Boundary layer preconditioners for finite-element discretizations of singularly perturbed reaction-diffusion problems

Thái Anh Nhan · Scott MacLachlan · Niall Madden

Received: date / Accepted: date

Abstract We consider the iterative solution of linear systems of equations arising from the discretization of singularly perturbed reaction-diffusion differential equations by finite-element methods on boundary-fitted meshes. The equations feature a perturbation parameter, which may be arbitrarily small, and, correspondingly, their solutions feature layers: regions where the solution changes rapidly. Therefore, numerical solutions are computed on specially designed, highly anisotropic layer-adapted meshes. Usually, the resulting linear systems are ill-conditioned, and, so, careful design of suitable preconditioners is necessary in order to solve them in a way that is robust, with respect to the perturbation parameter, and efficient. We propose a *boundary layer preconditioner*, in the style of that introduced by MacLachlan and Madden for a finite-difference method [14]. We prove the optimality of this preconditioner and establish a suitable stopping criterion for one-dimensional problems. Numerical results are presented which demonstrate that the ideas extend to problems in two dimensions.

Keywords singularly perturbed · layer-adapted meshes · robust multigrid · preconditioning

The work of SM was partially supported by an NSERC Discovery Grant; the work of TAN was supported by the Irish Research Council under Grant No. RS/2011/179.

T.A. Nhan
Department of Mathematics, Ohlone College,
43600 Mission Blvd., Fremont, CA 94539, USA
E-mail: anhan@ohlone.edu

S. MacLachlan
Department of Mathematics and Statistics
Memorial University of Newfoundland
St John's, NL, Canada
E-mail: smaclachlan@mun.ca

N. Madden
School of Mathematics, Statistics and Applied Mathematics,
National University of Ireland, Galway
Galway, Ireland
E-mail: Niall.Madden@NUIGalway.ie

1 Introduction

We consider the solution of linear systems of equations, by iterative methods, that arise in the discretisations of singularly perturbed reaction-diffusion differential equations by *finite-element* (FE) methods. Our model problems are:

$$-\varepsilon^2 u'' + b(x)u = f(x) \text{ on } \Omega := (0,1), \qquad u(0) = u(1) = 0, \tag{1}$$

and

$$-\varepsilon^2 \Delta u + b(x,y)u = f(x,y) \text{ on } \Omega := (0,1)^2, \quad u(x,y) = 0 \quad \text{for } (x,y) \in \partial\Omega. \tag{2}$$

Here, ε is a positive parameter, which may be arbitrarily small. We will assume that there are positive constants β_0 and β_1 such that, at all points in $\bar{\Omega}$, we have $0 < \beta_0^2 \leq b \leq \beta_1^2$. We are concerned with the case where the problem is *singularly perturbed*: $\varepsilon \ll 1$. The solution to (1) will typically possess two boundary layers: near $x = 0$ and $x = 1$. The two-dimensional problem (2) typically features four edge and four corner layers.

Efficient and accurate solution of these problems is not possible using standard techniques, such as classical finite-element methods applied on uniform meshes, as these methods require unreasonable assumptions concerning the number of degrees of freedom required in order to adequately resolve the layers expected in the solution. However, there is a large, and rapidly growing, literature on the design of techniques for accurately solving these problems, and their variants (such as coupled systems, time-dependent problems, semilinear equations, etc.); see, e.g., [8, 12, 18], and the many references therein. Most modern studies focus on the application of standard discretizations on special layer-adapted meshes which resolve any layers present, and have convergence properties that do not depend adversely on the perturbation parameter [12]. Such methods are termed "*parameter robust*".

The design of suitable linear solvers for parameter robust methods has received very little attention. This is an oversight since, as demonstrated in [14, §4.1], the computation time taken by a standard direct solver applied to a finite-difference discretization of (2) depends badly on ε, due to the propagation of subnormal numbers. (Although we don't repeat the arguments of [14, §4.1], the same phenomenon occurs for finite-element discretizations; see Tables 7 and 8). Furthermore, standard direct solvers are unlikely to be useful for high-dimensional problems, so one must employ iterative solvers in such cases. However, as we show in Section 2.3, the condition number of the linear system, constructed to solve (1) on a layer-adapted mesh, is inversely proportional to ε and, so, careful construction of robust preconditioners is required. The same is true for the linear system associated with (2), as demonstrated in Section 4.5.1.

There are a small number of published papers on the topic of linear solvers for singularly perturbed problems, mostly concerned with convection-diffusion problems; we refer the reader to the literature review in [14, §1] for an extensive discussion. The present study may be seen as a companion to [14], which proposed and analysed a robust *boundary layer preconditioner* for a finite-difference discretization of reaction-diffusion problems in one and two dimensions on layer adapted meshes. We

wish to extend this approach to finite-element discretizations based on linear elements for the one-dimensional problem, and bilinear elements on tensor-product grids for the two-dimensional problem. This presents a number of challenges.

1. For a finite-difference discretization, the zeroth-order term contributes only to the diagonal entry of the system matrix. Away from boundaries, this term dominates and, so, the application of a diagonal preconditioner in this region is both natural and easy to analyse. However, the corresponding term in the finite-element discretization gives non-zero off-diagonal terms in the system matrix. This complicates the analysis in Section 3, necessitating the introduction and optimization of a parameter in the proof of Theorem 2 that is not necessary in the finite-difference case.
2. For two-dimensional problems, the finite-difference method analysed in [14] has a five-point stencil, whereas the finite-element method we study here has a nine-point stencil, again complicating the method and its analysis.
3. The condition number of the (unsymmetrised) linear systems yielded by finite-difference methods for Problems (1) and (2) can be bounded independently of ε (see [17, Remark 2]). In contrast, the condition number of the FE discretization depends badly on ε.

For these reasons, in this paper we focus our analysis on the one-dimensional problem. This then motivates the design and implementation of a preconditioner for the two-dimensional problem. Numerical results demonstrate that it is highly effective, and significantly more efficient, for small values of ε, than application of a standard direct solver, even for a moderate number of degrees of freedom. Moreover, it scales optimally with both ε and the mesh discretization parameter.

Our analysis and examples are specifically for Shishkin Meshes, which are ubiquitous in the literature on the parameter robust solution of singularly perturbed problems [15]. However, we emphasise that the approach we take here applies directly to other layer-adapted meshes, whether they are constructed using *a priori* knowledge, such as Shishkin meshes, or the graded meshes of Bakhvalov [3], or adaptively using *a posteriori* information (see, e.g., [10]).

1.1 Outline

In Section 2, we review the basics of piecewise linear finite-element discretization for the one-dimensional problem. The Shishkin mesh is introduced, and basic properties of the discretization matrix are shown. Since we consider iterative solution techniques in this paper, Section 2.4 details stopping criteria for achieving full (asymptotic) accuracy of the finite-element discretization for these problems. The block-structured preconditioner for the one-dimensional problem is introduced in Section 3. Here, the main spectral equivalence theorem is proven, along with several useful corollaries, and numerical results in 1D are given in Section 3.2.

In Section 4, we turn our attention to the two-dimensional problem, first providing a condition number estimate for the bilinear FEM discretization on tensor-product

fitted meshes. The extension of the boundary-layer preconditioner from one to two-dimensions is presented in Section 4.4. Numerical results are given in Section 4.5. Concluding remarks and directions for future work are given in Section 5.

1.2 Notation and assumptions

Throughout this paper, we shall use the letter C (with or without subscripts) to denote a generic positive constant that may stand for different values in different places, but is always independent of the perturbation parameter, ε, and mesh parameter N. We use $D = \text{diag}(M)$ to denote the diagonal matrix with entries $d_{i,i} = m_{i,i}$ for all i, but $d_{i,j} = 0$ for $i \neq j$.

In Sections 2 and 3, we shall focus on one-dimensional problems, where solutions to the weak formulation of (1) exist in $H_0^1(\Omega) = \{v \in H^1(\Omega) : v \mid_{\delta\Omega} = 0\}$, where $H^1(\Omega)$ is the usual Sobolev space (see, e.g., [18, §2.2]). The energy norm associated with (1) is

$$\|u\|_\varepsilon := \sqrt{\varepsilon^2 \|u'\|_0^2 + \beta_0^2 \|u\|_0^2}, \qquad \forall v \in H_0^1(\Omega), \tag{3}$$

where $(u,v) := \int_0^1 u(x)v(x)dx$, and $\|u\|_0^2 = (u,u)$.

For discretization, a mesh with N subintervals on $\Omega = [0,1]$ is denoted by $\omega^N := \{0 = x_0 < x_1 < \cdots < x_N = 1\}$, and the local mesh widths are $h_i = x_i - x_{i-1}$. We write $b_i = b((x_{i-1} + x_i)/2)$ to denote the midpoint values of b on the mesh ω^N. In 1D, we discretize (1) using piecewise linear finite elements on ω^N, denoting the approximation space as $\mathcal{V}^N$, and approximating $b(x) \approx b_i$ for $x_{i-1} < x < x_i$. As a consequence of this, for $u^N \in \mathcal{V}^N$, if U^N denotes the vector of coefficients of u^N (in terms of a standard nodal basis set for $\mathcal{V}^N$), then $\|u^N\|_\varepsilon \leq \|U^N\|_A \leq (\beta_1/\beta_0)\|u^N\|_\varepsilon$, where A is the (SPD) discretization matrix corresponding to the operator in (1) and the basis for $\mathcal{V}^N$ and $\|\cdot\|_A$ is the standard norm induced by A. For vectors, we denote the standard Euclidean norm by $\|\cdot\|_2$. When considering two-dimensional problems in Section 4, we use the natural generalizations of the above norms, both continuum and discrete.

Unless otherwise noted, we always assume that the problem considered is singularly perturbed relative to the mesh ω^N. While this can be generally expressed in an asymptotic sense by requiring that $\varepsilon N \ll 1$ (i.e., that a "typical" mesh-width of $1/N$ is not comparable to ε), we make use of two concrete assumptions to formalize this. A strict assumption can be made with no assumptions on ω^N other than that it has N mesh intervals, requiring that there is a constant, C, such that

$$\varepsilon N^2 \leq C. \tag{4}$$

In practice, a weaker assumption is usually sufficient. For a layer-adapted mesh with constant mesh-spacing h_I away from the boundary layers, we can define

$$\delta_h = \left(\varepsilon/(h_I\beta_0)\right)^2, \tag{5}$$

and make the assumption that $\delta_h \ll 1$.

