Todd Dupont, Johnny Guzmán and L. Ridgway Scott*

Obtaining higher-order Galerkin accuracy when the boundary is polygonally approximated

https://doi.org/10.1515/jnma-2023-0135
Received November 6, 2023; accepted March 29, 2024; published online October 7, 2024

Abstract: Inspired by the methods developed in J. H. Bramble, T. Dupont, and V. Thomée ("Projection methods for Dirichlet's problem in approximating polygonal domains with boundary-value corrections," *Math. Comput.*, vol. 26, no. 120, pp. 869–879, 1972), we introduce a new technique that yields a symmetric formulation and has similar performance. The new method is based on a Robin-type problem on an approximate polygonal domain. Optimal error estimates in the energy norm are proved for piecewise quadratics and cubics. We provide numerical experiments that show our theoretical results are sharp.

Keywords: curved boundaries; Galerkin method; high order

MSC 2010 Classification: 65M60

1 Introduction

When a Dirichlet problem on a smooth domain is approximated by a polygon, an error occurs that is suboptimal for piecewise quadratic approximations [1]–[4]. There are several approaches that have been developed to overcome the sub-optimality. Two common approaches are: (1) methods that use curved elements (see, for example [5]–[7], to name a few) and (2) methods that use augmented polygonal approximations. The method we develop here falls within the second category. A non-exhaustive list of such methods includes [8]–[19].

To put our contribution in context, we consider a model problem and previous numerical methods for this problem. Let Ω be a piecewise smooth, bounded, two-dimensional domain that is on only one side of its boundary. Consider the Poisson equation with Dirichlet boundary conditions:

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega. \tag{1}$$

We assume that f and g are sufficiently smooth that u can be extended to be in $W^{2,q}(\widehat{\Omega})$, where $\widehat{\Omega}$ contains a neighborhood of the closure of Ω, for some q in the range $1 < q \leqslant \infty$. For a Lipschitz domain, this is possible with $q < 4/3$.

One way to discretize (1) is to approximate the domain Ω by polygons Ω_h, where the edge lengths of $\partial\Omega_h$ are of order h in size. Then conventional finite elements can be employed on each Ω_h, with the Dirichlet boundary

***Corresponding author: L. Ridgway Scott**, The University of Chicago, Emeritus, Chicago, IL 60637, USA, E-mail: ridg@uchicago.edu. https://orcid.org/0000-0002-7885-7106
Todd Dupont, Departments of Computer Science and of Mathematics, The University of Chicago, Chicago, IL 60637, USA, E-mail: dupont@cs.uchicago.edu
Johnny Guzmán, Division of Applied Mathematics, Brown University, Box F, 182 George Street, Providence, RI 02912, USA, E-mail: johnny_guzman@brown.edu. https://orcid.org/0000-0002-6769-2393

Open Access. © 2024 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

conditions being approximated by the assumption that $u_h = \hat{g}$ on $\partial\Omega_h$ [20], with $\hat{g}$ appropriately defined. For example, let us suppose for the moment that $g \equiv 0$ and we take $\hat{g} \equiv 0$ as well. In particular, we assume that Ω_h is triangulated with a nondegenerate mesh $\mathcal{T}_h$ of maximum triangle size h, and the boundary vertices of Ω_h are in $\partial\Omega$. We define $\mathring{W}_h^k := H_0^1(\Omega_h) \cap W_h^k$, where

$$W_h^k = \{v \in C(\Omega_h):\ v|_T \in \mathcal{P}_k\ \forall T \in \mathcal{T}_h\} \tag{2}$$

and $\mathcal{P}_k$ is the set of polynomials of total degree $\leqslant k$. Then the standard finite element approximation finds $u_h \in \mathring{W}_h^k$ satisfying

$$a_h(u_h, v) = (f, v)_{L^2(\Omega_h)} \quad \forall v \in \mathring{W}_h^k, \tag{3}$$

where $a_h(u, v) := \int_{\Omega_h} \nabla u \cdot \nabla v \, dx$. Here we assume that f is extended smoothly outside of Ω.

This approach for $k = 1$ (piecewise linear approximation) leads to the error estimate

$$\|u - u_h\|_{H^1(\Omega_h)} \leqslant Ch\|u\|_{H^2(\hat{\Omega})}.$$

However, when this approach is applied with piecewise quadratic polynomials ($k = 2$), the best possible error estimate [20] is

$$\|u - u_h\|_{H^1(\Omega_h)} \leqslant Ch^{3/2}, \tag{4}$$

which is less than optimal order by a factor of $\sqrt{h}$. The reason of course is that we have made only a piecewise linear approximation of $\partial\Omega$. Table 1 summarizes some computational experiments for the test problem in Section 3. We see a significant improvement for quadratics over linears, but there is almost no improvement with cubics. Moreover, we will see that a significant improvement using quadratics can be obtained using simple approaches that modify the variational form.

The first method using a polygonal approximate domain to obtain higher order schemes was based on modifying the method of Nitsche [5] and appeared in 1972 [8]. There are two main ingredients in the scheme in ref. [8]:

- Nitsche's penalization to both enforce the boundary condition and stabilize the method;
- Taylor's expansions to give approximate boundary conditions on Ω_h.

The idea is to think of u as the solution of a PDE on the approximate domain, with boundary conditions that need to be approximated.

Table 1: Errors $u_h - u_I$ in $L^2(\Omega_h)$ and $H^1(\Omega_h)$, as a function of the maximum mesh size (hmax) for the polygonal approximation (3) for the test problem in Section 3 using various polynomial degrees k. Key: "M" is input parameter to `mshr` function `circle` used to generate the mesh, "seg" is the number of boundary edges. The approximate solutions were generated using (3). The interpolant u_I is defined in (16).

k	M	L2 err	rate	H1 err	rate	seg	hmax
1	2	1.84e+00	NA	6.25e+00	NA	10	1.05e+00
1	4	2.93e−01	2.65	1.89e+00	1.73	20	4.94e−01
1	8	9.55e−02	1.62	1.06e+00	0.83	40	2.61e−01
1	16	2.47e−02	1.95	5.45e−01	0.96	80	1.35e−01
2	2	4.18e−01	NA	1.41e+00	NA	10	1.05e+00
2	4	9.44e−02	2.15	4.26e−01	1.73	20	4.94e−01
2	8	2.30e−02	2.04	1.59e−01	1.42	40	2.61e−01
2	16	5.62e−03	2.03	5.45e−02	1.54	80	1.35e−01
3	2	3.17e−01	NA	8.25e−01	NA	10	1.05e+00
3	4	8.81e−02	1.85	2.94e−01	1.49	20	4.94e−01
3	8	2.22e−02	1.99	1.07e−01	1.46	40	2.61e−01
3	16	5.53e−03	2.01	3.82e−02	1.49	80	1.35e−01

Methods for any polynomial order are developed in ref. [8], and it is proved that the error estimates are optimal in the energy norm. The bulk of ref. [8] analyzes a non-symmetric extension of Nitsche's method, but a symmetrized version is defined [8, eq. (3.12)] for the lowest-order scheme for which the same estimates hold by trivial modifications, and estimates are described in the case that the distance to the boundary is only approximated.

Two years later [8], was extended [15] to cover both 2 and 3 dimensions, variable coefficients, and estimates in H^{-1} as well as H^1 and L^2. The paper [15] uses the symmetric form throughout, it allows both linear and non-linear approximations of the boundary (that is, the approximating domains can be curved), and it also applies the elliptic theory to nonlinear parabolic equations.

The work of refs. [8], [15] has been extended in many ways [9]–[11], [16], [21]–[24] over the intervening decades.

Recently, the related shifted-boundary method has been developed [17], [18] and analyzed [25], [26]. It uses the two main ingredients (i.e., Nitsche's penalization and Taylor's expansions) to develop methods for Poisson's problem, fluid flow problems and advection–diffusion problems. The method seems to be stable when the boundary of Ω_h is within $O(h)$ of the boundary of Ω. Cheung et al. [12] develop a method using average Taylor expansions, and they prove stability in the case the boundary of Ω_h is within $O(h^{3/2})$ of the boundary of Ω. Finally, a series of papers [13], [14], [19], develops hybridizable discontinuous Galerkin (HDG) methods using polygonal approximations. HDG has its own stabilization mechanism which they use. A distinctive feature is that they approximate both u and ∇u independently. The latter has been named the "transfer path method" (TPM) [27] and been developed not only for HDG methods, but also for mixed finite element methods [28], [29]. Therefore, it does not require a stabilization parameter.

In this paper, we introduce a new method that uses a first order Taylor expansion which naturally leads to a Robin-type boundary problem on Ω_h. However, we do not use Nitsche's stabilization, and thus we avoid the need to choose a penalty parameter. Another parameter-free method is developed in refs. [30]–[32].

The Robin-type method studied here is naturally symmetric, but it can be not positive definite. Despite this, we are able to prove optimal estimates for piecewise quadratics and cubics. The analysis is quite different from what is used for methods that are rooted in Nitsche's method, such as refs. [8], [15] and its descendents. It is not a simple perturbation of the standard arguments. For this reason, we restrict attention to the model problem (1) in order to make our analysis as transparent as possible. We assume for simplicity that the vertices of the boundary Ω_h belong to boundary of Ω (i.e., a fitted mesh). We give numerical results that show that our estimates are sharp.

2 The Bramble–Dupont–Thomée approach

We start by reviewing the method [8] of Bramble–Dupont–Thomée (BDT) which achieves high-order accuracy by modifying Nitsche's method [5] applied on Ω_h. We assume that $\Omega_h \subset \Omega$ and we do not necessarily assume that the boundary vertices of Ω_h belong to $\partial\Omega$. The bilinear form used in ref. [8] is

$$N_h(u,v) = a_h(u,v) - \int_{\partial\Omega_h} \frac{\partial u}{\partial n} v \, \mathrm{d}s - \int_{\partial\Omega_h} \left(u + \delta \frac{\partial u}{\partial n}\right)\left(\frac{\partial v}{\partial n} - \gamma h^{-1} v\right) \mathrm{d}s. \tag{5}$$

Here, n denotes the outward-directed normal to $\partial\Omega_h$ and

$$\delta(x) = \min\{s > 0 : \, x + sn \in \partial\Omega\}.$$

Contrast the definition of δ to the closely related function d defined by

$$d(x) = \min\{|x - y| : \, y \in \partial\Omega\},$$

see Figure 1.

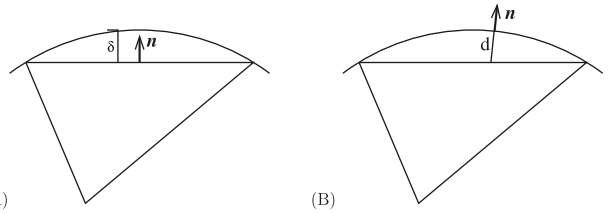

Figure 1: Definitions of (A) δ and (B) d.

For simplicity the assume that $g = 0$. Then the BDT method will find $u_h \in W_h^k$ such that

$$N_h(u_h, v) = \int_{\Omega_h} fv \, \mathrm{d}x \quad \forall v \in W_h^k.$$

If δ were 0, this would be Nitsche's method on Ω_h.

Corrections of arbitrary order, involving terms $\delta^\ell \frac{\partial^\ell u}{\partial n^\ell}$ for $\ell > 1$ are studied in ref. [8], but for simplicity we restrict attention to the first-order correction to Nitsche's method given in (5). The error estimates obtained in ref. [8] are as follows

$$|\!|\!| u - u_h |\!|\!|_1 \leqslant Ch^k \|u\|_{H^{k+1}(\Omega)} + Ch^{7/2} \|u\|_{W^2_\infty(\Omega)},$$

where

$$|\!|\!| v |\!|\!|_1 := \left(a_h(v, v) + h^{-1} \int_{\partial\Omega_h} v^2 \, \mathrm{d}s + h \int_{\partial\Omega_h} \left(\frac{\partial v}{\partial n} \right)^2 \mathrm{d}s \right)^{1/2}.$$

Thus using the variational form (5) leads to an approximation that is optimal-order with quadratics and cubics and is only suboptimal for quartics by a factor of $\sqrt{h}$.

3 An example: the circle

We consider a numerical example. Consider the case where Ω is a disc of radius R centered at the origin, in which case we have $d(\mathbf{x}) = R - |\mathbf{x}|$. However, it is more difficult to evaluate $\delta(\mathbf{x})$. We assume that the vertices of $\partial\Omega_h$ lie on $\partial\Omega$. We have $\mathbf{x} + \delta(\mathbf{x})\mathbf{n} \in \partial\Omega$ for $\mathbf{x} \in \partial\Omega_h$, where $\mathbf{n}$ denotes the outward normal to Ω_h. Let $\mathbf{t}$ denote the unit tangent vector to Ω_h such that $\mathbf{n} \times \mathbf{t}$ points up. We can write $\mathbf{x} = (\mathbf{x} \cdot \mathbf{n})\,\mathbf{n} + (\mathbf{x} \cdot \mathbf{t})\,\mathbf{t}$, and $(\mathbf{x} \cdot \mathbf{t})^2 = |\mathbf{x}|^2 - (\mathbf{x} \cdot \mathbf{n})^2$. Since $|\mathbf{x} + \delta(\mathbf{x})\mathbf{n}| = R$, we have

$$R^2 = (\mathbf{x} \cdot \mathbf{t})^2 + (\mathbf{x} \cdot \mathbf{n} + \delta(\mathbf{x}))^2 = |\mathbf{x}|^2 - (\mathbf{x} \cdot \mathbf{n})^2 + (\mathbf{x} \cdot \mathbf{n} + \delta(\mathbf{x}))^2.$$

Then

$$\delta(\mathbf{x}) = \pm\sqrt{R^2 - |\mathbf{x}|^2 + (\mathbf{x} \cdot \mathbf{n})^2} - \mathbf{x} \cdot \mathbf{n}.$$

Note that for $\mathbf{x} \in \partial\Omega_h$, $|\mathbf{x}| \leqslant R$ and $\mathbf{x} \cdot \mathbf{n} > 0$. Since $\delta(\mathbf{x}) \geqslant 0$, we must pick the plus sign, so

$$\delta(\mathbf{x}) = \sqrt{R^2 - |\mathbf{x}|^2 + (\mathbf{x} \cdot \mathbf{n})^2} - \mathbf{x} \cdot \mathbf{n}. \tag{6}$$

Note that this formula does not depend on the placement of the boundary vertices. If $\mathbf{x}_i$ is a boundary vertex, then $|\mathbf{x}_i| = R$, and $\delta(\mathbf{x}_i) = 0$.

It is not hard to see that $d - \delta = \mathcal{O}(h^4)$ in this case.

This problem is simple to implement using the FEniCS Project code `dolfin` [33]. We take $R = 1$, $u(x, y) = 1 - (x^2 + y^2)^3$, and $f = 36(x^2 + y^2)^2$ in the computational experiments described subsequently. Computational results for this example are given in Table 2, where we see optimal order approximation for $k \leqslant 3$, improvement for $k = 4$ over $k = 3$ (suboptimal by a factor $h^{-1/2}$), and no improvement for quintics. These errors are depicted in Figure 2.

