MEMOIRS OF NUMERICAL MATHEMATICS

NUMBER 8/9

1981/1982

The MEMOIRS OF NUMERICAL MATHEMATICS will publish original papers concerned with numerical mathematics, and normally will be published once a year.

All correspondence should be addressed to

Sin Hitotumatu (Executive Editor)
Research Institute for Mathematical Sciences
Kyoto University, Sakyo-ku, Kyoto
606, Japan

CONTENTS

An error analysis of some difference method for a singular perturbation problem

By

Koichi Niijima

1. Introduction

In his paper [2], J.J.H.Miller presented an exponentially fitted difference scheme for solving a boundary value problem with a small positive parameter ε;

$$\varepsilon y'' - b(x,\varepsilon)y = f(x,\varepsilon), \quad 0<x<1, \tag{1.1a}$$

$$y(0) = a_0, \quad y(1) = a_1. \tag{1.1b}$$

It has been shown in the paper that a solution of the scheme converges uniformly in ε to that of (1.1) on mesh points. However, several conditions which seem to be unnatural have been imposed on the coefficients. This is because the error analysis is due to the method of A.M.Il'in [1].

In this paper, we want to give a difference scheme of exponential type for the problem (1.1), and to propose a method of error estimates by which the uniform convergence in ε of our scheme can be established under week conditions.

2. Construction of a difference scheme

Let ε_0 be a small positive number, and define D by

$$D = \{(x,\varepsilon) \mid 0<x<1, \quad 0<\varepsilon<\varepsilon_0 \}.$$

We make the following assumptions on the coefficients;

A1. $b(x,\varepsilon)$ and $f(x,\varepsilon)$ are continuously differentiable with respect to x, and their derivatives are bounded on D.

A2. There exists a positive constant δ, not depending on ε, such that $b(x,\varepsilon) \geqq \delta$ in $\overline{D}$.

For constructing our difference scheme, we introduce the uniform mesh $x_i = ih$ for $i=0,\ldots,N$, where $Nh=1$, and use the abbreviations $b_i = b(x_i,\varepsilon)$ and $f_i = f(x_i,\varepsilon)$. We approximate the equation (1.1a) in subinterval (x_i, x_{i+1}) by

$$\varepsilon y_h^{(i)}{}'' - b_i y_h^{(i)} = f_i. \tag{2.1}$$

It is well known that this equation is solvable explicitly. Now, we shall connect at $x = x_i$ the solution $y_h^{(i)}(x)$ of (2.1) and that of (2.1) for i-1, together with their first derivatives. We continue this procedure from i=1 to i=N-1 to get

$$y_{h,i}^{(i-1)} = y_{h,i}^{(i)}, \quad i=1,\ldots,N-1, \tag{2.2a}$$

and

$$y_{h,i}^{(i-1)}{}' = y_{h,i}^{(i)}{}', \quad i=1,\ldots,N-1, \tag{2.2b}$$

where $y_{h,i}^{(j)}$ and $y_{h,i}^{(j)}{}'$ indicate $y_h^{(j)}(x_i)$ and $y_h^{(j)}{}'(x_i)$, respectively. For i=0,N, we impose the boundary conditions

$$y_{h,0}^{(0)} = a_0, \quad y_{h,N}^{(N-1)} = a_1. \tag{2.3}$$

We remark that conditions (2.2) and (2.3) yield a linear system concerning N-1 unknown numbers, since a solution of (2.1) contains two unknown numbers.

Although this system may be derived directly from these conditions, we try another approach for convenience of our error analysis. Define an inner product $(u,v)_i$ by

$$(u,v)_i=\int_{x_i}^{x_{i+1}} u(x)v(x)dx$$

and let $u^{(i)}(x)$ be a solution of

$$\varepsilon u^{(i)}{}'' - b_i u^{(i)} = 0. \qquad (2.4)$$

Then we always have, from (2.1),

$$(\ \varepsilon y_h^{(i)}{}'' - b_i y_h^{(i)}\ ,\ u^{(i)})_i = (f_i, u^{(i)})_i. \qquad (2.5)$$

On integrating the left hand side by parts, we obtain, by virtue of (2.4),

$$\varepsilon(u_{i+1}^{(i)}\ y_{h,i+1}^{(i)}{}' - u_i^{(i)} y_{h,i}^{(i)}{}') - \varepsilon(u_{i+1}^{(i)}{}' y_{h,i+1}^{(i)} - u_i^{(i)}{}' y_{h,i}^{(i)})$$

$$= (f_i, u^{(i)})_i. \qquad (2.6)$$

Since $u^{(i)}(x)$ may be expressed by a linear combination of the basis $\{\ \exp(-\rho_i(x-x_i)),\ \exp(\rho_i(x-x_i))\}$, where $\rho_i=\sqrt{b_i/\varepsilon}$, the equation (2.6) is equivalent to the following equations;

$$\varepsilon(\tau_i^{-1} y_{h,i+1}^{(i)}{}' - y_{h,i}^{(i)}{}') + \varepsilon\rho_i(\tau_i^{-1} y_{h,i+1}^{(i)} - y_{h,i}^{(i)}) = g_1^{(i)} \qquad (2.7a)$$

and

$$\varepsilon(\tau_i y_{h,i+1}^{(i)}{}' - y_{h,i}^{(i)}{}') - \varepsilon\rho_i(\tau_i y_{h,i+1}^{(i)} - y_{h,i}^{(i)}) = g_2^{(i)}, \qquad (2.7b)$$

where $\tau_i = \exp(\rho_i h)$, $g_1^{(i)} = (f_i, \exp(-\rho_i(x-x_i)))_i$ and $g_2^{(i)} = (f_i, \exp(\rho_i(x-x_i)))_i$. Since $\tau_i - \tau_i^{-1} \neq 0$, we may solve (2.7) in terms of the first derivatives to get

$$y_{h,i}^{(i)\prime} = \{\rho_i(2y_{h,i+1}^{(i)} - (\tau_i+\tau_i^{-1})y_{h,i}^{(i)}) + \frac{1}{\varepsilon}(\tau_i^{-1}g_2^{(i)} - \tau_i g_1^{(i)})\}/(\tau_i - \tau_i^{-1}) \tag{2.8a}$$

and

$$y_{h,i+1}^{(i)\prime} = \{\rho_i((\tau_i+\tau_i^{-1})y_{h,i+1}^{(i)} - 2y_{h,i}^{(i)}) + \frac{1}{\varepsilon}(g_2^{(i)} - g_1^{(i)})\}/(\tau_i - \tau_i^{-1}). \tag{2.8b}$$

By the matching condition (2.2), we obtain a difference scheme

$$-c_{i,i-1}y_{h,i-1} + c_{i,i}y_{h,i} - c_{i,i+1}y_{h,i+1} = d_i, \quad i=1,\ldots,N-1, \tag{2.9}$$

where we have used the abbreviation $y_{h,i} = y_{h,i}^{(i-1)} = y_{h,i}^{(i)}$. The coefficients occurred in (2.9) are defined by

$$c_{i,i-1} = 2\varepsilon\rho_{i-1}\frac{1}{\beta_{i-1}},$$

$$c_{i,i} = \varepsilon(\rho_{i-1}\frac{\alpha_{i-1}}{\beta_{i-1}} + \rho_i\frac{\alpha_i}{\beta_i}),$$

$$c_{i,i+1} = 2\varepsilon\rho_i\frac{1}{\beta_i}$$

and

$$d_i = \frac{2-\alpha_{i-1}}{\rho_{i-1}\beta_{i-1}} f_{i-1} + \frac{2-\alpha_i}{\rho_i\beta_i} f_i,$$

where $\alpha_j = \tau_j + \tau_j^{-1}$ and $\beta_j = \tau_j - \tau_j^{-1}$.

The solvability of the equation (2.9) is shown in the following lemma.

Lemma 1. The equation (2.9) subject to $y_{h,0}= a_0$ and $y_{h,N}= a_1$ has a unique solution.

Proof. Define a linear operator L by

$$Ly_h|_i= -c_{i,i-1}y_{h,i-1}+ c_{i,i}y_{h,i}- c_{i,i+1}y_{h,i+1}.$$

It is easy to show that $c_{i,i-1}$, $c_{i,i}$ and $c_{i,i+1}$ are positive. Moreover, since exp(t)+exp(-t)>2 holds for all t>0, we have $\rho_j\frac{\alpha_j}{\beta_j} > \rho_j\frac{2}{\beta_j}$, that is, $c_{i,i}>c_{i,i-1}+ c_{i,i+1}$. This shows that the operator L is positive. Therefore we can apply the maximum principle to (2.9) for getting our assertion.

Lemma 1 assures that there exists a differentiable function on (0,1) satisfying (2.1) in subinterval (x_i,x_{i+1}). In the later sections, we denote this function by $y_h(x)$.

3. Preliminaries for error estimates

The main aim of this paper is to prove the following theorem.

Theorem. Suppose that assumptions A1 and A2 are fulfilled. Then, for the solution $y_{h,i}$ of (2.9) satisfying $y_{h,0}= a_0$ and $y_{h,N}= a_1$, we have

$$\max_{i=0,\dots,N}|y_{h,i} - y_i|\leqq Mh,$$

where y_i denotes a solution of (1.1) at $x=x_i$, and M is a positive constant independent of ε and h.

For proving this theorem, we set $r(x) = y_h(x) - y(x)$ for a solution $y(x)$ of (1.1). Note that this $r(x)$ is differentiable on $(0,1)$, since $y_h(x)$ and $y(x)$ are so. Now, we always have, from (1.1a),

$$(\varepsilon y'' - by,\ u^{(i)})_i = (f, u^{(i)})_i \tag{3.1}$$

for a solution $u^{(i)}(x)$ of (2.4). Also, we may rewrite (2.5) using $y_h(x)$ as

$$(\varepsilon y_h'' - b_i y_h,\ u^{(i)})_i = (f_i, u^{(i)})_i. \tag{3.2}$$

On subtracting (3.1) from (3.2), we get

$$(\varepsilon r'' - b_i r,\ u^{(i)})_i = R_i(u^{(i)}) + S_i(u^{(i)}),$$

where $R_i(v) = ((b_i - b)y, v)_i$ and $S_i(v) = ((f_i - f)y, v)_i$. We integrate the left hand side by parts and employ (2.4) to obtain

$$\varepsilon(u_{i+1}^{(i)} r_{i+1}' - u_i^{(i)} r_i') - \varepsilon(u_{i+1}^{(i)\prime} r_{i+1} - u_i^{(i)\prime} r_i)$$
$$= R_i(u^{(i)}) + S_i(u^{(i)}). \tag{3.3}$$

The present purpose is to derive from this a relation between r_{k-1} and r_k for $k \geq 2$. We sum up (3.3) for $i=0,\ldots,k-1$ and arrange the summation to get

$$\varepsilon\{-u_0^{(0)} r_0' + \sum_{i=1}^{k-1} (u_i^{(i-1)} - u_i^{(i)}) r_i' + u_k^{(k-1)} r_k'\}$$
$$- \varepsilon\{\sum_{i=1}^{k-1} (u_i^{(i-1)\prime} - u_i^{(i)\prime}) r_i + u_k^{(k-1)\prime} r_k\} = \sum_{i=0}^{k-1} (R_i(u^{(i)}) + S_i(u^{(i)})), \tag{3.4}$$

where we have used $r_0 = 0$.

We now determine $u^{(j)}(x)$ for $j=0,\ldots,k-1$ by the matching conditions

$$u_i^{(i-1)} = u_i^{(i)}, \qquad u_i^{(i-1)\prime} = u_i^{(i)\prime}, \qquad i=1,\ldots,k-1$$

and by the boundary conditions $u_0^{(0)} = 0$ and $u_k^{(k-1)} = 1$. As seen in the preceding section, these conditions lead to

$$-c_{i,i-1}u_{i-1} + c_{i,i}u_i - c_{i,i+1}u_{i+1} = 0, \quad i=1,\ldots,k-1, \tag{3.5a}$$

$$u_0 = 0, \qquad u_k = 1, \tag{3.5b}$$

where we set $u_i = u_i^{(i-1)} = u_i^{(i)}$. The equation (3.5) is uniquely solvable by Lemma 1, and so there is a differentiable function $u(x)$ on $(0,x_k)$ which becomes a solution of (2.4) in subinterval (x_i,x_{i+1}). With this $u(x)$, (3.4) may be written as

$$\varepsilon r_k' - \varepsilon u_k' r_k = \sum_{i=0}^{k-1}(R_i(u) + S_i(u)). \tag{3.6}$$

On the other hand, we consider (3.3) for $i=k-1$. We now determine $u^{(k-1)}(x)$ so as to satisfy $u_{k-1}^{(k-1)} = 0$ and $u_k^{(k-1)} = 1$. By solving (2.4), we indeed get

$$u^{(k-1)}(x) = \{\exp(\rho_{k-1}(x-x_{k-1})) - \exp(-\rho_{k-1}(x-x_{k-1}))\}/(\tau_{k-1}-\tau_{k-1}^{-1}).$$

For brevity, we denote this function by $\hat{u}(x)$. Then (3.3) becomes

$$\varepsilon r_k' - \varepsilon(\hat{u}_k' r_k - \hat{u}_{k-1}' r_{k-1}) = R_{k-1}(\hat{u}) + S_{k-1}(\hat{u}). \tag{3.7}$$

Eliminating $\varepsilon r_k'$ from (3.6) and (3.7), we obtain

$$\varepsilon\hat{u}_{k-1}' r_{k-1} - \varepsilon(\hat{u}_k' - u_k') r_k = -\sum_{i=0}^{k-2}(R_i(u)+S_i(u)) - R_{k-1}(u-\hat{u}) - S_{k-1}(u-\hat{u}). \tag{3.8}$$

It is easily seen that $\hat{u}'_{k-1}= 2\rho_{k-1}/\beta_{k-1}$, $\hat{u}'_k= \rho_{k-1}\alpha_{k-1}/\beta_{k-1}$ and $u'_k= \rho_{k-1}\alpha_{k-1}/\beta_{k-1} - 2\rho_{k-1}u_{k-1}/\beta_{k-1}$. Therefore we have, from (3.8),

$$r_{k-1}- u_{k-1}r_k= \frac{\beta_{k-1}}{2\varepsilon\rho_{k-1}} T_{k-1}, \tag{3.9}$$

where T_{k-1} denotes the right hand side of (3.8). This is our desired relation between r_{k-1} and r_k.

In order to estimate r_j by (3.9), we need to know some properties of u_j which will be given in several lemmas below.

Lemma 2. The solution u_j of (3.5) satisfies

$$0= u_0<u_1<\ldots<u_{k-1}<u_k= 1. \tag{3.10}$$

Proof. Using the operator L introduced in Lemma 1, (3.5a) may be written as

$$Lu|_i = 0, \qquad i=1,\ldots,k-1. \tag{3.11}$$

Since L is positive, the maximum principle yields, by $u_0= 0$ and $u_k= 1$,

$$0 \leqq u_i \leqq 1, \qquad i=1,\ldots,k-1.$$

We further have $u_1>0$. Because, if $u_1=0$, then we get from (3.5a)

$$u_2= u_3=\ldots= u_k= 0,$$

which contradicts to $u_k= 1$.

Next, we show that the sequence $\{u_i\}$ is monotone increasing. Assume that for some i_0,

$$0<u_1\leqq u_2\leqq\ldots\leqq u_{i_0-1}>u_{i_0}.$$

But we can apply the maximum principle to (3.11) for $i=1,\ldots,i_0-1$ to get

$$0 \leqq u_i \leqq u_{i_0}, \quad i=1,\ldots,i_0-1.$$

This contradicts to $u_{i_0-1} > u_{i_0}$. Thus we have

$$0 < u_1 \leqq u_2 \leqq \cdots \leqq u_{k-1} \leqq 1.$$

The equality may be omitted since

$$u_i = \frac{c_{i,i-1}}{c_{i,i}} u_{i-1} + \frac{c_{i,i+1}}{c_{i,i}} u_{i+1}$$

$$\leqq \frac{c_{i,i-1} + c_{i,i+1}}{c_{i,i}} u_{i+1} < u_{i+1}.$$

The proof is completed.

Lemma 3. For the solution u_j of (3.5), we have

$$\frac{u_i}{u_{i+1}} < \frac{2}{\tau_i + \tau_i^{-1}}, \quad i=1,\ldots,k-1.$$

Proof. The inequalities $0 < u_{i-1} < u_i$ and $c_{i,i-1} > 0$ lead to

$$c_{i,i}u_i = c_{i,i-1}u_{i-1} + c_{i,i+1}u_{i+1}$$

$$< c_{i,i-1}u_i + c_{i,i+1}u_{i+1}$$

from which we get

$$\frac{u_i}{u_{i+1}} < \frac{c_{i,i+1}}{c_{i,i} - c_{i,i-1}}, \tag{3.12}$$

since $c_{i,i} > c_{i,i-1}$ and $u_{i+1} > 0$ hold. From the definition of $c_{i,j}$, it follows that

$$c_{i,i}-c_{i,i-1}=\varepsilon\rho_i\frac{\alpha_i}{\beta_i}+\varepsilon\rho_{i-1}\frac{\alpha_{i-1}-2}{\beta_{i-1}}$$

$$>\varepsilon\rho_i\frac{\alpha_i}{\beta_i}\ .$$

Combining this inequality and (3.12), we obtain

$$\frac{u_i}{u_{i+1}}<\frac{2\varepsilon\rho_i/\beta_i}{\varepsilon\rho_i\alpha_i/\beta_i}$$

$$=\frac{2}{\alpha_i}=\frac{2}{\tau_i+\tau_i^{-1}}$$

which completes the proof.

Lemma 4. Define v_i by

$$v_i=\frac{\tau^i-\tau^{-i}}{\tau^k-\tau^{-k}}\ ,\quad i=0,\ldots,k,$$

where $\tau=\exp(\rho h)$ and $\rho=\sqrt{\delta/\varepsilon}$. Then the solution u_j of (3.5) satisfies

$$u_i\lneq v_i,\quad i=1,\ldots,k-1.$$

Proof. The definition of v_i implies that $v_0=0$ and $v_k=1$. Therefore we have $v_0-u_0=0$ and $v_k-u_k=0$.