2 Discretization for one-dimensional problems

2.1 A simple finite-element method

We discretize (1) by first taking its variational formulation: find $u \in H_0^1(\Omega)$ such that

$$B_\varepsilon(u,v) := \varepsilon^2(u',v') + (bu,v) = (f,v) \quad \text{for all } v \in H_0^1(\Omega). \tag{6}$$

The finite-element formulation is arrived at by replacing $H_0^1(\Omega)$ with a suitably chosen finite-dimensional subspace. The natural choice is the space of piecewise linear functions on a mesh, ω^N, which we denote by $\mathscr{V}^N$. Since it may be that b is not easily integrated analytically, we use a quadrature rule equivalent to approximating it by a piecewise constant function on ω^N. Then the finite-element method on ω^N for (1) leads to the linear system

$$AU^N = F, \tag{7}$$

where U^N is the vector of coefficients for the expansion of the finite-element approximation, u^N, in terms of the chosen basis for $\mathscr{V}^N$. The system matrix is $A = S + M$, where (in stencil notation)

$$S = \left[-\frac{\varepsilon^2}{h_i} \quad \frac{\varepsilon^2}{h_i} + \frac{\varepsilon^2}{h_{i+1}} \quad -\frac{\varepsilon^2}{h_{i+1}} \right], \quad \text{and} \quad M = \left[\frac{h_i b_i}{6} \quad \frac{h_i b_i + h_{i+1} b_{i+1}}{3} \quad \frac{h_{i+1} b_{i+1}}{6} \right]. \tag{8}$$

Then, using standard finite-element analysis arguments, one can show that the following quasi-optimality result holds: there is a constant C, which is independent of ε, such that

$$\|u - u^N\|_\varepsilon \le C \|u - v^N\|_\varepsilon, \quad \text{for all } v^N \in \mathscr{V}^N.$$

Therefore, the error analysis is purely dependent on the approximation properties of the space $\mathscr{V}^N$ which, in turn, depends on ω^N.

2.2 Shishkin mesh

Numerous fitted meshes have been proposed for this problem, the most commonly studied ones being the piecewise uniform mesh of Shishkin [15], and the graded mesh of Bakhvalov [3]. To construct a piecewise uniform "Shishkin" mesh, we first define the *mesh transition point*

$$\tau = \min\left\{ \frac{1}{4}, 2\frac{\varepsilon}{\beta_0} \ln N \right\}, \tag{9}$$

where N is the number of mesh intervals, which is assumed to be an integer multiple of 4. Then, Ω is divided into three subintervals: $[0,\tau]$, $[\tau, 1-\tau]$ and $[1-\tau, 1]$. The mesh is constructed by subdividing $[\tau, 1-\tau]$ into $N/2$ equally sized mesh intervals, and subdividing each of $[0,\tau]$ and $[1-\tau,1]$ into $N/4$ equally sized mesh intervals, as shown in Figure 1.

For this mesh, one can show that there is a constant, C, independent of N and ε, such that

$$\|u - u^N\|_\varepsilon \le C\big(\varepsilon^{1/2} N^{-1} \ln N + N^{-2}\big). \tag{10}$$

For details, including a discussion of the effects of quadrature, see [12, Thm. 6.6].

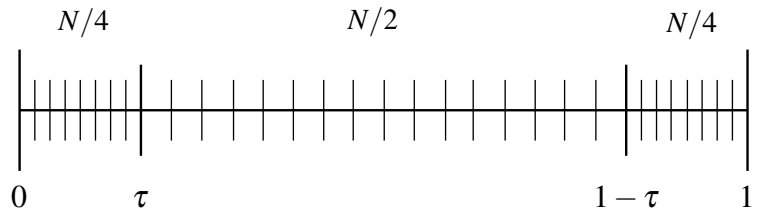

Fig. 1 A Shishkin mesh for a one-dimensional reaction-diffusion problem

Example 1 Consider the following one-dimensional reaction-diffusion problem:

$$-\varepsilon^2 u''(x) + u(x) = e^x \text{ on } (0,1), \qquad u(0) = u(1) = 0. \tag{11}$$

Table 1 shows computed errors, in the energy norm, by the approximation obtained by the finite-element method on a Shishkin mesh. This agrees with the error estimate given in (10). Note that, for small ε, the error is $\mathcal{O}(\varepsilon^{1/2} N^{-1} \ln N)$. One might expect that, if $\varepsilon \ll N^{-1}$, the bound in (10) would simplify to N^{-2}. However, the term $\varepsilon^{1/2} N^{-1} \ln N$ stems from first-order term in the energy norm and tends to dominate.

Table 1 $\|u - u^N\|_\varepsilon$ with u defined by (11) approximated using piecewise linear FEM on a Shishkin mesh

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{12}$
1	3.756e-03	1.878e-03	9.390e-04	4.695e-04	2.347e-04	1.174e-04
10^{-2}	1.449e-02	7.243e-03	3.621e-03	1.811e-03	9.054e-04	4.527e-04
10^{-4}	1.791e-02	1.024e-02	5.762e-03	3.201e-03	1.761e-03	9.604e-04
10^{-6}	5.664e-03	3.239e-03	1.822e-03	1.012e-03	5.568e-04	3.037e-04
10^{-8}	1.791e-03	1.024e-03	5.762e-04	3.202e-04	1.761e-04	9.605e-05
10^{-10}	5.667e-04	3.239e-04	1.822e-04	1.012e-04	5.568e-05	3.037e-05
10^{-12}	1.799e-04	1.025e-04	5.763e-05	3.202e-05	1.761e-05	9.605e-06

2.3 Condition number estimate

In general, layer-adapted meshes have mesh intervals of width $\mathcal{O}(N^{-1}\varepsilon)$ near the boundaries, but of width $\mathcal{O}(N^{-1})$ in the interior. As we now show, the condition number of the unpreconditioned discrete system arising from the finite-element discretization on such a mesh is unbounded as $\varepsilon \to 0$.

We denote the equidistant mesh width of the interior region of layer-adapted meshes by h_I, which we assume to be the largest mesh width and to be $\mathcal{O}(N^{-1})$ when $\varepsilon \ll 1$. Furthermore, we define

$$h_{\min} = \min_{i=1,\ldots,N} \{h_i\}.$$

Theorem 1 *Let A be the matrix associated with the linear system* (7)*, and assume that ω^N satisfies both the strict condition in* (4) *and that there exists $C_h > 0$ such that $h_{\min} \geq C_h \varepsilon / N$. Then, there is a constant C, independent of both N and ε, such that*

$$\kappa_2(A) \leq C(N h_{\min})^{-1}.$$

Proof Recall that the condition number of the matrix A, associated with the 2-norm, is $\kappa_2(A) := \|A\|_2 \|A^{-1}\|_2$. By examining the entries of A as defined in (8), and applying Geršgorin's Theorem, we easily see that

$$\begin{aligned} \|A\|_2 = \lambda_{\max}(A) &\leq \max_i \left\{ \frac{\beta_1^2}{2}(h_i + h_{i+1}) + 2\varepsilon^2 \left(\frac{1}{h_i} + \frac{1}{h_{i+1}}\right) \right\} \\ &\leq C \left(h_I + \frac{\varepsilon^2}{h_{\min}} \right) \leq C N^{-1}, \end{aligned} \tag{12}$$

where we use the assumptions on h_I and $h_{\min}$, as well as that in (4) to achieve the final bound.

In addition, we can bound the smallest eigenvalue of A from below by Geršgorin's Theorem, giving

$$\lambda_{\min}(A) \geq \min_i \left\{ \frac{h_i b_i + h_{i+1} b_{i+1}}{6} \right\} \geq \frac{\beta_0^2 h_{\min}}{3}. \tag{13}$$

A combination of (12) and (13) implies the estimate.

In practice, one finds that this bound is quite sharp for small ε, and the associated constant is $\mathcal{O}(1)$. Therefore, as $\varepsilon \to 0$ the system (7) is ill-conditioned. In particular, for the Shishkin mesh, $h_{\min} = 8\varepsilon \ln N / (N\beta_0)$, implying that $\kappa_2(A) \leq C(\varepsilon \ln N)^{-1}$. In Table 2, it is shown that this bound is sharp, for sufficiently small ε. For example, for the problem data of (11) one has $C \approx 3$.

Table 2 $\kappa_2(A)$ of the problem (11) discretized using piecewise linear FEM on a Shishkin mesh.

ε^2	$N = 2^4$	$N = 2^5$	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$
1	1.16e+02	4.64e+02	1.85e+03	7.42e+03	2.97e+04	1.19e+05
10^{-2}	1.04e+01	4.07e+01	1.62e+02	6.47e+02	2.59e+03	1.03e+04
10^{-4}	1.54e+01	2.07e+01	5.82e+01	1.72e+02	5.30e+02	1.68e+03
10^{-6}	1.59e+02	1.35e+02	1.16e+02	1.72e+02	5.30e+02	1.68e+03
10^{-8}	1.59e+03	1.35e+03	1.16e+03	1.01e+03	8.94e+02	1.68e+03
10^{-10}	1.59e+04	1.36e+04	1.17e+04	1.01e+04	8.95e+03	7.98e+03
10^{-12}	1.59e+05	1.36e+05	1.17e+05	1.01e+05	8.95e+04	7.98e+04

2.4 Stopping criteria

Since we consider iterative methods for the solution of $AU^N = F$, we need to derive suitable stopping criteria for the preconditioned conjugate-gradient algorithm. The approach presented here is similar in spirit to [14, §4.6] which was concerned with finite-difference approximations and maximum norm estimates. We now adapt that reasoning to the setting of finite-element discretizations and energy norm estimates.

We require any stopping criterion to be feasible, in the sense of not needing to compute a residual (for example) with entries comparable to floating point unit round-off. However, as we now show, this may not be possible for an unpreconditioned problem for the cases of interest, where $\varepsilon \ll N^{-1}$. This motivates the analysis of the preconditioned residual approach which, as our numerical experiments show, is effective.

Recall that U^N is the solution of the discrete problem (7). Let $U^{(k)}$ be the k^{th} iterate computed by an iterative procedure applied to (7), and let $u^{(k)} \in \mathcal{V}^N$ be the function whose coefficients in the finite-element basis are given by the vector $U^{(k)}$. Naturally, we wish to choose k so that $u^{(k)}$ is as good an approximation to u as u^N. That is, we ask for

$$\|u - u^{(k)}\|_\varepsilon \simeq \|u - u^N\|_\varepsilon.$$

Since

$$\|u - u^{(k)}\|_\varepsilon \leq \|u - u^N\|_\varepsilon + \|u^N - u^{(k)}\|_\varepsilon,$$

and $\|u^N - u^{(k)}\|_\varepsilon \leq \|U^N - U^{(k)}\|_A$, this means finding $U^{(k)}$ such that

$$\|U^N - U^{(k)}\|_A \leq C\|u - u^N\|_\varepsilon,$$

where C is some moderately small positive constant. Of course, in practice, U^N is unknown, so we must estimate the solver error $E^{(k)} = U^N - U^{(k)}$. This can be done with the residual

$$R^{(k)} = F - AU^{(k)} = F - A(U^N - E^{(k)}) = AE^{(k)},$$

giving $E^{(k)} = A^{-1}R^{(k)}$. Since A is symmetric and positive definite, so too is $A^{-1/2}$, the principle square root of A^{-1}. Thus $\|A^{-1/2}\|_2 = \|A^{-1}\|_2^{1/2}$, giving

$$\begin{aligned}\|E^{(k)}\|_A = \sqrt{(E^{(k)})^T A E^{(k)}} = \sqrt{(E^{(k)})^T A^T A^{-1} A E^{(k)}} \\ = \|A^{-1/2}R^{(k)}\|_2 \leq \|A^{-1/2}\|_2 \|R^{(k)}\|_2.\end{aligned}$$

The bounds on $\|u - u^N\|_\varepsilon$ and $\|A^{-1}\|_2$, from (10) and (13), respectively, lead to the stopping criterion that

$$\|R^{(k)}\|_2 \leq C(\varepsilon N^{-3/2} \ln^{3/2} N + \varepsilon^{1/2} N^{-5/2} \ln^{1/2} N), \tag{14}$$

to ensure $\|E^{(k)}\|_A \leq C\|u - u^N\|_\varepsilon$. Considering the case where ε is very small, this leads to a required residual reduction that may not be feasible in a finite-precision setting (particularly when generalized to two dimensions as in Section 4.3).

