A study on the conditioning of finite element equations with arbitrary anisotropic meshes via a density function approach

Lennard Kamenski[1], Weizhang Huang[2]

submitted: December 18, 2013

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
email: Lennard.Kamenski@wias-berlin.de

[2] The University of Kansas
Department of Mathematics
Lawrence, KS 66045
US
email: whuang@ku.edu

Abstract

The linear finite element approximation of a general linear diffusion problem with arbitrary anisotropic meshes is considered. The conditioning of the resultant stiffness matrix and the Jacobi preconditioned stiffness matrix is investigated using a density function approach proposed by Fried in 1973. It is shown that the approach can be made mathematically rigorous for general domains and used to develop bounds on the smallest eigenvalue and the condition number that are sharper than existing estimates in one and two dimensions and comparable in three and higher dimensions. The new results reveal that the mesh concentration near the boundary has less influence on the condition number than the mesh concentration in the interior of the domain. This is especially true for the Jacobi preconditioned system where the former has little or almost no influence on the condition number. Numerical examples are presented.

1 Introduction

Mesh adaptation is a common tool for use in the numerical solution of partial differential equations (PDEs) to enhance computational efficiency. It often results in nonuniform meshes whose elements vary significantly in size and shape from place to place on the physical domain. Nonuniform meshes could lead to ill-conditioned linear systems and their solution may deteriorate the efficiency of the entire computation. It is thus important in practice as well as in theory to understand how mesh nonuniformity affects the conditioning of linear systems resulting from discretization of PDEs on nonuniform meshes.

The issue has been studied by a number of researchers mostly for the linear finite element approximation of the Laplace operator or a general diffusion operator by developing bounds on the extremal eigenvalues on the resultant stiffness matrix, e.g., see [4, 6, 7, 13, 14, 15] for second-order elliptic PDEs or [2, 3, 9] for a more general setting of elliptic bilinear forms on Sobolev spaces of real index $m \in [-1, 1]$.

The estimation of the largest eigenvalue is well understood and it is easy to show that the largest eigenvalue is bounded by a multiple (with a constant depending on mesh connectivity) of the maximum of the largest eigenvalues of the local stiffness matrices [7]. Moreover, the largest diagonal entry of the stiffness matrix is a good estimate for the largest eigenvalue: it is tight within a factor of $d + 1$ for any mesh, where d is the dimension of the physical domain [13]. Sharp estimates in terms of mesh geometry are available for both isotropic [3, 6, 9, 14] and anisotropic [13] diffusion.

The estimation of the smallest eigenvalue is more challenging. Currently there are two approaches for this purpose. The first approach utilizes Sobolev's inequality and was first used by Bank and

Scott [4] for the Laplace operator with isotropic meshes in $d \geq 2$ dimensions. They developed a lower bound on the smallest eigenvalue of a diagonally scaled stiffness matrix and showed that the condition number of the scaled stiffness matrix is comparable to that with a uniform mesh. A similar result for elliptic bilinear forms on Sobolev spaces of real index $m \in [-1, 1]$ with shape-regular meshes in $d \geq 2$ dimensions was derived by Ainsworth, McLean, and Tran [2, 3]. Their result was later generalized to locally quasi-uniform meshes[1] in $d \geq 3$ dimensions by Graham and McLean [9]. Recently, Kamenski, Huang, and Xu [13] derived a similar bound for second-order elliptic PDEs which is valid for arbitrary meshes (i.e., without imposing any conditions on the mesh regularity) for any d; the established bound for the condition number depends on three factors, one representing the condition number of the linear finite element equations for the Laplace operator on a uniform mesh and the other factors arising from the nonuniformity of the mesh viewed in the metric specified by the inverse of the diffusion matrix $\mathbb{D}$ ($\mathbb{D}^{-1}$-*nonuniformity*) and the mesh nonuniformity in volume measured in the Euclidean metric (*volume-nonuniformity*). Further, it was shown in [13] that the Jacobi preconditioning — an optimal diagonal scaling for a symmetric positive definite sparse matrix — eliminates the effect of the mesh volume-nonuniformity and reduces the effect of the mesh $\mathbb{D}^{-1}$-nonuniformity. This result can be seen as a further generalization of [3, 4, 9] towards arbitrary anisotropic meshes and general diffusion coefficients.

In the second approach (hereafter referred to as *the density function approach*), a lower bound on the smallest eigenvalue of the stiffness matrix is obtained through a lower bound of the smallest eigenvalue of a continuous generalized eigenvalue problem involving an auxiliary density function. It was first employed by Fried [7] for the Laplace operator. Fried showed how to obtain the lower bound for the continuous problem for spherical domains and claimed a physical argument that the bound is also valid for general domains. The obtained bound for the smallest eigenvalue of the stiffness matrix is valid for any mesh in any dimension but in three and higher dimensions it is less sharp than bounds obtained in [9, 13]: Fried's bound is proportional to power $(d - 2)/d$ of the volume of the smallest mesh element whereas those in [9, 13] are proportional to an average of the volumes of all elements (cf. (31) and (35)).

In this paper we investigate the density function approach and develop bounds on the smallest eigenvalue and the condition number of both the stiffness matrix and the Jacobi preconditioned stiffness matrix for the linear finite element approximation of a general diffusion problem with arbitrary nonuniform meshes. We show that the approach yields a bound on the smallest eigenvalue that is much sharper than in the original work by Fried [7]. Moreover, the new results are even sharper than those obtained in [13] in one and two dimensions and comparable in three and higher dimensions. In particular, they involve a factor describing the distance of a given element to the boundary and reflect the fact that the mesh concentration near the boundary has less influence on the condition number than the mesh concentration in the interior of the domain. This is especially true for the Jacobi preconditioned system where the former has little or almost no influence on the condition number.

The outline of the paper is as follows. The linear finite element approximation of the general diffusion problem is described in Sect. 2. The lower bound for the smallest eigenvalue of the stiffness matrix is developed in Sect. 3, followed by the derivation of the bounds on the condition

---

[1] They are meshes where neighboring elements always comparable size and shape.

number in Sect. 4. Section 4 also contains comparison of the obtained bounds to those in [7] and [13]. Numerical examples are presented in Sect. 5. The conclusions are drawn in Sect. 6.

2 Finite element approximation

We consider the boundary value problem (BVP) of a prototype anisotropic diffusion equation in the form

$$
\begin{cases}
-\nabla \cdot (\mathbb{D}\nabla u) = f, & \text{in } \Omega, \\
u = 0, & \text{on } \partial\Omega,
\end{cases}
\tag{1}
$$

where Ω is a simply connected polygonal or polyhedral domain in $\mathbb{R}^d$ ($d \geq 1$) and $\mathbb{D} = \mathbb{D}(\boldsymbol{x})$ is the diffusion matrix. We assume that $\mathbb{D}$ is symmetric and positive definite and there exist two positive constants $d_{\min}$ and $d_{\max}$ such that

$$
d_{\min}I \leq \mathbb{D}(\boldsymbol{x}) \leq d_{\max}I, \qquad \forall \boldsymbol{x} \in \Omega,
\tag{2}
$$

where I is the $d \times d$ identity matrix and the less-than-or-equal sign means that the difference between the right-hand side and left-hand side terms is positive semidefinite. The weak formulation of the BVP (1) is to find $u \in V \equiv H_0^1(\Omega)$ such that

$$
\mathcal{A}(u, v) = \mathcal{F}(v), \qquad \forall v \in V,
\tag{3}
$$

where

$$
\mathcal{A}(u, v) = \int_\Omega \nabla v \cdot \mathbb{D}\nabla u \, d\boldsymbol{x} \qquad \text{and} \qquad \mathcal{F}(v) = \int_\Omega f v \, d\boldsymbol{x}.
$$

Let an affine family of simplicial meshes $\{\mathcal{T}_h\}$ for Ω be given. For any element $K \in \mathcal{T}_h$, let $F_K \colon \hat{K} \to K$ be the affine mapping from the reference element $\hat{K}$ to the mesh element K and F_K' the Jacobian matrix of F_K. For notational simplicity, we assume that $|\hat{K}| = 1$. Note that F_K' is constant on K.

