4-5-2011

Stable Generalized Finite Element Method (SGFEM)

I. Babuska
University of Texas at Austin

U. Banerjee
Syracuse University

Stable Generalized Finite Element Method (SGFEM)

I. Babuška * U. Banerjee †

Abstract

The Generalized Finite Element Method (GFEM) is a Partition of Unity Method (PUM), where the trial space of standard Finite Element Method (FEM) is augmented with non-polynomial shape functions with compact support. These shape functions, which are also known as the enrichments, mimic the local behavior of the unknown solution of the underlying variational problem. GFEM has been successfully used to solve a variety of problems with complicated features and microstructure. However, the stiffness matrix of GFEM is badly conditioned (much worse compared to the standard FEM) and there could be a severe loss of accuracy in the computed solution of the associated linear system. In this paper, we address this issue and propose a modification of the GFEM, referred to as the Stable GFEM (SGFEM). We show that the conditioning of the stiffness matrix of SGFEM is not worse than that of the standard FEM. Moreover, SGFEM is very robust with respect to the parameters of the enrichments. We show these features of SGFEM on several examples.

Keywords: Generalized finite element method (GFEM); partition of unity (PU); Extended Finite Element Method (XFEM); approximation; condition number, loss of accuracy, linear system; Validation and Verification

1 Introduction

During the last decade, the Generalized Finite Element Method (GFEM) and the eXtended Finite Element Method (XFEM) – two approaches based on the Partition of Unity Method (PUM) – were developed independently and have been widely used to solve various types of problems. Only recently, it was clearly recognized that these two methods are same and were referred to as XFEM/GFEM ([23]). Hence we believe that it is interesting to briefly describe the early development of these methods. It was also recognized that, though

*ICES, University of Texas at Austin, Austin, TX.

†Department of Mathematics, 215 Carnegie, Syracuse University, Syracuse, NY 13244. E-mail address: banerjee@syr.edu. This research was partially supported by IMA, University of Minnesota, Minneapolis, MN and J. T. Oden Faculty Fellowship, ICES, University of Texas at Austin, Austin, TX.

these methods have excellent convergence properties, the stiffness matrices associated with these methods could be ill-conditioned. In this paper, we especially address this issue and propose an easy modification, which we call the Stable Generalized Finite Element Method (SGFEM), that does not have the above mentioned conditioning problem and is very robust.

We start with a brief history of the early development of the methods based on PUM. Since this is history and not a survey, it is important to provide not only the publication date, but in addition, the submission dates for various papers, and pay careful attention to nomenclature for the various methods.

Brief early history: The idea of adding non-polynomial basis functions into the trial space of the FEM started in 1970's ([10, 12, 21]). However, these basis functions had global support and the associated stiffness and mass matrices lost their local structure.

Three Special FEMs, which used non-polynomial shape functions, were proposed in [5](1994, sub: Mar.1992) to solve second order problems with rough coefficients. In particular, the shape functions used in the Special FEM #3 have compact supports and are products of piecewise linear FE hat-functions and a non-polynomial function that mimic the special features of the unknown solution. This idea was further generalized with detailed mathematical theory and applications in the Ph.D. dissertation of J. M. Melenk [32](1995), where it was shown that the hat-functions could be replaced by any PU (with compact support). This method was referred to as PUM and PUFEM in [33](1996, sub: Apr.1996) and [6](1997, sub: Jul.1995), respectively; these papers contain major results on the method and its application to the problems with highly oscillatory solutions, problems with solutions with boundary layers, differential equations with rough coefficients, etc.

The PUM was referred to as the GFEM in [47](2000, sub: Jul.1998), [48](2000, sub: Nov.1998), [49](2001, sub: Jul.2000), where the hat-functions were used as the PU (similar to the Special FEM #3 in [5]). In these papers, GFEM is interpreted as an FEM augmented with non-polynomial shape functions with compact support, and it is shown that the use of only a few of these shape functions is enough to address the problems with singular solutions. Moreover, the idea of obtaining the non-polynomial shape functions by solving certain local problems is also introduced in these papers in the context of the analysis of a perforated plate.

In a parallel development, but independent of [47], [48], [49], PUM with hat-functions serving as the PU was also investigated in [9](1999, sub: Jul.1998) and [35],(1999, sub: Feb.1999) in the context of crack propagation problems. This method is similar to GFEM as it also uses the standard FE trial space augmented with non-polynomial shape functions. However, the major contributions of these papers is to show that the method does not need remeshing as the crack propagates. Also shape functions with jump discontinuities were used in these papers. The method was first referred to as the XFEM in the Ph.D. thesis of J. Dolbow [16](1999) and almost simultaneously in [50](2000, sub: Sep.1999), [13](2000, sub: Sep.1999], and [15](2000, sub: Sep.1999). We mention that

XFEM was employed in [50, 13, 15] to address crack propagation problems in 3-d, problems with branched cracks, and fracture in Reissner-Mindlin plates.