Table 2: Errors $u_h - u_I$ in $L^2(\Omega_h)$ and $H^1(\Omega_h)$ as a function of mesh size (hmax) for the test problem in Section 3 using the BDT approximation in Section 2, with $\gamma = 100$, for various polynomial degrees k. Key: M is the value of the `meshsize` input parameter to the `mshr` function `circle` used to generate the mesh. The number of boundary edges was set to $5M$, and hmax is the maximum mesh size. The interpolant u_I is defined in (16).

k	*M*	hmax	L2 error	rate	H1 error	rate
1	8	0.261	0.0947	1.61	1.06	0.82
1	16	0.135	0.0245	1.95	0.544	0.96
1	32	0.0688	0.00639	1.94	0.277	0.97
1	64	0.0353	0.00158	2.02	0.137	1.02
2	8	0.261	2.81e−03	2.61	0.103	1.57
2	16	0.135	3.70e−04	2.93	0.0277	1.89
2	32	0.0688	4.77e−05	2.96	0.00717	1.95
2	64	0.0353	5.91e−06	3.01	0.00179	2.00
3	8	0.261	1.56e−04	3.92	5.31e−03	2.54
3	16	0.135	9.44e−06	4.05	7.06e−04	2.91
3	32	0.0688	5.81e−07	4.02	9.23e−05	2.94
3	64	0.0353	3.57e−08	4.02	1.15e−05	3.00
4	8	0.261	1.49e−04	3.96	7.41e−04	3.42
4	16	0.135	9.29e−06	4.00	6.63e-05	3.48
4	32	0.0688	5.80e−07	4.00	5.90e−06	3.49
4	64	0.0353	3.63e−08	4.00	5.22e−07	3.50
5	8	0.261	1.47e−04	3.96	7.10e−04	3.41
5	16	0.135	9.27e−06	3.99	6.44e−05	3.46
5	32	0.0688	5.80e−07	4.00	5.77e−06	3.48
5	64	0.0353	3.62e−08	4.00	5.12e−07	3.49

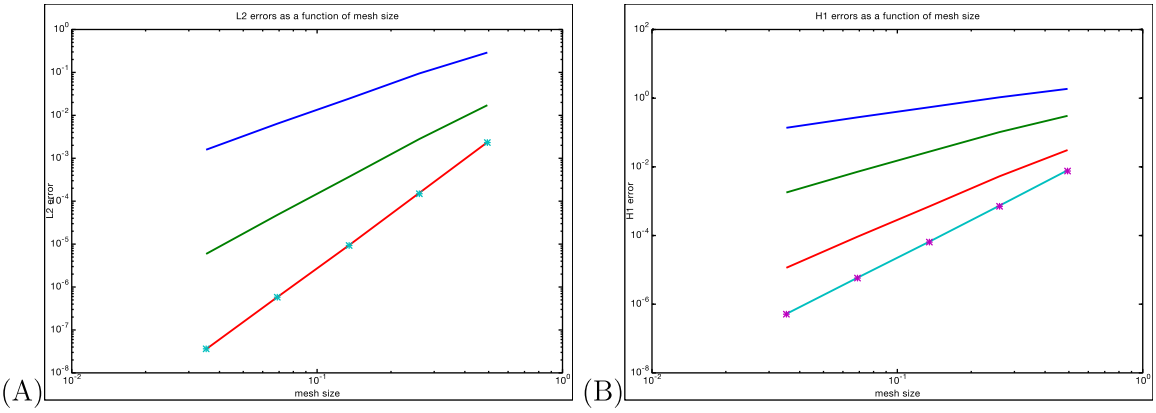

Figure 2: Errors $u_h - u_I$ in (A) $L^2(\Omega_h)$ and (B) $H^1(\Omega_h)$ as a function of the maximum mesh size for the BDT method with $\gamma = 100$. The asterisks indicate data for (A) $k = 4$ and (B) $k = 5$. The interpolant u_I is defined in (16).

4 A new method based on a Robin-type approach

One issue with the BDT method is that the resulting linear system is not symmetric, although it is easily symmetrized as we discuss in Section 11. Here we develop a technique that leads directly to a symmetric system. More importantly, this method does not require the parameter(s) from Nitsche's method. For Nitsche's method to succeed, γ must be chosen appropriately [34]. Naturally, there is a price to pay for having a parameter-free method. We will have to make some restrictive assumptions about the domain boundary that would not be required for the success of BDT.

We first separate $\partial\Omega$ into its piecewise linear part and its curvilinear part. We will assume that

$$\partial\Omega = \Gamma^0 \cup S_1 \cup \ldots \cup S_{\varkappa+1}, \quad \varkappa \geqslant 0, \tag{7}$$

where Γ^0 is a finite union of piecewise linear segments and the S_i's are C^2 and nowhere linear. The rectilinear part Γ_0 of the boundary may be empty, as is the cases for our numerical examples. Denote by

$$\mathbf{y}_i, \quad i = 1, \ldots, \varkappa, \tag{8}$$

the intersection points (if any) where S_i and S_{i+1} meet. We make the following assumption on the domain.

Assumption 1. We assume that the curvature is nonzero (and thus of one sign) in the interior of each S_i.

In Figure 3, we show a domain with two S_i's and exactly one point $\mathbf{y}_i$, where the two circles meet tangentially, the point (2, 1). The domain may be written as

$$\Omega = \big([0,2] \times [0,3]\big) \cup D(2,2) \backslash D(2,0), \tag{9}$$

where $D(x, y)$ denotes the unit disc centered at (x, y). The dashed lines indicate this construction via constructive solid geometry.

It may be that different S_i are disconnected and have no end points, as in our example in Section 9.2. So the set of $\mathbf{y}_i$'s could be empty ($\varkappa = 0$).

For the method in this section we assume that the vertices of Ω_h belong to $\partial\Omega$, and hence Ω_h might not be a subdomain of Ω. Thus, we need to define δ in this case. We assume that for every $\mathbf{x} \in \partial\Omega_h$ there is a unique smallest number $\delta(\mathbf{x})$ in absolute value such that

$$\mathbf{x} + \delta(\mathbf{x})\mathbf{n}(\mathbf{x}) \in \partial\Omega.$$

For $\mathbf{x} \in \Gamma^0$, we have $\delta(\mathbf{x}) = 0$ provided that $\Gamma^0 \subset \partial\Omega_h$.

We assume that the approximate domain boundary $\partial\Omega_h$ can be decomposed into three parts, as follows. Let $\mathcal{E}_h$ be the edges of $\partial\Omega_h$. Define

$$\Gamma^{\pm} = \bigcup \{e \in \mathcal{E}_h : \pm\delta|_{e^o} > 0\}, \tag{10}$$

where e^o denotes the interior of e. Let $\Gamma = \Gamma^+ \cup \Gamma^-$. We make the following assumption on the boundary approximations.

Assumption 2. We assume that the polygonal part Γ^0 of the boundary, defined in (7), is a subset of $\partial\Omega_h$. We also assume that all the vertices of $\partial\Omega_h$ belong to $\partial\Omega$ and that each $\mathbf{y}_i$ (defined just before Assumption 1) is a vertex

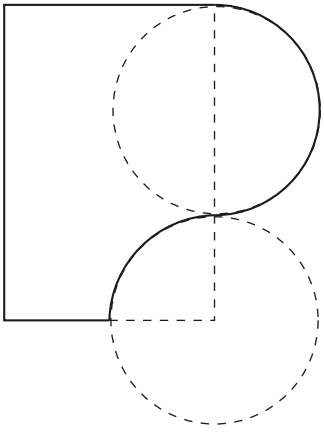

Figure 3: A P-shaped domain (solid lines) with one point $\mathbf{y}_i$. The dashed lines indicate the definition (9) via constructive solid geometry.

of $\partial\Omega_h$ and that the mesh is fine enough so that $\mathbf{y}_i$ and $\mathbf{y}_j$ are not the vertices of a single edge in $\partial\Omega_h$ when $i \neq j$. Finally, we assume that

$$\partial\Omega_h = \Gamma^0 \cup \Gamma.$$

This assumption means that δ cannot change sign in the interior of an edge.

Note that Γ depends on h, but we omit a subscript to simplify notation.

Our method resembles a Robin-type of boundary condition on Γ. It is similar to the closely related problem:

$$\begin{aligned} -\Delta w &= f \qquad \text{on } \Omega, \\ w &= g \qquad \text{on } \Gamma^0, \\ w + \delta\frac{\partial w}{\partial n} &= \hat{g} \qquad \text{on } \Gamma. \end{aligned}$$

Here we define

$$\hat{g}(\mathbf{x}) = g(\mathbf{x} + \delta(\mathbf{x})\mathbf{n}(\mathbf{x})) \tag{11}$$

for $\mathbf{x} \in \Gamma$. The key here is that, using that $u = g$ on $\partial\Omega$, for $\mathbf{x} \in \Gamma$ ($\mathbf{x}$ not a vertex of $\partial\Omega_h$) we have

$$Eu(\mathbf{x}) + \delta(\mathbf{x})\frac{\partial Eu}{\partial n}(\mathbf{x}) = \hat{g}(\mathbf{x}) + \int_0^{\delta(\mathbf{x})} (s - \delta(\mathbf{x}))\frac{\partial^2 Eu}{\partial n^2}(\mathbf{x} + s\mathbf{n})\,\mathrm{d}s. \tag{12}$$

Here Eu denotes an extension of u outside of Ω to allow for the possibility that $\Omega_h \not\subset \Omega$. The function $\mathcal{S}$ defined by

$$\mathcal{S}(\mathbf{x}) = \delta(\mathbf{x})^{-1}(Eu(\mathbf{x}) - \hat{g}(\mathbf{x})) = -\frac{\partial Eu}{\partial n}(\mathbf{x}) + \delta^{-1}(\mathbf{x})\int_0^{\delta(\mathbf{x})} (s - \delta(\mathbf{x}))\frac{\partial^2 Eu}{\partial n^2}(\mathbf{x} + s\mathbf{n})\,\mathrm{d}s \tag{13}$$

is piecewise smooth for smooth u and will play a significant role in the study of quadrature error in Section 6.3. We postpone to Section 13 a discussion of the smoothness of $\mathcal{S}$. But suffice it to say that

$$\mathcal{S}(\mathbf{x}) = -\frac{\partial Eu}{\partial n}(\mathbf{x}) + \delta(\mathbf{x})\widehat{\mathcal{S}}(\mathbf{x}) \tag{14}$$

and $\widehat{\mathcal{S}}$ is piecewise smooth.

Now we can define the method. Assume that Ω_h is triangulated by a nondegenerate mesh, and we denote by h_Γ the maximum edge length on the boundary and by h_Ω the maximum mesh size over the entire domain. By nondegenerate, we mean the following. For each triangle T in a mesh $\mathcal{T}_h$, let $\rho^T_{\min}$ be the radius of the largest circle lying inside T and let $\rho^T_{\max}$ be the radius of the smallest circle containing T. Define

$$\rho_{\mathcal{T}_h} = \inf_{T\in\mathcal{T}_h} \frac{\rho^T_{\min}}{\rho^T_{\max}}. \tag{15}$$

A nondegenerate mesh family is one for which $\rho_{\mathcal{T}_h} \geqslant \rho > 0$ for each mesh in the family. In the following, many important constants will depend on $\rho_{\mathcal{T}_h}$, but we will suppress mentioning the dependence to simplify the discussion.

Let $\{\psi_j\}$ be the usual Lagrange nodal basis for W^k_h, and define the interpolant

$$u_I = \sum_j u(x_j)\psi_j \tag{16}$$

for any continuous function u on the closure of Ω_h.

We start by defining one finite element space we will use:

$$\mathring{V}^k_h = \left\{ v \in W^k_h : \ v = 0 \text{ on } \Gamma^0, \ v(x) = 0 \ \text{ for all vertices } x \text{ of } \partial\Omega_h \right\}, \tag{17}$$

where W_h^k is defined in (2). Also define

$$\mathring{V}_h^k(g) = \left\{ v \in W_h^k : \ v = g_I \text{ on } \Gamma^0, \ v(x) = g_I(x) \ \text{ for all vertices } x \text{ of } \partial\Omega_h \right\}, \tag{18}$$

where $g_I \in C(\partial\Omega_h)$ is a suitable approximation of g and is a piecewise polynomial of degree at most k on $\partial\Omega_h$. For example, we can take g_I to be the interpolant, defined in (16), of $\hat{g}$ defined in (11). However, note that the definition (18) does not require values of g_I except on Γ^0 and at vertices of $\partial\Omega_h$ where we do know these values.

The bilinear form for the new method is given by

$$\tilde{b}_h(u, v) := a_h(u, v) + \tilde{c}_h(u, v), \tag{19}$$

where

$$a_h(u, v) = \int_{\Omega_h} \nabla u \cdot \nabla v \, \mathrm{d}x, \qquad \tilde{c}_h(u, v) = \int_{\Gamma} \delta^{-1} uv \, \mathrm{d}s. \tag{20}$$

The form $\tilde{c}_h(\cdot, \cdot)$ depends on h because Γ depends on h. Then the method solves: Find $u_h \in \mathring{V}_h^k(g)$ such that, for all $v \in \mathring{V}_h^k$,

$$\tilde{b}_h(u_h, v) = \int_{\Omega_h} (Ef) v \, \mathrm{d}x + \tilde{c}_h(\hat{g}, v), \tag{21}$$

where Ef is an extension of f outside of Ω, not necessarily the same extension used for u. For the moment, we assume that we can compute $\tilde{c}_h(\hat{g}, v)$ exactly, based on the formula (11). But this is not a practical method because it requires us to work with a space $\mathring{V}_h^k$ whose functions are forced to vanish at prescribed points. However, it is easier to analyze this limited method and from this we learn the key issues for more complicated (and practical) versions.

Integration by parts gives

$$\tilde{b}_h(Eu, v) - \int_{\Omega_h} (Ef) v \, \mathrm{d}x = \int_{\Omega_h \setminus \Omega} \big(-\Delta Eu - Ef \big) v \, \mathrm{d}x + \int_{\Gamma} \delta^{-1} (Eu) v - \frac{\partial Eu}{\partial n} v \, \mathrm{d}s. \tag{22}$$

Define the extension error

$$\mathcal{E} = \| -\Delta Eu - Ef \|_{L^\infty(\Omega_h \setminus \Omega)}. \tag{23}$$

For the two-circle problem in Section 9.2, we have analytical expressions for u and f, and using these for the extensions, we get $-\Delta Eu - Ef = 0$, so the term $\mathcal{E}$ can typically be ignored. Combining (12), (21), and (22), we get

$$|\tilde{b}_h(Eu - u_h, v)| \leqslant \mathcal{E} \|v\|_{L^1(\Omega_h \setminus \Omega)} + \left| \int_{\Gamma} \delta^{-1} \left(Eu + \delta \frac{\partial Eu}{\partial n} - \hat{g} \right) v \, \mathrm{d}s \right|. \tag{24}$$

We can estimate the second term via

$$\left| \int_{\Gamma} \delta^{-1} \left(Eu + \delta \frac{\partial Eu}{\partial n} - \hat{g} \right) v \, \mathrm{d}s \right| \leqslant C h^{2-2/q} \left\| \frac{\partial^2 Eu}{\partial n^2} \right\|_{L^q(\Omega \Delta \Omega_h)} \|v\|_{L^p(\Gamma)}, \qquad \frac{1}{p} + \frac{1}{q} = 1, \tag{25}$$

where $\Omega \Delta \Omega_h = \Omega \setminus \Omega_h \cup \Omega_h \setminus \Omega$. Here, the normal direction is the normal to Ω_h. We postpone the proof of (25) to Section 12. Our method thus appears at first to be a standard variational crime with small error. However, the stability of the method is not standard.

