A Non-conforming Combination of the Finite Element & Finite Volume Methods for a Convection-Diffusion Problem

Song Wang
University of Western Australia
mailto:swang@maths.uwa.edu.au

joint work with Z.C. Li
Outline

• The continuous problem (linear convection-diffusion equation with a singular perturbation parameter $\varepsilon$)
Outline

• The continuous problem (linear convection-diffusion equation with a singular perturbation parameter $\varepsilon$)

• A priori estimates for the solution and derivatives
Outline

- The continuous problem (linear convection-diffusion equation with a singular perturbation parameter $\varepsilon$)
- A priori estimates for the solution and derivatives
- Formulation of the nonlinear combination – the standard p.i. FEM in one subregion and the exponentially fitted FV in the rest
Outline

- The continuous problem (linear convection-diffusion equation with a singular perturbation parameter $\varepsilon$)
- A priori estimates for the solution and derivatives
- Formulation of the nonlinear combination – the standard p.l. FEM in one subregion and the exponentially fitted FV in the rest
- Coercivity analysis – major difficulty is the discontinuity in the FE spaces
Outline

- The continuous problem (linear convection-diffusion equation with a singular perturbation parameter $\varepsilon$)
- A priori estimates for the solution and derivatives
- Formulation of the nonlinear combination – the standard p.l. FEM in one subregion and the exponentially fitted FV in the rest
- Coercivity analysis – major difficulty is the discontinuity in the FE spaces
- Error estimates – almost uniform convergence $O \left( h \left| \ln \varepsilon / \ln h \right|^{1/2} \right)$. 
Outline

- The continuous problem (linear convection-diffusion equation with a singular perturbation parameter $\varepsilon$)
- A priori estimates for the solution and derivatives
- Formulation of the nonlinear combination – the standard p.l. FEM in one subregion and the exponentially fitted FV in the rest
- Coercivity analysis – major difficulty is the discontinuity in the FE spaces
- Error estimates – almost uniform convergence $O\left(h \left| \ln \varepsilon / \ln h \right|^{1/2} \right)$.

Warning: Notation bombing!
Consider stationary, linear, convection-diffusion problems of the form

\[- \nabla \cdot \mathbf{f}_u + Gu = F \quad \text{in} \quad \Omega := (0, 1)^2, \quad (1)\]

\[\mathbf{f}_u = \varepsilon \nabla u - au, \quad (2)\]

\[u|_{\partial \Omega} = 0, \quad (3)\]

where
Consider stationary, linear, convection-diffusion problems of the form

\[- \nabla \cdot f_u + Gu = F \quad \text{in} \quad \Omega := (0,1)^2,\]

\(f_u = \varepsilon \nabla u - au,\)

\(u|_{\partial \Omega} = 0,\)

where

\(\partial \Omega\) denotes the boundary of \(\Omega,\)

\(\varepsilon > 0\) is a positive parameter,

\(a = (a_1, a_2)\) is a known vector-valued function and

\(F\) is given function.
We assume that $\mathbf{a} \in (C^1(\bar{\Omega}))^2$, $G \in C(\Omega) \cap H^1(\Omega)$ and $F \in L^\infty(\Omega)$. 
We assume that \( a \in (C^1(\Omega))^2 \), \( G \in C(\Omega) \cap H^1(\Omega) \) and \( F \in L^\infty(\Omega) \).

Also, \( a \) satisfies
\[
\frac{1}{2} \nabla \cdot a + G \geq \alpha > 0 \quad \text{in} \quad \Omega
\]
(4)

for some positive number \( \alpha \), and

\[
a_1 \geq \alpha_1 > 0, \quad a_2 \geq \alpha_2 > 0 \quad \text{in} \quad \Omega.
\]
(5)
We assume that $a \in (C^1(\bar{\Omega}))^2$, $G \in C(\Omega) \cap H^1(\Omega)$ and $F \in L^\infty(\Omega)$.

Also, $a$ satisfies

$$\frac{1}{2} \nabla \cdot a + G \geq \alpha > 0 \quad \text{in} \quad \Omega$$

for some positive number $\alpha$, and

$$a_1 \geq \alpha_1 > 0, \quad a_2 \geq \alpha_2 > 0 \quad \text{in} \quad \Omega.$$  \hfill (5)

In this case, the solution to the problem has two exponential boundary layers of width $O(\varepsilon)$ at $x = 1$ and $y = 1$. 
A Example

\[ \frac{d}{dx} \left( \varepsilon \frac{du}{dx} + au \right) = 0 \quad \text{in} \quad (0, 1) \tag{6} \]
\[ u(0) = 1, \quad u(1) = 0 \tag{7} \]

with nonzero constant \( a \). This is a 2nd order constant coefficient, homogeneous ODE and the general solution is

\[ u = c_1/a + c_2 e^{-ax/\varepsilon} \]

Using the boundary conditions we can fix \( c_1 \) and \( c_2 \), and so

\[ u = \frac{e^{a(1-x)/\varepsilon} - 1}{e^{a/\varepsilon} - 1} \]

with the derivative

\[ \frac{du}{dx} = - \frac{ae^{a(1-x)/\varepsilon}}{\varepsilon(e^{a/\varepsilon} - 1)}. \]
Application of the 3-point central difference scheme to the example problem

Divide $[0, 1]$ into $n$ equal subintervals:

$$[0, 1] = [x_0, x_1] \cup [x_1, x_2] \cup \cdots \cup [x_{n-1}, x_n]$$

with step length $h$. The 3-point central difference scheme on this uniform mesh has the form:

$$\frac{d^2u}{dx^2} \approx \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2}, \quad \frac{du}{dx} \approx \frac{u_{i+1} - u_{i-1}}{2h}$$

