A SEMILINEAR FINITE ELEMENT METHOD*

SUN JIA-CHANG (孙家昶)
(*Computing Center, Academia Sinica, Beijing, China*)

Abstract

In the Ritz-Galerkin method the linear subspace of the trial solution is extended to a closed subset. Some results, such as orthogonalization and minimum property of the error function, are obtained. A second order scheme is developed for solving a linear singular perturbation elliptic problem and error estimates are given for a uniform mesh size. Numerical results for linear and semilinear singular perturbation problems are included.

§ 1. Introduction

The development of finite element methods has been successful in various fields. From a mathematical point of view, the methods are an extension of the Rayleigh-Ritz-Galerkin technique ([1], [11]—[13]). Usual finite element schemes, choosing piecewise polynomials as trial functions, are very efficient when there are no steep gradients in the true solution. Otherwise, poor results might occur. In order to get accurate numerical data, one may use the adaptive mesh technique or a higher precision scheme such as h-version or p-version[8]. Besides usual polynomials, rational elements[17] and exponential elements[6] have been introduced to enrich the trial subspace to reduce a number of parameters for a given precision. One thing in common among these techniques is that they are all reduced to a discrete linear system if the original differential equation is linear.

This paper proposes finite element methods of Ritz and Galerkin types for linear elliptic equations where the shape functions depend nonlinearly on a finite set of parameters. So the arising minimization problem is solved on a subset instead of a linear subspace (as it would be the case for piecewise linear shape functions). This approach allows for instance the use of exponential shape functions with the parameters occurring in the exponent. So in this case one probably obtains a significantly better approximation which justifies the additional labour.

In Sections 2 and 3, we generalize respectively the Ritz and the Galerkin methods from linear trial subspaces to subsets, and derive some results such as orthogonalization and error estimates. In Section 4, the semilinear finite element technique is applied to solve singular perturbation problems in one dimension: $-\varepsilon u''+pu'+qu=f,\ u(0)=u(1)=0$, which has been studied by various authors[4-9]. Our analysis shows an improvement over the scheme of using the piecewise linear subspace by one higher order of precision. Moreover, the constraint of mesh size h is relaxed from $O(\varepsilon^2)$ to $O(\varepsilon)$. The numerical tests including a linear and a

* Received July 1, 1983.

semilinear test singular perturbation problems are given in Section 5. Computational results show good agreement with the above theoretical analysis.

Some research results on the same topic in two-dimensions will be reported separately[16].

§ 2. A Ritz Method on Subsets

First we consider a self-adjoint elliptic linear differential equation

$$Lu=f. \tag{1}$$

Suppose $a(u, v)=(Lu, v)$ is a positive quadratic form in a real Hilbert space H with an inner product $(*, *)$ and a norm $\|*\|$:

$$C_2\|u\|^2\leqslant a(u, u)\leqslant C_1\|u\|^2 \quad \text{for all } u\in H, \tag{2}$$

where C_1 and C_2 are positive constants. u is defined as a weak solution of (1) if it satisfies

$$a(u, v)=(f, v) \quad \text{for all } v\in H. \tag{3}$$

It is well-known that u is a weak solution of (1) if and only if it is the unique minimum solution of a quadratic functional I, i.e.,

$$I(u)=\inf_{v\in H} I(v)=\inf_{v\in H}\{a(v, v)-2(f, v)\}. \tag{4}$$

In dealing with the variational problem (4), a well-known discretization is used to replace the space H with a sequence of finite-dimensional subspaces V^h contained in H such that

$$I(u^h)=\inf_{v\in V^h} I(v),$$

which is equivalent to the following weak solution

$$a(u^h, v^h)=(f, v^h) \quad \text{for all } v^h\in V^h. \tag{5}$$

Now we try to replace H in (4) with a sequence of closed subsets S^h with the same number of finite-dimensional parameters. Let T be a one-to-one differentiable map[10] from an open convex set V_1^h of V^h onto S^h: $TV_1^h=S^h$. In particular, $T'=T$ if T is a linear map.

Consider a restricted variational problem on the closed subset S^h:

$$I(u_s)=\inf_{v\in S^h} I(v). \tag{6}$$

Since S^h is closed, there exists a solution of (6) in S^h. If u_s minimizes I over S^h, $u_s=Tw$, then for any $\eta\in V_1^h$ and small $\alpha, I(u_s)\leqslant I(T(w+\alpha\eta))$ as $w+\alpha\eta\in V_1^h$. Let $T(w+\alpha\eta)=Tw+\alpha T\eta+\varkappa(\alpha)$, where T is positively homogeneous and

$$\varkappa(\alpha)=T(w+\alpha\eta)-Tw-\alpha T\eta.$$

We see that

$$\begin{aligned}I(T(w+\alpha\eta))=&I(u_s)+2\alpha[a(u_s, T\eta)-(f, T\eta)]+2[a(u_s, \varkappa(\alpha))-(f, \varkappa(\alpha))]\\&+\alpha^2a(T\eta, T\eta)+2\alpha a(T\eta, \varkappa(\alpha))+a(\varkappa(\alpha), \varkappa(\alpha))\equiv I(\alpha).\end{aligned}$$

For u_s to minimize I over S^h, it requires that $\lim\limits_{a\to 0} I'(\alpha)=0$. Observing that $\varkappa(0)=0$, $\varkappa'(0)=(T'(T^{-1}u_s)-T)\eta$, and

$$0=I'(\alpha)|_{\alpha\to 0}=2\{a(u_s, T\eta)-(f, T\eta)+a(u_s, \varkappa'(0))-(f, \varkappa'(0))\},$$

we have

$$a(u_s, T'(T^{-1}u_s)\eta)=(f, T'(T^{-1}u_s)\eta)\quad \text{for all } \eta\in V_1^h. \tag{7}$$

Therefore, we have the following theorem:

Theorem 1. *If* (i) V^h *is a subspace of* H, (ii) S^h *is a closed subset of* H, (iii) T *is a one-to-one positively homogeneous and differentiable map from an open convex set* V_1^h *of* V^h *onto* S^h: $TV_1^h=S^h$, *then there exists a solution* u_s *of* (6) *so that* (7) *holds.*

The above theorem shows that the nonlinear system (7) has at least one solution which minimizes the variational problem (6). Usually, it does not mean the equivalence of (6) and (7). In general, the uniqueness of the solution cannot be ensured. However, we have the following conclusion.

Theorem 2. *If* V_h^1 *contains* u^h *which is defined by* (5), *then for the map* T *which is sufficiently close to a linear map, i.e.,* $\|T-T'\|$ *is sufficiently small in the sense that for a fixed* $\varepsilon>0, \|(I-T'(T^{-1}u))v\|<\varepsilon$, $\forall u, v\in V_1^h$, *the nonlinear system* (7) *has a unique solution which minimizes the variational problem* (6).

Proof. In fact, (7) can be rewritten as

$$a(u_s, v^h)=(f, v^h)+Q(u_s, v^h),$$

where $$Q(u_s, v^h)=a(u_s, [I-T'(T^{-1}u_s)]v^h)-(f, [I-T'(T^{-1}u_s)]v^h).$$

Since there exists a unique solution in (5), the above system of equations must also have a unique solution if $\|T-T'\|$ is sufficiently small.