The role of the form $\tilde{c}_h(\cdot, \cdot)$ appears to be of two parts. Due to the singularity, it forces adoption of the boundary conditions. But in addition, it forms the required correction in view of (25).

4.1 Required bounds

The stability of the method rests on some inequalities. The first pair of these express continuity and a type of coercivity on an appropriate space. Define $\mathring{\mathcal{B}}_h^k$ to be the span of the basis functions ψ_j in (16) for which the corresponding nodes x_j are on Γ but not vertices. Then $\mathring{\mathcal{B}}_h^k$ is a complementary space:

$$\mathring{V}_h^k = \mathring{W}_h^k \oplus \mathring{\mathcal{B}}_h^k. \tag{26}$$

The choice of complementary space does not affect the computational method, but we now examine its effect on the subsequent analysis of the method.

Now assume that c_h is a general bilinear form defined on Γ and that $\mathcal{B}_h^k$ is a space of functions on Γ. Assume that there is an inner-product $\langle\cdot,\cdot\rangle_c$ so that $|v|_c^2 = \langle v, v\rangle_c$ and

$$|c_h(u,v)| \leqslant C|u|_c|v|_c \tag{27}$$

for all $u, v \in \mathcal{B}_h^k$, and

$$0 < \beta \leqslant \inf_{u\in\mathcal{B}_h^k} \sup_{v\in\mathcal{B}_h^k} \frac{|c_h(u,v)|}{|u|_c|v|_c}. \tag{28}$$

Here we are thinking of general forms c_h such as $\tilde{c}_h$ and spaces $\mathcal{B}_h^k$ such as $\mathring{\mathcal{B}}_h^k$. We will later apply these ideas to other, similar, forms and spaces For example, for the form in (20), we choose

$$\langle u, v\rangle_c = \int_\Gamma |\delta|^{-1} uv \, \mathrm{d}s.$$

For the spaces $\mathring{V}_h^k$ in (17) and $\mathring{\mathcal{B}}_h^k$ in (26), we can prove (28) by taking

$$v = \begin{cases} u & \text{on } \Gamma^+, \\ -u & \text{on } \Gamma^-, \\ 0 & \text{on } \Gamma^0, \end{cases} \tag{29}$$

in which case $c_h(u,v) = |u|_c^2 = |v|_c^2$. Thus for the form in (20) and the space $\mathring{\mathcal{B}}_h^k$ defined in (26), we have (27) with $C = 1$ and (28) with $\beta = 1$.

Now define V_h^k and $V_h^k(g)$ by

$$V_h^k = \left\{ v \in W_h^k : \, v = 0 \quad \text{on } \Gamma^0 \right\}, \qquad V_h^k(g) = \left\{ v \in W_h^k : \, v = g_I \quad \text{on } \Gamma^0 \right\}. \tag{30}$$

Define $\mathcal{B}_h^k$ to be the span of all basis functions ψ_j in (16) for which the corresponding nodes x_j are on Γ. Then $\mathcal{B}_h^k$ is a complementary space:

$$V_h^k = \mathring{W}_h^k \oplus \mathcal{B}_h^k. \tag{31}$$

However, if we do not require functions in V_h^k to vanish at vertices (see Section 4.4), then a problem can arise at a vertex where δ changes sign. If $\psi \in \mathcal{B}_h^k$ is the basis function for that vertex, then $\tilde{c}_h(\psi,\psi) \approx 0$ by symmetry. In such a case, we modify the form $c_h(\cdot,\cdot)$ as follows:

$$\bar{c}_h(u,v) = \int_\Gamma \delta^{-1} uv \, \mathrm{d}s + C\sum_i u(\mathbf{y}_i)v(\mathbf{y}_i) = \tilde{c}_h(u,v) + C\sum_i u(\mathbf{y}_i)v(\mathbf{y}_i), \tag{32}$$

where $\mathbf{y}_i$ denotes the boundary points where δ changes sign and C is a constant to be specified. Adding this sum does not change the exactness of satisfaction in (24), since u satisfies the boundary conditions exactly at all vertices of Ω_h. On the other hand, such a modification can enhance stability on the approximation space. We will show that (28) holds with suitable conditions.

The form $\bar{c}_h$ is more stable but still not quite practical, since it involves boundary integrals with a variable coefficient. We will later modify it further to use numerical quadrature.

Define the norm

$$\|v\|_a = \sqrt{a_h(v,v)}. \tag{33}$$

where the bilinear form a_h is defined in (20). The final inequality required for proving stability is a linking lemma of the form

$$\|v\|_a \leqslant c_1 h^{\ell} \, |v|_c \quad \text{for all } v \in \mathcal{B}_h^k, \tag{34}$$

for some $\ell \geqslant 0$. The latter will be proved in Lemma 2 for the form in (20) and the space $\mathcal{B}_h^k$ defined in (26) with $\ell = 1/2$.

4.2 The linking lemma

Due to the singularity of δ^{-1}, $\tilde{c}_h(u,v)$ may not be well defined for all $u, v \in \mathring{V}_h^k$. Therefore, we first show that this is not the case if we make Assumption 1.

Lemma 1. *Under Assumption* 1,

$$|\tilde{c}_h(u,v)| < \infty \quad \forall u, v \in \mathring{V}_h^k. \tag{35}$$

Proof. Assumption 1 implies that δ is non-zero in the interior of each edge. If $\delta' \neq 0$ at the end of each edge, then $\delta^{-1}v$ is bounded for all $v \in V_h^k$. The condition $\delta' \neq 0$ is obvious for edges whose vertices are in the interior of each segment S_i, since the curvature of S_i in the interior would be bounded away from zero, but at a boundary, the segment could be flat. But examining Figure 4 shows that $\delta' \neq 0$ there as well. The tangent to S_i at the left vertex has zero slope, but the chord connecting that vertex and the next has a positive slope, due to the strict monotonicity of the boundary arc. The difference of slopes is δ'. □

Assumption 1 does not hold for a domain of the form

$$\Omega_f = \{(x,y) : \, |x| < 1, \; f(x) < y < 1\} \tag{36}$$

if f is defined by

$$f(x) = \begin{cases} 0, & x \leqslant 0, \\ \sin(\pi/x)e^{-1/x}, & x > 0. \end{cases} \tag{37}$$

In particular, if we take boundary mesh points at (0,0) and $(1/n, 0)$, then the corresponding edge e with these vertices in the polygonal approximation will be on the x-axis. Thus $\delta = f$ in this interval and $c_h(v,v) < \infty$ only if v vanishes in that interval. Although we must rule out such domains for the new method, the BDT method would not be deterred by such boundaries.

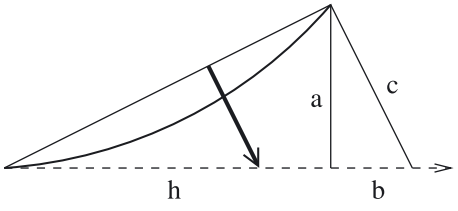

Figure 4: Bound for δ for the example in (38).

On the other hand, if f is monotone, Assumption 1 does hold for Ω_f. For f defined by

$$f(x) = \begin{cases} 0, & x \leqslant 0, \\ e^{-1/x}, & x > 0, \end{cases} \tag{38}$$

the slope of δ in $[0, 1/n]$ is e^{-n} at the origin. Therefore Assumption 1 holds for reasonable domains.

Lemma 2. *Under Assumption* 1*, there exists a constant* $c_1 > 0$ *such that*

$$\|v\|_a \leqslant c_1 \sqrt{h}\, |v|_c \quad \forall v \in \mathcal{B}_h^k. \tag{39}$$

Proof. Let $\mathcal{E}_h^\Gamma$ be the collection of edges that are a subset of Γ and let $\mathcal{T}_h^\Gamma$ be triangles T such that T has an edge in $\mathcal{E}_h^\Gamma$. Then, if $v \in \mathcal{B}_h^k$ and using inverse estimates [20] we have

$$\begin{aligned} \|v\|_a^2 &= \sum_{T \in \mathcal{T}_h^\Gamma} \|\nabla v\|_{L^2(T)}^2 \leqslant \sum_{T \in \mathcal{T}_h^\Gamma} \frac{C}{h_T^2} \|v\|_{L^2(T)}^2 \leqslant \sum_{e \in \mathcal{E}_h^\Gamma} \frac{C}{h_e} \|v\|_{L^2(e)}^2 \\ &\leqslant \sum_{e \in \mathcal{E}_h^\Gamma} \frac{C}{h_e} \Big(\max_{x \in e} |\delta(x)| \Big) \| \, |\delta|^{-1/2} v \|_{L^2(e)}^2. \end{aligned} \tag{40}$$

The result is complete after we use that

$$\max_{x \in e} |\delta(x)| \leqslant C h_e^2 \tag{41}$$

for all $e \in \mathcal{E}_h^\Gamma$, which holds since δ is essentially the error in a piecewise linear approximation of the boundary. □

4.3 Bounding δ

We will need an upper bound on δ for certain estimates. Return to the example in (36). In Figure 4, we depict a way to give a bound on δ. We have $\delta \leqslant c$, where c is the length of the edge indicated in Figure 4. Using the similarity of the triangles indicated in Figure 4, we find

$$\frac{a}{b} = \frac{h}{a} \quad \Rightarrow \quad b = \frac{a^2}{h}.$$

By the Pythagorean theorem,

$$c = \sqrt{a^2 + b^2} = a\sqrt{1 + a^2/h^2} = f(h)\sqrt{1 + f(h)^2/h^2}.$$

Therefore, we have proved the following result.

Lemma 3. *Suppose* Ω *is as in* (36) *where* $f(0) = 0$ *and* f *is strictly increasing for* $x \geqslant 0$*. Then*

$$\|\delta\|_{L^\infty([0,h])} \leqslant C \|f\|_{L^\infty([0,h])}, \tag{42}$$

where $C = \sqrt{1 + f(h)^2/h^2} \leqslant \sqrt{1 + \|f'\|_{L^\infty([0,h])}^2}$.

Thus if f is exponentially small, so is δ. More generally,

$$f^{(i)}(0) = 0 \quad \forall\, 0 \leqslant i \leqslant k \quad \Rightarrow \quad \|\delta\|_{L^\infty([0,h])} \leqslant C h^{k+1}. \tag{43}$$

More specifically, let $f(x) = x^{k+1}$, $k > 0$. Note that the slope of the line in Figure 4 is $a/h = h^k$. Then

$$\delta(x) = \left(1 + h^{2k}\right)^{-1}\left(h^k x - x^{k+1}\right), \quad x \in [0, h].$$

Therefore

$$\delta'(x) = \left(1 + h^{2k-2}\right)^{-1}\left(h^k - (k+1)x^k\right), \quad x \in [0, h].$$

Thus the maximum of δ on $[0, h]$ occurs when $x = (k+1)^{-1/k}h$, and hence

$$\|\delta\|_{L^\infty([0,h])} = \left(1 + h^{2k}\right)^{-1} h^{k+1}\left((k+1)^{-1/k} - (k+1)^{-(k+1)/k}\right). \tag{44}$$

This implies that $c_h(\cdot,\cdot)$ could be quite large in some cases. On the other hand, the bound (42) implies that we can modify certain approximations on the boundary when f is unusually small.

4.4 Relaxing the requirements

We consider three types of algorithmic modifications. The algorithm as presented so far requires the exact evaluation of boundary integrals with singular integrands. Here we consider the use of quadrature for computing the form $c_h(u, v)$. This has two benefits: it avoids the requirement for exact evaluation of boundary integrals, and it also avoids the singularities.

In addition, we allow for the approximation of δ. For example, the function δ may come from a mesh generator, and we must make provision for some inaccuracies. As a motivating example, we take

$$\delta_h = \epsilon \operatorname{sign}(\delta) + \delta, \tag{45}$$

where ϵ is either fixed or depends on h, even locally. This example has the added benefit of removing the singularity. We present computational results for this example, but our analysis applies to much more general δ_h.

Finally, we allow the boundary functions in $\mathcal{B}_h^k$ to be nonzero at vertices. This makes the implementation much easier. On the other hand, we still assume that the boundary functions in $\mathcal{B}_h^k$ vanish on the piecewise linear part Γ^0 of the boundary. If we also require boundary functions in $\mathcal{B}_h^k$ to vanish at all points $\mathbf{y}_i$, then we can use the definition of $c_h(\cdot,\cdot)$ as before. But if we want to allow them to be nonzero at such points, we need to make a modification of the boundary form as indicated in (32). More precisely, we assume that the form c_h and the related inner product and norm are now defined by

$$\begin{aligned} c_h(u, v) &= Q_\Gamma\left(\delta_h^{-1} uv\right) + \sum_i \mu_i u(\mathbf{y}_i) v(\mathbf{y}_i), \\ \langle u, v\rangle_c &= Q_\Gamma\left(|\delta_h|^{-1} uv\right) + \sum_i \mu_i u(\mathbf{y}_i) v(\mathbf{y}_i), \qquad |v|_c = \sqrt{\langle v, v\rangle_c}, \end{aligned} \tag{46}$$

where Q_Γ is a quadrature rule approximating the integral over Γ, and

$$\mu_i \geqslant 10\, Q_{E_i}\left(|\delta_h|^{-1}\psi_i^2\right), \tag{47}$$

where ψ_i is the Lagrange basis function that is 1 at $\mathbf{y}_i$ and E_i is the support of ψ_i.

Note that the analog of (27) holds:

$$|c_h(u, v)| \leqslant |u|_c |v|_c \quad \forall u, v \in \mathcal{B}_h^k. \tag{48}$$

For finite-element functions, we can evaluate the c_h form in (46) using

$$u(\mathbf{y}_i) v(\mathbf{y}_i) = Q_\Gamma\left(\chi_i^h uv\right),$$

where χ_i^h is an approximate identity centered at $\mathbf{y}_i$. This may simplify implementation on variational-form based systems such as `dolfin` [33]. The calculation of μ_i can be done in the same way.

4.5 Quadrature assumptions

Regarding the quadrature rule, we assume that

$$Q_\Gamma(f) = \sum_e Q_e(f),$$

where Q_e is a quadrature rule with positive weights approximating the integral over e without quadrature points at the vertices of e. More precisely, we assume that Q_e is generated from a single rule for [0,1] with the usual change of variables:

$$Q_{[0,1]}(f) = \sum_{i=1}^{q} \omega_i f(\xi_i), \qquad Q_{[a,b]}(f) = (b-a)\sum_{i=1}^{q} \omega_i f(a + \xi_i(b-a)).$$

Without loss of generality, we assume that $0 < \xi_1 < \xi_2 < \ldots < \xi_q < 1$ and that $\xi_1 \leqslant 1 - \xi_q$. For the approximate identity, we pick χ_i^h to satisfy

$$Q_\Gamma(\chi_i^h uv) = u(\mathbf{y}_i)v(\mathbf{y}_i) \quad \forall u, v \in \mathcal{B}_h^k.$$

In particular, we assume that χ_i^h is supported in E_i, the union of the two edges meeting at $\mathbf{y}_i$.

Let us make some assumptions about δ_h and Q_e. First of all we assume that

$$\delta\delta_h \geqslant 0, \tag{49}$$

that is, we assume that δ_h has the same sign in each element as δ. Recall that we assume that the points (8) are vertices in the mesh. For the example in (45), $\delta\delta_h = \epsilon|\delta| + \delta^2 \geqslant 0$.

