A Petrov-Galerkin Mixed Finite Element Method with Exponential Fitting

R. R. P. van Nooyen*
CWI, Centre for Mathematics and Information Sciences,
P. O. Box 4079, 1009 AB Amsterdam, Holland

Received 3 December 1991; revised manuscript received 1 May 1993

We discuss a Petrov-Galerkin mixed finite element formulation of the semiconductor continuity equations on a rectangular domain. We give error estimates for equations that are in principle degenerate in the singularly perturbed case. We give arguments that indicate that the method is also effective in the singularly perturbed case. We develop a discretization that gives a higher-order accurate solution for use in an *a posteriori* error estimator. © 1995 John Wiley & Sons, Inc.

I. INTRODUCTION

The use of a form of exponential fitting for the semiconductor continuity equation is suggested by the success of the Scharfetter–Gummel discretization [1] in one dimension and variations on that discretization in two dimensions. Numerous derivations of Scharfetter–Gummel type discretizations are given in the literature, for instance by Selberherr [2], Markowich [3], Bank et al. [4], Brezzi et al. [5], and others. This article extends a one-dimensional exponential fitting technique, discussed by Hemker [6], to the two-dimensional context.

In Section II we introduce a model equation for the semiconductor continuity equations and several bilinear forms, related to the coefficients in this equation. In Sections III and IV we treat the discretization. In Section V we collect some technical results, and in Section VI we derive two error estimates. These error estimates are based on the techniques used by Douglas and Roberts [7]. The proofs in Section VI take all characteristics of our special discrete system into account, in particular the quadrature rule for the approximation of certain integrals in the discrete system. Note that the error estimates in Section VI are degenerate if the problem is singularly perturbed, i.e., if the convection dominates in the problem. On the other hand, an indication for good behavior of the method for singular problems is that—for constant coefficients—it reproduces the solution $C \exp(-\beta_1 x_1 - \beta_2 x_2)$ exactly. In Section IX, we develop an *a posteriori* error estimator, and in the last section we discuss our findings.

*Present address: Fac TWI/Gebouw ET, Ye Etage, Vakgroep Toegepaste Analyse, Delft University of Technology, Mekelweg 4, Delft, Holland.

CCC 0749-159X/95/050501-24

II. EQUATION

We consider the following problem, find $u \in \mathrm{H}^2(\Omega)$ such that

$$-\mathrm{div}\left(\frac{1}{\alpha}(\mathbf{grad}\ u + u\boldsymbol{\beta})\right) + \gamma u = f \ in\ \Omega \quad \text{and}$$

$$u = -g \quad \text{on } \partial\Omega\,, \tag{1}$$

where Ω is a bounded rectangular domain in $\mathbf{R}^2$. We impose the following restrictions on the coefficients:

$$\alpha \in \mathrm{W}_1^\infty(\Omega) \quad \text{and} \quad \exists\, A \in \mathbf{R}\colon \alpha \geq A > 0 \ in\ \Omega\,, \tag{2}$$

$$\frac{1}{\alpha} \in \mathrm{W}_1^\infty(\Omega)\,, \tag{3}$$

$$\boldsymbol{\beta} = (\beta_1, \beta_2)^T \quad \text{with } \beta_1, \beta_2 \in \mathrm{W}_1^\infty(\Omega)\,, \tag{4}$$

$$\gamma \in \mathrm{W}_1^\infty(\Omega) \quad \text{and} \quad \gamma \geq 0 \ in\ \Omega\,, \tag{5}$$

where $\mathrm{W}_1^\infty(\Omega)$, $\mathrm{H}^2(\Omega)$ are the usual Sobolev spaces [8], and

$$\mathrm{H}(\mathrm{div}, \Omega) := \{\boldsymbol{\tau} \in \mathrm{L}^2(\Omega)^2 \,|\, \mathrm{div}\ \boldsymbol{\tau} \in \mathrm{L}^2(\Omega)\},$$

with scalar product

$$(\boldsymbol{\sigma}, \boldsymbol{\tau})_{\mathrm{H}(\mathrm{div},\Omega)} = \int_\Omega \boldsymbol{\sigma} \cdot \boldsymbol{\tau}\, d\mu + \int_\Omega \mathrm{div}\ \boldsymbol{\sigma}\ \mathrm{div}\ \boldsymbol{\tau}\, d\mu\,,$$

is a Hilbert space (see also Girault and Raviart [9] formula 2.15 in Section 2.2). We assume that the equation has a solution and that $f \in \mathrm{L}^2(\Omega)$, $g \in \mathrm{H}^{3/2}(\partial\Omega)$.

The stationary semiconductor continuity equations take the form (1). Here $\boldsymbol{\beta}$ corresponds to the electric field, the term γu corresponds to a linear approximation to the recombination term, and $1/\alpha$ corresponds to the electron or hole mobility. The exact correspondence depends on the choice of scaling [10].

In order to formulate the weak mixed form of this equation, we use the following bilinear forms

$$(s, t) = \int_\Omega st\, d\mu \quad \forall\, s, t \in \mathrm{L}^2(\Omega)\,,$$

$$a(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \int_\Omega \alpha\boldsymbol{\sigma} \cdot \boldsymbol{\tau}\, d\mu \quad \forall\, \boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathrm{H}(\mathrm{div}, \Omega)\,,$$

$$b(\boldsymbol{\sigma}, t) = \int_\Omega \boldsymbol{\beta} \cdot \boldsymbol{\sigma} t\, d\mu \quad \forall\, \boldsymbol{\sigma} \in \mathrm{H}(\mathrm{div}, \Omega), \qquad t \in \mathrm{L}^2(\Omega)\,,$$

$$c(s, t) = \int_\Omega \gamma st\, d\mu \quad \forall\, s, t \in \mathrm{L}^2(\Omega)\,,$$

$$\langle g, h\rangle = \int_{\partial\Omega} gh\, d\lambda \quad \forall\, g, h \in \mathrm{L}^2(\partial\Omega)\,.$$

Given these definitions, we see immediately that any solution $u \in \mathrm{H}^2(\Omega)$ of (1) generates a solution $(\boldsymbol{\sigma}, u) \in \mathrm{H}(\mathrm{div}, \Omega) \times \mathrm{L}^2(\Omega)$ of

$$a(\boldsymbol{\sigma}, \boldsymbol{\tau}) - (\mathrm{div}\ \boldsymbol{\tau}, u) + b(\boldsymbol{\tau}, u) = \langle g, \boldsymbol{\tau} \cdot \mathbf{n}_{\partial\Omega}\rangle \quad \forall\, \boldsymbol{\tau} \in \mathrm{H}(\mathrm{div}, \Omega)\,, \tag{6a}$$

$$(\mathrm{div}\ \boldsymbol{\sigma}, t) + c(u, t) = (f, t) \quad \forall\, t \in \mathrm{L}^2(\Omega)\,, \tag{6b}$$

where

$$\boldsymbol{\sigma} = -\frac{1}{\alpha}(\mathbf{grad}\ u + u\boldsymbol{\beta}).$$

In order to simplify the notation, we denote the Cartesian product of a normed linear space E with itself by $\mathbf{E}$ in boldface type, $\mathbf{E} := E \times E$. We define

$$\|(\mu_1, \mu_2)^T\|_{\mathbf{E}} := \left(\sum_{i=1}^{2} \|\mu_i\|_E^2\right)^{1/2} \quad \forall\ (\mu_1, \mu_2)^T \in \mathbf{E}.$$

III. PREPARATIONS

We introduce a partition of the domain and define the adjoint problem of (1), which we use in the derivation of one of our error estimates. Next, we introduce several special projections that are needed in the definition of our approximation spaces and in the derivation of the error estimates. Finally, we give an error estimate for the projections.

A. Partitioning the Domain

We assume that our domain Ω is rectangular. On Ω, we use Cartesian coordinates, with the unit vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ parallel to the edges of Ω. We define $\tau_i := \boldsymbol{\tau} \cdot \mathbf{e}_i$ for $\boldsymbol{\tau} \in \mathbf{L}^2(\Omega)$ and $x_i := \mathbf{x} \cdot \mathbf{e}_i$ for $\mathbf{x} \in \mathbf{R}^2$. Before we treat our discretization, we define our approximation space. We assume that our partition is the cartesian product of partitions

$$P = \{0 = p_0 < p_1 < \cdots < p_{N_1} = L_1\}, \tag{7}$$

and

$$Q = \{0 = q_0 < q_1 < \cdots < q_{N_2} = L_2\} \tag{8}$$

of the sides of our domain. We define the index set K,

$$K = \{(i + 1/2, j + 1/2) \mid i = 0, 1, \ldots, N_1 - 1, j = 0, 1, \ldots, N_2 - 1\},$$

with the obvious index pair for a given cell,

$$\Omega_{i+1/2, j+1/2} = \{\mathbf{x} \mid p_i < x_1 < p_{i+1}, q_j < x_2 < q_{j+1}\}.$$

We define $\mathbf{x}_k$ to be the center of Ω_k and $\mathbf{h}_k$ to be the diagonal of Ω_k, with the notation χ_k for the characteristic function of Ω_k. (The characteristic function of a set is the function that is equal to one in all points of the set and zero elsewhere.) The edges of Ω_k are the sets:

$$\Gamma_{k,i,j} = \{\mathbf{x} \in \overline{\Omega}_k \mid \mathbf{x} \cdot \mathbf{e}_i = (\mathbf{x}_k + (j - 1/2)\mathbf{h}_k) \cdot \mathbf{e}_i\} \quad \text{for } i = 1, 2, j = 0, 1. \tag{9}$$

$\chi_{k,i,j}$ is the characteristic function of edge $\Gamma_{k,i,j}$. So $(i,j) = (1,0), (1,1), (2,0), (2,1)$ denote the left, right, bottom and top edges.

B. Adjoint Problem

We use the following definition for the adjoint problem of (1) (cf. Douglas and Roberts [7]), $w \in \mathrm{H}^2(\Omega)$,

$$-\mathrm{div}\left(\frac{1}{\alpha}\,\mathbf{grad}\ w\right) + \frac{\boldsymbol{\beta}}{\alpha}\cdot\mathbf{grad}\ w + \gamma w = f \ \mathit{in}\ \Omega\,, \qquad w = 0 \quad \text{on } \partial\Omega\,. \tag{10}$$

The adjoint problem is called regular, if there is a unique solution w for every $f \in \mathrm{L}^2(\Omega)$ and this solution satisfies $\|w\|_{\mathrm{H}^2(\Omega)} \le C\|f\|_{\mathrm{L}^2(\Omega)}$ for every $f \in \mathrm{L}^2(\Omega)$.

Throughout this article the upper case C, without a subscript, denotes a generic constant, which may have a different value at each appearance.

The weak mixed form of the adjoint problem is:

$$(\boldsymbol{\tau}, w) \in \mathrm{H}(\mathrm{div},\Omega) \times \mathrm{L}^2(\Omega)\,, \tag{11}$$

$$a(\boldsymbol{\tau},\boldsymbol{\sigma}) - (\mathrm{div}\ \boldsymbol{\sigma}, w) = 0 \quad \forall\ \boldsymbol{\sigma} \in \mathrm{H}(\mathrm{div},\Omega) \quad \text{and} \tag{11a}$$

$$(\mathrm{div}\ \boldsymbol{\tau}, t) - b(\boldsymbol{\tau}, t) + c(w,t) = (f,t) \quad \forall\ t \in \mathrm{L}^2(\Omega)\,. \tag{11b}$$

Any solution $w \in \mathrm{H}^2(\Omega)$ of (10) generates a solution $(-1/\alpha\ \mathbf{grad}\ w, w)$ of this problem. If (10) is regular, then this solution satisfies $\|w\|_{\mathrm{H}^2(\Omega)} + \|\boldsymbol{\tau}\|_{\mathrm{H}^1(\Omega)} \le C\|f\|_{\mathrm{L}^2(\Omega)}$.

C. Projections

We introduce several local projections, and use these to define four global mappings, P_h, $\mathbf{P}_h$, Π_h, and $\tilde{\Pi}_h$ that map function spaces to finite dimensional function spaces. First, we define $P[\Omega_k]$ to be the orthogonal projection from $\mathrm{L}^2(\Omega_k)$ to the space of constant functions on Ω_k, and we define $P[\Gamma_{k,i,j}]$ to be the orthogonal projection from $\mathrm{L}^2(\Gamma_{k,i,j})$ to the space of constant functions on $\Gamma_{k,i,j}$.

