AN EXPONENTIAL FITTING SCHEME FOR GENERAL CONVECTION-DIFFUSION EQUATIONS ON TETRAHEDRAL MESHES

R. D. LAZAROV AND L. T. ZIKATANOV

Abstract. This paper contains construction and analysis a finite element approximation for convection dominated diffusion problems with full coefficient matrix on general simplicial partitions in $\mathbb{R}^d$, $d \geq 2$. This construction is quite close to the scheme of Xu and Zikatanov [22] where a diagonal coefficient matrix has been considered. The scheme is of the class of exponentially fitted methods that does not use upwind or checking the flow direction. It is stable for sufficiently small discretization step-size assuming that the boundary value problem for the convection-diffusion equation is uniquely solvable. Further, it is shown that, under certain conditions on the mesh the scheme is monotone. Convergence of first order is derived under minimal smoothness of the solution.

1. Introduction

We consider the following convection-diffusion-reaction problem: Find $u = u(x)$ such that

$$\text{(1.1)} \qquad \begin{cases} Lu \equiv -\nabla \cdot (D\nabla u + \mathbf{b}u) + \gamma u &= f \quad \text{in } \ \Omega, \\ u &= 0 \quad \text{on } \ \Gamma_D, \\ -D(\nabla u + \mathbf{b}u) \cdot \mathbf{n} &= g \quad \text{on } \Gamma_N^{in}, \\ D\nabla u \cdot \mathbf{n} &= 0 \quad \text{on } \Gamma_N^{out}. \end{cases}$$

Here Ω is a bounded polygonal domain in $\mathbb{R}^d$, $d = 2, 3$, $D = D(x)$ is $d \times d$ symmetric, bounded and uniformly positive definite matrix in Ω, $\mathbf{b}^t = (b_1(x), \dots, b_d(x))$ is a given vector function, $\mathbf{n}$ is the unit outer vector normal to $\partial\Omega$, and f is a given source function. We have also used the notation ∇u for the gradient of a scalar function u and $\nabla \cdot \mathbf{b}$ for the divergence of a vector function $\mathbf{b}$ in $\mathbb{R}^d$. The boundary of Ω, $\partial\Omega$, is split into Dirichlet, Γ_D, and Neumann, Γ_N, parts. Further, the Neumann boundary is divided into two parts: $\Gamma_N = \Gamma_N^{in} \cup \Gamma_N^{out}$, where $\Gamma_N^{in} = \{x \in \Gamma_N : \mathbf{n}(x) \cdot \mathbf{b}(x) > 0\}$ and $\Gamma_N^{out} = \{x \in \Gamma_N : \mathbf{n}(x) \cdot \mathbf{b}(x) \leq 0\}$. We assume that Γ_D has positive surface measure. The case $D(x) = \epsilon I$, where I is the

Date: Begin: Aug 12, 2003; today is October 14, 2004.

1991 *Mathematics Subject Classification.* 65N30, 65N15.

The authors thank the Center of Applied Scientific Computing at Livermore National Laboratory, University of California, for the partial financial support for this work through contract B529214.

identity matrix in $\mathbb{R}^d$ and $\epsilon > 0$ is a small parameter, corresponds to the important and difficult class of isotropic singularly perturbed convection-diffusion problems.

Various generalizations have wide practical applications. For example, γu could be replaced by nonlinear reaction term $\gamma(u)$ or the linear convective flux $\mathbf{b}u$ could be replaced by a nonlinear advection flux $\mathbf{b}(u)$. Finally, u could be a vector function describing the concentration of various chemicals or biological components so that (1.1) is a system of equations coupled through the absorption/reaction term. Now γ is a matrix that models the chemical reactions or the biological interaction of the components. All these cases give rise to mathematical problems of convection dominated processes with possibly anisotropic diffusion.

Our study of numerical method for solving (1.1) is motivated by the fact that the above problem is the simplest model of transport and dispersion of a passive contaminant in porous media. If the pressure $p(x)$ in the aquifer is known (or already has been computed by solving a standard diffusion problem) then the pressure gradient forces the ground water to flow. The transport of a contaminant dissolved in the water, is described by the dispersion-reaction equation (1.1), where $u(x)$ represents the contaminant concentration, $\mathbf{b} = \mathbf{v} = A\nabla p$ is the Darcy velocity (up to a sign), A is the permeability of the porous media, γ is the biodegradation/absorption rate, and $D(x)$ is the diffusion-dispersion matrix given by

$$D(x) = k_d I + k_t \mathbf{b}\mathbf{b}^t/|\mathbf{b}| + k_l(|\mathbf{b}|I - \mathbf{b}\mathbf{b}^t/|\mathbf{b}|). \tag{1.2}$$

Here k_d, k_t, and k_l are coefficients of diffusion, transverse dispersion, and longitudinal dispersions, respectively (cf. [8]). In dispersive underground flows $k_t > k_l$ which implies that $D(x)$ is positive definite matrix, but possibly ill-conditioned. This problem exhibits all difficulties associated with this class: monotione solutions that are highly localized due to internal and boundary layers, material heterogeneities and orthotropy, complex geometry, etc.