Next, we assume that, for all boundary edges $e \subset \Gamma$,

$$\|(\delta - \delta_h)|\delta_h|^{-1/2}\|_{L^\infty(e)} \leqslant C_D h_e^D \tag{50}$$

for some $D \geqslant 1$. As a result,

$$\|\delta_h\|_{L^\infty(e)} \leqslant \|\delta - \delta_h\|_{L^\infty(e)} + \|\delta\|_{L^\infty(e)} \leqslant C_D h_e^D \|\delta_h\|_{L^\infty(e)}^{1/2} + C_\Gamma h_e^2,$$

for all $e \subset \Gamma$. Thus the arithmetic-geometric mean inequality implies that, for some constant C,

$$\|\delta_h\|_{L^\infty(e)} \leqslant Ch_e^2 \quad \forall e \subset \Gamma. \tag{51}$$

To get a sense of the restrictiveness of the assumption (50), consider our motivating example (45). In this case, we have $|\delta_h| \geqslant \epsilon$, and $|\delta - \delta_h| = \epsilon$. Therefore (50) holds with $C = 1$ if $\epsilon \leqslant h_e^{2D}$ for all $e \subset \Gamma$. Secondly, we assume that

$$\left| \int_e \phi \, \mathrm{d}s - Q_\Gamma(\phi) \right| \leqslant Ch^m \|\phi\|_{W^{m,1}(e)}, \quad m > 2k, \tag{52}$$

for all boundary edges e.

4.6 Revised method

We now replace Lemma 1 by the following. Recall the definitions from (30). The revised method is the following: Find $u_h \in V_h^k(g)$ such that, for all $v \in V_h^k$,

$$b_h(u_h, v) = \int_{\Omega_h} (Ef)v \, \mathrm{d}x + c_h(\hat{g}, v). \tag{53}$$

We need a replacement for Lemma 2. Recall the definition (31) of $\mathcal{B}_h^k$. Note that (52) implies

$$\int_e v^2 \, ds = Q_e(v^2) \quad \forall v \in \mathcal{B}_h^k. \tag{54}$$

Lemma 4. *Suppose that the quadrature has positive weights and that* (52) *holds. Then*

$$\|v\|_{L^2(\Gamma)} \leqslant c_1 h |v|_c, \qquad \|v\|_a \leqslant c_1 \sqrt{h} |v|_c, \tag{55}$$

for all $v \in \mathcal{B}_h^k$.

Proof. From (54),

$$\begin{aligned} \sum_{e \in \mathcal{E}_h^\Gamma} \|v\|_{L^2(e)}^2 = \sum_{e \in \mathcal{E}_h^\Gamma} Q_e(v^2) &\leqslant \sum_{e \in \mathcal{E}_h^\Gamma} \Big(\max_{x \in e} |\delta_h(x)| \Big) \, Q_e\big(|\delta_h|^{-1} v^2 \big) \\ &\leqslant \Big(\max_e \max_{x \in e} |\delta_h(x)| \Big) Q_\Gamma\big(|\delta_h|^{-1} v^2 \big). \end{aligned} \tag{56}$$

Recall from (46) that

$$Q_\Gamma\big(|\delta_h|^{-1} v^2 \big) = |v|_c^2 - \sum_i \mu_i v(\mathbf{y}_i)^2 \leqslant |v|_c^2, \tag{57}$$

since the μ_i's are positive. From the proof of Lemma 2, we have

$$\|v\|_a^2 \leqslant C h^{-1} \|v\|_{L^2(\Gamma)}^2.$$

The result follows from (51). □

We will prove (28) for the relaxed algorithm in Section 6.

5 Error analysis

5.1 Stability analysis

Unfortunately, the bilinear form b_h is not positive definite. However, we are still able to prove stability of the method.

Theorem 1. *Assume that* (27), (28), *and* (34) *hold. Suppose that G is a bounded linear functional on* V_h^k *and suppose that* $u_h \in V_h^k$ *solves*

$$b_h(u_h, v) = G(v) \quad \forall v \in V_h^k.$$

Then, assuming

$$c_1 h^\ell \leqslant \frac{1}{2} \sqrt{\beta}, \tag{58}$$

we have

$$\|u_h\|_a \leqslant \frac{4}{3} \left(\sup_{v_h \in \mathring{W}_h^k} \frac{|G(v_h)|}{\|v_h\|_a} \right) + \frac{4 c_1 h^\ell}{3\beta} \left(\sup_{v_h \in \mathcal{B}_h^k} \frac{|G(v_h)|}{|v_h|_c} \right)$$

and

$$|u_h|_c \leqslant \frac{1}{\sqrt{\beta}} \left(\sup_{v_h \in \mathring{W}_h^k} \frac{|G(v_h)|}{\|v_h\|_a} \right) + \frac{2}{\beta} \left(\sup_{v_h \in \mathcal{B}_h^k} \frac{|G(v_h)|}{|v_h|_c} \right).$$

In particular, Theorem 1 proves that the system (53) is invertible.

Proof. We can write $u_h = w_h + s_h$, where $w_h \in \mathring{W}_h^k$ and $s_h \in \mathcal{B}_h^k$. Use (28) to define $\phi_h \in \mathcal{B}_h^k$ satisfying $|\phi_h|_c = |s_h|_c$ and $\beta|s_h|_c^2 \leqslant c_h(s_h, \phi_h)$. Then

$$\beta|s_h|_c^2 \leqslant c_h(s_h, \phi_h) = b_h(u_h, \phi_h) - a_h(u_h, \phi_h) = G(\phi_h) - a_h(u_h, \phi_h).$$

Hence, we have

$$\begin{aligned}
\beta|s_h|_c^2 &\leqslant \left(\sup_{v_h \in \mathcal{B}_h^k} \frac{|G(v_h)|}{|v_h|_c}\right)|\phi_h|_c + \|u_h\|_a \|\phi_h\|_a \\
&\leqslant \left(\sup_{v_h \in \mathcal{B}_h^k} \frac{|G(v_h)|}{|v_h|_c}\right)|s_h|_c + \left(\|w_h\|_a + c_1 h^\ell |s_h|_c\right)\|\phi_h\|_a \\
&\leqslant \left(\sup_{v_h \in \mathcal{B}_h^k} \frac{|G(v_h)|}{|v_h|_c}\right)|s_h|_c + c_1 h^\ell \left(\|w_h\|_a + c_1 h^\ell |s_h|_c\right)|\phi_h|_c \\
&= \left(\sup_{v_h \in \mathcal{B}_h^k} \frac{|G(v_h)|}{|v_h|_c}\right)|s_h|_c + c_1 h^\ell \left(\|w_h\|_a + c_1 h^\ell |s_h|_c\right)|s_h|_c.
\end{aligned} \tag{59}$$

Here we used (34) twice. In particular, we used

$$\|u_h\|_a \leqslant \|w_h\|_a + \|s_h\|_a \leqslant \|w_h\|_a + c_1 h^\ell |s_h|_c.$$

Assuming $h^\ell c_1 \leqslant \frac{1}{2}\sqrt{\beta}$ we have

$$\frac{3}{4}\beta|s_h|_c^2 \leqslant \left(\sup_{v_h \in \mathcal{B}_h^k} \frac{|G(v_h)|}{|v_h|_c}\right)|s_h|_c + \frac{1}{2}\sqrt{\beta}\,\|w_h\|_a |s_h|_c.$$

Hence,

$$\beta|s_h|_c \leqslant \frac{4}{3}\left(\sup_{v_h \in \mathcal{B}_h^k} \frac{|G(v_h)|}{|v_h|_c}\right) + \frac{2}{3}\sqrt{\beta}\,\|w_h\|_a. \tag{60}$$

Next,

$$\begin{aligned}
\|w_h\|_a^2 &= a_h(w_h, w_h) = a_h(u_h, w_h) - a_h(s_h, w_h) \\
&= b_h(u_h, w_h) - a_h(s_h, w_h) = G(w_h) - a_h(s_h, w_h).
\end{aligned}$$

We therefore have

$$\|w_h\|_a^2 \leqslant \left(\sup_{v_h \in \mathring{W}_h^k} \frac{|G(v_h)|}{\|v_h\|_a}\right)\|w_h\|_a + \|s_h\|_a \|w_h\|_a.$$

Dividing by $\|w_h\|_a$ and applying (34) and (60), we obtain

$$\begin{aligned}
\|w_h\|_a &\leqslant \left(\sup_{v_h \in \mathring{W}_h^k} \frac{|G(v_h)|}{\|v_h\|_a}\right) + \|s_h\|_a \\
&\leqslant \left(\sup_{v_h \in \mathring{W}_h^k} \frac{|G(v_h)|}{\|v_h\|_a}\right) + c_1 h^\ell |s_h|_c \\
&\leqslant \left(\sup_{v_h \in \mathring{W}_h^k} \frac{|G(v_h)|}{\|v_h\|_a}\right) + \frac{4c_1 h^\ell}{3\beta}\left(\sup_{v_h \in \mathcal{B}_h^k} \frac{|G(v_h)|}{|v_h|_c}\right) + \frac{2c_1 h^\ell}{3\sqrt{\beta}}\|w_h\|_a.
\end{aligned}$$

Thus for $h^\ell c_1 \leqslant \frac{1}{2}\sqrt{\beta}$ we have

$$\begin{aligned}\|w_h\|_a &\leqslant \frac{3}{2}\left(\sup_{v_h\in\mathring{W}_h^k}\frac{|G(v_h)|}{\|v_h\|_a}\right)+\frac{2c_1h^\ell}{\beta}\left(\sup_{v_h\in\mathring{\mathcal{B}}_h^k}\frac{|G(v_h)|}{|v_h|_c}\right)\\ &\leqslant \frac{3}{2}\left(\sup_{v_h\in\mathring{W}_h^k}\frac{|G(v_h)|}{\|v_h\|_a}\right)+\frac{1}{\sqrt{\beta}}\left(\sup_{v_h\in\mathring{\mathcal{B}}_h^k}\frac{|G(v_h)|}{|v_h|_c}\right).\end{aligned} \tag{61}$$

From (60) and (61) we get

$$\beta|u_h|_c=\beta|s_h|_c\leqslant 2\left(\sup_{v_h\in\mathring{\mathcal{B}}_h^k}\frac{|G(v_h)|}{|v_h|_c}\right)+\sqrt{\beta}\left(\sup_{v_h\in\mathring{W}_h^k}\frac{|G(v_h)|}{\|v_h\|_a}\right).$$

Finally, (34) and (60) again imply

$$\begin{aligned}\|u_h\|_a &\leqslant \|w_h\|_a+\|s_h\|_a\leqslant \|w_h\|_a+c_1h^\ell|s_h|_c\\ &\leqslant \left(1+\frac{2c_1h^\ell}{3\sqrt{\beta}}\right)\|w_h\|_a+\frac{4c_1h^\ell}{3\beta}\left(\sup_{v_h\in\mathring{\mathcal{B}}_h^k}\frac{|G(v_h)|}{|v_h|_c}\right)\\ &\leqslant \frac{4}{3}\left(\sup_{v_h\in\mathring{W}_h^k}\frac{|G(v_h)|}{\|v_h\|_a}\right)+\frac{4c_1h^\ell}{3\beta}\left(\sup_{v_h\in\mathring{\mathcal{B}}_h^k}\frac{|G(v_h)|}{|v_h|_c}\right)\end{aligned}$$

for $c_1h^\ell\leqslant\frac{1}{2}\sqrt{\beta}$. □

5.2 Quasi-optimal error estimates

For clarity, we begin with the algorithm based on $\mathring{V}_h^k$ in (17) and exact quadrature for the forms in (20). Note that in this case we have the inf-sup constant $\beta=1$.

Theorem 2. *Assume Assumptions* 1 *and* 2 *hold, and assume that u solves* (1) *and that* $\frac{\partial^2 Eu}{\partial n^2}\in L^q(\Omega\Delta\Omega_h)$ *where* $1<q\leqslant\infty$. *Define*

$$r_q=\min\left\{\frac{7}{2}-\frac{2}{q},\ 4-\frac{3}{q}\right\}.$$

Let $u_h\in\mathring{V}_h^k(g)$ *solve* (21). *Then we have*

$$\|Eu-u_h\|_a\leqslant\inf_{w\in\mathring{V}_h^k(g)}\left(\left(\frac{7}{3}+\frac{4c_1^2h}{3\beta}\right)\|w-Eu\|_a+\frac{4c_1\sqrt{h}}{3\beta}|w-Eu|_c\right)+C\left(h^3\mathcal{E}+h^{r_q}K_q\right),$$

where C depends on β^{-1}, $\mathcal{E}$ *is defined in* (23), *and*

$$K_q=\left\|\frac{\partial^2 Eu}{\partial n^2}\right\|_{L^q(\Omega\Delta\Omega_h)}, \tag{62}$$

cf. (25). *Similarly,*

$$|Eu-u_h|_c\leqslant\inf_{w\in\mathring{V}_h^k(g)}\left(\left(1+\frac{1}{\beta}\right)|Eu-w|_c+\left(\frac{1}{\sqrt{\beta}}+\frac{2c_1\sqrt{h}}{\beta}\right)\|Eu-w\|_a\right)+C\left(h^{5/2}\mathcal{E}+h^{r_q-1/2}K_q\right).$$

In the ideal case when $Eu\in W^{2,\infty}(\hat{\Omega})$, the term h^{r_q} simplifies to $h^{7/2}$.

Proof. Let $w\in\mathring{V}_h^k(g)$ be arbitrary and define $e_h=w-u_h\in\mathring{V}_h^k$. Then we see that

$$b_h(e_h,v)=G(v)\quad\forall v\in\mathring{V}_h^k,$$

where $G(v) = G_1(v) + G_2(v)$, $G_1(v) = b_h(Eu - u_h, v)$ and $G_2(v) = b_h(w - Eu, v)$. We can apply Theorem 1 to get bounds for $e_h = w - u_h$ for arbitrary w in terms of bounds for the form G. Thus we need only estimate the forms G_i.

Applying (25) to (24), we find

$$|G_1(v)| \leqslant \mathcal{E}\|v\|_{L^1(\Omega_h\setminus\Omega)} + Ch^{2-2/q}K_q\|v\|_{L^p(\Gamma)}, \qquad \frac{1}{p} + \frac{1}{q} = 1, \tag{63}$$

where K_q is defined in (62). Recall that $\mathring{W}_h^k = W_h^k \cap H_0^1(\Omega_h)$. For $v \in \mathring{W}_h^k$,

$$\|v\|_{L^1(\Omega_h\setminus\Omega)} \leqslant Ch^4\|v\|_{W_\infty^1(\Omega_h\setminus\Omega)} \leqslant Ch^3\|v\|_a,$$

using a standard inverse estimate. Thus

$$\sup_{v\in\mathring{W}_h^k} \frac{|G_1(v)|}{\|v\|_a} \leqslant Ch^3\mathcal{E} \tag{64}$$

since the boundary term in (63) vanishes.

