# A first-order system Petrov-Galerkin discretisation for a reaction-diffusion problem on a fitted mesh

James Adler[*]
Department of Mathematics
Tufts University
Medford, MA 02155

Scott MacLachlan[†]
Department of Mathematics and Statistics
Memorial University of Newfoundland
St John's, NL, Canada

and

Niall Madden[‡]
School of Mathematics, Statistics, and Applied Mathematics
National University of Ireland, Galway
Ireland

August 5, 2015

**Abstract**

We consider the numerical solution, by a Petrov-Galerkin finite-element method, of a singularly perturbed reaction-diffusion differential equation posed on the unit square. In Lin & Stynes (2012), it is argued that the natural energy norm, associated with a standard Galerkin approach, is not an appropriate setting for analysing such problems, and there they propose a method for which the natural norm is "balanced". In the style of a first-order system least squares (FOSLS) method, we extend the approach of Lin & Stynes (2012) by introducing a constraint which simplifies the associated finite-element space and the method's analysis. We prove robust convergence in a balanced norm on a piecewise uniform (Shishkin) mesh, and present supporting numerical results. Finally, we demonstrate how the resulting linear systems are solved optimally using multigrid methods.

## 1 Introduction

This paper considers the robust numerical solution of a linear two-dimensional reaction-diffusion problem, posed on the unit square:

$$Lu := -\varepsilon^2 \Delta u + bu = f \quad \text{in } \Omega := (0,1)^2, \qquad u = 0 \ \text{ on } \partial\Omega, \tag{1}$$

where $\varepsilon \in (0,1]$ is a positive parameter, and $b$ is a continuous positive function on $\Omega$ (precise details of our assumptions on the data are given below in Section 1.1). The problem is *singularly perturbed* in the sense that, as $\varepsilon \to 0$, it becomes ill-posed. More interestingly, from the point of view of computing accurate numerical solutions, if $\varepsilon$ is small, the solution features boundary and corner layers. Unless one makes some unreasonable assumptions on a lower bound for $\varepsilon$, the accuracy of classical methods is greatly compromised in the singularly perturbed regime: since derivatives of the solution are large, so too are the errors in the numerical method.

[*]Email: James.Adler@tufts.edu
[†]Email: smaclachlan@mun.ca
[‡]Corresponding author. Email: Niall.Madden@NUIGalway.ie

The challenge, then, of designing parameter robust methods (also known as $\varepsilon$-uniform methods) is to compute solutions that resolve all layers present, and with accuracy that does not depend adversely on $\varepsilon$. We refer to Roos *et al.* (2008) for references to the extensive literature on this topic.

Much of the published work in the field of robustly solving problems such as (1) centers on the use of *a priori* fitted meshes. In particular, seminal works on parameter robust methods for singularly perturbed problems, which were based on the use of standard finite-difference methods, involved the graded meshes of Bakhvalov (1969), and the piecewise-uniform meshes of Shishkin (1992). The latter have been studied greatly, with numerous refinements and extensions to various classes of convection-diffusion problems, time-dependent problems, coupled systems, etc; see (Miller *et al.*, 1996; Farrell *et al.*, 2000; Shishkin & Shishkina, 2009) and their references. A detailed analysis of finite-difference methods applied on a Shishkin mesh to (1) is given by Clavero *et al.* (2005). See also Andreev (2006), Kellogg *et al.* (2008b) and Linß (2010, Chap. 8).

The literature on numerical solutions of singularly perturbed reaction-diffusion problems also includes numerous studies of finite-element methods. Most directly relevant to this article is the analysis of the standard Galerkin method with bilinear elements on a Shishkin mesh applied to (1), which is provided in Liu *et al.* (2009), and includes improvements on some earlier works (Li & Navon, 1998; Apel, 1999), as well as an extension to a two-scale sparse grid method. For this method, the weak form of (1) is: *find* $u \in H_0^1(\Omega)$ *such that*

$$B(u,v) := \varepsilon^2(\vec{\nabla}u, \vec{\nabla}v) + (bu, v) = (f, v) \quad \forall v \in H_0^1(\Omega). \tag{2}$$

It is natural to conduct the analysis with respect to the energy norm,

$$\|v\|_E = \left(\varepsilon^2\|\vec{\nabla}v\|_0^2 + \|v\|_0^2\right)^{1/2}. \tag{3}$$

It is shown in Liu *et al.* (2009, Thm. 3.1) that if $u^N$ is the solution obtained using the standard Galerkin method on an $N \times N$ tensor-product mesh with bilinear basis functions on a Shishkin mesh (which we describe in detail in Section 3.1), then there exists a constant $C$, independent of $\varepsilon$ and $N$, such that

$$\|u - u^N\|_E \leq C(N^{-2} + \varepsilon^{1/2}N^{-1}\ln N). \tag{4}$$

This bound is robust in the sense that $C$ does not depend on $\varepsilon$. However, it is the contribution of the $\varepsilon^2\|\vec{\nabla}v\|_0^2$ term in the energy norm, (3), that provides the $\mathcal{O}(\varepsilon^{1/2}N^{-1}\ln N)$ quantity in (4). Thus, as $\varepsilon \to 0$, the energy norm does not express the boundary layers in the solution. Such a bound is not unique to estimates for bilinear elements on layer-adapted meshes; analogous results are given by Melenk (2002) for the *hp*-FEM on the so-called *spectral boundary layer mesh*. It has long been asserted, notably in (Farrell *et al.*, 2000, §1.2), that the energy norm (3) is inappropriate for reaction-diffusion problems. As support for this, they cite the result of Bagaev & Shaĭdurov (1998) showing that, for the one-dimensional analogue of (1), if $u^N$ is the solution obtained using the standard finite-element method with piecewise linear elements on a uniform $N \times N$ mesh, then, independently of $\varepsilon$,

$$\|u - u^N\|_E \leq CN^{-1/2}.$$

This holds even if $\varepsilon \ll N^{-1}$, when the solution's boundary layers are not resolved, and $\|u - u^N\|_\infty$ is $\mathcal{O}(1)$. Similar results are also possible for linear or bilinear finite-elements applied to the two-dimensional problem; see Schopf (2014, Chap. 2).

Standard finite-elements, applied on a suitable mesh, often yield solutions that converge (with optimal order) uniformly in the maximum norm, although full, general analysis is difficult. For example, Schatz & Wahlbin (1983) give a best approximation result in the maximum norm on quasi-uniform meshes. However, Kopteva (2014) notes that there is no such result for strongly-anisotropic triangulations, and gives an example of a triangulation on which the pointwise interpolation error is $\mathcal{O}(N^{-2})$ but the error in the finite-element solution is $\mathcal{O}(N^{-1})$. Leykekhman (2008) does provide an $\varepsilon$-uniform estimate for a problem in arbitrary dimensions, though for a problem with homogeneous Neumann boundary conditions, where layers are only weakly expressed. For a problem with strong layers, Kopteva (2007) proves second-order convergence for a two-dimensional semi-linear problem on a layer-adapted mesh. However, this is on a smooth domain and with finite elements only in the interior, and finite differences in the boundary-layer region. Similar results, though, using a different style of analysis for finite-element methods applied on graded meshes to problems on smooth domains are given by Blatov (1992a,b).

The weakness of the energy norm, (3), is discussed at length by Lin & Stynes (2012), who are then motivated to investigate what they term as a *balanced finite-element method*. It is balanced in the sense that

the natural norm associated with the method involves terms like $\varepsilon\|\vec{\nabla}v\|_0^2$, which are bounded independently of $\varepsilon$, unlike the term $\varepsilon^2\|\vec{\nabla}v\|_0^2$ in (3), which is not. The method is constructed by rewriting (1) as a system of first-order equations. A careful weighting of the components of this system leads to a new finite-element method, which is coercive with respect to a balanced norm, and which is then analysed on a tensor-product Shishkin mesh. However, this approach introduces several complications stemming from the fact that the solution to the reformulated problem is found in $H_0^1(\Omega) \times H(\mathrm{div})$. Whereas more classical methods on such grids can be based on, say, bilinear elements, in Lin & Stynes (2012), the $H(\mathrm{div})$ terms are discretised using Raviart-Thomas elements. This complicates both the analysis and the implementation.

An alternative approach, taken by Roos & Schopf (2014), is to analyse the standard Galerkin method with respect to a balanced norm, where the $\varepsilon$-weighting in (3) is reduced:

$$\|v\|_{\mathrm{bal}} := \left(\varepsilon\|\vec{\nabla}v\|_0^2 + \|v\|_0^2\right)^{1/2}.$$

Their analysis is based on the $L_\infty$-stability of the $L_2$-projection, and relies on delicate results of Oswald (2013). They prove that

$$\|u - u^N\|_{\mathrm{bal}} \leq CN^{-1}\ln^{3/2} N,$$

where, again, $u^N$ is the Galerkin solution on an $N \times N$ Shishkin mesh. A similar result for an *hp*-FEM approach on a *Spectral Boundary Layer* mesh has recently been established by Melenk & Xenophontos (2015). However, in the standard Galerkin setting, the bilinear form is not coercive with respect to the balanced norm. This, in part, motivates the construction of a $C^0$ interior penalty (CIP) method (Roos & Schopf, 2014, §3). By a careful local choice of the penalty parameter for each element edge—which depends on $N$ and $\varepsilon$, and is determined by the region of the fitted mesh where the edge is found—uniform convergence in a balanced norm is established.

In this paper, we present a Petrov-Galerkin finite-element discretisation based on a first-order system reformulation of the continuum partial differential equation. Thus, the approach is similar in spirit to that of Lin & Stynes (2012), in that the analysis is conducted with respect to a strong norm induced by the bilinear form. However, by the inclusion of a curl constraint, it has the advantage that it can be directly discretised using bilinear elements in a weighted $H^1$ product space, rather than requiring $H(\mathrm{div})$ elements as in Lin & Stynes (2012). The primary advantage of this is the same as that of First-Order System Least-Squares (FOSLS) finite-element methods (Cai *et al.*, 1994, 1997) over standard Galerkin or Petrov-Galerkin discretisations of second-order equations: ensuring coercivity and continuity in the weighted $H^1$ product space ensures both optimal approximation in that norm and the ability to solve the resulting linear systems with optimal computational cost through multigrid methods. Additionally, discretization in a weighted $H^1$ product space allows the potential of using existing *a posteriori* error estimates on anisotropic meshes for adaptive mesh refinement (Huang *et al.*, 2010). Such refinement strategies are necessary in the extension of the current work to problems where interior layers arise, such as with semilinear partial differential equations (PDEs).

The outline of this paper is as follows. In Section 1.1, we detail the assumptions we make on the problem data in (1) and, in Section 1.2, we introduce the notation that is used throughout the paper. In Section 2, we propose a new Petrov-Galerkin formulation, and establish the coercivity and continuity of the bilinear form. Next, we describe the piecewise uniform Shishkin mesh in Section 3.1 and the finite-element spaces used in Section 3.2, along with a decomposition that complements this partitioning in Section 3.3. Following some preliminary technical discussion in Section 3.4, we analyse the approximation to the solution to (1) in Section 3.5, which culminates in a proof of the uniform convergence of the method in Theorem 3.6. In Section 4, we confirm the theoretical findings and compare the quality of the solution obtained with that of both classical Galerkin and the method of Lin & Stynes (2012). We also investigate the construction and efficiency of a suitable multigrid solver. Finally, we conclude with a discussion of future work in Section 5.

## 1.1 Assumptions

As mentioned above, the reaction coefficient, $b$, in (1) is positive and continuous. Therefore, there is a positive $\beta$ such that $b \geq 2\beta^2$. The relative magnitudes of $\varepsilon$ and $\beta$ determine if the problem is classified as being singularly perturbed. Since we are interested in this case, we make the following assumption on $\varepsilon$.

**Assumption 1.1.** *Assume that*

$$\varepsilon \leq C\beta N^{-1}, \tag{5}$$

*where $C$ is independent of the problem data, as otherwise* (1) *may be solved successfully on a uniform mesh, and the analysis can be done using standard techniques, see, e.g., Cai* et al. *(1997).*

Next, to simplify the notation in Sections 2 and 3 below, we make the following assumptions on the reaction coefficient.

**Assumption 1.2.** *(i) Assume that $\beta$ is bounded away from zero. Specifically,*

$$\beta^{-1} \leq C,$$

*where $C$ is independent of the problem data.*

*(ii) Assume that $b$ is bounded above by a term that may be absorbed into other constants in the analysis.*

*If necessary, these assumptions may be realised with a suitable rescaling of* (1) *that is accommodated by Assumption 1.1, or by introducing suitable constants in the norms defined below in* (12).

Finally, to ensure the existence of a suitable decomposition of the solution to (1), which is used to perform the analysis, we require that the problem data satisfy the assumptions made in Liu *et al.* (2009, §2.1).

**Assumption 1.3.** *Assume that*

$$f, b \in C^{4,\alpha}(\bar{\Omega}),$$

*for some $\alpha \in (0, 1]$ and that $f$ vanishes at the corners of the domain.*

It may be possible to obtain the necessary decomposition employing weaker assumptions. See, for example, Andreev (2006) for a study of the convergence of a finite-difference method for (1) subject to weaker assumptions. A review of results and open problems relating to corner singularities is given by Kellogg & Stynes (2008). Discussions of the effect on weakened regularity assumptions for singularly-perturbed PDEs is discussed, e.g., by Ludwig & Roos (2012); that is in the context of superconvergence of convection-diffusion problems, but includes a model problem with a reaction term. Ludwig & Roos (2014) also consider the effects of relaxing compatibility conditions on non-square domains.

## 1.2 Notation

In this paper, we consider a two-dimensional domain, $\Omega = (0, 1)^2$, with boundary, $\partial\Omega$, and its edges denoted by

$$\Gamma_1 := \{(x, 0)| \quad 0 \leq x \leq 1\}, \quad \Gamma_2 := \{(0, y)| \quad 0 \leq y \leq 1\},$$
$$\Gamma_3 := \{(x, 1)| \quad 0 \leq x \leq 1\}, \quad \Gamma_4 := \{(1, y)| \quad 0 \leq y \leq 1\}.$$

The corners of $\bar{\Omega}$ are $c_1$, $c_2$, $c_3$, and $c_4$, and are labelled clockwise from $c_1 = (0, 0)$. See Figure 1. We consider the space of square-integrable functions $L_2(\Omega)$, defined on this domain along with its scalar product and norm denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|_0$, respectively. We also consider the associated Sobolev spaces and norms:

$$H^1(\Omega) = \{u \in L_2(\Omega) : \vec{\nabla} u \in (L_2(\Omega))^2\} \qquad \|u\|_1 = \left(||u||_0^2 + ||\vec{\nabla} u||_0^2\right)^{1/2},$$
$$H(\mathrm{div}) = \{\vec{w} \in (L_2(\Omega))^2 : \vec{\nabla} \cdot \vec{w} \in L_2(\Omega)\} \qquad \|\vec{w}\|_{\mathrm{div}} = \left(||\vec{w}||_0^2 + ||\vec{\nabla} \cdot \vec{w}||_0^2\right)^{1/2},$$
$$H(\mathrm{curl}) = \{\vec{w} \in (L_2(\Omega))^2 : \vec{\nabla} \times \vec{w} \in L_2(\Omega)\} \qquad \|\vec{w}\|_{\mathrm{curl}} = \left(||\vec{w}||_0^2 + ||\vec{\nabla} \times \vec{w}||_0^2\right)^{1/2}.$$

Since $\vec{w}$ is defined on $\Omega \subset \mathbb{R}^2$, the curl operator is interpreted by considering $\vec{w}$ to be defined on $\mathbb{R}^3$, but only supported on the $(x, y)$-hyperplane in $\mathbb{R}^3$ (see Cai *et al.* (1997)). This means that $\vec{\nabla} \times \vec{w} = \left(\frac{\partial}{\partial x} w_2 - \frac{\partial}{\partial y} w_1\right)$.

