

## GENERATING THE INTEGER NULL SPACE AND CONDITIONS FOR DETERMINATION OF AN INTEGER BASIS USING THE ABS ALGORITHMS

H. Esmaeili, N. Mahdavi-Amiri and E. Spedicato

*\*Department of Mathematical Sciences, Sharif University of Technology,  
Tehran, Iran*

*\*\*Department of Mathematics, Statistics and Computer Science, University of  
Bergamo, Bergamo, Italy*

**Abstract:** We have presented a method, based on the **ABS** class of algorithms, for solving the linear systems of Diophantine equations. The method provides the general solution of the system by computing an integer solution along with an integer matrix (generally rank deficient), named as the Abaffian, the integer row combinations of which generate the integer null space of the coefficient matrix. Here we show that, in general, one can not expect that any full set of linearly independent rows of the Abaffian form an integer basis for the integer null space. We determine the necessary and sufficient conditions under which a full rank Abaffian would serve as an integer basis.

---

<sup>0</sup>2000 MSC: 15A03, 15A06, 65F05, 65F30

<sup>0</sup>Keywords: ABS algorithms, Diophantine equation, Integer null space.

## 1. Introduction

Suppose  $\mathbb{Z}$  represents all integers. Consider the Diophantine linear system of equations

$$Ax = b, \quad x \in \mathbb{Z}^n \quad (1)$$

where  $A \in \mathbb{Z}^{m \times n}$ ,  $b \in \mathbb{Z}^m$ , and  $m \leq n$ . By solving the system (1), firstly we mean the determination of the existence of the solution. Secondly, if the system has a solution, then we mean the computation of an integer solution  $\bar{x}$  and an integer matrix  $H$  so that the rows of  $H$ , not necessarily independent, generate the integer null space of  $A$ . That is,

$$\text{Integer Null}(A) = \text{Integer Range}(H^T).$$

Having this, the integer solutions for (1) are determined by

$$x = \bar{x} + H^T y,$$

for integer vectors  $y$ . If the dimension of null space of  $A$  is  $r$ , and

$$H = [h_1, \dots, h_i, \dots, h_r]^T,$$

with  $h_i$ 's being linearly independent, then  $H^T$  is said to be an integer basis matrix for the integer null space of  $A$ .

Several methods, based on computing the Hermite normal form, have been introduced before ([3,5]). Recently, **ABS** methods have been used extensively for solving general linear systems. In [7], we have presented a method, based on the **ABS** class of algorithms, for solving the system (1). These methods produce an integer solution  $\bar{x}$ , if it exists, and an integer matrix, named Abaffian, whose integer row combinations span the integer null space of the coefficient matrix  $A$ ; hence the general integer solution of the system is readily at hand.

Section 2 explains the class of **ABS** methods and provides some of its properties. In section 3, we briefly discuss our algorithm for solving the Diophantine equations (1). In this section, we then show that, in general, one can not expect that any full set of linearly independent

rows of the Abaffian matrix form a basis for the integer null space of the coefficient matrix. In section 4, we present necessary and sufficient conditions on the Abaffian for the existence and hence the determination of an integer basis.

## 2. ABS Algorithms

**ABS** methods have been developed by Abaffy, Broyden and Spedicato [1]. Consider the system of linear equations

$$Ax = b, \tag{2}$$

where  $A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^m$  and  $rank(A) = m$ . Let  $A = (a_1, \dots, a_m)^T$ ,  $a_i \in \mathbb{R}^n$ ,  $i = 1, \dots, m$  and  $b = (b_1, \dots, b_m)^T$ . Also let  $A_i = (a_1, \dots, a_i)$  and  $b^{(i)} = (b_1, \dots, b_i)^T$ .

Assume  $x_1 \in \mathbb{R}^n$  arbitrary and  $H_1 \in \mathbb{R}^{n \times n}$ , Spedicato's parameter, arbitrary and nonsingular. Note that for any  $x \in \mathbb{R}^n$  we can write  $x = x_1 + H_1^T q$  for some  $q \in \mathbb{R}^n$ .

The **ABS** class of methods are of the direct iteration types of methods for computing the general solution of (2). In the beginning of the  $i$ th iteration,  $i \geq 1$ , the general solution of the first  $i - 1$  equation is at hand. We realize that if  $x_i$  is a solution for the first  $i - 1$  equations and if  $H_i \in \mathbb{R}^{n \times n}$ , with  $rank(H_i) = n - i + 1$ , is so that the columns of  $H_i^T$  span the null space of  $A_{i-1}^T$ , then

$$x = x_i + H_i^T q,$$

with arbitrary  $q \in \mathbb{R}^n$ , forms the general solution of the first  $i - 1$  equations. That is, with

$$H_i A_{i-1} = 0,$$

we have

$$A_{i-1}^T x = b^{(i-1)}.$$

Now, since  $rank(H_i) = n - i + 1$  and  $H_i^T$  is a spanning matrix for  $null(A_{i-1}^T)$ , by assumption (one that is trivially valid for  $i = 1$ ), then if we let

$$p_i = H_i^T z_i,$$

with arbitrary  $z_i \in \mathbb{R}^n$ , Broyden's parameter, then  $A_{i-1}^T p_i = 0$  and

$$x(\alpha) = x_i - \alpha p_i,$$

for any scalar  $\alpha$ , solves the first  $i - 1$  equations. We can set  $\alpha = \alpha_i$  so that  $x_{i+1} = x(\alpha_i)$  solves the  $i$ th equation as well. If we let

$$\alpha_i = \frac{a_i^T x_i - b_i}{a_i^T p_i},$$

with assumption  $a_i^T p_i \neq 0$ , then

$$x_{i+1} = x_i - \alpha_i p_i$$

is a solution for the first  $i$  equations. Now, to complete the **ABS** step,  $H_i$  must be updated to  $H_{i+1}$  so that  $H_{i+1} A_i = 0$ . It will suffice to let

$$H_{i+1} = H_i - u_i v_i^T \tag{3}$$

and select  $u_i, v_i$  so that  $H_{i+1} a_j = 0, j = 1, \dots, i$ . The updating formula (3) for  $H_i$  is a rank-one correction to  $H_i$ . The matrix  $H_i$  is generally known as the Abaffian. The **ABS** methods usually use  $u_i = H_i a_i$  and  $v_i = H_i^T w_i / w_i^T H_i a_i$ , where  $w_i$ , Abaffy's parameter, is an arbitrary vector satisfying

$$w_i^T H_i a_i \neq 0.$$

Thus, the updating formula can be written as below:

$$H_{i+1} = H_i - \frac{H_i a_i w_i^T H_i}{w_i^T H_i a_i}.$$

We can now give the general steps of an **ABS** algorithm [1,2]. In the algorithm below,  $r_{i+1}$  denotes the rank of  $A_i$  and hence the rank of  $H_{i+1}$  equals  $n - r_{i+1}$ .

### **ABS Algorithm for Solving General Linear Systems**

- (1) Choose  $x_1 \in \mathbb{R}^n$ , arbitrary, and  $H_1 \in \mathbb{R}^{n \times n}$ , arbitrary and non-singular. Let  $i = 1, r_1 = 0$ .

- (2) Compute  $t_i = a_i^T x_i - b_i$  and  $s_i = H_i a_i$ .
- (3) If ( $s_i = 0$  and  $t_i = 0$ ) then let  $x_{i+1} = x_i$ ,  $H_{i+1} = H_i$ ,  $r_{i+1} = r_i$  and go to step (7) (the  $i$ th equation is redundant). If ( $s_i = 0$  and  $t_i \neq 0$ ) then Stop (the  $i$ th equation and hence the system is incompatible).
- (4)  $\{s_i \neq 0\}$  Compute the search direction  $p_i = H_i^T z_i$ , where  $z_i \in \mathbb{R}^n$  is an arbitrary vector satisfying  $z_i^T H_i a_i = z_i^T s_i \neq 0$ . Compute

$$\alpha_i = t_i / a_i^T p_i$$

and let

$$x_{i+1} = x_i - \alpha_i p_i.$$

- (5)  $\{\text{Updating } H_i\}$  Update  $H_i$  to  $H_{i+1}$  by

$$H_{i+1} = H_i - \frac{H_i a_i w_i^T H_i}{w_i^T H_i a_i}$$

where  $w_i \in \mathbb{R}^n$  is an arbitrary vector satisfying  $w_i^T s_i \neq 0$ .

- (6) Let  $r_{i+1} = r_i + 1$ .
- (7) If  $i = m$  then Stop ( $x_{m+1}$  is a solution) else let  $i = i + 1$  and go to step (2).

We note that after the completion of the algorithm, the general solution of (2), if compatible, is written as  $x = x_{m+1} + H_{m+1}^T q$ , where  $q \in \mathbb{R}^n$  is arbitrary.

Below, we list certain properties of the **ABS** methods [2]. For simplicity, we assume  $\text{rank}(A_i) = i$ .

- $H_i a_i \neq 0$  if and only if  $a_i$  is linearly independent of  $a_1, \dots, a_{i-1}$ .
- Every row of  $H_{i+1}$  corresponding to a nonzero component of  $w_i$  is linearly dependent on other rows.
- The direction searches  $p_1, \dots, p_i$  are linearly independent.

- If  $L_i = A_i^T P_i$ , where  $P_i = (p_1, \dots, p_i)$ , then  $L_i$  is a nonsingular lower triangular matrix.
- The set of directions  $p_1, \dots, p_i$  together with independent columns of  $H_{i+1}^T$  form a basis for  $\mathbb{R}^n$ .
- The matrix  $W_i = (w_1, \dots, w_i)$  has full column rank and  $\text{Null}(H_{i+1}^T) = \text{Range}(W_i)$ , while  $\text{Null}(H_{i+1}) = \text{Range}(A_i)$ .
- If rows  $j_1, \dots, j_i$  of  $W_i$  are linearly independent then the same rows of  $H_{i+1}$  are linearly dependent and vice versa. Specially, each row of  $H_{i+1}$  corresponding to a nonzero element of  $w_i$  is dependent.
- If  $s_i \neq 0$ , then  $\text{rank}(H_{i+1}) = \text{rank}(H_i) - 1$ .
- The updating formula  $H_i$  can be written as:

$$H_{i+1} = H_1 - H_1 A_i (W_i^T H_1 A_i)^{-1} W_i^T H_1,$$

where  $W_i^T H_1 A_i$  is strongly nonsingular (the determinants of all of its main principal submatrices are nonzero).

### 3. Solving Linear Diophantine Equations

Consider the linear Diophantine system of equations

$$Ax = b, \quad x \in \mathbb{Z}^n \tag{4}$$

where  $A \in \mathbb{Z}^{m \times n}$ ,  $b \in \mathbb{Z}^m$ . The following results indicate how to choose  $H_1$ ,  $z_i$  and  $w_i$  within the **ABS** algorithms to obtain the integer solution of (4); see [7]. Assume  $\delta_i$  to be the greatest common divisor (*gcd*) of the components of  $H_i a_i$ .

**Theorem 1.** *Let  $A$  be full rank and suppose that the Diophantine system (4) is solvable. Consider the sequence of Abaffians generated by the basic ABS algorithm with the following parameter choices:*

- (a)  $H_1$  is unimodular (an integer matrix whose inverse is also integer with the modules of its determinant equal to 1).

- (b) For  $i = 1, \dots, m$ , the integer vector  $w_i$  is such that  $w_i^T H_i a_i = \delta_i$ ,  $\delta_i = \gcd(H_i a_i)$ .

Then the following properties are true:

- (c) The sequence of Abaffians generated by the algorithm is well-defined and consists of integer matrices.
- (d) If  $x_{i+1}$  is a special integer solution of the first  $i$  equations, then any integer solution  $x$  of the first  $i$  equations can be written in the form  $x = x_{i+1} + H_{i+1}^T q$  for some integer vector  $q$ .

**Theorem 2.** Let  $A$  be full rank and consider the sequence of matrices  $H_i$  generated by the basic ABS algorithm with parameter choices as in Theorem 1. Let the initial point  $x_1$  in the basic ABS algorithm be an arbitrary integer vector and let  $z_i$  be chosen such that  $z_i^T H_i a_i = \gcd(H_i a_i)$ . Then system (4) has integer solutions if and only if  $\gcd(H_i a_i)$  divides  $a_i^T x_i - b_i$  for  $i = 1, \dots, m$ .

**Note:** The computation of  $\delta_i$  and solving for an integer  $y$  in  $s_i^T y = \delta_i$ , where  $s_i = H_i a_i$ , can be achieved by Rosser's algorithm [9,10].

It follows from the above theorems that if there exists a solution for the system (4), then  $x = x_{m+1} + H_{m+1}^T q$ , with arbitrary  $q \in \mathbb{Z}^n$ , forms the general solution of (4). We continue by an analysis showing that, in general, any  $n-m$  independent columns of  $H_{m+1}^T$  would not be an integer basis for the integer null space of the matrix  $A$ . For simplicity, let  $x_1 = 0$ ,  $\bar{x} = x_{m+1}$ ,  $H = H_{m+1}$  and assume  $\text{rank}(A) = m$ . Let  $\bar{H} \in \mathbb{Z}^{(n-m) \times m}$  be a matrix composed of any set of  $n-m$  linearly independent rows of  $H$ . We show that, in general, it can not be expected that

$$x = \bar{x} + \bar{H}^T y, \quad y \in \mathbb{Z}^{n-m}, \quad (5)$$

provide all the integer solutions for (4).

Let  $P = (p_1, \dots, p_m)$  be the matrix of search directions obtained from the application of an **ABS** algorithm in solving the system (4). Let  $K = (P, \bar{H}^T)$ . We know that the matrix  $K$  is nonsingular and

$$AK = A(P, \bar{H}^T) = (AP, A\bar{H}^T) = (L, 0),$$

where  $L$  is lower triangular and nonsingular. The next theorem states conditions under which  $x = \bar{x} + \bar{H}^T y$ ,  $y \in \mathbb{Z}^{n-m}$ , forms the general solution of (4). We note that since  $x_1 = 0$  then we can write  $\bar{x} = Pq$  for some vector  $q$ . Hence we have  $b = A\bar{x} = APq = Lq$ . Since  $L$  is nonsingular then  $q = L^{-1}b$  and  $\bar{x} = PL^{-1}b$ .