So the discretised form of the DE is

$$\varepsilon \cdot \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + a \cdot \frac{u_{i+1} - u_{i-1}}{2h} = 0$$

$$u_0 = 1, \quad u_n = 0$$
\( i = 1, 2, \cdots, n - 1 \). This can be rewritten as

\[
\left( \frac{\varepsilon}{h^2} - \frac{a}{2h} \right) u_{i-1} - \frac{2\varepsilon}{h^2} u_i + \left( \frac{\varepsilon}{h^2} + \frac{a}{2h} \right) u_{i+1} = 0
\]

with \( u_0 = 1, u_n = 0 \), \( i = 1, 2, \cdots, n - 1 \), or in matrix form

\[
Au := \begin{pmatrix}
  d & c & 0 & \cdots & \cdots & 0 \\
  b & d & c & \cdots & \cdots & 0 \\
  \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
  \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
  0 & \cdots & \cdots & b & d & c \\
  0 & \cdots & \cdots & 0 & b & d
\end{pmatrix}
\begin{pmatrix}
  u_1 \\
  u_2 \\
  \vdots \\
  \vdots \\
  u_{n-2} \\
  u_{n-1}
\end{pmatrix}
= \begin{pmatrix}
  -b \\
  0 \\
  \vdots \\
  \vdots \\
  0 \\
  0
\end{pmatrix}
\]

where

\[
d = -\frac{2\varepsilon}{h}, \quad b = \frac{\varepsilon}{h} - \frac{a}{2}, \quad c = \frac{\varepsilon}{h} + \frac{a}{2}.
\]
Obvious $d \to 0$ and $b, c \to \mp a/2$ as $\varepsilon \to 0$, and thus $A \to \text{SINGULAR}$ as $\varepsilon \to 0$. This is numerically unstable.
The variational problem

Problem 2.1: Find \( u \in H_0^1(\Omega) \) such that for all \( v \in H_0^1(\Omega) \)

\[
A(u, v) = (F, v),
\]

where \( A(\cdot, \cdot) \) is a bilinear form on \((H_0^1(\Omega))^2\) defined by

\[
A(u, v) = (\varepsilon \nabla u - a u, \nabla v).
\]
The variational problem

Problem 2.1: Find $u \in H_0^1(\Omega)$ such that for all $v \in H_0^1(\Omega)$

$$A(u, v) = (F, v),$$

where $A(\cdot, \cdot)$ is a bilinear form on $(H_0^1(\Omega))^2$ defined by

$$A(u, v) = (\varepsilon \nabla u - au, \nabla v).$$

It is easy to show that

$$A(u, u) = ||u||^2_\varepsilon, \quad \forall u \in H_0^1(\Omega),$$

where

$$||v||^2_\varepsilon = (\varepsilon \nabla v, \nabla v) + ((\frac{1}{2} \nabla \cdot a + G)v, v).$$

is a norm on $H_0^1(\Omega)$.
We divide the solution region $\Omega$ into two parts $\Omega_1$ and $\Omega_2$ given respectively by

$$\Omega_1 = (0, 1 - \delta_1) \times (0, 1 - \delta_2),$$

$$\Omega_2 = (1 - \delta_1, 1) \times (0, 1) \cup (0, 1 - \delta_1) \times (1 - \delta_2, 1),$$

with $\delta_1, \delta_2 \in (0, 1)$ (cf. Figure 1). Obviously $\bar{\Omega}_1 \cup \bar{\Omega}_2 = \bar{\Omega}$. 
We divide the solution region $\Omega$ into two parts $\Omega_1$ and $\Omega_2$ given respectively by

$$\Omega_1 = (0, 1 - \delta_1) \times (0, 1 - \delta_2),$$

$$\Omega_2 = (1 - \delta_1, 1) \times (0, 1) \cup (0, 1 - \delta_1) \times (1 - \delta_2, 1),$$

with $\delta_1, \delta_2 \in (0, 1)$ (cf. Figure 1). Obviously $\bar{\Omega}_1 \cup \bar{\Omega}_2 = \bar{\Omega}$.

The region $\Omega_2$ consists of three subregions $\Omega_2^{(1)}, \Omega_2^{(2)}$ and $\Omega_2^{(3)}$ defined respectively by

$$\Omega_2^{(1)} = (1 - \delta_1, 1) \times (0, 1 - \delta_2),$$
$$\Omega_2^{(2)} = (0, 1 - \delta_1) \times (1 - \delta_2, 1),$$
$$\Omega_2^{(3)} = (1 - \delta_1, 1) \times (1 - \delta_2, 1).$$
Figure 1: Subdomains $\Omega_1$ and $\Omega_2 = \Omega_2^{(1)} \cup \Omega_2^{(2)} \cup \Omega_2^{(3)}$, $\Gamma = \Gamma_1 \cup \Gamma_2$. 

\[ \Omega_1 \quad \Gamma_1 \quad \Omega_2^{(1)} \quad \Omega_2 \quad \Gamma_2 \quad \Omega_2^{(2)} \quad \Omega_2^{(3)} \]
The choice of the transition parameters \( \delta_1 \) and \( \delta_2 \) is rather arbitrary, but it is required that \( \Omega_2 \) contains the boundary layers and \( \delta_1, \delta_2 = O(\varepsilon) \). One choice is

\[
\delta_1 = \frac{\beta}{\alpha_1} \varepsilon \ln(1/\varepsilon) \quad \text{and} \quad \delta_2 = \frac{\beta}{\alpha_2} \varepsilon \ln(1/\varepsilon), \quad (11)
\]

where \( \beta \geq 1 \) is a positive constant.
The choice of the transition parameters $\delta_1$ and $\delta_2$ is rather arbitrary, but it is required that $\Omega_2$ contains the boundary layers and $\delta_1, \delta_2 = O(\varepsilon)$. One choice is

$$
\delta_1 = \frac{\beta}{\alpha_1} \varepsilon \ln(1/\varepsilon) \quad \text{and} \quad \delta_2 = \frac{\beta}{\alpha_2} \varepsilon \ln(1/\varepsilon), \quad (11)
$$

where $\beta \geq 1$ is a positive constant.