Throughout this paper, $C$ always represents a constant that is independent of $\varepsilon$ and the discretisation parameter $N$, and may take different values in difference places.

## 2 The Petrov-Galerkin weak formulation

Similar to what is done in Cai *et al.* (1994, 1997) and Lin & Stynes (2012), we begin with rewriting (1) as a first-order system of equations, by defining $\vec{w} = \vec{\nabla} u$, $\mathcal{U} = (u, \vec{w})^T$, and writing the first-order system as $L_{\text{div}}\mathcal{U} = \mathcal{F}_{\text{div}}$,

$$L_{\text{div}}\mathcal{U} = \quad \sqrt{\varepsilon}\big(\vec{w} - \vec{\nabla} u\big) = \vec{0}, \tag{7a}$$

$$-\varepsilon^2 \vec{\nabla} \cdot \vec{w} + bu = f, \tag{7b}$$

plus appropriate boundary conditions (discussed below in Section 3.2). Rather than directly discretising these equations with a Galerkin finite-element method, leading to a saddle-point system for $\mathcal{U}$, in Lin & Stynes (2012), they propose a weak form of

$$a_{\text{div}}(\mathcal{U}, \mathcal{V}) := \langle L_{\text{div}}\mathcal{U}, M_{\text{div}}\mathcal{V}\rangle = \langle \mathcal{F}_{\text{div}}, M_{\text{div}}\mathcal{V}\rangle \quad \forall \mathcal{V} \in H^1(\Omega) \times H(\text{div}), \tag{8}$$

with $\mathcal{V} = (v, \vec{z})^T$, and

$$M_{\text{div}}\mathcal{V} = \quad \sqrt{\varepsilon}\big(\vec{z} - \vec{\nabla} v\big),$$
$$-\varepsilon \vec{\nabla} \cdot \vec{z} + bv.$$

As written, the natural Sobolev space for $\vec{w}$ and $\vec{z}$ is $H(\text{div})$, requiring approximations with specialised finite-element methods, such as the Raviart-Thomas elements used in Lin & Stynes (2012). However, following the continuum equation, since $\vec{w} = \vec{\nabla} u$, we can freely add a weighted "curl constraint", imposing that $\vec{\nabla} \times \vec{w} = \vec{0}$; with such a constraint, $\vec{w}$ is naturally represented in $\big(H^1(\Omega)\big)^2$, at least on domains, $\Omega$, where $H(\text{div}) \cap H(\text{curl}) = \big(H^1(\Omega)\big)^2$. This is the premise behind the FOSLS methodology, which allows one to get symmetric positive-definite discrete linear systems that are amenable to multigrid solution methods (Cai *et al.*, 1994, 1997). With this constraint, we write the first-order system as $L\mathcal{U} = \mathcal{F}$,

$$\sqrt{\varepsilon}\big(\vec{w} - \vec{\nabla} u\big) = \vec{0}, \tag{9a}$$

$$L\mathcal{U} = \quad -\varepsilon^2 \vec{\nabla} \cdot \vec{w} + bu = f, \tag{9b}$$

$$\varepsilon^2 \vec{\nabla} \times \vec{w} = \vec{0}. \tag{9c}$$

In a standard FOSLS discretisation, the $L_2$-norm of the residual of this system would be minimized over an appropriate $H^1$ product space. This would yield the bilinear form system,

$$\langle L\mathcal{U}, L\mathcal{V}\rangle = \langle \mathcal{F}, L\mathcal{V}\rangle \quad \forall \mathcal{V} \in \big(H^1(\Omega)\big)^3.$$

Alternately, we could use a mixed finite-element method to close the system, yielding the bilinear form $\langle L\mathcal{U}, \mathcal{W}\rangle = \langle \mathcal{F}, \mathcal{W}\rangle$. Since we have more equations in $L\mathcal{U} = \mathcal{F}$ than unknowns in $\mathcal{U}$, this necessitates a Petrov-Galerkin approach. Motivated by the FOSLS approach and Lin & Stynes (2012), as well as the error analysis that follows, we close the system by testing against a space different from the image of $\big(H^1(\Omega)\big)^3$ under $L$. Thus, we introduce the following operator,

$$\sqrt{\varepsilon}\big(\vec{z} - \vec{\nabla} v\big), \tag{10a}$$

$$M_k\mathcal{V} = \quad -\varepsilon \vec{\nabla} \cdot \vec{z} + bv, \tag{10b}$$

$$\varepsilon^k \vec{\nabla} \times \vec{z}, \tag{10c}$$

and use this to close the system with (9), obtaining the weak form,

$$a_k(\mathcal{U}, \mathcal{V}) := \langle L\mathcal{U}, M_k\mathcal{V}\rangle = \langle \mathcal{F}, M_k\mathcal{V}\rangle \quad \forall \mathcal{V} \in \big(H^1(\Omega)\big)^3. \tag{11}$$

Again, appropriate boundary conditions, discussed below, are easily incorporated. A natural choice for this formulation would be to set $k = 1$ in (10). Then, the $\varepsilon$-weighting is the same for the divergence term and the curl constraint. Although this seems intuitive, as we discuss below in Remark 2, and as supported by the analysis of Lemma 3.5, it can be argued that this leads to an "underweighting" of the curl constraint relative to the finite-element error analysis. To achieve equivalence with a norm that "balances" the curl term's contribution to the error analysis, we also consider (11) with $k = 0$. We note that the ellipticity

result that follows is valid for any $k \in \mathbb{R}$; however, we focus on the cases of $k = 0$ and $k = 1$, following the above arguments.

As is shown below, the two specific bilinear forms we consider, $a_0(\mathcal{U}, \mathcal{V})$ and $a_1(\mathcal{U}, \mathcal{V})$, have the associated energy norms,

$$|||\mathcal{U}|||_0^2 = \frac{\varepsilon}{2}\|\vec{\nabla}u\|_0^2 + \beta\|u\|_0^2 + \frac{\varepsilon}{2}\|\vec{w}\|_0^2 + \varepsilon^3\|\vec{\nabla}\cdot\vec{w}\|_0^2 + \varepsilon^2\|\vec{\nabla}\times\vec{w}\|_0^2, \tag{12a}$$

$$|||\mathcal{U}|||_1^2 = \frac{\varepsilon}{2}\|\vec{\nabla}u\|_0^2 + \beta\|u\|_0^2 + \frac{\varepsilon}{2}\|\vec{w}\|_0^2 + \varepsilon^3\|\vec{\nabla}\cdot\vec{w}\|_0^2 + \varepsilon^3\|\vec{\nabla}\times\vec{w}\|_0^2. \tag{12b}$$

Note that closing the system against the image of $\left(H^1(\Omega)\right)^3$ under $M_1$ leads to an equal weighting factor of $\varepsilon^3$ on both the divergence and curl terms in (12b), while closing against the image under $M_0$ gives the unequal weighting in (12a). As noted in Remark 3 below, the $k = 1$ formulation appears to offer superior performance when measured in the discrete maximum norm, although both $k = 0$ and $k = 1$ offer good convergence in their respective energy norms. We also note that the difference in weights on the divergence terms in (9b) and (10b) is necessary here in order to ensure the "balancing" of these norms, with weighting of $\varepsilon^3$ on the divergence term, as in Lin & Stynes (2012). The main difference between the formulation that arises from (9)–(10) and the approach of Lin & Stynes (2012) is the introduction on the curl constraints in (9c) and (10c). The coercivity and continuity of $a_k(\mathcal{U}, \mathcal{V})$ follows immediately from Lin & Stynes (2012, Theorem 3.1).

**Lemma 2.1.** *There exists a constant, $C \geq 1$, independent of $\varepsilon$ and $k$ such that*

$$|a_k(\mathcal{U}, \mathcal{V})| \leq C|||\mathcal{U}|||_k|||\mathcal{V}|||_k \quad \forall \mathcal{U}, \mathcal{V} \in \left(H^1(\Omega)\right)^3, \tag{13a}$$

$$a_k(\mathcal{U},\mathcal{U}) \geq |||\mathcal{U}|||_k^2 \qquad \forall \mathcal{U} \in \left(H^1(\Omega)\right)^3, \tag{13b}$$

*for $k \in \mathbb{R}$.*

*Proof.* Consider the simpler bilinear form, $a_{\text{div}}(\mathcal{U}, \mathcal{V})$, defined in (8), and note that it differs from that in (11) in omitting the curl term:

$$a_k(\mathcal{U}, \mathcal{V}) - a_{\text{div}}(\mathcal{U}, \mathcal{V}) = \varepsilon^{2+k}\langle\vec{\nabla}\times\vec{w}, \vec{\nabla}\times\vec{z}\rangle.$$

From (Lin & Stynes, 2012, Theorem 3.1), we know there is a positive constant, $C_{\text{div}}$, such that

$$|a_{\text{div}}(\mathcal{U}, \mathcal{V})| \leq C_{\text{div}}|||\mathcal{U}|||_{\text{div}}|||\mathcal{V}|||_{\text{div}} \quad \forall \mathcal{U}, \mathcal{V} \in \left(H^1(\Omega)\right)^3, \tag{14a}$$

$$a_{\text{div}}(\mathcal{U},\mathcal{U}) \geq |||\mathcal{U}|||_{\text{div}}^2 \qquad \forall \mathcal{U} \in \left(H^1(\Omega)\right)^3, \tag{14b}$$

for the corresponding norm,

$$|||\mathcal{U}|||_{\text{div}}^2 := \frac{\varepsilon}{2}\|\vec{\nabla}u\|_0^2 + \beta\|u\|_0^2 + \frac{\varepsilon}{2}\|\vec{w}\|_0^2 + \varepsilon^3\|\vec{\nabla}\cdot\vec{w}\|_0^2. \tag{15}$$

Thus,

$$a_k(\mathcal{U},\mathcal{U}) = a_{\text{div}}(\mathcal{U},\mathcal{U}) + \varepsilon^{2+k}\|\vec{\nabla}\times\vec{w}\|_0^2 \geq |||\mathcal{U}|||_k^2,$$

ensuring coercivity of $a_k$ with respect to the weighted energy norms in (12).

For continuity, note that

$$\begin{aligned}
|a_k(\mathcal{U}, \mathcal{V})| &\leq |a_{\text{div}}(\mathcal{U}, \mathcal{V})| + \varepsilon^{2+k}\left|\langle\vec{\nabla}\times\vec{w}, \vec{\nabla}\times\vec{z}\rangle\right| \\
&\leq C_{\text{div}}|||\mathcal{U}|||_{\text{div}}|||\mathcal{V}|||_{\text{div}} + \varepsilon^{2+k}\|\vec{\nabla}\times\vec{w}\|_0\|\vec{\nabla}\times\vec{z}\|_0 \\
&\leq C|||\mathcal{U}|||_k|||\mathcal{V}|||_k.
\end{aligned}$$

where $C = \max(C_{\text{div}}, 1)$, and the last inequality follows from the usual arithmetic-geometric mean inequality, noting that

$$|||\mathcal{U}|||_k^2 = |||\mathcal{U}|||_{\text{div}}^2 + \left(\varepsilon^{1+\frac{k}{2}}\|\vec{\nabla}\times\vec{w}\|_0\right)^2.$$

□

From here, the Lax-Milgram lemma establishes uniqueness of the continuum weak solution, $\mathcal{U}^*$, for any suitably smooth right-hand side, $\mathcal{F}$, by noting that $\langle \mathcal{F}, M_k \mathcal{V} \rangle$ is a continuous linear operator. Similarly, Céa's Lemma establishes the quasi-optimality of the finite-dimensional approximations: if $\mathbb{V}^N$ is a finite dimensional subspace of $H^1(\Omega)^3$, then there is $\mathcal{U}^N \in \mathbb{V}^N$ that satisfies

$$a_k(\mathcal{U}^N, \mathcal{V}^N) = \langle L\mathcal{U}^N, M_k \mathcal{V}^N \rangle = \langle \mathcal{F}, M_k \mathcal{V}^N \rangle \quad \forall \mathcal{V}^N \in \mathbb{V}^N, \tag{16}$$

giving

$$|||\mathcal{U}^* - \mathcal{U}^N|||_k \leq C \inf_{\mathcal{V}^N \in \mathbb{V}^N} |||\mathcal{U}^* - \mathcal{V}^N|||_k, \tag{17}$$

for the same constant, $C$, as in (13a), for all $k \in \mathbb{R}$. We note that the resulting formulation could also be viewed as a Galerkin discretization of the weak form in (11), since $\mathcal{U}^N, \mathcal{V}^N \in \mathbb{V}^N$; however, we retain the Petrov-Galerkin terminology to highlight the distinction from the FOSLS approach.

Further, we note that Lemma 2.1, the existence and uniqueness of both the continuum and discrete weak solutions, and the quasi-optimality bound in (17) all hold for more general domains than the unit square. As Theorem 3.1 of Lin & Stynes (2012), from which Lemma 2.1 follows, relies on integration-by-parts, all of these results hold under an assumption of a sufficiently smooth domain, $\Omega \in \mathbb{R}^d$, for $d = 2$ or 3, with ellipticity always being in $H^1 \times (H(\mathrm{div}) \cap H(\mathrm{curl}))$. In the case of the unit square and this paper, $H^1 \times (H(\mathrm{div}) \cap H(\mathrm{curl})) = (H^1)^3$.

# 3 The numerical method and analysis

From (17), one sees that the key step in the analysis of the accuracy of the discretisation on any mesh is to bound $\inf_{\mathcal{V}^N \in \mathbb{V}^N} |||\mathcal{U}^* - \mathcal{V}^N|||_k$. This is typically done by choosing a particular $\mathcal{V}^N \in \mathbb{V}^N$ and showing that $|||\mathcal{U}^* - \mathcal{V}^N|||_k$ is suitably small. In this section, we first introduce the Shishkin mesh and finite-element space that we consider for $\mathbb{V}^N$. Then, in Section 3.3, we review some properties of the continuum solution of (1) that are useful in the analysis, before describing the non-standard choice of the *modified* interpolant used in the analysis in Section 3.4. The analysis for bounding the error of this modified interpolant follows in Section 3.5.

## 3.1 The Shishkin mesh

The piecewise uniform mesh of Shishkin that we employ is very commonly used for problems such as (1). We provide a minimal description here, and refer the reader to, e.g., Liu *et al.* (2009) and Lin & Stynes (2012) for more details.

