SIAM J. NUMER. ANAL.
Vol. 52, No. 6, pp. 2957–2976

ROBUST PRECONDITIONING ESTIMATES FOR CONVECTION-DOMINATED ELLIPTIC PROBLEMS VIA A STREAMLINE POINCARÉ–FRIEDRICHS INEQUALITY*

OWE AXELSSON†, JÁNOS KARÁTSON‡, AND BALÁZS KOVÁCS§

Abstract. This paper is devoted to the streamline diffusion finite element method, combined with equivalent preconditioning, for solving convection-dominated elliptic problems. The preconditioner is obtained from the streamline diffusion inner product. It is proved that the obtained convergence is robust, i.e., bounded independently of the perturbation parameter ε, for proper convection vector fields. The key to the estimates is an improved "streamline" Poincaré–Friedrichs inequality.

Key words. streamline diffusion finite element method, robust preconditioning, Poincaré–Friedrichs inequality

AMS subject classifications. 65N30, 65F10

DOI. 10.1137/130940268

1. Introduction.

Convection-dominated elliptic equations form an important class in the modeling of stationary convection-diffusion problems and hence are the subject of intense research with vast literature; see, e.g., [12, 15, 16, 19] and the references therein. A common point is that standard finite element discretizations are inadequate for such problems and are hence replaced by some stabilized version, for which various approaches have been proposed. A widespread method in this class is the streamline diffusion finite element method (SDFEM); see, e.g., [12, Chap. 3].

The arising linear systems are generally solved by some preconditioned (conjugate gradient type) iterative method. The convergence of these iterations is also influenced by the convection-dominated character, i.e., the convergence becomes slow if the coefficient ε of the diffusion term is small. Our preconditioning approach is the equivalent operator preconditioning; see [13] for a solid foundation and [6] for a detailed survey for various elliptic problems. In particular, equivalent operator preconditioning has been applied in [14] for the SDFEM. However, just as for other such methods mentioned in [6], even for this preconditioned version the convergence estimates become arbitrarily slow if $\varepsilon \to 0$.

Our goal is to prove that the convergence using streamline diffusion preconditioning can in fact be robust, i.e., bounded independently of ε, for proper convection vector fields. We prove this via an improved "streamline" Poincaré–Friedrichs inequality. Altogether, our aim is to show that a proper combination of the two approaches (SDFEM and equivalent operators) results in a robustextension of the latter

*Received by the editors October 8, 2013; accepted for publication (in revised form) September 25, 2014; published electronically December 11, 2014. This research was supported by the Hungarian Scientific Research Fund OTKA, grant 112157.

http://www.siam.org/journals/sinum/52-6/94026.html

†Institute of Geonics AS CR, IT4 Innovations, Ostrava, Czech Republic (owea@it.uu.se).

‡Department of Applied Analysis and MTA-ELTE NumNet Research Group, ELTE University, Budapest H-1518, Hungary, and Department of Analysis, Technical University, Budapest, Hungary (karatson@cs.elte.hu).

§Department of Applied Analysis and MTA-ELTE NumNet Research Group, ELTE University, Budapest, Hungary (koboaet@cs.elte.hu).

to certain convection-dominated problems. The numerical tests reinforce the obtained theoretical estimates.

2. The problem and the SDFEM. For simplicity we present the results in detail for a simple class of problems with Dirichlet boundary conditions:

$$\begin{cases} -\varepsilon\Delta u + \mathbf{w}\cdot\nabla u = g, \\ u_{|\partial\Omega} = 0, \end{cases} \tag{2.1}$$

which satisfies the following.

Assumption 1.

(i) $\Omega \subset \mathbb{R}^d$ is a polyhedral domain.

(ii) $\mathbf{w} \in C^1(\overline{\Omega}, \mathbb{R}^n)$, $\operatorname{div}\mathbf{w} = 0$.

(iii) $g \in L^2(\Omega)$.

In section 4 we shall indicate the obvious modifications for more general boundary value problems, namely, allowing mixed boundary conditions (i.e., including boundary inflow), proper non-div-free convection fields, and lower order terms. The smoothness of the convection field can also be relaxed; see Remark 3.8. The homogeneity of the boundary conditions in (2.1) also serves only simplicity of exposition; the nonhomogeneous case can be reduced to this in a standard way.

For the weak formulation we use the real Hilbert space $H_0^1(\Omega)$ with inner product

$$\langle u, v\rangle_{H_0^1} = \int_\Omega \nabla u \cdot \nabla v.$$

Then the conditions ensure (see, e.g., [12, Chap. 3]) that problem (2.1) has a unique weak solution, i.e., $u \in H_0^1(\Omega)$ that satisfies

$$\int_\Omega \big(\varepsilon\,\nabla u\cdot\nabla v + (\mathbf{w}\cdot\nabla u)v\big) = \int_\Omega gv \qquad (\forall v \in H_0^1(\Omega)). \tag{2.2}$$

Let $\mathcal{T} = \{T_k\}_{k=1}^N$ be a triangulation of Ω into simplices and $V_h \subset H_0^1(\Omega)$ be the corresponding subspace of continuous, piecewise linear functions. The SDFEM is defined as follows [12, Chap. 3]. The usual finite element formulation is completed with a stabilizing term containing a set of parameters $\delta_k > 0$ $(k = 1, \dots, N)$:

$$\begin{aligned} &\int_\Omega \big(\varepsilon\,\nabla u_h\cdot\nabla v_h + (\mathbf{w}\cdot\nabla u_h)v_h\big) + \sum_{k=1}^N \delta_k \int_{T_k} (\mathbf{w}\cdot\nabla u_h)\,(\mathbf{w}\cdot\nabla v_h) \\ &\quad = \int_\Omega g(v_h + \delta\mathbf{w}\cdot\nabla v_h) \qquad (\forall v_h \in V_h). \end{aligned} \tag{2.3}$$

The left-hand side of (2.3) is called the streamline diffusion bilinear form:

$$\begin{aligned} a_{SD}(u_h, v_h) := &\int_\Omega \big(\varepsilon\,\nabla u_h\cdot\nabla v_h + (\mathbf{w}\cdot\nabla u_h)v_h\big) \\ &+ \sum_{k=1}^N \delta_k \int_{T_k} (\mathbf{w}\cdot\nabla u_h)\,(\mathbf{w}\cdot\nabla v_h) \qquad (u_h,\, v_h \in V_h). \end{aligned} \tag{2.4}$$

The corresponding streamline diffusion inner product is defined as

$$\langle u_h, v_h\rangle_{SD} := \int_\Omega \varepsilon\,\nabla u_h\cdot\nabla v_h + \sum_{k=1}^N \delta_k \int_{T_k} (\mathbf{w}\cdot\nabla u_h)\,(\mathbf{w}\cdot\nabla v_h) \qquad (u_h,\, v_h \in V_h) \tag{2.5}$$

with the induced norm

$$\|u_h\|_{SD}^2 = \int_\Omega \varepsilon \, |\nabla u_h|^2 + \sum_{k=1}^N \delta_k \int_{T_k} |\mathbf{w} \cdot \nabla u_h|^2 \, .$$

The streamline diffusion method involves a proper choice of the parameters δ_k. A widespread choice is $\delta_k = O\big(\frac{h_k}{|\mathbf{w}_k|}\big)$, where h_k denotes the diameter of T_k and $\mathbf{w}_k := \mathbf{w}_{|T_k}$; for a fixed convection field and uniform parameters on a regular mesh, this choice is simply $\delta = O(h)$. Then under proper assumptions [12, Chap. 3] the SDFEM converges, and its discretization error satisfies

$$\|u - u_h\|_{SD} \le C h^{3/2} \, \|D^2 u\|_{L^2},$$

where $D^2 u$ denotes the Hessian. The estimates and also our results in this paper involve the minimal parameter, which we will denote as

$$\delta_0 := \min_{k=1,\dots,N} \delta_k > 0 \, . \tag{2.6}$$

It is important to note that the parameters δ_k (in particular, δ_0) are chosen independently of ε.

Remark 2.1. The SDFEM can also be defined for general types of elements, in which case the bilinear form (2.4) is completed by an additional term

$$-\delta\varepsilon \sum_k \int_{T_k} (\Delta u_h) \mathbf{w} \cdot \nabla v_h;$$

see also [12]. (This term has vanished for the piecewise linear case above.) Even in this case one has a perturbation term of order $O(\delta\varepsilon)$. Since we assume that ε is small and $\delta = O(h)$, it turns out that this term is smaller than the discretization error, except possibly in layers, and is hence usually deleted from further considerations as well.

