First-Order System Least Squares Finite-Elements for Singularly Perturbed Reaction-Diffusion Equations

James H. Adler[1] , Scott MacLachlan[2(✉)] , and Niall Madden[3]

[1] Department of Mathematics, Tufts University, Medford, USA
James.Adler@tufts.edu
[2] Department of Mathematics and Statistics,
Memorial University of Newfoundland, St. John's, Canada
smaclachlan@mun.ca
[3] School of Mathematics, Statistics, and Applied Mathematics,
National University of Ireland Galway, Galway, Ireland
Niall.Madden@NUIGalway.ie

Abstract. We propose a new first-order-system least squares (FOSLS) finite-element discretization for singularly perturbed reaction-diffusion equations. Solutions to such problems feature layer phenomena, and are ubiquitous in many areas of applied mathematics and modelling. There is a long history of the development of specialized numerical schemes for their accurate numerical approximation. We follow a well-established practice of employing *a priori* layer-adapted meshes, but with a novel finite-element method that yields a symmetric formulation while also inducing a so-called "balanced" norm. We prove continuity and coercivity of the FOSLS weak form, present a suitable piecewise uniform mesh, and report on the results of numerical experiments that demonstrate the accuracy and robustness of the method.

Keywords: First-order system least squares (FOSLS) finite elements · Singularly perturbed differential equations · Parameter-robust discretizations

1 Introduction

The numerical solution of singularly perturbed differential equations (SPDEs) is of great interest to numerical analysts, given the importance of these equations in computational modelling, and the challenges they present for classical numerical schemes and the mathematical methods used to analyse them; see [15] for a survey. In this work, we focus on linear second-order *reaction-diffusion problems* of the form

$$-\varepsilon \Delta u + bu = f \text{ on } \Omega := (0,1)^d \qquad u|_{\partial\Omega} = 0, \tag{1}$$

for $d = 1, 2, 3$, where we assume there exist constants $0 < b_0 < b(\boldsymbol{x}) < b_1$ for every $\boldsymbol{x} \in \Omega$. Like all SPDEs, (1) is characterised by a small positive parameter that multiplies the highest derivative. It is "singular" in the sense that the problem is ill-posed if one formally sets $\varepsilon = 0$. As ε approaches this limit, the solution typically exhibits layers: regions of rapid change, whose length is determined by ε. The over-arching goal is to devise methods that resolve these layers, and for which the error (measured in a suitable norm) is independent of ε. Many classical techniques make the tacit assumption that derivatives of u are bounded, which does not hold, uniformly in ε, for solutions to (1). Numerous specialised methods, usually based around layer-adapted meshes, have been developed with the goal of resolving these layers and the attendant mathematical conundrums. The celebrated piecewise uniform meshes of Shishkin have been particularly successful in this regard; and analyses of finite-difference methods for (1) and its many variants is largely complete [13].

Finite-element methods (FEMs) applied on layer-adapted meshes have also been successfully applied to (1), but their analysis is more problematic. This is highlighted to great effect by Lin and Stynes who demonstrated that the usual energy norm associated with (1) is too weak to adequately express the layers present in the solution [10]. They proposed a first-order FEM (see Sect. 2) for which the associated norm is sufficiently strong to capture layers; they coined the term "balanced norm" to describe this.

A flurry of activity on balanced norms was prompted by [10], including the first-order system Petrov-Galerkin (FOSPeG) approach proposed by the authors [1], and we refer to its introduction for a survey of the progress up to 2015. Since then, developments have continued apace. Broadly speaking, studies can be classified as one of two types.

1. Those that give analyses of standard FEMs, but in norms that are not induced by the associated bilinear forms; see, e.g., [16] on sparse grid FEMs, and [12] on hp-FEMs.
2. Those that propose new formulations for which the associated norm is naturally "balanced"; see, e.g., the discontinuous Petrov-Galerkin method of Heuer and Karulik [7].

The present study belongs to the second of these classes: we propose a new FEM for which the induced norm is balanced. This method is related to our earlier work [1], but instead uses a weighted least-squares FEM to obtain a symmetric discrete system. In this first-order system least-squares (FOSLS) approach [4,5], care is taken in choosing the weight, so that the resulting norms are indeed balanced.

The remainder of the paper is outlined as follows. Section 2 gives a brief discussion on balanced norms, where the Lin and Stynes and FOSPeG methods are summarized. In Sect. 3, we discuss the weighted least-squares approach and provide the necessary analysis, which applies in one, two and three dimensions. In Sect. 4, we focus on the particular case of $d = 2$; we present a suitable Shishkin mesh of the problem, and present numerical results that support our findings. Some concluding remarks are given in Sect. 5.

