Conference Applications of Mathematics 2012
in honor of the 60th birthday of Michal Křížek.
Institute of Mathematics AS CR, Prague 2012

A CURIOUS PROPERTY OF OSCILLATORY FEM SOLUTIONS OF ONE-DIMENSIONAL CONVECTION-DIFFUSION PROBLEMS

Niall Madden[1], Martin Stynes[2]

[1] School of Mathematics, Statistics and Applied Mathematics
National University of Ireland, Galway, Ireland
niall.madden@nuigalway.ie
[2] Department of Mathematics
National University of Ireland, Cork, Ireland
m.stynes@ucc.ie

Abstract

Song, Yin and Zhang (*Int. J. Numer. Anal. Model.* 4: 127–140, 2007) discovered a remarkable property of oscillatory finite element solutions of one-dimensional convection-diffusion problems that leads to a novel numerical method for the solution of such problems. In the present paper this property is described using several figures, then a simple proof of the phenomenon is given which is much more intuitive than the technical analysis of Song et al.

1. The problem and the oscillation phenomenon

Consider the two-point boundary value problem

$$-\varepsilon u'' + au' + bu = f \quad \text{on } (0,1), \qquad u(0) = u(1) = 0, \tag{1}$$

where the parameter ε satisfies $0 < \varepsilon \ll 1$, while $a, b, f \in C[0,1]$ with $a > 0$ and $b \geq 0$. Problems such as this, where convection dominates diffusion, typically have solutions that are well-behaved away from $x = 1$ but near $x = 1$ change rapidly. We say that the solution has a boundary layer at $x = 1$. See Figure 1 for an example.

Remark 1. *All figures in this paper are for the particular example*

$$-\varepsilon u'' + u' = x \quad \textit{on } (0,1), \qquad u(0) = u(1) = 0, \tag{2}$$

with $\varepsilon = 5 \times 10^{-3}$. *Its solution behaves in a manner that is completely typical of this class of problems.*

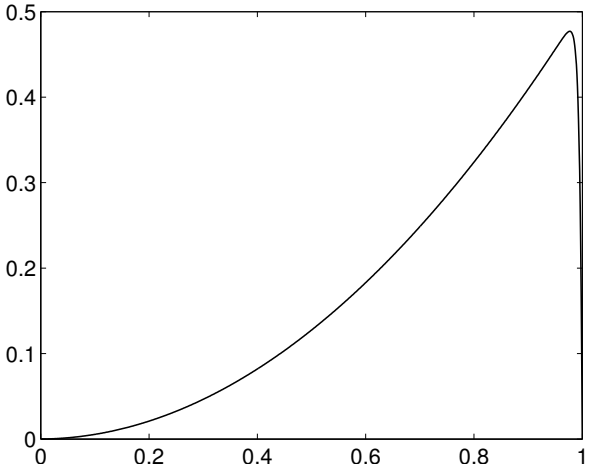

Figure 1: True solution of (2) with $\varepsilon = 5 \times 10^{-3}$.

Problems like (1) and their higher-dimensional analogues have many practical applications so much attention has been paid to their numerical solution. An overview of this area of research is given in [3].

In this paper we shall consider the solution of (1) using a Galerkin finite element method with piecewise linear test and trial functions that we now describe. First, write (1) in the following weak form: find $u \in H_0^1(0,1)$ satisfying

$$\int_0^1 [\varepsilon u'(x)v'(x) + a(x)u'(x)v(x) + b(x)u(x)v(x)]\,dx = \int_0^1 f(x)v(x)\,dx \quad \forall v \in H_0^1(0,1). \tag{3}$$

Let the mesh be $0 = x_0 < x_1 < x_2 < \cdots < x_N = 1$. For $i = 1, 2, \ldots, N-1$, let $\phi_i \in C[0,1]$ be the standard finite element piecewise linear function that satisfies $\phi_i(x_j) = \delta_{ij}$ and support $\phi_i = [x_{i-1}, x_{i+1}]$. Set $V_h = \text{span}\,\{\phi_1, \phi_2, \ldots, \phi_{N-1}\}$, so $V_h \subset H_0^1(0,1)$. Then our piecewise linear Galerkin finite element solution $u_h \in V_h$ is defined by the following discretization of (3):

$$\int_0^1 [\varepsilon u_h'(x)\phi_i'(x) + a_i u_h'(x)\phi_i(x) + b_i u_h(x)\phi_i(x)]\,dx = \int_0^1 f(x)\phi_i(x)\,dx \quad \text{for } i = 1, 2, \ldots, N-1. \tag{4}$$

Note here the nonstandard quadrature rule where $a(x)$ and $b(x)$ were replaced by constants $a_i := a(x_i)$ and $b_i := b(x_i)$ associated with the test function ϕ_i; this rule is introduced to ensure that our finite element method generates the same finite difference scheme as the papers [1, 2], whose results will be used in the proof of our Theorem 1. See also Remark 3 below.