Remark 2.2. Let $Lu := -\varepsilon\Delta u + \mathbf{w} \cdot \nabla u$ denote the convection-diffusion operator. A possible way to improve the approximation of a problem

$$Lu = f, \qquad u_{|\partial\Omega} = \psi, \tag{2.7}$$

in the boundary layers is to solve first the hyperbolic problem

$$\mathbf{w} \cdot \nabla u_0 = f, \qquad u_{0|\Gamma_-} = \psi,$$

using a method of characteristics (where Γ_- denotes the inflow boundary, i.e., where $\mathbf{w} \cdot \nu < 0$), and then to solve the defect-correction equation

$$\begin{cases} Lv = f - Lu_0, \\ v_{|\partial\Omega} = \psi - u_0. \end{cases} \tag{2.8}$$

Clearly $u := u_0 + v$ solves (2.7) but v is small away from the layer. Based on the residual $f - Lu_0$, one can adapt the mesh to better resolve the layers. After homogenization of the boundary condition, problem (2.8) is of the same type as (2.1), hence the results of the paper are valid.

If $\varphi_1, \dots, \varphi_n$ is a basis in V_h and we seek $u_h = \sum_{i=1}^n c_i \varphi_i$, then the finite element problem (2.3) leads to a linear algebraic system

$$\mathbf{A}\mathbf{c} = \mathbf{b}, \tag{2.9}$$

where $\mathbf{c} = (c_1, \dots, c_n)^T$ and

$$a_{ij} = a_{SD}(\varphi_j, \varphi_i), \qquad b_i = \int_\Omega g\varphi_i \qquad (i, j = 1, \dots, n).$$

3. The streamline diffusion preconditioner. Let $\mathbf{S}$ be the stiffness matrix corresponding to the inner product $\langle ., . \rangle_{SD}$:

$$s_{ij} = \langle \varphi_j, \varphi_i \rangle_{SD} \qquad (i, j = 1, \dots, n).$$

We propose $\mathbf{S}$ as preconditioner to system (2.9):

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{c} = \tilde{\mathbf{b}}, \tag{3.1}$$

where $\tilde{\mathbf{b}} := \mathbf{S}^{-1}\mathbf{b}$. The preconditioning matrix $\mathbf{S}$ will be called the "streamline diffusion preconditioner." Here $\mathbf{S}$ is a symmetric, positive definite (s.p.d.) matrix. The auxiliary systems hence can be solved with a variety of methods, discussed in subsection 3.4.

We note that $\mathbf{S}$ is the symmetric part of $\mathbf{A}$ under the conditions of problem (2.1), i.e., with div-free convection field and Dirichlet boundary conditions. (This is not the case for the more general problem to be mentioned in section 4; then $\mathbf{S}$ is a kind of shifted symmetric part.) Preconditioning with the symmetric part has long been a widespread strategy (see, e.g., [11, 21] and the authors' paper [5]); in particular, it allows a short one-step recurrence in a proper iterative solution method (see Remark 3.2 below). However, to the authors' knowledge it had not been considered for SDFEM before the paper [14], whose estimates will be revisited in section 3.1.

The convergence properties are studied w.r.t. the $\mathbf{S}$-inner product

$$\langle \mathbf{c}, \mathbf{d} \rangle_{\mathbf{S}} := \mathbf{S}\,\mathbf{c} \cdot \mathbf{d} \qquad (\mathbf{c}, \mathbf{d} \in \mathbb{R}^d)$$

and the corresponding $\mathbf{S}$-norm $|\mathbf{c}|_{\mathbf{S}}^2 := \mathbf{S}\mathbf{c} \cdot \mathbf{c}$.

The preconditioned problem is solved with a CG-type iterative method, designed either directly for the original system such as the GMRES, Orthomin, or GCG-LS methods [3, 10, 17] or for the normal (symmetrized) system such as the CGN method. The convergence of these methods depends on the coercivity bound and on the $\mathbf{S}$-norm of $\mathbf{S}^{-1}\mathbf{A}$:

$$\begin{aligned} &\lambda_0 := \lambda_0(\mathbf{S}^{-1}\mathbf{A}) := \inf\{\langle \mathbf{S}^{-1}\mathbf{A}\mathbf{c}, \mathbf{c}\rangle_{\mathbf{S}} : \ |\mathbf{c}|_{\mathbf{S}} = 1\} > 0, \\ &\Lambda := \Lambda(\mathbf{S}^{-1}\mathbf{A}) := \|\mathbf{S}^{-1}\mathbf{A}\|_{\mathbf{S}} = \sup\{\langle \mathbf{S}^{-1}\mathbf{A}\mathbf{c}, \mathbf{d}\rangle_{\mathbf{S}} : \ |\mathbf{c}|_{\mathbf{S}} = |\mathbf{d}|_{\mathbf{S}} = 1\}. \end{aligned} \tag{3.2}$$

Namely, for the GMRES, Orthomin, or GCG-LS methods, the rate of linear convergence is bounded by

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le \left(1 - \left(\frac{\lambda_0}{\Lambda}\right)^2\right)^{1/2} \qquad (k = 1, 2, \dots, n) \tag{3.3}$$

and for the CGN method by

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le 2^{1/k}\,\frac{\Lambda - \lambda_0}{\Lambda + \lambda_0} \qquad (k = 1, 2, \dots, n). \tag{3.4}$$

Convergence estimates can alternatively be described in terms of field-of-values; see e.g., [18].

Remark 3.1. The CGN method is often avoided since the normal equation may lead to a higher condition number. On the other hand, it involves a very simple recursion in contrast to the GMRES and Orthomin methods, and for many nonsymmetric problems it has proved to be efficient [9, 13]. We will also find in our tests that it converges as fast as those methods satisfying (3.3). A possible reason is that the convergence rate (3.4) is smaller than the rate (3.3) of GMRES, GCG-LS, etc., which may compensate the extra work. The main disadvantage of the CGN method may arise when the matrix is close to symmetric, whereas in our problem the matrix is strongly nonsymmetric.

Remark 3.2. The GCG-LS method, which is one of the CG-type iterations that avoid the normal equation and yield the convergence rate (3.3), is particularly efficient when symmetric part preconditioning is used; see [2, 21]. In this case the full GCG-LS algorithm reduces automatically to the truncated version GCG-LS(0), which consists of a simple one-step recurrence.

The convergence results that follow are based on the theory of equivalent preconditioning [6, 13].

3.1. Equivalent preconditioning. The main idea of equivalent preconditioning in the context of bilinear forms is that the bounds are inherited uniformly by the stiffness matrices as follows.

PROPOSITION 3.3. *Let the bilinear form a_{SD} be bounded and coercive w.r.t. the inner product $\langle .,. \rangle_{SD}$ with bounds M and m, that is,*

$$|a_{SD}(u_h, v_h)| \le M \|u_h\|_{SD} \|v_h\|_{SD}, \qquad a_{SD}(u_h, u_h) \ge m \|u_h\|_{SD}^2 \qquad (\forall u_h,\, v_h \in V_h).$$

Then $\mathbf{S}^{-1}\mathbf{A}$ inherits the same bounds w.r.t. the $\mathbf{S}$-norm, i.e.,

$$|\langle \mathbf{S}^{-1}\mathbf{A}\mathbf{c},\, \mathbf{d}\rangle_{\mathbf{S}}| \le M\, |\mathbf{c}|_{\mathbf{S}}\, |\mathbf{d}|_{\mathbf{S}}, \qquad \langle \mathbf{S}^{-1}\mathbf{A}\mathbf{c},\, \mathbf{c}\rangle_{\mathbf{S}} \ge m\, |\mathbf{c}|_{\mathbf{S}}^2 \qquad (\forall\, \mathbf{c}, \mathbf{d} \in \mathbb{R}^d).$$

Proof. It follows in a standard way [6, 14]. For completeness we give the following brief proof: for arbitrary $\mathbf{c}, \mathbf{d} \in \mathbb{R}^d$, letting $u_h = \sum_{j=1}^n c_j \varphi_j \in V_h$ and $v_h = \sum_{j=1}^n d_j \varphi_j \in V_h$, we obtain

$$|\langle \mathbf{S}^{-1}\mathbf{A}\mathbf{c},\, \mathbf{d}\rangle_{\mathbf{S}}| = |\mathbf{A}\,\mathbf{c}\cdot\mathbf{d}| = |a_{SD}(u_h, v_h)| \le M \|u_h\|_{SD} \|v_h\|_{SD} = M\, |\mathbf{c}|_{\mathbf{S}}\, |\mathbf{d}|_{\mathbf{S}},$$
$$\langle \mathbf{S}^{-1}\mathbf{A}\mathbf{c},\, \mathbf{c}\rangle_{\mathbf{S}} = \mathbf{A}\,\mathbf{c}\cdot\mathbf{c} = a_{SD}(u_h, u_h) \ge m \|u_h\|_{SD}^2 = m\, |\mathbf{c}|_{\mathbf{S}}^2. \qquad \square$$

