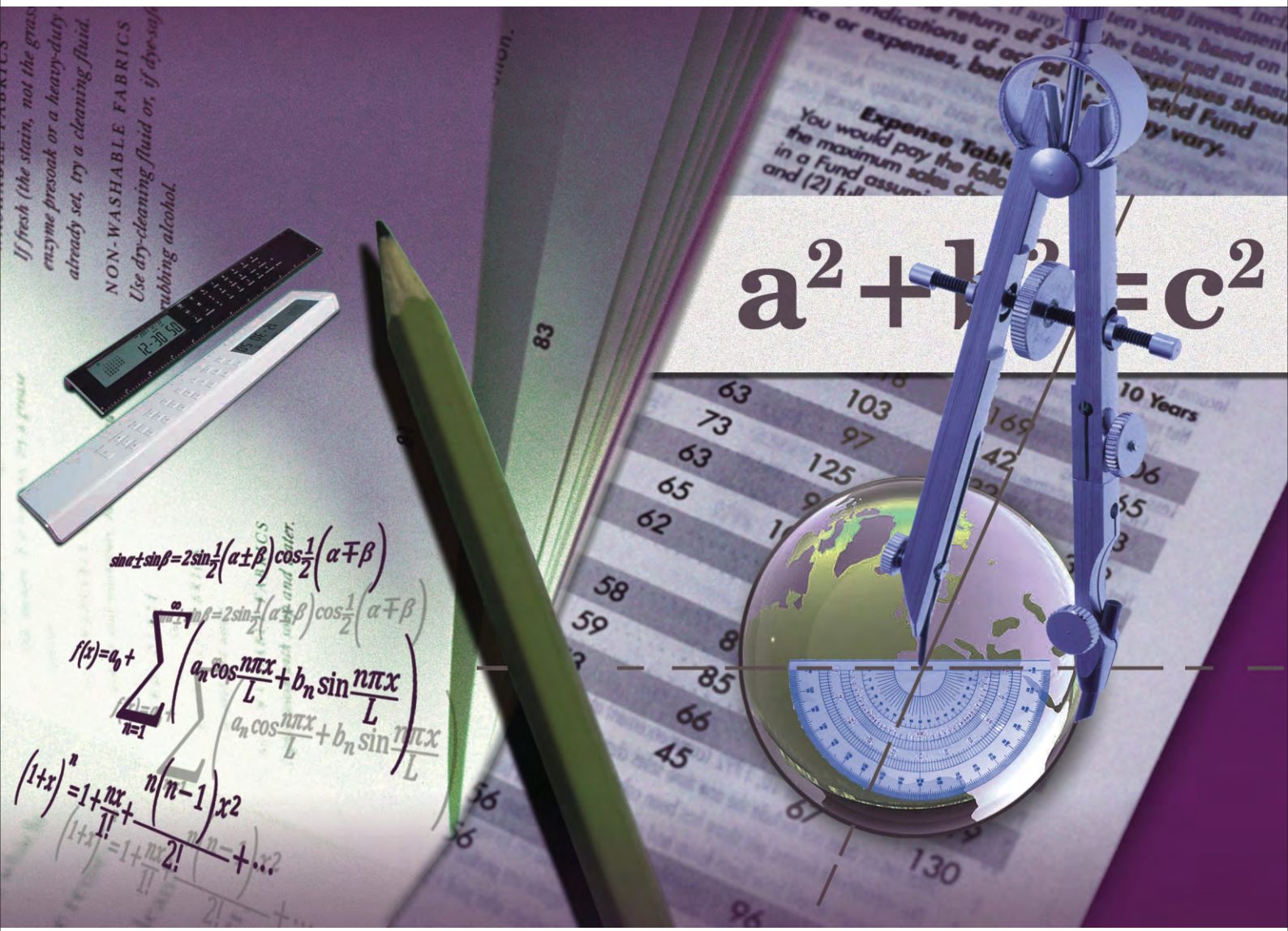
ISSN: 2152-7385 Vol.1, No.3, September 2010

Scientific Research

Applied Mathematics



ISSN: 2152-7385

9 772152 738001 03

www.scirp.org/journal/am

Journal Editorial Board

ISSN 2152-7385 (Print) ISSN 2152-7393 (Online)
http://www.scirp.org/journal/am

Editor-in-Chief

Editorial Board (According to Alphabet)

Editorial Assistant

Announcement

Owing to the large number of manuscripts that we are receiving, Applied Mathematics (AM) will increase the publication frequency from bimonthly to monthly as of September 2010.

AM Editorial Office

Applied Mathematics, **2010, 1, 137-249**
Published Online September 2010 in SciRes (http://www.SciRP.org/journal/am/)

TABLE OF CONTENTS

Volume 1 Number 3 September 2010

Applied Mathematics (AM)

Journal Information

SUBSCRIPTIONS

Applied Mathematics (Online at Scientific Research Publishing, www.SciRP.org) is published monthly by Scientific Research Publishing, Inc., USA.

Subscription rates:

Print: $50 per issue.
To subscribe, please contact Journals Subscriptions Department, E-mail: sub@scirp.org

SERVICES

Advertisements
Advertisement Sales Department, E-mail: service@scirp.org

Reprints (minimum quantity 100 copies)
Reprints Co-ordinator, Scientific Research Publishing, Inc., USA.
E-mail: sub@scirp.org

COPYRIGHT

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:
E-mail: am@scirp.org

Applied Mathematics, **2010, 1, 137-145**
doi:10.4236/am.2010.13018 Published Online September 2010 (http://www.SciRP.org/journal/am)

Some Models of Reproducing Graphs: I Pure Reproduction

Richard Southwell, Chris Cannings
School of Mathematics and Statistics, University of Sheffield, Sheffield, UK
E-mail: *bugzsouthwell@yahoo.com*
Received June 26, 2010; *revised August* 10, 2010; *accepted August* 15, 2010

Abstract

Many real world networks change over time. This may arise due to individuals joining or leaving the network or due to links forming or being broken. These events may arise because of interactions between the vertices which occasion payoffs which subsequently determine the fate of the nodes, due to ageing or crowding, or perhaps due to isolation. Such phenomena result in a dynamical system which may lead to complex behaviours, to self-replication, to chaotic or regular patterns, to emergent phenomena from local interactions. They give insight to the nature of the real-world phenomena which the network, and its dynamics, may approximate. To a large extent the models considered here are motivated by biological and social phenomena, where the vertices may be genes, proteins, genomes or organisms, and the links interactions of various kinds. In this, the first paper of a series, we consider the dynamics of pure reproduction models where networks grow relentlessly in a deterministic way.

Keywords: Reproduction, Graph, Network, Adaptive, Evolution

1. Introduction

There has been much recent interest in the way in which networks such as the World Wide Web grow, and the structures which result from various rules by which new vertices are added and link to the existing vertices. One of the most studied is the so called Preferential Attachment model whereby a new node is added at each time $t \in N^+$ (we use N and N^+ for the non-negative integers and positive integers respectively) and links to some set of existing vertices with probabilities which depend on the degrees of the latter. In the simplest case the probabilities are simply proportional to the degree, a model introduced by Yule [1], again by Simon [2], and then more recently by Barábasi and Albert [3]. The outcome of this process (see [4,5]) is a network in which the degree of a randomly selected node follows a power law distribution (*i.e.*, if X is that degree then the probability $Pr(X = k) = ak^{-b}$), and the network is scale-free in the sense that $Pr(X = l^*c) / Pr(X = c) = Pr(X = l^*d) / Pr(X = d)$ for all l, c and d.

On the other hand there has been relatively little attention paid to the growth of networks through the reproduction of existing vertices and the generation of links between these new vertices and the old vertices, although, of course, the preferential attachment model where a new vertex is linked to an existing vertex could be regarded as the production of an offspring by the latter. This is clearly a situation which arises in a biological population which reproduces itself and in which we track relatedness. In a population which reproduces asexually, if we join each individual to its parent, then we simply produce a tree for each clone. More interestingly, if in a sexually reproducing population we join each individual to their two parents we obtain a genealogy (see [6] for alternate ways of representing this network).

A further biological example happens when a genome duplication occurs [7]. The genes in the genome each code for some specific protein. If one considers the set of proteins of some organism as vertices in some network and joins any pair of vertices if the corresponding pair of proteins can bind then one obtains the protein-protein-interaction network. In a genome duplication every gene is essentially duplicated, so that there are now two copies producing the protein previously produced by one copy (we assume for simplicity that there is a simple one-one mapping of proteins to genes, ignore post-translational modification and other interactions, and splicing variation). If we then distinguish between the two copies of the genes and the proteins produced by those two copies

then we have a doubling of the set of vertices, and a quadrupling of the number of the links. This is our model 5 below.

More generally suppose that a set of entities are allowed to reproduce and that links which are produced in the new network are defined in terms of the existing links, and the relatedness of new and old verticies. In addition to the gene-duplication example above this might correspond to the establishment of the social network between individuals. For example, taking a gynocentric view, suppose that daughters of mothers who are friends are also always friends, and that mother and daughter also are treated as friends, then we obtain a particular set of relationships in a population as it reproduces, our model 6. Certainly it is well known in some species of apes and monkeys that social relationship is influenced by biological relatedness [8,9], and this is also well known in other groups e.g. spotted hyenas [10].

From now on we switch our terminology to that of graph theory, *i.e.*, refer to a graph rather than a network, and to an edge rather than a link. A graph, denoted $G(V,E)$ or just G for short, is a specification of some set V, whose elements are referred to as vertices, and some set $E \subset V \otimes^* V$ whose elements are called edges. $V \otimes^* V$ denotes the set of unordered pairs of elements (we adopt $\otimes$ for the direct product, *i.e.*, the set of ordered pairs, and elsewhere for the direct product of matrices) from V since we are restricting ourselves to undirected graphs, and we do not exclude the possibility of self-edges, that is choosing the same element of V twice. We will extensively use the notion of a graph product [11]. Suppose we have two graphs $G = (U,C)$ and $H = (V,D)$, then we define a new graph $K(W,E) = G\Delta H$ as a graph product of G and H, where $W = U \otimes V$, and the edge set E contains all $((u_1,v_1),(u_2,v_2))$, where $u_i \in U$ and $v_i \in V$, which satisfy some set of relationships which depend on the identity, adjacency or non-adjacency of u_1 and u_2, and of v_1 and v_2 [11].

We consider the following processes. The current graph is updated by adding to it a new vertex (the offspring) for each existing vertex (the parent). Each edge of the current graph is replaced by a subset of the edges of the complete graph formed on the pair of parent vertices and their two offspring; we always retain the edge between the parent vertices. Thus the "old" graph is always a subgraph of the "new" graph. The eight distinct ways in which this can be done constitute the set of models we consider (defined precisely below). Note further that there is no mortality in this model, all vertices and edges, once created are immortal. We shall discuss models in which the death of a vertex depends on the degree or the age of that vertex, and models in which interactions (games) between neighbours determine the survival in subsequent papers.

2. The Models

We are interested in a family of sequences of graphs $G_t(V_t,E_t)$, where $t \in N$ which we shall refer to as time, V_t is the set of vertices and $E_t \subseteq V_t \otimes^* V_t$ the set of edges. We define a set of functions $F_i()$ for $i = 0,..,7$, which map graphs to graphs. In general we consider the sequences defined recursively by specifying G_0 and function $F_i(G)$; then $G_{t+1} = F_i(G_t)$. In each case we form G_{t+1} as a graph product of the existing graph G_t with a simple two vertex graph.

Suppose that $H_i = H_i(W_i,K_i) = F_i(G)$ for $G = G(V,E)$. Then H_i has vertex set $W_i = V \otimes \{0,1\}$. Thus each vertex of V gives rise to two vertices in W_i. We shall refer to the vertices $(u,0)$ and $(u,1)$ arising from $u \in V$ as the offspring vertex and parent vertex respectively. H_i has edge set K_i. Now if *u* and *v* are distinct elements of *V*, then $(u,v) \notin E \Rightarrow ((u,j),(v,j)) \notin K_i$ for all j and $(u,v) \in E \Rightarrow ((u,1),(v,1)) \in K_i$. We introduce three indicators (functions taking values 0 or 1), α, β and γ to specify the additional edges which are added to K_i. The index i of the eight functions $F_i(G)$ are written in binary and these define the three indicators for that model e.g. F_6 has $\alpha = 1$, $\beta = 1$ and $\gamma = 0$. Thus $F_i(G)$ for $i = 4\alpha + 2\beta + \gamma$ has edges as follows. If $u \in V$ then $((u,0),(u,1)) \in K_i$ if, and only if, $\beta = 1$. If $(u,v) \in E$ then $((u,0),(v,0)) \in K_i$ if, and only if, $\alpha = 1$. Finally $((u,0),(v,1)) \in K_i$ $((u,1),(v,0)) \in K_i$ if, and only if, $\gamma = 1$.

These models are illustrated in **Figure 1**.

We shall discuss here only the details of the eight models described above by using the eight F_i repeatedly. We could generalize the model in various ways, e.g. by taking some sequence $\{x_t\}$, possibly generated at random, of elements from $\{0,1,2,\ldots,7\}$ and using the function F_{x_t} as the transition from G_t, or we could

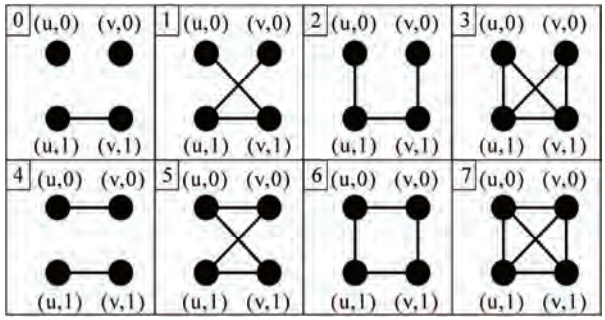


Figure 1. The motif that an edge {u, v} is replaced by under each model. The code for the models is shown at the top left of each panel.

take the α, β and γ themselves to be probabilities that the corresponding edges are included at each stage.

We shall derive results for our models for the number of vertices with degree *d*, the total number of edges, the chromatic number, diameter and average distance, and comment briefly on the automorphisms. A further paper will explore additional graph entities such as cliques, and cycles, the number of polygons of given size arising from a given polygon at the previous time step, and add further information regarding some of the entities explored here.

As stated above the operations introduced here are all equivalent to taking a graph product of G_T with some simple graph *Z*. **Table 1** specifies the products for each of the models.

3. Binary String Representation

As mentioned above, a useful, alternative way to define the rules is in terms of the notion of parent and offspring vertices, and binary strings. If we begin with a graph $G_0(V_0, E_0)$, then if $u \in V_0$, this vertex gives rise to $(u,1)$ and $(u,0)$ in G_1, which we write as $u0$ and $u1$, and to $((u,1),1)$, $((u,1),0)$ $((u,0),1)$ and $((u,0),0)$ in G_2, which are written in the obvious way as $u11$, $u10$, $u01$ and $u00$, and so on. In G_t there will be 2^t vertices arising from *u*. We denote these vertices as strings of length $(t+1)$ written $u\tau$ where τ runs over the binary strings of length *t*. We refer to these representations as vertex strings.

The eight models, all of which at time *t* have $|V_0| \times 2^t$ vertex strings, give rise to distinct edge sets. We now specify precisely the edge set for each model at time *t*. Consider two vertices ux_t and vy_t (possibly identical) at time *t*, so x_t and y_t are two binary strings of length *t*, whose *I*'th elements are denoted by x_t^i and y_t^i. Now we define a third string z_t, where the *i*'th element of z_t, z_t^i, is determined by the pair (x_t^i, y_t^i). The purpose of z_t is to specify the sequence of edges which need to be added to *u* and *v* in order to reach ux_t and vy_t. In specifying the models earlier we introduced α, β and γ, as indicators for the three distinct types of new edge. Here we identify the elements of z_t with α, β, γ and two new terms γ^* and δ. Thus if we have $(x_t^i, y_t^i) = (0,0)$, indicating that an edge must be placed between the offspring of the individuals ux_{t-1} and vy_{t-1}, then we record $z_t^i = \alpha$. Similarly we track the other edges, as is detailed below. Note for ease we introduce a δ corresponding to the choice $(x_t^i, y_t^i) = (1,1)$, and differentiate between $(x_t^i, y_t^i) = (0,1)$ and $(x_t^i, y_t^i) = (1,0)$ by using γ and γ^* respectively. When we use the z_t's to specify which edges exist in G_t, we shall in fact take $\delta = 1$ always, and $\gamma = \gamma^*$.

Table 1. The products are denoted by a single letter K = Kronecker, C = Cartesian, H = Comb, S = Strong = AND, N = non-standard.

Model	α	β	γ	Product	Edges of Z
0	0	0	0	K	{(1,1)}
1	0	0	1	K	{(0,1), (1,1)}
2	0	1	0	H	{(0,1)}
3	0	1	1	N	N
4	1	0	0	K	{(0,0), (1,1),}
5	1	0	1	K	{(0,0), (0,1), (1,1)}
6	1	1	0	C	{(0,1)}
7	1	1	1	S	{(0,1)}

If

1) $(x_t^i, y_t^i) = (0,0)$ then $z_t^i = \alpha$,
2) $(x_t^i, y_t^i) = (1,1)$ then $z_t^i = \delta$,
3) $(x_t^i, y_t^i) = (0,1)$ or $(x_t^i, y_t^i) = (1,0)$ and $ux_{t-1} = vy_{t-1}$ then $z_t^i = \beta$,
4) $(x_t^i, y_t^i) = (0,1)$ and $ux_{t-1} \neq vy_{t-1}$ then $z_t^i = \gamma$,
5) $(x_t^i, y_t^i) = (1,0)$ and $ux_{t-1} \neq vy_{t-1}$ then $z_t^i = \gamma^*$.

The string $(u,v)z_t$ specifies the start and the sequence of operations which need to take place to progress from (u,v) to (ux_t, vy_t). As examples consider (A) vertices $(u0010101010)$ and $(u0011001110)$, then $(u,v)z_{10} = ((u,u)\alpha\alpha\delta\beta\gamma\alpha\delta\gamma^*\delta\alpha)$, and (B) $(u0011001011)$ and $(v0011001011)$ gives rise to $((u,v),\alpha\alpha\delta\delta\alpha\alpha\delta\alpha\delta\delta)$. Further note that z_t can contain at most one β, and then only if $u = v$.

Now we assert that ux_t and vy_t are linked for a specific model if, and only if, each of the entries in z_t (such as α, β, etc.) is equal to 1. If we start with $u = v$ then we obtain sequences of the form $z_t = ((u,u) < \alpha, \delta >^k \beta < \alpha, \gamma, \gamma^*, \delta >^{t-k-1})$, where the powers of the sets <> are the direct products. Note here that α's in front of the β must be taken as having value 1 in every model since they relate to the same vertex. If we start with $u \neq v$ then we obtain sequences $z_t = ((u,v) < \alpha, \gamma, \gamma^*, \delta >^t)$.

There is one additional complication in the case where we have self-edges in G_0. We need to consider the ambiguities which may arise if $\beta = 1$ and $\alpha = 1$, since the former acting on a vertex, and the latter acting on a self edge at that vertex will result in the same edge. We can deal with this case efficiently by ensuring that any vertex with a self-edge at any time *t* is only subjected to one of the operators.

For our examples above we have that (A) requires $\alpha = \beta = \gamma = 1$ (recall $\delta = 1$ and $\gamma = \gamma^*$ always) so

there is an edge only in model 7, while (B) requires only that $\alpha = 1$ and $u \neq v$ (note that in the absence of a β, $u = v$ would only lead to two copies of the same vertex) so there is an edge in models 4, 5, 6 and 7.

4. Homogeneity

Definition. Merger of Two Graphs

Given two graphs $G(U,E)$ and $H(V,F)$ then we define the merger of G and H as the graph $J(U \cup V, E \cup F)$, and denote this by $G \uplus H$

Theorem

For each of the models specified above given some $G_0(V_0,E_0)$ then with $G_{t+1} = F(G_t)$ and writing in the obvious way $G_t = F^t(G_0)$ we have that, $G_t = \uplus_{(u,v)\in E_0} F^t\left(L(u,v)\right)$ where $L(u,v)$ is the graph with vertices u and v, and one edge (u,v).

Proof

This follows immediately from the definition of the functions $F_i(G)$. It is clear that for any $G(V,E)$ we have $F_i\left(G \uplus L(u,v)\right) = F_i(G) \uplus F_i\left(L(u,v)\right)$ for each of the possible cases 1) $u \in V$, $v \in E$ and $(u,v) \in E$, 2) $u \in V$, $v \in V$ and $(u,v) \notin E$, 3) $u \in V$ and $v \notin V$, and 4) $u \notin V$ and $v \notin V$. Then by induction the result follows.

In view of the theorem above much of the information is captured by considering the case where G_0 consists of a single edge. We consider how a single edge (and sometimes other equally simple structures) evolve under our models.

4.1. Models 0, 1, 4 & 5; Kronecker Products

As **Table 1** shows, four of the models use the Kronecker product, in fact those for which $\beta = 0$. For such products, which we denote by $\otimes$, we have $A(G \otimes H) = A(G) \otimes A(H)$, where $A(G)$ denotes the adjacency matrix of G, and $\otimes$ also denotes the Kronecker product when applied to matrices. The adjacency matrices for the Z's of models are 0, 1, 4 and 5, respectively,

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

The t th Kronecker powers of these, $2^t \times 2^t$ matrices, are easy to obtain. That for model 0 has a single 1 in the $(2^t, 2^t)$ position, and zeroes elsewhere, model 4 gives the identity matrix, model 5 gives a matrix of 1's. Only model 1 has an interesting pattern, which is essentially the bitwise AND pattern exhibited in [12], and which is discussed below in the Section 8.

4.2. Model 2

Model 2 is particularly simple as we have a tree structure; we simply add a new branch at each vertex. Here we can capture all the structure by starting with G_0 as a single isolated vertex. There are various ways to describe the resulting tree, and these will be explored in more detail in a subsequent paper. For the moment we give only one such description. Starting from a single vertex labeled u, we obtain vertices $u1$ and $u0$ which are linked, then vertices $u11$, $u10$, $u01$ and $u00$. After t steps we have a tree with $2^t - 1$ edges on the vertices of the cube of dimension t. This tree is necessarily a spanning tree. As we make an extra step we take the current cube, with its spanning tree, and make a copy of the cube, join vertices of the $t-1$ dimensional cube to the matching vertex in the copy. An alternate way of expressing this is to consider a t-dimensional cube with all edges present. Choose a particular coordinate and remove all the edges from the cube for which this coordinate is 0, then move to the cube which has this coordinate equal to 1, and within this cube repeat the process. At each stage one removes all the edges of a cube, whose dimension is one smaller than at the previous step.

4.3. Models 3, 6 and 7

Model 3 is by far the most complex and interesting of the models. We shall discuss several aspects of this model, along with the other models, but shall postpone a fuller discussion to subsequent papers.

In model 6 the graph arising from a single edge after t steps is the $(t+1)$-dimensional cube.

In model 7 the graph arising from a single edge after t steps is the complete graph on 2^{t+1} vertices.

5. Numbers of Vertices and Edges

For a general $G_0(V_0,E_0)$ we have immediately that the number of vertices at time t is $|V_0| * 2^t$ for all models. The number of edges on the other hand depends on the particular model, and can be relatively easily deduced from the z_t possibilities and the α etc. appropriate for each model. For example, for model 3, we have $\beta = \delta = \gamma = \gamma^* = 1$, so for $u \neq v$ we have $z^t = ((u,v) < \delta, \gamma, \gamma^* >^t)$ so there are clearly 3^t edges for each (u,v). We have $z_t = \left((u,u) < \alpha, \delta >^k \beta < \delta, \gamma, \gamma^* >^{t-k-1}\right)$ and this results in $(3^t - 2^t)$ edges for each u. The complete set of formulae are given in **Table 2**.

6. Chromatic Number

A vertex colouring of a graph G is the assignment of a colour to each vertex in such a way that no adjacent vertices in G have the same colour. The minimal number of colours required to achieve this is the chromatic number,

Table 2. Formulae describing the number of edges after *t* time steps within the different models.

Model	Number of Edges
0	$\|E_0\|$
1	$3^t \|E_0\|$
2	$(2^t-1)\|V_0\|+\|E_0\|$
3	$(3^t-2^t)\|V_0\|+3^t\|E_0\|$
4	$2^t\|E_0\|$
5	$4^t\|E_0\|$
6	$t*2^{t-1}\|V_0\|+2^t\|E_0\|$
7	$(2*4^{t-1}-2^{t-1})\|V_0\|+4^t\|E_0\|$

which we denote by $\chi(G)$. A colouring which achieves this minimum will be refered to as a minimal colouring. With the exception of model 7, the chromatic number of the growing graphs are easy to obtain.

6.1. Models 0, 1, 4 and 5

Suppose that we have a minimal colouring for G_0 with $\chi(G_0)$ colours. For models 0, 1, 4 and 5, each offspring can be given the same colour as its parent without violating the condition that we have a colouring, and so the chromatic number remains equal to $\chi(G_0)$.

6.2. Models 2 and 6

For models 2 and 6, suppose we have a minimal colouring of G_0 using the set $\{c_0, c_1, \ldots, c_{\chi(G_0)-1}\}$ of $\chi(G_0)$ colours. Now suppose that the offspring of a vertex coloured with c_i is coloured $c_{(i+1) mod \chi(G_0)}$. Then, provided $\chi(G_0) > 1$ this will constitute a minimal colouring for $R_i(G_0)$ for $i = 2$ and for $i = 6$. Thus

$$\chi(G_t) = max(2, \chi(G_0)).$$

6.3. Model 3

In this model $\chi(G_{t+1}) = \chi(G_t)+1$. This is because in any minimal, proper colouring of G_t there will be an individual with $\chi(G_t)-1$ distinct colours amongst its neighbours (actually there will be at least $\chi(G_t)$ such individuals, and since this individual's offspring is joined both to the individual and all its neighbours, this offspring must be given a new colour. This colour can then be given to every offspring.

6.4. Model 7

This is by far the most difficult case, and will be treated more fully elsewhere. We observe only that

$$\chi(G_t)+1 \le \chi(G_t) \le 2\chi(G_t).$$

The first inequality follows since model 3 produces a subgraph of model 7 when they act on the same *G*. The latter inequality is evident since giving a minimal, proper colouring of G_t using colours $\{c_0, c_1, \ldots, c_{\chi(G_0)-1}\}$ we can colour the offspring of any vertex coloured c_i with some c_i^*, from a set $\{c_0^*, c_1^*, \ldots, c_{\chi(G_0)-1}^*\}$ of completely new colours. It is clear that if G_t is a clique, or bipartite, then the chromatic number doubles, but this doubling is not general. For example the chromatic number of any polygonal graph Q of odd degree > 3 is 3, while $\chi(F_7(Q)) = 5$.

7. The Distance Structure

We now turn to the details of the distances between vertices. The distance between vertices *u* and *v* is denoted by $d(u,v)$, the diameter of a graph *g* by $\vartheta(G)$. For a graph G_t we denote the numbers of pairs of vertices with distance *x* as $\eta_t(x)$. For each models we shall derive the recursions for the distances through time. Models 0 and 4 are excluded since they lead to disconnected components for which the notion of distance is inappropriate. We also suppose that our initial graph is connected so that all subsequent graphs are.

7.1. Model 2

We begin with model 2 since this will allow an easy demonstration of our methods. It is clear that

$\vartheta(G_t) = \vartheta(G_{t-1}) + 2$. As in every model if $u \in V_t$ then

$d(u0,u1) = 1$, while if $(u,v) \in E_t$ with $d(u,v) = d$ then

$d(u1,v1) = d$, $d(u1,v0) = d(u1,v0) = d+1$ and

$d(u0,v0) = d+2$. We can then write

$\eta_{t+1}(0) = 2\eta_t(0)$,

$\eta_{t+1}(1) = \eta_t(0) + \eta_t(1)$,

$\eta_{t+1}(2) = 2\eta_t(1) + \eta_t(2)$ and

$\eta_{t+1}(k) = \eta_t(k-2) + 2\eta_t(k-1) + \eta_t(k)$ if $k \ge 3$.

This enables us to derive closed form expressions for the $\eta_t(i)$, for example

$\eta_t(0) = 2^t \eta_0(0)$,

$\eta_t(1) = \eta_0(1) + (2^t-1)\eta_0(0)$,

and

$\eta_t(2) = \eta_0(2) + 2t\eta_0(1) + 2(2^t - t - 1)\eta_0(0)$ but the forms rapidly become somewhat unmanageable.

We can specify the total number of distances, for all models, $L_t = (4^t |V_0|^2 + 2^t |V_0|)/2$ being simply the number of pairs of vertices plus the number of vertices, and being the same for all the models. The total of the distances

$$L_t^* = 4L_{t-1}^* + 4L_{t-1} - 3\eta_{t-1}(0)$$
$$= 4^t L_0^* + 2^{2t-1} t\left(\eta_0(0)\right)^2 + \left(2^{t-1} - 2^{2t-1}\right)\eta_0(0).$$

The average distance $d_t = L_t^*/L_t$ and for t large this gives $d_t \approx c + t$ where $c = (2L_0^* - \eta_0(0))/(\eta_0(0))^2$. The average distance thus increases by 1 at each stage.

We can derive this average result in another, more direct way. Suppose at some stage we have n vertices and average distance μ_t. Now we add the extra offspring. The average distance is just the distance between a randomly selected pair of vertices. We consider the possible pairs in the new graph. There are $n(2n+1)$ such pairs, n are of type $(u0,u1)$ and contribute a distance 1. Of the remaining $2n^2$, 1/4 are of type $(u1,v1)$, 1/2 are of type $(u1,v0)$, and 1/4 are of type $(u0,v0)$, which contribute $d(u1,v1)$, $d(u1,v1)+1$ and $d(u1,v1)+2$ respectively. The average over these $2n^2$ latter will thus be $\mu_t + 1$ and overall we will therefore have

$$\mu_{t+1} = \frac{n + 2n^2(\mu_t + 1)}{n(2n+1)} = 1 + 2n\mu_t/(2n+1).$$

A similar argument can be used to obtain an expression for the variance of the distances which asymptotically increases by 1/2 per time step.

7.2. Model 1

In order to derive the appropriate expressions for model 1 we need to track not only the distances but also the number of edges which belong to triangles. Accordingly we define set $\nabla(G)$ of the edges which belong to a triangle (an edge $(u,v) \in E$ belongs to a triangle if there exists some k such that $(i,k) \in E$ and $(j,k) \in E$). Define $\eta_t^{\nabla}(1) = |\nabla(G)|$ and $\eta_t^*(1) = \eta_t(1) - \eta_t^{\nabla}(t)$. The necessity of considering triangles arises because the offspring of two linked parents will be distance 3 apart if the parents' link is not part of a triangle, but only 2 apart if there is a triangle since they are linked through the common neighbour of their parents.

In detail we have

For all $u \in V$ we have $d(u0,u1) = 2$,

and for $(u,v) \in E$,

if $d(u,v) > 1$ we have

$d(u0,v0) = d(u0,v1) = d(u1,v0) = d(u1,v1) = d(u,v)$,

if $(u,v) \in \nabla(G)$ then $d(u0,v0) = 2$, $d(u0,v1) = d(u1,v0) = 1$, $(u0,v1)\nabla(G)$ and $(u1,v0) \in \nabla(G)$,

and

if $(u,v) \in V \setminus \nabla(G)$ then $d(u0,v0) = 3$, $d(u0,v1) = d(u1,v0) = 1$, $(u0,v1) \notin \nabla(G)$ and $(u1,v0) \notin \nabla(G)$.

We have immediately that

$\vartheta(G_t) = max(\vartheta(G_{t-1}),3) = max(\vartheta(G),3)$.

From these expressions we obtain recursions for the η's, as follows,

$\eta_t(0) = 2\eta_{t-1}(0)$,

$\eta_t^{\nabla}(1) = 3\eta_{t-1}^{\nabla}(1)$

$\eta_t^*(1) = 3\eta_{t-1}^*(1)$

$\eta_t(2) = \eta_{t-1}(0) + \eta_t^{\nabla}(1) + 4\eta_{t-1}(2)$

$\eta_t(3) = \eta_{t-1}^*(1) + 4\eta_{t-1}(3)$

$\eta_t(k) = 4\eta_{t-1}(k)$ if $k > 3$.

From the above recursions we can find simple closed form expressions for the η's, as follows,

$\eta_t(0) = 2^t\eta_0(0)$,

$\eta_t^{\nabla}(1) = 3^t\eta_0^{\nabla}(1)$

$\eta_t^*(1) = 3^t\eta_0^*(1)$

$\eta_t(2) = (4^t - 2^t)\eta_0(0)/2 + (4^t - 3^t)\eta_0^{\nabla}(1) + 4^t\eta_0(2)$

$\eta_t(3) = (4^t - 3^t)\eta_0^*(1) + 4^t\eta_0(3)$

$\eta_t(k) = 4\eta_{t-1}(k)$ if $k > 3$.

The total distance

$L_t^* = 4^t L_0^* + (4^t - 3^t)(2\eta_0^*(1) + \eta_0^{\nabla}(1)) + (4^t - 2^t)\eta_0(0)$,

so that the average distance converges to a constant. The variance of the distances also converges to a constant.

7.3. Model 3

We obtain, in a straightforward manner that

$\eta_t(0) = 2\eta_{t-1}(0)$,

$\eta_t(1) = \eta_{t-1}(0) + 3\eta_{t-1}(1)$,

$\eta_t(2) = \eta_{t-1}(1) + 4\eta_{t-1}(2)$,

$\eta_t(k) = 4\eta_{t-1}(k)$ for $k > 2$.

We have

$\vartheta(G_t) = max(\vartheta(G_{t-1}),2) = max(\vartheta(G),2)$,

and the total distance

$L_t^* = 4^t(L_0^* + \eta_0(0) + \eta_0(1)) - 3^t(\eta_0(0) + \eta_0(1))$ so that the average distance converges to a constant. The variance of the distances also converges to a constant.

7.4. Model 5

We obtain $\eta_t(0) = 2\eta_{t-1}(0)$,

$\eta_t(1) = 4\eta_{t-1}(1)$,

$\eta_t(2) = \eta_{t-1}(0) + 4\eta_{t-1}(2)$,

$\eta_t(k) = 4\eta_{t-1}(k)$ for $k > 2$.

We have

$\vartheta(G_t) = \vartheta(G)$,

and the total distance

$L_t^* = 4^t(L_0^* + \eta_0(0)) - 2^t\eta_0(0)$ so that the average distance converges to a constant. The variance of the distances also converges to a constant.

7.5. Model 6

We obtain $\eta_t(0) = 2\eta_{t-1}(0)$,
$\eta_t(1) = \eta_{t-1}(0) + 2\eta_{t-1}(1)$,
$\eta_t(k) = 2\eta_{t-1}(k-1) + 2\eta_{t-1}(k)$ for $k \geq 2$.
We have
$\vartheta(G_t) = \vartheta(G_{t-1}) + 1$, and
$L_t^* = 4L_{t-1}^* + 2L_{t-1} - \eta_{t-1}(0)$, from which we can easily demonstrate that the average distance L_t^* / L_t increases by 1/2 per time step for large *t*. The variance of the distances increases by 1/4 per time step for large *t*.

7.6. Model 7

We obtain
$\eta_t(0) = 2\eta_{t-1}(0)$,
$\eta_t(1) = \eta_{t-1}(0) + 4\eta_{t-1}(1)$,
$\eta_t(k) = 4\eta_{t-1}(k)$ for $k \geq 2$.
From this we have that the diameter is constant, and
$L_t^* = 4_t L_0^* + (4^t - 2^t)\eta_0(0)/2$,
so the average distance converges to a constant. The variance of the distances also converges to a constant.

8. Automorphism

In models 0, 1, 4 and 5 we have Kronecker products and this allows us to determine the automorphisms on vertices in G_t. In these models for a pair of vertices ur and vs, where *u* and *v* belong to G_0, and *r* and *s* are binary strings, to be automorphic in G_t, they must have *u* and *v* automorphic in G_0 and *r* and *s* to be automorphic in the Kronecker product of the appropriate Z. This is straightforward in each of the models. The most interesting case is model 1. For this the adjacency matrix is

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

The *n* th Kronecker product, A_k, of this matrix have its rows and columns naturally indexed by the *n* th Kronecker products of vectors $(0,1)$ and $(0,1)^T$, and will have a zero wherever the row index and column index have a 0 in the same position. It immediately follows that the matrix A_k is invariant under any permutation to the elements of the row and column indices. Accordingly the permutations induce an equivalence relation over the binary strings; two strings being automorphic if they contain the same number of 0's.

9. Generating All Graphs

Now [4] proved that for any graph $H(U,F)$ of order *n* there exists a set *W* of size $\leq n^2/4$ such that one can associate distinct subsets W_i with the n vertices, such that $(i,j) \in F$ if, and only if, $W_i \cap W_j = \phi$. Consider the case where we evolve a single vertex, linked to itself, for *t* time steps under model 1. The vertex set of the resulting graph will be the set of *t* length binary strings. Suppose each string x_t is associated with a set $S(x_t) = \{i \mid x_t^i = 0\}$. Then two vertices, x_t and y_t, are joined if, and only, $S(x_t) \cap S(y_t) = \phi$. It follows that every graph with *n* vertices is isomorphic to some subgraph of the *t*'th iterate when $t \geq \lceil n^2/4 \rceil$.

Now as pointed out in [4] the bound is exact only when the graph *H* is bipartite with vertices partitioned equally, when *n* is even, and differing by 1 when *n* is odd. Thus we will observe many graphs will appear at earlier stages, for example the graph K_n, the complete graph on *n* vertices, will appear at time $t = n-1$, since we may take $W_i = \{i\}$. We shall investigate this phenomenon elsewhere.

10. The Degree Distribution

Since we have a deterministic process the degrees of the vertices are well defined. We denote by $D(G,x)$ the number of vertices of degree *x* in graph *G*, and refer to the degree of any vertex *x* as $deg(x)$. We shall refer to the degree distribution by which we mean the distribution of the degree of a randomly chosen vertex.

10.1. Degree Distribution; Models 0, 1, 4, 5

We begin with the models which are Kronecker products. Given two graphs $J = (V,E)$ and $K = (W,F)$ with $y \in V$ with $deg(y)$ and $z \in W$ with $deg(z)$, then for $(y,z) \in J \otimes K$ we have $deg((y,z)) = deg(y) * deg(z)$. It follows that

$$D(J \otimes K, x) = \Sigma_{j \in \zeta(i)} D(J,j) D(K, i/j)$$

where $\zeta(i) = \{j \mid j \in N, j \mid i\}$, $j \mid i$ having the usual meaning that *j* divides *i*.

Now in these models we have $G_t = G_0 \otimes Z_t$, where Z_t is the graph obtained by taking the graph Z (as described in **Table 1**) and taking the Kronecker product of it with itself *t* times. By knowing the degree distribution of Z_t one can easily determine the distribution starting from a generic initial graph. Under model 0, Z_t has $(2^t - 1)$ isolated vertices and a single vertex with degree 1, model 1 has ${}_tC_r$ vertices with degree 2^r, model 4 has every vertex with degree 1, while model 5 has every vertex with degree 2^t.

10.2. Degree Distribution; Model 2

At each stage all the vertices of G_t increase their de-

gree by 1, and a set of $|G_t|$ vertices of degree 1 are added. Thus $D(G_t, x) = \sum_{r=0}^{t} ({}_tC_r) D(G_0, x-r)$.

10.3. Degree Distribution; Model 6

For model 6 we have a Cartesian product (which we denote $\odot$). For graphs $J = (V, E)$ and $K = (W, F)$ with $y \in V$ with $deg(y)$ and $z \in W$ with $deg(z)$, then for $(y, z) \in J \odot K$ we have $deg((y, z) = deg(y) + deg(z)$. Thus

$$D(J \odot K, x) = \sum_{j=0}^{i} D(J, j) + D(K, i/j)$$

and since the Cartesian power $\odot^t G(\{0,1\},(0,1))$ is simply the *t* dimensional cube we have that

$$D(G_t, x) = 2^t D(G_0, x-t).$$

10.4. Degree Distribution; Models 3 and 7

As usual model 3 is the most interesting, and the most difficult model to deal with. A vertex $v \in G_t$ with degree *x* gives rise to $v1$ with degree $2x+1$ and to $v0$ with degree $x+1$. Thus

$$D(G_t, x) = D(G_{t-1}, x-1) + D(G_{t-1}, (x-1)/2$$

A plot of the frequency distribution of degrees for t = 17 is shown in **Figure 2**.

This distribution will be explored further in subsequent papers.

In model 7 a vertex $v \in G_t$ with degree *x* gives rise to two vertices, $v1$ and $v0$, each with degree $2x+1$. If follows that

$$D(G_t, x) = 2^t D\left(G_0, \left(x+1-2^t\right)/2^t\right)$$

11. Discussion

We have presented a variety of models for the growth of networks based on parent-offspring links and suggested that these might be used to describe the growth of interactions between individuals within a population.

This description might well be used to represent the binding of proteins under gene- and or genome-duplications. Alternately we might be describing the interactions within a population of organisms where these interactions depend on the interactions which existed amongst those in the previous generation, such as dominance relations (though these might require a directed graph approach).

The model as formulated has deliberately been kept as simple as possible. Thus the model is deterministic. The deterministic assumption would rarely apply in a biological context, but might in a computer context. We have indicated some ways in which a stochastic element can be introduced. One can vary the transition function applied as a stochastic process, or one can vary the links made by making the parameters of the models probabilities, rather than 0 or 1.

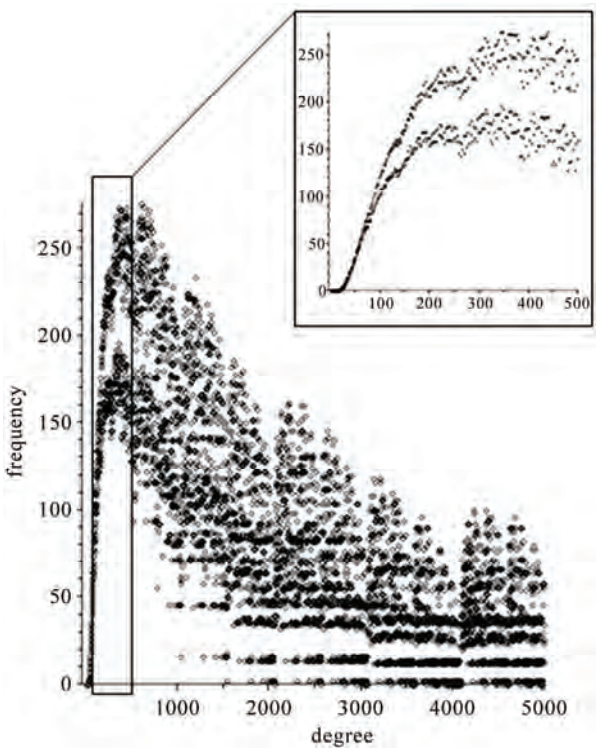


Figure 2. A truncated plot showing the frequencies of vertices with degree ≤ 5000 within the graph obtained by evolving a single edge for 17 time steps under model 3.

The simplicity of our model has allowed us to derive many results. Perhaps the most important of these is the theorem from Section 4. This theorem highlights the fact that the growth of any subgraph is independent of the nature of its exterior surroundings. By using this result and exploiting relations with graph products and binary strings we have derived formulas that describe chromatic number, distance structure and degree distributions.

The model has immortal vertices and edges. In subsequent papers we shall consider models with a similar reproductive structure, but allow for the death of vertices. In the next paper we shall treat the case where individuals have a fixed lifetime, and in the third we shall apply a threshold to the degree of a vertex, nodes which accumulate too high a degree will die. Naturally both of these processes could be made stochastic. In these models edges, once established through the reproductive phase disappear solely because of the death of one of their vertices. Additional features which we shall add in the future include sexual reproduction, and the embedding of the graph in space.

12. Acknowledgements

The authors gratefully acknowledge support from the EPSRC (Grant EP/D003105/1) and helpful input from Dr Jonathan Jordan, and others in the Amorph Research Project (www.amorph.group.shef.ac.uk).

13. References

[1] G. U. Yule, "A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S.," *Philosophical Transactions of the Royal Society of London* (*Series B*), Vol. 213, 1925, pp. 21-87.

[2] H. A. Simon, "On a Class of Skew Distribution Functions," *Biometrika*, Vol. 42, No. 3-4, 1955, pp. 425-440.

[3] A. L. Barab´asi and R. Albert, "Emergence of Scaling in Random Networks," *Science*, Vol. 286, No. 5439, 1999, pp. 509-512.

[4] B. Bollob'as, O. Riordan and G. Spenser, "The Degree Sequence of a Scale-Free Random Graph," *Random Structures and Algorithms*, Vol. 18, No. 3, 2001, pp. 279-290.

[5] J. Jordan, "The Degree Sequences and Spectra of Scale-Free Random Graphs," *Random Structures and Algorithms*, Vol. 29, No. 2, 2006, pp. 226-242.

[6] C. Cannings and A. W. Thomas, "Handbook of Statistical Genetics," John Wiley and Sons Ltd, New York, 2007.

[7] J. S. Taylor and J. Raes, "Duplication and Divergence: The Evolution of New Genes and Old Ideas," *Annual Review of Genetics*, Vol. 38, No. 1, 2004, pp. 615-643.

[8] A. Widdig, P. Nurnberg, M. Krawczak, W. J. Streich and F. B. Bercovitch, "Paternal Relatedness and Age Proximity Regulate Social Relationships among Adult Female Rhesus Macaques," *Proceedings of the National Academy of Science USA*, Vol. 98, No. 24, 2001, pp. 13769-13773.

[9] G. Hausfater, "Dominance and Reproduction in Baboons (Papio Cynocephalus)," *Contributions to Primatology*, Vol. 7, 1975, pp. 1-150.

[10] L. Frank, K. Holekamp and L. Smale, "Serengeti II: Dynamics, Management, and Conservation of an Ecosystem," University of Chicago Press, Chicago, 1995.

[11] W. Imrich and S. Klavar, "Product Graphs: Structure and Recognition," John Wiley and Sons Ltd, New York, 2000.

[12] S. Wolfram, "A New Kind of Science," Wolfram Media, Inc., Champaign, 2002.

[13] P. Erd¨os, A. W. Goodman and L. Posa, "The Representation of a Graph by Set Intersections," *Canadian Journal of Mathematics*, Vol. 18, No. 1, 1966, pp. 106-112.

Applied Mathematics, **2010, 1, 146-152**
doi:10.4236/am.2010.13019 Published Online September 2010 (http://www.SciRP.org/journal/am)

Predefined Exponential Basis Set for Half-Bounded Multi Domain Spectral Method

Fahhad Alharbi
King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia
E-mail: *alharbif@kacst.edu.sa*
Received March 25, 2010; *revised June* 29, 2010; *accepted July* 2, 2010

Abstract

A non-orthogonal predefined exponential basis set is used to handle half-bounded domains in multi domain spectral method (MDSM). This approach works extremely well for real-valued semi-infinite differential problems. It spans simultaneously wide range of exponential decay rates with multi scaling and does not suffer from zero crossing. These two conditions are necessary for many physical problems. For comparison, the method is used to solve different problems and compared with analytical and published results. The comparison exhibits the strengths and accuracy of the presented basis set.

Keywords: Multi-Domain Spectral Method, Meshfree Numerical Method, Non-Orthogonal Predefined, Exponential Basis Set, Half-Bounded Domain, Exponential Decay, Quantum Wells, Optical Waveguide

1. Introduction

With the growing complexities of the numerically studied problems in natural and applied sciences, spectral method (SM) starts gaining more attention mainly because of its high level of analyticity. This is resulted from its meshfree nature, which reduces the computational memory and time requirements where a major part of the problem is solved analytically. Different spectral methods have been developed since 1970s mainly from applied mathematical perspective. Despite of their superior mathematical performance, applying them was very limited when compared to finite difference (FDM) and finite element methods (FEM).

As known, SM is a special family of the weighted residual methods where the unknown functions are approximated by either an expansion of or interpolation using preselected basis sets. In this paper, the expansion method is used. Spectral methods work very well for homogeneous and smooth computational windows. But, it suffers from the Gibbs phenomenon if the structural function of the studied problem is not analytical. The Gibbs phenomenon is generally a peculiarity of any functional approximation at simple discontinuity. To avoid this problem, multi domain spectral method (MDSM) is developed where the computational window is divided into homogeneous domains where the discontinuities lie at the boundaries. Then, the spectral method is applied in each domain alone to build the matrices and vectors. These are then joined by applying the proper boundary conditions between domains [1-4].

2. The Exponential Basis Set

In many real-valued physical system, the extensions toward infinities decay exponentially as:

$$f \propto e^{\pm \alpha x} \tag{1}$$

where ± is used to cover both $\mp\infty$ with positive α. In SM and MDSM, this is one of the main problems [5-9]. A review paper by Shen and Wang discusses this in further details [10]. To overcome this problem, many techniques where used. They can be classified in the following three main categories:

1) Using exponentially decaying functions such as physical Hermite and Laguerre functions and rational Chebyshev and Legendre polynomials. Some other basis are used as well. Physical Hermite and Laguerre functions are decaying exponentially toward infinities as $e^{-x^2/2}$ and $e^{-x/2}$ respectively. This predefines a narrow ranged decay rate and hence limit the generality. They were adopted in studying phenomena that were known that they can be analyzed using such functions. For example, Laguerre function is the base for radial extension of

electron wave functions in hydrogenic atoms. So, it is expected to work for electronic distributions of some hydrogenic like atoms.

In rational Chebyshev and Legendre polynomials, the coordinates are transformed rationally to map $(-\infty,\infty)$ or $[0,\infty)$ into $[-1,1]$. Other forms of coordinate transformation were used as well. However, in general, this method suffers from the narrow ranged predefined decay rate. Also, mapping infinite intervals into small finite ones adds often some complexity and approximation errors.

Beside the narrow ranged predefined decay rate and because a finite number of bases can be used practically, this approach inherently forces many zero crossing since most of the used bases are forms of Jacobi polynomials which has N zeros for the N^{th} order polynomial. In many physical problems, this is expected and allowed. However, in physical system where the decay is behaving as described by Equation (1), this should not be the case.

2) Truncating the numerical window; this is used widely as well. The unbounded window is truncated and additional boundary conditions are used to force an asymptotic exponential behavior, *i.e.*, the function and its first derivative vanish at the truncating points. This reduces the analytical accuracy of SM by adding the truncation error. Also, this does not eliminate zero crossing and hence it doesn't fit the system with Equation (1) exponential decay.

3) Single scaling of the coordinates; this is similar to the first category; but with coordinate scaling where

$$x \Rightarrow cx \tag{2}$$

Therefore the predefined decay rate is also scaled. The scaling factor c is chosen intuitively to fit the studied system. However, this results in losing the generality and missing many eigen solutions in eigenvalue problems where the decaying rates for the different eigenvalue solutions are generally different.

The presented method overcomes zero-crossing and single scaling problems by approximating the decaying domain functions by exponential basis set which spans wide range of decaying rates as follows:

$$f = \sum_{n=0}^{N} a_n u_n = \sum_{n=0}^{N} a_n e^{-\alpha_n x} \tag{3}$$

where

$$\alpha_n = b^{d_s + \frac{n}{N}(d_e - d_s)} \tag{4}$$

b is the used exponential base and d_s and d_e are the smallest and largest used powers respectively. They should be predefined intuitively based on the studied problem. But, they allow many possible decay rates with very small number of bases. For example, by setting N = 10, b = 10, d_s = -5, and d_e = 5, 11 basis can be used to approximate any exponential function with decay rates between 0.00001 and 100000.

This basis set is then used directly within the frame of MDSM where Tao method is used for boundary conditions [1,4]. Since the base functions are only exponentials, differentiation and integration can be handled easily and analytically. For detailed implementation of MDSM, we would refer the readers to [1,4] for the details. However, listed below are some of the main implementation issues of the presented basis set within the frame of MDSM.

2.1. Coordinate Transformation

The selected bases form (Eg. 3) assumes that $x \in [\kappa,\infty)$. For simplicity, κ is set to 0. As standardly done in MDSM, coordinate transformation is used to extend the range of the selected basis set. The presented basis set can be used only for domains that extended into $\pm\infty$. So, the following coordinate transformations can be used.

1) For domains where $x \in [\upsilon,\infty)$,

$$x \Rightarrow x' + \upsilon \tag{5}$$

2) For domains where $x \in (-\infty,\omega]$,

$$x \Rightarrow x' + \omega \tag{6}$$

2.2. Scalar Products

As known, MDSM is highly associated with calculating the scalar products $(u_m|H|u_n)$. Hereunder listed are some of the common scalar products that are used in MDSM:

1) if H is a multiplication by a constant d, then:

$$(u_m|d|u_n) = \frac{d}{\alpha_n + \alpha_m} \tag{7}$$

2) if H is a multiplication by x'^k, then:

$$\left(u_m\left|x'^k\right|u_n\right) = \frac{k!}{(\alpha_n + \alpha_m)^{k+1}} \tag{8}$$

3) if H is k^{th} derivative, then:

$$\left(u_m\left|\frac{d^k}{dx'^k}\right|u_n\right) = \frac{(-\alpha_n)^k}{\alpha_n + \alpha_m} \tag{9}$$

3. Comparisons

The presented set was used to analyze successively many optical and quantum electronics. To illustrate its accuracy and validity, two application examples are shown.

The first one is to approximate a wide range of exponentially decaying functions using the same basis sets, while the second is to obtain the quantized energy states in semiconductor quantum well (QW).

3.1. Approximating Wide Range of Exponentially Decaying Functions

In this subsection, the presented set is applied to approximate five exponentially decaying functions and is compared with an approximation using physical Laguerre basis set. The used decaying rates are 0.00535, 0.0632, 0.752, 81.2, and 926. These were picked randomly. **Figure 1** shows the obtained approximation using an unscaled physical Laguerre basis set. It is clear that with 25 physical Laguerre functions, only two functions out of the five are approximated adequately. Yet, the method is converging but rather slowly for the remaining functions. For $e^{-0.752x}$, very few bases are needed to approximate the function to an adequate accuracy. If scaling was used, only two or mostly three functions would be approximated adequately depending on the scaling. For many eigenvalue problems, this is a very serious limitation where different eigenvalues have difierent decaying rates. So, only a small range of the eigenvalues can be obtained accurately.

The same five functions are approximated using the presented exponential basis set with $b = 10$, $d_s = -4$, and $d_e = 4$. The resulted approximations and their associated errors are shown in **Figure 2**. All the five functions were approximated adequately using the same exponential basis set and the convergence is geometric as can been seen. In many applications, it is crucial to find wide range of eigenvalues. By using the presented basis set,

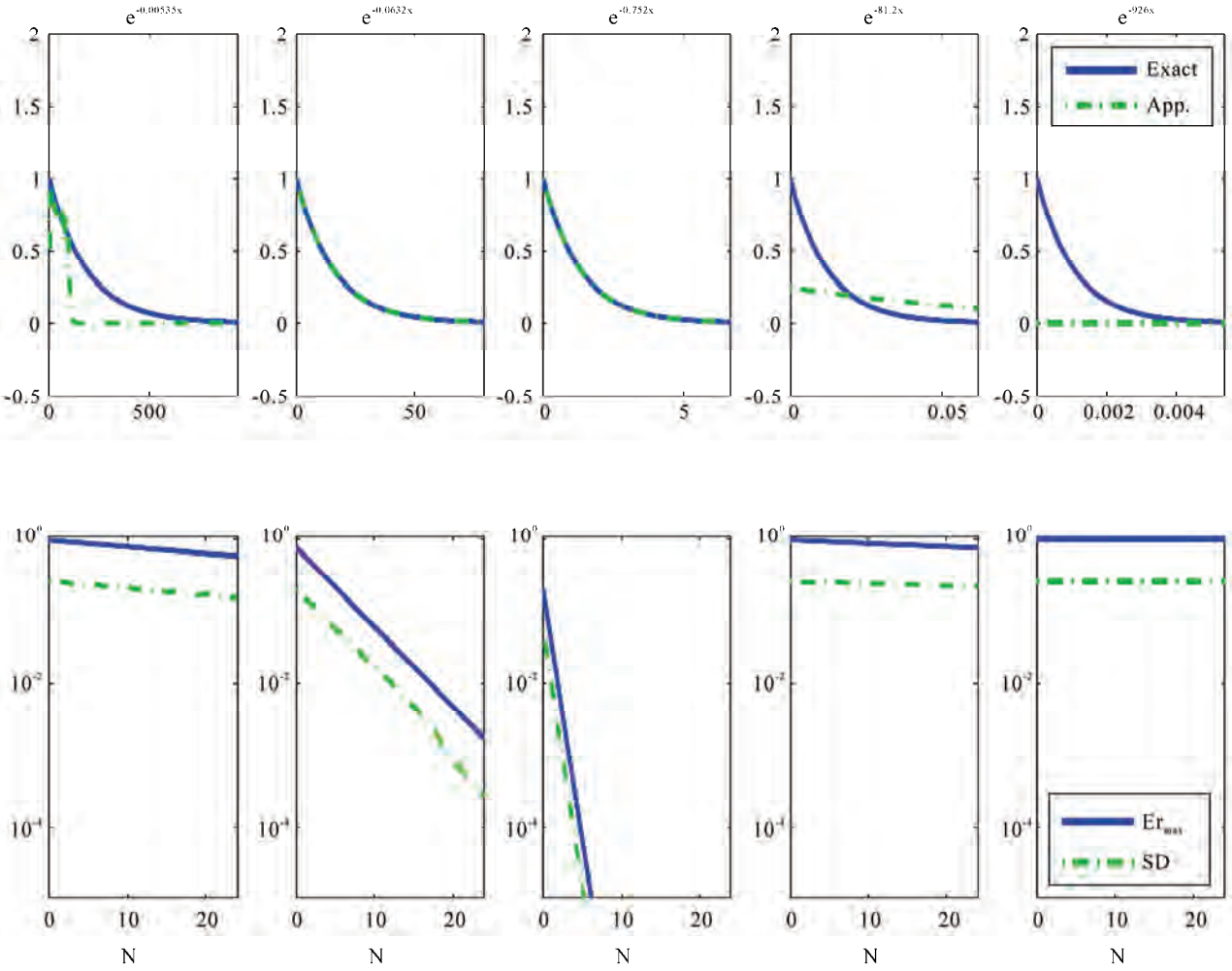


Figure 1. The approximations (top) and the maximum approximation errors and their standard deviations (bottom) of the five exponentially decaying functions using unscaled physical Laguerre basis set. The used decaying rates are 0.00535, 0.0632, 0.752, 81.2, and 926.

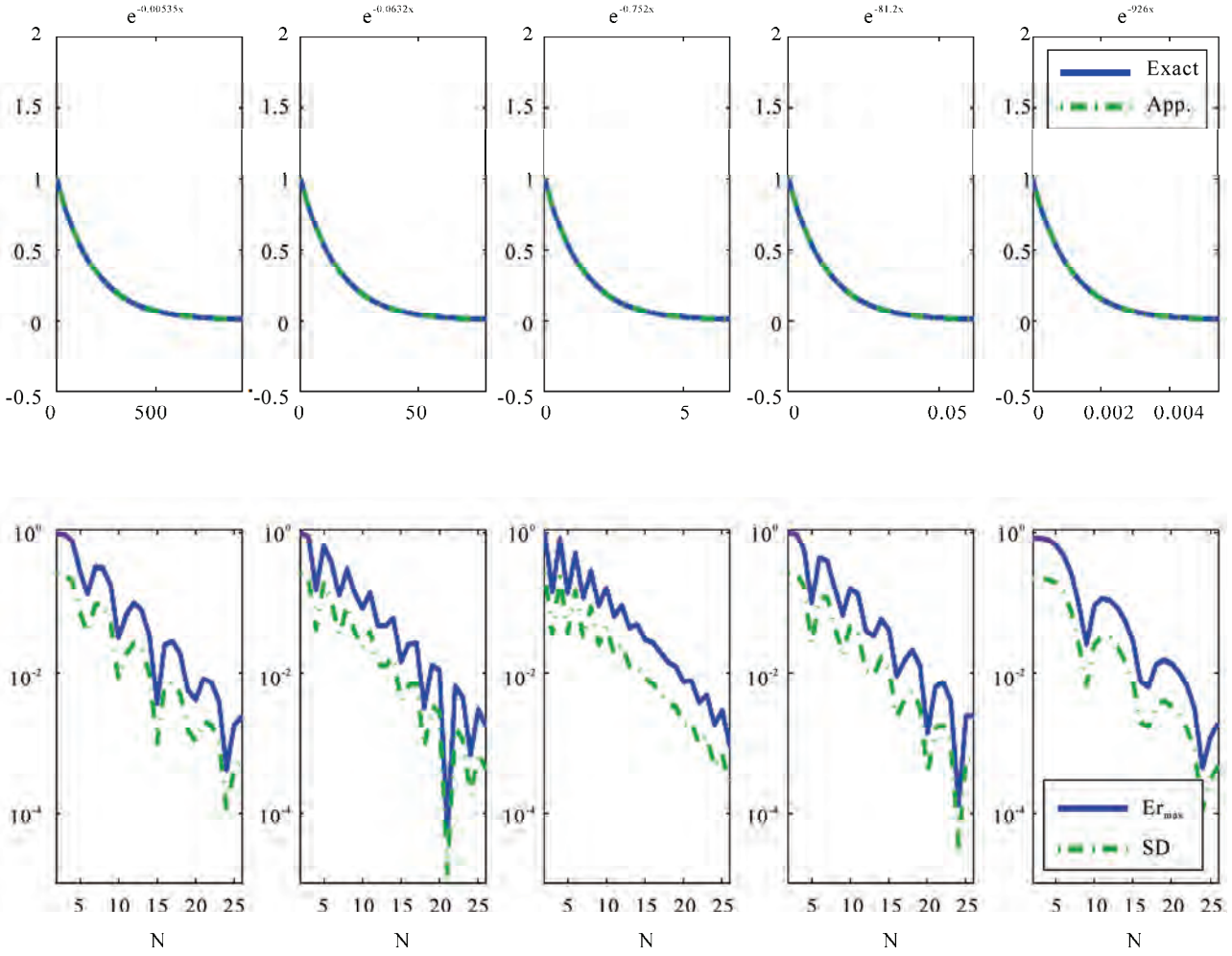


Figure 2. The approximations (top) and the maximum approximation errors and their standard deviations (bottom) of the five exponentially decaying functions using the presented basis set with the same bases. The used decaying rates are 0.00535, 0.0632, 0.752, 81.2, and 926.

this can be done explicitly.

The presented set is used also to approximate wider spectrum of decaying rates using the same bases. The used decaying rates in this analysis are 1:35 × 10^{-6}, 2.32 × 10^{-4}, and 3:2, and 4120, and 526000. Again, these were picked randomly. In this analysis, the used parameters are b = 10, d_s = −6, and d_e = 6. **Figure 3** shows the obtained approximations and their associated errors. It is clear that the approximation is very efficient for all the functions even with this wider spectrum of decaying rates.

3.2. Single Semiconductor QW without Biasing Field

In this subsection, a semiconductor quantum well is analyzed by MDSM and using the presented basis set. Electronic states in QWs are described by the envelopes of the Bloch wave function. These are the solutions of the effective mass approximation of Schrodinger wave equation, which is:

$$\frac{-\hbar^2}{2}\frac{d}{dx}\left(\frac{1}{m^*(x)}\frac{d\phi(x)}{dx}\right)+V(x)\phi(x)=\mathcal{E}\phi(x) \qquad (10)$$

where $\hbar$ is the normalized Planck constant, $m^*(x)$ is the effective mass, $V(x)$ is the QW potential in the growth direction, which is assumed to be the x-axis, $\mathcal{E}$ is the quantized energy level, and $\phi(x)$ is the quantized state. The studied structure is simply a thin layer of GaAs sandwiched in $Al_{0.3}Ga_{0.7}As$. The width of the QW layer is 20 nm. The structure is divided into three domains and the structural parameters and the selected expansion basis sets in each domain are shown in the **Table 1**. This structure can be analyzed analytically, where the energy states are the solutions of the following characteristic

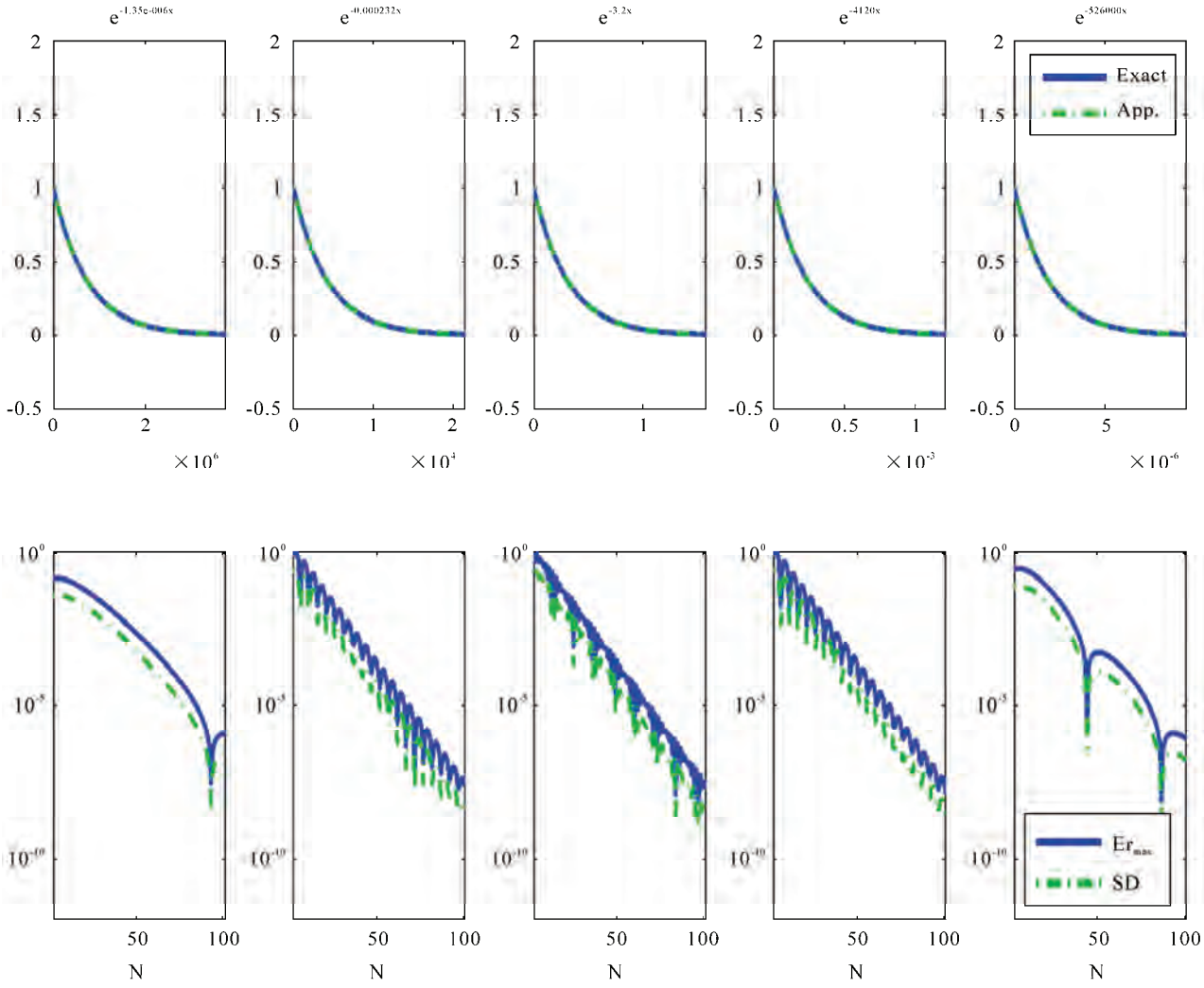


Figure 3. Approximations (top) and maximum approximation errors and their standard deviations (bottom) of the five exponentially decaying functions using the presented basis set with the same bases. The used decaying rates are 1.35×10^{-6}, 2.32×10^{-4}, and 3.2, and 4120, and 526000.

equation:

$$\left(1+\gamma^2\right)e^{i\delta L}-\left(1-\gamma^2\right)e^{-i\delta L}=0 \tag{11}$$

where

$$\delta=\sqrt{\mathcal{E}\frac{2m_w}{\hbar^2}} \tag{12}$$

$$\beta=\sqrt{\left(V_b-\mathcal{E}\right)\frac{2m_b}{\hbar^2}} \tag{13}$$

and

$$\gamma=i\frac{m_b}{m_w}\frac{\delta}{\beta} \tag{14}$$

where m_b and m_w are the effective masses in the barrier and the well and L is the width of the well. **Table 2** shows the obtained results and compare them to the analytical solutions and the results obtained recently by Huang using collocation spectral method (CSM) [11]. For these results, 15 basis are used for each domain. This is really overkilling in accuracy; but it is shown to reveal the accuracy of the presented method. The relative errors of the results obtained using the presented method and the exact solution are shown in **Figure 4** against the number of the used basis in each domain. It is clear that acceptable results can be achieved with only 9 bases in each domain. In QWs, an accuracy tolerance of 0.001 meV is usually suffcient. The speed of the method mainly depends on the largest used matrix in the analysis. Since three domains used in this analysis and the used number of bases is the same on all of them (this is not necessary), the largest matrix size is $3N \times 3N$ where N is the number of the used bases in each domain. We reach

Table 1. The structural parameters of the studied QW.

	Interval (nm)	$m^* = m_0$	$V(x)$ (meV)	Used basis set
D1	$(-\infty,-10)$	0.0919	225	The presented exponential basis set
D2	$(-10,10)$	0.067	0	Chebyshev basis set (the first kind)
D3	$(10,\infty)$	0.0919	225	The presented exponential basis set

Table 2. Energy levels of the first studied QW in meV.

Energy level	Exact	This work	CSM
E_1	9.965713824282	9.965713824282	9.965616696787
E_2	39.766000321692	39.766000321691	39.765614440578
E_3	88.920714571890	88.920714571958	88.9198632662562
E_4	155.586554286161	155.586554286589	155.585132856243

the machine accuracy with 15 basis in each domain where the largest matrix is only 45 × 45. This is handled very easily and rapidly. The whole analysis lasted about half a second. The envelopes of the four quantized states are shown in **Figure 5**.

4. Conclusions

A non-orthogonal basis set for half-bounded domain is presented and applied within the frame of MDSM. It is applied and compared with other methods and exhibits very high accuracy and geometric convergence. In MDSM, approximating exponentially decaying functions is one of the main problems mainly because of zero crossing and narrow-ranged decaying rates. The presented basis set overcomes this for real-valued functions.

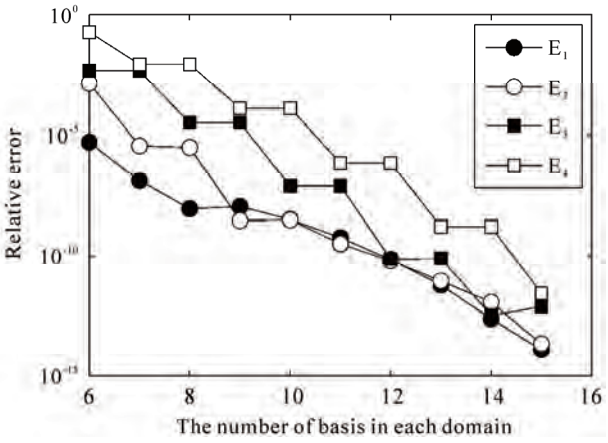


Figure 4. The relative error in energy levels between MDSM using the presented exponential set and the exact solutions of the studied QW.