Now we suppose that the generalized coordinates (real parameters) of the subset S^h are $q_1, \cdots, q_n$. Then for a minimum solution in S^h the first-variational equations of $I(w)$ must be eliminated

$$\frac{1}{2}\frac{\partial I}{\partial q_i}=a\left(w, \frac{\partial w}{\partial q_i}\right)-\left(f, \frac{\partial w}{\partial q_i}\right)=0\quad \text{for } i=1,\cdots, n. \tag{8}$$

The determinant of the second-variational matrix at the solution point is positive

$$\det\left(\frac{\partial^2 I}{\partial q_i\partial q_j}\right)>0. \tag{9}$$

Let $\{B_j\}$ be a basis. Then for each $w\in S^h$,

$$w=T^{-1}w+w^*,\quad \frac{\partial w}{\partial q_i}=B_i+\frac{\partial w^*}{\partial q_i},$$

where $$T^{-1}w=\sum_j q_jB_j,\quad w^*=w-T^{-1}w.$$

Substituting the above formulas into (8) yields

$$\sum_j a(B_i, B_j)q_j=(f, B_i)+G_i(q),$$

where $$G_i=\left(f, \frac{\partial w^*}{\partial q_i}\right)-a\left(w^*, \frac{\partial w}{\partial q_i}\right)-\sum_j q_j a\left(B_j, \frac{\partial w^*}{\partial q_i}\right).$$

Hence, the equations of the weak solution in subsets differ from those in subspaces only by the last extra term which tends to zero when the subset S^h tends to a subspace, i.e. for a fined $\varepsilon>0$, there exists $h_0>0$, for all $h\leqslant h_0$, such that $\|u^h-v^h\|<\varepsilon$, $\forall u^h\in S^h$, $v^h\in V^h$. Also, system (8) can be written as

$$a(w, B_i)=(f, B_i)+G_i^*(q), \tag{10}$$

where
$$G_i^* = \left(f, \frac{\partial w^*}{\partial q_i}\right) - a\left(w, \frac{\partial w^*}{\partial q_i}\right).$$

Hence, for each $v \in V^h$, ignoring the extra terms, we get an approximate equation
$$a(u_s, v) = (f, v) \text{ for all } v \in V^h. \tag{11}$$

The w in (10) corresponds to the unique solution of the variational problem (6) for the positive quadratic form $a(u, u)$ restricted in the subset S^h. Owing to the continuity of solutions with the system, a solution u_s of system (11) in S^h still exists, provided the distance between V^h and S^h, i.e. $\sup\limits_{\substack{u^s \in S^h \\ v^h \in V^h}} \|u^s - v^h\|$, is sufficiently small. Geometrically, it is obvious. Expressions (8) and (9) imply that a hypersurface in n dimensions $(q_1, \cdots, q_n)$, $z = (\partial I / \partial q_i)$, is separated by a hyperplane $z = 0$ and they have only one intersection point. Even if this hypersurface is moved, there still exists a unique intersection point if the distance of moving is sufficiently small.

There is another different approximation version from (11) which requires the finding of a $u_s \in S^h$ such that
$$a(u_s, u_s - v^h) = (f, u_s - v^h) \quad \text{for all } v^h \in V^h. \tag{12}$$

Suppose u_s is the unique solution of (12). From (3), for any v^h in V^h, $a(u, u_s - v^h) = (f, u_s - v^h)$. Subtracting (12) from the above formula leads to $a(u - u_s, u_s - v^h) = 0$. It also implies that
$$a(u - v^h, u - v^h) = a(u - u_s, u - u_s) + a(v^h - u_s, v^h - u_s).$$

Using (2), for any v^h in V^h, we find
$$C_2 \|u - u_s\|^2 \leqslant a(u - u_s, u - u_s) \leqslant a(u - v^h, u - v^h) \leqslant C_1 \|u - v^h\|^2.$$

Similar formulas exist for (11). Thus, we have proved the following fundamental theorem of the Ritz method on subsets which is an extension of Theorem 1.1 in [13] for subspaces.

Theorem 3. *Suppose u_s is the unique solution of* (12) *or* (11) *in a closed subset S^h. Then it satisfies the following properties:*

(a) *Minimization*
$$a(u - u_s, u - u_s) = \inf_{v^h \in V^h} a(u - v^h, u - v^h),$$

or
$$a(u - u_s, u - u_s) = \inf_{v^h \in V^h} a(u - u_s - v^h, u - u_s - v^h),$$

and
$$\|u - u_s\| \leqslant C \inf_{v^h \in V^h} \|u - v^h\|, \tag{13}$$

or
$$\|u - u_s\| \leqslant C \inf_{v^h \in V^h} \|u - u_s - v^h\|, \tag{14}$$

where C is a constant.

(b) *Orthogonalization*
$$a(u - u_s, u_s - v^h) = 0 \quad \text{for all } v^h \text{ in } V^h,$$

or
$$a(u - u_s, v^h) = 0 \quad \text{for all } v^h \text{ in } V^h. \tag{15}$$

As a system for the weak solution, (11) is more practical than (12). And the difference between them is small if the subset is "not far" from a subspace in some sense.

§ 3. A Galerkin Method on a Closed Nonlinear Subset

Now we extend the analysis of the Ritz method to the Galerkin method. Assume that the operator L in (1) is not self-adjoint in which derivatives of odd order spoil the self-adjointness of an elliptic equation and the associated quadratic functional $I(v)$ defined in (4) is not positive definite. The problem now is to find a stationary point instead of a minimum of $I(v)$.

Theorem 4[1]. *Suppose that H_1 and H_2 are two real Hilbert spaces with inner products $(*, *)_{H_1}$ and $(*, *)_{H_2}$, respectively, and that (f, v) is a continuous linear functional on H_2 and $a(u, v)=(Lu, v)_{H_2}$ a bilinear form with two inequalities*

(i) $|a(u, v)|\leqslant C_1\|u\|_{H_1}\|v\|_{H_2}$ *for all $u\in H_1$ and $v\in H_2$,*

where $C_1<\infty$.

(ii) $|a(u, u)|\geqslant C_2\|u\|^2_{H_1}$, $C_2>0$ *for all $u\in H_1$.*

Then there exists one and only one weak solution u_0 of the functional equation $Lu=f$ such that

$$a(u_0, v)=(f, v) \quad \text{for all } v\in H_2. \tag{16}$$

Galerkin's method is a natural discretization of weak form. In general, it involves two families of functions: a subspace S^h of the solution space (or trial space) H_1 and a subspace V^h of the test space H_2. Then the Galerkin solution u^h satisfying (11) is an element of S^h. Let $\{s_j\}$ be a basis for S^h and $\{v_j\}$ a basis for V^h. The solution $u^h=\sum_j q_j s_j$ satisfies a linear system

$$Aq=d, \tag{17}$$

where $A=(a(s_i, v_j))$ and $d=(f, v_j)$. If A^{-1} exists, there is a unique solution u^h of (11). However, if there is an odd-derivative term of bilinear form with significant size, the usual Galerkin method is unsatisfactory in general.