Therefore our task is to estimate m and M. As is well known [12, Chap. 3], the coercivity bound equals $m = 1$ under Assumption 1(ii) on $\mathbf{w}$: the divergence theorem implies

$$\int_\Omega (\mathbf{w}\cdot\nabla u_h) u_h = -\frac{1}{2}\int_\Omega (\operatorname{div}\mathbf{w}) u_h^2 = 0,$$

and hence

$$a_{SD}(u_h, u_h) = \int_\Omega \left(\varepsilon\, |\nabla u_h|^2 + (\mathbf{w}\cdot\nabla u_h) u_h\right) + \sum_{k=1}^N \delta_k \int_{T_k} |\mathbf{w}\cdot\nabla u_h|^2 \,= \|u_h\|_{SD}^2 .$$

On the other hand, the straightforward estimate of the upper bound will depend on ε. Since

$$a_{SD}(u_h, v_h) = \langle u_h, v_h \rangle_{SD} + \int_\Omega (\mathbf{w} \cdot \nabla u_h) v_h \,,$$

a natural upper estimate is

$$|a_{SD}(u_h, v_h)| \le \|u_h\|_{SD} \|v_h\|_{SD} + \|\mathbf{w} \cdot \nabla u_h\|_{L^2(\Omega)} \|v_h\|_{L^2(\Omega)}.$$

Here from (2.6)

$$\|\mathbf{w} \cdot \nabla u_h\|^2_{L^2(\Omega)} \le \frac{1}{\delta_0} \sum_{k=1}^{N} \delta_k \int_{T_k} |\mathbf{w} \cdot \nabla u_h|^2 \le \frac{1}{\delta_0} \|u_h\|^2_{SD} \,, \tag{3.5}$$

and hence

$$|a_{SD}(u_h, v_h)| \le \|u_h\|_{SD} \|v_h\|_{SD} + \frac{1}{\sqrt{\delta_0}} \|u_h\|_{SD} \|v_h\|_{L^2(\Omega)}$$

$$\le \left(1 + \frac{1}{\sqrt{\delta_0}} \sup_{v_h \in V_h} \frac{\|v_h\|_{L^2(\Omega)}}{\|v_h\|_{SD}}\right) \|u_h\|_{SD} \|v_h\|_{SD}. \tag{3.6}$$

Now we use the Poincaré–Friedrichs inequality

$$\|v\|_{L^2(\Omega)} \le C_\Omega \, \|\nabla v\|_{L^2(\Omega)} \qquad (v \in H^1_0(\Omega)).$$

As pointed out in the recent paper [14], since by definition

$$\|v_h\|^2_{SD} \ge \varepsilon \, \|\nabla v_h\|^2_{L^2(\Omega)},$$

one obtains the estimate

$$\|v_h\|_{L^2(\Omega)} \le \frac{C_\Omega}{\sqrt{\varepsilon}} \, \|v_h\|_{SD} \qquad (\forall v_h \in V_h). \tag{3.7}$$

Hence

$$|a_{SD}(u_h, v_h)| \le \left(1 + \frac{C_\Omega}{\sqrt{\delta_0 \varepsilon}}\right) \|u_h\|_{SD} \|v_h\|_{SD},$$

i.e., the upper bound is estimated as

$$M \le 1 + \frac{C_\Omega}{\sqrt{\delta_0 \varepsilon}} \,. \tag{3.8}$$

However, this bound deteriorates, i.e., tends to $+\infty$ as $\varepsilon \to 0$. Our goal is to replace the estimate (3.7) by one that is independent of ε. For this purpose we must compare $\|v\|_{L^2(\Omega)}$ to the streamline term instead of the first ε-dependent diffusion term in $\|.\|_{SD}$. This is an analogue of the standard Poincaré–Friedrichs inequality, and hence it will be called the streamline Poincaré–Friedrichs inequality.

3.2. A streamline Poincaré–Friedrichs inequality. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. Let us consider a vector field $\mathbf{w} \in C^1(\Omega, \mathbb{R}^d)$ and the corresponding system of ordinary differential equations

$$\dot{\gamma}(t) = \mathbf{w}(\gamma(t)). \tag{3.9}$$

The solutions of (3.9) are called characteristic curves corresponding to the vector field $\mathbf{w}$.

THEOREM 3.4 (see [1]). *A vector field $\mathbf{w} \in C^1(\Omega, \mathbb{R}^d)$ for which $\mathbf{w}(\mathbf{x}) \neq 0$ $(\mathbf{x} \in \Omega)$ is locally rectifiable. This means that any $\mathbf{x} \in \Omega$ has a neighborhood $V_{\mathbf{x}}$ such that the characteristic curves can be locally parametrized by a diffeomorphism $f_{\mathbf{x}} : U_{\mathbf{x}} \to V_{\mathbf{x}}$ on some open set $U_{\mathbf{x}}$, where parametrization means a mapping*

$$f_{\mathbf{x}}(s_1, \dots, s_{n-1}, t) := \gamma_{s_1,\dots,s_{n-1}}(t) \qquad \big((s_1, \dots, s_{n-1}, t) \in U_{\mathbf{x}}\big)$$

such that $t \mapsto \gamma_{s_1,\dots,s_{n-1}}(t)$ are a local family of characteristic curves.

DEFINITION 3.5. *A vector field $\mathbf{w} \in C^1(\overline{\Omega}, \mathbb{R}^d)$ for which $\mathbf{w}(\mathbf{x}) \neq 0$ $(\mathbf{x} \in \Omega)$ is called globally rectifiable on $\overline{\Omega}$ if the above local diffeomorphisms can be replaced by a global one onto $\overline{\Omega}$, i.e., there exists a diffeomorphism $f : K \to \overline{\Omega}$ on a compact set K such that*

$$f(s_1, \dots, s_{n-1}, t) := \gamma_{s_1,\dots,s_{n-1}}(t) \qquad \big((s_1, \dots, s_{n-1}, t) \in K\big),$$

where $t \mapsto \gamma_{s_1,\dots,s_{n-1}}(t)$ are the family of characteristic curves covering $\overline{\Omega}$. (Diffeomorphism means that $f \in C^1(K, \overline{\Omega})$ is one-to one and $f^{-1} \in C^1(\overline{\Omega}, K)$.)

We will briefly denote $\mathbf{s} := (s_1, \dots, s_{n-1})$; thus the above formula becomes

$$f(\mathbf{s}, t) := \gamma_{\mathbf{s}}(t) \qquad \big((\mathbf{s}, t) \in K\big). \tag{3.10}$$

An example of a globally rectifiable vector field in two dimensions is illustrated in Figure 1, where $\mathbf{w}(\mathbf{x}) \neq 0$ $(\mathbf{x} \in \Omega)$ and the inflow and outflow boundaries Γ_- (where $\mathbf{w} \cdot \nu < 0$) and Γ_+ (where $\mathbf{w} \cdot \nu \geq 0$), respectively, are connected.

We establish our theoretical results rigorously for globally rectifiable vector fields. We note that this property is restrictive, but the result can be extended to more general problems, as will be discussed in Remark 3.7.

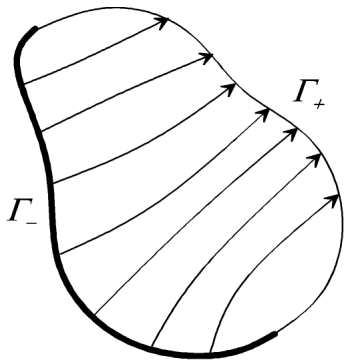

FIG. 1. *A globally rectifiable vector field.*

We will make use of the change of variables formula for integrating on Ω: for any function $z \in L^1(\Omega)$

$$\int_\Omega z(x_1,\dots,x_n)\,dx_1\dots\,dx_n = \int_K z(f(s_1,\dots,s_{n-1},t))\,\Big|\det\frac{\partial f(s_1,\dots,s_{n-1},t)}{\partial s_1\dots\partial s_{n-1}\partial t}\Big|\,ds_1\dots ds_{n-1}dt.$$

From Definition 3.5, using notation (3.10) (and the same for $\mathbf{x}$ and for the integration variables), and further using notation

$$J_{\mathbf{w}}(\mathbf{s},t) := \Big|\det\frac{\partial f(s_1,\dots,s_{n-1},t)}{\partial s_1\dots\partial s_{n-1}\partial t}\Big|$$

(which expresses that this Jacobian is ultimately determined by $\mathbf{w}$), we obtain

$$\int_\Omega z(\mathbf{x})\,d\mathbf{x} = \int_K z(\gamma_{\mathbf{s}}(t))\,J_{\mathbf{w}}(\mathbf{s},t)\,d\mathbf{s}\,dt. \tag{3.11}$$

The diffeomorphism property implies that

$$0 < \mu \le J_{\mathbf{w}}(\mathbf{s},t) \le \widetilde{\mu} \qquad \big((\mathbf{s},t)\in K\big), \tag{3.12}$$

where μ and $\widetilde{\mu}$ are independent of $(\mathbf{s},t)$.