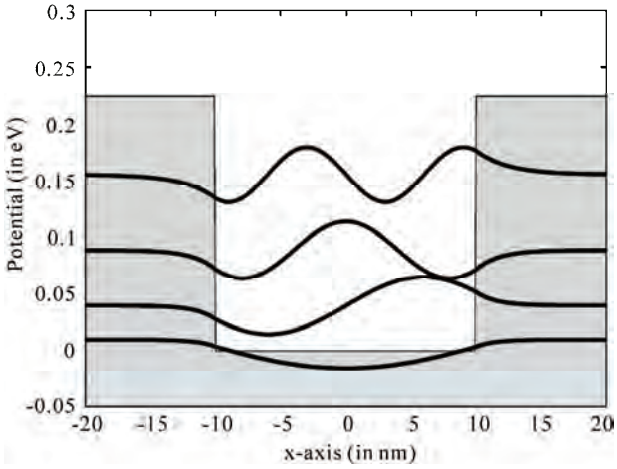


Figure 5. The envelopes of the four quantized states in the studied QW.

With the proper selection of parameters, it covers huge spectrum of decaying rates as shown.

5. References

[1] P. Grandclement and J. Novak, "Spectral Methods for Numerical Relativity," *Living Reviews in Relativity*, Vol. 12, No. 1, 2009, pp. 1-107.

[2] J. Boyd, "Chebyshev and Fourier Spectral Methods," Dover Publications, Mineola, 2001.

[3] C. G. Canuto, M. Y. Hussaini, A. M. Quarteroni and T. A. Zang, "Spectral Methods: Fundamentals in Single Domains," 1st Edition, Springer, New York, 2006.

[4] A. Toselli and O. Widlund, "Domain Decomposition Methods," Springer, Berlin, 2004.

[5] D. Fructus, D. Clamond, J. Grue and Â. Kristiansen, "An Efficient Model for Three-Dimensional Surface Wave

Simulations: Part I: Free Space Problems," *Journal of Computational Physics*, Vol. 205, No. 2, 2005, pp. 665-685.

[6] B.-Y. Guo and J. Shen, "Irrational Approximations and their Applications to Partial Differential Equations in Exterior Domains," *Advances in Computational Mathematics*, Vol. 28, No. 3, 2008, pp. 237-267.

[7] B.-Y. Guo, "Jacobi Spectral Approximations to Differential Equations on the Half Line," *Journal of Computational Mathematics*, Vol. 18, No. 1, 2000, pp. 95-112.

[8] J. Valenciano and M. Chaplain, "A Laguerre-Legendre Spectral-Element Method for the Solution of Partial Differential Equations on Infinite Domains: Application to the Diffusion of Tumour Angiogenesis Factors," *Mathematical and Computer Modelling*, Vol. 41, No. 2-3, 2005, pp. 1171-1192.

[9] V. Korostyshevskiy and T. Wanner, "A Hermite Spectral Method for the Computation of Homoclinic Orbits and Associated Functionals," *Journal of Computational and Applied Mathematics*, Vol. 206, No. 2, 2007, pp. 986-1006.

[10] J. Shen and L.-L. Wang, "Some Recent Advances on Spectral Methods for Unbounded Domains," *Communications in Computational Physics*, Vol. 5, No. 2-4, 2009, pp. 195-241.

[11] C.-C. Huang, "Semiconductor Nanodevice Simulation by Multidomain Spectral Method with Chebyshev, Prolate Spheroidal and Laguerre Basis Functions," *Computer Physics Communications*, Vol. 180, No. 1, 2009, pp. 375-383.

Applied Mathematics, **2010, 1, 153-158**
doi:10.4236/am.2010.13020 Published Online September 2010 (http://www.SciRP.org/journal/am)

Scientific Research

Modified Efficient Families of Two and Three-Step Predictor-Corrector Iterative Methods for Solving Nonlinear Equations

Sanjeev Kumar[1], Vinay Kanwar[2], Sukhjit Singh[3]
[1]*Department of Applied Sciences, ICL Institute of Engineering and Technology, Sountli, India*
[2]*University Institute of Engineering and Technology, Panjab University, Chandigarh, India*
[3]*Department of Mathematics, Sant Longowal Institute of Engineering and Technology, Longowal, India*
E-mail: *sanjeevbakshi1@gmail.com, vmithil@yahoo.co.in, sukhjit_d@yahoo.com*
Received June 9, 2010; *revised July* 13, 2010; *accepted July* 16, 2010

Abstract

In this paper, we present and analyze modified families of predictor-corrector iterative methods for finding simple zeros of univariate nonlinear equations, permitting $f'(x)=0$ near the root. The main advantage of our methods is that they perform better and moreover, have the same efficiency indices as that of existing multipoint iterative methods. Furthermore, the convergence analysis of the new methods is discussed and several examples are given to illustrate their efficiency.

Keywords: Nonlinear Equations, Iterative Methods, Multipoint Iterative Methods, Newton's Method, Traub-Ostrowski's Method, Predictor-Corrector Methods, Order of Convergence

1. Introduction

One of the most important and challenging problems in computational mathematics is to compute approximate solutions of the nonlinear equation

$$f(x)=0 \tag{1}$$

Therefore, the design of iterative methods for solving the nonlinear equation is a very interesting and important task in numerical analysis. Assume that Equation (1) has a simple root *r* which is to be found and let x_0 be our initial guess to this root. To solve this equation, one can use iterative methods such as Newton's method [1,2] and its variants namely, Halley's method [1-6], Chebyshev's method [1-6], Chebyshev-Halley type methods [6] etc. The requirement of $f'(x)\neq 0$ is an essential condition for the convergence of Newton's method. The above-mentioned variants of Newton's method have also two problems which restrict their practical applications rigorously. The first problem is that these methods require the computation of second order derivative. The second problem is that like Newton's method, these methods require the condition that $f'(x)\neq 0$ in the vicinity of the root.

For the first problem, Nedzhibov *et al.* [5] derived many families of multipoint iterative methods by discretizing the second order derivative involved in Chebyshev-Halley type methods [6]. We mention below only one root-finding technique (2.1) from [5], namely

$$\left.\begin{aligned} z_n &= x_n - \frac{f(x_n)}{f'(x_n)}, \\ x_{n+1} &= z_n - \frac{f(x_n)}{f'(x_n)}\frac{f(z_n)}{f(x_n)-2\lambda f(z_n)} \end{aligned}\right\}, \tag{2}$$

where $\lambda\in\Re$. For different specific values of λ, various multipoint iterative methods may result from (2).

For $\lambda=1$ in (2), we get

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}\left\{\frac{f(x_n)-f(z_n)}{f(x_n)-2f(z_n)}\right\}. \tag{3}$$

This is the famous Traub-Ostrowski's formula [1,2,4,5, 7,8], which is an order four formula. This method requires one evaluation of the function and two evaluations of its derivative per iteration. Thus the efficiency index [2] of this method is equal to $\sqrt[3]{4}\cong 1.587$ which is better than the one of Newton's method $\sqrt[2]{2}\cong 1.414$. Furthermore, Sharma and Guha [8] have developed a variant

of Traub-Ostrowski's method (3) which is defined by

$$\left.\begin{aligned} z_n &= x_n - \frac{f(x_n)}{f'(x_n)}, \\ y_n &= z_n - \frac{f(z_n)}{f'(x_n)}\frac{f(x_n)}{f(x_n)-2f(z_n)}, \\ x_{n+1} &= y_n - \frac{f(y_n)}{f'(x_n)}\frac{f(x_n)+a f(z_n)}{f'(x_n)+(a-2)f(z_n)}, \end{aligned}\right\} \quad (4)$$

where $a \in \Re$ is a parameter. This family requires an additional evaluation of function $f(x)$ at the point iterated by Traub-Ostrowski's method (3), consequently, the local order of convergence is improved from four to six. For $a = 0$, we obtain the method developed by Grau and Díaz-Barrero [7] defined by

$$\left.\begin{aligned} z_n &= x_n - \frac{f(x_n)}{f'(x_n)}, \\ y_n &= z_n - \frac{f(z_n)}{f'(x_n)}\frac{f(x_n)}{f(x_n)-2f(z_n)}, \\ x_{n+1} &= y_n - \frac{f(y_n)}{f'(x_n)}\frac{f(x_n)}{f'(x_n)-2f(z_n)}, \end{aligned}\right\}. \quad (5)$$

All these multipoint iterative methods are variants of Newton's method. Therefore, they require sufficiently good initial approximation and fail miserably like Newton's method if at any stage of computation, the derivative of the function is zero or very small in the vicinity of the root.

Recently, Kanwar and Tomar [3,4] proposed an alternative to the failure situation of Newton's method and its various variants. They also derived modifications over the different families of Nedzhibov *et al.* [5] multipoint iterative methods. Unfortunately, the various families introduced by Kanwar and Tomar [3] produces only multipoint iterative methods of order three.

Recently, Mir *et al.* [9] have proposed a new predictor-corrector method (designated as Simpson-Mamta method (SM)), which is defined by

$$\left.\begin{aligned} z_n &= x_n - \frac{2f(x_n)}{f'(x_n)+p\sqrt{f'^2(x_n)+4f^2(x_n)}}, \\ x_{n+1} &= x_n - \frac{6f(x_n)}{f'(x_n)+4f'\left((z_n+x_n)/2\right)+f'(z_n)}, \end{aligned}\right\}. \quad (6)$$

where *p* is chosen as a positive or negative sign so as to make the denominator largest in magnitude. This method is obtained by combining the quadratically convergent method due to Mamta *et al.* [10] and cubically convergent method due to Hasnov *et al.* [11]. This method will not fail like existing methods if $f'(x)$ is very small or even zero in the vicinity of the root. This method requires one evaluation of the function and three evaluations of its derivative per iteration. Thus the efficiency index of this method is equal to $\sqrt[4]{3} \cong 1.316$ which is not better than the one of Newton's method $\sqrt[2]{2} \cong 1.414$ or Traub-Ostrowski's method $\sqrt[3]{4} \cong 1.587$.

More recently, Gupta *et al.* [12] have developed a family of ellipse methods given by

$$x_{n+1} = x_n \pm \frac{f(x_n)}{\sqrt{f'^2(x_n)+p^2 f^2(x_n)}}, \quad (7)$$

where $p \neq 0 \in \Re$ and in which $f'(x) = 0$ is permitted at some points in the vicinity of the root. The beauty of this method is that it converges quadratically and moreover, has the same error equation as Newton's method. Therefore, this method is an efficient alternative to Newton's method.

In this paper, we present two families of predictor-corrector iterative methods based on quadratically convergent ellipse method (7), Nedihzbov *et al.* family (2) and the well-known Traub-Ostrowski's Formula (3).

2. Development of Methods

2.1. Two-Step Iterative Method and its Order of Convergence

Our aim is to develop a scheme that retains the order of convergence of Nedzhibov *et al.* family (3) and which can be used as an alternative to existing techniques or in cases where existing techniques are not successful. Thus we begin with the following predictor-corrector iterative scheme

$$\left.\begin{aligned} z_n &= x_n \pm \frac{f(x_n)}{\sqrt{f'^2(x_n)+p^2 f^2(x_n)}}, \\ x_{n+1} &= z_n \pm \frac{f(z_n)}{\sqrt{f'^2(x_n)+p^2 f^2(x_n)}}\frac{f(x_n)}{f(x_n)-2\lambda f(z_n)}, \end{aligned}\right\}, \quad (8)$$

where the positive sign is taken if $x_0 < r$ and the negative sign is taken if $x_0 \geq r$. Geometrically, if slope of the curve $\{f'(x_0)\}$ at the point $\{x_0, f(x_0)\}$ is negative, then take positive sign otherwise, negative. It is interesting to note that by ignoring the term in p, proposed family (8) reduces to Nedzhibov *et al.* family (2).

For $\lambda = 1$ in (8), we get

$$x_{n+1} = x_n \pm \frac{f(x_n)}{\sqrt{f'^2(x_n) + p^2 f^2(x_n)}} \left\{ \frac{f(x_n) - f(z_n)}{f(x_n) - 2f(z_n)} \right\}, \quad (9)$$

This is the modification over the Formula (3) of Traub-Ostrowski [2,5,7], and is also an order four formula. This method requires same evaluation of the function and its derivative as Traub-Ostrowski's method per iteration. Thus the efficiency index [2] of this method is equal to $\sqrt[3]{4} \cong 1.587$ which is better than the one of Newton's method $\sqrt[2]{2} \cong 1.414$ or SM method $\sqrt[4]{3} \cong 1.316$. More importantly, this method will not fail even if the derivative of the function is small or even zero in the vicinity of the root.

The asymptotic order of this method is presented in the following theorem.

Theorem 1. Suppose $f(x)$ is sufficiently differentiable function in the neighborhood of a simple root *r* and that x_0 is close to *r*, then the iteration scheme (8) has 3rd and 4th order convergence for

1) $\lambda \neq 1$,
2) $\lambda = 1$, respectively.

Proof: Since $f(x)$ is sufficiently differentiable, expanding $f(x_n)$ and $f'(x_n)$ about $x = r$ by Taylor's expansion, we have

$$f(x_n) = f'(r)\left[e_n + c_2 e_n^{\,2} + c_3 e_n^{\,3} + c_4 e_n^{\,4} + c_5 e_n^{\,5} + O\left(e_n^{\,6}\right) \right] \quad (10)$$

and

$$f'(x_n) = f'(r)\left[1 + 2c_2 e_n + 3c_3 e_n^{\,2} + 4c_4 e_n^{\,3} + 5c_5 e_n^{\,4} + O\left(e_n^{\,5}\right) \right] \quad (11)$$

where $e_n = x_n - r$ and $c_k = \frac{1}{k!}\frac{f^k(r)}{f'(r)}, k = 2,3,...$

Using Equations (10) and (11), we have

$$\frac{f(x_n)}{f'(x_n)} = e_n - c_2 e_n^{\,2} - 2\left(c_3 - c_2^2\right) e_n^{\,3} - \left(3c_4 - 7c_2 c_3 + 4c_2^3\right) e_n^{\,4} + O\left(e_n^{\,5}\right) \quad (12)$$

Therefore,

$$\frac{f(x_n)}{\sqrt{f'^2(x_n) + p^2 f^2(x_n)}} = \frac{f(x_n)}{f'(x_n)\sqrt{1 + p^2\left\{\frac{f(x_n)}{f'(x_n)}\right\}^2}} = e_n - c_2 e_n^{\,2} - \frac{1}{2}\left(-4c_2^{\,2} + 4c_3 + p^2\right) e_n^{\,3} - \frac{1}{2}\left(8c_2^{\,3} - 14c_2 c_3 + 6c_4 - 3c_2 p^2\right) e_n^{\,4} + O\left(e_n^{\,5}\right). \quad (13)$$

$$f(z_n) = f'(r)\left\{ c_2 e_n^{\,2} + \frac{1}{2}\left(-4c_2^{\,2} + 4c_3 + p^2\right) e_n^{\,3} + \frac{1}{2}\left(10c_2^{\,3} - 14c_2 c_3 + 6c_4 - 3c_2 p^2\right) e_n^{\,4} + O\left(e_n^{\,5}\right) \right\}. \quad (14)$$

$$f(x_n) - 2\lambda f(z_n) = f'(r)\left\{ e_n + (1 - 2\lambda) c_2 e_n^{\,2} + \left[c_3 - \lambda\left(-4c_2^{\,2} + 4c_3 + p^2\right) \right] e_n^{\,3} + O\left(e_n^{\,4}\right) \right\}. \quad (15)$$

Using Equations (12), (13) and (15), we obtain

$$\frac{f(z_n)}{f(x_n) - 2\lambda f(z_n)} = c_2 e_n + \frac{1}{2}\left(-6c_2^{\,2} + 4c_3 - 4c_2\lambda + p^2\right) e_n^{\,2} + \left[8c_2^{\,3} + 3c_4 - 2c_2\left(5c_3 + p^2\right) + \lambda\left(4c_2^{\,2} - 4c_3 - p^2\right) \right] e_n + O\left(e_n^{\,4}\right). \quad (16)$$

Using Equations (13)-(16) in Equation (8), we obtain,

$$e_{n+1} = \left[2(1 - \lambda) c_2^{\,2} e_n^{\,3} + c_2 \left\{ c_3(7 - 8\lambda) + \frac{p^2}{2}(3 - 4\lambda) + c_2^{\,2}\left(-9 + 14\lambda - 4\lambda^2\right) \right\} e_n^{\,4} + O\left(e_n^{\,5}\right) \right]. \quad (17)$$

While for $\lambda = 1$ in (18), we have

$$e_{n+1} = \left[{c_2}^3 - \frac{1}{2} c_2 \left(2c_3 + p^2 \right) \right] {e_n}^4 + O\left({e_n}^5 \right). \qquad (18)$$

Thus Equation (18) establishes the maximum order of convergence equal to four, for iteration scheme (8). This completes the proof of the theorem.

2.2. Three-Step Iterative Method and its Order of Convergence

On similar lines, we also propose a modification over the Formula (4) of Sharma and Guha [8]. Mir and Zaman [13] have considered three-step quadrature based iterative methods with sixth, seventh and eight order of convergence for finding simple zeros of nonlinear equations. Milovanović and Cvetković [14] further presented modifications over three-step iterative methods considered by Mir and Zaman [13]. Also Rafiq *et al.* [15] have presented similar three-step iterative method based on Newton's method with sixth-order convergence. All these modifications are targeted at increasing the local order of convergence with a view of increasing their efficiency index. But all these methods are variants of Newton's method and will not work if $f'(x)$ is very small or zero in the vicinity of the root. To overcome this problem, now we begin with the following predictor-corrector iterative scheme

$$\left.\begin{aligned} z_n &= x_n \pm \frac{f(x_n)}{\sqrt{f'^2(x_n) + p^2 f^2(x_n)}}, \\ y_n &= z_n \pm \frac{f(z_n)}{\sqrt{f'^2(x_n) + p^2 f^2(x_n)}} \frac{f(x_n)}{f(x_n) - 2\, f(z_n)}, \\ x_{n+1} &= y_n \pm \frac{f(y_n)}{\sqrt{f'^2(x_n) + p^2 f^2(x_n)}} \frac{f(x_n) + a f(z_n)}{f(x_n) + b\, f(z_n)}, \end{aligned}\right\} \qquad (19)$$

where *a* and *b* are parameters to be determined from the following convergence theorem.

Theorem 2. Let $f : I \to R$ denote a real valued function defined on *I*, where *I* is a neighborhood of simple root *r* of $f(x)$. Assume that $f(x)$ is sufficiently differentiable function in *I*. Then the iteration scheme (19) defines a one-parameter (*i.e.*, *a*) family of sixth order convergence if $b = a - 2$ and satisfies the following error equation:

$$e_{n+1} = \frac{1}{4} c_2 \left(2{c_2}^2 - 2c_3 - p^2 \right) \left\{ 4(1+a){c_2}^2 - 2c_3 - p^2 \right\} {e_n}^6 + O\left({e_n}^7 \right). \qquad (20)$$

Proof: follows on the similar steps as given in the previous theorem.

The proposed scheme (19) is now given by

$$\left.\begin{aligned} z_n &= x_n \pm \frac{f(x_n)}{\sqrt{f'^2(x_n) + p^2 f^2(x_n)}}, \\ y_n &= z_n \pm \frac{f(z_n)}{\sqrt{f'^2(x_n) + p^2 f^2(x_n)}} \frac{f(x_n)}{f(x_n) - 2\, f(z_n)}, \\ x_{n+1} &= y_n \pm \frac{f(y_n)}{\sqrt{f'^2(x_n) + p^2 f^2(x_n)}} \frac{f(x_n) + a f(z_n)}{f(x_n) + (a-2) f(z_n)}, \end{aligned}\right\} \qquad (21)$$

where $a \in \Re$. Note that for $p = 0$, we obtain method (4) obtained by Sharma and Guha [8] and for $(p, a) = (0, 0)$, we obtain method (5) developed by Grau and Díaz-Barrero [7].

3. Numerical Results

In this section, we shall present the numerical results obtained by employing the iterative methods namely Newton's method (NM), Traub-Ostrowski's method (3) (TOM), Simpson-Mamta method (6) (SM), modified Traub-Ostrowski's method (9) (MTOM), method (4) for $a = 1$ (M_3) and modified method (21) for $a = 1$ (MM_3) respectively to solve nonlinear equations given in **Table 1**. The results are summarized in **Table 2**. We use $\in = 10^{-15}$ as tolerance. Computations have been performed using C^{++} in double precision arithmetic. Here all the formulas are tested for $p = 1/2$. The following stopping criteria are used for computer programs:

1) $\left| x_{n+1} - x_n \right| < \in$,
2) $\left| f(x_{n+1}) \right| < \in$.

The behaviors of existing multipoint iterative schemes and proposed modifications can be compared by their corresponding correction factors. The correction factor $\frac{f(x_n)}{f'(x_n)}$, which appears in the existing multipoint iterative schemes is now modified by $\pm \frac{f(x_n)}{\sqrt{f'^2(x_n) + p^2 f^2(x_n)}}$, where $p \neq 0 \in \Re$. This is always well defined, even if $f'(x_n) = 0$. It is investigated that formulas (8) and (21) give very good approximation to the root when p is taken in between $0 < p < 1$. This is because that for small value of p, the ellipse will shrink in the vertical direction and extend along horizontal direction. This means that our next approximation will move faster towards the desired root. However, for $p > 1$ but not very large, the formulas work if the initial guess is very close

Table 1. Test problems.

No	Examples	[*a*,*b*]	Initial guess x_0	Root (*r*)
1	$(x-1)^6-1=0$	[1,3]	1.1 3.0	2.000000000000000
2	$x^3+4x^2-10=0$	[0,2]	0.0 0.1 2.0	1.3652300134140969
3	$\cos x - x = 0$	[0,2]	0.0 2.0	0.7390851332151600
4	$\tan^{-1} x = 0$	[-2,2]	−2.0 −1.0 2.0	0.000000000000000
5	$x^3+4x^2+\cos(x-1)-6=0$	[0.5,3]	1.8 3.0	1.000000000000000
6	$e^{x^2+7x-30}-1=0$	[2,4]	2.0 2.5 2.8 3.5	3.000000000000000
7	$e^{(1-x)}-1=0$	[-1,3]	−1.0 3.0	1.000000000000000
8	$\sin x = 0$	[0,2]	1.5	0.000000000000000

Table 2. Comparison table.

Examples	***Number of iterations***						***Number of functions evaluations***					
	NM	*TOM*	*SM*	*MTOM*	M_3	MM_3	*NM*	*TOM*	*SM*	*MTOM*	M_3	MM_3
		(3)	(6)	(9)	(4)	(21)		(3)	(6)	(9)	(4)	(21)
1	58	25	6	3	D	5	116	75	24	9	-	20
	8	4	5	4	3	3	16	12	20	12	12	12
	F	F	4	3	F	3	-	-	16	9	-	12
2	9	4	4	2	8	2	18	12	16	6	32	8
	4	2	3	2	2	2	8	6	12	6	8	8
3	3	2	3	2	2	2	6	6	12	6	8	8
	3	2	3	2	2	2	6	6	12	6	8	8
	D	5	5	3	8	3	-	15	20	9	32	12
4	5	3	3	3	3	2	10	9	12	9	12	8
	D	5	5	3	8	3	-	15	20	9	32	12
5	5	2	3	2	2	2	10	6	12	6	8	8
	6	3	4	2	3	2	12	9	16	6	12	8
	D	D	D	2	D	2	-	-	-	6	-	8
6	D	D	D	6	D	D	-	-	-	18	-	-
	15	5	21	4	D	D	30	15	84	12	-	-
	11	5	7	5	4	4	22	15	28	15	16	16
7	6	3	4	3	3	2	12	9	16	9	12	8
	9	4	5	3	10	3	18	12	20	9	40	12

to the required root. For larger value of p, the formulas do not work. This is perhaps due to the occurrence of numerical instability in the process of computation.

Example 8. $\sin x = 0$.

This equation has an infinite number of roots. Newton's method and Traub-Ostrowski's method with initial $x_0 = 1.5$ converge to -4π far away from the required root zero. Method (4) (M_3) converges to -6π. Our methods and SM method do not exhibit this type of behavior and converge to the nearest root zero.

4. Conclusions

The presented results indicate that the new proposed methods are more efficient and perform better than classical existing methods. The computational results in **Table 2** show that the modified Traub-Ostrowski's method (MTOM) (9) requires a smaller number of function evaluations than Newton's method (NM) and Traub-Ostrowski's method (3) (TOM). The computational results in **Table 2** also show that modified method (21) (MM_3) requires smaller number of function evaluations than method (4) (M_3). On similar lines, we can also modify Mir and Zaman [13], Milovanović and Cvetković [14] three-step iterative methods. Now a reasonably close starting value x_0 is not required for these methods to converge. This condition, however applies to practically all existing iterative methods for solving equations. Moreover, they have same efficiency indices as that of existing methods and do not fail if the derivative of the function is either zero or very small in the vicinity of the root. Therefore, these techniques have a definite practical utility.

5. Acknowledgements

We are grateful to the reviewer for the constructive remarks and suggestions which enhanced our work.

6. References

[1] A. M. Ostrowski, "Solution of Equations in Euclidean and Banach Space," 3rd Edition, Academic Press, New York, 1973.

[2] J. F. Traub, "Iterative Methods for the Solution of Equations," Prentice Hall, Englewood Cliffs, New Jersey, 1964.

[3] V. Kanwar and S. K. Tomar, "Modified Families of Newton, Halley and Chebyshev Methods," *Applied Mathematics and Computation*, Vol. 192, No. 1, September 2007, pp. 20-26.

[4] V. Kanwar and S. K. Tomar, "Exponentially Fitted Variants of Newton's Method with Quadratic and Cubic Convergence," *International Journal of Computer Mathematics*, Vol. 86, No. 9, September 2009, pp. 1603-1611.

[5] G. H. Nedzhibov, V. I. Hasanov and M. G. Petkov, "On Some Families of Multi-Point Iterative Methods for Solving Nonlinear Equations," *Numerical Algorithms*, Vol. 42, No. 2, June 2006, pp. 127-136.

[6] W. Werner, "Some Improvement of Classical Methods for the Solution of in Nonlinear Equations in Numerical Solution of Nonlinear Equations," *Lecture Notes Mathematics*, Vol. 878, 1981, pp. 426-440.

[7] M. Grau and J. L. Díaz-Barrero, "An Improvement to Ostrowski Root-Finding Method," A*pplied Mathematics and Computation*, Vol. 173, No. 1, February 2006, pp. 369-375, 450-456.

[8] J. R. Sharma and R. K. Guha, "A Family of Modified Ostrowski Methods with July Accelerated Sixth Order Convergence," *Applied Mathematics and Computation*, Vol. 190, No. 1, 2007, pp. 11-115.

[9] N. A. Mir, K. Ayub and A. Rafiq, "A Third-Order Convergent Iterative Method for Solving Non-Linear Equations," *International Journal of Computer Mathematics*, Vol. 87, No. 4, March 2010, pp. 849-854.

[10] Mamta, V. Kanwar, V. K. Kukreja and S. Singh, "On a Class of Quadratically Convergent Iteration Formulae," *Applied Mathematics and Computation*, Vol. 166, No. 3, July 2005, pp. 633-637.

[11] V. I. Hasnov, I. G. Ivanov, and G. Nedzhibov, "A New Modification of Newton's Method," *Application of Mathematics in Engineering*, Heron, Sofia, Vol. 27, 2002, pp. 278-286.

[12] K. C. Gupta, V. Kanwar and Sanjeev Kumar, "A Family of Ellipse Methods for Solving Nonlinear Equations," *International Journal of Mathematical Education in Science and Technology*, Vol. 40, No. 4, January 2009, pp. 571-575.

[13] N. A. Mir, K. Ayub and T. Zaman, "Some Quadrature Based Three-Step Iterative Methods for Nonlinear Equations," *Applied Mathematics and Computation*, Vol. 193, No. 2, November 2007, pp. 366-373.

[14] G. V. Milovanović and A. S. Cvetković, "A Note on Three-Step Iterative Methods for Nonlinear Equations," Studia University "Babes-Bolyai", *Mathematica*, Vol. LII, No. 3, 2007, pp. 137-146.

[15] A. Rafiq, S. Hussain, F. Ahmad, M. Awais and F. Zafar, "An Efficient Three-Step Iterative Method with Sixth-Order Convergence for Solving Nonlinear Equations," *International Journal of Computer Mathematics*, Vol. 84, No. 3, March 2007, pp. 369-375.

Applied Mathematics, **2010, 1, 159-178**
doi:10.4236/am.2010.13021 Published Online September 2010 (http://www.SciRP.org/journal/am)

Solidification and Structuresation of Instability Zones

Evgeniy Alexseevich Lukashov[1], Evgeniy Vladimirovich Radkevich[2]
[1]*Soyuz Aircraft-Engine Scientific and Technical Complex* (*AMNTK SOUYUZ*), *Moscow*, *Russia*
[2]*Moscow State University*, *Moscow*, *Russia*
E-mail: *elukashov@yandex.ru*, *evrad*07*@gmail.ru*
Received May 1, 2010; *revised July* 15, 2010; *accepted July* 17, 2010

Abstract

Two mathematical crystallization models describing structure formations in instability zones are proposed and justified. The first model, based on a phase field system, describes crystallization processes in binary alloys. The second model, based on a modified Biot model of a porous medium and the convective Cahn–Hilliard model, governs oriented crystallization. Physical interpretation and numerical analysis are discussed.

Keywords: Solidification, Structuresation of Instability Zones, Phase Field Model

1. Introduction

Unlike the main properties of oriented crystallization, properties responsible for the alloy structure have not yet been studied well. At the same time, owing to recent experimental results, many details of crystallization become known. In this paper, we propose the so-called "reconstruction" of oriented crystallization processes, *i.e.*, a detailed theoretical description based on the known main properties.

To reconstruct a process of binary alloy crystallization, one should begin with the question why the process "can live" in the stochastic instability. Perhaps, like in the case of complicated systems [1], the crystallization process can exist for a long time only due to solid structure formations in instability zones. Moreover, taking into account such structure formations, we are able to explain the solid phase growth – the crystallization mechanism.

It is known [2] that the structure formation in an alloy obtained by the oriented crystallization method is characterized by the following properties.

1) The process proceeds in a solid–liquid domain – a dynamic porous medium–where the solid phase is represented by growing dendrites, whereas the liquid phase occupies the space between these dendrites. According to experimental results, the solid phase growth is of order $O\left(\sqrt{t}\right)$, where t is time.

2) In the case of overlapping dendrites (in particular, their secondary branches), the melt solidification can lead to the contraction of melt and formation of internal stresses and micropores.

3) In turn, a solid-liquid crystallization zone appears because of the instability of the crystallization front which can be caused by the following reasons:

- concentration overcooling,
- segregation of the melt components in view of the spinodal decomposition (*i.e.*, phase transition with instable states) when the melt deeply penetrates into the metastable (or even labile) domain under high-speed (high-gradient) cooling in the interphase zone.

4) Properties of a new alloy are encoded in a seed crystal (a small piece of the solid phase) which, like the genetic code, determines the required properties of the crystallized part.

The experimental results concerning the distribution of crystallization centers over the blank surface are represented in **Figure 1**, where it is seen that crystallization centers are concentrated on convex parts of the surface, but not on its concave parts. In both cases, one of the phases grows in time, whereas the other decreases. We also note that the picture demonstrates the structure ordering.

The goal of this paper is to construct mathematical models reflecting Properties 1–4 and simulating the structure formation in alloys and, first of all, in instability zones. We propose two models (cf. Sections 3 and 4) with banding structure in the zone of instability.

But, first, we emphasize that, within the frameworks of models where structure formations in the instability zone are not taken into account, it is impossible to obtain the experimental order $O\left(\sqrt{t}\right)$ of the solid phase growth (cf. Property 1). We illustrate this fact by considering the well-known statistical Kolmogorov model [3] describing

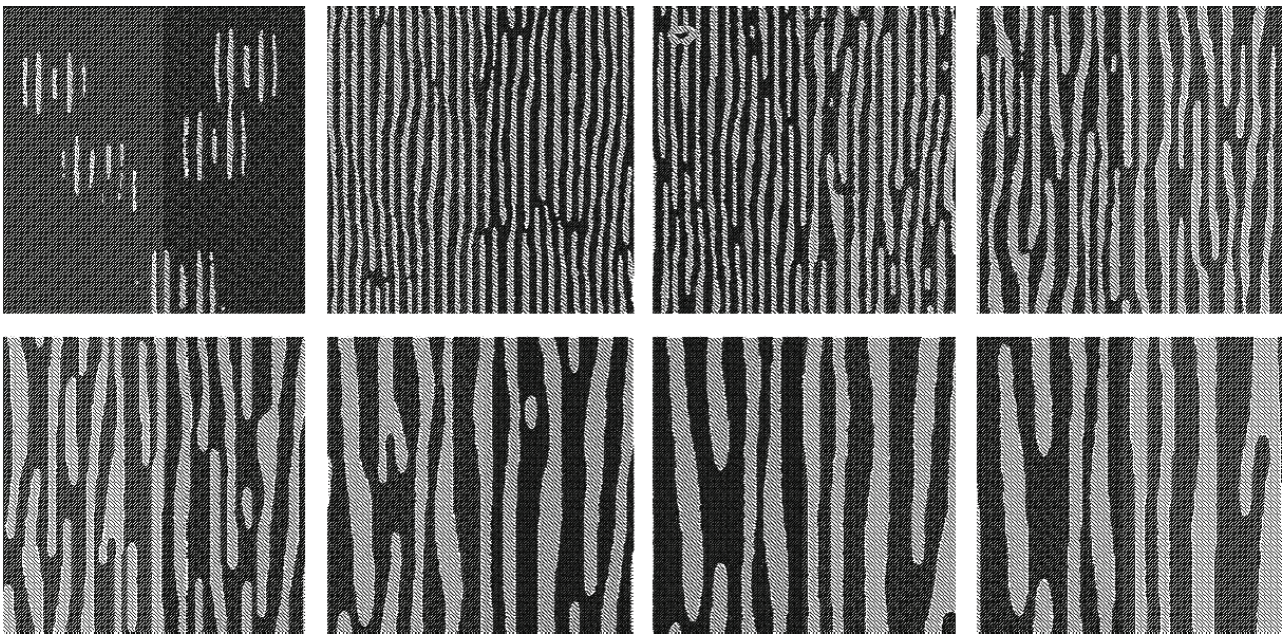


Figure 1. The experimental results are represented the distribution of crystallization centers over the blank surface.

the process of metal crystallization (cf. Section 2).

2. Kolmogorov's Model of Metal Crystallization

2.1. Physical Interpretation

In metallurgy, it is important to know the crystal growth velocity under a *random* formation of crystallization centers. Under rather general assumptions, Kolmogorov [3] derived an expression for the probability $p(t)$ that a randomly taken point P gets into the crystallized mass during the crystallization time-interval. With rather good approximation, we can assume that the mass crystallized in time t is equal to $p(t)$. Then it is possible to find the number of crystallization centers formed during the whole process of crystallizaton.

2.2. Mathematical Statement

Consider a domain $V \subset \mathbb{R}^d$, $d = 2,3$. Assume that at the initial time $t = 0$, the domain V is occupied by the so-called mother phase. At time t, some part $V_1(t)$ of V is occupied by a crystallized matter. Moreover, $V_1(t)$ enlarges in t as follows.

1) In a free part V/V_1 of V, new crystallization centers appear, so that for any domain $V' \subset V/V_1$ the probability of appearing a single crystallization center in V' during time Δt is equal to

$$\alpha(t)V'\Delta t + o(\Delta t),$$

whereas the probability of appearing more than one crystallization centers is of order $o(\Delta t)$, where $o(\Delta t)$ is infinitesimal in comparison with Δt. These probabilities are independent of the distribution of crystallization centers that are formed before time t (the process is Markovian) if only the *freedom* of V' from the crystallized mass at time t is not guaranteed.

2) Around the new-formed crystallization centers and around the entire crystallized mass, the mass grows with linear velocity

$$c(t,n) = k(t)c(n)$$

depending on time t and direction n, $|n| = 1$. It is assumed that the endpoints of vectors $c(n)n$ started at the origin and directed towards n form a *convex* surface.

Note that the *homogeneous* dependence of the linear velocity $c(t, n)$ on the direction n at all points is an essential restriction. In other words, we obtain formulas that are valid either

- in the case where the growth is uniform along all directions, or
- in the case of crystals of arbitrary shape but with the same spatial orientation.

We also mention the case where all crystallization centers are formed at initial times, in mean, β per volume unit. We obtain the corresponding formulas by taking into account that, in this case, $\alpha(t)$ is the Dirac function $\beta\delta(0)$ concentrated at the origin.

2.3. Formulas

We introduce the mean (over all directions) velocity of

the growth of crystallized mass c by the formula

$$c^d = \frac{1}{|S|}\int_S c^d(n)ds, \quad d = 2,3,$$

where the integral is taken over the surface of unit sphere S in $\mathbb{R}^d$ with center at the origin, $|S| = 4\pi$ if $d = 3$ and $|S| = 2\pi$ if $d = 2$. Then the following assertions hold.

1) For a sufficiently large (in comparison with the size of a crystallization center) domain V the domain $V_1(t)$ occupied by the crystallized phase takes the form

$$V_1(t) = V\left(1 - e^{-A_d c_d^d \Omega_d}\right) \tag{1}$$

If $\alpha(t)$ and $c(t, n)$ are time-independent, we can set $\alpha(t) = \alpha$, $k(t) = 1$. In this case,

$$\Omega_d = \frac{\alpha t^{d+1}}{d+1}, \tag{2}$$

which implies

$$V_1(t) = V\left(1 - e^{-\frac{A_d}{d+1}c_d^d \alpha t^{d+1}}\right) \tag{3}$$

2) If all crystallization centers are formed at initial times, then

$$\Omega_d = \int_0^t \alpha(t')\left(\int_{t'}^t k(\tau)d\tau\right)^d dt' = \beta\left(\int_0^t k(\tau)d\tau\right)^d \tag{4}$$

If, in addition, $k = 1$, *i.e.*, $c(t, n)$ is independent of t, then

$$\Omega_d = \beta t^d, \tag{5}$$

which implies

$$V_1(t) = V\left(1 - e^{-A_d c_d^d \beta t^d}\right) \tag{6}$$

We see that the mass growth is of power-like order $O(t^\alpha)$, $\alpha = 1, 2, 3$, $d = 1, 2, 3$.

2.4. Conclusions

The Kolmogorov model is not suitable for describing crystallization of twocomponent mixtures. Indeed, within the frameworks of the Kolmogorov model, the fact that the mass growth is of power order implies that the velocity is finite at $t = 0$, which contradicts the initial stage of the spinodal decomposition generating an initial distribution of crystallization centers.

3. Model of Binary Alloy Crystallization

Based on the phase field system proposed in [4] and [5], we construct a model of binary alloy crystallization with structure formation in the zone of instability.

A crystallization model based on the phase field conception was constructed in [6], where, in particular, a sawtooth solution to the temperature distribution problem in the phase transition domain was obtained. This result agrees with the qualitative description of autocrystallization phenomena in [7,8].

The goal of this section is to obtain a sawtooth solution to the temperature distribution problem for the following phase field system:

$$\frac{\partial\varphi}{\partial t} + \frac{\partial\theta}{\partial t} = \Delta\theta, \ (x,t) \in Q, \tag{7}$$

$$\varepsilon^2 \frac{\partial\varphi}{\partial t} = \varepsilon^2\Delta\varphi + \varphi - \varphi^3 + \varepsilon\varkappa\theta, \tag{8}$$

$$\varphi\big|_{t=0} = \varphi^0(x,\varepsilon), \ \theta\big|_{t=0} = \theta^0(x,\varepsilon), \tag{9}$$

$$\varphi|_\Sigma = 1, \quad \theta|_\Sigma = \theta_b, \tag{10}$$

where θ is the temperature; φ is the specific concentration of the order function, equal to 1 in the liquid phase and to -1 in the solid phase; $\varkappa = const$; $Q = (0, T) \times \Omega$, where $\Omega \subset \mathbb{R}^n$ is a bounded domain with C^∞ - boundary, $n \leqslant 3$; $\sum = [0,T] \times \partial\Omega$; the functions φ^0 and θ^0 are sufficiently smooth for $\varepsilon \geqslant const > 0$, and the function θ_b is also sufficiently smooth.

The system (7)–(10) describes slow crystallization processes [9] with an instable domain of intermediate aggregate state, where a structure formation appears.

3.1. Wave Train Type Solutions and Singular Limit Problem

Here, we consider a more general case where $\overline{\varphi^0} \in BV$, but $\overline{\varphi^0} \notin BVC$ [1] (cf. [6]). In the case of diffusion, we say that Ω is a *domain of intermediate aggregate state* (an IAS-domain) if $\varphi^0(x,\varepsilon) \rightharpoonup 0$ weakly as $\varepsilon \to 0$ in some subdomain $\Omega_{cr}^0 \subset \Omega$ of nonzero measure.

In accordance with [6], an IAS-domain is formed by a large number M of domains of pure (solid and liquid) phases of small volume of order υ_ε (*i.e.*, $M = M(\varepsilon) \to \infty$ and $\upsilon_\varepsilon \to 0$ as $\varepsilon \to 0$). The macroscopic description of an IAS-domain can be obtained by computing the weak limit of wave train type solutions as $\varepsilon \to 0$.

We formulate conditions imposed on IAS-domains.

1) The weak limit of the order functions $\varphi(x,t,\varepsilon)$ as $\varepsilon \to 0$ vanishes identically in the transition zone Ω_t^*.

2) In the domain $\Omega_{t,\varepsilon}^*$ corresponding to the regularization of the IAS-domain, the solution to the phase field system can be described in terms of the wave train

[1] A function $\overline{\varphi^0}$ belongs to the class *BVC* if $\overline{\varphi^0}$ is a function of bounded variation ($\overline{\varphi^0} \in BV$) and $\left|\overline{\varphi^0}\right| = 1$.

structure. In this case, the domain Ω is divided into a large number of domains of "small" volume occupied by "pure phases" and transition zones between them.

Remark 3.1 Condition (1) means, in particular, that for almost all t the limit order function $\overline{\varphi}$ belongs to $BV(\Omega)$, but $\overline{\varphi} \notin BVC(\Omega)$. Condition (2) is based on the conception proposed in [6]. According to this conception, the wave train structure is described by a chain of modified Stefan problems in domains of "small" volume occupied by "pure phases" and can be used for approximating the temperature in an IAS-domain. Such a structure is called the diffusion of the IAS-domain.

Remark 3.2 A situation where the limit order function $\overline{\varphi}$ vanishes on a set of nonzero measure is not good since this case corresponds to instable solutions to the isothermal diffusion equation. It is clear that such solutions can exist only under rather special conditions. Therefore, we need to impose rather restrictive conditions on the geometry of domains Ω, Ω_t^*, as well as on the initial and boundary conditions.

Remark 3.3 From the point of view of the theory of distributions, free boundary problems are problems about singularity propagation. Indeed, in the rigid-front situation, the limit order function is a Heaviside type function ($\overline{\varphi} = 1$ on Ω_t^+ and $\overline{\varphi} = -1$ on Ω_t^-) and the limit temperature remains continuous, but with weak discontinuity on the free boundary $\Gamma_t = \overline{\Omega_t^+} \cap \overline{\Omega_t^-}$.

To formulate the singular limit problem, we suppose that Γ_0 is a smooth surface of codimension 1, $\Gamma_0 \cap \partial\Omega = \varnothing$, dividing Ω into two parts $\Omega_0^{\pm}$ so that

$$\Omega = \Omega_0^+ \cup \Gamma_0 \cup \Omega_0^-$$

Let φ^0 and θ^0 be the initial data such that

$$\varphi^0 = \pm 1 + \mathcal{O}(\varepsilon)$$

outside an ε-neighborhood of the surface Γ_0 and $\theta^0 \in C(\Omega)$ (cf. details in [6,9]). The singular limit problem is written as

$$\frac{\partial \theta^{\pm}}{\partial t} = \Delta \theta^{\pm}, \quad x \in \Omega_t^{\pm}, \quad t > 0, \tag{11}$$

$$\theta^{\pm}|_{t=0} = \theta_{\pm}^0(x), \quad x \in \Omega_0^{\pm}, \quad \theta^+|_{\Sigma} = \theta_b, \tag{12}$$

$$\left[\theta^{\pm}\right]|_{\Gamma_t} = 0, \quad \left[\frac{\partial \theta^{\pm}}{\partial \nu}\right]|_{\Gamma_t} = -2V_{\nu}, \tag{13}$$

$$\varkappa_1 \theta^{\pm}|_{\Gamma_t} = \mathcal{K}_t - V_{\nu}. \tag{14}$$

This problem is the well-known modified Stefan problem with the Gibbs-Thomson condition (14) on the free boundary. Here,

$$\theta_{\pm}^0(x) = \overline{\theta^0}(x), \quad x \in \Omega_0^{\pm},$$

where $[f]|_{\Gamma_t}$ denotes the jump in f across the free boundary Γ_t; ν is the outward (relative to Ω_t^-) normal to Γ_t, V_{ν} is the normal velocity of the front Γ_t, $\mathcal{K}_t = -div(\nu)|_{\Gamma_t}$ is the mean curvature of the surface Γ_t, and $\varkappa_1 = 3\varkappa/\sqrt{2}$.

We assume that

$$\Gamma_t \cap \partial\Omega = \varnothing \quad \forall\, t \geqslant 0,$$

i.e., the front does not intersect the fixed boundary $\partial\Omega$.

Remark 3.4 The boundary conditions (13) and (14) can be interpreted as the Hugoniot type conditions corresponding to the problem of propagation of strong discontinuities of the limit order function $\overline{\varphi}$ and the problem of propagation of weak discontinuities of the limit temperature $\overline{\theta}$. This interpretation can be justified as follows. As is known, the necessary conditions for the existence of a shock wave type solution to a quasilinear hyperbolic equation generate an instable chain of Hugoniot type conditions. The same instability conditions (cf. [6]) are obtained for the boundary conditions on the free boundary if we use the classical definition (in $\mathcal{D}'$) of a weak solution to the phase field system. The boundary conditions in the interpretation of an IAS-domain as the limit of wave train type solutions are referred to as *Hugoniot type conditions.*

Let us describe the geometric structure. Assume that, at $t = 0$, the domain Ω contains domains of pure (liquid or solid) phase $\Omega_{0,\varepsilon}^{\pm}$ and also the melt domain $\Omega_{0,\varepsilon}^*$ occupied by a large number of pure phase domains of small volume $\Omega_{0,\varepsilon}^i$, $i = 1,2,\ldots,M$, where M is even. For the sake of simplicity, we consider the case of quasispherical symmetry. Let $\Gamma_{0,\varepsilon}^i$, $i = 1,\ldots,M-1$, be interfaces of domains $\Omega_{0,\varepsilon}^i$ so that

$$\partial\Omega_{0,\varepsilon}^i = \Gamma_{0,\varepsilon}^{i-1} \cup \Gamma_{0,\varepsilon}^i,$$

$$\Gamma_{0,\varepsilon}^0 = \partial\Omega_{0,\varepsilon}^-, \quad \partial\Omega_{0,\varepsilon}^+ = \Gamma_{0,\varepsilon}^M \cup \partial\Omega.$$

We denote by $D_{0,\varepsilon}^i$ the domains bounded by $\Gamma_{0,\varepsilon}^i$ and assume that

$$D_{0,\varepsilon}^i \subset D_{0,\varepsilon}^{i+1}, \quad i = 0,\ldots,M,$$

where

$$D_{0,\varepsilon}^0 = \Omega_{0,\varepsilon}^-, \quad D_{0,\varepsilon}^{M+1} \equiv \Omega.$$

Assume that $\Gamma_{0,\varepsilon}^i$ are smooth surfaces of codimension 1 such that

$$\begin{aligned} &c_1\varepsilon^{\alpha} \leq dist\left(\Gamma_{0,\varepsilon}^{k-1}, \Gamma_{0,\varepsilon}^k\right) \leq c_2\varepsilon^{\alpha}, \\ &c_1^{\pm} \leq \left|\Omega_{0,\varepsilon}^{\pm}\right| \leq c_2^{\pm}, \; dist\left(\Gamma_{0,\varepsilon}^M, \partial\Omega\right) \geq c_3, \end{aligned} \tag{15}$$

where $k = 1,\ldots,M$, $\alpha \in (0,1)$ and the constants $c_l^{\pm}, c_j > 0$ are independent of ε.

Assume that $\Gamma_{0,\varepsilon}^i$, $i = 0,\ldots,M$, satisfy the following geometric condition.

$\Gamma_{0,\varepsilon}^i \in C^3$ uniformly in $\varepsilon \in [0,\varepsilon_0]$, $M \to \infty$ and

$M\varepsilon^{\alpha} \to L = \text{const}$ as $\varepsilon \to 0$; moreover, the surfaces obtained after the limit passage occupy the mixture domain Ω_0^* bounded by C^3-surfaces Γ_0^- and Γ_0^+.

Remark 3.5 If Condition (A) holds, then there exists a function $s^0(x,\varepsilon) \in C^3(\overline{\Omega})$ such that any $\Gamma_{0,\varepsilon}^i$ is a level surface of this function.

Formula (15) shows that there are no interactions (up to $\mathcal{O}(\varepsilon^{\infty})$) between neighboring waves such that the distance between them is not less than $\mathcal{O}(\varepsilon^{1-\delta})$ with any constant $\delta > 0$. Thus, for sufficiently small t an asymptotic solution is expressed as the superposition of local solutions to the rigid-front problems (with one front $\Gamma_{0,\varepsilon}^i$) (cf. [6,8]) (as shown in Formula (16)).

As in the case of rigid-front solution, $\theta_{i,c}$ is a smooth extension of the auxiliary function $\theta_i = \theta_i(x,t,h)$,

$$\eta_i = \left(s^{(n_i)}(x,t,\varepsilon) - ih\right) \Big/ \left(\left|\nabla s_0^{(n_i)}\right| \varepsilon\right), \quad h = \varepsilon^{\alpha}, \quad \varepsilon \in [0, \varepsilon_0].$$

We recall that the family of functions $\{\theta_i\}$ and $s_0^{(j)}$, $j = 1,2$, is defined as a solution to the chain of modified Stefan problems with the Gibbs-Thomson condition

$$\frac{\partial \theta_i}{\partial t} = \Delta \theta_i, \quad x \in \Omega_{t,\varepsilon}^i, \quad t > 0, \tag{17}$$

$$\theta_{i-1}\,|_{\Gamma_{t,\varepsilon}^{i-1}-0} = \theta_i\,|_{\Gamma_{t,\varepsilon}^{i-1}+0}, \tag{18}$$

$$\theta_i\,|_{\Gamma_{t,\varepsilon}^{i}-0} = \theta_{i+1}\,|_{\Gamma_{t,\varepsilon}^{i}+0}, \tag{19}$$

$$\frac{\partial \theta_{i-1}}{\partial \nu_{i-1}}|_{\Gamma_{t,\varepsilon}^{i-1}-0} - \frac{\partial \theta_i}{\partial \nu_{i-1}}|_{\Gamma_{t,\varepsilon}^{i-1}+0} = (-1)^{i+1} 2V_{\nu_{i-1}}, \tag{20}$$

$$\frac{\partial \theta_i}{\partial \nu_i}|_{\Gamma_{t,\varepsilon}^{i}-0} - \frac{\partial \theta_{i+1}}{\partial \nu_i}|_{\Gamma_{t,\varepsilon}^{i}+0} = (-1)^{i} 2V_{\nu_i}, \tag{21}$$

$$(-1)^{i-1} \varkappa_1 \theta_i\,|_{\Gamma_{t,\varepsilon}^{i-1}-0} = \mathcal{K}_t^{i-1} - V_{\nu_{i-1}}, \tag{22}$$

$$(-1)^{i} \varkappa_1 \theta_i\,|_{\Gamma_{t,\varepsilon}^{i}+0} = \mathcal{K}_t^{i} - V_{\nu_i}, \tag{23}$$

with the initial and boundary (on $\partial\Omega$) conditions. Here, $i = 0,\ldots,M+1$. We set

$$\Gamma_{t,\varepsilon}^{-1} = \Gamma_{t,\varepsilon}^{M+1} = \varnothing,$$

so that the condition (18) (the condition (21)) vanishes for $i = 0$ ($i = M+1$). Furthermore,

$$V_{\nu_i} = -(|\nabla s_0^{(n_i)}|)^{-1} \frac{\partial s_0^{(n_i)}}{\partial t}|_{\Gamma_{t,\varepsilon}^i}, \quad \mathcal{K}_t^i = -div\,\nu_i\,|_{\Gamma_{t,\varepsilon}^i}$$

$\Omega_{t,\varepsilon}^0 = \Omega_{t,\varepsilon}^-$ denotes the domain bounded by $\Gamma_{t,\varepsilon}^0$ and $\Omega_{t,\varepsilon}^{M+1} = \Omega_{t,\varepsilon}^+$ denotes the domain bounded by $\Gamma_{t,\varepsilon}^M$ and $\partial\Omega$. The small corrections $c_1^{(j)}(x,t,h)$ are simultaneously corrections of order $\mathcal{O}(\varepsilon)$ for temperature which can be computed as solutions to the linearized chain of modified Stefan problems with the Gibbs-Thomson condition (cf. [6]).

For the sake of convenience, we impose the following condition (cf. [6]).

(A') There exist functions $s^{(1)}(x,t,\varepsilon)$ and $s^{(2)}(x,t,\varepsilon)$ that describe respectively the surfaces $\Gamma_{t,\varepsilon}^i$ with even and odd superscripts for $t \geqslant 0$. We denote by $\Omega_{t,\varepsilon}^i$ the domain bounded by the surfaces $\Gamma_{t,\varepsilon}^{i-1}$ and $\Gamma_{t,\varepsilon}^i$, $i = 1,\ldots,M$, and introduce the notation

$$\Omega_{t,\varepsilon}^* = \bigcup_{i=1}^{M} \Omega_{t,\varepsilon}^i.$$

Constructing formal asymptotic solutions, we find

$$s^{(j)}(x,t,\varepsilon) = s_0^{(j)}(x,t,h) + \varepsilon c_1^{(j)}(x,t,h),$$

$$h = \varepsilon^{\alpha}, \quad \varepsilon \in [0, \varepsilon_0], \quad j = 1,2,$$

so that $|\nabla_x s^{(j)}| > 0$ uniformly with respect to $x \in \Omega_{t,\varepsilon}^*$ for any $h \in [0, h_0 = \varepsilon_0^{\alpha}]$ and

$$\Gamma_{t,\varepsilon}^i = \left\{x, s_0^{(n_i)}(x,t,h) = ih\right\}, n_i = 1, i = 2k,$$

$$n_i = 2, i = 2k+1, 0 \leq i \leq M.$$

It is obvious that

$$s^{(1)}\,|_{t=0} = s^{(2)}\,|_{t=0} = s^0(x,\varepsilon)$$

and, with accuracy $\mathcal{O}(\varepsilon)$,

$$\nu_i = \nabla s_0^{(n_i)} / |\nabla s_0^{(n_i)}|\,|_{\Gamma_{t,\varepsilon}^i}$$

are outward normals to $D_{t,\varepsilon}^i$.

For fixed $\varepsilon > 0$ and sufficiently small $t > 0$ the classical solvability of the chain of modified Stefan problems with the Gibbs–Thomson condition is established in the same way as in [10]. At the same time, based only on the limit problem below, it is impossible to formulate the initial conditions for the temperature $\theta^0(x,\varepsilon)$ in such a way that these conditions have sense as $\varepsilon \to 0$ because the classical solvability of the modified Stefan problem with the Gibbs-Thomson condition assumes conjugate conditions on the initial surface $\Gamma_{0,\varepsilon}^i$ for any $M \to \infty$. However, we can overcome these

$$\begin{aligned} \varphi_1^{as}(x,t,\varepsilon) &= \sum_{i=0}^{M} (-1)^i \chi(\eta_i) + \varepsilon\{\frac{\varkappa}{2}\theta_0^{as}(x,t,h) + \sum_{i=0}^{M} \omega_i(\eta_i, x)\}, \\ \theta_0^{as}(x,t,\varepsilon) &= \frac{1}{2}(\theta_{i-1,c} + \theta_{i,c}) + \frac{1}{2}(\theta_{i-1,c} - \theta_{i,c})\chi(\eta_i), \quad x \in \Omega_{t,\varepsilon}^{i-1} \cup \Omega_{t,\varepsilon}^i. \end{aligned} \tag{16}$$

difficulties if find a model problem for a weak limit of temperature as $\varepsilon \to 0$. Thus, we choose the initial data

$$\varphi|_{t=0} = \varphi_1^{as}(x,0,\varepsilon) + \mathcal{O}(\varepsilon^2), \tag{24}$$

$$\theta|_{t=0} = \theta_0^{as}(x,0,\varepsilon) + \mathcal{O}(\varepsilon), \tag{25}$$

$$s^{(j)}|_{t=0} = s^0(x,\varepsilon), \tag{26}$$

where $\theta_0^{as}(x,0,\varepsilon)$, $\varphi_1^{as}(x,0,\varepsilon)$, and the smooth functions $s^0(x,\varepsilon)$ are such that conjugate conditions are satisfied for fixed $\varepsilon > 0$. We will be able to specify these conditions by obtaining the limit problem.

3.2. Limit Problem

The evolution of solutions can proceed in two different ways depending on the initial data:

$$\partial_t s^{(1)}|_{t=0}\, \partial_t s^{(2)}|_{t=0} < 0, \tag{27}$$

$$\partial_t s^{(1)}|_{t=0}\, \partial_t s^{(2)}|_{t=0} > 0, \tag{28}$$

where $s^{(j)}$, $j = 1,2$, are the functions in Condition (A').

In the case (27), the boundaries move in the opposite directions. Consequently, the wave train type structure exists only during a small time interval since the domain $\Omega_{t,\varepsilon}^{2k}$ or $\Omega_{t,\varepsilon}^{2k+1}$ vanishes for $t \sim \varepsilon^{\alpha}$. A similar situation for the classical Stefan problem was treated in [6]. In teh case of the phase field system, from (27) it follows that an "overheated" or "overcooled" domain appears in Ω_t^*.

To find conditions for the existence of wave train type solutions in some finite time interval independent of ε, we consider the case (28), where the boundaries move in the same direction. Assume that the following condition holds.

(B) There exists $T > 0$ such that for any $0 \le t \le T$ there exist functions $\theta_i(x,t,h)$, $i = 0,\dots,M+1$, such that the function $\tilde{\theta}(x,t,\varepsilon)$ (defined by $\tilde{\theta} = \theta_i$ for $x \in \overline{\Omega_{t,\varepsilon}^i}$) is continuous and is uniformly bounded for $\varepsilon \in [0,\varepsilon_0]$. Furthermore, $\theta_i \in C^1(Q_\varepsilon^i)$ uniformly for $\varepsilon \in [0,\varepsilon_0]$, where $Q_\varepsilon^i = \bigcup_{t\in[0,T]} \overline{\Omega_{t,\varepsilon}^i}$, and $\Gamma_{t,\varepsilon}^i \in C^3$.

We list some consequences of Condition (B). Since the functions θ_i are smooth, it is obvious that

$$\theta_i|_{\Gamma_{t,\varepsilon}^i - 0} - \theta_i|_{\Gamma_{t,\varepsilon}^{i-1} + 0} = \mathcal{O}(h).$$

Therefore, taking into account the Gibbs--Thomson law (22), (23), we find

$$\mathcal{K}_t^i - V_{\nu_i} + \mathcal{K}_t^{i-1} - V_{\nu_{i-1}} = \mathcal{O}(h).$$

Since the surfaces $\Gamma_{t,\varepsilon}^i$ are smooth and $V_{\nu_{i-1}} V_{\nu_i} > 0$, we have

$$s_0^{(j)}(x,t,h) = s_0(x,t) + h\tilde{s}_0^{(j)}(x,t,h), \quad j = 1,2 \tag{29}$$

where the functions s_0, $\tilde{s}_0^{(j)}$ and their third order derivatives are uniformly bounded for $h \in [0,h_0]$. As a consequence, we find

$$V_{\nu_i} = \mathcal{K}_t^i + \mathcal{O}(h). \tag{30}$$

By (30) and (22), (23), we have

$$(-1)^{i-1} \varkappa_1 \theta_i|_{\Gamma_{t,\varepsilon}^{i-1} - 0} = \mathcal{K}_t^{i-1} - V_{\nu_{i-1}},$$

$$(-1)^{i} \varkappa_1 \theta_i|_{\Gamma_{t,\varepsilon}^{i} + 0} = \mathcal{K}_t^{i} - V_{\nu_i},$$

which implies

$$\theta_i|_{\Gamma_{t,\varepsilon}^i} = \mathcal{O}(h).$$

From Condition (B) it follows that

$$\tilde{\theta}(x,t,\varepsilon) = h\tilde{\theta}^1, \tilde{\theta}^1 = \mathcal{O}(1), x \in \bar{\Omega}_{t,\varepsilon}^*, t \in [0,T], \tag{31}$$

where $\tilde{\theta}^1 = \tilde{\theta}_i^1$ for $x \in \overline{\Omega_{t,\varepsilon}^i}$.

By the Gibbs-Thomson law,

$$\frac{\mathcal{K}_t^i - \mathcal{K}_t^{i-1}}{h} - \frac{V_{\nu_i} - V_{\nu_{i-1}}}{h} = (-1)^i \varkappa_1 \left(\tilde{\theta}_i^1|_{\Gamma_{t,\varepsilon}^{i} - 0} + \tilde{\theta}_i^1|_{\Gamma_{t,\varepsilon}^{i-1} - 0} \right).$$

Since $\Gamma_{t,\varepsilon}^i \in C^3$ uniformly with respect to h, we find $\tilde{\theta}_i^1 \in C^1(Q_\varepsilon^i)$.

We need the following assertion.

Lemma 3.1 1) *Let ζ_i be partition points in the interval $[0,L]$, $\zeta_0 < \zeta_1 < \dots < \zeta_M$, and let $h = \max_i(\zeta_i - \zeta_{i-1})$. Suppose that M is even, $F(\zeta) \in C([0,L])$, and $F(\zeta) \in C^1([\zeta_{i-1},\zeta_i])$ for any $i = 1,\dots,M$. Then*

$$\left| \sum_{i=0}^{M} (-1)^i F(\zeta_i) \right| \le \text{const} \quad \textit{uniformly for } M \ge 2.$$

2) Assume that $F(\zeta) \in C([0,L])$ and $F(\zeta) \in C^2([\zeta_{i-1},\zeta_i])$ for any $i = 1,\dots,M$. Then

$$\sum_{i=0}^{M} (-1)^i F(\zeta_i) = \frac{1}{2}\left(F(\zeta_0) + F(\zeta_M) \right) + \mathcal{O}(h)$$

uniformly for even $M \ge 2$.

To prove the lemma, it suffices to group the terms in $F(\zeta_i) - F(\zeta_{i-1})$ in such a way that to represent them as differences of derivatives.

Note that for passing to the limit in the wave train as $\varepsilon \to 0$, we need a suitable well-defined notion of a weak solution. We give such a definition in accordance with [6].

Definition 3.1 *A pair of functions*

$$\theta \in L^2\left(0,T;W_2^1(\Omega)\right) \cap L^\infty\left(0,T;L^2(\Omega)\right),$$

$$\varphi \in W_2^{2,1}(Q) \cap L^\infty\left(0,T;W_2^1(\Omega) \cap L^4(\Omega)\right)$$

is called a *weak solution* to the problem (49) if for any test functions $\xi(x,t)$, $g(x,t)=(g_1(x,t),\dots,g_n(x,t))$ satisfying (38) the functions θ and φ satisfy the equation

$$I_\theta = \int_Q \left(\langle \nabla\theta, \nabla\xi\rangle - (\theta+\varphi)\xi_t\right)dxdt + \int_\Omega \left(\theta^0+\varphi^0\right)\xi(x,0))dx = 0 \tag{32}$$

and the integral identity (33)
where

$$e_\varepsilon(\varphi) = \frac{1}{2}\varepsilon|\nabla\varphi|^2 + \frac{1}{\varepsilon}W(\varphi), W(\varphi) = \left(\varphi^2-1\right)^2/4,$$

and g_x is the matrix with entries $(g_x)_{ik} = \partial g_i/\partial x_k$.

We set

$$\mathcal{T}_\varepsilon^i = \bigcup_{t\in[0,T]} \Gamma_{t,\varepsilon}^i, \quad i=0,\dots,M,$$

and $\mathcal{T}_\varepsilon^{M+1} \equiv \mathcal{T}^{M+1} = \partial\Omega\times[0,T]$.

Then we substitute (34) in the integral identity (33). We need the following assertion.

Lemma 3.2 *Suppose that* $\omega(\eta,x)\in\mathbf{S}$, $\psi(x)\in C^2(\overline{\Omega})$, $|\nabla\psi|\neq 0$, *and*

$$dist\left(\Gamma_t,\partial\Omega\right)\geq const>0.$$

Then for any function $g\in C^1\left(\overline{Q}\right)$

$$\lim_{\varepsilon\to 0}\varepsilon^{-1}\int_Q \omega\left(\frac{s}{\varepsilon},x\right)g(x,t)\,dxdt = \int_{\Omega_T} A_\omega(x)\beta^{-1}(x)g(x,-\psi)\,dx,$$

where $s=(t+\psi)\beta+\varepsilon s_1$, $\beta=|\nabla\psi|^{-1}$,

$$A_\omega = \int_{-\infty}^{\infty}\omega(\eta,x)\,d\eta,$$

and Ω_T is the domain bounded by Γ_0 and Γ_T.

By Lemma 3.2,

$$J_\varphi = \sum_{i=0}^M(\langle g,\nabla s_0\rangle(\mathcal{K}_t^i - V_{\nu_i})A_{\dot\chi^2},\delta(\mathcal{T}_\varepsilon^i)) - \sum_{i=0}^M(-1)^i(\langle g,\nabla s_0\rangle\theta A_{\dot\chi},\delta(\mathcal{T}_\varepsilon^i)) + \mathcal{O}(\varepsilon h^{-1}+h) = 0.$$

Applying assertion (a) of Lemma 3.1 to the second sum and using (31) together with Condition (B), we find

$$J_\varphi = \sum_{i=0}^M(\langle g,\nabla s_0\rangle(\mathcal{K}_t^i - V_{\nu_i})A_{\dot\chi^2},\delta(\mathcal{T}_\varepsilon^i)) + \mathcal{O}(\varepsilon h^{-1}+h) = 0. \tag{35}$$

We again obtain the relation (30) since the first sum in (35) has order $\mathcal{O}(h^{-1})$. Taking into account (29) and passing to the limit as $\varepsilon\to 0$, we see that (35) implies (30) in the entire domain

$$\Omega_t^* = \lim_{\varepsilon\to 0}\Omega_{t,\varepsilon}^*.$$

Consequently,

$$|\nabla s_0|^{-1}\frac{\partial s_0}{\partial t} = div\left(\frac{\nabla s_0}{|\nabla s_0|}\right), \quad x\in\Omega_t^*, \quad t>0. \tag{36}$$

We consider the integral identity (32). We first compute the weak limit of wave train in the derivative $-\varphi_t$ in the heat equation.

Lemma 3.3 *Let*

$$\varphi(x,t,\varepsilon) = \varphi_1^{as}(x,t,\varepsilon) + \mathcal{O}(\varepsilon^2),$$

where φ_1^{as} *is defined by Formula* (34), *and*

$$c_1 h \leqslant dist(\mathcal{T}_\varepsilon^i,\mathcal{T}_\varepsilon^{i+1}) \leqslant c_2 h, \quad i=0,\dots,M-1,$$

where the constants c_1 *and* c_2 *are independent of* ε. *Then*

$$(\frac{\partial\varphi}{\partial t},\xi) = 2(\{\sum_{j=0}^M(-1)^{j+1}V_{\nu_j}\delta(\mathcal{T}_\varepsilon^j)\},\xi) + C_1 + \mathcal{O}(\varepsilon h^{-1}+h) \tag{37}$$

for any functions $\xi\in C^1\left(\overline{Q}\right)$ *such that*

$$\xi, g\in C^1\left(\overline{Q}\right), \quad \xi|_\Sigma = g|_\Sigma = 0, \quad \xi|_{t=T} = g|_{t=T} = 0 \tag{38}$$

Here,

$$C_1 = \mathcal{O}\left(\left(\tilde s_0^{(1)} - \tilde s_0^{(2)}\right)\big|_{h=0}\right)$$

is a possible contribution of the terms depending on the first corrections to the phase s_0 relative to h.

We set

$$F(\zeta_k) = \int_{\mathcal{T}_\varepsilon^k} V_{\nu_k}\,d\sigma_k.$$

Applying assertion (b) of Lemma 3.1 to (37), we find

$$J_\varphi = \varepsilon\int_Q \varphi_t\langle g,\nabla\varphi\rangle dxdt - \int_Q e_\varepsilon(\varphi)divg\,dxdt + \int_Q\left(\varepsilon\langle\nabla\varphi, g_x\nabla\varphi\rangle + \varkappa\varphi div(g\theta)\right)dxdt = 0, \tag{33}$$

$$\varphi_1^{as}(x,t,\varepsilon) = \sum_{i=0}^M(-1)^i\chi(\eta_i) + \varepsilon\left\{\frac{\varkappa}{2}\theta_0^{as}(x,t,h) + \sum_{i=0}^M\omega_i(\eta_i,x)\right\}. \tag{34}$$

$$(\frac{\partial\varphi}{\partial t},\xi) = -\int\limits_{\mathcal{T}_\varepsilon^0}\xi V_{\nu_0}\,d\sigma_0 - \int\limits_{\mathcal{T}_\varepsilon^M}\xi V_{\nu_M}\,d\sigma_M + C_1 + \mathcal{O}(\varepsilon h^{-1} + h). \tag{39}$$

We recall that, by Condition (B), the family $\{\tilde\theta(x,t,\varepsilon)\}$ is bounded in $L^\infty(0,T;W_2^1(\Omega_t^*))$, uniformly with respect to ε and, consequently, $*$-weakly converges in $L^\infty(0,T;W_2^1(\Omega_t^*))$; moreover, by (31), we have $\tilde\theta \to 0$ as $\varepsilon \to 0$ for $x \in \Omega_t^*$ in the sense of the $L^2((0,T)\times\Omega_t^*)$-convergence. Thus,

$$\overline{\theta}(x,t) \overset{def}{=} \lim_{\varepsilon\to 0}\tilde\theta(x,t,\varepsilon) = 0, x \in \Omega_t^*.$$

It is obvious that (31) does not contradict (39) if only the sign of the leading term of corrections (depending on $\tilde s_0^{(j)}$) of velocities is independent of j and then $C_1 = 0$. On the other hand, in the domain Ω_t^*, the limiting heat equation has the free term C_1. To verify this fact, one should prove that

$$\tilde s_0^{(1)} = \tilde s_0^{(2)} + \mathcal{O}(h).$$

The proof is given below in the spherically symmetric case. Now, we continue computations in the integral identity (32). Integrating by parts

$$\tilde I_\theta \overset{def}{=} \int\limits_Q \{-\xi_t\theta + \langle\nabla\xi,\xi\theta\rangle\}\,dxdt + \int\limits_\Omega \xi(x,0)\theta^0 dx, \tag{40}$$

we find (41)
where

$$\theta_{(i)} = \theta|_{Q_\varepsilon^i}.$$

By (31) the integrals over $\Omega_{t,\varepsilon}^i$ and $\mathcal{T}_\varepsilon^i$, $i = 1,\ldots,M$, converge to zero as $\varepsilon \to 0$. We recall that, by Definition 3.1,

$$I_\theta = \tilde I_\theta - \int\limits_Q \varphi\xi_t\,dxdt + \int\limits_\Omega \varphi^0\xi(x,0)\,dx = 0. \tag{42}$$

Taking into account (30), (39), and (41), we arrive at the required result as $\varepsilon \to 0$:

$$\frac{\partial\overline{\theta}}{\partial t} = \Delta\overline{\theta}, \quad x \in \Omega\setminus\Omega_t^*, \quad t > 0, \tag{43}$$

$$\overline{\theta} = 0, \quad x \in \Omega_t^*, \quad t \ge 0, \tag{44}$$

$$\frac{\partial s_0}{\partial t} = |\nabla s_0|\,div\left(\frac{\nabla s_0}{|\nabla s_0|}\right), \quad x \in \Omega_t^*, \quad t > 0, \tag{45}$$

$$\overline{\theta}|_{\partial\Omega_t^*} = 0, \quad \frac{\partial\overline{\theta}}{\partial n}|_{\partial\Omega_t^*} = V_n, \quad t \ge 0, \tag{46}$$

$$\begin{aligned}&\overline{\theta}|_{t=0} = \overline{\theta^0}(x), \quad x \in \Omega\setminus\Omega_0^*,\\ &s_0|_{t=0} = s^0(x), \quad x \in \Omega_0^*, \quad \overline{\theta}|_{\partial\Omega} = \theta_b,\end{aligned} \tag{47}$$

where

$$\begin{aligned}&\partial\Omega_t^* = \Gamma_t^- \cup \Gamma_t^+,\\ &\Gamma_t^- = \{x \in \Omega, s_0(x,t) = 0\},\\ &\Gamma_t^+ = \{x \in \Omega, s_0(x,t) = L\},\end{aligned}$$

n denotes the outward normal to Ω_t^*, $V_n = |\nabla s_0|^{-1}\,\partial s_0/\partial t|_{\partial\Omega_t^*}$, and $s^0(x) \equiv s^0(x,0)$.