We use $P[\Omega_k]$ to create two global mappings, P_h: $\mathrm{L}^2(\Omega) \to \mathrm{L}^2(\Omega)$,

$$P_h f = \sum_{k\in K} \chi_k P[\Omega_k](f) \quad \forall\ f \in \mathrm{L}^2(\Omega)\,, \tag{12a}$$

and $\mathbf{P}_h$: $\mathbf{L}^2(\Omega) \to \mathbf{L}^2(\Omega)$,

$$\mathbf{P_h}\boldsymbol{\beta} = \sum_{k\in K} \chi_k(P[\Omega_k](\boldsymbol{\beta}\cdot\mathbf{e}_1)\mathbf{e}_1 + P[\Omega_k](\boldsymbol{\beta}\cdot\mathbf{e}_2)\mathbf{e}_2) \quad \forall\ \boldsymbol{\beta} \in \mathbf{L}^2(\Omega)\,. \tag{12b}$$

Next, we introduce two mappings, based on $P[\Gamma_{k,i,j}]$. These have as their domain the space Σ,

$$\Sigma := \{\boldsymbol{\tau} \in \mathrm{H}(\mathrm{div},\Omega)\,|\boldsymbol{\tau}|_{\partial\Omega_k}\cdot\mathbf{n}_{\partial\Omega_k} \in \mathrm{L}^2(\partial\Omega_k) \quad \forall\ k \in K\}\,.$$

This space is similar to that introduced by Roberts and Thomas in formula (1.10) of their report [11].

To simplify the definition of these mappings, we introduce local coordinates on each cell Ω_k,

$$\vec{\xi}_k := \begin{pmatrix} \dfrac{x_1 - x_{k,1}}{h_{k,1}} + \dfrac{1}{2} \\ \dfrac{x_2 - x_{k,2}}{h_{k,2}} + \dfrac{1}{2} \end{pmatrix}. \tag{13}$$

The mappings are defined as follows:

$$\Pi_h \boldsymbol{\tau} = \sum_{k\in K} \chi_k \sum_{i=1}^{2} ((1 - \xi_{k,i})P[\Gamma_{k,i,0}](\tau_i) + \xi_{k,i}P[\Gamma_{k,i,1}](\tau_i))\mathbf{e}_i , \tag{14}$$

$$\tilde{\Pi}_h \boldsymbol{\tau} = \sum_{k\in K} \chi_k \sum_{i=1}^{2} ((1 - \zeta_{k,i})P[\Gamma_{k,i,0}](\tau_i) + \zeta_{k,i}P[\Gamma_{k,i,1}](\tau_i))\mathbf{e}_i , \tag{15}$$

where

$$\zeta_{k,i} = \begin{cases} \dfrac{\exp(\xi_{k,i}h_{k,i}P[\Omega_k](\beta_i)) - 1}{\exp(h_{k,i}P[\Omega_k](\beta_i)) - 1} & \mathit{if}\ P[\Omega_k](\beta_i) \neq 0, \\ \xi_{k,i} & \mathit{if}\ P[\Omega_k](\beta_i) = 0. \end{cases}$$

For $\Pi_h\boldsymbol{\tau}$ we get the i^{th} component on Ω_k by linear interpolation between the projections of this component on the two sides orthogonal to $\mathbf{e}_i$. For $\Pi_h\boldsymbol{\tau}$, however, we obtain the same component by using an exponential function to interpolate between the projections of this component on the two sides orthogonal to $\mathbf{e}_i$.

The following finite dimensional function spaces are now introduced as the ranges of the above projections:

$$V_h = \Pi_h(\Sigma), \qquad W_h = P_h(\mathrm{L}^2(\Omega)), \qquad \text{and} \quad X_h = \tilde{\Pi}_h(\Sigma).$$

$V_h \times W_h$ is the lowest order Raviart–Thomas–Nedelec space for rectangles. This space and the above projections were described by Douglas and Roberts [7], Raviart and Thomas [12], and, for $\Omega \subset \mathbf{R}^3$, by Nedelec [13]. In this article, we use the usual space, $V_h \times W_h$ as the trial function space and $X_h \times W_h$ as the test function space. In effect, we use exponential test functions instead of the usual linear test functions. Thus, we obtain a Petrov–Galerkin mixed finite element discretization.

D. Error Estimates for Projections

A lemma on the accuracy of our projections is now found. Considering the number and diversity of articles on error estimates, e.g., the classical projection estimates from Ciarlet and Raviart [14], this may seem superfluous, but we shall see that the relative simplicity of the case under consideration makes it possible to derive sharp error estimates under minimal assumptions.

Lemma 1. *If f is a square integrable function with square integrable derivatives on a rectangle $\Omega = [0,h_1] \times [0,h_2]$ with sides $\Gamma_{1,1} = \{h_1\} \times [0,h_2]$, $\Gamma_{2,1} = [0,h_1] \times \{h_2\}$, $\Gamma_{1,0} = \{0\} \times [0,h_2]$, and $\Gamma_{2,0} = [0,h_1] \times \{0\}$, then the following inequalities hold:*

$$\|f - P[\Omega]f\|_{\mathrm{L}^2(\Omega)} \le (2h_1^2 + 2h_2^2)^{1/2}\|\mathbf{grad}\ f\|_{\mathrm{L}^2(\Omega)} . \tag{16a}$$

If s is a continuous function with domain $[0,h_1]$ and range $[0,1]$, then we have

$$\|f - (1 - s)\Pi[\Gamma_{1,0}]f - s\Pi[\Gamma_{1,1}]f\|_{\mathrm{L}^2(\Omega)} \le (2h_1^2 + 2h_2^2)^{1/2}\|\mathbf{grad}\ f\|_{\mathrm{L}^2(\Omega)} . \tag{16b}$$

If $f \in L^\infty(\Omega)$, $\mathbf{grad}\ f \in \mathbf{L}^\infty(\Omega)$, then

$$\|f - P[\Omega]f\|_{\mathrm{L}^\infty(\Omega)} \le (h_1 + h_2)\,\|\mathbf{grad}\ f\|_{\mathrm{L}^\infty(\Omega)} . \tag{16c}$$

Proof. We start by proving the above inequalities for $f \in C^1(\Omega)$. We can then extend them by the usual density argument to $\mathrm{H}^1(\Omega)$. To prove the first inequality, we write

$$\|f - P[\Omega]f\|^2_{\mathrm{L}^2(\Omega)} = \int_{x=0}^{h_1}\int_{y=0}^{h_2}\left(\frac{1}{h_1h_2}\int_{w=0}^{h_1}\int_{z=0}^{h_2} f(x,y) - f(w,z)\,dw\,dz\right)^2 dx\,dy\,,$$

by definition,

$$f(x,y) - f(w,z) = \int_{a=w}^{x}\frac{\partial f}{\partial a}(a,z)\,da + \int_{b=z}^{y}\frac{\partial f}{\partial b}(x,b)\,da\,.$$

If we substitute this into the above expression, then we find

$$\|f - P[\Omega]f\|^2_{\mathrm{L}^2(\Omega)} =$$

$$\int_{x=0}^{h_1}\int_{y=0}^{h_2}\left(\frac{1}{h_1h_2}\int_{w=0}^{h_1}\int_{z=0}^{h_2}\left[\int_{a=w}^{x}\frac{\partial f}{\partial a}(a,z)\,da + \int_{b=z}^{y}\frac{\partial f}{\partial b}(x,b)\,db\right]dw\,dz\right)^2 dx\,dy\,.$$

We apply the Hölder inequality to the inner integrals and extend the integrations over a and b, where appropriate,

$$\|f - P[\Omega]f\|^2_{\mathrm{L}^2(\Omega)} \le$$

$$\int_{x=0}^{h_1}\int_{y=0}^{h_2}\left(\frac{h_1^{1/2}}{h_2^{1/2}}\|\partial f/\partial x_1\|_{\mathrm{L}^2(\Omega)} + h_2^{1/2}\left(\int_{b=0}^{h_2}\left(\frac{\partial f}{\partial b}(x,b)\right)^2 db\right)^{1/2}\right)^2 dx\,dy\,.$$

We use $(|A| + |B|)^2 \le 2(A^2 + B^2)$ to write this as

$$\|f - P[\Omega]f\|^2_{\mathrm{L}^2(\Omega)} \le 2\int_{x=0}^{h_1}\int_{y=0}^{h_2}\frac{h_1}{h_2}\|\partial f/\partial x_1\|^2_{\mathrm{L}^2(\Omega)}\,dx\,dy$$
$$+ 2\int_{y=0}^{h_2} h_2\|\partial f/\partial x_2\|^2_{\mathrm{L}^2(\Omega)}\,dy\,.$$

This reduces to,

$$\|f - P[\Omega]f\|^2_{\mathrm{L}^2(\Omega)} \le 2h_1^2\|\partial f/\partial x_1\|^2_{\mathrm{L}^2(\Omega)} + 2h_2^2\|\partial f/\partial x_2\|^2_{\mathrm{L}^2(\Omega)}\,.$$

Now, we consider the second inequality, (16b), we write,

$$\|f - (1 - s)\Pi[\Gamma_{1,0}]f - s\Pi[\Gamma_{1,1}]f\|^2_{\mathrm{L}^2(\Omega)} =$$

$$\int_{x=0}^{h_1}\int_{y=0}^{h_2}\left(\frac{1}{h_2}\int_{z=0}^{h_2}[(1 - s(x))(f(x,y) - f(0,z)) + s(x)(f(x,y) - f(h_1,z))]dz\right)^2 dx\,dy\,.$$

We use partial derivatives to rewrite the expression,

$$\|f - (1 - s)\Pi[\Gamma_{1,0}]f - s\Pi[\Gamma_{1,1}]f\|^2_{\mathrm{L}^2(\Omega)} =$$

$$\int_{x=0}^{h_1}\int_{y=0}^{h_2}\left(\frac{1}{h_2}\int_{z=0}^{h_2}\left[(1 - s(x))\left\{\int_{a=0}^{x}\frac{\partial f}{\partial a}(a,z)\,da + \int_{b=z}^{y}\frac{\partial f}{\partial b}(x,b)\,db\right\}\right.\right.$$
$$\left.\left. + s(x)\left\{\int_{a=h_1}^{x}\frac{\partial f}{\partial a}(a,z)\,da + \int_{b=z}^{y}\frac{\partial f}{\partial b}(x,b)\,db\right\}\right]dz\right)^2 dx\,dy\,.$$

Using the Hölder inequality we extend the integrals where appropriate, so that

$$\|f - (1 - s)\Pi[\Gamma_{1,0}]f - s\Pi[\Gamma_{1,1}]f\|^2_{L^2(\Omega)} \leq$$
$$\int_{x=0}^{h_1}\int_{y=0}^{h_2}\left(\frac{h_1^{1/2}}{h_2^{1/2}}\|\partial f/\partial x_1\|_{L^2(\Omega)} + h_2^{1/2}\left(\int_{b=0}^{h_2}\left(\frac{\partial f}{\partial b}(x,b)\right)^2 db\right)^{1/2}\right)^2 dx\,dy\,.$$

We use $(|A| + |B|)^2 \leq 2(A^2 + B^2)$ to write this as

$$\|f - (1 - s)\Pi[\Gamma_{1,0}]f - s\Pi[\Gamma_{1,1}]f\|^2_{L^2(\Omega)} \leq$$
$$2\int_{x=0}^{h_1}\int_{y=0}^{h_2}\left(\frac{h_1}{h_2}\|\partial f/\partial x_1\|_{L^2(\Omega)} + h_2\int_{b=0}^{h_2}\left(\frac{\partial f}{\partial b}(x,b)\right)^2 db\right) dx\,dy\,,$$

which reduces to

$$\|f - (1 - s)\Pi[\Gamma_{1,0}]f - s\Pi[\Gamma_{1,1}]f\|^2_{L^2(\Omega)} \leq 2h_1^2\|\partial f/\partial x_1\|^2_{L^2(\Omega)} + 2h_2^2\|\partial f/\partial x_2\|^2_{L^2(\Omega)}\,.$$

Lastly we verify (16c),

$$f(x,y) - f(w,z) = \int_{a=w}^{x}\frac{\partial f}{\partial a}(a,z)\,da + \int_{b=z}^{y}\frac{\partial f}{\partial b}(x,b)\,da\,.$$

So,

$$\frac{1}{h_1h_2}\int_{x=0}^{h_1}\int_{y=0}^{h_2} f(x,y) - f(w,z)\,dx\,dy =$$
$$\frac{1}{h_1h_2}\int_{x=0}^{h_1}\int_{y=0}^{h_2}\left(\int_{a=w}^{x}\frac{\partial f}{\partial a}(a,z)\,da + \int_{b=z}^{y}\frac{\partial f}{\partial b}(x,b)\,da\right) dx\,dy$$
$$\leq (h_1 + h_2)\|\mathbf{grad}\ f\|_{\mathbf{L}^{\infty}(\Omega)}\,.$$

■

Note that the above inequalities imply

$$\|\boldsymbol{\sigma} - \Pi_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)} \leq \max_{k\in K}(2h_{k,1}^2 + 2h_{k,2}^2)^{1/2}\|\boldsymbol{\sigma}\|_{\mathbf{H}^1(\Omega)}\,, \tag{17a}$$

$$\|\boldsymbol{\sigma} - \tilde{\Pi}_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)} \leq \max_{k\in K}(2h_{k,1}^2 + 2h_{k,2}^2)^{1/2}\|\boldsymbol{\sigma}\|_{\mathbf{H}^1(\Omega)}\,, \tag{17b}$$

$$\|u - P_hu\|_{L^2(\Omega)} \leq \max_{k\in K}(2h_{k,1}^2 + 2h_{k,2}^2)^{1/2}\|u\|_{H^1(\Omega)}\,, \tag{17c}$$

for suitable u and $\boldsymbol{\sigma}$.