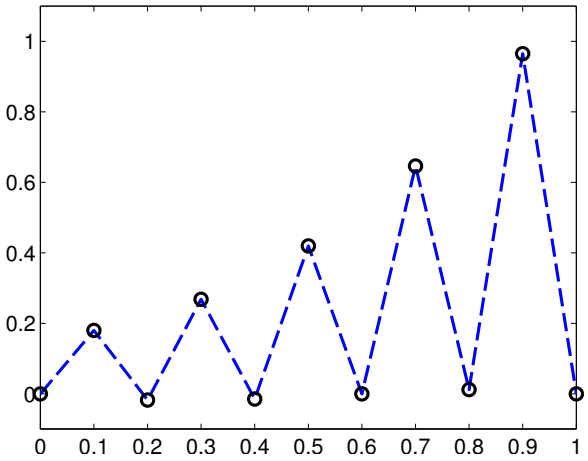

Figure 2: Computed solution on a uniform mesh with 10 intervals.

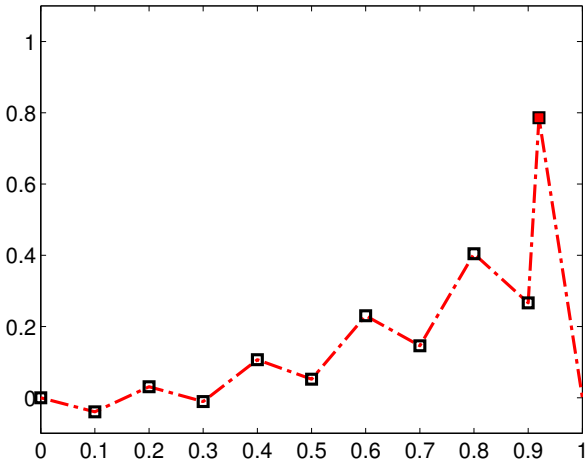

Figure 3: Computed solution with additional mesh point 0.92.

Throughout the paper, when we say "piecewise linear Galerkin method" we mean the finite element method just described.

First, we solve the boundary value problem on a uniform mesh containing N mesh intervals where $N \ll \varepsilon^{-1}$; this relationship between N and ε is usual in practical problems. When (2) is solved by the piecewise linear Galerkin method on a uniform mesh with $N = 10$, the solution is shown in Figure 2. This oscillatory and inaccurate solution is typical of what happens when one applies the piecewise linear Galerkin method to a convection-diffusion problem on a coarse mesh.

In [4] Song, Yin and Zhang modified the mesh in the Galerkin method by adding an arbitrarily-chosen mesh point to the mesh interval where the boundary layer lies. This is the interval $(0.9, 1)$ in our numerical example. Figures 3 and 4 show the computed solutions when the additional mesh points are 0.92 and 0.95 respectively.

Even though the oscillations have diminished, these two computed solutions are

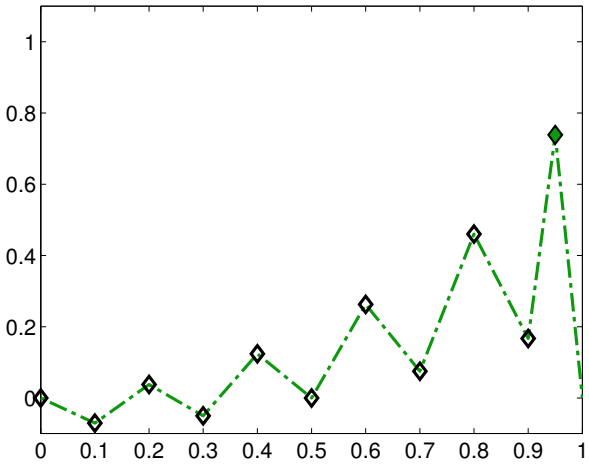

Figure 4: Computed solution with additional mesh point 0.95.