Thus, the problem (43)-(46) can be interpreted as two classical one-phase Stefan problems joined by Equation (45). Such an interpretation leads to the problem about the mixture domain for processes with surface tension (cf. [6,8]). The conditions (45), (46) and $\overline{\theta} = 0$ on Ω_t^* are conditions of Hugoniot type since they should be satisfied for the existence of the solution under consideration. The operator on the right-hand side of (45) degenerates along the direction ∇s_0, *i.e.*, along y_1 if we introduce the new coordinates $y_1 = s_0, y_2, \ldots, y_n$, where $y_2,\ldots,y_n$ are the coordinates on the surface $s_0 = const$. Equation (45) is ultraparbolic. As is known [11], a homogeneous ultraparabolic equation has no real analytic solutions with respect to t and y_1, except for the case where the solution is independent of the tangent variables. Further, we need to solve the Cauchy problem (46) for the heat Equation (43) relative to y_1 with the initial conditions on the surface $\partial\Omega_t^*$. For sufficiently small y_1 and t this ill-posed problem has a solution only for real analytic surfaces and initial data [11]; moreover, in this case, the values of $\overline{\theta}$ on the external boundary and at the initial time are uniquely determined by the values on

$$\begin{aligned}\tilde I_\theta = &\int_0^T\left\{\int\limits_{\Omega_{t,\varepsilon}^0}\xi\left(\frac{\partial\theta_{(0)}}{\partial t} - \Delta\theta_{(0)}\right)dx + \int\limits_{\Omega_{t,\varepsilon}^{M+1}}\xi\left(\frac{\partial\theta_{(M+1)}}{\partial t} - \Delta\theta_{(M+1)}\right)dx\right\}dt\\ &+ \int\limits_{\Omega_{t,\varepsilon}^0}\xi\frac{\partial\theta_{(0)}}{\partial\nu_0}d\sigma_0 - \int\limits_{\mathcal{T}_\varepsilon^M}\xi\frac{\partial\theta_{(M+1)}}{\partial\nu_M}d\sigma_M + \sum_{i=1}^M\int_0^T\int\limits_{\Omega_{t,\varepsilon}^i}\theta_{(i)}\left(-\frac{\partial\xi}{\partial t} - \Delta\xi\right)dxdt\\ &+ \int\limits_{\mathcal{T}_\varepsilon^i}\theta_{(i)}\frac{\partial\xi}{\partial\nu_i}d\sigma_i - \int\limits_{\mathcal{T}_\varepsilon^{i-1}}\theta_{(i)}\frac{\partial\xi}{\partial\nu_{i-1}}d\sigma_{i-1} + \int\limits_{\Omega_{0,\varepsilon}^*}\xi(x,0)\theta^0 dx,\end{aligned} \tag{41}$$

3.3. Example of a Structured Domain

Assume that $n = 3$, $\Omega = \{x, R_- < r < R_+\}$, where $r = |x|$, $R_- > 0$, and $\Omega_0^* = \{r_-(0) < r < r_+(0)\}$. Then Equation (45) becomes the first order equation

$$\frac{\partial s_0}{\partial t} = \frac{2}{r}\frac{\partial s_0}{\partial r}, \quad r \in \Omega_t^* = \{r_-(t) < r < r_+(t)\}, \quad t > 0. \qquad (48)$$

It is easy to solve the problem (48) with the initial condition

$$s_0|_{t=0} = s^0(r).$$

Namely,

$$s_0(r,t) = s^0\left(r^0\right)$$

along the characteristics

$$r\left(r^0,t\right) = \sqrt{\left(r^0\right)^2 - 4t}, \quad r_-(0) \le r^0 \le r_+(0)$$

for any smooth function $s^0(r)$ such that $s_r^0 > 0$.

Now, (43), (46) with

$$V_n = 2/r|_{\partial\Omega_t^*}$$

is the Cauchy problem (with respect to r) in two domains

$$Q_1 = \{R_- < r < r_-(t), t > 0\},$$
$$Q_2 = \{r_+(t) < r < R_+, t > 0\}.$$

To formulate the solvability conditions for this ill-posed problem, we recall the well-known fact (cf., for example, [11]): for the local existence of a solution to (43), (46) it is sufficient that the curves $r_\pm(t)$ be real analytic functions with respect to t, *i.e.*, $r_-(0) > 0$ and $t < r_-^2(0)/4$. Consequently, for sufficiently small $\delta_0 > 0$ and $T_0 = T_0(\delta_0)$, in the domains

$$Q_1^* = \{r_-(0) - \delta_0 < r < r_-(t), t < T_0\},$$
$$Q_2^* = \{r_+(t) < r < r_+(0) + \delta_0, t < T_0\}$$

there exists a real analytic solution $\overline{\theta}$ to the corresponding Cauchy problem. Thus, in order to solve the limit problem (43)-(46), we need to impose the following condition.

(C) Suppose that Ω is a spherically symmetric layer in $\mathbb{R}^3$, the initial and boundary data of the problem

$$\begin{aligned} &\partial_t\varphi + \partial_t\theta = \Delta\theta, \\ &\varepsilon^2\partial_t\varphi = \varepsilon^2\Delta\varphi + \varphi - \varphi^3 + \varepsilon\varkappa\theta \\ &\varphi|_{t=0} = \varphi^0(x,\varepsilon), \ \theta|_{t=0} = \theta^0(x,\varepsilon), \\ &\varphi|_\Sigma = 1, \ \theta|_\Sigma = \theta_b \end{aligned} \qquad (49)$$

are spherically symmetric, and

$$\Gamma_{0,\varepsilon}^i = \{x \in \Omega, |x| = r_i^0\},$$

where $0 < R_- < r_0^0 < r_1^0 < \cdots < r_M^0 < R_+$. Assume that $r_{j+1}^0 - r_j^0 = h$ and the differences $r_0^0 - R_-$ and $R_+ - r_M^0$ are sufficiently small; moreover, $s^0(r)$ is real analytic, $\partial s^0/\partial r > 0$, $\overline{\theta^0}(x)$ and θ_b are special data corresponding to the solution to the Cauchy problem for the heat Equations (43), (46).

We show that Condition (C) implies Condition (B) and the equality $C_1 = 0$ in (39). For this purpose, we return to the main problem (cf. (17)-(23))

$$\frac{\partial\theta_i}{\partial t} = \Delta\theta_i, \quad x \in \Omega_{t,\varepsilon}^i, \quad t > 0, \qquad (50)$$

$$\theta_{i-1}|_{\Gamma_{t,\varepsilon}^{i-1}-0} = \theta_i|_{\Gamma_{t,\varepsilon}^{i-1}+0}, \qquad (51)$$

$$\theta_i|_{\Gamma_{t,\varepsilon}^i-0} = \theta_{i+1}|_{\Gamma_{t,\varepsilon}^i+0}, \qquad (52)$$

$$\frac{\partial\theta_{i-1}}{\partial\nu_{i-1}}|_{\Gamma_{t,\varepsilon}^{i-1}-0} - \frac{\partial\theta_i}{\partial\nu_{i-1}}|_{\Gamma_{t,\varepsilon}^{i-1}+0} = (-1)^{i+1}2V_{\nu_{i-1}}, \qquad (53)$$

$$\frac{\partial\theta_i}{\partial\nu_i}|_{\Gamma_{t,\varepsilon}^i-0} - \frac{\partial\theta_{i+1}}{\partial\nu_i}|_{\Gamma_{t,\varepsilon}^i+0} = (-1)^i 2V_{\nu_i}, \qquad (54)$$

$$(-1)^{i-1}\varkappa_1\theta_i|_{\Gamma_{t,\varepsilon}^{i-1}-0} = \mathcal{K}_t^{i-1} - V_{\nu_{i-1}}, \qquad (55)$$

$$(-1)^i\varkappa_1\theta_i|_{\Gamma_{t,\varepsilon}^i+0} = \mathcal{K}_t^i - V_{\nu_i}. \qquad (56)$$

Let $\rho_i = \rho_i(t,h)$ be functions such that $\Gamma_{t,\varepsilon}^i = \{r, r = \rho_i(t,h)\}$. In the spherically symmetric case, we have

$$\mathcal{K}_t^i = -2/\rho_i.$$

Therefore, taking into account (30) and choosing ν_i directed in the opposite direction relative to the normals (with respect to $D_{t,\varepsilon}^i$), we find

$$V_{\nu_i} = -2/\rho_i + \mathcal{O}(h).$$

We make the change of variables $\theta_i = w_i/r$. Then the equality

$$\partial_t\theta = \Delta\theta, \ x \in \Omega_{t,\varepsilon}^i, \ t > 0$$

takes the form

$$\frac{\partial w_i}{\partial t} = \frac{\partial^2 w_i}{\partial r^2}, \ r \in \left(\rho_{i-1}(t), \rho_i(t)\right), \ t > 0. \qquad (57)$$

Since

$$V_{\nu_i} = -2\rho_i^{-1}(1 + hv_{\nu_i}), \quad v_{\nu_i} \overset{def}{=} \rho_i(\mathcal{K}_t^i - V_{\nu_i})/2h,$$

the conditions (51), (54) can be written as follows:

$$\frac{\partial w_{i-1}}{\partial r}|_{\Gamma_{t,\varepsilon}^{i-1}-0} - \frac{\partial w_i}{\partial r}|_{\Gamma_{t,\varepsilon}^{i-1}+0} = (-1)^i 4\left(1 + hv_{\nu_{i-1}}\right), \qquad (58)$$

$$\frac{\partial w_i}{\partial r}|_{\Gamma_{t,\varepsilon}^i-0} - \frac{\partial w_{i+1}}{\partial r}|_{\Gamma_{t,\varepsilon}^i+0} = (-1)^{i+1}4\left(1 + hv_{\nu_i}\right), \qquad (59)$$

$$w_{i-1}\big|_{\Gamma^{i-1}_{t,\varepsilon}-0} = w_i\big|_{\Gamma^{i-1}_{t,\varepsilon}+0}, \tag{60}$$

$$w_i\big|_{\Gamma^{i}_{t,\varepsilon}-0} = w_{i+1}\big|_{\Gamma^{i}_{t,\varepsilon}+0}. \tag{61}$$

We show that the problem (57), (58) has a solution satisfying the following properties:

1) $w_i = \mathcal{O}(h)$ uniformly with respect to i,

2) for any t the values $\hat{w}_i = (-1)^i w_i\big|_{r=p_i}$ are determined, with accuracy $\mathcal{O}(h^2)$, by the values of some function $\hat{w}_i \in C^1[\rho_0, \rho_M]$ on the grid $\{\rho_0, \ldots, \rho_M\}$.

We note that the first property is related to (56) and (30).

We look for a solution w_i to the problem (57), (58) in the form

$$w_i = a_i(r-\rho_{i-1}) + b_i(t) + u_i(t,r,h), \tag{62}$$

where the first two terms correspond to the Stefan condition (58) and u_i is a solution to the following chain of problems:

$$\frac{\partial u_i}{\partial t} - \frac{\partial^2 u_i}{\partial r^2} = a_i\dot{\rho}_{i-1} - \dot{b}_i, \quad i = 1,\ldots,M, \tag{63}$$

$$(u_j - u_{j+1})|_{r=\rho_j} = 0, \quad \left(\frac{\partial u_j}{\partial r} - \frac{\partial u_{j+1}}{\partial r}\right)\Big|_{r=\rho_j} = 0, \tag{64}$$
$$j = 0,\ldots,M.$$

We note that this chain is similar to that considered in [12] and differs by only the dependence of $f_i = a_i\dot{\rho}_{i-1} - \dot{b}_i$ in (63) on t. However, because of this dependence, it is obvious that the contribution of this chain to the solution is of order $\mathcal{O}(h^2)$.

To solve Equation (63), we first compute the coefficients a_i and b_i. From (58) and (63) it follows that

$$a_i = 2(-1)^{i+1}(1+hv_{\nu_i}), \quad b_1 = 0,$$

$$b_j = 2\sum_{k=2}^{j}(-1)^k\left[(1+hv_{\nu_{k-1}})\rho_{k-1} - (1+hv_{\nu_{k-2}})\rho_{k-2}\right],$$
$$j = 2,\ldots,M.$$

Assume that

$$\tilde{s}_0^{(j)}(x,t,h) = s_1(x,t) + \mathcal{O}(h), \quad j = 1,2, \tag{65}$$

where the functions $\tilde{s}_0^{(j)}$ are defined in (29). At the first glance, this assumption can lead to a contradiction in the equation for velocity correction (the linearized Gibbs-Thomson equation for $\tilde{s}_0^{(j)}$) if the functions ω_i computed under this assumption do not satisfy Conditions 1) and 2) However, it turns out that there is no contradiction.

Denote by $R(z,t,h)$ a solution to the equation

$$s_0(R,t) + hs_1(R,t) = z.$$

By construction, $\rho_i = R(ih,t,h)$ and, uniformly with respect to i up to order $\mathcal{O}(h)$, the functions v_{ν_i} are traces of some C^1-function v on the surfaces $r = \rho_i$. We note that $\partial R/\partial z > 0$. Furthermore,

$$b_j = 2h\sum_{k=2}^{j}(-1)^k \frac{\partial R}{\partial z}\big|_{z=h(k-2)} + \mathcal{O}(h^2) = \mathcal{O}(h).$$

By Lemma 3.1, the last estimate is uniform with respect to j. Further,

$$b_{j+2} - b_j = 2(-1)^{j+1}(\rho_{j+1} - 2\rho_j + \rho_{j-1}) + \mathcal{O}(h^3) = \mathcal{O}(h^2), \tag{66}$$

and this estimate is also uniform with respect to j. Now, we see (67)

Furthermore, from (66) and Lemma 3.1 it follows that

$$b_{j+2l} - b_j = \mathcal{O}(h^2)$$

uniformly with respect to j and l. In particular, from this estimate, the equality (67), and the condition $b_1 = 0$ we find

$$b_{2l} = 2h\frac{\partial R}{\partial z}\big|_{z=(2l-1)h} + \mathcal{O}(h^2), \quad b_{2l+1} = \mathcal{O}(h^2).$$

We consider a broken line $\mathcal{L}$ such that its linear parts are defined as $a_i(r-\rho_{i-1}) + b_i$ on the segments $[\rho_{i-1}, \rho_i]$. It is obvious that b_i are the values of $\mathcal{L}$ at the points $r = \rho_{i-1}$. Consequently, $\mathcal{L}$ is not symmetric with respect to the zero line (it is directed toward to the domain of positive values). However, the broken line can be centered by decreasing its values $h\frac{\partial R}{\partial z}|_{z=h(i-1)}$ in each segment $[\rho_{i-1}, \rho_i]$. It is obvious that this is equivalent to the existence of functions

$$m = h\frac{\partial R}{\partial z}\big|_{z=z(r,t,h)}$$

in $\mathcal{L}$. Here, $z = z(r,t,h)$ satisfies the equation $R(z,t,h) = r$.

We set

$$\mathcal{L}_1 = \mathcal{L} - m, \quad U_i = u_i + m.$$

Then for U_i we have the problem of the form (63) with the right-hand sides

$$G_i = a_i\dot{\rho}_{i-1} - \dot{b}_i + \frac{\partial m}{\partial t} - \frac{\partial^2 m}{\partial r^2}, \quad r \in (\rho_{i-1}, \rho_i). \tag{68}$$

$$b_{j+1} - b_j = 2(-1)^{j+1}\left(h\frac{\partial R}{\partial z} + \frac{h^2}{2}\frac{\partial^2 R}{\partial z^2} + h^2\frac{\partial}{\partial z}(Rv)\right)\big|_{z=(j-1)h} + \mathcal{O}(h^3). \tag{67}$$

To construct an asymptotic expansion of U_i, we solve a chain of problems. We look for a solution in the form

$$U_i = c_i(r-\rho_{i-1})(\rho_i - r) + c_{i1}(r-\rho_{i-1})^2 (\rho_i - r) + c_{1i}(r-\rho_{i-1})(\rho_i - r)^2 + \cdots,$$

where dots denote polynomials of higher degree. We note that polynomials of degree higher than 2 admit the estimate $O(h^3)$ and the coefficients c_i are determined by the relations

$$c_i = 2(-1)^{i+1}\dot{\rho}_{i-1} + \mathcal{O}(h), \quad i = 1,\ldots,M.$$

The contribution to the solution U_i of terms of order $O(h)$ in G_i is estimated by $\mathcal{O}(h^3)$. Consequently,

$$U_i = \widehat{U}_i + \mathcal{O}\left(h^3\right)$$

and the function

$$\widehat{U}_i = c_i(r-\rho_{i-1})(\rho_i - r)$$

is defined by a sequence that is symmetric with respect to the zeros of parabolas of order $\mathrm{mod}\,\mathcal{O}(h^3)$ because

$$a_i\dot{\rho}_{i-1} + a_{i+1}\dot{\rho}_i = \mathcal{O}(h).$$

Hence $\widehat{U}_i = \mathcal{O}(h^2)$ for $r \in (\rho_{i-1}, \rho_i)$ and the values of $\mathcal{L}_1$ at the points ρ_j are given by the relation

$$\mathcal{L}_1|_{r=\rho_j} = (-1)^j h\frac{\partial R}{\partial z}|_{z=(j-1)h} + \mathcal{O}(h^2), \quad j = 1,\ldots,M. \tag{69}$$

Thus, the problem (57), (58) has a solution with properties 1) and 2).

It remains to construct θ in the domains $R_- \le r \le \rho_0(t)$ and $\rho_M(t) \le r \le R_+$. We note that constructing $\mathcal{L}_1$, we defined $\mathrm{mod}\,\mathcal{O}(h)$ the values of θ and $\dfrac{\partial\theta}{\partial r}$ at the points $r = \rho_0(t)$ and $r = \rho_M(t)$. As in the case (43)-(46), this fact completes the construction of θ. Now, it is again required to solve the Cauchy problem with respect to r for the heat equation. Nevertheless, by Condition (C), the analyticity condition (necessary for solvability) is already valid.

Thus, by (69), the functions

$$\hat{\theta}_i(t) = (-1)^i\,\theta\big|_{r=\rho_i},$$

with accuracy $\mathcal{O}(h^2)$, are traces on the surfaces $\Gamma^i_{t,\varepsilon}$ of some function

$$\hat{\theta}(x,t,h) = \mathcal{O}(h)$$

of class C^1. Owing to this fact, we can compute the first correction for the phase $s_0(r,t)$. Indeed, substituting (29) into (56), we obtain the linearized Gibbs-Thomson conditions

$$\left(\frac{\partial\tilde{s}_0^{(n_j)}}{\partial t} - \frac{2}{r}\frac{\partial\tilde{s}_0^{(n_i)}}{\partial r}\right)\Big|_{r=\rho_i} = (-1)^i\,\theta_i\,\frac{\varkappa_1}{h}\frac{\partial s_0}{\partial r}\Big|_{r=\rho_i} + \mathcal{O}(h). \tag{70}$$

Our analysis shows that, with accuracy $\mathcal{O}(h)$, the right-hand side of (70) is the trace of a function of class C^1. Therefore, from (69) and the conditions

$$\tilde{s}_0^{(1)}\big|_{t=0} = \tilde{s}_0^{(2)}\big|_{t=0} = 0$$

we find

$$\left(\frac{\partial s_1}{\partial t} - \frac{2}{r}\frac{\partial s_1}{\partial r}\right)\Big|_{r=\rho_i} = \frac{\varkappa_1}{h}\frac{\partial R}{\partial z}\Big|_{z=(i-1)h}\frac{\partial s_0}{\partial r}\Big|_{r=\rho_i} + \mathcal{O}(h). \tag{71}$$

Let

$$\rho_i(t,h) = r_i(t) + h\tilde{r}_i(t,h),$$

so that

$$\frac{\tilde{r}_i(t,h)}{r_i(t)} = \mathcal{O}(1)$$

uniformly with respect to $i = 0,\ldots,M$. Taking into account Equation (48), we obtain

$$r_i = \sqrt{g^2(ih) - 4t},$$

where $g(z)$ is the inverse of s^0, *i.e.*, $s^0(g(z)) = z$. Thus, ignoring terms of order $\mathcal{O}(h)$, we can transform (71) as follows:

$$\frac{\partial s_1}{\partial t} = \frac{2}{r}\frac{\partial s_1}{\partial r} + \frac{\varkappa_1}{r}, \quad s_1|_{t=0} = 0,$$

i.e., our assumption about $s_1(r,t)$ is valid.

We note that, in view of (65), the value C_1 in (37), (39) is equal to zero and consequently, right-hand side of the heat equation in Ω_t^* vanishes.

Thus, Condition (C) implies the validity of Condition (B). As a result, we find (43)-(46) as the limit of the chain of Stefan problems with the Gibbs-Thomson condition.

We formulate the initial conditions. We assume that Conditions (A) and (C) are satisfied. Let

$$\varphi|_{t=0} = \varphi_1^{as}(x,0,\varepsilon) + O\left(\varepsilon^2\right), \quad S^j|_{t=0} = s^0(x,\varepsilon),$$

where $s^0(x,\varepsilon) = s^0(r)$. Let $\theta|_{t=0}$ in the domains $\Omega^i_{0,\varepsilon} = \{r^0_{i-1} < r < r^0_i\}$, $i = 1,\ldots,M$, is defined by

$$\theta_{(i)}|_{t=0} = (-1)^{i+1}\frac{2}{r}\left(-\frac{h}{2(s^0)'_r} + (r - r^0_{i-1}) + \mathcal{O}(h^2)\right),$$

and, in the domains $R_- < r < r^0_0$ and $r^0_M < r < R_+$, we set

$$\theta|_{t=0} = \Xi|_{t=0},$$

where Ξ is a solution to a special Cauchy problem (relative to r) for the heat Equation (43).

Theorem 3.1 *Under the above assumptions, there exists an asymptotic solution to the phase field system satisfying Condition* (B), *and it is possible to pass to the limit in* (49) *as* $\varepsilon \to 0$ *in the sense of Definition* 3.1. *The limit problem*

$$\frac{\partial \overline{\theta}}{\partial t} = \Delta \overline{\theta}, \quad x \in \Omega \setminus \Omega_t^*, \quad t > 0, \tag{72}$$

$$\overline{\theta} = 0, \quad x \in \Omega_t^*, \quad t \geq 0, \tag{73}$$

$$\frac{\partial s_0}{\partial t} = |\nabla s_0| div\left(\frac{\nabla s_0}{|\nabla s_0|}\right), \quad x \in \Omega_t^*, \quad t > 0, \tag{74}$$

$$\overline{\theta}|_{\partial\Omega_t^*} = 0, \quad \frac{\partial \overline{\theta}}{\partial n}|_{\partial\Omega_t^*} = V_n, \quad t \geq 0, \tag{75}$$

$$\begin{aligned} &\overline{\theta}|_{t=0} = \overline{\theta^0}(x), \quad x \in \Omega \setminus \Omega_0^*, \\ &s_0|_{t=0} = s^0(x), \quad x \in \Omega_0^*, \quad \overline{\theta}|_{\partial\Omega} = \theta_b, \end{aligned} \tag{76}$$

where

$$\begin{aligned} &\partial\Omega_t^* = \Gamma_t^- \cup \Gamma_t^+, \\ &\Gamma_t^- = \{x \in \Omega, s_0(x,t) = 0\}, \\ &\Gamma_t^+ = \{x \in \Omega, s_0(x,t) = L\}, \end{aligned}$$

n denotes the outward normal to Ω_t^*, $V_n = |\nabla s_0|^{-1} \partial s_0 / \partial t|_{\partial\Omega_t^*}$ *and* $s^0(x) \equiv s^0(x,0)$, *possesses a solution, at least for sufficiently small* (*but independent of* ε) *time.*

The above case can be explained by the fact that, outside the layer $r_0 \leq r \leq r_M$, the order function of the original problem takes different values: $\varphi \sim -1$ for $r < r_0$ and $\varphi \sim 1$ for $r > r_M$. It is obvious that all the arguments remain valid in the case where φ takes the same values ($\varphi \sim -1$ or $\varphi \sim 1$) for $r \notin [r_0, r_M]$. This means that M is odd. Then we again obtain a limit problem of the form (72)-(75). The limit passage can be justified in the same way as above, by solving a chain of problems which can be reduced to the chain of problems (63). In both cases (M is even or odd), the problems are ill-posed. However, as was noted in [6], such a wave train type structure appears in numerical experiments as solutions to the phase field system with the initial data $\theta^0 = 0$ for $R_- \leq r \leq R_+$ and

$$\varphi^0 = \begin{cases} 1 + \sum_{j=0}^{M} (-1)^j th\left(\dfrac{r - r_j^0}{\varepsilon}\right), & \text{M is odd,} \\ \sum_{j=0}^{M} (-1)^j th\left(\dfrac{r - r_j^0}{\varepsilon}\right), & \text{M is even.} \end{cases}$$

Figure 2(b) presents the graphs of solutions to the phase field system with spherically symmetric initial data for $M = 19$ and $\varepsilon = 10^{-2}$ at different times. One can see that the temperature in the mixture domain is of sawtooth form. Such a function is the leading part of the asymptotic expansion (62) of the solution to the chain of modified Stefan problems with the Gibbs-Thomson condition. In the numerical analysis performed by O. A. Vasil'eva, $\theta = 0$ on the external boundaries. This leads to an effect presented in the figure for time $t = 0.02$: the sawtooth structure begins to break down under the influence of boundary data. However, the order function is more stable and preserves its shape.

Figure 2(a) presents the graphs of solutions to the phase field system with spherically symmetric initial data for $M = 7$ and $\varepsilon = 10^{-2}$ at different times. The temperature has sawtooth shape in the IAS-domain, whereas it is periodic with amplitude $l = 1$ at center. Such a function is the leading part of the asymptotic expansion (62) of the solution to the chain of modified Stefan problems with the Gibbs−Thomson condition. The sawtooth structure "moves" to the center and begins to break down under the influence of nonspecial boundary data. The order function preserves its shape in this case.

Figure 3(a) presents the graphs of solutions to the phase field system with spherically symmetric initial data for $M = 7$ and $\varepsilon = 10^{-2}$ at different times. **Figure 3(b)** presents the graphs of solutions to the phase field system with spherically symmetric initial data for $M = 19$ and $\varepsilon = 10^{-2}$ at different times.

3.4. Comments and Conclusions

Based on the phase field system, it is possible to detect a banding structure formation in instability zones. However, to construct the mathematical model, we need to impose some restricted conditions.

1) The existence conditions are very restrictive, which can be explained by the geometry of domain Ω and the initial and boundary conditions. Note that the initial and boundary data are determined by the solution to the limit problem.

2) A standard definition of a weak solution can turn out to be not suitable. However, we can avoid these difficulties by introducing a special definition of a weak solution, which is important for nonlinear problems.

3) As was shown, a wave train type solution exists only for special boundary and initial data providing the existence of an asymptotic solution to the chain of Stefan problems with the Gibbs-Thomson condition for sufficiently small (but independent of ε) times. This fact allows us to pass to the limit of the chain of Stefan problems with the Gibbs−Thomson condition (in the sense of

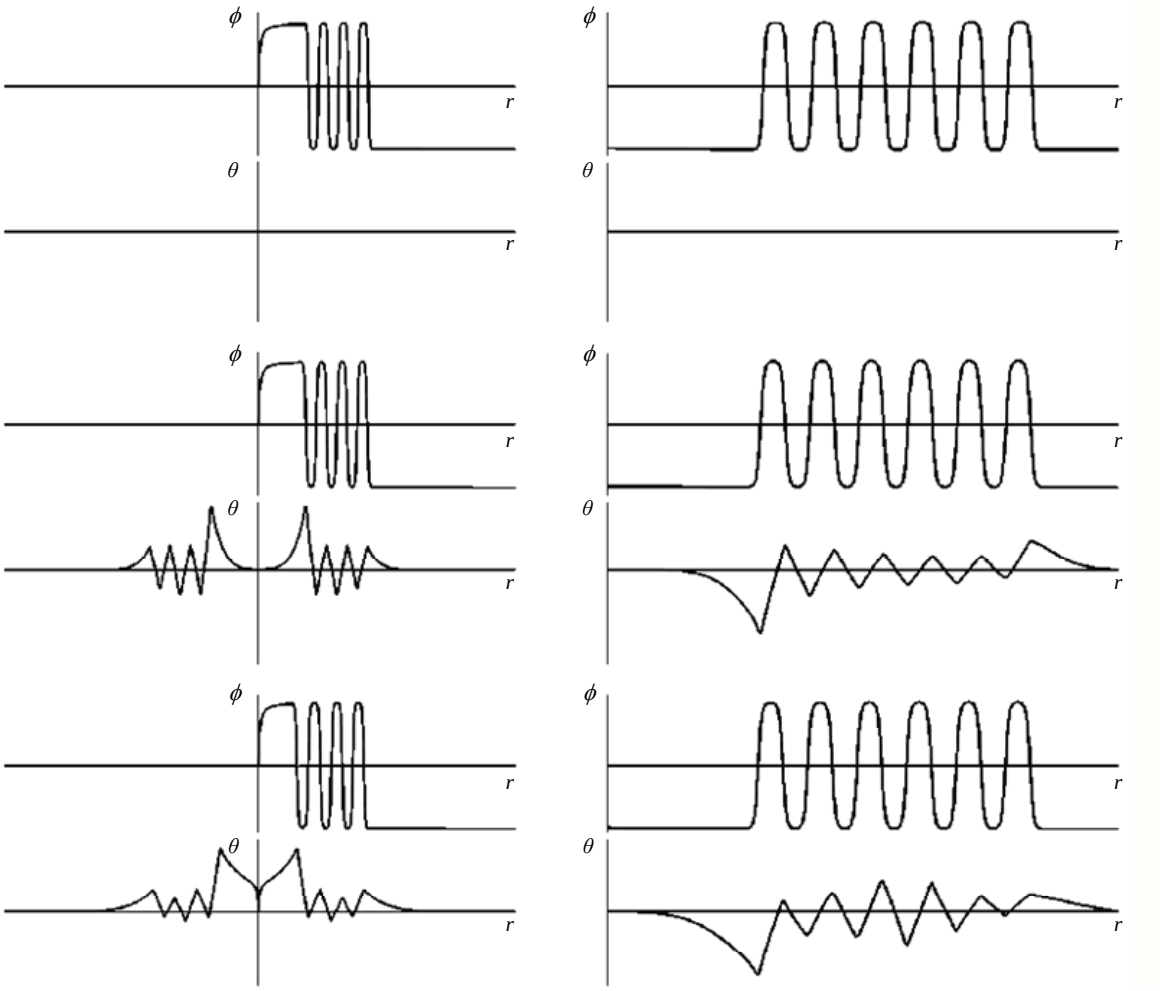


Figure 2. (a) Presents the graphs of solutions to the phase field system with spherically symmetric initial data for $M = 7$ and $\varepsilon = 10^{-2}$ at different times; (b) Presents the graphs of solutions to the phase field system with spherically symmetric initial data for $M = 19$ and $\varepsilon = 10^{-2}$ at different times.

Definition 3.1) and derive the limit problem (43)-(46).

4) As we shown in the above examples, the temperature $\theta(x,t,\varepsilon)$ is small ($\theta \to 0$ as $\varepsilon \to 0$) and has special "periodic" structure in the stratified domain.

5) Even in the rigid-front case, the solid phase growth is of order $\ln(1/t)$, which is lower than the order obtained in experimental way.

Thus, a banding structure in the phase stratification domain of a binary alloy was constructed under extremely restrictive conditions on the geometry of domain Ω and the initial and boundary conditions. Furthermore, the order $O(\sqrt{t})$ of the solid phase growth obtained in experiments is not achieved in this model. In view of these facts, it is necessary to look for other mathematical models describing qualitative experimental properties of crystallization. In the following section, for such a model we consider the convective Cahn-Hilliard equations in a porous medium of an overcooled melt.

4. Oriented Crystallization Model

There is a huge experimental literature on various structure formations in melt crystallization. Based on experimental results, one can conjecture that complex structure formations in crystallization are caused by the evolution of instabilities during phase transition processes which, in turn, is caused by different reasons and can be realized in different ways. We list some of such reasons.

1) concentration overcooling,

2) convective flows deforming the temperature field (gravity and thermocapillary convection),

3) phase stratification.

In addition, elastic properties of the solid phase, thin phase boundary, and adsorption phenomena can also

contribute to this effect.

4.1. Modified Convective Cahn-Hilliard Model in a Porous Melt

To construct a mathematical model governing the reconstruction of oriented crystallization (cf. [7,8]), a modified Biot model of a porous medium [13] was used for describing a liquid-solid zone and the convective Cahn-Hilliard model of spinodal decomposition [14,15] was used for describing segregation. In the model, we consider a binary eutectic alloy. For variables we take

the concentration of the component *A* or the component *B* of the binary alloy,

the temperature,

the growth velocity of the solid phase,

the contraction,

the convection velocity of the liquid phase.

The model includes the laws of conservation of mass and impulse for liquid and the law of conservation of total impulse for liquid and solid phases.

In accordance with the physical interpretation, the model also includes a modified Cahn-Hilliard equation [14] and the heat equation [7], regarded as a generalization of the Stefan problem [9]. Using a nonisothermic modification of the Cahn-Hilliard model, proposed in [7], we can construct a model that take into account the following physical effects.

Because of crystallization and melting, the temperature can vary. In turn, variations of temperature lead to variations of velocity and changes of the medium composition.

An equilibrium phase transition is realized at the melting temperature, whereas a nonequilibrium phase transition can be realized at different temperatures depending on the depth of penetration into metastable or labile regions. This fact shows that the modified Cahn-Hilliard model should include temperature-dependent parameters. Then both heat-mass transfer equations will govern mutually dependent processes.

The model reflects the structure of a liquid-to-solid transition zone of the crystallization front. It consists of an outer viscous layer (the hydrodynamic Prandtl layer) and a diffuse layer (the Nernst diffusion layer). In the case of a condensed system, the thickness of the Nernst layer is less than the thickness of the Prandtl layer by three orders and the heat-mass transfer laws can be assumed to be linear (the Fick and Fourier laws). On the boundary of the diffuse layer, near the solid phase, the volume strongly varies while a liquid-to-solid transition. Therefore, it is necessary to take into account elastic forces, which can be done within the framework of continuum mechanics.

We introduce the following notation:

c is the mole concentration of the component B in the binary alloy (In our case, the mole concentration of Sn in the liquid phase),

z is the contraction,

w_l is the convection velocity of the liquid phase,

u is the averaged displacement in the solid phase,

w_s is the mean growth velocity of the solid phase,

v is the averaged fictitious displacement in the liquid relative to the solid phase,

T is the temperature.

Furthermore, we set

$$w = w_l - w_s.$$

4.2. One-Dimensional Case

In this case, the model is represented (cf. [8]) by a system of differential equations which can be divided into the three subsystems:

$$\begin{cases} u_t = w^s, \quad v_t = w, \\ \rho\left(w^s\right)_t + \rho^l w_t = \left[\left(\lambda + 2\mu + \alpha^2 M\right)u_x + \alpha M v_x\right]_x + \rho g, \\ \rho^l\left(w^s\right)_t + \rho_{add} w_t = -Dw + \left[\alpha M u_x + M v_x\right]_x + \rho^l g, \end{cases} \tag{77}$$

$$\left[\left(\rho^l\right)_t + \left(\rho^l w\right)_x + \rho^s\left(w^s\right)_x = 0, \right. \tag{78}$$

$$\begin{cases} c_t + w(c - c_{kr})c_x &= [M_D(F(c,T) + u_x^2 \partial_c(\lambda + 2\mu) \\ &\quad - \varepsilon^2 (F_1(c,T)c_x)_x + \varepsilon^4 (F_2(c,T)c_{xx})_{xx})_x]_x, \\ &\quad (T + \varkappa c)_t = D_0 T_{xx}, \end{cases} \tag{79}$$

This system describes processes in the diffuse and Prandtl layers in dimensionless variables c, T, w^l, w^s, u, v, z.

The system (77) is a model of wave propagation in a porous skeleton filled with a liquid (a simplified version of the Biot model).

The system (78) is the continuity equation and describes the evolution of contraction.

The system (79) presented by the convective Cahn-Hilliard model and the heat transfer equation describes the formation and growth of Gibbs grains.

Note that we use equations of continuum mechanics to describe processes in the Prandtl layer, whereas for diffusion and heat processes we use the modified Cahn-Hilliard model where hydrodynamic processes and elastic-plastic state of the solid phase are taken into account. Let's note that all constructions of the previous chapter were maded for this one-dimensional case but more technically.

The model contains a number of dimensionless parameters. The elasticity modulus of the solid phase $\Lambda = \lambda + 2\mu$ is assumed to be a function of concentration and temperature: $\Lambda = \Lambda(c,T)$.

The parameter z is expressed by the formula

$$z = \frac{V - V^s - V^i}{V},$$

where V is the total melt volume, V^s is the current volume of the solid phase, and V^l is the current volume of the liquid phase.

The concentration c is expressed as

$$c = \frac{m_B^s / M_B}{m^l / M_l},$$

where M_A (M_B) is the atomic mass of the component A (B) and M^l is the averaged atomic mass of the melt: $M^l = (1-c)M_A + cM_B$.

The extra variable $y(c)$ of the form $y = m_B^l / m_B$ is expressed in terms of concentration as follows:

$$y(c) = \frac{qc}{(1-c)Q}, \tag{80}$$

which implies

$$\rho^l(c,z) = \frac{q + y(c)}{1 - z - R(1 - y(c))} R,$$

where $(1-y(c)))^{-2/3}$ is a bound for the surface of the solid phase in the liquid-solid region, Γ_1 is a parameter,

$$R = \frac{m_{Sn}}{\rho_s V} = 0,74,\ q = \frac{m_{Pl}}{m_{Sn}} = 0,25,\ Q = \frac{M_{Pl}}{M_{Sn}} = 1,75$$

are constants, and we adopt the normalization condition

$$\rho^s = \rho_B = 1.$$

Further,

ρ is the mean density.

$M = 6,2$ is the mobility (fluidity) of the liquid.

α^2 is the inverse of the relaxation time of fluidity (estimated as $\alpha \sim 10^{-3}$),

$g = -1/30$ is the acceleration of gravity,

D is the interphase friction coefficient, estimated as

$$D = \Gamma_1 e^{-\Xi/T} \left(1 - e^{-1/T}\right) \left((1 - y(c))\right)^{-2/3},$$

where $\Xi = 1$, $\Gamma_1 = 5$,

M_D is the diffuse mobility of the component B (estimated as $M_D = 1/T$),

$\varkappa$ is the ratio of the melting enthalpy to the heat capacity of the solid phase at a constant pressure.

The function F determines steady, metastable, and labile states of the system "melt-alloy" depending on the composition and temperature. The function F can be approximated by a cube polynomial in c at a fixed temperature:

$$F(c,T) = \begin{cases} (c - c^-)(c - c_{cr})(c^+), & T \le T_0 \\ (c - c_{cr})^3, & T \ge T_0, \end{cases}$$

where $T_0 = 400K$ and $c^{\pm}, c_{cr}$ are functions of temperature T which will be specified below,

$$T_{\min} \le T \le T_{\max},\quad T_{\min} = 233,15K,\quad T_{\max} = 456,15K.$$

We define three concentration values:

$c_{\min}$ equals to $c(T_{\min})$,

c_{mid} equals to $c_{cr}(T_{\min})$,

c_0 equals to c_{mid} in our experiment.

We set

$$c_{\min} = 0,04,\quad c_{mid} = 0,43.$$

We introduce $c^{\pm}(T)$ as the roots of the equation

$$T = \alpha_{clust} c^2 + \beta_{clust} c + \gamma_{clust},$$

where

$$\alpha_{clust} = \frac{T_{\min} - T_0}{(c_{\min} - c_0)^2},\ \beta_{clust} = -2\alpha_{clust} c_0,$$

$$\gamma_{clust} = T_0 + \alpha_{clust} c_0^2$$

and define $c_{cr}(T)$ as a linear function.

The function $F_1(c,z,T) \equiv 1000$ is interpreted as viscosity. At the first step, it is assumed to be constant. The structure of the interphase boundary at atomic level is characterized by the function F_2. We set $F_2 \equiv 0$ and $\varepsilon = 10^{-4}$ in the numerical experiment.

The model also contains some additional relations dictated by the physical interpretation of the problem. In particular, the model contains the "extra" density ρ_{add} such that

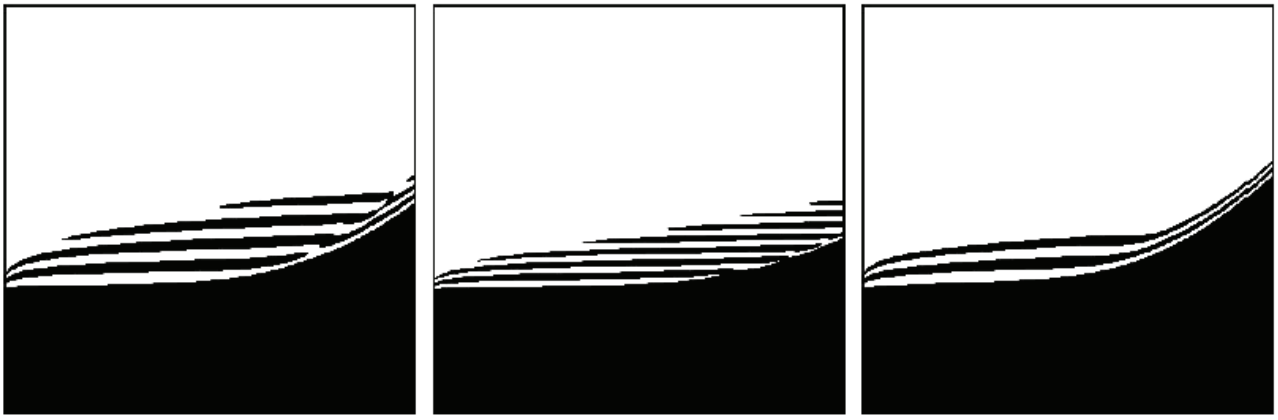
$$\rho_{add} = \frac{\beta \zeta(z) \rho^l}{(1 - y(c))^{1/3}}, \tag{81}$$

where $\beta = 0,055$ and, as a rule, $\zeta(z)$ is small for small z.

4.3. Numerical Analysis of the Model Describing Oriented Crystallization. One-Dimensional Case

The numerical results obtained by Rykov and Zaitsev [16] are presented in **Figures 3(a-c)**. Note, that the spatial x-axis is directed upward, whereas the t-axis is directed rightward along the horizontal line.

The systems presented in **Figures 3(a-c)** differ by the value of the parameter $\varkappa$. The numerical results show



(a) (b) (c)

Figure 3. Numerical simulation of the model describing oriented crystallization (one-dimensional case). The spatial *x*-axis is directed upward, whereas the *t*-axis is directed rightward along the horizontal line.

that the balance of convective and diffusive terms generates a modulated wave of formation of crystal grains, which differentiate the spinodal decomposition mechanism from the classical case where the Cahn-Hilliard model possesses a periodic solution.

4.4. Comments

1) In our model, for the sake of simplicity, we assume that porosity is constant, passing its functions to the contraction *z*. On this stap of the model construction we will elucidate the change range of the porosity when the modification of Biot model don't lose the hyperbolicity. It allow us on the next stap to pass to porosity as a problem variable, expressed the contraction as the function of porosity.

2) In the model (77)-(79), the convention is equal to zero at the initial time, $w|_{t=0} = 0$, and then it can be regarded as reaction to 1) the force of interphase friction between liquid and solid phases and 2) the gravity force. Thereby we specify the effective force in the convective Cahn-Hilliard model [14,15].

3) The initial distribution of crystal grains (which, unlike [14], is not given here) depends on only contraction, whereas the further distribution is determined by the process. So, no restrictive conditions are imposed on the initial-boundary data, unlike the case of the phase field system and the one-dimensional convective Cahn-Hilliard model.

4) In the subsystem (79), we took into account the results of [17]. Note that the above constructions remain also valid for the modified model (77)-(79) obtained from the two-dimensional model (cf. below) in the radial-symmetric case.

4.5. Two-Dimensional Case

Introduce the notation:

c is the mole concentration of Sn in the liquid phase,

z is contraction,

$\overline{w^l}$ is the liquid phase velocity

$\overline{u^l}$ is the averaged displacement in the solid phase,

$\overline{w^s}$ is the mean growth velocity of the solid phase (the averaged velocity of microfronts),

$\overline{w} = \overline{w^s} - \overline{w^l}$ is the averaged fictitious displacement in the liquid phase relative to the solid phase,

T is the temperature.

The system of two-dimensional equations can be written in the form

$$(u_{s1})_t = w_{s1}, \quad (u_{s2})_t = w_{s2}, \tag{82}$$

$$(u_1)_t = w_1, \quad (u_2)_t = w_2, \tag{83}$$

$$\begin{aligned}&\rho(w_{s1})_t + \rho_l(w_1)_t \\ &= [(\lambda(c,T) + 2\mu(c,T) + \alpha^2 M(T))(u_{s1})_x \\ &\quad + (\lambda(c,T) + \alpha^2 M(T))(u_{s2})_y \\ &\quad + \alpha M(T)((u_1)_x + (u_2)_y)]_x \\ &\quad + [\mu(c,T)((u_{s1})_y + (u_{s2})_x)]_y,\end{aligned} \tag{84}$$

$$\begin{aligned}&\rho(w_{s2})_t + \rho_l(w_2)_t - \rho g \\ &= [\mu(c,T)((u_{s1})_y + (u_{s2})_x]_x \\ &\quad + (\lambda(c,T) + \alpha^2 M(T))(u_{s1})_x \\ &\quad + [(\lambda(c,T) + 2\mu(c,T) + \alpha^2 M(T))(u_{s2})_y \\ &\quad + \alpha M(T)((u_1)_x + (u_2)_y)]_y,\end{aligned} \tag{85}$$

$$\begin{aligned}&\rho_l(w_{s1})_t + \rho_{add}\rho_l(w_1)_t = -D(c,T)w_1 \\ &+[M(T)(\alpha((u_{s1})_x + (u_{s2})_y) + (u_1)_x + (u_2)_y)))]_x,\end{aligned} \tag{86}$$

$$\begin{aligned}&\rho_l(w_{s2})_t + \rho_{add}\rho_l(w_2)_t - \rho_l g = -D(c,T)w_2 \\ &+[M(T)(\alpha((u_{s1})_x + (u_{s2})_y) + (u_1)_x + (u_2)_y)))]_y,\end{aligned} \tag{87}$$

$$(\rho_l)_t + (\rho_l w_1)_x + (\rho_l w_2)_y + \rho_s (w_{s1})_x + \rho_s (w_{s2})_y = 0 \quad (88)$$

$$\begin{aligned} & c_t + w_1 \cdot f(c,T)_x + w_2 \cdot f(c,T)_y \\ & = \{M_D(T)[F(c,T) - \varepsilon^2 (F_1^x(c,\overline{u_s},T)c_x)_x \\ & - \varepsilon^2 (F_1^y(c,\overline{u_s},T)c_y)_y + \varepsilon_{el}\frac{\partial}{\partial c}E_{el} + \varepsilon^6 \Delta^{(6)}]_x\}_x \quad (89) \\ & + \{M_D(T)[F(c,T) - \varepsilon^2 (F_1^x(c,\overline{u_s},T)c_x)_x \\ & - \varepsilon^2 (F_1^y(c,\overline{u_s},T)c_y)_y + \varepsilon_{el}\frac{\partial}{\partial c}E_{el} + \varepsilon^6 \Delta^{(6)}]_y\}_y \end{aligned}$$

$$(T + \varkappa c)_t = D_0 (T_{xx} + T_{yy}) \quad (90)$$

where

$$\begin{aligned} E_{el} &= (\lambda(c,T) + \alpha^2 M(T)((u_{s1})_x + (u_{s2})_y)^2 \\ &- \alpha M(T)[(u_1 - u_{s1})_x + (u_2 - u_{s2})_y]((u_{s1})_x + (u_{s2})_y) \\ &+ 2\mu(c,T)[((u_{s1})_x)^2 + \frac{1}{2}((u_{s1})_y + (u_{s2})_x)^2 + ((u_{s2})_y)^2], \end{aligned}$$

$$\Delta^{(6)} = \left(F_2^x\left(c,\overline{u_s},T\right)c_{xx}\right)_{xx} + \left(F_2^y\left(c,\overline{u_s},T\right)c_{yy}\right)_{yy}.$$

The system is considered in the rectangle $\Pi = [0,1]\times[0,2]$. The boundary conditions are specified by numerical experiments. Here, we write out general boundary conditions.

1) The vector-valued functions $\overline{u}$ and $\overline{w}$ satisfy the initial conditions

$$\overline{u}_s(0,x,y) = \overline{u}(0,x,y) = 0,$$

which corresponds to

$$\overline{w}_s(0,x,y) = \overline{w}(0,x,y) = 0.$$

Based on the one-dimensional model, we impose the boundary conditions

$$\overline{u}_s = \overline{w}_s = 0 \quad \text{for } x = 0 \text{ and } y = 0,$$

$$\partial_n \overline{u}_s = \partial_n \overline{w}_s = 0 \quad \text{for } x = 1 \text{ and } y = 2.$$

We assume that the displacements and velocities satisfy the following conditions on all four boundaries:

$$\partial_n \overline{u} = \partial_n \overline{w} = 0.$$

At the same time, it is natural to impose the impermeability condition on all the boundaries. Therefore, the boundary conditions can be modified as follows:

$$\begin{aligned} & u_x = w_x = (u_s)_x = (w_s)_x = \partial_x u_y = \partial_x w_y \\ & = \partial_x u_{sy} = \partial_x w_{sy} = 0 \text{ for } x = 0 \text{ and } x = 1, \end{aligned}$$

$$\begin{aligned} & u_y = w_y = (u_s)_y = (w_s)_y = \partial_y u_x = \partial_y w_x \\ & = \partial_y u_{sx} = \partial_y w_{sx} = 0 \text{ for } y = 0 \text{ and } y = 2 \end{aligned}$$

or, in the other notation of the axes,

$$\begin{aligned} & u_1 = w_1 = u_{s1} = w_{s1} = \partial_x u_2 = \partial_x w_2 = \partial_x u_{s2} \\ & = \partial_x w_{s2} = 0 \text{ for } x = 0 \text{ and } x = 1, \end{aligned}$$

$$\begin{aligned} & u_2 = w_2 = u_{s2} = w_{s2} = \partial_y u_1 = \partial_y w_1 = \partial_y u_{s1} \\ & = \partial_y w_{s1} = 0 \text{ for } y = 0 \text{ and } y = 2. \end{aligned}$$

2) The initial and boundary conditions on z have the form

$$z(0,x,y) = 0, \quad \partial_n z|_{\partial\Pi} = 0.$$

3) The initial conditions on c are as follows:

$$c(0,x,y) = c^-$$

for $(x,y) \in (x_{z1}, x_{z2}) \times (0, y_z)$, $x_{z1} = 1/3$, $x_{z2} = 2/3$, $y_z = 1/3$ and

$$c(o,x,y) = c_{cr}(T_{int}),$$

where $T_{int} = 300$, in the remaining domain. The boundary conditions on C are written as

$$\partial_n c|_{\partial\Pi} = 0, \quad \partial_n \mu^*|_{\partial\Pi} = 0,$$

where

$$\begin{aligned} \mu^* &= F(c,T) - \varepsilon^2 \left(F_1^x\left(c,\overline{u_s},T\right)c_x\right)_x \\ &- \varepsilon^2 \left(F_1^y\left(c,\overline{u_s},T\right)c_y\right)_y + \varepsilon_{el}\frac{\partial}{\partial c}E_{el}, \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial c}E_{el} &= \frac{\partial\lambda}{\partial c}\left((u_{s1})_x + (u_{s2})_y\right)^2 \\ &+ \frac{\partial\mu}{\partial c}\left[2\left((u_{s1})_x\right)^2 + \left((u_{s1})_y + (u_{s2})_x\right)^2 + 2\left((u_{s2})_y\right)^2\right], \end{aligned}$$

$$\frac{\partial}{\partial c}\lambda(c,T) = 20\left(c^+(T) + c^-(T)\right),$$

$$\frac{\partial}{\partial c}\mu(c,T) = \left(c^+(T) + c^-(T)\right),$$

i.e., for $x = 0$ and $x = 1$ the second condition takes the form

$$\begin{aligned} & \partial_x F(c,T) - \varepsilon^2 vis\, c_{xxx} - \varepsilon^2 c_{xyy} \\ & + \varepsilon_{el}\partial_x\left[\frac{\partial\lambda}{\partial c}\left((u_{s1})_x + (u_{s2})_y\right)^2\right] \\ & + \varepsilon_{el}\frac{\partial\mu}{\partial c}\left[2\left((u_{s1})_x\right)^2 + \left((u_{s1})_y + (u_{s2})_x\right)^2 + 2\left((u_{s2})_y\right)^2\right] \\ & = 0, \end{aligned}$$

where

$$\begin{aligned} \partial_x F(c,T) &= \left(\partial_x c - \frac{dc^-}{dT}\partial_x T\right)(c - c_{cr})(c - c^+) \\ &+ (c - c^-)\left(\partial_x c - \frac{dc_{cr}}{dT}\partial_x T\right)(c - c^+) \\ &+ (c - c^-)(c - c_{cr})\left(\partial_x c - \frac{dc^+}{dT}\partial_x T\right) \quad \text{for } T \le T_0 \end{aligned}$$

and

$$\partial_x F(c,T) = 3(c-c_{cr})^2\left(\partial_x c - \frac{dc_{cr}}{dT}\partial_x T\right) \text{ for } T > T_0;$$

$$\frac{dc^{\pm}}{dT} = m\frac{1}{\sqrt{\beta_{clust}^2 - 4\alpha_{clust}(\gamma_{clust}-T)}},$$

$$\frac{dc_{cr}}{dT} = \frac{c_0 = c_{mid}}{T_0 - T_{\min}}$$

Similarly, for $y = 0$ and $y = 2$ the second boundary condition on c takes the form

$$\partial_y \mu^* = 0.$$

4) For the temperature T we impose the initial conditions

$$T(0,x,y) = T_{\min} + (T_{\max} - T_{\min})\, y / y_{end}\,,\ y_{end} = 2$$

and the boundary conditions

$$T(t,0,y) = T(t,1,y) = T_{\min} + (T_{\max} - T_{\min})/(\theta t + 1),$$

$$\partial_n T\,|_{y=0,\,y=y_{end}} = 0,$$

where $T_{\min} = 233.15K$, $T_{\max} = 456.15K$, $\theta = 100$ is a parameter.

4.6. Numerical Analysis of the Model of Oriented Crystallization, Two-Dimensional Case

N. A. Zaitsev, Yu. G. Rykov, and V. Lysov, based on the methods of [16,18], performed a numerical analysis of the model. The numerical results are reproduced here under their kind permission.

Figures 4(a-d) and **5(a-d)** illustrate the numerical results and show a complicated dynamics of the crystallization process.

To test the crystallization model (82)-(89), the following dimensionless values of the main parameters were taken on the basis of their physical sense:

$$\rho = 6,73; \quad R = 0,74; \quad q = 0,25;$$

$$Q = 1,75; \quad M = 6,2; \quad \alpha = 1\times10^{-3};$$

$$g = 0; \quad \rho^s = 1; \quad D_0 = 1; \quad \varkappa = 7,2; \quad \varepsilon = 1\times10^{-4};$$

$$F_2(x,y) \equiv 0; \quad F_1(x,y) = 1\times10^3.$$

Time-development of crystallization process in the case of isotropic surface tension of crystal grains. The banding structure is transformed to the equiaxial structure. (a) $t = 4$, (b) $t = 8$, (c) $t = 18$, (d) $t = 22$ (**Figure 4**).

Figures 4(a-d) presents the situation where the surface tension of crystal grains is isotropic. In this case, the chemical potential has the form

$$\mu^* = F(c,T) + \varepsilon^2 \Delta c. \qquad (91)$$

The banding structure is formed at initial times and

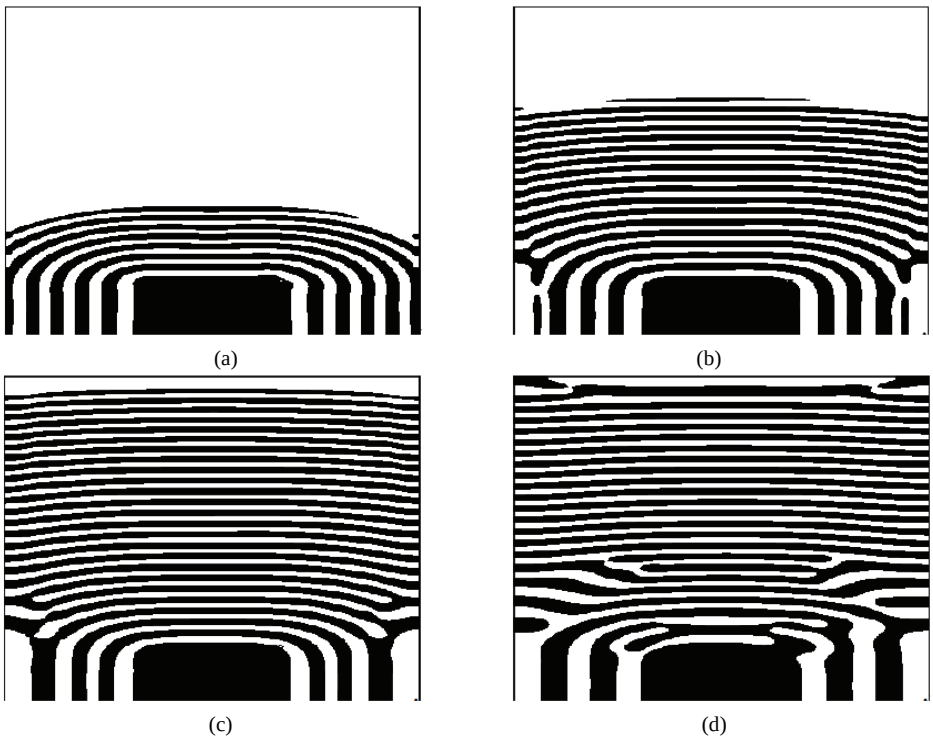


Figure 4. Time-development of crystallization process in the case of isotropic surface tension of crystal grains. The banding structure is transformed to the equiaxial structure: (a) t = 4; (b) t = 8; (c) t = 18; (d) t = 22.

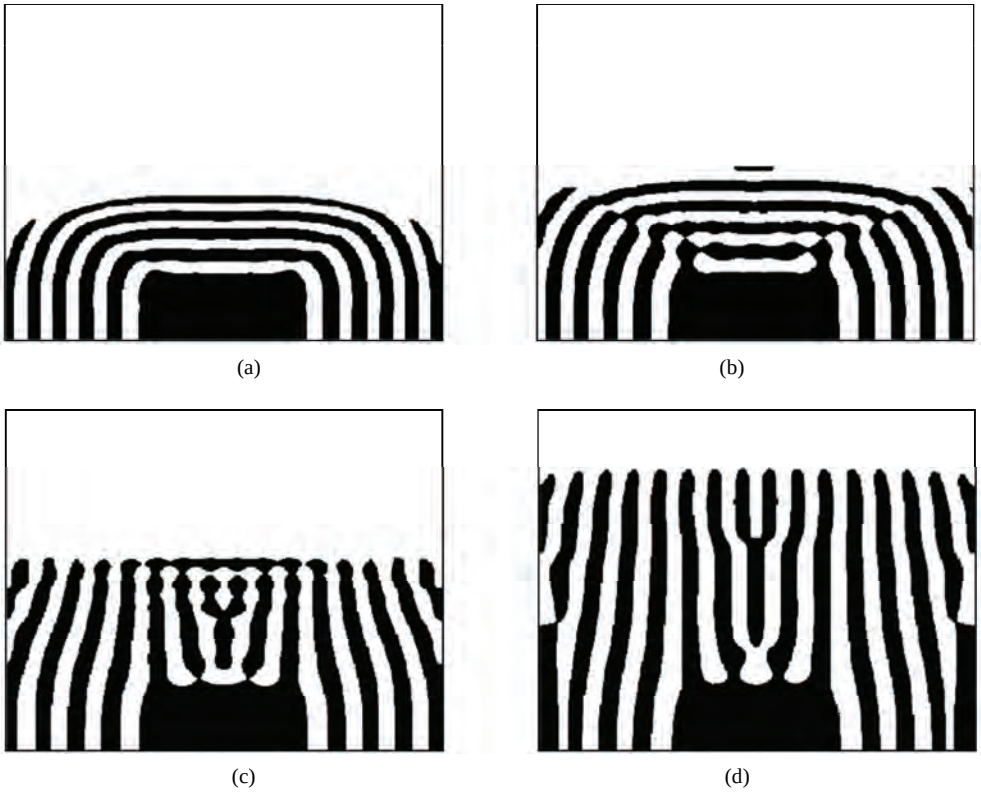


Figure 5. Time-development of crystallization process in the case of anisotropic surface tension of crystal grains (formation of a dendrite liquid-solid damain): (a) t = 1; (b) t = 2; (c) t = 4; (d) t = 8.

then is developed to an equiaxial like structure. There is a certain analogy with the crystallization of eutectics when one of the phases splits into small cells [19] or the dendrite growth [20] when the distance between secondary branches of dendrites in a perfectly solidified mass is much larger than that at initial times. We also can imagine a similar situation where a jet breaks down into drops when the absolute value of the surface tension is rather large.

Time-development of crystallization process in the case of anisotropic surface tension of crystal grains (formation of a dendrite liquid-solid domain): (a) $t = 1$, (b) $t = 2$, (c) $t = 4$, (d) $t = 8$ (**Figure 5**).

Figures 5(a-d) illustrates the numerical experiment in the case of an anisotropic surface tension, In this case, the following formula is used instead of (91):

$$\mu^* = F(c,T) + \varepsilon^2 \nabla_x \left[\left(A^- + \left(A^+ - A^- \right) \frac{c - c^-}{c^+ - c^-} \right) \nabla_x c \right],$$

$$A^+ = 1,5E; \quad A^- = \begin{pmatrix} 1 & 0 \\ 0 & 14 \end{pmatrix};$$

where E is the identity matrix. In this case, the banding structure, deformed because of overcrystallization, is developed to the dendrite structure.

4.7. Conclusions

The aforesaid shows that the proposed mathematical model of crystallization can be viewed as a mathematical reconstruction of various experiments. In particular, the following result of numerical analysis agrees with experimental observations: it is seen in **Figures 4(a), (b)** and **Figures 5(a), (b)** that the banding structure is the first structure formation to appear in the instability zone and the subsequent reformation of structure proceeds because of arising waves similar to the Marangoni instability wave at the boundaries of bands (cf. [2,4,5,21,22]). We also note that the numerical results concerning the influence of the crystallographic orientation of a growing crystal on the structure formation in alloys (cf., for example, [19]) can be regarded as confirmation of the verifiability of our mathematical crystallization model.

5. Acknowledgements

The work was financially supported by the Russian Foundation for Basic Research (grants No. 09-01-12024

and 09-01-00288).

6. References

[1] V. A. Avetisov, "*p*-Adic Description of Characteristic Relaxation in Complex System," *Journal of Physics A*: *Mathematical and General*, Vol. 36, No. 15, 2003, pp. 4239-4246.

[2] E. N. Kablov, "Cast Gas-Turbine Engine Blades (Alloys, Technology, Covering) [in Russian]," Moscow, 2001.

[3] A. N. Kolmogorov, "On the Statistical Theory of Crystallization of Metals [in Russian]," *Izv. Akad. Nauk SSSR, Ser. Mat.*, No. 3, 1937, pp. 355-359.

[4] F. M. Shemyakin and P. F. Mikhalev, "Physic-Chemical Periodic Processes [in Russian]," *Akad. Nauk SSSR*, Moscow, 1938.

[5] I. Z. Bezbakh, B. G. Zakharov and I. A. Prokhorov, "Radiographical Characterization of Microsegregation in Crystals [in Russian]," In: *Proceedings of the* 6*th International Conference "Growth of Monocrystals and Heat-Mass Transfer*", Obninsk, Vol. 2, 2005, pp. 352-361

[6] V. G. Danilov, G. A. Omel'yanov and E. V. Radkevich, "Hugoniot-Type Conditions and Weak Solutions to the phase Field System," *European Journal of Applied Mathematics*, Vol. 10, No. 1, 1999, pp. 55-77.

[7] E. V. Radkevich, "Mathematical Aspects of Nonequilibrium Processes [in Russian]," Tamara Rozhkovskaya Publisher, Novosibirsk, 2007.

[8] N. N. Yakovlev, E. A. Lukashev and E. V. Radkevich, "Problems of Reconstruction of the Process of Directional Solidification [in Russian]," *Dokl. Akad. Nauk, Ross. Akad. Nauk*, Vol. 421, No. 5, 2008, pp. 625-629; English translation: *Doklady Physics, Technical Physics*, Vol. 53, No. 8, 2008, pp. 442-446.

[9] V. Visintin, "Models of Phase Transitions," Birkhauser, Boston, 1996.

[10] E. V. Radkevich, "The Gibbs--Thomson Effect and Existence Conditions of Classical Solution for the Modified Stefan Problem," In: *Free Boundary Problems Involving Solids*, Science and Technology, Harlow, 1993, pp. 135-142.

[11] O. A. Oleynik and E. V. Radkevich, "Second Order Equations with Nonnegative Characteristic Form," *Am. Math. Soc., Providence*, 1973.

[12] A. A. Lacey and A. B. Tayler, "A Mushy Region in a Stefan Problem," *IMA Journal of Applied Mathematics*, Vol. 30, No. 3, 1983, pp. 303-313.

[13] M. A. Biot, "Mechanics of Deformation and Acoustic Propagation in Porous Media," *Journal of Applied Physics*, Vol. 33, No. 4, 1962, pp. 1482-1498.

[14] S. J. Watson, F. Otto, B. Y. Rubinstein and S. H. Davis, "Coarsening Dynamics for the Convective Cahn-Hilliard Equation," University of Bonn, Preprint, 2003.

[15] A. A. Golovin, S. H. Davis and A. A. Nepomnyashchy, "A Convective Cahn-Hilliard Model for the Formation of Facets and Carners in Crystal Growth," *Physical D*, Vol. 118, 1998, pp. 202-230.

[16] N. A. Zaitsev and Yu. G. Rykov, "Numerical Analysis of a Model Describing Metal Crystallization I. One-Dimensional Case," Preprint, Keldysh Institute of Applied Physics RAS, No. 72, 2007.

[17] W. Dreyer and B. Wagner, "Sharp-Interface Model for Eutectic Alloys. Part I. Concentration Dependent Surface Tension," Preprint, 2003.

[18] N. A. Zaitsev and Y. G. Rykov, "Numerical Analysis to a New Model of the Matel Solidification, 1-D Case," *Mathematical Simulation*, 2010, to Appear.

[19] N. P. Lyakishev and G. S. Burkhanov, "Metal Monocrystals [in Russian]," Eliz, Moscow, 2002.

[20] P. E. Shalin, *et al.*, "Monocrystals of Heat-Resistance Nickel Alloys [in Russian]," Mashinostroenie, Moscow, 1997.

[21] J. W. Matthews and A. E. Blacslee, "Defects in Epitaxial Multilayers. I. Misfit Dislocations," *Journal of Crystal Growth*, Vol. 27, No. 1, 1974, pp. 118-125.

[22] V. N. Vigdorovich, A. E. Vol'pyan and G. M. Kurdyumov, "Oriented Crystallization and Physic-Chemical Analysis [in Russian]," Chemistry, Moscow, 1976.

Applied Mathematics, **2010, 1, 179-188**
doi:10.4236/am.2010.13022 Published Online September 2010 (http://www.SciRP.org/journal/am)

Scientific Research

A Retrospective Filter Trust Region Algorithm for Unconstrained Optimization*

Yue Lu, Zhongwen Chen
School of Mathematics Science, Suzhou University, Suzhou, China
E-mail: *yueyue*403@gmail.com, *zwchen@suda.edu.cn*
Received June 8, 2010; *revised July* 18, 2010; *accepted July* 21, 2010

Abstract

In this paper, we propose a retrospective filter trust region algorithm for unconstrained optimization, which is based on the framework of the retrospective trust region method and associated with the technique of the multi-dimensional filter. The new algorithm gives a good estimation of trust region radius, relaxes the condition of accepting a trial step for the usual trust region methods. Under reasonable assumptions, we analyze the global convergence of the new method and report the preliminary results of numerical tests. We compare the results with those of the basic trust region algorithm, the filter trust region algorithm and the retrospective trust region algorithm, which shows the effectiveness of the new algorithm.

Keywords: Unconstrained Optimization, Retrospective, Trust Region Method, Multi-Dimensional Filter Technique

1. Introduction

Consider the following unconstrained optimization problem

$$\min \ f(x) \tag{1}$$

where $x \in R^n$, $f: R^n \to R$ is a twice continuously differentiable function.

The trust region method for unconstrained optimization is first presented by Powell [1], which, in some sense, is equivalent to the Levenberg-Marquardt method which is used to solve the least square problems and which was given by Levenberg [2] and Marquardt [3]. The basic idea of trust region methods works as follows. In the neighborhood of the current iterate (which is called the trust region), we define a model function that approximates the objective function in the trust region and compute a trial step within the trust region for which we obtain a sufficient model decrease. Then we compare the achieved reduction in $f(x)$ to the predicted reduction in the model for the trial step. If the ratio of achieved versus predicted reduction is sufficiently positive, we define our next guess to be the trial point. If this ratio is not sufficiently positive, we decrease the trust region radius in order to make the trust region smaller. Otherwise, we may increase it or possibly keep it unchanged. Since the trust region method is of the naturalness, the strong convergence and robustness, it has been concerned by many people, such as Powell [1,4,5], Schultz *et al.* [6], Sorensen [7], Moŕe [8], Yuan [9] and so on. In recent years, the trust region method has been applied to the optimization problems with equality constraints [10], simple bound constraints [11], convex constraints [12] and so on. Many of convergence results have been obtained, which can be seen in [13].