We next show that $L(v-u)|_i$ is nonnegative. An easy calculation gives

$$L(v-u)|_i=Lv|_i$$

$$=\frac{\varepsilon}{\tau^k-\tau^{-k}}\{-\frac{2\rho_{i-1}}{\beta_{i-1}}(\tau^{i-1}-\tau^{-(i-1)})$$

$$+(\rho_{i-1}\frac{\alpha_{i-1}}{\beta_{i-1}}+\rho_i\frac{\alpha_i}{\beta_i})(\tau^i-\tau^{-i})-\frac{2\rho_i}{\beta_i}(\tau^{i+1}-\tau^{-(i+1)})\}.$$

We now introduce two functions;

$$p(t) = \frac{t(e^{t}+ e^{-t})}{e^{t}- e^{-t}}$$

and

$$q(t) = \frac{t}{e^{t}- e^{-t}} .$$

It is easily seen that $p(t)$ is positive and monotone increasing for $t>0$, and $q(t)$ is positive and monotone decreasing for $t>0$. Using these auxiliary functions, $L(v-u)|_i$ may be written as

$$\begin{aligned} L(v-u)|_i = \frac{\varepsilon}{h(\tau^k-\tau^{-k})}\{&-2q(\rho_{i-1}h)(\tau^{i-1}-\tau^{-(i-1)}) \\ &+(p(\rho_{i-1}h)+ p(\rho_i h))(\tau^{i}-\tau^{-i}) \\ &-2q(\rho_i h)(\tau^{i+1}-\tau^{-(i+1)})\}. \end{aligned} \tag{3.13}$$

We can establish $L(v-u)|_i \geqq 0$ as follows;

(i) When $\rho_i \geqq \rho_{i-1}$, then $p(\rho_i h) \geqq p(\rho_{i-1}h)>0$ and $0<q(\rho_i h) \leqq q(\rho_{i-1}h)$ are valid. Therefore we have from (3.13)

$$\begin{aligned} L(v-u)|_i &\geqq \frac{\varepsilon}{h(\tau^k-\tau^{-k})}\{-2q(\rho_{i-1}h)(\tau^{i-1}-\tau^{-(i-1)})+ 2p(\rho_{i-1}h)(\tau^{i}-\tau^{-i}) \\ &\qquad - 2q(\rho_{i-1})(\tau^{i+1}-\tau^{-(i+1)})\} \\ &=\frac{2\varepsilon q(\rho_{i-1}h)}{h(\tau^k-\tau^{-k})}\{-(\tau^{i-1}-\tau^{-(i-1)})+ (\tau_{i-1}+\tau_{i-1}^{-1})(\tau^{i}-\tau^{-i}) \\ &\qquad -(\tau^{i+1}-\tau^{-(i+1)})\} . \end{aligned}$$

We further have, by virtue of $\tau_{i-1}+\tau_{i-1}^{-1} \geqq \tau+\tau^{-1}$,

$$L(v-u)|_i \geq \frac{2\varepsilon q(\rho_{i-1}h)}{h(\tau^k-\tau^{-k})}\{-(\tau^{i-1}-\tau^{-(i-1)})+(\tau+\tau^{-1})(\tau^i-\tau^{-i})$$
$$-(\tau^{i+1}-\tau^{-(i+1)})\}$$
$$= 0.$$

(ii) When $\rho_i<\rho_{i-1}$, then $0<p(\rho_i h)<p(\rho_{i-1}h)$ and $q(\rho_i h)>q(\rho_{i-1}h)>0$ are valid. We now have from (3.13)

$$L(v-u)|_i > \frac{\varepsilon}{h(\tau^k-\tau^{-k})}\{-2q(\rho_i h)(\tau^{i-1}-\tau^{-(i-1)})+ 2p(\rho_i h)(\tau^i-\tau^{-i})$$
$$- 2q(\rho_i h)(\tau^{i+1}-\tau^{-(i+1)})\}$$
$$=\frac{2\varepsilon q(\rho_i h)}{h(\tau^k-\tau^{-k})}\{-(\tau^{i-1}-\tau^{-(i-1)})+(\tau_i+\tau_i^{-1})(\tau^i-\tau^{-i})$$
$$-(\tau^{i+1}-\tau^{-(i+1)})\}.$$

Using again $\tau_i+\tau_i^{-1}\geq\tau+\tau^{-1}$, we get $L(v-u)|_i>0$.

Thus we can apply the maximum principle to get our assertion.

4. Proof of Theorem

First of all, we notice that all constants to be appeared are independent of ε and h. For proving our theorem, we return to the recurrence relation (3.9). From this relation, we get

$$|r_{k-1}|\leq u_{k-1}|r_k|+ \frac{\beta_{k-1}}{2\varepsilon\rho_{k-1}}|T_{k-1}|. \tag{4.1}$$

We can seek the bound of $|T_{k-1}|$ as follows. In the first, we obtain, from the definition of T_{k-1},

$$|T_{k-1}| \leq \sum_{i=0}^{k-2} (|R_i(u)|+|S_i(u)|)+|R_{k-1}(u-\hat{u})|+|S_{k-1}(u-\hat{u})| \quad .$$

By assumption A1, we have $|b_i - B(x,\varepsilon)| \leqq K_1 h$ on subinterval (x_i, x_{i+1}) for a positive constant K_1. We also know that a solution $y(x)$ of (1.1) satisfies $|y(x)| \leqq K_2$ on $(0,1)$ for a positive constant K_2. Moreover, it holds by Lemma 2 that in subinterval (x_i, x_{i+1}),

$$u(x) = [\{\tau_i \exp(-\rho_i(x-x_i))-\tau_i^{-1}\exp(\rho_i(x-x_i))\}u_i$$

$$+\{\exp(\rho_i(x-x_i))- \exp(-\rho_i(x-x_i))\}u_{i+1}]/(\tau_i-\tau_i^{-1})>0.$$

Accordingly, we have

$$|R_i(u)|=|\int_{x_i}^{x_{i+1}}(b_i- b(x,\varepsilon))y(x)u(x)dx|$$

$$\leqq K_1K_2h\int_{x_i}^{x_{i+1}}u(x)dx$$

$$\leqq Kh \frac{\alpha_i - 2}{\rho_i\beta_i}(u_i+ u_{i+1})$$

for a positive constant K. Similarly, we obtain

$$|S_i(u)| \leqq Kh \frac{\alpha_i - 2}{\rho_i\beta_i}(u_i+ u_{i+1}),$$

but now by using $|f_i - f(x,\varepsilon)| \leqq K_1 h$.

We turn to the estimate of $R_{k-1}(u-\hat{u})$ and $S_{k-1}(u-\hat{u})$. Since it holds on (x_{k-1}, x_k) that

$$u(x) - \hat{u}(x) = \frac{1}{\beta_k}\{\tau_{k-1}\exp(-\rho_{k-1}(x-x_{k-1})) - \tau_{k-1}^{-1}\exp(\rho_{k-1}(x-x_{k-1}))\}$$

$$\geqq 0,$$

we have

$$|R_{k-1}(u-\hat{u})| \leqq Kh\int_{x_{k-1}}^{x_k}(u(x) - \hat{u}(x))dx$$

$$= Kh\,\frac{\alpha_{k-1} - 2}{\rho_{k-1}\beta_{k-1}}\,u_{k-1}$$

and

$$|S_{k-1}(u-\hat{u})| \leqq Kh\,\frac{\alpha_{k-1} - 2}{\rho_{k-1}\beta_{k-1}}\,u_{k-1}.$$

From these estimations, it follows that

$$|T_{k-1}| \leqq 2Kh\{\sum_{i=0}^{k-2}\frac{\alpha_i - 2}{\rho_i\beta_i}(u_i + u_{i+1}) + \frac{\alpha_{k-1} - 2}{\rho_{k-1}\beta_{k-1}}u_{k-1}\}$$

$$= 4Kh\sum_{i=1}^{k-1}\left(\frac{\alpha_{i-1} - 2}{\rho_{i-1}\beta_{i-1}} + \frac{\alpha_i - 2}{\rho_i\beta_i}\right)u_i,$$

where we have used $u_0 = 0$. With $p(t)$ introduced in proving Lemma 4, we may write as

$$\frac{\alpha_j - 2}{\rho_j\beta_j} = \frac{\tau_j^{1/2} - \tau_j^{-1/2}}{\rho_j(\tau_j^{1/2} + \tau_j^{-1/2})} = \frac{h}{2p(\rho_j h/2)},$$

and already know that $p(t)$ is momotone increasing, say, $p(\rho_j h/2) \geqq p(\rho h/2)$. This leads to

$$\frac{\alpha_j - 2}{\rho_j \beta_j} \leqq \frac{h}{2p(\rho h/2)} = \frac{\tau^{1/2} - \tau^{-1/2}}{\rho(\tau^{1/2} + \tau^{-1/2})} ,$$

and so

$$|T_{k-1}| \leqq 8Kh \frac{\tau^{1/2} - \tau^{-1/2}}{\rho(\tau^{1/2} + \tau^{-1/2})} \sum_{i=1}^{k-1} u_i .$$

Combining this and (4.1), we finally get

$$|r_{k-1}| \leqq u_{k-1}|r_k| + 4Kh \frac{\tau^{1/2} - \tau^{-1/2}}{\tau^{1/2} + \tau^{-1/2}} \frac{\tau_{k-1} - \tau_{k-1}^{-1}}{\varepsilon \rho \rho_{k-1}} \sum_{i=1}^{k-1} u_i . \qquad (4.2)$$

We are now in a position to prove our theorem. It suffices to consider two cases;

(i) $h/\sqrt{\varepsilon} = h^{-\sigma}$ for $\sigma \geqq 0$.

(ii) $h/\sqrt{\varepsilon} = h^{\sigma}$ for $0 < \sigma \leqq 1$.

In case of (i), we rewrite the summation $\sum_{i=1}^{k-1} u_i$ as

$$\sum_{i=1}^{k-1} u_i = u_{k-1}\Big(1 + \sum_{i=1}^{k-2} u_i/u_{k-1}\Big) = u_{k-1}\Big(1 + \sum_{i=1}^{k-2} \prod_{j=i}^{k-2} u_j/u_{j+1}\Big) .$$

On applying Lemma 3 and on using the fact that $2/(\tau_j + \tau_j^{-1}) \leqq 2/(\tau + \tau^{-1})$, we obtain

$$\sum_{i=1}^{k-1} u_i < u_{k-1}\Big(1 + \sum_{i=1}^{k-2} \prod_{j=i}^{k-2} 2/(\tau_j + \tau_j^{-1}) \Big)$$

$$\leqq u_{k-1}\{1 + \sum_{i=1}^{k-2} (2/(\tau + \tau^{-1}))^{k-1-i}\}$$

$$< u_{k-1}/\Big(1 - \frac{2}{\tau + \tau^{-1}} \Big) .$$

Since $u_{k-1} = u_{k-1}/u_k < 2/(\tau_{k-1}+\tau_{k-1}^{-1})$ follows from Lemma 3, we further have

$$\sum_{i=1}^{k-1} u_i < \frac{2}{\tau_{k-1}+\tau_{k-1}^{-1}} / \left(1 - \frac{2}{\tau+\tau^{-1}} \right).$$

We combine this inequality and (4.2) to get

$$|r_{k-1}| < \frac{2}{\tau_{k-1}+\tau_{k-1}^{-1}} |r_k| + \frac{8Kh}{\varepsilon\rho\rho_{k-1}} \frac{\tau^{1/2}-\tau^{-1/2}}{\tau^{1/2}+\tau^{-1/2}} \frac{\tau_{k-1}-\tau_{k-1}^{-1}}{\tau_{k-1}+\tau_{k-1}^{-1}}$$

$$/ \left(1 - \frac{2}{\tau+\tau^{-1}} \right)$$

from which we obtain, by virtue of the inequalities $\varepsilon\rho\rho_{k-1} \geqq \delta$ and $\tau_{k-1}+\tau_{k-1}^{-1} \geqq \tau+\tau^{-1}$,

$$|r_{k-1}| < \frac{2}{\tau+\tau^{-1}} |r_k| + m_1 h / \left(1 - \frac{2}{\tau+\tau^{-1}} \right) \tag{4.3}$$

where $m_1 = 8K/\delta$. Using (4.3) successively by starting with $r_N = 0$, we get

$$|r_{k-1}| < m_1 h / \left(1 - \frac{2}{\tau+\tau^{-1}} \right)^2, \quad k=2,\ldots,N. \tag{4.4}$$

But since we now concern case (i), it holds that

$$\begin{aligned} \tau+\tau^{-1} &= \exp(\sqrt{\delta}/h^{\sigma}) + \exp(-\sqrt{\delta}/h^{\sigma}) \\ &\geqq \exp(\sqrt{\delta}) + \exp(-\sqrt{\delta}) \\ &\equiv \kappa > 2. \end{aligned}$$

Therefore we have from (4.4)

$$|r_{k-1}| < M_1 h, \qquad k=2,\ldots,N,$$

where $M_1 = m_1/(1-\kappa)^2$. The constant M_1 is trivially independent of ε and h.

We proceed to case (ii). In this case, we estimate the right hand side of (4.2) by using the results of Lemma 4.

It holds by Lemma 4 that

$$\sum_{i=1}^{k-1} u_i \leqq \sum_{i=1}^{k-1} \frac{\tau^i - \tau^{-i}}{\tau^k - \tau^{-k}}$$

$$= \frac{1}{\tau^k - \tau^{-k}} \left\{ \frac{\tau(\tau^{k-1}-1)}{\tau - 1} - \frac{\tau^{-1}(1-\tau^{-(k-1)})}{1-\tau^{-1}} \right\}$$

$$< \frac{1}{\tau^k - \tau^{-k}} \frac{\tau^k - \tau^{-k} - (\tau^{k-1} - \tau^{-(k-1)})}{\tau + \tau^{-1} - 2}$$

$$= \frac{1 - v_{k-1}}{\tau + \tau^{-1} - 2}$$

and $u_{k-1} \leqq v_{k-1} < 1$. Applying these estimates to (4.2), we obtain, because of $\varepsilon\rho\rho_{k-1} \geqq \delta$,

$$|r_{k-1}| < v_{k-1}|r_k| + \frac{4Kh}{\delta} \frac{\tau^{1/2} - \tau^{-1/2}}{\tau^{1/2} + \tau^{-1/2}} \frac{\tau_{k-1} - \tau_{k-1}^{-1}}{\tau + \tau^{-1} - 2} (1 - v_{k-1}).$$

Under the present condition, we have

$$\frac{\tau^{1/2} - \tau^{-1/2}}{\tau^{1/2} + \tau^{-1/2}} \frac{\tau_{k-1} - \tau_{k-1}^{-1}}{\tau + \tau^{-1} - 2} = \frac{\tau_{k-1} - \tau_{k-1}^{-1}}{\tau - \tau^{-1}}$$

$$= (\exp(\sqrt{b_{k-1}}h^\sigma) - \exp(-\sqrt{b_{k-1}}h^\sigma))/(\exp(\sqrt{\delta}h^\sigma) - \exp(-\sqrt{\delta}h^\sigma)).$$

The Taylor's expansion of exp(z) implies that the last term is bounded by a positive constant m_2 independent of ε and h. Therefore we get

$$|r_{k-1}| < v_{k-1}|r_k| + M_2(1- v_{k-1})h,$$

where $M_2 = 4Km_2/\delta$. The desired result is obtained by applying this estimate successively. Indeed, we get by starting with $r_N=0$

$$|r_{k-1}| < M_2\left(1- \prod_{j=k-1}^{N-1} v_j\right)h$$

$$< M_2 h.$$

This completes the proof of our theorem.

References

[1] A.M.Il'in, Differencing scheme for a differential equation with a small parameter affecting the highest derivative, Math. Notes Acad. Sci. USSR 6 (1969), 596-602.

[2] J.J.H.Miller, On the convergence, uniformly in ε, of difference schemes for a two point boundary singular perturbation problem, Proc. of Conf. on " The numerical Analysis of Singular Perturbation Problems ", Academic Press (1979), 467-474.

Department of Mathematics
Fukuoka Women's University
1-1 Kasumigaoka, Higashiku
Fukuoka 813, Japan

Construction of a difference scheme for some singular perturbation problem by a Liouville-Green transformation

By

Koichi Niijima

1. Introduction

Let ε be a small positive parameter and consider a singular perturbation problem of the form

$$\varepsilon y'' + a(x,\varepsilon)y' + b(x,\varepsilon)y = f(x,\varepsilon), \qquad 0\leqq x\leqq 1, \tag{1.1a}$$

$$y(0) = d_0, \quad y(1) = d_1. \tag{1.1b}$$

On $a(x,\varepsilon)$, $b(x,\varepsilon)$ and $f(x,\varepsilon)$, we make the following assumptions;

(i) The functions $a(x,\varepsilon)$, $b(x,\varepsilon)$ and $f(x,\varepsilon)$ are twice continuously differentiable with respect to x on $D = \{(x,\varepsilon) \mid 0\leqq x\leqq 1,\ 0<\varepsilon\leqq\varepsilon_0,\ \varepsilon_0\text{: small}\}$, and they are bounded on D together with partial derivatives with respect to x up to second order,

(ii) The condition $a(x,\varepsilon)\geqq\delta>0$ is satisfied on $\bar{D}$.

It is well known that a reasonable difference approximation to this problem may give inaccurate results for small ε. For example, the centered three-point difference scheme has this property. So we desire difference schemes whose solution converges to a solution of (1.1) uniformly in ε. Difference schemes of this type have been considered by A.M.Il'in [2] and recently by M.van Verdhuizen [7], in case of $b(x,\varepsilon)=0$, and by J.J.H.Miller [3] and K.V.Emel'yanov [1], in case of $b(x,\varepsilon)\leqq 0$ in D. However, the rate of convergence is of order h except for the method of Verdhuizen, where h denotes a mesh step.

The aim of this paper is to construct a difference scheme for the problem (1.1), whose solution converges to that of (1.1) uniformly in ε with order h^2, by using a Liouville-Green transformation. In our analysis, $b(x,\varepsilon)$ is not assumed to be nonpositive in D. The procedure of constructing our scheme is exactly the same as in K.Niijima [5],[6], but the error analysis requires detailed estimates concerning a solution of an analogous problem to (1.1). These estimates can be derived not from the maximum-principle but from Lemma 1 in K.Niijima [4].

In our approach, we can also obtain the value of the approximation at any points between the nodes, and its accuracy is of order h^2 uniformly in ε. This point is different from the results of Veldhuizen [7].

Throughout this paper, C_i denotes a constant independent of ε and the symbol C is used in common as a positive constant not depending on ε and h.

2. Estimates for a solution of an analogous problem to (1.1)

In this section, we give detailed estimates concerning the solution $\bar{y}(x)$ of the problem

$$\varepsilon\bar{y}'' + \bar{a}(x,\varepsilon)\bar{y}' + \bar{b}(x,\varepsilon)\bar{y} = \bar{f}(x,\varepsilon), \quad 0 \leq x \leq 1,$$
$$\bar{y}(0) = \bar{d}_0, \quad \bar{y}(1) = \bar{d}_1, \tag{2.1}$$

where $\bar{a}(x,\varepsilon)$, $\bar{b}(x,\varepsilon)$ and $\bar{f}(x,\varepsilon)$ are continuous and bounded on D, and the condition $\bar{a}(x,\varepsilon) \geq \bar{\delta} > 0$ is satisfied in $\bar{D}$.

Lemma 1. For the solution $\bar{y}(x)$ of (2.1), we have

$$|\bar{y}(x)| \leq C_1 \left(\int_0^1 \int_0^t \frac{|\bar{f}(s,\varepsilon)|}{\varepsilon} \exp\left(-\frac{\bar{\delta}}{\varepsilon}(t-s)\right) ds\,dt + |\bar{d}_0| + |\bar{d}_1| \right).$$

Proof. Applying Lemma 1 in [4] to an initial value problem associated with (2.1), we obtain

$$\bar{y}(x) = p(x)\left(\bar{y}'(0)\lambda(x) + \frac{1}{\varepsilon}\mu(x) + \bar{d}_0\right).$$

Here $\lambda(x)$ and $\mu(x)$ are

$$\lambda(x) = \int_0^x \frac{q(t)}{p^2(t)}\, dt$$

and

$$\mu(x) = \int_0^x \frac{q(t)}{p^2(t)} \int_0^t \bar{f}(s,\varepsilon) \frac{p(s)}{q(s)}\, ds\,dt,$$

respectively, where

$$q(x) = \exp\left(-\int_0^x \frac{\bar{a}(t,\varepsilon)}{\varepsilon} dt\right)$$

and

$$p(x) = \exp\left(\int_0^x \alpha(t)\,dt\right),$$

$\alpha(x)$ being a solution of

$$\alpha'(x) + \frac{\bar{a}(x,\varepsilon)}{\varepsilon}\alpha(x) + \alpha^2(x) + \frac{\bar{b}(x,\varepsilon)}{\varepsilon} = 0$$

subject to

$$\alpha(0) = 0.$$

The value $\bar{y}'(0)$ is determined by the end condition to get

$$\bar{y}(x) = (1 - \frac{\lambda(x)}{\lambda(1)})p(x)\bar{d}_0 + \frac{\lambda(x)}{\lambda(1)}\frac{p(x)}{p(1)}\bar{d}_1 + \frac{p(x)}{\varepsilon}(\mu(x) - \frac{\lambda(x)}{\lambda(1)}\mu(1)). \tag{2.2}$$

The bound for $|\bar{y}(x)|$ is derived as follows. As was shown in the proof of Lemma 1 in [4], there exists a constant C_2 such that

$$|\alpha(x)| \leqq C_2,$$

from which the bound

$$0 < C_3 \leqq p(x) \leqq C_4$$

is obtained. Combining this with the estimate $0 < q(x) \leqq \exp(-\frac{\bar{\delta}}{\varepsilon}x)$, we get

$$|\mu(x)| \leqq C_5 \int_0^x \int_0^t |\bar{f}(s,\varepsilon)| \exp(-\frac{\bar{\delta}}{\varepsilon}(t-s))\,ds\,dt.$$

Noticing here that $0 \leqq \frac{\lambda(x)}{\lambda(1)} \leqq 1$, we have the desired result.