IV. DISCRETIZATION

We describe our discretization. The basic idea of mixed finite elements with a lowest order Raviart–Thomas trial space and exponentially fitted test subspace for the vector valued functions is complicated by the use of a quadrature rule, needed to keep the M-matrix property for the system without Lagrange multipliers for nonzero γ. This quadrature rule is discussed in Section IVA. Another complication is the approximation of the coefficients by piecewise constant functions, as described below. In Section IVB we give the resulting discretization.

We replace the coefficients α, $\boldsymbol{\beta}$ and γ by two-dimensional step functions. To write our modified problem in weak form, we need to define three new bilinear forms:

$$\begin{aligned}\overline{a}(\boldsymbol{\sigma},\boldsymbol{\tau}) &= \int_\Omega \boldsymbol{\sigma}\cdot\boldsymbol{\tau}P_h\alpha\,d\mu \quad \forall\,\boldsymbol{\sigma},\boldsymbol{\tau}\in\Sigma,\\ \overline{b}(\boldsymbol{\sigma},t) &:= \int_\Omega t\boldsymbol{\sigma}\cdot \mathrm{P}_h\boldsymbol{\beta}\,d\mu \quad \forall\,\boldsymbol{\sigma}\in\Sigma, t\in \mathrm{L}^2(\Omega),\\ \overline{c}(s,t) &:= \int_\Omega stP_h\gamma\,d\mu \quad \forall\, s,t\in\mathrm{L}^2(\Omega).\end{aligned}$$

The bar on the bilinear forms denotes that the coefficients are replaced by their cell-wise averages. We then replace $\overline{a}$ by $\overline{a}_q$, the subscript q indicates that a—not yet specified—quadrature rule will be used in the evaluation of this bilinear form.

A. Quadrature Rule

We construct a quadrature rule $\overline{a}_{h,1}$ by imposing the condition that, if $\alpha,\boldsymbol{\beta}$ are constant, $\gamma\equiv 0$, and the solution satisfies $u = C\exp(-\beta_1x_1-\beta_2x_2)+K$, with $C,K\in\mathbf{R}$, then the discrete solution should satisfy $\boldsymbol{\sigma}_h = \Pi_h\boldsymbol{\sigma}$ and $u_h = P_hu$. We see that for the u given above $\boldsymbol{\sigma} = -K\boldsymbol{\beta}/\alpha$, so $\boldsymbol{\sigma}$ is constant. We define α_h separately for each basis function $\boldsymbol{\eta}_{i,j+1/2}$ where

$$\boldsymbol{\eta}_{i,j+1/2} = \begin{cases} \zeta_{(i-1/2,j+1/2)}\mathbf{e}_1 \in \Omega_{i-1/2,j+1/2}\,,\\ (1-\zeta_{(i+1/2,j+1/2)})\mathbf{e}_1\in\Omega_{i+1/2,j+1/2}\,,\\ 0 \text{ elsewhere}\,,\end{cases}$$

and $\boldsymbol{\eta}_{i+1/2,j}$, where

$$\boldsymbol{\eta}_{i+1/2,j} = \begin{cases} \zeta_{(i+1/2,j-1/2)}\mathbf{e}_2 \in \Omega_{i-1/2,j-1/2}\,,\\ (1-\zeta_{(i+1/2,j-1/2)})\mathbf{e}_2\in\Omega_{i+1/2,j+1/2}\,,\\ 0 \text{ elsewhere}\,.\end{cases}$$

We denote the set of all possible indices for the basis functions $\boldsymbol{\eta}$ by

$$\begin{aligned}E =&\{e = (i,j-1/2)\,|\,i = 0,1,2,\ldots,N_1, j = 1,2,\ldots,N_2\}\bigcup\\ &\{e = (i-1/2,j)\,|\,i = 1,2,\ldots,N_1, j = 0,1,2,\ldots,N_2\}.\end{aligned}$$

Our quadrature rule should satisfy the following condition:

$$\overline{a}_{h,1}(\boldsymbol{\sigma},\boldsymbol{\eta}_r) = \overline{a}(\boldsymbol{\sigma},\boldsymbol{\eta}_r)\,, \tag{A}$$

for all constant $\boldsymbol{\sigma}$ and all $r\in E$. Due to our assumption that the coefficients are constant, we have $a = \overline{a}$ and $b = \overline{b}$. The above condition guarantees that for constant coefficients and constant $\boldsymbol{\sigma}$,

$$\begin{aligned}a(\boldsymbol{\sigma},\boldsymbol{\tau}_h) - (u,\operatorname{div}\boldsymbol{\tau}_h) + b(\boldsymbol{\tau}_h,u) = \overline{a}_{h,1}(\Pi_h\boldsymbol{\sigma},\boldsymbol{\tau}_h) - (P_hu,\operatorname{div}\boldsymbol{\tau}_h)&\\ + b(\boldsymbol{\tau}_h,P_hu) \quad \forall\,\boldsymbol{\tau}_h\in X_h\,,&\end{aligned}$$

and we also have

$$(\operatorname{div}\boldsymbol{\sigma},t) = (\operatorname{div}\Pi_h\boldsymbol{\sigma},t) = 0 \quad \forall\, t\in\mathrm{L}^2(\Omega)\,.$$

So our condition (A) on $\bar{a}_{h,1}$ is sufficient for our purposes. We now select the quadrature rule by taking the following definition for $\bar{a}_{h,1}$,

$$\bar{a}_{h,1}(\boldsymbol{\sigma},\boldsymbol{\tau}) = \sum_{k\in K}\sum_{i=1}^{2} \mu(\Omega_k)P[\Omega_k](\alpha)\,(P[\Omega_k](\zeta_{k,i})P[\Gamma_{k,i,1}](\sigma_i\tau_i) + P[\Omega_k](1-\zeta_{k,i})P[\Gamma_{k,i,0}](\sigma_i\tau_i))\,. \tag{18a}$$

We introduce a new problem dependent norm on X_h

$$\|\boldsymbol{\tau}_h\|_h = \sum_{k\in K}\sum_{i=1}^{2} \mu(\Omega_k)\,(P[\Omega_k](\zeta_{k,i})P[\Gamma_{k,i,1}](\tau_{h,i}^2) + P[\Omega_k](1-\zeta_{k,i})P[\Gamma_{k,i,0}](\tau_{h,i}^2))^{1/2}. \tag{18b}$$

From this point onward, we take $\bar{a}_q = \bar{a}_{h,1}$.

B. Discrete System

We approximate the solution $(\boldsymbol{\sigma}, u)$ of (6) by $(\boldsymbol{\sigma}_h, u_h) \in V_h \times W_h$, where

$$\bar{a}_q(\boldsymbol{\sigma}_h,\boldsymbol{\tau}) - (u_h, \operatorname{div}\boldsymbol{\tau}) + \bar{b}(\boldsymbol{\tau}, u_h) = \langle \boldsymbol{\tau}\cdot\boldsymbol{\eta}_{\partial\Omega}, g\rangle \quad \forall\, \boldsymbol{\tau}\in X_h\,, \tag{19a}$$

$$(\operatorname{div}\boldsymbol{\sigma}_h, t) + \bar{c}(u_h, t) = (f, t) \quad \forall\, t \in W_h\,. \tag{19b}$$

If we use $\bar{a}$ instead of $\bar{a}_q$, then this means that our discrete problem does not always yield an M-matrix for u_h. Consider, for instance, the corresponding discretization on a uniform mesh with mesh width h in one dimension with $\alpha = 1$, $\boldsymbol{\beta} = \vec{0}$, and γ constant. If $\alpha\gamma h^2/6 > 1$, then the off-diagonal elements of the discretization matrix for u_h after elimination of $\boldsymbol{\sigma}_h$ through static condensation have the same sign as the elements on the diagonal.

The idea of using linear trial functions and exponential test functions was used by Hemker for singularly perturbed two-point boundary problems [6]. For the one-dimensional case, the introduction of exponential test functions follows from the requirement that it must be possible to approximate the Green's function of the problem by the test functions.

In the following sections, we prove that the solution of our discretization (19) is an $O(h)$ approximation to the solution of our original problem.

V. TECHNICAL RESULTS

This section contains some technical results, collected for later reference.

Lemma 2.

$$\tilde{\Pi}_h \circ \Pi_h = \tilde{\Pi}_h\,, \tag{20a}$$

$$\Pi_h \circ \tilde{\Pi}_h = \Pi_h\,, \tag{20b}$$

$$(\operatorname{div}\sigma, P_h t) = (\operatorname{div}\Pi_h\sigma, t) \quad \forall\, \sigma \in \Sigma, t \in L^2(\Omega)\,, \tag{20c}$$

$$\Pi_h\, \tau \cdot \vec{n}_{\partial\Omega} = \tilde{\Pi}_h \tau \cdot \vec{n}_{\partial\Omega} \quad \forall\, \tau \in \Sigma\,. \tag{20d}$$

Proof. Both mappings are based on the same projections $P[\Gamma_{k,i,j}]$, so (20a) and (20b) are trivial.

To prove (20c) we use a special case of Green's theorem:

$$\int_{\Omega_k} \operatorname{div} \sigma \, d\mu = \sum_{i=1}^{2} \frac{\mu(\Omega_k)}{h_{k,i}} (P[\Gamma_{k,i,1}](\sigma_i) - P[\Gamma_{k,i,0}](\sigma_i)) .$$

If we combine this with the definition of Π_h, the proof of (20c) is complete. Equation (20d) follows immediately from the definitions. ■

Lemma 3. *If $\boldsymbol{\sigma} \in \Sigma$ and we define $a_{k,i} = P[\Gamma_{k,i,0}](\boldsymbol{\sigma}_i)$ and $b_{k,i} = P[\Gamma_{k,i,1}](\boldsymbol{\sigma}_i)$, then the following inequalities hold for $\|\Pi_h \boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega_k)}$ and $\|\tilde{\Pi}_h \boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega_k)}$:*

$$\frac{\mu(\Omega_k)}{6} \sum_{i=1}^{2} (a_{k,i}^2 + b_{k,i}^2) \leq \|\Pi_h \boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega_k)}^2 \leq \frac{\mu(\Omega_k)}{2} \sum_{i=1}^{2} (a_{k,i}^2 + b_{k,i}^2) . \tag{21a}$$

$$\|\tilde{\Pi}_h \boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega_k)}^2 \leq 2\|\tilde{\Pi}_h \boldsymbol{\sigma}\|_h^2 \leq 12\|\Pi_h \boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega_k)}^2 . \tag{21b}$$

Proof. Formula (21a) follows immediately from

$$(\Pi_h \boldsymbol{\sigma}, \Pi_h \boldsymbol{\sigma}) = \sum_{i=1}^{2} \sum_{k \in K} \int_{\Omega_k} ((1 - \xi_{k,i}) a_{k,i} + \xi_{k,i} b_{k,i})^2 \, d\mu .$$

Next, we derive (21b) from

$$(\tilde{\Pi}_h \boldsymbol{\sigma}, \tilde{\Pi}_h \boldsymbol{\sigma}) = \sum_{i=1}^{2} \sum_{k \in K} \int_{\Omega_k} ((1 - \zeta_{k,i}) a_{k,i} + \zeta_{k,i} b_{k,i})^2 \, d\mu .$$

We see immediately that

$$\int_{\Omega_k} ((1 - \zeta_{k,i}) a_{k,i} + \zeta_{k,i} b_{k,i})^2 \, d\mu \leq \int_{\Omega_k} 2(1 - \zeta_{k,i})^2 a_{k,i}^2 + 2\zeta_{k,i}^2 b_{k,i}^2 \, d\mu \leq$$

$$2 \int_{\Omega_k} (1 - \zeta_{k,i}) a_{k,i}^2 + \zeta_{k,i} b_{k,i}^2 \, d\mu = 2\mu(\Omega_k) (P[\Omega_k](1 - \zeta_{k,i}) a_{k,i}^2 + P[\Omega_k](\zeta_{k,i}) b_{k,i}^2) .$$

This implies (21b). ■

Lemma 4 shows, that $\overline{a}$ is $\mathbf{L}^2(\Omega)$-bounded and $\mathbf{L}^2(\Omega)$-elliptic.