Instead, we can use a natural stopping criterion for the preconditioned conjugate gradient algorithm. Let $\hat{A}^{-1}$ be a good preconditioner for the matrix A in the sense that $\hat{A}^{-1}A \approx I$. Let $Z^{(k)} = \hat{A}^{-1}R^{(k)}$ be the preconditioned residual. Then, the inner product of residual and preconditioned residual can be used to estimate $\|E^{(k)}\|_A$ because

$$\left(Z^{(k)}\right)^T R^{(k)} = \left(E^{(k)}\right)^T A\hat{A}^{-1}AE^{(k)} \approx \|E^{(k)}\|_A^2.$$

Assuming the approximation to hold, it is straightforward to see that the stopping criterion needed to imply that $\|E^{(k)}\|_A \leq C\|u-u^N\|_\varepsilon$ for a Shishkin mesh ω^N is

$$\sqrt{\left(Z^{(k)}\right)^T R^{(k)}} \leq C(\varepsilon^{1/2}N^{-1}\ln N + N^{-2}). \tag{15}$$

Adapting these arguments to other meshes is straightforward, as long as analogous bounds to those in (10) and (13) are known for the mesh generation scheme to allow the necessary inequalities to be made. Adapting to other norms is also possible, as was done for the maximum norm in the finite-difference case in [14], although this may require further analysis to relate the continuum norm of $u-u^N$ to the relevant discrete norm of $U^N - U^{(k)}$.

3 A boundary layer preconditioning approach

3.1 Motivation and analysis

A boundary layer-adapted mesh is very fine in the region close to the boundary, but coarse and (typically) uniform in the interior. The scaling of the problems in these regions is very different and, so, it is natural to precondition them differently. Close to the boundaries, the linear system in (7) resembles that of a classical (i.e., non-singularly perturbed) problem, which is amenable to solution using standard techniques, such as multigrid methods. In the interior, the entries are dominated by the contribution from the reaction term, so we employ a diagonal scaling in this region.

Following the ideas in [14], we partition the mesh into two pieces:

1. the left and right layer regions (including their end points), $\omega_B^N := \omega^N \cap ([0,\tau] \cup [1-\tau,1])$, where the subscript B denotes the boundary region; and
2. the complement of ω_B^N, $\omega_I^N := \omega^N \backslash \omega_B^N$, where the subscript I denotes the interior region.

Then, the re-ordered mesh is denoted by $\tilde{\omega}^N := [\omega_B^N \;\; \omega_I^N]$. This ordering is used to partition the matrix $A = S + M$ as

$$A = \begin{pmatrix} A_{BB} & A_{BI} \\ A_{IB} & A_{II} \end{pmatrix} = \begin{pmatrix} S_{BB} & S_{BI} \\ S_{IB} & S_{II} \end{pmatrix} + \begin{pmatrix} M_{BB} & M_{BI} \\ M_{IB} & M_{II} \end{pmatrix}. \tag{16}$$

The submatrices A_{BI} and A_{IB} contain only two nonzero entries each, and make only a very modest contribution to the system. The matrix A_{II} may be approximated, in the spectral sense, by a suitably chosen diagonal matrix, noting that the entries in S_{II} are dominated by those in M_{II}. We approximate M_{II} by

$$D_{II} = m\mathrm{diag}(M_{II}), \tag{17}$$

where m is a positive parameter whose choice is informed by the analysis below. Defining

$$A_D = \begin{pmatrix} A_{BB} & 0 \\ 0 & D_{II} \end{pmatrix}, \tag{18}$$

we use A_D as an "ideal" preconditioner for A in the sense that (as Theorem 2 will show) it is spectrally equivalent to A with modest constants, but not yet a practical preconditioner in terms of efficiency (at least when generalized to higher dimensions, as discussed in Section 4).

Recall that we have assumed that $0 < \beta_0^2 \le b(x) \le \beta_1^2$ for all $x \in [0,1]$. Let

$$\gamma := \frac{\beta_1^2}{\beta_0^2 + \beta_1^2} < 1. \tag{19}$$

In (5), we defined

$$\delta_h = \left(\varepsilon/(h_I \beta_0)\right)^2,$$

Since we are interested in the case where $\varepsilon N \ll 1$, we consider only the case where $\delta_h \ll 1$.

Theorem 2 *Let A be the system matrix in* (7) *for the finite-element solution on a layer adapted mesh, reordered as in* (16)*, and let A_D be as defined in* (18)*. Let m be the parameter in* (17) *and q be any number such that $2m < q < 2m/\gamma$. Then, for all vectors V,*

$$\left(\theta_q - \frac{3q}{m}\sqrt{2\gamma}\delta_h - \frac{9q}{m}\delta_h^2\right) V^T A_D V \le V^T A V \le \left(1 + \frac{3}{2m} + \frac{6}{m}\delta_h + \frac{9}{m}\delta_h^2\right) V^T A_D V, \tag{20}$$

where

$$\theta_q = \min\left\{1 - \frac{\gamma q}{2m}, \frac{1}{2m} - \frac{1}{q}\right\} > 0.$$

Proof In the same way that we partitioned $\tilde{\omega}^N = [\omega_B^N \quad \omega_I^N]$, we partition a generic vector, V, as $V = [V_B \quad V_I]^T$, and note that

$$V^T A V^T = V_B^T A_{BB} V_B + 2V_B^T A_{BI} V_I + V_I^T A_{II} V_I,$$

and

$$V^T A_D V^T = V_B^T A_{BB} V_B + V_I^T D_{II} V_I.$$

Therefore, we require bounds for $V_I^T A_{II} V_I$ and $V_B^T A_{BI} V_I$ in terms of $V_B^T A_{BB} V_B$ and $V_I^T A_{II} V_I$.

Firstly, to bound $V_I^T A_{II} V_I$, we see that

$$\frac{1}{2m} V_I^T D_{II} V_I \le V_I^T A_{II} V_I, \tag{21}$$

since $V_I^T \left(A_{II} - \frac{1}{2m} D_{II}\right) V_I \ge 0$ for all V_I. We get an upper bound on $V_I^T A_{II} V_I$ in two parts, writing $A_{II} = S_{II} + M_{II}$, as in (16). By Geršgorin's Theorem, S_{II} can be

bounded, in the sense of spectral equivalence, by the diagonal matrix whose nonzero entries are $(4\varepsilon^2/h_I)$. Also, for any i,

$$\frac{4\varepsilon^2}{h_I} = \frac{4\varepsilon^2 h_I \beta_0^2}{h_I^2 \beta_0^2} \leq \delta_h \frac{6}{m} \frac{m h_I (b_i + b_{i+1})}{3}.$$

So, for any V_I,

$$V_I^T S_{II} V_I \leq \delta_h \frac{6}{m} V_I^T D_{II} V_I.$$

By Geršgorin's Theorem again, M_{II} can be bounded by the diagonal matrix whose i^{th} diagonal entry is

$$\frac{h_I(b_i + b_{i+1})}{2} = \frac{3}{2m} \left(\frac{m h_I (b_i + b_{i+1})}{3} \right).$$

Hence,

$$V_I^T M_{II} V_I \leq \frac{3}{2m} V_I^T D_{II} V_I.$$

Thus, combining this with (21), we get that

$$\frac{1}{2m} V_I^T D_{II} V_I \leq V_I^T A_{II} V_I \leq \left(\frac{3}{2m} + 6\frac{\delta_h}{m} \right) V_I^T D_{II} V_I.$$

We proceed to finding bounds for $\left| V_B^T A_{BI} V_I \right|$. Set $D_{BB} = m\text{diag}(M_{BB})$. By the Cauchy-Schwarz inequality,

$$\left| V_B^T A_{BI} V_I \right| \leq \left\| D_{BB}^{-1/2} A_{BI} V_I \right\|_2 \left\| D_{BB}^{1/2} V_B \right\|_2,$$

for any V_B and V_I. Since $2m V_B^T S_{BB} V_B \geq 0$, and noting that

$$2mM_{BB} - D_{BB} = \frac{1}{3} \begin{bmatrix} m h_i b_i & m(h_i b_i + h_{i+1} b_{i+1}) & m h_{i+1} b_{i+1} \end{bmatrix},$$

which implies $V_B^T \left(2mM_{BB} - D_{BB} \right) V_B \geq 0$, so we have $V_B^T D_{BB} V_B \leq 2m V_B^T A_{BB} V_B$ for all V_B. Therefore,

$$\left| V_B^T A_{BI} V_I \right| \leq \sqrt{2m} \left\| D_{BB}^{-1/2} A_{BI} V_I \right\|_2 \left(V_B^T A_{BB} V_B \right)^{1/2}, \tag{22}$$

for any V_B and V_I.

To bound $\left\| D_{BB}^{-1/2} A_{BI} V_I \right\|_2$, we use that

$$\left\| D_{BB}^{-1/2} A_{BI} V_I \right\|_2 = \left\| D_{BB}^{-1/2} (S_{BI} + M_{BI}) V_I \right\|_2 \leq \left\| D_{BB}^{-1/2} S_{BI} V_I \right\|_2 + \left\| D_{BB}^{-1/2} M_{BI} V_I \right\|_2.$$

There are only two nonzero entries in each of S_{BI} and $S_{IB} = S_{BI}^T$. They are in the first and last columns, and on different rows. So $S_{IB} D_{BB}^{-1} S_{BI}$ has only two nonzero entries,

$$s_1 := \frac{3\varepsilon^4}{m(h_{N/4} b_{N/4} + h_{N/4+1} b_{N/4+1}) h_{N/4+1}^2},$$

and

$$s_2 := \frac{3\varepsilon^4}{m(h_{3N/4}b_{3N/4} + h_{3N/4+1}b_{3N/4+1})h_{3N/4}^2},$$

which are the first and last entries on the diagonal, and with $h_{N/4+1} = h_{3N/4} = h_I$. Then, s_1 and s_2 can be bounded from above by

$$\frac{3\varepsilon^4}{h_I^3 m\beta_0^2} = \frac{9\varepsilon^4}{2m^2h_I^4\beta_0^4}\frac{2mh_I\beta_0^2}{3} \leq \delta_h^2\frac{9}{2m^2}\frac{mh_I(b_{i+1}+b_{i+2})}{3},$$

for any i. Thus,

$$V_I^T S_{IB}D_{BB}^{-1}S_{BI}V_I \leq \delta_h^2\frac{9}{2m^2}V_I^T D_{II}V_I,$$

and so,

$$\left\|D_{BB}^{-1/2}S_{BI}V_I\right\|_2 \leq \delta_h\frac{3}{m\sqrt{2}}\left\|D_{II}^{1/2}V_I\right\|_2.$$