Suppose that $S^h\in H_1$ is a closed subset with the same number of freedoms as V^h and that there exists an element $u^h\in S^h$ satisfying (11). Following a similar derivation in Section 2, we reach the following conclusion parallel to Theorem 3:

Theorem 5. *Assume that all conditions in Theorem 4 hold and let u^h be a solution of (11) in a closed subset S^h. Then (15) is still true. Moreover, (14) can be expressed by*

$$\|u-u^h\|_{H_1}\leqslant\frac{C_1}{C_2}\inf_{w\in V^h}\|u-u^h-w\|_{H_1}. \tag{18}$$

Let u_I denote an interpolation of any $u\in H_1$ in the subspace V^h. For any $u_J\in S^h$,

$$a(u-u^h, u-u^h)=a(u-u^h, u-u_J)+a(u-u^h, u_J-u^h).$$

By (15) we have

$$a(u-u^h, u_J-u^h)=a(u-u^h, (u_J-u^h)-(u_J-u^h)_I),$$

or

$$a(u-u^h, u_J-u^h)=a(u-u^h, (u-u^h)-(u-u^h)_I)-a(u-u^h, (u-u_J)-(u-u_J)_I).$$

An application of the inequalities of Theorem 4 gives

$$C_2\|u-u^h\|^2_{H_1}\leqslant C_1\|u-u^h\|_{H_1}\{\|u-u_J\|_{H_2}+\|(u_J-u^h)-(u_J-u^h)_I\|_{H_2}\}.$$

Therefore, we have shown the following error estimate.

Corollary 6. Let u_I denote an interpolation of any $u\in H_1$ in the subspace V^h. Then

$$\|u-u^h\|_{H_1}\leqslant\frac{C_1}{C_2}\{\|(u-u^h)-(u-u^h)_I\|_{H_2}+\inf_{u_J\in S^h}[\|u-u_J\|_{H_2}$$
$$+\|(u-u_J)-(u-u_J)_I\|_{H_2}\},\tag{19}$$

and

$$\|u-u^h\|_{H_1}\leqslant\frac{C_1}{C_2}\{\|u-u_{Jh}\|_{H_2}+\|(u-u^h)-(u-u^h)_I\|_{H_2}$$
$$+\|(u-u_{Jh})-(u-u_{Jh})_I\|_{H_2}\}.\tag{20}$$

The bounds (19)—(20) will play a central role in error analysis. It is clear that the subset S^h may be so chosen as to tend to a denumerable dense set as h tends to zero in the true solution space H_1, as V^h does in H_2. In this case, the limiting behaviors of the error in energy norm depends mainly on the approximation of the subset S^h as $h\to 0$.

For a non self-adjoint $a(u, v)$, the existence of the stationary point in the whole space H_1 is ensured by Theorem 4. Hence, from geometric intuition, there exists at least one stationary point in the sense of (11) for sufficiently small h. When the subspace S^h coincides with the subspace V^h, we assume that there exists a unique stationary point of (11). Hence, the unique stationary point still exists provided the subset S^h is "very close" to the subspace V^h. In general, we have the following theorem:

Theorem 7. *Suppose there exists a subspace SL^h with a basis $\{s_j\}$ in which the linear system* (11) *has a unique solution. Let T be a map from the subset S^h to the subspace SL^h such that for a basis $\{v_j\}$ of the test subspace V^h,*

$$\rho(A^{-1}J(G))<1,\tag{21}$$

where the notation ρ denotes the spectral radius of a matrix, A is defined in (17), *and $J(G)$ is a Jacobi matrix of the vector G, defined by*

$$G=(a(u^h-Tu^h,\ v_j)).$$

Then, there also exists a unique solution of the nonlinear system (11) *on the subset S^h.*

Proof. Let $Tu^h=\sum_j q_j s_j$. Since $a(u^h,\ v_j)=a(Tu^h,\ v_j)+a(u^h-Tu^h,\ v_j)$, from the orthogonalization property of (15), (11) becomes $\sum_j a(s_i,\ v_j)=(f,\ v_j)+G_j$, it can be written in matrix form as

$$Aq=d+G(q).\tag{22}$$

The system of equations can be solved by a "simple" iterative procedure

$$Aq^{(0)}=d,$$
$$Aq^{(k)}=d+G(q^{(k-1)}),\tag{23}$$

which is a contraction map if condition (21) is satisfied. Q.E.D.

Remark. (22) is very useful not only for the proof of the existence, but also for the computation of the solution.

§ 4. An Application to a Singular Perturbation Boundary Value Problem

Consider the following boundary value problem

$$Lu = -\varepsilon u'' + p(x)u' + q(x)u = f(x), \qquad u(0) = u(1) = 0, \tag{24}$$

where ε is a small positive parameter and $p(x)$, $q(x)$ and $f(x)$ are so smooth that their derivatives including the second-order one are uniformly bounded for all x in $[0, 1]$. In addition, $p(x) \geqslant p^* > 0$, $q(x) \geqslant \max(0, p'(x))$ on $[0, 1]$.

Let H_m be a Sobolev space of order m with the norm such that

$$\|u\|_m = \left\{\int_0^1 \sum_{i \leqslant m} (D^i u)^2 dx\right\}^{1/2}$$

and $a(u, v)$ be the non-symmetric bilinear form

$$a(u, v) = \int_0^1 \{\varepsilon u'v' + pu'v + quv\} dx. \tag{25}$$

With these notations the weak solution of (24) can be stated as: Find $u \in H_1^0[0, 1]$ such that

$$a(u, v) = (f, v) \quad \text{for all } v \in H_1^0[0, 1], \tag{26}$$

where

$$H_1^0[0, 1] = \{v \mid v \in H_1[0, 1] \text{ and } v(0) = v(1) = 0\}.$$

Existence and uniqueness of the solutions to (26) follow from Theorem 4 by the following lemma:

Lemma 8[9]. *There exists a positive constant C independent of ε such that*

$$|a(u, v)| \leqslant C\|u\|_{1,\varepsilon}\|v\|_1 \quad \text{for all } u,\ v \in H_1^0,$$

$$|a(u, v)| \leqslant C\|u\|_{1,\varepsilon}\|v\|_{1,\varepsilon,1/\varepsilon} \quad \text{for all } u,\ v \in H_1^0,$$

and

$$|a(u, u)| \geqslant C^{-1}\|u\|_{1,\varepsilon}^2 \quad \text{for all } u \in H_1^0,$$

where

$$\|u\|_{1,\varepsilon} = \left\{\int_0^1 (\varepsilon u'^2 + u^2)\, dx\right\}^{1/2}, \tag{27}$$

$$\|u\|_{1,\varepsilon,1/\varepsilon} = \left\{\int_0^1 \left(\varepsilon u'^2 + \frac{1}{\varepsilon} u^2\right) dx\right\}^{1/2}. \tag{28}$$

Now we apply the generalized Galerkin method described in Section 3 to solve problem (24). Let Δ_h denote a partition of the interval $[0, 1]$ into N subintervals $[x_{j-1}, x_j]$, $j = 1, 2, \cdots, N$, with $x_0 = 0$, $x_N = 1$. For convenience, first we only consider a uniform mesh: $x_j - x_{j-1} = h$, $j = 1, 2, \cdots, N$. Associated with Δ_h we have two subsets with the same freedom in $H_1^0[0, 1]$; one is the usual piecewise linear space P^h, the other is called SP_1^h which is defined by the following: if $u_s^h \in SP_1^h$, then for $x_{j-1} \leqslant x \leqslant x_j$, $t = (x - x_{j-1})/h$,

$$u_s^h(x) = \begin{cases} u_{j-1}(1-t) + u_j t, & \text{if } |u_j - u_{j-1}|/h < dl, \\ (u_{j-1} + c)\{(u_j + c)/(u_{j-1} + c)\}^t - c, & \text{otherwise}, \end{cases} \tag{29}$$

where c is a parameter to be chosen such that it well-defines the formula and makes a better approximation for the special problem, and dl is a control constant.