Among the deficiencies of the standard finite element, finite volume, and finite difference approximations are loss of monotonicity, so that the numerical solution often exhibits nonphysical oscillations, loss of solvability of the resulting algebraic problem, poor local resolution, fast dissipation of the energy, etc. A.A. Samarskii was one of the first to encounter the difficulties that arise in the numerical solution of such problems. In the early 60-ies A.A. Samarskii addressed most of the issues for one-dimensional problems that resulted in a new scheme described in his monograph [14, Chapter 4].

In the past 40 years many special approximation techniques have been developed for multidimensional problems, for structured and unstructured grids, for general second order elliptic operators, etc. These techniques include monotone and upwind finite difference, finite volume, and finite element methods (e.g. [2], [6], [9], [14], [15], [16], [17], and [21]), streamline diffusion stabilization of the finite element method (e.g. [1], [3], [4], [11], and [12]), and special functional spaces setting (e.g. [5] and [18]). For more information regarding numerical methods and analytical techniques in solving and studying convection-diffusion equations, especially convection dominated problems, we refer to the monograph of Ross, Stynes and Tobiska [13].

On a continuous level many convection-diffusion satisfy maximum principle. This is a desirable property of the solution of the resulting discrete problem as well. Scheme that satisfies maximum principle is often called *monotone scheme*. Among the several aforementioned schemes, upwind schemes are often monotone provided that the coefficient matrix $D(x)$ is diagonal. In the recent works [15], [16] Samarskii and his co-workers were able to derive monotone schemes on rectangular meshes for the problem (1.1) when $D(x)$ is a full matrix. These schemes are second order accurate on uniform meshes and solution in C^3.

The idea of construction of monotone schemes for singularly perturbed convection-diffusion problem goes back to the work by Scharfetter and Gummel [19], where the monotonicity has been a very desired property in numerical semiconductor device modeling. Exponentially fitted scheme for a general convection-diffusion problem with diagonal matrix $D(x)$ on an arbitrary simplicial mesh was derived and studied in [22], [24], and successfully used in [25] for semiconductor device simulation. Under mild conditions on the mesh, the partition has to satisfy certain angle condition, it has been shown that the scheme is monotone. Further, in [22] it was proved that the scheme converges with first order provided that the solution $u \in W^{1,p}$ and the flux $D(x)\nabla u + \mathbf{b}(x)u \in (W^{1,p})^d$ for $p > d$. Note, that these are very mild conditions on smoothness of the solution of problem (1.1).

The aim of this note is to construct an exponentially fitted finite element approximation of (1.1) on general simplicial partitions, for symmetric positive definite matrices $D(x)$, and for arbitrary vector-functions $\mathbf{b}$. The proposed scheme is a generalization of the discretization derived in [22], [24] for problems with diagonal matrices $D(x)$. Important role in the construction and the analysis plays the expansion of a constant over each element vector-flux using the lowest order Nedeleč basis for simplicial finite elements. This allows to present the bilinear form through differences of the vertex values of the test functions and the exponentially weighted local solution. This representation ensures the consistence of the method.

The scheme has several interesting features. It is a finite element scheme with a standard variational formulation (but with a modified bilinear form); it does not use explicitly the standard upwind techniques, such as checking the flow directions; it can be applied to very general unstructured grid in any spatial dimension. It would be difficult to expect that in such generality the scheme will be monotone. Nevertheless, we were able to find conditions that involve the geometry of the finite elements in a metric associated with D so that the scheme is monotone. Further, for sufficiently small step-size of the finite element partition we prove existence and uniqueness of the solution of the discrete problem by using a fundamental result of Schatz [20].

The paper is organized as follows. In Section 2 we introduce the necessary notations for Sobolev spaces, finite element partition and the discrete space. Section 3 contains the main results of the paper. In Subsection 3.1 we present the rationale used in derivation of exponentially fitting finite element scheme. An important concept here is the edge based interpolation of the total flux that uses an ordinary differential equation along the edge. In Subsection 3.2 we present the scheme itself as a consequence of this special interpolation. The main result here is contained in Lemma 3.1 where certain properties of the discrete bilinear form are obtained. Finally, in Subsection 3.3 we prove the stability of the scheme for sufficiently small mesh-size and derive an estimate for the error under minimal smoothness of the solution.

2. Notations

In this section, we introduce the necessary notation and describe some basic properties of finite element partitions and finite element spaces.