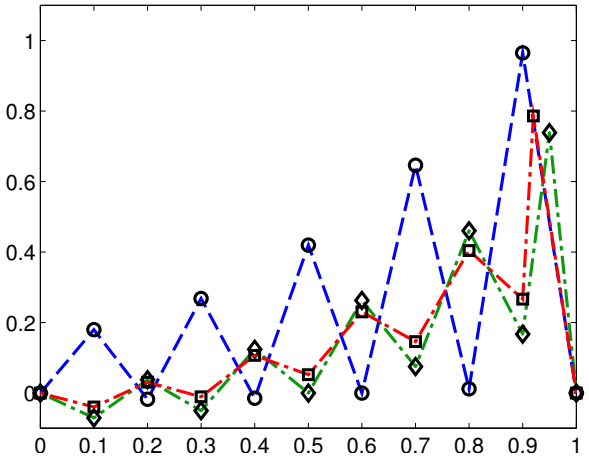

Figure 5: Superimposed computed solutions.

not significantly better than the solution of Figure 2 and little seems to have been gained. But now Song et al. had the clever idea of superimposing all three computed solutions, as shown in Figure 5.

This figure reveals that although the oscillations differ greatly, nevertheless all the computed solutions intersect at a common point in each of the mesh intervals $(0.1, 0.2)$, $(0.2, 0.3)$, $\ldots$, $(0.8, 0.9)$! Further numerical experiments confirm this fact: when a mesh point is added anywhere in the interval $(0.9, 1)$, each computed solution passes through the same fixed point in each of the mesh intervals between 0.1 and 0.9. Indeed, when more than one mesh point is added in (0.9, 1), the piecewise linear Galerkin solution still passes through the same fixed points.

And even more is true: in Figure 6 we superimpose the true solution of Figure 1 on the computed solutions of Figure 5, and clearly the common intersection points of the computed solutions are good approximations of the true solution!

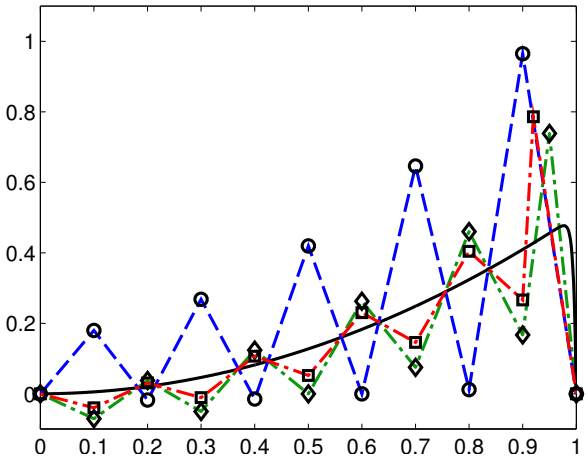

Figure 6: True solution with computed solutions.

The behaviour of Figures 5 and 6 is replicated if one varies N or ε (while keeping $N \ll \varepsilon^{-1}$) and when other test problems of the form (1) are considered.

The question now is: why does this happen?

2. Theoretical explanation

In [4], Song et al. give a complete theoretical explanation of the two phenomena that we have described: common intersection points of all piecewise linear Galerkin solutions when extra mesh point(s) are added inside the mesh interval containing the layer, and the proximity of these common points to the true solution. This analysis is 3 pages long and deals only with the special case of constant a and $b \equiv 0$ (it is stated in [4] that their arguments can be extended to the general case of (1)). Their arguments are somewhat intricate and consequently yield only a limited intuitive understanding of what we have observed experimentally.

We shall now give a much simpler and shorter argument that explains Figures 5 and 6 for the general case of $a, b \in C[0,1]$ and reveals the fundamental reason that these phenomena occur.

Suppose that we solve (1) using the piecewise linear Galerkin method on a uniform mesh with N subintervals, where $N \ll \varepsilon^{-1}$. Set $h = 1/N$. Denote the computed solution by $u_h \in C[0,1]$. The boundary layer in the true solution u lies inside the interval $(1-h, 1)$ because $N \ll \varepsilon^{-1}$; see [3]. We now introduce an arbitrary additional mesh point (or points) in the interval $(1-h, 1)$. Let $\hat{u}_h$ denote the piecewise linear Galerkin solution computed on this modified mesh.