We start with a one-dimensional mesh, $\omega_x$, with $N$ intervals, where we assume that $N$ is an integer multiple of 4. This mesh is uniform on each of the three subintervals: $[0, \tau]$, $[\tau, 1-\tau]$ and $[1-\tau, 1]$, which are partitioned into, respectively, $N/4$, $N/2$ and $N/4$ subintervals. The mesh parameter $\tau$ is defined to be

$$\tau = \min\{\frac{1}{4}, 2\varepsilon\beta^{-1} \ln N\}.$$

In the case of interest, $\varepsilon \leq C\beta N^{-1}$, so $\tau = 2\varepsilon\beta^{-1} \ln N \ll 1$. Thus, it represents a transition between a fine mesh near the boundaries, where the local mesh-width is $h_B = 8(\varepsilon/\beta)N^{-1} \ln N$, and a coarse mesh in the interior, where the mesh-width is $h_I = 2(1-2\tau)N^{-1} = \mathcal{O}(N^{-1})$. We then let $\omega_y^N = \omega_x^N$, and create the two-dimensional mesh, $\Omega^N$, by taking the tensor product of $\omega_x^N$ and $\omega_y^N$, as shown in Figure 1. That figure also shows the labelling for the interior, edge and boundary regions that we employ: $\bar{\Omega} = \Omega_{II} \cup \Omega_{BI} \cup \Omega_{IB} \cup \Omega_{BB}$, where

$$\begin{aligned}
\Omega_{II} &= [\tau, 1-\tau] \times [\tau, 1-\tau], \\
\Omega_{IB} &= [\tau, 1-\tau] \times ([0, \tau] \cup [1-\tau, 1]), \\
\Omega_{BI} &= ([0, \tau] \cup [1-\tau, 1]) \times [\tau, 1-\tau], \\
\Omega_{BB} &= ([0, \tau] \cup [1-\tau, 1]) \times ([0, \tau] \cup [1-\tau, 1]).
\end{aligned} \tag{18}$$

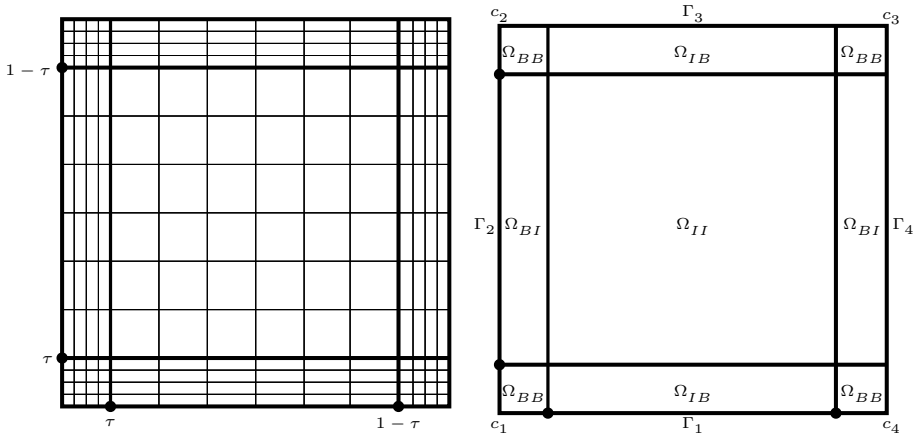



Figure 1: A Shishkin mesh for problem (1) (left). The labelling of the domain as used in the analysis (right).

### 3.2 The finite-element space

On the one-dimensional mesh, $\omega_x^N$, defined above, let $\{\zeta_0, \zeta_1, \zeta_2, \ldots, \zeta_{N-1}, \zeta_N\}$ represent the usual basis for the space of piecewise linear functions on $\omega_x$. That is, each $\zeta_i$ is piecewise linear on $\omega_x^N$, and

$$\zeta_i(x_j) = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases}$$

Let $\{\eta_0, \eta_1, \eta_2, \ldots, \eta_{N-1}, \eta_N\}$ be the corresponding basis for piecewise linear functions on $\omega_y^N$.

Now let $V^N$ be the space of piecewise bilinear functions on the tensor-product Shishkin mesh, $\Omega^N$, defined above, excluding boundaries (where homogeneous Dirichlet data for $u(x, y)$ is prescribed). Thus, $V^N$ is the space generated by the basis functions $\{\zeta_i(x)\eta_j(y)\}_{i,j=1,\ldots,N-1}$. For the vector field, $\vec{w}$, we prescribe homogeneous tangential components of $\vec{w}$ along the boundary, since the tangential component of $\vec{\nabla} u$ must also be zero. Thus, we define a vector space, $\left(V^N\right)^2$, of piecewise bilinear functions on $\Omega^N$ as the space generated by the functions,

$$\left\{ \begin{pmatrix} \zeta_i(x)\eta_j(y) \\ 0 \end{pmatrix} \right\}_{i=0,\ldots,N;j=1,\ldots,N-1}, \qquad \left\{ \begin{pmatrix} 0 \\ \zeta_i(x)\eta_j(y) \end{pmatrix} \right\}_{i=1,\ldots,N-1;j=0,\ldots,N}.$$

Defining $\left(H_0^1(\Omega)\right)^3$ to be the Sobolev space of three-component vectors under these boundary conditions, we have that $V^N \times \left(V^N\right)^2 \subset \left(H_0^1(\Omega)\right)^3$. Our finite-element method is then obtained by taking $\mathbb{V}^N = V^N \times \left(V^N\right)^2$ in (16).

Recalling (17), the accuracy of the method depends on the approximation properties of $\mathbb{V}^N$, and the analysis is conducted by estimating $|||\mathcal{U} - \mathcal{U}^N|||_k$, where $\mathcal{U}^N$ is the image of $\mathcal{U}$ under some suitable projection from $\left(H_0^1(\Omega)\right)^3$ to $\mathbb{V}^N$. The natural choice of projection is the nodal interpolation for each component: for any $\phi \in H_0^1(\Omega)$, let $\phi^I \in V^N$ be the element of $V^N$ such that $\phi^I(x_i, y_i) = \phi(x_i, y_i)$ for all $(x_i, y_i) \in \Omega^N$. The definition extends easily to an interpolant from $\left(H_0^1(\Omega)\right)^3$ to $\mathbb{V}^N$.

One can avail of standard interpolation estimates in order to establish a bound for $|||\mathcal{U} - \mathcal{U}^N|||_k$. However, since the Shishkin mesh $\Omega^N$ is highly anisotropic, the classical estimates of, say, Ciarlet (2002, Thm. 3.1.6) are insufficient (see Roos *et al.*, 2008, Remark 3.105). Instead, we require the following results, originally due to Apel (1999), which hold independently of the aspect ratio of mesh rectangles (see also, Roos *et al.* (2008, (3.124)) and Lin & Stynes (2012, Lemma 4.2)).

**Lemma 3.1.** *Let $K$ be a rectangle in $\Omega^N$ with sides $h_x$ and $h_y$. Then, for any $\phi \in H^2(K)$,*

$$\|\phi - \phi^I\|_{0,K} \leq C(h_x \|\phi_x\|_{0,K} + h_y \|\phi_y\|_{0,K}), \tag{19a}$$

$$\|\phi - \phi^I\|_{0,K} \leq C(h_x^2 \|\phi_{xx}\|_{0,K} + h_y^2 \|\phi_{yy}\|_{0,K}), \tag{19b}$$

$$\|(\phi - \phi^I)_x\|_{0,K} \leq C(h_x \|\phi_{xx}\|_{0,K} + h_y \|\phi_{xy}\|_{0,K}), \tag{19c}$$

$$\|(\phi - \phi^I)_y\|_{0,K} \leq C(h_x \|\phi_{xy}\|_{0,K} + h_y \|\phi_{yy}\|_{0,K}). \tag{19d}$$

### 3.3 Solution decomposition

The Shishkin mesh described above in Section 3.1, which distinguishes the corner, edge, and interior regions of the domain, motivates a decomposition of $u$, the solution to (1), in a way that complements the partitioning of the domain in (18). This decomposition was first established by Shishkin (1992) for problems in arbitrary dimensions (see also Chap. 3 of Shishkin & Shishkina (2009)), where it is shown that $u$ can be expressed as the sum of a regular component (i.e., with no strong dependency on $\varepsilon$), boundary layer components that are supported only along the edges of $\Omega$, and corner layer components on each corner of $\Omega$. Clavero *et al.* (2005, Theorem 2.2) provide full technical details for the case $d = 2$. Their arguments can be adapted to give the variant of decomposition we use, which is taken directly from Lemmas 1.1 and 1.2 of Liu *et al.* (2009).

**Lemma 3.2** (Liu *et al.* (2009, Lemmas 1.1 and 1.2))**.** *Under Assumption 1.3, the solution $u$ of (1) can be decomposed as*

$$u = V + W + Z = V + \sum_{i=1}^{4} W_i + \sum_{i=1}^{4} Z_i, \tag{20a}$$

*where each $W_i$ is a layer associated with the edge $\Gamma_i$ and each $Z_i$ is a layer associated with the corner $c_i$. There exists a constant $C$ such that*

$$\left|\frac{\partial^{m+n} V}{\partial x^m \partial y^n}(x,y)\right| \leq C(1+\varepsilon^{2-m-n}), \qquad 0 \leq m+n \leq 4, \tag{20b}$$

$$\left|\frac{\partial^{m+n} W_1}{\partial x^m \partial y^n}(x,y)\right| \leq C(1+\varepsilon^{2-m})\varepsilon^{-n} e^{-\beta y/\varepsilon} \qquad 0 \leq m+n \leq 3, \tag{20c}$$

$$\left|\frac{\partial^{m+n} W_2}{\partial x^m \partial y^n}(x,y)\right| \leq C\varepsilon^{-m}(1+\varepsilon^{2-n}) e^{-\beta x/\varepsilon} \qquad 0 \leq m+n \leq 3, \tag{20d}$$

$$\left|\frac{\partial^{m+n} Z_1}{\partial x^m \partial y^n}(x,y)\right| \leq C\varepsilon^{-m-n} e^{-\beta(x+y)/\varepsilon} \qquad 0 \leq m+n \leq 3, \tag{20e}$$

*with analogous bounds for $W_3$, $W_4$, $Z_2$, $Z_3$ and $Z_4$.*

Recalling Assumption 1.1, we have that

$$\tau = 2\varepsilon\beta^{-1} \ln N,$$

and so the mesh is not uniform. The significance of this, and the value of $\tau$, is that, for any point $(x, y) \in \Omega_{II}$,

$$e^{-\beta x/\varepsilon} \leq e^{-\beta\tau/\varepsilon} = N^{-2}, \qquad e^{-\beta y/\varepsilon} \leq e^{-\beta\tau/\varepsilon} = N^{-2}. \tag{21a}$$

We use these bounds in the remaining analysis. Furthermore, we have the following inequalities:

$$\|e^{-\beta y/\varepsilon}\|_{0,\Omega_{II}\cup\Omega_{BI}} = \|e^{-\beta x/\varepsilon}\|_{0,\Omega_{II}\cup\Omega_{IB}} \leq \sqrt{\frac{\varepsilon}{2\beta}} N^{-2}, \tag{21b}$$

$$\|e^{-\beta y/\varepsilon}\|_{0,\Omega_{BB}\cup\Omega_{IB}} = \|e^{-\beta x/\varepsilon}\|_{0,\Omega_{BB}\cup\Omega_{BI}} \leq \sqrt{\frac{\varepsilon}{2\beta}}, \tag{21c}$$

$$\|e^{-\beta(x+y)/\varepsilon}\|_{0,\Omega/\Omega_{BB}} \leq \frac{\varepsilon}{2\beta} N^{-2}, \qquad \text{and} \qquad \|e^{-\beta(x+y)/\varepsilon}\|_{0,\Omega_{BB}} \leq \frac{\varepsilon}{2\beta}. \tag{21d}$$

### 3.4 The modified interpolant

The standard nodal interpolant from $\left(H_0^1(\Omega)\right)^3$ to $\mathbb{V}^N$ is not sufficient to derive all the results we need for higher order terms in (12). This is because we require that certain components in the decomposition (20) decay rapidly in the interior of the domain, and satisfy bounds derived from those in (21). However, in the interior of the domain, interpolants of such components do not decay as rapidly. Therefore, we use an idea outlined in Lin & Stynes (2012) (see the discussion leading up to Corollary 4.6), and define a specialised operator that maps $\mathcal{U}$ onto $\mathbb{V}^N$.

Let $u$ be a function in $H_0^1(\Omega)$ that possesses a decomposition that satisfies (20). Recall that we write $\vec{w} = \vec{\nabla} u$, so, given the decomposition of Lemma 3.2, we decompose $w_1$ as

$$w_1 = u_x = V_x + W_x + Z_x = V_x + \sum_{i=1}^{4} (W_i)_x + \sum_{i=1}^{4} (Z_i)_x, \tag{22}$$

Then, there is a piecewise bilinear function, $w_1^{\widetilde{I}}$ in $V^N$, such that

$$w_1^{\widetilde{I}} = V_x^{\widetilde{I}} + \sum_{i=1}^{4} (W_i)_x^{\widetilde{I}} + \sum_{i=1}^{4} (Z_i)_x^{\widetilde{I}}, \tag{23}$$

where $V_x^{\widetilde{I}} = V_x^I$, the standard nodal interpolant of $V_x$, and

$$(W_k)_x^{\widetilde{I}}(x_i, y_j) = \begin{cases} (W_k)_x(x_i, y_j) & (x_i, y_j) \in \Omega^N / (\Omega_{II} \cup \Omega_{BI}), \\ 0 & (x_i, y_j) \in \Omega_{II} \cup \Omega_{BI}, \end{cases} \quad \text{for } k = 1, 3,$$

$$(W_k)_x^{\widetilde{I}}(x_i, y_j) = \begin{cases} (W_k)_x(x_i, y_j) & (x_i, y_j) \in \Omega^N / (\Omega_{II} \cup \Omega_{IB}), \\ 0 & (x_i, y_j) \in \Omega_{II} \cup \Omega_{IB}, \end{cases} \quad \text{for } k = 2, 4,$$

and

$$(Z_k)_x^{\widetilde{I}}(x_i, y_j) = \begin{cases} (Z_k)_x(x_i, y_j) & (x_i, y_j) \in \Omega^N / (\Omega_{II} \cup \Omega_{IB} \cup \Omega_{BI}), \\ 0 & (x_i, y_j) \in \Omega_{II} \cup \Omega_{IB} \cup \Omega_{BI}, \end{cases} \quad \text{for } k = 1, 2, 3, 4.$$

Define $w_2^{\widetilde{I}}$ analogously, and then set $\mathcal{U}^{\widetilde{I}} := (u^I, w_1^{\widetilde{I}}, w_2^{\widetilde{I}})^T \in \mathbb{V}^N$. Note that, by $\phi_x^I$, we always mean $(\phi_x)^I$, and never $(\phi^I)_x$.