**Theorem 3.** *The expression  $x = \bar{x} + \bar{H}^T y$  is the general solution for the Diophantine system (4) when the matrix  $K = (P, \bar{H}^T)$  is unimodular.*

**Proof:** The vector  $x = \bar{x} + \bar{H}^T y$  for any integer vector  $y$  is integer. For such  $x$  we have  $Ax = b$ , since  $A\bar{H}^T = 0$ . Now suppose  $x \in \mathbb{Z}^n$  satisfies  $Ax = b$ . Let  $K^{-1}x = u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ . Since  $K$  is unimodular then  $u$  is an integer vector with  $u_1 \in \mathbb{Z}^m$  and  $u_2 \in \mathbb{Z}^{n-m}$ . Now, we can write

$$b = Ax = AKK^{-1}x = AKu = (L, 0) \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = Lu_1, \quad u_1 = L^{-1}b.$$

Hence

$$x = Ku = (P, \bar{H}^T) \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = Pu_1 + \bar{H}^T u_2 = PL^{-1}b + \bar{H}^T u_2 = \bar{x} + \bar{H}^T u_2. \quad \square$$

**Note:** Using the geometry of numbers, the converse of the above theorem is also established (see [4]).

Consider the single Diophantine equation  $a^T x = 0$ ,  $x \in \mathbb{Z}^n$ . Assume  $H_1$  is unimodular. Let  $\delta = \gcd(s)$ , where  $s = H_1 a$ , and assume  $z$  is so that  $s^T z = \delta$ . We know from Rosser's algorithm that the first component of  $s$  has the largest magnitude in  $s$  and is nonzero, since  $s \neq 0$ . Thus  $\|s\|_\infty = |s^T e_1| \neq 0$ , where  $e_1$  is the first column of the identity matrix. Suppose  $p = H_1^T z$  is the search direction of the **ABS** method for solving  $a^T x = 0$ , and  $w$  is an integer vector so that  $w^T e_1 \neq 0$  and  $s^T w = \delta$ . Then, from the **ABS** properties, the first row of  $H_2$  is dependent and we can define  $\bar{H}$  to represent the independent rows of  $H_2$  by

$$\bar{H} = E \left( H_1 - \frac{H_1 a w^T H_1}{w^T H_1 a} \right) = E \left( I - \frac{s w^T}{\delta} \right) H_1,$$

where  $E$  is the identity matrix with its first row deleted. From Theorem 3, the vector  $x = \bar{H}^T y$ , where  $y$  is an arbitrary integer vector, is the general solution for  $a^T x = 0$ ,  $x \in \mathbb{Z}^n$ , when  $M = (p, \bar{H}^T)$  is unimodular. Since  $H_1$  is unimodular, then

$$M = (p, \bar{H}^T) = \left( H_1^T z, H_1^T \left( I - \frac{ws^T}{\delta} \right) E^T \right) = H_1^T \left( z, \left( I - \frac{ws^T}{\delta} \right) E^T \right)$$

is unimodular if and only if the matrix

$$\left( z, \left( I - \frac{ws^T}{\delta} \right) E^T \right)$$

is unimodular. Let

$$K = (z, B),$$

where

$$B = \left( I - \frac{ws^T}{\delta} \right) E^T.$$

We note that  $K$  is nonsingular. We shall make use of the determinant of  $K$  in subsequent discussions. Note the following lemma.

**Lemma 1.** *If  $K = (z, B)$ , where  $B = (I - ws^T/\delta)E^T$  and  $E$  is obtained from the identity matrix with its first row deleted, then  $\det K = w^T e_1$ .*

**Proof:** The matrix  $K = (z, B)$  is a rank one correction to the matrix  $\bar{K} = I - ws^T/\delta = (\bar{K}e_1, B)$ , since

$$K = \bar{K} + (z - \bar{K}e_1)e_1^T.$$

Note that

$$\begin{aligned} K &= I - e_1 e_1^T - \frac{ws^T}{\delta} + \frac{e_1^T s}{\delta} w e_1^T + z e_1^T \\ &= K_1 + (z - e_1) e_1^T, \end{aligned}$$

where

$$K_1 = I + w v^T,$$

and

$$v^T = \frac{1}{\delta}[(e_1^T s)e_1^T - s^T].$$

Hence

$$\det K_1 = 1 + v^T w = \frac{(e_1^T s)(e_1^T w)}{\delta} \neq 0.$$

Thus  $K_1$  is always properly defined and always nonsingular and we may write  $K = K_1 K_2$ , where

$$K_2 = I + K_1^{-1}(z - e_1)e_1^T.$$

Therefore

$$\det K_2 = 1 + e_1^T K_1^{-1}(z - e_1).$$

Now, expanding  $K_1^{-1}$  by the Sherman-Morrison-Woodbury formula [8] gives  $\det K_2 = \delta / (e_1^T s)$ . Since  $\det K = \det K_1 \det K_2$ , the result follows.  $\square$

Therefore,  $K$  and hence  $M$  are unimodular if and only if the integer vector  $w$  satisfies  $w^T e_1 = 1$  and  $w^T s = \delta$ , conditions not expected to hold in general. Egervary's method [6] is a special **ABS** method with the selections  $H_1 = I$ ,  $x_1 = 0$  and  $w = z$ . Since the Diophantine system  $s^T z = \delta$ ,  $z^T e_1 = 1$ , lacks integer solutions in general, then Egervary's claim that any set of independent columns of  $H^T$  provides an integer basis for the general integer solutions is refuted. The next example validates this statement.

**Example 1.** If we use Egervary's method for solving the homogeneous Diophantine equation

$$a^T x = 0 \quad , \quad a^T = (1, 1, 1), \quad (6)$$

with  $z = (2, 2, -3)^T$ , we obtain:

$$H = H_2^T = I - za^T = \begin{bmatrix} -1 & -2 & -2 \\ -2 & -1 & -2 \\ 3 & 3 & 4 \end{bmatrix}.$$

There are three possible choices for  $\bar{H}^T$ .

- (a)  $\bar{H}^T = (H_2^T e_1, H_2^T e_2)$ . The vector  $\bar{x} = (-2, 1, 1)^T$  satisfies (6).  
The only solution for  $\bar{H}^T t = \bar{x}$  is  $t = (-4/3, 5/3)^T$ .
- (b)  $\bar{H}^T = (H_2^T e_1, H_2^T e_3)$ . The vector  $\bar{x} = (-2, 1, 1)^T$  satisfies (6).  
The only solution for  $\bar{H}^T t = \bar{x}$  is  $t = (-3, 5/2)^T$ .
- (c)  $\bar{H}^T = (H_2^T e_2, H_2^T e_3)$ . The vector  $\bar{x} = (-3, 2, 1)^T$  satisfies (6).  
The only solution for  $\bar{H}^T t = \bar{x}$  is  $t = (5, -7/2)^T$ .

We see that in all the possible three cases there is at least one integer solution  $\bar{x}$  for (6) not being generated by an integer combinations of columns of  $\bar{H}^T$ .

**Note:** For the homogeneous Diophantine system (case  $b = 0$  in (4)), Egervary [6] presented a method being now a special version of the **ABS** algorithms with  $H_1 = I$ ,  $x_1 = 0$ ,  $z_i = w_i$  for all  $i$ . We realize that the general solution in this case is written as  $x = H_{m+1}^T y$ , where  $y \in \mathbb{Z}^n$  is arbitrary. Egervary believed that with  $r$  being the rank of  $A$ , any set of  $n - r$  independent columns of  $H_{m+1}^T$  would form an integer basis for the integer solutions of the system. The results given above clearly invalidates this belief (see also [7]).

In the next section, we introduce the necessary and sufficient conditions for producing an integer basis from the Abaffian matrix. There, we return to Example 1 again and show how to determine an integer basis using these conditions.

#### 4. The Necessary and Sufficient Conditions

Assume  $rank(A) = m$ . We now determine conditions under which one can eliminate  $m$  columns of  $H_{m+1}^T$  and obtain an integer basis, composed of  $n - m$  linearly independent columns, for  $Null(A) \cap \mathbb{Z}^n$ . For convenience, let  $H = H_{m+1}$ . Let  $W = (w_1, \dots, w_m) \in \mathbb{Z}^{n \times m}$  be the matrix with Abaffian parameters as its columns. We know that  $rank(W) = m$ . According to **ABS** properties, the rows of  $H$  corresponding to  $m$  linearly independent rows of  $W$  are linearly dependent. Since  $rank(H) = n - m$

then, without loss of generality, we can write  $H = \begin{pmatrix} \bar{H} \\ U\bar{H} \end{pmatrix}$ , where  $\bar{H} \in \mathbb{Z}^{(n-m) \times n}$  corresponds to the  $n - m$  linearly independent rows of  $H$  and  $U \in \mathbb{R}^{m \times (n-m)}$ . We can now let  $W^T = (V^T, T^T)$ , where  $T \in \mathbb{Z}^{m \times m}$  is nonsingular. Since

$$0 = H^T W = \bar{H}^T V + \bar{H}^T U^T T$$

then  $\bar{H}^T U^T = -\bar{H}^T V T^{-1}$  and whereof

$$U^T = -V T^{-1}.$$

We emphasize that  $U$  is not necessarily an integer matrix. Fix an arbitrary vector  $y \in \text{Null}(A) \cap \mathbb{Z}^n$ . The full column rank system

$$\bar{H}^T t = y \tag{7}$$

has a unique solution. The following lemma gives the correspondence between  $t$ , the unique solution of (7), and the solutions of the system

$$H^T x = y. \tag{8}$$

**Lemma 2.**  *$x$  is a solution of (8) if and only if we have*

$$x = \begin{pmatrix} t \\ 0 \end{pmatrix} + \begin{pmatrix} -U^T \\ I_m \end{pmatrix} q,$$

with  $t$  being the unique solution of (7) and  $q \in \mathbb{R}^m$ .

**Proof:** Let  $x = \begin{pmatrix} x_{n-m} \\ x_m \end{pmatrix}$  be a solution of (8). Then

$$y = H^T x = \bar{H}^T x_{n-m} + \bar{H}^T U^T x_m = \bar{H}^T (x_{n-m} + U^T x_m).$$

Since (7) has a unique solution then  $t = x_{n-m} + U^T x_m$  and hence

$$x = \begin{pmatrix} t \\ 0 \end{pmatrix} + \begin{pmatrix} -U^T \\ I_m \end{pmatrix} x_m.$$

Conversely, let  $t$  be the unique solution of (7) and  $q \in \mathbb{R}^m$ . Consider

$$x = \begin{pmatrix} t \\ 0 \end{pmatrix} + \begin{pmatrix} -U^T \\ I_m \end{pmatrix} q.$$

We have

$$H^T x = \bar{H}^T t - \bar{H}^T U^T q + \bar{H}^T U^T q = \bar{H}^T t = y.$$

Therefore,  $x$  is a solution of (8).  $\square$

We saw before that for any  $y \in \text{Null}(A) \cap \mathbb{Z}^n$ , the integer vector  $x = H_1^{-T} y$  solves (8). Let  $H_1^{-1} = (H_{11}^T, H_{21}^T)$ .  $H_1$  being unimodular, both  $H_{11}$  and  $H_{21}$  are integer matrices. Applying Lemma 2, for some  $q \in \mathbb{R}^m$  and  $t$ , the unique solution of (7), we must have:

$$x = H_1^{-T} y = \begin{pmatrix} H_{11} \\ H_{21} \end{pmatrix} y = \begin{pmatrix} t \\ 0 \end{pmatrix} + \begin{pmatrix} -U^T \\ I_m \end{pmatrix} q.$$

Hence we have:

$$H_{11} y = t - U^T q$$

$$H_{21} y = q.$$

Now, for  $x = H_1^{-T} y$  since  $y$  is an integer vector then both  $q$  and  $t - U^T q$  must be integer vectors. Therefore, to have  $t$  integer it would suffice that  $\bar{H}^T$  be constructed from  $H^T$  in such a way that the corresponding matrix  $U$  be integer (or can be reduced to an integer matrix). On the other hand, we realize that no column of  $\bar{H}^T$  should have a common divisor other than one (that is, the greatest common divisor for every column should be one), since the system  $\bar{H}^T t = y$  will have noninteger solutions, otherwise. Having this in mind, we consider reducing the matrix  $H^T = (\bar{H}^T, \bar{H}^T U^T)$  accordingly. To make the columns of  $\bar{H}^T$  be relatively prime, we multiply  $H^T$  by  $D$  on the right, where  $D$  is a diagonal matrix as below

$$D = \begin{pmatrix} \bar{D} & 0 \\ 0 & I_m \end{pmatrix},$$

with  $\bar{D}_{ii} = 1/\gcd(H^T e_i)$ . Therefore, we have

$$\tilde{H}^T = H^T D = (\hat{H}^T, \hat{H}^T \hat{U}^T),$$

where

$$\hat{H}^T = \bar{H}^T \bar{D}, \quad \hat{U}^T = \bar{D}^{-1} U^T.$$

Now, let  $\text{adj}(T)$  be the classical adjoint of  $T$  (that is,  $T^{-1} = \text{adj}(T)/\det T$ ). Since  $U^T = -VT^{-1}$ , we can write

$$\hat{U}^T = -\bar{D}^{-1} V T^{-1} = -\bar{D}^{-1} V \text{adj}(T)/\det T.$$

The following theorem states the necessary and sufficient conditions for the solution of the system  $\hat{H}^T t = \bar{H}^T \bar{D} t = y$  to be integer.

**Theorem 4.** *Let  $y \in \text{Null}(A) \cap \mathbb{Z}^n$  be arbitrary. The solution  $\hat{t}$  for the full column rank system  $\hat{H}^T t = y$  is an integer vector if and only if  $\det T \mid \bar{D}^{-1} V \text{adj}(T)$ . ( $a \mid b$  means  $a$  divided by  $b$  is an integer.)*

**Proof:** We saw that  $x = H_1^{-T} y \in \mathbb{Z}^n$ , for any  $y \in \text{Null}(A) \cap \mathbb{Z}^n$ , satisfies  $H^T x = y$ . Thus, for  $\tilde{x} = D^{-1} H_1^{-T} y \in \mathbb{Z}^n$  we have  $\tilde{H}^T \tilde{x} = y$ . Let  $\tilde{x} = (\tilde{x}_{n-m}^T, \tilde{x}_m^T)^T$  and suppose that  $\det T \mid \bar{D}^{-1} V \text{adj}(T)$ . Then  $\hat{U}$  is an integer matrix and

$$\tilde{x} = \begin{pmatrix} \tilde{x}_{n-m} \\ \tilde{x}_m \end{pmatrix} = D^{-1} H_1^{-T} y = \begin{pmatrix} \bar{D}^{-1} & 0 \\ 0 & I_m \end{pmatrix} \begin{pmatrix} H_{11} y \\ H_{21} y \end{pmatrix} = \begin{pmatrix} \bar{D}^{-1} H_{11} y \\ H_{21} y \end{pmatrix}.$$

Thus, the vector

$$\hat{t} = \tilde{x}_{n-m} + \hat{U}^T \tilde{x}_m = \bar{D}^{-1} H_{11} y + \bar{D}^{-1} U^T H_{21} y = \bar{D}^{-1} (H_{11} y + U^T H_{21} y)$$

is integer and

$$\begin{aligned} \hat{H}^T \hat{t} &= \bar{H}^T \bar{D} \bar{D}^{-1} (H_{11} y + U^T H_{21} y) = \bar{H}^T (I_{n-m}, U^T) \begin{pmatrix} H_{11} y \\ H_{21} y \end{pmatrix} \\ &= (\bar{H}^T, \bar{H}^T U^T) H_1^{-T} y = H^T x = y. \end{aligned}$$

Conversely, suppose that for any  $y \in \text{Null}(A) \cap \mathbb{Z}^n$ , the solution  $\hat{t}$  for  $\hat{H}^T t = y$  be integer. Applying Lemma 2, the integer solutions for  $\tilde{H}^T x = y$  can be written as

$$\tilde{x} = \begin{pmatrix} \hat{t} \\ 0 \end{pmatrix} + \begin{pmatrix} -\hat{U}^T \\ I_m \end{pmatrix} q,$$

where  $q \in \mathbb{Z}^m$ . Since  $\hat{H}^T \hat{x} = y$ , then

$$\begin{aligned} \hat{t} &= \tilde{x}_{n-m} + \hat{U}^T \tilde{x}_m = \bar{D}^{-1} H_{11} y + \bar{D}^{-1} U^T H_{21} y \\ &= (\bar{D}^{-1}, \bar{D}^{-1} U^T) \begin{pmatrix} H_{11} y \\ H_{21} y \end{pmatrix} = (\bar{D}^{-1}, \bar{D}^{-1} U^T) H_1^{-T} y. \end{aligned}$$

Note that, for any  $y \in \text{Null}(A) \cap \mathbb{Z}^n$ ,  $H_1^{-T} y$ ,  $\hat{t}$  and  $D^{-1}$  are integers. Therefore,  $\bar{D}^{-1} U^T$  must also be an integer matrix because the rows of  $H_{21}$  are relatively prime. From  $\bar{D}^{-1} U^T = -\bar{D}^{-1} \text{Vadj}(T) / \det T$ , it follows that  $\det T \mid \bar{D}^{-1} \text{Vadj}(T)$ .  $\square$

We now return to case (b) in Example 1. We have,

$$\begin{aligned} \bar{H}^T &= \begin{bmatrix} -1 & -2 \\ -2 & -2 \\ 3 & 4 \end{bmatrix}, \quad W = \begin{pmatrix} 2 \\ 2 \\ -3 \end{pmatrix}, \quad T = (2), \\ V &= \begin{pmatrix} 2 \\ -3 \end{pmatrix}, \quad U = \frac{-1}{2}(2, -3), \quad \bar{D} = \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}. \end{aligned}$$

We see that

$$\hat{H}^T = \begin{bmatrix} -1 & -1 \\ -2 & -1 \\ 3 & 2 \end{bmatrix},$$

and

$$\hat{U} = (-1, 3),$$

an integer vector now. The solution for  $\hat{H}^T t = y$ , with  $y = (-2, 1, 1)^T$ , is the integer vector  $\hat{t} = (-3, 5)^T$ . On the other hand, any  $y \in \text{Null}(a^T) \cap \mathbb{Z}^3$  can be written as  $y = (-\alpha - \beta, \alpha, \beta)^T$ , where  $\alpha$  and  $\beta$  are arbitrary integers. For any such  $y$ , the solution for  $\hat{H}^T t = y$  is given by  $\hat{t} = (-2\alpha - \beta, 3\alpha + 2\beta)^T$ . Therefore, in this case, we have

$$\{x \in \mathbb{Z}^3 \mid a^T x = 0\} = \{\hat{H}^T q \mid q \in \mathbb{Z}^2\}.$$

Similar developments for case (c) will also result in an integer basis for  $\text{Null}(a^T) \cap \mathbb{Z}^3$ . At the same time, we note that no integer basis can be

obtained from the matrix in case (a). Therefore, we observe that an integer basis can not necessarily be obtained from any set of linearly independent columns of  $H^T$ .

