COMPARISON OF VARIATIONAL DISCRETIZATIONS FOR A CONVECTION-DIFFUSION PROBLEM

CONSTANTIN BACUTA, CRISTINA BACUTA, and DANIEL HAYES

Communicated by Gabriela Marinoschi

For a model convection-diffusion problem, we obtain new error estimates for a general upwinding finite element discretization based on bubble modification of the test space. The key analysis tool is finding representations of the optimal norms on the trial spaces at the continuous and discrete levels. We analyze and compare three methods: the standard linear discretization, the saddle point least square and the upwinding Petrov–Galerkin methods. We conclude that the bubble upwinding Petrov–Galerkin method is the most performant discretization for the one-dimensional model. Our results for the model convection-diffusion problem can be extended for creating new and efficient discretizations for the multi-dimensional cases.

AMS 2020 Subject Classification: 35K57, 65N12, 65N22, 65N30, 74S05.

Key words: Petrov–Galerkin, upwinding, convection dominated problem, singularly perturbed problems.

1. INTRODUCTION

We consider the model of a singularly perturbed convection diffusion problem: Given data represented by $f \in L^2(\Omega)$, we look for a solution to the problem

$$\left\{ \begin{array}{rcl} -\varepsilon \Delta u + b \cdot \nabla u &=& f \quad \text{in} \quad \Omega, \\ u &=& 0 \quad \text{on } \partial\Omega, \end{array} \right. \tag{1}$$

for a positive constant ε and a bounded domain $\Omega \subset \mathbb{R}^d$. We assume $\varepsilon \ll 1$, and b is a given vector chosen such that a unique solution exists.

For the one-dimensional case, we assume that f is a continuous function on $[0,1]$, and we look for a solution $u = u(x)$ such that

$$\begin{cases} -\varepsilon u''(x) + b\, u'(x) = f(x), \quad 0 < x < 1 \\ u(0) = 0, \ u(1) = 0, \end{cases} \tag{2}$$

The authors' work on this material was supported by NSF-DMS 2011615.

where b is a positive constant. Without loss of generality, we further assume that $b = 1$. The model problems (1) and (2) arise when solving heat transfer problems on thin domains, as well as when using small step sizes in implicit time discretizations of parabolic convection diffusion type problems, see [27]. The solutions of these two problems are characterized by boundary layers, see e.g., [24, 28, 32, 34]. Approximating such solutions poses numerical challenges due to the ε-dependence of the stability constants and of the error estimates. There is a tremendous amount of literature addressing these types of problems, see e.g. [24, 28, 31, 32, 34, 21, 5, 14]. In this paper, we analyze mixed variational discretizations of the model convection diffusion problem (2), based on the concept of optimal trial norms at the continuous and the discrete levels. The concept of optimal trial norm was developed and used before, in e.g., [3]–[5, 18, 19, 21, 23, 30]. In our study, for certain discrete test spaces, we find new representations of such norms that allow for sharp error estimates and new analysis for saddle point or mixed variational formulations.

We start by reviewing the standard finite element discretization and two mixed variational formulations that are known as the Saddle Point Least Square (SPLS) and the Upwinding Petrov–Galerkin (UPG) methods. We present new error analysis results for the UPG method and discuss the advantages and disadvantages of the two mixed methods. The goal of the paper is to develop a set of tools and ideas for robust discretization of (2) towards building efficient new methods for the multi-dimensional version of convection dominated problems, such as (1).

In Section 2, we review the main concepts and notation for the general standard and mixed variational formulation and discretization. The general concept of optimal trial space and the main related results about optimal trial norms is reviewed in Section 3. We review approximation results for the standard linear and SPLS discretizations of (2) on uniformly distributed nodes in Section 4. We justify the *oscillatory behavior* of the SPLS method for certain data, in Section 4.3. We present a general approximation result for the UPG method in Section 5. In Section 6, we apply the general approximation result of Section 5 to particular test spaces constructed with quadratic bubbles and exponential type bubbles. In Section 7, we present a summary of the ideas and a conclusion for the standard and mixed variational formulation of (2).

2. THE GENERAL MIXED VARIATIONAL APPROACH

In this section, we review the main concepts and notation for the mixed variational formulation and discretization. This includes the Saddle Point Least Squares (SPLS) method and the particular case of the Petrov–Galerkin

(PG) discretization. We follow the Saddle Point Least Squares (SPLS) terminology that was introduced in [7]–[9, 10, 12, 13].

2.1. **The abstract variational formulation at the continuous level**

We consider the abstract mixed formulation: Find $u \in Q$ such that

$$b(v,u) = \langle F, v\rangle, \text{ for all } v \in V, \tag{3}$$

where $b(\cdot,\cdot)$ is a bilinear form, Q and V are possible different separable Hilbert spaces, and F is a continuous linear functional on V. We denote the dual of V by V^* and the dual pairing on $V^* \times V$ by $\langle\cdot,\cdot\rangle$. We assume that the inner products $a_0(\cdot,\cdot)$ and $(\cdot,\cdot)_Q$ induce the norms $|\cdot|_V = |\cdot| = a_0(\cdot,\cdot)^{1/2}$ and $\|\cdot\|_Q = \|\cdot\| = (\cdot,\cdot)_Q^{1/2}$. The bilinear form $b(\cdot,\cdot)$ is a continuous bilinear form on $V \times Q$ satisfying the $\sup - \sup$ condition

$$\sup_{u\in Q}\sup_{v\in V}\frac{b(v,u)}{|v|\,\|u\|} = M < \infty, \tag{4}$$

and the $\inf - \sup$ condition

$$\inf_{u\in Q}\sup_{v\in V}\frac{b(v,u)}{|v|\,\|u\|} = m > 0. \tag{5}$$

We assume that the functional $F \in V^*$ satisfies the *compatibility condition*

$$\langle F, v\rangle = 0 \quad \text{for all } v \in V_0 := \{v \in V |\ b(v,q) = 0 \quad \text{for all } q \in Q\}. \tag{6}$$

The following result about the existence and the uniqueness of the solution of (3) can be found in e.g., [1, 2, 16, 17].

Proposition 2.1. *If the form $b(\cdot,\cdot)$ satisfies* (4) *and* (5)*, and the data $F \in V^*$ satisfies the compatibility condition* (6)*, then the problem* (3) *has a unique solution that depends continuously on the data F.*

It is also known, see e.g., [11]–[13, 21], that under the *compatibility condition* (6), solving the mixed problem (3) reduces to solving a standard saddle point reformulation: Find $(w,u) \in V \times Q$ such that

$$\begin{array}{lcll} a_0(w,v) & + \quad b(v,u) & = \langle F, v\rangle & \text{for all } v \in V, \\ b(w,q) & & = 0 & \text{for all } q \in Q. \end{array} \tag{7}$$

In fact, we have that u is the unique solution of (3) *if and only if* $(w = 0, u)$ solves (7).

2.2. **PG and SPLS discretizations**

Let $b(\cdot,\cdot) : V \times Q \to \mathbb{R}$ be a bilinear form as defined in Section 2.1. Let $V_h \subset V$ and $\mathcal{M}_h \subset Q$ be finite-dimensional approximation spaces. We assume that the following discrete $\inf - \sup$ condition holds for the pair of spaces $(V_h, \mathcal{M}_h)$:

$$\inf_{u_h \in \mathcal{M}_h} \sup_{v_h \in V_h} \frac{b(v_h, u_h)}{|v_h|\,\|u_h\|} = m_h > 0. \tag{8}$$

We define

$$V_{h,0} := \{v_h \in V_h \,|\, b(v_h, q_h) = 0, \quad \text{for all } q_h \in \mathcal{M}_h\},$$

and let $F_h \in V_h^*$ to be the restriction of F to V_h, i.e., $\langle F_h, v_h\rangle := \langle F, v_h\rangle$ for all $v_h \in V_h$. Consider the following discrete compatibility condition

$$\langle F, v_h\rangle = 0 \quad \text{for all } v_h \in V_{h,0}. \tag{9}$$

As a direct consequence of Proposition (2.1), we have the following result.

PROPOSITION 2.2. *If the form $b(\cdot,\cdot)$ satisfies condition* (8) *on $V_h \times \mathcal{M}_h$, and the data $F_h \in V_h^*$ satisfies the compatibility condition* (9)*, then the problem of finding $u_h \in \mathcal{M}_h$ such that*

$$b(v_h, u_h) = \langle F, v_h\rangle, \;\; v_h \in V_h, \tag{10}$$

has a unique solution $u_h \in \mathcal{M}_h$ that depends continuously on the data F_h.