The key insight of our analysis is that because u_h and $\hat{u}_h$ share the same mesh on $[0, 1-h]$, one should compare them there instead of considering them on $[0,1]$.

On the interval $[0, 1-h]$, the computed solution u_h is the piecewise linear Galerkin solution of the two-point boundary problem

$$-\varepsilon v'' + av' + bv = f \quad \text{on } (0, 1-h), \qquad v(0) = 0, \quad v(1-h) = u_h(1-h),$$

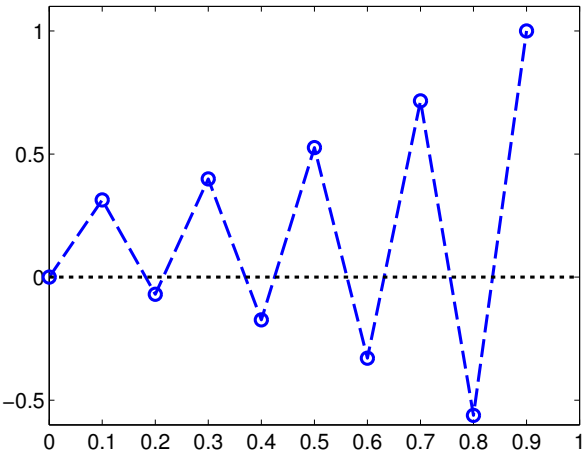

Figure 7: Numerical solution to (6) with $h = 1/10$.

and the computed solution $\hat{u}_h$ is the Galerkin solution of the boundary value problem

$$-\varepsilon w'' + aw' + bw = f \quad \text{on } (0, 1-h), \qquad w(0) = 0, \quad w(1-h) = \hat{u}_h(1-h).$$

Consequently their difference $u_h - \hat{u}_h$ is the piecewise linear Galerkin solution of the boundary value problem

$$-\varepsilon z'' + az' + bz = 0 \quad \text{on } (0, 1-h), \qquad z(0) = 0, \quad z(1-h) = u_h(1-h) - \hat{u}_h(1-h). \tag{5}$$

In Lemma 1 we shall prove that on a uniform mesh with intervals of width h, the piecewise linear Galerkin solution of the problem

$$-\varepsilon \zeta'' + a\zeta' + b\zeta = 0 \quad \text{on } (0, 1-h), \qquad \zeta(0) = 0, \quad \zeta(1-h) = 1 \tag{6}$$

oscillates about zero, in the sense that the computed solution equals zero at one point in each of the mesh intervals $(h, 2h)$, $(2h, 3h), \ldots, (1-2h, 1-h)$ and is otherwise non-zero in $(0, 1-h]$. This statement should be immediately plausible to those with experience in the numerical solution of convection-diffusion problems; it is demonstrated in Figure 7 for the differential operator of (2).

Denote the zeros of the Galerkin solution of (6) by $\zeta_2, \zeta_3, \ldots, \zeta_{N-1}$, where $(j-1)h < \zeta_j < jh$ for each j. Note that the Galerkin solution of (5) is a constant multiple of the Galerkin solution of (6); the multiplier is $u_h(1-h) - \hat{u}_h(1-h)$. Hence the piecewise linear Galerkin solution of (5) also vanishes at the points $\zeta_2, \zeta_3, \ldots, \zeta_{N-1}$. That is, $u_h(\zeta_j) = \hat{u}_h(\zeta_j)$ for each j, which means that these two computed solutions cross at each ζ_j ; and since the ζ_j are generated by problem (6), they are independent of the presence or absence of mesh points in the interval $(1-h, 1)$.

We conclude that all piecewise linear Galerkin solutions of (1) that are computed on a uniform mesh $\{0, h, 2h, \ldots, 1\}$ that is modified by possibly adding mesh point(s) to the interval $(1-h, 1)$ will cross at the fixed points ζ_j for $j = 2, 3, \ldots, N-1$. Thus the phenomenon of Figure 5 has been explained.

Lemma 1. *Consider the two-point boundary value problem* (6). *Subdivide the interval* $[0, 1-h]$ *by a uniform mesh with intervals of width* h *and assume that*

$$\min_{[0,1]} \left(\frac{a}{2} - \left| \frac{hb}{6} - \frac{\varepsilon}{h} \right| \right) > 0. \tag{7}$$

Then the piecewise linear Galerkin solution of (6) *oscillates about zero, in the sense that the computed solution equals zero at one point in each of the mesh intervals* $(h, 2h),\ (2h, 3h), \ldots, (1-2h, 1-h)$ *and is otherwise non-zero in* $(0, 1-h]$.