Denote the linear finite element space associated with mesh $\mathcal{T}_h$ by $V^h \subset V$. Then, a linear finite element solution $u_h \in V^h$ of BVP (1) is defined by

$$
\mathcal{A}(u_h, v_h) = \mathcal{F}(v_h), \qquad \forall v_h \in V^h.
\tag{4}
$$

Let N and N_{vi} be the numbers of the elements and the interior vertices of $\mathcal{T}_h$, respectively. If we order the vertices in such a way that the first N_{vi} vertices are interior vertices and denote by ϕ_j the standard linear basis function associated with the jth vertex, then we can express V^h and u_h as

$$
V^h = \text{span}\{\phi_1, \ldots, \phi_{N_{vi}}\} \qquad \text{and} \qquad u_h = \sum_{j=1}^{N_{vi}} u_j \phi_j.
\tag{5}
$$

In the following we will use the function form $u_h = \sum_j u_j \phi_j$ and the vector form $\boldsymbol{u} = [u_1, \ldots, u_{N_{vi}}]^T$ synonymously. Using this, we can write (4) in a matrix form as

$$
A\boldsymbol{u} = \boldsymbol{f},
$$

3

where the stiffness matrix A and the right-hand side term $\boldsymbol{f}$ are given by

$$A_{ij} = \mathcal{A}(\phi_j, \phi_i) = \sum_{K \in \mathcal{T}_h} |K| \, \nabla \phi_i \cdot \mathbb{D}_K \nabla \phi_j, \qquad i, j = 1, \ldots, N_{vi}, \qquad (6)$$

$$f_i = \sum_{K \in \mathcal{T}_h} \int_K f \phi_i \, d\boldsymbol{x}, \qquad\qquad i = 1, \ldots, N_{vi},$$

and $\mathbb{D}_K$ is the average of $\mathbb{D}(\boldsymbol{x})$ over K, i.e.,

$$\mathbb{D}_K = \frac{1}{|K|} \int_K \mathbb{D}(\boldsymbol{x}) \, d\boldsymbol{x}.$$

We are interested in the condition numbers of matrices A and $S^{-1}AS^{-1}$, where $S = \sqrt{\mathrm{diag}(A)}$ is the Jacobi preconditioner.

3   Smallest eigenvalue of the stiffness matrix

In this section we develop lower bounds for $\lambda_{\min}(A)$ and $\lambda_{\min}(S^{-1}AS^{-1})$ using the density function approach proposed by Fried [7], where $S = \sqrt{\mathrm{diag}(A)}$ is the Jacobi preconditioner. The approach utilizes an auxiliary eigenvalue problem

$$\begin{cases} -\Delta u = \lambda \rho u, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \qquad (7)$$

where ρ is a density function (to be chosen) satisfying

$$0 < \rho_{\min} \le \rho \le \rho_{\max} < \infty \qquad \text{and} \qquad \int_\Omega \rho(\boldsymbol{x}) \, d\boldsymbol{x} = 1. \qquad (8)$$

The corresponding Galerkin formulation reads as

$$(\nabla u, \nabla v) = \lambda(\rho u, v), \qquad \forall v \in H_0^1(\Omega).$$

Let λ_ρ be the smallest eigenvalue of this eigenvalue problem and $\lambda_{\min}(B_\rho)$ the smallest eigenvalue of the Galerkin mass matrix B_ρ associated with the density ρ. Then, for any vector $\boldsymbol{u}$,

$$\begin{aligned} \boldsymbol{u}^T A \boldsymbol{u} &= (\mathbb{D}\nabla u_h, \nabla u_h) \\ &\ge d_{\min}(\nabla u_h, \nabla u_h) \\ &\ge d_{\min}\lambda_\rho \, (\rho u_h, u_h) \\ &\ge d_{\min}\lambda_\rho \lambda_{\min}(B_\rho) \, \|\boldsymbol{u}\|_2^2, \end{aligned}$$

which leads to

$$\lambda_{\min}(A) \ge d_{\min}\lambda_\rho \lambda_{\min}(B_\rho). \qquad (9)$$

Thus, the key to the approach is to estimate λ_ρ and choose the density function so that the lower bound (9) is as large as possible.

4

Hereafter, we assume that ρ is piecewise constant, i.e., $\rho|_K = \rho_K = const$ for all $K \in \mathcal{T}_h$. We denote

$$|K|_\rho = \rho_K |K|, \qquad |\omega_j|_\rho = \sum_{K \in \omega_j} |K|_\rho, \qquad |\omega_{\min}|_\rho = \min_{j=1,\dots,N_{vi}} |\omega_j|_\rho.$$

It is known [13, Sect. 3] that

$$\lambda_{\min}(B_\rho) \geq \frac{|\omega_{\min}|_\rho}{(d+1)(d+2)}. \tag{10}$$

Inserting this into (9) yields

$$\lambda_{\min}(A) \geq \frac{d_{\min} |\omega_{\min}|_\rho}{(d+1)(d+2)} \lambda_\rho. \tag{11}$$

3.1   Smallest eigenvalue of the continuous eigenvalue problem

We now derive a lower bound on λ_ρ by means of a Green's function. The derivation consists of three lemmas, with the bound being given in Lemma 3.3.

First, we need the property of the strict positivity (or negativity) of eigenfunctions associated with the smallest eigenvalue of (7).

Lemma 3.1. *For any density distribution ρ satisfying (8), the smallest eigenvalue λ_ρ of the eigenvalue problem (7) is simple and positive. Any corresponding eigenfunction is either strictly positive or strictly negative in Ω.*

Proof. The positiveness of λ_ρ follows from

$$\lambda_\rho = \min_{u \neq 0} \frac{(\nabla u, \nabla u)}{(\rho u, u)} \geq \min_{u \neq 0} \frac{C_P \|u\|_2^2}{\rho_{\max} \|u\|_2^2} = \frac{C_P}{\rho_{\max}} > 0,$$

where C_P is the constant associated with Poincaré's inequality. The rest of the proof is the same as for the standard eigenvalue problem with $\rho \equiv 1$ (e.g., see [8, Theorem 8.38]), except that $\frac{(\nabla u, \nabla u)}{(u, u)}$ is replaced by $\frac{(\nabla u, \nabla u)}{(\rho u, u)}$. □

Lemma 3.2. *Let $G(\boldsymbol{x}, \boldsymbol{\xi})$ be the Green's function for $-\Delta$ subject to a homogeneous Dirichlet boundary condition on Ω and $d(\boldsymbol{x})$ the distance from $\boldsymbol{x}$ to the boundary $\partial\Omega$, i.e., $d(\boldsymbol{x}) = \min_{\boldsymbol{y} \in \partial\Omega} |\boldsymbol{x} - \boldsymbol{y}|$. Then,*

$$0 \leq G(\boldsymbol{x}, \boldsymbol{\xi}) \leq C \times \begin{cases} \sqrt{d(\boldsymbol{x})d(\boldsymbol{\xi})}, & \text{for } d = 1, \\ \ln\left(1 + \frac{d(\boldsymbol{x})d(\boldsymbol{\xi})}{|\boldsymbol{x}-\boldsymbol{\xi}|^2}\right), & \text{for } d = 2, \\ |\boldsymbol{x} - \boldsymbol{\xi}|^{2-d} \min\left\{1, \frac{d(\boldsymbol{x})d(\boldsymbol{\xi})}{|\boldsymbol{x}-\boldsymbol{\xi}|^2}\right\}, & \text{for } d \geq 3. \end{cases} \tag{12}$$

Proof. For example, see [1, 5, 10, 12]. □

5

Lemma 3.3. *Assume that ρ is piecewise constant and $p \in \left(1, \frac{d}{d-2}\right)$. Let $d_K = \max_{\boldsymbol{x} \in K} d(\boldsymbol{x})$. Then the smallest eigenvalue λ_ρ of the problem (7) is bounded from below by*

$$
\lambda_\rho \geq C \times \begin{cases} \left(\sum_K |K|_\rho \, d_K\right)^{-1}, & \text{for } d = 1, \\[2ex] \left(1 + \sum_K |K|_\rho \ln^2\left(1 + d_K \rho_{\max}\right)\right)^{-\frac{1}{2}}, & \text{for } d = 2, \quad (13) \\[2ex] \left(\frac{d}{d-2} - p\right)^{\frac{d}{d+2p}} \left(\sum_K |K|_\rho^{\frac{p}{p-1}} |K|^{-\frac{1}{p-1}} d_K^{\frac{p}{p-1} \cdot \frac{d-(d-2)p}{d+2p}}\right)^{-\frac{p-1}{p}}, & \text{for } d \geq 3. \end{cases}
$$

Proof. For a positive eigenfunction u_ρ associated with λ_ρ we have

$$
u_\rho(\boldsymbol{x}) = \lambda_\rho \int_\Omega G(\boldsymbol{x}, \boldsymbol{\xi}) \rho(\boldsymbol{\xi}) u_\rho(\boldsymbol{\xi}) \, d\boldsymbol{\xi}. \tag{14}
$$

For $1 \leq d \leq 2$, applying the Cauchy-Schwarz inequality to (14),

$$
u_\rho(\boldsymbol{x}) \leq \lambda_\rho \left(\int_\Omega \rho(\boldsymbol{\xi}) G(\boldsymbol{x}, \boldsymbol{\xi})^2 \, d\boldsymbol{\xi}\right)^{\frac{1}{2}} \left(\int_\Omega \rho(\boldsymbol{\xi}) u_\rho(\boldsymbol{\xi})^2 \, d\boldsymbol{\xi}\right)^{\frac{1}{2}}.
$$