We denote by $L^p(K)$, $1 \leq p \leq \infty$ the space of p-integrable real-valued functions over $K \subset \Omega$ (with the usual modification for $p = \infty$), by $(\cdot,\cdot)_K$ and $\|\cdot\|_K$, respectively, the inner product and the norm in $L^2(K)$. Further $|\cdot|_{1,p,K}$ and $\|\cdot\|_{1,p,K}$, respectively denote the semi-norm and norm of the Sobolev space $W^{1,p}(K)$. For $p = 2$ we use $H^1(K) := W^{1,2}(K)$ and if $K = \Omega$ often we suppress the index K so that $(\cdot,\cdot)_\Omega := (\cdot,\cdot)$ and $\|\cdot\|_\Omega := \|\cdot\|$, and $\|\cdot\|_{1,\Omega} := \|\cdot\|_1$. Further, we use the Hilbert space

$$H^1_D(\Omega) = \{v \in H^1(\Omega) : \ v|_{\Gamma_D} = 0\}.$$

We introduce the bilinear form $a(\cdot,\cdot)$ defined on $H^1_D(\Omega) \times H^1_D(\Omega)$:

$$a(u,v) := (D\nabla u + \mathbf{b}u, \nabla v) + (\gamma u, v) - \int_{\Gamma_N^{out}} \mathbf{b}\cdot\mathbf{n}\, u\, v\, ds. \tag{2.1}$$

Then (1.1) has the following weak form: Find $u \in H^1_D(\Omega)$ such that

$$a(u,v) = F(v) := (f,v) + \int_{\Gamma_N^{in}} gv\, ds \quad \text{for all } v \in H^1_D(\Omega). \tag{2.2}$$

Further in the paper we assume that the following inf-sup condition is valid: there is a constant $c_0 > 0$, such that

$$\sup_{v \in H^1_D(\Omega)} \frac{a(w,v)}{\|v\|_1} \geq c_0 \|w\|_1, \quad \forall w \in H^1_D(\Omega). \tag{2.3}$$

We shall also assume that the bilinear form $a(w,v)$ is bounded on $H^1_D(\Omega)$ and the lienar form $F(v)$ is continuous in $H^1_D(\Omega)$. Then the above problem has unique solution (cf. [10]).

Remark 2.1. A sufficient condition for (2.3) and continuity of $a(u,v)$ and $F(v)$ are, for example, $\gamma(x) + 0.5\nabla\cdot\mathbf{b}(x) \geq 0$ for all $x \in \Omega$, boundedness of the coefficients $D(x)$, $\mathbf{b}(x)$, and $\gamma(x)$ in Ω.

Let $\mathcal{T}_h$ be a family of simplicial finite element triangulations of Ω that are shape regular and satisfy the usual conditions (see [7, Chapter 2]). For simplicity of the exposition, we assume that the triangulation covers Ω exactly. Associated with each $\mathcal{T}_h$, let $V_h \subset H^1_D(\Omega)$ be the finite element space of piece-wise linear functions.

By $v_I \in V_h$ we denote the standard finite element Lagrange interpolant which assumes the values of $v \in C^0$ at the vertexes in the partition $\mathcal{T}_h$.

Given $T \in \mathcal{T}_h$, we introduce the following notation. By q_j, $j = 1, \dots, 4$ we denote the vertices of T, E is the edge connecting two vertices q_i and q_j, $\delta_{\mathrm{E}}\phi = \phi(q_i) - \phi(q_j)$ for any continuous function ϕ on E, and $\tau_{\mathrm{E}} = \delta_{\mathrm{E}}\, x = q_i - q_j$ is a directional vector of E (not assumed unitary).

3. Exponential fitting scheme for general convection-diffusion equations

3.1. Preliminaries.

Introduce a notation for the scaled flux

$$J(u) = \nabla u + \boldsymbol{\beta}(x) u, \qquad \boldsymbol{\beta} = D^{-1}\mathbf{b}. \tag{3.1}$$

We will assume further that $J(u) \in [W^{1,p}(\Omega)]^d$, $p > n$, D, $D^{-1} \in [W^{1,\infty}(\Omega)]^{d\times d}$ and $\mathbf{b} \in W^{1,\infty}(\Omega)$. These assumptions on the coefficients smoothness can be relaxed to hold element-wise (i.e. for each $T \in \mathcal{T}_h$) and the considerations below will still hold with changes of some of the norms used in the error estimate to be taken element-wise as well.

The basic idea which we use in the construction of the exponentially fitted scheme is to approximate the flux vector $J(u)$ with a constant vector field $J_T(u)$ on each element T of the partition $\mathcal{T}_h$. Apparently, if $J_T(u)$ is a constant on each simplex, then we can expand it using the Nedeleč basis as follows:

$$J_T = \sum_{E\in T} J_T \cdot \tau_{\mathrm{E}}\, \varphi_{\mathrm{E}}(x).$$

Here φ_{E} are the Nedeleč basis functions, which in terms of the barycentric coordinates λ_i are given by

$$\varphi_{\mathrm{E}} := \lambda_i \nabla\lambda_j - \lambda_j \nabla\lambda_i, \quad \mathrm{E} = (q_i, q_j).$$