THEOREM 3.6 (streamline Poincaré–Friedrichs inequality). *Let $\mathbf{w} \in C^1(\overline{\Omega},\mathbb{R}^d)$, for which $\mathbf{w}(\mathbf{x}) \ne 0$ ($\mathbf{x}\in\Omega$), be a globally rectifiable vector field on $\overline{\Omega}$. Then there exists a constant $C_{\mathbf{w}} > 0$ (depending on $\mathbf{w}$ but independent of v) such that*

$$\|v\|_{L^2(\Omega)} \le\ C_{\mathbf{w}}\,\|\mathbf{w}\cdot\nabla v\|_{L^2(\Omega)} \qquad (v\in H_0^1(\Omega)).$$

Proof. Let $v \in H_0^1(\Omega)$. Then (3.11) yields

$$\|v\|^2_{L^2(\Omega)} = \int_\Omega |v(\mathbf{x})|^2\,d\mathbf{x} = \int_K |v(\gamma_{\mathbf{s}}(t))|^2\,J_{\mathbf{w}}(\mathbf{s},t)\,d\mathbf{s}\,dt. \tag{3.13}$$

For fixed $(\mathbf{s},t)\in K$ let $t_0(\mathbf{s}) < t$ be such that $\gamma_{\mathbf{s}}(t_0(\mathbf{s})) \in \partial\Omega$, i.e., where the curve intersects the inflow boundary. Then the boundary condition implies $v(\gamma_{\mathbf{s}}(t_0(\mathbf{s}))) = 0$; hence the Newton–Leibniz formula and (3.9) yield

$$\begin{aligned} v(\gamma_{\mathbf{s}}(t)) &= \int_{t_0(\mathbf{s})}^t \nabla v(\gamma_{\mathbf{s}}(\tau))\cdot\dot\gamma_{\mathbf{s}}(\tau)\,d\tau \\ &= \int_{t_0(\mathbf{s})}^t \nabla v(\gamma_{\mathbf{s}}(\tau))\cdot\mathbf{w}(\gamma_{\mathbf{s}}(\tau))\,d\tau = \int_{t_0(\mathbf{s})}^t (\mathbf{w}\cdot\nabla v)(\gamma_{\mathbf{s}}(\tau))\,d\tau. \end{aligned}$$

Multiplying the integrand with $J_{\mathbf{w}}(\mathbf{s},\tau)^{1/2}J_{\mathbf{w}}(\mathbf{s},\tau)^{-1/2}$, the Cauchy–Schwarz inequality then implies

$$|v(\gamma_{\mathbf{s}}(t))|^2 \le \int_{t_0(\mathbf{s})}^t |(\mathbf{w}\cdot\nabla v)(\gamma_{\mathbf{s}}(\tau))|^2\,J_{\mathbf{w}}(\mathbf{s},\tau)\,d\tau\cdot\int_{t_0(\mathbf{s})}^t \frac{1}{J_{\mathbf{w}}(\mathbf{s},\tau)}\,d\tau.$$

Now let $t_1(\mathbf{s}) > t$ be such that $\gamma_{\mathbf{s}}(t_1(\mathbf{s})) \in \partial\Omega$, i.e., where the curve intersects the outflow boundary. Then, also using (3.12) and that $t_1(\mathbf{s}) - t_0(\mathbf{s}) < diam(K)$ (where $diam(K)$ denotes the diameter of K),

$$|v(\gamma_{\mathbf{s}}(t))|^2 \le \int_{t_0(\mathbf{s})}^{t_1(\mathbf{s})} |(\mathbf{w}\cdot\nabla v)(\gamma_{\mathbf{s}}(\tau))|^2\, J_{\mathbf{w}}(\mathbf{s},\tau)\, d\tau \cdot \int_{t_0(\mathbf{s})}^{t_1(\mathbf{s})} \frac{1}{J_{\mathbf{w}}(\mathbf{s},\tau)}\, d\tau$$

$$\le \int_{t_0(\mathbf{s})}^{t_1(\mathbf{s})} |(\mathbf{w}\cdot\nabla v)(\gamma_{\mathbf{s}}(\tau))|^2\, J_{\mathbf{w}}(\mathbf{s},\tau)\, d\tau \cdot \frac{diam(K)}{\mu}\ . \tag{3.14}$$

Here the set K can be given as $K = \{(\mathbf{s},t) \in \mathbb{R}^d :\ \mathbf{s}\in S,\ t_0(\mathbf{s}) < t < t_1(\mathbf{s})\}$, where the proper compact set $S \subset \mathbb{R}^{n-1}$ parametrizes the family of curves. Then we can rewrite (3.13) as

$$\|v\|^2_{L^2(\Omega)} = \int_S \int_{t_0(\mathbf{s})}^{t_1(\mathbf{s})} |v(\gamma_{\mathbf{s}}(t))|^2\, J_{\mathbf{w}}(\mathbf{s},t)\, dt\, d\mathbf{s}.$$

Here the first factor in the integrand in (3.14) is a function of $\mathbf{s}$ but not of t, hence we obtain

$$\|v\|^2_{L^2(\Omega)} \le \frac{diam(K)}{\mu} \int_S \int_{t_0(\mathbf{s})}^{t_1(\mathbf{s})} J_{\mathbf{w}}(\mathbf{s},t)\, dt \int_{t_0(\mathbf{s})}^{t_1(\mathbf{s})} |(\mathbf{w}\cdot\nabla v)(\gamma_{\mathbf{s}}(\tau))|^2\, J_{\mathbf{w}}(\mathbf{s},\tau)\, d\tau\, d\mathbf{s}$$

$$\le \frac{\widetilde{\mu}\, diam(K)^2}{\mu} \int_S \int_{t_0(\mathbf{s})}^{t_1(\mathbf{s})} |(\mathbf{w}\cdot\nabla v)(\gamma_{\mathbf{s}}(\tau))|^2\, J_{\mathbf{w}}(\mathbf{s},\tau)\, d\tau\, d\mathbf{s}$$

$$= \frac{\widetilde{\mu}\, diam(K)^2}{\mu} \int_K |(\mathbf{w}\cdot\nabla v)(\gamma_{\mathbf{s}}(\tau))|^2\, J_{\mathbf{w}}(\mathbf{s},\tau)\, d\mathbf{s}\, d\tau\ .$$

In view of (3.11), we obtain

$$\|v\|^2_{L^2(\Omega)} \le \frac{\widetilde{\mu}\, diam(K)^2}{\mu} \int_\Omega |(\mathbf{w}\cdot\nabla v)(\mathbf{x})|^2\, d\mathbf{x} = C^2_{\mathbf{w}}\, \|\mathbf{w}\cdot\nabla v\|^2_{L^2(\Omega)},$$

where $C_{\mathbf{w}} := diam(K)\sqrt{\widetilde{\mu}/\mu}$ is determined by the diffeomorphism f and thus by the field $\mathbf{w}$ but is independent of v. ∎

Remark 3.7. The condition $\mathbf{w}(\mathbf{x}) \neq 0$ and the global rectifiability in Theorem 3.6 are restrictive, but it can be seen from the proof that these are not necessary. As seen from (3.14), it suffices to have a parametrization of Ω such that the determinants satisfy

$$\max_{\mathbf{s}\in S} \int_{t_0(\mathbf{s})}^{t_1(\mathbf{s})} \frac{1}{J_{\mathbf{w}}(\mathbf{s},\tau)}\, d\tau\ <\ \infty.$$

For example, let us consider a two-dimensional vector field

$$\mathbf{w}(x_1,x_2) = (x_1\, g(x_1,x_2) - x_2,\ \ x_2\, g(x_1,x_2) - x_1)$$

with some given scalar function $g \in C^1(\overline{\Omega})$. We look for the parametrized solutions $\gamma_s(t) = (\gamma_s^{(1)}(t),\ \gamma_s^{(2)}(t))$ in the form

$$\gamma_s(t) = (\gamma_s^{(1)}(t),\ \gamma_s^{(2)}(t)) = \big(r_s(t)\, \cos t,\ r_s(t)\, \sin t\big).$$