Taking square, multiplying with $\rho(\boldsymbol{x})$, and integrating both sides over Ω gives

$$
\int_\Omega \rho(\boldsymbol{x}) u_\rho(\boldsymbol{x})^2 \, d\boldsymbol{x} \leq \lambda_\rho^2 \int_\Omega \rho(\boldsymbol{x}) \left[\int_\Omega \rho(\boldsymbol{\xi}) G(\boldsymbol{x}, \boldsymbol{\xi})^2 \, d\boldsymbol{\xi}\right] d\boldsymbol{x} \cdot \int_\Omega \rho(\boldsymbol{\xi}) u_\rho(\boldsymbol{\xi})^2 \, d\boldsymbol{\xi}.
$$

This yields

$$
\lambda_\rho \geq \left(\int_\Omega \rho(\boldsymbol{x}) \left[\int_\Omega \rho(\boldsymbol{\xi}) G(\boldsymbol{x}, \boldsymbol{\xi})^2 \, d\boldsymbol{\xi}\right] d\boldsymbol{x}\right)^{-\frac{1}{2}}. \tag{15}
$$

For $d = 1$, using this and Lemma 3.2,

$$
\lambda_\rho \geq C \left(\int_\Omega \rho(\boldsymbol{x}) \left[\int_\Omega \rho(\boldsymbol{\xi}) d(\boldsymbol{x}) d(\boldsymbol{\xi}) \, d\boldsymbol{\xi}\right] d\boldsymbol{x}\right)^{-\frac{1}{2}} = C \left(\int_\Omega \rho(\boldsymbol{x}) d(\boldsymbol{x}) \, d\boldsymbol{x}\right)^{-1}
$$

$$
\geq C \left(\sum_K |K|_\rho \, d_K\right)^{-1},
$$

which gives (13) for $d = 1$.

In 2D, from Lemma 3.2 we have

$$
\int_\Omega \rho(\boldsymbol{\xi}) G(\boldsymbol{x}, \boldsymbol{\xi})^2 \, d\boldsymbol{\xi} \leq C \int_\Omega \rho(\boldsymbol{\xi}) \ln^2\left(1 + \frac{d(\boldsymbol{x}) d(\boldsymbol{\xi})}{|\boldsymbol{x} - \boldsymbol{\xi}|^2}\right) d\boldsymbol{\xi}
$$

$$
\leq C \int_\Omega \rho(\boldsymbol{\xi}) \ln^2\left(1 + \frac{d(\boldsymbol{x}) h_\Omega}{|\boldsymbol{x} - \boldsymbol{\xi}|^2}\right) d\boldsymbol{\xi},
$$

6

where h_Ω denotes the diameter of Ω. Let $B_\varepsilon(\boldsymbol{x})$ be the disk with center $\boldsymbol{x}$ and radius ε. Then, using property (8) we have

$$
\begin{aligned}
\int_\Omega \rho(\boldsymbol{\xi}) G(\boldsymbol{x},\boldsymbol{\xi})^2 \, d\boldsymbol{\xi} &\leq C \int_{\Omega\setminus B_\varepsilon(\boldsymbol{x})} \rho(\boldsymbol{\xi}) \ln^2\left(1 + \frac{d(\boldsymbol{x})h_\Omega}{\varepsilon^2}\right) d\boldsymbol{\xi} \\
&\quad + C \int_{B_\varepsilon(\boldsymbol{x})} \rho_{\max} \ln^2\left(1 + \frac{d(\boldsymbol{x})h_\Omega}{|\boldsymbol{x}-\boldsymbol{\xi}|^2}\right) d\boldsymbol{\xi} \\
&\leq C \ln^2\left(1 + \frac{d(\boldsymbol{x})h_\Omega}{\varepsilon^2}\right) + C\rho_{\max} \int_0^\varepsilon r \ln^2\left(1 + \frac{d(\boldsymbol{x})h_\Omega}{r^2}\right) dr \\
&\leq C \ln^2\left(1 + \frac{d(\boldsymbol{x})h_\Omega}{\varepsilon^2}\right) + C\rho_{\max}\varepsilon^2 \left[1 + \ln^2\left(1 + \frac{d(\boldsymbol{x})h_\Omega}{\varepsilon^2}\right)\right].
\end{aligned}
$$

Taking $\varepsilon = \rho_{\max}^{-\frac{1}{2}}$, we get

$$
\int_\Omega \rho(\boldsymbol{\xi}) G(\boldsymbol{x},\boldsymbol{\xi})^2 \, d\boldsymbol{\xi} \leq C\left[1 + \ln^2\left(1 + d(\boldsymbol{x})h_\Omega\rho_{\max}\right)\right].
$$

Inserting this into (15), using (8) and the fact that $d(\boldsymbol{x}) \leq d_K$ for all $\boldsymbol{x} \in K$, we have

$$
\lambda_\rho \geq C\left(1 + \sum_K |K|_\rho \ln^2\left(1 + d_K h_\Omega \rho_{\max}\right)\right)^{-\frac{1}{2}}.
$$

This gives (13) for the 2D case upon absorbing h_Ω in the generic constant C.

For $d \geq 3$, we need a slightly different version of the estimate (15). In this case, Lemma 3.2 implies that $G(\boldsymbol{x},\cdot) \in L^p(\Omega)$ for any $\boldsymbol{x} \in \Omega$ and $p \in \left(1, \frac{d}{d-2}\right)$. Using Hölder's inequality,

$$
u_\rho(\boldsymbol{x}) \leq \lambda_\rho \left(\int_\Omega G(\boldsymbol{x},\boldsymbol{\xi})^p \, d\boldsymbol{\xi}\right)^{\frac{1}{p}} \left(\int_\Omega \rho(\boldsymbol{\xi})^q u_\rho(\boldsymbol{\xi})^q \, d\boldsymbol{\xi}\right)^{\frac{1}{q}},
$$

where q satisfies $\frac{1}{p} + \frac{1}{q} = 1$. By multiplying with $\rho(\boldsymbol{x})$, taking power q, and integrating both sides over Ω, we get

$$
\int_\Omega \rho(\boldsymbol{x})^q u_\rho(\boldsymbol{x})^q \, d\boldsymbol{x} \leq \lambda_\rho^q \int_\Omega \rho(\boldsymbol{x})^q \left(\int_\Omega G(\boldsymbol{x},\boldsymbol{\xi})^p \, d\boldsymbol{\xi}\right)^{\frac{q}{p}} d\boldsymbol{x} \int_\Omega \rho(\boldsymbol{\xi})^q u_\rho(\boldsymbol{\xi})^q \, d\boldsymbol{\xi}
$$

and therefore

$$
\lambda_\rho \geq \left(\int_\Omega \rho(\boldsymbol{x})^q \left(\int_\Omega G(\boldsymbol{x},\boldsymbol{\xi})^p \, d\boldsymbol{\xi}\right)^{\frac{q}{p}} d\boldsymbol{x}\right)^{-\frac{1}{q}}. \tag{16}
$$

With Lemma 3.2 we have, using a similar strategy as in 2D,

$$
\int_\Omega G(\boldsymbol{x}, \boldsymbol{\xi})^p \, d\boldsymbol{\xi} \le C \int_\Omega |\boldsymbol{x} - \boldsymbol{\xi}|^{(2-d)p} \min \left\{ 1, \frac{d(\boldsymbol{x})^p h_\Omega^p}{|\boldsymbol{x} - \boldsymbol{\xi}|^{2p}} \right\} d\boldsymbol{\xi}
$$

$$
\le C \int_{\Omega \cap B_\varepsilon(\boldsymbol{x})} |\boldsymbol{x} - \boldsymbol{\xi}|^{(2-d)p} \, d\boldsymbol{\xi} + C \int_{\Omega \setminus B_\varepsilon(\boldsymbol{x})} |\boldsymbol{x} - \boldsymbol{\xi}|^{(2-d)p} \frac{d(\boldsymbol{x})^p h_\Omega^p}{|\boldsymbol{x} - \boldsymbol{\xi}|^{2p}} \, d\boldsymbol{\xi}
$$

$$
\le C \int_0^\varepsilon r^{d-1} r^{(2-d)p} \, dr + C d(\boldsymbol{x})^p h_\Omega^p \, \varepsilon^{-dp}
$$

$$
\le C \frac{1}{d - (d-2)p} \varepsilon^{d-(d-2)p} + C d(\boldsymbol{x})^p h_\Omega^p \, \varepsilon^{-dp}.
$$