We use a similar argument to bound the term involving M_{BI}. From the definition of γ in (19), we see that $1/2 \leq \gamma < 1$. Also, because

$$\frac{b_{i+1}}{b_i + b_{i+1}} \leq \frac{b_{i+1}}{\beta_0^2 + b_{i+1}} \leq \frac{\beta_1^2}{\beta_0^2+\beta_1^2} = \gamma,$$

it follows that $b_{i+1} \leq \gamma(b_i + b_{i+1})$ for all i. Again we note that there are only two nonzero entries in each of M_{BI} and $M_{IB} = M_{BI}^T$, so that there are only two nonzero entries in $M_{IB}D_{BB}^{-1}M_{BI}$, given by

$$m_1 := \frac{3h_{N/4+1}^2b_{N/4+1}^2}{36m(h_{N/4}b_{N/4} + h_{N/4+1}b_{N/4+1})},$$

and

$$m_2 := \frac{3h_{3N/4}^2b_{3N/4}^2}{36m(h_{3N/4}b_{3N/4} + h_{3N/4+1}b_{3N/4+1})}.$$

From here, m_1 can be bounded from above by

$$\frac{h_I b_{N/4+1}}{12m} \leq \frac{\gamma}{4m^2}\frac{mh_I(b_{N/4+1}+b_{N/4+2})}{3},$$

with a similar bound coming from bounding m_2 by $h_I b_{3N/4}/(12m)$. Thus,

$$V_I^T M_{IB}D_{BB}^{-1}M_{BI}V_I \leq \frac{\gamma}{4m^2}V_I^T D_{II}V_I,$$

and so,

$$\left\|D_{BB}^{-1/2}M_{BI}V_I\right\|_2 \leq \frac{\sqrt{\gamma}}{2m}\left\|D_{II}^{1/2}V_I\right\|_2.$$

Hence,

$$\left\|D_{BB}^{-1/2}A_{BI}V_I\right\|_2 \leq \left(\frac{3}{m\sqrt{2}}\delta_h + \frac{\sqrt{\gamma}}{2m}\right)\left\|D_{II}^{1/2}V_I\right\|_2.$$

Recalling (22), this gives that

$$\begin{aligned}\left|V_B^T A_{BI} V_I\right| &\leq \sqrt{2m}\left(\frac{3}{m\sqrt{2}}\delta_h + \frac{\sqrt{\gamma}}{2m}\right)\left(V_I^T D_{II} V_I\right)^{1/2}\left(V_B^T A_{BB} V_B\right)^{1/2}, \\ &= \frac{1}{\sqrt{2m}}\left(3\sqrt{2}\delta_h + \sqrt{\gamma}\right)\left(V_I^T D_{II} V_I\right)^{1/2}\left(V_B^T A_{BB} V_B\right)^{1/2},\end{aligned}$$

for all V_B and V_I. Since

$$2ab \leq a^2/q + b^2 q,$$

for any real a, b, and $q > 0$, we have

$$2\left|V_B^T A_{BI} V_I\right| \leq \frac{1}{q} V_I^T D_{II} V_I + \frac{q}{2m}\left(3\sqrt{2}\delta_h + \sqrt{\gamma}\right)^2 V_B^T A_{BB} V_B, \tag{23}$$

for all V_B and V_I. Then, the lower bound for $V^T A V$ is

$$\begin{aligned}V^T A V &\geq V_B^T A_{BB} V_B - 2\left|V_B^T A_{BI} V_I\right| + V_I^T A_{II} V_I \\ &\geq V_B^T A_{BB} V_B - \frac{1}{q} V_I^T D_{II} V_I - \frac{q}{2m}\left(3\sqrt{2}\delta_h + \sqrt{\gamma}\right)^2 V_B^T A_{BB} V_B \\ &\quad + \frac{1}{2m} V_I^T D_{II} V_I \\ &\geq \left(1 - \frac{q}{2m}\left(3\sqrt{2}\delta_h + \sqrt{\gamma}\right)^2\right) V_B^T A_{BB} V_B + \left(\frac{1}{2m} - \frac{1}{q}\right) V_I^T D_{II} V_I \\ &= \left(1 - \frac{q}{2m}\left(18\delta_h^2 + 6\sqrt{2\gamma}\delta_h + \gamma\right)\right) V_B^T A_{BB} V_B \\ &\quad + \left(\frac{1}{2m} - \frac{1}{q}\right) V_I^T D_{II} V_I \\ &\geq \left(\min\left\{1 - \frac{\gamma q}{2m}, \frac{1}{2m} - \frac{1}{q}\right\} - \frac{3q}{m}\sqrt{2\gamma}\delta_h - \frac{9q}{m}\delta_h^2\right) V^T A_D V,\end{aligned} \tag{24}$$

in which q is naturally chosen such that $2m < q < 2m/\gamma$ to guarantee $\theta_q = \min\{1 - \gamma q/2m, 1/(2m) - 1/q\} > 0$. Recognizing that q is a parameter of the proof and not the method, for a fixed choice of m, q can be freely chosen in this range to optimize the resulting coercivity bound, as discussed in Corollary 1.

For the corresponding upper bound, we can choose q in (23) more freely, since we need not worry about positivity as in the lower bound. While choosing $q = 2$ yields a tighter bound (see Remark 1), the upper bound in (20) comes from taking $q = 1$,

yielding

$$\begin{aligned} V^T AV &\leq V_B^T A_{BB} V_B + 2\left|V_B^T A_{BI} V_I\right| + V_I^T A_{II} V_I \\ &\leq V_B^T A_{BB} V_B + V_I^T D_{II} V_I + \frac{1}{2m}\left(3\sqrt{2}\delta_h + \sqrt{\gamma}\right)^2 V_B^T A_{BB} V_B \\ &\quad + \frac{1}{m}\left(\frac{3}{2} + 6\delta_h\right) V_I^T D_{II} V_I \\ &= \left(1 + \frac{\gamma}{2m} + \frac{3}{m}\sqrt{2\gamma}\delta_h + \frac{9}{m}\delta_h^2\right) V_B^T A_{BB} V_B + \left(1 + \frac{3}{2m} + \frac{6}{m}\delta_h\right) V_I^T D_{II} V_I \\ &\leq \left(1 + \frac{3}{2m} + \frac{6}{m}\delta_h + \frac{9}{m}\delta_h^2\right) V^T A_D V. \end{aligned} \tag{25}$$

Combining (24) and (25) completes the proof.

Remark 1 (Tightness of upper bound) In practice, the most benefit is to be gained by choosing m and q primarily to optimize the lower bound in (20). For the upper bound, however, one could also optimize the parameter in (23), yielding

$$V^T AV \leq \left(\max\left\{1 + \frac{q\gamma}{2m}, \frac{1}{q} + \frac{3}{2m}\right\} + \mathscr{O}(\delta_h)\right) V^T A_D V.$$

The optimal upper bound, then, is achieved by choosing q as a function of γ and m such that the two $\mathscr{O}(1)$ terms are in balance, and any fixed choice of q can be seen as a suboptimal bound. In Theorem 2, we choose $q = 1$ to yield a simple bound, where the second term in the max always dominates the first. In practice, this can overestimate the optimal bound by as much as 40%; this does not change the qualitative result of the theorem (that A_D is a good preconditioner for A), but does have a notable quantitative effect in the bound achieved. For $q > 1$, an improved bound is possible when considering restricted ranges of m. For example, taking $q = 2$ for $1/2 < m < 2$ yields a bound that is accurate to within about 10% of the optimal bound. Of course, the optimal choice can also be made, similarly to the optimization of the lower bound discussed below, but such a choice does not seem to yield enough improvement to be worthwhile.

An important practical detail comes from choosing q to maximize the $\mathscr{O}(1)$ term in the lower bound and, subsequently, choosing m to offer the best-possible bound on the condition number of $A_D^{-1}A$. The following results present these optimizations.

Corollary 1 *For any value of m, θ_q is maximized by*

$$q^\star := \frac{2m-1}{2\gamma} + \sqrt{\left[\frac{2m-1}{2\gamma}\right]^2 + \frac{2m}{\gamma}}. \tag{26}$$

This yields the optimized lower spectral equivalence bound,

$$\left(\frac{1}{2m} - \frac{1}{q^\star} - \frac{3q^\star}{m}\sqrt{2\gamma}\delta_h - \frac{9q^\star}{m}\delta_h^2\right) V^T A_D V \leq V^T AV. \tag{27}$$

Proof We maximize θ_q by selecting $q^\star$ such that

$$1-\frac{\gamma q^\star}{2m}=\frac{1}{2m}-\frac{1}{q^\star}.$$

Rewriting this as a quadratic equation in $q^\star$ yields two roots. It is easy to see that the only positive one is that given in (26).

To see that $2m<q^\star<2m/\gamma$, we show that

$$4\gamma m<2m-1+\sqrt{(2m-1)^2+8m\gamma}<4m.$$

The right-hand inequality follows from the fact that $\gamma<1$, yielding

$$2m-1+\sqrt{(2m-1)^2+8m\gamma}<2m-1+\sqrt{(2m+1)^2}=4m.$$

The left-hand inequality follows from $1/2\le\gamma<1$, so that $(4\gamma-2)^2m^2<4m^2$, yielding

$$(4\gamma m-(2m-1))^2<(2m-1)^2+8m\gamma,$$

from which the left-hand inequality follows by taking square roots of both sides and rearranging.

Corollary 2 *Let $\eta(m)$ be the dominant term in the condition number bound of $A_D^{-1}A$ as a function of m, defined by the ratio between the $\mathscr{O}(1)$ term in the upper bound* (25)*, and the $\mathscr{O}(1)$ term of the optimized lower bound* (27)*, i.e.,*

$$\eta(m):=\frac{1+3/(2m)}{1/(2m)-1/q^\star}=\frac{2m+3}{1-2m/q^\star},$$

where $q^\star$ is defined in (26)*. Then, the optimized minimum value of $\eta(m)$ is attained when*

$$m=m^\star:=\frac{3\gamma-3-2\sqrt{3\gamma}}{2(\gamma-3)}. \tag{28}$$

Proof Using elementary calculus, it is easily verified that $\eta'(m^\star)=0$, and furthermore that $\eta''(m^\star)>0$.

It is interesting to note that the optimal value for m depends only mildly on γ, and is found in the interval $[(\sqrt{6}+3/2)/5,\sqrt{3}/2]\approx[0.79,0.87]$. Thus, while it is important to optimize θ_q in the lower bound of Theorem 2 to achieve a good picture of the true performance of the preconditioner, the choice of a fixed value of m, independent of γ, is reasonable from a practical point of view. Indeed, fixing m to be the mid-point in this range yields a leading-order spectral equivalence bound that is within 0.1% of that with the optimal choice of m for all γ.