Corollary. The derivative $\bar{y}'(x)$ satisfies the following estimate:

$$|\bar{y}'(x)| \leqq c_6 (1 + \frac{1}{\varepsilon} \exp(- \frac{\bar{\delta}}{\varepsilon}x)).$$

Proof. This estimate is obtained by differentiating $\overline{y}(x)$ in (2.2) and by noting that $\lambda(1) = O(\varepsilon)$.

3. Approximation to the problem (1.1)

We begin by approximating $a(x,\varepsilon)$, $b(x,\varepsilon)$ and $f(x,\varepsilon)$ which appear in the equation (1.1a). Let N be a positive integer and define a mesh step h by $h=1/N$. We also define equidistant mesh points x_i by $x_i=ih$, $i=0,1,\cdots,N$. For later use, we shall introduce two functions $g(x,\varepsilon)=b(x,\varepsilon)/a^2(x,\varepsilon)$ and $c(x,\varepsilon)= 1/a(x,\varepsilon)$. In subinterval $[x_i,x_{i+1}]$, we approximate $a(x,\varepsilon)$, $b(x,\varepsilon)$ and $f(x,\varepsilon)$ by

$$A(x,\varepsilon) = 1/(\alpha_i(x-x_i)^2 + \beta_i(x-x_i) + \gamma_i),$$

$$B(x,\varepsilon) = -(\alpha_i(x-x_i) + m_i - 1)A^2(x,\varepsilon)$$

and

$$F(x,\varepsilon) = A(x,\varepsilon)^{3/2}\exp(\frac{k_i}{2}\int_{x_i}^{x} A(t,\varepsilon)dt)$$

$$\cdot(\zeta_i(\phi(x)-\phi(x_i)) + \omega_i), \tag{3.1}$$

respectively. Here α_i, β_i, γ_i, m_i and k_i are

$$\alpha_i = -(g_{i+1}-g_i)/h,$$

$$\beta_i = (c_{i+1}-c_i)/h + g_{i+1}-g_i,$$

$$\gamma_i = c_i,$$

$$m_i = 1 - g_i$$

and

$$k_i = 2 - (c_{i+1}-c_i)/h - (g_{i+1}+g_i),$$

where $g_i=g(x_i,\varepsilon)$ and $c_i=c(x_i,\varepsilon)$. The function $\phi(x)$ included in (3.1) denotes

$$\phi(x) = \phi(x_i) + \int_{x_i}^{x} A(t,\varepsilon)dt \tag{3.2}$$

for $x_i \leqq x \leqq x_{i+1}$, and ζ_i and ω_i denote

$$\zeta_i = \frac{1}{s_i}\left(\exp(-\frac{k_i s_i}{2})\gamma_{i+1}^{3/2} f_{i+1} - \gamma_i^{3/2} f_i\right)$$

and

$$\omega_i = \gamma_i^{3/2} f_i,$$

respectively, where $s_i=\int_{x_i}^{x_{i+1}} A(t,\varepsilon)dt$. It follows at once that for $i=0,1,\cdots,N$,

$$A(x_i,\varepsilon) = a(x_i,\varepsilon),$$
$$B(x_i,\varepsilon) = b(x_i,\varepsilon) \tag{3.3}$$

and

$$F(x_i,\varepsilon) = f(x_i,\varepsilon).$$

This shows that the functions $A(x,\varepsilon)$, $B(x,\varepsilon)$ and $F(x,\varepsilon)$ are continuous on D. We further have the following lemma:

Lemma 2. The estimates

$$|A(x,\varepsilon) - a(x,\varepsilon)| \leqq Ch^2,$$

$$|B(x,\varepsilon) - b(x,\varepsilon)| \leqq Ch^2$$

and

$$|F(x,\varepsilon) - f(x,\varepsilon)| \leq Ch^2$$

hold on D.

Proof. From (3.3), it suffices to prove the above estimates only in subinterval $[x_i, x_{i+1}]$. In the first, we have, by virtue of the Taylor's theorem,

$$\begin{aligned}\alpha_i(x-x_i)^2+\beta_i(x-x_i)+\gamma_i &= c_i+((c_{i+1}-c_i)/h+g_{i+1}-g_i)(x-x_i) \\ &\qquad + O(h^2) \\ &= c_i+c_i'(x-x_i)+O(h^2) \\ &= c(x,\varepsilon) + O(h^2)\end{aligned}$$

which establishes the first estimate. This fact and the Taylor's theorem again lead to

$$\begin{aligned}B(x,\varepsilon) &= (-\alpha_i(x-x_i) + 1 - m_i)a^2(x,\varepsilon) + O(h^2) \\ &= \left(\frac{b(x,\varepsilon)}{a^2(x,\varepsilon)} + O(h^2)\right)a^2(x,\varepsilon) + O(h^2) \\ &= b(x,\varepsilon) + O(h^2)\end{aligned}$$

which asserts the second estimate. Before proving the last estimate, we note that $\phi(x)-\phi(x_i)= O(h)$ holds for $x_i \leq x \leq x_{i+1}$, and $\gamma_{i+1}^{3/2} f_{i+1} - \gamma_i^{3/2} f_i = O(h)$ and $k_i = O(1)$. By noticing further that $\phi(x)-\phi(x_i)=\int_{x_i}^{x} A(t,\varepsilon)dt$ and $\exp(-\frac{k_i s_i}{2})=1-\frac{k_i s_i}{2} + O(h^2)$, we have

$$
\begin{aligned}
F(x,\varepsilon) &= A(x,\varepsilon)^{3/2}\{1 + \frac{k_i}{2}(\phi(x)-\phi(x_i)) + O(h^2)\} \\
&\quad \cdot [\{(1-\frac{k_i s_i}{2}+O(h^2))\gamma_{i+1}^{3/2} f_{i+1}-\gamma_i^{3/2} f_i\}\frac{\phi(x)-\phi(x_i)}{s_i} + \gamma_i^{3/2} f_i] \\
&= A(x,\varepsilon)^{3/2}\{1 + \frac{k_i}{2}(\phi(x)-\phi(x_i))\} \\
&\quad \cdot [\{1-\frac{k_i}{2}(\phi(x)-\phi(x_i))-(1-\frac{\phi(x)-\phi(x_i)}{s_i})\}\gamma_{i+1}^{3/2} f_{i+1} \\
&\quad + (1-\frac{\phi(x)-\phi(x_i)}{s_i})\gamma_i^{3/2} f_i] + O(h^2) \\
&= A(x,\varepsilon)^{3/2}\{(1+O(h^2))\gamma_{i+1}^{3/2} f_{i+1} \\
&\quad - (1+\frac{k_i}{2}(\phi(x)-\phi(x_i)))(1-\frac{\phi(x)-\phi(x_i)}{s_i})(\gamma_{i+1}^{3/2} f_{i+1}-\gamma_i^{3/2} f_i)\} \\
&\quad + O(h^2) \\
&= A(x,\varepsilon)^{3/2}\{\gamma_i^{3/2} f_i+\frac{\phi(x)-\phi(x_i)}{s_i}(\gamma_{i+1}^{3/2} f_{i+1}-\gamma_i^{3/2} f_i)\} + O(h^2) \\
&= a(x,\varepsilon)^{3/2}\{\gamma_i^{3/2} f_i + \frac{x-x_i}{h}(\gamma_{i+1}^{3/2} f_{i+1}-\gamma_i^{3/2} f_i)\}+ O(h^2),
\end{aligned}
$$

where, in the last step, we have used the equality $(\phi(x)-\phi(x_i))/s_i=(x-x_i)/h + O(h)$. Since the term in the brackets gives a polygonal interpolation to $f(x,\varepsilon)/a(x,\varepsilon)^{3/2}$, we get the last estimate.

Corresponding to (1.1), we consider the problem

$$
\varepsilon Y'' + A(x,\varepsilon)Y' + B(x,\varepsilon)Y = F(x,\varepsilon), \quad 0\leq x\leq 1, \tag{3.4a}
$$

$$
Y(0) = d_0, \quad Y(1) = d_1. \tag{3.4b}
$$

We have the following theorem.

Theorem. Let $y(x)$ be a solution of (1.1), and $Y(x)$ a solution of (3.4). Then we have, in D,

$$|Y(x) - y(x)| \leqq Ch^2.$$

Proof. Putting $u(x) = Y(x) - y(x)$, it satisfies

$$\begin{aligned} &\varepsilon u'' + A(x,\varepsilon)u' + B(x,\varepsilon)u \\ &\qquad = F(x,\varepsilon) - f(x,\varepsilon) - (A(x,\varepsilon) - a(x,\varepsilon))y' - (B(x,\varepsilon) - b(x,\varepsilon))y \end{aligned} \tag{3.5}$$

and

$$u(0) = u(1) = 0.$$

Since $A(x,\varepsilon)$, $B(x,\varepsilon)$ and the right hand side of (3.5) are continuous on D and there exists a constant $\bar{\delta}$ such that $A(x,\varepsilon) \geqq \bar{\delta}$ in $\bar{D}$ for $\delta \geqq \bar{\delta} > 0$, we can apply Lemma 1 to the above problem to conclude that

$$|u(x)| \leqq Ch^2 \int_0^1 \int_0^t \frac{1+|y'(s)|}{\varepsilon} \exp(-\frac{\bar{\delta}}{\varepsilon}(t-s))\,ds\,dt,$$

where the results of Lemma 2 have also been used. By virtue of the estimate

$$|y'(x)| \leqq C_6 (1 + \frac{1}{\varepsilon} \exp(-\frac{\delta}{\varepsilon} x))$$

which follows by applying Corollary in Section 2 to (1.1), we have further

$$|u(x)| \leqq Ch^2 \int_0^1 \int_0^t (\frac{1}{\varepsilon} + \frac{1}{\varepsilon^2} \exp(-\frac{\delta}{\varepsilon}s)) \exp(-\frac{\bar{\delta}}{\varepsilon}(t-s))\,ds\,dt.$$

The proof is completed by noticing that the integral term is bounded by C because of $\delta \geq \bar{\delta}$.

4. A three-point difference scheme

The purpose of this section is to derive a three-point difference scheme between $Y_{i-1} = Y(x_{i-1})$, $Y_i = Y(x_i)$ and $Y_{i+1} = Y(x_{i+1})$, where $Y(x)$ is a solution of (3.4). On $[x_i, x_{i+1}]$, we change (3.4a), by the Liouville-Green transformation

$$z = \phi_i(x), \qquad v_i = \psi_i(x) Y(x),$$

into

$$\varepsilon \frac{d^2 v_i}{dz^2} + \{\frac{A(x,\varepsilon)}{\phi_i'} + \frac{\varepsilon}{\phi_i'^2}(\phi_i'' - 2\frac{\psi_i'}{\psi_i}\phi_i')\}\frac{dv_i}{dz}$$

$$+ \{\frac{1}{\phi_i'^2}(B(x,\varepsilon) - \frac{\psi_i'}{\psi_i}A(x,\varepsilon)) + \frac{\varepsilon}{\phi_i'^2}((\frac{\psi_i'}{\psi_i})^2 - (\frac{\psi_i'}{\psi_i})')\}v_i$$

$$= \frac{\psi_i}{\phi_i'^2} F(x,\varepsilon). \tag{4.1}$$

We shall determine ϕ_i' and ψ_i by the differential equations

$$\phi_i'' - 2\frac{\psi_i'}{\psi_i}\phi_i' = k\phi_i'^2$$

and

$$(\frac{\psi_i'}{\psi_i})^2 - (\frac{\psi_i'}{\psi_i})' = \ell\phi_i'^2$$

with some constants k and ℓ. As was shown in [6], ϕ_i' and ψ_i'/ψ_i are given by

$$\phi_i'(x) = 1/(\alpha(x-x_i)^2 + \beta(x-x_i) + \gamma)$$

and

$$\psi_i'(x)/\psi_i(x) = -(\alpha(x-x_i)+ m)/(\alpha(x-x_i)^2+\beta(x-x_i)+ \gamma),$$

respectively, where $m=(\beta+k)/2$ and $\ell=\alpha\gamma-m(\beta-m)$. Now, we choose $\alpha=\alpha_i$, $\beta=\beta_i$, $\gamma=\gamma_i$ and $k=k_i$. Then we have

$$A(x,\varepsilon) = \phi_i'(x) \tag{4.2}$$

and

$$B(x,\varepsilon) = \phi_i'^2(x) + \frac{\psi_i'(x)}{\psi_i(x)} A(x,\varepsilon)$$

which imply that (4.1) may be written as

$$\varepsilon \frac{d^2v_i}{dz^2} + (1 + \varepsilon k_i)\frac{dv_i}{dz} + (1 + \varepsilon\ell_i)v_i = \frac{\psi_i(x)}{\phi_i'^2(x)} F(x,\varepsilon). \tag{4.3}$$

Notice here that $\phi(x)$ in (3.2) is a function obtained by solving (4.2) for each i and by connecting them continuously. It follows from the inequality $A(x,\varepsilon)\geqq\bar{\delta}>0$ that $z=\phi(x)$ is continuous and monotonically increasing on $[0,1]$, and so it has an inverse $x=\phi^{-1}(z)$. Therefore the right hand side of (4.3) may be written as $\zeta_i(z-\phi(x_i))+ \omega_i$ with ζ_i and ω_i which have already been given in Section 3. Hence (4.3) is solvable analytically and the solution $v_i(z)$ takes the form

$$v_i(z) = K_1\exp(r_1^{(i)}(z-\phi(x_i)))+ K_2\exp(r_2^{(i)}(z-\phi(x_i)))+ w_i(z),$$

where $r_1^{(i)}$ and $r_2^{(i)}$ are roots of the equation

$$\varepsilon r^2 + (1 + \varepsilon k_i)r + 1 + \varepsilon \ell_i = 0,$$

and $w_i(z)$ denotes a particular solution of (4.3). Thus the solution $Y(x)$ of (3.4) may be written, in $[x_i, x_{i+1}]$, as

$$\begin{aligned} Y(x) &= v_i(\phi(x))/\psi_i(x) \\ &= (\alpha_i(x-x_i)^2+\beta_i(x-x_i)+\gamma_i)^{1/2} \\ &\quad \cdot [K_1\exp\{(r_1^{(i)}+\tfrac{k_i}{2})(\phi(x)-\phi(x_i))\}+K_2\exp\{(r_2^{(i)}+\tfrac{k_i}{2})(\phi(x)-\phi(x_i))\} \\ &\qquad + \exp(\tfrac{k_i}{2}(\phi(x)-\phi(x_i))w_i(\phi(x))]. \end{aligned}$$

The first step of obtaining our difference scheme is to express K_1 and K_2 by Y_i and Y_{i+1}. This is accomplished by solving the system

$$\sqrt{\gamma_i}(K_1 + K_2 + w_{i,i}) = Y_i,$$

$$\sqrt{\gamma_{i+1}}(K_1\exp(r_1^{(i)}s_i) + K_2\exp(r_2^{(i)}s_i) + w_{i,i+1}) = \exp(-\frac{k_i s_i}{2})Y_{i+1},$$

where $w_{i,j} = w_i(\phi(x_j))$. The next step is to match the first derivative of $Y(x)$ in $[x_i, x_{i+1}]$ with that of $Y(x)$ in $[x_{i-1}, x_i]$ at the node x_i. The matching condition yields

$$\begin{aligned} &(r_1^{(i)} + m_i)K_1 + (r_2^{(i)} + m_i)K_2 + m_i w_{i,i} + \frac{dw_{i,i}}{dz} \\ &= \{(r_1^{(i-1)}+\alpha_{i-1}h+m_{i-1})\exp(r_1^{(i-1)}s_{i-1})L_1 \\ &\quad + (r_2^{(i-1)}+\alpha_{i-1}h+m_{i-1})\exp(r_2^{(i-1)}s_{i-1})L_2 + (\alpha_{i-1}h+m_{i-1})w_{i-1,i} \\ &\qquad + \frac{dw_{i-1,i}}{dz}\}\xi_{i-1}, \end{aligned} \tag{4.5}$$

where $m_i=(\beta_i+k_i)/2$, $\xi_i=\exp(\frac{k_i s_i}{2})$, $\frac{dw_{i,j}}{dz} = \frac{dw_i(z)}{dz}\Big|_{z=\phi(x_j)}$ and

L_1 and L_2 are, respectively, equal to K_1 and K_2 in which the index i is replaced by i-1. Substituting these K_1, K_2, L_1 and L_2 into (4.5), we have

$$\kappa^{(i-1)}\tau_1^{(i-1)}\tau_2^{(i-1)}\xi_{i-1}a_{i-1}^{1/2}\,Y_{i-1} + (\eta^{(i-1)}+\eta^{(i)}+r_1^{(i)}+r_2^{(i)})\cdot a_i^{1/2}\,Y_i$$

$$+ \kappa^{(i)}/\xi_i\cdot a_{i+1}^{1/2}\,Y_{i+1} = \kappa^{(i-1)}\tau_1^{(i-1)}\tau_2^{(i-1)}\xi_{i-1}w_{i-1,i-1}$$

$$+ \eta^{(i-1)}\xi_{i-1}w_{i-1,i} + (\eta^{(i)}+r_1^{(i)}+r_2^{(i)})w_{i,i}$$

$$+ \kappa^{(i)}w_{i,i+1} + \xi_{i-1}\frac{dw_{i-1,i}}{dz} - \frac{dw_{i,i}}{dz},$$

where $\tau_j^{(i)}=\exp(r_j^{(i)}s_i)$, $\kappa^{(i)}=(r_2^{(i)}-r_1^{(i)})/(\tau_2^{(i)}-\tau_1^{(i)})$ and $\eta^{(i)}=(r_1^{(i)}\tau_1^{(i)}-r_2^{(i)}\tau_2^{(i)})/(\tau_2^{(i)}-\tau_1^{(i)})$. To calculate the right hand side, we must seek $w_i(z)$. Since the right hand side of (4.3) is a linear function of z and since $r_1^{(i)}r_2^{(i)}=(1+\varepsilon\ell_i)/\varepsilon\neq 0$ for sufficiently small ε, $w_i(z)$ takes the form

$$w_i(z) = \frac{\zeta_i}{\varepsilon r_1^{(i)}r_2^{(i)}}(z-\phi(x_i)) + \frac{\zeta_i(r_1^{(i)}+r_2^{(i)})}{\varepsilon r_1^{(i)2}r_2^{(i)2}} + \frac{\omega_i}{\varepsilon r_1^{(i)}r_2^{(i)}}.$$

After a careful computation, we finally get

$$\frac{t_{i-1}}{2\varepsilon\sinh(\sigma_{i-1})}\exp(-\frac{s_{i-1}}{2\varepsilon})a_{i-1}^{1/2}\,Y_{i-1}$$

$$+\{(c_{i+1}-2c_i+c_{i-1})/2h+(g_{i+1}-g_{i-1})/2-\frac{t_{i-1}}{2\varepsilon}\coth(\sigma_{i-1})-\frac{t_i}{2\varepsilon}\coth(\sigma_i)\}$$

$$\cdot a_i^{1/2}Y_i$$

$$+\ \frac{t_i}{2\varepsilon\sinh(\sigma_i)}\ \exp(\frac{s_i}{2\varepsilon})a_{i+1}^{1/2}Y_{i+1}$$

$$=\frac{t_{i-1}}{2\varepsilon\sinh(\sigma_{i-1})}\ \exp(-\frac{s_{i-1}}{2\varepsilon})(\omega_{i-1}/\bar{\ell}_{i-1}\ -\ n_{i-1}\zeta_{i-1})$$

$$-(\ \frac{t_{i-1}}{2\varepsilon}\ \coth(\sigma_{i-1})-\ \frac{\bar{k}_{i-1}}{2\varepsilon})(\omega_i/\bar{\ell}_{i-1}\ -\ n_{i-1}\xi_{i-1}\zeta_{i-1})$$

$$-(\ \frac{t_i}{2\varepsilon}\ \coth(\sigma_i)+\ \frac{\bar{k}_i}{2\varepsilon})(\omega_i/\bar{\ell}_i\ -\ n_i\zeta_i)\ +\ \frac{t_i}{2\varepsilon\sinh(\sigma_i)}\exp(\frac{s_i}{2\varepsilon})$$

$$\cdot(\omega_{i+1}/\bar{\ell}_i-\ n_i\xi_i\zeta_i)$$

$$+\ \xi_{i-1}\zeta_{i-1}/\bar{\ell}_{i-1}\ -\ \zeta_i/\bar{\ell}_i,$$

where $\sigma_i = t_i s_i/2\varepsilon$, t_i being $\sqrt{\bar{k}_i^{\,2}-4\varepsilon\bar{\ell}_i}$, and $\bar{k}_i=1+\varepsilon k_i$, $\bar{\ell}_i=1+\varepsilon\ell_i$ and $n_i=\bar{k}_i/\bar{\ell}_i^{\,2}$. This gives our difference scheme.