### Determining an Integer Basis

Considering the **ABS** properties, instead of deleting the  $m$  dependent columns of  $H^T$  all at the same time, deletions can be made in steps. From the **ABS** properties, at the end of the  $i$ -th iteration, an independent column of  $H_{i+1}^T$  can be identified and subsequently deleted. We know that any column of  $H_{i+1}^T$  corresponding to a nonzero component of  $w_i$  is linearly dependent on the other columns. Let  $w_i = (w_{1i}, \dots, w_{ni})^T$  and suppose  $w_{ki} \neq 0$ , for some  $k$ ,  $1 \leq k \leq n$ . Then  $T = w_{ki}$ ,  $V = (w_{1i}, \dots, w_{k-1,i}, w_{k+1,i}, \dots, w_{ni})^T$  and  $U^T = -V/w_{ki}$ . We can now state the following rule for the deletion of a dependent column of  $H_{i+1}^T$  at the end of the  $i$ -th iteration.

### Deletion Rule For a Dependent Column

Let  $\delta_j = \gcd(H_{i+1}^T e_j)$  for all  $j$ . Delete the  $k$ -th column of  $H_{i+1}^T$ , where  $w_i^T e_k \neq 0$  and  $w_i^T e_k | \delta_j w_i^T e_j$  for all  $j$ .

We note that one can not expect the satisfaction of the above conditions in all cases. Thus, the **ABS** approach may signal the failure by recognizing that an integer basis may not be obtained. Nevertheless, the columns of  $H^T$  span  $\text{Null}(A) \cap \mathbb{Z}^n$  and, as such, the general integer solutions may be obtained using  $H$ . The following example illustrates the point.

**Example 2.** Consider the Diophantine system below:

$$a^T x = 0 \quad , \quad a^T = (1, 3, -2). \quad (9)$$

With the choice  $z = (2, 3, 5)^T$ , we have:

$$H^T = I - za^T = \begin{bmatrix} -1 & -6 & 4 \\ -3 & -8 & 6 \\ -5 & -15 & 11 \end{bmatrix}.$$

We see that every column of  $H^T$  is relatively prime. Consider the system  $\bar{H}^T t = y$ , where  $y = (-1, 1, 1)^T$  is a solution of (9).

(a) By selecting  $\bar{H}^T = (H^T e_1, H^T e_2)$ , we have  $t = (-7/5, 2/5)^T$ .

(b) By selecting  $\bar{H}^T = (H^T e_1, H^T e_3)$ , we have  $t = (-5/3, -2/3)^T$ .

(c) By selecting  $\bar{H}^T = (H^T e_2, H^T e_3)$ , we have  $t = (5/2, 7/2)^T$ .

We see that in all the possible three cases there is at least one integer solution for (9) not being generated by an integer combinations of columns of  $\bar{H}^T$ .

## 5. Conclusions

We saw how an integer Abaffian (not necessarily full rank) matrix is obtained by use of the **ABS** methods for solving a linear Diophantine system of equations. The integer combinations of the rows of the Abaffian span the integer null space of the coefficient matrix. We proved that, in general, it can not be expected that the resulting Abaffian would contain an integer basis for this integer null space. Finally, we specified the necessary and sufficient conditions under which the Abaffian would present an integer basis.

***Acknowledgements.** The work of the first two authors has been supported by the Research Council of Sharif University of Technology. The authors are grateful to the referees for their constructive comments, specially to one anonymous referee who provided a more concise proof of Lemma 1.*

## References

- [1] J. Abaffy, C.G. Broyden, E. Spedicato, A class of direct methods for linear equations, *Numer. Math.*, 45 (1984) 361-376.

- [2] J. Abaffy, E. Spedicato, *ABS Projection Algorithms: Mathematical Techniques for Linear and Nonlinear Equations*, Ellis Horwood, Chichester, 1989.
- [3] G.H. Bradley, Algorithms for Hermite and Smith normal matrices and linear Diophantine equations, *Math. Comp.*, 25 (1971) 897-907.
- [4] J.W.S. Cassels, *An Introduction to the Geometry of Numbers*, Springer, Berlin, 1959.
- [5] T.J. Chou, E.E. Collins, Algorithms for the solution of systems of linear Diophantine equations, *SIAM J. Comp.*, 11 (1982) 786-708.
- [6] E. Egervary, On rank-diminishing operations and their applications to the solution of linear equations, *ZAMP*, 9 (1960) 376-386.
- [7] H. Esmaeili, N. Mahdavi-Amiri, E. Spedicato, A class of ABS algorithms for Diophantine linear systems, to appear in *Numerische Mathematik*.
- [8] G.H. Golub, C.F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 1989.
- [9] S. Morito, H.M. Salkin, Using the Blankinship algorithm to find the general solution of a linear Diophantine equation, *Acta Inf.*, 13 (1980) 379-382.
- [10] J.B. Rosser, A note on the linear Diophantine equation, *Amer. Math. Monthly*, 48 (1941) 662-666.

### تولید فضای پوچ صحیح و شرایط لازم و کافی برای تعیین یک پایه صحیح بر اساس روشهای $ABS$

بر اساس رده روشهای  $ABS$ ، روشی برای حل دستگاه معادلات دیوفانتی خطی ارائه کرده‌ایم. این روش جواب عمومی دستگاه را با ایجاد یک جواب صحیح برای دستگاه و یک ماتریس صحیح رتبه ناقص (ماتریس ابافی) که ترکیبهای صحیح سطرهايش فضای پوچ صحیح ماتریس ضرایب را تولید می‌کنند، بدست می‌دهد. در این مقاله ابتدا نشان می‌دهیم که در حالت کلی نمی‌توان انتظار داشت که هر مجموعه کامل مستقل خطی از سطرهاى ماتریس ابافی یک پایه صحیح برای فضای پوچ صحیح ماتریس ضرایب تشکیل دهد. شرایطی لازم و کافی را تعیین می‌کنیم که تحت آنها می‌توان از سطرهاى ماتریس ابافی یک ماتریس پایه صحیح برای فضای پوچ صحیح تشکیل داد.

## COMMON FIXED POINT THEOREMS USING GENERALIZED CONTRACTIVE CONDITIONS

M. Akkouchi, A. Bounabat and A. Hadik

*Département de Mathématiques, Faculté des Sciences-Semlalia. Université  
Cadi Ayyad, Avenue du prince My. Abdellah. B.P. 2390, Morocco  
makkouchi@caramail.com*

**Abstract:** The aim of this note is to prove some common fixed point theorems in complete metric spaces using generalized contractive conditions.

### 1. Introduction

In the paper [4], the authors gave some fixed point theorems generalizing and unifying many fixed point theorems obtained by Delbosco in [1], Skof in [8], Rakotch in [5], Reich in [7], and Fisher in [3], (see also the references for other related results). Precisely in [4] the following theorem was established:

**1.1 Theorem.** *Let  $T$  be a self-map of a complete metric space  $(X, d)$  and let  $\phi$  be a function verifying*

- (i)  $\phi : [0, \infty[ \rightarrow [0, \infty[$  is continuous and increasing in  $[0, \infty[$ , and*
- (ii)  $\phi(t) = 0 \iff t = 0$ .*

---

<sup>o</sup>2000 MSC: 47H10, 54H25.

<sup>o</sup>Keywords: Common fixed point theorem, generalized contractive condition.

We suppose that  $T$  satisfies the following condition:

$$\begin{aligned} \phi(d(Tx, Ty)) \leq & a(d(x, y))\phi(d(x, y)) + \\ & b(d(x, y))[\phi(d(x, Tx)) + \phi(d(y, Ty))] \\ & + c(d(x, y)) \min \{ \phi(d(x, Ty)), \phi(d(y, Tx)) \} \quad (K) \\ & \forall x, y \in X \quad \text{with } x \neq y, \end{aligned}$$

where  $a, b, c$  are three decreasing functions from  $]0, \infty[$  into  $[0, 1[$  such that  $a(t) + 2b(t) + c(t) < 1$ , for every  $t > 0$ . Then  $T$  has a unique fixed point.

In the second section of this paper we shall prove some common fixed point theorems for sets of self-mappings verifying contractive conditions close to the relation (K). These results complete and unify the main results obtained in the papers [4] and [6]. In section 3, we establish a common fixed point theorem in compact metric spaces using another type of contractive conditions. We may consider this theorem as a generalization of a theorem established by B. Fisher in [3]. Our generalization is different from that one given in the paper [4] by M.S. Khan, M. Swaleh and S. Sessa.

## 2. Main theorems

We shall denote by  $\Phi$  the set of functions  $\phi$  verifying conditions (i) and (ii). Many authors (see the references) were interested by fixed point theorems by altering the distances between the points with the use of functions belonging to the class  $\Phi$ . The purpose of this section is to contribute in this field of investigations. One of the main results of this paper is the following theorem.

**2.1 Theorem.** *Let  $\phi \in \Phi$  be a convex function and let  $S, T$  be two self-maps of a complete metric space  $(X, d)$  such that*

$$\begin{aligned} \phi(d(Sx, Ty)) \leq & a(d(x, y))\phi(d(x, y)) + b(d(x, y))\phi(d(x, Sx)) \\ & + c(d(x, y))\phi(d(y, Ty)) + e(d(x, y))\phi(\gamma d(x, Ty)) \\ & + f(d(x, y))\phi(\delta d(Sx, y)), \quad (A) \end{aligned}$$

for all distinct  $x, y$  in  $X$ , where  $\gamma, \delta$  are two fixed numbers such that  $0 \leq \gamma, \delta \leq \frac{1}{2}$ , and  $a, b, c, e, f$  are five decreasing functions from  $]0, \infty[$  into  $[0, 1[$  verifying  $a(t) + b(t) + c(t) + e(t) + f(t) < 1$ , for every  $t > 0$ . We suppose also that  $\max\{B, C\} < 1$ , where,  $B := \sup\{b(t) + \delta f(t) : t > 0\}$  and  $C := \sup\{c(t) + \gamma e(t) : t > 0\}$ . Then  $S$  and  $T$  have a unique common fixed point  $z \in X$ . Moreover  $\text{Fix}(S) = \text{Fix}(T) = \{z\}$ .

**Proof.** (I) We shall prove that the pair  $\{S, T\}$  has a common fixed point. Let  $x_0$  be some point in  $X$ , and define the sequence  $\{x_n\}$  by

$$\begin{aligned} x_{2n} &= Sx_{2n-1}, & n = 1, 2, \dots \\ x_{2n+1} &= Tx_{2n}, & n = 0, 1, 2, \dots \end{aligned}$$

We put  $t_n := d(x_n, x_{n+1})$  for all integer  $n$ . (I) is proved if  $t_{n_0} = 0$  for some integer  $n_0$ . Therefore, we may assume that  $t_n > 0$  for all integer  $n$ . We see that for an even integer  $n$ , we have

$$\phi(t_n) = \phi(d(Sx_{n-1}, Tx_n)) \leq \Psi_1 + \Psi_2 + \Psi_3 + \Psi_4 + \Psi_5,$$

where

$$\begin{aligned} \Psi_1 &= a(d(x_{n-1}, x_n))\phi(d(x_{n-1}, x_n)), \\ \Psi_2 &= b(d(x_{n-1}, x_n))\phi(d(x_{n-1}, Sx_{n-1})), \\ \Psi_3 &= c(d(x_{n-1}, x_n))\phi(d(x_n, Tx_n)), \\ \Psi_4 &= e(d(x_{n-1}, x_n))\phi(\gamma d(x_{n-1}, Tx_n)), \\ \Psi_5 &= f(d(x_{n-1}, x_n))(\delta d(x_n, Sx_{n-1})) \end{aligned}$$

So

$$\phi(t_n) \leq a(t_{n-1})\phi(t_{n-1}) + b(t_{n-1})\phi(t_{n-1}) + c(t_{n-1})\phi(t_n) + e(t_{n-1})\phi(\gamma[t_{n-1} + t_n]).$$

Hence by using the convexity of  $\phi$ , we get

$$\phi(t_n) \leq \frac{a(t_{n-1}) + b(t_{n-1}) + \gamma e(t_{n-1})}{1 - c(t_{n-1}) - \gamma e(t_{n-1})} \phi(t_{n-1}) < \phi(t_{n-1}). \quad (1)$$

In a similar manner, one can prove (for the same even integer) that

$$\phi(t_{n-1}) \leq \frac{a(t_{n-2}) + c(t_{n-2}) + \delta f(t_{n-2})}{1 - b(t_{n-2}) - \delta f(t_{n-1})} \phi(t_{n-2}) < \phi(t_{n-2}). \quad (1')$$

Since  $\phi$  is increasing, (1) and (1') show that sequence  $t_n$  is decreasing. Let  $t$  be the limit of  $t_n$ . We shall prove that  $t = 0$ . Indeed, suppose that  $t > 0$ . Then  $t \leq t_{2n}$  and by (1), we have

$$\phi(t_{2n}) \leq \frac{a(t) + b(t) + \gamma e(t)}{1 - c(t) - \gamma e(t)} \phi(t_{2n-1}).$$

Now we let  $n \rightarrow \infty$  and use the continuity of  $\phi$  to obtain

$$\phi(t) \leq \frac{a(t) + b(t) + \gamma e(t)}{1 - c(t) - \gamma e(t)} \phi(t) < \phi(t),$$

which is a contradiction, hence  $t = 0$ .