In Fletcher and Leyffer [14] a new technique for globalizing methods for nonlinear programming (NLP) is presented. The idea is referred to as an NLP filter, motivated by the aim of avoiding the need to choose penalty parameters, and considered the relationship between the objective function and the constraint violation in the view of multi-objective optimization. They make the values of the objective function and the constraint violation to be a pair (which is called the filter), construct a sophisticated filter mechanism by comparing the relationship between the pairs, and control the algorithm to converge to the stationary point of the problem (1). The results of numerical tests show that the filter methods are very effective. Fletcher *et al.* [14,15], Toint *et al.* [16], Ulbrich *et al.* [17], Wächter *et al.* [18,19] have combined the idea with SQP method, trust region method, interior-point method, line search methods, respectively, and obtained some interesting results about the filter method.

*This work was supported by Chinese NSF Grant 60873116.

Fletcher, Leyffer and Toint [20] review the ideas above and mention the application of the filter method in practice. In [14], they study the problem of the following form

$$\min \quad f(x),$$
$$s.t. \quad c(x) \le 0,$$

where $c: R^n \to R^m$ is continuously differentiable function. Define the measure of the constraint violation

$$h(c(x)) = \sum_{j=1}^{m} \max(0, c_j(x)).$$

We use $(f^{(k)}, h^{(k)})$ to denote values of $f(x)$ and $h(c(x))$ evaluated at x_k. Now, we give the following definitions about the filter methods.

Definition 1.1 A pair $(f^{(k)}, h^{(k)})$ obtained on iteration k is said to dominate another pair $(f^{(l)}, h^{(l)})$ if and only if

$$f^{(k)} \le f^{(l)} \text{ and } h^{(k)} \le h^{(l)}.$$

Definition 1.2 A filter is a list of pairs $(f^{(l)}, h^{(l)})$ such that no pair dominates any other.

We use F_k to denote the set of iteration indices $j(j \le k)$ such that $(f^{(j)}, h^{(j)})$ is an entry in the current filter.

Definition 1.3 A pair $(f^{(l)}, h^{(l)})$ is said to be acceptable for inclusion in the filter if it is not dominated by any pair in the filter, that is, for any pair $(f^{(l)}, h^{(l)}) \in F_k$, $(f^{(k)}, h^{(k)})$ satisfies

$$f^{(k)} \le f^{(l)} \text{ or } h^{(k)} \le h^{(l)} \tag{2}$$

In order to obtain the global convergence of the algorithm, we should make *f*, *h* satisfy the sufficient reduction condition, so we strengthen the acceptable rule (2) as

$$f^{(k)} \le f^{(l)} - \gamma_h h^{(l)} \text{ or } h^{(k)} \le (1-\gamma_h) h^{(l)} \tag{3}$$

where $\gamma_h \in (0,1)$. When $\gamma_h \in (0,1)$ is very small, there is negligible difference in practice between (2) and (3).

Definition 1.4 When the pair $(f^{(k)}, h^{(k)})$ is added to the list of pairs in the filter, any pairs in the filter that are dominated by the new pair are removed, that is, we remove the pair $(f^{(l)}, h^{(l)}) \in F_k$ which satisfies

$$f^{(k)} < f^{(l)} \text{ and } h^{(k)} < h^{(l)}.$$

This is called the modification of the filter.

Gould *et al.* [21] and Miao *et al.* [22] applies the filter technique to unconstrained optimization, whose characteristic is to relax the condition of accepting a trial step for the usual trust region method, which improves the effectiveness of the algorithm in some sense. The nonmonotonic algorithm also has the algorithm in some sense. The nonmonotonic algorithm also has the characteristic [23,24].

Recently, Bastin *et al.* [25] presents a retrospective trust region method for unconstrained optimization. Comparing their algorithm with the basic trust region algorithm, the updating way of the trust region radius is different, and the retrospective ratio

$$\tilde{\rho}_{k+1} = \frac{f(x_{k+1}) - f(x_{k+1} - s_k)}{m_{k+1}(x_{k+1}) - m_{k+1}(x_{k+1} - s_k)} = \frac{f(x_k) - f(x_k + s_k)}{m_{k+1}(x_k) - m_{k+1}(x_k + s_k)}$$

is mentioned, where $m_k(x_k + s)$ is the approximated quadratic model of the objective function $f(x)$ at x_k. s_k is a solution of the following trust region subproblem

$$\begin{aligned} &\min \ m_k(x_k + s), \\ &s.t. \ \|s\| \le \Delta_k. \end{aligned} \tag{4}$$

Δ_k is the trust region radius at the current iterate point. In the basic trust region algorithm, the ratio ρ_k

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}$$

plays the following two roles.

1) Determine the trial step to be accepted by the algorithm or not.

2) Adjust the trust region radius correspondingly.

In the retrospective trust region method, the two roles are played by the ratio ρ and $\tilde{\rho}$, respectively. In the basic trust region algorithm, the determination of trust region radius is an important and difficult problem. Sartenaer [26] and Zhang *et al.* [27] present the self-adaptive trust region methods and give some discussions about the determination of trust region radius. Bastin *et al.* [25] presents a retrospective trust region method for unconstrained optimization. The retrospective ratio in this method uses the information at the current iterate and the last iterate point, which can give the more effective estimation of trust region radius. Hence, the number of solving trust region subproblem may be decreased, which improves the effectiveness of the method.

In this paper we present a new algorithm for unconstrained optimization, which is based on the framework of the retrospective trust region method [25] and associated with the technique of the multi-dimensional filter [21,22]. Under reasonable assumptions, we analyze the global convergence of the new method and report the preliminary results of numerical tests. We compare the results with those of the basic trust region algorithm, the filter trust region algorithm and the retrospective trust region algorithm, which shows the effectiveness of the

new algorithm. This paper is organized as follows. The new algorithm is described in Section 2. Basic assumptions and some lemmas are given in Section 3. The analysis of the first order and second order convergence is given in Section 4 and Section 5, respectively. Section 6 reports the numerical results. Finally, we give some final remarks on this approach.

2. Algorithm

In this paper, we define $g(x)=\nabla_x f(x)$. At the current iterate x_k, $f_k = f(x_k)$, $g_k = g(x_k)$, g_{ki} denotes the ith component of the vector g_k. Throughout this paper, $\|\cdot\|$ denotes the Euclidean norm. Now, we review the basic-trust region algorithm as follows.

2.1. Algorithm BTR (Basic Trust Region Algorithm)

Step 0. (Initialization) Given an initial point $x_0 \in R^n$ and an initial trust-region radius $\Delta_0 > 0$. $0 < \eta_1 \le \eta_2 < 1$ and $0 < \gamma_1 \le \gamma_2 < 1$. Set $k := 0$.

Step 1. (Model definition) Define a model function m_k in $\Im_k$, where $\Im_k = \{x \in R^n \mid \|x - x_k\| \le \Delta_k\}$.

Step 2. (Step calculation) Compute a trial step s_k for solving trust region subproblem (4).

Step 3. (Updating iterate point) Compute $f(x_k + s_k)$ and the ratio

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}.$$

then,

$$x_{k+1} = \begin{cases} x_k + s_k, & \text{if } \rho_k \ge \eta_1, \\ x_k, & \text{if } \rho_k < \eta_1. \end{cases}$$

Step 4. (Updating trust-region radius) Set

$$\Delta_{k+1} \in \begin{cases} [\Delta_k, \infty), & \text{if } \rho_k > \eta_2, \\ [\gamma_2 \Delta_k, \Delta_k), & \text{if } \rho_k \in [\eta_1, \eta_2], \\ [\gamma_1 \Delta_k, \gamma_2 \Delta_k), & \text{if } \rho_k < \eta_1. \end{cases}$$

Set $k := k + 1$, and go to Step 1.

In the algorithm BTR, we do not give a formal stopping criterion. In practice, the stopping criterion can be installed in Step 1, such as

$$\|g_k\| \le eps \text{ or } k > k_{\max},$$

where eps denotes the precision, and $k_{\max}$ denotes the maximal number of iterations.

If x^* is a local minimizer of the problem (1), then $g(x^*) = 0$. Motivated by the filter method, we set $\|g(x)\|$ to be the measure of the iterate. Now we introduce some definitions about the multi-dimensional filter.

Definition 2.1 We say that a point x_k dominates another point x_l, if

$$|g_{kj}| \le |g_{lj}| \text{ for all } j = 1, 2, \cdots n.$$

Definition 2.2 A multi-dimensional filter F is a list of n-tuples of the form $(g_{k1}, \cdots, g_{kn})$ and if $g_k, g_l \in F$, then there exist two indices $j_1, j_2 \in \{1, 2, \cdots n\}$, $j_1 \ne j_2$ such that

$$|g_{lj_1}| < |g_{kj_1}| \text{ and } |g_{kj_2}| < |g_{lj_2}|.$$

Definition 2.3 The iterate point x_k is acceptable for the filter F if and only if for all $g_l \in F$, there exists $j \in \{1, 2, \cdots n\}$ such that

$$|g_{kj}| \le |g_{lj}| - \gamma_g \|g_l\|, \ \gamma_g \in (0, 1/\sqrt{n}).$$

Definition 2.4 If the iterate point x_k is acceptable, then it is added to the filter and any iterates in the filter that are dominated by the new iterate are removed, which is called the modification of the filter.

Combining with the filter technique and the retrospective idea, we describe our algorithm as follows.

2.2. Algorithm RFTR (Retrospective Filter Trust-Region Algorithm)

Step 0. (Initialization) An initial point $x_0 \in R^n$ and an initial trust-region radius $\Delta_0 > 0$ are given.

$$\gamma_g \in (0, 1/\sqrt{n}), \ 0 < \eta_1 < 1, 0 < \tilde{\eta}_1 \le \tilde{\eta}_2 < 1,$$
$$0 < \gamma_1 \le \gamma_2 < 1 \le \gamma_3.$$

Set the initial filter F to be the empty set and set $k := 0$.

Step 1. (Model definition) Define a model function m_k in $\Im_k$, where $\Im_k = \{x \in R^n \mid \|x - x_k\| \le \Delta_k\}$.

Step 2. (Updating retrospective trust-region radius) If $k = 0$, then go to Step 3. If $x_k = x_{k-1}$, then choose $\Delta_k \in [\gamma_1 \Delta_{k-1}, \gamma_2 \Delta_{k-1})$. Otherwise, compute the retrospective ratio

$$\tilde{\rho}_k = \frac{f(x_{k-1}) - f(x_k)}{m_k(x_{k-1}) - m_k(x_k)}.$$

Choose

$$\Delta_k \in \begin{cases} [\gamma_1 \Delta_{k-1}, \gamma_2 \Delta_{k-1}), & \text{if } \tilde{\rho}_k < \tilde{\eta}_1, \\ [\gamma_2 \Delta_{k-1}, \Delta_{k-1}), & \text{if } \tilde{\rho}_k \in [\tilde{\eta}_1, \tilde{\eta}_2], \\ [\Delta_{k-1}, \gamma_3 \Delta_{k-1}), & \text{if } \tilde{\rho}_k > \tilde{\eta}_2. \end{cases}$$

Step 3. (Step calculation) Compute a trial step s_k for solving trust region subproblem (4) and $x_k^+ = x_k + s_k$.

Step 4. (Updating iterate point) Compute

$$\rho_k = \frac{f(x_k) - f(x_k^+)}{m_k(x_k) - m_k(x_k^+)}.$$

Case 1: If $\rho_k \geq \eta_1$, then $x_{k+1} = x_k^+$.

Case 2: If $\rho_k < \eta_1$ and x_k^+ is accepted by the filter F, then $x_{k+1} = x_k^+$ and add $g_k^+ = g(x_k^+)$ into the filter F.

Case 3: If $\rho_k < \eta_1$ and x_k^+ is not accepted by the filter F, then $x_{k+1} = x_k$.

Step 5. Set $k := k+1$, go to Step 1.

Similar to the algorithm BTR, the stopping criterion can be installed in Step 1, such as

$$\|g_k\| \leq eps \text{ or } k > k_{\max},$$

where eps denotes the precision, and $k_{\max}$ denotes the maximal number of iterations. The Hessian matrix in the model function m_k can be obtained by BFGS updating formula or set $\nabla_{xx} m_k(x_k + s) = \nabla_{xx} f(x_k)$ for all s such that $x_k + s \in \Im_k$.

In the algorithm RFTR, the retrospective idea and the filter technique are two important characteristics. The retrospective ratio uses the information at the current iterate and the last iterate to adjust the trust-region radius, which can give the more effective estimation of trust region radius. The filter technique relaxes the condition of accepting a trial step comparing with the usual trust region method, which improves the effectiveness of the algorithm in some sense. From the algorithm RFTR, if the trial point is not accepted (Case 3 in Step 5 occurs), then the algorithm is similar to the basic trust-region algorithm, whose difference is just that we use the retrospective idea in the algorithm RFTR. However, if the trial point is accepted by the algorithm (Case 1 or Case 2 in Step 5 occurs), the retrospective idea and the filter technique all play the roles.

At the iterate x_k, if $x_{k+1} = x_k + s_k \neq x_k$, then the iterate is called the successful iterate and the iteration index x_k is called successful iteration.

3. Basic Assumptions and Lemmas

In this section, we present the global convergence analysis of the algorithm RFTR. We make the following assumptions.

A1 The all iterates x_k remain in a closed and bounded convex set Ω.

A2 $f: R^n \to R$ is a twice continuously differentiable function.

A3 The model function m_k is first-order coherent with the function f at the iterate x_k, *i.e.*, their values and gradients are equal at x_k for all k,

$$m_k(x_k) = f(x_k),\ \nabla_x m_k(x_k) = \nabla_x f(x_k).$$

A4 The Hessian matrix of the model function $\nabla_{xx} m_k$ is uniformly bounded, *i.e.*, there exists a constant $k_{umh} \geq 1$ such that

$$\|\nabla_{xx} m_k(x)\| \leq k_{umh} - 1,$$

holds for all $x \in R^n$ and all k.

Generally speaking, we do not need the global solution of the trust region subproblem. We only expect to decrease the model at least as much as at the Cauchy point. Therefore, we make the following assumption on the solution s_k of the trust region subproblem (4).

A5 There exists a constant k_{mdc}, for all k,

$$m_k(x_k) - m_k(x_k + s_k) \geq k_{mdc} \max\left\{\|g_k\| \min\left\{\frac{\|g_k\|}{\beta_k}, \Delta_k\right\}, |\tau_k| \min\left\{\tau_k^2, \Delta_k^2\right\}\right\},$$

where

$$\beta_k = 1 + \max_{x \in \Im_k} \|\nabla_{xx} m_k\|,$$
$$\Im_k = \left\{x \in R^n \mid \|x - x_k\| \leq \Delta_k\right\},$$
$$\tau_k = \min\left\{0, \lambda_{\min}(\nabla_{xx} m_k(x_k))\right\}.$$

By the assumptions A1 and A2, the Hessian matrix $\nabla_{xx} f(x)$ is uniformly bounded on Ω, *i.e.*, there exists a positive constant k_{ufh} such that, for all $x \in \Omega$,

$$\|\nabla_{xx} f(x)\| \leq k_{ufh}.$$

Now we study the global convergence of the algorithm RFTR. First we give a bound on the difference between the objective function and the model function m_k at the iterate x_{k-1} and x_k. The proof of the following result is similar to Theorem 3.1 in [25].

Lemma 3.1 [25] Suppose A1-A4 hold, then exists a positive constant k_{ubh},

$$|f(x_k) - m_{k-1}(x_{k-1})| \leq k_{ubh} \Delta_{k-1}^2 \tag{5}$$

and if iteration $k-1$ is successful, that

$$|f(x_{k-1}) - m_k(x_{k-1})| \leq k_{ubh} \Delta_{k-1}^2 \tag{6}$$

where $k_{ubh} = \max\{k_{ufh}, k_{umh}\}$.

As the retrospective ratio $\tilde{\rho}_k$ uses the reduction in m_k instead of the reduction in m_{k-1}, we need to compare their difference, which is provided by the next Lemma.

Lemma 3.2 [25] Suppose A1-A4 hold, then for every successful iteration $k-1$,

$$\left|\left(m_{k-1}(x_{k-1}) - m_{k-1}(x_k)\right) - \left(m_k(x_{k-1}) - m_k(x_k)\right)\right| \leq 2k_{ubh} \Delta_{k-1}^2$$

We conclude from this result that the denominators in the expression of $\tilde{\rho}_k$ and ρ_k are the same order as the error between the objective function and the model function. Similar to Theorem 6.4.2 in [13], we obtain the next result.

Lemma 3.3 Suppose A1-A5 hold, $g_{k-1} \neq 0$ and Δ_{k-1} satisfies that

$$\Delta_{k-1} \leq \min\left\{1-\eta_1, \frac{1-\tilde{\eta}_2}{3-2\tilde{\eta}_2}\right\} \frac{k_{mdc}}{k_{ubh}} \left\|g_{k-1}\right\| \tag{7}$$

Then iteration $k-1$ is successful and

$$\Delta_k \geq \Delta_{k-1}.$$

Proof. It follows from $k_{mdc}, \eta_1 \in (0,1)$ and the assumptions A3 and A5 that $\beta_{k-1} \leq k_{ubh}$. By (7),

$$\Delta_{k-1} < \frac{\left\|g_{k-1}\right\|}{\beta_{k-1}}.$$

By the assumption A5, we have that

$$\begin{aligned} m_{k-1}\left(x_{k-1}\right) - m_{k-1}\left(x_k\right) &\geq k_{mdc}\left\|g_{k-1}\right\| \min\left\{\frac{\left\|g_{k-1}\right\|}{\beta_{k-1}}, \Delta_{k-1}\right\} \\ &\geq k_{mdc}\left\|g_{k-1}\right\| \Delta_{k-1} \end{aligned} \tag{8}$$

On the other hand, it follows from (5) and (7) that

$$\begin{aligned} \left|\rho_{k-1} - 1\right| &= \left|\frac{f\left(x_k\right) - m_{k-1}\left(x_k\right)}{m_{k-1}\left(x_{k-1}\right) - m_{k-1}\left(x_k\right)}\right| \\ &\leq \frac{k_{ubh}}{k_{mdc}\left\|g_{k-1}\right\|} \Delta_{k-1} \leq 1-\eta_1. \end{aligned}$$

Thus, $\rho_{k-1} \geq \eta_1$, *i.e.*, the iteration $k-1$ is successful. Next we proof the second part of the conclusion.

By $k_{mdc}, \tilde{\eta}_2 \in (0,1)$, we have

$$\frac{1-\tilde{\eta}_2}{3-2\tilde{\eta}_2} < \frac{1}{2} \text{ and } k_{mdc}\frac{1-\tilde{\eta}_2}{3-2\tilde{\eta}_2} < 1 \tag{9}$$

The conditions (7), (9) and the definition of β_{k-1} in the assumption A5 imply that

$$\Delta_{k-1} < \frac{1}{2}\frac{k_{mdc}}{k_{ubh}}\left\|g_{k-1}\right\| \text{ and } \Delta_{k-1} < \frac{\left\|g_{k-1}\right\|}{\beta_{k-1}}.$$

Combining (8) and Lemma 3.2, we can conclude that

$$\begin{aligned} \left|m_k\left(x_{k-1}\right) - m_k\left(x_k\right)\right| &\geq \left|m_{k-1}\left(x_{k-1}\right) - m_{k-1}\left(x_k\right)\right| - 2k_{ubh}\Delta_{k-1}^2 \\ &\geq k_{mdc}\left\|g_{k-1}\right\|\Delta_{k-1} - 2k_{ubh}\Delta_{k-1}^2 \end{aligned} \tag{10}$$

By (6) and (10),

$$\left|\tilde{\rho}_k - 1\right| = \left|\frac{f\left(x_{k-1}\right) - m_k\left(x_{k-1}\right)}{m_k\left(x_{k-1}\right) - m_k\left(x_k\right)}\right| \leq \frac{k_{ubh}\Delta_{k-1}}{k_{mdc}\left\|g_{k-1}\right\| - 2k_{ubh}\Delta_{k-1}}.$$

(7) implies that

$$\left(3-2\tilde{\eta}_2\right)k_{ubh}\Delta_{k-1} \leq \left(1-\tilde{\eta}_2\right)k_{mdc}\left\|g_{k-1}\right\|.$$

i.e.,

$$k_{ubh}\Delta_{k-1} \leq \left(1-\tilde{\eta}_2\right)\left(k_{mdc}\left\|g_{k-1}\right\| - 2k_{ubh}\Delta_{k-1}\right).$$

Therefore, $\left|\tilde{\rho}_k - 1\right| \leq \left(1-\tilde{\eta}_2\right)$, *i.e.*, $\tilde{\rho}_k \geq \tilde{\eta}_2$.

As a consequence of this property, we may now prove that the trust region radius cannot become too small as long as a first-order critical point is not approached. The technique of the proof is similar to Theorem 3.4 in [25] and Theorem 6.4.3 in [13].

Lemma 3.4 [13,25] Suppose A1-A5 hold. Suppose, furthermore, that there exists a constant $k_{lbg} > 0$ such that $\left\|g_k\right\| \geq k_{lbg}$ for all k. Then there is a constant k_{lbd} such that $\Delta_k \geq k_{lbd}$ for all k.

4. First Order Convergence

Assume that $\{x_k\}$ is an infinite sequence generated by Algorithm RFTR. Under the assumptions (A1)-(A5), we discuss the first order convergence of the sequence $\{x_k\}$. At first, we define the following sets.

The set of successful iteration index

$$S = \left\{k \mid x_{k+1} = x_k + s_k\right\}.$$

The set of the iteration index which is added to the filter

$$A = \left\{k \mid g_k^+ \text{ or the iterate } x_k^+ = x_k + s_k \text{ is added to the filter } F\right\}.$$

The set of the iteration index which satisfies sufficient descent condition

$$D = \left\{k \mid \rho_k \geq \eta_1\right\}.$$

$|S|$ denotes the cardinal number of the set *S*. We now establish the criticality of the limit point of the sequence of the iterates when there are only finitely many successful iterations.

Theorem 4.1 Suppose A1-A5 hold and $|S| < +\infty$, then there exists an index k_0 such that $x_{k_0} \equiv x^*$ and x^* is a first-order critical point.

Proof. Let k_0 be the last successful iteration index, then

$$x_{k_0} = x_{k_0+j} \text{ and } \rho_{k_0+j} < \eta_1, \ \forall j = 1, 2, \cdots.$$

From Step 2 of the algorithm RFTR, we have that

$$\Delta_{k_0+j} \le \gamma_2 \Delta_{k_0+j-1} \le \cdots \le \gamma_2^j \Delta_{k_0}.$$

Thus,

$$\lim_{j\to\infty} \Delta_{k_0+j} = 0.$$

It follows from Lemma 3.4 that $\|g_{k_0}\| = 0$.

Next, we consider the case that there are infinitely many successful iteration. From the algorithm RFTR, we know that $A \subseteq S$. Therefore we consider the following two cases.

1) There are infinitely many filter iterations, *i.e.*, $|A| = +\infty$.

2) There are finitely many filter iterations, *i.e.*, $|A| < +\infty$.

First, we have the following result.

Theorem 4.2 Suppose A1-A5 hold and $|A| = |S| = +\infty$, then

$$\liminf_{k\to\infty} \|g_k\| = 0.$$

Proof. Suppose, by contradiction, that the result is not true, then there exists a positive constant k_{lbg} such that

$$\|g_k\| \ge k_{lbg} > 0 \tag{11}$$

holds for all k. Denote the index set $A = \{k_i\}$. It follows from the assumption A1-A5 that $\{\|g_k\|\}$ is bounded. So there exists a subsequence $\{k_{i_l}+1\}$ of $\{k_i+1\}$ which satisfies

$$\lim_{l\to\infty} \|g_{k_{i_l}+1}\| = \|g_\infty\| \ge k_{lbg}.$$

By the definition of k_{i_l}, the iterate point $x_{k_{i_l}+1}$ is accepted by the filter F, and for every $l > 1$ there exists $j_l \in \{1,2,\cdots,n\}$ such that

$$\left|g_{k_{i_l}+1,j_l}\right| - \left|g_{k_{i_{l-1}}+1,j_l}\right| \le -\gamma_g \left\|g_{k_{i_{l-1}}+1}\right\|. \tag{12}$$

Since there is only finite choices of j_l, without loss of generality, we set $j_l = j$. In (12), we follows from $l \to +\infty$ and (11) that

$$0 \leftarrow \left|g_{k_{i_l}+1,j}\right| - \left|g_{k_{i_{l-1}}+1,j}\right| \le -\gamma_g k_{lbg} < 0,$$

which is a contradiction. Thus the result is proved.

Now, we give the result when $|A| < +\infty$.

Theorem 4.3 Suppose A1-A5 hold and $|S| = +\infty$, $|A| < +\infty$, then

$$\liminf_{k\to\infty} \|g_k\| = 0.$$

Proof. Suppose, by contradiction, that the result is not true, then there exists a positive constant k_{lbg} such that (11) holds. Since $|A| < +\infty$, by the algorithm RFTR, we have that $\rho_k \ge \eta_1$ holds for all sufficiently large index $k \in S$. Denote

$$\sigma_k = \left|\{p, p+1, \cdots, k\} \cap S\right|.$$

It follows from the assumption A5, Lemma 3.4 and $\rho_k \ge \eta_1$ that

$$f(x_p) - f(x_{k+1}) = \sum_{\substack{j=p \\ j\in S}}^{k} \left(f(x_j) - f(x_{j+1}) \right)$$

$$\ge \sigma_k \eta_1 k_{mdc} k_{lbg} \min\left\{ \frac{k_{lbg}}{k_{umh}}, k_{lbd} \right\}.$$

as long as p, k is sufficiently large. $|S| = +\infty$ and $|A| < +\infty$ imply that σ_k may be large arbitrarily, which contradicts with the fact that $\{f(x_k)\}$ is bounded.

By Theorem 4.1, Theorem 4.2 and Theorem 4.3, there exists at least one limit point of the sequence $\{x_k\}$ of iterates generated by the algorithm RFTR which is a first-order critical point.

5. Second Order Convergence

We now exploit second-order information on the objective function to discuss the second order convergence of the sequence. We therefore introduce the following additional assumptions.

A6 The model is asymptotically second-order coherent with the objective function near first-order critical points, *i.e.*,

$$\lim_{k\to\infty} \|\nabla_{xx} f(x_k) - \nabla_{xx} m_k(x_k)\| = 0 \text{ where } \lim_{k\to\infty} \|g_k\| = 0.$$

A7 There exists a constant k_{lch} such that, for all k,

$$\|\nabla_{xx} m_k(x) - \nabla_{xx} m_k(y)\| \le k_{lch} \|x - y\|.$$

for all $x, y \in \Im_k$.

Lemma 5.1 Suppose that A1-A7 hold. Suppose also that there exists a sequence $\{k_i\}$ and a constant $k_{mqd} > 0$ such that

$$m_{k_i}(x_{k_i}) - m_{k_i}(x_{k_i} + s_{k_i}) \ge k_{mqd} \|s_{k_i}\|^2 > 0 \tag{13}$$

for all i sufficiently large. Finally, suppose that $\lim_{i\to\infty} \|s_{k_i}\| = 0$, then iteration k_i is successful and

$$\tilde{\rho}_{k_i+1} \ge \tilde{\eta}_2 \text{ and } \Delta_{k_i+1} \ge \Delta_{k_i} \tag{14}$$

for all i sufficiently large.

Proof. We first deduce that every iterations k_i is successful for i sufficiently large. By the mean value theorem and (13), for some ξ_k and ς_k in the segment $\left[x_{k_i}, x_{k_i+1}\right]$,

$$\begin{aligned}\left|\rho_{k_i}-1\right| &= \left|\frac{f\left(x_{k_i+1}\right)-m_{k_i}\left(x_{k_i+1}\right)}{m_{k_i}\left(x_{k_i}\right)-m_{k_i}\left(x_{k_i+1}\right)}\right| \\ &\le \frac{1}{2k_{mqd}\left\|s_{k_i}\right\|^2}\left|s_{k_i}^T\left(\nabla_{xx}f\left(\xi_{k_i}\right)-\nabla_{xx}m_{k_i}\left(\varsigma_{k_i}\right)\right)s_{k_i}\right| \\ &\le \frac{1}{2k_{mqd}}\Big(\left\|\nabla_{xx}f\left(\xi_{k_i}\right)-\nabla_{xx}f\left(x_{k_i}\right)\right\| \\ &\quad+\left\|\nabla_{xx}f\left(x_{k_i}\right)-\nabla_{xx}m_{k_i}\left(x_{k_i}\right)\right\| \\ &\quad+\left\|\nabla_{xx}m_{k_i}\left(x_{k_i}\right)-\nabla_{xx}m_{k_i}\left(\varsigma_{k_i}\right)\right\|\Big),\end{aligned}$$

When i goes to infinity, by our assumption that $\left\|s_{k_i}\right\|$ tends to 0, and the bounds

$$\left\|\xi_{k_i}-x_{k_i}\right\|\le\left\|s_{k_i}\right\| \text{ and } \left\|\varsigma_{k_i}-x_{k_i}\right\|\le\left\|s_{k_i}\right\|.$$

Combining the assumption A2 and A7, the first and third terms of the last right-hand side tend to 0. Meanwhile, the second tends to 0 because of the assumption A6 and Theorem 4.1, 4.2, 4.3. As a consequence, ρ_{k_i} tends to 1. when i goes to infinity, and thus larger than η_1 for i sufficiently large.

The residual proof is similar to Lemma 3.8 in [25].

Theorem 5.2 Suppose that A1-A7 hold and that the complete sequence of iterates $\{x_k\}$ converges to the unique limit point x^*. Then x^* is a second order critical point of (1).

Proof. By Theorem 4.1, 4.2, 4.3, $g\left(x^*\right)=0$. We suppose, by contradiction, that $\tau_*=\lambda_{\min}\left(\nabla_{xx}f\left(x^*\right)\right)<0$, by the assumption A6, there exists k_0 such that, for all $k\ge k_0, \tau_k=\lambda_{\min}\left(\nabla_{xx}m_k\left(x_k\right)\right)<\frac{1}{2}\tau_*$. It follows from the assumption A5 that

$$m_k\left(x_k\right)-m_k\left(x_k+s_k\right)\ge\frac{1}{2}k_{mdc}\left|\tau_*\right|\min\left\{\frac{1}{4}\left|\tau_*^2\right|,\Delta_k^2\right\}$$

hold for all $k\ge k_0$. Meanwhile, by the assumption we know that $s_k=x_{k+1}-x_k$ tends to 0. Thus, there exists $k_1\ge k_0$ such that, for all $k\ge k_1$, $\left\|s_k\right\|\le\min\left\{1/2\left|\tau_*\right|,\Delta_k\right\}$. By Lemma 5.1, there exists $k_2\ge k_1$ such that, for all $k\ge k_2$, we deduce that $\rho_k\ge\eta_1$, $\tilde{\rho}_{k+1}\ge\tilde{\eta}_2$ and $\Delta_{k+1}\ge\Delta_k$. Thus,

$$\begin{aligned}f\left(x_k\right)-f\left(x_{k+1}\right) &\ge \eta_1\left(m_k\left(x_k\right)-m_k\left(x_k+s_k\right)\right) \\ &\ge \frac{1}{2}\eta_1 k_{mdc}\left|\tau_*\right|\min\left\{\frac{1}{4}\left|\tau_*^2\right|,\Delta_k^2\right\},\end{aligned}$$

which follows from $\lim_{k\to\infty}x_k=x^*$ that $\lim_{k\to\infty}\Delta_k=0$. This contradicts with $\Delta_{k+1}\ge\Delta_k$ for all $k\ge k_2$.Thus, $\tau_*\ge 0$ and therefore x^* is a second-order critical point of (1).

6. Numerical Experiments

In this section, a preliminary numerical test of the algorithm BTR, the algorithm FTR [22], the algorithm RTR [25] and the algorithm FTR are given. The Matlab codes (Version 7.4.0.287 R2007a) were written corresponding to those algorithms. For the numerical tests, we use the following trust-region radius updated form which is proposed in Conn *et al.* [13].

$$\Delta_{k+1}=\begin{cases}\gamma_1\left\|s_k\right\|, & \text{if } \tilde{\rho}_{k+1}<\tilde{\eta}_1,\\ \Delta_k, & \text{if } \tilde{\rho}_{k+1}\in\left[\tilde{\eta}_1,\tilde{\eta}_2\right],\\ \max\left(\gamma_3\left\|s_k\right\|,\Delta_k\right), & \text{if } \tilde{\rho}_{k+1}>\tilde{\eta}_2.\end{cases}$$

and the following parameter settings [21,28].

$$\Delta\gamma_1=0.25,\ \gamma_3=3.5,\ \tilde{\eta}_1=\eta_1=0.0001,\ \tilde{\eta}_2=\eta_2=0.99,$$

$$\Delta_0=1,\ \gamma_g=\min\left\{0.001,1\Big/\left(2\sqrt{n}\right)\right\}.$$

The Hessian matrix of the model function is $\nabla_{xx}m_k\left(x\right)=\nabla_{xx}f\left(x_k\right)$. The termination criterion is as following,

$$\left\{\left\|g_k\right\|\le 10^{-6}\sqrt{n}\ \text{ or }\ k>k_{\max}\right\},\ k_{\max}=1000,$$

where $k_{\max}$ denotes the maximal iteration number.

We choose 24 test problems from [29], where "S201" means problem 201 in Schittkowski (1987) collection [29], 12 test problems from CUTE [25,30] and the famous Extended Rosenbrock test problem. In the following tables, "n" means the test problem's dimension, "nBTR, nFTR, nRTR, nRFTR" mean the number of iterations of the algorithm BTR, the algorithm FTR, the algorithm RTR and the algorithm RFTR, respectively. "ng1, ng2, ng3, ng4" mean the number of gradient evaluations of the algorithm BTR, the algorithm FTR, the algorithm RTR and the algorithm RFTR, respectively.

"r" means the rank of the number of iterations of the algorithm RFTR among the four algorithms, whose values is in {1, 2, 3, 4}, where "1" means that the number of iterations of the algorithm RFTR is the smallest among the four algorithms, so the algorithm RFTR is the best one among the four algorithms. "4" means that the number of iterations of the algorithm RFTR is the largest among the four algorithms, so the algorithm RFTR is the worst one among the four algorithms. "F" means that the algorithm does not stop when the maximal iteration number is achieved.

In **Table 1**, there are 20 test problems whose iteration number is the smallest, 2 test problems whose rank is second, 2 test problems whose iteration number is the largest among 24 test problems. The numerical results

Table 1. Reports the numerical results on 24 test problems from [29].

Problem	n	nBTR	nFTR	nRTR	nRFTR	ng1	ng2	ng3	ng4	r
S201	2	34	34	34	34	35	35	35	35	1
S202	2	6	6	6	6	7	7	7	7	1
S203	2	6	6	6	6	7	7	7	7	1
S205	2	8	8	8	8	9	9	9	9	1
S206	2	4	4	4	4	5	5	5	5	1
S207	2	9	7	11	7	8	9	10	9	1
S208	2	39	17	27	17	27	23	26	23	1
S209	2	108	45	99	53	86	55	93	61	2
S210	2	424	169	460	176	365	195	450	185	2
S211	2	37	13	37	12	29	17	32	16	1
S212	2	13	14	12	14	11	16	11	16	4
S213	2	27	27	27	27	28	28	28	28	1
S240	3	5	5	5	5	6	6	6	6	1
S241	3	13	13	13	13	14	14	14	14	1
S246	3	10	7	10	7	10	9	10	9	1
S256	4	15	15	15	15	16	16	16	16	1
S260	4	69	33	53	33	47	37	48	37	1
S261	4	13	15	13	15	13	17	13	17	4
S271	6	2	2	2	2	3	3	3	3	1
S273	6	10	10	10	10	11	11	11	11	1
S274	2	2	2	2	2	3	3	3	3	1
S275	4	2	2	2	2	3	3	3	3	1
S276	6	2	2	2	2	3	3	3	3	1
S308	2	11	9	11	9	10	11	10	11	1

Table 2. Reports the numerical results on the famous Extended Rosenbrock test problem.

n	nBTR	nFTR	nRTR	nRFTR	ng1	ng2	ng3	ng4	r
2	39	17	27	17	27	23	26	23	1
10	36	22	31	22	28	29	29	29	1
20	67	36	68	36	49	43	67	43	1
30	83	41	60	42	62	56	58	55	2
40	111	52	100	52	81	70	98	70	1
50	150	69	130	69	405	93	127	91	1
60	170	82	147	82	122	109	144	107	1
70	199	102	162	101	141	140	159	135	1
80	218	121	171	120	157	155	170	147	1
90	248	128	209	121	177	171	207	150	1
100	278	141	220	141	179	182	218	180	1
150	397	207	296	213	284	276	293	274	2
200	532	302	427	283	378	391	423	361	1
250	647	374	507	373	464	478	504	474	1
300	769	419	722	419	551	542	719	537	1
350	900	507	F	501	644	653	996	627	1
400	F	585	F	500	715	748	997	536	1
450	F	640	878	630	712	834	873	798	1
500	F	719	F	710	712	938	994	900	1

Table 3. Reports the numerical results on 12 test problems from CUTE [25,30].

Problem	n	nBTR	nFTR	nRTR	nRFTR	ng1	ng2	ng3	ng4	r
ARHEAD	100	5	5	5	5	6	6	6	6	1
ROSNSCHN	50	66	39	100	39	51	46	98	46	1
COSINE	100	F	F	34	36	986	1008	32	40	2
DQDRTIC	100	4	4	4	4	5	5	5	5	1
ERRINROS	50	52	67	48	27	39	78	44	33	1
FLERCHCR	100	29	29	29	29	30	30	30	30	1
LIARWHD	300	13	13	13	13	14	14	14	14	1
LOGHAIRY	2	61	64	55	31	51	66	50	33	1
NONDIA	100	24	24	24	24	25	25	25	25	1
PFIT4LS	3	271	291	309	233	217	323	300	245	1
POWELLS	4	15	15	15	15	16	16	16	16	1
WOOD	4	69	33	53	33	47	37	48	37	1

show that the number for the algorithm RFTR to solve trust region subproblem is the smallest in total.

In **Table 2**, There are only 2 cases whose rank is second, the others all are the best. Moreover, the algorithm RFTR is more and more effective as the increase of the problem's dimension.

In **Table 3**, There are only 1 case whose rank is second, the others all are the best. Moreover, The retrospective idea takes effects on the Problems COSINE, ERRINROS, LOGHAIRY clearly.

7. Conclusions and Perspectives

Trust region method is very reliable and robust and has very strong convergence properties. It is a class of very effective algorithms for solving unconstrained optimization now. The basic trust region algorithm is the monotone descent algorithm, *i.e.*, the value of the object function in the iterate sequence $\{x_k\}$ strictly decreases monotonically. If the iterates follow the bottom of curved narrow valleys, then the monotone descent algorithm converges very slowly. The idea of non-monotone method [23,24] abandons the restriction of the descent property of the value of the object function, which allows the sequence of iterates to follow the bottom of curved narrow valleys much more loosely, which hopefully results in longer and more efficient steps.

Trust region method combines with the filter technique, which, in some sense, relaxes the monotonicity condition which accepts the trial step. The filter technique improves the numerical effect for some problems.

The new algorithm RFTR presented in this paper combines with the filter technique and the retrospective idea, which the number of the algorithm RFTR to solve trust region subproblem is decreased in total. On the other hand, our algorithm also looks like a self-adaptive method based on the trust-region framework. Meanwhile, our algorithm is not like the other algorithms about self-adaptive method [26,27] which need to compute the gradient value and function value at the auxiliary point, but may measure the acceptance of the previous iterate and the current iterate for the new and old model function, respectively, which keep the robustness property of the trust-region method.

8. References

[1] M. J. D. Powell, "A New Algorithm for Unconstrained Optimization," In: J. B. Rosen, O. L. Mangasarian and K. Ritter, Eds., *Nonlinear Programming*, Academic Press, New York, 1970.

[2] K. Levenberg, "A Method for The Solution of Certain Nonlinear Problems in Least Squares," *The Quarterly of Applied Mathematics*, Vol. 2, No. 2, 1944, pp. 164-168.

[3] D. W. Marquardt, "An Algorithm for Least Squares Estimation of Nonlinear Inequalities," *SIAM Journal on Applied Mathematics*, Vol. 11, No. 2, 1963, pp. 431-441.

[4] M. J. D. Powell, "Convergence Properties of a Class of Minimization Algorithms," In: O. L. Mangasarian, R. R. Meyer and S. M. Robinson, Eds., *Nonlinear Programming*, Academic Press, New York, 1975, pp. 1-27.

[5] M. J. D. Powell, "On the Global Convergence of Trust-Region Algorithms for Unconstrained Optimization," *Mathematical Programming*, Vol. 29, No. 3, 1984, pp. 297-303.

[6] G. A. Schultz, R. B. Schnabei and R. H. Byrd, "A Family of Trust-Region-Based Algorithms for Unconstrained Minimization with Strong Global Convergence," *SIAM Journal on Numerical Analysis*, Vol. 22, No. 1, 1985, pp. 47-67.

[7] D. C. Sorensen, "Newton's Method with a Model Trust Region Modifications," *SIAM Journal on Numerical Analysis*, Vol. 19, No. 2, 1982, pp. 409-426.

[8] J. J. Moŕe, "Recent Developments in Algorithms and Software for Trust Region Methods," In A. R. Bachem, M. Grotshel and B. Korte, Eds., *Mathematical Programming*: *The State of the Art*, Springer-Verlag, Berlin, 1983, pp. 258-287.

[9] Y. X. Yuan, "On the Convergence of Trust Region Algorithm," *Mathematica Numerica Sinica*, Vol. 16, No. 3, 1996, pp. 333-346.

[10] M. Lalee, J. Nocedal and T. Plantenga, "On the Implentation of an Algorithm for Large-Scale Equality Constrained Optimization," *SIAM Journal on Optimization*, Vol. 8, No. 3, 1998, pp. 682-706.

[11] A. Friedlander, J. M. Martinez and S. A. Santos, "A New Trust Region Algorithm for Bound Constrained Minimization," *Applied Mathematics and Optimization*, Vol. 30, No. 3, 1994, pp. 235-266.

[12] A. R. Conn, N. I. M. Gould and P. L. Toint, "Convergence Properties of Minimization Algorithms for Convex Constraints Using a Structured Trust Region," *SIAM Journal on Optimization*, Vol. 6, No. 4, 1996, pp. 1059-1086.

[13] A. R. Conn, N. I. M. Gould and P. L. Toint, "Trust Region Methods," MPS-SIAM Series on Optimization, SIAM, Philadelphia, 2000.

[14] R. Fletcher and S. Leyffer, "Nonlinear Programming without a Penalty Function," *Mathematical Programming*, Vol. 91, No. 2, 2002, pp. 239-269.

[15] R. Fletcher, N. I. M. Gould, S. Leyffer, P. L. Toint and A. Wächter, "Global Convergence of a Trust-Region SQP-Filter Algorithm for General Nonlinear Programming," *SIAM Journal on Optimization*, Vol. 13, No. 3, 2002, pp. 635-659.

[16] R. Fletcher, S. Leyffer and P. L. Toint, "On the Global Convergence of a Filter-SQP Algorithm," *SIAM Journal on Optimization*, Vol. 13, No. 1, 2002, pp. 44-59.

[17] M. Ulbrich, S. Ulbrich and L. N. Vicente, "A Globally Convergent Primal Dual Interior-Point Filter Method for Nonconvex Nonlinear Programming," *Mathematical Programming*, Vol. 100, No. 2, 2003, pp. 379-410.

[18] A. Wächter and L. T. Biegler, "Line Search Filter Methods for Nonlinear Programming: Motivation and Global Convergence," *SIAM Journal on Optimization*, Vol. 16, No. 1, 2005, pp. 1-31.

[19] A. Wächter and L. T. Biegler, "Line Search Filter Methods for Nonlinear Programming: Local Convergence," *SIAM Journal on Optimization*, Vol. 16, No. 1, 2005, pp. 32-48.

[20] R. Fletcher, S. Leyffer and P. L. Toint, "A Brief History of Filter Methods," *SIAG*/*OPT Views and News*, Vol. 18, No. 1, 2006, pp. 2-12.

[21] N. I. M. Gould, C. Sainvitu and P. L. Toint, "A Filter-Trust-Region Method for Unconstrained Optimization," *SIAM Journal on Optimization*, Vol. 16, No. 2, 2005, pp. 341-357.

[22] W. H. Miao and W. Y. Sun, "A Filter Trust-Region Method for Unconstrained Optimization," *Numerical Mathematics − A Journal of Chinese Universities*. *Gaodeng Xuexiao Jisuan Shuxue Xuebao*, Vol. 29, No. 1, 2007, pp. 88-96.

[23] L. Grippo, F. Lampariello and S. Lucidi, "A Nonmonotone Line Search Technique for Newton's Methods," *SIAM Journal on Numerical Analysis*, Vol. 23, No. 4, 1986, pp. 707-716.

[24] P. L. Toint, "Non-Monotone Trust-Region Algorithms for Nonlinear Optimization Subject to Convex Constraints," *Mathematical Programming*, Vol. 77, No. 1, 1997, pp. 69-94.

[25] F. Bastin, V. Malmedy, M. Mouffe, P. L. Toint and D. Tomanos, "A Retrospective Trust-Region Method for Unconstrained Optimization," *Mathematical Programming*, Vol. 123, No. 2, 2010, pp. 395-418.

[26] A. Sartenaer, "Automatic Determination of an Initial Trust Region in Nonlinear Programming," *SIAM Journal on Scientific Computing*, Vol. 18, No. 6, 1997, pp. 1788-1803.

[27] X. S. Zhang, Z. W. Chen and J. L. Zhang, "A Self-Adaptive Trust Region Method Unconstrained Optimization," *Operations Research Transactions*, Vol. 5, No. 1, 2001, pp. 53-62.

[28] N. I. M. Gould, D. Orban, A. Sartenaer and P. L. Toint, "Sensitivity of the Trust-Region Algorithms to Their Parameters," 4*OR*: *A Quarterly Journal of Operations Research*, Vol. 3, No. 3, 2005, pp. 227-241.

[29] K. Schittkowski, "More Test Examples for Nonlinear Programming Codes," *Lecture Notes in Economics and Mathematical Systems*, Springer Verlag, Berlin, Heideberg, Vol. 282, 1987.

[30] H. Y. Benson, "Nonlinear Optimization Models by AMPL: Cute Set." http://www.princeton.edu/~rvdb/ampl/nlmodels

Applied Mathematics, **2010, 1, 189-199**
doi:10.4236/am.2010.13023 Published Online September 2010 (http://www.SciRP.org/journal/am)

Uncertainty Theory Based Novel Multi-Objective Optimization Technique Using Embedding Theorem with Application to R & D Project Portfolio Selection

Rupak Bhattacharyya, Amitava Chatterjee, Samarjit Kar
Department of Mathematics, National Institute of Technology, Durgapur, India
E-mail: {*mathsrup, amitavamath*}*@gmail.com, kar_s_k@yahoo.com*
Received June 6, 2010; *revised July* 19, 2010; *accepted July* 22, 2010

Abstract

This paper introduces a novice solution methodology for multi-objective optimization problems having the coefficients in the form of uncertain variables. The embedding theorem, which establishes that the set of uncertain variables can be embedded into the Banach space $C[0, 1] \times C[0, 1]$ isometrically and isomorphically, is developed. Based on this embedding theorem, each objective with uncertain coefficients can be transformed into two objectives with crisp coefficients. The solution of the original m-objectives optimization problem with uncertain coefficients will be obtained by solving the corresponding 2 m-objectives crisp optimization problem. The R & D project portfolio decision deals with future events and opportunities, much of the information required to make portfolio decisions is uncertain. Here parameters like outcome, risk, and cost are considered as uncertain variables and an uncertain bi-objective optimization problem with some useful constraints is developed. The corresponding crisp tetra-objective optimization model is then developed by embedding theorem. The feasibility and effectiveness of the proposed method is verified by a real case study with the consideration that the uncertain variables are triangular in nature.

Keywords: Uncertainty Theory, Uncertain Variable, Embedding Theorem, α-Optimistic and α-Pessimistic Value, R & D Project Portfolio Selection

1. Introduction

An important problem in topology is to decide when a space X can be embedded into another space Y, *i.e.*, when there exists an embedding from X into Y. This problem is called embedding problem. Theorems asserting the embedding of a space into some other space which is more manageable than the original space are known as embedding theorems. On the other hand, a theorem which asserts that a certain space cannot be embedded into some other space is known as non-embedding theorems. Non-embedding theorems are often quite deep and require methods well beyond the general topology. For example it is by no means trivial to prove that the 2-sphere (S^2) cannot be embedded into the Euclidean space.

The embedding theorems in crisp and fuzzy environments are already established. The fuzzy embedding theorem shows that each fuzzy number can be identified isometrically and isomorphically with an element in C [0, 1] × C[0, 1] where C [0, 1] is the set of all real valued continuous functions on [0, 1]. Puri and Ralescu [1] and Kaleva [2] have proved that the set of all fuzzy numbers can be embedded into a Banach space isometrically and isomorphically. Wu and Ma [3] provide a specific Banach space, which shows that the set of all fuzzy numbers can be embedded into the Banach space $C[0, 1] \times C[0, 1]$. Wu [4] propose an (α, β)-optimal solution concept of fuzzy optimization problem based on the possibility and necessity measures. To do so, the fuzzy optimization problem is transformed into a bi-objective programming problem by applying the embedding theorem. Wu [5] shows that the optimal solution of the crisp optimization problem obtained from the fuzzy optimization problem by using embedding theorem is also an optimal solution of the original fuzzy optimization problem under the set of core values of fuzzy numbers.

With increasing competition and limitations of financial resources, the way of selection of R & D projects

that maximize some measure of utility or benefit to the organization has become a critical one. The purposes of project portfolio decision are to allocate a limited set of resources to projects in a way that balance risk, reward and alignment with corporate strategy. Poor selection of R & D projects could have a significantly negative impact on organizations for decades. The R & D project portfolio decision deals with future events and opportunities, much of the information required to make portfolio decisions is at best uncertain and at worst very unreliable. Project selection is usually described in term of constraint optimization problem. Given a set of project proposals, the goal is to select a subset of projects to maximize some objective without violating the constraints. An R & D project usually involves several phases. Therefore, each phase is an option that is contingent on earlier of other options. Some attempts already exist for R & D project portfolio selection. Rabbani *et al.* [6,7], Fang *et al.* [8], Riddell and Wallace [9], Eilat *et al.* [10], Stummer and Heidenberger [11], Linton *et al.* [12], Ringuest *et al.* [13], Schmidt [14] and others have done significant works in the field of R & D project portfolio selection. To model uncertainty and vagueness, fuzzy set theory is used by many to characterize uncertain R & D project information. Pereira and junior [15], Coffin and Taylor [16], Machacha and Bhattacharyya [17], Kuchta [18], Mohamed and McCowan [19], Hsu *et al.* [20], Wang and Hwang [21], Kim *et al.* [22], Karsak [23] have applied fuzzy set theory in the field of R & D project portfolio selection. Unfortunately, R & D project managers have been unable to adopt many of these mechanisms.

But in reality, sometimes investors have to deal with the uncertainty which acts neither randomness nor fuzziness. In order to deal with such type of uncertainty, Liu [24] founds uncertainty theory as a branch of mathematics. Subsequently, Liu [25] proposes uncertain process and uncertain differential equation to deal with dynamic uncertain phenomena. In addition, uncertain calculus is introduced by Liu [26] to describe the function of uncertain processes, uncertain inference is introduced by Liu [26] via the tool of conditional uncertainty and uncertain logic is proposed by Li and Liu [27] to deal with uncertain knowledge. Liu [28] proposes an uncertain programming including expected value model, chance constrained programming and dependent-chance programming to model several optimization problems. Till now, several research works [29,30] have been done in this area, but none has considered the R & D project portfolio selection problem in the uncertain environment. Basically, till date, no embedding theorem based optimization technique is proposed in uncertainty theory.

In this paper, in Section 2, we provide some preliminaries required to develop the paper. In Section 3, an uncertain embedding theorem is proved and an uncertain single/multiple objective optimization method using the embedding theorem is established. In Section 4, we develop an uncertain linear bi-objective R & D project portfolio selection model. The objectives are 1) maximization of project benefit and 2) minimization of project risk. The risk is defined as the maximum loss that the decision maker may face in the worst case. This is considered as the projected maximum loss in case of failure of the project. Constraints on budget and resources are also considered. Using the embedding theorem established in Section 3, we convert the bi-objective uncertain optimization problem into a tetra-objective crisp optimization problem which is further transformed into a deterministic convex optimization model by global criteria approach. In Section 5 of this paper a real case study is provided to illustrate our method. The optimization software LINGO is used for the simulation. Finally in Section 6 some concluding remarks are presented.

2. Preliminaries

Before discussing the embedding theorem and its relevance in uncertain optimization problem we would like to discuss some basic concepts related with metric space, topology and uncertainty theory.

Definition 2.1 (Metric Space) A non-empty set X is said to be a metric space if to every pair of elements x, y of this set, there corresponds a non-negative real number $\rho(x, y)$ for which the following conditions hold.

1) $\rho(x, y) > 0$ and $\rho(x, y) = 0$ if and only if $x = y$

2) $\rho(x, y) = \rho(y, x)$

3) for any three elements x, y, z in X,

$$\rho(x, y) \leq \rho(x, z) + \rho(z, y).$$

The number $\rho(x, y)$ is called the difference or metric between the elements **x, y** and the above three conditions are called metric axioms and a metric space is sometimes written as (X, ρ).

Definition 2.2 A sequence $\{x_n\}$ of elements of a metric space X is called a Cauchy sequence if for every $\varepsilon > 0$ there exists a positive integer N such that $\rho(x_n, x_m) < \varepsilon$ whenever $n, m \geq N$.

If every Cauchy sequence of a metric space X has a finite limit in X then X is called complete. By Cauchy's general principle of convergence it can be shown that the real line and the complex plane with usual metric are the complete metric spaces.

Definition 2.3 A set E is called a normed linear space if

1) E is a linear space with real or complex numbers as scalars and

2) to every element x of E there is associated a unique

real number, called the norm of the element x and denoted by $||x||$.

The norm of an element x has to satisfy the following axioms.

1) $||x|| > 0$ and $||x|| = 0$ if and only if $x = \theta$

2) $||\alpha x|| = |\alpha|\ ||x||$ where α is a scalar

3) $||x + y|| \le ||x|| + ||y||$ for every x, y in E.

Note: In a normed linear space we can introduce a metric by $\rho(x, y) = ||x - y||$. The metric axioms are fulfilled as

1) $\rho(x, y) > 0$ and $\rho(x, y) = 0$ if and only if $||x - y|| = 0$,*i.e.*, if and only if $x - y = \theta$, *i.e.*, if and only if $x = y$.

2) $\rho(x, y) = ||x - y|| = ||(-1)(y - x)|| = |-1|\ ||y - x|| = ||y - x|| = \rho(y, x)$

3) $\rho(x, y) = ||x - y|| = ||(x - z) + (z - x)|| \le ||x - z|| + ||z - y|| = \rho(x, z) + \rho(z, y)$.

Definition 2.4 (Banach Space) If a normed linear space is complete in the sense of the convergence in norm, then it is called a Banach space.

Every finite dimensional normed linear space E is complete (that is a Banach space) and bounded. Every finite dimensional linear space can be made a Banach space.

Definition 2.5 Two objects A and B are said to be congruent (or isometric) if there exists a bijection f: $A \to B$ which preserves all distances in the sense that $d(x, y) = d(f(x), f(y))$ for all pairs (x, y) of points in A, d being used to denote the distance between points. Such a bijection, when it exists, is called congruence (or an isometry).

Definition 2.6 Let (X, τ), (Y, Ψ) be topological spaces. An embedding (or imbedding) theorem of X into Y is a function e: $X \to Y$ which is a homeomorphism when considered as a function from (X, τ) onto $(e(X), \Psi/e(X))$.

Definition 2.7 A function e: $X \to Y$ is an embedding function if and only if it is continuous and one-one and for every open set V in X there exists an open subset W of Y such that $e(V) = W \cap Y$.

Definition 2.8 The space $C[0, 1]$ is the set of all real valued continuous functions f on $[0, 1]$, such that f is left-continuous for any $t \in (0, 1]$ and right-continuous at 0, and f has a right limit for any $t \in [0, 1)$. The norm is defined as $\|f\| = \sup_{t \in [0,1]} |f(t)|$.

Definition 2.9 Let Γ be a non-empty set and $\mathring{A}$ a σ-algebra over Γ. Each element $\Lambda \in \mathring{A}$ is called an event. Let M be a set function over $\mathring{A}$. Then M is called an uncertain measure (Liu, [24]) if it satisfies the following four axioms.

Axiom 1. (Normality) $M\{\Gamma\} = 1$;

Axiom 2. (Monotonicity) $M\{\Lambda_1\} \le M\{\Lambda_2\}$ whenever $\Lambda_1 \subset \Lambda_2$;

Axiom 3. (Self-Duality) $M\{\Lambda\} + M\{\Lambda^c\} = 1$ for any event Λ;

Axiom 4. (Countable Subadditivity) $M\left\{\bigcup_{i=1}^{\infty} \Lambda_i\right\} \le \sum_{i=1}^{\infty} M\{\Lambda_i\}$, for every countable sequence of events $\{\Lambda_i\}$.

The triplet $(\Gamma, \mathring{A}, M)$ is called an uncertainty space.

Definition 2.10 An uncertain variable is a measurable function $\breve{\xi}$, from an uncertain space $(\Gamma, \mathring{A}, M)$ to the set of all real numbers, *i.e.*, for any Borel set B of real numbers, the set $\{\breve{\xi} \in B\} = \{\gamma \in \Gamma \,|\, \breve{\xi}(\gamma) \in B\}$ is an event.

Definition 2.11 (Liu, [24]) An uncertain variable $\breve{\xi}$ is said to have a first identification function λ if

1) $\lambda(x)$ is a non-negative function on $\mathbb{R}$ such that $\sup_{x \ne y}(\lambda(x) + \lambda(y)) = 1$;

2) for any set B of real numbers, we have,

$$M\{\breve{\xi} \in B\} = \begin{cases} \sup_{x \in B} \lambda(x) & \text{if } \sup_{x \in B} \lambda(x) < 0.5 \\ 1 - \sup_{x \in B^c} \lambda(x) & \text{if } \sup_{x \in B} \lambda(x) \ge 0.5. \end{cases}$$

Definition 2.12 A rectangular uncertain variable is defined to be the uncertain variable which is fully determined by the pair (a, b) of crisp numbers with $a < b$, and whose first identification function is

$$\lambda(x) = 0.5, \ \mathrm{a} \le \mathrm{x} \le \mathrm{b}.$$

Definition 2.13 A triangular uncertain variable is defined to be the uncertain variable which is fully determined by the triplet (a, b, c) of crisp numbers with $a < b < c$, and whose first identification function is

$$\lambda(x) = \begin{cases} \dfrac{x - a}{2(b - a)} & \text{if } \mathrm{a} \le \mathrm{x} \le \mathrm{b} \\ \dfrac{c - x}{2(c - b)} & \text{if } \mathrm{b} \le \mathrm{x} \le \mathrm{c}. \end{cases}$$

Definition 2.14 A trapezoidal uncertain variable is defined to be the uncertain variable which is fully determined by the quadruplet (a, b, c, d) of crisp numbers with $a < b < c < d$, and whose first identification function is

$$\lambda(x) = \begin{cases} \dfrac{x - a}{2(b - a)} & \text{if } \mathrm{a} \le \mathrm{x} \le \mathrm{b} \\ 0.5 & \text{if } \mathrm{b} \le \mathrm{x} \le \mathrm{c} \\ \dfrac{\mathrm{d} - x}{2(d - c)} & \text{if } \mathrm{c} \le \mathrm{x} \le \mathrm{d}. \end{cases}$$

Definition 2.15 (Liu, [24]) An uncertain variable $\breve{\xi}$ is said to have a second identification function ρ if

1) $\rho(x)$ is a nonnegative and integrable function on $\mathbb{R}$ such that $\int_{\mathbb{R}} \rho(x)dx \ge 1$;

2) for any set B of real numbers, we have,

$$M\{\xi \in B\} = \begin{cases} \int_B \rho(x)dx & \text{if } \int_B \rho(x)dx < 0.5 \\ 1 - \int_{B^c} \rho(x)dx & \text{if } \int_B \rho(x)dx \ge 0.5. \end{cases}$$

Definition 2.16 An exponential uncertain variable is defined to be the uncertain variable having the second identification function

$$\rho(x) = \frac{1}{\beta}\exp\left(-\frac{x}{\alpha}\right),\ x \geq 0,$$

and is denoted by *EXP*(α, β), where α, β are real numbers with $\alpha \geq \beta > 0$. Note that $\int_{-\infty}^{\infty} \rho(x)dx = \frac{\alpha}{\beta} \geq 1$.

Definition 2.17 A bell-shaped uncertain variable is defined to be the uncertain variable having the second identification function

$$\rho(x) = \frac{1}{\beta\sqrt{\pi}}\exp\left(-\frac{(x-m)^2}{\alpha^2}\right),\ x \in \mathbb{R},$$

and is denoted by *B*(m, α, β), where α, β are real numbers with $\alpha \geq \beta > 0$. Note that $\int_0^{\infty} \rho(x)dx = \frac{\alpha}{\beta} \geq 1$.

Definition 2.18 (Liu, [24]) The uncertain variables $\breve{\xi}_1, \breve{\xi}_2, \ldots, \breve{\xi}_n$ are said to be independent if

$$M\left\{\bigcap_{i=1}^{n}\left\{\breve{\xi}_i \in B_i\right\}\right\} = \min_{1 \leq i \leq n} M\left\{\breve{\xi}_i \in B_i\right\}$$

for Borel sets B_1, B_2,…, B_n of real numbers.

Definition 2.19 (Liu, [24]) The uncertainty distribution Φ: $\mathbb{R} \to$ [0, 1] of an uncertain variable $\breve{\xi}$ is defined by

$$\Phi(x) = M\{\breve{\xi} \leq x\}.$$

Definition 2.20 An uncertain variable $\breve{\xi}$ is called normal if its distribution function Φ is given by

$$\Phi(x) = \left(1 + \exp\left(\frac{\pi(m-x)}{\sqrt{3}\sigma}\right)\right)^{-1},\ x \in \mathbb{R}.$$

It is then denoted by $N(m, \sigma)$, where m, $\sigma(>0)$ are real numbers.

Definition 2.21 (Chen [8]) Let $\breve{\xi}$ be an uncertain variable and $\alpha \in (0, 1]$. Then

$$\breve{\xi}_{opt}^{\alpha} = \sup\left\{r \mid M\{\breve{\xi} \geq r\} \geq \alpha\right\}$$

is called the α – optimistic value of ξ, and

$$\breve{\xi}_{pes}^{\alpha} = \inf\left\{r \mid M\{\breve{\xi} \geq r\} \geq \alpha\right\}$$

is called the α – pessimistic value of $\breve{\xi}$.

Example 2.22 Let $\breve{\xi} = (a, b)$ be a rectangular uncertain variable. Then its α–optimistic and α–pessimistic values are

$$\breve{\xi}_{opt}^{\alpha} = \begin{cases} b & if\ \alpha \leq 0.5 \\ a & if\ \alpha > 0.5, \end{cases} \quad \breve{\xi}_{pes}^{\alpha} = \begin{cases} b & if\ \alpha \leq 0.5 \\ a & if\ \alpha > 0.5. \end{cases}$$

Example 2.23 Let $\breve{\xi} = (a, b, c)$ be a triangular uncertain variable. Then its α–optimistic and α–pessimistic values are

$$\breve{\xi}_{opt}^{\alpha} = \begin{cases} 2\alpha b + (1-2\alpha)c & if\ \alpha \leq 0.5 \\ (2\alpha-1)a + (2-2\alpha)b & if\ \alpha > 0.5, \end{cases}$$

$$\breve{\xi}_{pes}^{\alpha} = \begin{cases} (1-2\alpha)a + 2\alpha b & if\ \alpha \leq 0.5 \\ (2-2\alpha)b + (2\alpha-1)c & if\ \alpha > 0.5. \end{cases}$$

Example 2.24 Let $\breve{\xi} = (a, b, c, d)$ be a trapezoidal uncertain variable. Then its α–optimistic and α–pessimistic values are

$$\breve{\xi}_{opt}^{\alpha} = \begin{cases} 2\alpha c + (1-2\alpha)d & if\ \alpha \leq 0.5 \\ (2\alpha-1)a + (2-2\alpha)b & if\ \alpha > 0.5, \end{cases}$$

$$\breve{\xi}_{pes}^{\alpha} = \begin{cases} (1-2\alpha)a + 2\alpha b & if\ \alpha \leq 0.5 \\ (2-2\alpha)c + (2\alpha-1)d & if\ \alpha > 0.5. \end{cases}$$

Example 2.25 Let $\breve{\xi} = EXP(a, b)$ be an exponential uncertain variable. Then its α–optimistic and α–pessimistic values are

$$\breve{\xi}_{opt}^{\alpha} = \begin{cases} a.\ln\left(\frac{a}{a-\alpha b}\right) & if\ \alpha < 0.5 \\ a.\ln\left(\frac{a}{b-\alpha b}\right) & if\ \alpha \geq 0.5, \end{cases}$$

$$\breve{\xi}_{pes}^{\alpha} = \begin{cases} a.\ln\left(\frac{a}{\alpha b}\right) & if\ \alpha < 0.5 \\ a.\ln\left(\frac{a}{a-(1-\alpha)b}\right) & if\ \alpha \geq 0.5. \end{cases}$$

Example 2.26 Let $\breve{\xi} = \mathrm{B}(e, a, b)$ be a bell–shaped uncertain variable. Then its α–optimistic and α–pessimistic values are

$$\breve{\xi}_{opt}^{\alpha} = \begin{cases} \frac{a}{\sqrt{2}}\Phi^{-1}\left(\frac{\alpha b}{a}\right) + e & if\ \alpha < 0.5 \\ \frac{a}{\sqrt{2}}\Phi^{-1}\left(\frac{a-(1-\alpha)b}{a}\right) + e & if\ \alpha \geq 0.5, \end{cases}$$

$$\breve{\xi}_{pes}^{\alpha} = \begin{cases} \frac{a}{\sqrt{2}}\Phi^{-1}\left(\frac{a-\alpha b}{a}\right) + e & if\ \alpha < 0.5 \\ \frac{a}{\sqrt{2}}\Phi^{-1}\left(\frac{(1-\alpha)b}{a}\right) + e & if\ \alpha \geq 0.5. \end{cases}$$

Example 2.27 Let $\breve{\xi} = \aleph(m, \sigma)$ be a normal uncertain variable. Then its α–optimistic and α–pessimistic values are

$$\breve{\xi}_{opt}^{\alpha} = m - \frac{\sqrt{3}\sigma}{\pi}\ln\left(\frac{\alpha}{1-\alpha}\right),\ \breve{\xi}_{pes}^{\alpha} = m + \frac{\sqrt{3}\sigma}{\pi}\ln\left(\frac{\alpha}{1-\alpha}\right).$$

Theorem 2.28 Let $\breve{\xi}$ be an uncertain variable. Then $\breve{\xi}_{opt}^{\alpha}$ is an increasing and left continuous function of α.

Also $\breve{\xi}^{\alpha}_{pes}$ is a decreasing and left-continuous function of α (Liu [24]).

3. Uncertain Multiple Objective Optimization Method Using Embedding Theorem

In this section, we introduce a solution methodology for multiple objective programming problems in uncertain environment by using the concept of optimistic and pessimistic values of uncertain variables.

Definition 3.1 Let $\breve{\xi}$, $\breve{\zeta}$ be two uncertain variables. We write $\breve{\xi} \succeq \breve{\zeta}$ if and only if

$$\breve{\xi}^{\alpha}_{opt} \geq \breve{\zeta}^{\alpha}_{opt}, \breve{\xi}^{\alpha}_{pes} \geq \breve{\zeta}^{\alpha}_{pes} \forall \alpha \in [0,1].$$

We also write $\breve{\xi} \preceq \breve{\zeta}$ if and only if $\breve{\zeta} \succeq \breve{\xi}$.

On the other hand, we write that $\breve{\xi} \succ \breve{\zeta}$ if and only if

$$\breve{\xi}^{\alpha}_{opt} \geq \breve{\zeta}^{\alpha}_{opt},\ \breve{\xi}^{\alpha}_{pes} > \breve{\zeta}^{\alpha}_{pes} \forall \alpha \in [0,1] \quad \text{or,}$$
$$\breve{\xi}^{\alpha}_{opt} > \breve{\zeta}^{\alpha}_{opt},\ \breve{\xi}^{\alpha}_{pes} \geq \breve{\zeta}^{\alpha}_{pes} \forall \alpha \in [0,1] \quad \text{or,}$$
$$\breve{\xi}^{\alpha}_{opt} > \breve{\zeta}^{\alpha}_{opt},\ \breve{\xi}^{\alpha}_{pes} > \breve{\zeta}^{\alpha}_{pes} \forall \alpha \in [0,1].$$

We also write $\breve{\xi} \preceq \breve{\zeta}$ if and only if $\breve{\zeta} \prec \breve{\xi}$.

Definition 3.2 Let $\breve{\xi}$ be an uncertain variable. Then the norm of $\breve{\xi}$ is defined by $\left\|\breve{\xi}\right\| = E\left|\breve{\xi}\right|$.

Justification:

$$1)\ \left\|\breve{\xi}+\breve{\eta}\right\| = E\left|\breve{\xi}+\breve{\eta}\right|$$
$$= \int_0^{\infty} M\{\left|\breve{\xi}+\breve{\eta}\right| \geq r\}dr$$
$$\leq [\int_0^{\infty} M\{\left|\breve{\xi}\right| \geq r/2\}dr + \int_0^{\infty} M\{|\breve{\eta}| \geq r/2\}dr]$$
$$= [2E\left|\breve{\xi}\right| + 2E|\breve{\eta}|] = 2\left\|\breve{\xi}\right\| + 2\left\|\breve{\eta}\right\|$$

(Triangular inequality as defined by Liu [24]).

$$2)\ \left\|c\breve{\xi}\right\| = E\left|c\breve{\xi}\right| = |c|E\left|\breve{\xi}\right| = |c|\left\|\breve{\xi}\right\|.$$

Theorem 3.3 (Embedding theorem) Let the function $\pi : U(\mathbb{R}) \to C[0,1] \times C[0,1]$ is defined by

$$\pi(\breve{\xi}) = \begin{cases} (\breve{\xi}^{\alpha}_{opt}, \breve{\xi}^{\alpha}_{pes}), if\ \alpha > 0.5 \\ (\breve{\xi}^{\alpha}_{pes}, \breve{\xi}^{\alpha}_{opt}), if\ \alpha \leq 0.5 \end{cases}$$

Then the following properties hold good.

1) π is injective.

2) $\pi((1\{s\}\times\breve{\xi})+(1\{t\}\times\breve{\eta})) = s\pi(\breve{\xi}) + t\pi(\breve{\eta})\ \forall\ \breve{\xi},\breve{\eta} \in U(\mathbb{R})$, $s \geq 0$, $t \geq 0$.