Lemma 4. *Let $\alpha \in W_1^\infty(\Omega)$, $\alpha \geq A > 0 \in \Omega$ and $\overline{a}(\boldsymbol{\sigma}, \boldsymbol{\tau}) := \int_\Omega P_h(\alpha) \boldsymbol{\sigma} \cdot \boldsymbol{\tau} \, d\mu$ $\forall \ \boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathbf{L}^2(\Omega)$, then*

$$\overline{a}(\boldsymbol{\sigma}, \boldsymbol{\tau}) \leq \|\alpha\|_{\mathbf{L}^\infty(\Omega)} \|\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)} \|\boldsymbol{\tau}\|_{\mathbf{L}^2(\Omega)} \quad \forall \ \boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathbf{L}^2(\Omega) , \tag{22a}$$

and

$$\overline{a}(\boldsymbol{\tau}, \boldsymbol{\tau}) \geq A \|\boldsymbol{\tau}\|_{\mathbf{L}^2(\Omega)} \quad \forall \ \boldsymbol{\tau} \in \mathbf{L}^2(\Omega) . \tag{22b}$$

Proof. From (2) it follows that

$$A \leq \frac{\int_{\Omega_k} \alpha \, d\mu}{\mu(\Omega_k)} \leq \|\alpha\|_{\mathbf{L}^\infty(\Omega)} ,$$

together with the definitions of P and $\overline{a}$, this implies (22a) and (22b). ■

We introduce the minimum mesh width $h_{\min}$ and the maximum mesh width $h_{\max}$,

$$h_{\min} = \min_{k \in K} \min_{i=1,2} |h_{k,i}| , \tag{23a}$$

$$h_{\max} = \max_{k \in K} \max_{i=1,2} |h_{k,i}| . \tag{23b}$$

A. Properties of $\overline{a}_q$

We discuss the properties of the quadrature rule $\overline{a}_q$ and assume that $\overline{a}_q = \overline{a}_{h,1}$, where $\overline{a}_{h,1}$ is given by (18a). We also discuss the interaction between Π, $\tilde{\Pi}$, and $\overline{a}_q$. We show that $\overline{a}_q$ is $\mathbf{L}^2(\Omega)$-bounded on V_h, and we also show that $\overline{a}_q$ is $\mathbf{L}^2(\Omega)$-elliptic on V_h and X_h.

Lemma 5. *If* $\boldsymbol{\sigma}, \boldsymbol{\tau} \in \Sigma$, *then*

$$\overline{a}_q(\Pi_h\boldsymbol{\sigma}, \Pi_h\boldsymbol{\tau}) = \overline{a}_q(\Pi_h\boldsymbol{\tau}, \Pi_h\boldsymbol{\sigma}) = \overline{a}_q(\boldsymbol{\sigma}, \Pi_h\boldsymbol{\tau}) = \overline{a}_q(\Pi_h\boldsymbol{\sigma}, \boldsymbol{\tau}) = \overline{a}_q(\boldsymbol{\sigma}, \tilde{\Pi}_h\boldsymbol{\tau}) = \overline{a}_q(\tilde{\Pi}_h\boldsymbol{\sigma}, \boldsymbol{\tau}) = \overline{a}_q(\tilde{\Pi}_h\boldsymbol{\sigma}, \tilde{\Pi}_h\boldsymbol{\tau}), \tag{24a}$$

$$\|\alpha\|_{\mathrm{L}^\infty(\Omega)}\|\tilde{\Pi}_h\boldsymbol{\sigma}\|_h^2 \geq \overline{a}_q(\tilde{\Pi}_h\boldsymbol{\sigma}, \tilde{\Pi}_h\boldsymbol{\sigma}) \geq \frac{1}{2}\overline{a}(\tilde{\Pi}_h\boldsymbol{\sigma}, \tilde{\Pi}_h\boldsymbol{\sigma}) \geq \frac{A}{2}\|\tilde{\Pi}_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)}^2, \tag{24b}$$

$$\overline{a}_q(\Pi_h\boldsymbol{\sigma}, \Pi_h\boldsymbol{\tau}) \leq 6\|\alpha\|_{\mathrm{L}^\infty(\Omega)}\|\Pi_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)}\|\tilde{\Pi}_h\boldsymbol{\tau}\|_h, \tag{24c}$$

$$A\|\tilde{\Pi}_h\boldsymbol{\tau}\|_h^2 \leq \overline{a}_q(\boldsymbol{\tau}, \tilde{\Pi}_h\boldsymbol{\tau}) \leq \|\alpha\|_{\mathrm{L}^\infty(\Omega)}\|\tilde{\Pi}_h\boldsymbol{\tau}\|_h^2. \tag{24d}$$

Proof. The definitions of Π_h, $\tilde{\Pi}_h$, and $\overline{a}_q$ imply (24a). Inequality (24b) follows immediately from (18a), (18b), and (21b). To prove (24c), we need some auxiliary variables, $a_{k,i} = P[\Gamma_{k,i,0}](\boldsymbol{\sigma})$, $b_{k,i} = P[\Gamma_{k,i,1}](\boldsymbol{\sigma})$, $c_{k,i} = P[\Gamma_{k,i,0}](\boldsymbol{\tau})$, and $d_{k,i} = P[\Gamma_{k,i,1}](\boldsymbol{\tau})$. Cauchy–Schwarz is used twice, we obtain

$$\begin{aligned}\overline{a}_q(\Pi\boldsymbol{\sigma}, \tilde{\Pi}\boldsymbol{\tau}) &= \sum_{k\in K} P[\Omega_k](\alpha)\mu(\Omega_k)\sum_{i=1}^{2}(P[\Omega_k](1-\zeta_{k,i})a_{k,i}c_{k,i} + P[\Omega_k](\zeta_{k,i})b_{k,i}d_{k,i})\\ &\leq \sum_{k\in K} P[\Omega_k](\alpha)\mu(\Omega_k)\left(\sum_{i=1}^{2}(a_{k,i}^2 + b_{k,i}^2)\right)^{1/2}\\ &\quad\times\left(\sum_{i=1}^{2}[P[\Omega_k](1-\zeta_{k,i})^2c_{k,i}^2 + P[\Omega_k](\zeta_{k,i})^2d_{k,i}^2\right)^{1/2}.\end{aligned}$$

We use

$$P[\Omega_k](f)^2 \leq P[\Omega_k](f^2)$$

to rewrite the term in c and d, and we use (21a) to replace the term in a and b by $\|\Pi_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)}$,

$$\begin{aligned}\overline{a}_q(\Pi_h\boldsymbol{\sigma}, \tilde{\Pi}_h\boldsymbol{\tau}) \leq \sum_{k\in K} P[\Omega_k](\alpha)\mu(\Omega_k)6\frac{\|\Pi_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)}}{\mu(\Omega_k)^{1/2}}\\ \times\left(\sum_{i=1}^{2}(P[\Omega_k]((1-\zeta_{k,i})^2)c_{k,i}^2 + P[\Omega_k]((\zeta_{k,i})^2)d_{k,i}^2)\right)^{1/2}.\end{aligned}$$

We see immediately that this implies

$$\overline{a}_q(\Pi_h\boldsymbol{\sigma}, \tilde{\Pi}_h\boldsymbol{\tau}) \leq 6\|\alpha\|_{\mathrm{L}^\infty(\Omega)}\|\Pi_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)}\|\tilde{\Pi}_h\boldsymbol{\tau}\|_h.$$

This proves (24c). Inequality (24d) follows immediately from (18). ■

B. Difference between $\bar{a}$ and $\bar{a}_q$

For our error estimates, we need an upper bound for the difference between the value of $a(\boldsymbol{\sigma}_h, \boldsymbol{\tau})$ and that of $\bar{a}_q(\boldsymbol{\sigma}_h, \boldsymbol{\tau})$ for $\boldsymbol{\sigma}_h \in V_h$, $\boldsymbol{\tau} \in \mathbf{H}^1(\Omega)$. As we already know from (16c) (see also Lemmas 8 and 9) that

$$|a(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) - \bar{a}(\boldsymbol{\sigma}, \boldsymbol{\tau}_h)| \leq 2h_{\max} \|\alpha\|_{W_1^\infty(\Omega)} \|\boldsymbol{\sigma}\|_{\mathrm{L}^2(\Omega)} \|\boldsymbol{\tau}_h\|_{\mathrm{L}^2(\Omega)},$$

an estimate for $|\bar{a}(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) - \bar{a}_q(\boldsymbol{\sigma}, \boldsymbol{\tau}_h)|$ suffices. Such an estimate is derived in Lemma 6.

Lemma 6. *Let $\boldsymbol{\tau}_h \in X_h$ and $\boldsymbol{\sigma} \in \mathbf{H}^1(\Omega)$, then*

$$|\bar{a}(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) - \bar{a}_q(\boldsymbol{\sigma}, \boldsymbol{\tau}_h)| \leq 2\|\alpha\|_{\mathrm{L}^\infty(\Omega)} h_{\max} \|\boldsymbol{\tau}_h\|_h \|\boldsymbol{\sigma}\|_{\mathbf{H}^1(\Omega)}. \tag{25}$$

Proof. To simplify our notation, we introduce $a_{k,i} = P[\Gamma_{k,i,0}](\boldsymbol{\tau}_h)$, $b_{k,i} = P[\Gamma_{k,i,1}](\boldsymbol{\tau}_h)$, $\sigma_{k,i,0} = P[\Gamma_{k,i,0}](\sigma_i)$, and $\sigma_{k,i,1} = P[\Gamma_{k,i,1}](\sigma_i)$. We prove the lemma for $\boldsymbol{\sigma}$ with $\sigma_1, \sigma_2 \in C^1(\Omega)$, and extend by density.

We consider the difference between the two forms on one subdomain Ω_k with $P[\Omega_k](\alpha) = 1$.

$$\begin{aligned}
&\left| \int_{\Omega_k} \boldsymbol{\sigma} \cdot \boldsymbol{\tau}_h \, d\mu - \mu(\Omega_k) \sum_{i=1}^{2} (P[\Omega_k](1 - \zeta_{k,i}) P[\Gamma_{k,i,0}](\sigma_i \tau_{h,i}) \right. \\
&\left. + P[\Omega_k](\zeta_{k,i}) P[\Gamma_{k,i,1}](\sigma_i \tau_{h,i})) \right| = \left| \int_{\Omega_k} \sum_{i=1}^{2} ((1 - \zeta_{k,i}) a_{k,i} + \zeta_{k,i} b_{k,i}) \sigma_i \, d\mu - \mu(\Omega_k) \right. \\
&\quad \left. \times \sum_{i=1}^{2} (P[\Omega_k](1 - \zeta_{k,i}) P[\Gamma_{k,i,0}](a_{k,i}\sigma_i) + P[\Omega_k](\zeta_{k,i}) P[\Gamma_{k,i,1}](b_{k,i}\sigma_i)) \right| = \\
&\quad \left| \int_{\Omega_k} \sum_{i=1}^{2} ((1 - \zeta_{k,i}) a_{k,i} \sigma_i + \zeta_{k,i} b_{k,i} \sigma_i - P[\Omega_k](1 - \zeta_{k,i}) a_{k,i} \sigma_{k,i,0} \right. \\
&\quad \left. - P[\Omega_k](\zeta_{k,i}) b_{k,i} \sigma_{k,i,1}) \, d\mu \right| = \left| \int_{\Omega_k} \sum_{i=1}^{2} ((1 - \zeta_{k,i}) a_{k,i} (\sigma_i - \sigma_{k,i,0}) \right. \\
&\qquad\qquad \left. + \zeta_{k,i} b_{k,i} (\sigma_i - \sigma_{k,i,1})) \, d\mu \right| .
\end{aligned}$$

The application of the Cauchy–Schwarz inequality to this last term and insertion of α yields the following result:

$$|\bar{a}(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) - \bar{a}_q(\boldsymbol{\sigma}, \boldsymbol{\tau}_h)| \leq h_{\max} \|\alpha\|_{\mathrm{L}^\infty(\Omega)} \|\boldsymbol{\tau}_h\|_h \left(\sum_{k \in K} \sum_{i=1}^{2} \sum_{j=0}^{1} \|\sigma_i - \sigma_{i,k,j}\|_{\mathrm{L}^2(\Omega_k)}^2 \right)^{1/2} .$$

If we take $s \equiv j$ in (16b) then this implies

$$\begin{aligned}
|\bar{a}(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) - \bar{a}_q(\boldsymbol{\sigma}, \boldsymbol{\tau}_h)| &\leq \|\alpha\|_{\mathrm{L}^\infty(\Omega)} \|\boldsymbol{\tau}_h\|_h \left(\sum_{i=1}^{2} 4h_{\max}^2 \|\mathbf{grad}\ \sigma_i\|_{\mathrm{L}^2(\Omega)} \right)^{1/2} \\
&\leq 2h_{\max} \|\alpha\|_{\mathrm{L}^\infty(\Omega)} \|\boldsymbol{\tau}_h\|_h \|\boldsymbol{\sigma}\|_{\mathbf{H}^1(\Omega)} .
\end{aligned}$$

Because $C^1(\overline{\Omega})$ is dense in $\mathrm{H}^1(\Omega)$, the formula also holds for $\sigma_1, \sigma_2 \in \mathrm{H}^1(\Omega)$. ■

VI. ERROR ESTIMATES

We use the standard estimates for $\|\boldsymbol{\sigma} - \Pi_h\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)}$ and $\|u - P_h u\|_{L^2(\Omega)}$, as described in Section IIID, to reduce the problem to deriving bounds for $\|P_h u - u_h\|_{L^2(\Omega)}$ and $\|\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{L}^2(\Omega)}$. We discuss two possible derivations of an $\mathbf{O}(h)$ error bound. The first needs the assumption that $h_{\max}$ is "small enough," while the second places a condition on an approximation of the discrete version of the adjoint problem.