The variational formulation (10) is the *Petrov–Galerkin* (PG) discretization of (3). We note that for the case $V_{h,0} = \{0\}$, the compatibility condition (9) is trivially satisfied. In this case, assuming that $b(\cdot,\cdot)$ satisfies (8), the discretization (10) leads to a square linear system. Thus, we do not need to consider the SPLS discretization of (3).

In general, $V_{h,0}$ might not be a subset of V_0. Consequently, even though the continuous problem (3) has unique solution, the discrete problem (10) might not be well-posed if F_h does not satisfy the *compatibility condition* (9). However, if the form $b(\cdot,\cdot)$ satisfies (8) on $V_h \times \mathcal{M}_h$, then the problem of finding $(w_h, u_h) \in V_h \times \mathcal{M}_h$ satisfying

$$\begin{array}{lcll} a_0(w_h, v_h) & + \; b(v_h, u_h) & = \langle f, v_h\rangle & \text{for all } v_h \in V_h, \\ b(w_h, q_h) & & = 0 & \text{for all } q_h \in \mathcal{M}_h, \end{array} \tag{11}$$

does have a unique solution. We call the component u_h of the solution (w_h, u_h) of (11) the *saddle point least squares* approximation of the solution u of the original mixed problem (3).

The following error estimate for $\|u - u_h\|$ was proved in [12].

THEOREM 2.3. *Let $b : V \times Q \to \mathbb{R}$ satisfy* (4) *and* (5) *and assume that $F \in V^*$ is given and satisfies* (6). *Assume that u is the solution of* (3) *and $V_h \subset V$, $\mathcal{M}_h \subset Q$ are chosen such that the discrete* $\inf - \sup$ *condition* (8) *holds. If (w_h, u_h) is the solution of* (11), *then the following error estimate holds:*

$$\|u - u_h\| \leq \frac{M}{m_h} \inf_{q_h \in \mathcal{M}_h} \|u - q_h\|. \tag{12}$$

Remark 2.4. We note that the estimate (12) holds true if u_h is, in particular, the unique PG solution of (10). This is due to the fact that, if u_h is the solution of (10), then $(0, u_h)$ is the unique solution of (11).

For our analysis of the PG discretization of (1), we have a norm $\|\cdot\|_*$ on Q and a different norm $\|\cdot\|_{*,h}$ on the discrete trial space $\mathcal{M}_h$. For this case, the following version of Theorem 2.3 was proved in [5].

THEOREM 2.5. *Let $|\cdot|$, $\|\cdot\|_*$ and $\|\cdot\|_{*,h}$ be the norms on V, Q, and $\mathcal{M}_h$, respectively, such that they satisfy* (4), (5), *and* (8). *Assume that for some constant $c_0 > 0$, we have*

$$\|v\|_* \leq c_0 \|v\|_{*,h} \qquad \text{for all } v \in Q. \tag{13}$$

Let u be the solution of (3), *and let u_h be the unique solution of problem* (10). *Then, the following error estimate holds:*

$$\|u - u_h\|_{*,h} \leq c_0 \frac{M}{m_h} \inf_{p_h \in \mathcal{M}_h} \|u - p_h\|_{*,h}. \tag{14}$$

3. OPTIMAL TRIAL NORM FOR THE CONVECTION DIFFUSION PROBLEM

We consider the variational formulation of (1): Find $u \in H_0^1(\Omega)$ such that

$$(\varepsilon \nabla u, \nabla v) + (b \cdot \nabla u, v) = (f, v) \quad \text{for all } v \in H_0^1(\Omega). \tag{15}$$

Define $V = Q = H_0^1(\Omega)$ and $b : V \times Q \to \mathbb{R}$, $F \in V^*$ by

$$b(v, u) := (\varepsilon \nabla u, \nabla v) + (b \cdot \nabla u, v), \quad \text{and} \quad \langle F, v \rangle := (f, v).$$

For the analysis purpose, we allow different norms on the test and trial spaces. On the test space $V := H_0^1(\Omega)$, we consider the norm induced by $a_0(u, v) := (\nabla u, \nabla v)$. We can represent the *antisymmetric part* in the *symmetric* $a_0(\cdot, \cdot)$ inner product. First, we define the representation operator $T : Q \to Q$ by

$$a_0(Tu, v) = (b \cdot \nabla u, v), \quad \text{for all } v \in V.$$

In the multi-dimensional case, we have that

$$|Tu| = \|b \cdot \nabla u\|_{H^{-1}(\Omega)} \leq \|b\| \|u\|_{L^2(\Omega)}.$$

For the one-dimensional case and $b = 1$, we have

$$-((Tu)'', q) = a_0(Tu, q) = (u', q), \text{ for all } q \in Q.$$

By solving the corresponding differential equation, one can find that

$$Tu = x\overline{u} - \int_0^x u(s)\, ds, \tag{16}$$

where $\overline{u} = \int_0^1 u(s)\, ds$. Thus, $(Tu)'(x) = \overline{u} - u(x)$ and

$$|Tu|^2 = \int_0^1 |u(s) - \overline{u}|^2\, ds = \|u - \overline{u}\|^2 = \|u\|^2 - \overline{u}^2 \leq \|u\|^2. \tag{17}$$

Next, the optimal continuous trial norm on Q is defined by

$$\|u\|_* := \sup_{v \in V} \frac{b(v,u)}{|v|} = \sup_{v \in V} \frac{\varepsilon a_0(u,v) + a_0(Tu,v)}{|v|}.$$

Using the Riesz representation theorem and the fact that $a_0(Tu, u) = 0$, we obtain that the optimal trial norm on Q is given by

$$\|u\|_*^2 = \varepsilon^2 |u|^2 + |Tu|^2. \tag{18}$$

Thus, we have

$$\|u\|_*^2 := \epsilon^2 (\nabla u, \nabla u) + \|b \cdot \nabla u\|_{H^{-1}}^2.$$

Using (17) for the one-dimensional case, we get

$$\|u\|_*^2 = \varepsilon^2 |u|^2 + \|u\|^2 - \overline{u}^2. \tag{19}$$

3.1. **Discrete optimal trial norm**

We assume that $V_h \subset V = H_0^1(\Omega)$ and $\mathcal{M}_h \subset Q = H_0^1(\Omega)$ are discrete finite element spaces and that $\mathcal{M}_h \subset V_h$. For the purpose of obtaining a discrete optimal norm on $\mathcal{M}_h$, we let $P_h : Q \to V_h$ be the standard elliptic projection defined by

$$a_0(P_h\, u, v_h) = a_0(u, v_h) \quad \text{for all } v_h \in V_h.$$

The optimal trial norm on $\mathcal{M}_h$ is

$$\|u_h\|_{*,h} := \sup_{v_h \in V_h} \frac{b(v_h, u_h)}{|v_h|}. \tag{20}$$

Similarly to the continuous case,

$$\|u_h\|_{*,h} := \sup_{v_h \in V_h} \frac{\varepsilon a_0(u_h, v_h) + a_0(Tu_h, v_h)}{|v_h|} = \sup_{v_h \in V_h} \frac{\varepsilon a_0(u_h, v_h) + a_0(P_h Tu_h, v)}{|v_h|}.$$

From the definition of P_h and the anti-symmetry of T, we have

$$a_0(P_h T u_h, u_h) = a_0(T u_h, u_h) = 0.$$

Thus, by using the Riesz representation theorem on V_h, we get

$$\|u_h\|_{*,h}^2 = \varepsilon^2 |u_h|^2 + |P_h T u_h|^2 := \varepsilon^2 |u_h|^2 + |u_h|_{*,h}^2. \tag{21}$$

Note that for the given trial spaces $\mathcal{M}_h$ and Q, the above norm is well defined for any $u \in Q$. Hence, the continuous and discrete optimal trial norms can be compared on Q.

The advantage of using the optimal trial norm on Q and $\mathcal{M}_h$ resides with the fact that both $inf - sup$ and $sup - sup$ are equal to one at both the continuous and the discrete levels. As a direct consequence of Theorem 3.1, we obtain the following result.

THEOREM 3.1. *Let $\|\cdot\|_*$ and $\|\cdot\|_{*,h}$ be the norms on Q, and $\mathcal{M}_h$ and assume that* (13) *holds. Let u be the solution of* (27) *and let u_h be the unique solution of problem* (10). *Then the following error estimate holds:*

$$\|u - u_h\|_{*,h} \leq c_0 \inf_{p_h \in \mathcal{M}_h} \|u - p_h\|_{*,h}. \tag{22}$$

3.2. **Discrete optimal trial norm for the one-dimensional case**

We review some formulas and results from [5, 15].