3) $d_U(\breve{\xi},\breve{\eta}) = \left\|\pi(\breve{\xi}) - \pi(\breve{\eta})\right\|$.

That is, $U(\mathbb{R})$ can be embedded into $C[0,1] \times C[0,1]$ isometrically and isomorphically.

Proof. Let $\alpha \leq 0.5$.

1) Let, if possible, $\breve{\xi}$, $\breve{\zeta}$ be two distinct uncertain variables such that $\pi(\breve{\xi}) = \pi(\breve{\zeta})$. Then $(\breve{\xi}^{\alpha}_{pes}, \breve{\xi}^{\alpha}_{opt}) = (\breve{\zeta}^{\alpha}_{pes}, \breve{\zeta}^{\alpha}_{opt})$.

Since the two real open intervals are equal, the corresponding boundary points must be the same. Then $\breve{\xi}^{\alpha}_{pes} = \breve{\zeta}^{\alpha}_{pes}$ and $\breve{\xi}^{\alpha}_{opt} = \breve{\zeta}^{\alpha}_{opt}$, which contradicts the fact that $\breve{\xi} \neq \breve{\zeta}$. Hence our assumption is wrong and consequently the mapping π is injective.

2) We have the bottom equation.

3) We have,

$$d_U(\xi,\eta) = E\left|\xi-\eta\right|$$
$$= E(\left|\xi-\eta\right|^{\alpha}_{pes}, \left|\xi-\eta\right|^{\alpha}_{opt}) = (E(\left|\xi-\eta\right|^{\alpha}_{pes}), E(\left|\xi-\eta\right|^{\alpha}_{opt}))$$
$$= (\left\|\xi-\eta\right\|^{\alpha}_{pes}, \left\|\xi-\eta\right\|^{\alpha}_{opt}) = (\pi\left\|\xi-\eta\right\|) = \left\|\pi(\xi)-\pi(\eta)\right\|.$$

For $\alpha > 0.5$ the proof is similar.

Proposition 3.4 (Order preserving) Let π be the function defined in theorem 3.3 and let $\xi,\zeta \in U(\mathbb{R})$. Then $\xi \preceq \zeta$ if and only if $\pi(\xi) \leq \pi(\zeta)$. We also have $\xi \prec \zeta$ if and only if $\pi(\xi) < \pi(\zeta)$.

Proof. We note that

$$\xi \leq \zeta \Leftrightarrow \xi^{\alpha}_{opt} \leq \zeta^{\alpha}_{opt},\ \xi^{\alpha}_{pes} < \zeta^{\alpha}_{pes} \forall \alpha \in [0,1]$$
$$\Leftrightarrow (\xi^{\alpha}_{opt}, \xi^{\alpha}_{pes}) \leq (\zeta^{\alpha}_{opt},\ \zeta^{\alpha}_{pes})\ \forall \alpha \in [0, \frac{1}{2}],$$
$$(\xi^{\alpha}_{pes}, \xi^{\alpha}_{opt}) \leq (\zeta^{\alpha}_{pes},\ \zeta^{\alpha}_{opt})\ \forall \alpha \in [\frac{1}{2}, 1]$$
$$\Leftrightarrow \pi(\xi) \leq \pi(\zeta).$$

Similarly it can be shown that $\xi < \zeta$ if and only if $\pi(\xi) < \pi(\zeta)$.

Definition 3.5 Let $\breve{f}_1, \breve{f}_2, \breve{g}_1, \breve{g}_2$ be real valued functions defined by $\breve{f}_1, \breve{f}_2, \breve{g}_1, \breve{g}_2 : V \to U(\mathbb{R})$, where V is a real vector space. We say that $(\breve{f}_1, \breve{g}_1) \preceq (\breve{f}_2, \breve{g}_2)$ if and

---

$$\pi((1\{s\}\times\xi)+(1\{t\}\times\zeta))$$
$$= ([(1\{s\}\times\xi)+(1\{t\}\times\zeta)]^{\alpha}_{pes},\ [((1\{s\}\times\xi)+(1\{t\}\times\zeta)]^{\alpha}_{opt})$$
$$= ([(1\{s\}\times\xi)^{\alpha}_{pes} + (1\{t\}\times\zeta)^{\alpha}_{pes}],\ [(1\{s\}\times\xi)^{\alpha}_{opt} + (1\{t\}\times\zeta)^{\alpha}_{opt}])$$
$$= (([1\{s\}]^{\alpha}_{pes}.\xi^{\alpha}_{pes} + [1\{t\}]^{\alpha}_{pes}.\zeta^{\alpha}_{pes}), ([1\{s\}]^{\alpha}_{opt}.\xi^{\alpha}_{opt} + [1\{t\}]^{\alpha}_{opt}.\zeta^{\alpha}_{opt})$$
$$= (s.\xi^{\alpha}_{pes} + t.\zeta^{\alpha}_{pes}, s.\xi^{\alpha}_{opt} + t.\zeta^{\alpha}_{opt}) = s\pi(\xi) + t\pi(\zeta).$$

only if $\breve{f}_1 \preceq \breve{f}_2, \breve{g}_1 \preceq \breve{g}_2$. We say that $(\breve{f}_1, \breve{g}_1) \prec (\breve{f}_2, \breve{g}_2)$ if and only if $\forall \alpha \in [0,1]$,

$(\breve{f}_1)^{\alpha}_{opt} \le (\breve{g}_1)^{\alpha}_{opt}$, $(\breve{f}_2)^{\alpha}_{pes} < (\breve{g}_2)^{\alpha}_{pes}$

or, $(\breve{f}_1)^{\alpha}_{opt} < (\breve{g}_1)^{\alpha}_{opt}$, $(\breve{f}_2)^{\alpha}_{pes} \le (\breve{g}_2)^{\alpha}_{pes}$

or, $(\breve{f}_1)^{\alpha}_{opt} < (\breve{g}_1)^{\alpha}_{opt}$, $(\breve{f}_2)^{\alpha}_{pes} < (\breve{g}_2)^{\alpha}_{pes}$.

Let $\breve{f}$ be a function defined by $\breve{f}: V \to U(\mathbb{R})$. Let $g_i(x)$, $i = 1,2,...,m$ be real-valued functions defined on the same real vector space *V* and *X* be any subspace of *V*. Then let us now consider the following optimization problem as follows.

$$\begin{cases} \min & \breve{f}(x) \\ \textit{subject to} & g_i(x) \le 0, \quad i = 1,2,...,m, \\ & x \in X. \end{cases} \tag{3.1}$$

Then x^* is an optimal solution of the problem (3.1) if there exists no x ($\neq x^*$) such that $\breve{f}(x) \prec \breve{f}(x^*)$.

Let π be the function defined in theorem 3.3. Now we consider the following optimization problem by applying the embedding function π to problem (3.1).

$$\begin{cases} \min & \pi(\breve{f}(x)) \\ \textit{subject to} & g_i(x) \le 0, \quad i = 1,2,...,m. \\ & x \in X. \end{cases} \tag{3.2}$$

Then x^* is an optimal solution of the problem (3.2) if there exists no x ($\neq x^*$) such that $\pi(\breve{f}(x)) < \pi(\breve{f}(x^*))$.

Therefore, x^* is an optimal solution of the problem (3.2) if there exists no x ($\neq x^*$) such that $((\breve{f}(x)^{\alpha}_{pes}, (\breve{f}(x)^{\alpha}_{opt}) < ((\breve{f}(x^*)^{\alpha}_{pes}, (\breve{f}(x^*)^{\alpha}_{opt})$, (follow definition (3.1).

Proposition 3.6 x^* is an optimal solution of the problem (3.1) if and only if x^* is an optimal solution of the problem (3.2).

Proof. Proposition 3.4 states that $\breve{f}(x) \prec \breve{f}(x^*)$ if and only if $\pi(\breve{f}(x)) < \pi(\breve{f}(x^*))$, and so, the proof is obvious.

Therefore, the optimal solution of problem (3.1) is same as the optimal solution of the following problem:

$$\begin{cases} \min & \{\breve{f}(x)^{\alpha}_{pes}, \breve{f}(x)^{\alpha}_{opt}\} \\ \textit{subject to} & g_i(x) \le 0, \quad i = 1,2,...,m. \\ & x \in X. \end{cases} \tag{3.3}$$

Proposition 3.7 If x^* is a Pareto optimal solution of the problem (3.3) for some $\alpha^* \in [0,1]$, then x^* is an optimal solution of the problem (3.2).

Proof. Since x^* is a Pareto optimal solution of the problem (3.3), x^* is a feasible solution of the problem (3.2). If possible, let x^* is not an optimal solution of problem (3.2). Then there exists a feasible solution x ($\neq x^*$) such that $\pi(\breve{f}(x)) < \pi(\breve{f}(x^*)) \forall \alpha \in [0,1]$.

Then $\forall \alpha \in [0,1]$, we have

either $\breve{f}(x)^{\alpha}_{pes} \le \breve{f}(x^*)^{\alpha}_{pes}$, $\breve{f}(x)^{\alpha}_{opt} < \breve{f}(x^*)^{\alpha}_{opt}$

or, $\breve{f}(x)^{\alpha}_{pes} < \breve{f}(x^*)^{\alpha}_{pes}$, $\breve{f}(x)^{\alpha}_{opt} \le \breve{f}(x^*)^{\alpha}_{opt}$

or, $\breve{f}(x)^{\alpha}_{pes} < \breve{f}(x^*)^{\alpha}_{pes}$, $\breve{f}(x)^{\alpha}_{opt} < \breve{f}(x^*)^{\alpha}_{opt}$.

Since $\alpha^* \in [0,1]$, we then should have, either

$\breve{f}(x)^{\alpha^*}_{pes} \le \breve{f}(x^*)^{\alpha^*}_{pes}$, $\breve{f}(x)^{\alpha^*}_{opt} < \breve{f}(x^*)^{\alpha^*}_{opt}$ s

or, $\breve{f}(x)^{\alpha^*}_{pes} < \breve{f}(x^*)^{\alpha^*}_{pes}$, $\breve{f}(x)^{\alpha^*}_{opt} \le \breve{f}(x^*)^{\alpha^*}_{opt}$

or, $\breve{f}(x)^{\alpha^*}_{pes} < \breve{f}(x^*)^{\alpha^*}_{pes}$, $\breve{f}(x)^{\alpha^*}_{opt} < \breve{f}(x^*)^{\alpha^*}_{opt}$.

This shows that x^* is not an optimal solution of the problem (3.3); a contradiction with the assumption that it is a pareto-optimal solution of (3.3). So, our assumption is wrong and we are with the theorem.

Theorem 3.8 If x^* is a Pareto optimal solution of the problem (3.3) for some $\alpha^* \in [0,1]$, then x^* is an optimal solution of the problem (3.1).

Proof. The theorem is obvious from proposition 3.6 and 3.7.

Now we consider the uncertain multi-objective optimization problem as follows:

$$\begin{cases} \min(\breve{f}_1(x), \breve{f}_2(x), ..., \breve{f}_n(x)) \\ \text{subject to } g_j(x) \le 0,\ j = 1,2,...,m \\ x \in X \end{cases} \tag{3.4}$$

$$\begin{cases} \min\{\pi(\breve{f}_1(x), \breve{f}_2(x), ..., \breve{f}_n(x))\} \\ = \min\{\pi(\breve{f}_1(x)), \pi(\breve{f}_2(x)), ..., \pi(\breve{f}_n(x))\} \\ = \min[\{\breve{f}_1(x)^{\alpha}_{pes}, \breve{f}_1(x)^{\alpha}_{opt}\}, \{\breve{f}_2(x)^{\alpha}_{pes}, \breve{f}_2(x)^{\alpha}_{opt}\}, ..., \\ \qquad \{\breve{f}_n(x)^{\alpha}_{pes}, \breve{f}_n(x)^{\alpha}_{opt}\}] \\ = \min[\breve{f}_1(x)^{\alpha}_{pes}, \breve{f}_1(x)^{\alpha}_{opt}, \breve{f}_2(x)^{\alpha}_{pes}, \breve{f}_2(x)^{\alpha}_{opt}, ..., \breve{f}_n(x)^{\alpha}_{pes}, \breve{f}_n(x)^{\alpha}_{opt}] \\ \text{subject to } g_j(x) \le 0,\ j = 1,2,....,m \\ x \in X,\ 0 \le \alpha \le 1 \end{cases} \tag{3.5}$$

We say that x^* is an optimal solution of problem (3.4) if there exists no $x \neq x^*$ such that $\tilde{f}_i(x) \prec (\tilde{f}_i(x^*)$.

Let π be the embedding function defined in theorem 3.3. Then we consider the multi-objective optimization problem (3.5) by applying the embedding function π to problem (3.4).

We say that x^* is an optimal solution of problem (3.5), if there exists no $x \neq x^*$ such that $\pi(\bar{f}_i(x)) < \pi((\bar{f}_i(x^*))$.

Theorem 3.9 If x^* is a Pareto optimal solution of (3.5) for some $\alpha^* \in [0,1]$, then x^* is an optimal solution of the uncertain multi-objective optimization problem (3.4).

***Note*:** To solve uncertain multi-objective problem by using embedding theorem we first have to transform the uncertain multi-objective optimization problem into the crisp multi-objective optimization problem (3.5). The Pareto optimal solution of this problem is the optimal solution of the original uncertain multi-objective problem.

4. Uncertain R & D Project Portfolio Selection model

In this section, we first describe the notations used in the construction of the R & D project portfolio selection model. Then the objective function of the models will be constructed in the second subsection. In the third subsection we will discuss the constraints used in our portfolio selection model. The next subsection will include final mathematical model.

4.1. Notations

N = Number of candidate projects.
T = Number of periods.
I = Interest rate.

$$x_{it} = \begin{cases} 1 \text{ if project i is selected in period } t \\ 0 \text{ otherwise} \end{cases},$$

$x = (x_{it})_{N \times T}$ =decision matrix.

$\breve{v}_{it}$ = Projected uncertain outcome of project i in period t.

$\breve{r}_{it}$ = Projected uncertain risk of implementing project i in period t.

$\breve{c}_{it}$ = Expected uncertain cost required by i^{th} project in period t.

B_t = Budget available for stage t.

R_{it}^s = Amount of resource of type s required for implementation of project i individually in period t.

$R_{st}^{'}$ = Amount of available resources of type s in period t.

$\bar{R}_s$ = Total amount of available resources of type s.

4.2. Formulation of Objective Functions

In this R & D project portfolio selection problem we have considered two objectives: maximization of the benefit and minimization of the project risk.

4.2.1. Maximization of Benefit

The total outcome from the projects will be obtained by considering the total individual. If the interest rate for each period is I, the total outcome is

$$\breve{Z}_O(x) = \sum_{t=1}^{T} \frac{1}{(1+I)^t} \sum_{i=1}^{N} \breve{v}_{it} x_{it}.$$

The total cost will be obtained by the total individual costs for each project. Then the total cost is

$$\breve{Z}_C(x) = \sum_{t=1}^{T} \sum_{i=1}^{N} \breve{c}_{it} x_{it}.$$

Thus the benefit of the project portfolio is

$$\breve{Z}_B(x) = \breve{Z}_O(x) - \breve{Z}_C(x) = \sum_{t=1}^{T} \sum_{i=1}^{N} [\frac{1}{(1+I)^t} \breve{v}_{it} x_{it} - \breve{c}_{it} x_{it}].$$

Our objective will be to maximize the benefit.

4.2.2. Minimization of Project Risk

For successfully implementation of R&D project portfolio, the risk attached with the projects must be as less as possible. Here, we have defined risk as the opposite of expected profit. As the futures of all the projects are uncertain, implementation of a project may or may not yield us success. In case of failure, the decision maker may loose their money and time and resource. Let $r_{it} \geq 0$ is the amount the decision maker may loose in worst case for the i^{th} project at period t. Then the total risk involved in the project portfolio is $\sum_{t=1}^{T} \sum_{i=1}^{N} \breve{r}_{it} x_{it}$.

Therefore, the objective is to minimize total risk

$$\breve{Z}_R(x) = \sum_{t=1}^{T} \sum_{i=1}^{N} \breve{r}_{it} x_{it}.$$

Thus we are with the following bi-objective optimization problem

$$\begin{cases} Max & \breve{Z}_B(x) \\ Min & \breve{Z}_R(x). \end{cases}$$

4.3. Formulation of the Constraints

In this subsection we will formulate the constraints required to model the problem realistically.

4.3.1. Outcome Constraints

As the minimum expected outcome for the projects in period t is V_t, we have,

$$\sum_{i=1}^{N} \breve{v}_{it} x_{it} \succeq V_t \ \forall\ t,$$

$$i.e.,\ \sum_{i=1}^{N} (\breve{v}_{it})^{\alpha}_{pes} x_{it} \geq V_t, \sum_{i=1}^{N} (\breve{v}_{it})^{\alpha}_{opt} x_{it} \geq V_t \ \forall\ t.$$

4.3.2. Resource Constraints

The projects are implemented by using limited amount of resources. As the available resources are always finite, the required resource with particular type should be within the resource available of that type for each period. Thus we have

$$\sum_{i=1}^{N} R^{s}_{it} x_{it} \leq R^{'}_{st} \ \forall s,t.$$

The total amount of resources available is limited. So, the amount of resource required should not be more than the total resource available for each type of resources. Thus we have,

$$\sum_{t=1}^{T} \left(\sum_{i=1}^{N} R^{s}_{it} x_{it} \right) \leq \bar{R}_s \ \forall\ s.$$

4.3.3. Budget Constraints

The project expenses during the planning horizon should not exceed the predetermined budget for each stage or period. So, we have

$$\sum_{i=1}^{N} \breve{c}_{it} x_{it} \preceq B_t \ \forall t,$$

$$i.e.,\ \sum_{i=1}^{N} (\breve{c}_{it})^{\alpha}_{pes} x_{it} \leq B_t, \sum_{i=1}^{N} (\breve{c}_{it})^{\alpha}_{opt} x_{it} \leq B_t \ \forall t.$$

4.4. The Set of Feasible Solutions

In this subsection we construct the set X of feasible solutions $x = (x_{it})_{N\times T}$. Then we have problem (4.1).

4.5. The R & D Project Portfolio Selection Mode *l*

Keeping in mind the objectives and constrained obtained in the last two subsections, the R & D project portfolio selection problem is modeled as

$$\begin{cases} Max & \breve{Z}_B(x) \\ Min & \breve{Z}_R(x) \\ s.t. & x \in X. \end{cases} \tag{4.2}$$

By applying embedding theorem, the uncertain multiob jective optimization problem (4.2) is converted into the crisp multi-objective problem

$$\begin{cases} Max & \{\breve{Z}_B(x)^{\alpha}_{pes}, \breve{Z}_B(x)^{\alpha}_{opt}\} \\ Min & \{\breve{Z}_R(x)^{\alpha}_{pes}, \breve{Z}_R(x)^{\alpha}_{opt}\} \\ s.t. & x \in X, 0 \leq \alpha \leq 1. \end{cases} \tag{4.3}$$

The global criteria methods developed in the context of multi-objective optimization problem are really handy for obtaining the Pareto optimal solution. Let,

$$B^{+}_{pes} = \max\{\breve{Z}_B(x)^{\alpha}_{pes}\}, B^{+}_{opt} = \max\{\breve{Z}_B(x)^{\alpha}_{opt}\},$$
$$R^{+}_{pes} = \min\{\breve{R}_B(x)^{\alpha}_{pes}\}, R^{+}_{opt} = \min\{\breve{R}_B(x)^{\alpha}_{opt}\},$$
$$B^{-}_{pes} = \min\{\breve{Z}_B(x)^{\alpha}_{pes}\}, B^{-}_{opt} = \min\{\breve{Z}_B(x)^{\alpha}_{opt}\},$$
$$R^{-}_{pes} = \max\{\breve{R}_B(x)^{\alpha}_{pes}\}, R^{-}_{opt} = \max\{\breve{R}_B(x)^{\alpha}_{opt}\}.$$

Then the problem (4.3) is further converted into the following single objective convex programming problem.

$$\begin{cases} Min \left[\left(\dfrac{B^{+}_{pes} - \breve{Z}_B(x)^{\alpha}_{pes}}{B^{+}_{pes} - B^{-}_{pes}} \right)^2 + \left(\dfrac{B^{+}_{opt} - \breve{Z}_B(x)^{\alpha}_{opt}}{B^{+}_{opt} - B^{-}_{opt}} \right)^2 + \right. \\ \left. \left(\dfrac{\breve{R}_B(x)^{\alpha}_{pes} - R^{+}_{pes}}{R^{-}_{pes} - R^{+}_{pes}} \right)^2 + \left(\dfrac{\breve{R}_B(x)^{\alpha}_{opt} - R^{+}_{opt}}{R^{-}_{opt} - R^{+}_{opt}} \right)^2 \right]^{\frac{1}{2}} \\ such\ \text{that} \qquad x \in X, 0 \leq \alpha \leq 1. \end{cases} \tag{4.4}$$

5. Case Study

In this section a model is developed and solved based on data from the large scale organization B. M. Enterprise, Berhampore, West Bengal, India. The R & D wing of this organization is involved in different structural works in civil, mechanical and electrical fields. During the year 2009 the organization gets 10 project proposals from private as well as public sectors. All the proposals accompany data on the estimated outcome, estimated cost, funds, workers, budget and risk. After first round of

$$X = \left\{ x = (x_{it})_{N\times T} : \sum_{i=1}^{N} (\breve{v}_{it})^{\alpha}_{pes} x_{it} \leq V_t, \sum_{i=1}^{N} (\breve{v}_{it})^{\alpha}_{opt} x_{it} \leq V_t, \sum_{i=1}^{N} R^{s}_{it} x_{it} \leq R^{'}_{st}, \right.$$
$$\left. \sum_{i=1}^{N} (\breve{c}_{it})^{\alpha}_{pes} x_{it} \leq B_t, \ \sum_{i=1}^{N} (\breve{c}_{it})^{\alpha}_{opt} x_{it} \leq B_t, \sum_{t=1}^{T} \left(\sum_{i=1}^{N} R^{s}_{it} x_{it} \right) \leq \bar{R}_s \ \forall\ s,t \right\} \tag{4.1}$$

scrutiny 5 project proposals are short listed. All the five projects are scheduled over two periods and each period lasts one year. They are renamed as projects I, II, III, IV and V due to privacy. The estimated outcome, risk and projected cost are considered in the form of triangular uncertain variables. Interest rate is 5%. The estimated data for outcome, costs, risks, funds and workers are given in **Table 1**.

Constraints on fund, workers and budget for each period are given in **Table 2**.

As discussed in Section 4, the uncertain optimization model (4.2) is constructed which is converted into the model (4.3) by using the embedding theorem. The model (4.4) is then constructed as

$$\begin{cases} Min\left[\left(\dfrac{16.000-\breve{Z}_B(x)^{\alpha}_{pes}}{5.375}\right)^2+\left(\dfrac{18.475-\breve{Z}_B(x)^{\alpha}_{opt}}{7.375}\right)^2\right. \\ \qquad \left.+\left(\dfrac{\breve{R}_B(x)^{\alpha}_{pes}-2.075}{1.325}\right)^2+\left(\dfrac{\breve{R}_B(x)^{\alpha}_{opt}-2.2}{1.4}\right)^2\right]^{\frac{1}{2}} \quad (5.1) \\ \quad such\ \text{that} \qquad x\in X,\ 0\le\alpha\le 1. \end{cases}$$

The solution of the model (5.1) is done by using the *LINGO* software and the obtained solution is as follows: $x_{11} = 0$, $x_{21} = 1$, $x_{31} = 0$, $x_{41} = 1$, $x_{51} = 1$; $x_{12} = 0$, $x_{22} = 1$, $x_{32} = 0$, $x_{42} = 0$, $x_{52} = 1$.

It means that B. M. Enterprise should select the 2nd, 4th and 5th projects in 1st stage and 2nd and 5th projects in the second stage to get the optimum result.

For this portfolio, the benefit is estimated as (6.5, 11.4, 13.7) million rupees and the corresponding risk is (1.95, 2.4, 3.2) million rupees.

6. Conclusions

This paper introduces the concept of multiple objective uncertain optimization problems. In particular, this paper concentrates on the problems where the coefficients of the decision variables are uncertain variables. To do so, we propose and prove the uncertain embedding theorem from the space of uncertain variables to the Banach space C [0, 1] × C [0, 1]. By applying embedding theorem, each uncertain objective function can be converted into two deterministic objectives functions. The Pareto optimal solution of both the deterministic objectives is the optimal solution of the uncertain objective.

Table 1. Estimated project data.

Projects		I	II	III	IV	V
Outcome (in Million Rupees)	1st period	(4, 7, 9)	(1, 3, 5)	(0.2, 1.4, 2.8)	(0, 1, 1.4)	(5, 6, 7)
	2nd period	(7, 10, 12)	(2, 3, 4)	(2.5, 4, 5.2)	(1, 2, 3)	(6, 7.5, 9)
Cost (in Million Rupees)	1st period	(1, 2.2, 3)	(0.4, 1.2, 2)	(0.8,1, 1.2)	(0.1, 0.3, 0.8)	(2, 2.5, 2.9)
	2nd period	(2, 2.8, 3.7)	(1.0, 1.1, 2)	(1.5, 1.7, 2)	(1.6, 1.8, 2.1)	(3, 4, 5)
Risk (in Million Rupees)	1st period	(0.6, 0.8, 1)	(0.1, 0.2, 0.4)	(0.5, 0.7, 0.9)	(0.4, 0.5, 0.6)	(0.35, 0.4, 0.5)
	2nd period	(0, 0.4, 0.5)	(0.6, 0.7, 1)	(0.4, 0.5, 0.6)	(0.4, 0.5, 0.6)	(0.5, 0.6, 0.7)
Fund (in Million Rupees)	1st period	0.6	0.4	0.25	0.11	0.21
	2nd period	0.9	0.2	0.21	0.09	0.2
Workers (in numbers)	1st period	31	15	20	21	16
	2nd period	10	12	17	18	9

Table 2. Constraints.

Category	1st period	2nd period	Total
Outcome (in Million Rupees)	> 10.0	> 10.0	-
Budget (in Million Rupees)	< 8.0	< 9.0	-
Fund (in Million Rupees)	< 0.9	< 0.8	< 1.5
Workers (in numbers)	< 85	< 80	< 150

This paper also introduces a new model of R & D project portfolio selection by identifying project information like estimated future outcome, risk or estimated cost of the projects as uncertain variables. An uncertain bi-objective optimization model, that maximizes the benefit and minimizes the risk, is constructed. Constraints on budget, resources and outcomes are also included in the model to make it more realistic. The uncertain optimization method by embedding theorem is used to solve it. A real case study is provided for illustration.

In future, we will use the uncertain optimization approach to other real optimization problems like portfolio selection problem, supply chain management problem, poverty management problem etc.

For large data sets, meta-heuristic algorithms such as tabu search, simulated annealing, ant-colony optimization, and particle swarm optimization may be employed to solve the non-linear programming problem (4.4).

7. References

[1] M. L. Puri and D. A. Ralescu, "Differentials for Fuzzy Functions," *Journal of Mathematical Analysis and its Application*, Vol. 91, No. 2, 1983, pp. 552-558.

[2] O. Kaleva, "The Cauchy Problem for Fuzzy Differential Equations," *Fuzzy Sets and Systems*, Vol. 35, No. 3, 1990, pp. 389-396.

[3] C. X. Wu and M. Ma, "Embedding Problem of Fuzzy Number Space: Part I," *Fuzzy Sets and Systems*, Vol. 44, No. 1, 1991, pp. 33-38.

[4] C. X. Wu, "An (α, β)-Optimal Solution Concept in Fuzzy Optimization Problems," *Optimization*, Vol. 53, No. 2, 2004, pp. 203-221.

[5] C. X. Wu, "Evaluate Fuzzy Optimization Problems Based on Bi Objective Programming Problems," *Computer and Mathematics with Applications*, Vol. 47, No. 5, 2004, pp. 893-902.

[6] M. Rabbani, M. A. Bajestani and G. B. Khoshkhou, "A Multi-Objective Particle Swarm Optimization for Project Selection Problem," *Expert Systems with Applications*, Vol. 37, No. 1, 2010, pp. 315-321.

[7] M. Rabbani, R. T. Moghaddam, F. Jolai and H. R. Ghorbani, "A Comprehensive Model for R and D Project Portfolio Selection with Zero - One Linear Goal Programming," *IJE Transactions A*: *Basic*, Vol. 19, No. 1, 2006, pp. 55-66.

[8] Y. Fang, L. Chen and M. Fukushima, "A Mixed R & D Projects and Securities Portfolio Selection Model," *European Journal of Operational Research*, Vol. 185, No. 2, 2008, pp. 700-715.

[9] S. Riddell and W. A. Wallace, "The Use of Fuzzy Logic and Expert Judgment in the R & D Project Portfolio Selection Process," *Proceedings*: *PICMET*, Portland, 2007, pp. 1228- 1238.

[10] H. Eilat, B. Golany and A. Shtub, "Constructing and Evaluating Balanced Portfolios of R & D Projects with Interactions: A DEA Based Methodology," *European Journal of Operational Research*, Vol. 172, No. 3, 2006, pp. 1018-1039.

[11] C. Stummer and K. Heidenberger, "Interactive R&D Portfolio Selection Considers Multiple Objectives, Project Interdependencies, and Time: A Three Phase Approach," *Proceedings*: *PICMET*, Portland, 2001, pp. 423-428.

[12] J. D. Linton, S. T. Walsh, B. A. Kirchhoff, J. Morabito and M. Merges, "Selection of R & D Projects in a Portfolio," *Proceedings of IEEE Engineering Management Society*, Washington, 2000, pp. 506-511.

[13] J. L. Ringuest, S. B. Graves and R. H. Case, "Conditional Stochastic Dominance in R & D Portfolio Selection," *IEEE Transactions on Engineering Management*, Vol. 47, No. 4, 2000, pp. 478-484.

[14] R. L. Schmidt, "A model for R & D Project Selection with Combined Benefit, Outcome, and Resource Interactions," *IEEE Transactions on Engineering Management*, Vol. 40, No. 4, 1993, pp. 403-410.

[15] O. Pereira and D. Junior, "The R & D Project Selection Problem with Fuzzy Coefficients," *Fuzzy Sets and Systems*, Vol. 26, No. 3, 1988, pp. 299-316.

[16] M. A. Coffin and B. W. Taylor, "Multiple Criteria R & D Selection and Scheduling Using Fuzzy Logic," *Computers Operations Researches*, Vol. 23, No. 3, 1996, pp. 207-220.

[17] L. L. Machacha and P. Bhattacharya, "A Fuzzy-Logic-Based Approach to Project Selection," *IEEE Transactions on Engineering Management*, Vol. 47, No. 1, 2000, pp. 65-73.

[18] D. Kuchta, "A Fuzzy Mode for R & D Project Selection with Benefit: Outcome and Resource," *The Engineering Economist*, Vol. 46, No. 3, 2001, pp. 164-180.

[19] S. Mohamed and A. K. McCowan, "Modelling Project Investment Decisions under Uncertainty Using Possibility Theory," *International Journal of Project Management*, Vol. 19, No. 4, 2001, pp. 231-241.

[20] Y. G. Hsu, G. H. Tzeng and J. Z. Shyu, "Fuzzy Multiple Criteria Selection of Government-Sponsored Frontier Technology R & D Projects," *R & D Management*, Vol. 33, No. 5, 2003, pp. 539-551.

[21] J. Wang and W. L. Hwang, "A Fuzzy Set Approach for R&D Portfolio Selection Using a Real Options Valuation Model," *Omega*, Vol. 35, No. 3, 2007, pp. 247-257.

[22] S. S. Kim, Y. Choi, N. M. Thang, E. R. Ramos and W. J. Hwang, "Development of a Project Selection Method on Information System Using ANP and Fuzzy Logic," World Academy of Science, *Engineering and Technology*, Vol. 53, No. 6, 2009, pp. 411-415.

[23] E. E. Karsak, "A Generalized Fuzzy Optimization Framework for R&D Project Selection Using Real Options Valuation," *Proceedings of ICCCSA*, Berlin, Heidelberg, New York, 2006, pp. 918-927.

[24] B. Liu, "Uncertainty Theory," 2nd Edition, Springer-

Verlag, Berlin, 2007.

[25] B. Liu, "Fuzzy Process, Hybrid Process and Uncertain Process," *Journal of Uncertain Systems*, Vol. 2, No. 1, 2008, pp. 3-16.

[26] B. Liu, "Some Research Problems in Uncertainty Theory," *Journal of Uncertain Systems*, Vol. 3, No. 1, 2009, pp. 3-10.

[27] X. Li and B. Liu, "Hybrid Logic and Uncertain Logic," *Journal of Uncertain Systems*, Vol. 3, No. 2, 2009, pp. 83-94.

[28] B. Liu, "Uncertainty Theory," 3rd Edition. http://orsc.edu.cn/liu/ut.pdf

[29] X. Gao, "Some Properties of Continuous Uncertain Measure," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 18, No. 4, 2007, pp. 383-390.

[30] C. You, "Some Convergence Theorems of Uncertain Sequences," *Mathematical and Computer Modelling*, Vol. 49, No. 3-4, 2009, pp. 482-487.

Applied Mathematics, **2010, 1, 200-210**
doi:10.4236/am.2010.13024 Published Online September 2010 (http://www.SciRP.org/journal/am)

On Complete Bicubic Fractal Splines

Arya Kumar Bedabrata Chand[1], María Antonia Navascués[2]
[1]*Department of Mathematics, Indian Institute of Technology Madras, Chennai, India*
[2]*Departmento de Matemática Aplicada, Centro Politécnico Superior de Ingenieros, Universidad de Zaragoza, Zaragoza, Spain*
E-mail: *chand@iitm.ac.in, manavas@unizar.es*
Received June 6, 2010; *revised July* 20, 2010; *accepted July* 23, 2010

Abstract

Fractal geometry provides a new insight to the approximation and modelling of experimental data. We give the construction of complete cubic fractal splines from a suitable basis and their error bounds with the original function. These univariate properties are then used to investigate complete bicubic fractal splines over a rectangle Ω. Bicubic fractal splines are invariant in all scales and they generalize classical bicubic splines. Finally, for an original function $f \in C^4[\Omega]$, upper bounds of the error for the complete bicubic fractal splines and derivatives are deduced. The effect of equal and non-equal scaling vectors on complete bicubic fractal splines were illustrated with suitably chosen examples.

Keywords: Fractals, Iterated Function Systems, Fractal Interpolation Functions, Fractal Splines, Surface Approximation.

1. Introduction

Schoenberg [1] introduced "spline functions" to the mathematical literature. In the last 60 years, splines have proved to be enormously important in different branches of mathematics such as numerical analysis, numerical treatment of differential, integral and partial differential equations, approximation theory and statistics. Also, splines play major roles in field of applications, such as CAGD, tomography, surgery, animation and manufacturing. In this paper, we discuss on complete fractal splines that generalize the classical complete splines.

Fractal interpolation functions (FIFs) were introduced by Barnsley [2,3] based on the theory of iterated function system (IFS). The attractor of the IFS is the graph of FIF that interpolates a given set of data points. Fractal interpolation constitutes an advance in the sense that the functions used are not necessarily differentiable and show the rough aspect of real-world signals [3,4]. A specific feature is the fact that the graph of these interpolants possesses a fractal dimension and this parameter provides a geometric characterization of the measured variable which may be used as an index of the complexity of a phenomenon. Barnsley and Harrington [5] first constructed a differentiable FIF or C^r-FIF f that interpolates the prescribed data if values of $f^{(k)}$, $k = 1, 2, \ldots, r$, at the initial end point of the interval are given. In this construction, specifying boundary conditions similar to those of classical splines was found to be quite difficult to handle. The fractal splines with general boundary conditions are studied recently [6,7]. The power of fractal methodology allows us to generalize almost any other interpolation techniques, see for instance [8-10].

Fractal surfaces have proved to be useful functions in scientific applications such as metallurgy, physics, geology, image processing and computer graphics. Massopust [11] was first to put forward the construction of fractal interpolation surfaces (FISs) on triangular domains, where the interpolation points on the boundary of the domain are coplanar. Geronimo and Hardin [12], and Zhao [13] generalized this construction in different ways. The general bivariate FIS on rectangular grids are treated for instance in references [14,15]. Recently, Bouboulis and Dalla constructed fractal interpolation surfaces from FIFs through recurrent iterated function systems [16].

In this paper we approach the problem of complete cubic spline surface from a fractal perspective. In Section 2, we construct cardinal cubic fractal splines through moments and estimate the error bound of the complete cubic spline with the original function. The construction of bicubic fractal splines is carried out in Section 3 through tensor products. Finally, for an original function $f \in C^4[\Omega]$, upper bounds of the error for the complete

bicubic fractal splines and derivatives are deduced. The effect of scaling factors on bicubic fractal splines are demonstrated in the last section through various examples.

2. Complete Cubic Fractal Splines

We discuss on fractal interpolation based on IFS theory in Subsection 2.1 and construct cardinal cubic fractal spline through moments in Subsection 2.2. Upper bounds of L_∞-norm of the error of a complete cubic spline FIF with respect to the original function are deduced in Subsection 2.3.

2.1. Fractal Interpolation Functions

Let $\Delta_t : t_0 < t_1 < \ldots < t_N$ be a partition of the real compact interval $I = [t_0, t_N]$. Let a set of data points $\mathcal{D} = \{(t_n, x_n) \in I \times \mathbb{R} : n = 0,1,2,\ldots,N\}$ be given. Set $I_n = [t_{n-1}, t_n]$ and let $L_n : I \to I_n, \quad n = 1,2,\ldots,N$ be contractive homeomorphisms such that

$$\begin{aligned} &L_n(t_0) = t_{n-1},\ L_n(t_N) = t_n, \\ &\left|L_n(c_1) - L_n(c_2)\right| \le l\,|c_1 - c_2| \quad \forall\, c_1, c_2 \in I, \end{aligned} \tag{1}$$

for some $0 \le l < 1$. Let $C = I \times \mathbb{R}$ and N continuous mappings, $F_n : C \to \mathbb{R}$, satisfying

$$\begin{aligned} &F_n(t_0, x_0) = x_{n-1},\ F_n(t_N, x_N) = x_n, \\ &n = 1,2,\ldots,N, \\ &\left|F_n(t,x) - F_n(t,y)\right| \le |\alpha_n|\,|x - y|, \\ &t \in I, \quad x, y \in \mathbb{R}, \quad -1 < \alpha_n < 1. \end{aligned} \tag{2}$$

Now, define functions

$\forall\, n = 1,2,\ldots,N, w_n(t,x) = \left(L_n(t), F_n(t,x)\right), w_n : C \to I_n \times \mathbb{R}.$

Proposition 2.1 (Barnsley [2]) *The Iterated Function System* (*IFS*) $\{C; w_n : n = 1,2,\ldots,N\}$ *defined above admits a unique attractor* G. G *is the graph of a continuous function* $f : I \to \mathbb{R}$ *which obeys* $f(t_n) = x_n$ *for* $n = 0,1,2,\ldots,N$.

The above function f is called a Fractal Interpolation Function (FIF) corresponding to the IFS $\{(L_n(t), F_n(t,x))\}_{n=1}^N$. Let

$\mathcal{G} = \{g : I \to \mathbb{R} \mid g \text{ iscontinuous}, g(t_0) = x_0 \text{ and } g(t_N) = x_N\}$

$\mathcal{G}$ is a complete metric space respect to the uniform norm. Define, the Read-Bajraktarevi c′ operator T on $(\mathcal{G}, \|.\|_\infty)$ by

$$Tg(t) = F_n\left(L_n^{-1}(t), g\left(L_n^{-1}(t)\right)\right), t \in I_n, n = 1,2,\ldots,N. \tag{3}$$

According to (1)-(2), Tg is continuous on the interval $I_n; n = 1,2,\ldots,N$ and at each of the points t_1, $t_2,\ldots,t_{N-1}$. T is a contraction mapping on the metric space $(\mathcal{G}, \|.\|_\infty)$ *i.e.*

$$\|Tf - Tg\|_\infty \le |\alpha|_\infty \|f - g\|_\infty, \tag{4}$$

where $|\alpha|_\infty = \max\{|\alpha_n| : n = 1,2,\ldots,N\}$. Since $|\alpha|_\infty < 1$, T possesses a unique fixed point f (say) on $\mathcal{G}$, that is to say, there is $f \in \mathcal{G}$ such that $(Tf)(t) = f(t)\ \forall\, t \in I$. This function is the FIF corresponding to w_n and according to (3), the FIF satisfies the functional equation:

$$f(t) = F_n\left(L_n^{-1}(t), f \circ L_n^{-1}(t)\right), \ t \in I_n, \ n = 1,2,\ldots,N. \tag{5}$$

The most widely studied fractal interpolation functions so far are defined by the IFS

$$\begin{cases} L_n(t) = a_n t + b_n \\ F_n(t,x) = \alpha_n x + q_n(t) \end{cases} \tag{6}$$

where $-1 < \alpha_n < 1$ and $q_n : I \to \mathbb{R}$ are suitable continuous functions such that (2)are satisfied. α_n is called a vertical scaling factor of the transformation w_n and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)$ is the scale vector of IFS. The scale factors give a degree of freedom to the FIF and allow us to modify its properties. If $q_n(t)$ are affine in (6) for $t \in I$, then the FIF is called affine [3]. Based on the principle of construction [6] of a C^r-FIF, $r \in \mathbb{N}$, complete cardinal cubic fractal splines are constructed through their moments in the following.

2.2. Complete Cardinal Cubic Fractal Splines

A cubic spline is called complete if the values of its first derivative are prescribed at the end points. A function $h(t)$ defined on the grid $\Delta_t : t_0 < t_1 < \ldots t_N$ is called an interpolating cubic spline function if the function. 1) is a cubic polynomial on each partial segment $[t_{n-1}, t_n]$, $n = 1,2,\ldots,N$. 2) the function is of class $C^2[t_0, t_2]$. 3) satisfies the conditions $h(t_n) = x_n$, $n = 0,1,\ldots,N$. Two conditions are given in the form of restriction on the spline values and/or the values of its derivatives at the end points of the segment $[t_0, t_N]$.

Definition 2.1 *A function* $f_m(t)$ *is called a cardinal cubic fractal spline if* 1) f_m *is a FIF associated with the set of data points* $(t_n, \delta_{m,n})$ *with mesh* Δ_t, *that is to say*

$$\left.\begin{aligned} f_m(t_n) = \delta_{m,n} \quad &= 1, \quad m = n, \\ &= 0, \quad m \ne n, \end{aligned}\right\} \forall\, m, n = 0,1,2,\ldots,N. \tag{7}$$

Besides, 2) $f_m \in C^2[t_0, t_N]$, 3) *the corresponding IFS* $\omega_{m,n}(t,x) = (L_n(t), F_{m,n}(t,x))$ *is such that* $L_n(t)$ *is defined by* (6) *and* $F_{m,n}(t,x) = a_n^2 \alpha_{m,n} x + a_n^2 q_{m,n}(t), |\alpha_{m,n}| < 1$, *where* $q_{m,n}(t)$ *is a suitable cubic polynomial so that the polynomial associated with the fractal function* f_m'' *on the mesh* Δ_t *is affine.*

In the construction of cardinal cubic fractal splines, we have taken $\alpha_{m,n} = \alpha_n$, $n = 1,2,\ldots,N$; $m = 0,1,2,\ldots,N$.

A derivation of the defining equations for a cubic fractal spline through moments $M_{m,n} = f_m''(t_n)\ n = 0,1,\ldots,N$ can be found in [6], but for completeness and to set the terminology, it is outlined in the appendix.

For a basis of complete cubic fractal spline space on *I*, we need $f'_m(t_0) = f'_m(t_N) = 0$ for $m = 0,1,\ldots,N$ in the construction of cubic spline FIF, and two more complete cubic fractal splines f_{-1} and f_{N+1} such that

$$\begin{aligned} &f_{-1}(t_n) = 0, n = 0,1,\ldots,N;\ f'_{-1}(t_0) = 1,\ f'_{-1}(t_N) = 0,\\ &f_{N+1}(t_n) = 0, n = 0,1,\ldots,N;\ f'_{N+1}(t_0) = 0,\ f'_{N+1}(t_N) = 1. \end{aligned} \tag{8}$$

Denote, $x_{-1} = f(t_{-1}) = f'(t_0),\ x_{N+1} = f(t_{N+1}) = f'(t_N)$. Let f be the original function providing the data $\{(t_n, x_n)\}_{n=-1}^{N+1}$ and f_c be the complete cubic fractal spline corresponding to this data. Let $\mathcal{U}(I,\Delta_t) = \{h \mid h$ is a complete cubic fractal spline on $\Delta_t\}$. If $h \in \mathcal{U}(I,\Delta_t)$ interpolating the same data $\{(t_n, x_n)\}_{n=0}^{N}$, then due to the uniqueness of fixed point of Read-Bajraktarevi c′ operator, $h(t) = \sum_{m=-1}^{N+1} x_m f_m(t)$. Also, none of the f_m is a linear combination of other cardinal splines and hence $\{f_m\}_{m=-1}^{N+1}$ is a basis for $\mathcal{U}(I,\Delta_t)$. Define a complete cubic fractal spline operator $\mathcal{F}: C^2(I) \to \mathcal{U}(I,\Delta_t)$ as $\mathcal{F}(f) = f_c$ such that

$$\begin{aligned} f_c\left(L_n(t)\right) &= \sum_{m=-1}^{N+1} f(t_m)\, f_m\left(L_n(t)\right)\\ &= \sum_{m=-1}^{N+1} x_m f_m\left(L_n(t)\right), t \in I, n = 1,2,\ldots,N. \end{aligned} \tag{9}$$

It is easy to check that $\mathcal{F}$ is linear and bounded operator on $C^2(I)$. According to ((7)) and ((8)), we have

$$f_c(t_0) = f_c(L_1(t_0)) = \sum_{m=-1}^{N+1} x_m f_m(L_1(t_0)) = \sum_{m=-1}^{N+1} x_m \delta_{m,0} = x_0$$

and for $i = 1,2,\ldots,N$,

$$\begin{aligned} f_c(t_n) &= f_c\left(L_n(t_N)\right) = \sum_{m=-1}^{N+1} x_m f_m\left(L_N(t_N)\right)\\ &= \sum_{m=-1}^{N+1} x_m f_m(x_n) = \sum_{n=-1}^{N+1} x_m \delta_{m,n} = x_n \end{aligned}$$

Also, $f_c'(t_0) = \sum_{m=-1}^{N+1} x_m f_m'(t_0) = x_{-1} = f'(t_0)$ and $f_c'(t_N) = \sum_{m=-1}^{N+1} x_m f_m'(t_N) = x_{N+1} = f'(t_N)$. If we choose $\alpha_n = 0; n = 1,2,\ldots,N$, then from (26), it is clear that right side of cardinal spline f_m reduces to a cubic polynomial in t and hence, in this case f_m reduces to a classical complete cardinal spline S_m such that $S_m(t_n) = \delta_{m,n}$. The classical complete cubic spline $S(t)$ for the data $\{(t_n, x_n)\}_{n=0}^{N}$ is given by

$$S\left(L_n(t)\right) = \sum_{m=-1}^{N+1} x_m S_m\left(L_n(t)\right), t \in I, n = 1,2,\ldots,N. \tag{10}$$

2.3. Error Estimation with Univariate Fractal Splines

To estimate error bounds for the complete bicubic fractal spline, we need error bounds between a cubic fractal spline and the original function $f \in C^p(I)$, $p = 2,3,4$. For given moments $\{M_{m,n}\}_{n=0}^{N}$, we can observe that $q_{m,n}$ is a function of the scaling factors $\alpha_n; n = 1,2,\ldots,N$ for the cubic fractal spline equation (cf. (26)). We need the following proposition with the assumption $|\alpha_n| \le \kappa < 1$, for fixed κ.

Proposition 2.2 *Let f_m and S_m $(m = -1,0,\ldots,N+1)$ be the cardinal cubic fractal spline and the classical cardinal cubic spline respectively to the same set of data $\{(t_m, \delta_{m,n})\}_{m=0}^{N}$. Let $h_t = \max\{t_n - t_{n-1} : n = 1,2,\ldots,N\}$, $|\alpha|_\infty = \max\{|\alpha_n| : n = 1,2,\ldots,N\}$, and $|I|$ is the length of the interval I. Suppose the cubic polynomial $q_{m,n}(\alpha_n, t)$ associated with the IFS corresponding to the cardinal fractal spline f_m satisfies*

$$\left|\frac{\partial^{1+u} q_{m,n}(\tau_n, t)}{\partial \alpha_n \partial t^u}\right| \le \Theta_{u,m}$$

for $|\tau_n| \in (0, \kappa a_m^u), t \in I_n, u = 0,1,2$ and $n = 1,2,\ldots,N$. Then,

$$\left\| f_m^{(u)} - S_m^{(u)} \right\|_\infty \le \frac{h_t^{2-u} |\alpha|_\infty}{|I|^{2-u} - h_t^{2-u} |\alpha|_\infty} \left(\left\| S_m^{(u)} \right\|_\infty + \Theta_{u,m} \right), \tag{11}$$

$u = 0,1,2$.

The proof of the above proposition can be seen in [6]. Now, we derive an upper bound for the error between the classical complete cubic spline and a complete cubic fractal spline for the same set of interpolation data. According to (9) and (10), we get the bottom equation

$$\begin{aligned} \left| f_c^{(u)}\left(L_n(t)\right) - S^{(u)}\left(L_n(t)\right) \right| &= \left| \sum_{m=-1}^{N+1} f(t_m) \left(f_m^{(u)} - S_m^{(u)} \right) \left(L_n(t)\right) \right| \le \sum_{m=-1}^{N+1} \|f\|_1 \left\| f_m^{(u)} - S_m^{(u)} \right\|_\infty\\ &\le \sum_{m=-1}^{N+1} \|f\|_1 \frac{h_t^{2-u} |\alpha|_\infty}{|I|^{2-u} - h_t^{2-u} |\alpha|_\infty} \left(\Lambda_u + \Theta_u \right),\ u = 0,1,2, \end{aligned}$$

where 1-norm of f is $\|f\|_1 = Max\{\|f\|_\infty, \|f'\|_\infty\}$, $\Lambda_u = \max\{\|S_m^{(u)}\|_\infty : m = -1, 0, \ldots, N+1\}$ and $\Theta_u = \max\{\Theta_{u,m} : m = -1, 0, \ldots, N+1\}$. Set, $\Gamma_{u,\alpha,N} = \dfrac{|\alpha|_\infty (N+3)(\Lambda_u + \Theta_u)}{|I|^{2-u} - h_t^{2-u}|\alpha|_\infty}$. Since the above inequality is true for $n = 1, 2, \ldots, N$, we have the following estimate.

$$\left\| f_c^{(u)} - S^{(u)} \right\|_\infty \le \|f\|_1 \Gamma_{u,\alpha,N} h_t^{2-u}, \ u = 0,1,2. \quad (12)$$

We need the error bound between the complete cubic fractal spline f_c and the original function $f \in C^p(I)$, $p = 2,3,4$.

Proposition 2.3 [17] *Let S be the complete cubic spline interpolant of $f \in C^p[t_0, t_N]$ for $p = 2, 3,$ or 4 with the assumption $h_t \le 1$. Then*

$$\left\| (S-f)^{(u)} \right\|_\infty \le \varepsilon_{p,u} \left\| f^{(p)} \right\|_\infty h_t^{p-u}, \ 0 \le u \le \min\{p, 3\}, \quad (13)$$

where $\varepsilon_{p,u}$ are given in **Table 1** with $\eta_t = h_t / \min(t_n - t_{n-1})$.

If the original function f is such that $f^n \in C^p[t_0, t_N]$ with p-norm $\|f\|_p = Max\{\|f\|_\infty, \|f'\|_\infty, \ldots, \|f^{(p)}\|_\infty\}$, according to (12) and (13), we have the following upper bound estimation for the error.

$$\left\| (f_c - f)^{(u)} \right\|_\infty \le \|f\|_1 \Gamma_{u,\alpha,N} h_t^{2-u} + \left\| f^{(p)} \right\|_\infty \varepsilon_{p,u} h_t^{p-u}, \quad 0 \le u \le 2,$$

and

$$\left\| (f_c - f)^{(u)} \right\|_\infty \le \|f\|_p \left(\Gamma_{u,\alpha,N} h_t^{2-u} + \varepsilon_{p,u} h_t^{p-u} \right), \ 0 \le u \le 2. \quad (14)$$

3. Fractal Splines in Two Variables

Using univariate complete cubic fractal spline results, we construct complete bicubic fractal splines in Subsection 3.1 through tensor product and the upper bounds of the L_∞-norm of its error with the original functions in Subsection 3.2.

3.1. Construction of Complete Bicubic Fractal Splines

Suppose that $\Delta_t : a = t_0 < t_1 < \ldots < t_N = b$ and $\Delta_s : c = s_0 < s_1 < \ldots < s_J = d$ orm a rectangular mesh $\pi : \Delta_t \times \Delta_s$ for a rectangular region $\Omega = [a,b] \times [c,d]$. Let $f(t,s)$ be a sufficiently smooth function in the domain Ω. Let $\rho_f(t,s)$ be the complete bicubic spline fractal interpolation surface associated with the function $f(t,s)$ and the mesh π. Then, $\rho_f(t,s)$ is a tensor product of univariate cubic fractal splines such that

$$\begin{aligned}
&\rho_f(t_n, s_j) = f(t_n, s_j); \ n = 0,1,\ldots,N; \ j = 0,1,\ldots,J,\\
&\rho_f^{(1,0)}(t_n, s_j) = f^{(1,0)}(t_n, s_j); \ n = 0, N; \ j = 0,1,\ldots,J,\\
&\rho_f^{(0,1)}(t_n, s_j) = f^{(0,1)}(t_n, s_j); \ n = 0,1,\ldots,N; \ j = 0, J,\\
&\rho_f^{(1,1)}(t_n, s_j) = f^{(1,1)}(t_n, s_j); \ n = 0, N, \ j = 0, J,
\end{aligned} \quad (15)$$

where $\rho_f^{(\mu,\nu)} = \partial^{\mu+\nu} \rho_f / \partial t^\mu \partial s^\nu$. This definition is analogous to that of the classical complete bicubic spline in [19]. In the construction, we need two sets of nodal bases for univariate cubic fractal splines. Let $\{f_m(t)\}_{m=-1}^{N+1}$ be a nodal basis for the complete cubic fractal spline space $\mathcal{U}([a,b], \Delta_t)$ (cf. Section 2) and $\{\tilde{f}_i(s)\}_{i=-1}^{J+1}$ be a nodal basis for the complete cubic fractal spline space $\mathcal{V}([c,d], \Delta_s)$ with a choice of scaling parameters $\beta_j, j = 1, \ldots, J$ and $\tilde{L}_j : [c,d] \to [s_{j-1}, s_j]$, $\tilde{L}_j(s) = c_j s + d_j$, where $|c_j| < 1$ for $j = 1,2,\ldots,J$. Define generic transformations P and Q on $C^2[\Omega]$ respectively as

$$P\left(g\left(L_n(t), \tilde{L}_j(s)\right)\right) = \sum_{m=-1}^{N+1} g\left(t_m, \tilde{L}_j(s)\right) f_m\left(L_n(t)\right) \quad (16)$$

$$Q\left(h\left(L_n(t), \tilde{L}_j(s)\right)\right) = \sum_{i=-1}^{J+1} h\left(L_n(t), s_i\right) \tilde{f}_i\left(\tilde{L}_j(s)\right) \quad (17)$$

The $(N+3)(J+3)$-dimensional subspace $\mathcal{U}([a,b], \Delta_t) \otimes \mathcal{V}([c,d], \Delta_s)$ of $\mathcal{C}^2[\Omega]$ defined by the bottom equation is called the fractal tensor product of the spaces $\mathcal{U}([a,b], \Delta_t)$ and $\mathcal{V}([c,d], \Delta_s)$ with the basis $\{f_m(t)\tilde{f}_i(s) \mid m = -1,0,1,\ldots,N+1; i = -1,0,1,\ldots,J+1\}$. Now, we define complete cubic spline fractal surface $\rho_f = (P \circ Q) f$ for $f \in \mathcal{C}^2[\Omega]$ as

Table 1. Coefficients associated with the error of classical complete cubic spline.

$\varepsilon_{p,u}$	$u = 0$	$u = 1$	$u = 2$	$u = 3$
$p = 2$	9/8	4	10	-
$p = 3$	71/216	31/27	5	$(63 + 8\eta_t^2)/9$
$p = 4$	5/384	$1/24^a$	$3/8^a$	$(\eta_t + \eta_t^{-1})/2^a$

[a]See [18]

$$\mathcal{U}([a,b], \Delta_t) \otimes \mathcal{V}([c,d], \Delta_s) = \left\{ \sum_{m=-1}^{N+1} \sum_{i=-1}^{J+1} y_{m,i} f_m\left(L_n(t)\right) \tilde{f}_i\left(\tilde{L}_j(s)\right) : y_{m,i} \in \mathbb{R} \right\}$$

$$\rho_f\left(L_n(t),\tilde{L}_j(s)\right)=\sum_{m=-1}^{N+1}\sum_{i=-1}^{J+1} f(t_m,s_i)\, f_m\left(L_n(t)\right)\tilde{f}_i\left(\tilde{L}_j(s)\right),(t,s)\in\Omega, \tag{18}$$

where $f(t_{-1},s)=\partial f(t_0,s)/\partial t$, $f(t,s_{-1})=\partial f(t,s_0)/\partial s$ with the analogues for $f(t_{N+1},s)$, $f(t,s_{J+1})$, $f(t_{-1},s_{-1})$ etc. We now show that the function ρ_f satisfies the interpolation conditions. According to (18), for $n\in\{1,2,\ldots,N\}$ and $j\in\{1,2,\ldots,J\}$,

$$\begin{aligned}\rho_f(t_n,s_j)&=\left((P\circ Q)f\right)\left(L_n(t_N),\tilde{L}_j(s_J)\right)\\&=\sum_{m=-1}^{N+1}\sum_{i=-1}^{J+1} f(t_m,s_i)\, f_m(t_n)\,\tilde{f}_i(s_j)\\&=\sum_{m=-1}^{N+1}\sum_{i=-1}^{J+1} f(t_m,s_i)\,\delta_{m,n}\delta_{i,j}=f(t_n,s_j)\end{aligned}$$

Similarly, $\rho_f(t_0,s_j)=f(t_0,s_j)$, $j=1,2,\ldots,J$; $\rho_f(t_n,s_0)=f(t_n,s_0)$, $n=1,2,\ldots,N$. and $\rho_f(t_0,s_0)=f(t_0,s_0)$. Since $f_m'(t_0)=f_m'(t_N)=0$ for $m=0,1,\ldots,N$, using (8), we have

$$\rho_f^{(1,0)}(t_n,s_j)=\sum_{m=-1}^{N+1}\sum_{i=-1}^{J+1} f(t_m,s_i)\, f_m'(t_n)\,\tilde{f}_i(s_j)=f^{(1,0)}(t_n,s_j);$$

$n=0,N$; $j=0,1,\ldots,J$. Analogously, the rest of conditions of definition (15) are true. Since, $f_m=S_m$ if $\alpha_n=0\ \forall\ n=1,2,\ldots,N$ and $\tilde{f}_i=\tilde{S}_i$ if $\beta_j=0\ \forall\ j=1,2,\ldots,J$, we can retrieve classical complete bicubic spline S_f to the original function f from (18).

3.2. Upper Bounds of L_∞-Norm of Error

We will prove the L_∞-norm error of complete bicubic splines with the original function by using the following notations analogous to those of Proposition 2.2 for the *t*-variable.

Suppose $\tilde{q}_{i,j}$; $j=1,2,\ldots,J$ are cubic polynomials associated with the IFSs for $\tilde{f}_i$ such that

$$\left|\frac{\partial^{1+\nu}\tilde{q}_{i,j}(\tilde{\tau}_j,s)}{\partial\beta_j\partial s^{\nu}}\right|\le\tilde{\Theta}_{\nu,i}\quad\text{for}\quad i=-1,0,1,\ldots,J+1,$$

$|\tilde{\tau}_j|\in(0,\kappa^* c_j^{\nu})$ with $|\beta_j|\le\kappa^*<1$ for fixed real κ^*. Let, $\tilde{\Lambda}_\nu=\max\{\left\|\tilde{S}_i^{(\nu)}\right\|_\infty : i=-1,0,1,\ldots,J+1\}$, $\tilde{\Theta}_\nu=\max\{\tilde{\Theta}_{\nu,i} : i=-1,0,1,\ldots,J+1\}$ and

$$\tilde{\Gamma}_{\nu,\beta,J}=\frac{|\beta|_\infty(J+3)(\tilde{\Lambda}_\nu+\tilde{\Theta}_\nu)}{|J|^{2-\nu}-h_s^{2-\nu}|\beta|_\infty}.$$

Suppose

$C^4[\Omega]=\{g:\Omega\to\mathbb{R}: g^{(u,v)}\in C[\Omega],0\le u+v\le 4\}$ and the norm corresponding to this space is

$\|g\|_4=\max\{\left\|g^{(u,v)}\right\|_\infty;0\le u+v\le 4\}$.

Theorem 3.1 *Let ρ_f be the complete bicubic fractal spline to the original function $f\in C^4[\Omega]$. Then for an arbitrary sequence of partitions,*

$$\begin{aligned}&\left\|\left(f-\rho_f\right)^{(u,v)}\right\|_\infty\le\|f\|_4\left\{\Gamma_{u,\alpha,N}h_t^{2-u}+\varepsilon_{4-v,u}h_t^{4-u-v}\right.\\&+\tilde{\Gamma}_{v,\beta,J}h_s^{2-v}+\varepsilon_{4-u,v}h_s^{4-u-v}\\&\left.+\left(\Gamma_{u,\alpha,N}h_t^{2-u}+\varepsilon_{4-v,u}h_t^{4-u-v}\right)\left(\tilde{\Gamma}_{v,\beta,J}h_s^{2-v}+\varepsilon_{4-u,v}h_s^{4-u-v}\right)\right\},\\&0\le u,v\le 2.\end{aligned}$$

Proof. In order to calculate the error, we will use the generic transformations P and Q. Suppose that

$$\begin{aligned}f-\rho_f&=\left(f-Q(f)\right)-\left\{f-Q(f)-\left(P(f)-\rho_f\right)\right\}\\&\quad+\left(f-P(f)\right)\end{aligned} \tag{19}$$

Consider

$$\begin{aligned}&\left(f-P(f)\right)\left(L_n(t),\tilde{L}_j(s)\right)=f\left(L_n(t),\tilde{L}_j(s)\right)\\&-\sum_{m=-1}^{N+1} f\left(t_m,\tilde{L}_j(s)\right)f_m\left(L_n(t)\right)\end{aligned}$$

For $s=s^*$ fixed, $(P(f))^{(0,v)}(L_n(t),\tilde{L}_j(s^*))$ is the spline of $f^{(0,v)}(L_n(t),\tilde{L}_j(s^*))$ (with respect to the first variable) and we can apply (14) for $p=4-v$ since $f^{(0,v)}(.,\tilde{L}_j(s^*))\in C^{(4-v)}[a,b]$. For $0\le u\le 2$,

$$\begin{aligned}&\left\|f^{(u,v)}\left(.,\tilde{L}_j(s^*)\right)-P(f)^{(u,v)}\left(.,\tilde{L}_j(s^*)\right)\right\|_\infty\\&\le\left\|f^{(0,v)}\left(.,\tilde{L}_j(s^*)\right)\right\|_{4-v}\left(\Gamma_{u,\alpha,N}h_t^{2-u}+\varepsilon_{4-v,u}h_t^{4-v-u}\right)\\&\le\|f\|_4\left(\Gamma_{u,\alpha,N}h_t^{2-u}+\varepsilon_{4-v,u}h_t^{4-v-u}\right)\end{aligned} \tag{20}$$

Since the last term of (20) does not depend on s^*,

$$\left\|f^{(u,v)}-P(f)^{(u,v)}\right\|_\infty\le\|f\|_4\left(\Gamma_{u,\alpha,N}h_t^{2-u}+\varepsilon_{4-v,u}h_t^{4-v-u}\right) \tag{21}$$

In the same way,

$$\left\|f^{(u,v)}-Q(f)^{(u,v)}\right\|_\infty\le\|f\|_4\left(\tilde{\Gamma}_{v,\beta,J}h_s^{2-v}+\varepsilon_{4-u,v}h_s^{4-u-v}\right) \tag{22}$$

Consider $P(f)-\rho_f$ and using their definitions,

$$\begin{aligned}&\left(P(f)-\rho_f\right)\left(L_n(t),\tilde{L}_j(s)\right)\\&=\sum_{m=-1}^{N+1}\left\{f\left(t_m,\tilde{L}_j(s)\right)-\sum_{i=-1}^{J+1} f(t_m,s_i)\,\tilde{f}_i\left(\tilde{L}_j(s)\right)\right\}f_m\left(L_n(t)\right)\\&=\sum_{m=-1}^{N+1} R_1\left(t_m,\tilde{L}_j(s)\right)f_m\left(L_n(t)\right)=P(R_1)\left(L_n(t),\tilde{L}_j(s)\right)\end{aligned}$$

where $R_1=f-Q(f)$. Hence, we have $f-Q(f)-(P(f)-\rho_f)=R_1-P(R_1)$. For $s=s^*$ fixed, $P(R_1)^{(0,v)}$ is the cubic spline FIF of $R_1^{(0,v)}$ with respect to the first variable and we can apply (14) taking $p=4-v$.

$$\begin{aligned}&\left\|R_1^{(u,v)}\left(.,\tilde{L}_j\left(s^*\right)\right)-P(R_1)^{(u,v)}\left(.,\tilde{L}_j(s^*)\right)\right\|_\infty\\&\le\left\|R_1^{(0,v)}\left(.,\tilde{L}_j\left(s^*\right)\right)\right\|_{4-v}\left(\Gamma_{u,\alpha,N}h_t^{2-u}+\varepsilon_{4-v,u}h_t^{4-v-u}\right)\end{aligned}\quad(23)$$

For $0\le v\le 2$,

$$R_1^{(0,v)}\left(.,\tilde{L}_j\left(s^*\right)\right)=f^{(0,v)}\left(.,\tilde{L}_j\left(s^*\right)\right)-Q(f)^{(0,v)}\left(.,\tilde{L}_j\left(s^*\right)\right)$$

and similarly to (22) for $0\le u\le 2$,

$$\left\|R_1^{(u,v)}\left(.,\tilde{L}_j\left(s^*\right)\right)\right\|_\infty\le\|f\|_4\left(\tilde{\Gamma}_{v,\beta,J}h_s^{2-v}+\varepsilon_{4-u,v}h_s^{4-u-v}\right)$$

and by (23), we get the first equation of the bottom ones.

The inequality of Theorem 3.1 follows from (21)-(24).

Remark. From Theorem 3.1, it can be observed that the convergence of the bicubic fractal spline ρ_f is slower than that of the case of the classical bicubic spline S_f (see [20]). Since the classical bicubic spline is a particular case of bicubic fractal spline, Theorem 3.1 generalizes the classical result. If there exist positive reals k and l such that $\Gamma_{u,\alpha,N}<\frac{1}{(N+3)^k}$ and $\Gamma_{v,\beta,J}<\frac{1}{(J+3)^l}$, then the complete bicubic fractal spline ρ_f converges to the original function f in C^2-norm for uniform partitions.

4. Examples

First, we construct three different bases for complete cubic fractal spline space with $I=[0,3]$, $N=3$ and three different sets of scaling vectors. These scaling vectors play important role over classical splines in overall shape of fractal approximants. The scaling vectors are chosen for a basis of complete fractal splines as 1) Set I: $\alpha_n=0.9, i=1,2,3$; 2) Set II: $\alpha_n=-0.9, i=1,2,3$; 3) Set III: $\alpha_1=-0.5, \alpha_2=0.9, \alpha_3=0.7$. In our examples, $L_1(t)=\frac{1}{3}t, L_2(t)=\frac{1}{3}t+\frac{1}{3}$, and $L_3(t)=\frac{1}{3}t+\frac{2}{3}$. We compute the moments $M_{m,n}$, $n=1,2,3$; $m=-1,0,\ldots,4$ from Equations (26)-(28). These values of moments for three sets of bases are given in **Table 2**. These moments are then used in IFS $\{\mathbb{R}^2; w_n(t,x)=(L_n(t),F_n(x,t)), n=1,2,3\}$ to compute $F_n(x,t)$. From (5) and (25), we have the bottom equation. where x_n depends on the cardinal condition of the basis elements f_m, *i.e.*, $x_n=\delta_{m,n}$, $n=0,1,2,3; m=-1,0,\ldots,4$.

Using the above IFS, we compute basis elements for complete cubic fractal spline space with non-zero and zero scaling vectors. When scaling factors are same in Set I and Set II, the values of moments of f_{-1} and f_4; f_0 and f_3; f_1 and f_2 follow a particular pattern (see **Table 2**). This pattern is very close to the moments pattern of classical complete cardinal splines (with zero scale vector). That's why pair-wise similarity between complete cardinal fractal splines f_{-1} and f_4; f_0 and f_3; f_1 and f_2 in such cases (see **Figure 1(a)** and **Figure 1(b)**) observed as in classical complete cardinal splines (see **Figure 1(d)**). But for unequal scaling factors in Set III, there is no pattern between moments of complete cardinal fractal splines and hence, their shapes are completely different (see **Figure 1(c)**). The unequal scaling factors provides an additional advantage of complete cardinal fractal splines over their classical counterparts in smooth object modelling in engineering applications like computer graphics, CAD/CAM.

The non-zero scale vectors gives irregular shape to fractal splines because f_n'' are typical fractal functions, *i.e.*, fractal dimension of graph of f_n'' is non-integer. These cardinal fractal splines differ from their classical interpolants in the sense that they obey a functional relation related to self-similarity on smaller scales. Hence, cardinal fractal splines are defined globally on the entire domain. Moreover, classical complete splines are defined piecewisely between consecutive nodes and hence, their shapes can be defined locally. Importantly, if $\alpha_n=0$, $n=1,2,3$, then we can retrieve the basis elements (see **Figure 1(d)**) for the classical complete cubic spline space. Using these three sets of vertical scaling vectors, we have constructed complete bicubic splines in the following.

Some complete bicubic fractal splines are constructed using the tensor product of univariate fractal splines for the interpolation data given in **Table 3**. In all our examples, we assume the same boundary conditions for complete

$$\left\|R_1^{(u,v)}-P\left(R_1\right)^{(u,v)}\right\|_\infty\le\|f\|_4\left(\tilde{\Gamma}_{v,\beta,J}h_s^{2-v}+\varepsilon_{4-u,v}h_s^{4-u-v}\right).\left(\Gamma_{u,\alpha,N}h_t^{2-u}+\varepsilon_{4-v,u}h_t^{4-v-u}\right).\quad(24)$$

$$\begin{aligned}F_n(t,x)=\frac{1}{9}\Bigg\{&\alpha_n x+\frac{\left(M_{m,n}-\alpha_n M_{m,3}\right)t^3}{18}+\frac{\left(M_{m,n-1}-\alpha_n M_{m,0}\right)(3-t)^3}{18}-\frac{\left(M_{m,n}-\alpha_n M_{m,3}\right)t}{2}\\&-\frac{\left(M_{m,n-1}-\alpha_n M_{m,0}\right)(3-t)}{2}+\left(9x_{n-1}-\alpha_n x_0\right)\frac{3-t}{3}+\left(9x_n-\alpha_n x_3\right)\frac{t}{3}\Bigg\},\ n=1,2,3,\end{aligned}$$

Table 2. Moments for cardinal complete cubic fractal splines.