2 Balanced Norms

In [10], Lin and Stynes propose a first-order system reformulation of (1), writing the equivalent system as

$$L_{\text{div}}\mathcal{U} := \begin{pmatrix} \varepsilon^{1/4}(\boldsymbol{w} - \nabla u) \\ -\varepsilon\nabla \cdot \boldsymbol{w} + bu \end{pmatrix} = \begin{pmatrix} 0 \\ f \end{pmatrix} =: \mathcal{F}_{\text{div}}, \qquad (2)$$

for $\mathcal{U} = (u, \boldsymbol{w})^T$. Rather than forming a least-squares finite-element discretization as in [4,5], they choose to close the system in a nonsymmetric manner, defining $\mathcal{V} = (v, \boldsymbol{z})^T$ and

$$M_{\text{div}}\mathcal{V} := \begin{pmatrix} \varepsilon^{1/4}(\boldsymbol{z} - \nabla v) \\ -\varepsilon^{1/2}b^{-1}\nabla \cdot \boldsymbol{z} + v \end{pmatrix},$$

then writing the solution of (1) as that of the weak form

$$a_{\text{div}}(\mathcal{U}, \mathcal{V}) := \langle L_{\text{div}}\mathcal{U}, M_{\text{div}}\mathcal{V} \rangle = \langle \mathcal{F}_{\text{div}}, M_{\text{div}}\mathcal{V} \rangle \quad \forall \mathcal{V} \in H^1(\Omega) \times H(\text{div}). \quad (3)$$

In [10], it is shown that a_{div} is coercive and continuous with respect to the norm,

$$|||\mathcal{U}|||_{\text{div}}^2 := b_0\|u\|_0^2 + \frac{\varepsilon^{1/2}}{2}\|\nabla u\|_0^2 + \frac{\varepsilon^{1/2}}{2}\|\boldsymbol{w}\|_0^2 + \varepsilon^{3/2}\|\nabla \cdot \boldsymbol{w}\|_0^2, \qquad (4)$$

which is shown to be a *balanced* norm for the problem, in the sense that all the components in (4) have the same order of magnitude with respect to the perturbation parameter, ε.

In [1], the authors augmented the first-order system approach proposed by Lin and Stynes to include a curl constraint, in the same style as [5], leading to the first-order system reformulation of (1) as

$$L\mathcal{U} := \begin{pmatrix} \varepsilon^{1/4}(\boldsymbol{w} - \nabla u) \\ -\varepsilon\nabla \cdot \boldsymbol{w} + bu \\ \varepsilon\nabla \times \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ f \\ \boldsymbol{0} \end{pmatrix} =: \hat{\mathcal{F}}. \qquad (5)$$

Then, writing

$$M_k\mathcal{V} := \begin{pmatrix} \varepsilon^{1/4}(\boldsymbol{z} - \nabla v) \\ -\varepsilon^{1/2}b^{-1}\nabla \cdot \boldsymbol{z} + v \\ \varepsilon^{k/2}\nabla \times \boldsymbol{z} \end{pmatrix}, \qquad (6)$$

leads to the weak form

$$a_k(\mathcal{U}, \mathcal{V}) := \langle L\mathcal{U}, M_k\mathcal{V} \rangle = \langle \hat{\mathcal{F}}, M_k\mathcal{V} \rangle \quad \forall \mathcal{V} \in \left(H^1(\Omega)\right)^{1+d}. \qquad (7)$$

Building on the theory of [10], this form is shown to be coercive and continuous with respect to the balanced norm

$$|||\mathcal{U}|||_k^2 = b_0\|u\|_0^2 + \frac{\varepsilon^{1/2}}{2}\|\nabla u\|_0^2 + \frac{\varepsilon^{1/2}}{2}\|\boldsymbol{w}\|_0^2 + \varepsilon^{3/2}\|\nabla \cdot \boldsymbol{w}\|_0^2 + \varepsilon^{1+k/2}\|\nabla \times \boldsymbol{w}\|_0^2. \quad (8)$$

Furthermore, in [1], the authors show that, when discretized using piecewise bilinear finite elements on a tensor-product Shishkin mesh, this weak form leads to a *parameter-robust* discretization, with an error estimate independent of the perturbation parameter ε.

3   First-Order System Least Squares Finite-Element Methods

While theoretical and numerical results in [1] show the effectiveness of the first-order system Petrov-Galerkin approach proposed therein, the non-symmetric nature of the weak form also has disadvantages. Primary among these is that the weak form no longer can be used as an accurate and reliable error indicator, contrary to the common practice for FOSLS finite-element approaches [2–6]. Standard techniques to symmetrize the weak form in (7) fail, however, either sacrificing the balanced nature of the norm (and, thus, any guarantee of parameter robustness of the resulting discretization) or coercivity or continuity of the weak form (destroying standard error estimates). Here, we propose a FOSLS approach for the problem in (1), made possible by considering a weighted norm with spatially varying weight function. Weighted least-squares formulations have been used for a wide variety of problems including those with singularities due to the domain [8,9].

To this end, we define the weighted inner product on both scalar and vector $H^1(\Omega)$ spaces, writing

$$\langle u, v \rangle_\beta = \int_\Omega \beta(\boldsymbol{x}) u(\boldsymbol{x}) v(\boldsymbol{x}) \, d\boldsymbol{x},$$

with the associated norm written as $\|u\|_\beta$. Slightly reweighting the first-order system from (5), we have

$$\mathcal{L}\mathcal{U} := \begin{pmatrix} \varepsilon^{1/2}\big(\boldsymbol{w} - \nabla u\big) \\ -\varepsilon b^{-1/2}\nabla \cdot \boldsymbol{w} + b^{1/2}u \\ \varepsilon^{k/2}\nabla \times \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ b^{-1/2}f \\ \boldsymbol{0} \end{pmatrix} =: \mathcal{F}. \tag{9}$$

and pose the weighted FOSLS weak form as

$$a(\mathcal{U}, \mathcal{V}) = \langle \mathcal{L}\mathcal{U}, \mathcal{L}\mathcal{V} \rangle_\beta = \langle \mathcal{F}, \mathcal{L}\mathcal{V} \rangle_\beta \quad \forall \mathcal{V} \in \big(H^1(\Omega)\big)^{1+d}.$$