The goal then is to write out $J_T(u) \cdot \tau_{\mathrm{E}}$ in terms of u, for all edges E and thus determine the approximation. To find the moments of the tangential flux, we use the same technique as in [22]. Let $u \in H_0^1(\Omega) \cap C^0(\bar{\Omega})$. Consider an edge $\mathrm{E} \subset T$. Taking the Euclidean inner product with τ_{E} we obtain

$$(\nabla u \cdot \tau_{\mathrm{E}}) + (\boldsymbol{\beta} \cdot \tau_{\mathrm{E}})u = (J(u) \cdot \tau_{\mathrm{E}}).$$

A change of variables in this ordinary differential equation then gives:

$$e^{-\psi_{\mathrm{E}}}\partial_{\mathrm{E}}(e^{\psi_{\mathrm{E}}}u) = \frac{1}{|\tau_{\mathrm{E}}|}(J(u)\cdot\tau_{\mathrm{E}}), \quad \text{where} \quad \partial_{\mathrm{E}}\psi_{\mathrm{E}} = \frac{1}{|\tau_{\mathrm{E}}|}(\boldsymbol{\beta}\cdot\tau_{\mathrm{E}}) \tag{3.2}$$

and $\partial_{\mathrm{E}} v := \nabla v \cdot \tau_{\mathrm{E}}/|\tau_{\mathrm{E}}|$ is the directional derivative along the edge E. After integration over E we obtain that

$$\delta_{\mathrm{E}}(e^{\psi_{\mathrm{E}}}u) = \frac{1}{|\tau_{\mathrm{E}}|}\int_{\mathrm{E}} e^{\psi_{\mathrm{E}}}(J(u)\cdot\tau_{\mathrm{E}})ds.$$

Let $\mathcal{H}_{\mathrm{E}}(\boldsymbol{\beta})$ be the harmonic average of $e^{\psi_{\mathrm{E}}}$ over E defined as follows:

$$\mathcal{H}_{\mathrm{E}}(\boldsymbol{\beta}) = \left[\frac{1}{|\tau_{\mathrm{E}}|}\int_{\mathrm{E}} e^{\psi_{\mathrm{E}}} ds\right]^{-1}. \tag{3.3}$$

The constant approximation J_T is then obtained by using the mean value theorem $J^* \cdot \tau_{\mathrm{E}} \int_{\mathrm{E}} e^{\psi_{\mathrm{E}}}\, ds = \int_{\mathrm{E}} J \cdot \tau_{\mathrm{E}} e^{\psi_{\mathrm{E}}}\, ds$, and the definition then is

$$J_T(u)\cdot \tau_{\mathrm{E}} := J^* \cdot \tau_{\mathrm{E}} = \mathcal{H}_{\mathrm{E}}(\boldsymbol{\beta})\delta_{\mathrm{E}}(e^{\psi_{\mathrm{E}}}u).$$

3.2. **Discrete problem.** We now have all the ingredients needed to define the discrete approximation to (2.2). Based on the above considerations, we shall define two approximate bilinear forms. The first one is used in the formulation of the discrete problem and the second is used in an intermediate step needed to prove the error estimate.

On a fixed element $T \subset \mathcal{T}_h$, we first introduce

$$a_{h,T}(u_h, v_h) = \sum_{\mathrm{E}\subset T} \omega_{\mathrm{E}}^T(D)\mathcal{H}_{\mathrm{E}}(\boldsymbol{\beta})\delta_{\mathrm{E}}(e^{\psi_{\mathrm{E}}}u_h)\delta_{\mathrm{E}}v_h, \tag{3.4}$$

where

$$\omega_{\mathrm{E}}^T(D) = -\int_T D\nabla\lambda_i \cdot \nabla\lambda_j\, dx, \quad \mathrm{E} = (q_i, q_j).$$

Note, that $\omega_{\mathrm{E}}^T(D)$ give the element stiffness matrix for the diffusion part of the differential equation, $-\nabla\cdot(D(x)\nabla u)$.

Next, we use the expansion via the Nedeleč basis functions, to define

$$b_{h,T}(u_h, v_h) = \sum_{\mathrm{E}\subset T} \mathcal{H}_{\mathrm{E}}(\boldsymbol{\beta})\delta_{\mathrm{E}}(e^{\psi_{\mathrm{E}}}u_h)\int_T D\varphi_{\mathrm{E}}\cdot\nabla v_h\, dx. \tag{3.5}$$

The global bilinear form is then obtained by summing over all elements of the triangulation the local forms (3.4) and adding the contributions from the boundary Γ_N^{out}, as follows:

$$a_h(u_h, v_h) = \sum_{T\in\mathcal{T}_h} a_{h,T}(u_h, v_h) + \int_\Omega \gamma u_h v_h dx - \sum_{E\subset\Gamma_N^{out}}\int_E \mathbf{b}\cdot\mathbf{n}\, u_h v_h ds. \tag{3.6}$$

Finally, the finite element approximation of the problem (1.1) reads as follows: Find $u_h \in V_h$ such that

$$a_h(u_h, v_h) = F(v_h), \quad \text{for all} \quad v_h \in V_h. \tag{3.7}$$

The following lemma is the main tool used in the analysis of the above scheme.