An important limit to consider is the case of $\gamma\to1$, where $\beta_0/\beta_1\to0$. In this case, it is easy to see that $q^\star\to2m$ and, consequently, the spectral equivalence bound $\eta(m)\to\infty$. Indeed, while $\eta(m)$ increases slowly from a value of approximately 12.9 at $\gamma=1/2$ to 28 at $\gamma=3/4$, fast growth is seen beyond that point, with $\eta(m)\approx100$ for $\gamma=0.93$ and $\eta(m)\approx250$ for $\gamma=0.97$. Considering the proof of Theorem 2, we

see that γ appears only in bounding the contributions from the two non-zero entries in the matrix $M_{IB}D_{BB}^{-1}M_{BI}$. Consequently, if the values of the reaction coefficient can be bounded more closely in the neighbourhood of the mesh transition points, a tighter bound is possible here, and it is possible to avoid the divergence of $\eta(m)$. Whether this is possible clearly depends on the precise form of the reaction coefficient, b. We note that, while the bound on $\eta(m)$ depends badly on γ, it is independent of both ε and N; thus, the preconditioner always satisfies the parameter robustness that is sought for these problems.

Example 2 To demonstrate the benefit of optimizing q in the lower bound, recall Example 1, where $b \equiv 1$ and, therefore, $\gamma = 1/2$. We first compare two spectral equivalence constants when taking $q = 1$ in (24),

$$\mu(m) = \frac{1+3/(2m)}{\min\{1-1/(4m), 1/(2m)-1\}},$$

and when $q = q^\star$ in (27) corresponding to $\eta(m)$. On the left of Figure 2, we plot the functions $\mu(m)$ as dashed red line for $1/4 < m < 1/2$ (since the restriction for m with respect to q is $\gamma q/2 < m < q/2$) and $\eta(m)$ in solid blue for $0 < m \leq 1$. Note that two functions $\mu(m)$ and $\eta(m)$ obtain the same value only when $q = q^\star = 1$. This yields $m = 3/8$ (the corresponding range for valid q is $3/4 < q < 3/2$), and $\mu(3/8) = \eta(3/8) = 15$. Away from this point $m = 3/8$, the condition number bound $\mu(m)$ is significantly worse than that of $\eta(m)$ over the interval $1/4 < m < 1/2$. Furthermore, as for $\eta(m)$, from (28) the best choice of m when $\gamma = 1/2$ is

$$m^\star = \frac{3}{10} + \frac{\sqrt{6}}{5},$$

which attains the minimum of $\eta(m)$ for $m > 0$, marked as diamond in the left plot of Figure 2, and its value is

$$\eta\left(\frac{3}{10}+\frac{\sqrt{6}}{5}\right) = \frac{2}{25}\frac{(-2+2\sqrt{6}+5A)(9+\sqrt{6})}{-1+A} \approx 12.89,$$

$$A = \sqrt{\left(-\frac{2}{5}+\frac{2}{5}\sqrt{6}\right)^2 + \frac{6}{5}+\frac{4}{5}\sqrt{6}}.$$

Note that while the optimal condition number bound for the case of $q = 1$ is not significantly worse compared to that using $q^\star$, the prediction made by considering $\mu(m^\star)$ is completely invalid, since $m^\star$ does not fall in the range of $1/4 < m < 1/2$, so the lower bound with $q = 1$ is not positive. This highlights the importance of including q in the bound in Theorem 2, allowing us to consider a much broader range of preconditioner parameters, m, than would be otherwise possible with fixed values of q.

Secondly, we want to show the benefit of having $q^\star$ from another point of view, i.e., fixing $m = m^\star$, but letting q vary over the range $2m^\star < q < 4m^\star$ (since $\gamma = 1/2$ in this example). Let

$$\nu(q) = \frac{1+3/(2m^\star)}{\min\{1-q/(4m^\star), 1/(2m^\star)-1/q\}}, \quad 2\left(\frac{3}{10}+\frac{\sqrt{6}}{5}\right) < q < 4\left(\frac{3}{10}+\frac{\sqrt{6}}{5}\right).$$

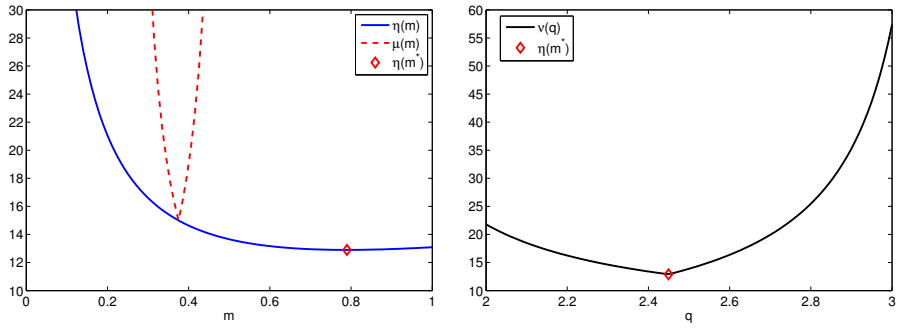

Fig. 2 Left: the leading-order condition-number bound $\eta(m)$ (with optimal proof parameter $q^\star$) versus that for fixed $q=1$, denoted by $\mu(m)$, when $b\equiv$ const. Right: the leading-order condition-number bound as a function of the proof parameter, q, with $m=m^\star$ (the value that optimizes the optimal leading-order condition-number bound $\eta(m)$). Note that $q^\star\approx 2.45$ is the value given by (26) for the optimal value of $m^\star$.

On the right of Figure 2, we show the function $\nu(q)$ for $2\le q\le 3$. It is clear that away from $q=q^\star$, the value of $\nu(q)$ is dramatically increased. For example, $\nu(2)$ is about 69% higher than $\eta(m^\star)$, while $\nu(3)$ is about 345% higher.

The theoretical results above establish that A is spectrally equivalent to A_D and, so, it follows that A_D is an excellent preconditioner for A, in the sense that it provides a good approximation to A. Using it as such requires that one solves linear systems involving A_D; in 1D this can be done optimally with a direct solver (of course, the same is true of A in 1D). However, efficient extensions of this approach to layer-adapted meshes in two (or more) dimensions rely on further approximations so that linear systems with A_{BB} can be (approximately) solved inexpensively. Noting that A_{BB} resembles the discretization matrix coming from a classical diffusion-dominated problem, we now consider approximating A_{BB} within A_D by a multigrid method (or any other suitable preconditioner). We make this precise in the following corollary.

Corollary 3 *Under the assumptions of Theorem 2, if $\hat{A}_{BB}$ is spectrally equivalent to A_{BB}, meaning that there exists constants $\tilde{C}_0$ and $\tilde{C}_1$ such that*

$$\tilde{C}_0 V_B^T \hat{A}_{BB} V_B \le V_B^T A_{BB} V_B \le \tilde{C}_1 V_B^T \hat{A}_{BB} V_B, \quad \text{for all } V_B,$$

then the matrix

$$\hat{A}=\begin{pmatrix} \hat{A}_{BB} & 0 \\ 0 & D_{II}\end{pmatrix}.$$

satisfies

$$C_0 V^T \hat{A} V \le V^T A V \le C_1 V^T \hat{A} V,$$

for all V where

$$C_0=\min\left\{\frac{1}{2m}-\frac{1}{q},\tilde{C}_0\left(1-\frac{q}{2m}\left(\gamma+18\delta_h^2+6\sqrt{2\gamma}\delta_h\right)\right)\right\}.$$

and

$$C_1=\max\left\{1+\frac{3}{2m}+\frac{6}{m}\delta_h,\tilde{C}_1\left(1+\frac{\gamma}{2m}+\frac{3}{m}\sqrt{2\gamma}\delta_h+\frac{9}{m}\delta_h^2\right)\right\},$$

Clearly the optimization of q considered above could be repeated here if the values of $\tilde{C}_0$ and $\tilde{C}_1$ were well-known from analysis of the preconditioner. In what follows, we do not consider the optimization again.

3.2 Numerical results

We begin by studying the application of unpreconditioned CG to solve the linear system (7) which comes from the finite-element discretization of Example 1 applied on a Shishkin mesh. Since the stopping criteria discussed in Section 2.4 only make sense if a good preconditioner is used, we use the criterion given in (14), with $C = 1$. Table 3 shows the iteration counts for the unpreconditioned CG applied to Example 1. It can be seen that, for a fixed N, the iteration counts depend badly on ε. For example, when $N = 2^{10}$, the number of iterations required for $\varepsilon^2 = 10^{-8}$ is 173, but it increases up to 665 for $\varepsilon^2 = 10^{-12}$. This agrees with Theorem 1, where the condition number is shown to be proportional to $(\varepsilon \ln N)^{-1}$ on a Shishkin mesh.

Table 3 Iteration counts for unpreconditioned CG applied to the problem in Example 1 discretized using piecewise linear FEM on a Shishkin Mesh.

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{12}$
10^{-6}	24	43	82	–	–	–
10^{-8}	71	85	88	173	339	661
10^{-10}	94	169	256	291	348	691
10^{-12}	124	239	423	665	908	991

We now consider a preconditioner following Corollary 3, where the exact solution with A_{BB} required to invert A_D is replaced by applying a single multigrid V-cycle [4, 19, 21] to the boundary sub-matrix. In what follows, we refer to the preconditioner as *MG-BLPCG*, to denote a boundary-layer preconditioned CG algorithm using multigrid as a component. Recall that when $\delta_h > 0.1$, the boundary layers are not developed, and the problem is effectively diffusion-dominated. In this case, multigrid methods are well-known to be effective by themselves, and there is no need for the preconditioning approach considered here. When $\delta_h \leq 0.1$, we employ the MG-BLPCG algorithm detailed as follows. For the interior component where our preconditioner is diagonal, we simply apply a diagonal scaling operation. For the boundary component, we use a standard multigrid implementation based on the description of [4, Chapter 3]. On each level, a single Gauss-Seidel sweep is used as both pre- and post-relaxation. Linear interpolation and its transpose define the grid-transfer operators; in the setting where $b(x) \equiv$ const, Galerkin coarsening is equivalent to rediscretization of the coarse-grid operators. The coarsest grid in the resulting multigrid hierarchy is fixed to have 8 points, where a direct solver is used.

Table 4 shows the iteration counts for the algorithm using the stopping criterion based on the preconditioned residual given in (15) with $C = 1/2$. Computed errors are shown in Table 5, confirming that the expected error behaviour in the energy

norm is, indeed, achieved. The iteration counts for the algorithm are steady and show only very small dependency on both N and ε. Notably, they are far less than those of Table 3 and exactly match those of the "ideal" preconditioner using exact inversion of A_{BB} (not shown here).

Table 4 Iteration counts for MG-BLPCG, using the energy norm stopping criterion, applied to the problem in Example 1 discretized using piecewise linear FEM on a Shishkin Mesh.

ε^2	$N=2^7$	$N=2^8$	$N=2^9$	$N=2^{10}$	$N=2^{11}$	$N=2^{12}$
10^{-6}	5	5	5	–	–	–
10^{-8}	6	6	7	7	7	6
10^{-10}	7	7	7	8	8	8
10^{-12}	8	8	8	8	9	9

Table 5 $\|u-u^{(k)}\|_\varepsilon$ for the solution computed using MG-BLPCG and the energy-norm stopping criterion, applied to the problem in Example 1 discretized using piecewise linear FEM on a Shishkin Mesh.