Another idea similar to the PUM was used in the h-p Cloud method in [17](1996, sub: Jun.1995), [18](1996, sub: Apr.1996), where the shape functions were the products of a PU and polynomials. The goal of this method is to obtain h-p FEM like approximation without using a FE mesh, in the spirit of meshless methods. The use of the "customized function" (which mimicked the exact solution) for crack problems was also suggested in [38](1997, sub: Dec.1996), under this framework. Later, the hat-functions were also used as the PU in the h-p Cloud method in [39](1998, sub: Dec. 1996).

Lot of work has been done in the area of GFEM and XFEM since these early work, described above. We will comment on some of the recent developments near the end of this section.

GFEM and the problem with conditioning: PUM is a flexible framework to design Galerkin methods that accurately approximate solutions of variational problems. The framework involves (a) accurately approximating the solution, locally, using functions in a local approximation space, and (b) gluing the local approximations, using a PU, to construct a globally conforming approximate solution. The GFEM, which is a PUM with FE hat functions serving as the PU, retains the important flexibility of choosing the local approximation space. The efficiency of GFEM lies in the fact that is requires only modifying an existing FE code to incorporate special shape functions with compact support. The GFEM, with appropriate choice of special shape functions, leads to excellent convergence properties. However, the use of hat-functions as PU may result into almost linearly dependent shape functions in GFEM, and the stiffness matrix could be severely ill-conditioned; the ill-conditioning could be much worse than the conditioning of the stiffness matrix of the FEM. This results into the loss of accuracy in the solution of the linear system associated with the GFEM. In fact, the shape functions could be linearly dependent yielding a singular stiffness matrix.

Various ad-hoc approaches have been developed in the literature to address this issue. For example, the stiffness matrix of GFEM was perturbed by an identity matrix of size ϵ (small) in [47, 49] and an iterative method was used to solve the perturbed linear system. Preconditioning of the stiffness matrix, based on domain decomposition, have been recently suggested in [34] to address the conditioning problem. In [25, 43], a flat-hat PU (modified FE hat functions with flattened top) was used in the PUM instead of hat-functions. The use of flat-hat PU certainly avoids the problem of loss of accuracy in the linear system, but it requires developing a code from the scratch.

Naturally, it is pertinent to ask if GFEM could be modified so that it retains the excellent convergence properties of the GFEM, and the loss of accuracy in the computed solution of the linear system of the modified GFEM is of the same order as that of the standard FEM. In this paper, we will show that the SGFEM has both of these features. We have chosen a 1-d problem to present the idea of the SGFEM primarily for the clarity of exposition and not to obscure the

analysis with details that are not directly related to the SGFEM. However, the ideas and the associated analysis (including the notational machinery) could be easily generalized to higher dimensions and will be reported in a future publication.

Indicator of the loss of accuracy in computed solution of the linear system: Consider the linear system $Ax = b$, associated with FEM, GFEM, or SGFEM, where A is an $n \times n$ sparse symmetric positive definite matrix. Let $\hat{x}$ be the computed solution of the linear system, obtained from an elimination method encoded in a linear algebra package and the computations follow the IEEE standard for floating point arithmetic (with guard digits). Set $\eta := \|x - \hat{x}\|_2/\|x\|_2$ – the relative error that measures the loss of accuracy in the computed solution. η depends on the round-off, but in general, it also depends on the elimination algorithm and its implementation in the package, the compiler, the processor, and the computing platform with single or multiple processors. η is related to the relative error in the approximate solution due to round-off.

We seek an indicator that reliably indicates the loss of accuracy in the computed solution, characterized by η, and is practically independent of other factors mentioned above. Let $H = DAD$, where D is a diagonal matrix with $D_{ii} = A_{ii}^{-1/2}$. Define the *scaled condition number* $\mathfrak{K}(A)$ of A by $\mathfrak{K}(A) := \kappa_2(H)$, where $\kappa_2(H) = \|H\|_2\|H^{-1}\|_2$ is the condition number of H based on the $\|\cdot\|_2$ vector norm. We hypothesize that $\mathfrak{K}(A)$ is the indicator, which we formalize as follows:

Hypothesis H:

$$\eta \approx Cn^{\beta}\mathfrak{K}(A)\epsilon;\ \beta \approx 0, \tag{1.1}$$

where ϵ is the machine precision.

We will elaborate on the precise meaning of the hypothesis and *validate* it in the Appendix, borrowing the ideas from the area of *Validation and Verification*. The indicator $\mathfrak{K}(A)$ will be used to compare various GFEMs with respect to the loss of accuracy in the computed solution, which will allow us to choose a preferable GFEM. In particular, we will show in this paper that $\mathfrak{K}(A^{SGFEM}) \leq \mathfrak{K}(A^{GFEM})$, where A^{SGFEM} and A^{GFEM} are the stiffness matrices of SGFEM and GFEM, respectively, and therefore the SGFEM is preferable over the GFEM.