This form leads to a natural weighted product norm given by

$$|||\mathcal{U}|||_{\beta,k}^2 = \|u\|_\beta^2 + \varepsilon\|\nabla u\|_\beta^2 + \varepsilon\|\boldsymbol{w}\|_\beta^2 + \varepsilon^2\|\nabla \cdot \boldsymbol{w}\|_\beta^2 + \varepsilon^k\|\nabla \times \boldsymbol{w}\|_\beta^2.$$

As shown below, under a reasonable assumption on the weight function, β, the FOSLS weak form is coercive and continuous with respect to this norm.

Theorem 1. *Let $\beta(\boldsymbol{x})$ be given such that there exists $C > 0$ for which*

$$\nabla\beta \cdot \nabla\beta < \frac{b_0\beta^2(\boldsymbol{x})}{\varepsilon(1+C)^2},$$

for every $\boldsymbol{x} \in \Omega$, and let $k \in \mathbb{R}$ be given. Then,

$$|a(\mathcal{U}, \mathcal{V})| \leq \big(3 + 2\max(b_0^{-1}, b_1)\big) |||\mathcal{U}|||_{\beta,k}|||\mathcal{V}|||_{\beta,k}$$

$$\min\left(\frac{C\min(1, b_0)}{1+C}, b_1^{-1}, 1\right) |||\mathcal{U}|||_{\beta,k}^2 \leq a(\mathcal{U}, \mathcal{U})$$

for all $\mathcal{U}, \mathcal{V} \in \big(H^1(\Omega)\big)^{1+d}$.

Proof. For the continuity bound, we note that

$$a(\mathcal{U}, \mathcal{V}) = \varepsilon \langle \boldsymbol{w} - \nabla u, \boldsymbol{z} - \nabla v \rangle_\beta$$
$$+ \langle -\varepsilon b^{-1/2} \nabla \cdot \boldsymbol{w} + b^{1/2} u, -\varepsilon b^{-1/2} \nabla \cdot \boldsymbol{z} + b^{1/2} v \rangle_\beta$$
$$+ \varepsilon^k \langle \nabla \times \boldsymbol{w}, \nabla \times \boldsymbol{z} \rangle_\beta.$$

Thus, by the Cauchy-Schwarz and triangle inequalities, we have

$$|a(\mathcal{U}, \mathcal{V})| \le \varepsilon \left(\|\boldsymbol{w}\|_\beta + \|\nabla u\|_\beta \right) \left(\|\boldsymbol{z}\|_\beta + \|\nabla v\|_\beta \right)$$
$$+ \left(\varepsilon b_0^{-1/2} \|\nabla \cdot \boldsymbol{w}\|_\beta + b_1^{1/2} \|u\|_\beta \right) \left(\varepsilon b_0^{-1/2} \|\nabla \cdot \boldsymbol{z}\|_\beta + b_1^{1/2} \|v\|_\beta \right)$$
$$+ \varepsilon^k \|\nabla \times \boldsymbol{w}\|_\beta \|\nabla \times \boldsymbol{z}\|_\beta$$
$$\le \left(3 + 2\max(b_0^{-1}, b_1) \right) |||\mathcal{U}|||_{\beta,k} |||\mathcal{V}|||_{\beta,k}.$$

For the coercivity bound, we note

$$a(\mathcal{U}, \mathcal{U}) = \varepsilon \|\boldsymbol{w} - \nabla u\|_\beta^2 + \varepsilon^2 \|b^{-1/2} \nabla \cdot \boldsymbol{w}\|_\beta^2 + \|b^{1/2} u\|_\beta^2 + \varepsilon^k \|\nabla \times \boldsymbol{w}\|_\beta^2$$
$$- 2\varepsilon \langle \nabla \cdot \boldsymbol{w}, u \rangle_\beta$$
$$\ge \varepsilon \|\boldsymbol{w} - \nabla u\|_\beta^2 + \varepsilon^2 b_1^{-1} \|\nabla \cdot \boldsymbol{w}\|_\beta^2 + b_0 \|u\|_\beta^2 + \varepsilon^k \|\nabla \times \boldsymbol{w}\|_\beta^2$$
$$- 2\varepsilon \langle \nabla \cdot \boldsymbol{w}, u \rangle_\beta.$$

Now consider

$$-2\varepsilon \langle \nabla \cdot \boldsymbol{w}, u \rangle_\beta = -2\varepsilon \int_\Omega (\nabla \cdot \boldsymbol{w}) \, u \beta d\boldsymbol{x}$$
$$= 2\varepsilon \int_\Omega \boldsymbol{w} \cdot \nabla (u\beta) d\boldsymbol{x}$$
$$= 2\varepsilon \int_\Omega (\boldsymbol{w} \cdot \nabla u) \, \beta d\boldsymbol{x} + 2\varepsilon \int_\Omega (\nabla \beta \cdot \boldsymbol{w}) \, u d\boldsymbol{x}$$
$$= 2\varepsilon \langle \boldsymbol{w}, \nabla u \rangle_\beta + 2\varepsilon \int_\Omega (\nabla \beta \cdot \boldsymbol{w}) \, u d\boldsymbol{x},$$