5. Numerical experiments

In this section, some numerical experiments are performed with our difference scheme and the computed solution is compared with an exact one. In each table below, only the maximum error at the nodes is listed. The experiments were carried out for N=8, 16 and 32.

The first problem is

$$\varepsilon y'' + \left(\frac{2\varepsilon}{1+x} + \frac{2}{(1+x)^2} \right) y' = -\frac{2\pi}{(1+x)^4}\left(\sin(\frac{\pi s}{2}) + \frac{\pi\varepsilon}{2}\cos(\frac{\pi s}{2})\right),$$

$$y(0) = y(1) = 0$$

with the exact solution

$$y(x) = \cos(\frac{\pi s}{2}) + \frac{\exp(-1/\varepsilon)-\exp(-s/\varepsilon)}{1-\exp(-1/\varepsilon)},$$

where $s=2x/(1+x)$. This problem was used in experiments by Veldhuizen [7].

	N = 8	N = 16	N = 32
$\varepsilon=10^{-1}$	0.76(-2)	0.20(-2)	0.54(-3)
10^{-2}	0.65(-2)	0.22(-2)	0.60(-3)
10^{-3}	0.64(-2)	0.22(-2)	0.63(-3)
10^{-4}	0.65(-2)	0.22(-2)	0.64(-3)
10^{-5}	0.64(-2)	0.22(-2)	0.63(-3)
10^{-6}	0.64(-2)	0.22(-2)	0.65(-3)
10^{-7}	0.64(-2)	0.22(-2)	0.63(-3)
10^{-8}	0.64(-2)	0.22(-2)	0.63(-3)

The next problem is

$$\varepsilon y'' + \frac{2}{(1+x)^2} y' - \frac{4}{(1+x)^3} y = -\frac{4}{(1+x)^4}\Big\{(1+x)\frac{\exp(-1/\varepsilon)}{1-\exp(-1/\varepsilon)}$$

$$+ (1+x + \frac{\pi^2\varepsilon}{4})\cos(\frac{\pi s}{2})$$

$$+ \frac{\pi}{2}(1-(1+x)\varepsilon)\sin(\frac{\pi s}{2})\Big\},$$

$$y(0) = y(1) = 0$$

with the same solution as above, where $s=2x/(1+x)$.

	N = 8	N = 16	N = 32
$\varepsilon=10^{-1}$	0.22(-2)	0.59(-3)	0.18(-3)
10^{-2}	0.22(-2)	0.91(-3)	0.28(-3)
10^{-3}	0.22(-2)	0.99(-3)	0.39(-3)
10^{-4}	0.22(-2)	0.99(-3)	0.39(-3)
10^{-5}	0.22(-2)	0.99(-3)	0.40(-3)
10^{-6}	0.22(-2)	0.99(-3)	0.40(-3)
10^{-7}	0.22(-2)	0.99(-3)	0.40(-3)
10^{-8}	0.22(-2)	0.99(-3)	0.40(-3)

References

[1] K.V.Emel'yanov, A difference scheme for an ordinary differential equation with a small parameter, U.S.S.R. Comput. Maths Math. Phys, 18 (1979), 72-80.

[2] A.M.Il'in, Differencing scheme for a differential equation with a small parameter affecting the highest derivative, Math. Notes Acad. Sci., U.S.S.R., 6 (1969), 596-602.

[3] J.J.H.Miller, On the convergence, uniformly in ε, of difference schemes for a two point boundary singular perturbation problem, Proc. of Conf. on "The numerical Analysis of Singular Perturbation Problems", Academic Press (1979), 467-474.

[4] K.Niijima, On the behavior of solutions of a singularly perturbed boundary value problem with a turning point, SIAM J. Math. Anal., 9 (1978), 298-311.

[5] K.Niijima, On a three-point difference scheme for a singular perturbation problem without a first derivative term. I, to appear in Memoir of Numerical Mathematics.

[6] K.Niijima, On a three-point difference scheme for a singular perturbation problem without a first derivative term. II, to appear in Memoir of Numerical Mathematics.

[7] M.van Veldhuizen, Higher order schemes of positive type for singular perturbation problems, Proc. of Conf. on " The numerical Analysis of Singular Perturbation Problems", Academic Press (1979), 361-383.

Department of Mathematics
Fukuoka Women's University
1-1 Kasumigaoka, Higashiku
Fukuoka 813, Japan

On the Viscous Shallow - Water Equations I
-- Derivation and Conservation Laws --

By

Hiroshi KANAYAMA* and Teruo USHIJIMA**

Synopsis

In this paper, we derive the two-dimensional viscous shallow-water equations, starting from the three-dimensional Reynolds equations. It is also shown that the derived equations have the horizontal viscosity terms hereditary from the original Reynolds equations and that mass and energy conservation laws are satisfied under physically plausible conditions.

1. Introduction

The aim of this survey paper is to clear the derivation of the two-dimensional viscous shallow-water equations, to show the complete forms of the horizontal viscosity terms, and to establish conservation properties of mass and energy in the derived equations.

---

* Systems Research Section, FUJIFACOM CORPORATION, 1-1, Tanabeshinden, Kawasaki-shi, Kanagawa, 210, Japan.

** Department of Information Mathematics, The University of Electro-Communications, 1-5-1, Chofugaoka, Chofu-shi, Tokyo, 184, Japan.

Computer simulation of the tidal motion in a bay has often been studied in recent years. It is well known that the two-dimensional viscous shallow-water equations are the basic ones for the tidal current.[1] However, a newcomer into this field would be puzzled by the number of apparently different mathematical models.[5] Especially, the horizontal viscosity terms are written in various forms. It seems that one of the reasons for this confusion is that there are many researchers who do not pay much attention to these terms because they are usually small in comparison with the vertical viscosity terms.[1] In the above situation, the authors believe that it is very important to clear the derivation of the viscous shallow-water equations and to show the complete forms of the horizontal viscosity terms. This is partly because, from the mathematical point of view, it is necessary to set suitable boundary conditions for the basic equations to be well-posed, which may depend on the forms of the viscosity terms.

Also, this is partly because, from the numerical point of view, it is important to distinguish the artificial viscosity produced by numerical schemes from the original viscosity in the basic equations. In fact, these necessity and importance of mathematical and numerical analysis of shallow-water phenomena are stressed by many engineers and scientists.

For the first step of this analysis, it is indispensable to settle a sound mathematical model which describes the phenomena by fairly plausible derivation. Our motivation of this paper is in this point.

Now, let us summarize the contents of this paper. First, we derive a general form of the two-dimensional viscous shallow-water equations, starting from the three-dimensional Reynolds equations for the time-smoothed velocity of the turbulent flow. The derived general form includes, as the special cases, a few typical horizontal viscosity terms used by several investigators, for example, Dronkers[1)], Leendertse-Liu[7)], Wang[12)] and Kawahara[11)]. Furthermore, the derived equations have the following two principal features:

A) the horizontal viscosity terms are hereditary from the original Reynolds equations,

B) mass and energy conservation laws[7)] are satisfied under physically plausible conditions.

As far as we know, Juncosa first discussed mass and energy conservation laws in simplified equations.[7)] Later, Gustafsson-Sundström[6)] demonstrated the well-posedness of a linearized system by the energy method. Our demonstration of conservation laws in the full system seems to be new.

Remark 1

It is noted that Zienkiewicz-Heinrich[5)] gives a completely different derivation of the viscous shallow-water equations from others.

2. Vertical averaging of the Reynolds equations under the hydrostatic pressure assumption

As the starting point, we adopt the three-dimensional Reynolds equations (R)[8] for the time-smoothed velocity of the turbulent flow in the following form :

$$\frac{Du}{Dt} = -\frac{1}{\rho}\frac{\partial P}{\partial x} + \frac{1}{\rho}\left(\frac{\partial \tau xx}{\partial x} + \frac{\partial \tau yx}{\partial y} + \frac{\partial \tau zx}{\partial z}\right) + Fx, \quad \ldots \quad (1)$$

$$\frac{Dv}{Dt} = -\frac{1}{\rho}\frac{\partial P}{\partial y} + \frac{1}{\rho}\left(\frac{\partial \tau xy}{\partial x} + \frac{\partial \tau yy}{\partial y} + \frac{\partial \tau zy}{\partial z}\right) + Fy, \quad \ldots \quad (2)$$

$$\frac{Dw}{Dt} = -\frac{1}{\rho}\frac{\partial P}{\partial z} + \frac{1}{\rho}\left(\frac{\partial \tau xz}{\partial x} + \frac{\partial \tau yz}{\partial y} + \frac{\partial \tau zz}{\partial z}\right) + Fz, \quad \ldots \quad (3)$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

where

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y} + w\frac{\partial}{\partial z},$$

x, y and z are Cartesian coordinates, t is time, u, v and w are the time-smoothed velocity components in the x, y and z directions respectively, P is the time-smoothed pressure, $\rho(>0)$ is the density of seawater, Fx, Fy and Fz are the x, y and z components of the forcing term, finally τxx, τyy, τzz, τxy, τyx, τyz, τzy, τzx, and τxz are stress components which are symmetric, that is $\tau xy = \tau yx$, $\tau yz = \tau zy$ and $\tau zx = \tau xz$. The density of seawater ρ is, for simlicity, assumed to be positive constant in this paper. Unless we clearly show the dependence on x, y, z and t, the same remark holds in the following.

To advance the description of the viscous shallow-water phenomena in which representative sizes of the quantities in the horizontal plane are quite larger than those in the vertical direction, we adopt the so-called hydrostatic pressure assumption. This implies that $\frac{Dw}{Dt}$ and

$\frac{1}{\rho}\left(\frac{\partial\tau xz}{\partial x} + \frac{\partial\tau yz}{\partial y} + \frac{\partial\tau zz}{\partial z}\right)$ are neglected in (3). Furthermore, it is assumed that the forcing term takes the following form : $(Fx, Fy, Fz) = (fv, -fu, -g)$, where f is the Coriolis coefficient and $g(>0)$ is the acceleration of gravity. The horizontal components of the forcing term (fv and -fu) repressent the principal parts of the Coriolis force.

The Coriolis coefficient f equals $2\omega \sin\phi$ where $\omega(>0)$ is the angular velocity of the earth and ϕ is the geographical latitude which is positive in the Northern Hemisphere and negative in the Southern Hemisphere. Hence, the equations of motion (1) - (3) are reduced to the following equations (5), (6) and (7) :

$$\frac{Du}{Dt} = -\frac{1}{\rho}\frac{\partial P}{\partial x} + \frac{1}{\rho}\left(\frac{\partial\tau xx}{\partial x} + \frac{\partial\tau yx}{\partial y} + \frac{\partial\tau zx}{\partial z}\right) + fv, \quad \ldots\ldots \quad (5)$$

$$\frac{Dv}{Dt} = -\frac{1}{\rho}\frac{\partial P}{\partial y} + \frac{1}{\rho}\left(\frac{\partial\tau xy}{\partial x} + \frac{\partial\tau yy}{\partial y} + \frac{\partial\tau zy}{\partial z}\right) - fu, \quad \ldots\ldots \quad (6)$$

$$0 = -\frac{1}{\rho}\frac{\partial P}{\partial z} - g. \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (7)$$

In this paper, the system (4) - (7) is called the Reynolds Hydrostatic system (RH). The concrete expressions of stress components will be described later.

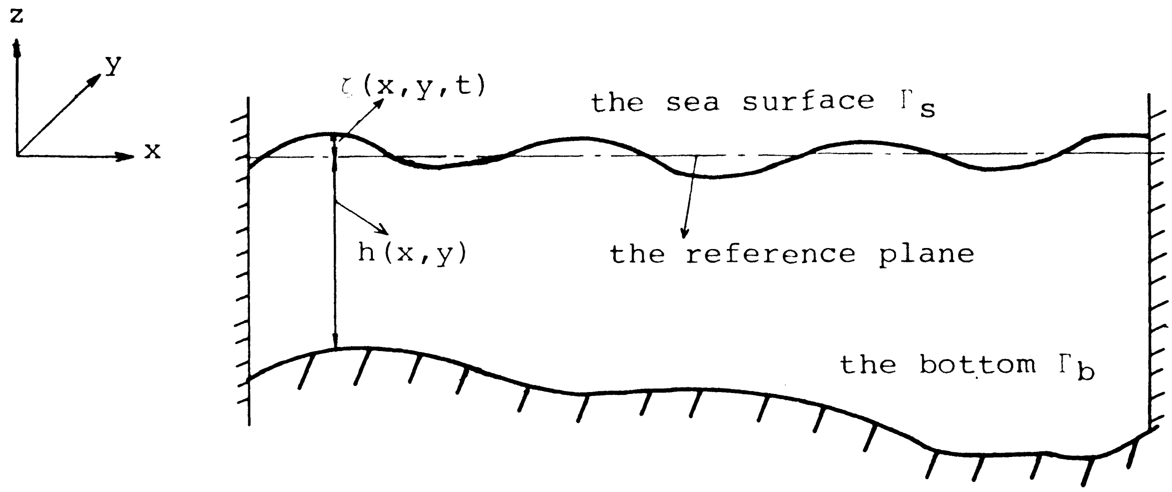

Fig.1 The cross section of a bay

Throughout this paper, the mean sea level which is assumed to be horizontal is taken as the reference plane (see Figure 1). Let us denote the sea surface and the bottom by Γs and Γb respectively. We assume that the two boundaries Γs and Γb are uniquely expressed as $z = \zeta(x,y,t)$ and $z = -h(x,y) < 0$, respectively, where the smooth function ζ may be considered as unknown in the following, while the smooth function h is known. Then, it is assumed that the set of smooth functions (u,v,w) of RH(4) - (7) satisfies the following two boundary conditions :

$$w = \frac{\partial \zeta}{\partial t} + u\frac{\partial \zeta}{\partial x} + v\frac{\partial \zeta}{\partial y}, \quad (x,y,z) \in \Gamma s, \quad \ldots\ldots\ldots \quad (8)$$

$$w = - u\frac{\partial h}{\partial x} - v\frac{\partial h}{\partial y}, \quad (x,y,z) \in \Gamma b. \quad \ldots\ldots\ldots\ldots \quad (9)$$

The conditions (8) and (9) mean that the normal velocity component vanishes both on the free boundary Γs and on the fixed boundary Γb.

In fact, they are obtained from the condition :

$\frac{DG}{Dt} = 0$, where $G = z-\zeta$ for (8) and $G = -h-z$ for (9) respectively.

Let us introduce the following notations :

$$H \equiv h+\zeta, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (10.0)$$

$$U \equiv \frac{1}{H}\int_{-h}^{\zeta} u dz, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (10.1)$$

$$V \equiv \frac{1}{H}\int_{-h}^{\zeta} v dz. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (10.2)$$

It is reasonable to restrict our consideration to the case which satisfies the condition :

$$H > 0. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (10.4)$$

Let $e(x,y,z,t)$ be an arbitrary mathematical expression dependent on the symbols x, y, z and t. For simplicity, we introduce the following notation :

$$e(x,y,z,t)\Big|_{z=-h}^{z=\zeta} = e(x,y,\zeta(x,y,t),t)-e(x,y,-h(x,y),t).$$

With these notations, the first proposition comes from (4).

Proposition 1.1

If the set of smooth functions (u,v,w) satisfies (4), then

$$\frac{\partial(HU)}{\partial x} + \frac{\partial(HV)}{\partial y} + (w-u\frac{\partial}{\partial x}z - v\frac{\partial}{\partial y}z)\Big|_{z=-h}^{z=\zeta} = 0. \quad \ldots(10)$$

Corollary 1.2

If the set of smooth functions (u,v,w), ζ and h satisfies (4), (8) and (9), then

$$\frac{\partial\zeta}{\partial t} + \frac{\partial(HU)}{\partial x} + \frac{\partial(HV)}{\partial y} = 0, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (11)$$

which is equivalent to

$$\frac{\partial H}{\partial t} + \frac{\partial(HU)}{\partial x} + \frac{\partial(HV)}{\partial y} = 0. \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (12)$$

Proof

The integration of (4) with respect to z from -h to ζ yields

$$\int_{-h}^{\zeta} \frac{\partial u}{\partial x}\, dz + \int_{-h}^{\zeta} \frac{\partial v}{\partial y}\, dz + w \Big|_{z=-h}^{z=\zeta} = 0. \quad \ldots\ldots\ldots\ldots \quad (13)$$

Remember the following equalities :

$$\int_{-h}^{\zeta} \frac{\partial u}{\partial x}\, dz = \frac{\partial}{\partial x} \int_{-h}^{\zeta} u dz - \left(u \frac{\partial}{\partial x} z\right) \Big|_{z=-h}^{z=\zeta} , \quad \ldots\ldots\ldots \quad (14)$$

$$\int_{-h}^{\zeta} \frac{\partial v}{\partial y}\, dz = \frac{\partial}{\partial y} \int_{-h}^{\zeta} v dz - \left(v \frac{\partial}{\partial y} z\right) \Big|_{z=-h}^{z=\zeta} . \quad \ldots\ldots\ldots \quad (15)$$

The substitution of (14) and (15) into (13) and the use of the boundary conditions (8) and (9) result in

$$\frac{\partial}{\partial x} \int_{-h}^{\zeta} u dz + \frac{\partial}{\partial y} \int_{-h}^{\zeta} v dz + \frac{\partial \zeta}{\partial t} = 0, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (16)$$

which means (11) due to the definitions (10.1) and (10.2).

Q.E.D.

The similar integration of (5) - (7) produces the second proposition.