(II) Now, we shall prove that  $\{x_n\}$  is a Cauchy sequence. Since  $t = 0$  one needs only to see that  $\{x_{2n}\}$  is a Cauchy sequence. To get a contradiction, let us suppose that there is a number  $\epsilon > 0$  and two sequences  $\{2n(k)\}$ ,  $\{2m(k)\}$  with  $2k \leq 2m(k) < 2n(k)$ , ( $k \in \mathbb{N}$ ) verifying

$$d(x_{2n(k)}, x_{2m(k)}) > \epsilon. \quad (2)$$

For each integer  $k$ , we shall denote  $2n(k)$  the least even integer exceeding  $2m(k)$  for which (2) holds. Then

$$d(x_{2m(k)}, x_{2n(k)-2}) \leq \epsilon \quad \text{and} \quad d(x_{2m(k)}, x_{2n(k)}) > \epsilon.$$

For each integer  $k$ , we shall put  $p_k := d(x_{2m(k)}, x_{2n(k)})$ ,  $q_k := d(x_{2m(k)+1}, x_{2n(k)+1})$ , and  $r_k := d(x_{2m(k)+1}, x_{2n(k)+2})$ , then we have

$$\begin{aligned} \epsilon &< p_k = d(x_{2m(k)}, x_{2n(k)}) \\ &\leq d(x_{2m(k)}, x_{2n(k)-2}) + d(x_{2n(k)-2}, x_{2n(k)-1}) + d(x_{2n(k)-1}, x_{2n(k)}) \\ &\leq \epsilon + t_{2n(k)-2} + t_{2n(k)-1}. \end{aligned} \quad (3)$$

Since the sequence  $\{t_n\}$  converges to 0, we deduce from (3) that  $\{p_k\}$  converges to  $\epsilon$ . Furthermore, the sequence  $\{q_k\}$  has also  $\epsilon$  as limit. Indeed, this fact results from the following estimates obtained by triangular inequality

$$\begin{aligned} -t_{2m(k)} - t_{2n(k)} + p_k &< d(x_{2m(k)+1}, x_{2n(k)+1}) \\ &\leq t_{2m(k)} + t_{2n(k)} + p_k \end{aligned} \quad (4)$$

We claim that the sequence  $\{r_k\}$  converges to  $\epsilon$ . Indeed, we get by using triangular inequality the following

$$r_k \leq t_{2m(k)} + p_k + t_{2n(k)} + t_{2n(k)+1}. \quad (5)$$

On the other hand, by definition of the integer  $2n(k)$  and by using triangular inequality we have the following

$$\begin{aligned} \epsilon &\leq d(x_{2m(k)}, x_{2n(k)-2}) \\ &\leq t_{2m(k)} + r_k + t_{2n(k)+1} + t_{2n(k)} + t_{2n(k)-1} + t_{2n(k)-2} \end{aligned} \quad (6)$$

(5) and (6) imply our claim. One can deduce that there exists an integer  $k_0$  such that  $d(x_{2n(k)+1}, x_{2m(k)}) > 0$ , and  $t_{2n(k)} < \frac{\epsilon}{2}$  for each integer  $k \geq k_0$ . Thus, by using (2) and the relation  $p_k - t_{2k} \leq d(x_{2n(k)+1}, x_{2m(k)})$ , we deduce (for all  $k \geq k_0$ ) that

$$\begin{aligned} \phi(r_k) = \phi(d(x_{2n(k)+2}, x_{2m(k)+1})) &= \phi(d(Sx_{2n(k)+1}, Tx_{2m(k)})) \\ &\leq \Gamma_1 + \Gamma_2 + \Gamma_3 + \Gamma_4 + \Gamma_5 \\ &\leq G_1 + G_2 + G_3 + G_4 + G_5, \end{aligned}$$

where

$$\begin{aligned} \Gamma_1 &= a(d(x_{2n(k)+1}, x_{2m(k)}))\phi(d(x_{2n(k)+1}, x_{2m(k)})), \\ \Gamma_2 &= b(d(x_{2n(k)+1}, x_{2m(k)}))\phi(d(x_{2n(k)+1}, x_{2n(k)+2})), \\ \Gamma_3 &= c(d(x_{2n(k)+1}, x_{2m(k)}))\phi(d(x_{2m(k)}, x_{2m(k)+1})), \\ \Gamma_4 &= e(d(x_{2n(k)+1}, x_{2m(k)}))\phi(\gamma d(x_{2n(k)+1}, x_{2m(k)+1})), \\ \Gamma_5 &= e(d(x_{2n(k)+1}, x_{2m(k)}))\phi(\delta d(x_{2m(k)}, x_{2n(k)+2})), \\ G_1 &= a(p_k - t_{2n(k)})\phi(t_{2n(k)} + p_k), \\ G_2 &= \phi(t_{2n(k)+1}), \\ G_3 &= \phi(t_{2m(k)}), \\ G_4 &= e(p_k - t_{2n(k)})\phi(\gamma q_k), \text{ and} \\ G_5 &= f(p_k - t_{2n(k)})\phi(\delta[t_{2m(k)} + r_k]). \end{aligned}$$

Let  $k \rightarrow \infty$ . Then by using the continuity of  $\phi$  and the fact that  $a, b, c, e, f$  are decreasing on  $]0, +\infty[$ , we obtain

$$\phi(\epsilon) \leq a\left(\frac{\epsilon}{2}\right)\phi(\epsilon) + e\left(\frac{\epsilon}{2}\right)\phi(\gamma\epsilon) + f\left(\frac{\epsilon}{2}\right)\phi(\delta\epsilon) < [a\left(\frac{\epsilon}{2}\right) + e\left(\frac{\epsilon}{2}\right) + f\left(\frac{\epsilon}{2}\right)]\phi(\epsilon) < \phi(\epsilon).$$

This gives a contradiction. Hence  $\{x_n\}$  is a Cauchy sequence in a complete metric space  $(X, d)$ , then one may find a point  $z = z(S, T) \in X$  such that  $x_n \rightarrow z$  as  $n \rightarrow \infty$ . Next, we shall prove that  $z$  is a common fixed point for  $S$  and  $T$ .

(III) Since  $t_n > 0$  for all integer  $n$ , we see that both subsequences  $(x_{2n})_n$  and  $(x_{2n+1})_n$  are not stationary. Therefore, we may find a subsequence  $(x_{2n(k)})_k$  such that  $x_{2n(k)+1} \neq z$  for every integer  $k$ . Let us suppose that  $Tz \neq z$ . In this case we are allowed to apply the inequality (A) and obtain for all  $k \in \mathbb{N}$ ,

$$\begin{aligned} \phi(d(x_{2n(k)+2}, Tz)) &= \phi(d(Sx_{2n(k)+1}, Tz)) \\ &\leq a(d(x_{2n(k)+1}, z))\phi(d(x_{2n(k)+1}, z)) + b(d(x_{2n(k)+1}, z))\phi(d(x_{2n(k)+1}, x_{2n(k)+2})) \\ &\quad + c(d(x_{2n(k)+1}, z))\phi(d(z, Tz)) + e(d(x_{2n(k)+1}, z))\phi(\gamma d(x_{2n(k)+1}, Tz)) \\ &\quad + f(d(x_{2n(k)+1}, z))\phi(\delta d(x_{2n(k)+2}, z)). \end{aligned} \tag{7}$$

By using the convexity of  $\phi$ , we deduce from (7) the following inequality:

$$\begin{aligned} \phi(d(x_{2n(k)+2}, Tz)) &\leq \phi(d(x_{2n(k)+1}, z)) + \phi(t_{2n(k)+1}) + \phi(\delta d(x_{2n(k)+2}, z)) \\ &\quad + [c(d(x_{2n(k)+1}, z)) + \gamma e(d(x_{2n(k)+1}, z))] \phi(d(z, Tz)) \\ &\quad + \phi(d(x_{2n(k)+1}, z)), \end{aligned} \tag{8}$$

which gives, after letting  $k \rightarrow \infty$  :

$$\begin{aligned} \phi(d(z, Tz)) &\leq \max\{c(t) + \gamma e(t) : t > 0\} \phi(d(z, Tz)) \\ &= C \phi(d(z, Tz)) < \phi(d(z, Tz)), \end{aligned} \tag{9}$$

which is a contradiction. Hence  $z = Tz$ , and in a similar way, it can be shown that  $z = Sz$ .

(IV) Suppose that there exists another point  $\xi \neq z$  fixed, for instance, by  $S$ . Then, by inequality (A), we have

$$\begin{aligned} \phi(d(\xi, z)) &= \phi(d(S\xi, Tz)) \\ &\leq [a(d(\xi, z)) + e(d(\xi, z)) + f(d(\xi, z))] \phi(d(\xi, z)) \\ &< \phi(d(\xi, z)) \end{aligned}$$

a contradiction. Therefore, we deduce that there exists a unique point  $z \in X$  such that  $Fix(S) = \{z\} = Fix(T) = Fix(\{S, T\})$ . This completes the proof of our theorem.  $\square$

Using the basic ideas in the proof of the previous theorem, one can establish the following theorem which generalizes the main result of the paper [6].

**2.2 Theorem.** *Let  $(X, d)$  be a complete metric space,  $\mathcal{A}$  a (finite or infinite) set of self-maps of  $X$  and  $\phi$  an element of  $\Phi$ . We suppose that for all  $S, T \in \mathcal{A}$  the following generalized contractive condition holds true:*

$$\begin{aligned} \phi(d(Sx, Ty)) \leq & \alpha(d(x, y))\phi(d(x, y)) + \beta(d(x, y))\phi(d(x, Sx)) \\ & + \gamma(d(x, y))\phi(d(y, Ty)) + \theta(d(x, y))\min\{\phi(d(x, Ty)), \\ & \phi(d(y, Sx))\} \quad \forall x, y \in X \quad \text{with } x \neq y, \end{aligned}$$

where  $\alpha, \beta, \gamma, \theta$  are four decreasing functions from  $]0, +\infty[$  into  $[0, 1[$  such that  $\alpha(t) + \beta(t) + \gamma(t) + \theta(t) < 1$ , for every  $t > 0$ , and  $\sup\{\max(\beta(t), \gamma(t)) : t > 0\} < 1$ . Then there exists a unique point  $z \in X$  such that  $\text{Fix}(S) = \{z\}$  for all  $S \in \mathcal{A}$ .

### 2.3 Remarks.

(a) If we take  $\mathcal{A} = \{S, T\}$  and  $\beta = \gamma$ , then we obtain the result by R. A. Rashwan and A. M. Sadeek in [6].

(b) If we take  $\beta = \gamma$  and  $\mathcal{A} = \{T\}$ , then we obtain one of the main results established by M. S. Khan et al. in the paper [4].

(c) If  $\theta = 0$ ,  $\mathcal{A} = \{T\}$  and the functions  $\alpha, \beta$  and  $\gamma$  are constants, then we get the results obtained by D. Delbosco in [1] and by F. Skof in the paper [8].

(d) Example: We give here an example where we discuss the validity of the assumptions of Theorem 2.2. We take  $X = \{1, 2, 3, 4\}$  and define a metric  $d$  on  $X$  by setting  $d(1, 2) = 1$ , and  $d(1, 3) = d(1, 4) = d(2, 3) = d(2, 4) = d(3, 4) = 2$ . We put  $\mathcal{A} = \{S, T, V\}$ , where  $S1 = S2 = S3 = S4 = 1$ ;  $T1 = T2 = T3 = 1, T4 = 2$ ; and  $V1 = V2 = V4 = 1, V3 = 2$ . For all  $t \geq 0$ , we put  $\alpha(t) = 2/5$ ,  $\beta(t) = 1/20$ ,  $\gamma(t) = 7/20$ ,  $\theta(t) = 1/6$ , and  $\phi(t) = t^2$ . Then all the conditions of Theorem 2.2 are satisfied for the set  $\mathcal{A} = \{S, T, V\}$ , which has 1 as unique common fixed point.

The following result is an easy consequence of our Theorem 2.2.

**2.4 Corollary.** *Let  $(X, d)$  a complete metric space,  $\mathcal{A}$  a (finite or infinite) set of self-maps of  $X$  and  $\phi$  an element of  $\Phi$ . We suppose that for all  $S, T \in \mathcal{A}$  the following generalized contractive condition holds true*

$$\begin{aligned} \phi(d(Sx, Ty)) \leq & \alpha(d(x, y))\phi(d(x, y)) + \beta(d(x, y))\min\{\phi(d(x, Sx)), \\ & \phi(d(y, Ty))\} + \gamma(d(x, y))\min\{\phi(d(x, Ty)), \phi(d(y, Sx))\} \\ & \forall x, y \in X \text{ with } x \neq y, \end{aligned} \quad (C)$$

where  $\alpha, \beta, \gamma$  are three decreasing functions from  $]0, \infty[$  into  $[0, 1[$  such that  $\alpha(t) + \mu\beta(t) + \gamma(t) < 1$ , for every  $t > 0$ , where  $\mu$  is a fixed constant in  $]1, +\infty[$ . Then there exists a unique point  $z \in X$  such that  $\text{Fix}(S) = \{z\}$  for all  $S \in \mathcal{A}$ .

### 3. A fixed point theorem in compact metric spaces

In a paper of Fisher [3], the following theorem has been established:

**3.1 Theorem.** *Let  $T$  be a continuous self-map of a compact metric space  $(X, d)$  such that*

$$d(Tx, Ty) < \frac{d(x, Tx) + d(y, Ty)}{2}, \quad (F)$$

for all distinct  $x, y$  in  $X$ . Then  $T$  has a unique fixed point .

Following the essential idea of our result presented in section 2 we shall generalize Theorem 3.1 as follows:

**3.2 Theorem.** *Let  $S, T$  be two self-maps of a compact metric space  $(X, d)$  and let  $\phi \in \Phi$  be a convex function. We suppose that  $T, S \circ T$  are continuous and that  $S, T$  verify for all distinct  $x, y$  in  $X$  the inequality*

$$\phi(d(Sx, Ty)) < \max\left\{\phi(d(x, y)), \frac{\phi(d(x, Sx) + \phi(d(y, Ty)))}{c}, \phi\left(\frac{d(x, Ty) + d(Sx, y)}{c}\right)\right\} \quad (G)$$

where  $c \geq 2$  is a fixed constant. Then  $S$  and  $T$  have a unique common fixed point  $z \in X$ . Moreover  $\text{Fix}(S) = \text{Fix}(T) = \{z\}$ .

**Proof.** Let  $x_0$  be an element in  $X$ , an associate to it the sequence  $(x_n)_n$  given by

$$\begin{aligned}x_{2n} &= Sx_{2n-1}, \quad n = 1, 2, \dots \\x_{2n+1} &= Tx_{2n}, \quad n = 0, 1, 2, \dots\end{aligned}$$

Without loss of generality, we may assume that  $t_n \neq 0$  for every integer  $n$ . In this case, it is easy to see that the sequence  $(\phi(t_n))$  is decreasing and therefore it converges. Since  $X$  is compact, we may find a subsequence  $(x_{2n(k)})_k$  converging to some element  $z \in X$ . Then by using the continuity of the maps  $T$  and  $\phi$ , we get

$$\begin{aligned}\phi(d(z, Tz)) &= \lim_{k \rightarrow +\infty} \phi(t_{2n(k)}) = \lim_{k \rightarrow +\infty} \phi(t_{2n(k)+1}) \\&= \lim_{k \rightarrow +\infty} \phi(d(x_{2n(k)+1}, x_{2n(k)+2})) \\&= \lim_{k \rightarrow +\infty} \phi(d(Tx_{2n(k)}, (S \circ T)x_{2n(k)})) \\&= \phi(d(Tz, (S \circ T)z)).\end{aligned}\tag{10}$$

Suppose that  $z \neq Tz$ , then we can apply the inequality (G) to  $x = Tz$  and  $y = z$ . By using (10) and the fact that  $\phi$  is convex and increasing, we obtain

$$\begin{aligned}\phi(d(z, Tz)) &= \phi(d(S(Tz), Tz)) \\&< \max \left\{ \phi(d(Tz, z)), \frac{\phi(d(Tz, STz)) + \phi(d(z, Tz))}{c}, \phi\left(\frac{d(STz, Tz) + d(Tz, z)}{c}\right) \right\} \\&\leq \max \left\{ \phi(d(Tz, z)), \frac{2\phi(d(z, Tz))}{c} \right\} \\&\leq \max \left\{ 1, \frac{2}{c} \right\} \phi(d(Tz, z)) \leq \phi(d(Tz, z)).\end{aligned}$$

This is a contradiction. Therefore we must have  $Tz = z$ . The relation (10) will imply that  $Sz = z$ . To end the proof, let us suppose that there exists a point  $\xi \neq z$  fixed, for instance, by  $S$ . Then by applying the inequality (G), we get

$$\phi(d(\xi, z)) = \phi(d(S\xi, Tz)) < \max \left\{ \phi(d(\xi, z)), \frac{2}{c}\phi(d(\xi, z)) \right\} \leq \phi(d(\xi, z)),$$

which is a contradiction.  $\square$

## References

- [1] D. Delbosco, Un estensione di un teorema sul punto fisso di S. Reich., *Rend. Sem. Mat. Univers. Politecn. Torino*, 35 (1976- 77) 233-239.