where we use the fact that $u = 0$ on the boundary in the integration by parts step. Note that

$$\langle \boldsymbol{w}, \nabla u \rangle_\beta = \frac{1}{4} \|\boldsymbol{w} + \nabla u\|_\beta^2 - \frac{1}{4} \|\boldsymbol{w} - \nabla u\|_\beta^2,$$

and, consequently, that

$$\varepsilon \|\boldsymbol{w} - \nabla u\|_\beta^2 + 2\varepsilon \langle \boldsymbol{w}, \nabla u \rangle_\beta = \frac{\varepsilon}{2} \|\boldsymbol{w} + \nabla u\|_\beta^2 + \frac{\varepsilon}{2} \|\boldsymbol{w} - \nabla u\|_\beta^2 = \varepsilon \|\boldsymbol{w}\|_\beta^2 + \varepsilon \|\nabla u\|_\beta^2.$$

Thus,

$$a(\mathcal{U}, \mathcal{U}) \ge b_0 \|u\|_\beta^2 + \varepsilon \|\boldsymbol{w}\|_\beta^2 + \varepsilon \|\nabla u\|_\beta^2 + \varepsilon^2 b_1^{-1} \|\nabla \cdot \boldsymbol{w}\|_\beta^2 + \varepsilon^k \|\nabla \times \boldsymbol{w}\|_\beta^2$$
$$+ 2\varepsilon \int_\Omega (\nabla \beta \cdot \boldsymbol{w}) \, u d\boldsymbol{x}.$$

Finally, consider

$$2\varepsilon \left| \int_\Omega (\nabla\beta \cdot \boldsymbol{w})\, u d\boldsymbol{x} \right| = 2\varepsilon \left| \left\langle \boldsymbol{w}, \frac{u}{\beta}\nabla\beta \right\rangle_\beta \right| \leq 2\varepsilon \|\boldsymbol{w}\|_\beta \left\| \frac{u}{\beta}\nabla\beta \right\|_\beta.$$

By our assumption on β,

$$\left\| \frac{u}{\beta}\nabla\beta \right\|_\beta^2 \leq \frac{b_0}{\varepsilon(1+C)^2}\|u\|_\beta^2,$$

and, so,

$$2\varepsilon \left| \int_\Omega (\nabla\beta \cdot \boldsymbol{w})\, u d\boldsymbol{x} \right| \leq 2\frac{\varepsilon^{1/2}b_0^{1/2}}{1+C}\|\boldsymbol{w}\|_\beta\|u\|_\beta.$$

This gives

$$\begin{aligned}
a(\mathcal{U},\mathcal{U}) \geq & b_0\|u\|_\beta^2 + \varepsilon\|\boldsymbol{w}\|_\beta^2 + \varepsilon\|\nabla u\|_\beta^2 + \varepsilon^2 b_1^{-1}\|\nabla \cdot \boldsymbol{w}\|_\beta^2 + \varepsilon^k\|\nabla \times \boldsymbol{w}\|_\beta^2 \\
& - 2\frac{\varepsilon^{1/2}b_0^{1/2}}{1+C}\|\boldsymbol{w}\|_\beta\|u\|_\beta \\
\geq & b_0\left(1 - \frac{1}{(1+C)}\right)\|u\|_\beta^2 + \varepsilon\left(1 - \frac{1}{(1+C)}\right)\|\boldsymbol{w}\|_\beta^2 \\
& + \varepsilon\|\nabla u\|_\beta^2 + \varepsilon^2 b_1^{-1}\|\nabla \cdot \boldsymbol{w}\|_\beta^2 + \varepsilon^k\|\nabla \times \boldsymbol{w}\|_\beta^2 \\
\geq & \min\left(\frac{C\min(1,b_0)}{1+C}, b_1^{-1}, 1\right)|||\mathcal{U}|||_{\beta,k}^2.
\end{aligned}$$

A natural question, in light of this result, is whether a suitable choice of $\beta(\boldsymbol{x})$ exists. We now give a concrete construction of one such family of functions, $\beta(\boldsymbol{x})$, for which the assumption above is satisfied. This family is constructed for the case of $\Omega = [0,1]^d$ with boundary layers along each boundary adjacent to the origin (i.e., where $x_i = 0$ for some i). The extension to boundary layers along all $2d$ boundary faces is straightforward from the construction.

Theorem 2. *Let $C > 0$ be given, and define $\gamma = \frac{b_0^{1/2}}{(1+C)\sqrt{d}}$. Take*

$$\beta(\boldsymbol{x}) = \left(1 + \frac{1}{\sqrt{\varepsilon}}e^{-\gamma x_1/\sqrt{\varepsilon}}\right) \cdots \left(1 + \frac{1}{\sqrt{\varepsilon}}e^{-\gamma x_d/\sqrt{\varepsilon}}\right) \tag{10}$$

Then,

$$\nabla\beta \cdot \nabla\beta < \frac{b_0\beta^2(\boldsymbol{x})}{\varepsilon(1+C)^2},$$

for every $\boldsymbol{x} \in \Omega$.