Then, by definition, $(W_1)_x^I$ and $(W_1)_x^{\widetilde{I}}$ are identical on the subregion $[0,1] \times [0,\tau]$, except on the narrow strip $[0,1] \times [\tau - h_B, \tau]$. Therefore, $(W_1)_x^I - (W_1)_x^{\widetilde{I}}$ is a piecewise bilinear function that vanishes along $y = \tau - h_B$, and is equal to $(W_1)_x$ along $y = \tau$. Then, on an arbitrary rectangle $[x_i, x_{i+1}] \times [\tau - h_B, \tau]$, it is a simple calculation to show that

$$\|(W_1)_x^I - (W_1)_x^{\widetilde{I}}\|^2_{0,[x_i,x_{i+1}]\times[\tau-h_B,\tau]} \le \frac{1}{9} h_B (x_{i+1} - x_i) \big(W_1(x_i, \tau) + W_1(x_{i+1}, \tau)\big)^2.$$

A similar bound is possible on an arbitrary rectangle in $[0,1] \times [1-\tau, 1-\tau+h_B]$. Then the bounds in (20c) and (21) combine to show that

$$\|(W_1)_x^I - (W_1)_x^{\widetilde{I}}\|_{0,\Omega_{BB}\cup\Omega_{IB}} \le C\varepsilon^{1/2} N^{-5/2} \ln^{1/2} N. \tag{24a}$$

Similarly,

$$\|(W_2)_x^I - (W_2)_x^{\widetilde{I}}\|_{0,\Omega_{BB}\cup\Omega_{BI}} \le C\varepsilon^{-1/2} N^{-5/2} \ln^{1/2} N, \tag{24b}$$

and

$$\|(Z_1)_x^I - (Z_1)_x^{\widetilde{I}}\|_{0,\Omega_{BB}} \le C N^{-5/2} \ln N. \tag{24c}$$

These results are used below in the proof of Lemma 3.3. The proofs of Lemmas 3.4 and 3.5 rely on analogous results for derivatives of $(W_1)_x^I - (W_1)_x^{\widetilde{I}}$, and similar terms. A direct calculation shows that, on an arbitrary rectangle $[x_i, x_{i+1}] \times [\tau - h_B, \tau]$,

$$\|\big((W_1)_x^I - (W_1)_x^{\widetilde{I}}\big)_x\|^2_{0,[x_i,x_{i+1}]\times[\tau-h_B,\tau]} = \frac{1}{3} \frac{h_B}{(x_{i+1} - x_i)} \big(W_1(x_{i+1}, \tau) - W_1(x_i, \tau)\big)^2.$$

Consequently,

$$\|\big((W_1)_x^I - (W_1)_x^{\widetilde{I}}\big)_x\|_{0,\Omega_{BB}\cup\Omega_{IB}} \le C N^{-3/2}. \tag{25a}$$

Similarly,

$$\|\big((W_2)_x^I - (W_2)_x^{\widetilde{I}}\big)_x\|_{0,\Omega_{BB}\cup\Omega_{BI}} \le C\varepsilon^{-3/2} N^{-3/2} \ln^{-1/2} N, \tag{25b}$$

$$\|\big((W_1)_x^I - (W_1)_x^{\widetilde{I}}\big)_y\|_{0,\Omega_{BB}\cup\Omega_{IB}} \le C\varepsilon^{-1/2} N^{-3/2} \ln^{-1/2} N, \tag{25c}$$

$$\|\big((W_2)_x^I - (W_2)_x^{\widetilde{I}}\big)_y\|_{0,\Omega_{BB}\cup\Omega_{BI}} \le C\varepsilon^{-1} N^{-3/2}, \tag{25d}$$

and

$$\|\big((Z_1)_x^I - (Z_1)_x^{\widetilde{I}}\big)_x\|_{0,\Omega_{BB}} \le C\varepsilon^{-1} N^{-3/2}. \tag{25e}$$

Similar bounds exist for the norm of the difference between the nodal and modified interpolants of other terms in (22), and for terms in an analogous decomposition of $w_2 = u_y$.

### 3.5 Approximation of $\mathcal{U}$ in $\mathbb{V}^N$

In order to get the estimate for Céa's Lemma, we wish to establish a bound for

$$|||\mathcal{U}-\mathcal{U}^{\widetilde{I}}|||_k^2 = \frac{\varepsilon}{2}\|\vec{\nabla}(u-u^I)\|_0^2 + \beta\|(u-u^I)\|_0^2 + \frac{\varepsilon}{2}\|\vec{w}-\vec{w}^{\widetilde{I}}\|_0^2 + \varepsilon^3\|\vec{\nabla}\cdot(\vec{w}-\vec{w}^{\widetilde{I}})\|_0^2 + \varepsilon^{2+k}\|\vec{\nabla}\times(\vec{w}-\vec{w}^{\widetilde{I}})\|_0^2, \quad (26)$$

by providing a bound for each of the five normed expressions on the right-hand side of this inequality. From Liu *et al.* (2009, Lemma 2.3),

$$\|(u-u^I)\|_0 \le CN^{-2}, \quad (27)$$

and from Lin & Stynes (2012, Lemma 4.4),

$$\varepsilon^{1/2}\|\vec{\nabla}(u-u^I)\|_0 \le CN^{-1}\ln N. \quad (28)$$

We now prove bounds for the remaining terms, in the order in which they appear in (26).

**Lemma 3.3.** *There exists a constant $C$ such that*

$$\varepsilon^{1/2}\|\vec{w}-\vec{w}^{\widetilde{I}}\|_0 \le C(\varepsilon^{1/2}N^{-1} + N^{-2}\ln^2 N).$$

*Proof.* Recall that $\vec{w} = (u_x, u_y)^T$. We give the analysis for $\|u_x - w_1^{\widetilde{I}}\|_0$; the techniques for analysing $\|u_y - w_2^{\widetilde{I}}\|_0$ are the same.

From (19a), (20b), and (22), and noting that $V_x^I = V_x^{\widetilde{I}}$

$$\|V_x - V_x^{\widetilde{I}}\|_0 \le CN^{-1}. \quad (29)$$

Next, we consider $\|(W_1)_x - (W_1)_x^{\widetilde{I}}\|_0$. Recall that the mesh-widths in the boundary and interior regions are, respectively, $h_B = 8(\varepsilon/\beta)N^{-1}\ln N$, and $h_I = 2(1-2\tau)N^{-1}$. On $\Omega_{BB}\cup\Omega_{IB}$, we apply (19b), (20c), and (21c) to get

$$\begin{aligned}\|(W_1)_x - (W_1)_x^I\|_{0,\Omega_{BB}\cup\Omega_{IB}} &\le C(h_I^2\|(W_1)_{xxx}\|_{0,\Omega_{BB}\cup\Omega_{IB}} + h_B^2\|(W_1)_{xyy}\|_{0,\Omega_{BB}\cup\Omega_{IB}}\\ &\le C\big(h_I^2\varepsilon^{-1} + h_B^2\varepsilon^{-2}\big)\|e^{-y\beta/\varepsilon}\|_{0,\Omega_{BB}\cup\Omega_{IB}}\\ &\le C\big(N^{-2}\varepsilon^{-1} + N^{-2}\ln^2 N\big)(\varepsilon/\beta)^{1/2}\\ &\le C\varepsilon^{-1/2}N^{-2}.\end{aligned}$$

It follows from this, and (24a), that

$$\|(W_1)_x - (W_1)_x^{\widetilde{I}}\|_{0,\Omega_{BB}\cup\Omega_{IB}} \le \|(W_1)_x - (W_1)_x^I\|_{0,\Omega_{BB}\cup\Omega_{IB}} + \|(W_1)_x^I - (W_1)_x^{\widetilde{I}}\|_{0,\Omega_{BB}\cup\Omega_{IB}} \le C\varepsilon^{-1/2}N^{-2}. \quad (30a)$$

On $\Omega_{II}\cup\Omega_{BI}$, we note that the modified interpolant vanishes, and so, using (21b), we get

$$\|(W_1)_x - (W_1)_x^{\widetilde{I}}\|_{0,\Omega_{II}\cup\Omega_{BI}} = \|(W_1)_x\|_{0,\Omega_{II}\cup\Omega_{BI}} \le C\varepsilon^{1/2}N^{-2}. \quad (30b)$$

Thus, $\|(W_1)_x - (W_1)_x^{\widetilde{I}}\|_0 \le C\varepsilon^{-1/2}N^{-2}$. The same bound holds for $\|(W_3)_x - (W_3)_x^{\widetilde{I}}\|_0$.

For the term $\|(W_2)_x - (W_2)_x^{\widetilde{I}}\|_0$, we first use (19b) on $\Omega_{BB}\cup\Omega_{BI}$, to get

$$\begin{aligned}\|(W_2)_x - (W_2)_x^I\|_{0,\Omega_{BB}\cup\Omega_{BI}} &\le C(h_B^2\|(W_2)_{xxx}\|_{0,\Omega_{BB}\cup\Omega_{BI}} + h_I^2\|(W_2)_{xyy}\|_{0,\Omega_{BB}\cup\Omega_{BI}}\\ &\le C\big(h_B^2\varepsilon^{-3} + h_I^2\varepsilon^{-1}\big)\|e^{-x\beta/\varepsilon}\|_{0,\Omega_{BB}\cup\Omega_{BI}}\\ &\le C\big(\varepsilon^{-1}N^{-2}\ln^2 N + \varepsilon^{-1}N^{-2}\big)(\varepsilon/\beta)^{1/2}\\ &\le C\varepsilon^{-1/2}N^{-2}\ln^2 N.\end{aligned}$$

This, along with (24b), gives

$$\|(W_2)_x - (W_2)_x^{\widetilde{I}}\|_{0,\Omega_{BB}\cup\Omega_{BI}} \le C\varepsilon^{-1/2}N^{-2}\ln^2 N. \quad (31a)$$

On the remainder of the domain, $\Omega_{II} \cup \Omega_{IB}$,

$$\|(W_2)_x - (W_2)_x^{\widetilde{I}}\|_{0,\Omega_{II}\cup\Omega_{IB}} = \|(W_2)_x\|_{0,\Omega_{II}\cup\Omega_{IB}} \leq C\varepsilon^{-1/2}N^{-2}. \tag{31b}$$

Thus, $\|(W_2)_x - (W_2)_x^{\widetilde{I}}\|_0 \leq C\varepsilon^{-1/2}N^{-2}\ln^2 N$. The same bound holds for $\|(W_4)_x - (W_4)_x^{\widetilde{I}}\|_0$.

Finally, we consider $Z_1$, the term in the decomposition associated with the corner layer at the origin. First, on $\Omega_{BB}$, we use (19b) and (21d) to show

$$\begin{aligned}\|(Z_1)_x - (Z_1)_x^{I}\|_{0,\Omega_{BB}} &\leq Ch_B^2(\|(Z_1)_{xxx}\|_{\Omega_{BB}} + \|(Z_1)_{xyy}\|_{\Omega_{BB}}) \\ &\leq C(\varepsilon N^{-1}\ln N)^2\varepsilon^{-3}(\varepsilon/\beta) \leq CN^{-2}\ln^2 N.\end{aligned}$$

This combines with (24c) to show that $\|(Z_1)_x - (Z_1)_x^{\widetilde{I}}\|_{0,\Omega_{BB}} \leq CN^{-2}\ln^2 N$. On $\Omega/\Omega_{BB}$, $(Z_1)_x^{\widetilde{I}} = 0$, so (21d) gives that $\|(Z_1)_x - (Z_1)_x^{\widetilde{I}}\|_{0,\Omega/\Omega_{BB}} \leq CN^{-2}$, and so

$$\|(Z_1)_x - (Z_1)_x^{\widetilde{I}}\|_0 \leq CN^{-2}\ln^2 N. \tag{32}$$

The same bound holds for $\|(Z_2)_x - (Z_2)_x^{\widetilde{I}}\|_0$, $\|(Z_3)_x - (Z_3)_x^{\widetilde{I}}\|_0$, and $\|(Z_4)_x - (Z_4)_x^{\widetilde{I}}\|_0$.

Combining these results for $(Z_k)_x$ with (29), (30) and (31) completes the proof. □

**Remark 1.** *The result of Lemma 3.3 is sufficient for our analysis, and its contribution to the bound for* (26) *is dominated by* $\mathcal{O}(N^{-1}\ln N)$ *terms. Subject to stronger assumptions on the problem data and, in particular, the compatibility conditions in Assumption 1.3, one may be able to form a decomposition of* $u$ *as in* (20)*, but with the third derivatives of* $V$ *bounded independently of* $\varepsilon$*. Then,* (19b) *could be used instead of* (19a) *in the arguments leading to* (29)*, which would yield that*

$$\varepsilon^{1/2}\|\vec{w} - \vec{w}^{\widetilde{I}}\|_0 \leq CN^{-2}\ln^2 N.$$

*However, there would be no change in the consequent bounds for* $|||\mathcal{U} - \mathcal{U}^{\widetilde{I}}|||_k$.

**Lemma 3.4.** *There exists a constant* $C$ *such that*

$$\varepsilon^{3/2}\|\vec{\nabla}\cdot(\vec{w} - \vec{w}^{\widetilde{I}})\|_0 \leq CN^{-1}\ln N. \tag{33}$$

*Proof.* The arguments are essentially the same as in Theorem 3.3: appeal to (19c) and (19d) in place of (19a) and (19b), and use the bounds in (25) instead of (24). □

**Remark 2.** *Using exactly the same arguments as in Lemma 3.4, one can show that*

$$\varepsilon^{3/2}\|\vec{\nabla}\times(\vec{w} - \vec{w}^{\widetilde{I}})\|_0 \leq CN^{-1}\ln N. \tag{34}$$

*Then* (27), (28)*, Lemmas 3.3 and 3.4, with* (34)*, give that* $|||\mathcal{U} - \mathcal{U}^{\widetilde{I}}|||_k \leq CN^{-1}\ln N$ *for* $k \geq 1$*. Combined with Lemma 2.1, this is enough to show that there is a constant, independent of* $\varepsilon$ *and* $N$*, such that*

$$|||\mathcal{U} - \mathcal{U}^N|||_k \leq CN^{-1}\ln N, \text{ for } k \geq 1.$$

Note, however, that the $\mathcal{O}(\varepsilon^{-3/2})$ terms in $\|\vec{\nabla}\cdot(\vec{w} - \vec{w}^{\widetilde{I}})\|_0$ arise from the edge layer terms, $W_i$; in the decomposition that leads to (33), the regular and corner components, $V$ and $Z$, contribute terms that are $\mathcal{O}(\varepsilon^{-1})$. Furthermore, from (20), the $\varepsilon$-dependency of the pointwise bounds on derivatives of $V$ and $Z$ depends only on the order of the derivatives. In contrast, $\varepsilon$-dependency in $W$ is weaker for cross derivative terms than pure derivatives. Therefore, as we now show, it is possible to obtain sharper terms than those presented in (34).

**Lemma 3.5.** *There exists a constant* $C$ *such that*

$$\varepsilon\|\vec{\nabla}\times(\vec{w} - \vec{w}^{\widetilde{I}})\|_0 \leq CN^{-1}\ln N.$$

*Proof.* By the triangle inequality, $\|\vec{\nabla}\times(\vec{w} - \vec{w}^{\widetilde{I}})\|_0 \leq \|(u_x - w_1^{\widetilde{I}})_y\|_0 + \|(u_y - w_2^{\widetilde{I}})_x\|_0$. We present arguments to show that $\|(u_x - w_1^{\widetilde{I}})_y\|_0 \leq C\varepsilon^{-1}N^{-1}\ln N$; the analysis for $\|(u_y - w_2^{\widetilde{I}})_x\|_0$ is the same.