Some current work in GFEM/XFEM: These methods have been used in a variety of applications. For example, XFEM has been used recently to address two-phase fluid flow problems ([19]), mechanical behavior of nano-structures ([20]), and heterogeneous material with random interfaces ([36]); GFEM has been used to address heat transfer problems with sharp thermal gradient ([40]), grain boundary in polycrystals ([44]), and electromagnetic problems ([30]). Special shape functions for problems with locally periodic coefficients are constructed in [31] that yield exponential order of convergence. Also local problems to compute the shape functions for problems with rough coefficients are constructed in [3], and it has been proved that GFEM yields exponential order

of convergence. For an extensive collection of references in XFEM/GFEM, we refer to [23].

Organization of the paper: In Section 2, we give the model problem in 1-d. We intentionally chose the problem in 1-d so that we could communicate the main ideas of SGFEM, when applied to this problem, in a fairly general fashion, without the notational and other technical complexity associated with higher dimensions. We describe the PUM and GFEM, together with the approximation results, in Section 3 and show the conditioning problem in GFEM on an example. In Section 4, we first describe the SGFEM in a simpler setting, show that SGFEM retains the convergence properties of GFEM, and establish that the scaled condition numbers of the stiffness matrices of the SGFEM and FEM are of the same order. We chose the simpler setting primarily to communicate the main idea of the method and the associated analysis. We then describe the SGFEM and provide the analysis in full generality. We note that some of the ideas presented here could have been presented in a simpler fashion by using 1-d arguments. However we did not take this approach; the notations and framework of the analysis, developed in this section, could be easily generalized to higher dimensions. In Section 5, we applied SGFEM to three specific examples, namely, interface problems, problems with singular solutions, and problems with discontinuous solutions. In the Appendix, we discuss the validation of Hypothesis H and present many validation experiments. We note that the Appendix is a very important part of this paper

2 Model problem

Let $\Omega = (0,1)$ and, for an integer $k \geq 0$, we denote the standard Sobolev spaces by $H^k(\Omega)$ with the norm $\|\cdot\|_{H^k(\Omega)}$ and seminorm $|\cdot|_{H^k(\Omega)}$; for $k = 0$, $H^0(\Omega) = L^2(\Omega)$. We would also use the spaces $H^k(A)$, where A is a sub-domain of Ω. Consider the variational problem

$$u \in H^1(\Omega), \quad B(u,v) = F(v), \quad \forall\ v \in H^1(\Omega), \tag{2.1}$$

where

$$B(u,v) := \int_\Omega a u' v'\, dx \quad \text{and } F(v) := \int_\Omega f v\, dx \tag{2.2}$$

such that $F(1) = \int_\Omega f\, dx = 0$. We assume that the function $a(x)$ is bounded, i.e., there are constants α, β such that

$$0 < \alpha \leq a(x) \leq \beta, \quad \forall\ x \in \Omega \tag{2.3}$$

We note that $a(x)$ could be smooth, but it also could be rough. It is well known that the problem (2.1) has a unique solution, up to an additive constant.

We define the *Energy norm*, $\|v\|_{\mathcal{E}(A)}$, of $v \in H^1(A)$, where A is a sub-domain of Ω, by

$$\|v\|^2_{\mathcal{E}(A)} := B_A(v,v), \quad \text{where } B_A(w,z) := \int_A a w' z'\, dx.$$

It is well known that the solution u of (2.1) is also the solution of a boundary value problem (BVP), posed in the strong form as

$$-[a(x)u']' = f, \quad au'(0) = au'(1) = 0 \tag{2.4}$$

provided au' is differentiable.

3 Generalized Finite Element Method (GFEM):

Let $\mathcal{S}$ be a finite dimensional subspace of $H^1(\Omega)$. The Ritz-Galerkin method to approximate the solution u of (2.1) is given by

$$u_h \in \mathcal{S}, \quad B(u_h, v) = F(v), \quad \forall\ v \in \mathcal{S}. \tag{3.1}$$

The solution u_h is unique up to an additive constant. We can obtain a unique solution by imposing a *natural constraint* on u_h, namely, $u_h(0) = 0$.

A *Partition of Unity method* (PUM) is a Ritz-Galerkin method, where $\mathcal{S}$ is constructed employing a (a) *Partition of Unity* (PU) and (b) *Local approximating spaces*. A Generalized Finite Element method (GFEM) is a PUM with special PU. We first briefly described the PUM.

For a parameter $h > 0$, Let $I^h := \{i \in \mathbb{Z} : 0 \le i \le N\}$, where $N = N(h)$ is an integer. For $i \in I^h$, let $\omega_i^h := (a_i^h, b_i^h) \subset \Omega$ such that (i) $\Omega = \cup_{i\in I^h}\omega_i^h$, and (ii) any $x \in \Omega$ belongs to at most κ of the open intervals ω_i^h; κ is independent of i, h. The open interval ω_i^h is called a *patch*. Subordinate to the cover $\{\omega_i^h\}_{i\in I^h}$, let $\{N_i^h\}_{i\in I^h}$ be a C^0 PU satisfying

$$\sum_{i\in I^h} N_i^h(x) = 1,\ \forall\ x \in \Omega, \quad \|N_i^h\|_{L^\infty(\Omega)} \le C, \quad \mathrm{diam}\{\omega_i^h\}\|(N_i^h)'\|_{L^\infty(\Omega)} \le C,$$

where $C > 0$ is independent of i (for details, see [6, 33, 4]).