Proof. A direct calculation shows that

$$\frac{\partial\beta}{\partial x_i} = \frac{\frac{-\gamma}{\varepsilon}e^{-\gamma x_i/\sqrt{\varepsilon}}}{\left(1 + \frac{1}{\sqrt{\varepsilon}}e^{-\gamma x_i/\sqrt{\varepsilon}}\right)}\beta(\boldsymbol{x}).$$

Consequently,

$$\nabla\beta \cdot \nabla\beta = \sum_{i=1}^{d} \left(\frac{\frac{-\gamma}{\varepsilon} e^{-\gamma x_i/\sqrt{\varepsilon}}}{\left(1 + \frac{1}{\sqrt{\varepsilon}} e^{-\gamma x_i/\sqrt{\varepsilon}}\right)} \right)^2 \beta^2(\boldsymbol{x}).$$

Note, however, that

$$\left(\frac{\frac{-\gamma}{\varepsilon} e^{-\gamma x_i/\sqrt{\varepsilon}}}{\left(1 + \frac{1}{\sqrt{\varepsilon}} e^{-\gamma x_i/\sqrt{\varepsilon}}\right)} \right)^2 = \frac{\gamma^2}{\varepsilon} \left(\frac{\frac{1}{\sqrt{\varepsilon}} e^{-\gamma x_i/\sqrt{\varepsilon}}}{\left(1 + \frac{1}{\sqrt{\varepsilon}} e^{-\gamma x_i/\sqrt{\varepsilon}}\right)} \right)^2 \leq \frac{\gamma^2}{\varepsilon}.$$

This gives

$$\nabla\beta \cdot \nabla\beta \leq \frac{d\gamma^2}{\varepsilon} \beta^2(\boldsymbol{x}).$$

Substituting in the chosen value for γ gives the stated result.

The final question to be resolved is whether $\beta(\boldsymbol{x})$ as given in (10) is a "good" choice, in the sense of whether quasi-optimal approximation in the resulting norm is expected to give a good approximation to the layer structure in a typical solution. We consider the case of $d = 2$, the unit square. Following Lemmas 1.1 and 1.2 of [11], we require that the problem data satisfy the assumptions of [11, §2.1], specifically that $f, b \in C^{4,\alpha}(\bar{\Omega})$ and that f vanishes at the corners of the domain. Denoting the four edges of the domain by Γ_i, $1 \leq i \leq 4$, numbered clockwise with the edge $y = 0$ as Γ_1, and the four corners of the domain by c_i, $1 \leq i \leq 4$, numbered clockwise with the origin as c_1, we have the following result.

Lemma 1 ([11, Lemmas 1.1 and 1.2]). *The solution u of (1) can be decomposed as*

$$u = V + W + Z = V + \sum_{i=1}^{4} W_i + \sum_{i=1}^{4} Z_i, \qquad (11a)$$

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

$$\left| \frac{\partial^{m+n} V}{\partial x^m \partial y^n}(x,y) \right| \leq C(1 + \varepsilon^{1-m/2-n/2}), \qquad\qquad 0 \leq m+n \leq 4, \quad (11b)$$

$$\left| \frac{\partial^{m+n} W_1}{\partial x^m \partial y^n}(x,y) \right| \leq C(1 + \varepsilon^{1-m/2})\varepsilon^{-n/2}e^{-y\sqrt{b_0/(2\varepsilon)}}, \quad 0 \leq m+n \leq 3, \quad (11c)$$

$$\left| \frac{\partial^{m+n} W_2}{\partial x^m \partial y^n}(x,y) \right| \leq C\varepsilon^{-m/2}(1 + \varepsilon^{1-n/2})e^{-x\sqrt{b_0/(2\varepsilon)}}, \quad 0 \leq m+n \leq 3, \quad (11d)$$

$$\left| \frac{\partial^{m+n} Z_1}{\partial x^m \partial y^n}(x,y) \right| \leq C\varepsilon^{-m/2-n/2}e^{-(x+y)\sqrt{b_0/(2\varepsilon)}}, \qquad 0 \leq m+n \leq 3, \quad (11e)$$

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

Thus, as a "stereotypical" solution of (1) in the case where boundary layers only form along the edges $x = 0$ and $y = 0$ of $[0, 1]^2$, we can consider

$$u(x) = u_0(x) + c_1 e^{-x\sqrt{b_0/(2\varepsilon)}} + c_2 e^{-y\sqrt{b_0/(2\varepsilon)}} + c_3 e^{-(x+y)\sqrt{b_0/(2\varepsilon)}}.$$

Next, we check if $|||\mathcal{U}|||_{\beta,k}$ is "balanced", not only in the sense of all terms having the same order, but in addition that each component in the stereotypical solution above is well-represented in the norm. This means the norm can be bounded from above and below by ε-independent values, so that it is not seen as being well-approximated by zero in the norm (unless truly vanishingly small), nor that the norm blows up as $\varepsilon \to 0$. For this case, (10) simplifies as

$$\beta(x, y) = \beta_1(x)\beta_1(y) \quad \text{where} \quad \beta_1(x) = 1 + \frac{1}{\sqrt{\varepsilon}} e^{-\gamma x_1/\sqrt{\varepsilon}},$$

and the checks rely on two direct calculations:

$$\int_0^1 \beta_1(x)dx = 1 + \frac{1}{\gamma}\left(1 - e^{-\gamma/\sqrt{\varepsilon}}\right) \approx 1 + \frac{1}{\gamma},$$

$$\int_0^1 \beta_1(x)\left(e^{-x\sqrt{b_0/(2\varepsilon)}}\right)^2 dx = \frac{1}{\gamma + \sqrt{2b_0}}\left(1 - e^{-\gamma/\sqrt{\varepsilon} - \sqrt{2b_0/\varepsilon}}\right)$$
$$+ \sqrt{\frac{\varepsilon}{2b_0}}\left(1 - e^{-2\sqrt{b_0/(2\varepsilon)}}\right)$$
$$\approx \frac{1}{\gamma + \sqrt{2b_0}},$$

With this, assuming that $u_0(x)$ is $\mathcal{O}(1)$ over a nontrivial fraction of the domain, we conclude that

$$|||(u_0, \nabla u_0)^T|||_{\beta,k} \approx 1 + \frac{1}{\gamma},$$

because of the separable nature of the calculation. Thus, the regular part of the solution is well-represented in the norm.