For $V = Q = H_0^1(0,1)$, we consider the standard inner product given by $a_0(u,v) = (u,v)_V = (u',v')$. We divide the interval $[0,1]$ into n equal length subintervals using the nodes $0 = x_0 < x_1 < \cdots < x_n = 1$ and denote $h := x_j - x_{j-1}, j = 1, 2, \ldots, n$. We define the corresponding finite element discrete space $\mathcal{M}_h$ as the space of all *continuous piecewise linear functions* with respect to the given nodes, that are zero at $x = 0$ and $x = 1$. Next, we let $\mathcal{M}_h = V_h$ be the standard space of continuous piecewise linear functions.

For the purpose of error analysis, on V_h we consider the standard norm induced by $a_0(\cdot,\cdot)$, but on $\mathcal{M}_h$ we choose an optimal norm from the stability point of view. On $V_h \times \mathcal{M}_h$, we consider the bilinear from

$$b_d(v_h, u_h) = d\, a_0(u_h, v_h) + (u_h', v_h) \quad \text{for all } u_h \in \mathcal{M}_h, v_h \in V_h, \tag{23}$$

where $d = d_{\varepsilon,h}$ is a constant that might depend on h and ε. The same arguments used in Section 3.1 to deduce the formula (21), can be used here with $\varepsilon = d$ to obtain

$$\|u_h\|_{*,h}^2 = \sup_{v_h \in \mathcal{M}_h} \frac{(b_d(w_h, u_h))^2}{|w_h|^2} = d^2 |u_h|^2 + |P_h T u_h|^2. \tag{24}$$

Denoting $|u|_{*,h} := |P_h T u|$, as in [5], we obtain the explicit formula

$$|u|^2_{*,h} := |P_h T u|^2 = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{h}\int_{x_{i-1}}^{x_i} u(x)\,dx\Big)^2 - \Big(\int_0^1 u(x)\,dx\Big)^2. \tag{25}$$

Using a Poincare inequality, we get

$$\|u\|^2_* - \Big(\varepsilon^2 + \frac{h^2}{\pi^2}\Big)|u|^2 = \|u-\overline{u}\|^2 - \frac{h^2}{\pi^2}|u|^2 \le |u|^2_{*,h} \le \|u\|^2, \tag{26}$$

see [5] for details.

4. STANDARD AND SPLS FINITE ELEMENT VARIATIONAL FORMULATION AND DISCRETIZATION

In this section, we review results for the standard and the SPLS finite element discretization of (2). In addition, we justify the oscillatory behavior of the $P^1 - P^2$ SPLS discretization. We use the following notation:

$$a_0(u,v) = \int_0^1 u'(x)v'(x)\,dx,\ (f,v) = \int_0^1 f(x)v(x)\,dx,\ \text{and}$$
$$b(v,u) = \varepsilon\, a_0(u,v) + (u',v) \quad \text{for all } u,v \in V := H^1_0(0,1).$$

A variational formulation of (2), with $b = 1$, is as follows:

Find $u \in V := H^1_0(0,1)$ such that

$$b(v,u) = (f,v),\ \text{for all } v \in V = H^1_0(0,1). \tag{27}$$

4.1. Standard discretization with $C^0 - P^1$ test and trial spaces

We divide the interval $[0,1]$ into n equal length subintervals using the nodes $0 = x_0 < x_1 < \cdots < x_n = 1$ and denote $h := x_j - x_{j-1}, j = 1,2,\ldots,n$. For the above uniform distributed notes on $[0,1]$, we define the corresponding finite element discrete space $\mathcal{M}_h$ as the subspace of $H^1_0(0,1)$, given by

$$\mathcal{M}_h = \{v_h \in V \mid v_h \text{ is linear on each } [x_j, x_{j+1}]\},$$

i.e., $\mathcal{M}_h$ is the space of all *continuous piecewise linear functions* with respect to the given nodes, that are zero at $x = 0$ and $x = 1$. We consider the nodal basis $\{\varphi_j\}_{j=1}^{n-1}$ with the standard defining property $\varphi_i(x_j) = \delta_{ij}$. We couple the above discrete trial space with the discrete test space $V_h := \mathcal{M}_h$. Thus, the standard $C^0 - P^1$ variational formulation of (27) is: Find $u_h \in \mathcal{M}_h$ such that

$$b(v_h,u_h) = \varepsilon(u_h',v_h') + (u_h',w_h) = (f,v_h) \quad \text{for all } v_h \in V_h. \tag{28}$$

From (26) it is easy to obtain the following estimate

$$\|u\|_*^2 \le \Big(1 + \Big(\frac{h}{\pi\,\varepsilon}\Big)^2\Big)\|u\|_{*,h}^2 \quad \text{for all } u \in Q. \tag{29}$$

As a consequence of Theorem 3.1 and (29), we have the following result.

THEOREM 4.1. *If u is the solution of* (27)*, and u_h the solution of the linear discretization* (28)*, then*

$$\|u - u_h\|_{*,h} \le c(h,\varepsilon) \inf_{v_h \in V_h} \|u - v_h\|_{*,h}, \ \textit{where}$$

$$c(h,\varepsilon) = \sqrt{1 + \Big(\frac{h}{\pi\,\varepsilon}\Big)^2} \approx \frac{h}{\pi\,\varepsilon} \quad \textit{if} \ \ \varepsilon << h.$$

In the next sections, we show that the optimal discrete norm and $c(h,\varepsilon)$ improve as we consider different test spaces. Numerical tests for the case $\int_0^1 f(x)\,dx \neq 0$, show that as $\varepsilon << h$, the linear finite element solution of (28) presents non-physical oscillations, see [5]. The behavior of the standard linear finite element approximation of (28) motivates the use of non-standard discretization approaches, such as the *saddle point least square* or *Petrov–Galerkin* methods.

4.2. **SPLS discretization**

A *saddle point least square* (SPLS) approach for solving (27) has been used before, for example in [11, 21, 5].
For $V = Q = H_0^1(0,1)$, we look for finding $(w,u) \in V \times Q$ such that

$$\begin{array}{llll} a_0(w,v) & + \ b(v,u) & = (f,v) & \text{for all } v \in V, \\ b(w,q) & & = 0 & \text{for all } q \in Q, \end{array} \tag{30}$$

where

$$b(v,u) = \varepsilon\, a_0(u,v) + (u',v) = \varepsilon\,(u',v') + (u',v).$$

Numerical tests for the discretization of (30) with various degree polynomial test and trail spaces were done in [21, 22]. Following [5], we review the main error analysis results for $\mathcal{M}_h = C^0 - P^1 := \text{span}\{\varphi_j\}_{j=1}^{n-1}$, with the standard linear nodal functions φ_j, and $V_h = C^0 - P^2$ on given uniformly distributed nodes on $[0,1]$. To define a basis for V_h, we consider a bubble function for each interval $[x_{i-1}, x_i], i = 1,2,\ldots,n$, defined by

$$B_i := 4\,\varphi_{i-1}\,\varphi_i, \quad i = 1,2,\ldots,n,$$

where $\varphi_0(x) = 1 - \frac{x}{h}$ on $[0,h]$, $\varphi_n(x) = 1 + \frac{x-1}{h}$ on $[1-h,1]$ and are extended by zero to the rest of the interval $[0,1]$. Then, we have

$$V_h := \text{span}\{\varphi_j\}_{j=1}^{n-1} + \text{span}\{B_j\}_{j=1}^{n}.$$

The SPLS discretization of (30) is: Find $(w_h, u_h) \in V_h \times \mathcal{M}_h$ such that

$$\begin{array}{lcll} a_0(w_h, v_h) & + & b(v_h, u_h) & = (f, v_h) \qquad \text{for all } v_h \in V_h, \\ b(w_h, q_h) & & & = 0 \qquad \text{for all } q_h \in \mathcal{M}_h. \end{array} \tag{31}$$

In this case, note that the projection P_h defined in Section 3.1, is the projection on the space $V_h = C^0 - P^2$. For any piecewise linear function $u_h \in \mathcal{M}_h$, we have that

$$Tu_h = x\overline{u}_h - \int_0^x u_h(s)\, ds$$

is a continuous piecewise quadratic function. Consequently, $Tu_h \in V_h$, and $P_h\, Tu_h = Tu_h$. The optimal discrete norm on $\mathcal{M}_h$ becomes

$$\|u_h\|^2_{*,h} = \varepsilon^2 |u_h|^2 + |Tu_h|^2 = \|u_h\|^2_*.$$

Using the optimal norm on $\mathcal{M}_h$, a discrete inf $-$ sup condition is satisfied, and the problem (31) has a unique solution. In addition, for this $P^1 - P^2$ SPLS discretization, we can consider the same norm given by

$$\|u\|^2_* = \varepsilon^2 |u|^2 + \|u - \overline{u}\|^2 = \varepsilon^2 |u|^2 + \|u\|^2 - \overline{u}^2 = \|u\|^2_{*,h}$$

on both spaces Q and $\mathcal{M}_h$. As a consequence of the approximation Theorem 2.3, we get the following optimal error estimate.

