An Orthogonal Collocation Method for Sobolev Equations

Xiaofei Lu\textsuperscript{1, a} and Ning Ma\textsuperscript{2, b}\textasteriskcentered

\textsuperscript{1}China University of Petroleum, Beijing, China
\textsuperscript{2}China University of Petroleum, Beijing, China
\textsuperscript{a}luxf012@126.com,  \textsuperscript{b}bjmaning@163.com;

Keywords: sobolev equations; orthogonal collocation method; error estimate.

Abstract. In this article, an orthogonal collocation method for the initial-boundary value problem for sobolev equations is given. Optimal order estimates are obtained for the approximation of the method.

Introduction

Sobolev equation appears in many areas of physics. The study about sobolev equations has important significance for solving physical engineering problems. With the sobolev equations we can analyze the flow of fluids through fissured rock, and obtain the process of the migration of the moisture in soil and solve thermodynamics problems, etc. There have been many numerical modeling methods for sobolev equations, such as difference methods and finite element methods. A mixed finite element method was established and the error estimate was carried out in [1], and a characteristics mixed finite element method was given in [2].

The collocation methods that meet constraint condition of interpolation have been developed since the 1970s. The methods satisfy the differential equation and the boundary conditions at collocation points by the piecewise polynomial approximation. Because the collocation method is easy to get approximation equations, does not need to compute numerical integration and has high-order accuracy, it is widely used in mathematical physics and engineering problems. For example, the collocation method in quasilinear parabolic equations was used in [3]. The collocation method in heat conduction equations was considered in [4].

Consider the linear sobolev equation with constant coefficients

\begin{align}
\begin{cases}
\frac{\partial u}{\partial t} = \nabla (a \nabla u) + b \nabla u + f(x,y), & (x,y) \in I, t \in (0,T), \\
|u|_0 = 0, & x \in \partial I, t \in [0,T], \\
|u|_{x=0} = u_0(x,y), & (x,y) \in I,
\end{cases}
\end{align}

where $u_t$ denotes the time derivative of the function $u$, $\nabla u$ denotes the gradient of the function $u$. And $a$, $b$ are known constants, $I=[0,1] \times [0,1]$, $\partial I$ is the boundary of $I$.

The Orthogonal Collocation Method

The domain $I$ is dissected into equally spaced rectangular grid by grid points $(x_i, y_j)$, $i=0,1,\ldots,M$; $j=0,1,\ldots,N$ and $0=x_0<x_1<\ldots<x_M<1$; $0=y_0<y_1<\ldots<y_N<1$. Introduce the following notations:

$I_{ij} = [x_{i-1}, x_i] \times [y_{j-1}, y_j]$, 
$h_x = x_i - x_{i-1}$, $h_y = y_j - y_{j-1}$, $h = \max\{h_x, h_y\}$, 
$H_3 = \{v = v(x,y) \in C^2(I) | v|_{h} \text{ is a bi-cubic polynomial, } i=1,\ldots,M, j=1,2,\ldots,N\}$, 
$H_3^0 = \{v = v(x) \in H_3 | v|_h = 0\}$.

We take four Gauss points as collocation points in $I_{ij}$:

$x_{ik} = x_{i-1} + h_x \theta_k$, $k = 1,2$, $y_{jk} = y_{j-1} + h_y \theta_k$, $k = 1,2,1 = 2$, 
where

\[ \theta_1 = (3 - \sqrt{3})/6, \theta_2 = (3 + \sqrt{3})/6. \]
Then we introduce an intermediate variable \( q = a \nabla u + b \nabla u \). So we can building our orthogonal collocation method as follows:

Seeking for \((U, Q) : [0, T] \rightarrow H_3 \times H_3\), such that

\[
\begin{align*}
\{U_i - \nabla Q - f\}(x_{a_i}, y_{b_i}, t) &= 0; \\
\{Q - a \nabla U_i - b \nabla U\}(x_{a_i}, y_{b_i}, t) &= 0; i = 1, \ldots, M; j = 1, \ldots, N; t \in (0, T); k, l = 1, 2; \\
U \big|_{t=0} &= 0, t \in [0, T], \\
U \big|_{t=0} &= -u_0(x) \text{ is sufficiently small.}
\end{align*}
\]  

\( (2) \)

A Weighted Galerkin Formulation

We shall introduce some notation defined on \( I \) and \( I_{ij} \) as follows:

\[
\langle u, v \rangle = \sum_{i=1}^{M} \sum_{j=1}^{N} \langle u, v \rangle_{ij} = \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{h_i h_j}{4} \sum_{k,l=1}^{2} u(x_{ik}, y_{jl}),
\]

\[
\langle u, v \rangle_x = \sum_{i=1}^{M} \langle u, v \rangle_{x_i} = \sum_{i=1}^{M} \sum_{k,l=1}^{2} u(x_{ik}, y, t),
\]

\[
\langle u, v \rangle_y = \sum_{j=1}^{N} \langle u, v \rangle_{y_j} = \sum_{j=1}^{N} \sum_{k,l=1}^{2} u(x, y_{jl}, t),
\]

\[ \|u\|^2 = \langle u, u \rangle, \quad \|u\|^2 = \langle u, u \rangle. \]