Recall the definition (26) of $\mathring{\mathcal{B}}_h^k$. Now consider $v \in \mathring{\mathcal{B}}_h^k$. For $p \geqslant 2$, inverse estimates give

$$\|v\|_{L^p(\Gamma)} \leqslant Ch^{1/p-1/2}\|v\|_{L^2(\Gamma)} \leqslant Ch^{1/p+1/2}\||\delta|^{-1/2}v\|_{L^2(\Gamma)}.$$

For $p < 2$,

$$\begin{aligned}
\|v\|_{L^p(\Gamma)}^p &= \int_\Gamma |\delta|^{p/2}|\delta|^{-p/2}|v|^p \,\mathrm{d}x \\
&\leqslant \left(\int_\Gamma |\delta|^{t'p/2}\,\mathrm{d}x\right)^{1/t'}\left(\int_\Gamma |\delta|^{-tp/2}|v|^{tp}\,\mathrm{d}x\right)^{1/t} \qquad \left[\frac{1}{t} + \frac{1}{t'} = 1\right] \\
&= \left(\int_\Gamma |\delta|^{p/(2-p)}\,\mathrm{d}x\right)^{1-p/2}\left(\int_\Gamma |\delta|^{-1}|v|^2\,\mathrm{d}x\right)^{1/t} \qquad \left[\frac{1}{t} = \frac{p}{2},\ \frac{1}{t'} = 1 - \frac{p}{2}\right].
\end{aligned}$$

Taking the p-th root, we get

$$\|v\|_{L^p(\Gamma)} \leqslant \left(\int_\Gamma |\delta|^{p/(2-p)}\,\mathrm{d}x\right)^{(2-p)/(2p)} |v|_c \leqslant Ch|v|_c.$$

Thus for general p,

$$\|v\|_{L^p(\Gamma)} \leqslant Ch^{r_p}|v|_c, \qquad r_p = \begin{cases} 1, & p \leqslant 2, \\ \dfrac{1}{p} + \dfrac{1}{2}, & p \geqslant 2. \end{cases}$$

Applying this to (63) and using (34) and inverse estimates, we get

$$\begin{aligned}
|G_1(v)| &\leqslant h^2\mathcal{E}\|v\|_{L^\infty(\Omega_h\setminus\Omega)} + h^{2-2/q+r_p}K_q|v|_c \\
&\leqslant h^2\mathcal{E}\|v\|_a + h^{2-2/q+r_p}K_q|v|_c \\
&\leqslant c_1h^{5/2}\mathcal{E}|v|_c + h^{2-2/q+r_p}K_q|v|_c,
\end{aligned}$$

where K_q is defined in (62). Note that

$$2 - \frac{2}{q} + r_p = \begin{cases} 3 - \dfrac{2}{q}, & q \geqslant 2\ (p \leqslant 2) \\ \dfrac{7}{2} - \dfrac{3}{q}, & q \leqslant 2\ (p \geqslant 2) \end{cases} = \min\left\{3 - \frac{2}{q}, \frac{7}{2} - \frac{3}{q}\right\} = r_q - \frac{1}{2}.$$

Hence,

$$h^{1/2}\left(\sup_{v\in\mathring{B}_h^k}\frac{|G_1(v)|}{|v|_c}\right)\leqslant C\big(h^3\mathcal{E}+h^{r_q}K_q\big). \tag{65}$$

Now consider G_2. If we let $v\in\mathring{W}_h^k$ then

$$G_2(v)=a_h(w-Eu,v)\leqslant\|w-Eu\|_a\|v\|_a.$$

Hence,

$$\sup_{v\in\mathring{W}_h^k}\frac{|G_2(v)|}{\|v\|_a}\leqslant\|w-Eu\|_a. \tag{66}$$

For $v\in\mathring{B}_h^k$, (34) implies

$$|G_2(v)|\leqslant\|w-Eu\|_a\|v\|_a+|w-Eu|_c|v|_c\leqslant\Big(c_1\sqrt{h}\|w-Eu\|_a+|w-Eu|_c\Big)|v|_c.$$

Therefore, we have

$$\sqrt{h}\sup_{v\in\mathring{B}_h^k}\frac{|G_2(v)|}{|v|_c}\leqslant c_1h\|w-Eu\|_a+\sqrt{h}|w-Eu|_c. \tag{67}$$

Combining (64) and (66), we get

$$\sup_{v\in\mathring{W}_h^k}\frac{|G(v_h)|}{\|v_h\|_a}\leqslant Ch^3\mathcal{E}+\|w-Eu\|_a. \tag{68}$$

Combining (65) and (67), we get

$$\sqrt{h}\sup_{v\in\mathring{B}_h^k}\frac{|G(v)|}{|v|_c}\leqslant C\Big(h^3\mathcal{E}+h^{r_q}\|u\|_{W^{2,q}(\hat{\Omega})}\Big)+c_1h\|w-Eu\|_a+\sqrt{h}|w-Eu|_c. \tag{69}$$

Applying Theorem 1, we find

$$\begin{aligned}\|w-u_h\|_a&\leqslant\frac{4}{3}\|w-Eu\|_a+\frac{4c_1}{3\beta}\Big(c_1h\|w-Eu\|_a+\sqrt{h}|w-Eu|_c\Big)+C\big(h^3\mathcal{E}+h^{r_q}K_q\big),\\ \sqrt{h}|w-u_h|_c&\leqslant\left(\frac{\sqrt{h}}{\sqrt{\beta}}+\frac{2c_1h}{\beta}\right)\|w-Eu\|_a+C\big(h^3\mathcal{E}+h^{r_q}K_q\big)+\frac{\sqrt{h}}{\beta}|w-Eu|_c,\end{aligned} \tag{70}$$

where C depends on β^{-1}. Therefore

$$\begin{aligned}\|Eu-u_h\|_a&\leqslant\|Eu-w\|_a+\|w-u_h\|_a\\ &\leqslant\left(\frac{7}{3}+\frac{4c_1^2h}{3\beta}\right)\|w-Eu\|_a+\frac{4c_1\sqrt{h}}{3\beta}|w-Eu|_c+C\big(h^3\mathcal{E}+h^{r_q}K_q\big).\end{aligned} \tag{71}$$

Similarly,

$$\begin{aligned}|Eu-u_h|_c&\leqslant|Eu-w|_c+|w-u_h|_c\\ &\leqslant\left(1+\frac{1}{\beta}\right)|Eu-w|_c+\left(\frac{1}{\sqrt{\beta}}+\frac{2c_1\sqrt{h}}{\beta}\right)\|Eu-w\|_a+C\big(h^{5/2}\mathcal{E}+h^{r_q-1/2}K_q\big).\end{aligned} \tag{72}$$

Taking the infimum over w completes the proof. □

6 Proof of inf-sup

Now we prove (28) for the relaxed algorithm. The main difficulty comes from changing signs for δ_h. On parts of the boundary where it is of one sign, we can apply the idea behind (29).

Let us number the basis functions $\psi_1, \psi_2, \dots, \psi_N$ of $\mathcal{B}_h^k$ associated with boundary nodes x_i so that $\psi_1, \psi_2, \dots, \psi_\varkappa$ are the basis functions for the boundary vertices $\mathbf{y}_i$ where δ changes sign. Note that Assumption 2 implies that h is small enough so that these are not neighboring vertices in the mesh, and thus the supports of ψ_i and ψ_j do not overlap. Write

$$u = \sum_{i=1}^{N} a_i \psi_i = s_h + \sum_{i=1}^{\varkappa} a_i \psi_i.$$

Define

$$v = \phi_h + \sum_{i=1}^{\varkappa} a_i \psi_i, \tag{73}$$

where

$$\phi_h = \begin{cases} s_h & \text{on } \{x \in \Gamma\colon\ \delta_h(x) \geqslant 0\}, \\ -s_h & \text{on } \{x \in \Gamma\colon\ \delta_h(x) \leqslant 0\}, \\ 0 & \text{on } \Gamma^0. \end{cases} \tag{74}$$

Note that

$$c_h(s_h, \phi_h) = |s_h|_c^2.$$

Expanding, we find

$$c_h(u, v) = |s_h|_c^2 + \sum_{i=1}^{\varkappa} \left(a_i^2 c_h(\psi_i, \psi_i) + a_i c_h(s_h, \psi_i) + a_i c_h(\psi_i, \phi_h) \right)$$

and

$$|u|_c^2 = \langle u, u \rangle_c = |s_h|_c^2 + \sum_{i=1}^{\varkappa} \left(a_i^2 |\psi_i|_c^2 + 2 a_i \langle s_h, \psi_i \rangle_c \right). \tag{75}$$

Therefore

$$|u|_c^2 = c_h(u, v) + \sum_{i=1}^{\varkappa} \left(a_i^2 |\psi_i|_c^2 - a_i^2 c_h(\psi_i, \psi_i) + 2 a_i \langle s_h, \psi_i \rangle_c - \left(a_i c_h(s_h, \psi_i) + a_i c_h(\psi_i, \phi_h) \right) \right). \tag{76}$$

Using the definition of ϕ_h, we have

$$\begin{aligned} \langle s_h, \psi_i \rangle_c - c_h(\phi_h, \psi_i) &= Q_{E_i}\left(|\delta_h|^{-1} \psi_i s_h \right) - Q_{E_i}\left(\delta_h^{-1} \psi_i \phi_h \right) \\ &= Q_{E_i}(|\delta_h|^{-1} \psi_i (s_h - \operatorname{sign}(\delta_h) \phi_h)) = 0, \end{aligned} \tag{77}$$

where we recall that E_i is the union of the two edges adjacent to $\mathbf{y}_i$. Therefore

$$|u|_c^2 = c_h(u, v) + \sum_{i=1}^{\varkappa} \left(a_i^2 |\psi_i|_c^2 - a_i^2 c_h(\psi_i, \psi_i) + a_i \langle s_h, \psi_i \rangle_c - a_i c_h(s_h, \psi_i) \right). \tag{78}$$

First of all,

$$|\psi_i|_c^2 - c_h(\psi_i, \psi_i) = Q_{E_i}\left(\left(|\delta_h|^{-1} - \delta_h^{-1} \right) (\psi_i)^2 \right) \leqslant 2 Q_{E_i}(|\delta_h|^{-1} (\psi_i)^2).$$

Define temporarily $q_i = Q_{E_i}(|\delta_h|^{-1} (\psi_i)^2)$. Therefore

$$a_i^2 \left(|\psi_i|_c^2 - c_h(\psi_i, \psi_i) \right) \leqslant 2 a_i^2 q_i.$$

Similarly,

$$|\langle s_h, \psi_i\rangle_c - c_h(s_h, \psi_i)| \leqslant 2|Q_{E_i}(|\delta_h|^{-1}(s_h\psi_i))|.$$

Since $s_h = u - a_i\psi_i$ on E_i, we find

$$|Q_{E_i}(|\delta_h|^{-1}(s_h\psi_i))| = |Q_{E_i}(|\delta_h|^{-1}\big(u\psi_i - a_i\psi_i^2\big)| \leqslant |Q_{E_i}(|\delta_h|^{-1}(u\psi_i)| + |a_i|q_i.$$

Using the arithmetic–geometric mean (AGM) inequality,

$$2|Q_{E_i}(|\delta_h|^{-1}(u\psi_i)| \leqslant \frac{1}{2|a_i|}Q_{E_i}\big(|\delta_h|^{-1}u^2\big) + 2|a_i|q_i.$$

Therefore

$$|a_i|\,|\langle s_h, \psi_i\rangle_c - c_h(s_h, \psi_i)| \leqslant \frac{1}{2}Q_{E_i}\big(|\delta_h|^{-1}u^2\big) + 3a_i^2q_i.$$

Thus (78) becomes

$$\begin{aligned} |u|_c^2 &\leqslant |c_h(u,v)| + \sum_{i=1}^{\varkappa}\Big(\frac{1}{2}Q_{E_i}\big(|\delta_h|^{-1}u^2\big) + 5a_i^2q_i\Big) \\ &\leqslant |c_h(u,v)| + \frac{1}{2}|u|_c^2 + \sum_{i=1}^{\varkappa}a_i^2\Big(-\frac{1}{2}\mu_i + 5q_i\Big). \end{aligned} \tag{79}$$

Thus (47) implies

$$\frac{1}{2}|u|_c^2 \leqslant |c_h(u,v)| \tag{80}$$

for v defined in (73).

We have thus shown that

$$|u|_c \leqslant 2\frac{|c_h(u,v)|}{|u|_c} \leqslant 2\frac{|c_h(u,v)|}{|v|_c}\frac{|v|_c}{|u|_c} \tag{81}$$

for v defined in (73).

Recall that E_i is the union of the two edges meeting at $\mathbf{y}_i$. We will show in Section 6.1 that

$$a_i^2|\psi_i|_c^2 = a_i^2|\psi_i|_{c,E_i}^2 \leqslant C_k|u|_{c,E_i}^2 \tag{82}$$

for a constant C_k depending only on the polynomial degree k and the shape regularity of the mesh.

Thus (82) implies that

$$\begin{aligned} |v|_c &\leqslant |\phi_h|_c + \Big|\sum_{i=1}^{\varkappa}a_i\psi_i\Big|_c = |s_h|_c + \Big|\sum_{i=1}^{\varkappa}a_i\psi_i\Big|_c \\ &\leqslant |u|_c + 2\Big|\sum_{i=1}^{\varkappa}a_i\psi_i\Big|_c \leqslant \Big(1 + 2\sqrt{C_k}\Big)|u|_c. \end{aligned} \tag{83}$$

Therefore

$$|v|_c \leqslant C|u|_c, \tag{84}$$

as required, proving the inf-sup condition (28) with $\beta = 1/2C$ by using (81).

6.1 Proof of (82)

Let us focus on a particular zero crossing i and rename basis functions as ϕ_i and renumber indices so that $i = 0$. Number the other basis functions supported in $e_\pm$ as ϕ_j, $\pm j = 1, \dots, k-1$, with $\phi_{\pm k}$ being the nearest vertex basis functions. Let e_- and e_+ be the two edges on either side of $\mathbf{y}_i$, and let $E = e_- \cup e_+$. Suppose that the edges $e_\pm$ can be parameterized by $\pm x \in [0, h_\pm]$. Suppose that

$$u|_{e_\pm} = \sum_{j=0}^{\pm k} b_j\phi_j.$$

To prove (82), we will prove more generally that

$$\sum_{j=0}^{\pm k} b_i^2 |\phi_j|^2_{c,e_\pm} \leqslant C|u|^2_{c,e_\pm} \tag{85}$$

for a constant C to be made explicit. The estimate (85) shows equivalence of two different norms on a finite dimensional space. There may be many ways to prove this, but we choose a simplification to facilitate this.

Let us introduce a simplifying assumption about δ. For each triangle T having a boundary edge $e \subset \Gamma$, choose coordinates so that e corresponds to $\{(x,0)\colon\ 0 \leqslant x \leqslant h\}$.

Assumption 3. We assume that δ is such that, for all boundary edges e,

$$|\delta(x)| \geqslant \zeta x(h-x) \quad \forall\ 0 < x < h, \tag{86}$$

where $\zeta > 0$ is independent of e and h.

By construction, the domain depicted in Figure 3 satisfies Assumption 3; to prove this, we need only to evaluate ζ for a circle of radius 1. In our example in Section 3, we examine such a circle, and we can now show that $\zeta = 1/2$. In particular, for δ as defined in (6), in appropriate coordinates we have

$$\delta((\xi+h)/2) = \sqrt{1-\xi^2} - \sqrt{1-h^2} = \left(h^2-\xi^2\right) r(\xi), \qquad |\xi| \leqslant h,$$

where $r(\xi) = \left(\sqrt{1-\xi^2} + \sqrt{1-h^2}\right)^{-1}$ is analytic for $|\xi| < 1$ and $r(\xi) \geqslant 1/2$ for $|\xi| < 1$, provided that $h \leqslant 1$.