THEOREM 4.2. *If u is the solution of (27), and u_h is the SPLS solution for the $(P^1 - P^2)$ discretization, then*

$$\|u - u_h\|_* \leq \inf_{p_h \in \mathcal{M}_h} \|u - p_h\|_* \leq \|u - u_I\|_*,$$

where u_I is the interpolant of the exact solution on the uniformly distributed nodes on $[0, 1]$.

4.3. The oscillatory behavior of the $P^1 - P^2$ SPLS discretization

For $\int_0^1 f(x)\, dx = 0$, the $P^1 - P^2$ SPLS discretization improves on the standard linear discretization of (27) from both the error point of view, and from the presence of the non-physical oscillations point of view. In [5], a detailed numerical analysis and comparison concluded that for $\int_0^1 f(x)\, dx \neq 0$, the SPLS solution u_h approximates the shift by a constant of the solution u of (27). In addition, *non-physical oscillations* still appear in the plot of u_h at the ends of the interval $[0, 1]$. An explanation of this phenomenon can be done using simplified variational problems. More precisely, we can consider the following *continuous simplified* problem obtained from (27), by letting $\varepsilon \to 0$: Find $u \in Q = H^1_0(0, 1)$ such that

$$(u', v) = (f, v) \qquad \text{for all } v \in V = H^1_0(0, 1). \tag{32}$$

The problem is not well posed when $\int_0^1 f(x)\,dx \neq 0$. In order to have the existence and the uniqueness of the solution of (32), we can change the trial space Q to $L_0^2(0,1) := \{u \in L^2(0,1) |\ \int_0^1 u = 0\}$. Nevertheless, in this case, the solution space cannot see the boundary conditions of the original problem (2). On the other hand, the *discrete simplified* linear system obtained from (31) by letting $\varepsilon \to 0$ becomes: Find $(w_h, u_h) \in V_h \times \mathcal{M}_h$ such that

$$(33) \qquad \begin{array}{lcll} (w_h', v_h') & + & (u_h', v_h) & = (f, v_h) \qquad \text{for all } v_h \in V_h, \\ (w_h', q_h) & & & = 0 \qquad \text{for all } q_h \in \mathcal{M}_h, \end{array}$$

and has unique solution because a discrete inf $-$ sup condition holds when using the optimal trial norm on $\mathcal{M}_h$. Numerical tests in [5] showed that oscillation in the discrete simplified solution u_h of (33) predict oscillatory behavior of u_h-the SPLS discrete solution of (31). In fact, for $\varepsilon << h$, in the "eye ball measure", the two solutions are identical. Next, we justify why the component u_h of (33) oscillates in the case $\int_0^1 f(x)\,dx \neq 0$.

Let u be the solution of (32) with $Q = L_0^2(0,1)$ and $V = H_0^1(0,1))$ and let u_h be the second component of the solution of (33). It is easy to check that $u(x) = w(x) - \overline{w}$, where $w(x) = \int_0^x f(s)\,ds$. By eliminating w_h from the system (33), it follows that $u_h - \overline{u}_h$ is the L^2 projection of u onto $\overline{\mathcal{M}}_h := \{w_h - \overline{w}_h\ | w_h \in \mathcal{M}_h\}$. We note that $\overline{\mathcal{M}}_h$ is a space of continuous piecewise linear functions that have the same values at the end points of $[0,1]$, while u cannot have the same values at the end points if $\int_0^1 f(x)\,dx \neq 0$. This explains the *non-physical oscillations* of the SPLS discretization of (31).

In the next section, we present a particular SPLS discretization that is free of *non-physical oscillations*.

5. THE PETROV–GALERKIN METHOD WITH BUBBLE TYPE TEST SPACE

For improving the stability and approximability of the standard linear finite element approximation for solving (27), various Petrov–Galerkin discretizations were considered, see e.g., [5, 20, 29, 33, 34]. In this section, we analyze a general class of Upwinding Petrov–Galerkin (UPG) discretizations based on a bubble modification of the standard $C^0 - P^1$ test space. The idea is to define V_h by adding to each φ_j, a pair of polynomial bubble functions. According to Section 2.2.2 in [34], this idea was first suggested in [25] and used in the same year in [20] with quadratic bubble modification. The method is known in literature as *upwinding PG method* or *upwinding finite element method*, see [33, 34]. Next, we build on the description of UPG introduced in [15] emphasizing on a new error analysis of the method.

The standard variational formulation for solving (2) with $b = 1$, is: Find $u \in Q = H_0^1(0,1)$ such that

$$b(v,u) = \varepsilon\, a_0(u,v) + (u',v) = (f,v) \qquad \text{for all } v \in V = H_0^1(0,1). \tag{34}$$

A general Petrov–Galerkin method for solving (34) chooses a test space of type $V_h \subset V = H_0^1(0,1)$ that is different from the trial space $\mathcal{M}_h \subset Q = H_0^1(0,1)$.

For describing the general UPG discretization, we consider a continuous (bubble) function $B : [0,h] \to \mathbb{R}$ with the following properties:

$$B(0) = B(h) = 0, \tag{35}$$

$$\int_0^h B(x)\,dx = b_1 h, \text{ with } b_1 > 0. \tag{36}$$

$$\int_0^h (B'(x))^2\,dx = \frac{b_2}{h}, \text{ with } b_2 > 0. \tag{37}$$

By translating B, we generate n bubble functions that are locally supported. For $i = 1,2,\ldots,n$, we define $B_i : [0,1] \to \mathbb{R}$ by $B_i(x) = B(x - x_{i-1}) = B(x - (i-1)h)$ on $[x_{i-1}, x_i]$, and we extend it by zero to the entire interval $[0,1]$. Note that $B_1 = B$ on $[0,h]$. For $i = 1,2,\ldots,n$, we have

$$B_i(x_{i-1}) = B_i(x_i) = 0, \text{ and } B_i = 0 \text{ on } [0,1]\backslash(x_{i-1}, x_i), \tag{38}$$

$$\int_{x_{i-1}}^{x_i} B_i(x)\,dx = b_1 h, \text{ with } b_1 > 0, \tag{39}$$

and

$$\int_{x_{i-1}}^{x_i} (B_i'(x))^2\,dx = \frac{b_2}{h}. \tag{40}$$

Next, we consider a particular class of Petrov–Galerkin discretizations of the model problem (34) with trial space $\mathcal{M}_h = \text{span}\{\varphi_j\}_{j=1}^{n-1}$ and the test space V_h obtained by modifying M_h using the bubble functions B_i. We define the test space V_h by

$$V_h := \text{span}\{\varphi_j + (B_j - B_{j+1})\}_{j=1}^{n-1},$$

where $\{B_i\}_{i=1,\ldots,n}$ satisfy (38)-(40). We note that both $\mathcal{M}_h$ and V_h have the same dimension of $(n-1)$.

The upwinding Petrov–Galerkin discretization with general bubble functions for (2) with $b = 1$ is: Find $u_h \in \mathcal{M}_h$ such that

$$b(v_h, u_h) = \varepsilon\, a_0(u_h, v_h) + (u_h', v_h) = (f, v_h) \qquad \text{for all } v_h \in V_h. \tag{41}$$

As presented in [15], we show that the variational formulation (41) admits a reformulation that uses a new bilinear form defined on *standard linear finite*

element spaces. We let

$$u_h = \sum_{j=1}^{n-1} \alpha_j \varphi_j,$$

and consider a generic test function v_h defined by

$$v_h = \sum_{i=1}^{n-1} \beta_i \varphi_i + \sum_{i=1}^{n-1} \beta_i (B_i - B_{i+1}) = \sum_{i=1}^{n-1} \beta_i \varphi_i + \sum_{i=1}^{n} (\beta_i - \beta_{i-1}) B_i,$$

where, we define $\beta_0 = \beta_n = 0$. Next, we use the splitting of v_h in a linear part plus a bubble part:

$$v_h = w_h + B_h, \text{ with } w_h := \sum_{i=1}^{n-1} \beta_i \varphi_i \text{ and } B_h := \sum_{i=1}^{n} (\beta_i - \beta_{i-1}) B_i.$$

Based on formulas (38), (39) and (40), the fact that u_h', w_h' are constant on each of the intervals $[x_{i-1}, x_i]$, and that $w_h' = \frac{\beta_i - \beta_{i-1}}{h}$ on $[x_{i-1}, x_i]$, we obtain

$$(u_h', B_h) = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} u_h' (\beta_i - \beta_{i-1}) B_i = \sum_{i=1}^{n} h\, u_h'\, w_h' \int_{x_{i-1}}^{x_i} B_i = b_1 h \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} u_h' w_h'.$$