A. Errors Due to Approximation of the Bilinear Forms

As preparation for the derivation of *a priori* error estimates, we derive some upper bounds on the errors caused by the piecewise constant approximation of the coefficients α, $\boldsymbol{\beta}$, and γ. We use the following well-known notation. If V and W are normed linear spaces, then $\mathbf{L}(V, W; \mathbf{R})$ is the space of bounded bilinear forms on V and W, the standard norm of an element $b \in \mathbf{L}(V, W; \mathbf{R})$ is given by

$$\|b\|_{\mathbf{L}(V, W; \mathbf{R})} = \sup_{v \in V} \sup_{w \in W} \frac{|b(v, w)|}{\|v\|_V \|w\|_W} .$$

Lemma 7. *If $\alpha \in W_1^\infty(\Omega)$, then*

$$\|a - \overline{a}_q\|_{\mathbf{L}(\mathbf{H}^1(\Omega), (X_h \|\cdot\|_h); \mathbf{R})} \leq 6h_{\max}\|\alpha\|_{W_1^\infty(\Omega)} ,$$

where $(X_h, \|\cdot\|_h)$ is a normed linear space with, as elements, the elements of X_h but with $\|\cdot\|_h$ as norm.

Proof. From Eqs. (16c) and (21b) it follows that

$$|a(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) - \overline{a}(\boldsymbol{\sigma}, \boldsymbol{\tau}_h)| \leq 4h_{\max}\|\alpha\|_{W_1^\infty(\Omega)}\|\boldsymbol{\sigma}\|_{\mathbf{L}^2(\Omega)}\|\boldsymbol{\tau}_h\|_h .$$

When combined with Lemma 6, this implies

$$\|a - \overline{a}_q\|_{L(\mathbf{H}^1(\Omega), (X_h, \|\cdot\|_h); \mathbf{R})} \leq 6h_{\max}\|\alpha\|_{W_1^\infty(\Omega)} .$$

■

Lemma 8. *If $\boldsymbol{\beta} \in \mathbf{W}_1^\infty(\Omega)$, then*

$$\|b - \overline{b}_l\|_{L(\mathbf{L}^2(\Omega), L^2(\Omega); \mathbf{R})} \leq 4h_{\max}\|\boldsymbol{\beta}\|_{\mathbf{W}_1^\infty(\Omega)} .$$

Proof. This follows immediately from (16c). ■

Lemma 9. *If $\gamma \in W_1^\infty(\Omega)$, then*

$$\|c - \overline{c}_q\|_{L(\mathbf{L}^2(\Omega), L^2(\Omega); \mathbf{R})} \leq 2h_{\max}\|\gamma\|_{W_1^\infty(\Omega)} .$$

Proof. This follows immediately from (16c). ■

B. An *a priori* Error Estimate

The following two lemmas show nice properties of our discretization, which are needed to derive the error bound.

Lemma 10. *Let* $\boldsymbol{\tau} \in \Sigma$, $t \in L^2(\Omega)$, *then*

$$\overline{b}(\tilde{\Pi}_h\boldsymbol{\tau}, t - P_h t) - (\text{div } \tilde{\Pi}_h\boldsymbol{\tau}, t - P_h t) = 0\,. \tag{26}$$

Proof. A straightforward calculation shows that $\mathbf{P}_h(\boldsymbol{\beta}) \cdot \tilde{\Pi}_h\boldsymbol{\tau} - \text{div } \tilde{\Pi}_h\boldsymbol{\tau}$ is constant on Ω_k. From this, (26) easily follows. ■

Lemma 11. *If* $(\boldsymbol{\sigma}, u)$ *is a solution of* (6) *and* $(\boldsymbol{\sigma}_h, u_h)$ *is a solution of* (19), *then*

$$(\text{div}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), P_h t) + c(u - u_h, P_h t) = 0 \quad \forall\, t \in \mathrm{L}^2(\Omega)\,. \tag{27}$$

Proof. We take (19b),

$$(\text{div } \boldsymbol{\sigma}_h, P_h t) + \overline{c}(u_h, P_h t) = (f, P_h t)\,,$$

$\overline{c}$ is derived by orthogonal $\mathrm{L}^2(\Omega_k)$ projection, so this implies

$$(\text{div } \boldsymbol{\sigma}_h, P_h t) + c(u_h, P_h t) = (f, P_h t)\,.$$

If we subtract this from (6b), $(\text{div } \boldsymbol{\sigma}, P_h t) + c(u, P_h t) = (f, P_h t)$, then we find (27). ■

We are now ready to give an estimate for $\|\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h$.

Theorem 1. *If* $(\boldsymbol{\sigma}, u)$ *is the solution of* (6), $(\boldsymbol{\sigma}_h, u_h)$ *is the solution of* (19) *and* $(\boldsymbol{\sigma}, u) \in \mathbf{H}^1(\Omega) \times H^2(\Omega)$, *then there exist positive real numbers C and D such that*

$$C \leq \frac{12}{A} \max(1, \|\alpha\|_{\mathrm{W}_1^\infty(\Omega)}, \|\boldsymbol{\beta}\|_{\mathrm{W}_1^\infty(\Omega)}, \|\gamma\|_{\mathrm{W}_1^\infty(\Omega)}) \max(1, \|\boldsymbol{\sigma}\|_{\mathrm{H}^1(\Omega)}, \|u\|_{\mathrm{L}^2(\Omega)})\,, \tag{28}$$

$$D \leq 2\frac{\|\boldsymbol{\beta}\|_{\mathrm{L}^\infty(\Omega)}}{A}\,,$$

$$\begin{aligned}\|\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h^2 \leq{}& Ch_{\max}(\|\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h + \|P_h u - u_h\|_{\mathrm{L}^2(\Omega)})\\ &+ D\|\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h\|P_h u - u_h\|_{\mathrm{L}^2(\Omega)}\,.\end{aligned}$$

Proof. According to (24d), $A\|\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_h^2 \leq \overline{a}_q(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h))$. This is the starting point for the derivation of our error bound. Equations (6a) and (19a) imply that

$$\begin{aligned}\overline{a}_q(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)) ={}& (\overline{a}_q - a)(\boldsymbol{\sigma}, \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)) + a(\boldsymbol{\sigma}, \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h))\\ &- \overline{a}_q(\boldsymbol{\sigma}_h, \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h))\\ ={}& (\overline{a}_q - a)(\boldsymbol{\sigma}, \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)) + (\text{div } \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), u)\\ &- b(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), u) + \langle g, \mathbf{n}_{\partial\Omega} \cdot \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\rangle\\ &+ \overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), u_h) - (\text{div } \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), u_h)\\ &- \langle g, \mathbf{n}_{\partial\Omega} \cdot \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\rangle\\ ={}& (\overline{a}_q - a)(\boldsymbol{\sigma}, \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)) + (\text{div } \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), u)_{\mathrm{L}^2(\Omega)}\\ &- (b - \overline{b})(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), u) - \overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), u)\\ &+ \overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), u_h) - (\text{div } \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), u_h)\,,\end{aligned}$$

where we give $b - \overline{b}$, $\overline{a}_q - a$, etc., their obvious meaning. If we use Lemma 10, we find

$$\begin{aligned}A\|\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_h^2 \leq{}& (\overline{a}_q - a)(\boldsymbol{\sigma}, \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)) - (b - \overline{b})(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), u)\\ &+ (\text{div } \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), P_h u - u_h)_{\mathrm{L}^2(\Omega)} - \overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), P_h u - u_h)\,.\end{aligned}$$

If we use (20b) and (20c) to prepare the way, then the application of Lemma 11 to this expression results in

$$A\|\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_h^2 \le (\overline{a}_q - a)(\boldsymbol{\sigma}, \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)) - (b - \overline{b})(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), u)$$
$$- c(u - u_h, P_h u - u_h) - \overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), P_h u - u_h).$$

As γ is nonnegative according to (5), we may add $c(P_h u - u_h, P_h u - u_h)$ on both sides of the inequality, we find

$$A\|\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_h^2 + c(P_h u - u_h, P_h u - u_h) \le (\overline{a}_q - a)(\boldsymbol{\sigma}, \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h))$$
$$-(b - \overline{b})(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), u) - (c - \overline{c})(u - P_h u, P_h u - u_h)$$
$$- \overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), P_h u - u_h).$$

We use Lemmas 7, 8, and 9 to reduce this to

$$A\|\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_h^2 \le h_{\max}(6\|\alpha\|_{\mathrm{W}_1^\infty(\Omega)}\|\boldsymbol{\sigma}\|_{\mathrm{H}^1(\Omega)} + 4\|\boldsymbol{\beta}\|_{\mathrm{W}_1^\infty(\Omega)}\|u\|_{\mathrm{L}^2(\Omega)})\,\|\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_h$$
$$+ 2h_{\max}\|\gamma\|_{\mathrm{W}_1^\infty(\Omega)}\|u - P_h u\|_{\mathrm{L}^2(\Omega)}\|P_h u - u_h\|_{\mathrm{L}^2(\Omega)}$$
$$+ 2\|\boldsymbol{\beta}\|_{\mathrm{L}^\infty(\Omega)}\|\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_h\|P_h u - u_h\|_{\mathrm{L}^2(\Omega)}.$$

Note that for all $u \in \mathrm{L}^2(\Omega)$, $\|u - P_h u\|_{\mathrm{L}^2(\Omega)} \le \|u\|_{\mathrm{L}^2(\Omega)}$, and $\|\Pi_h\boldsymbol{\sigma}\|_h = \|\tilde{\Pi}_h\boldsymbol{\sigma}\|_h$. ■

Next, we prepare for the second part of our error estimate.