Lemma 3.1. *The following relations hold for any $v_h \in V_h$:*

1. *If $w \in C(\overline{T})$ then*

$$b_{h,T}(w_I, v_h) = \sum_{E\subset T}\left[\frac{\mathcal{H}_E(\boldsymbol{\beta})}{|\tau_E|}\int_E e^{\psi_E}J(w)\cdot\tau_E ds\right]\int_T D\varphi_E\cdot\nabla v_h; \tag{3.8}$$

2. *If J_T is a constant vector on T, then for any $v_h \in V_h$*

$$\sum_{E\subset T} J_T \cdot \tau_E \int_T D\varphi_E \cdot \nabla v_h \, dx = \sum_{E\subset T} \omega_E^T(D) J_T \cdot \tau_E \delta_E v_h; \tag{3.9}$$

3. *If $w \in C(\overline{T})$ and $J(w) \in [W^{1,p}(T)]^n$, $p > d$, then the following inequality holds for every $v_h \in V_h$ and $T \in \mathcal{T}_h$*

$$|a_T(w, v_h) - a_{h,T}(w_I, v_h)| \le Ch|J(w)|_{1,p,T} \|v_h\|_{1,T}. \tag{3.10}$$

where

$$a_T(w, v_h) = \int_T (D\nabla w + \mathbf{b}w) \cdot \nabla v_h \, dx.$$

Proof. The proof of 1. follows directly from the derivation.

The proof of 2. can be done as follows: Consider $\Phi := J_T \cdot x$ for $x \in T$ (here x is treated as a vector). It is obvious that Φ is linear and that $\nabla\Phi \cdot \tau_{\mathrm{E}} = J_T \cdot \tau_{\mathrm{E}}$. A simple computation, using the fact that the Nedelec̆ projection

$$\Pi_{\mathcal{N}} J_T = \sum_{\mathrm{E}\in T} (J_T \cdot \tau_{\mathrm{E}}) \varphi_{\mathrm{E}}$$

satisfies the commutativity property $\Pi_{\mathcal{N}} \nabla\Phi = \nabla\Phi_I = \nabla\Phi$, completes the proof of 2.

To prove 3. we use (3.8) and split the difference in the following way

$$a_T(w, v_h) - a_{h,T}(w_I, v_h) = \mathcal{E}_1(J(w), v_h) + \mathcal{E}_2(J(w), v_h) \tag{3.11}$$

where

$$\mathcal{E}_1(J(w), v_h) = a_T(w, v_h) - b_{h,T}(w_I, v_h),$$

and

$$\mathcal{E}_2(J(w), v_h) = b_{h,T}(w_I, v_h) - a_{h,T}(w_I, v_h).$$

From the relations and item 1. we can expand the forms $a_T(w, v_h)$, $a_{h,T}(w_I, v_h)$ and $b_{h,T}(w_I, v_h)$ to get that

$$\begin{aligned} \mathcal{E}_1(J(w), v_h) &= \int_T J(w) \cdot \nabla v_h dx \\ &\quad - \sum_{\mathrm{E}\subset T} \left[\frac{\mathcal{H}_{\mathrm{E}}(\boldsymbol{\beta})}{|\tau_{\mathrm{E}}|} \int_{\mathrm{E}} e^{\psi_{\mathrm{E}}} J(w) \cdot \tau_{\mathrm{E}} ds \right] \int_T D\varphi_E \cdot \nabla v_h \, dx \end{aligned} \tag{3.12}$$

and

$$\begin{aligned} \mathcal{E}_2(J(w), v_h) &= \sum_{\mathrm{E}\subset T} \left[\frac{\mathcal{H}_{\mathrm{E}}(\boldsymbol{\beta})}{|\tau_{\mathrm{E}}|} \int_{\mathrm{E}} e^{\psi_{\mathrm{E}}} J(w) \cdot \tau_{\mathrm{E}} ds \right] \int_T D\varphi_E \cdot \nabla v_h \, dx \\ &\quad - \sum_{\mathrm{E}\subset T} \omega_{\mathrm{E}}^T(D) \left[\frac{\mathcal{H}_{\mathrm{E}}(\boldsymbol{\beta})}{|\tau_{\mathrm{E}}|} \int_{\mathrm{E}} e^{\psi_{\mathrm{E}}} J(w) \cdot \tau_{\mathrm{E}} ds \right] \delta_{\mathrm{E}} v_h. \end{aligned} \tag{3.13}$$