On each patch ω_i^h, $i \in I^h$, we consider an $(n_i + 1)$-dimensional space V_i^h – the *local approximating space*, namely

$$V_i^h = \mathrm{span}\{\varphi_j^{[i],h}\}_{j=0}^{n_i},\ \varphi_j^{[i],h} \in H^1(\omega_i) \text{ and } \varphi_0^{[i],h} = 1, \tag{3.2}$$

where n_is are non-negative integers. The functions $\varphi_j^{[i],h}$, $j > 0$, are carefully chosen such the functions in V_i^h mimic the the exact solution u, locally in ω_i^h. We will further comment on this issue later. In the rest of the paper, we will write I, ω_i, N_i, V_i, $\varphi_j^{[i]}$ in place of I^h, ω_i^h, N_i^h, V_i^h, $\varphi_j^{[i],h}$, respectively, with an understanding that they depend on the parameter h. The PUM is precisely (3.1), with the finite dimensional space $\mathcal{S}$ is given by

$$\mathcal{S} = \sum_{i\in I} N_i V_i = \mathrm{span}\{N_i\,\varphi_j^{[i]},\ 0 \le j \le n_i,\ i \in I\} := \mathcal{S}_1 + \mathcal{S}_2, \tag{3.3}$$

where

$$\mathcal{S}_1 = \{\zeta :\ \zeta = \sum_{i\in I} y_0^{[i]}\varphi_0^{[i]}N_i\}, \quad \mathcal{S}_2 = \{\zeta :\ \zeta = \sum_{i\in I}\sum_{j=1}^{n_i} y_j^{[i]}\varphi_j^{[i]}N_i\}, \tag{3.4}$$

and $y_0^{[i]}$, $y_j^{[i]} \in \mathbb{R}$. The functions $\varphi_j^{[i]}$, $j \geq 1$, and the associated spaces V_i are sometimes referred to as *enrichments* and *enrichment spaces* respectively in the literature. We will refer to $\mathcal{S}_1$ as the *basic part of* $\mathcal{S}$; $\mathcal{S}_2$ will be referred to as the *enrichment part* of $\mathcal{S}$. Moreover, we will refer to the Galerkin method with $\mathcal{S} = \mathcal{S}_1$ as the *basic part of PUM*. Thus every PUM has a basic part based only on the PU.

We now present the main approximation result of PUM in the Energy norm (see [6, 33, 4]).

Theorem 3.1 *Suppose $u \in H^1(\Omega)$. Suppose for $i \in I$, there exists $\xi^i \in V_i$ and $C_1 > 0$, independent of i, such that*

$$\|u - \xi^i\|_{L^2(\omega_i)} \leq C_1 \, diam(\omega_i) \, \|u - \xi^i\|_{\mathcal{E}(\omega_i)} \quad \textit{and} \quad \|u - \xi^i\|_{\mathcal{E}(\omega_i)} \leq \epsilon_i.$$

Then there exists $v \in \mathcal{S}$ such that

$$\|u - v\|_{\mathcal{E}(\Omega)} \leq C \big[\textstyle\sum_{i \in I} \epsilon_i^2 \big]^{1/2}, \tag{3.5}$$

where the positive constant C depends on $\kappa, C_1, \beta/\alpha$.

It is immediate from Theorem 3.1 that the PUM solution $u_h \in \mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$ of (3.1) satisfies

$$\|u - u_h\|_{\mathcal{E}(\Omega)} \leq \inf_{v \in \mathcal{S}} \|u - v\|_{\mathcal{E}(\Omega)} \leq C \big[\textstyle\sum_{i \in I} \epsilon_i^2 \big]^{1/2}, \tag{3.6}$$

where u is the solution of (2.1). It is clear from above that the global accuracy of the PUM solution u_h depends on how accurately the solution u of (2.1) can be approximated by the functions in V_i, locally on the patches ω_i.

We mention that in higher dimensions, the patches ω_i are subdomains, which can have quite general shape. Theorem 3.1, as presented above, is also true is higher dimensions.

We now describe the GFEM. Recall that the choice of PU in PUM is arbitrary. The GFEM is a PUM, where (a) the patches ω_i are "FE stars" relative to a finite element (FE) triangulation of Ω, and (b) the piecewise linear FE hat-functions N_i, associated with the vertices of FE triangulation, serve as the PU.