As usual, we decompose $u_x$ as in (22). From (20), all partial derivatives of $V$ of order $k$ are bounded, pointwise, by $C\varepsilon^{2-k}$. Therefore, we imitate the arguments of (29) and (32), but using (25) instead of (24), to show that

$$\|(V_x - V_x^{\widetilde{I}})_y\|_0 \leq C\varepsilon^{-1}N^{-1}, \tag{35}$$

and

$$\|((Z_i)_x - (Z_i)_x^{\widetilde{I}})_y\|_0 \leq C\varepsilon^{-1}N^{-1}\ln N \quad \text{for } i = 1,2,3,4. \tag{36}$$

Therefore, it is only necessary to consider the $W$ terms in detail.

On the region, $\Omega_{BB} \cup \Omega_{IB}$, (19d) gives that

$$\begin{aligned}\|\big((W_1)_x - (W_1)_x^I\big)_y\|_{0,\Omega_{BB}\cup\Omega_{IB}} &\leq C(h_I\|(W_1)_{xxy}\|_{0,\Omega_{BB}\cup\Omega_{IB}} + h_B\|(W_1)_{xyy}\|_{0,\Omega_{BB}\cup\Omega_{IB}}\\ &\leq C\big(N^{-1}\varepsilon^{-1} + (\varepsilon N^{-1}\ln N)\varepsilon^{-2}\big)(\varepsilon/\beta)^{1/2}\\ &\leq C\varepsilon^{-1/2}N^{-1}\ln N.\end{aligned}$$

Recalling the bound for $\|\big((W_1)_x^I - (W_1)_x^{\widetilde{I}}\big)_y\|_{0,\Omega_{BB}\cup\Omega_{IB}}$ in (25c), we get that

$$\|\big((W_1)_x - (W_1)_x^{\widetilde{I}}\big)_y\|_{0,\Omega_{BB}\cup\Omega_{IB}} \leq C\varepsilon^{-1/2}N^{-1}\ln N. \tag{37a}$$

On $\Omega_{II} \cup \Omega_{BI}$, we use (21b) and (23) to get

$$\|\big((W_1)_x - (W_1)_x^{\widetilde{I}}\big)_y\|_{0,\Omega_{II}\cup\Omega_{BI}} = \|(W_1)_{xy}\|_{0,\Omega_{II}\cup\Omega_{BI}} \leq C\varepsilon^{-1/2}N^{-2}. \tag{37b}$$

Bounds analogous to those in (37) hold for $\|\big((W_3)_x - (W_3)_x^{\widetilde{I}}\big)_y\|_0$.

For the term $\|\big((W_2)_x - (W_2)_x^{\widetilde{I}}\big)_y\|_0$, recall (25d) gives that $\|\big((W_2)_x^I - (W_2)_x^{\widetilde{I}}\big)_y\|_{0,\Omega_{BB}\cup\Omega_{BI}} \leq C\varepsilon^{-1}N^{-3/2}$; then we use (19d) on $\Omega_{BB} \cup \Omega_{BI}$, to get

$$\begin{aligned}\|\big((W_2)_x - (W_2)_x^I\big)_y\|_{0,\Omega_{BB}\cup\Omega_{BI}} &\leq C(h_B\|(W_2)_{xxy}\|_{0,\Omega_{BB}\cup\Omega_{BI}} + h_I\|(W_2)_{xyy}\|_{0,\Omega_{BB}\cup\Omega_{BI}}\\ &\leq C\big(h_B\varepsilon^{-2} + h_I\varepsilon^{-1}\big)\|e^{-x\beta/\varepsilon}\|_{0,\Omega_{BB}\cup\Omega_{BI}}\\ &\leq C\varepsilon^{-1/2}N^{-1}\ln N.\end{aligned}$$

We now conclude that

$$\|\big((W_2)_x - (W_2)_x^{\widetilde{I}}\big)_y\|_{0,\Omega_{BB}\cup\Omega_{BI}} \leq C\varepsilon^{-1/2}N^{-1}\ln N,$$

while on the remainder of the domain, $\Omega_{II} \cup \Omega_{IB}$,

$$\|\big((W_2)_x - (W_2)_x^{\widetilde{I}}\big)_y\|_{0,\Omega_{II}\cup\Omega_{IB}} = \|(W_2)_{xy}\|_{0,\Omega_{II}\cup\Omega_{IB}} \leq C\varepsilon^{-1/2}N^{-2}.$$

Thus,

$$\|\big((W_2)_x - (W_2)_x^{\widetilde{I}}\big)_y\|_0 \leq C\varepsilon^{-1/2}N^{-1}\ln N.$$

The same bound holds for $\|\big((W_4)_x - (W_4)_x^{\widetilde{I}}\big)_x\|_0$. Combining this with (35), (36) and (37) completes the proof. □

Finally, using Lemma 2.1, along with the bounds in (27), (28), and Lemmas 3.3, 3.4, and 3.5, we now state our main theorem.

**Theorem 3.6.** *There exists a constant, $C$, which is independent of $\varepsilon$ and $N$, such that*

$$|||\mathcal{U} - \mathcal{U}^N|||_k \leq CN^{-1}\ln N, \text{ for } k \geq 0.$$

# 4 Numerical results

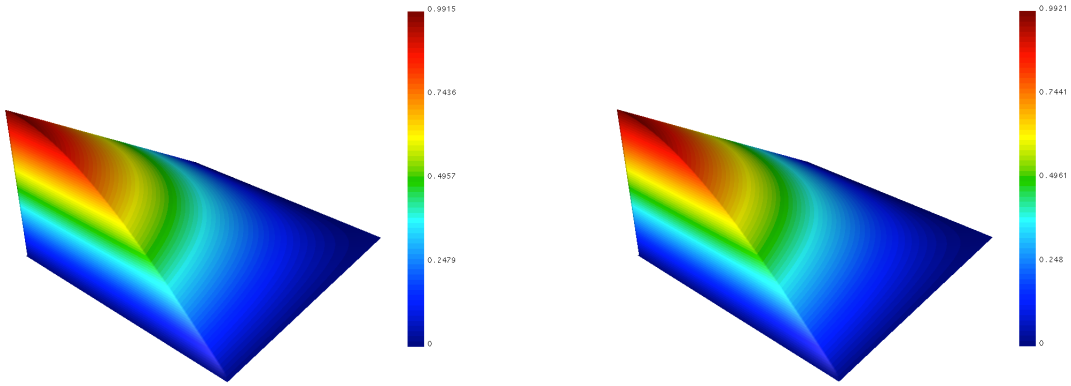
In this section, we verify the accuracy of our proposed method by presenting sets of numerical results obtained on various meshes. We begin by demonstrating the sharpness of the error bounds of Theorem 3.6 on the Shishkin mesh, in Section 4.1. This is followed, in Section 4.2, by comparisons to both the classical Galerkin method and the method of Lin & Stynes (2012). In Section 4.3, we provide some remarks from numerical investigation of its behaviour on uniform and Bakhvalov meshes. Since one of

our motivations for proposing the method is the possibility of designing a suitable solver, in Section 4.4, we consider the use of AMG-preconditioned GMRES as an iterative method with optimal cost for the discretisation matrices that arise.

We take as a test problem a specially constructed example, taken from Kopteva (2008) (see also Russell & Madden, 2014), for which $b = 1$, and $f$ is chosen such that the exact solution to (1) is

$$u = \left(\cos\left(\frac{\pi x}{2}\right) - \frac{e^{-x/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}}\right)\left(1 - y - \frac{e^{-y/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}}\right). \tag{38}$$

This problem exhibits exponential boundary layers near $x = 0$ and $y = 0$, and a corner layer close to the origin as shown in Figure 2; however, no other corner or boundary layers appear in this solution. Therefore, we have adjusted the construction of the mesh described in Section 3.1 slightly: we take $\tau = \min\{1/2, 2\varepsilon\beta^{-1} \ln N\}$ and form the one-dimensional meshes with $N/2$ intervals on the two sub-regions $[0, \tau]$ and $[\tau, 1]$.




(a) Solution (38) with $\varepsilon = 10^{-3}$.

(b) Solution (38) with $\varepsilon = 10^{-6}$.

Figure 2: Exact solution showing boundary and corner layers.

The Petrov-Galerkin formulation defined in (11) is discretised using bilinear elements for each component of solution, $\mathcal{U} = (u, w_1, w_2)^T$. All matrices and vectors for the tests were constructed using the modular finite-element library, MFEM. For the first set of results, the resulting linear systems were solved using the UMFPACK LU decomposition (Davis, 2004a,b; Davis & Duff, 1997, 1999).

## 4.1 Shishkin mesh

Tables 1 and 2 provide the data for discretisations with the two weak forms, for $k = 0, 1$, on the Shishkin mesh, corresponding to the analysis in Section 3.5. Writing the continuum (exact) solution $\mathcal{U}^* = (u^*, w_1^*, w_2^*)^T$ and the discrete solution in $\mathbb{V}^N = V^N \times \left(V^N\right)^2$ as $\mathcal{U} = (u^N, w_1^N, w_2^N)^T$, these tables (and that which follows for the uniform mesh) give three measures of the error:

- The error measured in the energy norm induced by the bilinear forms presented in (12), $|||\mathcal{U}^* - \mathcal{U}^N|||_k$. For the Shishkin mesh, it is these norms that are analysed in Section 3.5.
- The classical energy norm, $\|u^* - u^N\|_E$ (defined in (3)), of the error for the first component alone, for comparison with performance of a typical Galerkin discretisation.
- The discrete maximum norm, $\|u^* - u^N\|_{\ell_\infty}$, of the error of the first component alone, for comparison with performance of a typical finite-difference discretisation.

For ease of comparison to the theoretical analysis, we include reduction rates in the tables, giving the ratio of each entry to that of the previous column, comparing errors in $\mathcal{U}^N$ with those from $\mathcal{U}^{N/2}$. Standard first-order convergence, then should yield ratios approaching 0.5, while second-order convergence yields ratios approaching 0.25. Convergence at the rate of $N^{-1} \ln N$ would give ratios of

| $N$ | 128 | 256 | 512 | 1024 |
|---|---|---|---|---|
| $\frac{N^{-1}\ln(N)}{(N/2)^{-1}\ln(N/2)}$ | 0.58 | 0.57 | 0.56 | 0.56. |

The most obvious conclusion from these tables is that the error analysis is sharp, with reduction in the induced energy norms clearly matching the $N^{-1}\ln N$ prediction of that theory for both formulations. In both cases, for $\varepsilon = 1$ and $\varepsilon = 0.1$, convergence in the induced norm is fully first order (as expected for the standard FOSLS discretisation), but reduces to $N^{-1}\ln N$ for smaller $\varepsilon$. Also of note is that there remains slight variation in the values of $|||\mathcal{U}^* - \mathcal{U}^N|||_0$ for fixed $N$ and varying $\varepsilon$ with the $k = 0$ formulation, in contrast to the typical extremely uniform asymptotic performance observed with many schemes for singularly perturbed problems, and for the values of $|||\mathcal{U}^* - \mathcal{U}^N|||_1$ for fixed $N$ and varying $\varepsilon$ with the $k = 1$ formulation.

| | $\|\|\|\mathcal{U}^* - \mathcal{U}^N\|\|\|_0$ (Reduction Rate w.r.t. N) | | | | |
|---|---|---|---|---|---|
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 5.521e-03 | 2.761e-03(0.50) | 1.380e-03(0.50) | 6.901e-04(0.50) | 3.451e-04(0.50) |
| $10^{-1}$ | 4.342e-02 | 2.176e-02(0.50) | 1.089e-02(0.50) | 5.444e-03(0.50) | 2.722e-03(0.50) |
| $10^{-2}$ | 1.067e-01 | 6.403e-02(0.60) | 3.715e-02(0.58) | 2.104e-02(0.57) | 1.172e-02(0.56) |
| $10^{-3}$ | 9.880e-02 | 6.049e-02(0.61) | 3.598e-02(0.59) | 2.079e-02(0.58) | 1.172e-02(0.56) |
| $10^{-4}$ | 9.174e-02 | 5.518e-02(0.60) | 3.299e-02(0.60) | 1.947e-02(0.59) | 1.127e-02(0.58) |
| $10^{-5}$ | 8.988e-02 | 5.261e-02(0.59) | 3.050e-02(0.58) | 1.773e-02(0.58) | 1.033e-02(0.58) |
| $10^{-6}$ | 8.965e-02 | 5.216e-02(0.58) | 2.975e-02(0.57) | 1.681e-02(0.57) | 9.512e-03(0.57) |
| | $\|u^* - u^N\|_E$ (Reduction Rate w.r.t. N) | | | | |
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 1.593e-03 | 7.964e-04(0.50) | 3.982e-04(0.50) | 1.991e-04(0.50) | 9.955e-05(0.50) |
| $10^{-1}$ | 7.680e-03 | 3.838e-03(0.50) | 1.919e-03(0.50) | 9.594e-04(0.50) | 4.797e-04(0.50) |
| $10^{-2}$ | 6.889e-03 | 3.992e-03(0.58) | 2.272e-03(0.57) | 1.275e-03(0.56) | 7.081e-04(0.56) |
| $10^{-3}$ | 2.255e-03 | 1.307e-03(0.58) | 7.396e-04(0.57) | 4.126e-04(0.56) | 2.282e-04(0.55) |
| $10^{-4}$ | 7.104e-04 | 4.180e-04(0.59) | 2.386e-04(0.57) | 1.332e-04(0.56) | 7.320e-05(0.55) |
| $10^{-5}$ | 2.235e-04 | 1.301e-04(0.58) | 7.516e-05(0.58) | 4.255e-05(0.57) | 2.359e-05(0.55) |
| $10^{-6}$ | 7.915e-05 | 4.157e-05(0.53) | 2.339e-05(0.56) | 1.326e-05(0.57) | 7.444e-06(0.56) |
| | $\|u^* - u^N\|_{\ell_\infty}$ (Reduction Rate w.r.t. N) | | | | |
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 1.600e-05 | 4.014e-06(0.25) | 1.005e-06(0.25) | 2.515e-07(0.25) | 6.290e-08(0.25) |
| $10^{-1}$ | 2.389e-03 | 6.198e-04(0.26) | 1.579e-04(0.25) | 3.987e-05(0.25) | 1.002e-05(0.25) |
| $10^{-2}$ | 4.144e-02 | 1.730e-02(0.42) | 6.434e-03(0.37) | 2.144e-03(0.33) | 6.749e-04(0.31) |
| $10^{-3}$ | 1.513e-01 | 8.557e-02(0.57) | 4.091e-02(0.48) | 1.656e-02(0.40) | 5.821e-03(0.35) |
| $10^{-4}$ | 3.290e-01 | 2.467e-01(0.75) | 1.573e-01(0.64) | 8.623e-02(0.55) | 3.985e-02(0.46) |
| $10^{-5}$ | 4.281e-01 | 4.207e-01(0.98) | 3.504e-01(0.83) | 2.499e-01(0.71) | 1.554e-01(0.62) |
| $10^{-6}$ | 4.449e-01 | 4.832e-01(1.09) | 4.924e-01(1.02) | 4.467e-01(0.91) | 3.520e-01(0.79) |