If each function r_s solves the ODE

$$\dot{r}_s(t)\ =\ r_s(t)\ \ g\big(r_s(t)\, \cos t,\ r_s(t)\, \sin t\big),$$

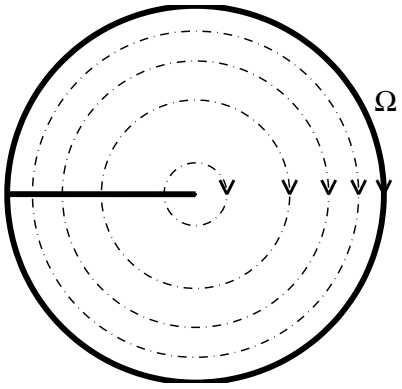

FIG. 2. *A circular vector field.*

then an elementary calculation yields that

$$\dot\gamma_s^{(1)}(t) = \dot r_s(t)\,\cos t - r_s(t)\,\sin t = r_s(t)\,\cos t\; g\big(r_s(t)\,\cos t,\; r_s(t)\,\sin t\big) - r_s(t)\,\sin t$$
$$= \gamma_s^{(1)}(t)\; g(\gamma_s^{(1)}(t), \gamma_s^{(2)}(t)) - \gamma_s^{(2)}(t) = w_1\big(\gamma_s^{(1)}(t),\; \gamma_s^{(2)}(t)\big)$$

and similarly $\dot\gamma_s^{(2)}(t) = w_2\big(\gamma_s^{(1)}(t),\; \gamma_s^{(2)}(t)\big)$. That is, each γ_s solves (3.9), i.e., they are the characteristic curves corresponding to the above vector field $\mathbf{w}$. We can then define

$$f(s,t) := \gamma_s(t) := \big(r_s(t)\,\cos t,\; r_s(t)\,\sin t\big)$$

from (3.10). Then

$$J_{\mathbf{w}}(s,t) = \begin{vmatrix} \frac{\partial r_s(t)}{\partial s}\,\cos t & -r_s(t)\,\sin t \\ \frac{\partial r_s(t)}{\partial s}\,\sin t & r_s(t)\,\cos t \end{vmatrix} = r_s(t)\frac{\partial r_s(t)}{\partial s} = \frac{1}{2}\frac{\partial}{\partial s}\Big(r_s^2(t)\Big).$$

This shows that if

$$\frac{\partial}{\partial s}\Big(r_s^2(t)\Big) \ge \text{const} > 0,$$

i.e., $r_s^2(t)$ grows at least linearly, then $J_{\mathbf{w}}(s,t)$ has a positive lower bound, and then the same proof works as in Theorem 3.6.

As a concrete simple example from the above class, we can consider the circular vector field

$$\mathbf{w}(x_1,x_2) = (-x_2,\, x_1);$$

see Figure 2. The circular characteristic curves can be parametrized as

$$h(s,t) = \gamma_s(t) = \big(\sqrt{s}\,\cos t,\; \sqrt{s}\,\sin t\big),$$

i.e., we now have $r_s(t) = \sqrt{s}$ and $g \equiv 0$. Then $\frac{\partial r_s(t)}{\partial s} = \frac{1}{2\sqrt{s}}$ and hence

$$J_{\mathbf{w}}(s,t) \equiv \frac{1}{2}.$$

Hence Theorem 3.6 is valid, i.e., the streamline Poincaré–Friedrichs inequality holds for the circular vector field.

Remark 3.8. The condition $\mathbf{w} \in C^1$ can also be relaxed. We can even allow a piecewise constant field $\mathbf{w}$, yielding piecewise smooth curves γ, provided that the restrictions of the corresponding mapping f are diffeomorphisms on the subdomains on which $\mathbf{w}$ is constant.

3.3. Robust preconditioning for the convection-diffusion problem. Now we can readily summarize the results. For any $v_h \in V_h$ we have

$$\|\mathbf{w} \cdot \nabla v_h\|_{L^2(\Omega)} \leq \frac{1}{\sqrt{\delta_0}} \|v_h\|_{SD} \tag{3.15}$$

from (3.5); hence Theorem 3.6 yields

$$\|v_h\|_{L^2(\Omega)} \leq \frac{C_{\mathbf{w}}}{\sqrt{\delta_0}} \|v_h\|_{SD} \qquad (v_h \in V_h).$$

Then (3.6) implies the following.

COROLLARY 3.9. *Let $\mathbf{w} \in C^1(\overline{\Omega}, \mathbb{R}^d)$, for which $\mathbf{w}(\mathbf{x}) \neq 0$ ($\mathbf{x} \in \Omega$), be a globally rectifiable vector field on $\overline{\Omega}$. Then*

$$|a_{SD}(u_h, v_h)| \leq \left(1 + \frac{C_{\mathbf{w}}}{\delta_0}\right) \|u_h\|_{SD} \|v_h\|_{SD} \qquad (\forall u_h,\, v_h \in V_h).$$

That is, the upper bound of a_{SD} satisfies

$$M \leq 1 + \frac{C_{\mathbf{w}}}{\delta_0}, \tag{3.16}$$

which is an estimate independent of ε. Since the lower bound is $m = 1$, we have altogether proved the next theorem.

THEOREM 3.10. *Let $\mathbf{w} \in C^1(\overline{\Omega}, \mathbb{R}^d)$, for which $\mathbf{w}(\mathbf{x}) \neq 0$ ($\mathbf{x} \in \Omega$), be a globally rectifiable vector field on $\overline{\Omega}$. Then the linear convergence of the conjugate gradient method for the preconditioned system* (3.1) *is bounded independently of ε. Namely, for the GCG-LS method, the residual satisfies*

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \leq \left(1 - \frac{1}{M^2}\right)^{1/2} = \frac{\sqrt{C_{\mathbf{w}}(C_{\mathbf{w}} + 2\delta_0)}}{C_{\mathbf{w}} + \delta_0} \qquad (k = 1, 2, \ldots, n) \tag{3.17}$$

and for the CGN method

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \leq 2^{1/k} \frac{M - 1}{M + 1} = 2^{1/k} \frac{C_{\mathbf{w}}}{C_{\mathbf{w}} + 2\delta_0} \qquad (k = 1, 2, \ldots, n), \tag{3.18}$$

where both estimates are independent of ε.

Remark 3.11.

(i) We note that the above theorem also holds for certain not globally rectifiable vector fields, discussed in Remark 3.7.

(ii) It is seen that to avoid a too slow rate of convergence, the parameter δ_0 must not be taken too small. On the other hand, its choice influences the discretization error as mentioned in section 2. Hence this choice must be a balance between obtaining fewer iterations and a small discretization error. A typical choice, as also mentioned in section 2, is $\delta_0 = O(h)$; see [12].

3.4. Solution of the auxiliary problems. The preconditioner $\mathbf{S}$ is an s.p.d. matrix; hence the auxiliary systems can be solved with a variety of methods. Furthermore, we will show that although the matrix $\mathbf{S}$ itself depends on ε, the conditioning properties of $\mathbf{S}$ are independent of ε. For this we assume that the mesh is regular, i.e., there exists $\theta_0 > 0$ such that the smallest angles of elements satisfy

$$\theta_K \geq \theta_0$$

for all elements K.

The auxiliary systems can be solved with a lot of efficient methods elaborated for s.p.d. problems: for instance, one can use algebraic multilevel methods [7, 8] or algebraic multigrid (see, e.g., [20]). A general description of iterative methods is given in [3].

The performance of the above methods is mainly determined by the range of eigenvalues and hence by the condition number of the system matrix. Since the problems in our case depend on the parameter ε, it is basically important to see how the matrix depends on ε. We now verify that the condition number of $\mathbf{S}$ is bounded independently of ε.

Let $\mathbf{M}$ denote the mass matrix w.r.t. the same mesh, i.e.,

$$m_{ij} = \langle \varphi_j, \varphi_i \rangle_{L^2} \qquad (i, j = 1, \dots, n).$$

It is well known [12] that its eigenvalues $\lambda(\mathbf{M})$ satisfy

$$\lambda_{min}(\mathbf{M}) \geq C_2 h^d, \qquad \lambda_{max}(\mathbf{M}) \leq C_1 h^d, \tag{3.19}$$

where d is the space dimension (i.e., $\Omega \subset \mathbb{R}^d$), the constants C_1, $C_2 > 0$ depend on the domain Ω and the regularity parameter θ_0, and hence the condition number of $\mathbf{M}$ is uniformly bounded, i.e.,

$$\kappa(\mathbf{M}) \leq \frac{C_1}{C_2}.$$

Let $-\Delta_h$ denote the discretization of the negative Laplacian w.r.t. the same mesh as used for our boundary value problem, i.e.,

$$(-\Delta_h)_{ij} = \langle \nabla \varphi_i, \nabla \varphi_j \rangle_{L^2} \qquad (i, j = 1, \dots, n).$$

Then, as is also well known (see, e.g., [12]),

$$\lambda_{max}(-\Delta_h) = \sup_{\substack{\mathbf{c} \in \mathbb{R}^n \\ \mathbf{c} \neq 0}} \frac{-\Delta_h \mathbf{c} \cdot \mathbf{c}}{|\mathbf{c}|^2} \leq C_3 h^{d-2}, \qquad \lambda_{min}(-\Delta_h) \geq C_4 h^d, \tag{3.20}$$

where $C_3, C_4 > 0$ depend on the domain Ω and the regularity parameter θ_0. Hence

$$\kappa(-\Delta_h) \le \frac{C_3}{C_4}\, h^{-2}. \tag{3.21}$$

First we show that $\kappa(\mathbf{M}^{-1}\mathbf{S})$ is uniformly bounded w.r.t. ε.