Basis Elements	Moments	Set I	Set II	Set III	Classical Case
f_{-1}	$M_{-1,0}$	−16.4058	−2.5546	−2.4586	−3.4667
	$M_{-1,1}$	−10.7536	1.4714	0.6318	0.9333
	$M_{-1,2}$	−12.5797	0.4584	−1.4137	−0.2667
	$M_{-1,3}$	−10.9275	0.4844	−0.5122	0.1333
f_0	$M_{0,0}$	−24.4348	−2.8448	−2.7653	−4.4000
	$M_{0,1}$	−14.5217	3.5448	2.0938	2.8000
	$M_{0,2}$	−19.4783	0.3500	−2.2451	−0.8000
	$M_{0,3}$	−15.5652	0.7396	−0.2621	0.4000
f_1	$M_{1,0}$	27.8261	3.3903	3.5365	5.6000
	$M_{1,1}$	11.3913	−5.7266	−3.6048	−5.2000
	$M_{1,2}$	22.6087	1.8319	3.6812	3.2000
	$M_{1,3}$	12.1739	−1.2850	−2.0094	−1.6000
f_2	$M_{2,0}$	12.1739	−1.2850	−2.0094	−1.6000
	$M_{2,1}$	22.6087	1.8319	1.3664	3.2000
	$M_{2,2}$	11.3913	−5.7266	−1.0007	−5.2000
	$M_{2,3}$	27.8261	3.3903	9.3235	5.6000
f_3	$M_{3,0}$	−15.5652	0.7396	1.2382	0.4000
	$M_{3,1}$	−19.4783	0.3500	0.1447	−0.8000
	$M_{3,2}$	−14.5217	3.5448	−0.4354	2.8000
	$M_{3,3}$	−24.4348	−2.8448	−7.0520	−4.4000
f_4	$M_{4,0}$	10.9275	−0.4844	−0.7738	−0.1333
	$M_{4,1}$	12.5797	−0.4584	−0.1938	0.2667
	$M_{4,2}$	10.7536	−1.4714	1.0400	−0.9333
	$M_{4,3}$	16.4058	2.5546	5.0306	3.4667

Table 3. Interpolation data for complete bicubic splines.

$f(t_n, s_j)$	$s_0 = 1$	$s_1 = 2$	$s_2 = 3$	$s_3 = 3$
$t_0 = 1$	−1	11	−5	10
$t_1 = 2$	2	14	3	15
$t_2 = 3$	0	9	1	17
$t_3 = 4$	1	10	3	13

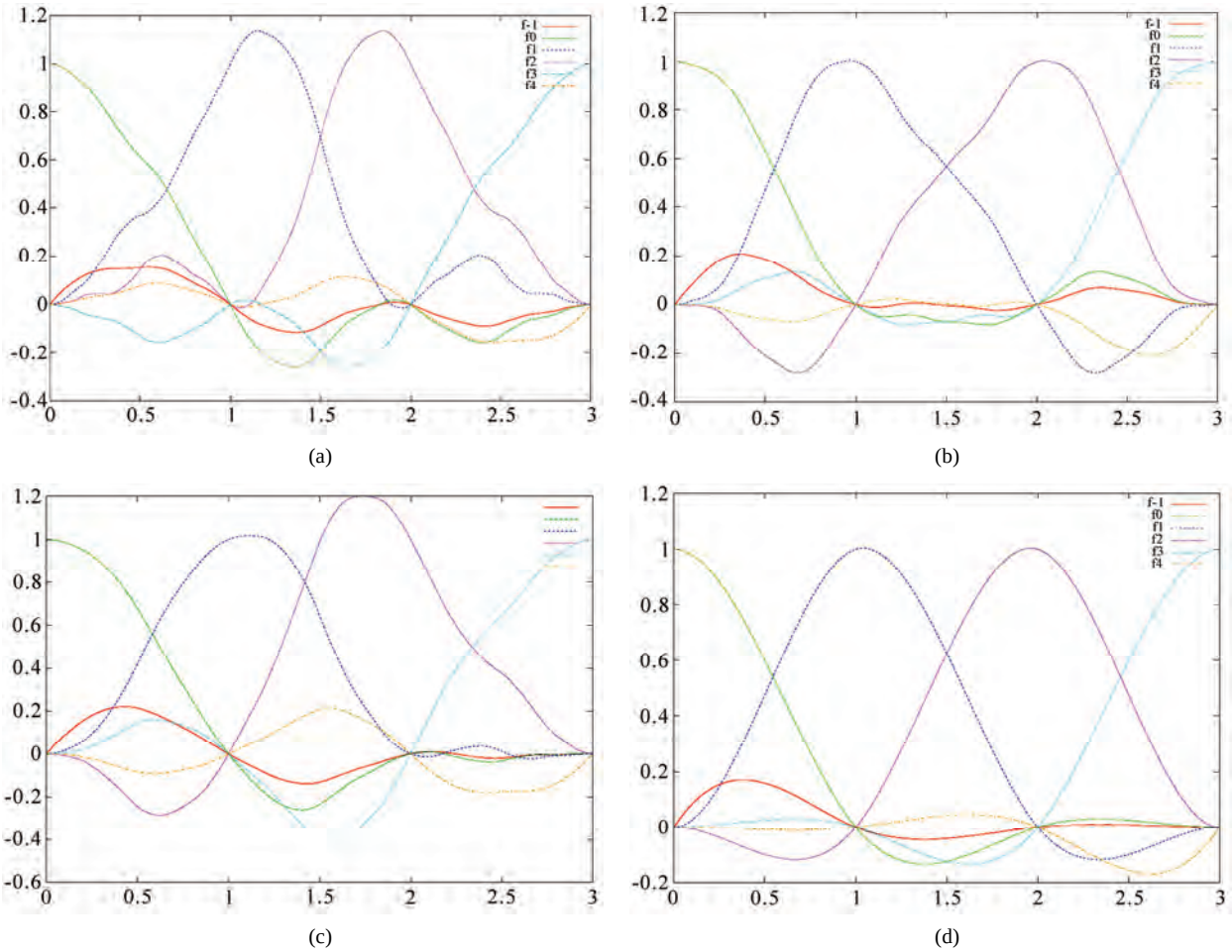


Figure 1. Bases for complete cubic fractal spline space. (a) Cardinal cubic fractal splines with Set I; (b) Cardinal cubic fractal splines with Set II; (c) Cardinal cubic fractal splines with Set III; (d) Classical cardinal cubic splines with $\alpha_n = 0, n = 1,2,3$.

bicubic splines: $f^{(1,0)}(t_n, s_j) = 5; n = 0,3,\ j = 0,1,2,3$, $f^{(0,1)}(t_n, s_j) = 3; n = 0,1,2,3,\ j = 0,3,$ and $f^{(1,1)}(t_n, s_j) = 2;$ $j = 0,3$. The scaling vectors are same in both directions in our first three examples, *i.e.*, $\alpha_n = \beta_j$ for $n = j$ and we take these scale vectors as Set I, Set II and Set III defined in above for univariate case. The cardinal splines $f_m = \tilde{f}_i$ if $i = m$ in these cases for $i, m = -1, 0, \ldots, 4$. Based on (18), the points of complete bicubic fractal splines are generated and plotted in **Figsures 2-4**. The effect of change in scaling factors from 0.9 to –0.9 on complete bicubic spline can be seen from **Figsures 2-3**. The difference in the shape of complete bicubic spline for an unequal scaling factors can be observed by comparing **Figure 4** with **Figsures 2-3**.

Next, we take scaling vectors in *t*-direction as Set I and in *s*-direction as Set III and the corresponding complete bicubic spline generated in **Figure 5**. It has similarity with both **Figure 2** and **Figure 4** due to self-similarity relation in *t* and *s* directions respectively. For **Figure 6**, we chose scaling vectors in *t*-direction as Set III and in *s*-direction as Set II. The distinct deviation in *s*-direction of complete bicubic spline is present in this case as in **Figure 3**. Finally, we chose $\alpha_n = \beta_j = 0 \quad \forall n, j$. Since $f_m = S_m$ and $\tilde{f}_i = \tilde{S}_i$ in this case, we retrieve the classical complete bicubic spline S_f in **Figure 7**. An infinite number of complete bicubic splines can be constructed interpolation the same data by choosing different sets of scaling vectors. Hence, the presence of scaling vectors in bicubic fractal splines provides an additional advantage over classical bicubic splines in smooth surface modelling. Since bicubic fractal splines are invariant in all scales, it can also be applied to bivariate image compression and zooming problems in image processing.

5. Conclusions

We introduced bases for complete cubic fractal splines through cardinal fractal splines in the present work. These cardinal fractal splines constructed through moments

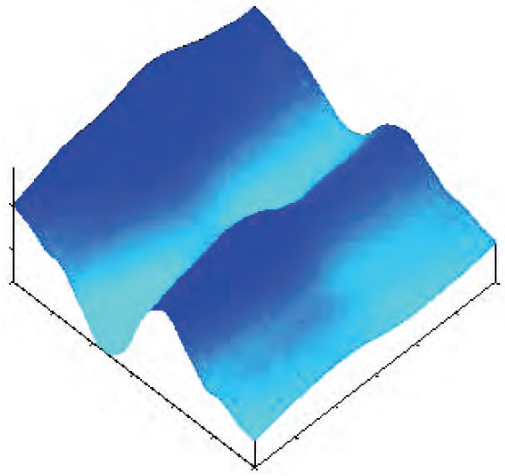


Figure 2. Complete bicubic fractal spline with scale vectors Set I in both directions.

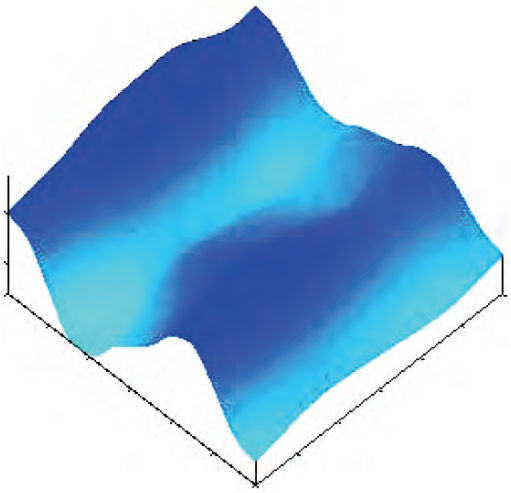


Figure 3. Complete bicubic fractal spline with scale vectors Set II in both directions.

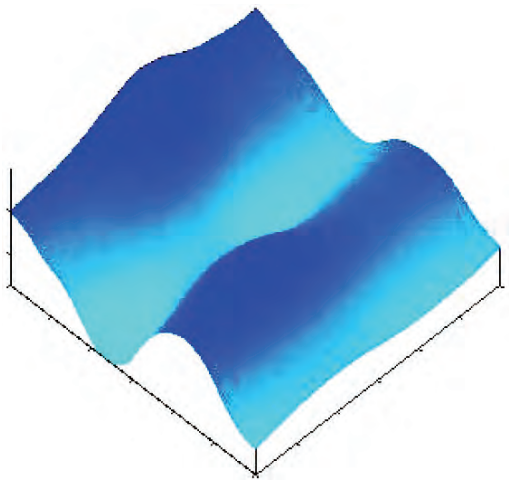


Figure 4. Complete bicubic fractal spline with scale vectors Set III in both directions.

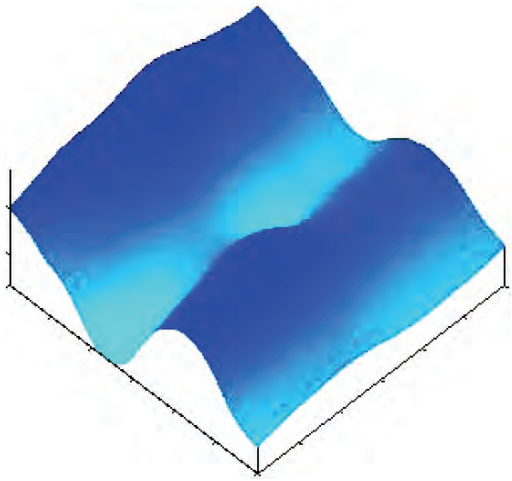


Figure 5. Complete bicubic fractal spline with scale vectors Set I, Set III.

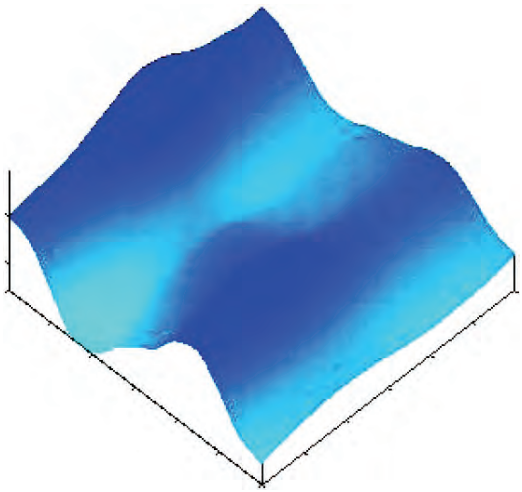


Figure 6. Complete bicubic fractal spline with scale vectors Set III, Set II.

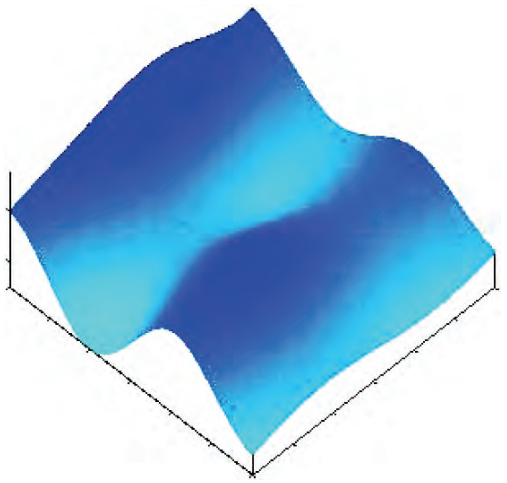


Figure 7. Classical complete bicubic spline with zero scale vector.

as in the classical case. Using tensor product of cardinal cubic fractal splines, bicubic fractal splines introduced over rectangular domains with rectangular partition. These bicubic fractal splines are invariant in all scales due to underlying fixed point equation. L_∞-norm of the error of complete cubic fractal spline with respect to the original function $f \in C^p[\Omega]$, $p = 2,3$ or 4 has been deduced. The presence of scaling factors can be exploited in bivariate optimization problems with prescribed interpolation conditions. The effect of equal and non-equal scaling factors in complete bicubic splines is explained. The present work may play important role in smooth surface modelling in computer graphics and image processing applications.

6. Acknowledgements

The work was supported by the project No: SB 2005-0199, Spain. The first author is thankful to the Institute of Mathematics and Applications, Bhubaneswar, India for its support after this project.

7. References

[1] I. J. Schoenberg, "Contribution to the Problem of Approximation of Equidistant Data by Analytic Functions, Part A and B," *Quarterly of Applied Mathematics*, Vol. 4, No. 2, 1946, pp. 45-99, 112-141.

[2] M. F. Barnsley, "Fractals Everywhere," Academic Press, Orlando, Florida, 1988.

[3] M. F. Barnsley, "Fractal Functions and Interpolation," *Constructive Approximation*, Vol. 2, No. 2, 1986, pp. 303-329.

[4] D. S. Mazel and M. H. Hayes, "Using Iterated Function Systems to Model Discrete Sequences," *IEEE Transactions on Signal Processing*, Vol. 40, No. 7, 1992, pp. 1724-1734.

[5] M. F. Barnsley and A. N. Harrington, "The Calculus of Fractal Interpolation Functions," *Journal of Approximation Theory*, Vol. 57, No. 1, 1989, pp. 14-34.

[6] A. K. B. Chand and G. P. Kapoor, "Generalized Cubic Spline Fractal Interpolation Functions," *SIAM Journal on Numerical Analysis*, Vol. 44, No. 2, 2006, pp. 655-676.

[7] M. A. Navascués and M. V. Sebastián, "Smooth Fractal Interpolation," *Journal of Inequalities and Applications*, 2006, pp. 1-20.

[8] M. A. Navascués, "Fractal Polynomial Interpolation," *Zeitschrift fur Analysis und ihre Anwendungen*, Vol. 25, No. 2, 2005, pp. 401-418.

[9] M. A. Navascués, "A Fractal Approximation to Periodicity," *Fractals*, Vol. 14, No. 4, 2006, pp. 315-325.

[10] A. K. B. Chand and M. A. Navascués, "Generalized Hermite Fractal Interpolation," *Academia de Ciencias, Zaragoza*, Vol. 64, 2009, pp. 107-120.

[11] P. R. Massopust, "Fractal Surfaces," *Journal of Mathematical Analysis and Applications*, Vol. 151, No. 1, 1990, pp. 275-290.

[12] J. S. Geronimo and D. Hardin, "Fractal Interpolation Functions from $\mathbb{R}^n \to \mathbb{R}^m$ and their Projections," *Zeitschrift fur Analysis und ihre Anwendungen*, Vol. 12, No. 3, 1993, pp. 535-548.

[13] N. Zhao, "Construction and Application of Fractal Interpolation Surfaces," *Visual Computer*, Vol. 12, No. 3, 1996, pp. 132-146.

[14] H. Xie and H. Sun, "The Study of Bivariate Fractal Interpolation Functions and Creation of Fractal Interpolation Surfaces," *Fractals*, Vol. 5, No. 4, 1997, pp. 625-634.

[15] A. K. B. Chand and G. P. Kapoor, "Hidden Variable Bivariate Fractal Interpolation Surfaces," *Fractals*, Vol. 11, No. 3, 2003, pp. 277-288.

[16] P. Bouboulis and L. Dalla, "Fractal Interpolation Surfaces derived from Fractal Interpolation Functions," *Journal of Mathematical Analysis and Applications*, Vol. 336, No. 2, 2007, pp. 919-936.

[17] R. E. Carlson and C. A. Hall, "Error Bounds for Bicubic Spline Interpolation," *Journal of Approximation Theory*, Vol. 7, 1973, pp. 41-47.

[18] C. A. Hall and W. W. Meyer, "Optimal Error Bounds for Cubic Spline Interpolation," *Journal of Approximation Theory*, Vol. 16, No. 2, 1976, pp. 105-122.

[19] C. de Boor, "Bicubic Spline Interpolation," *Journal of Mathematics and Physics*, Vol. 41, No. 4, 1962, pp. 212-218.

[20] J. H. Ahlberg, E. N. Nilson and J. L. Walsh, "The Theory of Splines and their Applications," Academic Press, New York, 1967.

Appendix

In this section, the defining equations for the construction of a cubic spline FIF f_m in Subsection 2.2 are given. Using Property 3) of Definition 2.1 and (5), the polynomials associated with $f_m^{''}$ is affine. So,

$$f_m^{''}\left(L_n(t)\right) = \alpha_n f_m^{''}(t) + \frac{c_n(t-t_0)}{t_N - t_0} + d_n, \quad n = 1,2,\ldots,N. \tag{25}$$

By (1) and (25), $c_n = M_n - M_{n-1} - \alpha_n(M_N - M_0)$ and $d_n = M_{n-1} - \alpha_n M_0$. Substituting c_n and d_n in (25) and integrating it twice, we will have two constants of integration. Solving these constants by (1), the cubic fractal spline in terms of moments can be written as

$$\begin{aligned} f_m\left(L_n(t)\right) = a_n^2 \Bigg\{ & \alpha_n f_m(t) + \frac{\left(M_{m,n} - \alpha_n M_{m,N}\right)(t-t_0)^3}{6(t_N - t_0)} \\ & + \frac{\left(M_{m,n-1} - \alpha_n M_{m,0}\right)(t_N - t)^3}{6(t_N - t_0)} \\ & - \frac{\left(M_{m,n-1} - \alpha_n M_{m,0}\right)(t_N - t_0)(t_N - t)}{6} \\ & - \frac{\left(M_{m,n} - \alpha_n M_{m,N}\right)(t_N - t_0)(t - t_0)}{6} \\ & + \left(\frac{x_{n-1}}{a_n^2} - \alpha_n x_0\right)\frac{t_N - t}{t_N - t_0} + \left(\frac{x_n}{a_n^2} - \alpha_n x_N\right)\frac{t - t_0}{t_N - t_0} \Bigg\}, \\ & n = 1,2,\ldots,N. \end{aligned} \tag{26}$$

Set, $h_n = t_n - t_{n-1}; n = 1,2,\ldots,N$. Now, use the condition that $f_{m'}(t)$ is continuous at the knots $t_1, t_2, \ldots, t_{N-1}$ to give the following result.

$$\begin{aligned} & -a_{n+1}\alpha_{n+1} f_m'(t_0) - \frac{\alpha_n h_n + 2\alpha_{n+1}h_{n+1}}{6} M_{m,0} \\ & + \frac{h_n}{6} M_{m,n-1} + \frac{h_n + h_{n+1}}{3} M_{m,n} + \frac{h_{n+1}}{6} M_{m,n+1} \\ & - \frac{2\alpha_n h_n + \alpha_{n+1} h_{n+1}}{6} M_{m,N} + a_n \alpha_n f_m'(t_N) \\ & = \frac{x_{n+1} - x_n}{h_{n+1}} - \frac{x_n - x_{n-1}}{h_n} - \left(a_{n+1}\alpha_{n+1} - a_n\alpha_n\right)\frac{x_N - x_0}{t_N - t_0}, \\ & n = 1,2,\ldots,N-1.; \end{aligned} \tag{27}$$

At the initial point t_0 of the interval I, we have the following relation for $f_m'(t_0)$.

$$\begin{aligned} & 6(1 - a_1\alpha_1) f_m'(t_0) + 2(1-\alpha_1) h_1 M_{m,0} + h_1 M_{m,1} \\ & -\alpha_1 h_1 M_{m,N} = \frac{6}{h_1}\left[x_1 - x_0 - \alpha_1 a_1^2 (x_N - x_0)\right] \end{aligned} \tag{28}$$

Similarly, at the final point t_N of the interval I for $f_m'(t_N)$, we have

$$\begin{aligned} & -\alpha_N h_N M_{m,0} + h_N M_{m,N-1} + 2(1-\alpha_N) h_N M_{m,N} \\ & -6(1 - a_N\alpha_N) f_m'(t_N) = \frac{-6}{h_N}\left[x_N - x_{N-1} - \alpha_N a_N^2 (x_N - x_0)\right] \end{aligned} \tag{29}$$

The moments $M_{m,n}; n = 0,1,\ldots,N$, $f_m'(t_0)$ and $f_m'(t_N)$ are evaluated from the system of Equations (27)-(29). The existence of these parameters is guaranteed by the uniqueness of the attractor from the fixed point theorem.

Applied Mathematics, **2010, 1, 211-214**
doi:10.4236/am.2010.13025 Published Online September 2010 (http://www.SciRP.org/journal/am)

Scientific Research

On the Behavior of the Residual in Conjugate Gradient Method

Teruyoshi Washizawa
Analysis Technology Center, Canon Inc., Tokyo, Japan
E-mail: *washizawa.teruyoshi@canon.co.jp*
Received March 21, 2010; *revised July* 22, 2010; *accepted July* 24, 2010

Abstract

In conjugate gradient method, it is well known that the recursively computed residual differs from true one as the iteration proceeds in finite arithmetic. Some work have been devoted to analyze this behavior and to evaluate the lower and the upper bounds of the difference. This paper focuses on the behavior of these two kinds of residuals, especially their lower bounds caused by the loss of trailing digit, respectively.

Keywords: Conjugate Gradient, Residual, Convergence, Finite Arithmetic, Lower Bound

1. Introduction

Conjugate gradient (CG) method and its varieties are popular as one of the best unsteady iterative methods for solving the following linear system:

$$Ax = b \tag{1}$$

In CG method, an approximate solution x_k is expected to approach the exact solution x^*. For the symmetric positive definite A, it is proved that the A-norm of the error monotonically decreases as the iteration proceeds in exact arithmetic. This will be called as A-norm monotonicity of the error in the remaining part of this article. It is obvious that we cannot calculate directly such a norm of the error without the solution. Therefore, almost algorithms employ the residual which is easily calculated as the difference between the left hand side (LHS) and the right hand side (RHS) of (1), $r_k := b\text{-}Ax_k$. In practice, the residual is calculated by the recursion formula because of the computational complexity of the matrix vector product Ax_k [1,2].

However, this recursion formula causes another problem in which the recursive residual differs from the true residual as the iteration proceeds. It can be also observed that the recursive residual decreases after the true one seams to reach its lower bound. We should terminate the CG steps just before the difference is too large to be neglected.

Ginsburg has proposed a simple criterion [1]:

For the true residual calculated as the difference between LHS and RHS of a linear system and the recursive residual calculated by using the recursion formula, the procedure is terminated when the 2-norm of their difference is greater than the 2-norm of the recursive residual:

$$\|r_k\| < \exp(k/n)^2 \|s_k - r_k\|$$

where n is dimensionarity of a linear system.

Several researchers have proposed the estimations of the lower and the upper bound of the norm of the error and the residual. Woźniakowski investigated the numerical stabilities and good-behaviors of three stationary iterative methods and CG method using the true residual $b\text{-}Ax_k$ [3,4]. Woźniakowski gave the upper bound of the ultimately attainable accuracies of the A-norm and the 2-norm of the error, and 2-norm of the true residual. Bollen gives the round-off error analysis of descent methods and lead a general result on the attainable accuracy of the approximate solution in finite arithmetic [2]. It has also shown that the general result is applied to the Gauss-Southwell method and the gradient method to obtain the decreasing rates of the A-norm of the error in finite arithmetic. Greenbaum have shown that for tiny perturbation ε_M, the eigenvalues and the A-norm of the error vectors generated over a fixed number of perturbed iterative steps are approximately the same as those quantities generated by the exact recurrences applied to a "nearby" matrix [5]. The lower bound of the true residual is pointed out in [6]. Two kinds of the estimates of the A-norm of the error at every step in CG algorithm has been proposed and verified that those estimates are the lower and the upper bound in [7]. The lower and the upper bounds of the A-norm of the error have been also given by Meurant [7] and Strakoš and Tichý [8]. Strakoš

and Tichý have proposed the tight estimate for the lower bound of both the *A*-norm and the 2-norm of the error in every step. This stepwise lower bound, however, keeps decreasing after the error reaches its 'global' lower bound. Therefore, the terminating criteria by using this stepwise lower bound cannot detect the global lower bound of the error. Calvetti *et al.* has proposed the estimates of the lower and upper bound of the *A*-norm of the error in CG method [9]. Those previous studies give the stepwise lower and the upper bound of the error and the residual but the global bounds.

In the remaining part of this article, we will first show that the true and the recursive residual almost monotonically decrease as the iteration proceeds. Then, these lower bounds will be shown.

2. Notations

We shall give the notations appeared throughout this article. *A* and *b* is, respectively, a coefficient matrix and a constant vector in a linear system. $\|\cdot\|$ in connection with a vector and a matrix, respectively, stands for the 2-norm and spectral norm, $\|\cdot\|_A$ in connection with a vector stands for the norm under the metric tensor *A*. The exact value of a variable *x* is denoted as $\overline{x}$. The floating point representation of a variable *x* is denoted simply as *x*. The computational error caused by the floating point representation is denoted as an operator $\varepsilon_M(x) \coloneqq x - \overline{x}$.

The exact solution of (1) is denoted as x^* which is described formally as $x^*=A^{-1}b$. At the *k*-th step, an approximate solution, the error, the true residual, and the recursive residual is, respectively, described as x_k, e_k, s_k, and r_k. They are computed in CG method as follows:

$$x_{k+1} \coloneqq x_k + \alpha_k p_k, e_k \coloneqq x^* - x_k,$$
$$s_k \coloneqq b - Ax_k, r_{k+1} \coloneqq r_k - \alpha A p_k$$

Since *A* is a constant matrix and $\|\varepsilon_M(A)\|/\|A\|$ is almost equal to ε_M without the dependence on the number of iterations, $\varepsilon_M(A)$ is out of our concern as well as $\varepsilon_M(b)$.

3. Almost Monotonicity of Residuals in Finite Arithmetic

In this section, we will see the true and the recursive residual has the 2-norm almost monotonicity in finite arithmetic, respectively.

3.1. The 2-Norm Almost Monotonicity of True Residual

The true residual is calculated as the difference between LHS and RHS. This is equivalent to multiplication of *A* to the error in finite arithmetic,

$$s_k = b - Ax_k = b - A\left(x_k - x^*\right) - Ax^* = -Ae_k$$

The behavior of true residual s_k is, therefore, equivalent to Ae_k.

The *A*-norm monotonicity of the error in finite arithmetic has been proved in theorem-3.1 of [2]. The following theorem shows the error has the *2*-norm almost monotonicity.

Theorem-1. If e_k has the *A*-norm monotonicity, e_k has the 2-norm almost monotonicity for a regular matrix *A*,

$$\exists k > j, \|e_k\| < \|e_j\| \tag{2}$$

Proof. The relationship between 2-norm and *A*-norm of error is

$$\|e_k\| = \left\|A^{-1/2}e_k\right\|_A \le \left\|A^{-1/2}\right\|\|e_k\|_A \tag{3}$$

Similarly,

$$\|e_k\|_A = \left\|A^{1/2}e_k\right\| \le \left\|A^{1/2}\right\|\|e_k\| \tag{4}$$

Substituting (3) and (4) into (2) and we yield

$$\left\|A^{-1/2}\right\|\|e_k\|_A < \left\|A^{1/2}\right\|^{-1}\|e_j\|_A \tag{5}$$

Equation (5) holds if *k* exists to satisfy the following relation

$$\|e_k\|_A < \kappa\left(A^{1/2}\right)^{-1}\|e_j\|_A \tag{6}$$

where $\kappa(A)$ is the condition number of a matrix *A*. From the *A*-norm monotonicity of the error e_k, since the following relation holds for any positive value *a*,

$$\exists k > j, \|e_k\|_A < a\|e_j\|_A$$

there exists $k > j$ that satisfies (6) and consequently (2). We have to notice that (2) does not hold when $a = 0$, *i.e.*, the inverse of the coefficient matrix *A* is singular.

Theorem-1 leads to the 2-norm almost monotonicity of the true residual using the relationship $\|s_k\| = \|Ae_k\|$.

Theorem-2. If e_k has the 2-norm almost monotonicity, s_k has the 2-norm almost monotonicity for a regular matrix *A*,

$$\exists k > j, \ \|s_k\| < \|s_j\| \tag{7}$$

Proof. The relationship between the 2-norm of the error and that of the residual is

$$\|s_k\| = \|Ae_k\| \le \|A\|\|e_k\| \tag{8}$$

Similarly,

$$\|e_j\| = \left\|A^{-1}s_j\right\| \le \left\|A^{-1}\right\|\|s_j\| \tag{9}$$

Substituting (8) and (9) into (7) and we yield

$$\|A\|\|e_k\| < \left\|A^{-1}\right\|^{-1}\|e_j\| \tag{10}$$

Equation (10) holds if k exists to satisfy the following relation

$$\|e_k\| < \kappa(A)^{-1}\|e_j\| \tag{11}$$

From the 2-norm almost monotonicity of the error e_k, since the following relation holds for any positive value a,

$$\exists k > j, \|e_k\| < a\|e_j\| \tag{12}$$

there exists $k > j$ that satisfies (11) and consequently (7).

3.2. The 2-Norm Almost Monotonicity of Recursive Residual

Before the proof of almost monotonicity of recursive residual in finite arithmetic, we first give the proof of almost monotonicity of recursive residual in exact arithmetic.

From the A-norm monotonicity of the error in exact arithmetic, the 2-norm almost monotonicity of the residual in exact arithmetic can be proved. We have to notice that the recursive residual $\overline{r}_j$ is identical to the true residual $\overline{s}_j$ in exact arithmetic.

Theorem-3. If $\forall n, \|\overline{e}_{n+1}\|_A < \|\overline{e}_n\|_A$, then the following proposition holds for a regular matrix A:

$$\exists k > j, \|\overline{r}_k\| < \|\overline{r}_j\| \tag{13}$$

Proof. The relationship between the error and the residual gives:

$$\|\overline{r}_k\| = \|A\overline{e}_k\| = \|A^{1/2}\overline{e}_k\|_A \tag{14}$$

Then we yield the lower and the upper bound of the 2-norm of the true residual:

$$\left\|A^{-1/2}\right\|^{-1}\|\overline{e}_k\|_A \le \|\overline{r}_k\| \le \left\|A^{1/2}\right\|\|\overline{e}_k\|_A \tag{15}$$

From above equation, the sufficient condition for (13) can be given as follows:

$$\left\|A^{1/2}\right\|\|\overline{e}_k\|_A < \left\|A^{-1/2}\right\|^{-1}\|\overline{e}_j\|_A \tag{16}$$

that is, the equation holds if there exists $k > j$ so that

$$\|\overline{e}_k\|_A < \kappa\left(A^{1/2}\right)^{-1}\|\overline{e}_j\|_A \tag{17}$$

From the A-norm monotonicity of the error $\overline{e}_k$, the following equation holds for any positive value a,

$$\exists k > j, \|\overline{e}_k\|_A < a\|\overline{e}_j\|_A \tag{18}$$

and (13) holds.

Now we show the almost monotonicity of the recursive residual in finite arithmetic.

Theorem-4. If the recursive residual has 2-norm almost monotonicity in exact arithmetic, then the recursive residual has the 2-norm almost monotonicity in finite arithmetic:

$$\exists k > j, \|r_k\| < \|r_j\| \tag{19}$$

Proof. Equation (19) is rewritten as

$$\exists k > j, \|\overline{r}_k + \varepsilon_M(\overline{r}_k)\| < \|\overline{r}_j + \varepsilon_M(\overline{r}_j)\| \tag{20}$$

The following relationship is one of its sufficient conditions

$$\max\left[\|\overline{r}_k + \varepsilon_M(\overline{r}_k)\|\right] < \min\left[\|\overline{r}_j + \varepsilon_M(\overline{r}_j)\|\right]$$

The evaluation of the maximum value of LHS is

$$\max\left[\|r_k\|\right] = (1+\varepsilon_M)\|\overline{r}_k\| \tag{21}$$

Similarly, the minimum value of RHS is evaluated as

$$\min\left[\|r_j\|\right] = (1-\varepsilon_M)\|\overline{r}_j\| \tag{22}$$

Substituting (21) and (22) into (20), the sufficient condition of (20) is given as

$$\exists k > j, \|\overline{r}_k\| < (1-\varepsilon_M)/(1+\varepsilon_M)\|\overline{r}_j\| \tag{23}$$

There exists $k > j$ for $(1-\varepsilon_M)/(1+\varepsilon_M) > 0$ from theorem-3 and (23) holds.

4. Lower Bounds of Error and Residual in Finite Arithmetic

It has been shown that the 2-norm of two kinds of residuals, respectively, decreases almost monotonically in finite arithmetic in the previous section.

Now we consider whether if the 2-norm of each variable stops decreasing before the approximate x_k does not reach its target x^*.

4.1. Lower Bound of Error

Theoem-1 shows the approximate solution x_j approaches the exact solution almost monotonically in finite arithmetic. The correction of the recursion formula of x_k, however, vanishes by the loss of trailing digits so that the error stops changing, *i.e.*,

$$\left|\Delta x_{stop}(n)\right| / \left|x_{stop}(n)\right| < \varepsilon_M \Rightarrow x_{stop+1} = x_{stop}$$

where $x_k(n)$ is the n-th component of x_k.

On the other hand, the solution in finite arithmetric x^* is not always identical to that in exact arithmetic $\overline{x}^*$.

Therefore, the target for iterative algorithms in finite

arithmetic should not be $\bar{x}^*$ but x^*. The error caused by the loss of trailing digits is described formally as $x^* - x_{stop}$. Since the true residual is given by multiplying A to the error, the lower bound of the true residual is given as $A\left(x^* - x_{stop}\right)$.

4.2. Lower Bound of Recursive Residual

Theorem-4 shows the recursive residual reduces its 2-norm almost monotonically. The next theorem proves that the change of the recursive residual stops only when $||r_{k+1}|| < \varepsilon_M ||r_k||$. It decreases almost monotonically until then.

Theorem-5. The recursive residual never have a lower bound caused by the loss of trailing digits.

Proof. The recursion formula of the residual is in general described as

$$r_{k+1} = r_k - \Delta r_k \tag{24}$$

where $\Delta r_k := \alpha_k A p_k$. The residual reaches its lower bound r_k if the following condition is satisfied:

$$\Delta r_k < \varepsilon_M \left(r_k\right) \tag{25}$$

We will show the condition of (25) never be satisfied in not only exact but also finite arithmetic.

In exact arithmetic, (24) satisfies the following relationship:

$$\begin{aligned}\left\|\bar{r}_{k+1}\right\|^2 &= \left\|\bar{r}_k - \Delta\bar{r}_k\right\|^2 = \left\|\bar{r}_k\right\|^2 - 2\left(\bar{r}_k, \Delta\bar{r}_k\right) + \left\|\Delta\bar{r}_k\right\|^2 \\ &= \left\|\bar{r}_k\right\|^2 - 2\left(\bar{r}_k, \bar{r}_k - \bar{r}_{k+1}\right) + \left\|\Delta\bar{r}_k\right\|^2 \\ &= \left\|\bar{r}_k\right\|^2 - 2\left\|\bar{r}_k\right\|^2 + \left\|\Delta\bar{r}_k\right\|^2 = -\left\|\bar{r}_k\right\|^2 + \left\|\Delta\bar{r}_k\right\|^2\end{aligned}$$

where using $\left(\bar{r}_k, \bar{r}_{k+1}\right) = 0$.

Therefore, three vectors in above recursion formula forms the right triangle so that the following relationship holds in exact arithmetic :

$$\left\|\Delta\bar{r}_k\right\| \geq \left\|\bar{r}_k\right\| \tag{26}$$

This shows that (25) never be satisfied in exact arithmetic.

Now we evaluate above in finite arithmetic.

$$\begin{aligned}\frac{\left\|\Delta r_k\right\|}{\left\|r_k\right\|} &\geq \frac{\min\left[\left\|\Delta\bar{r}_k + \varepsilon_M\left(\Delta\bar{r}_k\right)\right\|\right]}{\max\left[\left\|\bar{r}_k + \varepsilon_M\left(\bar{r}_k\right)\right\|\right]} \\ &= \frac{\left(1-\varepsilon_M\right)\left\|\Delta\bar{r}_k\right\|}{\left(1+\varepsilon_M\right)\left\|\bar{r}_k\right\|} = \left(1 - \frac{2\varepsilon_M}{1+\varepsilon_M}\right)\frac{\left\|\Delta\bar{r}_k\right\|}{\left\|\bar{r}_k\right\|}\end{aligned}$$

According to (26), we yield

$$\left\|\Delta r_k\right\| / \left\|r_k\right\| \geq \left\{1 - 2\varepsilon_M / \left(1+\varepsilon_M\right)\right\}$$

and prove that the recursive residual never have a lower bound caused by the loss of trailing digits.

The termination of the iterations is caused only when $\left\|r_{k+1}\right\| \leq \varepsilon_M \left\|r_k\right\|$ by the significant decrease of the recursive residual for $\alpha_k A p_k \approx r_k$.

5. Conclusions

In this article, the convergence behaviors of true and recursive residual have been analyzed.

The results obtained are summarized below:

1) In finite arithmetic, the 2-norm of the error and the residual, respectively, almost monotonically decreases.

2) 2-norm of the error has the lower bound in finite arithmetic as well as the true residual.

3) 2-norm of the recursive residual never has a non-zero lower bound caused by the loss of trailing digits in finite arithmetic.

6. References

[1] T. Ginsburg, "The Conjugate Gradient Method," *Numerische Mathematik*, Vol. 5, No. 1, 1963, pp. 191-200.

[2] J. A. M. Bollen, "Numerical Stability of Descenet Methods for Solving Linear Equations," *Numerische Mathematik*, Vol. 43, No. 3, 1984, pp. 361-377.

[3] H. Woźniakowski, "Round-off Error Analysis of Iteratinos for Large Linear Systems," *Numeriche Mathematik*, Vol. 30, No. 3, 1978, pp. 301-314.

[4] H. Woźniakowski, "Roundoff Error Analysis of New Class of Conjugate-Gradient Algorithms," *Linear Algebra and its Applications*, Vol. 29, 1980, pp. 507-529.

[5] A. Greenbaum, "Behavior of Slightly Perturbed Lanczos and Conjugate-Gradient Recurrences," *Linear Algebra and its Applications*, Vol. 113, 1989, pp. 7-63.

[6] A. Greenbaum, "Estimating the Attainable Accuracy of Recursively Computed Residual Methods," *SIAM Journal on Matrix Analysis and Applications*, Vol. 18, No. 3, 1997, pp. 535- 551.

[7] G. Meurant, "The Computation of Bounds for the Norm of the Error in the Conjugate Gradient Algorithm," *Numerical Algorithms*, Vol. 16, No. 3-4, 1997, pp. 77-87.

[8] Z. Strakoš and P. Tichý, "On Error Estimation in the Conjugate Gradient Method and Why it Works in Finite Precision Computations," *Electronic Transactions on Numerical Analysis*, Vol. 13, 2002, pp. 56-80.

[9] D. Calvetti, S. Morigi, L. Reichel and F. Sgallari, "Computable Error Bounds and Estimates for the Conjugate Gradient Method," *Numerical Algorithms*, Vol. 25, No. 1-4, 2000, pp. 75-88.

Applied Mathematics, 2010, 1, 215-221
doi:10.4236/am.2010.13026 Published Online September 2010 (http://www.SciRP.org/journal/am)

A Pest Management Epidemic Model with Time Delay and Stage-Structure

Yumin Ding[1], Shujing Gao[1], Yujiang Liu[1], Yun Lan[2]
[1]*Key Laboratory of Jiangxi Province for Numerical Simulation and Emulation Techniques, Gannan Normal University, Ganzhou, China*
[2]*Jiangxi Environmental Engineering Vocational College, Ganzhou, China*
E-mail: *gaosjmath*@126.*com*
Received June 12, 2010; *revised July* 22, 2010; *accepted July* 25, 2010

Abstract

In this paper, an SI epidemic model with stage structure is investigated. In this model, impulsive biological control which release infected pest to the field at a fixed time periodically is considered, and obtained the sufficient conditions for the global attractivity of pest-extinction periodic solution and permanence of the system. We also prove that all solutions of the model are uniformly ultimately bounded. The sensitive analysis on the two thresholds and to the changes of the releasing amounts of infected pest is shown by numerical simulations. Our results provide a reliable tactic basis for the practice of pest management.

Keywords: Pest Management, Stage Structure, Impulsive System, Permanence

1. Introduction

Pests outbreak often cause serious ecological and economic problems, and the warfare between human and pests has sustained for thousands of years. With the development of society and the progress of science and technology, a great deal of pesticides were used to control pests, because they can quickly kill a significant portion of pest population and sometimes provide the only feasible method for preventing economic loss. However, pesticide pollution is also recognized as a major health hazard to human beings and beneficial insects. At present, more and more people are concerned about the effects of pesticide residues on human health and on the environment [1].

In natural world, there are many insects whose individual members have a life history that takes them through two stages, larva and mature. Pathogens may not be effective against laver, that is, the disease only attacks the susceptible mature pest population. For example, saltcedar leaf beetle is such a pest. Pest control strategies have been attracted many experts over the past years. Recently, stage-structured models have received much attraction [2,3]. However, the epidemic models with stage-structure have been seldom studied. Zhang *et al.* [4] introduced the pest based on the stage-structure model which incorporates a discrete delay and pulses in order to investigate how epidemics influence the pest control process. An SI model with impulsive perturbations on diseased pest and spraying pesticides at fixed moment is proposed and investigated in [5], which obtained the sufficient conditions of the global attractivity of pest-extinction periodic solution and permanence of the system.

Incidence plays a very important role in research of epidemic models, bilinear and standard incidence rates have been frequently used in classical epidemic models [6]. Several different incidence rates have been proposed by researchers. Anderson *et al*. pointed out that standard incidence is more suitable than bilinear incidence [7,8]. Levin *et al.* have adopted the incidence form like $\beta S^q I^p$ or $\beta S^q I^p / N$ [9]. Lindstrom pointed out the crowed incidence $\beta S(t)/(1+aS(t)+bS^2(t))$ [10]. However there are seldom authors have concerned the stage-structured models under the simultaneous effect of disease and crowed incidence. A stage-structure model with the crowed incident rate is considered in [11]. According to the facts of pest management, we take the crowded effect as the incidence rate. Therefore, in this paper, a pest management epidemic model which with time delay and stage-structure is considered.

The present paper is organized as follows. In the next section, we formulate the pest management model. In Section 3, some essential lemmas which will be used to prove our main results are introduced. In Section 4, global attractivity of the susceptible pest-eradication pe-

riodic solution and the permanence of the model is analyzed. In the final section, we present some numerical simulations to illustrate the results and point out some future research directions.

2. Model Formulation

In this paper, we study the pest management epidemic model:

$$\left.\begin{cases} L'(t) = B(S(t))S(t) - \gamma L(t) - e^{-\gamma\tau}B(S(t-\tau))S(t-\tau), \\ S'(t) = e^{-\gamma\tau}B(S(t-\tau))S(t-\tau) - \dfrac{\beta S(t)I(t)}{1+aS(t)+bS^2(t)} - \eta S(t), \\ I'(t) = \dfrac{\beta S(t)I(t)}{1+aS(t)+bS^2(t)} - dI(t), \end{cases}\right\} t \neq nT, \tag{1}$$
$$\Delta I(t) = I(t^+) - I(t) = \mu, \quad t = nT, \ n \in Z_+,$$

with initial conditions

$$\begin{cases} (\varphi_1(t), \varphi_2(t), \varphi_3(t)) \in C_3^+ \ \text{ for } \ t \in [-\tau, 0], \ \varphi_i(0) > 0, \ i = 1,2,3, \\ \varphi_1(0) = \int_{-\tau}^{0} e^{r\theta} B(\varphi_2(\theta))\varphi_2(\theta)d\theta, \end{cases} \tag{2}$$

where all the coefficients of model (1) are nonnegative and $L(t), S(t), I(t)$ represent the larva, mature susceptible and infected pest population at time *t*, respectively. The model is derived from the following assumptions.

(H_1) The death rate of larva population is proportional to the existing larva population with proportionality constant γ, the death rate of mature susceptible and infected pest population is proportional to the existing mature susceptible and infected pest population with proportionality constants η and d, respectively.

(H_2) Only the susceptible pest population can reproduce. $B(S)$ is a birth rate function of the susceptible pest population for $S \in (0,\infty)$ with $B(S)$ is monotonically decreasing, $\lim_{S\to\infty} B(S) = B(\infty)$ exists and $B(0^+) > \eta > \bar{\delta} > B(\infty)$, where $\bar{\delta} = \frac{1}{2}\min\{\eta, \gamma, d\}$.

(H_3) τ represents a constant time to maturity, the product term $e^{-\gamma\tau}B(S(t-\tau))S(t-\tau)$ describes that immature pest who were laid at time $t-\tau$ and survive at time t.

(H_4) The incident rate is the crowded effect.

$$\frac{\beta S(t)}{1+aS(t)+bS^2(t)}.$$

(H_5) μ is the releasing amounts of infected pest at $t = nT, n = 1,2,\cdots,$ and T is the period of the impulsive effect.

Before going into any detail, we simplify model (1) and restrict our attention to the following model:

$$\left.\begin{cases} S'(t) = e^{-\gamma\tau}B(S(t-\tau))S(t-\tau) - \dfrac{\beta S(t)I(t)}{1+aS(t)+bS^2(t)} - \eta S(t), \\ I'(t) = \dfrac{\beta S(t)I(t)}{1+aS(t)+bS^2(t)} - dI(t), \\ \Delta I(t) = \mu, \quad t = nT. \end{cases}\right\} t \neq nT, \tag{3}$$

The initial conditions for (3) are

$$(\varphi_2(t), \varphi_3(t)) \in C^+ = C([-\tau, 0], R_+^2), \varphi_i(0) > 0, \ i = 2,3. \tag{4}$$

3. Some Useful Lemmas

The solution of system (1), denoted by $x(t) = (L(t), S(t), I(t))^T$ is a piecewise continuous function $x: R_+ \to R_+^3$, $x(t)$ is continuous on $(nT, (n+1)T)$, $n \in Z_+$ and $x(nT^+) = \lim_{t\to nT^+} x(t)$ exists. Before demonstrating the main results, we need to give some lemmas which will be used as follows.

Lemma 1. (see [11]). Let $(\varphi_1(t), \varphi_2(t), \varphi_3(t)) > 0$ for $-\tau < t < 0$. Then any solution of system (1) is strictly positive.

Lemma 2. Let the function $m \in PC'[R^+, R]$ satisfies the inequalities

$$\begin{cases} m'(t) \le p(t)m(t) + q(t), \ t \ge t_0, t \neq t_k, \ k = 1,2,\cdots, \\ m'(t_k^+) \le d_k m(t_k) + b_k, t = t_k, \end{cases}$$

where $p, q \in PC[R^+, R]$ and $d_k \ge 0$, b_k are constants. Then

$$m(t) \le m(t_0) \prod_{t_0<t_k<t} d_k \exp(\int_{t_0}^{t} p(s)ds) + \sum_{t_0<t_k<t} \left(\prod_{t_k<t_j<t} d_j \exp(\int_{t_0}^{t} p(s)ds) \right) b_k + \int_{t_0}^{t} \prod_{s<t_k<t} d_k \exp\left(\int_{s}^{t} p(\sigma)d\sigma\right) q(s)ds, \quad t \ge t_0.$$

The proof of this lemma is given in [12].

We now show that all solutions of (1) are uniformly ultimately bounded.

Lemma 3. Any solution $(L(t), S(t), I(t))$ of system (1) is uniformly ultimately bounded. That is, there exists a constant $M = \frac{\lambda}{\bar{\delta}} + \frac{\mu e^{\bar{\delta}T}}{e^{\bar{\delta}T} - 1} > 0$ such that $L(t) \le M$, $S(t) \le M$, $I(t) \le M$ for sufficiently large *t*.

Proof. Define $V(t) = L(t) + S(t) + I(t)$. By simple computation when $t \ne nT$, we calculate the derivative of *V* along the solution of system (1)

$$D^+V(t) = B(S)S - \gamma L - \eta S - dI \le B(S)S - 2\bar{\delta}(L + S + I),$$

for $t \in (nT, (n+1)T)$.

Then we derive

$$D^+V(t) + \bar{\delta}V \le B(S)S - \bar{\delta}S, \quad t \in (nT, (n+1)T).$$

Obviously, from conditions (H_1) and (H_2), we are easy to know that there exists a constant $\lambda > 0$ such that

$$D^+V(t) + \bar{\delta}V \le \lambda, \ t \in (nT, (n+1)T),$$

for *n* large enough.

When $t = nT$, we get

$$V(nT^+) = V(nT) + \mu.$$

According to Lemma 2, we derive

$$V(t) \le V(0)e^{-\bar{\delta}t} + \int_0^t \lambda e^{-\bar{\delta}(t-s)}ds + \sum_{0<nT<t} \mu e^{-\bar{\delta}(t-nT)} \to \frac{\lambda}{\bar{\delta}} + \frac{\mu e^{\bar{\delta}T}}{e^{\bar{\delta}T} - 1} \triangleq M \ \text{ as } \ t \to \infty.$$

Therefore by the definition of $V(t)$, we obtain that each positive solution of system (1) is uniformly ultimately bounded. This completes the proof.

Lemma 4. Consider the following delay differential equation:

$$x'(t) = a_1 x(t-\tau) - a_2 x(t). \tag{5}$$

where a_1, a_2 and τ are all positive constants and $x(t) > 0$ for $t \in [-\tau, 0]$. We have:

1) If $a_1 < a_2$, then $\lim_{x\to\infty} x(t) = 0$;

2) If $a_1 > a_2$, then $\lim_{x\to\infty} x(t) = +\infty$.

The proof of this lemma is given in [13].

Lemma 5 (see [11]). Consider the following impulsive system:

$$\begin{cases} v'(t) = -dv(t), & t \ne nT, \\ v(nT^+) = v(nT) + \mu, & t = nT, \quad n = 1, 2, \cdots, \end{cases} \tag{6}$$

where d, $\mu > 0$. Then there exists a unique positive periodic of system (6)

$$\tilde{v}(t) = v^* e^{-d(t-nT)}, \quad t \in (nT, (n+1)T], \ n \in Z_+,$$

which is globally asymptotically stable, where $v^* = \frac{\mu}{1 - e^{-dT}}$.

4. Main Results

In this section that follows we determine the global attractivity condition of the susceptible pest-extinction periodic solution and the permanence of the system (3).

4.1. Global Attractivity of the Susceptible Pest-Extinction Periodic Solution

Denote

$$R^* \triangleq \frac{\left(B(0)e^{-\gamma\tau} - \eta\right)(1 + aM + bM^2)(e^{dT} - 1)}{\mu\beta} \tag{7}$$

where $M = \frac{\lambda}{\bar{\delta}} + \frac{\mu e^{\bar{\delta}T}}{e^{\bar{\delta}T} - 1}$

Theorem 1. Let $(S(t), I(t))$ be any solution of system (3), the susceptible pest-extinction periodic solution $(0, \tilde{I}(t))$ of (3) is globally attractive provided that $R^* < 1$.

Proof. Since $R^* < 1$, we can choose ε_0 sufficiently small such that

$$B(0)e^{-\gamma\tau} < \frac{\beta}{1 + aM + bM^2}\left(\frac{\mu e^{-dT}}{1 - e^{-dT}} - \varepsilon_0\right) + \eta \tag{8}$$

Note that $I'(t) \ge -dI(t)$, from Lemma 2 and Lemma 5, we have that for the given ε_0 there exists an integer k_1 such that for $nT < t \le (n+1)T$, $n > k_1$

$$I(t) > \tilde{I}(t) - \varepsilon_0 \ge \left(\frac{\mu e^{-dT}}{1 - e^{-dT}} - \varepsilon_0\right) \triangleq \rho. \tag{9}$$

From condition (H_1), (3) and (9), we yield

$$\frac{dS(t)}{dt} \le B(0)e^{-\gamma\tau}S(t-\tau) - \left(\frac{\beta}{1 + aM + bM^2}\rho + \eta\right)S(t),$$

for $t > nT + \tau$, $n > k_1$.

Consider the following comparison differential system

$$\frac{dy(t)}{dt} = B(0)e^{-\gamma\tau}y(t-\tau) - \left(\frac{\beta}{1+aM+bM^2}\rho+\eta\right)y(t), \tag{10}$$

for $t > nT+\tau,\ n > k_1$.

From (8), we have $B(0)e^{-\gamma\tau} < \frac{\beta}{1+aM+bM^2}\rho+\eta$.

According to Lemma 4, we have $\lim_{t\to\infty} y(t) = 0$.

By the comparison theorem, we have $\limsup_{t\to\infty} S(t) < \lim_{t\to\infty} y(t) = 0$. Incorporating into the positivity of $S(t)$, we know that

$$\lim_{t\to\infty} S(t) = 0$$

Therefore, for any $\varepsilon_1 > 0$ (sufficiently small), there exists an integer k_2 $(k_2T > k_1T+\tau)$ such that $S(t) < \varepsilon_1$ for all $t > k_2T$.

Form system (3) and Lemma 5, we have

$$-dI(t) \le \frac{dI(t)}{dt} \le (-d+\beta\varepsilon_1)I(t).$$

Then we have $z_1(t) \le I(t) \le z_2(t)$ and $z_1(t) \to \tilde{I}(t)$, $z_2(t) \to \tilde{I}(t)$ as $t \to \infty$, while $z_1(t)$ and $z_2(t)$ are the solutions of

$$\begin{cases} z_1'(t) = -dz_1(t), & t \ne nT, \\ z_1(t^+) = z_1(t)+\mu, & t = nT, \\ z_1(0^+) = I(0^+), \end{cases}$$

and

$$\begin{cases} z_2'(t) = (-d+\beta\varepsilon_1)z_2(t), & t \ne nT, \\ z_2(t^+) = z_2(t)+\mu, & t = nT, \\ z_2(0^+) = I(0^+), \end{cases}$$

respectively, $\tilde{z}_2(t) = \frac{\mu\exp((-d+\beta\varepsilon_1)(t-nT))}{1-\exp((-d+\beta\varepsilon_1)T)}$ for $nT < t \le (n+1)T$. Therefore, for any $\varepsilon_2 > 0$, there exists an integer k_3, $n > k_3$ such that

$$\tilde{I}(t)-\varepsilon_2 < I(t) < \tilde{z}_2(t)+\varepsilon_2 \text{. for } t > nT.$$

Let $\varepsilon_2 \to 0$, we get $\tilde{z}_2(t) \to \tilde{I}(t)$ Hence $I(t) \to \tilde{I}(t)$ as $t \to \infty$. This completes the proof.

4.2. Permanence

Persistence (or permanence) is an important property of dynamical systems, in this section, we focus on the permanence of system (3).

Denote

$$R_* \triangleq \frac{(B(0)e^{-\gamma\tau}-\eta)(1-e^{(\beta S^*-d)T})}{\mu\beta}. \tag{11}$$

where $S^* = \frac{1}{\beta}\left[d+\frac{1}{T}\ln\left(1-\frac{\mu\beta}{B(0)e^{-\gamma\tau}-\eta}\right)\right] > 0.$

Theorem 2. Suppose $R_* > 1$. Then there is a positive constant q such that each positive solution $(S(t), I(t))$ of (3) satisfies $S(t) \ge q$, for sufficiently large t.

Proof. Let $(S(t), I(t))$ be the solution of system (3) with initial condition (4). Note that the first equation of (3) can be rewritten as

$$\frac{dS(t)}{dt} = e^{-\gamma\tau}B(S(t))S(t) - \eta S(t) - \frac{\beta S(t)I(t)}{1+aS+bS(t)^2} - e^{-\gamma\tau}\frac{d}{dt}\int_{t-\tau}^{t} B(S(\theta))S(\theta)d\theta.$$

In the following we define:

$$W(t) = S(t) + e^{-\gamma\tau}\int_{t-\tau}^{t} B(S(\theta))S(\theta)d\theta.$$

Then the derivative of $W(t)$ with respect to the solution of system (3) is governed by

$$\frac{dW}{dt} = \left(e^{-\gamma\tau}B(S(t)) - \eta - \frac{\beta I(t)}{1+aS(t)+bS^2(t)}\right)S(t).$$

Since $R_* > 1$, we can choose sufficiently small $S^*(< \frac{d}{\beta})$ and ξ such that

$$e^{-\gamma\tau}B(S^*) - \eta - \beta\left(\frac{\mu}{1-e^{(\beta S^*-d)T}}+\xi\right) > 0. \tag{12}$$

We claim that for any $t_0 > 0$, it is impossible that $S(t) < S^*$ for all $t \ge t_0$. Suppose that the claim is not valid. Then there is a $t_0 > 0$ such that $S(t) < S^*$ for all $t \ge t_0$. It follows from the second equation of system (3) that for

$$\frac{dI(t)}{dt} = \frac{\beta S(t)I(t)}{1+aS(t)+bS^2(t)} - dI(t) \le (-d+\beta S^*)I(t)$$

Consider the comparison impulsive system for $t \ge t_0$,

$$\begin{cases} z'(t) = (-d+\beta S^*)z(t), & t \ne nT, \\ z(t^+) = z(t)+\mu, & t = nT, \end{cases} \tag{13}$$

According to Lemma 1, we get the unique positive periodic solution of system (13)

$$\tilde{z}(t) = z^*e^{(\beta S^*-d)(t-nT)}, \quad nT < t \le (n+1)T,$$

is globally asymptotically stable, where $z^* = \frac{\mu}{1-e^{(\beta S^*-d)T}}$.

By the comparison theorem in impulsive differential equations, we know that for any sufficiently small $\varepsilon > 0$, there exists a $t_1 (> t_0 + \tau)$ such that

$$I(t) \le z^* + \varepsilon \triangleq \sigma, \tag{14}$$

for all $t > t_1$. It follows from (12) that $e^{-\gamma\tau} B(S^*) - \eta - \beta\sigma > 0$.

Further,

$$W'(t) > (e^{-\gamma\tau} B(S^*) - \eta - \beta\sigma) S(t) \quad \text{for} \quad t \ge t_1. \tag{15}$$

Set

$$S_m = \min_{t \in [t_1, t_1 + \tau]} S(t),$$

We will show that $S(t) \ge S_m$. for all $t \ge t_1 > t_0$. Otherwise, there is a $h \ge 0$ such that $S(t) \ge S_m$ for $t_1 \le t \le t_1 + \tau + h$, $S(t_1 + \tau + h) = S_m$ and $S'(t_1 + \tau + h) < 0$. Accordingly, from the first equation of (3) and the inequality (14) we yield

$$\begin{aligned} S'(t_1 + \tau + h) &= B(S(t_1 + h)) S(t_1 + h) e^{-\gamma\tau} \\ &\quad - \frac{\beta S(t_1 + \tau + h) I(t_1 + \tau + h)}{1 + aS(t_1 + \tau + h) + bS^2(t_1 + \tau + h)} \\ &\quad - \eta S(t_1 + \tau + h) \\ &\ge (B(S^*) e^{-\gamma\tau} - \beta\sigma - \eta) S_m > 0 \end{aligned}$$

which leads to a contradiction. Therefore $S(t) \ge S_m$ for all $t \ge t_1$. As a consequence, (15) leads to

$$W'(t) > (B(S^*) e^{-\gamma\tau} - \beta\sigma - \eta) S_m > 0$$

for $t \ge t_1$, which implies that $W(t) \to \infty$ as $t \to \infty$. This contradicts $W(t) \le M(1 + B(0)\tau e^{-\gamma\tau})$. The claim is proved.

By the claim, we need to consider two cases.

Case 1. $S(t) \ge S^*$ for all large t.

Case 2. $S(t)$ oscillates about S^* for that t is large enough. Define

$$q = \min\left\{\frac{S^*}{2}, q_1\right\}.$$

where $q_1 = S^* e^{-(\beta\sigma + \eta)T}$. We want to show that $S(t) \ge q$ for all large t. The conclusion is evident in the first case. For the second case, let $t^* > 0$ and $\upsilon > 0$ satisfy

$$S(t^*) = S(t^* + \upsilon) = S^* \quad \text{and} \quad S(t) < S^*,$$

for all $t \in (t^*, t^* + \upsilon)$, where t^* is sufficiently large such that

$$I(t) \le \sigma \quad \text{for } t^* < t < t^* + \upsilon,$$

$S(t)$ is uniformly continuous. The positive solutions of (3) are ultimately bounded and $S(t)$ is not affected by the impulses. Hence, there is a g ($0 < g < \tau$ and g is dependent of the choice of t^*) such that $S(t) > \frac{S^*}{2}$ for $t^* < t < t^* + g$.

If $\upsilon < g$, there is nothing to prove. Let us consider the case $g < \upsilon < \tau$. Since $S'(t) > -(\beta\sigma + \eta) S(t)$ and $S(t^*) = S^*$, hence $S(t) \ge q_1$ for $t^* + g \le t \le t^* + \tau$.

If $\upsilon > \tau$, it is obvious that $S(t) \ge q$ for $t \in [t^*, t^* + \tau]$. Then proceeding exactly as the proof for the above claim, we see that $S(t) \ge q$ for $t \in [t^* + \tau, t^* + \upsilon]$, because the kind of interval $t \in [t^*, t^* + \upsilon]$ is chosen in an arbitrary way (we only need t^* to be large). We concluded that $S(t) \ge q$ for all large t. In the second case, in view of our above discussion, the choice of q is independent of the positive solution, and we proved that any positive solution of (3) satisfies $S(t) \ge q$ for all sufficiently large t. This completes the proof.

Theorem 3. Suppose $R_* > 1$. Then system (1) is permanent.

Proof. Denote $(L(t), S(t), I(t))$ be any solution of system (1). From the second equation of system (3) and Theorem 2, we have

$$\frac{dI(t)}{dt} \ge I(t)\left(\frac{\beta q}{1 + aM + bM^2} - d\right).$$

Let $\frac{\beta q}{1 + aM + bM^2} \triangleq A$, it is easy to get $I(t) \to \infty$ if $A \ge d$, so we can always obtain the positive lower boundary by the theorem of differential equations. Otherwise, by the same argument as those in the proof of Theorem 1, we have $\liminf_{t \to \infty} I(t) \ge p$, where

$$p = \frac{\mu e^{(A-d)T}}{1 - e^{(A-d)T}} - \varepsilon.$$

In view of Theorem 2, the first equation of system (1) becomes

$$\frac{dL(t)}{dt} \ge B(M) q - MB(0) e^{-\gamma\tau} - \gamma L(t).$$

It is easy to obtain

$$\liminf_{t \to \infty} L(t) \ge \delta$$

where $\delta = \frac{B(M) q - MB(0) e^{-\gamma\tau}}{\gamma} - \varepsilon$. By Theorem 2 and the above discussion, system (1) is permanent. The proof of Theorem 3 is complete.

5. Numerical Analysis and Discussion

We have studied a delayed epidemic model with stage-structure and impulses, theoretically analyze the influ-

ence of impulsive releasing for the infected pest population, and also obtained that the pest-extinction periodic solution of system (1) is globally attractive if the control variable $R^* < 1$ given by (7), and the system is permanent with $R_* > 1$ which is given by (11). We know that, besides the release amount of infectious pests each time, the period of impulsive vaccination and the effective contact rate play an important role in the dynamical behavior of the system. In the following, we will specially analyze the influence of the release amount of infectious pests to the dynamical system. We assume $B(S) = e^{-S}$, and consider the hypothetical set of parameter values as $\gamma = 0.45$, $\tau = 1$, $a = 0.1$, $b = 0.01$, $\beta = 0.75$, $\eta = 0.5$, $d = 0.6$, $T = 4$, $\lambda = 0.001$.

By Theorem 1 and Theorem 3, we know that when R^*=0.9929 < 1, the pest-extinction periodic solution of system (3) is globally attractive and the susceptible pest population becomes extinct (see **Figure 1**). When the release amount of infected pest reaches a certain value μ=0.1 such that $R_* = 1.6681 > 1$, the system (3) is permanent (see **Figure 2**). So far we have only discussed two cases: $R^* < 1$ and $R_* > 1$. But for $R_* \le 1 \le R^*$, the susceptible pest population either becomes extinct (see **Figure 3**) or coexists to the infect pest population (see **Figure 4**). According to the above numerical simulation, we think there exists a threshold parameter to decide the extinction of the susceptible pest population and the permanence of the system. These issues will be considered in our future research.

From **Figure 5(a)**, we can observe that: R_* is sensitive to μ value as μ is small enough, whereas not sensitive to μ value as $\mu > 1.5$. We can also get the similar phenomenon from **Figure 5(b)**. We hope that our results will provide an insight to pest management practicing.

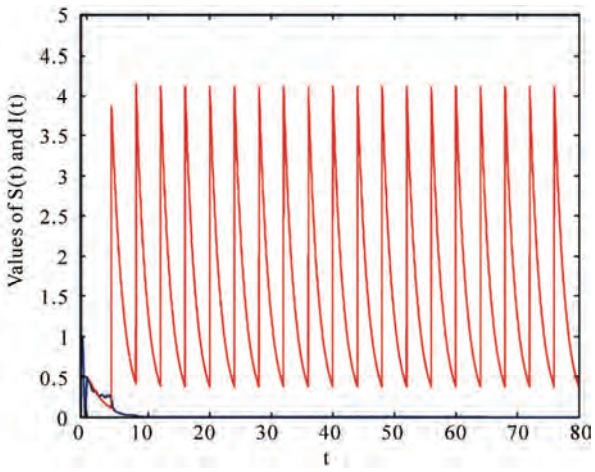


Figure 1. Dynamical behavior of system (3) with $\mu = 3.8$, $R^* = 0.9929 < 1$.

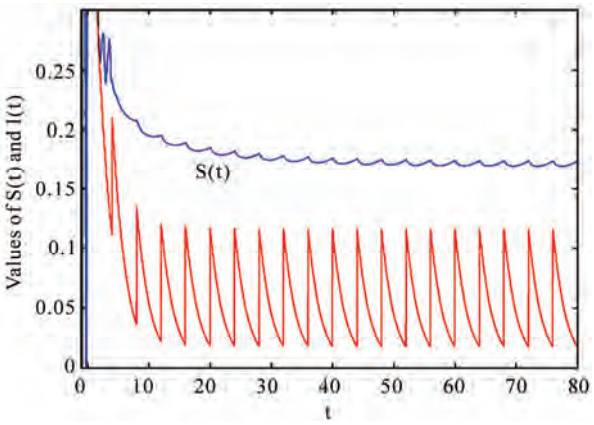


Figure 2. Dynamical behavzior of system (3) with $\mu = 0.1$, $R_* = 1.6681 > 1$.

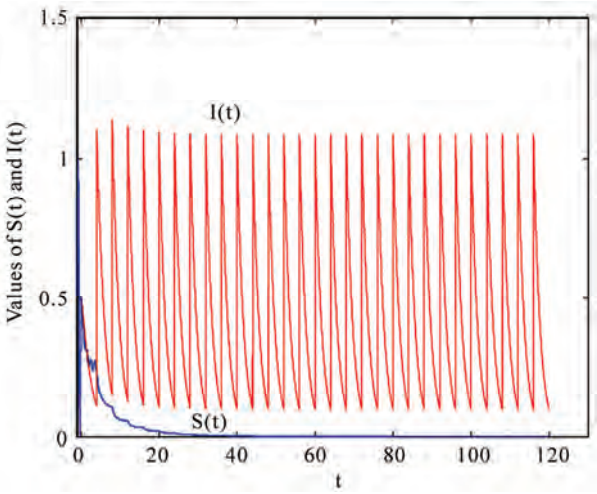


Figure 3. Dynamical behavior of system (3) with $\mu = 1$, $R_* = 0.1669 < 1$ and $R^* = 2.2026 > 1$.

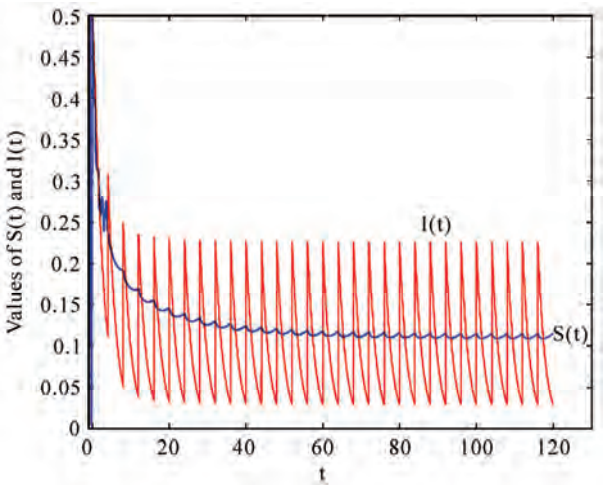
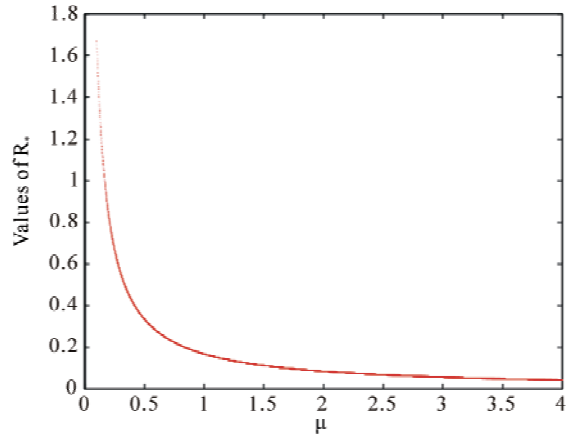
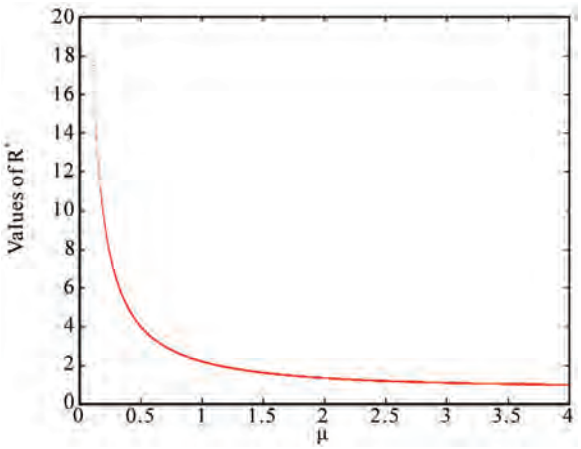


Figure 4. Dynamical behavior of system (3) with $\mu = 0.2$, $R_* = 0.8343 < 1$ and $R^* = 9.5212 > 1$.