Returning to the proof of (82), define

$$\|v\|_{Q,e_\pm} = \max_j \left| v\left(\xi_j^\pm\right)\right|. \tag{87}$$

Then

$$Q_{e_\pm}\left(|\delta_h|^{-1}\right) = \sum_j \omega_j^\pm \left|\delta_h\left(\xi_j^\pm\right)\right|^{-1} \leqslant \left(\sum_j \omega_j^\pm\right) \|\delta_h^{-1}\|_{Q,e_\pm} = h_\pm \|\delta_h^{-1}\|_{Q,e_\pm}. \tag{88}$$

If (50) holds, then for all $x \in e_\pm$ we have

$$|\delta(x)| \leqslant |\delta_h(x)| + |\delta(x) - \delta_h(x)| \leqslant |\delta_h(x)| + C_D h_{e_\pm}^D |\delta_h(x)|^{1/2} \leqslant |\delta_h(x)| + C_D h_{e_\pm}^{D+1}.$$

Therefore Assumption 3 implies

$$|\delta_h(x)| \geqslant |\delta(x)| - C_D h_{e_\pm}^{D+1} \geqslant \zeta x(h_{e_\pm} - x) - C_D h_{e_\pm}^{D+1}.$$

Thus

$$\|\delta_h^{-1}\|_{Q,e_\pm} \leqslant \left(\zeta\xi_1(1-\xi_1)h_\pm^2 - C_D h_{e_\pm}^{D+1}\right)^{-1} = h_\pm^{-2}\left(\zeta\xi_1(1-\xi_1) - C_D h_{e_\pm}^{D-1}\right)^{-1}. \tag{89}$$

If $D > 1$, then

$$\|\delta_h^{-1}\|_{Q,e_\pm} \leqslant \frac{2h_\pm^{-2}}{\zeta\xi_1(1-\xi_1)}, \tag{90}$$

provided that

$$C_D h_\pm^{D-1} \leqslant \frac{1}{2}\xi_1(1-\xi_1)\zeta. \tag{91}$$

Therefore (88) implies

$$Q_{e_\pm}\left(|\delta_h|^{-1}\right) \leqslant \frac{2h_\pm^{-1}}{\zeta\xi_1(1-\xi_1)} \tag{92}$$

for $h_\pm$ satisfying (91).

By the equivalence of norms on a finite dimensional vector space, we have

$$\sum_{j=0}^{\pm k} b_j^2 \|\phi_j\|^2_{L^2(e_\pm)} \leqslant c_k \|u\|^2_{L^2(e_\pm)}. \tag{93}$$

This is proved by scaling separately by $h_\pm$ on each interval $e_\pm$ to the unit interval, as follows. Define

$$\phi_j|_{e_\pm}(\pm x h_\pm) = \hat{\phi}_j(x) \quad \forall x \in [0,1].$$

Then c_k is chosen so that

$$\sum_{j=0}^{k} b_i^2 \|\hat{\phi}_j\|^2_{L^2([0,1])} \leqslant c_k \left\| \sum_{j=0}^{k} b_j \hat{\phi}_j \right\|^2_{L^2([0,1])}$$

for all $\{b_j\} \in \mathbb{R}^{k+1}$. This completes the proof of (93). Note that (54) implies

$$\|u\|^2_{L^2(e_\pm)} = Q_{e_\pm}(u^2) \leqslant \|\delta_h\|_{L^\infty(e_\pm)} Q_{e_\pm}\left(|\delta_h|^{-1}u^2\right) = \|\delta_h\|_{L^\infty(e_\pm)} |u|^2_{c,e_\pm}.$$

Thus (93) and (51) imply

$$\sum_{j=0}^{\pm k} b_i^2 \|\phi_j\|^2_{L^2(e_\pm)} \leqslant c_k \|\delta_h\|_{L^\infty(e_\pm)} |u|^2_{c,e_\pm} \leqslant C c_k h_\pm^2 |u|^2_{c,e_\pm}, \tag{94}$$

where C is the constant in (51). Recalling (54), (90), and the norm defined in (87), we have

$$|\phi_j|^2_{c,e_\pm} \leqslant \|\delta_h^{-1}\|_{Q,e_\pm} Q\left(\phi_j^2\right) = \|\delta_h^{-1}\|_{Q,e_\pm} \|\phi_j\|^2_{L^2(e_\pm)} \leqslant \frac{2h_\pm^{-2}}{\zeta\xi_1(1-\xi_1)} \|\phi_j\|^2_{L^2(e_\pm)}. \tag{95}$$

Combining (94) and (95) yields

$$\sum_{j=0}^{\pm k} b_i^2 |\phi_j|^2_{c,e_\pm} \leqslant \frac{2Cc_k}{\zeta\xi_1(1-\xi_1)} |u|^2_{c,e_\pm}$$

proving (85). Thus

$$a_i^2 |\psi_i|_c^2 = b_0^2 \left(|\phi_0|^2_{c,e_+} + |\phi_0|^2_{c,e_-} \right) \leqslant C_k |u|^2_{c,E},$$

where $E = e_- \cup e_+$, as claimed, with

$$C_k = \frac{4Cc_k}{\zeta\xi_1(1-\xi_1)},$$

provided that the constant D in (50) satisfies $D > 1$ and that the mesh size is small enough that (58) holds. Thus we have proved the following theorem.

Theorem 3. *Suppose that δ_h satisfies* (49) *and* (50)*, with the constant $D > 1$, that the mesh size is small enough that* (58) *holds, that the mesh is nondegenerate, that the quadrature rule satisfies* (52)*, and that Assumptions* 1–3 *hold. Then the inf-sup bound* (28) *holds.*

6.2 Estimates for the general algorithm

The main change to the arguments in the proof of Theorem 2 is that (22) and subsequent estimates need to be augmented. The replacement for (22) reads

$$b_h(Eu, v) - \int_{\Omega_h} (Ef)v \, \mathrm{d}x = \int_{\Omega_h \setminus \Omega} \left(-\Delta Eu - Ef\right) v \, \mathrm{d}x + Q_\Gamma(\delta_h^{-1}(Eu)v) - \int_\Gamma \frac{\partial Eu}{\partial n} v \, \mathrm{d}s. \tag{96}$$

Note that

$$Q_\Gamma(\delta_h^{-1}(Eu)v) - \int_\Gamma \frac{\partial Eu}{\partial n} v \, \mathrm{d}s - c_h(\hat{g}, v) = Q_\Gamma(\delta_h^{-1}(Eu - \hat{g})v) - \int_\Gamma \frac{\partial Eu}{\partial n} v \, \mathrm{d}s.$$

Thus subtracting (53) from (96) revises (24) to be

$$|G_1(v)| = |b_h(Eu - u_h, v)| \leqslant \mathcal{E}\|v\|_{L^1(\Omega_h\setminus\Omega)} + \left|\int_\Gamma \delta^{-1}(Eu-\hat{g})v\,\mathrm{d}s - Q_\Gamma(\delta_h^{-1}(Eu-\hat{g})v)\right| + Ch^{2-2/q}\|Eu\|_{W^{2,q}(\hat{\Omega})}\|v\|_{L^p(\Gamma)}, \qquad \frac{1}{p}+\frac{1}{q}=1, \tag{97}$$

where we have used (25). Therefore

$$|G_1(v)| \leqslant \mathcal{E}\|v\|_{L^1(\Omega_h\setminus\Omega)} + Q_h|v|_c + Ch^{2-2/q}\|Eu\|_{W^{2,q}(\hat{\Omega})}\|v\|_{L^p(\Gamma)}, \tag{98}$$

where the quadrature error Q_h is defined by

$$Q_h = \sup_{v\in B_h^k}\frac{1}{|v|_c}\left|\int_\Gamma \delta^{-1}(Eu-\hat{g})v\,\mathrm{d}s - Q_\Gamma(\delta_h^{-1}(Eu-\hat{g})v)\right|. \tag{99}$$

The estimate (64) is unchanged since the boundary terms again vanish. Using (55), estimate (65) becomes

$$\sqrt{h}\sup_{v\in B_h^k}\frac{|G_1(v)|}{|v|_c} \leqslant C\Big(h^3\mathcal{E} + h^{r_q}\|u\|_{W^{2,q}(\hat{\Omega})}\Big) + \sqrt{h}Q_h. \tag{100}$$

The estimates for G_2 are unchanged. Thus (68) is also unchanged. Combining (100) and (67), we get the following analog of (69):

$$\sqrt{h}\sup_{v\in B_h^k}\frac{|G(v)|}{|v|_c} \leqslant C\Big(h^3\mathcal{E} + h^{r_q}\|u\|_{W^{2,q}(\hat{\Omega})}\Big) + c_1h\|w-Eu\|_a + \sqrt{h}|w-Eu|_c + \sqrt{h}Q_h. \tag{101}$$

Applying Theorem 1, we get

$$\|Eu-u_h\|_a \leqslant C\inf_{w\in V_h^k(g)}\Big(\|w-Eu\|_a + \sqrt{h}|w-Eu|_c\Big) + C\Big(h^3\mathcal{E} + h^{r_q}\|u\|_{W^{2,q}(\hat{\Omega})} + \sqrt{h}Q_h\Big). \tag{102}$$

Similarly,

$$|Eu-u_h|_c \leqslant C\inf_{w\in V_h^k(g)}\big(|Eu-w|_c + \|Eu-w\|_a\big) + C\Big(h^{5/2}\mathcal{E} + h^{r_q-1/2}\|u\|_{W^{2,q}(\hat{\Omega})}\Big) + Q_h. \tag{103}$$

6.3 Quadrature error

The quadrature error (99) can be written in terms of the piecewise smooth function $S = \delta^{-1}(Eu-\hat{g})$ as

$$Q_h = \sup_{v\in B_h^k}\frac{1}{|v|_c}\left|\int_\Gamma Sv\,\mathrm{d}s - Q_\Gamma\big(\delta\delta_h^{-1}Sv\big)\right|. \tag{104}$$

To estimate Q_h, consider

$$\left|\int_\Gamma Sv\,\mathrm{d}s - Q_\Gamma\big(\delta\delta_h^{-1}Sv\big)\right| \leqslant \left|\int_\Gamma Sv\,\mathrm{d}s - Q_\Gamma(Sv)\right| + \left|Q_\Gamma\big(\big(1-\delta\delta_h^{-1}\big)Sv\big)\right|. \tag{105}$$

Introduce the broken Sobolev norms $\widetilde{W}_p^m(\Gamma)$ in the usual way for functions that are smooth on each element of Γ.

The first term on the right-hand side of (105) can be estimated using (52) and inverse estimates:

$$\begin{aligned}\left|\int_\Gamma \mathcal{S}v\,\mathrm{d}s - Q_\Gamma(\mathcal{S}v)\right| &\leqslant Ch^m\|\mathcal{S}v\|_{\widetilde{W}_1^m\Gamma} \leqslant Ch^m\|\mathcal{S}\|_{\widetilde{W}_\infty^m\Gamma}\|v\|_{\widetilde{W}^{k,1}(\Gamma)} \\ &\leqslant Ch^{m-k+1/2}\|\mathcal{S}\|_{\widetilde{W}^{m,\infty}(\Gamma)}|v|_{L^2(\Gamma)} \leqslant Ch^{m-k+3/2}\|\mathcal{S}\|_{\widetilde{W}^{m,\infty}(\Gamma)}|v|_c.\end{aligned} \tag{106}$$

For the second term on the right-hand side of (105), we have

$$\left|Q_\Gamma\big(\big(1-\delta\delta_h^{-1}\big)\mathcal{S}v\big)\right| = \left|Q_\Gamma((\delta_h-\delta)\delta_h^{1/2}\mathcal{S}\delta_h^{-1/2}v)\right| \leqslant C_D h^D\|\mathcal{S}\|_{L^\infty(\Gamma)}\,|v|_c. \tag{107}$$

Combining (106) and (107) yields

$$\mathcal{Q}_h \leqslant \|\delta-\delta_h\|_{L^\infty(\Gamma)}|\mathcal{S}|_c + Ch^{m-k+3/2}\|\mathcal{S}\|_{\widetilde{W}^{m,\infty}(\Gamma)}. \tag{108}$$

Combining (102), (103), and (108), we have proved the following theorem.

Theorem 4. *Assume that the inverse estimates*

$$\|v\|_{L^p(\Gamma)} \leqslant Ch^{1/p-1/2}\|v\|_{L^2(\Gamma)}, \qquad \|v\|_{W_1^k(\Gamma)} \leqslant Ch^{-k+1/2}\|v\|_{L^2(\Gamma)}$$

hold for $p \geqslant 2$, $k \geqslant 1$, *and* $v \in \mathcal{B}_h^k$. *Assume that Assumption* 2 *and* (50) *hold. Suppose that the quadrature has positive weights and that* (52) *holds, and that u solves* (1) *and that* $Eu \in W^{2,\infty}(\hat{\Omega})$. *Let* $u_h \in V_h^k(g)$ *solve* (53). *Then we have*

$$\begin{aligned}\|Eu-u_h\|_a &\leqslant \inf_{w\in V_h^k(g)}\Big(\Big(\frac{7}{3}+c_1h\Big)\|Eu-w\|_a + \sqrt{h}|Eu-w|_c\Big) \\ &\quad + C\Big(h^3\mathcal{E} + h^{7/2}\|Eu\|_{W^{2,\infty}(\hat{\Omega})}\Big) + Ch^D|\mathcal{S}|_c + Ch^{m-k+2}\|\mathcal{S}\|_{\widetilde{W}^{m,\infty}(\Gamma)}, \\ |Eu-u_h|_c &\leqslant \inf_{w\in V_h^k(g)}\Big(2|Eu-w|_c + (1+c_1\sqrt{h})\|Eu-w\|_a\Big) \\ &\quad + C\Big(h^{5/2}\mathcal{E} + h^3\|Eu\|_{W^{2,\infty}(\hat{\Omega})}\Big) + Ch^D|\mathcal{S}|_c + Ch^{m-k+2}\|\mathcal{S}\|_{\widetilde{W}^{m,\infty}(\Gamma)},\end{aligned} \tag{109}$$

where $\mathcal{E}$ *is defined in* (23) *and* $\mathcal{S}$ *is defined in* (13).

If $D \geqslant 7/2$, the estimates in (109) are as good as in (63), at least for appropriate quadrature rules. However, our computations in Section 9 suggest that there may be benefit to taking $\|\delta-\delta_h\|_{L^\infty(\Gamma)}$ even smaller.

7 Approximation results

We have proved stability and quasi-optimal approximation in certain norms, but we need to translate this into more conventional representations. A key point is that, so far, the form $c_h(v,v)$ could be arbitrarily large.

In Theorems 2 and 4 we can take $w = (Eu)_I$, but the issue is to estimate

$$\int_e \delta^{-1}(Eu-(Eu)_I)^2\,\mathrm{d}s.$$

We can write $Eu = \delta\,\mathcal{S} + \hat{g}$, and so it suffices to estimate

$$\int_e \delta^{-1}(\delta\, S - (\delta\, S)_I)^2\, \mathrm{d}s, \qquad \int_e \delta^{-1}(\hat{g} - (\hat{g})_I)^2\, \mathrm{d}s.$$

For the first of these, it is tempting to investigate how interpolation commutes with multiplication by δ. When δ is very small, we can just take $(\delta\, S)_I$ to be zero, and we get a small result. For the second, when δ is small, $\hat{g}$ is very close to g, which may provide some benefit. But first we limit our discussion to a simpler situation.