ε^2	$N=2^7$	$N=2^8$	$N=2^9$	$N=2^{10}$	$N=2^{11}$	$N=2^{12}$
10^{-6}	5.680e-03	3.250e-03	1.824e-03	–	–	–
10^{-8}	1.795e-03	1.028e-03	5.765e-04	3.204e-04	1.762e-04	9.629e-05
10^{-10}	5.673e-04	3.245e-04	1.828e-04	1.013e-04	5.573e-05	3.042e-05
10^{-12}	1.800e-04	1.026e-04	5.773e-05	3.211e-05	1.762e-05	9.615e-06

4 Generalization to two-dimensional problems

We turn to a finite-element discretization of the two-dimensional reaction-diffusion problem (2) on a tensor product mesh with bilinear elements. After introducing the resulting system, in Section 4.1, we present a model problem whose boundary layers motivate a two-dimensional analogue of the mesh from Section 2.2. Next, in Section 4.2 we consider application of a direct solver and observe that, although the theoretical error estimate is validated, solve times scale poorly with the perturbation parameter. Thus iterative methods must employed. However, in Section 4.3 we show that the resulting linear system is ill-conditioned. Therefore, we extend our boundary layer preconditioner of Section 3 to the two-dimensional setting. In Section 4.5 we discuss implementations issues, including derivation of sharp stopping criteria, and present numerical results demonstrating the efficiency of the method.

4.1 The two-dimensional finite-element method

Let $\omega_x^N = \{0 = x_0 < x_1 < \ldots < x_N = 1\}$ and $\omega_y^N = \{0 = y_0 < y_1 < \ldots < y_N\}$ be arbitrary meshes, each with N intervals on $[0,1]$. Set $\Omega^{N,N} = \{(x_i, y_j)\}_{i,j=0}^N$ to be

the Cartesian product of ω_x^N and ω_y^N, and define $h_i = x_i - x_{i-1}$ and $k_j = y_j - y_{j-1}$. For any fixed ε, the discretization of $-\varepsilon^2 \Delta u$ is straightforward; in order to avoid issues of quadrature in evaluating the weighted finite-element mass matrix entries, we assume that b is approximated as a piecewise constant on each element, writing $b_{i,j} = b(x_{i-1/2}, y_{j-1/2})$, for $x_{i-1/2} = (x_i + x_{i-1})/2$ and $y_{j-1/2} = (y_j + y_{j-1})/2$. The matrix decomposes into three terms,

$$A = S^{(x)} + S^{(y)} + M, \tag{29}$$

whose stencils are

$$S^{(x)} = \varepsilon^2 \begin{bmatrix} -\dfrac{k_{j+1}}{6h_i} & \dfrac{k_{j+1}}{6h_i} + \dfrac{k_{j+1}}{6h_{i+1}} & -\dfrac{k_{j+1}}{6h_{i+1}} \\ -\dfrac{k_{j+1}+k_j}{3h_i} & \dfrac{k_{j+1}+k_j}{3h_i} + \dfrac{k_{j+1}+k_j}{3h_{i+1}} & -\dfrac{k_{j+1}+k_j}{3h_{i+1}} \\ -\dfrac{k_j}{6h_i} & \dfrac{k_j}{6h_i} + \dfrac{k_j}{6h_{i+1}} & -\dfrac{k_j}{6h_{i+1}} \end{bmatrix},$$

$$S^{(y)} = \varepsilon^2 \begin{bmatrix} -\dfrac{h_i}{6k_{j+1}} & -\dfrac{h_{i+1}+h_i}{3k_{j+1}} & -\dfrac{h_{i+1}}{6k_{j+1}} \\ \dfrac{h_i}{6k_{j+1}} + \dfrac{h_i}{6k_j} & \dfrac{h_{i+1}+h_i}{3k_{j+1}} + \dfrac{h_{i+1}+h_i}{3k_j} & \dfrac{h_{i+1}}{6k_{j+1}} + \dfrac{h_{i+1}}{6k_j} \\ -\dfrac{h_i}{6k_j} & -\dfrac{h_{i+1}+h_i}{3k_j} & -\dfrac{h_{i+1}}{6k_j} \end{bmatrix},$$

and

$$M = \begin{bmatrix} \dfrac{m_{i,j+1}}{36} & \dfrac{m_{i,j+1}+m_{i+1,j+1}}{18} & \dfrac{m_{i+1,j+1}}{36} \\ \dfrac{m_{i,j+1}+m_{i,j}}{18} & \dfrac{m_{i,j+1}+m_{i,j}+m_{i+1,j}+m_{i+1,j+1}}{9} & \dfrac{m_{i+1,j}+m_{i+1,j+1}}{18} \\ \dfrac{m_{i,j}}{36} & \dfrac{m_{i,j}+m_{i+1,j}}{18} & \dfrac{m_{i+1,j}}{36} \end{bmatrix}, \tag{30}$$

where we set $m_{i,j} := h_i k_j b_{i,j}$.

To motivate the construction of a suitable piecewise uniform mesh, we consider the following example (see also [14, §1.4]).

Example 3 Consider (2) with $b \equiv 1$ and f chosen so that

$$u(x,y) = x^3(1+y^2) + \sin(\pi x^2) + \cos(\pi y/2) + (1+x+y)\left(e^{-2x/\varepsilon} + e^{-2y/\varepsilon}\right). \tag{31}$$

All of the results in this section are for a finite-element method, with bilinear elements, applied on a tensor product Shishkin mesh, $\omega^{N\times N}$, constructed as follows. Since the solution, u, has boundary layers only along the edges $x = 0$ and $y = 0$, and a corner layer at $(0,0)$, a suitable Shishkin mesh for this problem is constructed by taking the transition point as

$$\tau = \left\{ \frac{1}{2}, 2\frac{\varepsilon}{\beta_0} \ln N \right\},$$

forming a one-dimensional mesh, ω^N, by subdividing both $[0,\tau]$ and $[\tau,1]$ into $N/2$ equally sized mesh intervals. (Note that this is slightly simpler than the mesh described in Section 2.2, in that it condenses only near one boundary). Then the nodes in the mesh, $\omega^{N\times N}$, are $(x_i,y_j)=(\omega_i^N,\omega_j^N)$ for $i,j=0,1,\ldots,N$ (see [14, Fig. 2.1(a)]).

The error analysis of reaction-diffusion problems in two dimensions on this mesh can be found in, e.g., [13] and also [18, pages 404–406], where it is shown that the bound given in (10) for the one-dimensional problem also holds here. That is, there exists a constant independent of both ε and N such that

$$\|u-u^N\|_\varepsilon \le C(\varepsilon^{1/2}N^{-1}\ln N+N^{-2}). \tag{32}$$

4.2 Direct solvers

In order to make computations comparable between different discretizations, we use the same setting as that of [14,16], i.e., the program was coded in C and executed using a single core of a node with an AMD Opteron 2427, 2200 MHz processor and 32Gb of RAM. We use CHOLMOD Version 1.7.1 to solve the sparse symmetric positive-definite linear systems; see [5,6].

First we use the direct solver to verify that the estimate in (32) is sharp. To this end, errors in the energy norm, computed element-wise using tensor-product 3-point Gaussian quadrature, are shown in Table 6; they are, indeed, consistent with the estimate in (32). So clearly, and as expected, the direct solver provides adequate resolution of the system (29). However, this approach is highly, and surprisingly, inefficient, even for relatively small values of N. In Table 7, we show the time in seconds, averaged over three runs, required to solve the linear systems that correspond to the results in Table 6. For a fixed N, say $N=2^{11}$, it is easily observed that the amount of time required to solve the linear system depends quite badly on the perturbation parameter.

Table 6 $\|u-u^N\|_\varepsilon$ for Example 3. The discrete solution is found using the direct solver, CHOLMOD.

ε^2	$N=2^7$	$N=2^8$	$N=2^9$	$N=2^{10}$	$N=2^{11}$
1	2.372e-02	1.186e-02	5.931e-03	2.966e-03	1.483e-03
10^{-2}	2.964e-02	1.483e-02	7.417e-03	3.708e-03	1.854e-03
10^{-4}	2.670e-02	1.533e-02	8.636e-03	4.800e-03	2.641e-03
10^{-6}	8.478e-03	4.868e-03	2.743e-03	1.524e-03	8.386e-04
10^{-8}	2.684e-03	1.540e-03	8.677e-04	4.823e-04	2.653e-04
10^{-10}	8.535e-04	4.876e-04	2.744e-04	1.525e-04	8.390e-05
10^{-12}	2.847e-04	1.558e-04	8.697e-05	4.825e-05	2.653e-05

If the results of Table 7 are compared with the corresponding results for the finite-difference method, as given in [14, Table 4.1] and also [16, Table 3], two issues become apparent:

1. When ε is $\mathcal{O}(1)$, the solve times for the finite-element discretization are about twice those for the finite-difference discretization. This is because the finite-element discretization has a 9-point stencil rather than 5-point stencil of the finite

Table 7 Cholesky (CHOLMOD) solve times for linear systems generated using piecewise bilinear FEM on a tensor-product Shishkin mesh.

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$
1	0.0978	0.718	5.58	35.22	353.02
10^{-2}	0.0972	0.718	5.58	35.16	352.95
10^{-4}	0.0971	0.717	5.58	35.17	353.00
10^{-6}	0.0971	0.718	6.61	101.32	1243.37
10^{-8}	0.0971	0.718	6.25	98.78	1311.67
10^{-10}	0.0971	0.718	6.46	101.24	1321.89
10^{-12}	0.0971	0.718	6.78	104.82	1337.34

difference discretization, and so the system matrix has roughly twice the number of nonzero entries.

2. For the finite-difference case, the solve times, as shown in [14, Table 4.1] and [16, Table 3], initially increase, and then decrease when ε becomes smaller. In contrast, the solve times for the finite-element case increase initially, but then stabilize. Although we do not offer an analysis of Cholesky factorization in this case, the framework of [16] could be used to investigate this.

In Table 8, we give the number of nonzero entries in the Cholesky factors produced by CHOLMOD for a range of values of N and ε, as well as the number of subnormal entries. This agrees completely with the results of Table 7. For small ε and large N, we observe a significant increase in the number of subnormal numbers arising in the Cholesky factors, as well as a decrease in the number of nonzero entries in the factors, due to underflow-zeros.

Table 8 Number of nonzero entries (left) and subnormal numbers (right) in Cholesky factors generated by CHOLMOD for Example 3. No subnormal numbers appear in the factors for data points not included in the bottom table.

	Number of nonzero entries					Number of subnormals	
ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{10}$	$N = 2^{11}$
1	573163	3239141	17011189	63549693	304900961	–	–
10^{-2}	573163	3239141	17011189	63549693	304900961	–	–
10^{-4}	573163	3239141	17011189	63549693	304900961	–	–
10^{-6}	573163	3239141	17011189	63166392	300992678	74982	441440
10^{-8}	573163	3239141	17011189	63276869	293538627	69508	955686
10^{-10}	573162	3239134	17011179	63263046	293598199	71831	957773
10^{-12}	573160	3239136	17011171	63234561	293268356	75100	934242

Analogous results are observed with other implementations of Cholesky factorization, and LU-based direct solvers. Consequently, it is clear that direct solvers have limitations for this problem, and an iterative solver is required.