Proposition 2

If the set of smooth functions (u,v,w), P, τ_{xx}, τ_{yx}, τ_{zx}, τ_{xy}, τ_{yy} and τ_{zy} satisfies (4) - (7), and if the atmospheric pressure P (x,y,ζ,t) is constant, then,

$$\frac{\partial}{\partial t} \int_{-h}^{\zeta} u\, dz + \frac{\partial}{\partial x} \int_{-h}^{\zeta} u^2\, dz + \frac{\partial}{\partial y} \int_{-h}^{\zeta} uv dz + \left\{u\left(w - \frac{\partial}{\partial t} z - u \frac{\partial}{\partial x} z - v \frac{\partial}{\partial y} z\right)\right\} \Big|_{z=-h}^{z=\zeta}$$

$$= -gH\frac{\partial \zeta}{\partial x}+\frac{1}{\rho}\ \{\frac{\partial}{\partial x}\int_{-h}^{\zeta} \tau_{xx}dz+\frac{\partial}{\partial y}\int_{-h}^{\zeta} \tau_{yx}dz-(\tau_{xx}\frac{\partial}{\partial x}z+\tau_{yx}\frac{\partial}{\partial y}z-\tau_{zx})\Bigg|_{z=-h}^{z=\zeta}\}$$

$$+ f\int_{-h}^{\zeta} v\ dz, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (17)$$

$$\frac{\partial}{\partial t}\int_{-h}^{\zeta} v\ dz+\frac{\partial}{\partial x}\int_{-h}^{\zeta} uv\ dz+\frac{\partial}{\partial y}\int_{-h}^{\zeta} v^2dz+\{v(w-\frac{\partial}{\partial t}z-u\frac{\partial}{\partial x}z-v\frac{\partial}{\partial y}z)\}\Bigg|_{z=-h}^{z=\zeta}$$

$$= -gH\frac{\partial \zeta}{\partial y}+\frac{1}{\rho}\ \{\frac{\partial}{\partial x}\int_{-h}^{\zeta} \tau_{xy}dz+\frac{\partial}{\partial y}\int_{-h}^{\zeta} \tau_{yy}dz-(\tau_{xy}\frac{\partial}{\partial x}z+\tau_{yy}\frac{\partial}{\partial y}z-\tau_{zy})\Bigg|_{z=-h}^{z=\zeta}\}$$

$$- f\int_{-h}^{\zeta} u\ dz. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (18)$$

<u>Proof</u>

It is sufficient to prove only (17). The integration of (7) with respect to z from z to ζ leads

$$P\Bigg|_{z=\zeta}^{z=z} = \rho g(\zeta-z). \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (19)$$

Then, we have

$$\frac{\partial P}{\partial x} = \rho g\ \frac{\partial \zeta}{\partial x}\ ,\ \frac{\partial P}{\partial y} = \rho g\ \frac{\partial \zeta}{\partial y}\ , \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (20)$$

because the atmospheric pressure is assumed to be constant. Therefore, the similar integration of (5) yields

$$\int_{-h}^{\zeta} \frac{\partial u}{\partial t}\ dz + \int_{-h}^{\zeta} u\frac{\partial u}{\partial x}\ dz + \int_{-h}^{\zeta} v\frac{\partial u}{\partial y}\ dz + \int_{-h}^{\zeta} w\frac{\partial u}{\partial z}\ dz$$

$$= -gH\frac{\partial \zeta}{\partial x} + \frac{1}{\rho}(\int_{-h}^{\zeta} \frac{\partial \tau_{xx}}{\partial x}\ dz + \int_{-h}^{\zeta} \frac{\partial \tau_{yx}}{\partial y}\ dz + \tau_{zx}\Bigg|_{z=-h}^{z=\zeta}) + f\int_{-h}^{\zeta} vdz.$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (21)$$

As (14) and (15), the following equalities hold :

$$\int_{-h}^{\zeta} \frac{\partial u}{\partial t}\, dz = \frac{\partial}{\partial t} \int_{-h}^{\zeta} u dz - \left(u \frac{\partial}{\partial t} z\right) \Bigg|_{z=-h}^{z=\zeta}, \quad \ldots\ldots\ldots\ldots\ldots \quad (22)$$

$$\int_{-h}^{\zeta} u \frac{\partial u}{\partial x}\, dz = \frac{1}{2} \frac{\partial}{\partial x} \int_{-h}^{\zeta} u^2 dz - \frac{1}{2} \left(u^2 \frac{\partial}{\partial x} z\right) \Bigg|_{z=-h}^{z=\zeta}, \quad \ldots\ldots \quad (23)$$

$$\int_{-h}^{\zeta} v \frac{\partial u}{\partial y}\, dz = \int_{-h}^{\zeta} \frac{\partial}{\partial y}(uv)\, dz - \int_{-h}^{\zeta} u \frac{\partial v}{\partial y}\, dz$$

$$= \frac{\partial}{\partial y} \int_{-h}^{\zeta} uv dz - \left(uv \frac{\partial}{\partial y} z\right) \Bigg|_{z=-h}^{z=\zeta} - \int_{-h}^{\zeta} u \frac{\partial v}{\partial y}\, dz, \quad \ldots \quad (24)$$

$$\int_{-h}^{\zeta} w \frac{\partial u}{\partial z}\, dz = (wu) \Bigg|_{z=-h}^{z=\zeta} - \int_{-h}^{\zeta} u \frac{\partial w}{\partial z}\, dz, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (25)$$

$$\int_{-h}^{\zeta} \frac{\partial \tau_{xx}}{\partial x}\, dz = \frac{\partial}{\partial x} \int_{-h}^{\zeta} \tau_{xx} dz - \left(\tau_{xx} \frac{\partial}{\partial x} z\right) \Bigg|_{z=-h}^{z=\zeta}, \quad \ldots\ldots\ldots \quad (26)$$

$$\int_{-h}^{\zeta} \frac{\partial \tau_{yx}}{\partial y}\, dz = \frac{\partial}{\partial y} \int_{-h}^{\zeta} \tau_{yx} dz - \left(\tau_{yx} \frac{\partial}{\partial y} z\right) \Bigg|_{z=-h}^{z=\zeta}. \quad \ldots\ldots\ldots \quad (27)$$

Furthermore, by (4), we have

$$- \int_{-h}^{\zeta} u \left(\frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) dz = \int_{-h}^{\zeta} u \frac{\partial u}{\partial x}\, dz. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (28)$$

The substitution of (22) - (28) into (21) completes the proof of Proposition 2.

Q.E.D.

In the following, the system (10), (17) and (18) is called the vertical Averaging Reynolds Hydrostatic system (ARH).

3. The viscous shallow-water equations

Let us define $\tilde{u}$ and $\tilde{v}$ as follows :

$$\tilde{u} = u-U, \quad \tilde{v} = v-V.$$

Obviously, from the definitions (10.1) and (10.2), it follows that

$$\int_{-h}^{\zeta} \tilde{u}\, dz = 0, \quad \int_{-h}^{\zeta} \tilde{v}\, dz = 0.$$

Furthermore, we define the two-diemensional stresses τ_{xx}', τ_{yx}', τ_{xy}' and τ_{yy}' as follows :

$$\tau_{xx}' = \frac{1}{H}\int_{-h}^{\zeta} \tau_{xx} dz - \frac{\rho}{H}\int_{-h}^{\zeta} \tilde{u}^2 dz, \quad \tau_{yx}' = \frac{1}{H}\int_{-h}^{\zeta} \tau_{yx} dz - \frac{\rho}{H}\int_{-h}^{\zeta} \tilde{u}\tilde{v} dz,$$

$$\tau_{xy}' = \frac{1}{H}\int_{-h}^{\zeta} \tau_{xy} dz - \frac{\rho}{H}\int_{-h}^{\zeta} \tilde{u}\tilde{v} dz, \quad \tau_{yy}' = \frac{1}{H}\int_{-h}^{\zeta} \tau_{yy} dz - \frac{\rho}{H}\int_{-h}^{\zeta} \tilde{v}^2 dz.$$

τ_{xy}' is also symmetric because τ_{xy} is so. Then, a general form of the Viscous Shallow-Water equations (VSW) is the system which consits of (11) derived from (10) and the following (29) and (30) derived from (17) and (18) respectively:

$$\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} + V\frac{\partial U}{\partial y} = -g\frac{\partial \zeta}{\partial x} + \frac{1}{H\rho}\left\{\frac{\partial}{\partial x}(H\tau_{xx}') + \frac{\partial}{\partial y}(H\tau_{yx}')\right\}$$

$$+ \frac{1}{H\rho}(|\nabla_3 G_s|\tau_{sx} - |\nabla_3 G_b|\tau_{bx}) + fV, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (29)$$

$$\frac{\partial V}{\partial t} + U\frac{\partial V}{\partial x} + V\frac{\partial V}{\partial y} = -g\frac{\partial \zeta}{\partial y} + \frac{1}{H\rho}\left\{\frac{\partial}{\partial x}(H\tau_{xy}') + \frac{\partial}{\partial y}(H\tau_{yy}')\right\}$$

$$+ \frac{1}{H\rho}(|\nabla_3 G_s|\tau_{sy} - |\nabla_3 G_b|\tau_{by}) - fU, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (30)$$

where, $G_s = z-\zeta$, $\quad \nabla_3 G_s = (-\frac{\partial \zeta}{\partial x}, -\frac{\partial \zeta}{\partial y}, 1)$,

$|\nabla_3 G_s| = \{\nabla_3 G_s \cdot \nabla_3 G_s\}^{1/2}$, $G_b = (-h) - z$,

$\tau_{sx} = \vec{\tau}_x \cdot \vec{n}_s$, $\tau_{bx} = -\vec{\tau}_x \cdot \vec{n}_b$, $\tau_{sy} = \vec{\tau}_y \cdot \vec{n}_s$, $\tau_{by} = -\vec{\tau}_y \cdot \vec{n}_b$,

$\vec{\tau}_x = (\tau_{xx}, \tau_{yx}, \tau_{zx})$, $\vec{\tau}_y = (\tau_{xy}, \tau_{yy}, \tau_{zy})$,

$\vec{n}_s = \dfrac{\nabla_3 G_s}{|\nabla_3 G_s|}$, $\vec{n}_b = \dfrac{\nabla_3 G_b}{|\nabla_3 G_b|}$.

The symbol $\cdot$ denotes the inner product of two vectors.

<u>Remark 2</u>

$$\{u(w - \frac{\partial}{\partial t}z \; u\frac{\partial}{\partial x}z - v\frac{\partial}{\partial y}z)\}\Big|_{z=-h}^{z=\zeta} = 0 \quad \text{and}$$

$$\{v(w - \frac{\partial}{\partial t}z - u\frac{\partial}{\partial x}z - v\frac{\partial}{\partial y}z)\}\Big|_{z=-h}^{z=\zeta} = 0$$ in the left hand sides of (17) and (18) hold due to the boundary conditions (8) and (9).

<u>Remark 3</u>

In the right hand sides of (17) and (18), we have

$$-\left(\tau_{xx}\frac{\partial}{\partial x}z + \tau_{yx}\frac{\partial}{\partial y}z - \tau_{zx}\right)\Big|_{z=-h}^{z=\zeta} = |\nabla_3 G_s|\tau_{sx} - |\nabla_3 G_b|\tau_{bx},$$

$$-\left(\tau_{xy}\frac{\partial}{\partial x}z + \tau_{yy}\frac{\partial}{\partial y}z - \tau_{zy}\right)\Big|_{z=-h}^{z=\zeta} = |\nabla_3 G_s|\tau_{sy} - |\nabla_3 G_b|\tau_{by}.$$

If the positive directions of stresses are taken as in Figure 2, τ_{sx} and τ_{bx} are equivalent to the x-direction stresses on the tangential planes of Γ_s and Γ_b respectively. τ_{sy} and τ_{by} are the corresponding y-direction stresses.

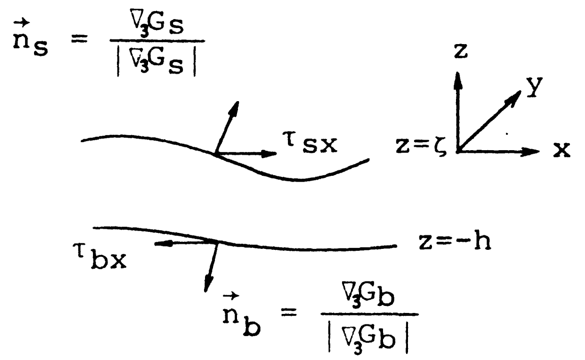

Fig.2 The positive directions of τ_{sx} and τ_{bx}

Usually, the following concrete expressions[1] for τ_{sx}, τ_{bx}, τ_{sy} and τ_{by} are used :

$$|\nabla_3 G_s|\tau_{sx} = \theta\rho_a W_x (W_x{}^2+W_y{}^2)^{1/2}, \quad |\nabla_3 G_s|\tau_{sy} = \theta\rho_a W_y (W_x{}^2+W_y{}^2)^{1/2},$$

$$|\nabla_3 G_b|\tau_{bx} = \frac{\rho g}{C^2} U(U^2+V^2)^{1/2}, \quad |\nabla_3 G_b|\tau_{by} = \frac{\rho g}{C^2} V(U^2+V^2)^{1/2},$$

where W_x and W_y are velocity components of the wind in the x and y directions respectively, $\theta(>0)$ is the coefficient of the wind effect, $\rho_a(>0)$ is the density of atmosphere, and $C(>0)$ is the Chezy's coefficient. Finally, it is also noted that $|\nabla_3 G_s| \simeq 1$ and $|\nabla_3 G_b| \simeq 1$ if the quadratic terms of $\frac{\partial\zeta}{\partial x}$, $\frac{\partial\zeta}{\partial y}$, $\frac{\partial h}{\partial x}$ and $\frac{\partial h}{\partial y}$ are sufficiently small compared with 1.[9]

Remark 4

From the definitions of $\tilde{u}$ and $\tilde{v}$, it can be easily shown that

$$\frac{\partial}{\partial x}\int_{-h}^{\zeta} u^2 dz+\frac{\partial}{\partial y}\int_{-h}^{\zeta} uvdz = \frac{\partial}{\partial x}(HU^2)+\frac{\partial}{\partial y}(HUV)+\frac{\partial}{\partial x}\int_{-h}^{\zeta} \tilde{u}^2 dz+\frac{\partial}{\partial y}\int_{-h}^{\zeta} \tilde{u}\tilde{v}dz,$$

$$\frac{\partial}{\partial x}\int_{-h}^{\zeta} uvdz+\frac{\partial}{\partial y}\int_{-h}^{\zeta} v^2 dz = \frac{\partial}{\partial x}(HUV)+\frac{\partial}{\partial y}(HV^2)+\frac{\partial}{\partial x}\int_{-h}^{\zeta} \tilde{u}\tilde{v}dz+\frac{\partial}{\partial y}\int_{-h}^{\zeta} \tilde{v}^2 dz,$$

in the left hand sides of (17) and (18). Also, we have, due to (12),

$$\frac{\partial}{\partial t}(HU) + \frac{\partial}{\partial x}(HU^2) + \frac{\partial}{\partial y}(HUV) = H\left(\frac{\partial U}{\partial t} + U\frac{\partial U}{\partial x} + V\frac{\partial U}{\partial y}\right),$$

$$\frac{\partial}{\partial t}(HV) + \frac{\partial}{\partial x}(HUV) + \frac{\partial}{\partial y}(HV^2) = H\left(\frac{\partial V}{\partial t} + U\frac{\partial V}{\partial x} + V\frac{\partial V}{\partial y}\right).$$

Remark 5

If $\tilde{u}^2$, $\tilde{u}\tilde{v}$ and $\tilde{v}^2$ are sufficiently small compared with U^2, UV and V^2 respectively, we have

$$\tau_{xx}' \simeq \frac{1}{H}\int_{-h}^{\zeta} \tau_{xx}dz \ , \quad \tau_{yx}' \simeq \frac{1}{H}\int_{-h}^{\zeta} \tau_{yx}dz \ ,$$

$$\tau_{xy}' \simeq \frac{1}{H}\int_{-h}^{\zeta} \tau_{xy}dz \ , \quad \tau_{yy}' \simeq \frac{1}{H}\int_{-h}^{\zeta} \tau_{yy}dz \ .$$

In the following Remarks 6, 7 and 8, we examine the relations between constitutive laws assumed in the original RH system and those assumed in the VSW system.

Remark 6

Now, we suppose the following constitutive law [5] in RH (5) and (6) :

$$\frac{1}{\rho}\tau_{xx} = 2\nu_H\frac{\partial u}{\partial x} \ , \ \frac{1}{\rho}\tau_{yy} = 2\nu_H\frac{\partial v}{\partial y} \ , \ \frac{1}{\rho}\tau_{xy} = \frac{1}{\rho}\tau_{yx} = \nu_H\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right),$$

$$\frac{1}{\rho}\tau_{zx} = \nu_V\frac{\partial u}{\partial z} \ , \quad \frac{1}{\rho}\tau_{zy} = \nu_V\frac{\partial v}{\partial z} \ ,$$

where $\nu_H(>0)$ and $\nu_V(>0)$ mainly denote the coefficients of eddy viscosity in the horizontal and vertical directions respectively. Then, for example, in the right hand side of (17), we have

$$\frac{1}{\rho}\left(\frac{\partial}{\partial x}\int_{-h}^{\zeta}\tau_{xx}dz+\frac{\partial}{\partial y}\int_{-h}^{\zeta}\tau_{yx}dz\right) = \frac{\partial}{\partial x}\left(2\nu_H\int_{-h}^{\zeta}\frac{\partial u}{\partial x}dz\right)+\frac{\partial}{\partial y}\left(\nu_H\int_{-h}^{\zeta}\frac{\partial u}{\partial y}dz+\nu_H\int_{-h}^{\zeta}\frac{\partial v}{\partial x}dz\right)$$

$$= \frac{\partial}{\partial x}\left[2\nu_H\left\{\frac{\partial}{\partial x}\int_{-h}^{\zeta}udz-\left(u\frac{\partial}{\partial x}z\right)\Big|_{z=-h}^{z=\zeta}\right\}\right]+\frac{\partial}{\partial y}\left[\nu_H\left\{\frac{\partial}{\partial y}\int_{-h}^{\zeta}udz-\left(u\frac{\partial}{\partial y}z\right)\Big|_{z=-h}^{z=\zeta}\right\}\right]$$

$$+\frac{\partial}{\partial y}\left[\nu_H\left\{\frac{\partial}{\partial x}\int_{-h}^{\zeta}vdz-\left(v\frac{\partial}{\partial x}z\right)\Big|_{z=-h}^{z=\zeta}\right\}\right]$$

$$= \frac{\partial}{\partial x}\left[2\nu_H\left\{\left(H\frac{\partial U}{\partial x}\right)+\left\{(U-u)\frac{\partial}{\partial x}z\right\}\Big|_{z=-h}^{z=\zeta}\right\}\right]$$

$$+\frac{\partial}{\partial y}\left[\nu_H\left\{\left(H\frac{\partial U}{\partial y}\right)+\left\{(U-u)\frac{\partial}{\partial y}z\right\}\Big|_{z=-h}^{z=\zeta}\right\}\right]+\frac{\partial}{\partial y}\left[\nu_H\left\{H\frac{\partial V}{\partial x}\right)+\left\{(V-v)\frac{\partial}{\partial x}z\right\}\Big|_{z=-h}^{z=\zeta}\right\}\right]$$

$$\simeq\frac{\partial}{\partial x}\left\{H\,2\nu_H\left(\frac{\partial U}{\partial x}\right)\right\}+\frac{\partial}{\partial y}\left\{H\nu_H\left(\frac{\partial U}{\partial y}+\frac{\partial V}{\partial x}\right)\right\}.$$

In the last part, we assumed that $\tilde{u}\frac{\partial H}{\partial x}$, $\tilde{u}\frac{\partial H}{\partial y}$, $\tilde{v}\frac{\partial H}{\partial x}$ and $\tilde{v}\frac{\partial H}{\partial y}$ are sufficiently small compared with $H\frac{\partial U}{\partial x}$, $H\frac{\partial U}{\partial y}$, $H\frac{\partial V}{\partial x}$ and $H\frac{\partial V}{\partial y}$ respectively. Also, it is remarked that the z-independence of ν_H was essentially used in the above calculation.

The horizontal viscosity terms of this type were adopted by, for example, Wang[12].

<u>Remark 7</u>

Remarks 5 and 6 suggest to ignore the following term in (17) :

$$\frac{\partial}{\partial x}\left[2\nu_H\left\{(U-u)\frac{\partial}{\partial x}z\right\}\Big|_{z=-h}^{z=\zeta}-\int_{-h}^{\zeta}\tilde{u}^2dz\right]+\frac{\partial}{\partial y}\left[\nu_H\left\{(U-u)\frac{\partial}{\partial y}z+(V-v)\frac{\partial}{\partial x}z\right\}\Big|_{z=-h}^{z=\zeta}-\int_{-h}^{\zeta}\tilde{u}\tilde{v}dz\right].$$

The same note also holds for (18). This disregard means that we assume the following two-dimensional constitutive law:

(*) $\frac{1}{\rho}\tau_{xx}' = 2\nu_H \frac{\partial U}{\partial x}$, $\frac{1}{\rho}\tau_{xy}' = \frac{1}{\rho}\tau_{yx}' = \nu_H(\frac{\partial U}{\partial y}+\frac{\partial V}{\partial x})$, $\frac{1}{\rho}\tau_{yy}' = 2\nu_H \frac{\partial V}{\partial y}$.