Thus,

$$(u_h', B_h) = b_1 h (u_h', w_h'), \quad \text{where } v_h = w_h + B_h. \tag{42}$$

In addition, since u_h' is constant on $[x_{i-1}, x_i]$, we have

$$(u_h', B_i') = u_h' \int_{x_{i-1}}^{x_i} B_i'(x)\, dx = \ 0 \text{ for all } i = 1, 2, \ldots, n.$$

Hence,

$$(u_h', B_h') = 0, \text{ for all } u_h \in \mathcal{M}_h, v_h = w_h + B_h \in V_h. \tag{43}$$

From (42) and (43), for any $u_h \in \mathcal{M}_h, v_h = w_h + B_h \in V_h$ we get

$$b(v_h, u_h) = (\varepsilon + b_1 h)\, (u_h', w_h') + (u_h', w_h). \tag{44}$$

Introducing the notation $d = d_{\varepsilon,h} = \varepsilon + h\, b_1$ and using the notation of Section 3.2, we have

$$b(v_h, u_h) = b_d(u_h, w_h), \text{ where } v_h = w_h + B_h, \text{ and } u_h, w_h \in \mathcal{M}_h. \tag{45}$$

Using (43) and (40), we note that for any $v_h = w_h + B_h \in V_h$ the energy norm

of v_h is a multiple of the energy of the linear part w_h. Indeed,

$$\begin{aligned}(v_h', v_h') &= (w_h' + B_h', w_h' + B_h') = (w_h', w_h') + (B_h', B_h') \\ &= (w_h', w_h') + \sum_{i=1}^{n} (\beta_i - \beta_{i-1})^2 (B_i', B_i') \\ &= (w_h', w_h') + b_2 h \sum_{i=1}^{n} \Big(\frac{\beta_i - \beta_{i-1}}{h} \Big)^2 \\ &= (w_h', w_h') + b_2 \sum_{i=1}^{n} \Big(\int_{x_{i-1}}^{x_i} (w_h')^2 \Big)^2 = (w_h', w_h') + b_2 (w_h', w_h').\end{aligned}$$

Consequently,

$$|v_h|^2 = (1 + b_2)|w_h|^2. \tag{46}$$

The formulas (45) and (46) lead to the following result.

THEOREM 5.1. *For the bilinear form $b(\cdot,\cdot)$ of (34) on $\mathcal{M}_h \times V_h$ with the bubble enriched test space V_h, the discrete optimal norm on $\mathcal{M}_h$ is given by*

$$\|u_h\|_{*,h}^2 = \frac{(\varepsilon + h\, b_1)^2}{1 + b_2} \, |u_h|^2 + \frac{1}{1 + b_2} \, |u_h|_{*,h}^2. \tag{47}$$

where $|u_h|_{*,h}^2$ is defined in (25).

Proof. Using the definition of $\|u_h\|_{*,h}$ along with the work of Section 3, we can reduce the supremum over V_h to a supremum over $\mathcal{M}_h$. Indeed, using the splitting $v_h = w_h + B_h$, the equations (45) and (46), we have

$$\|u_h\|_{*,h}^2 = \sup_{v_h \in V_h} \frac{(b(v_h, u_h))^2}{|v_h|^2} = \sup_{v_h \in V_h} \frac{(b_d(w_h, u_h))^2}{|v_h|^2} = \sup_{w_h \in \mathcal{M}_h} \frac{(b_d(w_h, u_h))^2}{(1 + b_2)|w_h|^2}.$$

Next, by combining this formula with (24) we obtain (47). □

PROPOSITION 5.2. *Assume that h is chosen such that*

$$\varepsilon^2 + \frac{h^2}{\pi^2} \le (\varepsilon + h\, b_1)^2. \tag{48}$$

Then, the following inequality between $\|u\|_*$ and $\|u\|_{*,h}$ holds on Q.

$$\|u\|_*^2 \le (1 + b_2) \, \|u\|_{*,h}^2, \ \text{for all } u \in Q = H_0^1(0,1). \tag{49}$$

Proof. Using the inequality (26) for $\|u\|_*$ and the formula (47) for $\|u\|_{*,h}$, we have

$$\|u\|_*^2 - (1 + b_2) \, \|u\|_{*,h}^2 \le \Big(\varepsilon^2 + \frac{h^2}{\pi^2} - (\varepsilon + h\, b_1)^2 \Big) |u|^2.$$

Now, under the assumption (48), we obtain (49). □

As a consequence, we have the following error estimate.

THEOREM 5.3. *If u is the solution of* (27)*, u_h the solution of the UPG formulation* (41)*, and h is chosen such that* (48) *holds, then*

$$\|u - u_h\|_{*,h} \leq \sqrt{1+b_2} \inf_{p_h \in \mathcal{M}_h} \|u - p_h\|_{*,h}. \tag{50}$$

Proof. The estimate is a direct consequence of the approximation Theorem 3.1 and Proposition 5.2. □

Remark 5.4. Based on (44), the linear system associated with the UPG method (41) is

$$\Big(\Big(\frac{\varepsilon}{h} + b_1\Big)S + C\Big)\,U = F_{pg}, \tag{51}$$

where $U, F_{pg} \in \mathbb{R}^{n-1}$ and

$$U := \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-1} \end{bmatrix}, \quad F_{PG} := \begin{bmatrix} (f,\varphi_1) \\ (f,\varphi_2) \\ \vdots \\ (f,\varphi_{n-1}) \end{bmatrix} + \begin{bmatrix} (f, B_1 - B_2) \\ (f, B_2 - B_3) \\ \vdots \\ (f, B_{n-1} - B_n) \end{bmatrix},$$

$$S = \text{tridiag}(-1, 2, -1), \text{ and } C = \text{tridiag}\Big(-\frac{1}{2}, 0, \frac{1}{2}\Big).$$

By using the notation $d = d_{\varepsilon,h} = \varepsilon + h\,b_1$, the matrix of the finite element system (51) is

$$M_{fe} = \text{tridiag}\Big(-\frac{d}{h} - \frac{1}{2},\ 2\,\frac{d}{h},\ -\frac{d}{h} + \frac{1}{2}\Big). \tag{52}$$

6. UPWINDING PG WITH PARTICULAR BUBBLE FUNCTIONS

6.1. Upwinding PG with quadratic bubble functions

We consider the model problem (27) with the following discrete space $\mathcal{M}_h = \text{span}\{\varphi_j\}_{j=1}^{n-1}$ and V_h a modification of $\mathcal{M}_h$ using *quadratic bubble functions*. The method can be found in e.g., [29]. In [15], we related the *quadratic bubble UPG* method to the general upwinding Finite Difference (FD) method and presented ways to improve the performance of upwinding FD methods. In this section, we establish error estimates for the quadratic bubble UPG method.

First, for a parameter $\beta > 0$, we define the bubble function B on $[0, h]$ by

$$B(x) = \frac{4\,\beta}{h^2} x(h - x).$$

Using the function B and the general construction of Section 5, we define the set of bubble functions $\{B_1, B_2, \ldots, B_n\}$ on $[0, 1]$ and

$$V_h := \text{span}\{\varphi_j + (B_j - B_{j+1})\}_{j=1}^{n-1}.$$

Elementary calculations show that (36) holds with $b_1 = \frac{2\beta}{3}$, and (37) holds with $b_2 = \frac{16\beta^2}{3}$. In this case, we have

$$d = d_{\varepsilon,h} = \varepsilon + h\, b_1 = \varepsilon + \frac{2\beta}{3} h, \text{ and } 1 + b_2 = \frac{19}{3}\beta^2.$$

According to (47), the optimal norm on $\mathcal{M}_h$ is given by

$$\|u_h\|_{*,h}^2 = \frac{3}{19\beta^2}\Big(\Big(\varepsilon + \frac{2\beta}{3}\, h\Big)^2 |u_h|^2 + |u_h|_{*,h}^2\Big). \tag{53}$$

In this case, we note that the restriction (48) is satisfied for any $h > 0$ if, for example, $\beta \geq \frac{\sqrt{3}}{2\,\pi} \approx 0.28$. As a consequence, we have the following result.