Table 1: Induced energy, classical energy, and discrete maximum norms for solutions on Shishkin meshes using the bilinear form, $a_0$, defined in (11)

For both formulations, classical energy norm convergence with respect to $N$ follows naturally as a corollary to Theorem 3.6, since

$$\|u^* - u^N\|_E \leq \frac{1}{\beta}|||\mathcal{U}^* - \mathcal{U}^N|||_k,$$

for $\varepsilon \leq 1/(2\beta)$ (to account for the scaling in (26)). However, we note (as expected) that the energy norms show $\varepsilon$-dependence, scaling like $\sqrt{\varepsilon}$. As for the uniform mesh case results that follow, this agrees with the scaling of classical energy-norm errors for the Galerkin discretisation, as detailed in Liu *et al.* (2009, Thm. 3.1) (see Equation (4)). An important distinction between the two formulations appears in the discrete maximum norm errors, however, with much better performance for the $k = 1$ formulation in comparison to the $k = 0$ formulation. For $k = 0$, when $\varepsilon$ is small, we see extremely poor convergence (particularly for small $N$) in the discrete maximum norm. In contrast, for $k = 1$, the discrete maximum norm of the error in $u$ converges perfectly, at a rate of $N^{-2}\ln^2 N$, matching the convergence of the finite-difference discretisation on the Shishkin mesh (Clavero *et al.*, 2005). We note, however, that the $k = 0$ formulation still offers convergence in the norm associated with $k = 1$, since $|||\mathcal{V}|||_1 \leq |||\mathcal{V}|||_0$ for any $\varepsilon \leq 1$ (and such convergence has been observed numerically as well).

| | $\|\|\|\mathcal{U}^* - \mathcal{U}^N\|\|\|_1$ (Reduction Rate w.r.t. N) | | | | |
|---|---|---|---|---|---|
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 5.521e-03 | 2.761e-03(0.50) | 1.380e-03(0.50) | 6.901e-04(0.50) | 3.451e-04(0.50) |
| $10^{-1}$ | 3.106e-02 | 1.554e-02(0.50) | 7.771e-03(0.50) | 3.886e-03(0.50) | 1.943e-03(0.50) |
| $10^{-2}$ | 8.311e-02 | 4.865e-02(0.59) | 2.784e-02(0.57) | 1.566e-02(0.56) | 8.704e-03(0.56) |
| $10^{-3}$ | 8.416e-02 | 4.927e-02(0.59) | 2.819e-02(0.57) | 1.586e-02(0.56) | 8.814e-03(0.56) |
| $10^{-4}$ | 8.426e-02 | 4.933e-02(0.59) | 2.823e-02(0.57) | 1.588e-02(0.56) | 8.825e-03(0.56) |
| $10^{-5}$ | 8.427e-02 | 4.934e-02(0.59) | 2.823e-02(0.57) | 1.589e-02(0.56) | 8.826e-03(0.56) |
| $10^{-6}$ | 8.427e-02 | 4.934e-02(0.59) | 2.823e-02(0.57) | 1.589e-02(0.56) | 8.826e-03(0.56) |
| | $\|u^* - u^N\|_E$ (Reduction Rate w.r.t. N) | | | | |
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 1.593e-03 | 7.964e-04(0.50) | 3.982e-04(0.50) | 1.991e-04(0.50) | 9.955e-05(0.50) |
| $10^{-1}$ | 7.674e-03 | 3.838e-03(0.50) | 1.919e-03(0.50) | 9.594e-04(0.50) | 4.797e-04(0.50) |
| $10^{-2}$ | 6.787e-03 | 3.962e-03(0.58) | 2.265e-03(0.57) | 1.274e-03(0.56) | 7.079e-04(0.56) |
| $10^{-3}$ | 2.181e-03 | 1.273e-03(0.58) | 7.279e-04(0.57) | 4.095e-04(0.56) | 2.275e-04(0.56) |
| $10^{-4}$ | 6.918e-04 | 4.034e-04(0.58) | 2.306e-04(0.57) | 1.297e-04(0.56) | 7.206e-05(0.56) |
| $10^{-5}$ | 2.218e-04 | 1.279e-04(0.58) | 7.296e-05(0.57) | 4.103e-05(0.56) | 2.279e-05(0.56) |
| $10^{-6}$ | 7.892e-05 | 4.141e-05(0.52) | 2.318e-05(0.56) | 1.299e-05(0.56) | 7.209e-06(0.55) |
| | $\|u^* - u^N\|_{\ell_\infty}$ (Reduction Rate w.r.t. N) | | | | |
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 1.600e-05 | 4.014e-06(0.25) | 1.005e-06(0.25) | 2.515e-07(0.25) | 6.290e-08(0.25) |
| $10^{-1}$ | 2.331e-03 | 6.124e-04(0.26) | 1.570e-04(0.26) | 3.975e-05(0.25) | 1.000e-05(0.25) |
| $10^{-2}$ | 1.352e-02 | 5.062e-03(0.37) | 1.754e-03(0.35) | 5.748e-04(0.33) | 1.810e-04(0.31) |
| $10^{-3}$ | 1.361e-02 | 5.096e-03(0.37) | 1.765e-03(0.35) | 5.786e-04(0.33) | 1.822e-04(0.31) |
| $10^{-4}$ | 1.362e-02 | 5.098e-03(0.37) | 1.766e-03(0.35) | 5.787e-04(0.33) | 1.823e-04(0.32) |
| $10^{-5}$ | 1.362e-02 | 5.098e-03(0.37) | 1.766e-03(0.35) | 5.787e-04(0.33) | 1.823e-04(0.32) |
| $10^{-6}$ | 1.362e-02 | 5.098e-03(0.37) | 1.766e-03(0.35) | 5.787e-04(0.33) | 1.823e-04(0.32) |

Table 2: Induced energy, classical energy, and discrete maximum norms for solutions on Shishkin meshes using the bilinear form, $a_1$, defined in (11)

**Remark 3.** *There is, thus, a fundamental difference between the solutions generated by the two formulations, with the $k = 1$ formulation offering superior performance when the discrete maximum norm is considered, as is typical for singularly perturbed problems. The analysis of the connection between these induced energy norms and the discrete maximum norm is a focus for future research.*

## 4.2 Comparison with other approaches

A natural comparison is with the accuracy of other approaches, such as the classical Galerkin discretization or the similar approach from Lin & Stynes (2012), using $H(\mathrm{div})$ elements for $\vec{w}$ and $\vec{z}$. For the classical Galerkin discretization using piecewise bilinear elements (to match the space $V^N$ used for $u^N$ in the formulation presented here), we can only compare results for the scalar variable, $u$, computing $\|u^* - u^N\|_E$ and $\|u^* - u^N\|_{\ell_\infty}$. These values closely match those reported in Table 2 for the $a_1$ formulation proposed here. We note, however, that the discretization proposed here provides much more information than the Galerkin discretization and, in particular, gives two representations of $\vec{\nabla} u$, as $\vec{\nabla} u^N$ (obtained by directly differentiating the bilinear approximation $u^N$) and as $\vec{w}^N \in \left(V^N\right)^2$ from the reformulation as a first-order system. Thus, while the $L^2$ errors, $\|u^* - u^N\|_0$ and $\|\vec{\nabla} u^* - \vec{\nabla} u^N\|_0$, are comparable between the methods, we see much better approximation of $\vec{\nabla} u^*$ by $\vec{w}^N$; see Table 3. Numerically, we observe that $\varepsilon^{1/2}\|\vec{w}^* - \vec{w}^N\|_0$ scales as $\mathcal{O}(N^{-2}\ln^2 N)$, as discussed in Remark 1, rather than the scaling of $\mathcal{O}(N^{-1}\ln N)$ that bounds $\varepsilon^{1/2}\|\vec{\nabla} u^* - \vec{\nabla} u^N\|$ for the Galerkin solution (Liu *et al.*, 2009) and both our solution and the method of Lin & Stynes (2012). Since the error in the gradient dominates the classical Galerkin energy norm, it can be argued that our approach provides a better approximation in this norm than the Galerkin solution, as observed numerically, since

$$\left(\|u^* - u^N\|_0^2 + \varepsilon^2\|\vec{w}^* - \vec{w}^N\|_0^2\right)^{1/2} = \mathcal{O}\left(N^{-2}\left(1 + \varepsilon^{1/2}\ln^2 N\right)\right).$$

| | $\varepsilon^{1/2}\|\vec{\nabla}u^* - \vec{\nabla}u^N\|_0$, classical Galerkin (Reduction Rate w.r.t. N) | | | | |
|---|---|---|---|---|---|
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 1.592e-03 | 7.965e-04(0.50) | 3.982e-04(0.50) | 1.991e-04(0.50) | 9.955e-05(0.50) |
| $10^{-1}$ | 2.424e-02 | 1.213e-02(0.50) | 6.067e-03(0.50) | 3.033e-03(0.50) | 1.517e-03(0.50) |
| $10^{-2}$ | 6.736e-02 | 3.953e-02(0.59) | 2.263e-02(0.57) | 1.274e-02(0.56) | 7.078e-03(0.56) |
| $10^{-3}$ | 6.846e-02 | 4.016e-02(0.59) | 2.300e-02(0.57) | 1.295e-02(0.56) | 7.194e-03(0.56) |
| $10^{-4}$ | 6.858e-02 | 4.023e-02(0.59) | 2.304e-02(0.57) | 1.297e-02(0.56) | 7.205e-03(0.56) |
| $10^{-5}$ | 6.858e-02 | 4.023e-02(0.59) | 2.304e-02(0.57) | 1.297e-02(0.56) | 7.207e-03(0.56) |
| $10^{-6}$ | 6.859e-02 | 4.023e-02(0.59) | 2.304e-02(0.57) | 1.297e-02(0.56) | 7.207e-03(0.56) |
| | $\varepsilon^{1/2}\|\|\vec{w}^* - \vec{w}^N\|\|_0$, method of Lin & Stynes (2012) (Reduction Rate w.r.t. N) | | | | |
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 1.405e-03 | 7.023e-04(0.50) | 3.512e-04(0.50) | 1.756e-04(0.50) | 8.779e-05(0.50) |
| $10^{-1}$ | 9.686e-03 | 4.840e-03(0.50) | 2.419e-03(0.50) | 1.210e-03(0.50) | 6.048e-04(0.50) |
| $10^{-2}$ | 1.180e-02 | 6.070e-03(0.51) | 3.192e-03(0.53) | 1.697e-03(0.53) | 9.059e-04(0.53) |
| $10^{-3}$ | 1.055e-02 | 5.094e-03(0.48) | 2.514e-03(0.49) | 1.258e-03(0.50) | 6.343e-04(0.50) |
| $10^{-4}$ | 1.043e-02 | 4.989e-03(0.48) | 2.438e-03(0.49) | 1.207e-03(0.50) | 6.014e-04(0.50) |
| $10^{-5}$ | 1.041e-02 | 4.979e-03(0.48) | 2.430e-03(0.49) | 1.202e-03(0.49) | 5.980e-04(0.50) |
| $10^{-6}$ | 1.041e-02 | 4.978e-03(0.48) | 2.429e-03(0.49) | 1.201e-03(0.49) | 5.977e-04(0.50) |
| | $\varepsilon^{1/2}\|\|\vec{w}^* - \vec{w}^N\|\|_0$, $a_1$ formulation (Reduction Rate w.r.t. N) | | | | |
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 1.796e-05 | 4.489e-06(0.25) | 1.122e-06(0.25) | 2.806e-07(0.25) | 7.014e-08(0.25) |
| $10^{-1}$ | 7.112e-04 | 1.780e-04(0.25) | 4.451e-05(0.25) | 1.113e-05(0.25) | 2.782e-06(0.25) |
| $10^{-2}$ | 4.163e-03 | 1.425e-03(0.34) | 4.665e-04(0.33) | 1.477e-04(0.32) | 4.560e-05(0.31) |
| $10^{-3}$ | 4.185e-03 | 1.433e-03(0.34) | 4.690e-04(0.33) | 1.485e-04(0.32) | 4.584e-05(0.31) |
| $10^{-4}$ | 4.187e-03 | 1.434e-03(0.34) | 4.692e-04(0.33) | 1.486e-04(0.32) | 4.587e-05(0.31) |
| $10^{-5}$ | 4.187e-03 | 1.434e-03(0.34) | 4.693e-04(0.33) | 1.486e-04(0.32) | 4.587e-05(0.31) |
| $10^{-6}$ | 4.187e-03 | 1.434e-03(0.34) | 4.693e-04(0.33) | 1.486e-04(0.32) | 4.587e-05(0.31) |

Table 3: Errors in gradient approximation from classical Galerkin, method of Lin & Stynes (2012), and $a_1$ formulation on Shishkin meshes.

Comparing with the method of Lin & Stynes (2012), we again see improvements in the approximation of the gradient terms.[1] While the method of Lin & Stynes (2012) offers full first-order convergence, $\mathcal{O}(N^{-1})$ in $\varepsilon^{1/2}\|\vec{w}^* - \vec{w}^N\|_0$ for this example (in contrast to the rates of $\mathcal{O}(N^{-2}\ln^2 N)$ observed for the simpler example in Lin & Stynes (2012)), the $a_1$ formulation proposed here retains the rate of $\mathcal{O}(N^{-2}\ln^2 N)$, as discussed in Remark 1. Furthermore, the method of Lin & Stynes (2012) offers poorer approximation of $u$ itself, when measured in the discrete maximum norm, as shown in Table 4. For small $\varepsilon$ and large $N$, convergence of $||u^* - u^N||_{\ell_\infty}$ degrades to $\mathcal{O}(N^{-1})$, in comparison to $\mathcal{O}(N^{-2}\ln^2 N)$ as observed for both the classical Galerkin formulation and the formulation using $a_1$. Similar degradation was also seen in the approximation of $\vec{w}$ when measured in the discrete maximum norm (data not included here), with $\mathcal{O}(N^{-1}\ln N)$ convergence for the method of Lin & Stynes (2012) (matching that of the computed $\vec{\nabla}u^N$ for the Galerkin formulation) compared to $\mathcal{O}(N^{-2}\ln^2 N)$ for the $a_1$ formulation proposed here.

## 4.3 Uniform and Bakhvalov mesh results

As discussed in Section 1, it is shown in Schopf (2014, Chap. 2) that the classical Galerkin method applied on a uniform mesh appears to converge, uniformly in $\varepsilon$, with a rate of $N^{-1/2}$ in the classical energy norm. This is surprising since, if layers are not resolved, one expects the pointwise error to be $\mathcal{O}(1)$. Table 5 displays the error in the numerical approximations generated using the $a_1$ bilinear form discretised on a truly uniform $N \times N$ mesh.