We let $\Gamma_1 = \bar{\Omega}_1 \cap \bar{\Omega}_2^{(1)}$ and $\Gamma_2 = \bar{\Omega}_1 \cap \bar{\Omega}_2^{(2)}$, and put $\Gamma = \Gamma_1 \cup \Gamma_2$. In the rest of this paper $\Gamma$ is sometimes regarded as an oriented curve. If $\Gamma$ is oriented counter-clockwise then it is denoted as $\Gamma^+$. Otherwise, we use $\Gamma^-$ to denote it.
Assumption 2.1: \( u \) has the representation

\[
    u = U_1 + U_2 + U_3 + U_4,
\]

such that

\[
    ||U_1||_{k,\infty,\Omega} \leq C \quad \text{for} \quad k = 0, 1, 2,
\]

and \( U_2, U_3 \) and \( U_4 \) satisfy

\[
    \left| \frac{\partial^{i+j} U_2}{\partial x^i \partial y^j} \right| \leq C \varepsilon^{-i} \exp \left( -\frac{\alpha_1 (1-x)}{\varepsilon} \right),
\]

\[
    \left| \frac{\partial^{i+j} U_3}{\partial x^i \partial y^j} \right| \leq C \varepsilon^{-j} \exp \left( -\frac{\alpha_2 (1-y)}{\varepsilon} \right),
\]

\[
    \left| \frac{\partial^{i+j} U_4}{\partial x^i \partial y^j} \right| \leq C \varepsilon^{-(i+j)} \exp \left( -\frac{\alpha_1 (1-x)}{\varepsilon} \right) \exp \left( -\frac{\alpha_2 (1-y)}{\varepsilon} \right),
\]

for \( 0 \leq i + j \leq 2 \) and some positive constant \( C \).
Theorem 2.1: If $\beta \geq 3/2$, then the solution $u$ to Problem 2.1 satisfies

$$||u||_{i, \Omega_1} \leq C \quad i = 0, 1, 2,$$  \hspace{1cm} (13)

for some positive constants $C$, independent of $u$.

Furthermore, if $\beta \geq 2$, then

$$||u||_{i, \infty, \Omega_1} \leq C \quad i = 0, 1, 2.$$  \hspace{1cm} (14)
Proof: Let $C$ be a generic positive constant independent of $u$ and $\varepsilon$. We only show that $\|u\|_{2,\Omega_1} \leq C$.

From the definition of $\Omega_1$ and (12) we have

\[
\|\partial^2 U_2/\partial x^2\|_{0,\Omega_1}^2 \leq C\varepsilon^{-4} \int_{\Omega_1} \exp \left( -\frac{2\alpha_1}{\varepsilon} (1 - x) \right) \, dx
\]

\[
= C(1 - \delta_2) \left. e^{-2\alpha_1/\varepsilon} e^{2\alpha_1 x/\varepsilon} \right|_{1-\delta_1}^{0}
\]

\[
= C(1 - \delta_2) \left( e^{-2\alpha_1(1-\delta_1)/\varepsilon} - e^{-2\alpha_1/\varepsilon} \right)
\]

\[
= C \frac{1 - \delta_2}{2\alpha_1 \varepsilon^3} \left( e^{-2\beta \ln \frac{1}{\varepsilon}} - e^{-2\alpha_1/\varepsilon} \right)
\]

\[
= C \frac{1 - \delta_2}{2\alpha_1 \varepsilon^3} \left( e^{2\beta - 3} - \varepsilon^{-3} e^{-2\alpha_1/\varepsilon} \right).
\]

Note that $\varepsilon^{-3} e^{-2\alpha_1/\varepsilon}$ is uniformly bounded above for all $0 < \varepsilon \leq O(1)$ since $\alpha_1 > 0$. Thus, if $\beta \geq 3/2$ we have from
the above inequality

\[ \left\| \frac{\partial^2 U_2}{\partial x^2} \right\|_{0, \Omega_1}^2 \leq C. \]

The proof of boundedness for the first and second seminorms of all other terms are analogous to this case. Therefore we have shown (13).
the above inequality

\[ \| \frac{\partial^2 U_2}{\partial x^2} \|_{0, \Omega_1}^2 \leq C. \]

The proof of boundedness for the first and second seminorms of all other terms are analogous to this case. Therefore we have shown (13).