PROPOSITION 3.12. *There exists $C > 0$ independently of h and ε such that*

$$\kappa(\mathbf{M}^{-1}\mathbf{S}) \le Ch^{-2}.$$

Proof. Let λ be an eigenvalue of $\mathbf{M}^{-1}\mathbf{S}$. Then some vector $\mathbf{c} \in \mathbb{R}^d,\ \mathbf{c} \neq 0$ satisfies

$$\mathbf{Sc} = \lambda\, \mathbf{Mc},$$

hence

$$\mathbf{Sc} \cdot \mathbf{c} = \lambda\, \mathbf{Mc} \cdot \mathbf{c}.$$

Let $u_h = \sum_{i=1}^n c_i \varphi_i$, where $\varphi_1, \ldots, \varphi_n$ is a basis in V_h as introduced in section 2. Then $u_h \neq 0$, and by definition

$$\mathbf{Sc} \cdot \mathbf{c} = \sum_{i,j=1}^{n} \langle \varphi_j, \varphi_i \rangle_{SD}\ c_j\, c_i = \|u_h\|_{SD}^2,$$

and similarly

$$\mathbf{Mc} \cdot \mathbf{c} = \|u_h\|_{L^2}^2, \tag{3.22}$$

therefore

$$\lambda = \frac{\|u_h\|_{SD}^2}{\|u_h\|_{L^2}^2}.$$

Hence we must give uniform upper and lower bounds for the above fraction. First, using (3.15) and Theorem 3.6,

$$\|u_h\|_{SD}^2 \ge \delta_0 \|\mathbf{w} \cdot \nabla u_h\|_{L^2(\Omega)}^2 \ge \frac{\delta_0}{C_{\mathbf{w}}^2} \|u_h\|_{L^2(\Omega)}^2,$$

hence

$$\lambda \ge \frac{\delta_0}{C_{\mathbf{w}}^2}.$$

On the other hand, since we study the case $\varepsilon \to 0$, we may assume that $\varepsilon \le \widetilde{\varepsilon}$, where $\widetilde{\varepsilon}$ is independent of ε. Also, let

$$\delta_{max} := \max_{k=1,\ldots,N} \delta_k; \tag{3.23}$$

then (since the standard chioce is $\delta_{max} = O(h) \to 0$) we may assume that $\delta_{max} \le \widetilde{\delta}$, where $\widetilde{\delta}$ is independent of h and ε. Then we have

$$\lambda = \frac{\int_\Omega \varepsilon\, |\nabla u_h|^2 + \sum_{k=1}^N \delta_k \int_{T_k} |\mathbf{w} \cdot \nabla u_h|^2}{\|u_h\|_{L^2}^2} \le \left(\widetilde{\varepsilon} + \widetilde{\delta} \|\mathbf{w}\|_\infty^2\right) \frac{\|\nabla u_h\|_{L^2}^2}{\|u_h\|_{L^2}^2}.$$

Here

$$\|\nabla u_h\|_{L^2}^2 = \sum_{i,j=1}^{n} \langle \nabla\varphi_j, \nabla\varphi_i\rangle_{L^2}\ c_j\, c_i = -\Delta_h \mathbf{c}\cdot\mathbf{c},$$

hence, using (3.22), (3.20), and (3.19), respectively,

$$\frac{\|\nabla u_h\|_{L^2}^2}{\|u_h\|_{L^2}^2} = \frac{-\Delta_h \mathbf{c}\cdot\mathbf{c}}{|\mathbf{c}|^2}\frac{|\mathbf{c}|^2}{\mathbf{M}\mathbf{c}\cdot\mathbf{c}} \le \frac{\lambda_{max}(-\Delta_h)}{\lambda_{min}(\mathbf{M})} \le \frac{C_3}{C_2} h^{-2}.$$

Altogether,

$$\kappa(\mathbf{M}^{-1}\mathbf{S}) = \frac{\lambda_{max}(\mathbf{M}^{-1}\mathbf{S})}{\lambda_{min}(\mathbf{M}^{-1}\mathbf{S})} \le \left(\widetilde{\varepsilon} + \widetilde{\delta}\|\mathbf{w}\|_\infty^2\right)\frac{C_3 C_{\mathbf{w}}^2}{C_2\,\delta_0}\, h^{-2} = C\, h^{-2},$$

where $C := (\widetilde{\varepsilon} + \widetilde{\delta}\|\mathbf{w}\|_\infty^2)\,\frac{C_3 C_{\mathbf{w}}^2}{C_2\,\delta_0}$ is independent of h and ε. ∎

THEOREM 3.13. *There exists $c > 0$ independently of h and ε such that*

$$\kappa(\mathbf{S}) \le c h^{-2}.$$

Proof. Using Proposition 3.12, we have

$$\kappa(\mathbf{S}) \le \kappa(\mathbf{M})\kappa(\mathbf{M}^{-1}\mathbf{S}) \le \frac{C_1}{C_2}\,\kappa(\mathbf{M}^{-1}\mathbf{S}) \le \frac{C_1 C}{C_2}\, h^{-2},$$

where $c := C_1 C/C_2$ is independent of h and ε. ∎

The obtained result means, when compared to (3.21), that the condition number of $\mathbf{S}$ behaves in the same way as that of the Laplacian, independently of ε. We note that this property also will be illustrated by the numerical tests in section 5. This implies that the performance of the mentioned multigrid and multilevel methods for the auxiliary systems involving $\mathbf{S}$ is qualitatively similar to the case of standard elliptic problems, in particular, of Poisson equations. We can then use a combined method, that is, precondition with $\mathbf{S}$ for the outer iterations while the arising systems with $\mathbf{S}$ are solved with a multigrid or multilevel method. This leads to a robust method with a rate of convergence independent of ε.

4. Generalizations to general mixed boundary value problems. It is straightforward to extend the above results to general mixed boundary value problems

$$\begin{cases} -\varepsilon\Delta u + \mathbf{w}\cdot\nabla u + qu = g, \\ u_{|\Gamma_D} = 0, \qquad \frac{\partial u}{\partial \nu} + \beta u_{|\Gamma_N} = 0, \end{cases} \tag{4.1}$$

that satisfy the following.

Assumption 2.

(i) $\Omega \subset \mathbb{R}^d$ is a polyhedral domain; $\partial\Omega = \Gamma_D \cup \Gamma_N$ is a decomposition of the boundary into nonoverlapping, relatively open subparts.
(ii) $\mathbf{w} \in C^1(\overline{\Omega}, \mathbb{R}^n)$, $q \in L^\infty(\Omega)$, $\beta \in L^\infty(\Gamma_N)$.
(iii) $q - \frac{1}{2}\operatorname{div}\mathbf{w} \ge 0$ in Ω, $\mathbf{w}\cdot\nu \ge 0$ on Γ_N.
(iv) $g \in L^2(\Omega)$.

Namely, the above assumptions also ensure the coercivity and boundedness of the corresponding bilinear form; see, e.g., [6]. Further, it is clear that the condition $u = 0$, used in Theorem 3.6, is only required in the proof on the inflow boundary Γ_-. Now, in assumption (iii) above, we have $\mathbf{w}\cdot\nu \ge 0$ on Γ_N, which means that $\Gamma_N \subset \Gamma_+$, i.e., $\Gamma_- \subset \Gamma_D$, that is, we have indeed $u = 0$ on Γ_-. Therefore Theorems 3.6 and 3.10 remain valid in the same form as proved above for Dirichlet problems.

5. Numerical experiments. We have run numerical tests for two model problems of the class (2.1). We have used Courant elements for the FEM subspace, and for simplicity the constants δ_n have all been fixed to a common value δ. The algorithm was carried out in MATLAB and the tests were run on a standard desktop PC. We have used the preconditioned CGN (PCGN) iteration for both test problems; see Remarks 3.1 and 3.2 for other iterations. For the first problem we have also run the GCGLS(0) iteration, which is a short recurrence avoiding normal equations. It is thus more directly suited to the studied symmetrically preconditioned system than the GMRES but yields the same convergence rate (3.3); see Remark 3.2. However, the computer work proved to be the same for GCGLS(0) as for the CGN method; hence for the second problem only PCGN was used. The auxiliary linear systems were solved with built-in solvers, due to the modest size of the problems.

The following numerical tests strengthen our theoretical results. As predicted by Theorem 3.10, the convergence of the iteration with streamline preconditioning is robust in ε, i.e., the iteration number to achieve a certain tolerance (Tol) is bounded independently of ε as $\varepsilon \searrow 0$. In the second test with enclosed flow, the actual iteration numbers in fact approach very close to our theoretical uniform bound, i.e., the latter can be thought of as realistic. Further, as predicted by Theorem 3.13, the conditioning properties of the preconditioning matrices are also bounded independently of ε, which is shown by the bounded amount of total computer work.