Let $N = 1/h$ and recalling $I = \{i : 0 \leq i \leq N\}$, let $\mathcal{T} := \{x_i = ih : i \in I$. Let $\{\tau_k\}_{k \in I \setminus \{0\}}$ be the *uniform mesh* on Ω, where $\tau_k := [x_{k-1}, x_k]$ are the *elements*; $\overset{o}{\tau}_k := (x_{k-1}, x_k)$ is the interior of τ_k. The points x_i are called the *vertices* of the mesh. The patches $\{\omega_i\}_{i \in I}$ are defined as $\omega_i := (x_{i-1}, x_{i+1})$, $i = 1, 2, \cdots, N-1$; also $\omega_0 := (x_0, x_1)$ and $\omega_N := (x_{N-1}, x_N)$. For $i \in I$, let N_i be the standard hat-functions associated with the vertex x_i; the support of N_i is $\overline{\omega}_i$. Note that $\overline{\omega}_0 = \tau_1$, $\overline{\omega}_N = \tau_N$ and $\overline{\omega}_i = \tau_i \cup \tau_{i+1}$ for $i = 1, 2, \ldots, N-1$. $\overline{\omega}_i$ is the FE star associated with the vertex x_i. Clearly, $\{N_i\}_{i \in I}$ form a PU subordinate to the patches $\{\omega_i\}_{i \in I}$. The associated GFEM is the Galerkin method (3.1) with $\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$ (see (3.3)). Clearly $\mathcal{S}_1$ is the standard FE space of piecewise linear functions, and consequently, the *basic part of the GFEM* is

the standard finite element method (FEM). Thus the trial space $\mathcal{S}$ of the GFEM is precisely the standard FE trial space, augmented with the space $\mathcal{S}_2$. Thus GFEM could be implemented by incorporating enrichments into an existing FE code. The name GFEM was first used in [47, 48] to highlight exactly this point. The description of GFEM is exactly same in higher dimensions; it is based on the standard FE triangulation of Ω.

Remark 3.2 We note that we have considered a uniform mesh only for the simplicity of exposition; in fact, the ideas and theory in this paper could also be presented for *locally quasi-uniform meshes*, i.e., when $C_1 \leq |\tau_{k+1}|/|\tau_k| \leq C_2$ for $k = 1, \cdots, N-1$, with $C_1, C_2 > 0$ independent of k. ∎

The accuracy of the GFEM (also PUM) solution depends on the choice of V_i, as mentioned before (see Theorem 3.1). The functions $\varphi_j^{[i]} \in V_i$ (see (3.2),(3.3)) are carefully chosen based on the available information on the unknown solution u of (2.1) to mimic the unknown solution locally in ω_i. Examples of V_i, suitable for specific applications are available in the literature (e.g., see [4]). We briefly mention some of the examples that we will consider in this paper:

• If the unknown solution u is smooth in ω_i, then the $\varphi_j^{[i]}$s are usually chosen to be polynomials in $\mathcal{P}^j(\omega_i)$ and the associated spaces V_i are spaces of polynomials of degree n_i. We note that n_i could could be different for different i, based on the available information on u.

• When $a(x)$ is a piecewise smooth and discontinuous function (interface problems), $\varphi_j^{[i]}$s are chosen such that $a[\varphi_j^{[i]}]'$ is smooth on ω_i. Clearly, $\varphi_j^{[i]}$ are continuous piecewise smooth functions with derivatives that are discontinuous at the discontinuities of $a(x)$.

• If the unknown solution u is singular, then $\varphi_j^{[i]}$ should be chosen as singular functions, mimicking the singularity of u.

• If u is discontinuous at $x = c$ in the domain, then $\varphi_j^{[i]}$s are chosen to be discontinuous functions on those ω_is that contain $x = c$. We note however that problems with discontinuous solutions cannot be cast as (2.1); we will address these problems in Section 5.3 of this paper.

Remark 3.3 GFEM provides a flexible framework to obtain various Galerkin methods. Many classical methods could be cast in this framework. For example, with $n_i = 0$ for $i \in I$ in (3.2), GFEM (with $\mathcal{S} = \mathcal{S}_1$) yields the classical FEM. Moreover, let $s(x)$ be a function (could be singular) defined on Ω. Consider $n_i = 1$ and $\varphi_1^{[i]} = s(x)|_{\omega_i}$ in the definition of $\mathcal{S}$ in (3.4). Then GFEM, with $y_1^{[i]} = b$ (a constant) for $i \in I$, yields the classical "singular FEM" (see [46, 21, 11, 12, 10, 41]), where the standard finite element trial space is augmented by the global function $s(x)$. Moreover, we note that n_i *in* (3.2) *could be different for different values of* i. In fact, one may use $n_i > 0$ only for a few patches ω_i, as needed for accuracy, based on the available information; for other patches,

$n_i = 0$, i.e., $V_i = \text{span}\{1\}$. This idea was also discussed and implemented in the original GFEM papers [47, 48].