- [2] M. Edelstein, On fixed and periodic points under contractive mappings, *J. London Math. Soc.*, 37 (1962) 74-79.
- [3] B. Fisher, A fixed point mapping, *Bull. Calcutta Math. Soc.*, 68 (1976) 265-266.
- [4] M.S. Khan, M. Swaleh and S. Sessa, Fixed point theorems by altering distance between the points, *Bull. Austral. Math. Soc.*, 30 (1984) 1-9.
- [5] E. Rakotch, A note on contractive mappings, *Proc. Amer. Math. Soc.*, 13 (1962) 459-465.
- [6] R.A. Rashwan, A.M. Sadeek, A common fixed point theorem in complete metric spaces, *Electronic Journal : Southwest. Jour. of Pure and Appl. Math.*, Vol. 1, (1996) 6-10.
- [7] S. Reich, Kannan's fixed point theorem, *Boll. U. M. I.*, Vol. 4, 4 (1971) 1-11.
- [8] F. Skof, Teorema di punti fisso per applicazioni negli spazi metrici, *Atti. Accad. Aci. Torino*, 111 (1977) 323-329.

## ON NONLINEAR WATER WAVES IN A CHANNEL

Z.R. Bhatti, I.R. Durrani and S. Asghar

\* *Department of Mathematics, Govt. College of Science, Wahdat Road,  
Lahore-54570, Pakistan*

\*\* *Center of Excellence in Solid State Physics, University of the Punjab,  
Quaid-i-Azam Campus, Lahore-540590, Pakistan*

\*\*\* *Department of Mathematics, Quaid-i-Azam University, Islamabad,  
Pakistan*

**Abstract:** This paper is concerned with some approximate equations for the study of nonlinear water waves in a channel of variable cross section. A system of shallow water equations for finite amplitude waves is given and a Korteweg deVries (KdV) equation with variable coefficients for small amplitude waves is also presented.

### 1. Introduction

One of the interesting problems of water waves in a sloping channel concerns the breaking of a wave moving toward a shoreline, the development of a bore, and the movement of the shoreline after the bore reaches it. For the two dimensional case corresponding to a rectangular channel of variable depth, the bore run-up problem was studied by Keller et al [1] on the basis of shallow water equations [2]. Later Gurtin [3] derived a criterion for the breaking of an acceleration wave in a two dimensional

---

<sup>o</sup>MSC: 76B15, 35Q35.

<sup>o</sup>Keywords: Nonlinear water waves, Korteweg deVries equation.

channel and his result was extended by Jeffery and Mvungi [4] to the case of a rectangular channel of variable width depth. We generalize Gurtin's result to predict the breaking point of an acceleration wave in a channel of variable cross section and review some existent results regarding the bore run-up problem for a rectangular channel with a uniformly sloping bottom. Up to date, the shallow water equation for a two dimensional channel with analytical initial data have been justified by Kano and Nashida [5] and for the three dimensional case with a priori assumptions on the free surface by Berger [6]. At present we may accept shallow water equations as model equations.

Another application of our results deals with the development of a solitary wave in a channel of variable cross section. Recently, there have been discussions on the so called infinite mass dilemma, which arises from the formation of a shelf behind the solitary wave. If the shelf were extended to infinity, then infinite mass would be created or annulled by a perturbation of the solitary wave. We shall establish a global existence theorem for the solution of the KdV equation for a general channel as a consequence of the existence results due to Kato [7]. It follows that the shelf, if formed behind the solitary wave in a general channel, can only be finite. A rigorous justification of the validity of the KdV equation here should be an important contribution to the theory of water waves.

## 2. Shallow Water Equations and the Breaking of a Wave

We consider the irrotational motion of an inviscid, incompressible fluid of constant density under gravity in a channel with a boundary defined by  $h^*(x^*, y^*, z^*) = 0$ , where  $z^*$  is positive upward and  $x^*$  is in the longitudinal direction (figure). The governing equations are

$$\nabla^* \cdot \bar{q}^* = 0 \quad (2.1)$$

$$\nabla^* \wedge \bar{q}^* = 0 \quad (2.2)$$

$$\rho(\bar{q}_{i^*}^* + \bar{q}^* \cdot \nabla^* q^*) = -\nabla^* p^* + \bar{g} \quad (2.3)$$

subject of the boundary conditions

$$n_{t^*}^* + \bar{g}^* + \nabla^* \xi^* = 0 \quad (2.4)$$

at

$$\xi^* = -\xi^* + \eta^*(t^*, x^*, y^*), \quad \rho^* = 0 \quad (2.5)$$

$$\bar{q}^* + \nabla^* h^* = 0 \quad \text{at} \quad h^* = 0 \quad (2.6)$$

Here

$$\nabla^* = \left( \frac{\partial}{\partial x^*}, \frac{\partial}{\partial y^*}, \frac{\partial}{\partial z^*} \right), \quad \bar{q}^* = (u^*, v^*, w^*)$$

is the velocity,  $t^*$  is the time,  $\bar{g} = (0, 0, -g)$  is the constant gravitational acceleration,  $\rho$  is the constant density,  $\rho^*$  is the pressure, and  $z^* = \eta^*$  is the equation of the free surface. To derive the shallow water equations, we make the following assumptions. The channel boundary is convex, sufficiently smooth, and varies slowly in the longitudinal direction; the magnitude of the transverse velocities is much smaller than that of the longitudinal velocity. As suggested by Friedrichs [8], we introduce non dimensional variables:

$$t = \frac{1}{\sqrt{B}} \frac{t^*}{\sqrt{h/g}}, \quad (x, y, z) = \left( \frac{1}{\sqrt{B}} \frac{x^*}{H}, \frac{y^*}{H}, \frac{z^*}{H} \right),$$

$$\eta = \left( \frac{\eta^*}{H} \right), \quad h = \left( \frac{h^*}{H} \right), \quad (u, v, w) = \left( \frac{u^*}{\sqrt{gH}}, \frac{\sqrt{\beta} v^*}{\sqrt{gH}}, \frac{\sqrt{\beta} w^*}{\sqrt{gH}} \right)$$

where  $\sqrt{B} = (L/H)$  and 'L' and 'H' are respectively the horizontal and transverse length scales. In terms of them (2.1) to (2.6) become

$$u_x + v_y + w_z = 0 \quad (2.7)$$

$$\beta u_y = v_x, \quad u_z = w_x, \quad v_z = w_y \quad (2.8)$$

$$u_t + uu_x + vu_y + wu_z + p_x = 0 \quad (2.9)$$

$$v_t + uv_x + vv_y + vw_z + \beta p_y = 0 \quad (2.10)$$

$$w_t + uw_x + vw_y + vw_z + \beta(p_z + 1) = 0 \quad (2.11)$$

$$\eta_t + u\eta_x + v\eta_y - w = 0, \quad \text{at } z = \eta \quad (2.12)$$

$$p = 0 \quad (2.13)$$

$$uh_x + vh_y + wh_z = 0 \quad \text{at } h = 0 \quad (2.14)$$

Assume that  $u, v, w$  and  $\beta$  possess an asymptotic expansion of the form

$$\phi \sim \phi_0 \beta^{-1} \phi_1 + \beta^{-2} \phi_2 + \dots \quad (2.15)$$

Substitute (2.15) into (2.7) to (2.14). The equations for the zeroth order approximation are

$$u_{0x} + v_{0y} + w_{0z} = 0 \quad (2.16)$$

$$u_{0y} = u_{0z} = 0 \quad (2.17)$$

$$u_{0t} + u_0 u_{0x} + p_{0x} + v_0 u_{0y} + w_0 u_{0z} = 0 \quad (2.18)$$

$$p_{0y} = 0, \quad p_{0z} = -1 \quad (2.19)$$

$$\eta_{0t} + u_0 \eta_{0x} + v_0 \eta_{0y} - w_0 = 0 \quad \text{at } z = 0 \quad (2.20)$$

$$p_0 = 0 \quad (2.21)$$

$$u_0 h_x + v_0 h_y + w_0 h_z = 0 \quad \text{at } h = 0 \quad (2.22)$$

Seen from (2.17), (2.19) and (2.21),  $u_0$  is a function of  $(t, x)$  only and

$$p_0 = (-z + \eta_0) \quad (2.23)$$

This implies  $\eta$  is also a function of  $(t, x)$  only. It follows from (2.17), (2.18) and (2.23) that

$$u_{0t} + u_0 u_{0x} + \eta_{0x} = 0 \quad (2.24)$$

Now we integrate (2.16) over a cross section  $D$  of the channel, apply the divergence theorem and make use of (2.20) and (2.22) to obtain:

$$\int_D (v_{0y} + w_{0z}) dy dz = -u_{0x} A(t, x) = -u_0 \int_{\Gamma} h_x (h_y^2 + h_z^2)^{-1/2} ds + (\eta_{0t} + u_0 \eta_{0x}) B(t, x)$$

Rearranging the terms, we have

$$\eta_{0t} + u_0 \eta_{0x} + u_{0x} \frac{A(t, x)}{B(t, x)} - \left[ \frac{u_0}{B(t, x)} \right] \int_{\Gamma} h_x (\sqrt{h_y^2 + h_z^2})^{-1} ds = 0 \quad (2.25)$$

where  $A(t, x)$  is the area,  $B(t, x)$  is the width and ' $\Gamma$ ' is the wetted boundary of the cross section  $D$  (figure). (2.24) and (2.25) form a system of nonlinear equations, which may be used to model bore formation and its subsequent development in a channel of variable cross section.

In the following we extend Gurtin's method to the case of a general channel. The assumptions made are the following:

- 1)  $u_0, \eta_0$  are continuous.
- 2) The first and the second derivatives of  $u_0$  and  $\eta_0$  possess at most jump discontinuities.
- 3)  $u_0 = \eta_0 = 0$  ahead of the wave.

Denote the value of a function ' $f$ ' immediately behind the wave front by  $\bar{f}$ . Thereafter we also drop the subscript ' $0$ ', from assumptions (1), (2), we have

$$\bar{u} = \bar{\eta} = 0 \quad (2.26)$$

By total differentiation

$$\bar{u}_t = -c \bar{u}_x, \quad \bar{\eta}_t = -c \bar{\eta}_x \quad (2.27)$$

Where 'c' is the speed of the wave front. From (2.24), (2.25) and (2.26) it follows that:

$$\overline{u}_1 + \overline{\eta}_x = 0, \quad \overline{\eta}_t + \frac{\overline{u}_x A}{B} = 0 \quad (2.28)$$

Comparing (2.27) and (2.28), we have:

$$c = \sqrt{\frac{\overline{A}}{\overline{B}\overline{u}_t = c^{-1}\overline{\eta}_x}} \quad (2.29)$$

Now we differentiate (2.24) with respect to 't' and (2.25) with respect to x, and evaluate the equations behind the wave front. Then we eliminate  $\overline{\eta}_{tx}$  and make use of the expression

$$c^2 \overline{u}_{xt} - \overline{u}_{tt} = c^2 \frac{d}{dx}(\overline{u}_x) - \frac{d}{dx}(\overline{u}_t)$$

to obtainq

$$-2c \frac{d}{dx}(\overline{\eta}_x)^{-1} + (\overline{\eta}_x)^{-1}[c^t - \overline{I}_1(\overline{B}c)] + \frac{3}{c} = 0$$

where

$$\overline{I}_1 = \int_{\Gamma} \frac{h_x}{\sqrt{h_y^2 + h_z^2}} ds$$

Hence

$$\overline{\eta}_x = \frac{a_0}{\sqrt{c}} \left[ \left( \frac{3}{2} a_0 \int_{x_0}^x c^{-5/2} \exp \int_{x_0}^{x'} \overline{I}_1(2\overline{A})^{-1} dx' dx + 1 \right)^{-1} x \exp \int_{x_0}^x \overline{I}_1(2\overline{A})^{-1} dx \right] \quad (2.30)$$

where  $a_0$  is the initial value of  $\eta_x$  at  $x = x_0$ . We call  $x = l$  a shoreline of  $\overline{A}(l) = 0$  but  $\overline{B}(l) \neq 0$ , and let

$$I(x) = \frac{3}{2} \int_{x_0}^x \exp(-5/2) \exp \int_{x_0}^{x'} \overline{I}_1(2\overline{A})^{-1} dx' dx$$

Suppose  $a_0 < 0$ . If  $I(l) = \infty$ , then  $\overline{\eta}_x$  and the wave breaks before it reaches the shoreline. If  $I(l) \neq \infty$ , then either the wave breaks before it reaches the shoreline or it breaks at the shoreline. Next suppose

$a_0 > 0, I(l) \neq \infty$ , then the wave breaks at the shoreline. Otherwise if  $I(l) = \infty$ , evaluate the limit of  $\bar{\eta}_x$  given by (2.30) as  $x \rightarrow l$  and obtain

$$\lim_{x \rightarrow l} \bar{\eta}_x = \frac{2}{3} \int - \left( \frac{(\bar{d})'}{4} + \frac{\bar{T}_1}{2\bar{B}} \right)_{x=l} \quad (2.31)$$

Here  $\bar{d} = (\bar{A}/\bar{B})$ , hence the wave will never break if  $(\bar{d})'$  is finite at  $x = l$ . However for the channels of variable cross section the equilibrium for surface may converge to a point and this case is also of interest. Assume again ( $a_0 > 0$ ), if  $I(l) = \infty$ . If  $\bar{B}(l) = \bar{d}(l)$  and  $(\bar{d})'$  is finite at  $x = l$ , we assume  $h(x, y, z) = -z + g(x, y)$

$$\bar{T}_1 = \int_{\Gamma} h_x (h_y^2 + h_z^2)^{-1/2} ds = \int_{-b_1}^{b_2} g_x dy$$

Here  $y = -b_1, b_2$  are the end points of the width  $\bar{B}(x)$ . It follows from (2.31) that

$$\lim \bar{\eta}_x = \left( \frac{2}{3} \right) \left[ - \left( \frac{(\bar{d})'}{4} + \frac{gx}{2} \right) \right]_{x=l}$$

then the wave will never break.