5.1. Layer near a segment of a square. Let our domain be the unit square $\Omega := [0,1]^2$ in $\mathbb{R}^2$ and the vector field be the constant $\mathbf{w} := (1,0)$. For better control of the error, we consider a problem where the exact solution is known. The function g in (2.1) is chosen such that the exact solution of the problem is

$$u(x,y) = \left(x - \frac{e^{x/\varepsilon} - 1}{e^{1/\varepsilon} - 1}\right) 4y(1-y). \tag{5.1}$$

Here the first factor of $u(x,y)$ is the exact solution of the well-known one-dimensional problem $-\varepsilon u'' + u' = 1, \quad u(0) = u(1) = 0$ that has a boundary layer near $x = 1$. Therefore the function $u(x,y)$ has a boundary layer as well near the segment $x = 1$, transformed such that the boundary condition is satisfied on the whole boundary of Ω (see also Figure 6). We have run the tests with $h = 2^{-8}$, $\delta = h$ and $Tol = 10^{-6}$.

In Figure 3 we can compare the number of iterations for $\varepsilon = 1, 10^{-1}, \ldots, 10^{-10}$ and the CPU times spent by the iteration. It can be seen that both the number of iterations, denoted by a circle, and the work spent on the auxiliary systems, denoted by a triangle, are bounded independently of ε.

On the other hand, we note that by decreasing ε, there is an initial increasing iteration error phase until reaching near the uniform bound. This is explained by the "old bound" (3.8), since the latter together with the "new bound" (3.16) implies that the overall bound on the error is of the form $\min\{\text{const}_1,\ 1 + \frac{const_2}{\sqrt{\varepsilon}}\}$. That is, the error increases as long as ε decreases to about 10^{-6}, where the error reaches near the new bound and thus it becomes approximately constant as ε decreases further. This behavior is clarified by Figure 4. It not only shows that the actual iteration numbers remain below both the old and new bounds as predicted by theory but also that the old bound deteriorates to infinity as ε tends to zero.

We have also run the same test using the GCGLS(0) iteration instead of PCGN (cf. Remark 3.2). The computer work proved to be essentially the same as for the PCGN method: one iteration step took less time, but more iterations were necessary,

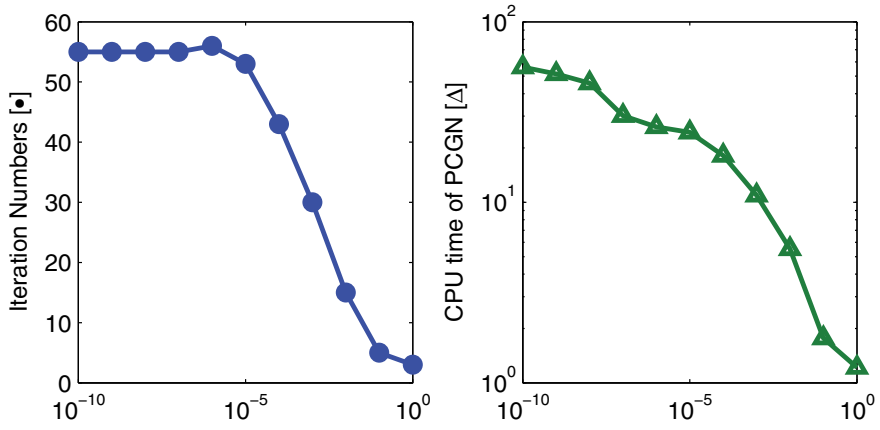

FIG. 3. *Iteration numbers and CPU times on the square.*

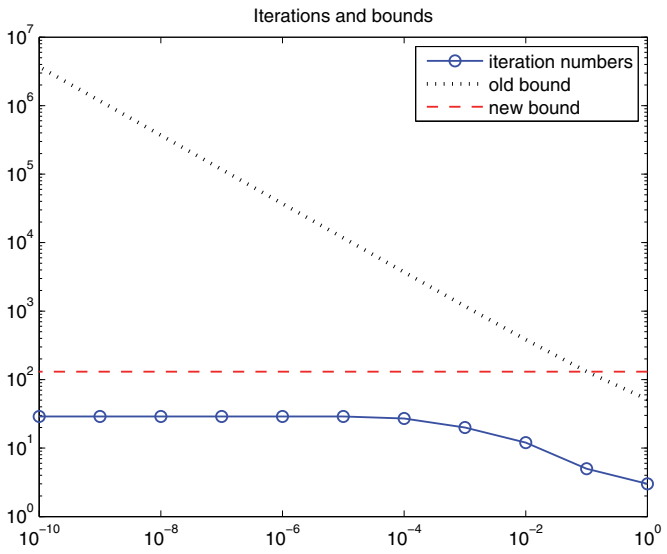

FIG. 4. *Actual iteration numbers versus the old and new bounds on the square.*

which led to a similar amount of total work; see Figure 5. As also discussed earlier in Remark 3.1, PCGN is thus a suitable method for the studied problem.

The numerical and exact solutions in the case $\varepsilon = 10^{-10}$ are plotted together with the distribution of error in Figure 6. The error, which is less than the tolerance 10^{-6}, comes essentially from the layer points; hence it could possibly be further decreased by involving adaptive mesh refinement as well.

5.2. Enclosed flow on a disc. Now let our domain Ω be the unit disc in $\mathbb{R}^2$ and the vector field be defined as the circular enclosed flow

$$\mathbf{w}(x, y) := (-y, x);$$

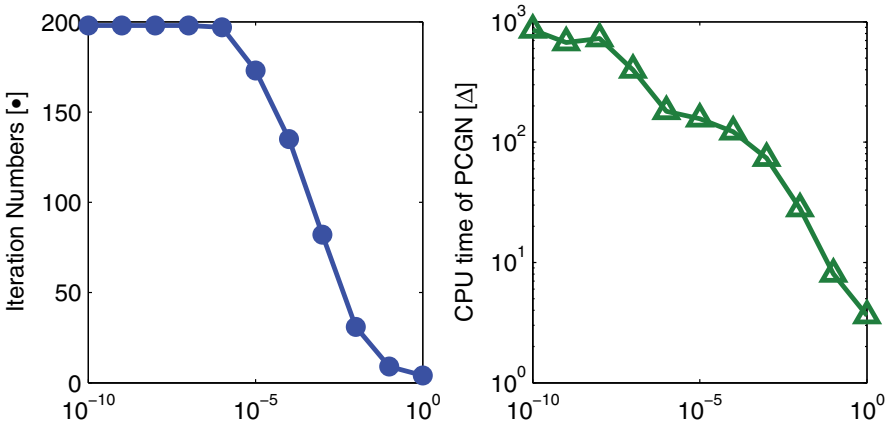

FIG. 5. *Iteration numbers and CPU times with GCGLS(0) on the square.*

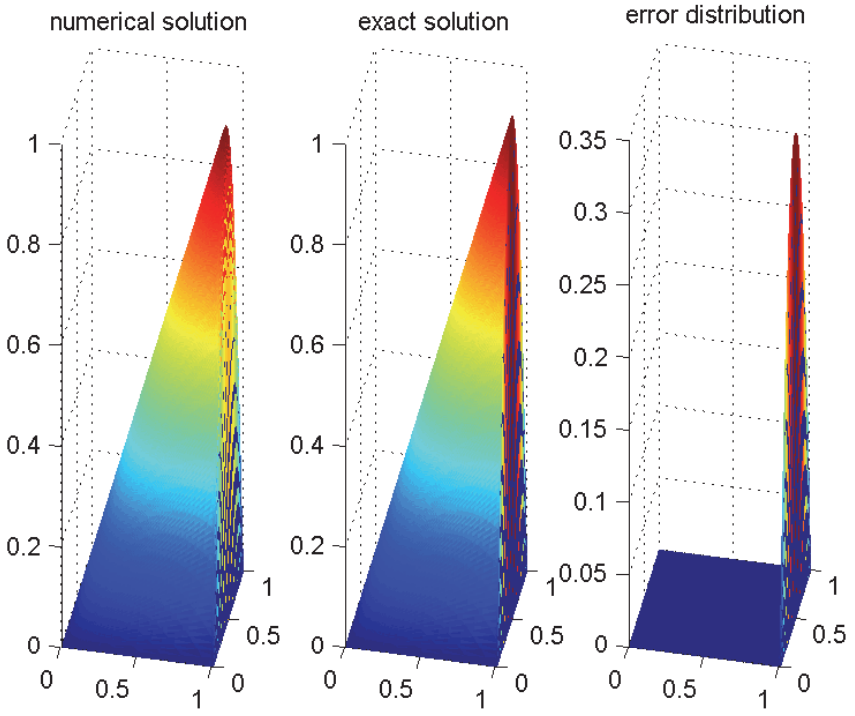

FIG. 6. *The numerical and exact solutions on the square.*

see Figure 2. The function g in (2.1) is chosen such that the exact solution of the problem is

$$u(x,y) = x^4\left(R^2 - \frac{e^{R^2/4\varepsilon}-1}{e^{1/4\varepsilon}-1}\right),$$