3.1 Scaled condition number of the stiffness matrix of GFEM

The stiffness matrix $\mathbf{A}$ of the GFEM is positive semi-definite. Even when the GFEM solution is naturally constrained with $u_h(0) = 0$, i.e., when $\mathbf{A}$ is positive definite, the condition number $\kappa_2(\mathbf{A})$ can be extremely large, specifically larger than the condition number of standard FE stiffness matrix, which is $O(h^{-2})$ for second order problems. However, according to Hypothesis H, the scaled condition number $\mathfrak{K}(\mathbf{A}) = \kappa_2(H)$ is a reliable indicator of the loss of accuracy in the computed solution of $\mathbf{A}x = b$. Recall $H = D\mathbf{A}D$, where D is a diagonal matrix with $D_{ii} = \mathbf{A}_{ii}^{-1/2}$. We now present an example where $\mathfrak{K}(\mathbf{A})$ is much larger than the scaled condition number of the standard FE stiffness matrix, which is again $O(h^{-2})$.

Suppose $a(x) = 1$ in (2.2) and let $u = x^\alpha$ with $1/2 < \alpha < 3/2$, $\alpha \neq 1$. Note that $x^\alpha \in H^1(\Omega)$ but $x^\alpha \notin H^2(\Omega)$. We consider a GFEM with $n_i = 1$ and $\varphi_1^{[i]} := x^\alpha|_{\omega_i}$, $i \in I$. From the definition of $\mathcal{S}$, any $v \in \mathcal{S}$ is of the form

$$v(x) = \sum_{i \in I \setminus \{0\}} a_i N_i(x) + \sum_{i \in I} b_i N_i(x) x^\alpha; \quad a_i, b_i \in \mathbb{R}. \tag{3.7}$$

We have set $a_0 = 0$ to impose the constraint $u_h(0) = 0$ on the GFEM solution u_h. It can be easily shown that $u_h = u$, i.e. *there is no approximation error.*

We let $\eta := [a_1, \ldots, a_N, b_0, \cdots, b_N]^T \in \mathbb{R}^{2N+1}$. Then $B(v, v) = \eta^T \mathbf{A} \eta$, where $\mathbf{A}$ is the $(2N+1)\times(2N+1)$ positive definite stiffness matrix of the GFEM. We note that $\mathbf{A}_{ii} = |N_i|^2_{H^1(\omega_i)}$ for $1 \leq i \leq N$ and $\mathbf{A}_{N+1+j,N+1+j} = |N_j x^\alpha|^2_{H^1(\omega_j)}$ for $0 \leq j \leq N$. Therefore by considering $v \in \mathcal{S}$ of the form

$$v(x) = \sum_{i \in I \setminus \{0\}} a_i \frac{N_i(x)}{|N_i|_{H^1(\Omega)}} + \sum_{i \in I} b_i \frac{N_i(x) x^\alpha}{|N_i x^\alpha|_{H^1(\Omega)}}, \quad a_i, b_i \in \mathbb{R}, \tag{3.8}$$

it is easy to see that $B(v, v) = \eta^T H \eta$, where H is as mentioned before.

We consider a $v \in \mathcal{S}$ of the form (3.8) with $a_i = 0$ for $1 \leq i \leq N-1$, $a_N = 1$, and $b_i = 0$ for $i \in I$. Then

$$B(v, v) = \int_\Omega v'^2 dx = 1 \quad \text{and } \|\eta\|^2 := \sum_{i \in I \setminus \{0\}} a_i^2 + \sum_{i \in I} b_i^2 = 1.$$

Therefore,

$$\frac{B(v, v)}{\|\eta\|^2} = 1 \leq \lambda_M, \tag{3.9}$$

where λ_M is the largest eigenvalue of H.

Let $g(x) \in H^2(\Omega)$ be a non-decreasing function with $g(x) = 0$ for $0 \leq x \leq 1/4$ and $0 < C \leq g(x_i) \leq 1$ for $i \geq \lceil N/2 \rceil$. For h small enough, let

$1/8 < x_k \leq 1/4$ be the vertex closest to $x = 1/4$. Clearly x^α and gx^α are in $H^2(\widehat{\Omega})$, where $\widehat{\Omega} := (1/8, 1)$. We now consider a $v \in \mathcal{S}$ of the form (3.8) with $a_i = -g(x_i)x_i^\alpha |N_i|_{H^1(\Omega)}$ and $b_i = g(x_i)|N_i x^\alpha|_{H^1(\Omega)}$. Then

$$v(x) = -\sum_{i=k}^{N} g(x_i)x_i^\alpha N_i(x) + \sum_{i=k}^{N} g(x_i)N_i(x)x^\alpha.$$

Thus $v = 0$ on $[0, x_k]$. Moreover, on τ_i, $i \geq k+1$, we have

$$v|_{\tau_i} = -\mathcal{I}_h^i(gx^\alpha) + x^\alpha \mathcal{I}_h^i(g),$$

where $\mathcal{I}_h^i(f)$ is the linear interpolant of f on τ_i, interpolating at x_{i-1} and x_i. Therefore,

$$\begin{aligned} |v|_{H^1(\overset{o}{\tau}_i)} &= \big|gx^\alpha - \mathcal{I}_h^i(gx^\alpha) - gx^\alpha + x^\alpha \mathcal{I}_h^i(g)\big|_{H^1(\overset{o}{\tau}_i)} \\ &\leq Ch\big[\,|gx^\alpha|_{H^2(\overset{o}{\tau}_i)} + \|x^\alpha\|_{H^2(\overset{o}{\tau}_i)}|g|_{H^2(\overset{o}{\tau}_i)}\big], \end{aligned}$$