Choosing ε such that the terms on the right-hand side are equal, we get

$$
\varepsilon = \left((d - (d-2)p) \, d(\boldsymbol{x})^p h_\Omega^p \right)^{\frac{1}{d+2p}}
$$

and

$$
\int_\Omega G(\boldsymbol{x}, \boldsymbol{\xi})^p \, d\boldsymbol{\xi} \le C(d - (d-2)p)^{-\frac{dp}{d+2p}} d(\boldsymbol{x})^{p \cdot \frac{d-(d-2)p}{d+2p}},
$$

where h_Ω has been absorbed into the generic constant C. Inserting this into (16) leads to

$$
\lambda_\rho \ge C \left(\int_\Omega \rho(\boldsymbol{x})^q (d - (d-2)p)^{-\frac{dq}{d+2p}} d(\boldsymbol{x})^{\frac{q(d-(d-2)p)}{d+2p}} \, d\boldsymbol{x} \right)^{-\frac{1}{q}}
$$

$$
\ge C(d - (d-2)p)^{\frac{d}{d+2p}} \left(\int_\Omega \rho(\boldsymbol{x})^q d(\boldsymbol{x})^{\frac{q(d-(d-2)p)}{d+2p}} \, d\boldsymbol{x} \right)^{-\frac{1}{q}}
$$

$$
\ge C(d - (d-2)p)^{\frac{d}{d+2p}} \left(\sum_K \rho_K^q \, |K| \, d_K^{\frac{q(d-(d-2)p)}{d+2p}} \right)^{-\frac{1}{q}}.
$$

Since $q = p/(p-1)$ and ρ is element-wise constant, this gives (13) for $d \ge 3$. $\qquad\square$

3.2 Smallest eigenvalue of the stiffness matrix

Having established a lower bound on λ_ρ, we can now proceed with the estimation of $\lambda_{\min}(A)$ and $\lambda_{\min}(S^{-1}AS^{-1})$. Combining (11) with Lemma 3.3, we have

$$
\lambda_{\min}(A) \ge C d_{\min} |\omega_{\min}|_\rho
$$

$$
\times \begin{cases}
\left(\sum_K |K|_\rho \, d_K \right)^{-1}, & \text{for } d = 1, \\[2ex]
\left(1 + \sum_K |K|_\rho \ln^2 \left(1 + d_K \rho_{\max} \right) \right)^{-\frac{1}{2}}, & \text{for } d = 2, \quad (17) \\[2ex]
\left(\frac{d}{d-2} - p \right)^{\frac{d}{d+2p}} \left(\sum_K |K|_\rho^{\frac{p}{p-1}} \, |K|^{-\frac{1}{p-1}} \, d_K^{\frac{p}{p-1} \cdot \frac{d-(d-2)p}{d+2p}} \right)^{-\frac{p-1}{p}}, & \text{for } d \ge 3.
\end{cases}
$$

8

The density function ρ is arbitrary so far. The optimal ρ is such that the right-hand side term of (17) attains the maximum value. It is difficult, if not impossible, to find the optimal ρ in general. We follow Fried [7] to choose ρ such that

$$|K|_\rho = const, \quad \forall K \in \mathcal{T}_h. \tag{18}$$

This gives

$$\rho(\boldsymbol{x})|_K = \frac{1}{N\,|K|} \tag{19}$$

and

$$|K|_\rho = \frac{1}{N} \quad \text{and} \quad |\omega_{\min}|_\rho \geq p_{\min}\,|K_{\min}|_\rho = \frac{p_{\min}}{N} \geq \frac{1}{N}, \tag{20}$$

where $p_{\min}$ is the minimum number of elements in a mesh patch. Inserting this into (17), using $|\bar{K}| = |\Omega|/N$ (the average element volume) and $\rho_{\max} = 1/(N\,|K_{\min}|)$, we arrive at the following estimate.

Lemma 3.4. *The smallest eigenvalue of the stiffness matrix A for the linear finite element approximation of BVP* (1) *is bounded from below by*

$$\lambda_{\min}(A) \geq \frac{Cd_{\min}}{N} \times \begin{cases} \left(\frac{1}{N}\sum_K d_K\right)^{-1}, & \text{for } d = 1, \\[2ex] \left(1 + \frac{1}{N}\sum_K \ln^2\left(1 + \frac{|\bar{K}|}{|K_{\min}|}d_K\right)\right)^{-\frac{1}{2}}, & \text{for } d = 2, \\[2ex] \left(\frac{d}{d-2} - p\right)^{\frac{d}{d+2p}}\left(\frac{1}{N}\sum_K\left(\frac{|\bar{K}|}{|K|}\right)^{\frac{1}{p-1}}d_K^{\frac{p}{p-1}\cdot\frac{d-(d-2)p}{d+2p}}\right)^{-\frac{p-1}{p}}, & \text{for } d \geq 3. \end{cases} \tag{21}$$

Remark 3.1 (Optimality of the density function). For $d \neq 2$, the choice of ρ in (19) is at least optimal for element-wise constant density functions. For $d = 1$, from (17) we have

$$\lambda_{\min}(A) \geq Cd_{\min}\frac{|\omega_{\min}|_\rho}{\sum_K |K|_\rho\,d_K} \geq Cd_{\min}\frac{p_{\min}\,|K_{\min}|_\rho}{\sum_K |K|_\rho\,d_K} = Cd_{\min}\frac{p_{\min}}{\sum_K \frac{|K|_\rho}{|K_{\min}|_\rho}d_K}.$$

Maximizing the bound on the right-hand side is equivalent to minimizing the sum $\sum_K \frac{|K|_\rho}{|K_{\min}|_\rho}d_K$. Since $\frac{|K|_\rho}{|K_{\min}|_\rho} \geq 1$ and $d_K > 0$ for all K, the optimal choice of ρ_K is such that $\frac{|K|_\rho}{|K_{\min}|_\rho} = 1$ for all K, which is equivalent to (18). The case of $d \geq 3$ is similar.

For the scaled stiffness matrix we have the following result.

Lemma 3.5. *The smallest eigenvalue of the Jacobi preconditioned stiffness matrix $S^{-1}AS^{-1}$ is*

bounded from below by

$$\lambda_{\min}(S^{-1}AS^{-1}) \geq \frac{C}{N^{\frac{2}{d}}}$$

$$\times \begin{cases} \left(\frac{1}{N^2}\sum_K |K|\,\beta_K d_K\right)^{-1}, & \text{for } d=1, \\[2ex] \left(\frac{1}{N}\sum_K |K|\,\beta_K\right)^{-\frac{1}{2}}\left(\frac{1}{N}\sum_K |K|\,\beta_K\left[1+\ln^2\left(1+d_K\gamma_h\right)\right]\right)^{-\frac{1}{2}}, & \text{for } d=2, \quad (22) \\[2ex] \left(\frac{d}{d-2}-p\right)^{\frac{d}{d+2p}}\left(\frac{1}{N^{\frac{2p}{d(p-1)}}}\sum_K |K|\,\beta_K^{\frac{p}{p-1}} d_K^{\frac{p}{p-1}\cdot\frac{d-(d-2)p}{d+2p}}\right)^{-\frac{p-1}{p}}, & \text{for } d\geq 3, \end{cases}$$

where β_K and γ_K, defined as

$$\beta_K = \frac{1}{d_{\min}}\left\|(F'_K)^{-1}\mathbb{D}_K(F'_K)^{-T}\right\|_2, \quad \forall K\in\mathcal{T}_h, \tag{23}$$

$$\gamma_h = \frac{\max_K \beta_K}{\sum_K |K|\,\beta_K} \tag{24}$$

reflect the non-uniformity of the mesh viewed in the metric specified by $\mathbb{D}^{-1}$ (see also Remark 4.3).

Proof. As for the non-scaled case we have (cf. (9) and (10))

$$\lambda_{\min}(S^{-1}B_\rho S^{-1}) \geq C\min_j \frac{|\omega_j|_\rho}{A_{jj}}$$

and therefore

$$\lambda_{\min}(S^{-1}AS^{-1}) \geq Cd_{\min}\lambda_\rho \min_j \frac{|\omega_j|_\rho}{A_{jj}}. \tag{25}$$

It is known [13] that

$$A_{jj} \leq C\sum_{K\in\omega_j} |K|\cdot\left\|(F'_K)^{-1}\mathbb{D}_K(F'_K)^{-T}\right\|_2 = Cd_{\min}\sum_{K\in\omega_j} |K|\,\beta_K, \quad j=1,\ldots,N_{vi}, \tag{26}$$

where C is a constant depending only on $\hat{K}$ and the linear basis functions on $\hat{K}$. Thus,

$$\min_j \frac{|\omega_j|_\rho}{A_{jj}} \geq \frac{C}{d_{\min}}\min_j \frac{\sum_{K\in\omega_j} |K|\,\rho_K}{\sum_{K\in\omega_j} |K|\,\beta_K}.$$

Once again, we choose ρ to get rid of the minimum sign[2], viz.,

$$\frac{\sum_{K\in\omega_j} |K|\,\rho_K}{\sum_{K\in\omega_j} |K|\,\beta_K} = const \quad \forall j.$$

---

[2]Although it is not quite clear if this choice is optimal, it is the best choice we could find so far.