Let \( U \) be the solution of (2). We can get the semi-discrete Galerkin method:

\[
\begin{align*}
\langle U_i, z_1 \rangle + \langle Q, \nabla z_1 \rangle &= \langle f, z_1 \rangle, \\
\langle Q, z_2 \rangle - \langle a \nabla U_i + b \nabla U, z_2 \rangle &= 0.
\end{align*}
\]

\( (3) \)

From the definition of \( \langle \cdot, \cdot \rangle \), the solutions of (2) and (3) are equivalent.

Let \( \{\zeta_l : l=1,2,\ldots,4MN\} = \{(x_{ik}, y_{jl})\}, i=1,2,\ldots, M, j=1,2,\ldots, N, k, l=1,2\} \) and \( \{Z_i\}_{4MN} \) is a group base of \( H_3^0 \). So any \( U(x, y) \) in \( H_3^0 \) can expressed as follows

\[
U(x, y) = \sum_{i=1}^{4MN} \beta_i(t) Z_i(x, y).
\]

Then relations given by (2) and (3) can now be written as

\[
F \beta + G \beta = R, C \beta + A \beta = S,
\]

where

\[
\begin{align*}
F &= (F_{ij})_{4MN \times 4MN}, F_{ij} = Z_j(\zeta_i) - a \Delta Z_j(\zeta_i); \\
G &= (G_{ij})_{4MN \times 4MN}, G_{ij} = -b \Delta Z_j(\zeta_i); \\
C &= (C_{ij})_{4MN \times 4MN}, C_{ij} = \langle Z_j - a \Delta Z_j, Z_i \rangle; \\
A &= (A_{ij})_{4MN \times 4MN}, A_{ij} = -b \langle \Delta Z_j, Z_i \rangle; \\
R &= (R_i)_{4MN}, R_i = f(\zeta_i); S = (S_i), S_i = \langle f, Z_i \rangle.
\end{align*}
\]

We notice that the \( \tau \) satisfied \( F \tau = 0 \) would also be the solution of \( C \tau = 0 \). If \( C \) is nonsingular, \( F \) is also surely nonsingular. To achieve our conclusion, we just need that \( C \) is nonsingular. So we can prove it by counterevidence as follows:

Assume that \( C \) is singular. So we can find a 4MN-dimensional nonzero vector \( \tau \) which makes \( C \tau = 0 \).

Let

\[
M(x, y) = \sum_{i=1}^{4MN} \tau_i Z_i(x, y), \quad \tau = (\tau_i)_{4MN},
\]

1502
we can get \( \langle M, M \rangle = 0 \) from \( C \tau = 0 \). Furthermore we can get \( M(\xi) = 0, l = 1, 2, \ldots, 4MN \). From lemma 1.2 of [4], there is \( M(x, y) = 0 \). So \( \tau = 0 \) and it conflicts to the assumption. Then \( C \) is nonsingular and the solution of (2) and (3) are equivalent.

**Convergence Analysis**

Then we can define a map \((W, V): [0, T] \rightarrow H_3 \times H_3\) such that

\[
\langle \nabla (q - V), z \rangle = 0, \forall z \in H_3,
\]

\[
\langle a \nabla (u_i - W) + b \nabla (u - W), \nabla z \rangle = 0, \forall z \in H_3,
\]

\[
\langle u_0 - W(0), z \rangle, \forall z \in H_3.
\]

We can get the following convergence conclusion for the above numerical Scheme.

**Theorem:** When \( u \in L^\infty(0;\mathbb{R}^3(I)) \cap L^2(0, T; H^1(I)), u_i \in L^2(0, T; H^1(I)), \) then there exists a constant \( K \) such that, for \( h \) sufficiently small

\[
\| u - U \| \leq Kh^4, \quad \| q - Q \| \leq Kh^4.
\]

**Proof:** Let \( v = W - U, \eta = u - W, w = V - Q, \xi = q - V, \) so \( u = V + \eta, q = V + \xi \). The following conclusion in [4] will be needed:

\[
\| \eta \| \leq c h^4 \sum_{i,j} \| u^{(4)} \|^{2/3}_i, \| \eta \| \leq c h^4 \sum_{i,j} \| u^{(4)} \|^{2/3}_i,
\]

\[
\| \xi \| \leq c h^4 \sum_{i,j} \| q^{(4)} \|^{2/3}_i, \| \xi \| \leq c h^4 \sum_{i,j} \| q^{(4)} \|^{2/3}_i,
\]

where \( c \) is a positive constant and \( u \) is sufficiently smooth. So we can get the error equations:

\[
\begin{align*}
\langle \eta, z_1 \rangle + \langle v, z_2 \rangle + \langle \xi, \nabla z_1 \rangle + \langle w, \nabla z_2 \rangle &= 0, \\
\langle \xi, z_2 \rangle + \langle w, z_2 \rangle - \langle a \nabla v, + b \nabla v, z_2 \rangle &= 0.
\end{align*}
\]