For any fixed \( \varepsilon \), the function in the right side of (12) is increasing in \( x \). Thus we have

\[ \| \frac{\partial^2 U_2}{\partial x^2} \|_{0, \infty, \Omega_1} \leq C \varepsilon^{-2} e^{-\alpha_1 \delta_1 / \varepsilon} = C \varepsilon^\beta^{-2}. \]

The last term in the above is uniformly bounded for all \( \beta \geq 2 \).
The finite element formulation
The finite element formulation

$T_h^{(1)}$ — a regular triangulation on $\Omega_1$. 
The finite element formulation

\[ T_h^{(1)} \] — a regular triangulation on \( \Omega_1 \).

\[ \{ x_i \}_{1}^{N_1} \] — vertices of \( T_h^{(1)} \) not on \( \partial \Omega \).
The finite element formulation

\[ T_h^{(1)} \] — a regular triangulation on \( \Omega_1 \).

\[ \{x_i\}_{1}^{N_1} \] — vertices of \( T_h^{(1)} \) not on \( \partial\Omega \).

\[ T_h^{(2)} \] — a mesh on \( \Omega_2 \) containing long, thin rectangles.
The finite element formulation

\( T_h^{(1)} \) — a regular triangulation on \( \Omega_1 \).

\( \{x_i\}_{1}^{N_1} \) — vertices of \( T_h^{(1)} \) not on \( \partial\Omega \).

\( T_h^{(2)} \) — a mesh on \( \Omega_2 \) containing long, thin rectangles.

\( \{x_i\}_{N_1+1}^{N_1+N_2} \) — vertices of \( T_h^{(2)} \) not on \( \partial\Omega \).
The finite element formulation

\( T_h^{(1)} \) — a regular triangulation on \( \Omega_1 \).

\( \{x_i\}_{1}^{N_1} \) — vertices of \( T_h^{(1)} \) not on \( \partial \Omega \).

\( T_h^{(2)} \) — a mesh on \( \Omega_2 \) containing long, thin rectangles.

\( \{x_i\}_{N_1+1}^{N_1+N_2} \) — vertices of \( T_h^{(2)} \) not on \( \partial \Omega \).

\( E_h^{(2)} \) — set of edges of \( T_h^{(2)} \). Each edge has a length of either \( \delta_1/M_1 \) or \( \delta_2/M_2 \).
The finite element formulation

\( T_h^{(1)} \) — a regular triangulation on \( \Omega_1 \).

\( \{ x_i \}_{1}^{N_1} \) — vertices of \( T_h^{(1)} \) not on \( \partial \Omega \).

\( T_h^{(2)} \) — a mesh on \( \Omega_2 \) containing long, thin rectangles.

\( \{ x_i \}_{N_1+1}^{N_1+N_2} \) — vertices of \( T_h^{(2)} \) not on \( \partial \Omega \).

\( E_h^{(2)} \) — set of edges of \( T_h^{(2)} \). Each edge has a length of either \( \delta_1/M_1 \) or \( \delta_2/M_2 \).

The meshes \( T_h^{(1)} \) and \( T_h^{(2)} \) form a conforming mesh on \( \Omega \) and we denote it by \( T_h \).
Figure 2: A typical hybrid mesh for $\Omega$. 
Associated with $T_h^{(2)}$, we define two meshes dual to it.

$D_h^{(2)}$ — the Dirichlet tessellation

$B_h^{(2)}$ — quadrilateral mesh by connecting the two end-points of the edge and the mid-points of the rectangles (or rectangle if the edge is on $\Gamma$) sharing the edge.
Figure 3: Elements and edges associated with the node $x_i$. 
Finite element subspaces:

$U_h^{(1)}$ — the conventional piecewise linear finite element space of dimension $N_1$ constructed on the partition $T_h^{(1)}$. 
Finite element subspaces:

$U_h^{(1)}$ — the conventional piecewise linear finite element space of dimension $N_1$ constructed on the partition $T_h^{(1)}$.

$V_h^{(2)} = \text{span}\{\xi_i\}$ — test space on $T_h^{(2)}$, where $\xi_i$ is piecewise constant given by

$$
\xi_i = \begin{cases} 
1 & \text{on } d_i \\
0 & \text{otherwise}
\end{cases}.
$$
$U_h^{(2)}$ — exponentially fitted trial space constructed below.

For each $e_{i,j} \in E_h^{(2)}$ connecting the two neighbouring nodes $x_i$ and $x_j$, we define an exponential function $\phi_{i,j}$ on $e_{i,j}$ by

$$
\frac{d}{de_{i,j}} \left( \varepsilon \frac{d\phi_{i,j}}{de_{i,j}} - \bar{a}_{i,j} \phi_{i,j} \right) = 0,
$$

$$
\phi_{i,j}(x_i) = 1, \quad \phi_{i,j}(x_j) = 0,
$$

$e_{i,j}$ — unit vector from $x_i$ to $x_j$

$\bar{a}_{i,j}$ — constant approximation to $a \cdot e_{i,j}$

$$
\phi_i = \begin{cases} 
\phi_{i,j} & \text{on } b_{i,j} \text{ if } j \in I_i \\
0 & \text{otherwise},
\end{cases}
$$
$b_{i,j}$ — the element of $B_h$ containing $e_{i,j}$,

$$I_i = \{j : e_{i,j} \in E_h\}$$  \hspace{1cm} (16)

denotes the index set of all neighbour nodes of $x_i$.

We put $U_h^{(2)} = \text{span}\{\phi_i\}_{1}^{N'}$.