THEOREM 6.1. *If u is the solution of* (27)*, u_h the solution of the upwinding PG formulation* (41)*, with quadratic bubble test space and $\beta \geq \frac{\sqrt{3}}{2\,\pi}$, then by using the discrete norm* (53)*, we have*

$$\|u - u_h\|_{*,h} \leq \sqrt{\frac{19}{3}}\, \beta \inf_{p_h \in \mathcal{M}_h} \|u - p_h\|_{*,h}. \tag{54}$$

Equivalently, by squaring and rescaling the estimate (54)*, we have*

$$\begin{aligned} &\Big(\varepsilon + \frac{2\beta}{3}\, h\Big)^2 |u - u_h|^2 + |u - u_h|_{*,h}^2 \\ &\qquad \leq \frac{19}{3}\beta^2 \inf_{p_h \in \mathcal{M}_h} \Big(\Big(\varepsilon + \frac{2\beta}{3}\, h\Big)^2 |u - p_h|^2 + |u - p_h|_{*,h}^2\Big). \end{aligned} \tag{55}$$

For implementation purposes, according to (52), the matrix of the finite element system (51) with quadratic bubble upwinding is

$$M_{fe}^q = \text{tridiag}\Big(-\frac{\varepsilon}{h} - \frac{2\beta}{3} - \frac{1}{2},\ \frac{2\,\varepsilon}{h} + \frac{4\beta}{3},\ -\frac{\varepsilon}{h} - \frac{2\beta}{3} + \frac{1}{2}\Big). \tag{56}$$

We note that for $\beta = 0$, we obtain the matrix corresponding to the standard finite element discretization (28). The case $\beta = 1$ was studied in [5]. The possibility of choosing $\beta = \beta(\varepsilon, h)$ allows for further simplification.

6.2. **Special cases for quadratic bubble upwinding**

Using the settings of Section 6.1, we choose β such that the upper diagonal in the matrix M_{fe}^{q} of (56) is zero. This implies

$$\beta = \frac{3}{4}\Big(1 - \frac{2\,\varepsilon}{h}\Big).$$

To satisfy $\beta > 0$ and (48) for a fixed ε, we restrict the range for h to

$$h > 2.6\,\varepsilon.$$

This case is interesting because the matrix M of the FE system (51) becomes a *bidiagonal lower triangular matrix*

$$M = \text{tridiag}(-1, 1, 0). \tag{57}$$

As a direct consequence of Theorem 6.3 and $\varepsilon + \frac{2\beta}{3}\,h = h/2$, we have that the solution u_h of the upwinding PG formulation (41), satisfies

$$h^2\,|u - u_h|^2 + 4|u - u_h|_{*,h}^2 \leq \frac{57}{4} \inf_{p_h \in \mathcal{M}_h} \left(h^2\,|u - p_h|^2 + 4|u - p_h|_{*,h}^2\right). \tag{58}$$

In addition, the system $M\,U = F_{pg}$ can be solved forward to obtain:

$$u_j = (f, \varphi_1 + \varphi_2 + \cdots + \varphi_j) + (f, B_1 - B_{j+1}),\ j = 1, 2, \ldots n-1. \tag{59}$$

We introduce the nodal function φ_0 corresponding to $x_0 = 0$, i.e., φ_0 is the continuous piecewise linear function such that $\varphi_0(x_j) = \delta_{0,j}$, $j = 1, 2, \ldots n$. Using that $\varphi_0 + \varphi_1 + \cdots + \varphi_j = 1$ on $[0, x_j]$, the formula (59) leads to

$$u_j = \int_0^{x_j} f(x)\ dx + \int_0^{x_1} f(B_1 - \varphi_0)\ dx + \int_{x_j}^{x_{j+1}} f(\varphi_j - B_{j+1})\ dx, \tag{60}$$

where

$$B_1(x) = 3\Big(1 - \frac{2\,\varepsilon}{h}\Big)\Big(\frac{x}{h}\Big)\Big(1 - \frac{x}{h}\Big),\ x \in [0, h],$$

and

$$B_{j+1}(x) = B_1(x - jh),\ x \in [x_j, x_{j+1}],\ j = 1, 2, \ldots n-1.$$

The next result shows that the discrete solution $u_h = \sum_{j=1}^{n-1} u_j\,\varphi_j$ is close to the interpolant of $w(x) := \int_0^x f(t)\,dt$, hence it is *free of non-physical oscillations.*

THEOREM 6.2. *If $u_h = \sum_{j=1}^{n-1} u_j\,\varphi_j$ is the solution of the UPG formulation* (41)*, with quadratic bubble test space and $\beta = \frac{3}{4}\left(1 - \frac{2\varepsilon}{h}\right)$, then*

$$\left|u_j - \int_0^{x_j} f(x)\ dx\right| \leq \|f\|_\infty \Big(2 - \frac{2\,\varepsilon}{h}\Big)h,\ j = 1, 2, \ldots, n-1.$$

Proof. We note that

$$\int_0^{x_1} B_1\,dx = \int_{x_j}^{x_{j+1}} B_{j+1}\,dx = \Big(1-\frac{2\,\varepsilon}{h}\Big)\frac{h}{2} \text{ and } \int_0^{x_1} \varphi_1\,dx = \int_{x_j}^{x_{j+1}} \varphi_j = \frac{h}{2}.$$

Thus, assuming f is continuous on $[0,1]$, by using the formulas (60) and the triangle inequality, we have

$$\Big|u_j - \int_0^{x_j} f(x)\,dx\Big|$$
$$= \Big|\int_0^{x_1} f(B_1-\varphi_0)\,dx + \int_{x_j}^{x_{j+1}} f(\varphi_j - B_{j+1})\,dx\Big| \le \|f\|_\infty\Big(2-\frac{2\,\varepsilon}{h}\Big)h. \quad \square$$

Theorem 6.2 proves that the components u_j of the PG solution (60) approximate $w(x_j) = \int_0^{x_j} f(x)\,dx$ with $\mathcal{O}(h)$. If f is independent of ε, then w is independent of ε, and consequently, the PG solution given by (60) is *free of non-physical oscillations.*

6.3. **Upwinding PG with exponential bubble functions**

As presented in [15], we consider the model problem (27) with the discrete space $\mathcal{M}_h = \text{span}\{\varphi_j\}_{j=1}^{n-1}$ and a basis for V_h obtained by modifying the basis of $\mathcal{M}_h$ using *exponential bubble functions.* We define the bubble function B on $[0,h]$ as the solution of

$$-\varepsilon B'' - B' = 1/h,\ B(0) = B(h) = 0. \tag{61}$$

Using the function B and the general construction of Section 5, we define the set of bubble functions $\{B_1, B_2, \ldots, B_n\}$ on $[0,1]$ by translations of the function B. The test space V_h is defined by

$$V_h := \text{span}\{\varphi_j + (B_j - B_{j+1})\}_{j=1}^{n-1} = \text{span}\{g_j\}_{j=1}^{n-1}, \tag{62}$$

where $g_j := \varphi_j + (B_j - B_{j+1})$, $j = 1, 2, \ldots, n-1$. The idea of using a *local dual problem* for building the trial space is also presented in Section 2.2.3 of [33], where an earlier reference [26] is acknowledged. However, in [26, 33] the concept of *discrete Green's function* was used to produce basis functions that span the test space V_h. Here, we managed to build a basis for our test space V_h using the general construction of Section 5 with the bubble B defined in equation (61).

In order to deal with efficient computations of coefficients and the finite element matrix of the *exponential bubble UPG method*, we introduce the following notation

$$g_0 := \tanh\Big(\frac{h}{2\varepsilon}\Big) = \frac{e^{\frac{h}{2\varepsilon}} - e^{-\frac{h}{2\varepsilon}}}{e^{\frac{h}{2\varepsilon}} + e^{-\frac{h}{2\varepsilon}}} = \frac{1-e^{-\frac{h}{\varepsilon}}}{1+e^{-\frac{h}{\varepsilon}}}, \tag{63}$$

$$l_0 := \frac{1+g_0}{2g_0} \text{ and } u_0 := \frac{1-g_0}{2g_0}. \tag{64}$$

The unique solution of (61) is

$$B(x) = l_0\big(1 - e^{-\frac{x}{\varepsilon}}\big) - \frac{x}{h},\ x \in [0,h]. \tag{65}$$

It is easy to check that

$$\int_0^h B(x)\,dx = \frac{h}{2g_0} - \varepsilon, \text{ and } \int_0^h \big(B'(x)\big)^2\,dx = \frac{1}{2\varepsilon}\Big(\frac{1}{g_0} - \frac{2\varepsilon}{h}\Big). \tag{66}$$

Thus, (36) holds with $b_1 = \frac{1}{2g_0} - \frac{\varepsilon}{h}$, and (37) holds with $b_2 = \frac{1}{g_0}\frac{h}{2\varepsilon} - 1$. In this case, we have

$$d = d_{\varepsilon,h} = \varepsilon + h\,b_1 = \frac{h}{2g_0} \text{ and } \ 1 + b_2 = \frac{h}{2\varepsilon}\frac{1}{g_0}.$$

According to (47), the optimal norm on $\mathcal{M}_h$ is given by

$$\|u_h\|^2_{*,h} = 2g_0\frac{\varepsilon}{h}\Big(\frac{h^2}{4g_0^2}|u_h|^2 + |u_h|^2_{*,h}\Big). \tag{67}$$

Since $tanh(x) \in (0,1)$ for $x > 0$, the condition (48) is satisfied with no restriction for h. Consequently, we have the following result.