Lemma 5. *Let e denote the interval $[0, h]$. Let $\hat{\delta}(x) = x(h - x)$. Suppose that $\phi \in H_0^1(e)$. Then*

$$\sup_{r\in e} \frac{\phi(r)^2}{\hat{\delta}(r)}\, \mathrm{d}r \leqslant \frac{1}{h}\int_e \phi'(t)^2\, \mathrm{d}t \tag{110}$$

for all $\phi \in H_0^1(e)$.

Proof. We begin with $h = 1$. For $\phi \in C_c^\infty(0,1)$,

$$\phi(r)^2 = \left(\int_0^r \phi'(t)\, \mathrm{d}t\right)^2 = \left(\int_r^1 \phi'(t)\, \mathrm{d}t\right)^2.$$

Therefore

$$\phi(r)^2 = (1-r)\left(\int_0^r \phi'(t)\, \mathrm{d}t\right)^2 + r\left(\int_r^1 \phi'(t)\, \mathrm{d}t\right)^2.$$

From the Cauchy–Schwarz inequality,

$$\begin{aligned}\phi(r)^2 &\leqslant r(1-r)\left(\int_0^r \phi'(t)^2\, \mathrm{d}t\right) + r(1-r)\left(\int_r^1 \phi'(t)^2\, \mathrm{d}t\right)\\ &= r(1-r)\left(\int_0^1 \phi'(t)^2\, \mathrm{d}t\right).\end{aligned} \tag{111}$$

The lemma follows from the density of $C_c^\infty(0,1)$ in $H_0^1(0,1)$. The result for $h \neq 1$ follows by scaling variables. □

Lemma 6. *Assume that Assumption 3 holds. Then*

$$\| |\delta|^{-1/2}(u - u_I)\|_{L^2(e)} \leqslant Ch^k \|u\|_{H^{k+1}(e)},$$

where the constant C does not depend on e.

Proof. Applying Lemma 5 with $\phi = u - u_I$, we find from (86) that

$$\begin{aligned}\| |\delta|^{-1/2}(u - u_I)\|_{L^2(e)} &\leqslant |e|^{1/2} \| |\delta|^{-1/2}(u - u_I)\|_{L^\infty(e)} \leqslant \frac{1}{\sqrt{\zeta}} |u - u_I|_{H^1(e)}\\ &\leqslant Ch^k |u|_{H^{k+1}(e)},\end{aligned} \tag{112}$$

since the restriction of the interpolant to e is the interpolant of the restriction, for Lagrange elements. □

Theorem 5. *Suppose that the assumptions of Theorem 4 hold and that Assumption 3 holds. Let $u_h \in V_h^k(g)$ solve (53). Then we have*

$$\begin{aligned}\|Eu - u_h\|_a \leqslant & Ch^k\big(\|Eu\|_{H^{k+1}(\Omega)} + \|Eu\|_{H^{k+1}(\partial\Omega)}\big) \\ & + C\Big(h^3\mathcal{E} + h^{7/2}\|Eu\|_{W^{2,\infty}(\hat{\Omega})}\Big) + Ch^D|\mathcal{S}|_c + Ch^{m-k+2}\|\mathcal{S}\|_{\widetilde{W}^{m,\infty}(\Gamma)}, \\ |Eu - u_h|_c \leqslant & Ch^k\big(\|Eu\|_{H^{k+1}(\Omega)} + \|Eu\|_{H^{k+1}(\partial\Omega)}\big) \\ & + C\Big(h^{5/2}\mathcal{E} + h^3\|Eu\|_{W^{2,\infty}(\hat{\Omega})}\Big) + Ch^D|\mathcal{S}|_c + Ch^{m-k+2}\|\mathcal{S}\|_{\widetilde{W}^{m,\infty}(\Gamma)},\end{aligned} \tag{113}$$

where $\mathcal{E}$ is defined in (23) and $\mathcal{S}$ is defined in (13).

If we require only $|u|_{H^{k+1}(T)}$ to be bounded, we get a sub-optimal estimate:

$$|u - u_I|_{H^1(e)} \leqslant Ch^{k-1/2}|u|_{H^{k+1}(T)}.$$

Thus

$$\| \, |\delta|^{-1/2}(u - u_I)\|_{L^2(e)} \leqslant Ch^{k-1/2}\|u\|_{H^{k+1}(T)}$$

is the best possible estimate.

8 Implementation

The modification of W_h^k to obtain the space $\mathring{V}_h^k$ of piecewise polynomials vanishing at boundary vertices is not trivial to implement in automated systems like FEniCS [33]. Thus we took the approach of modifying c_h as described in Section 4.4. The default approach to boundary integrals in `dolfin`, which it inherits from FIAT, is to choose a Gauss-type rule with order m sufficiently large that

$$\int_e v^2 \, \mathrm{d}s = Q_e(v^2) \quad \forall v \in \mathcal{B}_h^k. \tag{114}$$

We also experimented with δ_h given by (45) for various values of ϵ. The answers do not depend on ϵ for ϵ small, as indicated in Table 3. We were even able to have $\epsilon = 0$ for (45) using `dolfin`, as our estimates confirm.

Table 3: Unit disc domain. Errors $\|u_h - u_I\|_{L^2(\Omega_h)}$, $\|u_h - u_I\|_{H^1(\Omega_h)}$, and $\| \, |\delta|^{-1/2}(u_h - u_I)\|_{L^2(\partial\Omega_h)}$ as a function of ϵ and maximum mesh size (hmax) for the Robin-like approximation (21) but modified as in Section 4.4, for piecewise quadratic polynomials ($k = 2$). Key: *M* is the value of the `meshsize` input parameter to the `mshr` function `circle` used to generate the mesh; segs is the number of boundary edges.

k	*M*	segs	hmax	ϵ	L2 err	H1 err	bdry err
2	64	320	3.5e−02	1.0e−04	1.1e−03	2.1e−03	1.3e−01
2	64	320	3.5e−02	1.0e−05	1.1e−04	1.8e−03	2.5e−02
2	64	320	3.5e−02	1.0e−06	1.2e−05	1.8e−03	3.2e−03
2	64	320	3.5e−02	**1.0e−07**	6.0e−06	1.8e−03	3.2e−04
2	64	320	3.5e−02	1.0e−08	5.9e−06	1.8e−03	4.3e−05
2	64	320	3.5e−02	1.0e−10	5.9e−06	1.8e−03	3.1e−05
2	128	640	1.8e−02	1.0e−07	1.3e−06	4.4e−04	6.4e−04
2	128	640	1.8e−02	**1.0e−08**	7.3e−07	4.4e−04	6.5e−05
2	128	640	1.8e−02	1.0e−09	7.2e−07	4.4e−04	7.3e−06
2	128	640	1.8e−02	1.0e−10	7.2e−07	4.4e−04	3.9e−06
2	256	1,280	9.0e−03	**1.0e−09**	8.9e−08	1.1e−04	1.3e−05
2	256	1,280	9.0e−03	1.0e−10	8.9e−08	1.1e−04	1.3e−06

9 Computational experiments

Here we consider two examples. In the first, $\Omega_h \subset \Omega$ and $\delta > 0$. In the second, Ω is not convex and δ is of both signs.

9.1 Return to the circle

We return now to the computational test problem described in Section 3. By varying ϵ, we were able to assess the impact of an approximate δ as studied in Section 4.4. We see that there are visible effects. We have highlighted (in boldface) the smallest value of ϵ for which there is an impact in Table 3. Thus it appears that taking $\|\delta - \delta_h\|_{L^\infty(\Gamma)} \approx h^{k+1}$ is a good choice.

We see from Table 4 and Figure 5 that the $H^1(\Omega_h)$ error is optimal order for $k \leqslant 3$, consistent with Theorem 4. In these cases, the $L^2(\Omega_h)$ error is also optimal order, and the boundary error is higher order for quadratics. For $k \geqslant 4$ our numerical experiments seem to predict the error

$$\|u - u_h\|_{H^1(\Omega_h)} \approx C\big(h^{7/2} + h^k\big),$$

which coincides with Theorem 4.

It appears from Table 4 that the boundary error term

$$\| \, |\delta|^{-1/2}(u - u_h)\|_{L^2(\partial\Omega_h)} \approx Ch^3 \quad \forall k \geqslant 2,$$

which is consistent with Theorem 4.

Comparing Tables 2 and 4, we see that the errors are almost identical for degrees $k = 2$ and $k = 3$. The Robin method is only slightly less accurate for higher degrees. Note that the condition number for the Robin method can be quite large; we used direct methods to solve the linear systems in both cases.

Table 4: Unit disc domain. Errors $\|u_h - u_I\|_{L^2(\Omega_h)}$, $\|u_h - u_I\|_{H^1(\Omega_h)}$, and $\| \, |\delta^{-1/2}(u_h - u_I)\|_{L^2(\partial\Omega_h)}$ as a function of mesh size (hmax) for the method (53) for various polynomial degrees k. The fudge factor ϵ was taken to be 10^{-13}. Results were insignificantly different for smaller values, including $\epsilon = 0$. Key: M is the value of the `meshsize` input parameter to the `mshr` function `circle` used to generate the mesh. The number of boundary edges was set to $5M$, and hmax is the maximum mesh size.

k	M	hmax	L2 error	rate	H1 error	rate	bdry err	rate
1	16	0.135	0.0264	1.95	0.545	0.96	0.292	1.04
1	32	0.0688	0.00683	1.95	0.277	0.98	0.145	1.01
1	64	0.0353	0.00169	2.01	0.137	1.02	0.0724	1.00
2	16	0.135	3.71e−04	2.88	0.0278	1.90	0.00177	2.71
2	32	0.0688	4.80e−05	2.95	0.00719	1.95	2.52e−04	2.81
2	64	0.0353	5.94e−06	3.02	0.00179	2.00	3.12e−05	3.02
3	16	0.135	8.43e−06	3.94	7.07e−04	2.91	5.22e−04	2.98
3	32	0.0688	5.39e−07	3.97	9.25e−05	2.93	6.52e−05	3.00
3	64	0.0353	3.35e−08	4.00	1.15e−05	3.01	8.13e−06	3.01
4	16	0.135	8.43e−06	3.99	7.07e−05	3.45	5.34e−04	2.97
4	32	0.0688	5.27e−07	4.00	6.38e−06	3.47	6.74e−05	2.99
4	64	0.0353	3.29e−08	4.00	5.69e−07	3.49	8.47e−06	2.99
5	16	0.135	8.43e−06	3.99	6.80e−05	3.45	5.35e−04	2.97
5	32	0.0688	5.27e−07	4.00	6.11e−06	3.48	6.75e−05	2.99
5	64	0.0353	3.30e−08	4.00	5.45e−07	3.49	8.47e−06	2.99

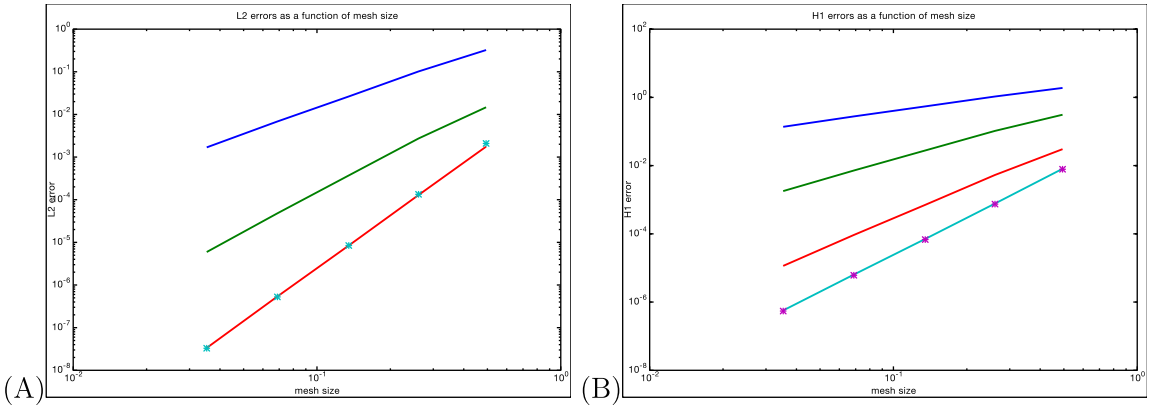

Figure 5: Errors $u_h - u_I$ in (A) $L^2(\Omega_h)$ and (B) $H^1(\Omega_h)$ as a function of the maximum mesh size for the method (21). The asterisks indicate data for (A) $k = 4$ and (B) $k = 5$.

9.2 An example with $\delta < 0$

Now consider the case where Ω is a disc of radius 1 centered at the origin, having a concentric disc of radius $R < 1$ removed.

For boundary value problem, we take $R = 1/2$ and $-\Delta u = f$, with

$$u(x, y) = (x^2 + y^2) - 5(x^2 + y^2)^2 + 4(x^2 + y^2)^3, \qquad f = -4 + 80(x^2 + y^2) - 144(x^2 + y^2)^2$$

in the computational experiments described in Table 5. Note that u vanishes on both boundary arcs, and that the computed errors are consistent with the error estimates in Theorem 4.

10 Boundary layers

It is natural to expect the error with various boundary approximations might be limited to a boundary layer, with the interior error of a smaller magnitude. Our observations indicate something like this, but the behavior is more complex. In Figure 6, we see two computations done on the same mesh based on a triangulation of Ω_h with $\partial\Omega_h$ having 80 segments and using piecewise-quadratic approximation. In Figure 6(A), we see the simple

Table 5: Disc with a disc removed. Errors $u_h - u_I$ measured in $L^2(\Omega_h)$ (L2 error), $H^1(\Omega_h)$ (H1 error), and $L^2(\partial\Omega_h)$ (bdry error) as a function of mesh size (hmax) for the method (53) for selected polynomial degrees k. Here $\epsilon = 10^{-9}$. Key: M is the value of the `meshsize` input parameter to the `mshr` function `circle` used to generate the mesh. The number of boundary edges for the outer boundary was set to $4M$, and the number of boundary edges for the inner boundary was set to $2M$.

k	*M*	hmax	L2 error	H1 error	bdry error
2	16	0.132	8.76e−04	6.87e−02	1.39e−04
2	32	0.070	1.20e−04	1.84e−02	9.64e−06
2	64	0.036	1.54e−05	4.68e−03	6.51e−07
3	16	0.132	2.90e−05	2.29e−03	6.59e−05
3	32	0.070	1.89e−06	3.07e−04	4.13e−06
3	64	0.036	1.17e−07	3.93e−05	2.47e−07
4	16	0.132	2.23e−05	3.37e−04	7.24e−05
4	32	0.070	1.39e−06	2.97e−05	4.57e−06
4	64	0.036	8.10e−08	2.61e−06	2.76e−07

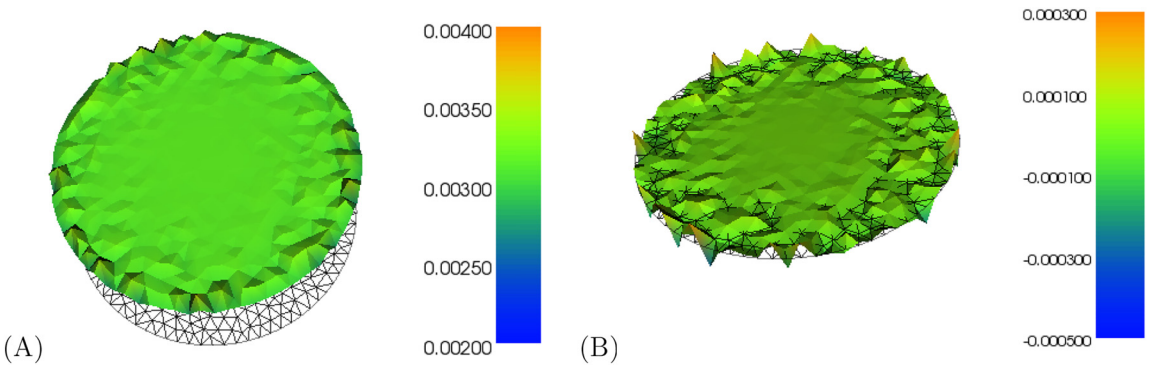

Figure 6: Error with piecewise quadratics on a mesh with $\partial\Omega_h$ having 80 segments. The mesh is drawn in the plane corresponding to zero error. (A) The method (3), no boundary integral corrections. The error is uniformly positive. (B) The Robin-like method (21). The error oscillates around zero. Note the factor of ten difference in scales in the error plots.

polygonal approximation (3). In this case, the error is somewhat larger near the boundary, but it does not decay to zero in the interior. Thus there is a significant pollution effect away from the boundary. On the other hand, Figure 6(B) shows what happens for the Robin-like method (21). Now we see that the error does decay towards zero in the interior, with the majority of the error concentrated at the boundary.