10

This and (8) lead to

$$\rho(\boldsymbol{x})|_K = \frac{\beta_K}{\displaystyle\sum_{\tilde{K}\in\mathcal{T}_h} |\tilde{K}|\beta_{\tilde{K}}}$$

and

$$\min_j \frac{\sum_{K\in\omega_j} |K|\,\rho_K}{\sum_{K\in\omega_j} |K|\,\beta_K} = \frac{1}{\displaystyle\sum_{\tilde{K}\in\mathcal{T}_h} |\tilde{K}|\beta_{\tilde{K}}}.$$

The estimate (22) follows from this, inequality (25) and Lemma 3.3. □

4 Condition numbers of the stiffness matrix and the diagonally scaled stiffness matrix

We first quote an estimate on the maximum eigenvalues of A and $S^{-1}AS^{-1}$ from [13].

Lemma 4.1 ([13])**.** *The largest eigenvalues of the stiffness matrix A and preconditioned stiffness matrix $S^{-1}AS^{-1}$ with the Jacobi preconditioner $S = \mathrm{diag}\left(\sqrt{A_{jj}}\right)$ for the linear finite element approximation of BVP* (1) *are bounded by*

$$\max_j A_{jj} \leq \lambda_{\max}(A) \leq (d+1)\max_j A_{jj} \leq d_{\min} \, d \left(\frac{\sqrt{d+1}}{d!}\right)^{\frac{2}{d}} \max_j \sum_{k\in\omega_j} |K|\beta_K, \quad (27)$$

$$1 \leq \lambda_{\max}(S^{-1}AS^{-1}) \leq d+1. \tag{28}$$

Combining Lemma 4.1 with Lemmas 3.4 and 3.5 we obtain our main theorem.

Theorem 4.1. *Denote the average element volume by $|\bar{K}| = \frac{\Omega}{N}$ and let $p \in (1, \frac{d}{d-2})$. Then the condition numbers of A and $S^{-1}AS^{-1}$ for the linear finite element approximation of BVP* (1) *are bounded by*

$$\kappa(A) \leq CN^{\frac{2}{d}}\left(N^{\frac{d-2}{d}}\max_j \sum_{K\in\omega_j} |K|\beta_K\right)$$

$$\times \begin{cases} \frac{1}{N}\sum_K d_K, & \text{for } d = 1, \\[2mm] \left(1 + \frac{1}{N}\sum_K \ln^2\left(1 + \frac{|\bar{K}|}{|K_{\min}|}d_K\right)\right)^{\frac{1}{2}}, & \text{for } d = 2, \\[2mm] \frac{1}{\left(\frac{d}{d-2}-p\right)^{\frac{d}{d+2p}}}\left(\frac{1}{N}\sum_K \left(\frac{|\bar{K}|}{|K|}\right)^{\frac{1}{p-1}} d_K^{\frac{p}{p-1}\cdot\frac{d-(d-2)p}{d+2p}}\right)^{\frac{p-1}{p}}, & \text{for } d \geq 3, \end{cases} \tag{29}$$

11

and

$$\kappa(S^{-1}AS^{-1}) \leq CN^{\frac{2}{d}}$$

$$\times \begin{cases} \frac{1}{N^2} \sum_K |K| \beta_K d_K, & \text{for } d = 1, \\ \left(\frac{1}{N} \sum_K |K| \beta_K \right)^{\frac{1}{2}} \left(\frac{1}{N} \sum_K |K| \beta_K \left[1 + \ln^2 (1 + d_K \gamma_h) \right] \right)^{\frac{1}{2}}, & \text{for } d = 2, \\ \frac{1}{\left(\frac{d}{d-2} - p \right)^{\frac{d}{d+2p}}} \left(\frac{1}{N^{\frac{2p}{d(p-1)}}} \sum_K |K| \beta_K^{\frac{p}{p-1}} d_K^{\frac{p}{p-1} \cdot \frac{d-(d-2)p}{d+2p}} \right)^{\frac{p-1}{p}}, & \text{for } d \geq 3, \end{cases} \tag{30}$$

where β_K and γ_h defined in (23) *and* (24).

Remark 4.1. From the theorem we can see that both bounds for $\kappa(A)$ and $\kappa(S^{-1}AS^{-1})$ contain the maximum distance d_K from element K to the boundary of the domain. Since d_K becomes smaller when K is closer to $\partial\Omega$, elements close to the boundary have less influence on the bounds than those away from the boundary. This information is useful for problems having boundary layers for which adaptive meshes are typically dense near the boundary.

Remark 4.2 (Comparison to [13]). The bounds in (29) and (30) are sharper than (in the cases of $d \leq 2$) or comparable to (in the case of $d \geq 3$) those obtained in [13][Theorem 5.2] using Sobolev's inequality,

$$\kappa(A) \leq CN^{\frac{2}{d}} \times \left(N^{\frac{d-2}{d}} \max_j \sum_{K \in \omega_j} |K| \beta_K \right) \times \begin{cases} 1, & \text{for } d = 1, \\ 1 + \ln \frac{|\bar{K}|}{|K_{\min}|}, & \text{for } d = 2, \\ \left(\frac{1}{N} \sum_K \left(\frac{|\bar{K}|}{|K|} \right)^{\frac{d-2}{2}} \right)^{\frac{2}{d}}, & \text{for } d \geq 3. \end{cases} \tag{31}$$

and

$$\kappa(S^{-1}AS^{-1}) \leq CN^{\frac{2}{d}} \times \begin{cases} \left(\frac{1}{N^2} \sum_K |K| \beta_K \right), & \text{for } d = 1, \\ \left(\frac{1}{N} \sum_K |K| \beta_K \right) (1 + |\ln \gamma_h|), & \text{for } d = 2, \\ \left(\frac{1}{N} \sum_K |K| \beta_K^{\frac{d}{2}} \right)^{\frac{2}{d}}, & \text{for } d \geq 3. \end{cases} \tag{32}$$

For $d \leq 2$, bounds (31) and (32) follow directly from Theorem 4.1 if we replace d_K by its largest possible value h_Ω (the diameter of Ω). For $d \geq 3$, if we replace d_K with h_Ω and take p close to $d/(d-2)$, the new bounds are very close to the bound (31) and (32). In this sense, they are comparable.

Remark 4.3 ($\mathbb{D}^{-1}$-uniform meshes). Like (31) and (32), bounds (29) and (30) involve three factors: the basic factor $N^{\frac{2}{d}}$ which corresponds to the condition number of the stiffness matrix for the Laplace operator on a uniform mesh, the factor involving β_K which reflects the mesh non-uniformity in the metric specified by $\mathbb{D}^{-1}$, and the factor involving the element volume which represents the mesh non-uniformity in the Euclidean metric. Since

$$\frac{1}{d} \text{tr} \left((F_K')^{-1} \mathbb{D}_K (F_K')^{-T} \right) \leq \left\| (F_K')^{-1} \mathbb{D}_K (F_K')^{-T} \right\|_2 \leq \text{tr} \left((F_K')^{-1} \mathbb{D}_K (F_K')^{-T} \right),$$

we can estimate β_K as

$$\left(\frac{N}{\sigma_h}\right)^{\frac{2}{d}} \left(\frac{\frac{1}{d}\operatorname{tr}\left((F_K')^{-1}\mathbb{D}_K(F_K')^{-T}\right)}{\det\left((F_K')^{-1}\mathbb{D}_K(F_K')^{-T}\right)^{\frac{1}{d}}}\right)\left(\frac{\frac{\sigma_h}{N}}{|K|\det(\mathbb{D}_K)^{-\frac{1}{2}}}\right)^{\frac{2}{d}}$$

$$\leq \beta_K \leq d\left(\frac{N}{\sigma_h}\right)^{\frac{2}{d}}\left(\frac{\frac{1}{d}\operatorname{tr}\left((F_K')^{-1}\mathbb{D}_K(F_K')^{-T}\right)}{\det\left((F_K')^{-1}\mathbb{D}_K(F_K')^{-T}\right)^{\frac{1}{d}}}\right)\left(\frac{\frac{\sigma_h}{N}}{|K|\det(\mathbb{D}_K)^{-\frac{1}{2}}}\right)^{\frac{2}{d}},$$

where $\sigma_h = \sum_K |K|\det(\mathbb{D}_K)^{-\frac{1}{2}}$. For any uniform mesh in the metric specified by $\mathbb{D}^{-1}$ we have (e.g., see Huang and Russell [11])

$$\frac{1}{d}\operatorname{tr}\left((F_K')^{-1}\mathbb{D}_K(F_K')^{-T}\right) = \det\left((F_K')^{-1}\mathbb{D}_K(F_K')^{-T}\right)^{\frac{1}{d}} \quad\text{and}\quad \frac{\sigma_h}{N} = |K|\det(\mathbb{D}_K)^{-\frac{1}{2}}$$

and therefore

$$\left(\frac{N}{\sigma_h}\right)^{\frac{2}{d}} \leq \beta_K \leq d\left(\frac{N}{\sigma_h}\right)^{\frac{2}{d}}.$$