Proof. Let $g \in C[0, 1-h]$ denote the piecewise linear Galerkin solution of (6) on the given mesh. From (4), after division by h the difference scheme defining the nodal values of g is

$$-\frac{\varepsilon}{h^2}\left(g_{i+1} - 2g_i + g_{i-1}\right) + \frac{a_i(g_{i+1} - g_{i-1})}{2h} + \frac{b_i}{6}(g_{i+1} + 4g_i + g_{i-1}) = 0$$

for $i = 1, \ldots, N-2$, with $g_0 = 0$ and $g_{N-1} = 1$, where $g_j := g(jh)$ for all j. This scheme can be rewritten as

$$\left(\frac{a_i}{2h} + \frac{b_i}{6} - \frac{\varepsilon}{h^2}\right) g_{i+1} + \left(\frac{4b_i}{6} + \frac{2\varepsilon}{h^2}\right) g_i + \left(-\frac{a_i}{2h} - \frac{\varepsilon}{h^2} + \frac{b_i}{6}\right) g_{i-1} = 0 \tag{8}$$

for $i = 1, \ldots, N-2$. The hypothesis (7) ensures that the coefficients of g_{i+1} and g_i are positive but the coefficient of g_{i-1} is negative.

Observe first that the solution of this difference scheme cannot have $g_1 = 0$ because then taking $i = 1$ in (8) would imply that $g_2 = 0$, and a similar inductive argument then leads to $g_{N-1} = 0$ which is false. Thus $g_1 \neq 0$.

If $g_1 > 0$, then taking $i = 1$ in (8) and recalling the signs of the coefficients there and $g_0 = 0$, we see that $g_2 < 0$. Similarly, $g_1 < 0$ implies that $g_2 > 0$. Thus in all cases one has $g_1 g_2 < 0$. One can now proceed inductively, invoking (8) for $i = 2, 3, \ldots, N-2$ and using the signs of its coefficients, to get $g_i g_{i+1} < 0$ for each i. The desired result follows. □

Remark 2. *Inequality* (7) *says that* h *is sufficiently small (so* $a/2$ *dominates* $hb/6$*) and that* ε *is small relative to* h. *Thus* (7) *is in practice a very mild restriction on the mesh.*

The accuracy of the computed solutions at the fixed crossing points that we observed in Figure 6 will now be justified. The argument resembles that of [4, Theorem 3.7], but see Remark 3 below.

Theorem 1. *Subdivide* $[0, 1]$ *by a uniform mesh of width* h. *Assume that* $h \geq \varepsilon\,|\ln \varepsilon|$ *and that* (7) *is satisfied. Then the piecewise linear Galerkin solution* u_h *of the two-point boundary value problem* (1) *satisfies*

$$|u(\zeta_i) - u_h(\zeta_i)| \leq Ch^2 \quad \textit{for } i = 2, 3, \ldots, N-1,$$

where the constant C *is independent of* ε *and* h.

Proof. Since $h \geq \varepsilon\,|\ln\varepsilon|$, one can insert extra mesh points in the interval $(1-h, 1)$ to construct a Bakhvalov mesh for problem (1). See [3] for a description of this mesh. It follows from [1] and [2] (the first paper proves the case $b \equiv 0$ and the second shows how such results can be extended to $b \geq 0$) that the piecewise linear Galerkin solution u_B on the Bakhvalov mesh satisfies $\max_{[0,1]} |u(x) - u_B(x)| \leq Ch^2$ for some constant C. But in particular this implies that $|u(\zeta_i) - u_B(\zeta_i)| \leq Ch^2$ for each i (since (7) holds true by hypothesis, Lemma 1 is valid and consequently the ζ_i are well defined). But our analysis earlier in the section showed that $u_h(\zeta_i) = u_B(\zeta_i)$ for each i, so we are done. □