For the W_2 layer term, we write $w_2(x, y) = e^{-x\sqrt{b_0}/2\varepsilon}$ and calculate from the above that

$$\|w_2\|_\beta^2 \approx \left(1 + \frac{1}{\gamma}\right)\frac{1}{\gamma + \sqrt{2b_0}}.$$

Noting that all derivatives of this term with respect to y are zero and that $\partial_x^\ell w_2 = (-\sqrt{b_0/(2\varepsilon)})^\ell w_2$, we compute

$$|||(w_2, \nabla w_2)^T|||_{\beta,k}^2 = \|w_2\|_\beta^2 + \varepsilon\|\nabla w_2\|_\beta^2 + \varepsilon\|\nabla w_2\|_\beta^2 + \varepsilon^2\|\nabla \cdot \nabla w_2\|_\beta^2$$
$$+ \varepsilon^k\|\nabla \times \nabla w_2\|_\beta^2$$
$$= \|w_2\|_\beta^2 + \frac{b_0}{2}\|w_2\|_\beta^2 + \frac{b_0}{2}\|w_2\|_\beta^2 + \left(\frac{b_0}{2}\right)^2\|w_2\|_\beta^2 + 0$$
$$\approx \left(1 + b_0 + \left(\frac{b_0}{2}\right)^2\right)\left(1 + \frac{1}{\gamma}\right)\frac{1}{\gamma + \sqrt{2b_0}}.$$

Again, this shows that the W_2 layer term is well-represented in the norm. Similar calculations show the same to be true for the W_1 layer and Z_1 corner terms in the stereotypical solution.

4 Numerical Results

To test the above approach, we consider a two-dimensional problem with constant $b = 1$ posed on the unit square. We construct a problem whose solution mimics the stereotypical solution discussed above, with two edge layers and one corner layer. Specifically, we choose f so that the solution is

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

We note that this has somewhat more complex layer behaviour than the stereotypical solution, but still obeys the bounds of Lemma 1. Also, the solution is constructed so as to obey the homogeneous Dirichlet boundary conditions. For numerical stability, we rescale the equations by defining $\boldsymbol{w} = \sqrt{\varepsilon}\nabla u$ and making corresponding changes in weights to preserve the balanced nature of the norm. With this, we pick k to match the powers of ε in the weighting terms of both $\|\nabla \cdot \boldsymbol{w}\|_{\beta}^2$ and $\|\nabla \times \boldsymbol{w}\|_{\beta}^2$ in $\|\|\mathcal{U}\|\|_{\beta,k}$, equivalent to taking $k = 2$ above.

We discretize the test problem on a tensor-product Shishkin mesh (see, e.g., [1, §3] for more details). To do this, we select a transition point, $\tau > 0$, and construct a one-dimensional mesh with $N/2$ equal-sized elements on each of the intervals $[0, \tau]$ and $[\tau, 1]$. The two-dimensional mesh is created as a tensor-product of this mesh with itself, with rectangular (quadrilateral) elements. For the choice of τ, we slightly modify the standard choice from the literature (see, for example, [1,10,11]) to account for both the layer functions present in the solution decomposition and in the definition of $\beta(\boldsymbol{x})$ in (10). As such, we take

$$\tau = \min\left\{ \frac{1}{2}, (p+1)\sqrt{\frac{2\varepsilon}{b_0}}\gamma^{-1}\ln N \right\}$$

where p is the degree of the polynomial space ($p = 1$ for bilinear elements, $p = 2$ for biquadratic, and $p = 3$ for bicubic), so that this factor matches the expected L^2 rate of convergence of the approximation, while the terms $\sqrt{2\varepsilon/b_0}\gamma^{-1}$ decrease appropriately as ε does, but increase (corresponding to increasing layer width) with decreases in b_0 or γ. In the results that follow, we take $\gamma = 0.5$, implying $C = \sqrt{2} - 1$. All numerical results were computed using the software for automating the finite-element method, Firedrake [14], for the discretization and a direct solver for the resulting linear systems.

Table 1 shows the expected reduction rates in errors with respect to the mesh parameter, N, if we were to have standard estimates of approximation error in the β-norm on the Shishkin meshes considered here. Tables 2, 3 and 4 show the measured errors (relative to the manufactured solution) for the bilinear, biquadratic, and bicubic discretizations, respectively. Expected behaviour for

Table 1. Expected error reduction rates on a Shishkin mesh.