From (29) and (30), we get

$$\kappa(A) \leq C\left(\frac{N}{\sigma_h}\right)^{\frac{2}{d}}(N\,|\omega_{\max}|)$$

$$\times\begin{cases} \frac{1}{N}\sum_K d_K, & \text{for } d = 1, \\[2mm] \left(\frac{1}{N}\sum_K\left[1 + \ln^2\left(1 + d_K\frac{|\bar{K}|}{|K_{\min}|}\right)\right]\right)^{\frac{1}{2}}, & \text{for } d = 2, \\[2mm] \frac{1}{\left(\frac{d}{d-2}-p\right)^{\frac{d}{d+2p}}}\left(\frac{1}{N}\sum_K\left(\frac{|\bar{K}|}{|K|}\right)^{\frac{1}{p-1}}d_K^{\frac{p}{p-1}\cdot\frac{d-(d-2)p}{d+2p}}\right)^{\frac{p-1}{p}}, & \text{for } d \geq 3, \end{cases} \quad (33)$$

and

$$\kappa(S^{-1}AS^{-1}) \leq C\left(\frac{N}{\sigma_h}\right)^{\frac{2}{d}}\times\begin{cases} \sum_K |K|\,d_K, & \text{for } d = 1, \\[2mm] \left(\sum_K |K|\left[1 + \ln^2\left(1 + d_K\right)\right]\right)^{\frac{1}{2}}, & \text{for } d = 2, \\[2mm] \frac{1}{\left(\frac{d}{d-2}-p\right)^{\frac{d}{d+2p}}}\left(\sum_K |K|\,d_K^{\frac{p}{p-1}\cdot\frac{d-(d-2)p}{d+2p}}\right)^{\frac{p-1}{p}}, & \text{for } d \geq 3. \end{cases} \quad (34)$$

If we replace d_K by h_Ω in (34), we get

$$\kappa(S^{-1}AS^{-1}) \leq C\left(\frac{N}{\sigma_h}\right)^{\frac{2}{d}},$$

which gives the same bound obtained in [13] for $\mathbb{D}^{-1}$-uniform meshes.

13

Remark 4.4. As mentioned before, the approach we used here to estimate $\lambda_{\min}(A)$ and $\lambda_{\min}(S^{-1}AS^{-1})$ was first proposed by Fried [7]. However, there is significant difference between our development and Fried's. First, Fried obtains a bound on $\lambda_{\min}^{\rho}(-\Delta)$ for balls where the analytical formula of the Green's function is available and claims using a physical intuition that the bound is also valid for other domains. Our derivation for the bound on $\lambda_{\min}^{\rho}(-\Delta)$ is mathematically rigorous, as seen in Lemmas 3.1–3.3. Second, the lower bound obtained in Fried [7] for $\lambda_{\min}(A)$ can be expressed in the current notation as

$$\lambda_{\min}(A) \geq \frac{Cd_{\min}}{N} \begin{cases} 1, & \text{for } d = 1, \\ \left(1 + \ln \frac{|\bar{K}|}{|K_{\min}|}\right)^{-1}, & \text{for } d = 2, \\ \left(\frac{|\bar{K}|}{|K_{\min}|}\right)^{\frac{2}{d}-1}, & \text{for } d \geq 3. \end{cases} \tag{35}$$

It is easy to see that our estimate (21) is sharper than (35). Third, Fried estimates the maximum eigenvalue of the stiffness matrix by the maximum of the maximum eigenvalues of the local stiffness matrices, with the latter being computed directly for the Laplace operator. On the other hand, our estimate on the maximum eigenvalue of the stiffness matrix in Lemma 4.1 is not only more general but also more accurate. Finally, we would like to mention that Fried [7] does not study preconditioning for the stiffness matrix.

5 Numerical experiments

The dependence of the conditioning on the diffusion matrix was extensively discussed in [13] and the purpose of this paper is to investigate the dependence of the conditioning on the mesh density throughout the domain. Thus, for the simplicity, we set $\mathbb{D} = I$ (Laplace operator) in all of our numerical examples.

Bounds on the smallest eigenvalue contain a constant C independent of the mesh but dependent on the dimension, domain, reference element, and basis functions on the reference element. In our computation we obtain its value empirically by calibrating the bounds for $\lambda_{\min}(S^{-1}AS^{-1})$ through comparing the exact and estimated values for a series of uniform meshes.

5.1 1D

In 1D, the new bounds (29) and (30) (with $\mathbb{D} = I$) become

$$\kappa(A) \leq C \sum_K d_K \cdot \max_j \sum_{K \in \omega_j} \frac{1}{|K|}, \tag{36}$$

$$\kappa(S^{-1}AS^{-1}) \leq C \sum_K \frac{d_K}{|K|}. \tag{37}$$

On the other hand, the bounds (31) and (32) from [13] reduce to

$$\kappa(A) \leq CN \cdot \max_j \sum_{K \in \omega_j} \frac{1}{|K|}, \tag{38}$$

$$\kappa(S^{-1}AS^{-1}) \leq C \sum_K \frac{1}{|K|}. \tag{39}$$

We consider two set of meshes formed by Chebyshev nodes and nodes simulating extreme boundary layer refinement.

5.1.1 Chebyshev nodes

The first 1D example is a non-uniform mesh given by the end points $x_0 = 0$ and $x_N = 1$ and the Chebyshev nodes,

$$x_j = \frac{1}{2}\big(1 - \cos(\xi_j)\big), \quad j = 1, \ldots, N-1, \tag{40}$$

where

$$\xi_0 = 0, \ \xi_N = 1 \quad \text{and} \quad \xi_j = \frac{\pi(2j-1)}{2(N-1)}, \quad j = 1, \ldots, N-1.$$

For this mesh, we have

$$|K_j| = x_j - x_{j-1} \sim \frac{1}{N}\sin(\xi_j), \quad d_{K_j} = \min\left(\frac{1}{2}\big(1 - \cos(\xi_j)\big), \frac{1}{2}\big(1 + \cos(\xi_{j-1})\big)\right).$$

Using these, we get

$$\max_j \sum_{K \in \omega_j} \frac{1}{|K|} \sim \max_j \frac{1}{|K_j|} \sim N^2$$

and

$$\frac{1}{N}\sum_K d_K \sim \frac{2}{N}\sum_{j=1}^{N/2} \frac{1}{2}\big(1 - \cos(\xi_j)\big) \sim \int_0^{\frac{\pi}{2}} \big(1 - \cos(\xi)\big)\, d\xi \sim 1.$$

Thus, from (36) and (38) we see that the new bound and the bound from [13] both lead to

$$\kappa(A) \leq CN^3.$$

For the scaled case, we have

$$\sum_K \frac{d_K}{|K|} \sim 2\sum_{j=1}^{N/2} \frac{\frac{1}{2}\big(1 - \cos(\xi_j)\big)}{\frac{\pi}{N}\sin(\xi_j)} \sim N^2 \int_0^{\frac{\pi}{2}} \frac{1 - \cos(\xi)}{\sin(\xi)}\, d\xi \sim N^2$$

and

$$\sum_K \frac{1}{|K|} \sim 2\sum_{j=1}^{N/2} \frac{1}{\frac{\pi}{N}\sin(\xi_j)} \sim N^2 \int_{\xi_1}^{\frac{\pi}{2}} \frac{1}{\sin(\xi)}\, d\xi \sim N^2 \ln|\tan(\xi/2)| \, |_{\xi_1}^{\frac{\pi}{2}} \sim N^2 \ln N.$$

15

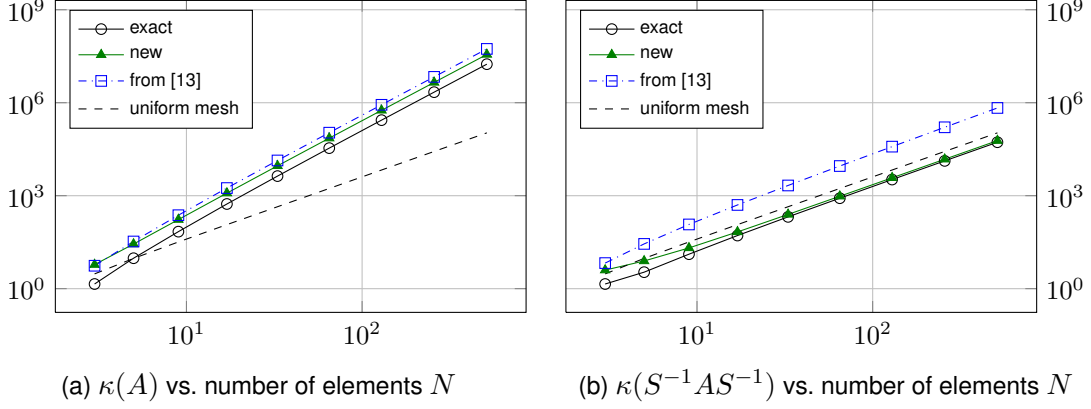(a) $\kappa(A)$ vs. number of elements N (b) $\kappa(S^{-1}AS^{-1})$ vs. number of elements N

Figure 1: Conditioning for the Chebyshev nodes (40)

Thus, we have

$$\kappa(S^{-1}AS^{-1}) \leq CN^2 \tag{41}$$

for the new bound (37) and

$$\kappa(S^{-1}AS^{-1}) \leq CN^2 \ln N$$

from the bound (39) from [13]. Notice that (41) has the same order as $\kappa(S^{-1}AS^{-1})$ for a uniform mesh as N increases.