For a fixed $u(x)$, divide the interval $[0, 1]$ into two subintervals such that $[0, 1]=I_r+I_s$, where I_r is a regular subinterval over which the first derivative of $u(x)$ is bounded by a control number and I_s is a singular subinterval over which $u'(x)$ could be very large (near boundary layer in this problem).

For fixed c and dl, SP_1^h, consisting of all admissible elements of (29), is a nonlinear subset in H_1^0. It differs from the corresponding linear space V^h only where the element has a large first derivative.

In the test function space, we keep $\{v_j\}$ as the "roof" basis:

$$v_j^h(x)=\begin{cases}(x-x_{j-1})/h, & x_{j-1}\leqslant x<x_j, \quad j=1, 2, \cdots, N-1,\\ (x_{j+1}-x)/h, & x_j\leqslant x\leqslant x_{j+1}.\end{cases}$$

For simplicity we first suppose that the coefficients p and q of (24) are constant. In order to integrate (25), we need the following lemma which can be verified by an integration by parts.

Lemma 9. *For $ab>0$,*

$$I_0=\int_0^1 a^{1-t}b^t\,dt=\frac{b-a}{\log(b/a)},$$

$$I_k=\int_0^1 a^{1-t}b^t t^k\,dt=I_0\,\frac{b-kI_{k-1}}{b-a}, \quad k=1, 2, \cdots.$$

In particular

$$I_1=\frac{1}{\log(b/a)}\left\{b-\frac{b-a}{\log(b/a)}\right\}.$$

We have derived some inequalities in [14] about I_0 and I_1 which will be useful for later discussion.

Lemma 10. *Suppose $a, b>0$. Then*

$$(ab)^{1/2}\leqslant I_0\leqslant\frac{a+b}{2},$$

$$\frac{1}{2}(ab)^{1/2}\min(1, a^{-1/4}b^{1/4})\leqslant I_1\leqslant\frac{a+b}{4}\max\left(1, \frac{b+(ab)^{1/2}}{b+a}\right),$$

with equality if and only if $a=b$.

The corresponding integral of linear interpolation to I_k is

$$LI_k=\int_0^1[a(1-t)+bt]\,t^k\,dt=\frac{a+(k+1)b}{(k+1)(k+2)}.$$

Therefore, it is not difficult to verify the following estimates:

$$0\leqslant LI_0-I_0\leqslant\frac{1}{2}(b^{1/2}-a^{1/2})^2,$$

$$\frac{1}{12}(b^{1/2}-a^{1/2})(b^{1/2}-2a^{1/2})\leqslant LI_1-I_1\leqslant\frac{1}{6}(b^{1/2}-a^{1/2})(2b^{1/2}-a^{1/2}), \quad b\geqslant a>0,$$

$$-\frac{1}{12}(a-b)\leqslant LI_1-I_1\leqslant\frac{1}{6}(a^{1/4}-b^{1/4})(a^{3/4}+a^{1/2}b^{1/4}+a^{1/4}b^{1/2}-2b^{3/4}), \quad a\geqslant b>0,$$

$$\sup_{0<a,b<1}|LI_0-I_0|=\frac{1}{2}, \quad \sup_{0<a,b<1}|LI_1-I_1|=\frac{1}{3}. \tag{30}$$

Integrating (25) from x_{j-1} to x_j yields

$$
\begin{aligned}
a(u_s^h, v_{j-}^h) &= \int_0^1 \left(\frac{\varepsilon}{h}+pt\right)(c+u_{j-1})^{1-t}(c+u_j)^t \log\frac{c+u_j}{c+u_{j-1}}\,dt \\
&\quad + hq\int_0^1 \{(c+u_{j-1})^{1-t}(c+u_j)^t - c\}t\,dt \\
&= \frac{\varepsilon}{h}(u_j-u_{j-1}) + p\left\{c+u_j - \frac{u_j-u_{j-1}}{\log((c+u_j)/(c+u_{j-1}))}\right\} \\
&\quad + hq\left\{-\frac{c}{2}+\frac{u_j+c}{\log((c+u_j)/(c+u_{j-1}))}-\frac{u_j-u_{j-1}}{(\log((c+u_{j-1})/(c+u_j)))^2}\right\}.
\end{aligned}
$$

Similarly, integrating (25) from x_j to x_{j+1} leads to

$$
\begin{aligned}
a(u_s^h, v_{j+}^h) &= \int_0^1 \left(-\frac{\varepsilon}{h}+p(1-t)\right)(c+u_j)^{1-t}(c+u_{j+1})^t \log\frac{c+u_{j+1}}{c+u_j}\,dt \\
&\quad + hq\int_0^1 \{(c+u_j)^{1-t}(c+u_{j+1})^t - c\}t\,dt \\
&= \frac{\varepsilon}{h}(u_j-u_{j+1}) - p\left\{(c+u_j) - \frac{u_{j+1}-u_j}{\log((c+u_{j+1})/(c+u_j))}\right\} \\
&\quad + hq_j\left\{-\frac{c}{2}+\frac{u_j+c}{\log((c+u_j)/(c+u_{j+1}))}-\frac{u_j-u_{j+1}}{(\log((c+u_{j+1})/(c+u_j)))^2}\right\}.
\end{aligned}
$$

For $[x_{j-1}, x_j] \in I_r$, a straightforward computation yields

$$
a(u_s^h, v_j^h) = \frac{\varepsilon}{h}[2u_j - u_{j-1} - u_{j+1}] + \frac{1}{2}p(u_{j+1}-u_{j-1}) + \frac{h}{6}q(u_{j+1}+4u_j+u_{j-1}). \tag{31}
$$

For $[x_{j-1}, x_j] \in I_s$,

$$
\begin{aligned}
a(u_s^h, v_j^h) &= a(u_s^h, v_{j-}^h) + a(u_s^h, v_{j+}^h) = \frac{\varepsilon}{h}(2u_j - u_{j-1} - u_{j+1}) \\
&\quad + p\left\{\frac{u_{j+1}-u_j}{\log((c+u_{j+1})/(c+u_j))} - \frac{u_j-u_{j-1}}{\log((c+u_j)/(c+u_{j-1}))}\right\} \\
&\quad + qh\left\{(c+u_j)\left[\frac{1}{\log((c+u_j)/(c+u_{j+1}))}+\frac{1}{\log((c+u_j)/(c+u_{j-1}))}\right]\right. \\
&\quad \left. - c - \frac{u_j-u_{j+1}}{(\log((c+u_{j+1})/(c+u_j)))^2} - \frac{u_j-u_{j-1}}{(\log((c+u_{j-1})/(c+u_j)))^2}\right\}.
\end{aligned} \tag{32}
$$

Comparing with (31), we rewrite the last formula (32) in the following type:

$$
a(u_s^h, v_j^h) = \frac{\varepsilon}{h}(2u_j - u_{j-1} - u_{j+1}) + \frac{1}{2}p(u_{j+1}-u_{j-1}) + \frac{h}{6}q(u_{j+1}+4u_j+u_{j-1}) + g_j, \tag{33}
$$

where g_j is the right-hand side difference between (32) and (31).