The projection of the flux of the $U_h^{(2)}$-interpolant $u_I$ of $u$ on $e_{i,j}$ satisfies

$$f_{i,j} := \varepsilon \frac{d u_I}{d e_{i,j}} - \bar{a}_{i,j} u_I = \frac{\varepsilon}{|e_{i,j}|} \left( B\left( \frac{\bar{a}_{i,j} |e_{i,j}|}{\varepsilon} \right) u_j - B\left( -\frac{\bar{a}_{i,j} |e_{i,j}|}{\varepsilon} \right) u_i \right)$$  \hspace{1cm} (17)

on the edge $e_{i,j}$, where $B$ denotes the Bernoulli function defined by

$$B(x) = \begin{cases} 
\frac{x}{e^x - 1}, & x \neq 0, \\
1, & x = 0.
\end{cases}$$  \hspace{1cm} (18)
Also, the approximation error in $f_{i,j}$ satisfies

$$\| f_u \cdot e_{i,j} - f_{i,j} \|_{\infty,e_{i,j}} \leq C \left( |f_u|_{1,\infty,e_{i,j}} + |a|_{1,\infty,e_{i,j}} \| u \|_{\infty,e_{i,j}} \right),$$

where $C$ is a positive constant independent of $h$, $u$ and $\varepsilon$. 
We now choose

\[ U_h = U_h^{(1)} \oplus U_h^{(2)} \], and

\[ V_h = U_h^{(1)} \oplus V_h^{(2)} \].
We now choose
\[ U_h = U_h^{(1)} \oplus U_h^{(2)} , \] and
\[ V_h = U_h^{(1)} \oplus V_h^{(2)} . \]

Both \( U_h \) and \( V_h \) are non-conforming as they are not continuous across \( \Gamma \).

We let
\[ \hat{a}|_{b_{i,j}} = \bar{a}_{i,j} e_{i,j} + \bar{a}_{i,j}^\perp l_{i,j}, \quad (20) \]

and introduce \( P : C(\bar{\Omega}_2) \mapsto V_h^{(2)} \) such that
\[ P(w)(x) = \sum_{i=N_1+1}^{N_2} w(x_i) \xi_i(x), \quad x \in \bar{\Omega}_2, \quad (21) \]
Problem 3.1: Find $u_h \in U_h$ such that

$$a(u_h, v_h) := a_1(u_h, v_h) + a_2(u_h, v_h) = (F, v_h), \quad \forall v_h \in V_h,$$

(22)

where $a_1(\cdot, \cdot)$ and $a_2(\cdot, \cdot)$ are bilinear forms on $\Omega_1$ and $\Omega_2$ defined respectively by

$$a_1(u_h, v_h) = \int_{\Omega_1} (\varepsilon \nabla u_h - a u_h) \cdot \nabla v_h \, dx + (G u_h, v_h)_{\Omega_1},$$

(23)

$$a_2(u_h, v_h) = - \sum_{d \in D_h^{(2)}} \int_{\partial d \setminus \partial \Omega_2} (\varepsilon \nabla u_h - \hat{a} u_h) \cdot n v_h |_d \, ds$$

$$+ (P(G u_h), v_h)_{\Omega_2}.$$

(24)

NB. $V_h$ and $U_h$ are equivalent under the mapping $P$, i.e.,

$P : U_h \mapsto V_h$ is surjective.
The Bubnov-Galerkin problem corresponding to Problem 3.1:

Problem 3.2: Find $u_h \in U_h$ such that for all

$$b(u_h, v_h) = (F, v_h)_{\Omega_1} + (F, P(v_h))_{\Omega_2}, \quad \forall v_h \in U_h, \quad (25)$$

where $(\cdot, \cdot)_{\Omega_k}$ denotes the inner product on $L^2(\Omega_k)$ for $k = 1, 2$ and $b(\cdot, \cdot)$ is a bilinear form on $U_h \times U_h$ defined by

$$b(u_h, v_h) := a_1(u_h, v_h) + a_2(u_h, P(v_h)). \quad (26)$$
Coercivity of the bilinear form $b(\cdot, \cdot)$
Coercivity of the bilinear form $b(\cdot, \cdot)$

Assumption 4.1: Let the mesh $T_h^{(2)}$ be sufficiently fine such that the inequality

$$\frac{1}{2} \int_{\partial d_i} \hat{a} \cdot n ds + G(x_i)|d_i| \geq \alpha_0 > 0$$

holds for all $d_i \in D_h^{(2)}$, where $\hat{a}$ is the approximation of $a$ defined in (20) and $x_i$ denotes the mesh node contained in $d_i$. 
Coercivity of the bilinear form $b(\cdot, \cdot)$

Assumption 4.1: Let the mesh $T_h^{(2)}$ be sufficiently fine such that the inequality

$$\frac{1}{2} \int_{\partial d_i} \hat{a} \cdot n ds + G(x_i)|d_i| \geq \alpha_0 > 0 \quad (27)$$

holds for all $d_i \in D_h^{(2)}$, where $\hat{a}$ is the approximation of $a$ defined in (20) and $x_i$ denotes the mesh node contained in $d_i$.

Furthermore, since all the mesh lines in $T_h^{(2)}$ are parallel to one of the axes and $a$ satisfies (5), it is obvious that

$$\min_{e_{i,j} \in E_h^{(2)}} |\hat{a} \cdot e_{i,j}| = \min_{e_{i,j} \in E_h^{(2)}} |a_{i,j}| \geq \min \{\alpha_1, \alpha_2\}. \quad (28)$$
Note that when restricted to $\Omega_1$, the test and trial spaces are equal to each other and thus from (23) we have

$$a_1(u, u) = \int_{\Omega_1} (\varepsilon \nabla u - au) \cdot \nabla u ds + \int_{\Omega_1} Gu^2 dx$$

$$= \varepsilon \|\nabla u\|^2_{0, \Omega_1} - \int_{\Omega_1} au \cdot \nabla u ds + (Gu, u)_{\Omega_1} \quad (29)$$

for any $u \in U_h$. Observing that $u \nabla u = \frac{1}{2} \nabla (u^2)$, a direct application of the Gauss-Green-Stokes formula then yields:

$$-\frac{1}{2} \int_{\Omega_1} a \cdot \nabla (u^2) dx = -\frac{1}{2} \int_{\partial \Omega_1} u^2 a \cdot n ds + \frac{1}{2} u^2 \nabla \cdot a dx.$$
Substituting the above into (29) we obtain using (4)

\[ a_1(u, u) = \varepsilon \| \nabla u \|_{0, \Omega_1}^2 + \frac{1}{2} (\nabla \cdot a u, u)_{\Omega_1} - \frac{1}{2} \int_{\partial \Omega_1} u^2 a \cdot n ds \]

\[ = \varepsilon \| \nabla u \|_{0, \Omega_1}^2 + ((\frac{1}{2} \nabla \cdot a + G)u, u)_{\Omega_1} \]

\[- \frac{1}{2} \int_{\Gamma^+} u^2 a \cdot n ds \]

\[ \geq \varepsilon \| \nabla u \|_{0, \Omega_1}^2 + \alpha \| u \|_{0, \Omega_1}^2 - \frac{1}{2} \int_{\Gamma^+} u^2 a \cdot n ds, \quad (30) \]

since \( u = 0 \) on \( \partial \Omega_1 \setminus \Gamma^+ \).
Now, for any \( u \in U_h \) it has been shown that

\[
a_2(u, P(u)) = - \sum_{d_i \in D_h^{(2)}} \int_{\partial d_i \setminus \partial \Omega_N} (\varepsilon \nabla u_h - \hat{a} u_h) \cdot n P(u_h) ds
\]

\[
+ \sum_{d_i \in D_h^{(2)}} G(x_i) u_i^2 |d_i|
\]

\[
= \sum_{e_{i,j} \in E_h^{(2)}} \sigma_{i,j} B(\rho_{i,j})(1 + e^{\rho_{i,j}})(u_j - u_i)^2 \frac{|b_{i,j}|}{|e_{i,j}|}
\]

\[
+ \sum_{e_{i,j} \in E_h^{(2)}} \frac{\bar{a}_{i,j}}{2} (u_i^2 - u_j^2) |l_{i,j}| + \sum_{d_i \in D_h^{(2)}} G(x_i) u_i^2 |d_i|, \tag{31}
\]

where \( l_{i,j} = \partial d_i \cap \partial d_j \), \( \sigma_{i,j} = \varepsilon / |e_{i,j}| \), \( \rho_{i,j} = \bar{a}_{i,j} / \sigma_{i,j} \) and \( B(\cdot) \) is the Bernoulli function defined in (18).
We can show that

$$\sum_{e_{i,j} \in E_h^{(2)}} \frac{\bar{a}_{i,j}}{2} (u_i^2 - u_j^2) |l_{i,j}| + \sum_{d_i \in D_h^{(2)}} G(x_i) u_i^2 |d_i| \geq \alpha_0 \sum_{d_i \in D_h^{(2)}} u_i^2 - \frac{1}{2} \sum_{d_i \in D_h^{(2)}} u_i^2 \int_{\partial d_i \cap \Gamma^-} \hat{a} \cdot n \, ds. \quad (32)$$
Using the definition of the Bernoulli function $B(\cdot)$ in (18) we have

$$\sigma_{i,j} B(\rho_{i,j})(1 + e^{\rho_{i,j}}) = \bar{a}_{i,j} \frac{e^{\rho_{i,j}} + 1}{e^{\rho_{i,j}} - 1} \geq |\bar{a}_{i,j}|.$$  

Using this inequality, (32) and (31) we obtain

$$a_2(u, P(u)) \geq C \left( \sum_{e_{i,j} \in E_h^{(2)}} |e_{i,j}| \left( \frac{u_j - u_i}{|e_{i,j}|} \right)^2 |b_{i,j}| + \sum_{d_i \in D_h^{(2)}} u_i^2 \right)$$  

$$- \frac{1}{2} \sum_{d_i \in D_h^{(2)}} u_i^2 \int_{\partial d_i \cap \Gamma^-} \mathbf{a} \cdot \mathbf{n} ds.$$  

(33)
Let $\| \cdot \|$ be a functional on the finite element space $U_h$ defined by

$$
\| v \| = \left( \| v \|_{\varepsilon, \Omega_1}^2 + \| v \|_{h, \Omega_2}^2 \right)^{1/2},
$$

where

$$
\| v \|_{\varepsilon, \Omega_1}^2 = \varepsilon \| \nabla v \|_{0, \Omega_1}^2 + \| v \|_{0, \Omega_1}^2,
$$

$$
\| v \|_{h, \Omega_2}^2 = \sum_{e_{i,j} \in E_h^{(2)}} \left| e_{i,j} \right| \left( \frac{v_j - v_i}{\left| e_{i,j} \right|} \right)^2 |b_{i,j}| + \sum_{d_i \in D_h^{(2)}} v_i^2.
$$

It is easy to see that $\| \cdot \|$ is a semi-discrete energy norm on $U_h$. 
Theorem 4.1: Let Assumption 4.1 be fulfilled. Then, for all $u \in U_h$, we have

$$b(u, u) \geq C\|u\|^2,$$  

(37)

where $C$ denotes a generic positive constant independent of $\varepsilon$, $h$ and $u$. 
PROOF. We can show that

\[ b(u, u) \geq C\|u\|^2 + \frac{1}{2} J, \]

where \( J \) denotes the sum of the two line integrals.