The numerical comparison of the estimated and exact values for $\kappa(A)$ and $\kappa(S^{-1}AS^{-1})$ are presented in Fig. 1. As expected, for the non-scaled case both estimates are comparable and very tight (Fig. 1a). On the other hand, the new estimate (29) is more accurate than the one from [13] if diagonal scaling is applied (Fig. 1b). For this example, the new bound on $\kappa(S^{-1}AS^{-1})$ seems to be asymptotically exact.

5.1.2 Boundary layer refinement

The second 1D example simulates boundary layer mesh refinement towards the boundary point $x = 0$ with the internal nodes

$$x_j = 2^j/2^N, \qquad j = 1, \ldots, N - 1. \tag{42}$$

For this mesh, we have

$$|K_j| = x_j - x_{j-1} = \frac{2^{j-1}}{2^N}, \quad d_{K_j} = \min\left(\frac{2^j}{2^N}, 1 - \frac{2^{j-1}}{2^N}\right).$$

Then, bounds (38) and (39) for [13] become

$$\kappa(A) \leq CN2^N \quad \text{and} \quad \kappa(S^{-1}AS^{-1}) \leq C2^N,$$

whereas the new bounds (36) and (37) give

$$\kappa(A) \leq C2^N \quad \text{and} \quad \kappa(S^{-1}AS^{-1}) \leq CN.$$

16

(a) $\kappa(A)$ vs. number of elements N (b) $\kappa(S^{-1}AS^{-1})$ vs. number of elements N
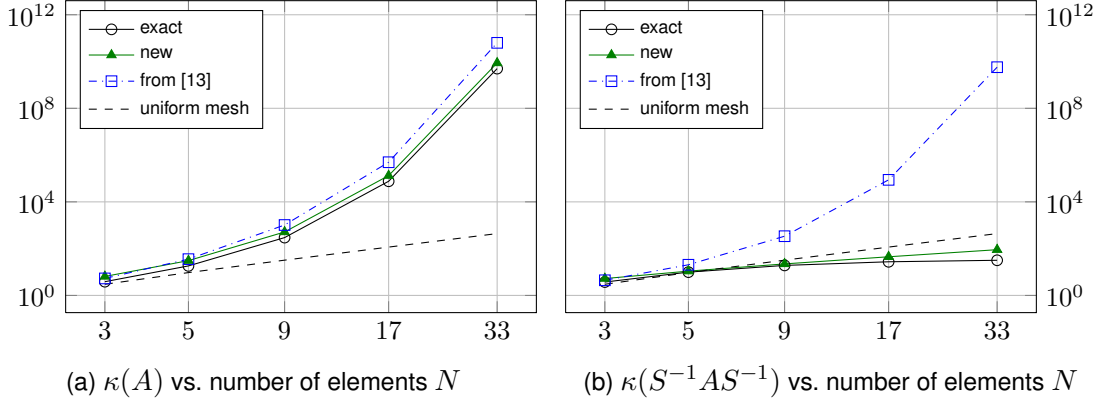
Figure 2: Conditioning for the boundary layer refinement (42)

This shows that, for boundary layer refinement, the new bound for the scaled case is a significant improvement. Note that $\kappa(S^{-1}AS^{-1}) = \mathcal{O}(N)$, which has a smaller order as $N \to \infty$ than the condition number for a uniform mesh (which is $\mathcal{O}(N^2)$). Thus, for 1D problems with steep boundary layers, strong mesh concentration towards the boundary not only increases the accuracy of the solution but, at the same time, improves the conditioning.

Numerical results are in perfect agreement with the analysis. Indeed, Fig. 2a shows that both new and old estimates are comparable for the non-scaled case, with the new one being slightly more accurate. After scaling (Fig. 2b), the situation is quite different: the new estimate is very close to the exact value whereas the estimate from [13] exhibits a dramatic overestimation.

5.2 2D

In 2D we consider a mesh for the unit square $[0, 1] \times [0, 1]$ with $\mathcal{O}(N^{1/2})$ skew elements near the boundary, as shown in Fig. 3a.

First, we set the maximum aspect ratio at $125 : 1$ and verify the dependence of the condition number on the number of mesh elements N (Fig. 4). Then, we set $N = 20\,000$ and change the maximum aspect ratio of the mesh elements to investigate the dependence of the conditioning on the mesh shape (Fig. 5).

The observation is that the new and the [13] estimates are comparable, with the new one being slightly more accurate, especially when changing the maximal mesh aspect ratio. However, neither of the estimates is as accurate as in 1D, meaning that our estimation of the smallest eigenvalue can be further improved. This observation is essentially valid in higher dimensions as well. Since the Green's function has a singularity for $d \geq 2$ (and depends on the shape of the domain), it is difficult to obtain estimates as sharp as in 1D in general. Interestingly, in our example, the exact condition number of the scaled stiffness matrix appears to be independent of the aspect ratio of the boundary layer elements. This suggests that the optimal estimate in 2D, similar to the 1D case, should be

$$\kappa(S^{-1}AS^{-1}) \leq C \sum_K |K|\, \beta_K \log\left(1 + \frac{d_K}{|K|}\right).$$ (43)
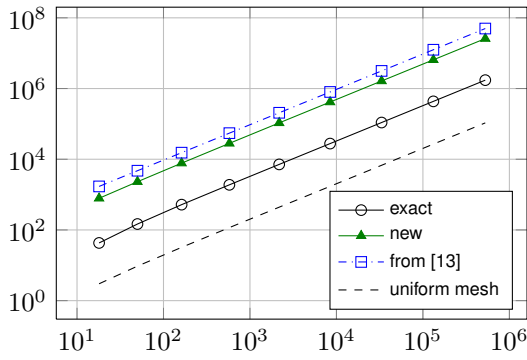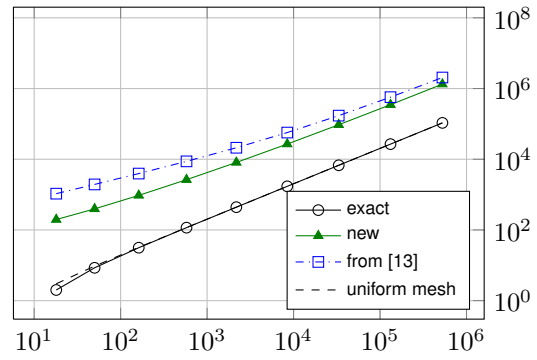
17
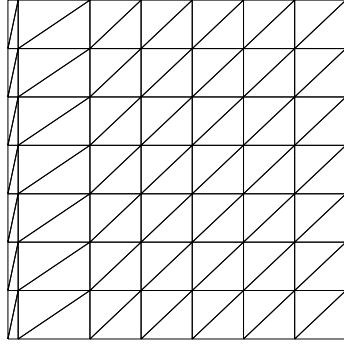
(a) 2D meshes used in Sect. 5.2
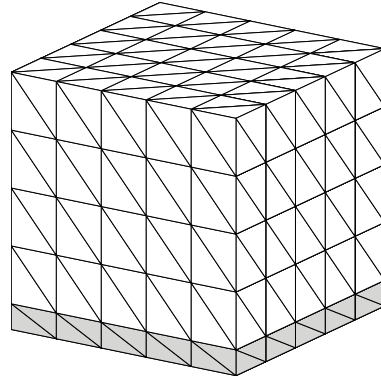
(b) 3D meshes used in Sect. 5.3

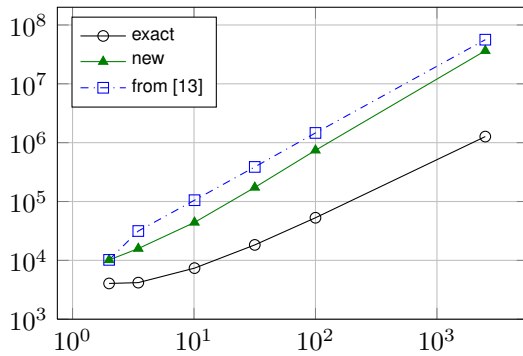Figure 3: Test meshes in two and three dimensions
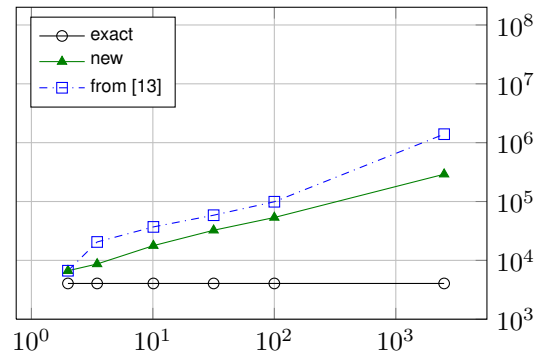