(a)



(b)

Figure 5. The sensitive analysis of μ to R_* and R^*. (a) $\mu - R_*$; (b) $\mu - R^*$.

6. Acknowledgements

The research have been supported by The Natural Science Foundation of China (10971037), The National Key Technologies R & D Program of China (2008BAI68B01), The Postgraduate Innovation Fund of Jiangxi Province (YC09A124).

7. References

[1] R. Q. Shi and L. S. Chen, "An Impulsive Predator-Prey Model with Disease in the Prey for Integrated Pest Management," *Communications in Nonlinear Science and Numerical Simulation*, Vol. 15, No. 2, 2010, pp. 421-429.

[2] J. A. Cui and X. Y. Song, "Permanence of a Predator-Prey System with Stage Structure," *Discrete Continuous Dynamical Systems-Series B*, Vol. 4, No. 3, 2004, pp. 547-554.

[3] Y. N. Xiao and L. S. Chen, "Global Stability of a Predator-Prey System with Stage Structure for the Predator," *Acta Mathematica Sinica*, Vol. 20, No. 1, 2004, pp. 63-70.

[4] H. Zhang, L. S. Chen and J. J. Nieto, "A Delayed Epidemic Model with Stage-Structure and Pulses for Pest Management Strategy," *Nonlinear Analysis*: *Real World Applications*, Vol. 9, No. 4, 2008, pp. 1714-1726.

[5] X. Wang, Y. D. Tao and X. Y. Song, "Mathematical Model for the Control of a Pest Population with Impulsive Perturbations on Diseased Pest," *Applied Mathematical Modelling*, Vol. 33, No. 7, 2009, pp. 3099-3106.

[6] H. W. Hethcote, "The Mathematics of Infectious Disease," *Siam Review*, Vol. 42, No. 4, 2002, pp. 599-653.

[7] R. M. Anderson, R. M. May and B. Anderson, "Infectious Diseases of Human: Dynamics and Control," Oxford Science Publications, Oxford, 1991.

[8] R. M. Anderson and R. M. May, "Population Biological of Infectious Diseases," Springer Berlin-Heidelberg, New York, 1982.

[9] W. M. Liu, S. A. Levin and Y, Lwasa, "Influence of Nonlinear Incidence Rates upon the Behavior of SIRS Epidemiological Models," *Journal of Mathematical Biology*, Vol. 23, No. 2, 1986, pp. 187-204.

[10] T. Lindstrom, "A Generalized Uniqueness Theorem for Limit Cycles in a Predator-Prey System," *Acta Academic Akoensis*, *Series B*, Vol. 49, No. 2, 1989, pp. 1-9.

[11] J. J. Jiao, X. Z. Meng ang L. S. Chen, "Global Attractivity and Permanence of a Stage-Structured Pest Management SI Model with Time Delay and Diseased Pest Impulsive Transmission," *Chaos, Solitons and Fractals*, Vol. 38, No. 3, 2008, pp. 658-668.

[12] V. Lakshmikantham, D. D. Bainov and P. Simeonov, "Theory of Impulsive Differential Equations," World Scientific, Singapor, 1989.

[13] Y. Kuang, "Delay Differential Equation with Application in Population Dynamics," Academic Press, New York, 1993.

Applied Mathematics, **2010, 1, 222-229**
doi:10.4236/am.2010.13027 Published Online September 2010 (http://www.SciRP.org/journal/am)

Solving Large Scale Nonlinear Equations by a New ODE Numerical Integration Method

Tianmin Han[1], Yuhuan Han[2]
[1]*China Electric Power Research Institute, Beijing, China*
[2]*Hedge Fund of America, San Jose, USA*
E-mail: *han_tianmin@yahoo.com.cn, ibmer.ibm@gmail.com, william.han@eMallGuide.com*
Received June 11, 2010; *revised July* 23, 2010; *accepted July* 26, 2010

Abstract

In this paper a new ODE numerical integration method was successfully applied to solving nonlinear equations. The method is of same simplicity as fixed point iteration, but the efficiency has been significantly improved, so it is especially suitable for large scale systems. For Brown's equations, an existing article reported that when the dimension of the equation N = 40, the subroutines they used could not give a solution, as compared with our method, we can easily solve this equation even when N = 100. Other two large equations have the dimension of N = 1000, all the existing available methods have great difficulties to handle them, however, our method proposed in this paper can deal with those tough equations without any difficulties. The sigularity and choosing initial values problems were also mentioned in this paper.

Keywords: Nonlinear Equations, Ordinary Differential Equations, Numerical Integration, Fixed Point Iteration, Newton's Method, Stiff, Ill-Conditioned

1. Introduction

The classic methods for solving nonlinear equations $F(X) = 0$ mainly have the following two types:

1) Fixed Point Iteration:

$$X_{n+1} = G(X_n) \quad \text{here } G(X) = F(X) + X \tag{1}$$

2) Newton Iteration:

$$X_{n+1} = X_n - J(X_n)^{-1} F(X_n) \quad \text{here } J(X) \text{ is the Jacobian of } F(X) \tag{2}$$

As the book [1] Pg. 17 described, the solution of the nonlinear system $F(X) = 0$ can be interpreted as steady states or equilibrium point of the dynamic system $\dot{X} = F(X)$ In fact, those two iterations are all equivalent to explicit Euler method in the field of ODE numerical integration.

For the differential equation:

$$\dot{X} = F(X) \tag{3}$$

The Euler method:

$$X_{n+1} = X_n + hF(X_n) \tag{4}$$

is a general expression of fixed point iteration [1] Pg.299

If we take $h = 1$, we can get (1)

As for Newton iteration, for the differential equation:

$$\dot{X} = -J(X)^{-1} F(X) \tag{5}$$

using explicit Euler method:

$$X_{n+1} = X_n - hJ(X)^{-1} F(X) \tag{6}$$

and taking $h = 1$ we get (2)

These relations can also be found in [2] Pg.768 or [3] §7.5.

We developed a set of numerical integration method in [4]. They have accuracy 1st-5th order. Among them, the simplest one is the 1st order PEC scheme. This scheme has very large stable region, so we can take it as a tool to integrate the differential equation and get fast convergence speed to solve $F(X) = 0$.

For the sake of completeness, we rederive the algorithm in the next section.

2. Algorithm

Consider the problem:

$$\begin{cases} \dot{X} = F(X) \\ X(0) = X_0 \end{cases} \tag{7}$$

Using implicit Euler method:

$$X_{n+1} = X_n + hF\left(X_{n+1}\right) \tag{8}$$

introducing variable

$$Z_{n+1} = hF\left(X_{n+1}\right) \tag{9}$$

we have

$$Z_{n+1} = X_{n+1} - X_n \tag{10}$$

Multiplying both sides of (9) by ε ($\varepsilon > 0$), we obtain

$$\varepsilon Z_{n+1} = \varepsilon hF\left(X_{n+1}\right) \tag{11}$$

Equation (11) can be reformulated as follows:

$$\left(\varepsilon/h\right)Z_{n+1} = \varepsilon F\left(X_{n+1}\right) \tag{12}$$

Equation (12) plus Equation (10), we obtain

$$\left(1+\varepsilon/h\right)Z_{n+1} = \varepsilon F\left(X_{n+1}\right) + X_{n+1} - X_n \tag{13}$$

Let

$$\omega = h/\left(h+\varepsilon\right) \tag{14}$$

Equation (13) can be rewritten as

$$Z_{n+1} = \omega\left(\varepsilon F\left(X_{n+1}\right) + X_{n+1} - X_n\right) \tag{15}$$

From (8) and (9), we have

$$X_{n+1} = X_n + Z_{n+1} \tag{16}$$

Combining (15) and (16), we obtain a new implicit integration method, which is fully equivalent to (8).

We use the simple iteration method to solve the implicit system (15) and (16), and choose the initial iteration value $X_{n+1}^{(0)} = X_n + Z_n$. Only one iteration applies to the implicit system (15) and (16), then we obtain an explicit integration scheme as follows:

$$\begin{cases} X_{n+1}^{(0)} = X_n + Z_n \\ Z_{n+1} = \omega\left(\varepsilon F\left(X_{n+1}^{(0)}\right) + Z_n\right) \\ X_{n+1} = X_n + Z_{n+1} \end{cases} \tag{17}$$

(17) is named as the EPS method in this article.

In order to investigate the stability of the EPS method (17), we consider the model equation

$$\dot{x} = \lambda x \tag{18}$$

where λ is a complex number. Then, we have

$$\begin{cases} x_{n+1} = x_n + z_{n+1} \\ z_{n+1} = \omega\varepsilon\lambda\left(x_n + z_n\right) + \omega z_n \end{cases} \tag{19}$$

or the matrix form

$$\begin{bmatrix} x_{n+1} \\ z_{n+1} \end{bmatrix} = \begin{bmatrix} 1+\omega\varepsilon\lambda & \omega+\omega\varepsilon\lambda \\ \omega\varepsilon\lambda & \omega+\omega\varepsilon\lambda \end{bmatrix} \begin{bmatrix} x_n \\ z_n \end{bmatrix} \tag{20}$$

The characteristic equation of (20) is given by

$$\mu^2 - \left(1+\omega+2\omega\varepsilon\lambda\right)\mu + \omega\left(1+\varepsilon\lambda\right) = 0 \tag{21}$$

Let

$$\varepsilon = \alpha h \tag{22}$$

From (21), we obtain

$$h\lambda = \frac{\mu^2 - \left(1+\omega\right)\mu + \omega}{\left(2\mu-1\right)\omega\alpha} \tag{23}$$

Giving α a special value, let μ vary and keep $|\mu| = 1$, then we obtain an enclosed curve, which is just the boundary of the absolute stability region in $h\lambda$-plane. Set $\mu = \cos\theta + j\sin\theta$, $j = \sqrt{-1}$, $0 \le \theta \le 2\pi$, then we rewrite (23) as follows:

$$\mathrm{Re}\left(h\lambda\right) = -\left(\left(1+\omega-2\sin^2\theta-\cos\theta-\omega\cos\theta\right)\left(1-2\cos\theta\right)-2\left(2\cos\theta-1-\omega\right)\sin^2\theta\right)\Big/\left(\left(5-4\cos\theta\right)\alpha\omega\right) \tag{24}$$

$$\mathrm{Im}\left(h\lambda\right) = -\left(2\left(1+\omega-2\sin^2\theta-\cos\theta-\omega\cos\theta\right)\sin\theta + \left(2\cos\theta-1-\omega\right)\sin\theta\left(1-2\cos\theta\right)\right)\Big/\left(\left(5-4\cos\theta\right)\alpha\omega\right) \tag{25}$$

The curve of the boundary of the absolute stability region is obtained when θ varies from 0 to 2π. If α is a small number, the stability region will be close to real axis and spreads far away towards the left-half plane. For example, when α = 0.01, as it is shown in **Figure 1**, the left end point of the stability region can reach −134, so the integration step size can be increased significantly.

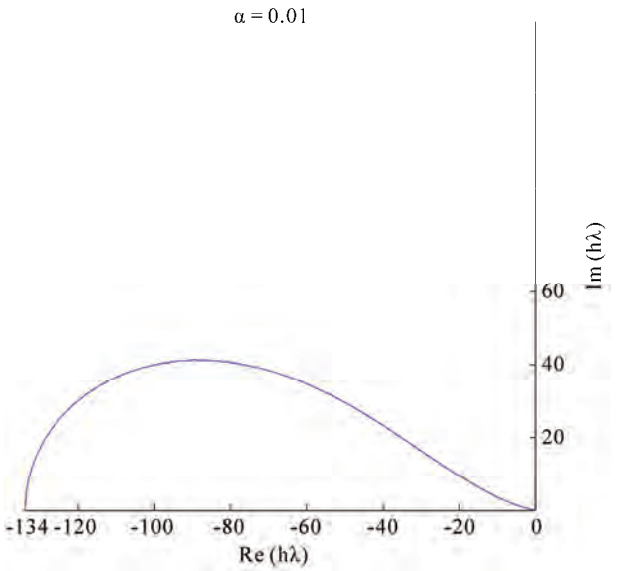


Figure 1. The absolute stability region of the EPS method for α = 0.01.

In the model Equation (18), if λ is very close to the imaginary axis, *i.e.*, $Im(\lambda) \gg Re(\lambda)$, α should be taken a bigger value. For α = 100, the stability region is shown by **Figure 4**. We can find that the region includes a section of the imaginary axis. This property is unusual for an explicit method.

When α = 1, *i.e.*, $h = \varepsilon$, then the stability region of the EPS method is all the same as the explicit Euler method. It is enclosed by a circle with center at (−1, 0) and its radius is 1. In fact, in (24) and (25) taking α = 1, then ω = 0.5, we have

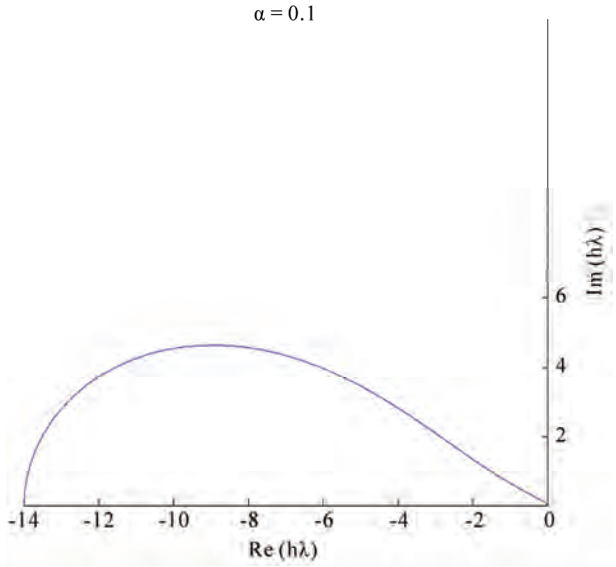


Figure 2. The absolute stability region of the EPS method for α = 0.1.

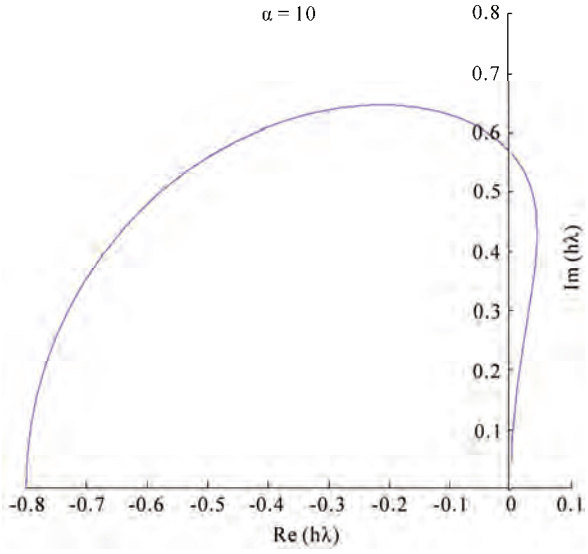


Figure 3. The absolute stability region of the EPS method for α = 10.

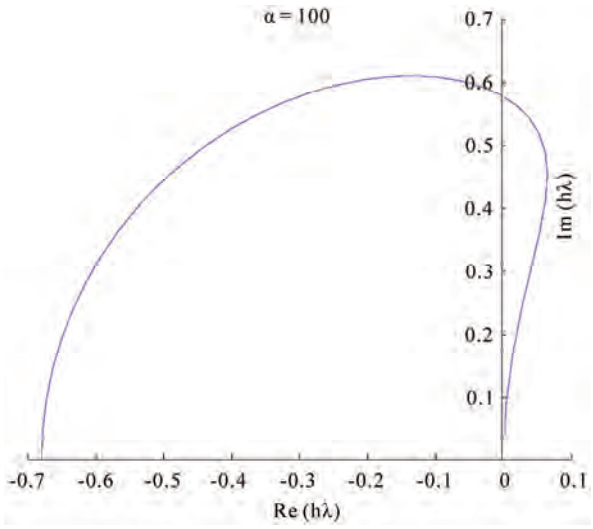


Figure 4. The absolute stability region of the EPS method for α = 100.

$$\begin{aligned}\mathrm{Re}(h\lambda) &= -\left(\left(1.5-2\sin^2\theta-1.5\cos\theta\right)\left(1-2\cos\theta\right)\right.\\ &\quad \left.-4\cos\theta\sin^2\theta+3\sin^2\theta\right)\Big/\left(2.5-2\cos\theta\right)\\ &= \cos\theta-1\end{aligned} \tag{26}$$

$$\begin{aligned}\mathrm{Im}(h\lambda) &= -\left(\left(3-4\sin^2\theta-3\cos\theta\right)\sin\theta\right.\\ &\quad \left.+\left(5\cos\theta-1.5-4\cos^2\theta\right)\sin\theta\right)\Big/\left(2.5-2\cos\theta\right)\\ &= \sin\theta\end{aligned} \tag{27}$$

3. Implementation of the Algorithm

In this article, we merely discuss how to use the EPS method to integrate the differential equation $\dot{X} = F(X)$. Usually ODE integration methods require the condition $\partial F/\partial X < 0$ holds. That is to say the eigenvalues of the Jacobian distribute in the left-half part of the complex plane. For our purpose, to solve $F(x)$ = 0 and to solve $-F(x)$ = 0 are equivalent. In other field the "half plane condition" is always said to be "positive definite", *i.e.*, the eigenvalues are in the right plane. This fact reminds us the differential equation to deal with is $\dot{X} = -F(X)$.

The EPS method can also be applied to the differential equation

$$\dot{X} = -J(X)^{-1} F(X) \tag{28}$$

In this case, if $F(X)$ is replaced with $-F(X)$ in (28), it does not change the form of (28). So the sign in front of $F(X)$ is meaningless at all. By the way, choosing $\varepsilon = h$ = 1, according to many numerical experiments have done by us, the numerical results of EPS are almost the same

as the numerical results of the Newton's method (the details are not given in this article).

Despite the EPS method is a Jacobian-Free method, if it is not difficult to obtain the diagonal matrix $D(X)$ of the Jacobian $J(X)$, then we can integrate differential equation

$$\dot{X} = -D(X)^{-1} F(X) = -G(X)$$

we can get even much better results, especially, when $J(X)$ is a diagonal dominant matrix. However, it needs to consider a strategy to avoid overflow when some elements of the matrix $D(X)$ are very small.

At present we have not developed a adaptive program which can automatically choose parameter ε and the step size h, but we give a strategy roughly as follows.

For non-stiff system, we pick up the parameter ε on [0.5, 1.0] and determine h by the size of $\|F(X)\|$. For stiff system, we need to estimate the spectral radius ρ of the Jacobian matrix $J(X)$ such that $\varepsilon\rho < 1$ is satisfied. In fact, if λ is a positive real number, for $\dot{x} = -\lambda x$, when $\varepsilon\lambda \le 4/3$, we can prove that the scheme (19) is stable for all h ($0 < h < \infty$). Small value ε can strengthen stability but will reduce the efficiency.

For some easy problems we can take fixed step size in the whole calculating process. Usually we divide the calculating process into three stages, in each stage, different step size will be taken.

To do this, we set three parameters TOL_1, TOL_2, TOL_3. At first, we choose step size h_1 to start the calculation till $\|F\| < TOL_1$ is satisfied, the first stage is completed. Taking current value of X as initial value, we start the second stage calculation with step size h_2 till $\|F\| < TOL_2$. Do the same as we have done till finally $\|F\| < TOL_3$, then we end our calculation. In this paper, the $\|\cdot\|$ means Euclidean norm.

Outline of the Algorithm

Step 1. Give an initial value X_0. Set ε, h and compute $\omega = h/(h+\varepsilon)$.

Step 2. Compute $F(X)$ and $D(X)$ if it is needed.

Step 3. Compute $G(X) = D(X)^{-1} F(X)$ or $G(X) = F(X)$. If an element $d_i(X)$ of matrix $D(X)$ is less than one, the division is omitted and we have $g_i(X) = f_i(X)$.

Step 4. Compute $Z_0 = hG(X_0)$.

Step 5. $X := X + Z$.

Step 6. Compute $F(X)$, $D(X)$, $G(X)$ by the way of Step 2 and Step 3.

Step 7. If $F(X) < TOL$, then stop, else do

$$X := X - Z$$
$$Z := \omega(\varepsilon G(X) + Z)$$
$$X := X + Z$$

Go to Step 5.

4. Numerical Experiments

We now present numerical results for five examples. Some of them have already had results in literature. So we can compare our results with theirs. We also compare the results of fixed point iteration (explicit Euler method) with ours as well. This is because we identify our method as an improvement for the fixed point iteration and the explicit Euler method was well represented in all explicit methods.

4.1. Example 1 [2]

$$f_1(X) = x_1^2 - x_2 + 1$$
$$f_2(X) = x_1 - \cos\left(\frac{\pi}{2}x_2\right)$$

The initial value $x_0 = (1, 0)$. The solution we want to seek is $x^* = (0, 1)$. The Jacobian of the system is:

$$J(X) = \begin{bmatrix} 2x_1 & -1 \\ 1 & \frac{\pi}{2}\sin\left(\frac{\pi}{2}x_2\right) \end{bmatrix}$$

and the determinant of the Jacobian is given by

$$\det(J(X)) = x_1\pi\sin\left(\frac{\pi}{2}x_2\right) + 1$$

So at the line $\sin\left(\frac{\pi}{2}x_2\right) = -\frac{1}{\pi x_1}$, the singularity occurs.

Newton method does not converge to x^* but rather, it crosses the singularity line and converges in eight iterations to $x^* = \left(-\frac{1}{2}\sqrt{2}, \frac{3}{2}\right)$.

The damped Newton method was also applied to this problem and it converged to x^* in 107 iterations. The total number of function evaluations is as many as 321.

In [2], there are 12 algorithms, all of them are based on trapezoid formula, have been tested for this example. Among them the PE_BCE_B is the best, here the E_B means using Broyden method to approximate J^{-1}. The iteration is 17 times and the evaluation is 36 times.

There are four algorithms, each of them needs iterate more than 100 times. The rest seven algorithms need to iterate 23~47 times and evaluate 68~282 times respectively. All those calculations use double precision.

This example was considered as a difficult problem, because the differential equation to deal with is $\dot{X} = -J(X)^{-1} F(X)$ and the Jacobian is singular.

If the differential equation to be handled is $\dot{X} = -F(X)$, all the trouble will disappear. In fact it is a non-stiff

equation, we can reach the equilibrium point easily. In our calculation the single precision was used. For the sake of comparing with [2], we take $|f_1| < 10^{-5}$, $|f_2| < 10^{-5}$ as convergence criteria.

Explicit Euler method and EPS method were tested for this example. For Euler method, taking step size h = 0.24, 0.25, 0.26, 0.27, 0.28, the results show h = 0.28 overflow happened. The numbers of function evaluation for other step size were 74,72,72,84 respectively. The best result was given by h = 0.25: x_1 = .1879 × 10^{-4}, x_2 = .1000 × 10^1, f_1 = −.6020 × 10^{-5}, f_2 = −.9568 × 10^{-5}. EPS method: take ε = 1 and h = 0.4, 0.5, 0.6, overflow happened at h = 0.6. For h = 0.4, 37 times evaluation was needed. The best result was give by h = 0.5: the numbers of evaluation is 31, x_1 = .2418 × 10^{-5}, x^2 = .1000 × 10^1, f_1 = −.5305 × 10^{-5}, f_2 = .7182 × 10^{-5}.

4.2. Example 2

We construct a large scale mild stiff system to test our method. For $F(X) = A(X) - b = 0$, the differential equation is $\dot{X} = -F(X) = -(A(X) - b)$. Here b is a constant vector and $A(X) = UDUC(X)$, $C(X) = (x_1^3, x_2^3, \ldots, x_N^3)^T$, $U = \left(I - \frac{2}{u^T u} u u^T\right)\left(U = U^T = U^{-1}\right)$, u = (1, 1,…, 1)$^{\mathrm{T}}$, N = 1000. D is block diagonal matrix: D = *diag* (D_i) and $D_i = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix}$. The solution of $F(X)$ = 0 is X^* = (1, 1, …, 1)$^{\mathrm{T}}$. For this X^*, when the value of D_i was given, the value of b can be calculated, and the equation can be entirely determined.

The characteristic of the system depends on the choice of D_i. In the following three types of the D are given, the initial values for all of them are X_0 = (0, 0, …, 0)$^{\mathrm{T}}$. In this problem, using the special form of U and D the function $F(X)$ can be easily computed, but to compute the Jacobian is no longer an easy task, we must compute every element of the matrix. Furthermore, the Jacobian is a dense matrix and the Newton method will lose all superiority for this large scale system.

Example 2.1 Take D as a diagonal matrix, *i.e.*, in D_i we put $b_i = c_i = 0$, $a_i = 2_i - 1$, $d_i = 2_i$, i = 1, 2, …, N/2. The results of EPS method and explicit Euler method are listed in the **Tables 1**, **2**. Here the NFE is the abbreviation of Number of Function Evaluations and the "Step Size" means the best step size, the highest efficiency was reached by this step size.

Compare **Tables 1** and **2**, for EPS Method in three stages the Step Size h_1, h_2, h_3 have the relation $h_3 = 2h_2 = 2(2h_1) = 4h_1$, but for Euler Method h_1, h_2, h_3 almost keep a constant. The ratio of NFE is 1244/12003 ≈ 0.1

Example 2.2 The subblocks of D have the following form:

$$D_i = \begin{bmatrix} 2i & i \\ -i & 2i \end{bmatrix},\ i = 1, 2, \ldots, N/2$$

i.e., the eigenvalues of matrix D distribute in a wedge region. The results of both methods are listed in **Table 3** and **Table 4**.

The ratio of NFE is 2219/8014 ≈ 0.28

Example 2.3 $D_i = \begin{bmatrix} 1 & i/100 \\ -i/100 & 1 \end{bmatrix},\ i = 1, 2, \ldots, N/2$.

The eigenvalues of D distribute in a line. The line is parallel with imaginary axis. The maximum ratio of imaginary part and real part is 5:1. The results of both methods are listed in the **Tables 5** and **6**.

The ratio of NFE is 499/1379 ≈ 0.36. We did not give the Jacobian of F, but according to the situation of matrix

Table 1. EPS method ε = 0.0004; $\|F(X_0)\|$ = .1827 × 10^5.

Tolrance	Step Size	NFE	$\|F\|$
TOL1 = 1D - 0	0.0025	119	0.99 × 10^0
TOL2 = 1D - 5	0.005	669	0.9810 × 10^{-5}
TOL3 = 1D - 10	0.01	1244	0.9610 × 10^{-10}

Table 2. Euler method $\|F(X_0)\|$ = .1827 × 10^5.

Tolrance	Step Size	NFE	$\|F\|$
TOL1 = 1D - 0	0.00055	597	0.9980 × 10^0
TOL2 = 1D - 5	0.00066	6099	0.9985 × 10^{-5}
TOL3 = 1D - 10	0.0066	12003	0.9685 × 10^{-10}

Table 3. EPS method ε = 0.00025, $\|F(X_0)\|$ = .2044 × 10^5.

Tolrance	Step Size	NFE	$\|F\|$
TOL1 = 1D - 0	0.001	273	0.9967 × 10^0
TOL2 = 1D - 5	0.002	1165	0.9890 × 10^{-5}
TOL3 = 1D - 10	0.004	2219	0.9977 × 10^{-10}

Table 4. Euler method $\|F(X_0)\|$ = .2044 × 10^5.

Tolrance	Step Size	NFE	$\|F\|$
TOL1 = 1D - 0	0.00044	636	0.9987 × 10^0
TOL2 = 1D - 5	0.000528	4223	0.9971 × 10^{-5}
TOL3 = 1D - 10	0.000528	8014	0.9860 × 10^{-10}

D, we can get a general conception for the distribution of eigenvalues of the Jacobian. Compare three cases above, we can conclude that if the eigenvalues are close to real axis the EPS method will be more efficient.

4.3. Example 3

Brown's Almost Linear Function

$$f_i(X) = X_i + \sum_{j=1}^{N} X_j - (N+1), i = 1, 2, \ldots, N-1$$

$$f_N(X) = \prod_{j=1}^{N} X_j - 1$$

the initial values are $X_i(0) = 0.5$. The solution to be searched is $X^* = (1, 1, \ldots, 1)^T$. This is a difficult problem. Brown in [5] reported his research work. For $N = 5$ Newton method converged to the root given approximately by (−.579, −.579, −.579, −.579, 8.90); however, for N = 10, 30 Newton method diverged quite rapidly.

Brown's method did an excellent work, for N = 5, 10, 30, after 6, 7, 9 times iteration they all converged to X^* For N = 10, 30, 40 the authors of [6] tested their elaborate subroutines *NEQ*1 and *NEQ*2 for this tough problem, unfortunately the test failed for N = 40. Let us take a look at the differential equation:

$$\dot{X} = -F(X)$$

The last row of th Jacobian is $\partial f_N(X)/\partial X_i = \prod_{j=1, j\neq i}^{N} X_j$. When N is large enough, at the neighborhood of the initial point this row almost equals zero vector, so the equation is considered a very stiff or ill-conditioned system for large N.

The differential equation virtually to deal with is:

$$\dot{X} = -D(X)^{-1} F(X)$$

The diagonal matrix $D(X)$ has elements $d_i(X) = 2.0$, i = 1, 2, …, $N-1$ and $d_N(X) = \prod_{j=1}^{N-1} X_j$. If the value of $d_N(X)$ is very small, the measures must be taken to avoid overflow (for the details see paragraph 3). As we mentioned before, we divided the calculation into three stages and took different step size for each stage. For N = 10, 30, 40, 100, the results of EPS method were listed in **Tables 7-10**.

Explicit Euler method (fixed point iteration) can also get the results, but the expense was very expensive. The change in step size is very small in different stage. We use:

$$N(h_1, h_2, h_3)M$$

to express the dimension of the equation, three different

Table 5. EPS method ε = 0.1, $\|F(X_0)\|$ = .8396 × 10^2.

Tolrance	Step Size	NFE	$\|F\|$
TOL1 = 1D - 0	0.01	217	0.9762 × 10^0
TOL2 = 1D - 5	0.02	401	0.9580 × 10^{-5}
TOL3 = 1D - 10	0.04	499	0.9789 × 10^{-10}

Table 6. Euler method $\|F(X_0)\|$ = :8396 × 10^2.

Tolrance	Step Size	NFE	$\|F\|$
TOL1 = 1D - 0	0.011	255	0.9894 × 10^0
TOL2 = 1D - 5	0.0132	790	0.9935 × 10^{-5}
TOL3 = 1D - 10	0.0132	1379	0.9835 × 10^{-10}

Table 7. N = 10, $\varepsilon = \frac{2}{10}$, $\|F(X_0)\|$ = .1653 × 10^2.

Tolrance	Step Size	NFE	$\|F\|$
TOL1 = 1D - 0	0.65	5	0.4991 × 10^0
TOL2 = 1D - 5	1.0	35	0.2714 × 10^{-5}
TOL3 = 1D - 10	1.2	119	0.8864 × 10^{-10}

Table 8. N = 30, $\varepsilon = \frac{2}{30}$, $\|F(X_0)\|$ = .8348 × 10^2.

Tolrance	Step Size	NFE	$\|F\|$
TOL1 = 1D - 0	0.3	6	0.3737 × 10^0
TOL2 = 1D - 5	0.9	61	0.4548 × 10^{-5}
TOL3 = 1D - 10	1.2	277	0.5100 × 10^{-10}

Table 9. N = 40, $\varepsilon = \frac{2}{40}$, $\|F(X_0)\|$ = .1280 × 10^3.

Tolrance	Step Size	NFE	$\|F\|$
TOL1 = 1D - 0	0.2	6	0.6212 × 10^0
TOL2 = 1D - 5	0.6	41	0.5133 × 10^{-5}
TOL3 = 1D - 10	1.2	293	0.1592 × 10^{-10}

Table 10. N = 100, $\varepsilon = \frac{2}{100}$, $\|F(X_0)\|$ = .5025 × 10^3.

Tolrance	Step Size	NFE	$\|F\|$
TOL1 = 1D - 0	0.1	7	0.2985 × 10^0
TOL2 = 1D - 5	0.3	57	0.5787 × 10^{-5}
TOL3 = 1D - 10	1.2	640	0.4534 × 10^{-10}

step size and the NFE. The results are as follows: 10 (0.2, 0.25, 0,3) 788, 30 (0.11, 0.11, 0.112) 4586, 40 (0.09, 0.09, 0.09) 7540, 100 (0.035, 0.035, 0.035) 42183.

The ratio of NFE for both methods are: 119/788 ≈ 0.15, 277/4586 ≈ 0.06, 293/7540 ≈ 0.04, 640/42183 ≈ 0.015. From the data above we can see that along with N increasing the ill-conditioned extent is becoming more severe and the superiority of EPS method compared with Euler method is even more obvious.

The evaluation of functions is main calculation in both methods. Despite EPS method needs some extra expenses, this part is relatively very small. For $N = 100$, as we listed above, the ratio of NFE for both methods is approximately 0.015. Even if the extra expense is added, as a conservative estimate, the work amount of EPS method does not reach 2% of Euler's.

We have no intention for $N = 10$, 30 to compare the NFE with NEQ1 and NEQ2 in [6]. It is because that the main expense in those two subroutines is solving linear equations, the expense for evaluation of functions only takes small part of the total.

4.4. Example 4

Two-Point Boundary Value Problem [7] P.80 For two-point boundary value problem

$$u^n(t) = \frac{1}{2}\left(u(t) + t + 1\right)^3, 0 < t < 1, u(0) = u(1) = 0$$

we apply the standard $O(h^2)$ discretization then we can get the following nonlinear equations:

$$f_i(X) = -x_{i-1} + 2x_i - x_{i+1} + \frac{1}{2}h^2\left(x_i + t_i + 1\right)^3$$

$1 \le i \le n$, taking $n = 10$, $x_0 = x_{n+1} = 0$, $t_i = ih$, $h = \frac{1}{n+1}$

It is well known that the initial values play an important role in the procedure of solving a nonlinear equations. As in [6,7] did, set standard starting vector x_s, which regarded as being close to the solution, then using x_s, $10x_s$, $100x_s$ as initial values to test the algorithm. Usually for most algorithm when $x_0 = x_s$ the test got success, when $x_0 = 100x_s$ the test failed.

Four algorithms with three initials x_s, $10x_s$, $100x_s$ were tested in [7], here $x_s = (\xi_1, \xi_2, \ldots, \xi_n)^T$ and $\xi_j = t_j(t_j - 1)$, $1 \le j \le n$.

Relatively speaking, this is a simple problem, every algorithm with any initial value had no trouble to get the solution. Same thing happened for EPS method, for each case mentioned above we get the solution without any trouble. For the sake of comparison, we take $\max_{1 \le i \le n} |f_i| < 10^{-15}$ as convergence criteria. We integrate differential equation:

$$\dot{X} = -D(X)^{-1} F(X)$$

and take $D = diag$ (2, 2, …, 2), $\varepsilon = 0.5$, $h = 2.0$ For initial value $X_0 = X_s$, $X_0 = 10X_s$, $X_0 = 100X_s$, the NEFs are 197, 237, 259 respectively. As compared with Newton method the corresponding figures are 34, 45, 100 [7]. Our goal is not to compare those two sets of figure, as we said before, the main work for Newton method is to solve the linear equations. What we want to do is to compare the following figures: 237/197 ≈ 1.2, 259/197 ≈ 1.3, 45/34 ≈ 1.3, 100/34 ≈ 2.9 Those figures mean that when initial value varied from $X_0 = X_s$ to $X_0 = 100X_s$ EPS method only increases work amount 30%, but for Newton method the work amount will increase 190%.

Another thing is worth mentioning here. For $X_0 = 100X_s$, using tridiagonal solver, after 10 times iteration Newton method got the result max $|f_i| < .2776 \times 10^{-16}$, ($1 \le i \le 10$). However, if at the starting stage using EPS method with $\varepsilon = 0.5$, $h = 1.6$, to make max $|f_i| < 1.0$, only 11 times function evaluation is needed. At the moment, taking current *X's* values as initial value and using Newton method merely 4 times iteration the almost same result was obtained. This fact shows that if the initial value is regarded as being far away from the solution, then EPS method can be chosen as a tool to improve it.

By the way, for Euler method the best step size is 0.9, the numbers of function evaluation for $X_0 = X_s X_0 = 10X_s$, $X_0 = 100X_s$ are 609, 685, 705, respectively.

4.5. Example 5

Broyden tridiagonal function [6] p.28

$$f_i(X) = -x_{i-1} + (3 - 2x_i)x_i - 2x_{i+1} + 1$$

where $x_0 = x_{n+1} = 0$ *and* $i = 1, 2, \ldots, N$; $N = 1000$ the diagonal elements of Jacobian are $3 - 4x_i$ $X_s = (-1, -1, \ldots, -1)^T$ In [6] $N = 10$, the numbers of function evaluation for NEQ1 and NEQ2 are 23 and 25.

We use Euler method and EPS method to integrate differential equation

$$\dot{X} = -D(X)^{-1} F(X)$$

The initial values are $X_0 = X_s X_0 = 10X_s$, $X_0 = 100X_s$. The results are almost the same for both methods. It can be found in **Table 11**.

From the results above, when the initial value $x_i(0)$ taking "negative" values, despite $X_0 = 100X_s$ is regarded as being far away form the solution, however, every method carried out smoothly. But when $x_i(0)$ taking "positive" values the situation would be totally different.

We tested Newton method, taking $\|F\| < 10^{-10}$ as convergence criteria. The results are as follows: for $x_i(0) =$

Table 11. The results of Euler and EPS methods.

Method	Tolrance	Step Size	NFE	$\|F\|$
Euler	$X_0 = X_s$	1.0	41	$.7468 \times 10^{-10}$
EPS	$X_0 = X_s$	1.0 ($\varepsilon = h$)	41	$.7468 \times 10^{-10}$
Euler	$X_0 = 10X_s$	0.5	108	$.9335 \times 10^{-10}$
EPS	$X_0 = 10X_s$	0.5 ($\varepsilon = h$)	108	$.9334 \times 10^{-10}$
Euler	$X_0 = 100X_s$	0.5	117	$.7573 \times 10^{-10}$
EPS	$X_0 = 100X_s$	0.5 ($\varepsilon = h$)	117	$.7572 \times 10^{-10}$

0.0, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01, 0.015, 0.016, 0.017, 0.018, the numbers of iteration are 16, 19, 24, 29, 35, 41, 47, 53, 59, 65, 72, 105, 112, 120, 127. When x_i (0) = 0.019 overflow happened.

For EPS method, taking the same convergence criteria, choosing $h = \varepsilon = 1.0$ and x_i (0) = 0.0, 0.5, 0.7, the numbers of function evaluation are merely 42, 43, 45. Overflow happened if x_i (0) = 0.8. This result may be expected because at the very beginning if x_i (0) = 0.75 the diagonal elements of the Jacobian are all equal to zero.

5. References

[1] P. Deuflhard, "Newton Methods for Nonlinear Problems Affine Invariance and Adaptive Algorithms," Springer-Verlag, Berlin, Heidelberg, 2004.

[2] P. T. Boggs, "The Solution of Nonlinear Systems of Equations by A-Stable Integration Technique," *SIAM Journal on Numerical Analysis*, Vol. 8, No. 4, 1971, pp. 767-785.

[3] J. M. Ortega and W. C. Rheinboldt, "Iterative Solution of Nonlinear Equations in Several Variables," Academic Press, New York, 1970.

[4] T. M. Han and Y. H. Han, "Solving Implicit Equations Arising from Adams-Moulton Methods," *BIT*, Vol. 42, No. 2, 2002, pp. 336-350.

[5] K. M. Brown, "Computer Oriented Algorithms for Solving Systems of Simultaneous Nonlinear Algebraic Equations," In: G. D. Byrne and C. A. Hall, Eds., *Numerical Solution of Systems of Nonlinear Algebraic Equations*, Academic Press, New York, 1973, pp. 281-348.

[6] J. J. Moré, B. S. Garbow and K. E. Hillstrom, "Testing Unconstrained Optimization Software," *ACM Transactions on Mathematical Software*, Vol. 7, No. 1, 1981, pp. 17-41.

[7] J. J. Moré and M. Y. Cosnard, "Numerical Solution of Nonlinear Equations," *ACM Transactions on Mathematical Software*, Vol. 5, No. 1, 1979, pp. 64-85.

Applied Mathematics, **2010, 1, 230-233**
doi:10.4236/am.2010.13028 Published Online September 2010 (http://www.SciRP.org/journal/am)

Ribbon Element on Co-Frobenius Quasitriangular Hopf Algebras

Guohua Liu
Department of Mathematics, Southeast University, Nanjing, China
E-mail: *liuguohua2000cn@yahoo.com.cn*
Received June 5, 2010; *revised July* 23, 2010; *accepted July* 29, 2010

Abstract

Let (*H*, *R*) be a co-Frobenius quasitriangular Hopf algebra with antipode *S*. Denote the set of group-like elements in *H* by *G* (*H*). In this paper, we find a necessary and sufficient condition for (*H*, *R*) to have a ribbon element. The condition gives a connection with the order of *G* (*H*) and the order of S^2.

Keywords: Co-Frobenius Hopf Algebra, Ribbon Element

1. Introduction

A Hopf algebra *H* is called co-Frobenius if *H* is either left or right co-Frobenius as a coalgebra, *i.e.*, if there exists a left or right H^* monomorphism from *H* to H^*. It turns out that *H* is co-Frobenius if and only if *H* has nonzero integrals [1,2]; in particular every finite dimensional Hopf algebra is co-Frobenius. Among the properties of finite dimensional Hopf algebras that hold for all co-Frobenius Hopf algebras are the bijectivity of the antipode, a bijective correspondence between the group-like elements of the Hopf algebra and the one dimensional ideals of the dual algebra, the existence of a distinguished group-like element, and a reasonable theory of Galois extensions.

The class of infinite dimensional co-Frobenius Hopf algebras includes cosemisimple Hopf algebras, such as the group algebra of an infinite group. Tensoring such a Hopf algebra *H* with a finite dimensional Hopf algebra *K*, yields an infinite dimensional Hopf algebra with non-zero integral obtained by tensoring the integrals of *H* and *K*. As well, a recent example of Van Daele [3] gives an infinite dimensional co-Frobenius Hopf algebra without normal Hopf subalgebra.

The topological motivation for this paper is supported by the fact that ribbon Hopf algebras (Hopf algebra with a distinguished ribbon element) can be used to construct invariants of framed links embedded in three dimensional space [4]. And the same structure can be used to produce invariant of three dimensional manifolds. These three dimensional manifolds are represented by surgery on framed links, and their invariants are special cases of invariants for the links. In the case of quantum group SLq(2), these invariants have been intensively investigated by Reshetukhin and Turaev [5], Kirby [6], and others.

In this paper, we give a necessary and sufficient condition for the co-Frobenius quasitriangular Hopf algebra to have a ribbon element. Based on the ideals and results of Beattie, Bulacu and others [7-9], we generalize the results of Kauffman and Radford [10] to co-Frobenius quasitriangular Hopf algebras. We find the group-like elements α and g which play a special role in the theory of ribbon Hopf algebras. Our main result is Theorem 5, which states that a co-Frobenius quasitriangular Hopf algebra (H,R) ($G(H)$ has odd order) has a ribbon element if and only if, S^2 has odd order.

Throughout this paper, *H* will denote a co-Frobenius Hopf algebra over a field *k*. All maps are assumed to be *k*-linear. We use the Sweedler-type notation for the comultiplication maps $\Delta(h) = h_1 \otimes h_2$ for all $h \in H$. As usual, the H^*-bimodule structure on *H* and the *H*-bimodule structure on H^* are given by

$$l^* \rightarrow h \leftarrow m^* = m^*(h_1)h_2 l^*(h_3)$$

$$(h \rightarrow m^* \leftarrow l)(m) = m^*(lmh)$$

for all $h,l,m \in H$ and $l^*, m^* \in H^*$. The antipode of H is denoted by *S* with composition inverse S^{-1}. The set of group-like elements in H is denoted by G(H) and the group-likes of H^0, namely the set of algebra maps from *H* to *k*, by $G(H^0)$.

Let *H* be a co-Frobenius Hopf algebra over a field *k*. Recall that a Hopf algebra *H* is co-Frobenius if H^{*rat},

the unique maximal rational submodule of H^*, is nonzero, or, equivalently, if the space of left or right integrals for H, denoted by $\int_l^{H^*}$ and $\int_r^{H^*}$ respectively, is nonzero. It was shown in [9] that H contains a distinguished group-like element g, which is also called the modular element of H, such that for all $\lambda \in \int_l^{H^*}$ and $h \in H$:

$$\lambda(h_1)h_2 = \lambda(h)g^{-1} \text{ and } \lambda \cdot S^2 = g^{-1} \to \lambda \leftarrow g$$

For Γ either a nonzero left or right integral for H in H^{*rat}, there are bijective maps from H to H^{*rat} given by

$$h \to (h \to \Gamma) \text{ and } h \to (\Gamma \leftarrow h).$$

Let χ denote the generalized Frobenius automorphism of H defined in [7], that is, for $\lambda \in \int_l^{H^*}$, χ is the algebra automorphism of H defined by

$$h \to \lambda = \lambda \leftarrow \chi(h), \text{ for all } h \in H.$$

Then the algebra map $\alpha = \varepsilon \cdot \chi \in H^*$ is called the modular element for H in H^*, and

$$\chi(h) = \alpha(h_2)S^{-2}(h_1), \text{ for all } h \in H.$$

Recall that, for a Hopf algebra H and $R = R^1 \otimes R^2 = r^1 \otimes r^2 \in H \otimes H$, then ($H$, R) is called quasitriangular if for all $h \in H$,

$$(QT1)\Delta(R^1) \otimes R^2 = R^1 \otimes r^1 \otimes R^2 r^2;$$

$$(QT2)\, R^1 \otimes \Delta(R^2) = R^1 r^1 \otimes r^2 \otimes R^2;$$

$$(QT3)\left(\Delta^{cop}(h)\right)R = R\left(\Delta(h)\right), \text{ for all } h \in H;$$

$$(QT4)\,\varepsilon(R^2)R^1 = 1, \varepsilon(R^1)R^2 = 1.$$

Set

$$u = S(R^2)R^1, c = uS(u).$$

By the result of Drinfeld [11] or Radford [12], S^2 is an inner automorphism induced by u and $S(u)^{-1}$, *i.e.*

$$S^2(a) = uau^{-1}, \text{ and } S^2(a) = S(u)^{-1}aS(u),$$

for all $a \in H$.

c is called the Casimir element of (H, R), and $S^2(a) = uau^{-1}$ implies that c is in the center of H.

If (H, R) is quasitriangular, Beattie and Bulauc [13] introduced two group homomorphisms from $G(H^0)$ to G (H) given by

$$\eta \to a_\eta := \eta(R^1)R^2, \eta \to b_\eta := \eta\left(S^{-1}(R^2)\right)R^1 = \eta^{-1}(R^2)R^1$$

They showed that $a_\alpha, b_\alpha \in G(H)$ and $a_\alpha b_{\alpha-1} \in G(H) \cap Z(H)$. Now set

$$g_\alpha = b_{\alpha-1} = (b_\alpha)^{-1}, h = b_{\alpha-1}g^{-1}.$$

By [13], we have $uS(u)^{-1} = S(u)^{-1}u = gg_\alpha$.

By [12] and [13], we have (g, α denote the modular elements),

$$c = u^2 h.$$

Since c is central, $S^2(a) = uau^{-1}, c = u^2 h$, implies that

$$S^4(a) = uS^2(a)u^{-1} = u^2 a(u^{-1})^2 = u^2 ac^{-1}(u^{-1})^2 = h^{-1}ah$$

We say that $\upsilon \in H$ is a quasi-ribbon element of (H, R) if the following conditions are satisfied:

$$(R.1)\upsilon^2 = c;$$

$$(R.2)S(\upsilon) = \upsilon;$$

$$(R.3)\varepsilon(\upsilon) = 1;$$

$$(R.4)\Delta(\upsilon) = (R_{21}R_{12})^{-1}(\upsilon \otimes \upsilon),$$

Drinfeld observed that u satisfies the last condition. A quasi-ribbon element in the center of H is called a ribbon element, and in this case (H, R, v) is called a ribbon Hopf algebra [14]. The reader is referred to [14] for a detailed discussion of ribbon Hopf algebras and their relationship to links and three-manifolds.

2. Ribbon Hopf Algebra

Let (H, R) be a finite dimensional quasitriangular Hopf algebra. In [10], the authors found a necessary and sufficient condition for the existence of ribbon elements on (H, R). The purpose of this section is to generalize their result to co-Frobenius quasitriangular Hopf algebras. We find that most of the results in [10] also hold for co-Frobenius quasitriangular Hopf algebras.

Lemma 1. *Suppose that* (H, R) *is a co-Frobenius quasitriangular Hopf algebra over a field k, H contains a distinguished group-like element g and that* $\upsilon \in H$ *is a quasi-ribbon element of H. Let*

$$g_\alpha = b_{\alpha-1} = (b_\alpha)^{-1}, h = b_{\alpha-1}g^{-1}, u = S(R^2)R^1$$

and set $l = u^{-1}\upsilon$. *Then:*

1) $l^2 = h$;

2) $l \in G(H)$.

Proof. 1) By (R.2) ($S(v) = v$), we have $S^2(\upsilon) = \upsilon$. Thus u and v commute by $S^2(\alpha) = uau^{-1}$ for all $a \in H$. By ($R.1$) $\upsilon^2 = c$ and $c = u^2 h$ we have $\upsilon^2 = u^2 h$. Thus $l^2 = h$.

2)

$$\begin{aligned}\Delta(l) &= \Delta\left(u^{-1}\upsilon\right) = \Delta\left(u^{-1}\right)\Delta(\upsilon)\\ &= \Delta(u)^{-1}\Delta(\upsilon)\\ &= \left(\left(R_{21}R_{12}\right)^{-1}u\otimes u\right)^{-1}\left(R_{21}R_{12}\right)^{-1}(\upsilon\otimes\upsilon)\\ &= u^{-1}\upsilon\otimes u^{-1}\upsilon = l\otimes l\end{aligned}$$

Theorem 2. *Suppose that* (*H*, *R*) *is a co-Frobenius quasitriangular Hopf algebra over a field k, and let u and v be as above. Then*:

1) $l\to ul$ *defines a one-one correspondence between* $\{l\in G(H)\,|\,l^2=h\}\leftrightarrow\{$*quasi-ribbon elements of* (*H*, *R*)$\}$;

2) *Suppose that* $l\in G(H)$ *and* $l^2=h$, *Then v = ul is a ribbon element of* (*H*,*R*) *if and only if* $S^2(a)=l^{-1}al$ *for all* $a\in H$.

Proof. 1) Recall that *u* commutes with the group-like elements of *H*. Thus *v* = *ul* = *lu*. Using $(u^2h=c)$ we see that $\upsilon^2=u^2l^2=u^2h=c$, so (R.1) holds for *v*. Now,

$$S(\upsilon)=S(ul)=S(l)S(u)=l^{-1}S(u),$$

by $l^2=h$ we have $l^{-1}=lh^{-1}$ and *hu* = *S*(*u*), which follows from $u^2h=c=uS(u)$, Therefore

$$S(\upsilon)=l^{-1}S(u)=l^{-1}hu=lh^{-1}hu=lu=ul=\upsilon.$$

Thus, (R.2) holds for *v*. Note that (R.3) is immediate since $\varepsilon(u)=1=\varepsilon(l)$. Also

$$\begin{aligned}\Delta(\upsilon)=\Delta(ul)&=\Delta(u)\Delta(l)=\left(R_{21}R_{12}\right)^{-1}(u\otimes u)(l\otimes l)\\ &=\left(R_{21}R_{12}\right)^{-1}\upsilon\otimes\upsilon\end{aligned}$$

and (R.4) holds for *v*. The proof of (1) is finished by Lemma 1.

2) If $\upsilon=ul$, $S^2(a)=uau^{-1}=(\upsilon l^{-1})a(\upsilon l^{-1})^{-1}=l^{-1}al$. On the other hand, $S^2(a)=l^{-1}al=uau^{-1}$ implies *va* = *av* for all $a\in H$.

Corollary 3. *Suppose that* (*H*, *R*) *is a co-Frobenius quasitriangular Hopf algebra. Let g and a be the distinguished group-like elements of H and* H^*, *respectively, and let h be as above. Then:*

1) *If h has odd order, or if g and a have odd order, Then* (*H*, *R*) *has a quasi-ribbon element*;

2) *If G* (*H*) *has odd order, Then* (*H*, *R*) *has a unique quasi-ribbon element.*

Proof. 1) By b_{a-1} commuting with all $a\in G(H)$, and $\eta\to b_\eta$ is a group homomorphism from $G(H^0)$ to *G* (*H*). We have that *h* has a square root in *G*(*H*), which must be unique if *G* (*H*) has odd order. Therefore the corollary follows by part 1 of the above Theorem.

Proposition 4. *Suppose that* (*H*, *R*) *is a co-Frobenius quasitriangular Hopf algebra. Let g and a be the distinguished group-like elements of H and* H^*, *respectively, and let h be as above. Then if either*

1) *If h and* S^2, *or*;

2) *If g, a and* S^2 *have odd order. Then* (*H*, *R*) *has a ribbon element.*

Proof. First, condition (2) implies condition (1). Suppose that *h* and S^2 have odd order. Let *l* be the unique square root of *h* having odd order. Define map $\tau(a):H\to H$ by $\tau(a)(b)=aba^{-1}$.

Then $\tau(l^{-1})^2=\tau(h^{-1})$ and $\tau(l^{-1})$ have odd order. Recall that $S^4(a)=h^{-1}ah=\tau(h^{-1})$. Since $l\in G(H)$, S^2 and $\tau(l^{-1})$ are two elements of odd order whose squares are equal. Consequently, $S^2=\tau(l^{-1})$, and the proposition follows by part (2) of Theorem 2.

Theorem 5. *Suppose that* (*H*, *R*) *is a co-Frobenius quasitriangular Hopf algebra with antipode S over a field k and assume that G* (*H*) *has odd order, Then* (*H*, *R*) *has a ribbon element (which is necessarily unique) if and only if* S^2 *has odd order.*

Proof. If (*H*, *R*) has a ribbon element, then there exists an $x\in G(H)$ such that $S^2(a)=x\,ax^{-1}$ for all $a\in H$ by Theorem 2. Since *x* has odd order it follows that S^2 does also.

Conversely, suppose that S^2 has odd order. Since *h* has odd order, it follows that (*H*, *R*) has a ribbon element by Proposition 4. This completes our proof.

When *H* is unimodular, We note that $h=g^{-1}$ since $a=\varepsilon$ in this case, by Theorem 2, the existence of ribbon (or quasi-ribbon) elements is determined by square roots of *g*.

Proposition 6. *Suppose that* (*H*, *R*) *is a co-Frobenius quasitriangular Hopf algebra with antipode S over a field k, Suppose further that H is unimodular and let g be the distinguished group-like element of H. Then*:

1) (*H*, *R*) *has a quasi-ribbon element if and only if* $l^2=g$ *for some* $l\in G(H)$;

2) (*H*, *R*) *has a ribbon element if and only if* $l^2=g$ *for some* $l\in G(H)$ *which satisfies* $S^2(a)=lal^{-1}$ *for all* $a\in H$.

Proposition 7. *Suppose that* (*H*, *R*) *is a co-Frobenius quasitriangular Hopf algebra with antipode S over a field k, suppose further that H and* H^* *are both unimodular Then*:

1) *u is a quasi-ribbon element of* (*H*, *R*);

2) *u is a ribbon element of* (*H*, *R*) *if and only if* $S^2=I$.

3. Acknowledgements

The author is supported by the NSFC project 10826037.

4. References

[1] B. I. P. Lin, "Semiperfect Coalgebras," *Journal of Algebra*, Vol. 49, No. 2, 1977, pp. 357-373.

[2] Y. Doi, "Homological Coalgebra," *Journal of the Mathematical Society of Japan*, Vol. 33, No. 1, 1981, pp. 31-50.

[3] A. van Daele, "An Algebraic Framework for Group Duality," *Advances in Mathematics*, Vol. 140, No. 2, 1998, pp. 323-366.

[4] N. Y. Reshetikhin and V. G. Turaev, "Ribbon Graphs and their Invariants Derived from Quantum Groups," *Communications in Mathematical Physics*, Vol. 127, No. 1, 1990, pp. 1-26.

[5] N. Y. Reshetikhin and V. G. Turaev, "Invariants of 3-Manifolds via Link Polynomials and Quantum Groups," *Invented Mathematics*, Vol. 103, No. 3, 1991, pp. 547-597.

[6] R. Kirby and P. Melvin, "The 3-Manifolds Invariants of Witten and Reshetikhin-Turaev for sl (2, C)," *Invented Mathematics*, Vol. 105, No. 3, 1991, pp. 473-545.

[7] M. Beattie, D. Bulacu and B. Torrecillas, "Radford's s4 Formula for Co-Frobenius Hopf Algebras," *Journal of Algebra*, Vol. 307, No. 1, 2007, pp. 330-342.

[8] M. Beattie, S. Dascalescu, L. Grunenfelder and C. Nastasescu, "Finiteness Conditions, Co-Frobenius Hopf Algebras and Quantum Groups," *Journal of Algebra*, Vol. 200, No. 1, 1998, pp. 312- 333.

[9] M. Beattie, S. Dascalescu and S. Raianu, "Galois Extensions for Co-Frobenius Hopf Algebras," *Journal of Algebra*, Vol. 198, No. 1, 1997, pp. 164-183.

[10] L. H. Kauffman and D. E. Radford, "A Necessary and Sufficient Condition for a Finite–Dim–Ensional Drinfeld Double to be a Ribbon Hopf Algebras," *Journal of Algebra*, Vol. 159, No. 1, 1993, pp. 98-114.

[11] V. G. Drinfeld, "On Almost Cocommutative Hopf Algebras," *Leningrad Mathematical Journal*, Vol. 1, No. 2, 1990, pp. 321-342.

[12] D. E. Radford, "On the Antipode of a Quasitriangular Hopf Algebras," *Journal of Algebra*, Vol. 151, No. 1, 1992, pp. 1-11.

[13] M. Beattie and D. Bulacu, "On the Antipode of a Co-Frobenius (Co) Quasitriangular Hopf Algebras," *Communications in Algebra*, Vol. 37, No. 9, 2009, pp. 2981-2993.

[14] N. Y. Reshetikhin and V. G. Turaev, "Invariants of 3-Manifolds via Link Polymonials and Quantum Groups," *Communications in Mathematical Physics*, Vol. 127, No. 1, 1990, pp. 7-26.

Applied Mathematics, **2010, 1, 234-243**
doi:10.4236/am.2010.13029 Published Online September 2010 (http://www.SciRP.org/journal/am)

Semi-Markovian Model of Monotonous System Maintenance with Regard to its Elements' Deactivation and Age

Yuriy E. Obzherin, Aleksey I. Peschansky
Sevastopol National Technical University, Sevastopol, Ukraine
E-mail: *vmsevntu@mail.ru*
Received June 18, 2010; *revised July* 26, 2010; *accepted August* 1, 2010

Abstract

An explicit form of reliability and economical stationary performance indexes for monotonous multicomponent system with regard to its elements' maintenance has been found. The maintenance strategy investigated supposes preventive maintenance execution for elements that has attained certain operating time to failure. Herewith for the time period of elements' maintenance or restoration operable elements, functionally connected with the failed ones, are deactivated. The problems of maintenance execution frequency optimization have been solved. For the model building the theory of semi-Markovian processes with a common phase field of states is used.

Keywords: Maintenance, Semi-Markovian Process, System Stationary Characteristics, System Performance Indexes Optimization

1. Introduction

One of the methods of the complex technical systems' reliability improvement is their maintenance. The review of the results concerning this subject can be found in the works [1-3]. One of the strategies of a single-component system maintenance is the strategy known in literature as "Depending-on-age restoration" [4-6]. This strategy being used, the system is considered to be completely restored after its failure. If the system has been operating without failures for the given time period τ, then its maintenance, after which it is completely restored, is executed. In [7] semi-Markovian model of the above-mentioned strategy for multicomponent monotonous system maintenance under assumption that any system's element failure does not result in deactivation of elements that are in up state, are functionally connected with the failed ones, and do not belong to any up-state path has been built.

The goal of the present article is to build semi-Markovian model of maintenance in age of a multicomponent system's elements with regard to their deactivation. On the basis of the model built it is necessary to define stationary reliability and economical performance indexes of the system and to solve the problem of elements' maintenance optimal terms determination.

2. The Problem Definition and Mathematical Model Building

Let us consider N-component system with a monotonous structure and describe the strategy of its elements' maintenance. At the time zero $t = 0$ system operation begins and an acceptable operating time to failure level (age) τ_i for each i-element of system is determined. On attaining this level element's planned maintenance is carried out. The failure-free operation time of system's i-element is a random value (RV) α_i with distribution function (DF) $F_i(t)$. Unless system's i-element fails by the moment τ_i, element's planned maintenance that restores it completely begins. The maintenance lasts random period of time β_i^p with DF $G_i^p(t)$.

If system's i-element has failed by the moment τ_i, its failure is discovered instantly and its emergency restoration (ER) begins. This restoration lasts RV β_i with DF $G_i(t)$. As a result of ER, an element is restored completely and the whole maintenance process occurs again.

Let us assume that due to emergency failure or to the beginning of some element's maintenance the operable

elements that do not belong to any other up-state path are deactivated. Besides, the elements in state of ER or maintenance, the restoration of which would not result in any up-state path formation, are deactivated.

The elements deactivated have the same operable level at the moment of their activation. The latter happens at the end of element's ER or maintenance under the condition of simultaneous up-state path formation.

Time diagram of system operation is shown in **Figure 1**.

Let us begin semi-Markovian (SM) model building of the system. To begin with the phase field of states should be defined. Each element of system can be in three physical states:

1 – in up state or deactivated in up state;

0 – in state of restoration or deactivated in state of restoration;

2 – in state of maintenance or deactivated in state of maintenance.

System's physical states will be indicated with a set of vectors $D = \left\{ \overline{d} = (d_1, ..., d_N),\ d_k = 0,1,2;\ k = \overline{1,N} \right\}$. The component d_k of vector $\overline{d}$ denotes the physical state of system's k-element.

The physical states to exhibit SM property, they should be extended. With this purpose we will indicate the number of element that was last to change its state. Let us add continuous components, denoting time periods of elements' dwelling in their states. In the code of extended state these time periods will be indicated by vector $\overline{x}^{(i)} = (x_1, ..., x_{i-1}, 0, x_{i+1}, ..., x_N)$.

Besides, in accordance with the chosen maintenance strategy we will introduce vector $\overline{u} = (u_1, ..., u_N)$, the components of which indicate elements' operating time since the last restoration of their up state, to the code of system's states.

Thus, the system's phase field of SM states with regard to its elements' maintenance execution is the following:

$$E^* = \left\{ i\overline{d}\,\overline{x}^{(i)}\,\overline{u},\ \ i = \overline{1,N} \right\}$$

The significance of the code of states:

i is the number of element that was last to change its physical state;

$d_k = 0,1,2$ is the code of system's k-element physical state;

x_k is time period between i-element's last state change and the nearest moment of k-element's change $(x_i = 0)$ regardless of deactivation time; and if $d_k = 1$ then x_k is the time period till the nearest emergency failure of k-element;

u_k is operating time to failure of k-element since the end of its last ER or maintenance. If $d_k = 2$ it is considered that $u_k = \tau_k$. At the moment of i-element's transition to up state after its maintenance or ER its operating time is equal zero: $u_i = 0$.

Let us indicate I_d a set of numbers of elements deactivated in the state $i\overline{d}\,\overline{x}^{(i)}\,\overline{u},\ \ i = \overline{1,N}$. System dwelling time periods are defined by ratios:

$$\theta_{i\overline{d}\,\overline{x}^{(i)}\,\overline{u}} = \gamma_i^{(d_i)} \wedge \underset{\substack{k \neq i \\ k \notin I_d}}{\Lambda} x_k \underset{\substack{k \in \Omega_d^1 \\ k \notin I_d}}{\Lambda} \left(\tau_k - u_k \right)$$

where Λ is a sign of minimum; Ω_d^1 is a set of numbers of vector $\overline{d}$ components that are equal to 1,

$$\gamma_i^{(d_i)} = \begin{cases} \alpha_i, & d_i = 1, \\ \beta_i, & d_i = 0, \\ \beta_i^p, & d_i = 2. \end{cases}$$

Let us describe the probabilities (probability densities) of embedded Markovian chain (EMC) $\{\xi_n,\ n \geq 0\}$ transition. It is necessary to note that *i*-element can change its physical state 1 into the state 0 (ER) and into the

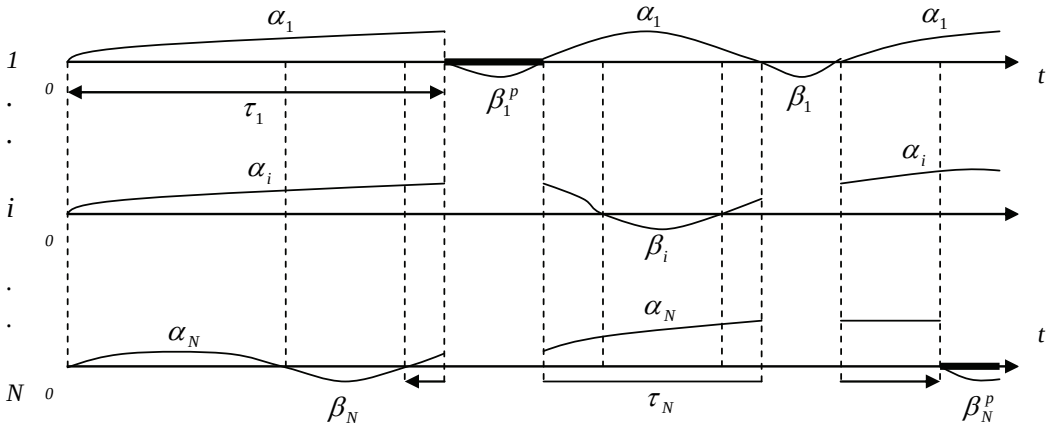


Figure 1. Time diagram of the system operation with elements' deactivation after the first element failure and with regard to their maintenance in age.

state 2 (maintenance) but the states 0 and 2 can be changed only into the state 1.

Let us indicate

$$z_{i,I_d} = \underset{\substack{k\neq i\\ k\notin I_d}}{\Lambda} x_k \wedge \underset{\substack{k\in\Omega_d^1\\ k\notin I_d}}{\Lambda} (\tau_k - u_k) \tag{1}$$

and let Ω_d^0, Ω_d^2 be sets of numbers of vector $\overline{d}$ components that are equal to 0 and 2 respectively.

The state $i\overline{d}\,\overline{x}^{(i)}\overline{u}$, $i=\overline{1,N}$ admits the following transitions:

1) to the set of states $i\overline{d'}\,\overline{x}^{(i)}\overline{u'}$, $d_i'\neq 2$ with the probability density of transition

$p^{i\overline{d'}\,\overline{x}^{(i)}\overline{u'}}_{i\overline{d}\,\overline{x}^{(i)}\overline{u}} = \psi_i^{(d_i)}(z_{i,I_d} - y)$, where $y < z_{i,I_d}$, $\psi_i^{(d_i)}(\cdot)$, is the density of probability distribution of RV $\gamma_i^{(d_i)}$, $d_k' = d_k$, $k\neq i$; $x_k' = x_k - (z_{i,I_d} - y)$, $k\neq i$, $k\notin I_d$; $x_k' = x_k$, $k\in I_d$;

$$u_k' = \begin{cases} u_k + z_{i,I_d} - y, & k\in\Omega_d^1,\ k\notin I_d,\\ u_k, & k\in\Omega_d^0,\ k\in I_d,\quad k\neq i,\\ \tau_k, & k\in\Omega_d^2, \end{cases}$$

$$u_i' = \begin{cases} u_i + z_{i,I_d} - y, & i\in\Omega_d^1,\\ 0, & i\in\Omega_d^0\cup\Omega_d^2; \end{cases}$$

2) to the set of states $i\overline{d'}\,\overline{x'}^{(i)}\overline{u'}$, $d_i = 1$, $d_i' = 2$ with transition probability $P^{i\overline{d'}\,\overline{x}^{(i)}\overline{u'}}_{i\overline{d}\,\overline{x}^{(i)}\overline{u}} = \overline{F}_i(\tau_i)$, where $d_k' = d_k$, $k\neq i$; $x_k' = x_k - \tau_i$, $k\neq i$, $k\notin I_d$; $x_k' = x_k$, $k\in I_d$;

$$u_k' = \begin{cases} u_k + \tau_i, & k\in\Omega_d^1,\ k\notin I_d,\\ u_k, & k\in\Omega_d^0,\ k\in I_d,\\ \tau_k, & k\in\Omega_d^2; \end{cases}$$

3) to the set of states $j\overline{d'}\,\overline{x'}^{(j)}\overline{u'}$, $j\neq i$, $j\notin I_d$ with the probability density of transition

$p^{j\overline{d'}\,\overline{x}^{(j)}\overline{u'}}_{i\overline{d}\,\overline{x}^{(i)}\overline{u}} = \psi_i^{(d_i)}(z_{i,I_d} + y)$, where $y>0$, $d_k' = d_k$, $k\neq j$, $x_i' = y$, $x_k' = x_k - z_{i,I_d}$, $k\neq i,j$,

$$u_j' = \begin{cases} u_j + z_{i,I_d}, & j\in\Omega_d^1,\ d_j' = 0,\\ \tau_j, & j\in\Omega_d^1,\ d_j' = 2,\\ 0, & j\in\Omega_d^0\cup\Omega_d^2, \end{cases}$$

$$u_k' = \begin{cases} u_k + z_{i,I_d}, & k\in\Omega_d^1,\ k\notin I_d,\\ u_k, & k\in\Omega_d^0,\ k\in I_d,\quad k\neq j.\\ \tau_k, & k\in\Omega_d^2, \end{cases}$$

Let us assume that the conditions of stationary distribution $\rho(\cdot)$ [8,9] existence and uniqueness for EMC $\{\xi_n, n\geq 0\}$ are fulfilled. The following theorem takes place.