	$N = 64$	$N = 128$	$N = 256$	$N = 512$
$N^{-1} \ln N$	0.60	0.58	0.57	0.56
$(N^{-1} \ln N)^2$	0.36	0.34	0.33	0.32
$(N^{-1} \ln N)^3$	0.22	0.20	0.19	0.18
$(N^{-1} \ln N)^4$	0.13	0.12	0.11	0.10

Table 2. β-weighted norm and discrete max norm errors for model problem with bilinear discretization.

$\|\|\mathcal{U}^* - \mathcal{U}^N\|\|_{\beta,2}$ (Reduction Rate w.r.t. N)				
ε/N	32	64	128	256	512
10^{-6}	3.086e−01	1.921e−01 (0.62)	1.137e−01 (0.59)	6.531e−02 (0.57)	3.680e−02 (0.56)
10^{-8}	3.086e−01	1.921e−01 (0.62)	1.137e−01 (0.59)	6.532e−02 (0.57)	3.681e−02 (0.56)
10^{-10}	3.086e−01	1.921e−01 (0.62)	1.137e−01 (0.59)	6.533e−02 (0.57)	3.681e−02 (0.56)
10^{-12}	3.086e−01	1.921e−01 (0.62)	1.137e−01 (0.59)	6.533e−02 (0.57)	3.681e−02 (0.56)

$\|u^* - u^N\|_{\ell_\infty}$ (Reduction Rate w.r.t. N)				
ε/N	32	64	128	256	512
10^{-6}	6.935e−02	1.981e−02 (0.29)	6.436e−03 (0.32)	2.051e−03 (0.32)	6.448e−04 (0.31)
10^{-8}	6.945e−02	1.983e−02 (0.29)	6.444e−03 (0.32)	2.053e−03 (0.32)	6.455e−04 (0.31)
10^{-10}	6.946e−02	1.984e−02 (0.29)	6.445e−03 (0.32)	2.054e−03 (0.32)	6.456e−04 (0.31)
10^{-12}	6.946e−02	1.984e−02 (0.29)	6.445e−03 (0.32)	2.054e−03 (0.32)	6.456e−04 (0.31)

Table 3. β-weighted norm and discrete max norm errors for model problem with biquadratic discretization.

$\|\|\mathcal{U}^* - \mathcal{U}^N\|\|_{\beta,2}$ (Reduction Rate w.r.t. N)				
ε/N	32	64	128	256	512
10^{-6}	9.307e−02	3.854e−02 (0.41)	1.394e−02 (0.36)	4.655e−03 (0.33)	1.484e−03 (0.32)
10^{-8}	9.306e−02	3.854e−02 (0.41)	1.394e−02 (0.36)	4.656e−03 (0.33)	1.485e−03 (0.32)
10^{-10}	9.306e−02	3.854e−02 (0.41)	1.394e−02 (0.36)	4.656e−03 (0.33)	1.485e−03 (0.32)
10^{-12}	9.306e−02	3.854e−02 (0.41)	1.394e−02 (0.36)	4.656e−03 (0.33)	1.485e−03 (0.32)

$\|u^* - u^N\|_{\ell_\infty}$ (Reduction Rate w.r.t. N)				
ε/N	32	64	128	256	512
10^{-6}	1.512e−02	1.817e−03 (0.12)	2.715e−04 (0.15)	3.609e−05 (0.13)	4.133e−06 (0.11)
10^{-8}	1.518e−02	1.823e−03 (0.12)	2.730e−04 (0.15)	3.622e−05 (0.13)	4.145e−06 (0.11)
10^{-10}	1.519e−02	1.823e−03 (0.12)	2.733e−04 (0.15)	3.624e−05 (0.13)	4.147e−06 (0.11)
10^{-12}	1.519e−02	1.823e−03 (0.12)	2.734e−04 (0.15)	3.625e−05 (0.13)	4.147e−06 (0.11)

the bilinear case is a reduction like $N^{-1} \ln N$ for $\|\|\mathcal{U}^* - \mathcal{U}^N\|\|_{\beta,2}$ (where $\mathcal{U}^*$ represents the manufactured solution, u^* and its gradient) and like $(N^{-1} \ln N)^2$ for the discrete maximum norm of the error, $\|u^* - u^N\|_{\ell_\infty}$, which is measured

Table 4. β-weighted norm and discrete max norm errors for model problem with bicubic discretization.

$|||\mathcal{U}^* - \mathcal{U}^N|||_{\beta,2}$ (Reduction Rate w.r.t. N)				
ε/N	32	64	128	256	512
10^{-6}	2.786e−02	7.800e−03 (0.28)	1.748e−03 (0.22)	3.419e−04 (0.20)	6.185e−05 (0.18)
10^{-8}	2.785e−02	7.800e−03 (0.28)	1.749e−03 (0.22)	3.420e−04 (0.20)	6.187e−05 (0.18)
10^{-10}	2.785e−02	7.800e−03 (0.28)	1.749e−03 (0.22)	3.420e−04 (0.20)	6.187e−05 (0.18)
10^{-12}	2.785e−02	7.800e−03 (0.28)	1.749e−03 (0.22)	3.420e−04 (0.20)	6.187e−05 (0.18)
$\|u^* - u^N\|_{\ell_\infty}$ (Reduction Rate w.r.t. N)				
ε/N	32	64	128	256	512
10^{-6}	4.364e−03	6.989e−04 (0.16)	9.807e−05 (0.14)	1.148e−05 (0.12)	1.198e−06 (0.10)
10^{-8}	4.370e−03	6.993e−04 (0.16)	9.812e−05 (0.14)	1.149e−05 (0.12)	1.199e−06 (0.10)
10^{-10}	4.371e−03	6.994e−04 (0.16)	9.813e−05 (0.14)	1.149e−05 (0.12)	1.199e−06 (0.10)
10^{-12}	4.371e−03	6.994e−04 (0.16)	9.813e−05 (0.14)	1.149e−05 (0.12)	1.199e−06 (0.10)