Define $\alpha \equiv \frac{\varepsilon}{h}$. Substituting (31) and (33) into the generalized Galerkin method, i.e.,

$$
a(u_s^h, v_j^h) = (f, v_j^h) \quad \text{for } j=1, 2, \cdots, N-1, \tag{34}
$$

gives

$$
L_h U^h = \begin{cases} (f, v_j^h), & \text{if } j \in I_r, \\ (f, v_j^h) - g(U_{j-1}^h, U_j^h, U_{j+1}^h), & \text{if } j \in I_s, \end{cases} \tag{35}
$$

where the left-hand side

$$L_hU^h=-\left(\alpha+\frac{p}{2}-\frac{h}{6}q\right)U^h_{j-1}+\left(2\alpha+\frac{2h}{3}q\right)U^h_j-\left(\alpha-\frac{p}{2}-\frac{h}{6}q\right)U^h_{j-1},$$

which is exactly the same as the scheme from usual piecewise linear subspace.

(35) can be rewritten in a special matrix form as (22)

$$AU=d+Q(U), \tag{36}$$

where $A=(\alpha_{i,j})$ is a tridiagonal matrix and

$$\alpha_{i,j}=\begin{cases}-\left(\alpha+\frac{p}{2}-\frac{h}{6}q\right), & i>j,\\ 2\alpha+\frac{2h}{3}q, & i=j,\\ -\left(\alpha-\frac{p}{2}-\frac{h}{6}q\right), & i<j.\end{cases} \tag{37}$$

Denote the determinants of the first j and the last $N-i$ principal determinants of A by D_j and $D_{i,N-1}$, respectively. Set

$$\beta_n=\frac{D_{n-1}}{D_n},\quad \beta_{j,N-1}=\frac{D_{j+1,N-1}}{D_{j,N-1}}.$$

Using the recursion formula

$$\beta_n=\left\{2\alpha+\frac{2h}{3}q-\left(\alpha-\frac{p}{2}-\frac{h}{6}q\right)\left(\alpha+\frac{p}{2}-\frac{h}{6}q\right)\beta_{n-1}\right\}^{-1},$$

we obtain the following lemma.

Lemma 11. *If* $\alpha=\frac{\varepsilon}{h}\geqslant\frac{p}{2}+\frac{h}{6}q$, *then*

$$\beta_n\leqslant\left\{\alpha+\frac{p}{2}-\frac{h}{6}q\right\}^{-1}\quad \textit{for all } n<N-1,$$

$$\beta_{n,N-1}\leqslant\left\{\alpha+\frac{p}{2}-\frac{h}{6}q\right\}^{-1}\quad \textit{for all } n<N-1.$$

Thus, we find a relationship between the elements of the inverse matrix A^{-1}.

Theorem 12. *When*

$$\alpha=\frac{\varepsilon}{h}\geqslant\frac{p}{2}+\frac{h}{6}q, \tag{38}$$

$A^{-1}=(\alpha^{-1}_{i,j})$ *is non-negative and*

$$\alpha^{-1}_{i,j}\geqslant\alpha^{-1}_{i,j-1},\quad \textit{if } i\geqslant j\quad \textit{or}\quad \alpha^{-1}_{i,j}\leqslant\alpha^{-1}_{i,j-1},\quad \textit{if } i<j. \tag{39}$$

Proof. Making use of (37), we only need to note that

$$\alpha^{-1}_{i,j}=\begin{cases}\left(\alpha+\frac{p}{2}-\frac{h}{6}q\right)^{i-j}D_{j-1}D_{N-1-i}/D_{N-1}, & \text{if } i\geqslant j,\\ \left(\alpha-\frac{p}{2}-\frac{h}{6}q\right)^{i-j}D_{i-1}D_{N-1-j}/D_{N-1}, & \text{if } i\leqslant j.\end{cases} \tag{40}$$

Q.E.D.

When A^{-1} exists, from (36),

$$U=A^{-1}(d+Q(U)). \tag{41}$$

Now we look for an estimate of $\|A^{-1}J(Q(U))\|$, where $J(Q)$ is the Jacobi matrix of Q. Let $v=u+\bar{c}$, where $\bar{c}$ is a constant vector consisting of the constant component c.

Note that $Q_j(v)$ is homogeneous for $j<N-1$. Applying the Euler theorem of homogeneous functions, we have

$$\{J(Q(u))(u+c)\}_j=\{Q(u)\}_j, \quad \text{if } j<N-1. \tag{42}$$

Since the singularity of the exact solution of (24) is only near $x=1$, the width of the boundary layer is less than $k\varepsilon$, where k is a constant. By inequalities (30), (39), (38) and (40), a straightforward computation yields

$$J(Q(u))(u+c)=\{0, \cdots, 0, Q_n, \cdots, Q_{N-2}, Q^*_{N-1}\},$$
$$\{A^{-1}J(Q(u))u\}_i=\sigma_i+hq\tau_i,$$

where

$$-\frac{3p}{4}<\left(\alpha+\frac{p}{2}+\frac{h}{6}q\right)\sigma_i<\frac{p}{2}, \quad -\frac{2}{3}qk\varepsilon<\left(\alpha+\frac{p}{2}-\frac{h}{6}q\right)hq\tau_i<\frac{2}{3}qk\varepsilon.$$

Therefore we have

Theorem 13. *If the mesh size satisfies condition* (38), *i.e.*,

$$h\leqslant\frac{2\varepsilon}{p}\left\{\frac{1}{2}+\left[\frac{1}{4}+\frac{2}{3}\left(\frac{\varepsilon q}{p}\right)^2\right]^{1/2}\right\}^{-1}\approx\frac{2\varepsilon}{p}\left\{1-\frac{2}{3}\left(\frac{\varepsilon q}{p}\right)^2\right\} \tag{43}$$

as well as

$$k\varepsilon\leqslant\frac{3p}{8q} \tag{44}$$

holds, then, the map $A^{-1}Q(u)$ is contractive, and the semilinear system (36) *can be solved by the following convergent "simple" iteration*

$$AU^{(0)}=d,$$
$$AU^{(k)}=d+Q(U^{(k-1)}), \quad k=1, 2, \cdots. \tag{45}$$

Remark. When ε is small, in practice, the mesh condition (43) can be simplified to $h<\frac{2\varepsilon}{p}$.

Now we derive an error estimate. Let u be the true solution of (24). Decompose u in the following way[9]:

$$u(x)=\gamma\{e^{-p(1)(1-x)/\varepsilon}+Z(x)\}, \tag{46}$$

where γ is a constant bounded uniformly for all $0<\varepsilon<1$, and

$$|Z(x)|\leqslant C, \quad |Z'(x)|\leqslant C, \quad |Z''(x)|\leqslant C\left\{1+\frac{1}{\varepsilon}e^{-\beta(1-x)/\varepsilon}\right\},$$

C being a constant independent of ε and $0<\beta<p^*$.