Lemma 12. *If* $(\boldsymbol{\sigma}, u)$ *is the solution of* (6), $(\boldsymbol{\sigma}_h, u_h)$ *is a solution of* (19), *and* $(\boldsymbol{\tau}, q)$ *is the solution of the adjoint problem for an arbitrary right-hand side* $p \in L^2(\Omega)$, *then*

$$(\operatorname{div} \boldsymbol{\tau}, P_h u - u_h) - b(\boldsymbol{\tau}, P_h u - u_h) = a(\boldsymbol{\sigma}, \tilde{\Pi}_h\boldsymbol{\tau}) - \overline{a}_q(\boldsymbol{\sigma}_h, \tilde{\Pi}_h\boldsymbol{\tau})$$
$$+ (b - \overline{b})(\tilde{\Pi}_h\boldsymbol{\tau}, u) + \overline{b}(\tilde{\Pi}_h\boldsymbol{\tau} - \boldsymbol{\tau}, P_h u - u_h) + (\overline{b} - b)(\boldsymbol{\tau}, P_h u - u_h).$$

Proof. We start by replacing b by $\overline{b}$,

$$(\operatorname{div} \boldsymbol{\tau}, P_h u - u_h) - b(\boldsymbol{\tau}, P_h u - u_h) = (\operatorname{div} \boldsymbol{\tau}, P_h u - u_h) - \overline{b}(\boldsymbol{\tau}, P_h u - u_h)$$
$$+ (\overline{b} - b)(\boldsymbol{\tau}, P_h u - u_h).$$

We use (20a) and (20c) to get

$$(\operatorname{div} \boldsymbol{\tau}, P_h u - u_h) - b(\boldsymbol{\tau}, P_h u - u_h) = (\operatorname{div} \tilde{\Pi}_h\boldsymbol{\tau}, P_h u - u_h) - \overline{b}(\tilde{\Pi}_h\boldsymbol{\tau}, P_h u - u_h)$$
$$+ \overline{b}(\tilde{\Pi}_h\boldsymbol{\tau} - \boldsymbol{\tau}, P_h u - u_h) + (\overline{b} - b)(\boldsymbol{\tau}, P_h u - u_h),$$

Lemma 10 to find

$$(\operatorname{div} \boldsymbol{\tau}, P_h u - u_h) - b(\boldsymbol{\tau}, P_h u - u_h) = (\operatorname{div} \tilde{\Pi}_h\boldsymbol{\tau}, u - u_h) - \overline{b}(\tilde{\Pi}_h\boldsymbol{\tau}, u - u_h)$$
$$+ \overline{b}(\tilde{\Pi}_h\boldsymbol{\tau} - \boldsymbol{\tau}, P_h u - u_h) + (\overline{b} - b)(\boldsymbol{\tau}, P_h u - u_h),$$

and Eqs. (6a) and (19a) to produce

$$(\operatorname{div} \boldsymbol{\tau}, P_h u - u_h) - b(\boldsymbol{\tau}, P_h u - u_h) = a(\boldsymbol{\sigma}, \tilde{\Pi}_h\boldsymbol{\tau}) - \langle g, \tilde{\Pi}_h\boldsymbol{\tau} \cdot \mathbf{n}_{\partial\Omega}\rangle - \overline{a}_q(\boldsymbol{\sigma}_h, \tilde{\Pi}_h\boldsymbol{\tau})$$
$$+ \langle g, \tilde{\Pi}_h\boldsymbol{\tau} \cdot \mathbf{n}_{\partial\Omega}\rangle + (b - \overline{b})(\tilde{\Pi}_h\tau, u) + \overline{b}(\tilde{\Pi}_h\boldsymbol{\tau} - \boldsymbol{\tau}, P_h u - u_h)$$
$$+ (\overline{b} - b)(\boldsymbol{\tau}, P_h u - u_h).$$

■

Lemma 13. *If* $(\boldsymbol{\sigma}, u)$ *is the solution of* (6), $(\boldsymbol{\sigma}_h, u_h)$ *is a solution of* (19), *and* $(\boldsymbol{\tau}, w)$ *is the solution of the adjoint problem for an arbitrary right-hand side* $p \in L^2(\Omega)$, *then*

$$c(P_h w, u - u_h) = -a(\tau, \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)) + \bar{b}(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), w - P_h w).$$

Proof. From Lemma 11,

$$c(P_h w, u - u_h) = -(\mathrm{div}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), P_h w),$$

and according to (20b) and (20c) we can rewrite the right-hand side,

$$c(P_h w, u - u_h) = -(\mathrm{div}\ \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), P_h w).$$

We wish to use Eq. (26) from Lemma 10 to remove P_h. To do this, we must add and subtract a term $\bar{b}(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), P_h w)$ on the right-hand side of our equation. We apply Lemma 10 and gather terms in $\bar{b}$ together,

$$c(P_h w, u - u_h) = -(\mathrm{div}\ \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), w) + \bar{b}(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), w - P_h w).$$

Finally, we use (11a),

$$c(P_h w, u - u_h) = -a(\tau, \tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)) + \bar{b}(\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h), w - P_h w).$$

■

Theorem 2. *Assume the adjoint problem* (11) *has a unique solution for all square integrable right-hand sides, and assume that there is a constant* C_r *such that, if* $(\boldsymbol{\tau}, w)$ *is the solution of* (11) *for a given right-hand side* f, *then*

$$\|\boldsymbol{\tau}\|_{\mathrm{H}^1(\Omega)} + \|w\|_{\mathrm{H}^1(\Omega)} \le C_r \|f\|_{\mathrm{L}^2(\Omega)}.$$

Now, if $(\boldsymbol{\sigma}, u) \in \mathbf{H}^1(\Omega) \times H^2(\Omega)$ *is the solution of* (6), *and* $(\boldsymbol{\sigma}_h, u_h)$ *is a solution of* (19), *then there are constants*

$$0 < C, D, E \le 4C_r(1 + 2h_{\max}) \max(\|\alpha\|_{\mathrm{W}_1^\infty(\Omega)}, \|\boldsymbol{\beta}\|_{\mathrm{W}_1^\infty(\Omega)}, \|\gamma\|_{\mathrm{W}_1^\infty(\Omega)}),$$

such that

$$\|P_h u - u_h\|_{\mathrm{L}^2(\Omega)} \le C h_{\max}(\|u\|_{\mathrm{L}^2(\Omega)} + \|\boldsymbol{\sigma}\|_{\mathrm{H}^1(\Omega)}) + D h_{\max} \|\tilde{\Pi}_h(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|_h + E h_{\max} \|P_h u - u_h\|_{\mathrm{L}^2(\Omega)}.$$

Proof. If we have an estimate for $(P_h u - u_h, p)$ for all $p \in \mathrm{L}^2(\Omega)$, then we can use

$$\|t\|_{\mathrm{L}^2(\Omega)} = \sup_{p \in \mathrm{L}^2(\Omega), p \neq 0} \frac{(p, t)}{\|p\|_{\mathrm{L}^2(\Omega)}},$$

to find $\|P_h u - u_h\|_{\mathrm{L}^2(\Omega)}$. We use the regularity of the adjoint problem (11) to find a solution $(\boldsymbol{\tau}, w) \in \mathbf{H}^1(\Omega) \times \mathrm{L}^2(\Omega)$ of (11) for a given right-hand side $p \in \mathrm{L}^2(\Omega)$. We may write

$$(p, P_h u - u_h) = (\mathrm{div}\ \boldsymbol{\tau}, P_h u - u_h) - b(\boldsymbol{\tau}, P_h u - u_h) + c(w, P_h u - u_h).$$

If we apply Lemma 12, we find

$$\begin{aligned}(p,P_hu-u_h)&=a(\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\tau})-\overline{a}_q(\boldsymbol{\sigma}_h,\tilde{\Pi}_h\boldsymbol{\tau})+(b-\overline{b})(\tilde{\Pi}_h\boldsymbol{\tau},u)\\&\quad+\overline{b}(\tilde{\Pi}_h\boldsymbol{\tau}-\boldsymbol{\tau},P_hu-u_h)+(\overline{b}-b)(\boldsymbol{\tau},P_hu-u_h)\\&\quad+c(w-P_hw,P_hu-u_h)+c(P_hw,P_hu-u_h).\end{aligned}$$

Using Lemma 13, we find that

$$\begin{aligned}(p,P_hu-u_h)&=a(\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\tau})-\overline{a}_q(\boldsymbol{\sigma}_h,\tilde{\Pi}_h\boldsymbol{\tau})+(b-\overline{b})(\tilde{\Pi}_h\boldsymbol{\tau},u)\\&\quad+\overline{b}(\tilde{\Pi}_h\boldsymbol{\tau}-\boldsymbol{\tau},P_hu-u_h)+(\overline{b}-b)(\boldsymbol{\tau},P_hu-u_h)\\&\quad+c(w-P_hw,P_hu-u_h)-a(\boldsymbol{\tau},\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h))\\&\quad+\overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),w-P_hw),\end{aligned}$$

which can be written as

$$\begin{aligned}(p,P_hu-u_h)&=(a-\overline{a}_q)(\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\tau})+\overline{a}_q(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h,\tilde{\Pi}_h\boldsymbol{\tau})+(b-\overline{b})(\tilde{\Pi}_h\tau,u)\\&\quad+\overline{b}(\tilde{\Pi}_h\boldsymbol{\tau}-\boldsymbol{\tau},P_hu-u_h)+(\overline{b}-b)(\boldsymbol{\tau},P_hu-u_h)\\&\quad+c(w-P_hw,P_hu-u_h)-a(\boldsymbol{\tau},\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h))\\&\quad+\overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),w-P_hw).\end{aligned}$$

We use (24a) to write this as

$$\begin{aligned}(p,P_hu-u_h)&=(a-\overline{a}_q)(\boldsymbol{\sigma},\tilde{\Pi}_h\boldsymbol{\tau})-(a-\overline{a}_q)(\boldsymbol{\tau},\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h))+(b-\overline{b})(\tilde{\Pi}_h\tau,u)\\&\quad+\overline{b}(\tilde{\Pi}_h\boldsymbol{\tau}-\boldsymbol{\tau},P_hu-u_h)+(\overline{b}-b)(\boldsymbol{\tau},P_hu-u_h)\\&\quad+c(w-P_hw,P_hu-u_h)+\overline{b}(\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h),w-P_hw).\end{aligned}$$

Use of the regularity of the adjoint problem (11), Lemmas 7, 8, and 9, and the projection error estimates (16a,b,c) leads to

$$\begin{aligned}\|P_hu-u_h\|_{L^2(\Omega)}&\le C_r(1+2h_{\max})2h_{\max}\|\alpha\|_{W_1^\infty(\Omega)}(\|\boldsymbol{\sigma}\|_{\mathbf{H}^1(\Omega)}+\|\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|_h)\\&\quad+4C_rh_{\max}\|\boldsymbol{\beta}\|_{\mathbf{W}_1^\infty(\Omega)}(1+2h_{\max})\|u\|_{L^2(\Omega)}+2C_rh_{\max}\|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)}\|P_hu-u_h\|_{L^2(\Omega)}\\&\quad+2C_rh_{\max}(h_{\max}\|\gamma\|_{W_1^\infty(\Omega)}\|P_hu-u_h\|_{L^2(\Omega)}+\|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)}\|\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|_{L^2(\Omega)}).\end{aligned}$$

This can be written as

$$\begin{aligned}\|P_hu-u_h\|_{L^2(\Omega)}&\le\tilde{C}h_{\max}(1+h_{\max})+\tilde{D}h_{\max}(1+h_{\max})\|\tilde{\Pi}_h(\boldsymbol{\sigma}-\boldsymbol{\sigma}_h)\|_h\\&\quad+\tilde{E}h_{\max}(1+h_{\max})\|P_hu-u_h\|_{L^2(\Omega)}.\end{aligned}$$

■

If $h_{\max}$ is small enough, Theorems 1 and 2 together give an $\mathbf{O}(h_{\max})$ error estimate.

An important limit on $h_{\max}$ is implied by the form of the estimates in Theorems 1 and 2. The main problem is that large values of $\|\alpha\|_{W_1^\infty(\Omega)}$, $\|\boldsymbol{\beta}\|_{W_1^\infty(\Omega)}$, and $\|\gamma\|_{W_1^\infty(\Omega)}$ decrease the range of $h_{\max}$ for which the estimate is valid. Specifically, in Theorem 2 we need to bring the term with coefficient E to the left-hand side, so we need, for example,

$$4h_{\max}C_r(1+2h_{\max})\max(\|\alpha\|_{W_1^\infty(\Omega)},\|\boldsymbol{\beta}\|_{W_1^\infty(\Omega)},\|\gamma\|_{W_1^\infty(\Omega)})\le 1/2.$$

Note that C_r, the condition of the dual problem, may also depend on α.