Let \( \bar{u}(x) \) — the piecewise constant function on \( \Gamma \) defined by \( \bar{u}(x) = u_i \) if \( x \in \partial d_i \cap \Gamma \), i.e., \( \bar{u} \) is the restriction of \( P(u|_{\Omega_2}) \) on \( \Gamma \). we have

\[
J = \sum_{k=1}^{N_{\Gamma}} \int_{\Gamma^+_k} (\hat{a}\bar{u}^2 - a u^2) \cdot n ds,
\]

\( \Gamma^+_k \) — an edge of the mesh on \( \Gamma^+ \)

\( N_{\Gamma} \) — the number of edges on \( \Gamma \).

Note \( a \cdot n \geq \alpha_1 \) or \( \alpha_2 \) on \( \Gamma \).
On $\Gamma_k$, $\hat{a} \cdot n = \overline{a}_{i,j} = \sup_{x \in e_{i,j}} a \cdot n$ by (20), and $u$ and $\overline{u}$ are defined respectively by

$$u = u_i + \frac{|x - x_i|}{|x_j - x_i|} (u_j - u_i),$$

$$\text{and } \overline{u} = \begin{cases} u_i, & x_i \leq x < (x_i + x_j)/2, \\ u_j, & (x_i + x_j)/2 < x \leq x_j. \end{cases}$$
Parametrizing $e_{i,j}$ as $\{s : 0 \leq s \leq |e_{i,j}|\}$, we have

$$J_{i,j} := \int_0^{|e_{i,j}|} (\bar{a}_{i,j} \bar{u}^2 - au^2 \cdot n)ds$$

$$= \bar{a}_{i,j} \int_0^{|e_{i,j}|} (\bar{u}^2 - u^2)ds + \int_0^{|e_{i,j}|} (a_{i,j} - a \cdot n)u^2ds$$

$$\geq \bar{a}_{i,j} \int_0^{|e_{i,j}|} (\bar{u}^2 - u^2)ds,$$

(40)

since $a_{i,j} \geq a \cdot n > 0$ on $e_{i,j}$. 
We can further show that
\[ \int_0^{\xi} (\bar{u}^2 - u^2) ds = \frac{1}{12} \left. \frac{d^2(u^2)}{ds^2} \right|_{s=\xi} |e_{i,j}|^3 = \frac{|e_{i,j}|}{6} (u_j - u_j)^2 \geq 0, \]

Therefore
\[ J = \sum_{e_{i,j} \subset \Gamma^+} J_{i,j} \geq 0. \]

This completes the proof. \( \square \)
Convergence
Lemma 5.1: Let Assumptions 2.1 be fulfilled. If $\beta \geq 2$ in (11) and $M_1 = M_2 = M$, a positive integer, then, for any element edge $e_{i,j} \in E_h^{(2)}$, there exists a positive integer $C$, independent of $h$, $u$ and $\varepsilon$, such that

$$
\int_{l_{i,j}} |f_u \cdot e_{i,j} - f_{i,j}| \, ds \leq \left\{
\begin{array}{ll}
C |l_{i,j}| hK_1, & e_{i,j} \subset \bar{\Omega}_2^{(1)} \cup \bar{\Omega}_2^{(2)}, \\
C |l_{i,j}| \frac{1}{M} \ln(\frac{1}{\varepsilon}), & e_{i,j} \subset \Omega_2^{(3)},
\end{array}\right.
$$

where $f_u$ and $f_{i,j}$ are defined in (2) and (17) respectively and

$$
K_1 = \max\{1, h^{-1} \varepsilon^{\beta/2M}, h^{-1} \varepsilon \ln \frac{1}{\varepsilon}\}. \quad (41)
$$
Convergence

Lemma 5.1: Let Assumptions 2.1 be fulfilled. If $\beta \geq 2$ in (11) and $M_1 = M_2 = M$, a positive integer, then, for any element edge $e_{i,j} \in E_h^{(2)}$, there exists a positive integer $C$, independent of $h$, $u$ and $\varepsilon$, such that

$$\int_{l_{i,j}} |\mathbf{f}_u \cdot e_{i,j} - f_{i,j}| ds \leq \begin{cases} C |l_{i,j}| h K_1, & e_{i,j} \subset \bar{\Omega}_2^{(1)} \cup \bar{\Omega}_2^{(2)}, \\ C |l_{i,j}| \frac{1}{M} \ln \left( \frac{1}{\varepsilon} \right), & e_{i,j} \subset \Omega_2^{(3)}, \end{cases}$$

where $\mathbf{f}_u$ and $f_{i,j}$ are defined in (2) and (17) respectively and

$$K_1 = \max\{1, h^{-1} \varepsilon^{\beta/2M}, h^{-1} \varepsilon \ln \frac{1}{\varepsilon}\}. \quad (41)$$

The proof is based on directly estimating the integral on three different types of edges.
Theorem 5.1: Let Assumptions 2.1 and 4.1 be fulfilled. If \( \beta \geq 3 \) in (11) and \( M_1 = M_2 = M \), a positive integer, then there exists a positive integer \( C \), independent of \( h, u \) and \( \varepsilon \), such that

\[
\| u_I - u_h \| \leq C h \left( K_1^{1/2} + K_2 \right),
\]  

where \( u_I \) and \( u_h \) denote respectively the \( U_h \)-interpolation of the solution \( u \) to Problem 2.1 and the solution to Problem 3.1, \( K_1 \) is defined in (41) and \( K_2 \) is defined as

\[
K_2 = \max\{ M^{1/2} K_1, M^{-1/2} \sqrt{\varepsilon \ln \frac{1}{\varepsilon}} \}. 
\]
PROOF. Sketch only. For any $v_h \in V_h$, 