4.3 Condition number estimate

Having established the need for an iterative solver, we now show that use of a suitable preconditioner is essential because, as with the one-dimensional reaction-diffusion problem, the linear system of the two-dimensional problem (2) is ill-conditioned when the problem is discretized by the finite-element method on boundary layer-adapted meshes.

Theorem 3 *Let A be the matrix defined in* (29)*, and assume that ω_x^N and ω_y^N satisfy the conditions* (4) *and that there exists $C_h > 0$ such that $h_{\min} \geq C_h \varepsilon / N$, where $h_{\min}$ is the minimum mesh width over both ω_x^N and ω_y^N. Then, there is a constant C, independent of both N and ε, such that*

$$\kappa_2(A) \leq C(N h_{\min})^{-2}.$$

Proof By Geršgorin's Theorem, we have

$$\begin{aligned} \|A\|_2 &\leq \max_{i,j} \left\{ \varepsilon^2 \left(\frac{k_j}{h_i} + \frac{k_{j+1}}{h_i} + \frac{k_j}{h_{i+1}} + \frac{k_{j+1}}{h_{i+1}} + \frac{h_i}{k_j} + \frac{h_i}{k_{j+1}} + \frac{h_{i+1}}{k_j} + \frac{h_{i+1}}{k_{j+1}} \right) \right. \\ &\qquad \left. + \frac{1}{9} \left(m_{i,j} + m_{i,j+1} + m_{i+1,j} + m_{i+1,j+1} \right) \right\} \\ &\leq 8\varepsilon^2 \frac{h_I}{h_{\min}} + \frac{4}{9} \beta_1^2 h_I^2 \leq C N^{-2}, \end{aligned} \tag{33}$$

where we use (4) and $h_{\min} \geq C\varepsilon/N$ for the last inequality.

To bound the smallest eigenvalue of A defined in (29) from below, we estimate the lower bound for the eigenvalue of the mass matrix M, since $(S^{(x)} + S^{(y)})$ is symmetric positive definite by construction. To this end, we use the approach of Wathen [20]. If λ is an any eigenvalue of M, then

$$\min_{V \neq 0} \frac{V^T M V}{V^T V} \leq \lambda. \tag{34}$$

Noting that the mass matrix for the element (i, j) is

$$E_{i,j} = m_{i,j} E, \quad \text{where} \quad E = \begin{bmatrix} 1/9 & 1/18 & 1/18 & 1/36 \\ 1/18 & 1/9 & 1/36 & 1/18 \\ 1/18 & 1/36 & 1/9 & 1/18 \\ 1/36 & 1/18 & 1/18 & 1/9 \end{bmatrix}. \tag{35}$$

Let $\mathrm{diag}_{i,j} E_{i,j}$ be the block-diagonal matrix of element matrices $E_{i,j}$, and let L be the Boolean assembly matrix, then we have

$$M = L^T (\mathrm{diag}_{i,j} E_{i,j}) L.$$

We also set $D_{i,j}$ to be the identity matrix associated with the degrees of freedom on element (i, j), and take the diagonal matrix, $D = L^T (\mathrm{diag}_{i,j} D_{i,j}) L$, to be the "finite-element assembly" of these submatrices. Consequently, the diagonal entries of D simply count how many elements each degree of freedom contributes to in the mesh.

Since $\Omega^{N,N}$ is a regular rectangular grid, each node (i,j) appears in at most four elements. Therefore, the values of each diagonal entry of D is also at most 4, and

$$I \le D \le 4I,$$

where I denotes the full-grid identity matrix. This yields that the inequality (34) is equivalent to

$$4 \min_{V \neq 0} \frac{V^T L^T (\mathrm{diag}_{i,j} E_{i,j}) LV}{V^T DV} \le \lambda,$$

or,

$$4 \min_{V \neq 0} \frac{V^T L^T (\mathrm{diag}_{i,j} E_{i,j}) LV}{V^T L^T (\mathrm{diag}_{i,j} D_{i,j}) LV} = 4 \min_{i,j} \min_{W \neq 0} \frac{W^T (\mathrm{diag}_{i,j} E_{i,j}) W}{W^T (\mathrm{diag}_{i,j} D_{i,j}) W} \le \lambda,$$

which gives

$$4 \min_{i,j} \lambda_{i,j}^{\min} \le \lambda,$$

where $\lambda_{i,j}^{\min}$ is the minimum eigenvalue of $E_{i,j}$ since $D_{i,j}$ is a 4-by-4 identity matrix. From direct computation for the eigenvalues of the matrix E in (35) and using the fact that $m_{i,j} = h_i k_j b_{i,j} \ge \beta_0^2 h_{\min}^2$, we get

$$\beta_0^2 h_{\min}^2 \frac{1}{36} \le \min_{i,j} \lambda_{i,j}^{\min}.$$

It follows then that

$$\|A^{-1}\|_2 \le 9/\left(\beta_0^2 h_{\min}^2\right). \tag{36}$$

Combining (33) and (36), we obtain

$$\kappa_2(A) \le C(Nh_{\min})^{-2}.$$

For the Shishkin mesh described in Section 4.1, where $h_{\min}$ behaves like $\varepsilon \ln N/(N\beta_0)$, this implies that $\kappa_2(A) \le C\varepsilon^{-2} \ln^{-2} N$. Table 9 gives computed values of $\kappa_2(A)$, and shows that this estimate is sharp with the constant $C \approx 0.5$. In particular, for a fixed N, $\kappa_2(A)$ is proportional to ε^{-2}.

Table 9 $\kappa_2(A)$ for the problem (2) discretized by piecewise bilinear FEM on a tensor-product Shishkin mesh.

ε^2	$N=16$	$N=32$	$N=64$	$N=128$	$N=256$	$N=512$
1	4.92e+01	1.97e+02	7.90e+02	3.16e+03	1.26e+04	5.06e+04
10^{-2}	8.83e+00	3.45e+01	1.37e+02	5.48e+02	2.19e+03	8.76e+03
10^{-4}	1.79e+02	2.08e+02	5.15e+02	1.34e+03	3.66e+03	1.03e+04
10^{-6}	1.94e+04	1.44e+04	1.09e+04	1.44e+04	4.02e+04	1.16e+05
10^{-8}	1.95e+06	1.45e+06	1.11e+06	8.62e+05	6.88e+05	1.17e+06
10^{-10}	1.95e+08	1.45e+08	1.11e+08	8.63e+07	6.88e+07	5.60e+07
10^{-12}	1.95e+10	1.45e+10	1.11e+10	8.63e+09	6.89e+09	5.60e+09

4.4 Boundary layer preconditioner

As in the finite-difference case [14, §4.5], we partition A into a corner region, where the mesh is highly resolved in both directions, the edge regions, where the mesh is highly resolved in one direction but not both, and the interior region. Further, we assume that the mesh spacing (in both directions) in the non-resolved portions of the grid is uniform, with spacing h_I. Thus, we write

$$A = \begin{pmatrix} A_{CC} & A_{CE} & A_{CI} \\ A_{EC} & A_{EE} & A_{EI} \\ A_{IC} & A_{IE} & A_{II} \end{pmatrix},$$

where the subscripts C, E and I indicate the block structure of corners, edge layers, and interior points, respectively. The preconditioner, A_D, will be defined using the same partitioning:

$$A_D = \begin{pmatrix} A_{CC} & 0 & 0 \\ 0 & T_{EE} & 0 \\ 0 & 0 & D_{II} \end{pmatrix}. \tag{37}$$

Here D_{II} is the diagonal matrix with entries based on the scaled diagonal of the mass matrix, i.e., $D_{II} = m\text{diag}(M_{II})$. The tridiagonal matrix T_{EE} is constructed so that it provides a good approximation to A_{EE}. The choice of T_{EE} stems from the following observation. Along the edges, rectangles are long and thin with one side of length $\mathscr{O}(\varepsilon N^{-1})$, and other side of length $\mathscr{O}(N^{-1})$. Therefore, depending on the orientation of the rectangles, some entries in A_{EE} are very small compared to others. A preconditioner can be formed by either neglecting these terms, or aggregating them.

For our implementation, here we consider the approach that T_{EE} is constructed by summing the coefficients along the direction of the large mesh-width. Considering the block associated with the edge along x-axis, we'll decompose $A_{EE} = S^{(x)}_{EE} + S^{(y)}_{EE} + M_{EE}$ as the (column-wise) tridiagonal terms in $S^{(y)}_{EE}$ and M_{EE} to give

$$T_{EE} = \begin{bmatrix} 0 & -\varepsilon^2 \dfrac{h_I}{k_{j+1}} + \dfrac{h_I k_{j+1}(b_{i,j+1} + b_{i+1,j+1})}{12} & 0 \\ 0 & \varepsilon^2 \left(\dfrac{h_I}{k_{j+1}} + \dfrac{h_I}{k_j} \right) + \dfrac{h_I k_{j+1}(b_{i,j+1} + b_{i+1,j+1}) + h_I k_j (b_{i,j} + b_{i+1,j})}{6} & 0 \\ 0 & -\varepsilon^2 \dfrac{h_I}{k_j} + \dfrac{h_I k_j (b_{i,j} + b_{i+1,j})}{12} & 0 \end{bmatrix}.$$

4.5 Implementation of the preconditioner

Here, we describe the implementation of the boundary layer preconditioner proposed in Section 4.4, and derive a suitable stopping criterion. Numerical results of Section 4.5.3 show that, in practice, the approach is robust and very promising when the perturbation parameter, ε, is small. Analysis of this preconditioner remains as future work.

4.5.1 Boundary layer preconditioned CG

Based on the structure of the boundary layer preconditioner defined in (37), we apply different strategies to efficiently solve the subsystems needed to compute $Z^{(k)} = A_D^{-1} R^{(k)}$.

- For the corner region, we use the black box multigrid method (BoxMG) of Dendy [1, 7], as in [14]. BoxMG is well-known as an optimized structured-grid multigrid algorithm for problems with heterogeneous meshes or coefficients, as occurs in the corner regions. As an approximation to A_{CC}^{-1}, we apply a single BoxMG V-cycle using an alternating-line relaxation scheme that is the default in BoxMG. For this scheme, the grid lines in both the x- and y-directions are split alternately into "red" and "black" lines. The pre-relaxation sweep relaxes first all red lines in the $x-$direction, then all black lines in the $x-$direction, then all red lines in the $y-$direction, then all black lines in the $y-$direction. The opposite ordering is used in the post-relaxation sweep to preserve symmetry of the preconditioner. The multigrid scheme coarsens each corner region down to a 3×3 coarse grid where a direct solver is used.
- For the edge region, T_{EE} is a tridiagonal matrix. Therefore, we use the tridiagonal solver *algorithm* `DPTTRS` [11] that is part of LAPACK library of subroutines for solving problems in numerical linear algebra [2].
- A simple diagonal scaling is used for the interior region since the corresponding preconditioner in this region is a diagonal matrix.