A change of variable from the standard reference element $\widehat{T}$ (of unit size) to T: $x = B\hat{x} + b_0$ and $w(x) = \hat{w}(\hat{x})$ gives the following scaled bilinear forms

$$\begin{aligned}
a_T(w, v_h) &= |\mathrm{det}B| \int_{\hat{T}} (B^{-1}\hat{D}\widehat{J(w)} \cdot \nabla \hat{v}_h)\, d\hat{x} \\
a_{h,T}(w, v_h) &= \sum_{\hat{\mathrm{E}}\subset\hat{T}} \omega_{\mathrm{E}}^T(D) \left[\frac{\mathcal{H}_{\mathrm{E}}(\boldsymbol{\beta})}{|\tau_{\hat{\mathrm{E}}}|} \int_{\hat{\mathrm{E}}} e^{\widehat{\psi_{\mathrm{E}}}} (\widehat{J(w)} \cdot \tau_{\hat{\mathrm{E}}}) d\hat{s} \right] \delta_{\mathrm{E}} \hat{v}_h \\
b_{T,h}(w, v_h) &= |\mathrm{det}B| \sum_{\hat{\mathrm{E}}\subset\hat{T}} \left[\frac{\mathcal{H}_{\mathrm{E}}(\boldsymbol{\beta})}{|\tau_{\hat{\mathrm{E}}}|} \int_{\hat{\mathrm{E}}} e^{\widehat{\psi_{\mathrm{E}}}} (\widehat{J(w)} \cdot \tau_{\hat{\mathrm{E}}}) d\hat{s} \right] \\
&\qquad \times \int_{\hat{T}} (B^{-1}\hat{D}\hat{\varphi}_{\hat{\mathrm{E}}} \cdot \nabla \hat{v}_h)\, d\hat{x}.
\end{aligned}$$

By our assumptions on the smoothness of $J(w)$, the corresponding error functionals $\hat{\mathcal{E}}_i(\widehat{J(w)}, \hat{v}_h)$, $i = 1, 2$, can be appropriately bounded:

$$\hat{\mathcal{E}}_i(\widehat{J(w)}, \hat{v}_h) \leq C_i \|\widehat{J(w)}\|_{0,\infty,\hat{T}} \|\hat{v}_h\|_{1,\hat{T}}, \tag{3.14}$$

where C_i might depend on D, but do not depend on $\boldsymbol{\beta}$. By the Sobolev inequality we have that

$$\|\widehat{J(w)}\|_{0,\infty,\hat{T}} \leq C \|\widehat{J(w)}\|_{1,p,\hat{T}}, \ p > d.$$

We observe that from (3.8) and (3.9), it follows that $\mathcal{E}_i(J(w), v_h) = 0$ if $J(w)$ is a constant on T. By applying the Bramble-Hilbert Lemma on $\widehat{T}$, and scaling back to T we obtain the desired result:

$$|\mathcal{E}_i(J(w), v_h)| \leq Ch|J(w)|_{1,p,T}|v_h|_{1,T}, \quad i = 1, 2. \tag{3.15}$$

□

3.3. **Solvability of the discrete problem and error estimate.** In this paragraph we state two lemmas related to the solvability of the problem and then a result related to the error bound. The first result, the proof of which follows straightforward from the definition, is related to the monotonicity of the scheme (i.e. discrete maximum principle). This amounts to a condition on the geometry of the mesh associated with the matrix D.

Lemma 3.2. *The stiffness matrix corresponding to the bilinear form (3.7) is an M-matrix for any continuous function $\boldsymbol{\beta}$ if and only if the following inequality holds for all edges E in the triangulation*

$$\sum_{T\supset E} \omega_E^T \geq 0 \tag{3.16}$$

One may check out easily that if D is a constant matrix and $d = 2$ (two spatial dimensions) this is equivalent to the statement that the triangulation is Delaunay in the metric introduced by D. Namely, instead of Euclidean inner product $\mathbf{b} \cdot \mathbf{n}$ of the vectors $\mathbf{b}$ and $\mathbf{n}$ in $\mathbb{R}^d$ we need to use the inner product $D\mathbf{b} \cdot \mathbf{n}$ (recall that D is a symmetric and positive definite matrix). In this case the global stiffness matrix of the finite element system is nonsingular and therefore the scheme (3.7) has unique solution.

Next result is about solvability of the discrete problem for sufficiently small characteristic mesh size h. Let us first consider an auxiliary discrete problem with $a(\cdot,\cdot)$ in place of $a_h(\cdot,\cdot)$ in (3.7). The latter problem is solvable, and a convincing (but not rigorous) argument to prove this claim is that the convection term is one order lower than the diffusion term and hence, decreasing h will make the diffusion term dominating and the problem weakly coercive. Some more detailed considerations and a rigorous arguments can be found in Schatz [20] or Xu [23].

Lemma 3.3. *For sufficiently small h the following inf-sup condition holds*

$$\sup_{v_h\in V_h}\frac{a_h(w_h,v_h)}{\|v_h\|_{1,\Omega}}\geq c_1\|w_h\|_{1,\Omega}\quad\forall w_h\in V_h \tag{3.17}$$

with a constant $c_1>0$ independent of mesh-size h.