We have shown in [14]

Lemma 14. *Let u be the true solution of* (24), *if h and ε are of the same order, then*

$$\|u^{(j)}\|_\infty=O(h^{-j}), \quad j=1,2, \cdots, .$$

$$\left\|u''-\frac{u'^2}{c+u}\right\|_\infty=O(h^{-1}), \quad \left\|\left\{u''-\frac{u'^2}{c+u}\right\}'\right\|_\infty=O(h^{-2}), \tag{47}$$

where $$c=\lim_{\varepsilon\to0}\lim_{x\to1}\varepsilon u'(x)/p(1)=\gamma.$$

In addition, using error estimates for such a sub-linear positive interpolation function u^h_I of $u(x)$ in SP^h_1, we have shown in [15] that

$$\|u^h_I-u\|_\infty=O(h), \quad \|u'^h_I-u'\|_\infty=O(1). \tag{48}$$

Since the width of the boundary layer is in the same order as ε, if h is also kept in the same order as ε, then,

$$\|u_J^h-u\|_0=O(h^{3/2}),\qquad \|u_J^h-u\|_1=O(h^{1/2}),$$
$$\|u_J^h-u\|_{1,\varepsilon}=O(h),\qquad \|u_J^h-u\|_{1,\varepsilon,1/\varepsilon}=O(h). \tag{49}$$

Let H_1 and H_2 be the Hilbert spaces having respective norms (27) and (28). Applying Lemma 8 and using (20), we obtain

$$\|u-u^h\|_{1,\varepsilon}\leqslant\frac{C_1}{C_2}\{\|u-u_{Jh}\|_{1,\varepsilon,1/\varepsilon}+\|(u-u^h)-(u-u^h)_I\|_{1,\varepsilon,1/\varepsilon}$$
$$+\|(u-u_{Jh})-(u-u_{Jh})_I\|_{1,\varepsilon,1/\varepsilon}\},$$

where the subscript I denotes the interpolation in the test space V^h– a piecewise linear function subspace. On the right-hand side of the above inequality, the first term is dominant. Hence, from (49), we get the main error estimate for scheme (34).

Theorem 15. *If the mesh size satisfies condition* (43), *then,*

$$\|u_s^h-u\|_{1,\varepsilon}\leqslant Ch, \tag{50}$$

where C is a constant which is uniformly bounded for all small ε satisfying (44).

Substituting the true solution u into scheme (35) and applying the Taylor expansion and using equations (24) and (30) yield

$$L_h u_j=(f,\ v_j^h)+O(h^2)+\frac{h^3}{12}\{pu^{(3)}+qu''\}+\cdots.$$

If $j\in I_r$ for $j\in I_s$

$$L_h u_j=(f,\ v_j^h)+O(h^2)-g_j(u_{j-1},\ u_j,\ u_{j+1})+Tr_j(u),$$
$$Tr_j(u)=\frac{h^3}{12}\left\{pu^{(3)}-p\Big(\frac{u'^2}{c+u}\Big)'+2qu''\right\}+\cdots.$$

Noting that the width of the boundary layer in only $k\varepsilon$, using (47) we have

$$\|A^{-1}Tr(x)\|_\infty=O(h). \tag{51}$$

Furthermore

$$\|u_s^h-u\|_\infty=O(h),$$
$$\|u_s'^h-u'\|_\infty=O(1).$$

Similarly,

$$\|u_s^h-u\|_0=O(h^{1.5}),$$
$$\|u_s^h-u\|_1=O(h^{0.5}).$$

Summarizing the above results gives the following theorem of error estimate.

Theorem 16. *For small ε satisfying* (44), *if the mesh size satisfies condition* (43), *then the generalized Galerkin method on the subset* (34) *has one more order of precision than its corresponding scheme of piecewise linear subspace, i.e., there exist constants C_0, C_1, C_∞ and C'_∞ which are uniformly bounded for all small ε such that*

$$\|u_s^h-u\|_0\leqslant C_0h^{1.5},\quad \|u_s^h-u\|_1\leqslant C_1h^{0.5},$$
$$\|u_s^h-u\|_\infty\leqslant C_\infty h,\quad \|u_s'^h-u'\|_\infty\leqslant C'_\infty. \tag{52}$$

For variable coefficients p and q, it can be similarly shown that the above conclusion remains valid for small ε if two additional inequalities are satisfied:

$$h<\frac{2}{\|p\|_\infty}\varepsilon,$$

and

$$\frac{1}{6}(q_{j-1}+4q_j+q_{j+1})\geqslant\frac{1}{2h}(p_{j+1}-p_{j-1}),\quad j=1,2,\cdots;\tag{53}$$

the latter inequality is a discrete form for the elliptic condition of $q(x)\geqslant p'(x)$.

In fact, we only need to point out that, owing to the smoothness of p and q, if we substitute their piecewise linear interpolations into (25), then, (31) becomes

$$\begin{aligned}a(u_\varepsilon^h, v_j^h)=&\frac{\varepsilon}{h}[2u_j-u_{j-1}-u_{j+1}]\\&+\frac{1}{6}[u_{j+1}(2p_j+p_{j+1})+u_j(p_{j-1}-p_{j+1})-u_{j-1}(2p_j+p_{j-1})]\\&+\frac{h}{12}[u_{j+1}(q_j+q_{j+1})\ +u_j(q_{j-1}+6q_j+q_{j+1})+u_{j-1}(q_j+q_{j-1})]+O(h^2).\end{aligned}$$

The associated tridiagonal matrix $A=(\alpha_{i,j})$ in (37) now is

$$\alpha_{i,j}=\begin{cases}-\left[\alpha+\frac{1}{6}(2p_j+p_{j-1})-\frac{h}{12}(q_j+q_{j-1})\right], & i=j+1,\\ 2\alpha-\frac{1}{6}(p_{j+1}-p_{j-1})+\frac{h}{12}(q_{j+1}+6q_j+q_{j-1}), & i=j.\\ -\left[\alpha-\frac{1}{6}(2p_j+p_{j+1})-\frac{h}{12}(q_j+q_{j+1})\right], & i=j-1.\end{cases}$$

The rest derivation is similar to the above, and we omit the details.

§ 5. Numerical Results

In the tables below, we adopt the following notations. Let $N=1/h$, SL represent the subset scheme (34) and L the corresponding linear scheme. Denote the maximum error with sign of the discrete solution by Er(Max) and the node where Er(Max) occurs by x_M. The notations Er(H1, eps), Er(H0) and Er(H1) represent the approximation errors in terms of $H_{1,\varepsilon}$, H_0 and H_1, respectively. The CPU time is expressed in seconds. The Fortran program was run in double precision, on a DEC-System 2060 computer. The iterative error control for (45) is set to 10^{-5} and the constant $dl=2$ in (29).

Example 1. Consider a linear singular perturbation problem with constant coefficients,

$$Lu=-\varepsilon u''+u'+(1+\varepsilon)u=f(x),\quad \text{in } (0,1),$$

$$u(0)=u(1)=0,$$

where $f(x)=(1+\varepsilon)(a-b)x-\varepsilon a-b$, $a=1+e^{-(1+\varepsilon)/\varepsilon}$, $b=1+e^{-1}$, with true solution

$$u(x)=e^{-(1+\varepsilon)(1-x)/\varepsilon}+e^{-x}-a+(a-b)x.$$

The results listed in Tables 1—4 show that:

Table 1-1 SL: $h/\varepsilon=1.5$

N	x_M	Er(Max)	Er(H1, eps)	Er(H0)	Er(H1)	CPU
25	0.920	−0.5837D−02	0.3508D−01	0.1740D−02	0.2146D+00	0.09
50	0.960	−0.6023D−02	0.2079D−01	0.1239D−02	0.1798D+00	0.16
100	0.970	−0.4562D−02	0.1555D−01	0.7205D−03	0.1902D+00	0.50
200	0.985	−0.1810D−02	0.8410D−02	0.2019D−03	0.1456D+00	1.07
400	0.993	−0.1239D−02	0.5495D−02	0.1031D−03	0.1346D+00	2.17
800	0.996	−0.5259D−03	0.3000D−02	0.3223D−04	0.1039D+00	4.47
1600	0.998	−0.3363D−03	0.1823D−02	0.1483D−04	0.8928D−01	8.56