(5)

Firstly, the choice \( z_1 = v, z_2 = \nabla v \) leads to the equations

\[
\begin{align*}
\langle \eta, v \rangle + \langle v, v \rangle + \langle \xi, \nabla v \rangle + \langle w, \nabla v \rangle &= 0, \\
\langle \xi, \nabla v \rangle + \langle w, \nabla v \rangle - \langle a \nabla v, + b \nabla v, \nabla v \rangle &= 0.
\end{align*}
\]

The first equation in (5) minus the second one. We can see

\[
\langle a \nabla v, + \nabla v \rangle + \langle v, v \rangle = - \langle b \nabla v, \nabla v \rangle - \langle \eta, v \rangle.
\]

Add Cauchy inequality and the equation

\[
\langle \nabla v, \nabla v \rangle = \frac{1}{2} \frac{d}{dt} \| \nabla v \|^2,
\]

we can obtain

\[
\frac{a}{2} \frac{d}{dt} \| \nabla v \|^2 + \frac{1}{2} \frac{d}{dt} \| v \|^2 = a \langle \nabla v, + \nabla v \rangle + \langle v, v \rangle + \frac{1}{2} \langle v, v \rangle,
\]

\[
\frac{d}{dt} \| \nabla v \|^2 + \frac{d}{dt} \| v \|^2 \leq K(\| \nabla v \|^2 + \| \eta \|^2 + \| v \|^2).
\]

where \( K \) is a positive constant. Integrating the above equation, we get

\[
\| \nabla v \|^2 + \| v \|^2 \leq K \int_0^T (\| \nabla v \|^2 + \| \eta \|^2 + \| v \|^2) ds.
\]

1503
It follows from the Gronwall lemma that
\[ \| \nabla v \|^2 + \| v \|^2 \leq K \int_0^T (\| \eta \|_t^2) \, ds. \] (6)

Secondly the choice \( z_1 = v_t, \) \( z_2 = \nabla v_t \) leads to the equations
\[
\begin{align*}
\langle \eta_t, v_t \rangle + \langle v_t, v_t \rangle + \langle \xi, \nabla v_t \rangle + \langle w, \nabla v_t \rangle &= 0, \\
\langle \xi, \nabla v_t \rangle + \langle w, \nabla v_t \rangle - \langle a \nabla v_t + b \nabla v, \nabla v_t \rangle &= 0.
\end{align*}
\] (7)

The first equation in (7) minus the second one. We can see:
\[ \langle a \nabla v_t, \nabla v_t \rangle + \langle v_t, v_t \rangle = - \langle b \nabla v_t, \nabla v_t \rangle - \langle \eta_t, v_t \rangle. \]

Add Cauchy inequality to obtain
\[ \| \nabla v_t \|_t^2 + \| v_t \|_t^2 \leq K (\| \eta_t \|_t^2 + \| \nabla v \|_t^2) \]
\[ \leq K (\| \eta_t \|_t^2 + \int_0^T \| \eta_t \|_t^2 \, ds). \] (8)

Then the choice \( z_2 = w \) leads to the equation
\[ \langle \xi, w \rangle + \langle w, w \rangle = \langle a \nabla v_t + b \nabla v, w \rangle. \]

It follows that
\[ \| w \|_t^2 = \langle a \nabla v_t + b \nabla v, w \rangle - \langle \xi, w \rangle \leq K (\| \xi \|_t^2 + \| \nabla v_t \|_t^2 + \| \nabla v \|_t^2) \]
\[ \leq K (\| \xi \|_t^2 + \| \eta_t \|_t^2 + \int_0^T \| \eta_t \|_t^2 \, ds). \] (9)

Then (2,2) of [5] implies that \( \forall f \in H_x, \| f \|_2 \leq K \| f \|. \) So combine with (4), (6), (8) and (9), so we can obtain
\[ \| u - U \|_2 \leq \| \eta \|_t^2 + \| v \|_2 \leq \| \eta \|_t^2 + \| \nabla v \|_t^2 + \int_0^T \| \eta_t \|_t^2 \, ds \leq K_h^8, \]
\[ \| q - Q \|_2 \leq \| \xi \|_t^2 + \| w \|_2 \leq \| \xi \|_t^2 + \| \nabla w \|_t^2 + \int_0^T \| \eta_t \|_t^2 \, ds \leq K_h^8, \]
for \( h \) sufficiently small, and the proof is complete.

Acknowledgements

We thank the fund "Basic Subjects Fund of China University of Petroleum (Beijing) (2462015YQ0604)".

References