In addition, to enhance the computational efficiency, we introduce three user-chosen parameters, c_1, c_2, and c_3 to appropriately scale the corner, edge, and interior components, respectively, of the preconditioned residual $Z^{(k)}$ in the MG-BLPCG algorithm. To avoid redundancy in the parameters, we fix $m = 1$ and set $D_{II} = \mathrm{diag}(M_{II})$ and rely on c_3 to scale this component of the preconditioner.

4.5.2 Stopping criteria

Arguments similar to those in Section 2.4 can be used to derive the stopping criterion associated with the energy norm for the two-dimensional problems. In particular, for a Shishkin mesh, the stopping criterion based on the preconditioned residual is

$$\sqrt{\left(Z^{(k)}\right)^T R^{(k)}} \leq C(\varepsilon^{1/2} N^{-1} \ln N + N^{-2}). \tag{38}$$

In contrast, the necessary stopping criterion for the unpreconditioned residual to achieve accuracy in the energy norm for a Shishkin mesh is

$$\|R^{(k)}\|_2 \leq C\left(\varepsilon^{3/2} N^{-2} \ln^2 N + \varepsilon N^{-3} \ln^3 N\right). \tag{39}$$

In practice, one interprets (39) as $\|R^{(k)}\|_2 \leq C_1 \varepsilon^{3/2} N^{-2} \ln^2 N + C_0 \varepsilon N^{-3} \ln^3 N$, because, since C_0 and C_1 come from the zero-order and first-order terms in the energy norm defined in (3). Though both are independent of ε and N, they may be of different orders of magnitude. In our experiments, whose results are reported in Tables 10 and 11, we have taken $C_1 = 0.4$, and $C_0 = 0.01 C_1$.

For small values of ε and large values of N, it is clear that we may not be able to compute $\|R^{(k)}\|_2$ accurately enough in double precision to reliably achieve the stopping tolerance on the unpreconditioned residual. For example, the bound (assuming $C = \mathcal{O}(1)$) for $N = 512$ and $\varepsilon = 10^{-6}$ is $\mathcal{O}(10^{-12})$, but the condition number reported in Table 9 is $\mathcal{O}(10^9)$. For $N = 4096$ and $\varepsilon = 10^{-8}$, the bound on $\|R^{(k)}\|$ falls below 10^{-16}; in contrast, the bound on the preconditioned residual for these extreme values is only $\mathcal{O}(10^{-7})$.

4.5.3 Numerical results

As in the analysis of Section 3, our boundary layer preconditioner is designed for singularly perturbed problems. Thus, we only report results for cases where $\delta_h \leq 0.1$. In Table 10, we give the CPU solve times of MG-BLPCG, together with the iteration counts. The preconditioner parameters are determined experimentally as $c_1 = 1$, $c_2 = 1$, and $c_3 = 0.65$, and the stopping criterion (38) is used. We emphasize that the iteration counts are optimal with respect to N. Moreover, they are essentially robust with respect to ε: the weak dependency on ε is explained by $\varepsilon^{1/2}$ term in the stopping criterion (38). Table 11 verifies that the computed solutions obtain the same level of accuracy as those with the direct solver reported in Table 6.

Table 10 CPU times (and iteration counts) for solution using MG-BLPCG, averaged over 3 runs.

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{12}$
10^{-6}	0.014 (6)	0.073 (6)	0.369 (7)	–	–	–
10^{-8}	0.014 (7)	0.080 (7)	0.390 (7)	2.35 (8)	13.16 (10)	82.35 (14)
10^{-10}	0.016 (8)	0.095 (8)	0.456 (8)	2.38 (8)	12.03 (9)	61.13 (10)
10^{-12}	0.019 (10)	0.112 (10)	0.547 (10)	2.87 (10)	13.16 (10)	61.05 (10)

Table 11 $\|u - u^{(k)}\|_\varepsilon$ for solution using MG-BLPCG

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{12}$
10^{-6}	8.479e-03	4.868e-03	2.743e-03	–	–	–
10^{-8}	2.684e-03	1.541e-03	8.679e-04	4.824e-04	2.655e-04	1.449e-04
10^{-10}	8.541e-04	4.879e-04	2.746e-04	1.526e-04	8.391e-05	4.578e-05
10^{-12}	2.848e-04	1.559e-04	8.701e-05	4.827e-05	2.654e-05	1.448e-05

Table 10 clearly shows that, for a small ε and a large N, the MG-BLPCG algorithm is far more efficient than the direct solver used to compute the results in Table 7. For example, when $\varepsilon^2 \leq 10^{-8}$, the MG-BLPCG is about 35 times faster than the direct solver for $N = 2^{10}$, and just over 100 times faster for $N = 2^{11}$, the largest value of N for which we can apply the direct solver on the hardware we are using. To compare the solve times by CHOLMOD and MG-BLPCG, in Figure 3, for the case $\varepsilon^2 = 10^{-12}$, we plot the solve times versus the degrees of freedom in the system.

While a nested dissection ordering would allow factorization of A in $\mathcal{O}(N^3)$ time [9], the CHOLMOD times scale more closely to $\mathcal{O}(N^4)$, in comparison to $\mathcal{O}(N^2)$ for MG-BLPCG. For comparison, we also include measured solve times for both BoxMG-preconditioned CG and AMG-preconditioned CG, as natural alternatives to the boundary-layer preconditioner considered here. For large meshes, the advantages of the MG-BLPCG approach are clearly noticeable.

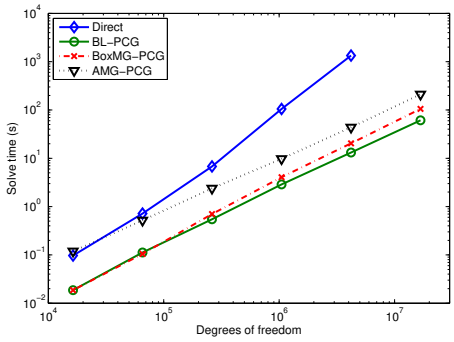

Fig. 3 Solve times for CHOLMOD, MG-BLPCG, BoxMG, and AMG schemes, taking $\varepsilon^2 = 10^{-12}$

5 Conclusions

This paper focuses on effective preconditioning strategies for the solution of the linear systems that arise from linear and bilinear finite-element discretizations of singularly perturbed reaction-diffusion equations on boundary-fitted meshes. We show that the boundary-layer preconditioned conjugate gradient algorithm proposed in [14] for finite-difference discretizations extends naturally to the finite-element case, but the analysis proposed therein must be extended to deal with the finite-element mass matrices that arise from the reaction term. The analysis presented in this paper applies only to the one-dimensional setting; its extension to two- and three-dimensional discretizations remains future work, complicated by the tensor-product structure of the higher-dimensional discretizations.

References

1. Alcouffe, R.E., Brandt, A., Dendy Jr., J.E., Painter, J.W.: The multigrid method for the diffusion equation with strongly discontinuous coefficients. SIAM J. Sci. Statist. Comput. **2**(4), 430–454 (1981). DOI 10.1137/0902035. URL `http://dx.doi.org/10.1137/0902035`
2. Anderson, E., Bai, Z., Bischof, C., Blackford, L.S., Demmel, J., Dongarra, J.J., Du Croz, J., Hammarling, S., Greenbaum, A., McKenney, A., Sorensen, D.: LAPACK Users' Guide (Third Ed.). Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, USA (1999)
3. Bakhvalov, N.: Towards optimization of methods for solving boundary value problems in the presence of boundary layers. Zh. Vychisl. Mat. i Mat. Fiz. **9**, 841–859 (1969)

4. Briggs, W.L., Henson, V.E., McCormick, S.F.: A multigrid tutorial, second edn. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2000). DOI 10.1137/1.9780898719505. URL http://dx.doi.org/10.1137/1.9780898719505
5. Chen, Y., Davis, T.A., Hager, W.W., Rajamanickam, S.: Algorithm 887: CHOLMOD, supernodal sparse Cholesky factorization and update/downdate. ACM Trans. Math. Softw. **35**(3), 22:1–22:14 (2008). DOI 10.1145/1391989.1391995. URL http://doi.acm.org/10.1145/1391989.1391995
6. Davis, T.A.: Direct methods for sparse linear systems, *Fundamentals of Algorithms*, vol. 2. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2006). DOI 10.1137/1.9780898718881. URL http://dx.doi.org/10.1137/1.9780898718881
7. Dendy Jr., J.E.: Black box multigrid. J. Comput. Phys. **48**(3), 366–386 (1982). DOI 10.1016/0021-9991(82)90057-2. URL http://dx.doi.org/10.1016/0021-9991(82)90057-2
8. Farrell, P.A., Hegarty, A.F., Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Robust Computational Techniques for Boundary Layers. No. 16 in Applied Mathematics. Chapman & Hall/CRC, Boca Raton, U.S.A. (2000)
9. George, A.: Nested dissection of a regular finite element mesh. SIAM J. Numer. Anal. **10**, 345–363 (1973)
10. Kopteva, N., Madden, N., Stynes, M.: Grid equidistribution for reaction-diffusion problems in one dimension. Numer. Algorithms **40**(3), 305–322 (2005). DOI 10.1007/s11075-005-7079-6
11. LAPACK version 3.5.0: Subroutine DPTTRS. www.netlib.org/lapack/
12. Linß, T.: Layer-adapted meshes for reaction-convection-diffusion problems. Lecture notes in Mathematics 1985. Berlin: Springer. xi, 320 p. (2010). DOI 10.1007/978-3-642-05134-0
13. 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). DOI 10.1093/imanum/drn048. URL http://dx.doi.org/10.1093/imanum/drn048
14. MacLachlan, S., Madden, N.: Robust solution of singularly perturbed problems using multigrid methods. SIAM J. Sci. Comput. **35**(5), A2225–A2254 (2013). DOI 10.1137/120889770. URL http://dx.doi.org/10.1137/120889770
15. 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, NJ (2012). DOI 10.1142/9789814390743. URL http://dx.doi.org/10.1142/9789814390743. Error estimates in the maximum norm for linear problems in one and two dimensions
16. Nhan, T.A., Madden, N.: Cholesky factorisation of linear systems coming from finite difference approximations of singularly perturbed problems. In: BAIL 2014–Boundary and Interior Layers, Computational and Asymptotic Methods, Lect. Notes Comput. Sci. Eng., pp. 209–220. Springer International Publishing (2015). DOI 10.1007/978-3-319-25727-3_16
17. Roos, H.G.: A note on the conditioning of upwind schemes on Shishkin meshes. IMA J. Numer. Anal. **16**(4), 529–538 (1996). DOI 10.1093/imanum/16.4.529. URL http://dx.doi.org/10.1093/imanum/16.4.529
18. 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-Verlag, Berlin (2008)
19. Trottenberg, U., Oosterlee, C.W., Schüller, A.: Multigrid. Academic Press, Inc., San Diego, CA (2001)
20. Wathen, A.J.: Realistic eigenvalue bounds for the Galerkin mass matrix. IMA J. Numer. Anal. **7**(4), 449–457 (1987). DOI 10.1093/imanum/7.4.449. URL http://dx.doi.org/10.1093/imanum/7.4.449
21. Wesseling, P.: An introduction to multigrid methods. Pure and Applied Mathematics (New York). John Wiley & Sons, Ltd., Chichester (1992)