Kawahara[11] also used this type of constitutive law. But, in his paper, the horizontal viscosity terms take the following forms in the right hand sides of (29) and (30) respectively:

$$\frac{\partial}{\partial x}\{2\nu_H(\frac{\partial U}{\partial x})\} + \frac{\partial}{\partial y}\{\nu_H(\frac{\partial U}{\partial y}+\frac{\partial V}{\partial x})\},$$

$$\frac{\partial}{\partial x}\{\nu_H(\frac{\partial V}{\partial x}+\frac{\partial U}{\partial y})\} + \frac{\partial}{\partial y}\{2\nu_H(\frac{\partial V}{\partial y})\} .$$

Namely, the space dependence of H is neglected in a sense. On the other hand, Connor-Brebbia[9] describes the following horizontal viscosity terms in (29) and (30), when ρ is constant :

$$\frac{1}{H}\left[\frac{\partial}{\partial x}\{2\nu_H\frac{\partial (HU)}{\partial x}\} + \frac{\partial}{\partial y}\{\nu_H(\frac{\partial (HU)}{\partial y} + \frac{\partial (HV)}{\partial x})\}\right],$$

$$\frac{1}{H}\left[\frac{\partial}{\partial x}\{\nu_H(\frac{\partial (HV)}{\partial x} + \frac{\partial (HU)}{\partial y})\} + \frac{\partial}{\partial y}\{2\nu_H\frac{\partial (HV)}{\partial y}\}\right],$$

assuming directly the two-dimensional constitutive law without describing the three-dimensional one. This means that

$$\frac{1}{\rho}\tau_{xx}'=2\nu_H(\frac{\partial U}{\partial x}+\frac{1}{H}\frac{\partial H}{\partial x}U),\ \frac{1}{\rho}\tau_{yx}'\doteq\nu_H(\frac{\partial U}{\partial y}+\frac{\partial V}{\partial x}+\frac{1}{H}\frac{\partial H}{\partial y}U+\frac{1}{H}\frac{\partial H}{\partial x}V),$$

$$\frac{1}{\rho}\tau_{xy}'=\frac{1}{\rho}\tau_{yx}',\ \frac{1}{\rho}\tau_{yy}'=2\nu_H(\frac{\partial V}{\partial y}+\frac{1}{H}\frac{\partial H}{\partial y}V).$$

Therefore, again, the neglect of the space dependence of H derives the same two-dimensional constitutive law as in Wang[12] and in Kawahara[11].

Remark 8

There is another assumption of constitutive law in RH(5) and (6).[7] That is

$$\frac{1}{\rho}\tau_{xx}=\nu_H\frac{\partial u}{\partial x},\quad \frac{1}{\rho}\tau_{yx}=\nu_H\frac{\partial u}{\partial y},\quad \frac{1}{\rho}\tau_{zx}=\nu_v\frac{\partial u}{\partial z}\ ,$$

$$\frac{1}{\rho}\tau_{xy}=\nu_H\frac{\partial v}{\partial x},\quad \frac{1}{\rho}\tau_{yy}=\nu_H\frac{\partial v}{\partial y},\quad \frac{1}{\rho}\tau_{zy}=\nu_v\frac{\partial v}{\partial z}\ .$$

It is noted that, in this case, τ_{xy} is not equivalent to τ_{yx}. The similar discussion to Remarks 5, 6 and 7 leads the following two-dimensional constitutive law which was adopted by, for example, Leendertse-Liu[7], Gustafsson-Sundström[6] and Walters-Cheng[14]:

$$(**)\quad \frac{1}{\rho}\tau_{xx}'=\nu_H\frac{\partial U}{\partial x},\ \frac{1}{\rho}\tau_{yx}'=\nu_H\frac{\partial U}{\partial y},\ \frac{1}{\rho}\tau_{xy}'=\nu_H\frac{\partial V}{\partial x},\ \frac{1}{\rho}\tau_{yy}'=\nu_H\frac{\partial V}{\partial y}\ .$$

Pinder-Gray[10] seems to use the horizontal viscosity terms of this type. Some formulas in page 268 of the first edition of their book should be corrected, in our opinion.

Dronkers[1] considered the abridged forms of this type, neglecting the space dependence of H. Namely, the horizontal viscosity terms in (29) and (30) take the following forms respectively :

$$\frac{\partial}{\partial x}(\nu H\frac{\partial U}{\partial x}) + \frac{\partial}{\partial y}(\nu H\frac{\partial U}{\partial y})\ ,$$

$$\frac{\partial}{\partial x}(\nu H\frac{\partial V}{\partial x}) + \frac{\partial}{\partial y}(\nu H\frac{\partial V}{\partial y})\ .$$

These abridged forms were also used by, for example, Kanayama-Ohtsuka[4] and Praagman[15]. Leendertse[2] and Kaneko et al.[3] completely neglected the horizontal viscosity terms.

Tanaka - Ono[13] describes the following terms in (29) and (30) of the same type as in Connor - Brebbia[9], using the second constitutive law :

$$\frac{1}{H}\left[\frac{\partial}{\partial x}\{\nu_H\frac{\partial (HU)}{\partial x}\} + \frac{\partial}{\partial y}\{\nu_H\frac{\partial (HU)}{\partial y}\}\right],$$

$$\frac{1}{H}\left[\frac{\partial}{\partial x}\{\nu_H\frac{\partial (HV)}{\partial x}\} + \frac{\partial}{\partial y}\{\nu_H\frac{\partial (HV)}{\partial y}\}\right].$$

4. Conservation of mass and energy

This chapter demonstrates that the derived viscous shallow-water equations satisfy mass and energy conservation laws under physically plausible conditions. The vector notation is used in the following. Then, the derived VSW system (11), (29) and (30) can be rewritten as

$$\frac{\partial\zeta}{\partial t} + \mathrm{div}\ (H\vec{U}) = 0, \quad \ldots\ldots \quad (31)$$

$$\frac{\partial\vec{U}}{\partial t} + \vec{U}\cdot\nabla\vec{U} + \vec{U}[f] + g\ \nabla\zeta$$

$$= \frac{1}{H\rho}\ (\mathrm{div}\ (H\tau) + |\nabla_3 G_s|\vec{\tau}_s - |\nabla_3 G_b|\vec{\tau}_b), \quad \ldots\ldots \quad (32)$$

where

$$\vec{U} = (U,\ V), \qquad [f] = \begin{bmatrix} 0 & f \\ -f & 0 \end{bmatrix},$$

$$\tau = \begin{pmatrix} \tau_{xx} & \tau_{yx} \\ \tau_{xy} & \tau_{yy} \end{pmatrix},$$

$$\vec{\tau}_x = (\tau_{xx},\ \tau_{yx}), \qquad \vec{\tau}_y = (\tau_{xy},\ \tau_{yy}),$$

$$\vec{\tau}_s = (\tau_{sx},\ \tau_{sy}), \qquad \vec{\tau}_b = (\tau_{bx},\ \tau_{by}).$$

It is noted that the dashes of the two-dimensional stresses are omitted here, and that we use the following notation :

$$\mathrm{div}(H\tau) = (\mathrm{div}(H\vec{\tau}_x),\ \mathrm{div}(H\vec{\tau}_y)).$$

The initial conditions are specified as

$$\zeta(x,\ y,\ 0) = \zeta_0(x,\ y), \quad \ldots\ldots\ldots\ldots \quad (33.1)$$

$$\vec{U}(x,\ y,\ 0) = \vec{U}_0(x,\ y), \quad \ldots\ldots\ldots\ldots \quad (33.2)$$

in the smooth bounded domain Ω. Γ denotes its boundary. We consider the following boundary condition :

$$U_n \equiv \vec{U}(x,\ y,\ t)\cdot\vec{n} = 0, \quad \ldots\ldots\ldots\ldots \quad (34)$$

where $\vec{n}$ denotes the outer normal unit vector on Γ. The initial data $\vec{U}_0(x,\ y)$ and $\zeta_0(x,\ y)$ are assumed to be sufficiently smooth and to satisfy the compatibility conditions associated with the boundary conditions (34), (49) and (55), and the condition (10.4). Under the above situation, the following mass conservation law holds.

<u>Proposition 3</u> (Mass conservation)

If the set of smooth functions ζ and $\vec{U}$ satisfies (31) and (34) for $t>0$, we have

$$\frac{d}{dt}\int_\Omega \zeta\ d\vec{x} = 0. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (35)$$

<u>Proof</u>

By the equation of continuity (31), the Gauss-Green formula and the boundary condition (34), the derivation of (35) is straightforward :

$$\frac{d}{dt}\int_\Omega \zeta\, d\vec{x} = \int_\Omega \frac{\partial\zeta}{\partial t}\, d\vec{x} = -\int_\Omega \mathrm{div}(H\vec{U})\, d\vec{x}$$

$$= -\oint_\Gamma H\vec{U}\cdot\vec{n}\, ds = 0.$$

Q.E.D.

Now, we consider the following energy form T(t) attached to our VSW system (31) and (32) :

$$T(t) = \frac{g}{2}\int_\Omega \zeta^2 d\vec{x} + \frac{1}{2}\int_\Omega H|\vec{U}|^2 d\vec{x}\ , \ \ldots\ldots. \qquad (36)$$

where it is required that $\zeta(x, y, t)$ is square integrable over Ω with respect to the usual Lebesgue measure for fixed t, and that $\vec{U}(x, y, t)$ is square integrable over Ω with respect to the measure $H(x, y, t)\, d\vec{x}$ $(H=h+\zeta>0)$ for fixed t.

Proposition 4 (Energy Conservation)

If the set of smooth functions ζ, $\vec{U}$ and τ satisfies (31), (32), (34) and (10.4) for t>0, we have

$$\frac{dT}{dt} = \frac{1}{\rho}\ (\mathrm{div}(H\tau)+|\nabla_3 G_s|\vec{\tau}_s-|\nabla_3 G_b|\vec{\tau}_b,\ \vec{U}),\ \ldots\ldots \qquad (37)$$

where (,) denotes the L^2-inner product over Ω.

Proof

Multiplying (31) and (32) by $g\zeta$ and by $H\vec{U}$ respectively, and integrating the results over Ω, we can get

$$(\frac{\partial\zeta}{\partial t},\ g\zeta) + (\mathrm{div}(H\vec{U}),\ g\zeta) = 0,\ \ldots\ldots\ldots\ldots\ldots\ldots \qquad (40)$$

$$(\frac{\partial\vec{U}}{\partial t},\ H\vec{U}) + (\vec{U}\cdot\nabla\vec{U}, H\vec{U}) + (\vec{U}[f], H\vec{U}) + (g\nabla\zeta, H\vec{U})$$

$$= (\frac{1}{H\rho}(\mathrm{div}(H\tau) + |\nabla_3 G_s|\vec{\tau}_s - |\nabla_3 G_b|\vec{\tau}_b),\ H\vec{U}).\ \ldots\ldots. \qquad (41)$$

The addition of (40) and (41) yields

$$(\frac{\partial\vec{U}}{\partial t},H\vec{U}) + (\vec{U}\cdot\nabla\vec{U},H\vec{U}) + (\frac{\partial\zeta}{\partial t},g\zeta)$$

$$= -(\vec{U}[f],H\vec{U}) - g\{(\nabla\zeta,H\vec{U}) + (\mathrm{div}(H\vec{U}),\zeta)\}$$

$$+ \frac{1}{\rho}(\mathrm{div}(H\tau) + |\nabla_3 G_s|\vec{\tau}_s - |\nabla_3 G_b|\vec{\tau}_b,\vec{U}). \quad \dots\dots\dots\dots \quad (42)$$

It is noted that

$$(\vec{U}[f], H\vec{U}) = 0, \quad \dots\dots\dots\dots\dots\dots\dots\dots \quad (43)$$

and, due to the Gauss-Green formula and the boundary condition (34),

$$(\nabla\zeta, H\vec{U}) + (\mathrm{div}(H\vec{U}),\zeta) = 0. \quad \dots\dots\dots\dots\dots\dots \quad (44)$$

Furthermore, we claim that the left hand side of (42) equals $\frac{dT}{dt}$. In fact, we have

$$(\vec{U}\cdot\nabla\vec{U}, H\vec{U}) = \frac{1}{2}(\frac{\partial\zeta}{\partial t}, |\vec{U}|^2), \quad \dots\dots\dots\dots\dots\dots \quad (45)$$

because, by the Gauss-Green formula, the boundary condition (34) and the equation of continuity (31),

$$(\vec{U}\cdot\nabla\vec{U},H\vec{U}) = (U\frac{\partial\vec{U}}{\partial x} + V\frac{\partial\vec{U}}{\partial y}, H\vec{U}) = \frac{1}{2}(\nabla(|\vec{U}|^2), H\vec{U})$$

$$= \frac{1}{2}\int_\Omega \mathrm{div}(H|\vec{U}|^2\vec{U})\, d\vec{x} - \frac{1}{2}\int_\Omega \mathrm{div}(H\vec{U})|\vec{U}|^2\, d\vec{x}$$

$$= \frac{1}{2}\oint_\Gamma H|\vec{U}|^2\vec{U}\cdot\vec{n}\, ds + \frac{1}{2}\int_\Omega \frac{\partial\zeta}{\partial t}|\vec{U}|^2\, d\vec{x} = \frac{1}{2}(\frac{\partial\zeta}{\partial t},|\vec{U}|^2).$$

Therefore, we have

$$\frac{dT}{dt} = (\frac{\partial\vec{U}}{\partial t},H\vec{U}) + (\vec{U}\cdot\nabla\vec{U},H\vec{U}) + (\frac{\partial\zeta}{\partial t},g\zeta)$$

$$= \frac{1}{\rho}(\mathrm{div}(H\tau) + |\nabla_3 G_s|\vec{\tau}_s - |\nabla_3 G_b|\vec{\tau}_b,\vec{U}).$$

Q.E.D.

Remark 9

If ζ, $\vec{U}$ and τ in Proposition 4 further satisfies

$$\frac{1}{\rho}\ (\mathrm{div}(H\tau) + |\nabla_3 G_s|\vec{\tau}_s - |\nabla_3 G_b|\vec{\tau}_b,\ \vec{U}) \leq 0, \quad \ldots\ldots \quad (46)$$

then, Proposition 4 implies

$$\frac{dT}{dt} \leq 0. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (47)$$

Let us adopt the concrete expressions for $\vec{\tau}_s$ and $\vec{\tau}_b$ in Remark 3 and choose the first constitutive law (*) in Remark 7. Then, the following additional conditions (48) and (49) are sufficient for (46) :

$$W_x = W_y = 0 \quad \text{in } \Omega, \quad \ldots\ldots\ldots\ldots\ldots \quad (48)$$

$$U_t = 0 \qquad \text{on } \Gamma, \quad \ldots\ldots\ldots\ldots\ldots \quad (49)$$

where U_t denotes the tangential velocity component of $\vec{U}$. In fact, from the expressions for $\vec{\tau}_s$ and $\vec{\tau}_b$, it is obvious that

$$\frac{1}{\rho}\ (|\nabla_3 G_s|\vec{\tau}_s,\ \vec{U}) = 0, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (50)$$

$$-\frac{1}{\rho}\ (|\nabla_3 G_b|\vec{\tau}_b,\ \vec{U}) \leq 0. \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (51)$$

Furthermore, we can also show

$$\frac{1}{\rho}\ (\mathrm{div}(H\tau),\ \vec{U}) \leq 0. \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (52)$$

The proof is as follows. By the Gauss-Green formula, we have

$$\frac{1}{\rho}\ (\mathrm{div}(H\tau),\vec{U}) = \frac{1}{\rho}\{(\mathrm{div}(H\vec{\tau}_x),U) + (\mathrm{div}(H\vec{\tau}_y),V)\}$$

$$= \frac{1}{\rho}\{\int_\Omega \mathrm{div}(HU\vec{\tau}_x)d\vec{x} + \int_\Omega \mathrm{div}(HV\vec{\tau}_y)d\vec{x} - (H\nabla U,\vec{\tau}_x) - (H\nabla V,\vec{\tau}_y)\}$$

$$= \oint_\Gamma H(U\frac{\vec{\tau}_x}{\rho} + V\frac{\vec{\tau}_y}{\rho})\cdot\vec{n}\ ds - (H\nabla U,\frac{\vec{\tau}_x}{\rho}) - (H\nabla V,\frac{\vec{\tau}_y}{\rho}). \quad \ldots \quad (53)$$

Since the boundary conditions (34) and (49) imply $U=V=0$, we have

$$\oint_{\Gamma} H(U\frac{\vec{\tau}x}{\rho} + V\frac{\vec{\tau}y}{\rho})\cdot\vec{n}\ ds = 0.$$

The final check is to show

$$(H\nabla U,\frac{\vec{\tau}x}{\rho}) + (H\nabla V,\frac{\vec{\tau}y}{\rho}) \geq 0. \quad \ldots\ldots\ldots\ldots \quad (54)$$

In fact, the first constitutive law (*) yields

$$(H\nabla U,\frac{\vec{\tau}x}{\rho}) + (H\nabla V,\frac{\vec{\tau}y}{\rho})$$

$$= \int_{\Omega} H\nu_H\{2(\frac{\partial U}{\partial x})^2 + \frac{\partial U}{\partial y}(\frac{\partial U}{\partial y}+\frac{\partial V}{\partial x}) + \frac{\partial V}{\partial x}(\frac{\partial V}{\partial x}+\frac{\partial U}{\partial y}) + 2(\frac{\partial V}{\partial y})^2\}\ d\vec{x}$$

$$= \int_{\Omega} H\nu_H\{2(\frac{\partial U}{\partial x})^2 + (\frac{\partial U}{\partial y}+\frac{\partial V}{\partial x})^2 + 2(\frac{\partial V}{\partial y})^2\}\ d\vec{x} \geq 0,$$

because of the condition (10.4) and $\nu_H > 0$.

Remark 10

Choose alternatively the second constitutive law (**) in Remark 8 instead of the first one (*) in Remark 7. Then, the following additional conditions (48) and (55) are sufficient for (46) :

$$W_x = W_y = 0 \text{ in } \Omega, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (48)$$

$$\lambda U_t + (1-\lambda)\frac{\partial U_t}{\partial n} = 0 \quad \text{on } \Gamma, \quad \ldots\ldots\ldots\ldots \quad (55)$$

where $\frac{\partial U_t}{\partial n}$ denotes the outer normal derivative of U_t and λ is a constant such that $0\leq\lambda\leq 1$. For the proof, it is sufficient to check only (52). The equality (53) yields

$$\frac{1}{\rho}\ (\mathrm{div}(H\tau),\ \vec{U})$$

$$= \oint_{\Gamma} H(U\frac{\vec{\tau}x}{\rho} + V\frac{\vec{\tau}y}{\rho})\cdot\vec{n}\ ds$$

$$- (H\nabla U, \frac{\vec{\tau}_x}{\rho}) - (H\nabla V, \frac{\vec{\tau}_y}{\rho})$$

$$= \oint_\Gamma H\nu_H (U\nabla U + V\nabla V)\cdot\vec{n}\, ds - (H\nu_H \nabla\vec{U}, \nabla\vec{U}).$$

Furthermore, due to the boundary conditions (34) and (55), the condition (10.4), and $\nu_H > 0$, we have

$$\oint_\Gamma H\nu_H (U\nabla U + V\nabla V)\cdot\vec{n}\, ds = \frac{1}{2}\oint_\Gamma H\nu_H \nabla(|\vec{U}|^2)\cdot\vec{n}\, ds$$

$$= \frac{1}{2}\oint_\Gamma H\nu_H \frac{\partial}{\partial n}(|\vec{U}|^2)\, ds = \frac{1}{2}\oint_\Gamma H\nu_H \frac{\partial}{\partial n}(U_n^2 + U_t^2)\, ds$$

$$= \oint_\Gamma H\nu_H (U_n \frac{\partial U n}{\partial n} + U_t \frac{\partial U t}{\partial n})\, ds = \oint_\Gamma H\nu_H U_t \frac{\partial U t}{\partial n}\, ds$$

$$= \begin{cases} \oint_\Gamma H\nu_H \frac{\lambda - 1}{\lambda}(\frac{\partial U t}{\partial n})^2\, ds \leq 0, & \text{if } 0<\lambda\leq 1 \\ 0 & , \text{ if } \lambda = 0 \end{cases}$$

which completes the proof.