If the above inequality holds, then Theorem 2 implies

$$\|P_h u - u_h\|_{L^2(\Omega)} \leq 2Ch_{\max}(\|u\|_{L^2(\Omega)} + \|\boldsymbol{\sigma}\|_{H^1(\Omega)}) + 2Dh_{\max}\|\tilde{\Pi}_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h,$$

with C, D as in Theorem 2. Insertion of this estimate in Theorem 1 leads to a similar problem. Here we need an assumption of the form

$$16h_{\max}C_r(1 + 2h_{\max}) \max(\|\alpha\|_{W_1^\infty(\Omega)}, \|\boldsymbol{\beta}\|_{W_1^\infty(\Omega)}, \|\gamma\|_{W_1^\infty(\Omega)}) \frac{\|\boldsymbol{\beta}\|_{L^\infty(\Omega)}}{A} \leq 1/2$$

to bring the square of the error from the right-hand side to the left-hand side. Theorem 1 then gives

$$\|\tilde{\Pi}_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h^2 \leq 2Kh_{\max}(\|\tilde{\Pi}_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h + 2Ch_{\max}(\|u\|_{L^2(\Omega)} + \|\boldsymbol{\sigma}\|_{H^1(\Omega)}) \\ + 2Dh_{\max}\|\tilde{\Pi}_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h),$$

with K the constant C from Theorem 1 and C, D as in Theorem 2. If $\|\tilde{\Pi}_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h$ is smaller than $h_{\max}$, then we find

$$\|\tilde{\Pi}_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h^2 \leq 2Kh_{\max}(h_{\max} + 2Ch_{\max}(\|u\|_{L^2(\Omega)} + \|\boldsymbol{\sigma}\|_{H^1(\Omega)}) + 2Dh_{\max}),$$

i.e., an $O(h_{\max})$ error estimate for $\|\tilde{\Pi}_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h$ and, if it is larger than $h_{\max}$, then we may divide terms in the right-hand side either by $\|\tilde{\Pi}_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h$ or by $h_{\max}$, so

$$\|\tilde{\Pi}_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h \leq 2Kh_{\max}(1 + 2C(\|u\|_{L^2(\Omega)} + \|\boldsymbol{\sigma}\|_{H^1(\Omega)}) + 2Dh_{\max}),$$

so again we find an $O(h_{\max})$ error estimate for $\|\tilde{\Pi}_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_h$.

C. Another Approach

To improve our estimate of $\|P_h u - u_h\|_{L^2(\Omega)}$, we consider the adjoint of the discrete problem. This means, that we look for $(\boldsymbol{\tau}_h, v_h) \in X_h \times W_h$, such that

$$\bar{a}_q(\boldsymbol{\tau}_h, \boldsymbol{\sigma}_h) - (\operatorname{div} \boldsymbol{\sigma}_h, v_h) = 0 \quad \forall\, \boldsymbol{\sigma}_h \in V_h, \tag{29a}$$

$$(\operatorname{div} \boldsymbol{\tau}_h, t_h) - \bar{b}(\boldsymbol{\tau}_h, t_h) + \bar{c}(v_h, t_h) = (f, t_h) \quad \forall\, t_h \in W_h. \tag{29b}$$

We call this system regular, if there is at least one solution for each $f \in P_h(L^2(\Omega))$, and that all solutions for a particular f satisfy

$$\|\boldsymbol{\tau}_h\|_h + \|v_h\|_{L^2(\Omega)} \leq C\|P_h f\|_{L^2(\Omega)}, \tag{29c}$$

with C independent of the mesh size. This is a somewhat less stringent regularity condition than that given for the continuous adjoint problem (10). Note, that $\boldsymbol{\tau}_h \in X_h$, so $\boldsymbol{\tau}_{h,i}$ is a piecewise exponential function on Ω_k for $i = 1, 2$.

An example of a general condition under which this system is regular is the following:

$$\alpha \geq A > 0, \qquad \gamma \geq C_0 > 0, \qquad \text{and} \quad AC_0 - \|\boldsymbol{\beta}\|_{L^\infty(\Omega)}^2 \geq C_1 > 0. \tag{30}$$

To show this, we need the following relations:

$$\int_\Omega \frac{P_h(\alpha)}{4}\boldsymbol{\tau}_h \cdot \boldsymbol{\tau}_h - \mathbf{P}_h(\boldsymbol{\beta}) \cdot \boldsymbol{\tau}_h v_h + P_h(\gamma) v_h v_h \, d\mu = \tag{31}$$

$$\int_\Omega \frac{P_h(\alpha)}{4}\left(\boldsymbol{\tau}_h - \frac{2\mathbf{P}_h(\boldsymbol{\beta})}{P_h(\alpha)} v_h\right)^2 + \left(P_h(\gamma) - \frac{\mathbf{P}_h(\boldsymbol{\beta})^2}{P_h(\alpha)}\right) v_h v_h \, d\mu \geq \tag{31a}$$

$$\int_\Omega P_h(\gamma)\left(v_h - \frac{\mathbf{P}_h(\boldsymbol{\beta}) \cdot \boldsymbol{\tau}_h}{2P_h(\alpha)}\right)^2 + \left(P_h(\alpha) - \frac{\mathbf{P}_h(\boldsymbol{\beta})^2}{P_h(\gamma)}\right) \frac{\boldsymbol{\tau}_h \cdot \boldsymbol{\tau}_h}{4} \, d\mu\,. \tag{31b}$$

We know, that (div $\tilde{\Pi}_h\boldsymbol{\sigma}, P_h t$) = (div $\Pi_h\boldsymbol{\sigma}, P_h t$), so, if we take the sum of (29a) and (29b) with $\boldsymbol{\sigma} = \Pi_h\boldsymbol{\tau}_h$ and $t = v_h$, we find

$$\overline{a}_q(\boldsymbol{\tau}_h, \Pi_h\boldsymbol{\tau}_h) - \overline{b}(\boldsymbol{\tau}_h, v_h) + \overline{c}(v_h, v_h) = (f, v_h)\,. \tag{32}$$

According to (24a), $\overline{a}_q(\boldsymbol{\tau}_h, \Pi_h\boldsymbol{\tau}_h) = \overline{a}_q(\tilde{\Pi}_h\boldsymbol{\tau}_h, \tilde{\Pi}_h\boldsymbol{\tau}_h)$, and by (24b) we have

$$\frac{1}{4}\overline{a}(\tilde{\Pi}_h\boldsymbol{\sigma}, \tilde{\Pi}_h\boldsymbol{\sigma}) \leq \overline{a}_q(\tilde{\Pi}_h\boldsymbol{\sigma}, \tilde{\Pi}_h\boldsymbol{\sigma})\,.$$

Hence,

$$\int_\Omega \frac{P_h(\alpha)}{4}\boldsymbol{\tau}_h \cdot \boldsymbol{\tau}_h - \mathbf{P}_h(\boldsymbol{\beta}) \cdot \boldsymbol{\tau}_h v_h + P_h(\gamma) v_h v_h \, d\mu \leq \int_\Omega P_h(f) v_h \, d\mu\,. \tag{33}$$

This expression is identical to (31), so (31a) is smaller than (f, v_h); combined with (30), this implies

$$\frac{C_1}{A}\|v_h\|_{\mathrm{L}^2(\Omega)} \leq \|f\|_{\mathrm{L}^2(\Omega)}\,. \tag{34a}$$

In the same way, we find that (31b) is smaller than (f, v_h); together with (30) and (34a), this implies

$$\frac{C_1}{(AC_0)^{1/2}}\|\boldsymbol{\tau}_h\|_{\mathrm{L}^2(\Omega)} \leq \|f\|_{\mathrm{L}^2(\Omega)}\,. \tag{34b}$$

From (32) we see that this implies

$$A\|\boldsymbol{\tau}_h\|_h^2 \leq a_q(\boldsymbol{\tau}_h, \boldsymbol{\tau}_h) \leq \|f\|_{\mathrm{L}^2(\Omega)}\|v_h\|_{\mathrm{L}^2(\Omega)} + \|\boldsymbol{\beta}\|_{\mathrm{L}^\infty(\Omega)}\|\boldsymbol{\tau}_h\|_{\mathrm{L}^2(\Omega)}\|v_h\|_{\mathrm{L}^2(\Omega)}$$
$$+ \|\gamma\|_{\mathrm{L}^\infty(\Omega)}\|v_h\|^2_{\mathrm{L}^2(\Omega)}\,,$$

this implies that there is a C such that

$$\|\boldsymbol{\tau}_h\|_h \leq C\|f\|_{\mathrm{L}^2(\Omega)}\,.$$

Theorem 3. *If we assume that* (29c) *holds, then*

$$\|P_h u - u_h\|_{\mathrm{L}^2(\Omega)} \leq h_{\max}(6\|\alpha\|_{\mathrm{W}_1^\infty(\Omega)}\|\boldsymbol{\sigma}\|_{\mathbf{H}^1(\Omega)} + 2(\|\boldsymbol{\beta}\|_{\mathrm{W}_1^\infty(\Omega)} + \|\gamma\|_{\mathrm{W}_1^\infty(\Omega)}\|u\|_{\mathrm{L}^2(\Omega)})\,. \tag{35}$$

Proof. We use (29b),

$$(P_h u - u_h, P_h f) = (\mathrm{div}\ \boldsymbol{\tau}_h, P_h u - u_h) - \overline{b}(\boldsymbol{\tau}_h, P_h u - u_h) + \overline{c}(P_h u - u_h, v_h)\,.$$

Hence, according to Lemma 10 and the definition of $\overline{c}$,

$$(P_h u - u_h, P_h f) = (\mathrm{div}\ \boldsymbol{\tau}_h, u - u_h) - \overline{b}(\boldsymbol{\tau}_h, u - u_h) + \overline{c}(u - u_h, v_h)\,.$$

We use (6a) and (19a) to find

$$\begin{aligned}(P_h u - u_h, P_h f) &= (\operatorname{div} \boldsymbol{\tau}_h, u - u_h) - (\bar{b} - b)(\boldsymbol{\tau}_h, u) - b(\boldsymbol{\tau}_h, u) + \bar{b}(\boldsymbol{\tau}_h, u_h)\\ &\quad + \bar{c}(u - u_h, v_h)\\ &= (\bar{b} - b)(\boldsymbol{\tau}_h, u) + a(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) - \bar{a}_q(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + (\bar{c} - c)(u, v_h)\\ &\quad + c(u - u_h, v_h).\end{aligned}$$

According to (24a) and Lemma 11, this implies

$$\begin{aligned}(P_h u - u_h, P_h f) &= (\bar{b} - b)(\boldsymbol{\tau}_h, u) + (a - \bar{a}_q)(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) + \bar{a}_q(\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\tau}_h)\\ &\quad + (\bar{c} - c)(u, v_h) - (\operatorname{div}(\Pi_h \boldsymbol{\sigma} - \boldsymbol{\sigma}_h), v_h).\end{aligned}$$

Now, (29a) implies

$$(P_h u - u_h, P_h f) = (\bar{b} - b)(\boldsymbol{\tau}_h, u) + (a - \bar{a}_q)(\boldsymbol{\sigma}, \boldsymbol{\tau}_h) + (\bar{c} - c)(u, v_h).$$

Finally, we use Lemmas 7, 8, and 9 and (29c) to obtain our error estimate (35). ■

VII. VERIFICATION OF THE LOCAL MAXIMUM PRINCIPLE

We use the discrete adjoint problem to show that, for this quadrature rule, the matrix after elimination of $\boldsymbol{\sigma}$ by static condensation is an M-matrix. The discrete adjoint problem is defined in (29).

We assume a regular uniform mesh. We denote the matrix corresponding to (29), after elimination of $\boldsymbol{\sigma}_h$, by A. We see that the matrix A has nonpositive off-diagonal elements. We shall show that A is an M-matrix. To do this, we use Theorem 5.12, Chapter 5, page 124 of [15]. This theorem states that, for irreducible matrices with nonpositive off-diagonal elements, the M-matrix property is equivalent to the existence of a positive vector with a nonnegative image that is not identically zero. In our case, the vector $(1, 1, \ldots, 1)^T$ has such an image, because all row sums are nonnegative, and any row corresponding to an edge or corner has a positive row-sum.

The fact that the matrix A is irreducible follows from Theorem 3.6 [15], which states that, for a square matrix, irreducibility is equivalent to its di-graph being strongly connected. Inspection shows that the di-graph of the matrix under consideration is indeed strongly connected.

According to Theorem 5.6 [15], A^T is an M-matrix, too. This implies that the discrete equations for the original u_h satisfy a local maximum principle.

The M-matrix property implies that the system for u_h has a unique solution. From the form of the equations for $\boldsymbol{\sigma}_h$, we see that a given u_h induces a unique $\boldsymbol{\sigma}_h$. This implies that our system is always uniquely solvable. A quick calculation of the coefficients of u_h in (19a) shows that, for constant coefficients and large $\boldsymbol{\beta}$, i.e., with large convection diffusion ratios, we get a relation between σ_h and u_h, where the "upwind" point is weighed more heavily. If $\boldsymbol{\beta}/\alpha$ remains bounded and we go to the limit $|\beta_1| + |\beta_2| \rightarrow \infty$, then we get a first-order upwind scheme. This suggests that the scheme, in which the coefficients are continuously dependent on this ratio, remains useful close to such a limit.

VIII. A POSTERIORI ESTIMATOR

We use a special quadrature rule and obtain a higher-order discretization. We seek an $\overline{a}_{h,3}(\cdot,\cdot)$, that minimizes $\overline{a} - \overline{a}_{h,3}$. To do this, we choose a special quadrature rule for each $\overline{a}(\cdot, \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is one of the basis functions introduced earlier. Due to the nature of our test functions, the quadrature rule is essentially a one-dimensional rule.