Proof. As we have pointed out, when the original bilinear form $a(\cdot,\cdot)$ is used in (3.7), the discrete problem is uniquely solvable (for sufficiently small h). Hence, there exists a constant c_2 such that

$$\sup_{v_h\in V_h}\frac{a(w_h,v_h)}{\|v_h\|_{1,\Omega}}\geq c_2\|w_h\|_{1,\Omega},\quad\forall w_h\in V_h. \tag{3.18}$$

Let $v_h, w_h\in V_h$. Then obviously

$$a_h(w_h,v_h)=a(w_h,v_h)+\left[a_h(w_h,v_h)-a(w_h,v_h)\right].$$

The first term is estimated using the condition (3.18). To estimate the second term we use (3.10) from lemma 3.1, sum up over all T and apply the Schwarz inequality to obtain that

$$|a(w_h,v_h)-a_h(w_h,v_h)|\leq Ch\left\{\sum_{T\in\mathcal{T}_h}|J(w_h)|^2_{1,p,T}\right\}^{1/2}\|v_h\|_{1,\Omega}.$$

Observing that $|w_h|_{2,T}=0$ for any $w_h\in V_h$, $T\in\mathcal{T}_h$ we get

$$|J(w_h)|_{1,p,T}\leq C\|\boldsymbol{\beta}\|_{1,\infty,T}\|w_h\|_{1,T}.$$

Summing over all the elements of the partition we have

$$|a(w_h,v_h)-a_h(w_h,v_h)|\leq C\,h\max_{T\in\mathcal{T}_h}\|\boldsymbol{\beta}\|_{1,\infty,T}\|w_h\|_{1,\Omega}\|v_h\|_{1,\Omega}, \tag{3.19}$$

and for h satisfying

$$h\leq h_0\equiv C\left[\max_{T\in\mathcal{T}_h}\|\boldsymbol{\beta}\|_{1,\infty,T}\right]^{-1}$$

the discrete problem has a unique solution. □

As a consequence of Lemma 3.1, Lemma 3.3 we get the following convergence result.

Theorem 3.4. *Let u be the solution of the problem (2.2). Assume that for all $T \in \mathcal{T}_h$, $D \in \left(W^{1,\infty}(T)\right)^{d\times d}$, $\boldsymbol{\beta} \in [W^{1,\infty}(T)]^d$, $u \in W^{1,p}(T)$, $\gamma \in C(\overline{T})$, and $J(u) \equiv \nabla u + \boldsymbol{\beta}(x)u \in \left(W^{1,p}(T)\right)^d$, $p > d$. Then for sufficiently small h, the following estimate holds:*

$$\|u_I - u_h\|_{1,\Omega} \leq Ch \left\{ \sum_{T\in\mathcal{T}_h} |J(u)|^2_{1,p,T} + \sum_{T\in\mathcal{T}_h} |u|^2_{1,p,T} \right\}^{1/2} \tag{3.20}$$

Remark 3.5. There are also other possibilities for expressing the flux $J(u)$. For example, instead of the flux (3.1) one can write

$$\begin{aligned} D(x)\nabla u + \mathbf{b}u &= D(x)\alpha(x)^{-1}\left(\alpha(x)\nabla(u) + \alpha(x)D^{-1}(x)\mathbf{b}u)\right) \\ &:= \widetilde{D}(x)\left(\alpha(x)\nabla(u) + \boldsymbol{\beta}u)\right), \end{aligned}$$

where

$$\boldsymbol{\beta} = \alpha(x)D^{-1}(x)\mathbf{b}, \quad \widetilde{D}(x) = D(x)\alpha(x)^{-1}$$

with $\alpha(x)$ a suitable positive function (or a positive diagonal matrix). Then define

$$J(u) = \alpha(x)\nabla(u) + \boldsymbol{\beta}u.$$

For such a choice of $J(u)$ the derivation and the analysis of an exponentially fitting scheme are essentially the same with some changes occurring in the harmonic averages used to define the discrete problem.

For example, one may choose

$$\alpha(x) = (\lambda_{min}(D(x)) + \lambda_{max}(D(x))/2,$$

where $\lambda_{min}(D(x))$ and $\lambda_{max}(D(x))$ are the minimum and maximum eigenvalues of $D(x)$. For such choice $\widetilde{D}^{-1} = \alpha D^{-1}$ is better conditioned. For example, problem (1.1), (1.2) with data such as $k_d = 0.0001$, $k_t = 21$, and $k_l = 2.1$, used in [6], might require such modification.

Remark 3.6. As we have pointed out in the introduction, in many cases $D(x)$ takes the form (1.2). Then introducing the orthogonal projection $\pi_{\mathbf{b}} = \mathbf{b}\mathbf{b}^t/|\mathbf{b}|^2$ along the vector $\mathbf{b}(x)$ we can rewrite $D(x)$ in the form

$$D(x) = k_d I + k_t|\mathbf{b}|\pi_{\mathbf{b}} + k_l|\mathbf{b}|(I - \pi_{\mathbf{b}}).$$

Now one easily finds that $D^{-1}\mathbf{b} = (k_d + k_t|\mathbf{b}|)^{-1}\mathbf{b}$, i.e. the evaluation of $D^{-1}\mathbf{b}$ is just a multiplication of $\mathbf{b}$ by a scalar.

Acknowledgment.