References

1) Dronkers, J.J. : Tidal computations in rivers and coastal waters, North-Holland Publishing Co., Amsterdam, 1964.

2) Leedertse, J.J. : Aspects of a computational model for long-period waterwave propagation, RAND RM-5294-PR, 1967.

3) Kaneko, Y. et al. : Numerical simulation on tidal currents and pollutant dispersion due to alternating direction implicit method - application to Osaka Bay- , Report of P.H.R.I., Vol.14, PP.3-61, 1975 (in Japanese).

4) Kanayama, H. and K. Ohtsuka : Finite element analysis on the tidal current and COD distribution in Mikawa Bay, Coastal Engineering in Japan, Vol.21, PP.157-171, 1978.

5) Zienkiewicz, O.C. and J.C. Heinrich : A unified treatment of steady-state shallow water and two-dimensional Navier-Stokes equations - finite element penalty function approach, Comp. Meth. in Appl. Mechs. and Eng., Vol.17/18, PP.673-698, 1979.

6) Gustafsson, B. and A. Sundström : Incompletely parabolic problems in fluid dynamics, SIAM J. Appl. Math., Vol.35, No.2, PP.343-357, 1978.

7) Leendertse, J.J. and S-K.Liu : A three-dimensional model for estuaries and coastal seas : Vol.II, aspects of computation, RAND R-1764-OWRT, 1975.

8) Bird, R.B. et al. : Transport phenomena, John Wiley & Sons, Inc., 1960.

9) Connor, J.J. and C.A. Brebbia : Finite element techniques for fluid flow, Newnes - Butterworths, London, 1976.

10) Pinder, G.E. and W.G. Gray : Finite element simulation in surface and subsurface hydrology, Academic Press, 1977.

11) Kawahara, M.: Steady and unsteady finite element analysis of incompressible viscous fluid, Finite Elements in Fluids, Vol.3, edited by Gallagher et al., John Wiley and Sons, PP.23-54, 1978.

12) Wang, H-P. : Multi-leveled finite element hydrodynamic model of Block Island Sound, Finite Elements in Water Resources, Vol.1, PP.469-493, Pentech, 1977.

13) Tanaka, T. and Y. Ono : Finite element analysis of typhoon surge in Ise Bay, Proc. of U.S. Japan Seminar on Interdisciplinary Finite Element Analysis, Cornell Univ., 1978.

14) Walters, R.A. and R.T. Cheng : A two-dimensional hydrodynamic model of a tidal estuary, Advances in Water Resources, Vol.2, PP.177-184, 1979.

15) Praagman, N. : A finite element solution of the shallow water equations, Ph. D., Delft, 1979.

Optimal error estimates for H^{-1}-Galerkin method for parabolic problems with time dependent coefficients

By

Mitsuhiro TOMONAGA

§1. Introduction.

An H^{-1}-Galerkin method was introduced by Rachford and Wheeler [5] as an approximation scheme for a numerical solution of boundary value problem of a differential equation. Wheeler [6] applied the procedure to a one space dimensional parabolic problem with time independent coefficients and obtained optimal L^2 and L^∞ error estimates. The collocation-Galerkin methods, which are mixed schemes of a collocation method and a Galerkin method, were introduced by Diaz, and several results were derived by Diaz [2], Dunn and Wheeler [4] and Wheeler [7] for a two point boundary value problem. In particular, using discontinuous piecewise polynomial spaces with collocation based on the Jacobi points, Diaz [3] analyzed the collocation-H^{-1}-Galerkin method for one space dimensional parabolic problem with time dependent coefficients and derived optimal L^2-estimates.

In this paper we clarify essential analogy between the H^{-1}-Galerkin method and the collocation-H^{-1}-Galerkin method, and, using this property, we extend the results in [6] to the case of time-dependent coefficients and also derive an optimal L^{∞}-estimate for the scheme in [3]. In the next section, we describe a parabolic problem and define an H^{-1}-Galerkin method. In §3, the collocation-H^{-1}-Galerkin method is described and it is explained how these two methods are connected with each other. In §4, we present L^2-and L^{∞}-global rates of convergence for the H^{-1}-Galerkin method, by the similar procedures to that of [3] and [4] for the collocation-H^{-1}-Galerkin method. Finally we derive L^{∞}-error estimate for the collocation-H^{-1}-Galerkin method using similar arguements in §4.

§2. A parabolic problem and H^{-1}-Galerkin method.

We consider the following initial boundary value problem to parabolic equations in a single space variable.

$$
(2.1) \qquad \begin{aligned}
\frac{\partial u}{\partial t} + Lu &= f(x,t), && (x,t) \in I \times J, \\
u(x,0) &= u_0(x), && x \in I \\
u(1,t) = u(0,t) &= 0, && t \in J
\end{aligned}
$$

where $I=(0,1)$, $J=(0,T]$, and the elliptic operator L is defined by

$$Lu = -\frac{\partial}{\partial x}\left(a(x,t)\frac{\partial u}{\partial x}\right) + b(x,t)\frac{\partial u}{\partial x} + c(x,t)u.$$

Assume that there exist positive constants a_0 and a_1 such that

$$a_0 \leqq a(x,t) \leqq a_1, \qquad (x,t) \in I \times J.$$

Also assume that a_x, a_{xx}, a_{xt}, a_t, b_x, b_{xx}, b_{xt}, b_t, c_x and c_t

are in $L^\infty(I)$ uniformly on t, $f(\cdot,t)$ is in $L^2(I)$ and Hölder continuous uniformly on t and u_0 is in $C_0^1(I)$.

In the following sections, the symbol u denotes the solution to (2.1) and a, b and c denote the coefficients of the operator L unless otherwise stated. We denote the adjoints of the operator L by

$$L^*v = -\frac{\partial}{\partial x}(a\frac{\partial v}{\partial x}) - \frac{\partial}{\partial x}(bv) + cv,$$

and also define the operator L_t by

$$L_t u = -\frac{\partial}{\partial x}(\frac{\partial a}{\partial t}\cdot\frac{\partial u}{\partial x}) + \frac{\partial b}{\partial t}\cdot\frac{\partial u}{\partial x} + \frac{\partial c}{\partial t}\cdot u$$

and denote its adjoint by

$$L_t^*v = -\frac{\partial}{\partial x}(\frac{\partial a}{\partial t}\cdot\frac{\partial}{\partial x}v) - \frac{\partial}{\partial x}(\frac{\partial b}{\partial t}\cdot v) + \frac{\partial c}{\partial t}\cdot v.$$

Now, we introduce the following notation. Let $\Delta = \{0=x_0 < x_1 < \cdots\cdots < x_N = 1\}$ be a partition of I and set $I_i=(x_{i-1}, x_i)$, $h_i=x_i-x_{i-1}$ and $h = \max_{1\le i\le N} h_i$. For $E \subset I$ and a positive integer r, denote by $P_r(E)$ the class of polynomials of degree not greater than r restricted to E. We define

$$\mathcal{M}_{-1}^r = \{v : v|_{I_i} \in P_r(I_i),\ 1\leq i\leq N\},$$

$$\mathcal{M}_k^r = \{C^k(I) \cap \mathcal{M}_{-1}^r \cap H_0^1(I),\quad k\geq 0\},$$

$$Z_k^r = \{v\in\mathcal{M}_k^r | \frac{d^j v}{dx^j}(x_i)=0,\quad 0\leq i\leq N,\quad 0\leq j\leq k\}$$

and

$$Z_k^r(I_i) = \{v\in P_r(I_i) | \frac{d^j v}{dx^j}(x_{i-1}) = \frac{d^j v}{dx^j}(x_i)=0,\quad 0\leq j\leq k\}.$$

Also let

$$(\phi,\psi) = \int_0^1 \phi(x)\psi(x)\,dx$$

and

$$(\phi,\psi)_E = \int_E \Phi(x)\psi(x)\,dx, \qquad E\subset I.$$

The H^{-1} continuous time Galerkin approximation is defined by a map $U:\bar{J} \to \mathcal{M}_{-1}^{r}$, where

$$(2.2) \qquad (\frac{\partial U}{\partial t}, v) + (U, L^*v) = (f, v) \qquad v\in\mathcal{M}_1^{r+2},\ t\in J.$$

$U(\cdot,0)$ will be defined later.

The above equality relation defines a system of ordinary differential equations for the time variable and determines U uniquely, once $U(\cdot,0)$ has been specified.

Now we introduce the following additional notation. For an open interval E and a nonnegative integer m, let $H^m(E)$ and $W_\infty^m(E)$ denote the closure of $C^\infty(\bar{E})$ in the norms

$$\|f\|_{H^m(E)} = \left(\sum_{j=0}^{m} \|f^{(j)}\|^2_{L^2(E)}\right)^{1/2}$$

and

$$\|f\|_{W_\infty^m(E)} = \sum_{j=0}^{m} \|f^{(j)}\|_{L^\infty(E)},$$

respectively.

We define $H^{-m}(E) = (H^m(E))'$, using the norm

$$\|f\|_{H^{-m}(E)} = \sup_{0\neq g\in H^m(E)} \frac{|(f,g)|}{\|g\|_{H^m(E)}} .$$

When E is equal to I or I_i, we shall suppress the symbol I or I_i and simply denote by i (e.g. $\|f\|_{L^2}$ and $\|f\|_{L^2(i)}$ mean $\|f\|_{L^2(I)}$ and $\|f\|_{L^2(I_i)}$, respectively).

Also, for X a normed space with norm $\|\cdot\|_X$, and for maps $\phi:\bar{J} \longrightarrow X$, we adopt the notation

$$\|\phi\|^2_{L^2(J:X)} = \int_J \|\nu(t)\|^2_X dt$$

and

$$\|\phi\|_{L^\infty(J:X)} = \operatorname*{ess\,sup}_{t\in J} \|\nu(t)\|_X .$$

§3. The relation between the H^{-1}-Galerkin method and the collocation-H^{-1}-Galerkin method.

The H^{-1}-Galerkin method is closely related with the collocation-H^{-1}-Galerkin method as described below. The collocation-H^{-1}-Galerkin method is defined as a map $U:\bar{J} \longrightarrow \mathcal{M}^r_{-1}$ satisfying the following collocation relations and weak form. That is, for $t \in J$

$$\text{(3.1 i)} \quad \begin{aligned} &\frac{\partial U}{\partial t}(x_{ij},t) + LU(x_{ij},t) = f(x_{ij},t), \quad 1\leq j\leq r-1, \quad 1\leq i\leq N, \\ &(\frac{\partial U}{\partial t},v) + (U,L^*v) = (f,v), \quad v\in\mathcal{M}^3_1, \end{aligned}$$

and for t=0

$$\text{(3.1 ii)} \quad \begin{aligned} &L(U-u_0)(x_{ij},0)=0 \quad 1\leq j\leq r-1, \quad 1\leq i\leq N, \\ &(U(\cdot,0)-u_0, L^*v) \quad\quad v\in\mathcal{M}^3_1, \end{aligned}$$

where x_{ij} are points in I_i which are defined by the following transformation of the Jacobi points ρ_j on I with weight function $x^2(1-x)^2$,

$$x_{ij} = x_{i-1} + h_i\rho_j, \qquad 1 \leqq j \leqq r-1, \qquad 1 \leqq i \leqq N.$$

It follows that the map U satisfying (3.1 i) is represented equivalently in the following semidiscrete bilinear form, that is,

$$(3.2) \qquad \langle U_t, v\rangle + \mathcal{L}(U,v) = \langle f,v\rangle, \qquad v \in \mathcal{M}_1^{r+2}, \qquad t \in J,$$

where $\qquad \mathcal{L}(U,v) = \langle LU, v_1\rangle + (U, L^* v_2)$

and $\qquad \langle g,v\rangle = \sum_{i=1}^{N} (\langle g,v_1\rangle_i + (g,v_2)_i),$

for $v = v_1 + v_2 \in \mathcal{M}_1^{r+2}$ such that $v_1 \in Z_1^{r+2}$, $v_2 \in \mathcal{M}_1^3$,

where for g defined on each I_i,

$$\langle g, v_1\rangle_i = h_i \sum_{j=1}^{r-1} w_j \frac{g(x_{ij}, t) v_1(x_{ij})}{\rho_j^2 (1-\rho_j)^2}, \qquad (w_j > 0).$$

Since the equations (2.2) and (3.1 i)(or (3.2)) are the same form for $v \in \mathcal{M}_1^3$ and $\mathcal{M}_1^{r+2} = Z_1^{r+2} \oplus \mathcal{M}_1^3$, the collocation-$H^{-1}$-Galerkin method completely coincides with the H^{-1}-Galerkin method, replacing the discrete inner product $\langle \cdot, v_1\rangle$ in the former with the continuous inner product $(\cdot, v_1)$. This equality does not hold in general, although $\langle \phi, v\rangle_i = (\phi, v)_i$ for $\phi \cdot v \in P_{2r+1}(I_i)$. Therefore, these two methods define different approximation schemes. However, there exist following correspondences between various properties derived by each scheme.

(1) If an equality for the collocation points holds, as in

most cases, the equality of the same type holds but it is transformed to weak form in the H^{-1}-Galerkin method. For example, if the equality

$$LU(x_{ij},t) = \phi(x_{ij},t), \quad 1 \leqq j \leqq r-1, \quad 1 \leqq i \leqq N, \quad t \in J,$$

holds, one will be able to predict that equality of the following type should be established for the H^{-1}-Galerkin method.

$$(LU,v) = (\phi,v), \qquad v \in Z_1^{r+2}, \qquad t \in J.$$

(2) The argument for the one method, based upon the weak forms whose test functions are in $\mathcal{M}_1^3$, are completely identical with the one of the other method.

(3) There are almost no differences between the results of the error estimation (including intermediate results) for both methods.

As described in the next section, because of the above reasoning, we can estimate the error for the H^{-1}-Galerkin method using the techniques analogous to that of [3] and [4] for the collocation-H^{-1}-Galerkin method.

§4. Optimal error estimates for the H^{-1}-Galerkin method.

4.1 L^2-error-estimate.

First we show the following lemma which is used often later.

Lemma 1. Let $f \in L^2(I_i)$, then for each $t \in \bar{J}$ there exists a unique polynomial $Y \in P_{r-2}(I_i)$ $(r \geq 2)$ satisfying

$$(4.1) \qquad (aY,v) = (f,v), \qquad v \in Z_1^{r+2}(I_i),$$

and there exists a constant C, independent of h_i, such that

$$(4.2) \qquad \|Y\|_{L^2(I_i)} \leq C\|f\|_{L^2(I_i)} \quad .$$

Proof. Clearly it suffices to show that the conclusion holds for a unit interval (i.e. $I_i = I$). Since $P_{r-2}(I)$ and $Z_1^{r+2}(I)$ are of the same dimension, existence of Y satisfying (4.1) follows from uniqueness. On the other hand, the uniqueness immediately results from (4.2). Thus, if we show (4.2), then the proof is completed.

Now we define a norm on $P_r(I)$ by

$$|||\phi||| = \|x(1-x)\phi\|_{L^2(I)} \quad .$$

Then, since all norms are equivalent on the (r+1)-dimensional space of $P_r(I)$, there exist constants C_1 and C_2 (positive and independent of ϕ) such that

$$C_1\|\phi\|_{L^2} \leq |||\phi||| \leq C_2\|\phi\|_{L^2} \quad .$$

Hence,

$$\frac{\|x(1-x)Y\|_{L^2}^2}{\|Y\|_{L^2}\|x^2(1-x)^2Y\|_{L^2}} = \frac{|||Y|||}{\|Y\|_{L^2}} \cdot \frac{\|x(1-x)Y\|_{L^2}}{|||x(1-x)Y|||} \geq C_1/C_2 \quad .$$

Therefore, by the assumptions of a and (4.1),

$$\|Y\|_{L^2} \leq C\frac{\|x(1-x)Y\|_{L^2}^2}{\|x^2(1-x)^2Y\|_{L^2}} \leq C\frac{|(aY,x^2(1-x)^2Y)|}{\|x^2(1-x)^2Y\|_{L^2}} \leq C\|f\|_{L^2} \quad ,$$

which concludes the proof.

Hereafter, without notice, we assume $r \geqq 2$, but it will be easily seen by the process of the proof that each proposition in this section (except lemma 1) is also valid for r=1.

Now, for the error analysis we define as in [4] the following projection $Y:\bar{J} \longrightarrow \mathcal{M}_{-1}^{r}$ satisfying for t $\bar{J}$.

(4.3) $\qquad (a(Y-u)'',v_1) = 0, \qquad v_1 \in Z_1^{r+2},$

(4.4) $\qquad (Y-u,L^*v_2) = 0, \qquad v_2 \in \mathcal{M}_1^3.$

Existence and uniqueness of Y satisfying (4.3) and (4.4), for h sufficiently small, immediately result from the following lemma.

Lemma 2. If $u(\cdot,t) \in H^{r+1}(I)$ $(t \in \bar{J})$, then, for h sufficiently small, there exists a constant C, independent of h, such that for Y satisfying (4.3) and (4.4)

$$\| Y - u \|_{L^2} \leqq C\left(\sum_{i=1}^{N} h_i^{2(r+1)} \|u\|^2_{H^{r+1}(I_i)}\right)^{1/2}, \qquad t \in \bar{J} .$$

Proof. For fixed t, from Lemma 1, the local projection $\Pi_i : H^{r-1}(I_i) \longrightarrow P_{r-2}(I_i)$ $(1 \leqq i \leqq N)$ determined by (4.3) is a continuous linear mapping. And if u" $P_{r-2}(I_i)$ then $\Pi_i u''=u''$. Hence, from the well known result of approximation theory (e.g. [1]) there exists a positive constant C such that

(4.5) $$\| Y'' - u'' \|_{L^2(I_i)} \leqq C h_i^{r-1} \|u\|_{H^{r+1}(I_i)} .$$

The rest of the proof is completed by an argument similar to that of lemma 2 in [4].

As in [6] we define an elliptic projection $W:\bar{J} \longrightarrow \mathcal{M}_{-1}^{r}$ satisfying for $t\in\bar{J}$

$$(4.6) \qquad (W-u,L^*v) = 0, \qquad v\in\mathcal{M}_{1}^{r+2} .$$

It was known in [5] that, for h sufficiently small, there exists a unique function W satisfying (4.6). We have the following estimation for u-W.

Theorem 1. Let W be the solution to (4.6). If $u(\cdot,t)\in H^{r+1}(I)$ $(t\in J)$, for h sufficiently small, there exists a constant C, independent of h, such that

$$\|u - W\|_{L^2} + (\sum_{i=1}^{N} h_i^4\|u - W\|^2_{H^2(I_i)})^{1/2} \leq C(\sum_{i=1}^{N} h_i^{2(r+1)}\|u\|^2_{H^{r+1}(I_i)})^{1/2} .$$

Proof. As the proof is similar to that of Theorem 1 in [4], we only show principal parts below.