$$
\begin{align*}
 a(u_h - u_I, v_h) &= a(u - u_I, v_h) + (Gu - P(Gu), v_h^{(2)})_{\Omega_2} \\
&- \int_{\Gamma_-} f_u \cdot n (v_h^{(2)} - v_h^{(1)}) \, ds \\
\leq & \left| \int_{\Omega_1} (f_u - f_{u_I}) \cdot \nabla v_h^{(1)} \, dx + (Gu - Gu_I, v_h^{(1)})_{\Omega_1} \right| \\
&+ \left| \int_{\Gamma_-} f_u \cdot n (v_h^{(2)} - v_h^{(1)}) \, ds \right| \\
&+ \left| \sum_{d \in D_h^{(2)}} v_h^{(2)} (f_u - f_{u_I}) \cdot n \, ds \right| + \left| (Gu - P(Gu), v_h^{(2)})_{\Omega_2} \right| \\
=: & \quad R_1 + R_2 + R_3 + R_4.
\end{align*}
$$

(44)
The terms $R_1$, $R_2$, $R_3$ and $R_4$ can be estimated as:

For the four terms we have (Lemma 5.1 is used here)

$$R_1 \leq Ch\|v_h^{(1)}\|_{\varepsilon,\Omega_1},$$
The terms $R_1, R_2, R_3$ and $R_4$ can be estimated as

For the four terms we have (Lemma 5.1 is used here)

$$R_1 \leq C h \|v_h^{(1)}\|_{\varepsilon, \Omega_1},$$

$$R_2 \leq C h^2 \|v_h\|_{\infty},$$
The terms $R_1$, $R_2$, $R_3$ and $R_4$ can be estimated as

For the four terms we have (Lemma 5.1 is used here)

\[ R_1 \leq Ch\|v_h^{(1)}\|_{\varepsilon,\Omega_1}, \]

\[ R_2 \leq Ch^2\|v_h\|_{\infty}, \]

\[ R_3 \leq ChK_2\|v\|_{h,\Omega_2}, \]
The terms $R_1$, $R_2$, $R_3$ and $R_4$ can be estimated as

For the four terms we have (Lemma 5.1 is used here)

\[ R_1 \leq Ch\|v_h^{(1)}\|_{\varepsilon, \Omega_1}, \]

\[ R_2 \leq Ch^2\|v_h\|_{\infty}, \]

\[ R_3 \leq ChK_2\|v\|_{h, \Omega_2}, \]

\[ R_4 \leq Ch\|v_h\|_{\infty}(\varepsilon \ln \frac{1}{\varepsilon} + \varepsilon^{\beta/2M}). \]
Finally, substituting these into (44) we have

\[ a(u_h - u_I, v_h) \leq C \left[ hK_2 ||v|| + h^2 K_1 ||v_h||_{\infty} \right], \]

Choosing \( v_h = P(u_h - u_I) \) and using (37) we obtain

\[ ||u_h - u_I||^2 \leq C \left[ hK_2 ||u_h - u_I|| + h^2 K_1 \right]. \]

This is of the form

\[ y^2 \leq CK_2 hy + CK_1 h^2 \text{ or } (y - \frac{1}{2} CK_2 h)^2 \leq CK_1 h^2 + \frac{(CK_2)^2}{4} h^2. \]

which reduces to

\[ y \leq h \sqrt{CK_1 + \frac{(CK_2)^2}{4}} + \frac{CK_2}{2} h \leq Ch(K_2 + K_1^{1/2}). \]
Replacing $y$ with $\|u_h - u_I\|$ we obtain

\[ \|u_h - u_I\| \leq Ch(K_1^{1/2} + K_2). \]

This completes the proof of the theorem. \qed
Corollary: Let the Assumptions in Theorem 5.1 be fulfilled. If $\varepsilon, h, M$ and $\beta$ are such that $h^{-1}\varepsilon^{\beta/2M} \leq O(1)$ and $h^{-1}\varepsilon \ln(1/\varepsilon) \leq O(1)$, then we have

$$\|u_I - u_h\| \leq ChM^{1/2}. \quad (45)$$

Furthermore, if choose $h^{-1}\varepsilon^{\beta/2M} = O(1)$, then

$$\|u_I - u_h\| \leq Ch\sqrt{\ln \frac{\varepsilon}{\ln h}}. \quad (46)$$

Note $h^{-1}\varepsilon^{\beta/2M} = 1 \implies M = \left\lceil \frac{\beta \ln \varepsilon}{2 \ln h} \right\rceil$.

($\beta$ can be chosen to be 2 according to Theorem 2.1 and Theorem 5.1.)
Remark: When \( M = 1 \) and \( \beta = 2 \), we have uniform convergence of order \( O(h) \). This is true because all the mesh points are outside the layers.

In this case the method does not resolv the layers.
Conclusions

- When $\varepsilon \ll h$, $\|u_I - u_h\|$ converges to zero at the rate of $h$ almost uniformly in $\varepsilon$. 
Conclusions

- When $\varepsilon << h$, $\|u_I - u_h\|$ converges to zero at the rate of $h$ almost uniformly in $\varepsilon$.

- A mixture of triangular and rectangular elements. (or a triangular mesh)
Conclusions

- When $\varepsilon \ll h$, $\|u_I - u_h\|$ converges to zero at the rate of $h$ almost uniformly in $\varepsilon$.

- A mixture of triangular and rectangular elements. (or a triangular mesh)

- Possibility of extending to singularly perturbed problems with non-rectangular, polygonal regions. (the a priori estimates corresponding to those in Assumption 2.1 need to be established on non-rectangular polygonal regions.)