at the nodes of the mesh corresponding to the finite-element degrees of freedom. These are both expected to be raised by one power in the biquadratic case, and a further one power for bicubics. In Tables 2, 3, and 4, we see convergence behaviour comparable to these rates, with the exception of the results for the discrete maximum norm in Table 3. These seem to show a superconvergence-type phenomenon, although we have no explanation for this observation at present.

5   Conclusions

In the paper, we propose and analyse a new weighted-norm first-order system least squares methodology tuned for singularly perturbed reaction-diffusion equations that lead to boundary layers. The analysis includes a standard ellipticity result for the FOSLS formulation in a weighted norm, and shows that this norm is suitably weighted to be considered a "balanced norm" for the problem. Numerical results confirm the effectiveness of the method. Future work includes completing the error analysis by proving the necessary interpolation error estimates, with respect to $||| \cdot |||_{\beta,2}$, investigating the observed superconvergence properties, generalizing the theory to convection-diffusion equations, and investigating efficient linear solvers for the resulting discretizations.

References

1. Adler, J.H., MacLachlan, S., Madden, N.: A first-order system Petrov-Galerkin discretisation for a reaction-diffusion problem on a fitted mesh. IMA J. Numer. Anal. **36**(3), 1281–1309 (2016)
2. Berndt, M., Manteuffel, T.A., McCormick, S.F.: Local error estimates and adaptive refinement for first-order system least squares (FOSLS). Electron. Trans. Numer. Anal. **6**, 35–43 (1997)

3. Brezina, M., Garcia, J., Manteuffel, T., McCormick, S., Ruge, J., Tang, L.: Parallel adaptive mesh refinement for first-order system least squares. Numer. Linear Algebra Appl. **19**, 343–366 (2012)

4. Cai, Z., Lazarov, R., Manteuffel, T., McCormick, S.: First-order system least squares for second-order partial differential equations: part I. SIAM J. Numer. Anal. **31**, 1785–1799 (1994)

5. Cai, Z., Manteuffel, T., McCormick, S.: First-order system least squares for second-order partial differential equations: part II. SIAM J. Numer. Anal. **34**(2), 425–454 (1997). https://doi.org/10.1137/S0036142994266066

6. De Sterck, H., Manteuffel, T., McCormick, S., Nolting, J., Ruge, J., Tang, L.: Efficiency-based h- and hp-refinement strategies for finite element methods. Numer. Linear Algebra Appl. **15**(2–3), 89–114 (2008). https://doi.org/10.1002/nla.567

7. Heuer, N., Karkulik, M.: A robust DPG method for singularly perturbed reaction-diffusion problems. SIAM J. Numer. Anal. **55**(3), 1218–1242 (2017). https://doi.org/10.1137/15M1041304

8. Lee, E., Manteuffel, T.A., Westphal, C.R.: Weighted-norm first-order system least squares (FOSLS) for problems with corner singularities. SIAM J. Numer. Anal. **44**(5), 1974–1996 (2006)

9. Lee, E., Manteuffel, T.A., Westphal, C.R.: Weighted-norm first-order system least-squares (FOSLS) for div/curl systems with three dimensional edge singularities. SIAM J. Numer. Anal. **46**(3), 1619–1639 (2008)

10. Lin, R., Stynes, M.: A balanced finite element method for singularly perturbed reaction-diffusion problems. SIAM J. Numer. Anal. **50**(5), 2729–2743 (2012). https://doi.org/10.1137/110837784

11. Liu, F., Madden, N., Stynes, M., Zhou, A.: A two-scale sparse grid method for a singularly perturbed reaction-diffusion problem in two dimensions. IMA J. Numer. Anal. **29**(4), 986–1007 (2009). https://doi.org/10.1093/imanum/drn048

12. Melenk, J.M., Xenophontos, C.: Robust exponential convergence of hp-FEM in balanced norms for singularly perturbed reaction-diffusion equations. Calcolo **53**(1), 105–132 (2016). https://doi.org/10.1007/s10092-015-0139-y

13. Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Fitted Numerical Methods for Singular Perturbation Problems, revised edn. World Scientific Publishing Co., Pte. Ltd., Hackensack (2012). https://doi.org/10.1142/9789814390743

14. Rathgeber, F., et al.: Firedrake: automating the finite element method by composing abstractions. ACM Trans. Math. Softw. (TOMS) **43**(3), 24 (2017)

15. Roos, H.G., Stynes, M., Tobiska, L.: Robust Numerical Methods for Singularly Perturbed Differential Equations. Springer Series in Computational Mathematics, vol. 24, 2nd edn. Springer, Berlin (2008). https://doi.org/10.1007/978-3-540-34467-4

16. Russell, S., Stynes, M.: Balanced-norm error estimates for sparse grid finite element methods applied to singularly perturbed reaction-diffusion problems. J. Numer. Math. **27**(1), 37–55 (2019). https://doi.org/10.1515/jnma-2017-0079