In Table 5, we see a distinct lack of convergence in both the induced energy and discrete maximum norms for small $\varepsilon$, contrasted with typical finite-element convergence (first-order in induced energy and

[1] As the method of Lin & Stynes (2012) could not be easily implemented in the software framework described above, results for this approach were computed using deal.II (Bangerth *et al.*, 2007, 2013), again using a direct solver. For consistency, results for the $a_1$ formulation in Table 3 also come from an implementation using deal.II.

| | $\|\|u^* - u^N\|\|_{\ell_\infty}$ (Reduction Rate w.r.t. N) | | | | |
|---|---|---|---|---|---|
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 1.600e-05 | 4.014e-06(0.25) | 1.005e-06(0.25) | 2.515e-07(0.25) | 6.290e-08(0.25) |
| $10^{-1}$ | 2.338e-03 | 6.133e-04(0.26) | 1.571e-04(0.26) | 3.977e-05(0.25) | 1.000e-05(0.25) |
| $10^{-2}$ | 1.352e-02 | 5.063e-03(0.37) | 1.754e-03(0.35) | 5.748e-04(0.33) | 1.810e-04(0.31) |
| $10^{-3}$ | 1.451e-02 | 5.321e-03(0.37) | 1.811e-03(0.34) | 5.847e-04(0.32) | 1.823e-04(0.31) |
| $10^{-4}$ | 1.488e-02 | 5.489e-03(0.37) | 1.879e-03(0.34) | 7.032e-04(0.37) | 2.827e-04(0.40) |
| $10^{-5}$ | 1.492e-02 | 5.513e-03(0.37) | 1.890e-03(0.34) | 9.313e-04(0.49) | 4.493e-04(0.48) |
| $10^{-6}$ | 1.493e-02 | 5.515e-03(0.37) | 1.920e-03(0.35) | 9.732e-04(0.51) | 4.903e-04(0.50) |

Table 4: Discrete maximum norms for solutions on Shishkin meshes using the formulation from Lin & Stynes (2012), defined in (8)

| | $\|\|\|\mathcal{U}^* - \mathcal{U}^N\|\|\|_1$ (Reduction Rate w.r.t. N) | | | | |
|---|---|---|---|---|---|
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 5.521e-03 | 2.761e-03(0.50) | 1.380e-03(0.50) | 6.901e-04(0.50) | 3.451e-04(0.50) |
| $10^{-1}$ | 3.106e-02 | 1.554e-02(0.50) | 7.771e-03(0.50) | 3.886e-03(0.50) | 1.943e-03(0.50) |
| $10^{-2}$ | 3.223e-01 | 1.717e-01(0.53) | 8.738e-02(0.51) | 4.389e-02(0.50) | 2.197e-02(0.50) |
| $10^{-3}$ | 8.384e-01 | 7.609e-01(0.91) | 6.032e-01(0.79) | 3.900e-01(0.65) | 2.145e-01(0.55) |
| $10^{-4}$ | 1.367e-01 | 3.908e-01(2.86) | 7.309e-01(1.87) | 8.433e-01(1.15) | 7.937e-01(0.94) |
| $10^{-5}$ | 7.148e-02 | 6.805e-02(0.95) | 8.034e-02(1.18) | 1.040e-01(1.29) | 2.728e-01(2.62) |
| $10^{-6}$ | 6.225e-02 | 4.637e-02(0.74) | 3.862e-02(0.83) | 3.971e-02(1.03) | 4.950e-02(1.25) |
| | $\|u^* - u^N\|_E$ (Reduction Rate w.r.t. N) | | | | |
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 1.593e-03 | 7.964e-04(0.50) | 3.982e-04(0.50) | 1.991e-04(0.50) | 9.955e-05(0.50) |
| $10^{-1}$ | 7.674e-03 | 3.838e-03(0.50) | 1.919e-03(0.50) | 9.594e-04(0.50) | 4.797e-04(0.50) |
| $10^{-2}$ | 2.795e-02 | 1.421e-02(0.51) | 7.138e-03(0.50) | 3.573e-03(0.50) | 1.787e-03(0.50) |
| $10^{-3}$ | 5.467e-02 | 3.489e-02(0.64) | 2.052e-02(0.59) | 1.109e-02(0.54) | 5.689e-03(0.51) |
| $10^{-4}$ | 6.090e-02 | 4.217e-02(0.69) | 2.896e-02(0.69) | 1.978e-02(0.68) | 1.288e-02(0.65) |
| $10^{-5}$ | 6.114e-02 | 4.329e-02(0.71) | 3.063e-02(0.71) | 2.164e-02(0.71) | 1.507e-02(0.70) |
| $10^{-6}$ | 6.114e-02 | 4.329e-02(0.71) | 3.063e-02(0.71) | 2.167e-02(0.71) | 1.532e-02(0.71) |
| | $\|u^* - u^N\|_{\ell_\infty}$ (Reduction Rate w.r.t. N) | | | | |
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 1.600e-05 | 4.014e-06(0.25) | 1.005e-06(0.25) | 2.515e-07(0.25) | 6.290e-08(0.25) |
| $10^{-1}$ | 2.331e-03 | 6.124e-04(0.26) | 1.570e-04(0.26) | 3.975e-05(0.25) | 1.000e-05(0.25) |
| $10^{-2}$ | 1.253e-01 | 4.712e-02(0.38) | 1.478e-02(0.31) | 4.172e-03(0.28) | 1.110e-03(0.27) |
| $10^{-3}$ | 8.386e-01 | 6.795e-01(0.81) | 4.176e-01(0.61) | 1.770e-01(0.42) | 6.660e-02(0.38) |
| $10^{-4}$ | 9.756e-01 | 9.722e-01(1.00) | 9.600e-01(0.99) | 8.559e-01(0.89) | 7.572e-01(0.88) |
| $10^{-5}$ | 9.986e-01 | 9.988e-01(1.00) | 9.988e-01(1.00) | 9.877e-01(0.99) | 9.730e-01(0.99) |
| $10^{-6}$ | 9.986e-01 | 9.988e-01(1.00) | 9.989e-01(1.00) | 9.989e-01(1.00) | 9.989e-01(1.00) |

Table 5: Induced energy, classical energy, and discrete maximum norms for solutions on uniform meshes using the bilinear form, $a_1$, defined in (11)

second-order in discrete maximum norm) for problems that are not singularly perturbed relative to the mesh. On the whole, this is quite expected: when $\varepsilon = 1$, the discretisation is the same as the standard FOSLS discretisation (Cai *et al.*, 1997), and for cases where $\varepsilon N$ is not small, the Shishkin mesh coincides with a uniform mesh. For these cases, the classical energy-norm error also behaves as expected, with first-order convergence. Interestingly, while we have no corresponding theory, for small $\varepsilon$, the convergence of the energy-norm error matches the $\varepsilon$-uniform $N^{-1/2}$ rate given by Schopf (2014, Chap. 2) for the solution generated by the standard Galerkin scheme. Similar results are seen for the formulation using the $a_0$ weak form.

In a seminal paper, Bakhvalov (1969) proposed a graded layer-adapted mesh for singularly perturbed problems. This mesh is more complicated to construct and analyse than the piecewise-uniform Shishkin mesh, but is superior: as proved in Kellogg *et al.* (2008a), a standard finite-difference method applied

on the Bakhvalov mesh to (1) has full second-order convergence without the spoiling logarithmic factor associated with the Shishkin mesh. We refer the reader to, e.g., Linß (2010, §2.1.1) for details on the construction of this mesh. We offer no analysis for our methods applied on a Bakhvalov mesh. Numerical investigation, however, shows that the $k = 1$ formulation yields first-order convergence in the induced energy norm, convergence in the classical energy norm that is first-order in $N$ but scales like $\sqrt{\varepsilon}$, and fully second-order convergence, independent of $\varepsilon$, in the discrete maximum norm.

## 4.4 Solving the discretised systems

Having established both the theoretical convergence rates with respect to the induced energy norm and numerical convergence in both the classical energy and discrete maximum norms, we now turn to the question of whether or not the solution for the $k = 1$ formulation can be efficiently computed using standard preconditioned iterative methods. We choose to focus on this formulation using the Shishkin mesh for the obvious reason that it offers the best accuracy across the three error measures considered above. Timings presented below are for runs using a single core of a 8-Core 3GHz Intel Xeon "Sandy Bridge" machine with 256GB of RAM; our driver code is written in C++, and all libraries and drivers are compiled using the GNU Compiler Collection with full optimizations, using optimized LAPACK and BLAS libraries.

We first consider timings of both the factorization and solution phases using the UMFPACK LU factorization routines (Davis, 2004a,b; Davis & Duff, 1997, 1999). Table 6 details the time required for the factorization and solve phases of the LU factorization for varying $N$ and $\varepsilon$. For small $N$, we see essentially constant factorization times across all $\varepsilon$; however, for $N = 512$ and 1024, we see some growth in the time required for setting up the factorization as $\varepsilon \to 0$. In MacLachlan & Madden (2013), such growth was associated with appearance of subnormal floating-point numbers in the LU factors, although we have not analysed causes here. More notable is that, for each $\varepsilon$, the growth in time required for computing the factors with $N$ is more than the $N^3$ we would expect if using a nested dissection-like ordering taking advantage of the 9-point grid connectivity of the system when considered in terms of nodal $3 \times 3$ blocks. For $\varepsilon = 1$, the ratios of CPU times for increasing successive values of $N$ are 8.92, 11.11, 14.89, and 11.85, showing clear growth beyond the factor of 8 that should be achievable. Solve times are more moderate, and scale well in $\varepsilon$ for all $N$, with increase by a factor of roughly 5-7 as $N$ increases for fixed $\varepsilon$, suggesting that the ordering chosen for the LU factorization is generally a good one.

| | Factorization Time in Seconds | | | | |
|---|---|---|---|---|---|
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 1.821e-01 | 1.625e+00 | 1.805e+01 | 2.688e+02 | 3.184e+03 |
| $10^{-1}$ | 1.563e-01 | 1.729e+00 | 1.859e+01 | 3.189e+02 | 3.163e+03 |
| $10^{-2}$ | 1.592e-01 | 1.714e+00 | 2.062e+01 | 3.038e+02 | 3.238e+03 |
| $10^{-3}$ | 1.592e-01 | 1.736e+00 | 2.044e+01 | 2.786e+02 | 3.104e+03 |
| $10^{-4}$ | 1.667e-01 | 1.639e+00 | 1.958e+01 | 2.963e+02 | 3.155e+03 |
| $10^{-5}$ | 1.679e-01 | 1.564e+00 | 2.290e+01 | 2.891e+02 | 3.656e+03 |
| $10^{-6}$ | 1.684e-01 | 1.667e+00 | 2.194e+01 | 3.386e+02 | 1.001e+04 |

| | Solve Time in Seconds | | | | |
|---|---|---|---|---|---|
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 4.130e-03 | 2.160e-02 | 1.095e-01 | 7.384e-01 | 3.802e+00 |
| $10^{-1}$ | 3.543e-03 | 2.027e-02 | 1.368e-01 | 7.555e-01 | 3.751e+00 |
| $10^{-2}$ | 3.608e-03 | 2.156e-02 | 1.127e-01 | 7.470e-01 | 3.719e+00 |
| $10^{-3}$ | 3.615e-03 | 2.111e-02 | 1.381e-01 | 7.464e-01 | 3.730e+00 |
| $10^{-4}$ | 3.671e-03 | 2.159e-02 | 1.092e-01 | 7.409e-01 | 3.727e+00 |
| $10^{-5}$ | 3.673e-03 | 2.102e-02 | 1.277e-01 | 7.384e-01 | 3.781e+00 |
| $10^{-6}$ | 3.674e-03 | 2.149e-02 | 1.380e-01 | 7.561e-01 | 3.961e+00 |

Table 6: Factorization and solve phase timings for $k = 1$ formulation discretised on Shishkin meshes, using UMFPACK LU decomposition.

We compare with the setup and solve times of using algebraic multigrid (AMG) as a preconditioner for GMRES, still for the $k = 1$ formulation, noting that the discretisation matrices that arise for both

formulations are non-symmetric, due to the Petrov-Galerkin formulation considered. We use the hypre package developed by Lawrence Livermore National Laboratory to provide both the GMRES routine and the AMG preconditioner, through the BoomerAMG code (Henson & Yang, 2002). While both hypre and BoomerAMG support large-scale parallelism, we consider only a single thread in these runs. Therefore, we use classical serial AMG parameters, making use of the nodal structure of the degrees of freedom, a strength of connection threshold of 0.5, classical (Ruge-Stüben) coarsening and interpolation, and a scaled symmetric Gauss-Seidel smoother. See Ruge & Stüben (1987) and Stüben (2001) for general discussions of choice of parameters within AMG, and hypre (2012); Henson & Yang (2002) for details of the specifics of the hypre and BoomerAMG packages. A single V(1,1) cycle of AMG is used as the preconditioner at each step of the GMRES iterations.

Table 7 presents timings for the AMG setup and solve phases for varying $N$ and $\varepsilon$. In contrast to the slight increase in setup times with decreasing $\varepsilon$ noted for LU factorization above, we now see some decrease in setup times for AMG with decreasing $\varepsilon$. This is expected, since more entries in the system matrices are deemed to be weak connections as $\varepsilon$ (and, consequently, the entries scaled by $\varepsilon$) decreases. For each fixed $\varepsilon$, we further see near-optimal scaling in the timings for the setup phase, with growth by factors of roughly 4 each time $N$ doubles. Furthermore, we note that the setup times are dramatically better for AMG across all problems; for $N = 1024$, the AMG setup time is less than 1/500 of that of LU decomposition for $\varepsilon = 1$, and less than 1/2000 of the time required for LU decomposition for $\varepsilon = 10^{-6}$.

| | Setup Time in Seconds | | | | |
|---|---|---|---|---|---|
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 2.392e-02 | 8.940e-02 | 3.684e-01 | 1.547e+00 | 6.191e+00 |
| $10^{-1}$ | 2.444e-02 | 8.691e-02 | 3.621e-01 | 1.516e+00 | 6.157e+00 |
| $10^{-2}$ | 1.820e-02 | 9.315e-02 | 3.986e-01 | 1.630e+00 | 5.840e+00 |
| $10^{-3}$ | 1.734e-02 | 6.759e-02 | 2.832e-01 | 1.223e+00 | 5.804e+00 |
| $10^{-4}$ | 1.555e-02 | 6.638e-02 | 2.832e-01 | 1.197e+00 | 5.089e+00 |
| $10^{-5}$ | 1.564e-02 | 6.177e-02 | 2.602e-01 | 1.229e+00 | 5.049e+00 |
| $10^{-6}$ | 1.502e-02 | 6.280e-02 | 2.651e-01 | 1.094e+00 | 4.950e+00 |

| | Solve Time in Seconds (GMRES Iterations) | | | | |
|---|---|---|---|---|---|
| $\varepsilon$/N | 64 | 128 | 256 | 512 | 1024 |
| $10^{-0}$ | 2.231e-02 (5) | 1.097e-01 (5) | 6.359e-01 (6) | 3.188e+00 (7) | 1.319e+01 (7) |
| $10^{-1}$ | 1.802e-02 (4) | 1.086e-01 (5) | 5.527e-01 (5) | 2.811e+00 (6) | 1.314e+01 (7) |
| $10^{-2}$ | 1.987e-02 (4) | 1.295e-01 (5) | 7.080e-01 (6) | 4.085e+00 (8) | 1.547e+01 (8) |
| $10^{-3}$ | 1.853e-02 (4) | 1.152e-01 (5) | 5.934e-01 (5) | 3.490e+00 (7) | 1.891e+01 (9) |
| $10^{-4}$ | 1.930e-02 (4) | 1.204e-01 (5) | 5.902e-01 (5) | 3.493e+00 (7) | 1.445e+01 (7) |
| $10^{-5}$ | 1.818e-02 (4) | 1.114e-01 (5) | 5.706e-01 (5) | 3.035e+00 (6) | 1.271e+01 (6) |
| $10^{-6}$ | 1.814e-02 (4) | 1.123e-01 (5) | 5.753e-01 (5) | 2.943e+00 (6) | 1.288e+01 (6) |

Table 7: Setup and solve phase timings (with GMRES iteration counts in parentheses) for $k = 1$ formulation discretised on Shishkin meshes using AMG-preconditioned GMRES.