Table 1-2 L: $h/\varepsilon=1.5$

N	x_M	Er(Max)	Er(H1, eps)	Er(H0)	Er(H1)	CPU
25	0.960	−0.8199D−01	0.1216D+00	0.1742D−01	0.7371D+00	0.40
50	0.980	−0.8112D−01	0.1183D+00	0.1223D−01	0.1019D+01	0.06
100	0.990	−0.8070D−01	0.1167D+00	0.8620D−02	0.1425D+01	0.33
200	0.995	−0.8048D−01	0.1159D+00	0.6084D−02	0.2004D+01	0.75
400	0.998	−0.8038D−01	0.1155D+00	0.4299D−02	0.2827D+01	1.53
800	0.999	−0.8033D−01	0.1153D+00	0.3038D−02	0.3992D+01	3.23
1600	0.999	−0.8030D−01	0.1152D+00	0.2148D−02	0.5641D+01	6.54

Table 2 SL: $h/\varepsilon=1.75$

N	x_M	Er(Max)	Er(H1, eps)	Er(H0)	Er(H1)	CPU
25	0.920	−0.1899D−01	0.4066D−01	0.4869D−02	0.2671D+00	0.11
50	0.960	−0.6682D−02	0.2341D−01	0.1254D−02	0.2187D+00	0.25
100	0.980	−0.4199D−02	0.1643D−01	0.6130D−03	0.2171D+00	0.51
200	0.990	−0.1913D−02	0.1043D−01	0.2120D−03	0.1950D+00	1.11
400	0.993	−0.1162D−02	0.6063D−02	0.8863D−04	0.1604D+00	2.48
800	0.996	−0.7638D−03	0.3817D−02	0.4342D−04	0.1428D+00	4.93
1600	0.998	−0.3355D−03	0.2080D−02	0.1374D−04	0.1101D+00	10.03

Table 3-1 SL: $h/\varepsilon=2.0$

N	x_M	Er(Max)	Er(H1, eps)	Er(H0)	Er(H1)	CPU
25	0.920	0.6469D−02	0.2969D−01	0.1423D−02	0.2097D+00	0.18
50	0.960	−0.7583D−02	0.2592D−01	0.1366D−02	0.2588D+00	0.36
100	0.970	0.3605D−02	0.1359D−01	0.3779D−03	0.1921D+00	0.66
200	0.990	−0.1235D−02	0.1032D−01	0.1033D−03	0.2064D+00	1.47
400	0.990	0.1452D−02	0.5153D−02	0.7843D−04	0.1457D+00	2.93
800	0.996	−0.6512D−03	0.4100D−02	0.3331D−04	0.1640D+00	5.81
1600	0.997	0.8136D−03	0.1324D−02	0.3186D−04	0.7485D−01	12.21

Table 3-2 *L*: $h/\varepsilon=2.0$

N	x_M	Er (Max)	Er (H1, eps)	Er (H0)	Er (H1)	CPU
25	0.960	−0.1366D+00	0.1647D+00	0.2753D−01	0.1149D+01	0.03
50	0.980	−0.1360D+00	0.1614D+00	0.1939D−01	0.1602D+01	0.21
100	0.990	−0.1356D+00	0.1597D+00	0.1369D−01	0.2250D+01	0.37
200	0.995	−0.1355D+00	0.1589D+00	0.9668D−02	0.3171D+01	0.77
400	0.998	−0.1354D+00	0.1584D+00	0.6833D−02	0.4477D+01	1.55
800	0.999	−0.1354D+00	0.1582D+00	0.4830D−02	0.6326D+01	3.35
1600	0.999	−0.1354D+00	0.1581D+00	0.3415D−02	0.8942D+01	6.61

Table 4 *SL*: $h/\varepsilon=2.25$

N	x_M	Er (Max)	Er (H1, eps)	Er (H0)	Er (H1)	CPU
25	0.920	0.1618D−01	0.2526D−01	0.3461D−02	0.1877D+00	0.17
50	0.940	0.2635D−01	0.2103D−01	0.5016D−02	0.2167D+00	0.48
100	0.980	−0.4804D−02	0.1985D−01	0.5968D−03	0.2976D+00	0.86
200	0.990	−0.1223D−02	0.1143D−01	0.9819D−04	0.2424D+00	1.46
400	0.985	0.1082D−01	0.7342D−02	0.9483D−03	0.2184D+00	3.86
800	0.988	0.2192D 02	0.4271D−02	0.9035D−04	0.1812D+00	8.98
1600	0.997	0.4134D−02	0.4955D−02	0.2310D−03	0.2970D+00	16.28

Table 5-1 *SL*: $h/\varepsilon=1.5$

N	x_M	Er (Max)	Er (H1, eps)	Er (H0)	Er (H1)	CPU
25	0.920	−0.1930D−01	0.3845D−01	0.4679D−02	0.2337D+00	0.34
50	0.960	−0.5220D−02	0.1843D−01	0.1040D−02	0.1593D+00	0.98
100	0.970	−0.4529D−02	0.1436D−01	0.6564D−03	0.1757D+00	2.42
200	0.985	−0.1742D−02	0.7681D−02	0.1787D−03	0.1330D+00	5.40
400	0.993	−0.1201D−02	0.5141D−02	0.9504D−04	0.1259D+00	11.58
800	0.996	−0.5050D−03	0.2820D−02	0.2971D−04	0.9767D−01	23.48
1600	0.998	−0.3337D−03	0.1738D−02	0.1397D−04	0.8515D−01	47.13

Table 5-2 *L*: $h/\varepsilon=1.5$

N	x_M	Er (Max)	Er (H1, eps)	Er (H0)	Er (H1)	CPU
25	0.960	−0.8244D−01	0.1217D+00	0.1773D−01	0.7377D+00	2.64
50	0.980	−0.8125D−01	0.1184D+00	0.1231D−01	0.1020D+01	9.63
100	0.990	−0.8100D−01	0.1167D+00	0.8685D−02	0.1426D+01	25.26
200	0.995	−0.8059D−01	0.1159D+00	0.6102D−02	0.2005D+01	56.46
400	0.998	−0.8043D−01	0.1155D+00	0.4305D−02	0.2827D+01	11.24
800	0.999	−0.8035D−01	0.1153D+00	0.3041D−02	0.3992D+01	23.49
1600	0.999	−0.8031D−01	0.1152D+00	0.2149D−02	0.5641D+01	48.78

Table 6-1 *SL*: $h/\varepsilon=1.75$

N	x_M	Er (Max)	Er (H1, eps)	Er (H0)	Er (H1)	CPU
25	0.920	−0.1968D−01	0.3901D−01	0.4724D−02	0.2562D+00	2.81
50	0.960	−0.6017D−02	0.2080D−01	0.1041D−02	0.1943D+00	1.14
100	0.980	−0.3879D−02	0.1503D−01	0.5494D−03	0.1986D+00	3.26
200	0.985	−0.3431D−02	0.1136D−01	0.3500D−03	0.2124D+00	7.61
400	0.993	−0.1143D−02	0.5662D−02	0.8080D−04	0.1498D+00	14.67
800	0.996	−0.7534D−03	0.3625D−02	0.4076D−04	0.1356D+00	29.78
1600	0.998	−0.3246D−03	0.1988D−02	0.1290D−04	0.1052D+00	65.90