### 3. Run-up Problem

We consider a bore propagating towards a shoreline in a rectangular channel with a uniformly slopping bottom. On the basis of the shallow water equations, we can find a fairly complete solution of the bore run up problem. The bore path at the point of breaking to the shoreline may be approximately determined by Whitham's rule [10]. Here we shall consider the movement of shoreline after the bore reaches to shore. The shallow water wave equations for a rectangular channel of variable depth are obtained from (2.24) and (2.25) as

$$u_t + uu_x + \eta_x = 0 \quad (3.1)$$

$$\eta_x + [u(\eta + d_0)]_x = 0 \quad (3.2)$$

Here we also drop the superscripts of  $u$  and  $\eta$  and  $d_0 = -\gamma x, \gamma > 0$ . We assume  $t = 0$  when the bore reaches the shoreline  $x = 0$ . Let

$$c^2 = \eta + d_0 \quad (3.3)$$

$$\alpha = 2c + u + \eta t = u^0, \quad \beta = 2c - u - \eta t + u^0 \quad (3.4)$$

In terms of 'a' and 'B', (3.1) and (3.2) can be expressed as

$$x_\alpha = (u - c)t_\alpha, \quad x_\beta = (u + c)t_\beta \quad (3.5)$$

By cross differentiation of equation (3.5) and making use of (3.4), we have

$$t_{\alpha\beta} + \frac{3(t_\alpha + t_\beta)}{[2(\alpha + \beta)]} = 0 \quad (3.6)$$

If we introduce the canonical variables

$$a = (\alpha + \beta)^{3/2}t_\alpha, \quad b = (\alpha + \beta)^{3/2}t_\beta \quad (3.7)$$

(3.6) yields as system of equations

$$(\alpha + \beta)a_\beta = -\frac{3b}{2}, \quad (\alpha + \beta)b_\alpha = -\frac{3a}{2} \quad (3.8)$$

Let

$$Y = a + b, \quad Z = a - b \quad (3.9)$$

It follows from (3.8) that

$$Y_{\alpha\beta} = \frac{15Y}{[4(\alpha + \beta)^2]}, \quad Z_{\alpha\beta} = \frac{3Z}{[4(\alpha + \beta)^2]} \quad (3.10)$$

In the  $\alpha\beta$ -plane we prescribe sufficiently smooth data. However, the precise nature of the data is immaterial. We require only  $t_\alpha(\alpha, \beta^*) > 0$ ,  $t_\beta(\theta, \beta) < 0$  and that as  $\beta \rightarrow 0^+$  along  $\alpha = 0$ .

$$\lim a = a^0 > 0, \quad \lim y = u^0 > 0 \quad (3.11)$$

$$\lim x = \lim t = 0, \quad b(0, \beta) = 0(\beta^{a/2}) \quad (3.12)$$

where the existence of the positive limit  $a^0$  and the behavior of  $b(0, \beta)$  for small  $\beta$  were established by Ho and Mayer [9].

#### 4. KdV Equation and the Development of a Solitary Wave

We only sketch the derivation of KdV equation for a channel of variable cross section; the details may be found in Shen and Zhong [10]. We introduce the non dimensional variables

$$t = \beta^{-3/2} \frac{t^*}{\sqrt{H/g}}, \quad (x, y, z) = (\beta^{-3/2} \frac{x^*}{H}, \frac{y^*}{H}, \frac{z^*}{H})$$

$(\eta, h, p)$  and  $(u, v, w)$  are the same as before. The method used here is the specialization of the procedure developed by Shen [2] and Keller [1]. We assume that  $u, v, w, p, \eta$  depend explicitly upon a new variable,  $\xi = \beta S(t, x)$ , where  $S$  is a function of  $t$  and  $x$  only, will be called a phase function. Then we assume that they possess an asymptotic expansion of the form:

$$\phi(\xi, t, x, y, z, \beta) \sim \phi_0 + \beta^{-1} \phi_1 + \beta^{-2} \phi_2 + \dots$$

and we assume that the zeroth order approximation is given by

$$(u_0, v_0, w_0) = 0, \quad p_0 = -z_0, \quad \eta_0$$

The equation for the first approximation determines a Hamilton-Jacobi equation for  $S$ . Let  $k = S_x, w = -S_t$ . Then

$$w = kG(x), G(x) = \pm \sqrt{\frac{a(x)}{b(x)}} \quad (4.1)$$

where  $a(x)$  is the area of the cross section  $D_0$ , and  $b(x)$  is the width of  $D_0$  of water wave at rest (figure). (4.1) may be solved by the method of characteristics and the corresponding characteristic equation are

$$\frac{dt}{d\sigma} = \mu, \quad \frac{dx}{d\sigma} = \mu G(x), \quad \frac{dk}{d\sigma} = -k\mu G'(x), \quad \frac{d\mu}{d\sigma} = \frac{dS}{d\sigma} = 0 \quad (4.2)$$

where ' $\mu$ ' is the proportionality factor. We choose  $\mu = 1$ , so that  $\sigma = t$ . The equation of (4.2) determine a one parameter family of bicharacteristics, called rays.

$$x = x(t, \sigma_1)$$

where  $\sigma_1$  is constant along a ray. The equations for the second approximation determine a KdV equation with variable coefficients:

$$m_0\eta_{1t} + m_1\eta_{1x} + m_2\eta_1 + m_3\eta_1\eta_{1\xi} + m_4\eta_{1\xi\xi\xi} = 0 \quad (4.3)$$

where

$$m_0 = 2b(x) \quad (4.4)$$

$$m_1 = \frac{2a(x)}{G(x)} \quad (4.5)$$

$$m_2 = -[G(x)]^{-1} \int_{\Gamma_0} \frac{h_x}{\sqrt{h_y^2 + h_z^2}} ds - G^{-2}(x)G'(x)a(x) \quad (4.6)$$

$$m_3 = 3k[G(x)]^{-1}b(x) - \frac{1}{w}[\phi_y(t, x, y_2, 0) - \phi_y(t, x, y_1, 0)] \quad (4.7)$$

$$m_4 = w^{-1} \int \int_{D_0} (\nabla\phi)^2 dydz \quad (4.8)$$

' $\Gamma_0$ ' is the wetted boundary of  $D_0$ ;  $y = y_1, y_2$  are the endpoints of the width of  $D_0$ ; and  $\phi_0$  is a solution of the Neumann problem.

$$\begin{aligned} \nabla^2\phi &= k^2 \quad \text{in } D_0 \\ \phi_z &= w^2 \quad \text{at } z = 0 \\ \phi_y h_y + \phi_z h_z &= 0 \quad \text{at } \Gamma_0 \end{aligned}$$

Since from equation (4.2)

$$\left(\frac{d}{d\sigma}\right) = \partial_t + G(x)\partial_x, \quad \left(\frac{dx}{d\sigma}\right) = G(x)$$

along a ray, we may express (4.3) in terms of  $\sigma$  and  $\xi$ .

$$m_0\eta_{1\sigma} + m_2\eta_1 + m_3\eta_1\eta_{1\xi} + m_4\eta_{1\xi\xi\xi} = 0$$

or in terms of  $x$  and  $\xi$

$$G(x) = +\sqrt{\frac{a(x)}{b(x)}}, \quad S = -t + \int_0^x \frac{1}{G(x)} dz$$

which is a solution of equation (4.2) and it follows that

$$w = 1, \quad k = G^{-1}(x)$$

For rectangular and triangular channels, the coefficients given in (4.4) to (4.8) can be explicitly evaluated, Shen and Zhong [10]. It is also remarked in passing that (4.3) has been used to study the fission of solutions in channel of variable cross section [10] and a justification of the asymptotic method used should also be of interest.

**Figure:** A cross section of the channel

## References

- [1] H.G. Keller, D.A. Levine, G.B. Whitham, Motion of a bore over a sloping beach, *J. Fluid Mech.*, 7 (1960) 302-316.
- [2] M.C. Shen and R.E. Meyer, Climb of a bore on a beach III. Run-up, *J. Fluid Mech.*, 16 (1963) 108.
- [3] M.E. Gurtin, *Quart. Appl. Math.*, 33 (1975) 187.
- [4] A. Jefferey, J. Mvungi, On the breaking of water waves in a channel of arbitrary varying depth and width, *J. Appl. Math. Phys. ZAMP*, Vol. 31, 4 (1980) 758-761.
- [5] T. Kano, T. Nishida, Sur les ondes de surface del eau des ondes en eau peu profonde (French), *J. Math. Kyoto Univ.*, Vol. 19, 2 (1979) 335-370.
- [6] N. Berger, Derivation of approximate long wave equations in a nearly uniform channel of approximately rectangular cross section, *SIAM J. Appl Math.*, Vol. 31, 3 (1976) 438-448.
- [7] T. Kato, The Cauchy problem for the korteweglmhy de Vries equation, *Res. Notes in Maths. Pitnam NY*, 53 (1980) 293-307.
- [8] K.O. Friedrichs, *Comm. Pure and Appl. Math.*, 1 (1948) 81.
- [9] D.V. Ho, R.E. Meyer, Climb of a bore on a beach. I. Uniform beach slope, *J. Fluid Mech.*, 14 (1962) 305-318.
- [10] M.C. Shen, X.C. Zhong, Derivation of K-dv equations for water waves in a channel with variable cross section, *J. de. Mec.*, Vol. 20, 4 (1981) 789.

## SHAPE-MEASURE METHOD FOR SOLVING ELLIPTIC OPTIMAL SHAPE PROBLEMS (FIXED CONTROL CASE)

A. Fakharzadeh J. and J. E. Rubio

*\*Department of Mathematics, Shahid Chamran University of Ahwaz, Ahwaz,  
Iran  
a\_fakharzadeh@hotmail.com*

*\*\*Department of Applied Mathematical Studies, University of Leeds, Leeds,  
LS2 9JT, UK*

**Abstract:** The aim of this paper is to introduce a new method for solving optimal shape problems which are defined with respect to a pair of geometrical elements. The problem is to find the optimal domain for a given functional that is involved with the solution of a linear or nonlinear elliptic equation with a boundary condition over a domain. By transferring the problem into a measure-theoretical form the shape-measure method, in Cartesian coordinates, will be used to find the optimal solution in two steps. First we will find the solution of the elliptic problem for a given domain by using the embedding method. Then the Shape-Measure method will be applied to find the optimal solution. Two examples are given for the linear and nonlinear cases of the elliptic problem.

---

<sup>0</sup>MSC: 49.

<sup>0</sup>Keywords: Elliptic equation, Radon measure, linear system, optimal Shape.

## 1. Introduction

Consider the optimal shape (OS) or optimal shape design (OSD) problems in which they are defined with respect to a pair of geometrical elements; this pair consists in a measurable set (in  $\mathbb{R}^2$ ), which can be regarded as a domain, and a simple closed curve containing a given point, which is the boundary of the set. Based on the simple property of curves, the related OSD problem depends on the geometry which is used. We solved the appropriate OS in [2] by introducing shape-measure method in Polar coordinates. But in Cartesian coordinates, it is difficult to introduce a linear condition which determines the property of a closed curve being simple; thus in this paper we consider those measurable sets  $D$  which its boundary consists in a variable part  $\Gamma$  and a fixed part between two given points, to be sure it is simple.

This paper deals with solving an OS or OSD problem with a fixed control, which is to find the optimal domain like  $D$  for a given function,  $I$ , that is involved with the solution of a linear or nonlinear elliptic partial differential equation with a boundary condition over  $D$ . The process of solution is achieved in two stages. First for a fixed domain, by using the idea of approximating a curve by broken lines,  $\Gamma$  can be determined with fixed number of  $M$  points. Then  $D$ , any integral on  $D$  and the variational form of elliptic equations can be considered as a function of  $M$  variables. By means of a well-known process of embedding, we transfer the problem into a measure-theoretical one. The history of this idea can be found, for instance, in [2] and [10]. Then we enlarge the underlying space to reach an infinite linear system of equations that the unknown is a measure. By the use of total sets and putting an appropriate discretization, one can approximate the solution of the problem with the solution of a finite linear system of equations. Hence the value of  $I$  is calculated as a function of  $M$  variables for any given domain  $D$ . In the second stage, considering the previous one, a vector function  $J : D \in \mathcal{D}_M \rightarrow I(D)$  is set up. Using a standard minimization algorithm on  $J$ , gives the minimizer domain; then Theorem 1, proves that this minimizer, is the optimal solution for the problem. Finally, two

examples for the linear and nonlinear cases of elliptic problem are given.

## 2. Problem

Let  $D \subset \mathbb{R}^2$  be a bounded domain with a piecewise-smooth, closed and simple boundary  $\partial D$ . We assume that some part of  $\partial D$  is fixed and the rest,  $\Gamma$ , with the given initial and final points  $A$  and  $B$  respectively, is not fixed. Here we suppose that the fixed part of  $\partial D$  is made by three segments, parts of lines  $y = 0, x = 0$  and  $y = 1$  between points  $A(1, 0), (0, 0), (0, 1), B(1, 1)$  (see Figure 1). For more general case, the reader is advised to see [1]. Thus we choose an appropriate (variable) curve  $\Gamma$  joining  $A$  and  $B$ , so that  $D$  is well-defined. Let  $X \in D \longrightarrow u(X) \in \mathbb{R}$ , that  $X = (x, y) \in \mathbb{R}^2$ , is a bounded solution of the following elliptic partial differential equation with the boundary condition on the domain  $D$ :

$$\Delta u(X) + f(X, u) = v(X), \quad u|_{\partial D} = 0, \quad (1)$$

where  $X \in D \longrightarrow v(X) \in \mathbb{R}$  is a bounded fixed control function; the function  $f$  is assumed to be a bounded and continuous real-valued function in  $L_2(D \times \mathbb{R})$ . A domain  $D$  as above, is called an *admissible domain* if the elliptic equation (1) has a bounded solution on  $D$ ; we denote by  $\mathcal{D}$  as the set of all such admissible domains. We are going to solve the problem of minimizing the functional  $I(D) = \int_D f_\circ(X, u) dX$ , on the set  $\mathcal{D}$  where  $f_\circ$  is a given continuous, nonnegative, real-valued function on  $D \times \mathbb{R}$ . To calculate the value of  $I(D)$  for a given domain  $D$ , it is necessary, first, to identify the solution of the partial differential equations (1).

## 3. Weak Solution and Metamorphosis

In general, it is difficult and sometimes impossible to identify a classical solution for the problem like (1); thus usually one tries to find a generalized or *weak* solution of them which is more applicable than the classical one in some branches. In our method, especially whenever one wants to change the problem into a measure-theoretical form,

this kind of solution is more appropriate. Hence the variational form of the problem (1) is introduced in the following proposition. We remind the reader that here  $H_0^1(D) = \{\psi \in H^1(D) : \psi|_{\partial D} = 0\}$ , where  $H^1(D) = \left\{h \in L_2(D) : \frac{\partial h}{\partial x} \in L_2(D), \frac{\partial h}{\partial y} \in L_2(D)\right\}$  is the Sobolev space of order 1.

**Proposition 1:** *Let  $u$  be the classical solution of (1), then we have the following equality:*

$$\int_D (u\Delta\psi + \psi f) dX = \int_D \psi v dX ; \forall \psi \in H_0^1(D). \quad (2)$$

**Proof:** Multiplying (1) by the function  $\psi \in H_0^1(D)$  and then integrating over  $D$ , with use of the Green's formula (see for instance [4]) gives:

$$\int_D (u\Delta\psi + \psi f - \psi v) dX = \int_{\partial D} \left(\psi \frac{\partial u}{\partial n} - u \frac{\partial \psi}{\partial n}\right) dS,$$

where  $n$  is the unit vector normal to the boundary  $\partial D$  and directed outward with respect to  $D$ . Because  $\psi|_{\partial D} = 0$  and  $u|_{\partial D} = 0$ , then (2) is satisfied.  $\square$

**Definition:** A function  $u \in H^1(D)$  is called a bounded weak solution of the problem (1) when it bounded and satisfies in the equality (2) for all functions  $\psi \in H_0^1(D)$ .

We remind the reader that conditions for the existence of the classical and of the weak solution of the problem (1), and also other properties of them such as boundedness and uniqueness, have been considered in many references, like [4] and [3].

Now we can apply our *Shape-Measure* method for solving the problem. The bounded weak solution can be represented by a positive Radon measure. Hence instead of looking for the weak solution on the given domain  $D$ , one prefers to seek for its related measure, defined on the appropriate space. For the rest of the paper, we suppose  $\Omega \equiv U \times \overline{D}$ , where  $U \subset \mathbb{R}$  is the smallest bounded set in which the bounded weak solution  $u(\cdot)$  takes values. By applying the Riesz Representation Theorem [11], similar to the Proposition 3.1 in [2], one can prove the following proposition.

**Proposition 2:** *Let  $u(X)$  be a bounded generalized solution of (1); there exists a unique positive Radon measure, say  $\mu_u$ , in  $\mathcal{M}^+(\Omega)$  so that:*

$$\mu_u(F) \equiv \int_{\Omega} F d\mu_u = \int_D F(X, u) dX ; \forall F \in C(\Omega). \quad (3)$$

Thus the equality (2) changes into the following:

$$\mu_u(F_{\psi}) = \gamma_{\psi} \quad ; \quad \forall \psi \in H_0^1(D) \quad (4)$$

where  $F_{\psi} = u\Delta\psi + f\psi$  and  $\gamma_{\psi} = \int_D \psi v dX$ . Also,  $I(D) = \mu_u(f_{\circ})$ .