A. Derivation of the Rule

For $\boldsymbol{\eta}_{i,j+1/2}$, we proceed as follows. We replace the two-dimensional integral by a repeated integral, we integrate exactly in the $\mathbf{e}_2$ direction, and then use a three-point rule to approximate the remaining integral. As nodes for the last integration, we take either the centers of $\Gamma_{i-1/2,j+1/2,0}$, $\Gamma_{i+1/2,j+1/2,0}$, and $\Gamma_{i+1/2,j+1/2,1}$. Or, if we are at a boundary, the edge center on the boundary and the two next closest edge centers. We choose the weights as follows:

$$\alpha_h(\Pi_h \boldsymbol{\sigma}, \boldsymbol{\eta}_{i,j+1/2}) = \alpha(\boldsymbol{\sigma}, \boldsymbol{\eta}_{i,j+1/2}),$$

for all $\boldsymbol{\sigma}$ with x_1-components that are second-order polynomials in x_1, i.e., for all $a, b, c \in \mathbf{R}$, and all $\boldsymbol{\eta}_{i,j+1/2}$, we have

$$\alpha_h(\Pi_h((ax_1^2 + bx_1 + c)\mathbf{e}_1), \boldsymbol{\eta}_{i,j+1/2}) = \alpha((ax_1^2 + bx_1 + c)\mathbf{e}_1, \boldsymbol{\eta}_{i,j+1/2}).$$

In a similar manner, we define the rule for $\boldsymbol{\eta}_{i+1/2,j}$.

B. Estimator for the Local Discretization Error and a Lower Bound for the Global Error

We use this rule to obtain an *a posteriori* estimator for the local discretization error and a lower bound for the global error. It is immediately obvious that

$$\overline{a}_{h,3}(\boldsymbol{\sigma}, \boldsymbol{\eta}_r) - \overline{a}_{h,1}(\boldsymbol{\sigma}, \boldsymbol{\eta}_r) \geq \mathrm{O}(h_{\max}^2),$$

where r is a possible index-tuple. Moreover,

$$\overline{a}(\boldsymbol{\sigma}, \boldsymbol{\eta}_r) - \overline{a}_{h,3}(\boldsymbol{\sigma}, \boldsymbol{\eta}_r) = \mathrm{O}(h_{\max}^3),$$

if $\boldsymbol{\sigma}$ is smooth enough. If

$$|\overline{a}_{h,3}(\boldsymbol{\rho}_h, \boldsymbol{\eta}_r) - (\operatorname{div} \boldsymbol{\eta}_r - \boldsymbol{\beta} \cdot \boldsymbol{\eta}_r, w_h)| \geq K,$$

then we have either

$$\|w_h\|_{\mathrm{L}^\infty(\Omega)} \geq C_1 K,$$

or

$$\|\boldsymbol{\rho}_h\|_{\mathrm{L}^\infty(\Omega)} \geq C_2 K.$$

We see immediately that, if $(\boldsymbol{\sigma}_h, u_h)$ is the solution of (19) with $\overline{a}_q = \overline{a}_{h,1}$, then

$$\overline{a}_{h,1}(\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\eta}_r) - ((\mathrm{div} - \boldsymbol{\beta})\boldsymbol{\eta}_r, P_h u - u_h) = \mathbf{O}(h^k),$$

with $k = 1$ or 2 depending on the coefficients in (1) and

$$\overline{a}_{h,3}(\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{\eta}_r) - ((\mathrm{div} - \boldsymbol{\beta})\boldsymbol{\eta}_r, P_h u - u_h) = \mathbf{O}(h_{\max}^{k+2}) + \overline{a}_{h,1}(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r) - \overline{a}_{h,3}(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r).$$

So, $(a_{h,1} - a_{h,3})(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r)$ is an estimate for the local discretization error. Moreover, this implies that there is a constant C such that

$$\|\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{L^2(\Omega)} + \|P_h u - u_h\|_{L^2(\Omega)} \geq C|\overline{a}_{h,1}(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r) - \overline{a}_{h,3}(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r)| + \mathbf{O}(h_{\max}^{k+2}).$$

If we assume that

$$\|\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{L^2(\Omega)} + \|P_h u - u_h\|_{L^2(\Omega)} = \mathbf{O}(h_{\max}^{k}),$$

we see that, for $h_{\max}$ small enough,

$$\|\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{L^2(\Omega)} + \|P_h u - u_h\|_{L^2(\Omega)} \geq \frac{1}{2} C|\overline{a}_{h,1}(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r) - \overline{a}_{h,3}(\boldsymbol{\sigma}_h, \boldsymbol{\eta}_r)|.$$

This provides a lower bound on the global discretization error. We expect the solution for $\overline{a}_{h,3}$ to be two orders of magnitude more accurate than the solution for $\overline{a}_{h,1}$.

IX. NUMERICAL RESULTS

We consider problem (1) with

$$u = \tanh(\alpha(\sqrt{2}x_1 - ((\sqrt{2} - 1)/2 + x_2))),$$

$$\alpha = 100, \qquad \beta_1 = 100, \qquad \beta_2 = 100\sqrt{2}, \qquad \Gamma_1 = \partial\Omega, \qquad g = u|_{\partial\Omega},$$

$$f = -\frac{\mathrm{div}(\mathbf{grad}\ u + u\boldsymbol{\beta})}{\alpha}.$$

The smallest upper bound on $h_{\max}$ now follows from the second condition given in Section VIB. This condition reduces to

$$3200h_{\max}C_r(1 + h_{\max}) < 1/2,$$

so, neglecting the factor $1 + h_{\max}$, for the theorems to apply we would need $h_{\max} \leq 1/(3200C_r)$.

We find the following results for the two discretizations. The directional components of the error vectors for the fluxes were the equal to the accuracy given.

The $\log_2$ of the errors for $\overline{a}_q = \overline{a}_{h,1}$

Meshwidth	$\log_2\|P_h u - u_h\|$	$\log_2\|(\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot\mathbf{e}_1\|$	$\log_2\|(\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot\mathbf{e}_2\|$
1/4	−1.3	−1.3	−1.1
1/8	−1.8	−1.7	−1.4
1/16	−2.5	−2.4	−2.1
1/32	−3.6	−3.4	−3.1
1/64	−5.1	−5.0	−4.6

The log$_2$ of the errors for $\overline{a}_q = \overline{a}_{h,3}$			
Meshwidth	$\log_2\|P_h u - u_h\|$	$\log_2\|(\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot \mathbf{e}_1\|$	$\log_2\|(\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot \mathbf{e}_2\|$
1/4	−2.2	−2.9	−2.6
1/8	−3.5	−4.0	−3.6
1/16	−5.4	−6.1	−5.6
1/32	−8.0	−9.2	−8.7
1/64	−10.5	−13.0	−12.4

We see that the order of convergence is indeed higher for the second method. We also see that the difference in order for the fluxes approaches 2. Deviations from the expected order may be caused by the steepness of the solution relative to the mesh.

To test the stability of the low order method, we applied it to problem (1) with

$$u = \tanh(\alpha(\sqrt{2}x_1 - ((\sqrt{2} - 1)/2 + x_2))),$$

with

$$\beta_1 = \alpha, \qquad \beta_2 = \alpha\sqrt{2},$$

for $\alpha = 1000$ and $\alpha = 1000000$, with

$$\Gamma_1 = \partial\Omega, \qquad g = u|_{\partial\Omega}, \qquad f = -\frac{\operatorname{div}(\mathbf{grad}\ u + u\boldsymbol{\beta})}{\alpha}.$$

This gave the following results for $\alpha = 1000$.

The log$_2$ of the errors for $\overline{a}_q = \overline{a}_{h,1}$			
Meshwidth	$\log_2\|P_h u - u_h\|$	$\log_2\|(\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot \mathbf{e}_1\|$	$\log_2\|(\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot \mathbf{e}_2\|$
1/4	−0.8	−1.0	−0.8
1/8	−1.0	−1.2	−0.8
1/16	−1.3	−1.4	−1.0
1/32	−1.6	−1.7	−1.2
1/64	−2.0	−2.0	−1.5

And for $\alpha = 1000000$.

The log$_2$ of the errors for $\overline{a}_q = \overline{a}_{h,1}$			
Meshwidth	$\log_2\|P_h u - u_h\|$	$\log_2\|(\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot \mathbf{e}_1\|$	$\log_2\|(\Pi_h\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot \mathbf{e}_2\|$
1/4	−0.8	−1.0	−0.8
1/8	−1.0	−1.2	−0.8
1/16	−1.3	−1.4	−1.0
1/32	−1.5	−1.6	−1.2
1/64	−1.8	−1.8	−1.4

The bound given in Section VIB reduces to

$$32\alpha h_{\max} C_r(1 + h_{\max}) < 1/2,$$

so, neglecting the factor $1 + h_{\max}$, for the theorem to apply we need $h_{\max} \leq 1/(32\alpha C_r)$. Evidently, this criterion is not met; nevertheless, convergence occurs, because the scheme reduces to an upwind discretization in cases of large α, i.e., for small diffusion constants.

X. CONCLUSIONS

The Petrov–Galerkin mixed finite element method with exponentially fitted test functions for the fluxes has several nice properties. For instance, just as for a finite volume method, if the true solution $\boldsymbol{\sigma}$ is divergence-free, then the same holds for $\boldsymbol{\sigma}_h$. Furthermore, we have a formal *a priori* error estimate, and, after elimination of $\boldsymbol{\sigma}_h$ by static condensation, the two-dimensional discretization results in an M-matrix for u_h. We can extend the method to three dimensions without additional difficulties. Section IX suggests that the scheme with the three-point quadrature rule $\overline{a}_{h,3}$ can serve as a source for *a posteriori* error estimates. To judge the effectiveness of the method for singularly perturbed problems is very difficult. However, the fact that it incorporates exponential fitting, copes well with the exponential solution of the constant coefficient case, and approaches a two-dimensional upwind scheme if the convection goes to infinity suggests that the method based on $\overline{a}_{h,1}$ can be applied to such problems.

References

1. D. L. Scharfetter and H. K. Gummel, "Large-signal analysis of a silicon read diode oscillator," *IEEE Transactions on Electron Devices,* Vol. ED-16, **1**, 64 (1969).
2. Siegfried Selberherr, *Analysis and Simulation of Semiconductor Devices,* Springer–Verlag, Wien, New York, 1984.
3. Peter A. Markowich, *The Stationary Semiconductor Device Equations,* Springer–Verlag, Wien, New York, 1986.
4. Wolfgang Fichtner, Donald J. Rose, and Randolph E. Bank, "Numerical methods for semiconductor device simulation," *SIAM J. Sci. Stat. Comp.* **4**, 416 (1983).
5. F. Brezzi, L. D. Marini, and P. Pietra, "Mixed exponential fitting schemes for current continuity equations," in *Proceedings of the Sixth International NASECODE Conference,* J. H. Miller, Ed., Press Ltd., Dublin, 1989.
6. P. W. Hemker, *A Numerical Study of Stiff Two-Point Boundary Problems,* Mathematical Centre Tracts, 80, Mathematical Centre, Amsterdam, 1977.
7. J. Douglas, Jr. and J. E. Roberts, "Global estimates for mixed methods for second order elliptic equations," *Math. Comp.* **44**, 39 (1985).
8. Robert A. Adams, *Sobolev Spaces,* Academic Press, New York, 1975.
9. V. Girault and P. Raviart, *Finite Element Methods for Navier–Stokes Equations,* Springer–Verlag, Berlin, 1986.
10. S. J. Polak, C. den Heijer, H. A. Schilders, and P. Markowich, "Semiconductor device modelling for the numerical point of view," *Int. J. Numer. Meth. Eng.* **24**, 763 (1987).
11. Jean E. Roberts and Jean–Marie Thomas, "Mixed and Hybrid Finite Element Methods," RR 737, *INRIA,* Rocquencourt, October 1987.
12. P. A. Raviart and J. M. Thomas, "A mixed finite element for 2-nd order elliptic problems," in *Mathematical Aspects of the Finite Element Method,* Lecture Notes in Mathematics, Springer, **606**, 292 (1977).
13. J. C. Nedelec, "Mixed Finite Elements in $\mathbf{R}^3$," *Numer. Math.* **35**, 315 (1980).
14. P. G. Ciarlet and P. A. Raviart, "General Lagrange and Hermite interpolation in $\mathbf{R}^n$ with applications to finite element methods," *Arch. Rational Mech. Anal.* **46**, 177 (1972).
15. M. Fiedler, *Special Matrices and Their Applications in Numerical Mathematics,* Martinus Nijhoff, Dordrecht, 1986.