Table 6-2 *L*: $h/\varepsilon=1.75$

N	x_M	Er (Max)	Er (H1, eps)	Er (H0)	Er (H1)	CPU
25	0.960	−0.1163D+00	0.1453D+00	0.2433D−01	0.9480D+00	2.73
50	0.980	−0.1077D+00	0.1411D+00	0.1569D−01	0.1312D+01	10.29
100	0.990	−0.1104D+00	0.1397D+00	0.1145D−01	0.1842D+01	26.24
200	0.995	−0.1076D+00	0.1387D+00	0.7839D−02	0.2590D+01	56.17
400	0.998	−0.1075D+00	0.1382D+00	0.5538D−02	0.3655D+01	119.92
800	0.999	−0.1073D+00	0.1380D+00	0.3908D−02	0.5163D+01	140.12
1600	0.999	−0.1072D+00	0.1379D+00	0.2760D−02	0.7297D+01	496.64

Table 7-1 *SL*: $h/\varepsilon=2.0$

N	x_M	Er (Max)	Er (H1, eps)	Er (HO)	Er (H1)	CPU
25	0.920	−0.1928D−01	0.3572D−01	0.4404D−02	0.2507D+00	3.42
50	0.960	−0.7551D−02	0.2358D−01	0.1238D−02	0.2354D+00	10.95
100	0.980	−0.3771D−02	0.1592D−01	0.4465D−03	0.2251D+00	25.85
200	0.985	−0.2508D−02	0.8006D−02	0.2353D−03	0.1601D+00	55.86
400	0.993	−0.1181D−02	0.6546D−02	0.8375D−04	0.1851D+00	116.21
800	0.996	−0.7179D−03	0.4017D−02	0.3538D−04	0.1607D+00	236.06

Table 7-2 *L*: $h/\varepsilon=2.0$

N	x_M	Er (Max)	Er (H1, eps)	Er (H0)	Er (H1)	CPU
25	0.960	−0.1421D+00	0.1653D+00	0.2877D−01	0.1151D+01	3.28
50	0.980	−0.1337D+00	0.1612D+00	0.1907D−01	0.1601D+01	11.33
100	0.990	−0.1255D+00	0.1590D+00	0.1264D−01	0.2242D+01	27.03
200	0.995	−0.1369D+00	0.1590D+00	0.9775D−02	0.3173D+01	58.68
400	0.998	−0.1337D+00	0.1583D+00	0.6743D−02	0.4474D+01	124.43
800	0.999	−0.1323D+00	0.1580D+00	0.4715D−02	0.6318D+01	248.85

1. The iteration of (45) converges monotonically if the ratio $h/\varepsilon<2$. The results agree with the theoretical analysis above, and the SL-scheme is more accurate than the L-scheme.

2. When $2\leqslant h/\varepsilon\leqslant 2.25$, the iteration is still convergent but with some oscillation, and the error is getting larger. CPU time costs more, too. If the ratio h/ε increases again, the iteration (45) does not converge.

3. For a required accuracy, the CPU time costs much less using the SL-scheme than the L-scheme. The smaller the ε is, the more advantages the SL-scheme has. For instance, given an admissible maximum error at knots $\leqslant 0.005$, their CPU time ratio are about 0.3:1.1 and 3:15, for $\varepsilon=0.01$ and 0.001, respectively.

Example 2. Consider a semilinear singular perturbation problem

$$Lu=-\varepsilon u''+u'+(1+\varepsilon)u=f(x, u), \quad \text{in } (0, 1),$$

$$u(0)=u(1)=0,$$

where

$$f(x, u)=a-b-(1+\varepsilon)\left\{e^{-x}-u+\frac{c}{u+a-(a-b)x-e^{-x}}\right\},$$

with the same constants a, b, c and the same solution as example 1.

In the semilinear case, the advantage of the SL-scheme over the L-scheme is more obvious than in the linear case. The results by the SL-scheme also agree with Theorem 15 and they are much better than those produced by the L-scheme for a same required accuracy (see Tables 5—7).

Acknowledgement. The author would like to express his gratitude to Professor Martin H. Schultz for his interest in this work and many valuable suggestions.

References

[1] A. Z. Aziz, ed., The Mathematical Foundations of the Finite Element with Applications to Partial Differential Equations, AP, 1972.

[2] I. Babuska, W. G. Szymczak, An Error Analysis for the Element Method Applied to Convection Diffusion Problems, Technical Note BN-962, 1981, March. Institute for Physical Science and Technology, University of Maryland.

[3] I. Babuska, B. A. Szabo, I. N. Katz, The p-version of the Finite Element Method, *SIAM J. Num. Anal.*, 18(1981), 515—545.

[4] J. W. Barrett, K. W. Morton, Optimal Finite Element Solutions to Diffusion-Convection Problems in One Dimension, *Int. J. Num. Math. in Engng.*, 15 (1980), 1457—1474.

[5] J. Christie, A. R. Mitchell, Upwinding of High Order Galerkin Methods in Conduction-Convection Problems, *Int. J. Num. Math. in. Engng.*, 12 (1978), 1764—1771.

[6] P. P. N. DeGroen, P. W. Hemker, Error Bounds for Exponentially Fitted Galerkin Methods Applied to Stiff Two-point Boundary Value Problems (Eds. P. W. Hemker and J. J. H. Miller), Academic Press, New York, 1979.

[7] D. F. Griffiths, J. Lorentz, An Analysis of the Petrov-Galerkin Finite Element Method, *Comp. Math. Appl. Mech. Engng.*, 14 (1978), 39—64.

[8] J. C. Heinrich, P. S. Huyakorn, O. C. Zienkienwicz, A. R. Mitchell, An "Upwind" Finite Element Scheme for Two-Dimensional Convective Transport Equation, *Int. J. Num. Math. in Engng.*, 11 (1977), 131—143.

[9] R. B. Kellcgg, Han Hon-de, The Finite Element Method for a Singular Perturbation Problem Using Enriched Subspaces, Technical Note BN-978, 1981, September. University of Maryland.

[10] F. Scarpini, Some Nonlinear Complementarity Systems Algorithms and Application to Unilateral

Boundary-value Problem, Bologna Nicola Zanchelli Editore, 1980.

[11] M. H. Schultz, Spline Analysis, Prentice-Hall Inc., 1973.

[12] M. H. Schultz, ed., Elliptic Problem Solvers, Academic Press, New York, 1981.

[13] G. Strang, G. J. Fix, An Analysis of the Finite Element Method, Prentice-Hall, Inc., 1973.

[14] Sun Jia-chang, Semi-linear Difference Schemes for Singular Perturbation Problems in One Dimension, Department of Computer Science, Yale University, Technical Report #216, 1982.

[15] Sun Jia-chang, A Galerkin Method on Nonlinear Subsets and Its Application to a Singular Perturbation Problem, Department of Computer Science, Yale University, Technical Report #217, 1982.

[16] Sun Jia-chang, Semi-linear Difference Schemes and Semi-linear Finite Element Methods for Singular Perturbation Problems in Two Dimensions.

[17] E. L. Wachspress, A Rational Finite Element Basis, Academic Press, New York, 1975.