Because the measure  $\mu_u$  projects on the  $(x, y)$ -space as the respective Lebesgue measure, we should have  $\mu_u(\xi) = a_{\xi}$ , where  $\xi : \Omega \rightarrow \mathbb{R}$  depends only on variable  $X$  (i.e.  $\xi \in C_1(\Omega)$ ), and  $a_{\xi}$  is the Lebesgue integral of  $\xi$  over  $D$ . Therefore the problem can be described as follows:

Find a measure  $\mu_u \in \mathcal{M}^+(\Omega)$  so that it satisfies the following constraints:

$$\begin{aligned} \mu_u(F_{\psi}) &= \gamma_{\psi}, & \forall \psi \in H_0^1(D); \\ \mu_u(\xi) &= a_{\xi}, & \forall \xi \in C_1(\Omega). \end{aligned} \quad (5)$$

As Rubio did in [9], to be sure that we do not miss any solution, consider a more general version of the problem by extending the underlying space; instead of finding a  $\mu_u \in \mathcal{M}^+(\Omega)$ , defined by Proposition 2, satisfying equalities (5), we seek a measure  $\mu \in \mathcal{M}^+(\Omega)$  which satisfies just the conditions

$$\begin{aligned} \mu(F_{\psi}) &= \gamma_{\psi}, & \forall \psi \in H_0^1(D); \\ \mu(\xi) &= a_{\xi}, & \forall \xi \in C_1(\Omega). \end{aligned} \quad (6)$$

Hence we have  $I(D) = \mu(f_{\circ})$ . The system (6) is linear because all the functions in the right-hand-side of equations are linear functions in their argument  $\mu$ . But the number of equations and the underlying space are not finite.

#### 4. Approximation

We shall develop the system (6) by requiring that only a finite number of the constraints are satisfied. This will be achieved by choosing countable sets of functions whose linear combinations are dense in the appropriate spaces. First we try to approximate the unknown part of the boundary,  $\Gamma$ , just by the finite number of points.

**Approximating  $\partial D$  with broken lines:** The idea of selecting a finite set of points instead of the curve  $\Gamma$ , comes from the approximation of a curve by broken lines. In general the curve  $\partial D$ , and hence  $\Gamma$ , can be regarded as an infinite set of points. More specifically, by applying the density property, one can regard  $\Gamma$  as a countable set. For the given  $D$  and hence for the given  $\Gamma$ , let  $A_m = (x_m, y_m)$ ,  $m = 0, 1, 2, \dots, M$ , be a finite number of these points (we suppose  $A_0 = A$ ). We link together each pair of consecutive points  $A_m$  and  $A_{m+1}$  for  $m = 0, 1, \dots, M - 1$  and close this curve by joining the points  $A_M$  and  $B$  together. Now the resulted shape, which is denoted by  $\partial D_M$ , is an approximation for  $\partial D$ ; we also call  $D_M$  to the domain which introduced by its boundary  $\partial D_M$ . The domain  $D_M$  is called a *M-approximated domain of D* (domains  $D, D_M$  and their boundaries are shown in Figure 1).

It is possible that by increasing the number of points,  $M$ , the curve  $\partial D_M$  will become closer and closer (in the Euclidean metric) to the curve  $\partial D$ , and hence one may conclude that the minimizer of  $I$  over  $\mathcal{D}_M$ , if one exists, tends to the minimizer of  $I$  over  $\mathcal{D}$ , if one exists. In the Appendix, we have explained some of the difficulties that arise. Thus, we will fix the number of points ( $M$ ) and look for the minimizer of  $I(D)$  amongst all admissible  $D_M$ 's.

Here we have actually  $2M$  unknowns to determine,  $x_1, x_2, \dots, x_M$ ,  $y_1, y_2, \dots, y_M$ . It would be more convenient if one, somehow, could reduce the number of unknowns, without losing the generality. For a given positive integer  $M$ , let the value of the components  $y_1, y_2, \dots, y_M$ , be fixed. Because  $x_m$  is a free term, the point  $A_m$  could be anywhere on the line  $y = Y_m, x \geq 0$  for every  $m$  (see Figure 1). Therefore points  $A_m$

and  $A_{m+1}$  can be chosen so that they belong to  $\Gamma$  and hence the part of  $\Gamma$  between the lines  $y = Y_m$  and  $y = Y_{m+1}$  can be approximated by the segment  $A_m A_{m+1}$  (especially whenever the number  $M$  is large). It means, we do not lose generality. Thus, from now on, we fix the components  $y_1, y_2, \dots, y_M$  with the values  $Y_1, Y_2, \dots, Y_M$ , respectively. Indeed the set  $\{A_m = (x_m, Y_m), m = 1, 2, \dots, M\}$ , which is called *M-representation of D*, determines the M-approximation domain  $D_M$ .

**First set of functions:** We are going to introduce the set  $\{\psi_i \in H_0^1(D) : i = 1, 2, \dots\}$  so that the linear combinations of the functions  $\psi_i$ s are uniformly dense - that is, dense in the topology of the uniform convergence - in the space  $H_0^1(D)$ . We know that the vector space of polynomials with the variable  $x$  and  $y$ ,  $P(x, y)$ , is dense in  $C^\infty(\overline{D})$ ; therefore the set  $P_0(x, y) = \{p(x, y) \in P(x, y) \mid p(x, y) = 0, \forall (x, y) \in \partial D\}$ , is dense (uniformly of course) in the following space:

$$\{h \in C^\infty(\overline{D}) : h|_{\partial D} = 0\} \equiv C_0^\infty(\overline{D}).$$

So the set  $Q(x, y) = \{1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3, \dots\}$  is a countable base for the vector space  $P(x, y)$  and hence each elements of  $P(x, y)$  and also  $P_0(x, y)$ , is a linear combination of the elements in  $Q(x, y)$ . By theorem 3 of Mikhailov [4] page 131, the space  $C^\infty(\overline{D})$  is dense in  $H^1(D)$ ; thus the space  $C_0^\infty(\overline{D})$  will be dense in  $H_0^1(D)$ . Consequently, the space  $P_0(x, y)$  is uniformly dense in  $H_0^1(D)$ . We define the function  $\psi_i$  for each  $i$  as:

$$\psi_i(x, y) = xy(y-1) \prod_{l=1}^M (x - x_l + y - y_l) q_i(x, y), \quad (7)$$

where  $q_i$  is an element of the countable set  $Q(x, y)$ . Therefor  $\psi|_\Gamma = 0$  and the set  $\{\psi_i(x, y) : i = 1, 2, \dots\}$ , is total in  $H_0^1(D)$ .

**Second set of functions:** Let  $L$  be a given positive integer number and divide  $D$  into  $L$  (not necessary equal) parts  $D_1, D_2, \dots, D_L$ , so that by increasing  $L$  the area of each  $D_s, s = 1, 2, \dots, L$ , will be decreased.

Then, for each  $s = 1, 2, \dots, L$ , we define:

$$\xi_s(x, y, u) = \begin{cases} 1 & \text{if } (x, y) \in D_s \\ 0 & \text{otherwise.} \end{cases}$$

These functions are not continuous, but each of them is the limit of an increasing sequence of positive continuous functions,  $\{\xi_{s_k}\}$ ; then if  $\mu$  is any positive Radon measure on  $\Omega$ ,  $\mu(\xi_s) = \lim_{k \rightarrow \infty} \mu(\xi_{s_k})$ . Now consider the set  $\{\xi_j : j = 1, 2, \dots, l\}$  of all such functions, for all positive integer  $L$ . The linear combination of these functions can approximate a function in  $C_1(\Omega)$  arbitrary well (see [9] chapter 5).

As a result, the problem (6) can be replaced by another one in which we are looking for the measure  $\mu \in \mathcal{M}^+(\Omega)$ , so that it satisfies the following constraints:

$$\begin{aligned} \mu(F_i) &= \gamma_i, & i &= 1, 2, \dots; \\ \mu(\xi_j) &= a_j, & j &= 1, 2, \dots \end{aligned} \quad (8)$$

where  $F_i \equiv F_{\psi_i}$ ,  $\gamma_i \equiv \gamma_{\psi_i}$ ,  $a_j \equiv a_{\xi_j}$ . To approximate the system of equations in (8) with a finite system of equations, first we choose a finite number of equations as follows:

$$\begin{aligned} \mu_{M_1, M_2}(F_i) &= \gamma_i, & i &= 1, 2, \dots, M_1; \\ \mu_{M_1, M_2}(\xi_j) &= a_j, & j &= 1, 2, \dots, M_2, \end{aligned} \quad (9)$$

where  $M_1$  and  $M_2$  are two positive integers. If we denote by  $Q(M_1, M_2)$  the set of positive Radon measures in  $\mathcal{M}^+(\Omega)$  which satisfy equalities (6), and also denote by  $Q$  the set of positive Radon measures in  $\mathcal{M}^+(\Omega)$  which satisfy equalities (6), by regarding the property of the total sets one can easily prove the following Proposition by considering the proof of Proposition III.1 in [9].

**Proposition 3:** *If  $M_1, M_2 \rightarrow \infty$ ; then  $Q(M_1, M_2) \rightarrow Q$ , hence for the large enough numbers  $M_1$  and  $M_2$  the set  $Q$  can be identified by  $Q(M_1, M_2)$ .*

But even if the number of equations in (6) is finite, the underlying space  $Q(M_1, M_2)$  is still infinite-dimensional. It is possible to define a finite linear system whose solutions can be used to approximate that for (6).

**Discretization:** By a result of Rosenbloom [8], which was proved in Theorem A.5 Appendix in [9], that  $\mu_{M_1, M_2}$  in (6) can be characterized as  $\mu_{M_1, M_2} = \sum_{n=1}^{M_1+M_2} \alpha_n \delta(Z_n)$ , with triples  $Z_n \in \Omega$  and the coefficients  $\alpha_n \geq 0$  for  $n = 1, 2, \dots, M_1+M_2$ , where  $\delta(z) \in \mathcal{M}^+(\Omega)$  is supposed to be a unitary atomic measure with support the singleton set  $\{z\}$ . This structural result points the way toward a further approximation scheme; the measure problem is equivalent to a nonlinear one in which the unknowns are the coefficients  $\alpha_n$  and supports  $\{Z_n\}$ . It would be more convenient if one could find the solution only with respect to the coefficients  $\alpha_n$ ; this would be a finite linear system of equations (a type of linear programming problem). The answer lies in approximating this support, by introducing a set dense in  $\Omega$ . Proposition III.3 of [9] Chapter 3, states that the measure  $\mu_{M_1, M_2}$  has the following form

$$\mu_{M_1, M_2} = \sum_{n=1}^N \alpha_n \delta(Z_n), \quad (10)$$

where  $Z_n, n = 1, 2, \dots, N$ , belongs to a dense subset of  $\Omega$ .

Now let put a discretization on  $\Omega$ , with the nodes  $Z_n = (x_n, y_n, u_n)$ , in a dense subset of  $\Omega$ ; then we can set up the following linear system in which the unknowns are the coefficients  $\alpha_n$ :

$$\begin{aligned} \alpha_n &\geq 0, & n &= 1, 2, \dots, N; \\ \sum_{n=1}^N \alpha_n F_i(Z_n) &= \gamma_i, & i &= 1, 2, \dots, M_1; \\ \sum_{n=1}^N \alpha_n \xi_j(Z_n) &= a_j, & j &= 1, 2, \dots, M_2. \end{aligned} \quad (11)$$

We remind the reader that the solution of (11) is not necessary unique, (even if the problem (1) satisfies the necessary conditions for

having a unique bounded weak solution, because of the approximation scheme. Each solution introduces a measure  $\mu_{M_1, M_2}$  via the equality (10) which has the same properties (approximately) as the measure  $\mu_u$ , the representative measure for the weak solution  $u(X)$ . Indeed we achieve an approximate solution for the elliptic problem in the given domain  $D$ . Therefore we are able to calculate the value of  $I(D)$  for each given domain  $D$ . In the next, we shall explain how one can find the optimal domain for the mentioned OS problem in  $\mathcal{D}_M$  by applying the above results.

## 5. The optimal solution

The main aim of the present section is to find an optimal domain  $D^* \in \mathcal{D}_M$  so that the value of  $I(D^*)$  will be the minimum on the set  $\mathcal{D}_M$ . By applying the result of the previous section, a solution of (1) can be found. This solution is approximated by a solution of the linear system (11) according to the variables,  $x_m, m = 1, 2, \dots, M$ . As mentioned, this solution is not necessary unique. Let us to specify one of them for each  $D$ ; there are some possibilities, for example, by solving the following linear programming problem, one may chose that one in which the value of  $\int_D f_\circ(X, u)dX$  (for a given  $D$ ) is minimum according to the variables  $\alpha_n, n = 1, 2, \dots, N$ :

$$\begin{aligned}
 \text{Minimize :} & \quad \sum_{n=1}^N \alpha_n f_\circ(Z_n) \\
 \text{Subject to :} & \quad \alpha_n \geq 0, \quad n = 1, 2, \dots, N; \\
 & \quad \sum_{n=1}^N \alpha_n F_i(Z_n) = \gamma_i, \quad i = 1, 2, \dots, M_1; \\
 & \quad \sum_{n=1}^N \alpha_n \xi_j(Z_n) = a_j, \quad j = 1, 2, \dots, M_2. \quad (12)
 \end{aligned}$$

As a result, for each  $D$ , the value  $I(D) = \int_D f_\circ(X, u) dX \equiv \mu(f_\circ) \simeq \mu_{M_1, M_2}(f_\circ)$ , is defined uniquely in terms of the variables  $x_m, m = 1, 2, \dots, M$ .

So, we set up a function,  $J$ , on  $\mathcal{D}_M$  defined by

$$J : D \in \mathcal{D}_M \longrightarrow I(D) \cong \mu_{M_1, M_2}(f_\circ) \in \mathbb{R}; \quad (13)$$

where  $\mu_{M_1, M_2}(f_\circ) = \sum_{n=1}^N \alpha_n f_\circ(Z_n)$ . By regarding the definition of M-representation of  $D$ , clearly  $J$  is a function of the variables  $x_1, x_2, \dots, x_M$ , and hence can be regarded as a vector function:

$$J : (x_1, x_2, \dots, x_M) \in \mathbb{R}^M \longrightarrow \mu_{M_1, M_2}(f_\circ) \in \mathbb{R}. \quad (14)$$

It is not possible in general to ascertain continuity properties of this function (see for instance [6]); we can say, however, that, since this is a real-valued function which is bounded below, and is defined on a compact set (since constraints are to be put in the variables), it is possible to find a sequence of points so that the value of the function along the sequence tends to the (finite) infimum of the function. The coordinate values corresponding to the points in the sequence are of course finite.

Now, suppose that  $(x_1^*, x_2^*, \dots, x_M^*)$  is the minimizer of the vector function  $J$ ; it can be identified by using one of the related minimization methods (for instance the method introduced by Nelder and Mead, see [12] and [5]). For this, one can apply standard Algorithms and Routines (like *AMOEBA* [7] or *EO4JAF-NAG* Library Routine). The introduced domain by the minimizer  $(x_1^*, x_2^*, \dots, x_M^*)$  is denoted by  $D^*$ . We assume in the following theoretical result that the minimization algorithm which is used, (such as *AMOEBA*) is perfect; that is, it comes out with the *global minimum* of  $J$  in its (compact) domain.