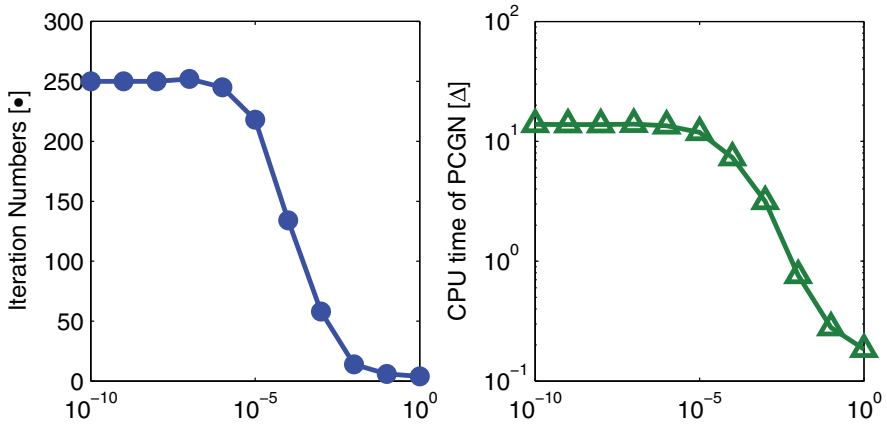

FIG. 7. *Iteration numbers and CPU times with PCGN on the disc.*

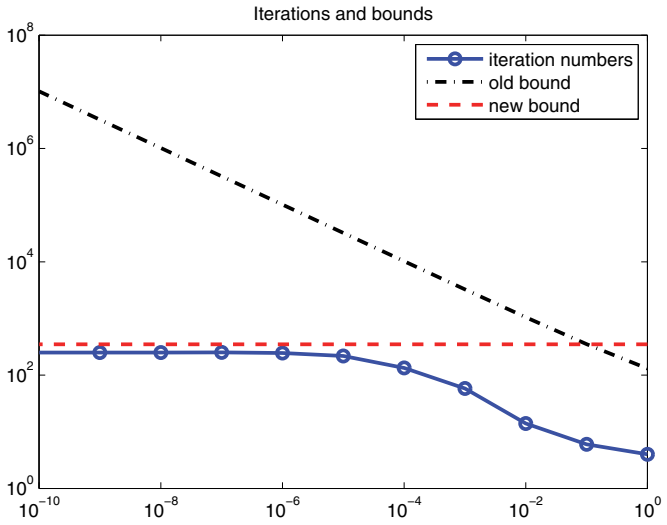

FIG. 8. *Actual iteration numbers versus the old and new bounds on the disc.*

where $R^2 = x^2 + y^2$. This solution exhibits a layer similar to (5.1) near the two opposite portions of the circle (see also Figure 9). We have run the test with $h = 2^{-9}$, $\delta = h$, and $Tol = 10^{-6}$.

In Figure 7 we can compare the number of PCGN iterations for $\varepsilon = 1, 10^{-1}, \ldots, 10^{-10}$ and the CPU times spent by the iteration. Similarly to the first test problem, both the number of iterations and the work spent on the auxiliary systems are bounded independently of ε.

Figure 8 is the counterpart of Figure 4, explaining the initial increasing iteration error phase until reaching near the uniform bound. Moreover, the actual iteration numbers now approach almost exactly the uniform bound as ε tends to zero, which suggests that our bound is close to sharp in this case.

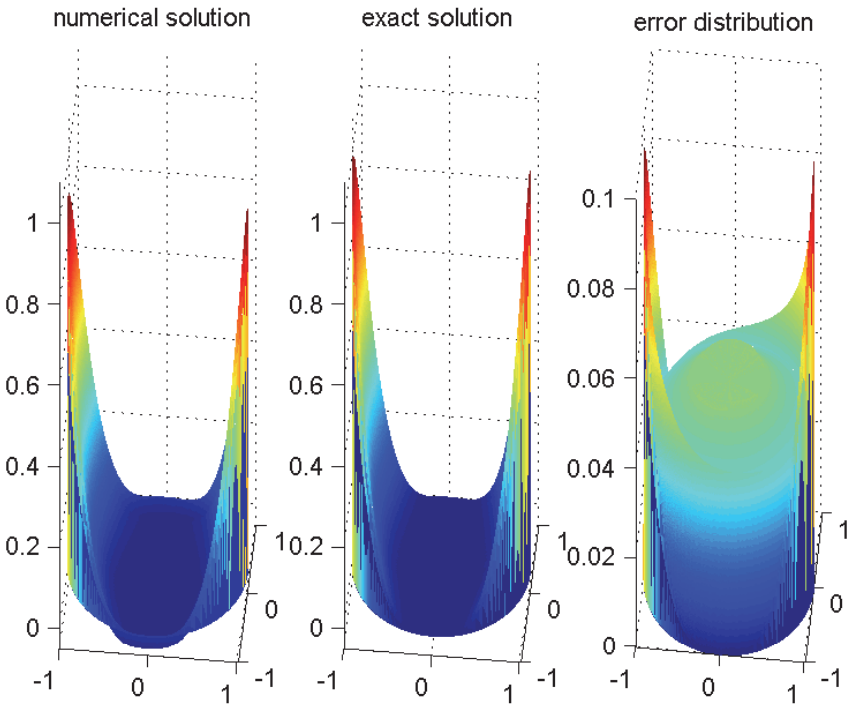

FIG. 9. *The numerical and exact solutions on the disc.*

Finally, the numerical and exact solutions in the case $\varepsilon = 10^{-6}$ are plotted together with the distribution of error in Figure 9, showing a similar behavior as for the first test problem.

Acknowledgment. The authors acknowledge the valuable comments of the anonymous referees, which helped to improve the paper.

REFERENCES

[1] V. I. ARNOLD, *Ordinary Differential Equations*, Springer, New York, 1992.
[2] O. AXELSSON, *A generalized conjugate gradient least square method*, Numer. Math., 51 (1987), pp. 209–227.
[3] O. AXELSSON, *Iterative Solution Methods,* Cambridge University Press, Cambridge, UK, 1994.
[4] O. AXELSSON, E. GLUSHKOV, AND N. GLUSHKOVA, *The local Green's function method in singularly perturbed convection-diffusion problems*, Math. Comp., 78 (2009), pp. 153–170.
[5] O. AXELSSON AND J. KARÁTSON, *Symmetric part preconditioning for the conjugate gradient method in Hilbert space*, Numer. Funct. Anal., 24 (2003), pp. 455–474.
[6] O. AXELSSON AND J. KARÁTSON, *Equivalent operator preconditioning for elliptic problems*, Numer. Algorithms, 50 (2009), pp. 297–380.
[7] O. AXELSSON AND P. S. VASSILEVSKI, *Algebraic multilevel preconditioning methods* I, Numer. Math., 56 (1989), pp. 157–177.
[8] O. AXELSSON AND P. S. VASSILEVSKI, *Algebraic multilevel preconditioning methods* II, SIAM J. Numer. Anal., 27 (1990), pp. 1569–1590.
[9] H. EGGER, *Preconditioning CGNE iteration for inverse problems*, Numer. Linear Algebra Appl., 14 (2007), pp. 183–196.

[10] S. C. Eisenstat, H. C. Elman, and M. H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 20 (1983), pp. 345–357.
[11] H. C. Elman and M. H. Schultz, *Preconditioning by fast direct methods for nonself-adjoint nonseparable elliptic equations*, SIAM J. Numer. Anal., 23 (1986), pp. 44–57.
[12] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*, Numer. Math. Sci. Comput., Oxford University Press, New York, 2005.
[13] V. Faber, T. Manteuffel, and S. V. Parter, *On the theory of equivalent operators and application to the numerical solution of uniformly elliptic partial differential equations*, Adv. in Appl. Math., 11 (1990), pp. 109–163.
[14] R. C. Kirby, *From functional analysis to iterative methods*, SIAM Rev., 52 (2010), pp. 269–293.
[15] K. W. Morton, *Numerical Solution of Convection-Diffusion Problems*, Appl. Math. Math. Comput. 12, Chapman & Hall, London, 1996.
[16] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations,* Springer-Verlag, Berlin, 1996.
[17] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[18] G. Starke, *Field-of-values analysis of preconditioned iterative methods for nonsymmetric elliptic problems*, Numer. Math., 78 (1997), pp. 103–117.
[19] M. Stynes, *Steady-state convection-diffusion problems*, Acta Numer., 14 (2005), pp. 445–508.
[20] P. S. Vassilevski, *Multilevel Block Factorization Preconditioners: Matrix-Based Analysis and Algorithms for Solving Finite Element Equations*, Springer, New York, 2008.
[21] O. Widlund, *A Lanczos method for a class of non-symmetric systems of linear equations*, SIAM J. Numer. Anal., 15 (1978), pp. 801–812.