11 Higher order and symmetric methods

The Robin-type method presented in the previous section is at most of $O(h^{7/2})$. High-order methods using the same technique do not lead to symmetric systems. For simplicity assume that $g \equiv 0$. Using that

$$\left| u|_{\partial\Omega_h} + \delta \frac{\partial u}{\partial n}\bigg|_{\partial\Omega_h} + \frac{\delta^2}{2}\frac{\partial^2 u}{\partial n^2}\bigg|_{\partial\Omega_h} \right| \leqslant C\delta^3 \|u\|_{W^3_\infty(\Omega)},$$

we define

$$b_h(u,v) = a_h(u,v) + \int_{\partial\Omega_h} \delta^{-1} uv \, \mathrm{d}s + \int_{\partial\Omega_h} \frac{\delta}{2}\frac{\partial^2 u}{\partial n^2} v \, \mathrm{d}s. \tag{115}$$

Unfortunately, b_h is not symmetric.

One way to have higher-order, symmetric methods is by symmetrizing the approach of Bramble–Dupont–Thomée. Recall that Bramble et al. [8] developed arbitrary order methods, but that the bilinear forms are not symmetric. The lowest order method was presented in Section 2, where the bilinear N_h is given by (5). One way to symmetrize N_h and maintain the same convergence rates is by introducing the bilinear form:

$$M_h(u,v) = N_h(u,v) + \int_{\partial\Omega_h} \mu\delta h^{-1} \frac{\partial v}{\partial n}\left(u + \delta\frac{\partial u}{\partial n}\right) \mathrm{d}s.$$

This is precisely what is done in ref. [8, eq. (3.12)]. We see that

$$M_h(u,v) = a_h(u,v) + \int_{\partial\Omega_h} \left(\mu\frac{\delta}{h} - 1\right)\left(\delta\frac{\partial u}{\partial n}\frac{\partial v}{\partial n} + \frac{\partial u}{\partial n}v + \frac{\partial v}{\partial n}u\right) \mathrm{d}s + \frac{\mu}{h}\int_{\partial\Omega_h} uv \, \mathrm{d}s$$

is clearly symmetric. It would be very interesting if one can symmetrize even higher order methods.

12 Proof of (25)

For each edge e in the triangulation on Γ, we can choose coordinates so that the normal direction in (12) is the y-coordinate:

$$\begin{aligned}|\delta(x)|^{-1}\left|Eu(x,0)+\delta(x)\frac{\partial Eu}{\partial n}(x,0)-\hat{g}(x,0)\right| &= |\delta(x)|^{-1}\left|\int_0^{\delta(x)}(s-\delta(x))\frac{\partial^2 Eu}{\partial n^2}(x,s)\,\mathrm{d}s\right| \\ &\leqslant |\delta(x)|^{-1}\left(\int_0^{\delta(x)}|s-\delta(x)|^p\,\mathrm{d}s\right)^{1/p}\left(\int_0^{\delta(x)}\left|\frac{\partial^2 Eu}{\partial n^2}(x,s)\right|^q\mathrm{d}s\right)^{1/q} \\ &\leqslant Ch^{2-2/q}\left(\int_0^{\delta(x)}\left|\frac{\partial^2 Eu}{\partial n^2}(x,s)\right|^q\mathrm{d}s\right)^{1/q}.\end{aligned} \tag{116}$$

Here we have used the simplified notation $\delta(x,0)=\delta(x)$. Recall from (41) that $\delta=\mathcal{O}\left(h^2\right)$. Therefore

$$\begin{aligned}\left|\int_e \delta^{-1}\left(Eu+\delta\frac{\partial Eu}{\partial n}-\hat{g}\right)v\,\mathrm{d}x\right| &\leqslant Ch^{2-2/q}\int_0^h\left(\int_0^{\delta(x)}\left|\frac{\partial^2 Eu}{\partial n^2}(x,s)\right|^q\mathrm{d}s\right)^{1/q}|v(x)|\,\mathrm{d}x \\ &\leqslant Ch^{2-2/q}\left(\int_0^h\int_0^{\delta(x)}\left|\frac{\partial^2 Eu}{\partial n^2}(x,s)\right|^q\mathrm{d}s\,\mathrm{d}x\right)^{1/q}\left(\int_e|v(x)|^p\,\mathrm{d}x\right)^{1/p}.\end{aligned} \tag{117}$$

Summing over all edges e and applying Hölder's inequality one more time completes the proof of (25).

In the case that $q=\infty$, this simplifies to

$$\left|\int_e \delta^{-1}\left(Eu+\delta\frac{\partial Eu}{\partial n}-\hat{g}\right)v\,\mathrm{d}x\right| \leqslant C\left\|\frac{\partial^2 Eu}{\partial n^2}\right\|_{L^\infty(\Omega\Delta\Omega_h)}\int_e|\delta(x)|\,|v(x)|\,\mathrm{d}x,$$

and thus we see there is no singularity due to the zeroes of δ.

13 Piecewise smoothness of $\mathcal{S}$

Recall the function $\mathcal{S}$ defined in (13). For each edge e in the triangulation on Γ, we can choose coordinates so that the normal direction in (12) is the y-coordinate.

The first term is clearly smooth if Eu is smooth, so we focus on the second. Choosing $\sigma=s/\delta(\mathbf{x})$ we see from (14) that

$$\begin{aligned}\widehat{\mathcal{S}}(\mathbf{x}) &= \delta^{-2}(\mathbf{x})\int_0^{\delta(\mathbf{x})}(s-\delta(\mathbf{x}))\frac{\partial^2 Eu}{\partial n^2}(\mathbf{x},s)\,\mathrm{d}s \\ &= \int_0^1(\sigma-1)\frac{\partial^2 Eu}{\partial n^2}(\mathbf{x},\sigma\delta(\mathbf{x}))\,\mathrm{d}\sigma.\end{aligned} \tag{118}$$

Thus if Eu is smooth, then $\widehat{\mathcal{S}}$ is piecewise smooth, and thus $\mathcal{S}=-\frac{\partial Eu}{\partial n}+\delta\widehat{\mathcal{S}}$ is also piecewise smooth.

14 Conclusions and perspectives

We have presented and analyzed a parameter-free method to impose boundary conditions for the Poisson equation with curved boundaries. The analysis involves a theory for general constraints which can be applied to other methods. For example, it can be applied to Nitsche's method. In this case, the exponent ℓ in the linking lemma is zero. The stability condition in the new analysis requires $c_1 h^{\ell}$ to be sufficiently small. This can be understood as explaining why γ in Nitsche's method needs to be sufficiently large.

Assumption 3 is used in only two places, in proving (85) and in approximation results in Section 7. It is possible that this assumption can be relaxed substantially. It would also be of interest to know if the fitted-mesh requirement, that the vertices of $\partial\Omega_h$ belong to $\partial\Omega$, can be relaxed. We have not studied L^2 error estimates, although these are known for the method in ref. [8]. Similarly, we have not considered extensions to 3D, but this would also of interest. However, we have been able to extend these results to vector-valued functions in the context of the Stokes equations [35].

Acknowledgments: We thank Rob Kirby and Anders Logg for valuable information regarding quadrature in `dolfin`.
Research ethics: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors state no conflict of interest.
Research funding: Guzmán was funded by NSF DMS 2309606.
Data availability: The raw data and codes can be obtained on request from the corresponding author.

References

[1] A. Berger, R. Scott, and G. Strang, "Approximate boundary conditions in the finite element method," in *Symposia Mathematica*, vol. 10, London, Academic Press, 1972, pp. 295–313.
[2] Z. Li, T. Lin, and X. Wu, "New cartesian grid methods for interface problems using the finite element formulation," *Numer. Math.*, vol. 96, no. 1, pp. 61–98, 2003.
[3] L. R. Scott, "Finite element techniques for curved boundaries," Ph.D. dissertation, Massachusetts Institute of Technology, 1973.
[4] R. Scott, "Interpolated boundary conditions in the finite element method," *SIAM J. Numer. Anal.*, vol. 12, no. 3, pp. 404–427, 1975.
[5] J. Nitsche, "Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind," *Abh. Math. Semin. Univ. Hambg.*, vol. 36, no. 1, pp. 9–15, 1971.
[6] A. Hansbo and P. Hansbo, "An unfitted finite element method, based on Nitsche's method, for elliptic interface problems," *Comput. Methods Appl. Mech. Eng.*, vol. 191, nos. 47–48, pp. 5537–5552, 2002.
[7] E. Burman, "Ghost penalty," *C. R. Math.*, vol. 348, nos. 21–22, pp. 1217–1220, 2010.
[8] J. H. Bramble, T. Dupont, and V. Thomée, "Projection methods for Dirichlet's problem in approximating polygonal domains with boundary-value corrections," *Math. Comput.*, vol. 26, no. 120, pp. 869–879, 1972.
[9] J. H. Bramble and J. T. King, "A robust finite element method for nonhomogeneous Dirichlet problems in domains with curved boundaries," *Math. Comput.*, vol. 63, no. 207, pp. 1–17, 1994.
[10] E. Burman, P. Hansbo, and M. Larson, "A cut finite element method with boundary value correction," *Math. Comput.*, vol. 87, no. 310, pp. 633–657, 2018.
[11] E. Burman, P. Hansbo, and M. G. Larson, "Dirichlet boundary value correction using Lagrange multipliers," *BIT Numer. Math.*, vol. 60, no. 1, pp. 235–260, 2020.
[12] J. Cheung, M. Perego, P. Bochev, and M. Gunzburger, "Optimally accurate higher-order finite element methods for polytopial approximations of domains with smooth boundaries," *Math. Comput.*, vol. 88, no. 319, pp. 2187–2219, 2019.
[13] B. Cockburn, W. Qiu, and M. Solano, "A priori error analysis for HDG methods using extensions from subdomains to achieve boundary conformity," *Math. Comput.*, vol. 83, no. 286, pp. 665–699, 2014.
[14] B. Cockburn and M. Solano, "Solving dirichlet boundary-value problems on curved domains by extensions from subdomains," *SIAM J. Sci. Comput.*, vol. 34, no. 1, pp. A497–A519, 2012.
[15] T. Dupont, "L_2 error estimates for projection methods for parabolic equations in approximating domains," in *Mathematical Aspects of Finite Elements in Partial Differential Equations*, London, Elsevier, 1974, pp. 313–352.

[16] R. H. W. Hoppe, "A penalty method for the approximate solution of stationary Maxwell equations," *Numer. Math.*, vol. 36, no. 4, pp. 389—403, 1981.
[17] A. Main and G. Scovazzi, "The shifted boundary method for embedded domain computations. Part I: Poisson and Stokes problems," *J. Comput. Phys.*, vol. 372, pp. 972—995, 2018.
[18] A. Main and G. Scovazzi, "The shifted boundary method for embedded domain computations. Part II: linear advection—diffusion and incompressible Navier—Stokes equations," *J. Comput. Phys.*, vol. 372, pp. 996—1026, 2018.
[19] M. Solano and F. Vargas, "A high order HDG method for Stokes flow in curved domains," *J. Sci. Comput.*, vol. 79, no. 3, pp. 1505—1533, 2019.
[20] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, vol. 15 TAM, 3rd ed., New York, Springer Science & Business Media, 2008.
[21] J. J. Blair, "Higher order approximations to the boundary conditions for the finite element method," *Math. Comput.*, vol. 30, no. 134, pp. 250—262, 1976.
[22] A. Hansbo and P. Hansbo, "A finite element method for the simulation of strong and weak discontinuities in solid mechanics," *Comput. Methods Appl. Mech. Eng.*, vol. 193, nos. 33—35, pp. 3523—3540, 2004.
[23] S. Bertoluzza, M. Pennacchio, and D. Prada, "High order VEM on curved domains," *arXiv preprint arXiv:1811.04755*, 2018.
[24] L. B. Da Veiga, F. Brezzi, L. D. Marini, and A. Russo, "Virtual elements and curved edges," *arXiv preprint arXiv:1910.10184*, 2019.
[25] N. M. Atallah, C. Canuto, and G. Scovazzi, "Analysis of the shifted boundary method for the Stokes problem," *Comput. Methods Appl. Mech. Eng.*, vol. 358, p. 112609, 2020.
[26] N. M. Atallah, C. Canuto, and G. Scovazzi, "The second-generation shifted boundary method and its numerical analysis," *arXiv preprint arXiv:2004.10584*, 2020.
[27] N. Sánchez, T. Sánchez-Vizuet, and M. E. Solano, "Afternote to "coupling at a distance": convergence analysis and a priori error estimates," *Comput. Methods Appl. Math.*, vol. 22, no. 4, pp. 945—970, 2022.
[28] R. Oyarzúa, M. Solano, and P. Zúñiga, "A high order mixed-fem for diffusion problems on curved domains," *J. Sci. Comput.*, vol. 79, no. 1, pp. 49—78, 2019.
[29] R. Oyarzúa, M. Solano, and P. Zúñiga, "A priori and a posteriori error analyses of a high order unfitted mixed-FEM for Stokes flow," *Comput. Methods Appl. Mech. Eng.*, vol. 360, p. 112780, 2020.
[30] L. Blank, A. Caiazzo, F. Chouly, A. Lozinski, and J. Mura, "Analysis of a stabilized penalty-free Nitsche method for the Brinkman, Stokes, and Darcy problems," *ESAIM: Math. Model. Numer. Anal.*, vol. 52, no. 6, pp. 2149—2185, 2018.
[31] T. Boiveau, E. Burman, and S. Claus, "Penalty-free Nitsche method for interface problems," in *Geometrically Unfitted Finite Element Methods and Applications*, New York, Springer, 2017, pp. 183—210.
[32] E. Burman, "A penalty-free nonsymmetric Nitsche-type method for the weak imposition of boundary conditions," *SIAM J. Numer. Anal.*, vol. 50, no. 4, pp. 1959—1981, 2012.
[33] A. Logg, K. A. Mardal, and G. Wells, Eds. *Automated Solution of Differential Equations by the Finite Element Method: The FEniCS Book*, vol. 84, New York, Springer Science & Business Media, 2012.
[34] L. R. Scott, *Introduction to Automated Modeling Using FEniCS*, Cedarville, MI, Computational Modeling Initiative, 2018.
[35] F. Eickmann, L. R. Scott, and T. Tscherpel, "High-order Stokes approximation on polygonally approximated curved boundaries," *in preparation*, 2024.