**Theorem 1:** *Let  $M, M_1$  and  $M_2$  be the given positive integer numbers which were defined in section 4, and  $D^*$  be the minimizer of (14) as mentioned above. Then  $D^*$  is the minimizer domain of the functional  $I$  over  $\mathcal{D}_M$  and the value of  $I(D^*)$  can be approximated by  $J(D^*)$ ; moreover  $J(D^*) \longrightarrow I(D^*)$  as  $M_1$  and  $M_2$  tend to infinity.*

**Proof:** Suppose  $D^*$  is not the minimizer of  $I$ ; hence there exists a domain, call  $D'$ , in  $\mathcal{D}_M$  so that  $I(D') < I(D^*)$ . Proposition 2 shows that there is a unique measure, call  $\mu'$ , in  $\mathcal{M}^+(\Omega)$  so that  $I(D') = \mu'(f_\circ)$ ,

and also Proposition 2 states that for sufficiently large numbers  $M_1$  and  $M_2$ ,  $\mu'(f_\circ)$  can be approximated by  $\mu'_{M_1, M_2}(f_\circ)$  in  $Q(M_1, M_2)$ . Thus,  $I(D') \cong \mu'_{M_1, M_2}(f_\circ) = J(D')$ . In the same way, one can show that  $J(D^*)$  approximates  $I(D^*)$ ; so  $I(D^*) \cong \mu^*_{M_1, M_2}(f_\circ) = J(D^*)$ . Hence  $J(D') < J(D^*)$ , which is contrary with the fact that  $D^*$  is the minimizer of  $J$ . Moreover, from Proposition 2 it follows that  $J(D^*)$  tends to  $I(D^*)$  as  $M_1, M_2 \rightarrow \infty$ .  $\square$

## 6. Numerical Examples

For the next two examples, we consider the elliptic equations (1) for which the function  $v(x, y)$  (the fixed control function) is defined as:

$$v(x, y) = \begin{cases} 1 & \text{if } (x, y) \in D \cap C \\ 0 & \text{otherwise,} \end{cases}$$

where  $C$  is the square  $[\frac{1}{4}, \frac{3}{4}] \times [\frac{1}{4}, \frac{3}{4}]$  ( see Figure 2 ). We also take  $M = 8$  and suppose  $Y_1, Y_2, \dots, Y_8$  are  $0.15, 0.25, \dots, 0.85$ , respectively. By extra constraints on  $x_2, x_3, \dots, x_7$ ,  $x_m \geq \frac{3}{4}$ ,  $m = 2, 3, \dots, 7$ , the value of  $\gamma_i$  for any  $D \in \mathcal{D}_M$  is defined as

$$\gamma_i = \int_{\frac{1}{4}}^{\frac{3}{4}} \int_{\frac{1}{4}}^{\frac{3}{4}} \psi_i(x, y) dx dy; i = 1, 2, \dots, M_1.$$

We also assume that the function  $u(\cdot)$  takes value in the bounded set  $U = [-1, 1]$  (one may obtain the set  $U$  by trial and error so as to be sure that the appropriate finite linear system in (11) has a solution).

Our way to find an optimal domain is an iterative method. For a given set of variables  $x_1 = X_1, x_2 = X_2, \dots, x_8 = X_8$ , we will set up the linear system (11) and calculate the value of  $I(D)$  according to the  $X_m$ 's. Then the standard minimization algorithm changes these  $X_1, X_2, \dots, X_8$ , to new ones for which the value of  $I(D)$  is supposed to be less than previous; these new values introduce a new domain. Again, in the next iteration, an appropriate linear system for the new domain will be solved to calculate the value of  $I(D)$  and see whether  $I(D)$  is smaller than the previous one in the former iteration or not. If the value

is not smaller, the Algorithm changes the domain with the suitable one; if it has been smaller, the Algorithm seeks again for the other domain like  $D' \in \mathcal{D}_M$  with the smaller value of  $I(D')$  than  $I(D)$ . The iteration will be stopped whenever the optimal domain is obtained; note that we assume in this discussion that the standard minimization Algorithm (*AMOEB*A) is qualified to obtain the global minimizer without any restriction (see Appendix C of [1]).

### 6.1. Nods and Equations

To establish the linear system (11) it is necessary to put a discretization on  $\Omega$ ; because our method is iterative, the discretizations depends on the values  $X_1, X_2, \dots, X_8$  at each iteration. Thus, we select  $N = 740$  nodes  $Z_n = (x_n, y_n, u_n)$  in  $\Omega$ , so that each component is a rational number; hence these nodes belong to a dense subset of  $\Omega$ . Since  $u|_{\partial D} = 0$ , for each  $(x_n, y_n) \in \partial D$ , we should have  $Z_n = (x_n, y_n, 0)$ . This fact has been taken into account in the discretization by choosing 36 related nodes. The rest of the nodes are related to the interior points of  $D$ . We consider  $Z_n = (x_n, y_n, u_n) \in D$  for  $n = 36 + 88(i - 1) + 11(j - 1) + k$  as

$$x_n = \frac{(i + 0.5)X_j}{10}, y_n = Y_j, u_n = \frac{2(k - 1)}{10} - 1$$

that  $1 \leq i \leq 8, 1 \leq j \leq 8, 1 \leq k \leq 11$ .

To set up the mentioned linear system in (11) we select  $M_1 = 10$  and  $M_2 = 8$ , and consider the polynomial  $q_i(x, y)$  form the set  $\{1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3\}$ . Also the domain  $D$  is divided into 8 parts, say  $D_1, D_2, \dots, D_8$ , as follows:  $D_1$  is the region of  $D$  between the lines  $y = 0$  and  $y = 0.2$  ( $O A e_1 o_2$  in Figure 2),  $D_2$  is the region of  $D$  between the lines  $y = 0.2$  and  $y = 0.3$  ( $o_1 e_1 e_2 o_2$  in Figure 2), and similarly  $D_3, D_4, \dots, D_7$ ; we define  $D_8$  as the region of  $D$  between the lines  $y = 0.8$  and  $y = 1$  ( $o_7 e_7 B E$  in Figure 2), where  $x_{e_l} = \frac{1}{2}(X_{l+1} - X_l) + X_l; l = 1, 2, \dots, 7$ . Therefore  $a_j = \int_D \xi_j(x, y) dX = \text{area of } D_j, \forall j = 1, 2, \dots, 8$ .

Hence in our case, the linear system (11) is

$$\begin{aligned} \alpha_n &\geq 0, & n &= 1, 2, \dots, 740; \\ \sum_{n=1}^{740} \alpha_n F_i(Z_n) &= \gamma_i, & i &= 1, 2, \dots, 10; \\ \sum_{n=1}^{740} \alpha_n \xi_j(Z_n) &= a_j, & j &= 1, 2, \dots, 8. \end{aligned} \quad (15)$$

To find the nonnegative unknowns  $\alpha_n$ 's we apply the *E04MBF* – *NAG* Library Routine Document. The result shows a nonnegative value for each  $\alpha_n, n = 1, 2, \dots, 740$ , that satisfy the linear system. By applying these values in (10), one can calculate the value of  $I(D)$  for a given function  $f_o$ , which is a function of the variables  $X_1, X_2, \dots, X_8$ ; thus we have set up the function  $J$  in (14). By applying a standard minimization algorithm on  $J$ , the optimal domain in  $\mathcal{D}_M$  is obtained. We remind the reader that the functions  $F_i$  and the values of  $\gamma_i, i = 1, 2, \dots, 10$ , have been calculated by the package “*Maple V.3*”.

## 6.2. Minimization

In minimization, we apply the Downhill Simplex Method in Multidimension by using the Subroutine *AMOEB*A (see [7]) with the conditions  $X_1 \geq 0, X_8 \geq 0$  and  $X_m \geq 0.75, m = 2, 3, \dots, 7$ ; besides, we also consider an upper bound for variables (suppose they are not higher than 2). These conditions are applied by means of a penalty method to change the constraint minimization problem into an unconstrained one (for instance see [12]).

To start, *AMOEB*A needs an initial value for variables  $X_m$ , when  $m = 1, 2, \dots, 8$ , (a given domain). At any iteration the new domain is illustrated and the new value for  $J$  is calculated; comparing this value with the previous one leads the algorithm to find a domain with a smaller value. This procedure is repeating till the optimal domain is characterized.

In the next, two examples are given; one for the linear case and the other for the nonlinear case of the elliptic equation. We chose the

function  $f_\circ$  as  $f_\circ = (u - 0.1)^2$ , this function, indeed, can be considered as a distribution of heat in the surface for the system governed by an elliptic equations.

### 6.3. Example 1

In the linear case defined by the partial differential equations (1) and  $f(x, y, u) = 0$ , the function  $F_i$  in (15) is  $F_i = u\Delta\psi_i$ ;  $i = 1, 2, \dots, 10$ . We used the initial values  $X_m = 1.0, m = 1, 2, \dots, 8$ , and the stopping tolerance for the program (variable  $ftol$  in the Subroutine *AMOEB*A) has been chosen as  $10^{-7}$ . Here are the results:

- The optimal value of  $I = 0.70469099432415$ ;
- The number of iterations = 827;
- The value of the variables in the final step:  
 $X_1 = 1.033028, X_2 = 1.390598, X_3 = 1.422364, X_4 = 0.97706,$   
 $X_5 = 1.017410, X_6 = 0.958974, X_7 = 1.018387, X_8 = 0.951333.$

These values represent the optimal domain. The initial and the final domain has been shown in the Figure 3, and also the alteration of the objective function, according to the number of iterations, has been plotted in the Figure 4.

### 6.4. Example 2

For the nonlinear case of the partial differential equations (1), we have taken  $f(x, y, u) = 0.25u^2$ , and used the same initial values and stopping tolerance as *Example 1*. The obtained results are:

- The optimal value of  $I = 0.45467920356379$ ;
- The number of iterations = 502;
- The value of the variables in the final step:  
 $X_1 = 1.050197, X_2 = 1.085212, X_3 = 0.750001, X_4 = 0.768701,$   
 $X_5 = 1.129861, X_6 = 1.137751, X_7 = 0.977838, X_8 = 1.615668,$

which represent the optimal domain, shown in the Figure 5. Also the change of the objective function, according to the number of iterations, has been plotted in the Figure 6.

## References

- [1] A. Fakharzadeh J. *Shapes, Measures And Elliptic Equations*, PhD thesis, Dept. of Applied Mathematical Studies, Leeds University, 1996.
- [2] A. Fakharzadeh J., J.E. Rubio, Shapes and Measures, *IMA Journal of Mathematical Control and Information*, 16 (1999) 207-220.
- [3] O.A. Ladyzhenskaya, N.N. Ural'tseva, *Linear and Quasilinear Elliptic Equations*. vol. 46, ACADEMIC PRESS, Mathematics in Science and Engineering, 1968.
- [4] V.P. Mikhailov, *Partial Differential Equation*, MIR Publisher, Moscow, 1978.
- [5] J.A. Nelder, R. Mead, A simplex method for function minimization, *The Computer Journal*, 7 (1964-65) 303-313.
- [6] O. Pironneau, *Optimal Shape Design for Elliptic System*, Springer-Verlag, New York - Berlin - Heidelberg - Tokyo, 1983.
- [7] W.H. Press, B.P. Flannery, S.A. Teukolsky, Vetterling. W. T. *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, 1986.
- [8] P.C. Rosenbloom, Quelques classes de problèmes extrémaux, *Bulletin de la Société Mathématique de France*, 80 (1952) 183-216.
- [9] J.E. Rubio, *Control and Optimization: The Linear Treatment of Nonlinear Problems*, Manchester University Press, Manchester, 1986.
- [10] J.E. Rubio, *Modern trends in the Calculus of Variations and Optimal Control Theory*, In R. Geise, editor, Vortagsauszuge, Mathematiker Kongress 1990, P.155-163, Berlin, Mathematische Gesellschaft.
- [11] W. Rudin, *Real and Complex Analysis*, Tata McGraw-Hill Publishing Co.Ltd, New Delhi, second edition, 1983.
- [12] G.R. Walsh, *Method of Optimization*, John Wiley and Sons Ltd., 1975.

- [13] L.C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders Company, 1969.

## 7. Appendix

### Why $\mathcal{D}_M$ instead of $\mathcal{D}$ ?

Based on the approximation of a closed and simple curve in  $\mathbb{R}^2$  by a set of broken lines, we decided to consider  $\mathcal{D}_M$  as the underlying space in which the minimization takes place. Indeed we approximated the variable part of any domain  $D \in \mathcal{D}_M$ ,  $\Gamma$ , by  $M$  number of segments (in other words by  $M + 1$  corners). As  $M \rightarrow \infty$ , if an appropriate optimal shape design problem in  $\mathcal{D}_M$  has a minimizer, then this may tend in some topology to the minimizer over  $\mathcal{D}$  if such exists. However things can go wrong; for instance: There may be no minimizer over  $\mathcal{D}_M$ , there may be no minimizer over  $\mathcal{D}$  (or both  $\mathcal{D}$  and  $\mathcal{D}_M$ ), the sequence of minimizer over  $\mathcal{D}_M$  may not be convergent or may tend in some sense towards a curve that does not define a shape.

On the other hand, let  $D_M^* \in \mathcal{D}_M$  be the optimal solution of the appropriate problem over  $\mathcal{D}_M$ , and  $\eta_M^* \in \mathcal{M}^+(\omega)$  be the optimal measure which represents the boundary of  $D_M^*$  ( $\partial D_M^*$ ); then because  $\mathcal{M}^+(\omega)$  is compact, the sequence  $\{\eta_M^*\}_{M=1}^\infty$  and hence  $\{\partial D_M^*\}_{M=1}^\infty$ , have a convergent subsequence even they are not convergent. Young in [13] has shown that their related subsequences of broken lines, tends to an infinitesimal zigzag (generalized curve). This is not (necessarily) an admissible curve (see [13] Chapter VI). So the solution over  $\mathcal{D}_M$  does not tend to the solution over  $\mathcal{D}$ , even in the weakly\*-sense. Also, there is the important point that too oscillatory boundaries (like the infinitesimal zigzag) sometimes cause problem; Pironneau in [6] shows some of these problems.

So, we prefer to fix the number of  $M$  in this paper, and search for the optimal solution of the appropriate optimal shape design problems over  $\mathcal{D}_M$ .

**8. Figures**

**Figure 1:**  $D$  and  $\partial D$  in the defined assumption.

**Figure 2:** An admissible domain  $D$  under the assumptions of the numerical work

**Figure 3:** The initial and the optimal domain for the linear case of elliptic equation.

**Figure 4:** Changes of the objective function according to iterations in the linear case.

**Figure 5:** The initial and the optimal domain for nonlinear case of elliptic equations.

**Figure 6:** Changes of the objective function according to iterations in the nonlinear case.

### روش شکل-اندازه برای حل مسائل بیضوی شکل بهینه (حالت کنترل ثابت)

هدف این نوشتار ارائه شیوه‌ای برای حل دسته مسائلی از نظریه شکل بهینه (یا طراحی شکل بهینه با فرض کنترل ثابت) است که بر پایه زوجی از عناصر هندسی تعریف شده‌اند. مسئله عبارت است از تعیین دامنه بهینه برای یک تابع داده شده بطوریکه در آن یک سیستم (خطی و یا غیر خطی) بیضوی با شرایط مرزی مورد نظر برقرار باشد. با تبدیل مسئله به مسئله‌ای از نظریه اندازه‌ها، روش شکل-اندازه، در دستگاه دکارتی قائم، بکار گرفته خواهد شد تا طی دو مرحله دامنه (شکل) بهینه شناسایی گردد. ابتدا جواب دستگاه بیضوی برای یک دامنه ثابت مفروض را با استفاده از روش نشانیدن به دست می‌آوریم. آنگاه در مرحله دوم شیوه شکل-اندازه بکار گرفته خواهد شد تا دامنه بهینه تعیین گردد. دو مثال بر حسب حالت‌های خطی و غیر خطی معادلات بیضوی نیز ارائه می‌گردد.