Let η=W-Y and ζ=u-Y, where Y is as in Lemma 2, then for $v\in Z_1^{r+2}(I_i)$, $(L\eta,v)_i=(L\zeta,v)_i$. Since $(a\zeta'',v)_i=0$, we have by Lemma 1

$$\|\eta''\|_{L^2(I_i)} \leq C(\|\eta'\|_{L^2(I_i)} + \|\eta\|_{L^2(I_i)} + \|\zeta'\|_{L^2(I_i)} + \|\zeta\|_{L^2(I_i)})$$

Using the well known inequality $\|f'\|_{L^2(I_i)} \leq C(h_i\|f''\|_{L^2(I_i)} + h_i^{-1}\|f\|_{L^2(I_i)})$, for h sufficiently small, we have

$$(4.7) \qquad \|\eta''\|_{L^2(i)} \leq C(h_i^{-1}\|\eta\|_{L^2(i)} + h_i\|\zeta''\|_{L^2(i)} + h_i^{-1}\|\zeta\|_{L^2(i)}) .$$

Thus, as in [4], we get the following estimate.

$$(4.8) \qquad \|\eta\|_{L^2} \leq C\ h(\sum_{i=1}^{N} h_i^{2(r+1)}\|u\|^2_{H^{r+1}(i)})^{1/2} .$$

Hence, by Lemma 2 we obtain

$$(4.9) \qquad \|u - W\|_{L^2} \leq C(\sum_{i=1}^{N} h_i^{2(r+1)}\|u\|^2_{H^{r+1}(i)})^{1/2} .$$

Moreover, from (4.5), (4.7), (4.8) and Lemma 2, for h sufficiently small, we have

$$(4.10) \qquad \sum_{i=1}^{N} h_i^4 \|(u - W)''\|^2_{L^2(i)} \leq C(\sum_{i=1}^{N} h_i^{2(r+1)} \|u\|^2_{H^{r+1}(i)}) .$$

Also,

$$(4.11) \qquad \sum_{i=1}^{N} h_i^4 \|(u-W)'\|^2_{L^2(i)} \leq C\ h^2\{(\sum_{i=1}^{N} h_i^4 \|(u-W)''\|^2_{L^2(i)}) + \|u-W\|^2_{L^2(i)}\} .$$

From (4.9)-(4.11) the conclusion immediately follows.

Next, we estimate an error for time derivative of W. One can easily verify that, if $u_t(\cdot,t)\in L^2(I)$ $(t\in J)$, from the definition of W and the assumption for the coefficients of L*, $\frac{\partial W}{\partial t}(\cdot,t)\in L^2(I)$.

Let $\tilde{W}:\bar{J} \longrightarrow \mathcal{M}_{-1}^{r}$ be the elliptic projection of u_t defined by

$$(4.12) \qquad (\tilde{W} - u_t, L^*v) = 0, \qquad v\in\mathcal{M}_1^{r+2}, \quad t\in\bar{J} .$$

We have the following lemma as in [3].

Lemma 3. Let W and W_t be the solutions to (4.6) and (4.12) respectively. For h sufficiently small, there exists a constant C, independent of h, such that

$$\|\tilde{W} - W_t\|_{L^2} \leqq C(\|u - W\|_{L^2} + (\sum_{i=1}^{N} h_i^4 \|u - W\|_{H^2(i)}^2)^{1/2}) .$$

Proof. Let $\eta = W_t - \tilde{W}$ and $=W-u$. Then, for $v \in Z^{r+2}(I_i)$ and each $t \in J$,

$$\begin{aligned}(L\eta, v)_i &= (L\eta, v)_i + (L(\tilde{W} - u_t), v)_i \\ &= ((L\xi)_t, v)_i - (L_t\xi, v)_i \\ &= -(L_t\xi, v)_i .\end{aligned}$$

Therefore,

$$(a\eta'', v)_i = ((-a_x + b)\eta' + c\eta - a_t\xi'' - (a_{tx} - b_t)\xi' + c_t\xi, v)_i , \qquad v \in Z_1^{r+2}(I_i) .$$

Hence, if h is sufficiently small, by Lemma 1

$$\begin{aligned}\|\eta''\|_{L^2(i)} &\leqq C(\|\eta'\|_{L^2(i)} + \|\eta\|_{L^2(i)} + \|\xi''\|_{L^2(i)} + \|\xi'\|_{L^2(i)} \\ &\qquad + \|\xi\|_{L^2(i)}) \\ &\leqq C(h_i^{-1}\|\eta\|_{L^2(i)} + \|\xi\|_{H^2(i)})\end{aligned}$$

By an argument similar to that of Lemma 3.2 in [3], the desired result is obtained.

We obtain the following result from Lemma 3 and Theorem 1.

Theorem 2. Let W be the solution to (4.6). If $u(\cdot,t)$ and $u_t(\cdot,t)\in H^{r+1}(I)$ $(t\in J)$, then, for h sufficiently small, there exists a constant C, independent of h, such that

$$\left\|\frac{\partial}{\partial t}(u - w)\right\|_{L^2} \leqq C\{\sum_{i=1}^{N} h_i^{2(r+1)}(\|u\|^2_{H^{r+1}(i)} + \left\|\frac{\partial u}{\partial t}\right\|^2_{H^{r+1}(i)})\}^{1/2} .$$

Now, let U and W be the solutions to (2.2) and (4.6), respectively. Also, let $\zeta=U-W$ and $\xi=u-W$. Then we have

$$(\frac{\partial \zeta}{\partial t},v) + (\zeta,L^*v) = (\xi_t,v), \qquad v\in\mathcal{M}_1^{r+2}, \quad t\in J.$$

Therefore, by Lemma 1 in [6], which is valid even if the coefficients of L^* depends on time variable, we obtain the following estimate.

$$(4.13) \qquad \|U-W\|_{L^\infty(J:L^2)} \leqq C(\|(U-W)(0)\|_{L^2} + \left\|\frac{\partial}{\partial t}(u-W)\right\|_{L^2(J:L^2)})$$

From the above results, we have the following main theorem for the L^2-error-estimate.

Theorem 3. Let u be the solution to (2.1) and U be the solution to (2.2) with $U(\cdot,0)=W(\cdot,0)$. If u and $u_t\in L^2(J:H^{r+1})$ and $\|u(\cdot,t)\|_{H^{r+1}(i)}\in H^1(J)$ $(1\leq i\leq N)$, then, for h sufficiently small, there exists a constant C, independent of h, such that

$$\|u-U\|_{L^\infty(J:L^2)} < C\{\sum_{i=1}^{N} h_i^{2(r+1)}(\|u\|^2_{L^2(J:H^{r+1}(i))} + \left\|\frac{\partial u}{\partial t}\right\|^2_{L^2(J:H^{r+1}(i))})\}^{1/2} .$$

Proof. Using Theorem 1, Theorem 2, (4.13) and the triangle inequality,

$$\|u-U\|_{L^\infty(J:L^2)} \leqq C\{\operatorname*{ess\,sup}_{t\in J} (\sum_{i=1}^{N} h_i^{2(r+1)} \|u\|^2_{H^{r+1}(i)})^{1/2}$$

$$+ (\sum_{i=1}^{N} h_i^{2(r+1)} (\|u\|^2_{L^2(J:H^{r+1}(i))} + \|\frac{\partial u}{\partial t}\|^2_{L^2(J:H^{r+1}(i))}))^{1/2}\} .$$

By the imbedding theorem, for each i ($1\leqq i\leqq N$)

$$\sup_{t\in J} \|u\|^2_{H^{r+1}(i)} \leqq C(\|u\|^2_{L^2(J:H^{r+1}(i))} + \int_J (\frac{d}{dt}\|u\|_{H^{r+1}(i)})^2 dt) .$$

One can easily verify that

$$\frac{d}{dt}\|u\|_{H^{r+1}(i)} \leqq \|\frac{\partial u}{\partial t}\|_{H^{r+1}(i)}, \qquad t\in J .$$

Hence, we obtain the result.

4.2 L^∞-error-estimate

In this section, we estimate the error with L^∞-norm. However, we need to assume that the partition Δ satisfy semi-uniformity. That is, there exists a constant K_0 such that

$$\max_{1\leqq i,j\leqq N} h_i h_j^{-1} \leqq K_0 h^{-1}$$

First, note that (4.13) may be improved to the following ([6]).

$$(4.14) \qquad \|U-W\|_{L^\infty(J:L^2)} \leqq C(\|(U-W)(0)\|_{L^2} + \|\frac{\partial}{\partial t}(u-W)\|_{L^2(J:H^{-1})}) .$$

Next, we have the following lemma corresponding to Theorem 2.

Lemma 4. Assume the hypothesis of Theorem 2. Then, for h sufficiently small,

$$\left\|\frac{\partial}{\partial t}(u-W)\right\|_{H^{-1}} \leq C\, h\Big(\sum_{i=1}^{N} h_i^{2(r+1)}\big(\|u\|^2_{H^{r+1}(i)} + \left\|\frac{\partial u}{\partial t}\right\|^2_{H^{r+1}(i)}\big)\Big)^{1/2}.$$

Proof. Let $\tilde{W}$ be the solution to (4.12). We consider for each $\psi \in H^1(I)$ the dual problem,

$$L^*\Phi = \psi, \qquad \text{on } I,$$

$$\Phi(0) = \Phi(1) = 0.$$

Now, let $\eta = W_t - \tilde{W}$ and $\xi = W-u$, then for each $\hat{\Phi} \in \mathcal{M}_1^3$, we have

$$(4.15) \qquad (\eta,\psi) = (\eta, L^*(\Phi-\hat{\Phi})) - (\xi, L_t^*\hat{\Phi})$$

Choosing $\hat{\Phi}$ appropriately, that is, $\hat{\Phi}^{(j)}(x_i) = \Phi^{(j)}(x_i)$, $j \in \{0,1\}$, $0 \leq i \leq N$, one obtains by elliptic regualarity,

$$|(\eta, L^*(\Phi-\hat{\Phi}))| \leq C \sum_{i=1}^{N} \|\eta\|_{L^2(i)} h_i \|\Phi\|_{H^3(i)} \leq C\, h\|\eta\|_{L^2}\|\Phi\|_{H^3}$$

$$\leq C\, h\|\eta\|_{L^2}\|\psi\|_{H^1}.$$

To estimate the second term of (4.15) we define the norms $|||\cdot|||_{H^s}$ and $|||\cdot|||_{H^{-s}}$ for a positive integer s by

$$|||f|||_{H^s} = \Big(\sum_{i=1}^{N} \|f\|^2_{H^s(i)}\Big)^{1/2}, \qquad f \in H^s(I_i), \quad i \leq i \leq N,$$

and

$$|||g|||_{H^{-s}} = \sup_{\substack{0 \neq f \in H^s(i) \\ 1 \leq i \leq N}} \frac{|(g,f)|}{|||f|||_{H^s}}, \qquad g \in L^2(I),$$

respectively.

Now we have

$$|(\xi, L_t^*\Phi)| \leq |||\xi|||_{H^{-1}} |||L_t^*\hat{\Phi}|||_{H^1} \leq C|||\xi|||_{H^{-1}} \left(\sum_{i=1}^{N} \|\Phi\|^2_{H^3(i)}\right)^{1/2} \leq C|||\xi|||_{H^{-1}} \|\Phi\|_{H^3}$$

$$\leq C|||\xi|||_{H^{-1}} \|\psi\|_{H^1}$$

Thus, by (4.15)

$$\|\eta\|_{H^{-1}} \leq C(h\|\eta\|_{L^2} + |||\xi|||_{H^{-1}}):$$

From Lemma 3 and Theorem 1,

$$\|\eta\|_{L^2} \leq C\left(\|\xi\|_{L^2} + \left(\sum_{i=1}^{N} h_i^4 \|\xi\|^2_{H^2(i)}\right)^{1/2}\right)$$

$$\leq C\left(\sum_{i=1}^{N} h_i^{2(r+1)} \|u\|^2_{H^{r+1}(i)}\right)^{1/2} \quad .$$

It follows that from the estimates for the negative norms in [5]

$$|||\xi|||_{H^{-1}} \leq C\ h\left(\sum_{i=1}^{N} h_i^{2(r+1)} \|u\|^2_{H^{r+1}(i)}\right)^{1/2} \quad .$$

Hence, $$\|\eta\|_{H^{-1}} \leq C\ h\left(\sum_{i=1}^{N} h_i^{2(r+1)} \|u\|^2_{H^{r+1}(i)}\right)^{1/2} \quad .$$

Therefore,

$$\left\|\frac{\partial}{\partial t}(u-W)\right\|_{H^{-1}} \leq \|u_t - \tilde{W}\|_{H^{-1}} + \|\tilde{W} - W_t\|_{H^{-1}}$$

$$\leq C\ h\left(\sum_{i=1}^{N} h_i^{2(r+1)} \left(\|u\|^2_{H^{r+1}(i)} + \|u_t\|^2_{H^{r+1}(i)}\right)\right)^{1/2} ,$$

which concludes the proof.

Next, by the semi-uniformity of the partition and the well-known property of the piecewise-polynomial, we have

$$\| U - W \|_{L^\infty(I\times J)} \leqq C\, h^{-1} \| U - W \|_{L^\infty(J:L^2)} .$$

Using the estimate in [5], we obtain for each $t \in J$

$$\| u - W \|_{L^\infty} \leqq C \max_{1\leqq i\leqq N} h_i^{r+1} \| u \|_{W_\infty^{r+1}(i)} .$$

Thus, if we choose $U(\cdot,0)=W(\cdot,0)$, from (4.14), Lemma 4 and the triangle inequality, we have the following L^∞-estimate.

Theorem 4. Let u be the solution to (2.1) and U be the solution to (2.2) with $U(\cdot,0)=W(\cdot,0)$. If $u \in L^\infty(J:W_\infty^{r+1})$ and $\frac{\partial u}{\partial t} \in L^2(J:H^{r+1})$, then, for h sufficiently small, there exists a constant C, independent of h, such that

$$\| u - U \|_{L^\infty(I\times J)} \leqq C \max_{1\leqq i\leqq N} h_i^{r+1} \| u \|_{L^\infty(J:W_\infty^{r+1}(i))}$$

$$+ C\Big\{ \sum_{i=1}^{N} h_i^{2(r+1)} \Big(\| u \|^2_{L^2(J:H^{r+1}(i))} + \Big\| \frac{\partial u}{\partial t} \Big\|^2_{L^2(J:H^{r+1}(i))} \Big) \Big\}^{1/2} .$$

§5. L_∞-error-estimate for the collocation-H^{-1}-Galerkin method.

In this section, applying the arguments as in the previous section, we estimate optimal L^∞ global rate of convergence for the collocation-H^{-1}-Galerkin method described in §3. We also assume that the partition of the interval I is semi-uniform as in §4.2. Since each argument is analogous to that of §4.2 or in [3], the detailed proof will be omitted.

As before let u denote the solution to (2.1), but the map

$U:\bar{J} \rightarrow \mathcal{M}^r_{-1}$ denote the solution to (3.1). In order to obtain estimate for u-U, we estimate the error for elliptic projection $W:\bar{J} \rightarrow \mathcal{M}^r_{-1}$ defined as follows.

(5.1) $\quad \mathcal{L}(W,v) = \mathcal{L}(u,v), \qquad v \in \mathcal{M}^{r+2}_1, \qquad t \in \bar{J}.$

First, we obtain the following lemma corresponding to lemma 4. Using the estimates derived in [4], the proof is quite similar to that of Lemma 4.

Lemma 4'. Let W be the solution to (5.1). If $u(\cdot,t)$ and $\frac{\partial u}{\partial t}(\cdot,t) \in H^{r+1}(I)$ $(t \in J)$, then, for h sufficiently small, there exists a constant C, independent of h, such that

$$\left\| \frac{\partial}{\partial t}(u-W) \right\|_{H^{-1}} \leq C\, h\Big\{ \sum_{i=1}^{N} h_i^{2(r+1)} \Big(\|u\|^2_{H^{r+1}(i)} + \left\| \frac{\partial u}{\partial t} \right\|^2_{H^{r+1}(i)} \Big) \Big\}^{1/2} .$$

Now, from Lemma 4' and Theorem 3.1 in [3], we have immediately the following result.

Lemma 5. Let W be the solution to (5.1). If $u(\cdot,t)$ and $\frac{\partial u}{\partial t}(\cdot,t) \in H^{r+1}(I)$ $(t \in J)$, then, for h sufficiently small, there exists a constant C, independent of h, such that

$$\left\| \frac{\partial}{\partial t}(u-W) \right\|^2_{H^{-1}} + \sum_{i=1}^{N} h_i^4 \left\| \frac{\partial^2}{\partial x \partial t}(u-W) \right\|^2_{L^2(i)} \leq C\, h^2 \Big(\sum_{i=1}^{N} h_i^{2(r+1)} \Big(\|u\|^2_{H^{r+1}(i)} + \left\| \frac{\partial u}{\partial t} \right\|^2_{H^{r+1}(i)} \Big) \Big).$$

The following lemma is an improvement of Theorem 4.2 in [3].

Lemma 6. For h sufficiently small, there exists a constant C, independent of h, such that

$$\|U - W\|_{L^\infty(J:L^2)} \leq C\left(\left\|\frac{\partial}{\partial t}(u-W)\right\|^2_{L^2(J:H^{-1})} + \sum_{i=1}^{N} h_i^4 \left\|\frac{\partial^2}{\partial x \partial t}(u-w)\right\|^2_{L^2(J:L^2(i))}\right)^{\frac{1}{2}} .$$

Proof. Let ζ=U-W and ξ=u-W, then from (3.2) and (5.1) it follows that

$$<\zeta_t, v> + \mathcal{L}(\zeta, v) = <\xi_t, v>, \qquad v \in \mathcal{M}_1^{r+2}, \quad t \in J.$$

Therefore, using the estimates in [3], we have

$$|<\xi_t, V>| \leq |(\xi_t, V)| + |<\xi_t, V> - (\xi_t, V)|$$

$$\leq \|\xi_t\|_{H^{-1}} \|V\|_{H^1} + C \sum_{i=1}^{N} h_i^2 \|\xi_{tx}\|_{L^2(i)} \left\|\frac{\partial V}{\partial x}\right\|_{L^2(i)}$$

$$\leq C(\|\xi_t\|^2_{H^{-1}} + \sum_{i=1}^{N} h_i^4 \|\xi_{tx}\|^2_{L^2(i)}) + C\left\|\frac{\partial V}{\partial x}\right\|^2_{L^2} .$$

The quite analogous argument of the proof for Theorem 4.2 in [3], except the above, enables one to verify the desired result.

Now, using Lemma 5, Lemma 6 and the estimate in [4], an argument similar to that of §4.2 yields the following L^∞-error-estimate corresponding to Theorem 4.

Theorem 4'. Let u be the solution to (2.1) and U be the solution to (3.1). If $u \in L^\infty(J:W_\infty^{r+1})$ and $\frac{\partial u}{\partial t} \in L^2(J:H^{r+1})$, then, for h sufficiently small, there exists a constant C, independent of h, such that

$$\|u - U\|_{L^\infty(I\times J)} \leq C \max_{1\leq i\leq N} h_i^{r+1}\|u\|_{L^\infty(J:W_\infty^{r+1}(i))}$$

$$+ C\{\sum_{i=1}^{N} h_i^{2(r+1)}(\|u\|^2_{L^2(J:H^{r+1}(i)}$$

$$+ \left\|\frac{\partial u}{\partial t}\right\|^2_{L^2(J:H^{r+1}(i))})\}^{1/2} .$$

The author is grateful to Professor Seiiti Huzino for his polite guidance and profitable suggestions throughout this work.

References

[1] Ciarlet, P.G. & Raviart, P.A., General Lagrange and Hermite interpolation in R^n with applications to finite element methods, Arch. Rational Mech. Anal., 46 (1972), 177-199.

[2] Diaz, J.C., A collocation-Galerkin method for the two point boundary value problem using continuous piecewise polynomial spaces, SIAM J. Numer. Anal., 14 (1977), 844-858.

[3] Diaz, J.C., Collocation-H^{-1}-Galerkin method for parabolic problems with time dependent coefficients, SIAM Journal of Numerical Analysis, 16 (1979), 911-922.

[4] Dunn, R.J. & Wheeler, M.F., Some collocation-Galerkin method for two-point boundary value problems, SIAM J. Numer. Anal., 13 (1976), 720-733.

[5] Rachford, H.H. & Wheeler, M.F., An H^{-1}-Galerkin procedure for two-point boundary value problem, Mathematical aspects

of finite elements in partial differential equations, C. de Boor, ed., Academic Press, New York, (1974), 353-382.

[6] Wheeler, M.F., An H^{-1}-Galerkin method for parabolic problems in a single space variable, SIAM J. Numer. Anal., 12 (1975), 803-817.

[7] Wheeler, M.F., A C^0-collocation-finite element method for two point boundary value problems and one space dimensional parabolic problems, SIAM J. Numer. Anal., 14 (1977), 71-90.