THEOREM 6.3. *If u is the solution of* (27)*, u_h the solution of the upwinding PG formulation* (41) *with exponential bubble test space, then using the discrete norm* (67)*, we have*

$$\|u - u_h\|_{*,h} \le \sqrt{\frac{h}{2\varepsilon}\frac{1}{g_0}}\ \inf_{p_h \in \mathcal{M}_h} \|u - p_h\|_{*,h}. \tag{68}$$

Equivalently, by squaring and rescaling the estimate (68)*, we have*

$$\begin{aligned} h^2\ |u-u_h|^2 + 4g_0^2\,|u-u_h|^2_{*,h} \\ \le \frac{h}{2\varepsilon}\frac{1}{g_0}\inf_{p_h\in\mathcal{M}_h}\left(h^2\,|u-p_h|^2 + 4g_0^2\,|u-p_h|^2_{*,h}\right). \end{aligned} \tag{69}$$

For implementation purposes, we include the matrix of the finite element system (51) with exponential bubble upwinding. From (52), we get

$$M^e_{fe} = \text{tridiag}\Big(-\frac{1+g_0}{2g_0},\ \frac{1}{g_0},\ -\frac{1-g_0}{2g_0}\Big) = \text{tridiag}\Big(-l_0, \frac{1}{g_0}, -u_0\Big). \tag{70}$$

As presented in [15, 33, 32], the exponential UPG method produces the exact solution at the nodes. Thus, the solution u_h of the UPG formulation (41) with exponential bubble test space, satisfies $u_h = I_h(u)$, with u the exact solution of (27). In spite of that, the error estimate (68) does not guaranty approximation in the energy norm, especially when $\varepsilon << h$.

First, we emphasize that for $\varepsilon << h$, the computed solution $u_{h,c}$ is close to the interpolant of $w(x) = \int_0^x f(t)\,dt$ on $[0, x_{n-1}]$ and the energy error $|u - u_{h_c}|$ could be large. Indeed, due to the behavior of $g_0 = g_0(\varepsilon, h)$ and $g_j = g_j(\varepsilon, h)$ as $\frac{\varepsilon}{h} \to 0$, we have

$$g_0 \to 1, \text{ and } g_j = \varphi_j + B_j - B_{j+1} \to \chi_{|_{[x_{j-1}, x_j]}},$$

and the convergence, in both cases, is *exponentially fast*. Here, $\chi_{|_{[x_{j-1}, x_j]}}$ is the characteristic function of $[x_{j-1}, x_j]$. This implies fast convergence in

$$M^e_{fe} \to \text{tridiag}(-1, 1, 0), \text{ and } (f, g_j) \to \int_{x_{j-1}}^{x_j} f(x)\,dx, \text{ as } \frac{\varepsilon}{h} \to 0. \tag{71}$$

Thus, if accurate quadratures are used, $u_{h,c}$ is very close to the solution of

$$[\text{tridiag}(-1, 1, 0)]\, U = \Big[\int_{x_0}^{x_1} f(x)\,dx, \ldots, \int_{x_{n-2}}^{x_{n-1}} f(x)\,dx \Big]^T. \tag{72}$$

An immediate forward solve for the system gives

$$u_j = \int_0^{x_j} f(x)\,dx, \ j = 1, 2, \ldots, n-1.$$

This implies that, when $\varepsilon << h$, the computed solution $u_{h,c} \in \mathcal{M}_h$ satisfies

$$u_{h,c}(x_j) \approx u_j = \int_0^{x_j} f(x)\,dx, \ j = 1, 2, \ldots, n-1.$$

Based on the above observation, we show that, for a particular example, the energy error for the computed solution $u_{h,c}$ does not present a well established error order. We solve (2) for $f = 1$ and $b = 1$. The exact solution is

$$u(x) = x - \frac{e^{\frac{x}{\varepsilon}} - 1}{e^{\frac{1}{\varepsilon}} - 1}.$$

In this case, $(f, g_j) = (1, g_j) = (1, \varphi_j) = h = \int_{x_{j-1}}^{x_j} f(x)\,dx$ is computed exactly. For $\varepsilon << h$, e.g., $\frac{h}{\varepsilon} \geq 36.05$, the computed matrix M^e_{fe} of the finite element system (51) is tridiag$(-1, 1, 0)$ and $w(x) = x$. Thus,

$$u_{h,c}(x_j) = x_j, \ j = 1, 2, \ldots, n-1.$$

Direct computation of $|u - u_{h,c}|^2$ shows that

$$|u - u_{h_c}|^2 = \frac{1}{2\varepsilon} \frac{1 + e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}} - \frac{2}{h} \frac{1 - e^{-h/\varepsilon}}{1 - e^{-1/\varepsilon}} + \frac{1}{h}.$$

Hence, for $\varepsilon << h$,

$$|u - u_{h_c}|^2 \approx \frac{1}{2\varepsilon} - \frac{1}{h}. \tag{73}$$

Thus, the error is *large*, and *decreases slowly* as h decreases.

Second, using the same arguments, fast convergence in (71) for $\varepsilon << h$ and the solution of the limit system (72), the discrete UPG solution $u_h = I_h(u)$ is also very close to $I_h(w)$, hence to $u_{h,c}$.

For the same problem with $f = 1$, $b = 1$, elementary calculations give

$$|u - I_h(u)|^2 = |u - u_h|^2 = \frac{1 + e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}} \Big(\frac{1}{2\varepsilon} - \frac{1}{h} \frac{1 - e^{-h/\varepsilon}}{1 + e^{-1/\varepsilon}} \Big),$$

which leads to

$$|u - u_h|^2 \approx \frac{1}{2\varepsilon} - \frac{1}{h}, \text{ for } \varepsilon << h.$$

Not surprisingly, we obtained the same estimate for $|u - u_{h,c}|^2$ in (73).

This example shows that, even though the exponential bubble UPG method reproduces the exact solution at the nodes, the energy error could be quite large for $\varepsilon << h$, for both the computed and the exact solution. The large energy error is mainly related to the inability of the interpolant to approximate well the exact solution on uniform meshes and is less related to the computation error in approximating u_h by $u_{h,c}$. The energy error improves if we compute the error on subdomains away from the boundary layer.

7. REMARKS AND CONCLUSIONS FOR EXTENDING THE RESULTS TO MUTI-DIMENSIONAL CASES

We analyzed and compared mixed variational formulations for a model convection-diffusion problem. The key ingredient in our analysis is the representation of the optimal norm on the trial spaces at the continuous and discrete levels. The ideas presented for the one-dimensional model problem can be used for developing new and efficient discretizations for the multi-dimensional cases of convection dominated problems.

Below, we list the most useful ideas that our study concluded it help designing new and more efficient discretization methods for convection dominated problems in the multi-dimensional case.

I) First, we note that for any type of discrete variational formulation we use to approximate the solution of (27), the discrete solution u_h is independent of the norms we choose on the test and trial spaces. However, for the *standard linear* test space method, the SPLS method and the UPG method, the *discrete optimal trial norm* identifies what can be approximated with the given choice of test and trial spaces. For example, for the *standard linear* test space method, only a weighted energy norm $\varepsilon|u|$ can be used to measure the error. The second part of the discrete optimal norm is a semi-norm that is weaker than

the L^2-norm, see (26). Consequently, we cannot expect an optimal L^2-error approximation for this discretization.

The weight for the energy norm improves from ε to $\varepsilon + \frac{2\beta}{3}h$ for the quadratic UPG, see (53), and to $\frac{h}{2g_0}$ for the exponential UPG, see (67). As shown by our results in (55) and (69), this improvement leads to *better norm estimates* for the UPG discretizations.

II) The *continuous and discrete optimal trial norms* and the dependence $c_0 = c_0(\varepsilon, h)$ in the error estimate (22) can *predict* approximability of the continuous solution for the *given choice of the discrete test and trial spaces*. For example, for the *standard linear* test space method, the norms (19) and (24) are weak when compared to the standard *unweighted* H^1-norm. In addition, for $\varepsilon << h$, $c_0 \approx \frac{h}{\varepsilon}$ could be very large. For the SPLS method $c_0 = 1$, but the optimal continuous and discrete norms given by (19) are still weak. However, the error approximation for the SPLS improves when compared with the *standard linear* case, as presented in [5, 6].