(a) $\kappa(A)$ vs. number of elements N

(b) $\kappa(S^{-1}AS^{-1})$ vs. number of elements N

Figure 4: 2D meshes with a fixed maximum aspect ratio of $125:1$ and changing N

(a) $\kappa(A)$ vs. maximum aspect ratio

(b) $\kappa(S^{-1}AS^{-1})$ vs. maximum aspect ratio

Figure 5: 2D meshes with a fixed $N = 20\,000$ and changing aspect ratio
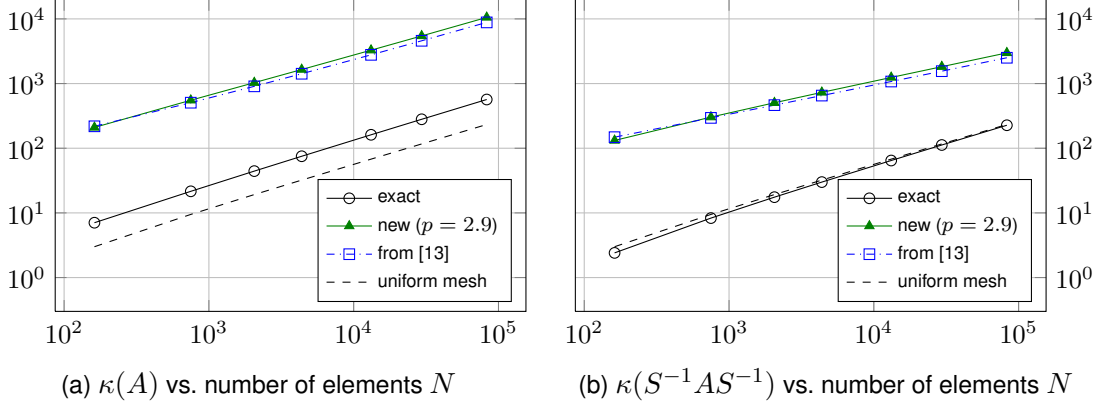
18

(a) $\kappa(A)$ vs. number of elements N (b) $\kappa(S^{-1}AS^{-1})$ vs. number of elements N

Figure 6: 3D meshes with a fixed aspect ratio of $25 : 25 : 1$ and changing N

(a) $\kappa(A)$ vs. number of elements N (b) $\kappa(S^{-1}AS^{-1})$ vs. number of elements N
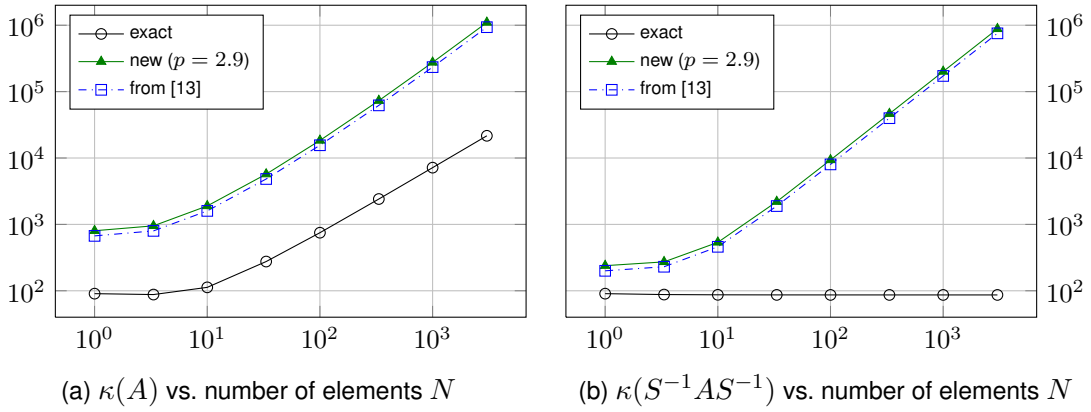
Figure 7: 3D meshes with a fixed $N = 29\,478$ and changing aspect ratio

5.3 3D

Similarly to the 2D case, we use a mesh for the unit cube $[0, 1]^3$ with $\mathcal{O}(N^{2/3})$ skew elements near the boundary (Fig. 3b) and consider two different settings: fixed anisotropy ($25 : 25 : 1$) with increasing number of elements and fixed $N = 29\,478$ paired with the changing anisotropy of the mesh. Numerical results for $p = 2.9$ are presented in Figs. 6 and 7.

First, we observe that the new estimate is comparable to the estimate from [13] for both non-scaled and scaled cases. Second, as in the 2D case, we observe that increasing the maximum aspect ratio of elements at the boundary has no impact on the exact condition number of the Jacobi preconditioned stiffness matrix (at least for the considered mesh type). This indicates that the obtained bounds for the scaled stiffness matrix are not optimal. The numerical results also suggest that the optimal bound should have a much stronger dependence on the distance d_K of an element K from the boundary of the domain.

6 Conclusions

In the previous sections we have studied the linear finite element approximation of the boundary value problem (1) with general nonuniform meshes and developed bounds for the condition numbers of the stiffness matrix and Jacobi preconditioned stiffness matrix. The main result (Theorem 4.1) shows that the density function approach of Fried [7] can be made mathematically rigorous for general domains and lead to estimates that provide more detail and are sharper than existing estimates for general adaptive meshes in one and two dimensions and comparable in three and higher dimensions. Moreover, the bounds in Theorem 4.1 involve a factor d_K which describes the maximum distance from element K to the boundary of the domain and becomes smaller when K is closer to $\partial\Omega$. They reveal that the mesh concentration near the boundary has less influence on the condition number than the mesh concentration in the interior of the domain. This is especially true for the Jacobi preconditioned system where the former has little or almost no effect on the condition number. The numerical results presented in Sect. 5 confirm the theoretical analysis although they also suggest that the new bounds could be further improved in two and higher dimensions.

References

[1] S. Agmon, A. Douglis, and L. Nirenberg. Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. I. *Comm. Pure Appl. Math.*, 12:623–727, 1959.

[2] M. Ainsworth, B. McLean, and T. Tran. Diagonal scaling of stiffness matrices in the Galerkin boundary element method. *ANZIAM J.*, 42(1):141–150, 2000.

[3] M. Ainsworth, W. McLean, and T. Tran. The conditioning of boundary element equations on locally refined meshes and preconditioning by diagonal scaling. *SIAM J. Numer. Anal.*, 36(6):1901–1932 (electronic), 1999.

[4] R. E. Bank and L. R. Scott. On the conditioning of finite element equations with highly refined meshes. *SIAM J. Numer. Anal.*, 26(6):1383–1394, 1989.

[5] A. Dall'Acqua and G. Sweers. Estimates for Green function and Poisson kernels of higher-order Dirichlet boundary value problems. *J. Differential Equations*, 205(2):466–487, 2004.

[6] Q. Du, D. Wang, and L. Zhu. On mesh geometry and stiffness matrix conditioning for general finite element spaces. *SIAM J. Numer. Anal.*, 47(2):1421–1444, 2009.

[7] I. Fried. Bounds on the spectral and maximum norms of the finite element stiffness, flexibility and mass matrices. *Internat. J. Solids and Structures*, 9:1013–1034, 1973.

[8] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

[9] I. G. Graham and W. McLean. Anisotropic mesh refinement: the conditioning of Galerkin boundary element matrices and simple preconditioners. *SIAM J. Numer. Anal.*, 44(4):1487–1513 (electronic), 2006.

[10] M. Grüter and K.-O. Widman. The Green function for uniformly elliptic equations. *Manuscripta Math.*, 37(3):303–342, 1982.

[11] W. Huang and R. D. Russell. *Adaptive Moving Mesh Methods*, volume 174 of *Applied Mathematical Sciences.* Springer, New York, 2011.

[12] A. Ifra and L. Riahi. Estimates of Green functions and harmonic measures for elliptic operators with singular drift terms. *Publ. Mat.*, 49(1):159–177, 2005.

[13] L. Kamenski, W. Huang, and H. Xu. Conditioning of finite element equations with arbitrary anisotropic meshes. *Math. Comp.* (forthcoming). arXiv: 1201.3651 [math.NA].

[14] J. R. Shewchuk. What is a good linear element? Interpolation, conditioning, and quality measures. In *Proceedings of the 11th International Meshing Roundtable*, pages 115–126, Sandia National Laboratories, 2002.

[15] A. J. Wathen. Realistic eigenvalue bounds for the Galerkin mass matrix. *IMA J. Numer. Anal.*, 7(4):449–457, 1987.