For the solve phase, we must first choose a stopping criterion for GMRES. Here, we use an experimentally determined criterion, requiring the $\ell_2$ norm of the residual of the linear system to be less than $0.25/N^3$ for the Shishkin mesh, which reliably achieves essentially the same accuracy in the discrete maximum norm as the direct solver across all values of $N$ and $\varepsilon$, varying only in the last digit reported in only a few cases. For most values of $\varepsilon$, we see growth in solve phase timing of roughly $N^2 \ln N$, corresponding to the tighter stopping tolerance for increasing $N$. While solve phase timings and iteration counts are roughly equal for both large and small values of $\varepsilon$, we see some growth, in both iteration counts and solve times for $\varepsilon = 10^{-2}$ and $10^{-3}$. For $N = 1024$, these represent cases in the middle ground, where the problem is not strongly singularly perturbed relative to the mesh, but the Shishkin mesh is also not a uniform mesh. In these cases, the stopping criterion used is not sharp, resulting in "oversolving" of the linear system, beyond the point where the error in the approximate solution is changing in any of the error measures considered above. By using a less-strict stopping tolerance, improved performance could be found for these cases, but at the cost of having to tune the algorithm for such special cases. Similar performance of both direct and iterative solvers is seen for the case of Bakhvalov meshes.

While the setup phase times for AMG are notably faster than those for LU factorization, even for large

meshes, the solve phase timings for the direct solver are faster than those for AMG, by factors of 3-5 times. While it could be argued that this ratio works in favor of the direct solver, by allowing a single expensive setup phase to be amortized over the solution of the linear system for many right-hand sides, this is not practically the case. For $N = 1024$ and $\varepsilon = 1$, the two approaches would break even only after solving the linear system for 339 right-hand sides. For $N = 1024$ and $\varepsilon = 10^{-6}$, they would break even after solving the linear system for 1122 right-hand sides.

# 5 Conclusions

We have proposed two new "balanced" finite-element methods for solving singularly perturbed reaction-diffusion problems, which simplify the approach of Lin & Stynes (2012). Theoretical analysis is provided that shows coercivity and continuity in induced "balanced" norms, and establishes uniform convergence on Shishkin meshes. A numerical investigation reveals that the formulation that gives the same weighting to the divergence and curl terms is superior in practise, offering good convergence of errors measured in the discrete maximum norm. Further theoretical and empirical investigations are needed in order to determine the precise reason for this.

Experimental results on the graded layer-adapted mesh of Bakhvalov on the unit square show that the method also works very well in this case: a more thorough study is needed to ground this in theory, and to optimise the mesh parameters for this method. While the formulation retains ellipticity in the induced balanced norm for any suitably smooth domain $\Omega$, additional work would be needed to extend the approximation properties proven here to other domains. First, a suitable solution decomposition is needed, along with pointwise bounds; these can be found for several cases in Shishkin & Shishkina (2009). Secondly, an analogue of the tensor-product Shishkin mesh used here (or any suitably graded mesh to resolve boundary and corner layers) is needed, for which analogues of the theorems presented in Section 3.5 or Lin & Stynes (2012) can be proven. Analogous results to Lemma 3.1 are known on triangular meshes, subject to a maximum angle condition, and could be used in more general settings in two dimensions (Roos *et al.*, 2008, Chapter 3). Likewise, similar results are known on tensor-product meshes in three dimensions (Apel, 1999).

Two main advantages arise from working in a weighted $H^1$ product space. First, as seen in Section 4.4, fast convergence of AMG-preconditioned GMRES is obtained for the resulting discretised systems. As a result, cost of solution to an $\varepsilon$-independent discretization error depends only very weakly on $\varepsilon$. Secondly, robust *a posteriori* error estimates are available for finite-element discretizations, such as the one proposed here, on $H^1$ spaces (Kunert, 2001; Formaggia & Perotto, 2001, 2003; Picasso, 2003), which can be used to drive adaptive mesh refinement algorithms, such as in Huang *et al.* (2010). Such adaptivity is considered a necessity for extending the approach developed here to semilinear reaction-diffusion problems, and will be the subject of future research.

# Acknowledgements

This work was partially supported by the National Science Foundation under grant DMS-1216972. The work of SM was partially supported by an NSERC Discovery Grant. The authors thank the anonymous referees for their helpful comments that improved the quality of this manuscript. We also thank Runchang Lin for his assistance verifying the results in Table 3.

# References

Andreev, V. (2006) On the accuracy of grid approximations of nonsmooth solutions of a singularly perturbed reaction-diffusion equation in the square. *Differ. Uravn.*, **42**, 895–906, 1005.

Apel, T. (1999) *Anisotropic finite elements: Local estimates and applications.* Advances in Numerical Mathematics. Stuttgart: B.G. Teubner.

Bagaev, B. M. & Shaĭdurov, V. V. (1998) *Setochnye metody resheniya zadach s pogranichnym sloem. Chast 1*. Novosibirsk: "Nauka", Sibirskoe Predpriyatie RAN, p. 199.

Bakhvalov, N. (1969) Towards optimization of methods for solving boundary value problems in the presence of boundary layers. *Zh. Vychisl. Mat. i Mat. Fiz.*, **9**, 841–859.

Bangerth, W., Hartmann, R. & Kanschat, G. (2007) deal.II – a general purpose object oriented finite element library. *ACM Trans. Math. Softw.*, **33**, 24/1–24/27.

Bangerth, W., Heister, T., Heltai, L., Kanschat, G., Kronbichler, M., Maier, M., Turcksin, B. & Young, T. D. (2013) The `deal.ii` library, version 8.0. *arXiv preprint* `http://arxiv.org/abs/1312.2266`.

Blatov, I. A. (1992a) On the Galerkin finite-element method for elliptic quasilinear singularly perturbed boundary value problems. I. *Differentsial'nye Uravneniya*, **28**, 1168–1177, 1285.

Blatov, I. A. (1992b) On the Galerkin finite-element method for elliptic quasilinear singularly perturbed boundary value problems. II. *Differentsial'nye Uravneniya*, **28**, 1799–1810, 1840.

Cai, Z., Lazarov, R., Manteuffel, T. & McCormick, S. (1994) First-order system least squares for second-order partial differential equations: Part I. *SIAM J. Numer. Anal.*, 1785–1799.

Cai, Z., Manteuffel, T. & McCormick, S. (1997) First-order system least squares for second-order partial differential equations. II. *SIAM J. Numer. Anal.*, **34**, 425–454.

Ciarlet, P. (2002) *The finite element method for elliptic problems*. Classics in Applied Mathematics, vol. 40. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), pp. xxviii+530. Reprint of the 1978 original [North-Holland, Amsterdam].

Clavero, C., Gracia, J. & O'Riordan, E. (2005) A parameter robust numerical method for a two dimensional reaction-diffusion problem. *Math. Comput.*, **74**, 1743–1758.

Davis, T. (2004a) Algorithm 832: UMFPACK, an unsymmetric-pattern multifrontal method. *ACM Trans. Math. Softw.*, **30**, 196–199.

Davis, T. (2004b) A column pre-ordering strategy for the unsymmetric-pattern multifrontal method. *ACM Trans. Math. Softw.*, **30**, 165–195.

Davis, T. & Duff, I. (1997) An unsymmetric-pattern mulitfrontal method for sparse LU factorization. *SIAM J. Matrix Anal. Appl.*, **18**, 140–158.

Davis, T. & Duff, I. (1999) A combine unifrontal/multifrontal method for unsymmetric sparse matrices. *ACM Trans. Math. Softw.*, **25**, 1–19.

Farrell, P., Hegarty, A., Miller, J., O'Riordan, E. & Shishkin, G. (2000) *Robust Computational Techniques for Boundary Layers*. Applied Mathematics. Boca Raton, U.S.A.: Chapman & Hall/CRC.

Formaggia, L. & Perotto, S. (2001) New anisotropic a priori error estimates. *Numerische Mathematik*, **89**, 641–667.

Formaggia, L. & Perotto, S. (2003) Anisotropic error estimates for elliptic problems. *Numerische Mathematik*, **94**, 67–92.

Henson, V. & Yang, U. (2002) BoomerAMG: a parallel algebraic multigrid solver and preconditioner. *Applied Numerical Mathematics*, **41**, 155–177.

Huang, W., Kamenski, L. & Lang, J. (2010) A new anisotropic mesh adaptation method based upon hierarchical a posteriori error estimates. *Journal of Computational Physics*, **229**, 2179 – 2198.

hypre (2012) High performance preconditioners. http://www.llnl.gov/CASC/hypre/.

Kellogg, R. B., Linss, T. & Stynes, M. (2008a) A finite difference method on layer-adapted meshes for an elliptic reaction-diffusion system in two dimensions. *Math. Comput.*, **77**, 2085–2096.

Kellogg, R. B., Madden, N. & Stynes, M. (2008b) A parameter-robust numerical method for a system of reaction-diffusion equations in two dimensions. *Numer. Methods Partial Differ. Equations*, **24**, 312–334.

Kellogg, R. B. & Stynes, M. (2008) Layers and corner singularities in singularly perturbed elliptic problems. *BIT*, **48**, 309–314.

Kopteva, N. (2007) Maximum norm error analysis of a 2D singularly perturbed semilinear reaction-diffusion problem. *Math. Comp.*, **76**, 631–646 (electronic).

Kopteva, N. (2008) Maximum norm a posteriori error estimate for a 2D singularly perturbed semilinear reaction-diffusion problem. *SIAM J. Numer. Anal.*, **46**, 1602–1618.

Kopteva, N. (2014) Linear finite elements may be only first-order pointwise accurate on anisotropic triangulations. *Math. Comp.*, **83**, 2061–2070.

Kunert, G. (2001) Robust a posteriori error estimation for a singularly perturbed reaction-diffusion equation on anisotropic tetrahedral meshes. *Advances in Computational Mathematics*, **15**, 237–259.

Leykekhman, D. (2008) Uniform error estimates in the finite element method for a singularly perturbed reaction-diffusion problem. *Math. Comp.*, **77**, 21–39 (electronic).

Li, J. & Navon, I. (1998) Uniformly convergent finite element methods for singularly perturbed elliptic boundary value problems. I. Reaction-diffusion type. *Comput. Math. Appl.*, **35**, 57–70.

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

Linss, T. (2010) *Layer-adapted meshes for reaction-convection-diffusion problems*. Lecture Notes in Mathematics, vol. 1985. Berlin: Springer-Verlag, pp. xii+320.

Liu, F., Madden, N., Stynes, M. & Zhou, A. (2009) A two-scale sparse grid method for a singularly perturbed reaction–diffusion problem in two dimensions. *IMA J. Numer. Anal.*, **29**, 986–1007.

Ludwig, L. & Roos, H.-G. (2012) Superconvergence for convection-diffusion problems with low regularity. *Applications of Mathematics 2012* (J. Brandts, J. Chleboun, S. Korotov, K. Segeth, J. Sístek & V. T. eds). Czech Academy of Sciences, pp. 173–187.

Ludwig, L. & Roos, H.-G. (2014) Finite element superconvergence on Shishkin meshes for convection-diffusion problems with corner singularities. *IMA J. Numer. Anal.*, **34**, 782–799.

MacLachlan, S. & Madden, N. (2013) Robust solution of singularly perturbed problems using multigrid methods. *SIAM J. Sci. Comput.*, **35**, A2225–A2254.

Melenk, J. M. (2002) *hp-finite element methods for singular perturbations*. Lecture Notes in Mathematics, vol. 1796. Berlin: Springer-Verlag, pp. xiv+318.

Melenk, J. & Xenophontos, C. (2015) Robust exponential convergence of hp-fem in balanced norms for singularly perturbed reaction-diffusion equations. *Calcolo*, 1–28.

MFEM (2011) Modular finite element methods library. http://mfem.googlecode.com.

Miller, J. J. H., O'Riordan, E. & Shishkin, G. I. (1996) *Fitted numerical methods for singular perturbation problems*. River Edge, NJ: World Scientific Publishing Co., Inc., pp. xiv+166.

Oswald, P. (2013) $L_\infty$-Bounds for the $L_2$-Projection onto Linear Spline Spaces. *Recent Advances in Harmonic Analysis and Applications* (D. Bilyk, L. D. Carli, A. Petukhov, A. Stokolos & B. Wick eds). Springer Proceedings in Mathematics & Statistics, vol. 25. New York: Springer, pp. 303–316.

Picasso, M. (2003) An anisotropic error indicator based on zienkiewicz–zhu error estimator: Application to elliptic and parabolic problems. *SIAM Journal on Scientific Computing*, **24**, 1328–1355.

Roos, H.-G., Stynes, M. & Tobiska, L. (2008) *Robust Numerical Methods for Singularly Perturbed Differential Equations*. Springer Series in Computational Mathematics, vol. 24, 2nd edn. Berlin: Springer-Verlag.

Roos, H.-G. & Schopf, M. (2014) Convergence and stability in balanced norms of finite element methods on Shishkin meshes for reaction-diffusion problems. *ZAMM, Z. Angew. Math. Mech.* doi: 10.1002/zamm.201300226.

Ruge, J. & Stüben, K. (1987) Algebraic multigrid (AMG). *Multigrid Methods* (S. McCormick ed.). Frontiers in Applied Mathematics, vol. 3. Philadelphia, PA: SIAM, pp. 73–130.

Russell, S. & Madden, N. (2014) A multiscale sparse grid finite element method for a two-dimensional singularly perturbed reaction-diffusion problem. *Submitted*.

Schatz, A. H. & Wahlbin, L. B. (1983) On the finite element method for singularly perturbed reaction-diffusion problems in two and one dimensions. *Math. Comp.*, **40**, 47–89.

Schopf, M. (2014) Error analysis of the Galerkin FEM in $L_2$-based norms for problems with layers. *Ph.D. thesis*, TU Dresden.

Shishkin, G. I. (1992) *Discrete Approximation of Singularly Perturbed Elliptic and Parabolic Equations*. Ekaterinburg: Russian Academy of Sciences, Ural Section. In Russian.

Shishkin, G. I. & Shishkina, L. P. (2009) *Difference methods for singular perturbation problems*. Monographs and Surveys in Pure and Applied Mathematics, vol. 140. Boca Raton, FL: CRC Press, pp. xvi+393.

Stüben, K. (2001) An introduction to algebraic multigrid. *Multigrid* (U. Trottenberg, C. Oosterlee & A. Schüller eds). London: Academic Press, pp. 413–528.