III) We can *choose the test space* to *create upwinding diffusion from the convection part* in the variational formulation as done in the bubble UPG method. We can see how this idea works by comparing (28) and (44). *The FE upwinding process can be done at the basis level* by adding locally supported upwinding functions to each nodal function of the trial space. The *upwinding process* leads to the elimination of the non-physical oscillation in the discrete solutions, and to better approximation. The idea can be extended to the multi-dimensional case.

IV) For the $P^1 - P^2$-SPLS method, we have that the continuous and the discrete optimal trial norms agree and have a simple representation, making the error analysis more elegant. However, the UPG method performs better in spite of the fact that the test space for the UPG is a subspace of the test space $C^0 - P^2$ for the SPLS. The construction of a test space that creates upwinding diffusion from the convection part and leads to a simple optimal discrete norm, remains to be investigated even in the one-dimensional case.

V) According to recent work in [5, 15], the UPG method, the Streamline Diffusion (SD) method, and the Upwinding Finite Difference (UFD) method may lead to the same matrix of the resulting linear systems. For an UPG versus SD comparison see [5], and for an UPG versus UFD comparison see [15]. In spite of that, for both comparisons it was observed that UPG performs better than SD, and better than UFD. This can be justified by the fact that the UPG is *a global variational* method, SD is *a local residual stabilization* method, while the UFD is a *a particular version* of the UPG method where the dual vector of UPG is approximated by using a low order quadrature, such as the trapezoid rule, see [15].

We conclude by declaring the bubble UPG method as the most performant discretization for the one-dimensional model, and we believe that the ideas of the bubble UPG approach can be successfully extended to the multi-dimenional case to outperform the existing methods for convection dominated problems, in particular, the stream-line diffusion method.

REFERENCES

[1] A. Aziz and I. Babuška, *Survey lectures on mathematical foundations of the finite element method*. In: A. Aziz (Ed.), *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations* (Proc. Sympos., Univ. Maryland, Baltimore). Academic Press, New York, London, 1972.

[2] C. Bacuta, *Schur complements on Hilbert spaces and saddle point systems*. J. Comput. Appl. Math. **225** (2009), *2*, 581–593.

[3] C. Bacuta, D. Hayes, and J. Jacavage, *Notes on a saddle point reformulation of mixed variational problems*. Comput. Math. Appl. **95** (2021), 4–18.

[4] C. Bacuta, D. Hayes, and J. Jacavage, *Efficient discretization and preconditioning of the singularly perturbed reaction-diffusion problem*. Comput. Math. Appl. **109** (2022), 270–279.

[5] C. Bacuta, D. Hayes, and T. O'Grady, *Saddle point least squares discretization for convection-diffusion*. Appl. Anal. **103** (2024), *12*, 2241–2268.

[6] C. Bacuta, D. Hayes, and T. O'Grady, *Notes on finite element discretization for a model convection-diffusion problem*. 2023, arXiv:2302.07809.

[7] C. Bacuta and J. Jacavage, *A non-conforming saddle point least squares approach for an elliptic interface problem*. Comput. Methods Appl. Math. **19** (2019), *3*, 399–414.

[8] C. Bacuta and J. Jacavage, *Saddle point least squares preconditioning of mixed methods*. Comput. Math. Appl. **77** (2019), *5*, 1396–1407.

[9] C. Bacuta and J. Jacavage, *Least squares preconditioning for mixed methods with non-conforming trial spaces*. Appl. Anal. **99** (2020), *16*, 2755–2775.

[10] C. Bacuta, J. Jacavage, K. Qirko, and F.J. Sayas, *Saddle point least squares iterative solvers for the time harmonic Maxwell equations*. Comput. Math. Appl. **70** (2017), *11*, 2915–2928.

[11] C. Bacuta and P. Monk, *Multilevel discretization of symmetric saddle point systems without the discrete LBB condition*. Appl. Numer. Math. **62** (2012), *6*, 667–681.

[12] C. Bacuta and K. Qirko, *A saddle point least squares approach to mixed methods*. Comput. Math. Appl. **70** (2015), *12*, 2920–2932.

[13] C. Bacuta and K. Qirko, *A saddle point least squares approach for primal mixed formulations of second order PDEs*. Comput. Math. Appl. **73** (2017), *2*, 173–186.

[14] Cr. Bacuta and C. Bacuta, *Connections between finite difference and finite element approximations*. Appl. Anal. **102** (2023), *6*, 1808–1820.

[15] Cr. Bacuta and C. Bacuta, *Connections between finite difference and finite element approximations for a convection-diffusion problem*. Rev. Roumaine Math. Pures Appl. **69** (2024), *3-4*. Preprint.

[16] D. Boffi, F. Brezzi, L. Demkowicz, R.G. Durán, R. Falk, and M. Fortin, *Mixed finite elements, compatibility conditions, and applications*. In: D. Boffi and L. Gastaldi (Eds.), Lecture Notes in Math. 1939, Springer, Berlin, Fondazione C.I.M.E., Florence, 2008.

[17] D. Boffi, F. Brezzi, and M. Fortin, *Mixed Finite Element Methods and Applications*. Springer Ser. Comput. Math. 44, Springer, Heidelberg, 2013.

[18] D. Broersen and R. Stevenson, *A robust Petrov–Galerkin discretisation of convection-diffusion equations*. Comput. Math. Appl. **68** (2014), *11*, 1605–1618.

[19] J. Chan, N. Heuer, T. Bui-Thanh, and L. Demkowicz, *A robust DPG method for convection-dominated diffusion problems II: adjoint boundary conditions and mesh-dependent test norms*. Comput. Math. Appl. **67** (2014), *4*, 771–795.

[20] I. Christie, D.F. Griffiths, A.R. Mitchell, and O.C. Zienkiewicz, *Finite element methods for second order differential equations with significant first derivatives*. Internat. J. Numer. Methods Engrg. **10** (1976), *6*, 1389–1396.

[21] A. Cohen, W. Dahmen, and G. Welper, *Adaptivity and variational stabilization for convection-diffusion equations*. ESAIM Math. Model. Numer. Anal. **46** (2012), *5*, 1247–1273.

[22] L. Demkowicz, T. Führer, N. Heuer, and X. Tian, *The double adaptivity paradigm (how to circumvent the discrete inf-sup conditions of Babuška and Brezzi)*. Comput. Math. Appl. **95** (2021), 41–66.

[23] L. Demkowicz and J. Gopalakrishnan, *A class of discontinuous Petrov–Galerkin methods. Part I: the transport equation*. Comput. Methods Appl. Mech. Engrg. **199** (2010), *23-24*, 1558–1572.

[24] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*. Cambridge Univ. Press, Cambridge, 1996.

[25] R.H. Gallagher, O.C. Zienkiewicz, and P. Hood, *Newtonian and non-newtonian viscous incompressible flow, temperature induced flows, finite element solutions, in the mathematics of finite elements and applications*. II. In: J.R. Whiteman (Ed.), *Proceedings of the Second Brunel University Conference of the Institute of Mathematics and its Applications* (Uxbridge, April 7–10, 1975), pp. 235–267. Academic Press, London, New York, 1976.

[26] P.W. Hemker, *A Numerical Study of Stiff Two-Point Boundary Problems*. Mathematical Centre Tracts 80, Mathematisch Centrum, Amsterdam, 1977.

[27] R. Lin and M. Stynes, *A balanced finite element method for singularly perturbed reaction-diffusion problems*. SIAM J. Numer. Anal. **50** (2012), *5*, 2729–2743.

[28] T. Linß, *Layer-Adapted Meshes for Reaction-Convection-Diffusion Problems*. Lecture Notes in Math. 1985, Springer, Berlin, 2010.

[29] A.R. Mitchell and D.F. Griffiths, *The Finite Difference Method in Partial Differential Equations*. A Wiley-Interscience Publication, John Wiley & Sons, Chichester, 1980.

[30] K.W. Morton and J.W. Barrett, *Optimal Petrov–Galerkin methods through approximate symmetrization*. IMA J. Numer. Anal. **1** (1981), *4*, 439–468.

[31] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical Mathematics*. Texts Appl. Math. **37**, Springer, Berlin, 2007.

[32] H.G. Roos and M. Schopf, *Convergence and stability in balanced norms of finite element methods on Shishkin meshes for reaction-diffusion problems*. ZAMM Z. Angew. Math. Mech. **95** (2015), *6*, 551–565.

[33] H.G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*. Springer Ser. Comput. Math. 24, Springer, Berlin, 1996.

[34] O.C. Zienkiewicz, R.L. Taylor, and P. Nithiarasu, *The Finite Element Method for Fluid Dynamics*. Elsevier/Butterworth Heinemann, Amsterdam, 2014.

University of Delaware
Department of Mathematical Sciences
501 Ewing Hall
Newark, DE, USA, 19716
bacuta@udel.edu
crbacuta@udel.edu
dphayes@udel.edu