This paper is dedicated to the 85-th birthday of Academician Alexander Andreevich Samarskii – a pioneer in numerical analysis and computational mathematics and computational physics. Under his longstanding leadership the Keldysh Institute of Applied Mathematics at the Russian Academy of Sciences and the Department of Computational Mathematics and Cybernetics at Moscow State University

had played fundamental role in establishing *Mathematical Modeling* as a dynamic branch of contemporary mathematics. We are pleased to acknowledge the vision, the dedication, and the seminal contributions of Acad. A.A. Samarskii to this important research area, which has become a the main link between science and engineering on the basis of mathematics and computer information technologies.

References

[1] R. Bank, J. Bürger, W. Fichtner, and R. Smith. Some up-winding techniques for finite element approximations of convection diffusion equations. *Numer. Math.*, 58:185–202, 1974.

[2] R. Bank and D. Rose. Some error estimates for the box method. *SIAM J. Numer. Anal.*, 24:777–787, 1987.

[3] F Brezzi and A. Russo. Choosing bubbles for advection-diffusion problems. *Math. Models Methods Appl. Sci.*, 32(1-3):571–587, 1994.

[4] A. Brooks and T.H. Hughes. Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comp. Meth. in Appl. Mech. Eng.*, 32:199–259, 1982.

[5] C. Canuto and A. Tabacco. An anisotropic functional setting for convection-diffusion problems. *East-West J. Numer. Math.*, 9(3):199–231, 2001.

[6] C. Carstensen, R.D. Lazarov, and S.T. Tomov. Explicit and averaging a posteriori error estimates for adaptive finite volume methods. *SIAM J. Numer. Anal.*, 42(6):to appear, 2004.

[7] P. Ciarlet. *The Finite Element Method for Elliptic Problems*, volume 4 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1978. Studies in Mathematics and its Applications, Vol. 4.

[8] G. Dagan. *Flow and Transport in Porous Formations*. Springer-Verlag, Berlin-Heidelberg, 1989.

[9] L. Durlofsky, B. Engquist, and S. Osher. Triangle based adaptive stencils for the solution of hyperbolic conservation laws. *J. Compt. Phys.*, 98(1):199–259, 1992.

[10] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*, volume 224 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 1983.

[11] T.H. Hughes. Greens functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubles and the origins of stabilized methods. *Comp. Meth. in Appl. Mech. Eng.*, 127:387–401, 1995.

[12] C. Johnson. *Numerical solution of partial differential equations by the finite element method.* Cambridge University Press, Cambridge, 1987.

[13] H.-O. Ross, M. Stynes, and L. Tobiska. *Numerical Methods for Singularly Perturbed Differential Equations.* Studies in Mathematics and its Applications. Springer, 1996.

[14] A.A. Samarskii. *Theory of Difference Schemes.* Nauka, Moscow, 1977.

[15] A.A. Samarskii, P.P. Matus, V.I. Mazhukin, and I.E. Mozolevski. Monotone difference schemes for equations with mixed derivatives. *Computers & mathematics with applications*, 44(3-4):501–510, 2002.

[16] A.A. Samarskii and P.N. Vabishchevich. Monotone difference schemes for the transport equation. *Doklady Academii Nauk, Russia*, 361(1):21–23, 1998.

[17] A.A. Samarskii and P.N. Vabishchevich. Monotone difference schemes on triangular grids. *Doklady Academii Nauk, Russia*, 371(6):742–746, 2000.

[18] G. Sangalli. Analysis of the advection-diffusion operator. *Numer. Math.*, 97(5):779–796, 2004.

[19] D. Scharfetter and H. Gummel. Large-signal analysis of a silicon read diod oscilator. *IEEE Trans. Electron Devices*, ED-16(205):959–962, 1969.

[20] A. H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Math. Comp.*, 28(205):952–962, 1974.

[21] M. Tabata. A finite element approximation corresponding to upwind finite differencing. *mem. Numer. math..*, 4:47–63, 1977.

[22] J. Xu and L. Zikatanov. A monotone finite element scheme for convection-diffusion equations. *Math. Comp.*, 68(228):1429–1446, 1999.

[23] Jinchao Xu. Two-grid discretization techniques for linear and nonlinear PDEs. *SIAM J. Numer. Anal.*, 33(5):1759–1777, 1996.

[24] L. T. Zikatanov. A modified Galerkin-Petrov method for modeling semiconductor devices on the basis of the finite element method. *Mat. Model.*, 4(5):85–99, 1992. In Russian.

[25] L. T. Zikatanov and M. S. Kaschiev. Finite element method for semiconductor device modeling. Technical Report R11-91-371, Communications of the Joint Institute for Nuclear Research, Dubna, 1991. In Russian.

Department of Mathematics, Texas A & M University, College Station, TX 77843, U.S.A.

E-mail address: lazarov@math.tamu.edu

Department of Mathematics, Pennsylvania State University, University Park, PA 16802, U.S.A.

E-mail address: ltz@math.psu.edu