Theorem. The stationary distribution of EMC $\{\xi_n, n\geq 0\}$ is defined by the following expressions:

$$\rho\left(i\overline{d}\,\overline{x}^{(i)}\overline{u}\right) = \begin{cases} \rho \prod\limits_{k\in\Omega_d^0} f_k(u_k)\overline{G}_k(x_k) \prod\limits_{k\in\Omega_d^1} f_k(u_k + x_k) \prod\limits_{k\in\Omega_d^2} \overline{F}_k(\tau_k)\overline{G}_k^p(x_k), & i\notin\Omega_d^1,\ x_i = 0,\\ \rho \prod\limits_{k\in\Omega_d^0} f_k(u_k)\overline{G}_k(x_k) \prod\limits_{\substack{k\in\Omega_d^1\\ k\neq i}} f_k(u_k + x_k) \prod\limits_{k\in\Omega_d^2} \overline{F}_k(\tau_k)\overline{G}_k^p(x_k), & i\in\Omega_d^1, \end{cases} \tag{2}$$

$$\rho = \left[\sum_{d\in D^*}\left[\sum_{\substack{i\in\Omega_d^1\\ i\notin I_d}}\prod_{\substack{k=1\\ k\neq i}}^{N} T_k^{(d_k)}(\tau_k) + \sum_{\substack{i\in\Omega_d^0\\ i\notin I_d}} F_i(\tau_i)\prod_{\substack{k=1\\ k\neq i}}^{N} T_k^{(d_k)}(\tau_k) + \sum_{\substack{i\in\Omega_d^2\\ i\notin I_d}} \overline{F}_i(\tau_i)\prod_{\substack{k=1\\ k\neq i}}^{N} T_k^{(d_k)}(\tau_k)\right]\right]^{-1}$$

$$T_k^{(1)}(\tau_k) = \int_0^{\tau_k} \overline{F}_k(t)\,dt,\quad T_k^{(0)}(\tau_k) = F_k(\tau_k)M\beta_k,\quad T_k^{(0)}(\tau_k) = \overline{F}_k(\tau_k)M\beta_k^p.$$

Theorem proving. The stationary distribution of probabilities $\rho(B)$ obeys the system of integral equations [8]

$$\rho(B) = \int_{E^*} \rho(dz)P(z,B).$$

For example, the equation of this system for the state $i\overline{d}\,\overline{x}^{(i)}\overline{u}$, $d_i = 0$, $i=\overline{1,N}$; $i\notin I_d$; is as follows:

$$\begin{aligned}\rho\left(i\overline{d}\,\overline{x}^{(i)}\overline{u}\right) &= f_m\left(x_m + u_{m,I}\right)\rho\left(m\overline{d'}\,\overline{x''}^{(m)}\overline{u''}\right)\\ &+ \sum_{\substack{j\in\Omega_{d'}^0\cup\Omega_{d'}^2\\ j\notin I_{d'}}} \int_0^{u_{m,I}} \psi_j^{(d_j')}(t + x_j)\,\rho\left(j\overline{d'}(\overline{x'}^{(j)}\overline{u'}\right)dt,\end{aligned} \tag{3}$$

$x_i = 0$, $d_i = 0$, $i=\overline{1,N}$; $i\notin I_d$;

$d_i' = 1,\ d_k' = d_k,\ k \neq i,\ u_{m,I} = \underset{\substack{k \in \Omega_{d'}^1 \\ k \neq I_{d'}}}{\Lambda} u_k$

By the direct substitution one can check that Formula (2) define the solution of this equation. For the state $i\overline{d}\overline{x}^{(i)}\overline{u}$ we deal with $d_i = 0$, $d_i' = 1$, $\Omega_{d'}^1 - \{i\} = \Omega_d^1$, $\Omega_{d'}^0 \cup \{i\} = \Omega_d^0$, $\Omega_{d'}^2 = \Omega_d^2$. Substituting (2) to the second member of Equation (3) we get the following results:

$$f_m(u_{m,I} + x_m) \prod_{k \in \Omega_{d'}^0} f_k(u_k) \overline{G}_k(x_k + u_{m,I}) \prod_{\substack{k \in \Omega_{d'}^1 \\ k \neq m}} f_k(u_k + x_k) \prod_{k \in \Omega_{d'}^2} \overline{F}_k(\tau_k) \overline{G}_k^p(x_k + u_{m,I}) +$$

$$\left[\sum_{\substack{j \in \Omega_{d'}^0 \\ j \neq I_{d'}}} \int_0^{u_{m,I}} g_j(x_j + t) f_j(u_j) \prod_{\substack{k \in \Omega_{d'}^0 \\ k \neq I_{d'}, k \neq j}} f_k(u_k) \overline{G}_k(x_k + t) \prod_{k \in \Omega_{d'}^1} f_k(u_k + x_k) \prod_{\substack{k \in \Omega_{d'}^2 \\ k \neq I_{d'}}} \overline{F}_k(\tau_k) \overline{G}_k^p(x_k + t) dt + \right.$$

$$\left. \sum_{\substack{j \in \Omega_{d'}^2 \\ j \neq I_{d'}}} \int_0^{u_{m,I}} g_j^p(x_j + t) \prod_{\substack{k \in \Omega_{d'}^0 \\ k \neq I_{d'}}} f_k(u_k) \overline{G}_k(x_k + t) \prod_{k \in \Omega_{d'}^1} f_k(u_k + x_k) \overline{F}_j(\tau_j) \prod_{\substack{k \in \Omega_{d'}^2 \\ k \neq I_{d'}, k \neq j}} \overline{F}_k(\tau_k) \overline{G}_k^p(x_k + t) dt \right] \times$$

$$\prod_{\substack{k \in \Omega_{d'}^0 \\ k \in I_{d'}}} f_k(u_k) \overline{G}_k(x_k) \prod_{\substack{k \in \Omega_{d'}^2 \\ k \in I_{d'}}} \overline{F}_k(\tau_k) \overline{G}_k^p(x_k) =$$

$$\prod_{k \in \Omega_{d'}^0} f_k(u_k) \overline{G}_k(x_k + u_{m,I}) \prod_{k \in \Omega_{d'}^1} f_k(u_k + x_k) \prod_{k \in \Omega_{d'}^2} \overline{F}_k(\tau_k) \overline{G}_k^p(x_k + u_{m,I}) -$$

$$\prod_{k \in \Omega_{d'}^1} f_k(u_k + x_k) \prod_{\substack{k \in \Omega_{d'}^0 \\ k \in I_{d'}}} f_k(u_k) \overline{G}_k(x_k) \prod_{\substack{k \in \Omega_{d'}^2 \\ k \in I_{d'}}} \overline{F}_k(\tau_k) \overline{G}_k^p(x_k) \times$$

$$\int_0^{u_{m,I}} \frac{\partial}{\partial t} \left[\prod_{\substack{k \in \Omega_{d'}^0 \\ k \neq I_{d'}}} f_k(u_k) \overline{G}_k(x_k + t) \prod_{\substack{k \in \Omega_{d'}^2 \\ k \neq I_{d'}}} \overline{F}_k(\tau_k) \overline{G}_k^p(x_k + t) \right] dt =$$

$$\prod_{k \in \Omega_{d'}^1} f_k(u_k + x_k) \prod_{k \in \Omega_{d'}^0} f_k(u_k) \overline{G}_k(x_k) \prod_{k \in \Omega_{d'}^2} \overline{F}_k(\tau_k) \overline{G}_k^p(x_k) =$$

$$f_i(u_i) \prod_{k \in \Omega_d^1} f_k(u_k + x_k) \prod_{\substack{k \in \Omega_d^0 \\ k \neq i}} f_k(u_k) \overline{G}_k(x_k) \prod_{k \in \Omega_d^2} \overline{F}_k(\tau_k) \overline{G}_k^p(x_k) = \frac{1}{\rho} \rho\left(i\overline{d}\overline{x}^{(i)}\overline{u}\right).$$

In the same way it can be checked that Formula (2) define the stationary distributions for the rest of system's states. The constant ρ is determined due to normalization condition.

3. Definition of System Stationary Characteristics

Let us define the following system stationary performance indexes: mean stationary operating time to failure $T_+^*(\tau_1, ..., \tau_N)$; mean stationary restoration time $T_-^*(\tau_1, ..., \tau_N)$; stationary steady state availability factor (SSAF) $K_u^*(\tau_1, ..., \tau_N)$; mean specific income $S^*(\tau_1, ..., \tau_N)$ per calendar time unit, and mean specific expenses $C^*(\tau_1, ..., \tau_N)$ per time unit of system's good state.

Let us divide the phase field E^* of system's states into two non-overlapping subsets E_+^* and E_-^*; E_+^* is a subset of up states, E_-^* is a subset of down states:

$$E_+^* = \left\{ i\overline{d}\overline{x}^{(i)}\overline{u},\ \overline{d} \in D_+^*,\ \ i = \overline{1, N} \right\}$$

$$E_-^* = \left\{ i\overline{d}\overline{x}^{(i)}\overline{u},\ \overline{d} \in D_-^*,\ \ i = \overline{1, N} \right\}$$

Here $D_+^*\left(D_-^*\right)$ is a set of vectors $\overline{d}$ the components of which are equal to the codes of physical states of system's elements; this system is in a subset of up (down) states $E_+^*\left(E_-^*\right)$.

Mean stationary operating time to failure T_+^*, mean stationary restoration time T_-^*, and stationary SSAF K_u^* of the system will be estimated with the help of formulas [8,9]

$$T_+^* = \frac{\int\limits_{E_+^*} m(z)\rho(dz)}{\int\limits_{E_-^*} \rho(dz)P\left(z,E_+^*\right)},\quad T_-^* = \frac{\int\limits_{E_-^*} m(z)\rho(dz)}{\int\limits_{E_-^*} \rho(dz)P\left(z,E_+^*\right)},\qquad (4)$$

$$K_u^* = \frac{T_+^*}{T_+^* + T_-^*}$$

where $\rho(\cdot)$ is the stationary distribution of EMC $\{\xi_n, n \ge 0\}$, $m(z)$ are mean time periods of system's dwelling in its states, $P(z,E_+^*)$ are probabilities of EMC $\{\xi_n, n \ge 0\}$ transition from down to up states.

To define the stationary indexes with the help of Formula (4) it is necessary to define the basic characteristics included in these formulas.

Let us begin with the integral $\int\limits_{E_+^*} m(z)\rho(dz)$. Mean time period of system's dwelling in the state $i\overline{d}\,\overline{x}^{(i)}\overline{u}$ is found by the formula $M\left[\theta_{i\overline{d}\,\overline{x}^{(i)}\overline{u}}\right] = \int\limits_0^{z_{i,I_d}} \overline{\Psi}_i^{(d_i)}(t)dt$, where z_{i,I_d} is given by (1). We have

$$\int\limits_{E_+^*} m(z)\rho(dz) = \sum_{d\in D_+^*}\sum_{\substack{i=1\\ i\notin I_d}}^{N}\int\limits_U d\hat{u}\int\limits_{R_+^{N,i}} \rho\left(i\overline{d}\,\overline{x}^{(i)}\overline{u}\right)d\overline{x}^{(i)}\int\limits_0^{z_{i,I_d}}\overline{\Psi}_i^{(d_i)}(t)dt =$$

$$-\sum_{d\in D_+^*}\prod_{\substack{k\in\Omega_d^1\\ k\in I_d}}\int\limits_0^{\tau_k}\overline{F_k}(s)ds\int\limits_0^{\tau^{1,I_d}}\frac{d}{dt}\left[\prod_{\substack{k\in\Omega_d^1\\ k\notin I_d}}\int\limits_t^{\tau_k}\overline{F_k}(s)ds\prod_{k\in\Omega_d^0}F_k(\tau_k)\int\limits_t^{\infty}\overline{G_k}(s)ds\prod_{k\in\Omega_d^2}\overline{F_k}(\tau_k)\int\limits_t^{\infty}\overline{G_k^p}(s)ds\right]dt =$$

$$\sum_{d\in D_+^*}\prod_{k\in\Omega_d^1}\int\limits_0^{\tau_k}\overline{F_k}(s)ds\prod_{k\in\Omega_d^0}M\beta_k F_k(\tau_k)\prod_{k\in\Omega_d^2}M\beta_k^p\overline{F_k}(\tau_k) = \sum_{d\in D_+^*}\prod_{k=1}^{N}T_k^{(d_k)}(\tau_k).$$

Here

$$\tau^{1,I_d} = \bigwedge_{\substack{k\in\Omega_d^1\\ k\notin I_d}}\tau_k,\quad R_+^{N,i} = \left\{\overline{x}^{(i)},\ x_k \ge 0,\ k=\overline{1,N}\right\},\quad U = \left\{\hat{\overline{u}} = \left(u_{i_1},\ldots u_{i_s}\right), 0 \le u_{i_r} \le \tau_k,\ i_r = k,\ k\in\Omega_d^0,\Omega_d^1\right\}.$$

The values $T_k^{(d_k)}(\tau_k)$ have the following significance: $T_k^{(1)}(\tau_k)$ is mean time period of k-element dwelling in up state, and $T_k^{(0)}(\tau_k)+T_k^{(2)}(\tau_k)$ is mean time period of this element dwelling in down state during its regeneration.

Analogically, we have

$$\int\limits_{E_-^*} m(z)\rho(dz) = \sum_{d\in D_-^*}\prod_{k\in\Omega_d^1}\int\limits_0^{\tau_k}\overline{F_k}(s)ds\prod_{k\in\Omega_d^0}M\beta_k F_k(\tau_k)\prod_{k\in\Omega_d^2}M\beta_k^p\overline{F_k}(\tau_k) = \sum_{d\in D_-^*}\prod_{k=1}^{N}T_k^{(d_k)}(\tau_k).$$

Let us calculate the integral in denominators of ratios (4). It is necessary to note that the transitions to E_+^* can occur from the subset $E_-^{*'}\subset E_-^*$ only with the probability equal to 1 where $E_-^{*'} = \left\{i\overline{d}\,\overline{x}^{(i)}\overline{u},\ \overline{d}\in D_-^*,\ i\in\Omega_d^0\cup\Omega_d^2,\ i\notin I_d\right\}$. We have

$$\int\limits_{E_-^*}\rho(dz)P(z,E_+^*) = \int\limits_{E_-^{*'}}\rho(dz) = \sum_{d\in D_-^*}\left[\sum_{\substack{i\in\Omega_d^0\\ i\notin I_d}}F_i(\tau_i)\prod_{\substack{k=1\\ k\ne i}}^{N}T_k^{(d_k)}(\tau_k) + \sum_{\substack{i\in\Omega_d^2\\ i\notin I_d}}\overline{F}_i(\tau_i)\prod_{\substack{k=1\\ k\ne i}}^{N}T_k^{(d_k)}(\tau_k)\right]$$

Thus, the Formula (4) are transformed into

$$T_+^*(\tau_1,...,\tau_N)=\frac{\sum\limits_{d\in D_+^*}\prod\limits_{k=1}^{N}T_k^{(d_k)}(\tau_k)}{\sum\limits_{d\in D_-^*}\left[\sum\limits_{\substack{i\in\Omega_d^0\\ i\neq I_d}}F_i(\tau_i)\prod\limits_{\substack{k=1\\ k\neq i}}^{N}T_k^{(d_k)}(\tau_k)+\sum\limits_{\substack{i\in\Omega_d^2\\ i\neq I_d}}\overline{F}_i(\tau_i)\prod\limits_{\substack{k=1\\ k\neq i}}^{N}T_k^{(d_k)}(\tau_k)\right]}, \tag{5}$$

$$T_-^*(\tau_1,...,\tau_N)=\frac{\sum\limits_{d\in D_-^*}\prod\limits_{k=1}^{N}T_k^{(d_k)}(\tau_k)}{\sum\limits_{d\in D_-^*}\left[\sum\limits_{\substack{i\in\Omega_d^0\\ i\neq I_d}}F_i(\tau_i)\prod\limits_{\substack{k=1\\ k\neq i}}^{N}T_k^{(d_k)}(\tau_k)+\sum\limits_{\substack{i\in\Omega_d^2\\ i\neq I_d}}\overline{F}_i(\tau_i)\prod\limits_{\substack{k=1\\ k\neq i}}^{N}T_k^{(d_k)}(\tau_k)\right]}, \tag{6}$$

$$K_u^*(\tau_1,...,\tau_N)=\frac{\sum\limits_{d\in D_+^*}\prod\limits_{k=1}^{N}T_k^{(d_k)}(\tau_k)}{\sum\limits_{d\in D^*}\prod\limits_{k=1}^{N}T_k^{(d_k)}(\tau_k)}. \tag{7}$$

Let us determine system stationary characteristics T_+^*, T_-^*, $K_u^*(\tau_1,...,\tau_N)$ by means of elements' SSAF $K_i(\tau_i)$ defined by the formulas [4,5]:

$$K_i(\tau_i)=\frac{T_i^{(1)}(\tau_i)}{T_i^{(1)}(\tau_i)+T_i^{(0)}(\tau_i)+T_i^{(2)}(\tau_i)},\ i=\overline{1,N}.$$

Let M_i, $i=\overline{1,\omega}$, be all the different sets of elements of system paths, and Φ_i, $i=\overline{1,s}$ be sets of elements of system [4] sections; $A(\Phi_i)\left(A(M_i)\right)$ is a set of deactivated elements of section Φ_i (of M_i path). One should pay attention that according to the definition the elements not belonging to the set of elements of path are in down state, *i.e.*, are in a state 0 or 2. The elements not belonging to the set of elements of section are in up state 1.

The Formulas' (5)–(7) transformation of averages products' sums lead to the following result:

$$T_+^*(\tau_1,...,\tau_N)=\frac{\sum\limits_{i=1}^{\omega}\prod\limits_{n\in M_i}K_n(\tau_n)\prod\limits_{\substack{n=1\\ n\notin M_i}}^{N}\left(1-K_n(\tau_n)\right)}{\sum\limits_{i=1}^{s}\sum\limits_{\substack{j\in\Phi_i\\ j\notin A(\Phi_i)}}\frac{1}{T_j^{(0)}+T_j^{(2)}}\prod\limits_{\substack{n=1\\ n\notin\Phi_i}}^{N}K_n(\tau_n)\prod\limits_{n\in\Phi_i}\left(1-K_n(\tau_n)\right)}, \tag{8}$$

$$T_-^*(\tau_1,...,\tau_N)=\frac{\sum\limits_{i=1}^{s}\prod\limits_{\substack{n=1\\ n\notin\Phi_i}}^{N}K_n(\tau_n)\prod\limits_{n\in\Phi_i}\left(1-K_n(\tau_n)\right)}{\sum\limits_{i=1}^{s}\sum\limits_{\substack{j\in\Phi_i\\ j\notin A(\Phi_i)}}\frac{1}{T_j^{(0)}+T_j^{(2)}}\prod\limits_{\substack{n=1\\ n\notin\Phi_i}}^{N}K_n(\tau_n)\prod\limits_{n\in\Phi_i}\left(1-K_n(\tau_n)\right)}, \tag{9}$$

$$K_u(\tau_1,...,\tau_N)=\frac{\sum\limits_{i=1}^{\omega}\prod\limits_{n\in M_i}K_n(\tau_n)\prod\limits_{\substack{n=1\\ n\notin M_i}}^{N}\left(1-K_n(\tau_n)\right)}{\sum\limits_{i=1}^{\omega}\prod\limits_{n\in M_i}K_n(\tau_n)\prod\limits_{\substack{n=1\\ n\notin M_i}}^{N}\left(1-K_n(\tau_n)\right)+\sum\limits_{i=1}^{s}\prod\limits_{\substack{n=1\\ n\notin\Phi_i}}^{N}K_n(\tau_n)\prod\limits_{n\in\Phi_i}\left(1-K_n(\tau_n)\right)}. \tag{10}$$

To define mean specific income $S(\tau_1,...,\tau_N)$ per calendar time unit and mean specific expenses $C^*(\tau_1,...,\tau_N)$ per time unit of system's up state the Formula [10] will be used

$$S^*=\frac{\int\limits_{E^*}m(z)f_s(z)\rho(dz)}{\int\limits_{E^*}m(z)\rho(dz)},\ C^*=\frac{\int\limits_{E^*}m(z)f_c(z)\rho(dz)}{\int\limits_{E_+^*}m(z)\rho(dz)} \tag{11}$$

where $f_s(z)$, $f_c(z)$ are functions defining income and expenses respectively in each state. These functions are as follows:

$$f_s(z)=\begin{cases}-\sum\limits_{\substack{k\in\Omega_d^0\\k\notin I_d}} c_k-\sum\limits_{\substack{k\in\Omega_d^2\\k\notin I_d}} c_k^p,\ z\in\left\{\ i\overline{\overline{dx}}^{(i)}\overline{u}\ \right\}\in E_-^*,\\ \sum\limits_{\substack{k\in\Omega_d^1\\k\notin I_d}} c_k^0-\sum\limits_{\substack{k\in\Omega_d^0\\k\notin I_d}} c_k-\sum\limits_{\substack{k\in\Omega_d^2\\k\notin I_d}} c_k^p,\ z\in\left\{\ i\overline{\overline{dx}}^{(i)}\overline{u}\ \right\}\in E_+^*,\end{cases}$$

$$f_c(z)=\sum_{\substack{k\in\Omega_d^0\\k\notin I_d}} c_k+\sum_{\substack{k\in\Omega_d^2\\k\notin I_d}} c_k^p,\quad z\in\left\{\ i\overline{\overline{dx}}^{(i)}\overline{u}\ \right\}\in E^*.$$

Here c_i^0, c_i and c_i^p, $i=\overline{1,N}$, are income per time unit of system's up state, expenses per time unit of ER, and expenses per time unit of system's i-element maintenance respectively.

The Formula (11) can be transformed into the following expressions:

$$S^*(\tau_1,...,\tau_N)=\frac{\sum\limits_{d\in D^*}\left[\sum\limits_{\substack{k\in\Omega_d^1\\k\notin I_d}} c_k^0-\sum\limits_{\substack{k\in\Omega_d^0\\k\notin I_d}} c_k-\sum\limits_{\substack{k\in\Omega_d^2\\k\notin I_d}} c_k^p\right]\prod\limits_{k=1}^{N}T_k^{(d_k)}(\tau_k)}{\sum\limits_{d\in D^*}\prod\limits_{k=1}^{N}T_k^{(d_k)}(\tau_k)}=$$

$$\left\{\sum_{i=1}^{\omega}\left[\sum_{\substack{j\in M_i\\ j\notin A(M_i)}} c_j^0\prod_{n\in M_i}K_n(\tau_n)\prod_{\substack{n=1\\n\notin M_i}}^{N}\left(1-K_n(\tau_n)\right)-\sum_{j\notin M_i}C_j(\tau_j)K_j(\tau_j)\prod_{n\in M_i}K_n(\tau_n)\prod_{\substack{n\in M_i\\n\neq j}}\left(1-K_n(\tau_n)\right)\right]-\right.$$

$$\left.\sum_{i=1}^{s}\sum_{\substack{j\in\Phi_i\\ j\notin A(\Phi_i)}}C_j(\tau_j)K_j(\tau_j)\prod_{\substack{n=1\\n\notin\Phi_i}}^{N}K_n(\tau_n)\prod_{\substack{n\in\Phi_i\\n\neq j}}\left(1-K_n(\tau_n)\right)\right\}/$$

$$\left[\sum_{i=1}^{\omega}\prod_{n\in M_i}K_n(\tau_n)\prod_{\substack{n=1\\n\notin M_i}}^{N}\left(1-K_n(\tau_n)\right)+\sum_{i=1}^{s}\prod_{\substack{n=1\\n\notin\Phi_i}}^{N}K_n(\tau_n)\prod_{n\in\Phi_i}\left(1-K_n(\tau_n)\right)\right] \tag{12}$$

$$C^*(\tau_1,...,\tau_N)=\frac{\sum\limits_{d\in D^*}\left[\sum\limits_{\substack{k\in\Omega_d^0\\k\notin I_d}} c_k+\sum\limits_{\substack{k\in\Omega_d^2\\k\notin I_d}} c_k^p\right]\prod\limits_{k=1}^{N}T_k^{(d_k)}(\tau_k)}{\sum\limits_{d\in D_+^*}\prod\limits_{k=1}^{N}T_k^{(d_k)}(\tau_k)}=\left\{\sum_{i=1}^{\omega}\sum_{j\in M_i}C_j(\tau_j)K_j(\tau_j)\prod_{n\in M_i}K_n(\tau_n)\prod_{\substack{n\notin M_i\\n\neq j}}\left(1-K_n(\tau_n)\right)+\right.$$

$$\left.\sum_{i=1}^{s}\sum_{\substack{j\in\Phi_i\\ j\notin A(\Phi_i)}}C_j(\tau_j)K_j(\tau_j)\prod_{\substack{n=1\\n\notin\Phi_i}}^{N}K_n(\tau_n)\prod_{\substack{n\in\Phi_i\\n\neq j}}\left(1-K_n(\tau_n)\right)\right\}/\left[\sum_{i=1}^{\omega}\prod_{n\in M_i}K_n(\tau_n)\prod_{\substack{n=1\\n\notin M_i}}^{N}\left(1-K_n(\tau_n)\right)\right] \tag{13}$$

Here $C_i(\tau_i)=\dfrac{c_i^pT_i^{(2)}(\tau_i)+c_iT_i^{(0)}(\tau_i)}{T_i^{(1)}(\tau_i)}$ are mean specific expenses per time unit of i-element's up state.

4. Optimization of Elements' Maintenance Terms

The task of defining optimal terms of elements' maintenance execution with the purpose of gaining the best system's performance index is reduced to the definition of the points of absolute extremum τ_i^u, τ_i^s, τ_i^c of the functions (10), (12) and (13) respectively. The attainment of function's extremums under some arguments $\tau_j\to\infty$ signifies that it is not expedient to execute maintenance of elements with respective numbers. In this case we should change $K_j(\infty)$ for $\dfrac{M\alpha_j}{M\alpha_j+M\beta_j}$, and $C_j(\infty)$

for $\frac{c_j M\beta_j}{M\alpha_j}$ in the Formulas (10), (12) and (13).

Let us write down formulas for the definition of stationary characteristics of multicomponent systems with concrete structures.

Stationary characteristics of serial system. The structure including N elements in series has one path $M_1 = \{1,...,N\}$ and N sections $\{\Phi_i\}_{i=1}^N = \{\{1\},\{2\},...,\{N\}\}$. System stationary performance indexes (8)–(10), (12) and (13) will be given by:

$$K_u^*(\tau_1,...,\tau_N) = \left[1+\sum_{i=1}^N \frac{1-K_i(\tau_i)}{K_i(\tau_i)}\right]^{-1},$$

$$S^*(\tau_1,...,\tau_N) = \frac{\sum_{i=1}^N c_i^0 - \sum_{i=1}^N C_i(\tau_i)}{1+\sum_{i=1}^N \frac{1-K_i(\tau_i)}{K_i(\tau_i)}},$$

$$C^*(\tau_1,...,\tau_N) = \sum_{i=1}^N C_i(\tau_i),$$

$$T_+^*(\tau_1,...,\tau_N) = \frac{1}{\sum_{i=1}^N \frac{1}{T_i^{(1)}(\tau_i)}},$$

$$T_-^*(\tau_1,...,\tau_N) = \frac{\sum_{i=1}^N \frac{T_i^{(0)}(\tau_i)+T_i^{(2)}(\tau_i)}{T_i^{(1)}(\tau_i)}}{\sum_{i=1}^N \frac{1}{T_i^{(1)}(\tau_i)}}.$$

Stationary characteristics of parallel-serial system. The block scheme of the parallel-serial system is shown in **Figure 2**.

For the system of the structure like this the Formulas (10), (12) and (13) for the system stationary characteristics definition are as follows:

$$K_u^*\left(\tau_{11},\ldots,\tau_{LN_L}\right) = 1-\prod_{i=1}^L\left(1-\left[1+\sum_{n=1}^{N_i}\frac{1-K_{in}\left(\tau_{in}\right)}{K_{in}\left(\tau_{in}\right)}\right]^{-1}\right),$$

$$S^*\left(\tau_{11},\ldots,\tau_{LN_L}\right) = \sum_{i=1}^L 1+\left[1+\sum_{n=1}^{N_i}\frac{1-K_{in}\left(\tau_{in}\right)}{K_{in}\left(\tau_{in}\right)}\right]^{-1}\sum_{n=1}^{N_i}\frac{S_{in}\left(\tau_{in}\right)}{K_{in}\left(\tau_{in}\right)},$$

$$C^*\left(\tau_{11},\ldots,\tau_{LN_L}\right) = \frac{1}{K_u^*\left(\tau_{11},\ldots,\tau_{LN_L}\right)}\sum_{i=1}^L\sum_{n=1}^{N_i}C_{in}\left(\tau_{in}\right)\left[1+\sum_{n=1}^{N_i}\frac{1-K_{in}\left(\tau_{in}\right)}{K_{in}\left(\tau_{in}\right)}\right]^{-1},$$

where $K_{in}\left(\tau_{in}\right), S_{in}\left(\tau_{in}\right), C_{in}\left(\tau_{in}\right)$ are SSAF, mean specific income of *i*-chain's *n*-element per calendar time unit, and mean specific expenses per time unit of element's up state respectively:

$$K_{in}\left(\tau_{in}\right) = \frac{T_{in}^{(1)}\left(\tau_{in}\right)}{T_{in}^{(1)}\left(\tau_{in}\right)+T_{in}^{(0)}\left(\tau_{in}\right)+T_{in}^{(2)}\left(\tau_{in}\right)},$$

$$S_{in}\left(\tau_{in}\right) = \frac{c_{in}^0 T_{in}^{(1)}\left(\tau_{in}\right)-c_{in}T_{in}^{(0)}\left(\tau_{in}\right)-c_{in}^p T_{in}^{(2)}\left(\tau_{in}\right)}{T_{in}^{(1)}\left(\tau_{in}\right)+T_{in}^{(0)}\left(\tau_{in}\right)+T_{in}^{(2)}\left(\tau_{in}\right)},$$

$$C_{in}\left(\tau_{in}\right) = \frac{c_{in}T_{in}^{(0)}\left(\tau_{in}\right)+c_{in}^p T_{in}^{(2)}\left(\tau_{in}\right)}{T_{in}^{(1)}\left(\tau_{in}\right)},$$

$$T_{in}^{(1)}\left(\tau_{in}\right) = \int_0^{\tau_{in}}\overline{F}_{in}\left(s\right)ds,$$

$$T_{in}^{(2)}\left(\tau_{in}\right) = M\beta_{in}^p\overline{F}_{in}\left(\tau_{in}\right),$$

$$T_{in}^{(0)}\left(\tau_{in}\right) = M\beta_{in}F_{in}\left(\tau_{in}\right).$$

Stationary characteristics of serial-parallel system. The block scheme of serial-parallel system is shown in **Figure 3**.

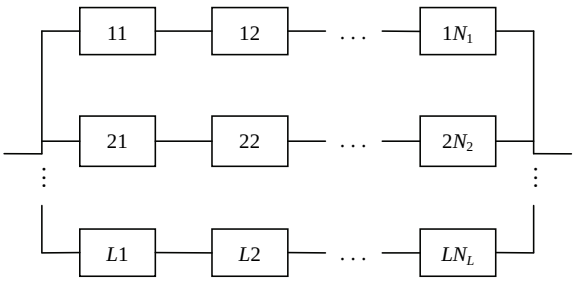


Figure 2. Block scheme of parallel-serial system.

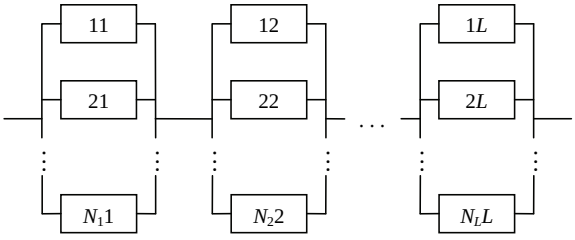


Figure 3. Block scheme of serial-parallel system.

System stationary performance indexes are defined by the formulas:

$$K_u^*\left(\tau_{11},\ldots,\tau_{LN_L}\right)=\left[1+\sum_{i=1}^{L}\frac{\prod_{n=1}^{N_i}\left(1-K_{ni}\left(\tau_{ni}\right)\right)}{1-\prod_{n=1}^{N_i}\left(1-K_{ni}\left(\tau_{ni}\right)\right)}\right]^{-1},$$

$$S_u^*\left(\tau_{11},\ldots,\tau_{LN_L}\right)=K_u^*\left(\tau_{11},\ldots,\tau_{LN_L}\right)\sum_{i=1}^{L}\frac{\sum_{n=1}^{N_i}S_{ni}\left(\tau_{ni}\right)}{1-\prod_{n=1}^{N_i}\left(1-K_{ni}\left(\tau_{ni}\right)\right)},$$

$$C_u^*\left(\tau_{11},\ldots,\tau_{LN_L}\right)=\sum_{i=1}^{L}\frac{\sum_{n=1}^{N_i}C_{ni}\left(\tau_{ni}\right)K_{ni}\left(\tau_{ni}\right)}{1-\prod_{n=1}^{N_i}\left(1-K_{ni}\left(\tau_{ni}\right)\right)}.$$

where $K_{ni}\left(\tau_{ni}\right), S_{ni}\left(\tau_{ni}\right), C_{ni}\left(\tau_{ni}\right)$ are SSAF, mean specific income of the *i*-chain's *n*-element per calendar time unit, and mean specific expenses per time unit of element's up state:

$$K_{ni}\left(\tau_{ni}\right)=\frac{T_{ni}^{(1)}\left(\tau_{ni}\right)}{T_{ni}^{(1)}\left(\tau_{ni}\right)+T_{ni}^{(0)}\left(\tau_{ni}\right)+T_{ni}^{(2)}\left(\tau_{ni}\right)},$$

$$S_{ni}\left(\tau_{ni}\right)=\frac{c_{ni}^{0}T_{ni}^{(1)}\left(\tau_{ni}\right)-c_{ni}T_{ni}^{(0)}\left(\tau_{ni}\right)-c_{ni}^{p}T_{ni}^{(2)}\left(\tau_{ni}\right)}{T_{ni}^{(1)}\left(\tau_{ni}\right)+T_{ni}^{(0)}\left(\tau_{ni}\right)+T_{ni}^{(2)}\left(\tau_{ni}\right)},$$

$$C_{ni}\left(\tau_{ni}\right)=\frac{c_{ni}T_{ni}^{(0)}\left(\tau_{ni}\right)+c_{ni}^{p}T_{ni}^{(2)}\left(\tau_{ni}\right)}{T_{ni}^{(1)}\left(\tau_{ni}\right)},$$

$$T_{ni}^{(1)}\left(\tau_{ni}\right)=\int_{0}^{\tau_{ni}}\bar{F}_{ni}\left(s\right)ds,$$

$$T_{ni}^{(2)}\left(\tau_{ni}\right)=M\beta_{ni}^{p}\bar{F}_{ni}\left(\tau_{ni}\right),$$

$$T_{ni}^{(0)}\left(\tau_{ni}\right)=M\beta_{ni}F_{ni}\left(\tau_{ni}\right).$$

Let us make concrete calculations to define optimal maintenance execution terms for three-component serial system. Let operating time to failure and restoration time are disposed according to Erlang with densities

$$f_i(t)=\lambda_i\frac{\left(\lambda_i t\right)^{m_i-1}}{(m_i-1)!}e^{-\lambda_i t},\quad g_i(t)=\mu_i\frac{\left(\mu_i t\right)^{m_i-1}}{(m_i-1)!}e^{-\mu_i t},$$

$i=1,2,3$. Initial data and calculation results are represented in the **Tables 1** and **2**.

In the **Table 2** K_u^{∞}, S^{∞}, C^{∞} denote system performance indexes in case if elements' maintenance is not carried out. If elements attain optimal time to failure, their maintenance execution increases these indexes for 4.5%, 6.8% and 38.3% respectively.

5. Conclusions

In the present paper semi-Markovian model of the multicomponent restorable system operation with regard to elements' deactivation and maintenance in age has been built. With the help of this model an explicit form of reliability and economical stationary performance indexes for the system with assumption of a general form of elements' time to failure and restoration time distributions has been defined. The system stationary characteristics found are explicitly dependent on the periodicity of its elements' maintenance execution. This fact allows solving the problems of the characteristics' improvement. In the limiting case (when the periodicity of elements'

Table 1. System initial data.

№	γ_i	θ_i	$M\alpha_i$, *h*	$M\beta_i$, *h*	$M\beta_i^p$, *h*	c_i^0, *c.u./h*	c_i, *c.u./h*	c_i^p, *c.u./h*
1	2	50	44.311	5	1	5	1	0.2
2	3	15	13.395	3	1	7	3	2
3	4	20	18.128	4	0.5	9	3	1

Table 2. Calculation results.

№	τ_i^k, *h*	$K_u^{\max}$	K_u^{∞}	τ_i^s, *h*	$S^{\max}$ *c.u./h*	S^{∞} *c.u./h*	τ_i^c, *h*	$C^{\max}$ *c.u./h*	C^{∞} *c.u./h*
1	25.533			23.131			15.608		
2	9.548	0.916	0.869	8.982	18.553	16.393	7.694	0.507	1.373
3	9.354			8.852			6.909		

maintenance execution increases infinitely) the stationary characteristics defined in the present work take the form of the well-known expressions for the characteristics of restorable system in case of the passive strategy of maintenance (elements' maintenance is not carried out) [8,9].

6. References

[1] C. Valdez-Flores and R. M. Feldman, "A Survey of Preventive Maintenance Models for Stochastically Deteriorating Single-Unit Systems," *Naval Research Logistics*, Vol. 36, No. 4, 1989, pp. 419-446.

[2] D. I. Cho and M. Parlar, "A Survey of Maintenance Models for Multi-Unit Systems," *European Journal of Operational Research*, Vol. 51, No. 2, 1991, pp. 1-23.

[3] R. Dekker and R. A. Wildeman, "A Review of Multi-Component Maintenance Models with Economic Dependence," *Mathematical Methods of Operations Research*, Vol. 45, No. 3, 1997, pp. 411-435.

[4] F. Beichelt and P. Franken, "Zuverlässigkeit und Instanphaltung," *Mathematische Methoden*, VEB Verlag Technik, Berlin, 1983.

[5] R. E. Barlow and F. Proschan, "Mathematical Theory of Reliability," John Wiley and Sons, New York, 1965.

[6] V. A. Kashtanov and A. I. Medvedev, "The Theory of Complex Systems' Reliability (Theory and Practice)," European Center for Quality, Moscow, 2002.

[7] A. I. Peschansky, "Monotonous System Maintenance with Regard to Operating Time to Failure of Each Element," *Industrial Processes Optimization*, Vol. 11, 2009, pp. 77-83.

[8] V. S. Korolyuk and A. F. Turbin, "Markovian Restoration Processes in the Problems of System Reliability," Naukova dumka, Kiev, 1982.

[9] A. N. Korlat, V. N. Kuznetsov, M. I. Novikov and A. F. Turbin, "Semi-Markovian Models of Restorable and Service Systems," Shtiintsa, Kishinev, 1991.

[10] V. M. Shurenkov, "Ergodic Markovian Processes," Nauka, Moscow, 1989.

Applied Mathematics, **2010, 1, 244-249**
doi:10.4236/am.2010.13030 Published Online September 2010 (http://www.SciRP.org/journal/am)

Reinforcing a Matroid to Have *k* Disjoint Bases

Hong-Jian Lai[1,2], Ping Li[2], Yanting Liang[2], Jinquan Xu[3]
[1]*College of Mathematics and System Sciences, Xinjiang University, Urumqi, China*
[2]*Department of Mathematics, West Virginia University, Morgantown, USA*
[3]*Department of Mathematics, Huizhou University, Huizhou, China*
E-mail: *hjlai@math.wvu.edu*
Received May 14, 2010; *revised July* 29, 2010; *accepted August* 2, 2010

Abstract

Let $\tau(M)$ denote the maximum number of disjoint bases in a matroid M. For a connected graph G, let $\tau(G) = \tau(M(G))$, where $M(G)$ is the cycle matroid of G. The well-known spanning tree packing theorem of Nash-Williams and Tutte characterizes graphs G with $\tau(G) \geq k$. Edmonds generalizes this theorem to matroids. In [1] and [2], for a matroid M with $\tau(M) \geq k$, elements $e \in E(M)$ with the property that $\tau(M-e) \geq k$ have been characterized in terms of matroid invariants such as strength and η-partitions. In this paper, we consider matroids M with $\tau(M) < k$, and determine the minimum of $|E(M')|-|E(M)|$, where M' is a matroid that contains M as a restriction with both $r(M') = r(M)$ and $\tau(M') \geq k$. This minimum is expressed as a function of certain invariants of M, as well as a min-max formula. These are applied to imply former results of Haas [3] and of Liu *et al.* [4].

Keywords: Disjoint Bases, Edge-Disjoint Spanning Trees, Spanning Tree Packing Numbers, Strength, Fractional Arboricity

1. Introduction

In this paper, we use N and Q_+ to denote the set of all natural numbers and the set of all positive fractional numbers, respectively, and consider finite matroids and graphs. Undefined notations and terminology can be found in [5] or [6] for matroids, and [7] for graphs. Thus for a connected graph G, $\omega(G)$ denotes the number of components of G. For a matroid M, r_M (or r, when the matroid M is understood from the context) denotes the rank function of M, and $E(M)$, $I(M)$, $C(M)$ and $B(M)$ denote the ground set of M, and the collections of independent sets, the circuits, and the bases of M, respectively. Furthermore, if M is a matroid with $E = E(M)$, and if $X \subset E$, then $M-X$ is the restricted matroid of M obtained by deleting the elements in X from M, and M/X is the matroid obtained by contracting elements in X from M. As in [5] or [6], we use $M-e$ for $M-\{e\}$ and M/e for $M/\{e\}$.

For a matroid M, let $\tau(M)$ denote the maximum number of disjoint bases of M. For a graph G, define $\tau(G) = \tau(M(G))$, where $M(G)$ denotes the cycle matroid of G. Thus if G is a connected graph, then $\tau(G)$ is the spanning tree packing number of G. Readers are referred to [8] for a survey on $\tau(G)$. The well-known spanning tree packing theorem of Nash-Williams [9] and Tutte [10] characterizes graphs with k edge-disjoint spanning trees, for any integer $k > 0$. Edmonds [11] proved the corresponding theorem for matroids.

Let $k > 0$ be an integer. For any matroid M with $\tau(M) \geq k$, we are interested in finding elements $e \in E(M)$ that have the property that $\tau(M-e) \geq k$. Characterizations of all such elements have been found in [1] and [2]. For a graph G, the problem of determining which edges should be added to G so that the resulting graph has k edge-disjoint spanning trees has been studied, see Haas [3] and Liu *et al.* [4], among others. As the arguments in these papers are involved vertices, it is natural to consider the possibility of extending these results to matroids. Since matroids in general do not have a concept corresponding to vertices, one can no longer add an element to a matroid as adding an edge in graphs. Therefore, we need to reformulate the problem so that it would fit the matroid setting while generalizing the graph theory results.

Let M be a matroid and $k \in N$. If there is a matroid M' with $r(M') = r(M)$ and $\tau(M') \geq k$ such that M' has a restriction isomorphic to M (we then

view M as a restriction of M'), then M' is a $(\tau \geq k)$-extension of M. We shall show that any matroid has a $(\tau \geq k)$-extension. We then define $F(M,k)$ to be the minimum integer $l > 0$ such that M has a $(\tau \geq k)$-extension M' with $|E(M')| - |E(M)| = l$. The main purpose of this paper is to determine $F(M,k)$ in terms of other invariants of M.

By definition, if M is a matroid with $r(M) = 0$, then $\forall k \in N$, $\tau(M) \geq k$. Accordingly, for a connected graph G, if $|V(G)| = 1$, then $\tau(G) \geq k$ for any $k \in N$. For a graph G, $a_1(G)$, the edge arboricity of G, is the minimum number of spanning trees of G whose union equals $E(G)$. For a matroid, we define the similar concept $\gamma_1(M)$, which is the minimum number of bases of M whose union equals $E(M)$. The following theorems are well known.

Theorem 1.1 *Let G be a connected graph with $|V(G)| > 1$, and let $k > 0$ be an integer. Each of the following holds.*

1) (Nash-Williams [9] and Tutte [10]) $\tau(G) \geq k$ if and only if $\forall X \subseteq E(G)$, $|X| \geq k(\omega(G - X) - 1)$.

2) (Nash-Williams [12]) $a_1(G) \leq k$ if and only if $\forall X \subseteq E(G)$, $|X| \leq k(|V(G[X])| - \omega(G[X]))$.

Theorem 1.2 (*Edmonds* [11]) *Let M be a matroid with $r(M) > 0$. Each of the following holds.*

1) $\tau(M) \geq k$ if and only if $\forall X \subseteq E(M)$, $|E(M) - X| \geq k(r(M) - r(X))$.

2) $\gamma_1(M) \leq k$ if and only if $\forall X \subseteq E(M)$, $|X| \leq kr(X)$.

Let M be a matroid with rank function r. For any subset $X \subseteq E(M)$ with $r(X) > 0$, the density of X is

$$d_M(X) = \frac{|X|}{r_M(X)}.$$

When the matroid M is understood from the context, we often omit the subscript M. We also use $d(M)$ for $d(E(M))$. Following [13] and [14], the strength $\eta(M)$ and the fractional arboricity $\gamma(M)$ of M are respectively defined as

$$\begin{aligned} &\eta(M) = \min\{d(M/X) : r(X) < r(M)\}, \\ &\text{and } \gamma(M) = \max\{d(X) : r(X) > 0\}. \end{aligned} \tag{1}$$

Thus Theorem 1.2 above indicates that

$$\tau(M) = \lfloor \eta(M) \rfloor, \text{ and } \gamma_1(M) = \lceil \gamma(M) \rceil. \tag{2}$$

We assume that M is a matroid with $r(M) > 0$. A subset $X \subseteq E(M)$ is an η-maximal subset and $M|X$ is an η-maximal restriction if for any subset $Y \subseteq E(M)$ that properly contains X, we have $\eta(M|Y) < \eta(M|X)$. In [1] and [2], it has been proved that any matroid M has a unique decomposition based on its η-maximal subsets.

Theorem 1.3 ([1] *and* [2]) *Let M be a matroid with $r(M) > 0$. Then each of the following holds.*

1) There exist an integer $m \in N$, and an m-tuple $(l_1, l_2, ..., l_m)$ of rational numbers in Q_+ such that

$$\eta(M) = l_1 < l_2 < ... < l_m = \gamma(G), \tag{3}$$

and a sequence of subsets

$$J_m \subset ... \subset J_2 \subset J_1 = E(M); \tag{4}$$

such that for each i with $1 \leq i \leq m$, $M|J_i$ is an η-maximal restriction of M with $\eta(M|J_i) = l_i$.

2) The integer m and the sequences (4) and (3) are uniquely determined by M.

For a matroid M, the m-tuple $(l_1, l_2, ..., l_m)$ and the sequence in (4) will be referred as the η-spectrum and the η-decomposition of M, respectively. For each subscript j with $1 \leq j \leq m$, we refer J_j to be the set corresponding to l_j. Our main result can now be stated as follows.

Theorem 1.4 *For $k \in N$, let M be a matroid with $\tau(M) \leq k$. If $\gamma(M) < k$, define $J_{i(k)} = \varnothing$; and if $\gamma(M) \geq k$, let $i(k)$ denote the smallest subscript in* (3) *such that $l_{i(k)} \geq k$. Then*

1) $F(M,k) = k(r(M) - r(J_{i(k)})) - |E(M) - J_{i(k)}|$.

2) $F(M,k) = \max_{X \subseteq E(M)}\{kr(M/X) - |M/X|\}$.

In the next section, we shall present some of the useful properties related to the strength and the fractional arboricity of a matroid M, and to the decomposition of M. Section 3 will be devoted to the proofs of the main results. In the last section, we shall show some applications of our main results.

2. Preliminaries

Both $\eta(M)$ and $\gamma(M)$ have been studied by many, see [14-16] and [17], among others. From the definition of $d(M), \eta(M)$ and $\gamma(M)$, we immediately have, for any matroid M with $r(M) > 0$,

$$\eta(M) \leq d(M) \leq \gamma(M). \tag{5}$$

A matroid M satisfying $\eta(M) = \gamma(M)$ is called a uniformly dense matroid. Both $\eta(M)$ and $\gamma(M)$ can also be described by their behavior in some parallel extension of the matroid M.

Definition 2.1 *Let M be a matroid and let $\phi : E(M) \mapsto N$ be a function. For each $e \in E(M)$, let $X_e = \{e^1, e^2, \cdots, e^{\varphi(e)}\}$ be a set such that $X_e \cap X_{e'} = \varnothing$, $\forall e, e' \in E(M)$ with $e \neq e'$. The ϕ-*parallel extension *of M, denoted by M_ϕ, is obtained from M by replacing each element $e \in E(M)$ by a class of $\phi(e)$ parallel elements X_e. Thus $E(M_\varphi) = \bigcup_{e \in E(M)} X_e$ such that a subset $Y \subseteq E(M_\phi)$ is independent in M_ϕ if and*

only if both $\{e \in E(M): X_e \cap Y \neq \varnothing\}$ *is independent in* M *and* $\forall e \in E(M)$, $|X_e \cap Y| \leq 1$. *For* $t \in N$, *if* $\forall e \in E(M)$, $\phi(e) = t$ *is a constant function, we write* M_t *for* M_ϕ, *and call* M_t *the* t-parallel extension *of* M.

Let $E' = \{e^1 : e \in E(M)\} \subseteq E(M_\varphi)$. Then the bijection $e \leftrightarrow e^1$ between $E(M)$ and E' yields a matroid isomorphism between M and $M_\phi | E'$. Under this bijection, we shall view that $M = M_\phi | E'$ is a restriction of M_ϕ.

Theorem 2.2 (*Theorem* 4 *of* [14]) *Let* M *be a matroid and let* $s \geq t > 0$ *be integers. Then*

1) $\eta(M) \geq \frac{s}{t}$ if and only if $\eta(M_t) \geq s$.

2) $\gamma(G) \leq \frac{s}{t}$ if and only if $\gamma(M_t) \leq s$.

Theorem 2.3 (*Theorem* 6 *of* [14]) *Let* M *be a matroid. The following are equivalent.*

1) $\eta(M) = d(M)$.
2) $\gamma(M) = d(M)$.
3) $\eta(M) = \gamma(M)$.
4) $\eta(M) = \frac{s}{t}$, for some integers $s \geq t > 0$, and M_t, the t-parallel extension of M, is a disjoint union of s bases of M.
5) $\gamma(M) = \frac{s}{t}$, for some integers $s \geq t > 0$, and M_t, the t-parallel extension of M, is a disjoint union of s bases of M.

Lemma 2.4 ([14], [1] *and* [2]) *Let* M *be a matroid with* $r(M) > 0$, *and let* $l \geq 1$ *be fractional number. Each of the following holds.*

1) (Lemma 10 [14]) If $X \subset E(M)$ and if $\eta(M | X) \geq \eta(M)$, then $\eta(M / X) = \eta(M)$.
2) (Theorem 17 of [14]) If $X \subset E(M)$ and if $d(X) = \gamma(M)$, then $\eta(M | X) = \gamma(M | X) = d(X) = \gamma(M)$.
3) A matroid M is uniformly dense if and only if any subset $X \subseteq E(M)$, $d(X) \leq \eta(M)$.
4) A matroid M is uniformly dense if and only if for any restriction N of M, $\eta(N) \leq \eta(M)$.
5) If $d(M) \geq l$, then there exists a subset $X \subseteq E(M)$ with $r(X) > 0$ such that $\eta(M | X) \geq l$.

For each rational number $l > 1$, define

$$S_l = \{M : \eta(M) \geq l\}. \quad (6)$$

Proposition 2.5 ([1] *and* [2]) *Let* $p > q > 0$ *be integers, and* $l = \frac{p}{q} \in Q_+$ *be a rational number. The matroid family* S_l *satisfies the following properties.*

(C1) If $r(M) = 0$, then $M \in S_l$.
(C2) If $M \in S_l$ and if $e \in E(M)$, then $M / e \in S_l$.
(C3) Let $X \subseteq E(M)$ and let $N = M | X$. If $M / X \in S_l$ and if $N \in S_l$, then $M \in S_l$.

Lemma 2.6 ([1] *and* [2]) *Let* W, $W' \subset E(M)$ *be subsets, and let* $l \in Q_+$. *If* $\eta(M | W) \geq l$ *and* $\eta(M | W') \geq l$, *then* $\eta(M | (W \cup W')) \geq l$.

Lemma 2.7 ([1] *and* [2]) *If* $X \subseteq E(M)$ *is an* η-*maximal subset, then* X *is a closed set in* M.

3. Characterization of the Must-Added Elements with Respect to Having *k* Disjoint Bases

The main purpose of this section is to prove Theorems 1.4. We will start with a lemma.

Lemma 3.1 *Let* M *be a matroid and let* $k > 0$ *be an integer. Each of the following holds.*

1) $\eta(M) \geq k$ if and only if $F(M,k) = 0$.
2) If $\gamma(M) \leq k$, then

$$F(M,k) = kr(M) - |E(M)|.$$

Moreover, there exists a map $\phi: E(M) \mapsto N$, such that M_ϕ is a matroid that contains M as a restriction with $\eta(M_\phi) = \gamma(M_\phi) = k$, and such that $|E(M_\varphi)| - |E(M)| = F(M,k)$.

Proof: 1) By (2), $\eta(M) \geq k$ if and only if $\tau(M) \geq k$. By the definition of $F(M,k)$, $\tau(M) \geq k$ if and only if $F(M,k) = 0$. This proves 1).

2) Since $\gamma(M) \leq k$, it follows by (2) that M has disjoint bases $B_1, \cdots, B_k$ such that $E(M) = \bigcup_{i=1}^{k} B_i$. Define $\phi(e) = |\{B_i : e \in B_i\}|$. Then $\phi: E(M) \mapsto N$. Let $L = M_\phi$ be the ϕ-parallel extension of M. Then by Definition 2.1, M is contained in L as a restriction. Moreover, both $|E(L)| = \sum_{i=1}^{k} |B_i| = kr(M)$ and $\tau(L) = k$. It follows by Theorem 2.3 that $\eta(L) = \gamma(L) = k$. Hence by Theorem 2.3 1) or 2), $|E(L)| = k\,r(L) = k\,r(M)$, and so $F(M,k) = |E(L)| - |E(M)| = kr(M) - |E(M)|$.

When $k = 2$, the cycle matroid version of Lemma 3.1 has been frequently applied in the study of supereulerian graphs, see Theorem 7 of [18] and Lemma 2.3 of [19], among others. (For a literature review on supereulerian graphs, see [20] and [21].)

Proof of Theorem 1.4 1): Let M be a matroid with $r(M) > 0$. If $\tau(M) \geq k$, then by (2) and by Theorem 1.3, $i(k) = i_1$, and so

$$E(M) = J_{i(k)}, \text{ and } F(M,k) = 0.$$

Thus Theorem 1.4 1) follows trivially with $\tau(M) \geq k$. Hence we assume that $\tau(M) < k$. If $\gamma(M) < k$, then Theorem 1.4 1) follows from Lemma 3.1.

Therefore, we may assume that $\eta(M) < k$ and $\gamma(M) \geq k$. By Theorem 1.3, we must have $m > 1$. Let

$i(k)$ be the smallest subscript in η-spectrum (3) of M such that $l_{i(k)} \geq k$. By Theorem 1.3, $\eta(M \mid J_{i(k)}) \geq k$. Let $M' = M / J_{i(k)}$. By the assumption that $\eta(M) < k$ and by Lemma 2.4 1), $\eta(M') = \eta(M)$. By the choice of $i(k)$, $\gamma(M') < k$, and so by Lemma 3.1,

$$F(M', k) = kr(M') - |E(M')|, \tag{7}$$

and there must be a function $\phi' : E(M') \mapsto N$ such that $M'_{\phi'}$ satisfies $\eta(M'_{\phi'}) = \gamma(M'_{\phi'}) = k$. Define $\phi : E(M) \mapsto N$ as follows:

$$\phi(e) = \begin{cases} \phi'(e) & \text{if } e \notin J_{i(k)} \\ 1 & \text{if } e \in J_{i(k)} \end{cases}.$$

Then M_ϕ is a matroid that contains M as a restriction, such that $J_k(M) \subset E(M_\phi)$. By the definition of ϕ, $M_\phi \mid J_{i(k)} = M \mid J_{i(k)} \in S_k$. Since $M_\phi / J_{i(k)} = M'_{\phi'} \in S_k$, it follows by Proposition 2.5(C3) that $M_\phi \in S_k$. Thus by (7) and by Lemma 2.7,

$$\begin{aligned} F(M,k) &= F(M',k) = kr(M') - |E(M')| \\ &= k\left(r(M) - r\left(J_{i(k)}\right)\right) - \left|E(M) - J_{i(k)}\right|, \end{aligned}$$

and so Theorem 1.4 1) is established.

To continue our proof for Theorem 1.4, we introduce the following function: for any $X \subseteq E(M)$, define

$$\begin{aligned} &f_k(M,X) = kr(M/X) - |M/X|, \\ &\text{and } F_k(M) = \max_{X \subseteq E(M)} \{f_k(M,X)\}. \end{aligned} \tag{8}$$

The function $f_k(M,X)$ was introduced by Bruno and Weinberg [22] to investigate the principal partition of matroids. They are closely related to the strength and fractional arboricity of matroids, as to be shown in Lemma 3.2 below.

Lemma 3.2 *Let M be a matroid with $r(M) > 0$, and let $k > 0$ be an integer. Each of the following holds.*

1) $F_k(M) = 0$ if and only if $\eta(M) \geq k$.

2) $F_k(M) = f_k(M, \varnothing)$ if and only if $\gamma(M) \leq k$.

3) Let $i(k)$ denote the smallest i_j in (3) such that $i(k) \geq k$, and $J_{i(k)}$ denote the corresponding set in the η-decomposition (4) of M. Then $F_k(M / J_{i(k)}) = F(M,k)$.

4) For any $e \in E(M)$, $F_k(M) \geq F_k(M / e)$. In particular, $F_k(M) \geq F(M,k)$.

5) If $X_0 \subset E(M)$ satisfies $F_k(M) = f_k(M, X_0)$, then $F_k(M) = f_k(M / X_0) = F_k(M / X_0) = f_k(M / X_0, \varnothing)$ and $\gamma(M / X_0) \leq k$.

Proof: 1) By definition (8), $F_k(M) = 0$ if and only if $\forall X \subseteq E(M)$, $f_k(M,X) = kr(M / X) - |E(M / X)| \leq 0$. By the definition of $\eta(M)$, $\forall X \subseteq E(M)$, $kr(M / X) - |E(M / X)| \leq 0$ if and only if $\eta(M) \geq k$.

2) By the definition of $F_k(M)$, $F_k(M) = f_k(M, \varnothing)$ if and only if $\forall X \subseteq E(M)$,

$$k(r(M) - r(X)) - |E - X| \leq kr(M) - |E|;$$

and so if and only if $\forall X \subseteq E(M)$ with $r(X) > 0$, $\dfrac{|X|}{r(X)} \leq k$. By the definition of $\gamma(M)$, this happens if and only if $\gamma(M) \leq k$.

3) By Theorem 1.3, $\gamma(M / J_{i(k)}) < k$. By 2) of this lemma, by Lemma 2.7, and by Theorem 1.4 1),

$$\begin{aligned} F_k\left(M/J_{i(k)}\right) &= f_k\left(M/J_{i(k)}, \varnothing\right) = kr\left(M/J_{i(k)}\right) - \left|M/J_{i(k)}\right| \\ &= k\left(r(M) - r\left(J_{i(k)}\right)\right) - \left(|E| - \left|J_{i(k)}\right|\right) = F(M,k). \end{aligned}$$

4) For any $e \in E(M)$, by the definition of $F_k(M)$ in (8), $F_k(M) \geq F_k(M / e)$. It follows by 3) of this lemma that $F_k(M) \geq f_k(M, X) = F(M,k)$.

5) By 4), and by the choice of X_0, we have

$$\begin{aligned} F_k(M) &\geq F_k(M/X_0) \geq f_k(M/X_0, \varnothing) \\ &= f_k(M, X_0) = F_k(M). \end{aligned}$$

Thus equalities must hold and so $F_k(M) = f_k(M / X_0) = F_k(M / X_0) = f_k(M / X_0, \varnothing)$. It follows by 2) that $\gamma(M / X_0) \leq k$. This proves 5).

Lemma 3.3 *Suppose that $X_0 \subseteq E(M)$ satisfies $f_k(M, X_0) = F_k(M)$. Then $\eta(M \mid X_0) \geq k$.*

Proof: By Lemma 3.1 1), it suffices to show that $F_k(M \mid X_0) = 0$. For any $Y \subseteq X_0$, as

$$\begin{aligned} &f_k(M \mid X_0, Y) = k\left(r(X_0) - r(Y)\right) - |X_0| + |Y|, \\ &\text{and } f_k(M, X_0) = k\left(r(M) - r(X_0)\right) - |E(M)| + |X_0|. \end{aligned}$$

It follows that $f_k(M \mid X_0, Y) + f_k(M, X_0) = f_k(M, Y) \leq F_k(M) = f_k(M, X_0)$. Thus by definition, $f_k(M \mid X_0, Y) \leq 0$. This implies that $F_k(M \mid X_0) = 0$, and so $\eta(M \mid X_0) \geq k$.

Proof of Theorem 1.4 2): By Lemma 3.2 4), it suffices to show that $F_k(M) \leq F(M,k)$. We shall argue by induction on $|E(M)|$ to proceed the proof.

Suppose first that $F_k(M) = 0$. Then by Lemma 3.2 1), $F_k(M) = 0$ if and only if $\eta(M) \geq k$. By Lemma 3.1 1), we have $F(M,k) = 0 = F_k(M)$ in this case. Thus we assume that $F_k(M) > 0$.

By Lemma 3.1 1), $F_k(M) > 0$ if and only if $\eta(M) < k$. If $\gamma(M) \leq k$, then by Lemma 3.1 2), and by Lemma 3.2 2),

$$F_k(M) = f_k(M, \varnothing) = kr(M) - |E(M)| = F(M,k).$$

Hence we may assume that Theorem 1.4 2) holds for smaller values of $|E(M)|$, and that

$$\eta(M) < k < \gamma(M). \tag{9}$$

By induction, we may assume that M does not have loops. By Theorem 1.3, and by (9), both $i(k)$, the smal-

lest j in (3) such that $l_j \geq k$, and $J_{i(k)}$, the corresponding set in (4), exist.

Let $X_0 \subset E(M)$ be a subset such that $F_k(M) = f_k(M, X_0)$. By (9), $X_0 \neq \varnothing$. Since $|E(M/X_0)| < |E(M)|$, by Lemma 3.2 5) and by induction, we have

$$F_k(M) = f_k(M/X_0) = F_k(M/X_0) = F(M/X_0, k),$$
$$\text{and } \gamma(M/X_0) \leq k.$$

Suppose that $F(M,k) = l$. Then there exists a matroid M' with $M' \in S_k$, which contains M as a restriction and satisfies $|E(M') - E(M)| = l$. Note that $X_0 \subseteq E(M) \subseteq E(M')$. Let $W = E(M') - E(M)$, and $W_0 = W - cl_{M'}(X_0)$. Then $|W_0| \leq |W|$.

Since $M' \in S_k$, it follows by Proposition 2.5 (C2) that $M'/X_0 \in S_k$. Since M is a restriction of M', M/X_0 is a restriction of M'/X_0. It follows by the definition of $F(M/X_0, k)$ and by (10) that

$$F_k(M) = F(M/X_0, k) \leq |E(M'/X_0) - E(M/X_0)|$$
$$\leq |W_0| \leq |W| = F(M,k).$$

This, together with Lemma 3.2 4), implies Theorem 1.4 2).

4. Applications

Let G be a graph, and $M = M(G)$ be the cycle matroid of G. Let $F(G,k) = F(M(G),k)$, and $f_k(G,X) = f_k(M(G),X)$, for any edge subset $X \subseteq E(G)$. Let $\omega(G)$ denote the number of connected components of G. The next theorem follows immediately from Theorem 1.4.

Theorem 4.1 (*Theorems* 3.4 *and* 3.10 *of* [4]) *For $k \in N$, let G be a connected graph with $\tau(M(G)) \leq k$ and let $i(k)$ denote the smallest i_j in* (3) *such that $i(k) \geq k$. Then*

1) $F(G,k) = k(|V(G)| - |V(G[J_{i(k)}])| + \omega(G[J_{i(k)}]) - 1) - |E(G) - J_{i(k)}|$.

2) $F(G,k) = \max_{X \subseteq E(G)}\{f_k(G,X)\}$.

The problem of reinforcing graphs to have k edge-disjoint spanning trees has also been investigated by others. In [3], the following is proved.

Theorem 4.2 (*Haas*, *Theorem* 1 *of* [3]) *The following are equivalent for a graph G, and integers $k > 0$ and $l > 0$.*

1) $|E(G)| = k(|V(G)| - 1) - l$ and for subgraphs H of G with at least 2 vertices, $|E(H)| \leq k(|V(H)| - 1)$.

2) There exists some l edges which when added to G result in a graph that can be decomposed into k spanning trees.

Proof: Assume that 1) holds. Then by 1), $\gamma(M(G)) \leq k$. It follows by the assumption that $|E(G)| = k(|V(G)| - 1) - l$ and by Lemma 3.1 2) that $F(G,k) = l$, and so 1) is obtained.

Assume 2) holds. Since adding l edges to G can result in a graph in S_k, by (1) and by (2), $\gamma(M(G)) \leq k$. By Lemma 3.1 2),

$$k(|V(G)| - 1) - |E(G)| = F(G,k) = l,$$

and so 2) must hold.

5. References

[1] H.-J. Lai, P. Li and Y. Liang, "Characterization of Removable Elements with Respect to Having *k* Disjoint Bases in a Matroid," Submitted.

[2] P. Li, Ph.D. Dissertation, West Virginia University, to be Completed in 2012.

[3] R. Haas, "Characterizations of Arboricity of Graphs," *Ars Combinatoria*, Vol. 63, 2002, pp. 129-137.

[4] D. Liu, H.-J. Lai and Z.-H. Chen, "Reinforcing the Number of Disjoint Spanning Trees," *Ars Combinatoria*, Vol. 93, 2009, pp. 113-127.

[5] D. J. A. Welsh, "Matroid Theory," Academic Press, London, New York, 1976.

[6] J. G. Oxley, "Matroid Theory," Oxford University Press, New York, 1992.

[7] J. A. Bondy and U. S. R. Murty, "Graph Theorym," Springer, New York, 2008.

[8] E. M. Palmer, "On the Spannig Tree Packing Number of a Graph, a Survey," *Discrete Mathematics*, Vol. 230, No. 1-3, 2001, pp. 13-21.

[9] C. St. J. A. Nash-Williams, "Edge-Disjoint Spanning Trees of Finite Graphs," *Journal of the London Mathematical Society*, Vol. 36, No. 1, 1961, pp. 445-450.

[10] W. T. Tutte, "On the Problem of Decomposing a Graph into *n* Connected Factors," *Journal of the London Mathematical Society*, Vol. 36, No. 1, 1961, pp. 221-230.

[11] J. Edmonds, "Lehman's Switching Game and a Theorem of Tutte and Nash-Williams," *Journal of Research of the National Bureau of Standards, Section B*, Vol. 69B, 1965, pp. 73-77.

[12] C. St. J. A. Nash-Williams, "Decomposition of Fininte Graphs into Forest," *Journal of the London Mathematical Society*, Vol. 39, No. 1, 1964, p. 12.

[13] W. H. Cunningham, "Optimal Attack and Reinforcement of a Network," *Journal of Associated Computer Machanism*, Vol. 32, 1985, pp. 549-561.

[14] P. A. Catlin, J. W. Grossman, A. M. Hobbs and H.-J. Lai, "Fractional Arboricity, Strength and Principal Partitions in Graphs and Matroids," *Discrete Applied Mathematics*, Vol. 40, No. 1, 1992, pp. 285-302.

[15] A. M. Hobbs, "Computing Edge-Toughness and Fractional Arboricity," *Contemporary Mathematics*, Vol. 89 1989, pp. 89-106.

[16] A. M. Hobbs, L. Kannan, H.-J. Lai and H. Y. Lai,

"Transforming a Graph into a 1-Balanced Graph," *Discrete Applied Mathematics*, Vol. 157, No. 1, 2009, pp. 300-308.

[17] A. M. Hobbs, L. Kannan, H.-J. Lai, H. Y. Lai and Q. W. Guo, "Balanced and 1-Balanced Graph Construction," *Discrete Applied Mathematics*, Accepted.

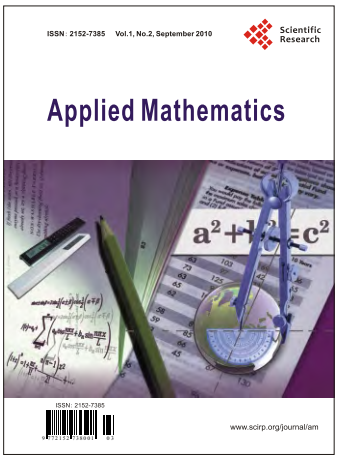
[18] P. A. Catlin, "Super-Eulerian Graphscollapsible Graphs, and Four-Cycles," *Congressus Numerantium*, Vol. 58, 1987, pp. 233-246.

[19] P. A. Catlin, Z. Han and H.-J. Lai, "Graphs without Spanning Closed Trails," *Discrete Mathematics*, Vol. 160, No. 1-3, 1996, pp. 81-91.

[20] P. A. Catlin, "Super-Eulerian Graphs - A Survey," *Journal of Graph Theory*, Vol. 16, No. 2, 1992, pp. 177-196.

[21] Z. H. Chen and H.-J. Lai, "Reduction Techniques for Super-Eulerian Graphs and Related Topics - A Survey," *Combinatorics and Graph Theory* 95, Vol. 1, World Science Publishing, River Edge, New York, 1995.

[22] J. Bruno and L. Weinberg, "The Principal Minors of a Matroid," *Linear Algebra and Its Applications*, Vol. 4, 1971, pp. 17-54.



Call for Papers

Applied Mathematics (AM)

ISSN 2152-7385 (Print) ISSN 2152-7393 (Online)
http://www.scirp.org/journal/am

***Applied Mathematics* (*AM*)** is an international journal dedicated to the latest advancement of applied mathematics. The goal of this journal is to provide a platform for scientists and academicians all over the world to promote, share, and discuss various new issues and developments in different areas.

Subject Coverage

This journal invites original research and review papers that address the following issues. Topics of interest include, but are not limited to:

- Approximation Theory
- Chaos Theory
- Combinatorics
- Complexity Theory
- Computability Theory
- Control Theory
- Cryptography
- Discrete Geometry
- Dynamical Systems
- Financial Mathematics
- Game Theory
- Graph Theory
- Information Theory
- Mathematical Biology
- Mathematical Chemistry
- Mathematical Economics
- Mathematical Physics
- Mathematical Psychology
- Mathematical Sociology
- Numerical Analysis
- Operations Research
- Optimization
- Probability Distribution
- Probability Theory
- Statistics
- Stochastic Processes
- Theoretical Computer Science

We are also interested in short papers (letters) that clearly address a specific problem, and short survey or position papers that sketch the results or problems on a specific topic. Authors of selected short papers would be invited to write a regular paper on the same topic for future issues of ***Applied Mathematics***.

Notes for Intending Authors

Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere. Paper submission will be handled electronically through the website. All papers are refereed through a peer review process. For more details about the submissions, please access the website.

Website and E-Mail

http://www.scirp.org/journal/am **E-mail: am@scirp.org**

Applied Mathematics, **2010, 1, 137-249**
Published Online September 2010 in SciRes (http://www.SciRP.org/journal/am/).

TABLE OF CONTENTS

Volume 1 Number 3 **September 2010**