where we have used standard interpolation estimates. Thus recalling that $v = 0$ on $[0, x_k]$, we have

$$\begin{aligned} B(v,v) = |v|^2_{H^1(\Omega)} &= \sum_{i=k+1}^{N} |v|^2_{H^1(\overset{o}{\tau}_i)} \\ &\leq Ch^2\big[|gx^\alpha|^2_{H^2(\widehat{\Omega})} + \|x^\alpha\|^2_{H^1(\widehat{\Omega})}\|g\|^2_{H^1(\widehat{\Omega})}\big] := Ch^2|||gx^\alpha|||^2. \quad (3.10) \end{aligned}$$

Also,

$$\|\eta\|^2 \geq \sum_{i=k}^{N}[g(x_i)]^2|N_i x^\alpha|^2_{H^1(\Omega)} \geq \sum_{i=\lceil N/2\rceil}^{N}[g(x_i)]^2|N_i x^\alpha|^2_{H^1(\Omega)} \geq \frac{C}{h^2},$$

where we have used that $|N_i x^\alpha|^2_{H^1(\Omega)} \geq C/h$ for $i \geq \lceil N/2\rceil$. Thus using (3.10), we have

$$\frac{B(v,v)}{\|\eta\|^2} \leq Ch^4|||gx^\alpha|||^2,$$

and hence,

$$\lambda_m \leq Ch^4|||gx^\alpha|||^2$$

where λ_m is the smallest eigenvalue of H. Finally, from (3.9), we get

$$\mathfrak{K}(\mathbf{A}) = \kappa_2(H) = \frac{\lambda_M}{\lambda_m} \geq \frac{Ch^{-4}}{|||gx^\alpha|||^2},$$

which is much bigger than the scaled condition number of the stiffness matrix of the standard FEM; we recall that the standard FEM is basic part of the GFEM. Thus from Hypothesis H, there will be severe loss of accuracy in the computed solution of $\mathbf{A}x = b$. We will show this feature in the Appendix.

It is interesting to note that using $v \in \mathcal{S}$ of the form (3.7) and following the same arguments as before, we can also show that the condition number $\kappa_2(\mathbf{A}) \geq Ch^{-4}/|||gx^\alpha|||^2$. We stated this property at the beginning of this subsection.

4 Stable Generalized Finite Element Method (SGFEM):

A GFEM will be referred to as an SGFEM if the GFEM satisfies the following property: the scaled condition number $\mathfrak{K}(\mathbf{A})$ of the associated stiffness matrix $\mathbf{A}$ is of the same order with respect to h as of the stiffness matrix of the basic part of the GFEM. Since the basic part of any GFEM is the standard FEM, therefore a GFEM is an SGFEM provided $\mathfrak{K}(\mathbf{A}) = O(h^{-2})$ for second order problems. As mentioned before, we will present the analysis for uniform meshes. However, the analysis is valid for locally quasi-uniform meshes.

We first present a particular example highlighting the ideas and results related to SGFEM in a simpler setting.

4.1 An example of the SGFEM:

Let $a(x) = 1$ in (2.1) and suppose the solution of (2.1) is smooth, in particular let $u \in H^3(\Omega)$. Since the solution is unique up to an additive constant, we seek u with $u(0) = 0$. It is well known that a function in $H^3(\Omega)$ could be accurately approximated, locally in ω_i, by polynomials of degree 2; recall that the patches ω_i have been defined in Section 3. Based on this information, we consider $V_i = \text{span}\{\varphi_j^{[i]}\}_{j=0}^{2}$ (i.e., $n_i = 2$), where $\varphi_1^{[i]} = (x - x_i)$ and $\varphi_2^{[i]} = (x - x_i)^2$, for $0 \le i \le N$. Recall that $\varphi_0^{[i]} = 1$. Thus $V_i = \mathcal{P}^2(\omega_i)$.

We let

$$\overline{\varphi}_j^{[i]} := \varphi_j^{[i]} - \mathcal{I}_{\omega_i}(\varphi_j^{[i]}), \quad \text{where } \mathcal{I}_{\omega_i}(\varphi_j^{[i]}) := \sum_{1-1\le k\le i+1} \varphi_j^{[i]}(x_k) N_k \big|_{\omega_i};$$

$\mathcal{I}_{\omega_i}(\varphi_j^{[i]})$ is the piecewise linear interpolant of $\varphi_j^{[i]}$ on the patch ω_i. We adjust the operators $\mathcal{I}_{\omega_0}$ and $\mathcal{I}_{\omega_1}$; they interpolate at $\{x_0, x_1\}$ and $\{x_{N-1}, x_N\}$ respectively. We define a *modified local approximation space* $\overline{V}_i = \text{span}\{\overline{\varphi}_j^{[i]}\}_{j=0}^{2}$, associated with V_i. Clearly, $\overline{\varphi}_j^{[i]} = 0$ for $j = 0, 1$ and thus $\overline{V}_i = \text{span}\{\overline{\varphi}_2^{[i]}\}$.