Remark 3. *The quadrature rule used in* (4) *was chosen to fit with the theory of [1], where the convective term* $(au')(x_i)$ *is approximated by the finite difference*

$$a(x_i)\,\frac{u_h(x_{i+1}) - u_h(x_{i-1})}{x_{i+1} - x_{i-1}}\,.$$

It is pointed out in [1, Remark 4] that, surprisingly, the convergence result for the Bakhvalov mesh that we invoked in our proof of Theorem 1 is no longer valid if instead one uses the slightly different difference approximation

$$\frac{a(x_i)}{2}\cdot\left[\frac{u_h(x_{i+1}) - u_h(x_i)}{x_{i+1} - x_i} + \frac{u_h(x_i) - u_h(x_{i-1})}{x_i - x_{i-1}}\right].$$

Thus it is not clear if Theorem 1 still holds true when we use some alternative quadrature rule in (4)*. This issue seems to have been overlooked in [4], where only constant-coefficient differential operators are analysed in detail and it is asserted that the results can be "readily generalized" to operators with smooth coefficients.*

3. Numerical results

We now describe an algorithm for recovering an accurate approximation to (1) from an oscillatory Galerkin finite element solution. It is equivalent to Algorithm 1 of [4], but closer in spirit to the analysis given in Lemma 1 and Theorem 1.

Step 1: Compute u_h, the Galerkin solution to (4) on a uniform mesh with N intervals of width $h = 1/N$.

Step 2: Compute ζ_h, the Galerkin solution to (6).

Step 3: Take $\zeta_2, \zeta_3, \ldots, \zeta_{N-1}$ to be the zeros of $\zeta_h(x)$ in $(h, 1-h)$. That is,

$$\zeta_i = \frac{x_{i-1}\zeta(x_i) - x_i\zeta(x_{i-1})}{\zeta(x_i) - \zeta(x_{i-1})} \quad \text{for } i = 2, 3, \ldots, N-1.$$

Return: $\{u_h(0), u_h(\zeta_2), u_h(\zeta_3), \ldots, u_h(\zeta_{N-1}), u_h(1)\}$.

We now compute the errors obtained when this algorithm is applied to our test problem (2), in order to demonstrate that the resulting solution is robust with respect to ε and converges as described in Theorem 1. In fact not only is the computed solution second-order accurate at the points ζ_i, but also its piecewise linear interpolant $\widetilde{u}_h$ (with knots at the ζ_i) is pointwise second-order accurate on the interval $[0, \zeta_{N-1}]$. In Table 1 we report the values of

$$\mathcal{E}_N := \|u - \widetilde{u}_h\|_{L_\infty[0,1-\zeta_{N-1}]}$$

for a range of values of ε and N. We consider only small ε since, when ε is large, the numerical solution is not oscillatory and consequently one would not have to apply the above recovery algorithm.

Table 1 shows that the method is second-order convergent.

ε	$N = 2^5$	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$
10^{-6}	4.88e-04	1.22e-04	3.05e-05	7.62e-06	1.90e-06	4.74e-07
10^{-7}	4.88e-04	1.22e-04	3.05e-05	7.63e-06	1.91e-06	4.77e-07
10^{-8}	4.88e-04	1.22e-04	3.05e-05	7.63e-06	1.91e-06	4.77e-07
10^{-9}	4.88e-04	1.22e-04	3.05e-05	7.63e-06	1.91e-06	4.77e-07
10^{-10}	4.88e-04	1.22e-04	3.05e-05	7.63e-06	1.91e-06	4.77e-07

Table 1: Errors $\mathcal{E}_N$ for the above algorithm applied to (2).

References

[1] Andreev, V. B. and Kopteva, N. V.: Investigation of difference schemes with an approximation of the first derivative by a central difference relation. Zh. Vychisl. Mat. i Mat. Fiz. **36** (1996), 101–117; translation in Comput. Math. Math. Phys. **36** (1996), 1065-1078.

[2] Kopteva, N. V.: On the convergence, uniform with respect to the small parameter, of a scheme with central difference on refined grids. Zh. Vychisl. Mat. Mat. Fiz. **39** (1999), 1662–1678; translation in Comput. Math. Math. Phys. **39** (10) (1999), 1594-1610.

[3] Roos, H.-G., Stynes, M., and Tobiska, L.: *Robust numerical methods for singularly perturbed differential equations. Convection-diffusion-reaction and flow problems. Second edition.* Springer Series in Computational Mathematics, vol. 24. Springer-Verlag, Berlin, 2008.

[4] Song, Q. S., Yin, G., and Zhang, Z.: An ε-uniform finite element method for singularly perturbed two-point boundary value problems. Int. J. Numer. Anal. Model. **4** (2007), 127–140.