It is well known (see [33, 49]) that the scaled condition number of the stiffness matrix of the GFEM, with V_i as the local approximation spaces, could be extremely large or even unbounded. We will use the GFEM with $\overline{V}_i$ precisely to address this issue, and show that the GFEM based on the approximation space

$$\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2, \quad \text{with } \mathcal{S}_1 = \sum_{i\in I\setminus\{0\}} a_i N_i \text{ and } \overline{\mathcal{S}}_2 = \sum_{i\in I} N_i \overline{V}_i$$

is an SGFEM. Note that $v(0) = 0$ for all $v \in \mathcal{S}$. We have chosen $a_0 = 0$ in the definition of $\mathcal{S}_1$ to impose the constraint $u_h(0) = 0$ to obtain a unique GFEM solution $u_h \in \mathcal{S}$.

It is easy to check that the assumptions in Theorem 3.1 hold; in fact, there exists $\xi_i \in V_i$ such that $\|u - \xi_i\|_{\mathcal{E}(\omega_i)} \le Ch^2 |u|_{H^3(\omega_i)}$. Therefore it is clear from (3.6) that $\|u - u_h\|_{\mathcal{E}(\Omega)} = \mathcal{O}(h^2)$, where u_h is the GFEM solution, based on

$\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$ (recall $\mathcal{S}_2 = \sum_{i\in I} N_i V_i$). We first show that the GFEM based on $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ also yields the same optimal order of convergence.

Proposition 4.1 *There exists a $v \in \mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$, such that*

$$\|u - v\|_{\mathcal{E}(\Omega)} \leq Ch^2 |u|_{H^3(\Omega)},$$

where the positive constant C independent of h.

Proof: Since $u \in H^3(\omega_i)$ for $0 \leq i \leq N$, it is well known that there exists $\xi^i \in V_i = \mathcal{P}^2(\omega_i)$ such that

$$\|u - \xi^i\|_{\mathcal{E}(\omega_i)} \leq Ch^2 |u|_{H^3(\omega_i)}. \tag{4.1}$$

Let $\mathcal{I}_h u = \sum_{i\in I} u(x_i) N_i$. It is clear that $\mathcal{I}_h u = \mathcal{I}_{\omega_i} u$ on ω_i, Therefore using standard interpolation results, we have

$$\begin{aligned} \|(u - \mathcal{I}_h u) - (\xi^i - \mathcal{I}_{\omega_i}\xi^i)\|_{L^2(\omega_i)} &= \|(u - \xi^i) - \mathcal{I}_{\omega_i}(u - \xi^i)\|_{L^2(\omega_i)} \\ &\leq C\mathrm{diam}(\omega_i)\, \|(u - \xi^i)\|_{\mathcal{E}(\omega_i)}, \end{aligned}$$

and similarly,

$$\|(u - \mathcal{I}_h u) - (\xi^i - \mathcal{I}_{\omega_i}\xi^i)\|_{\mathcal{E}(\omega_i)} \leq C\|u - \xi^i\|_{\mathcal{E}(\omega_i)} \leq Ch^2 |u|_{H^3(\omega_i)}.$$

Let $w := u - \mathcal{I}_h u$; clearly $w \in H^1(\Omega)$. From above, $\xi^i - \mathcal{I}_{\omega_i}\xi^i \in \overline{V}_i$ approximates w locally in ω_i. Therefore, from the Theorem 3.1, there is $\overline{v} \in \overline{\mathcal{S}}_2$ such that

$$\|w - \overline{v}\|^2_{\mathcal{E}(\Omega)} \leq C^2 \sum_{i\in I} h^4 |u|^2_{H^3(\omega_i)} \leq C^2 h^4 |u|^2_{H^3(\Omega)}. \tag{4.2}$$

Let $v = \mathcal{I}_h u - u(x_0) + \overline{v}$. Since $\{N_i\}_{i\in I}$ is a PU, we have $\mathcal{I}_h u - u(x_0) = \sum_{i\in I\setminus\{0\}} [u(x_i) - u(x_0)] N_i \in \mathcal{S}_1$. Thus $v \in \mathcal{S}$ and using (4.2), we get

$$\|u - v\|_{\mathcal{E}(\Omega)} = \|w - \overline{v}\|_{\mathcal{E}(\Omega)} \leq Ch^2 |u|_{H^3(\Omega)},$$

which is the desired result. □

Using Proposition 4.1, we immediately get that $\|u - u_h\|_{\mathcal{E}(\Omega)} = \mathcal{O}(h^2)$, where u_h is the GFEM solution based on $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$. We also note that we approximated $u - \mathcal{I}_h u$ by the functions in $\overline{V}_i$ in the proof of Proposition 4.1 – this is the main idea of SGFEM. Later, we will further comment on this issue.

We now address the scaled condition number of the stiffness matrix $\mathbf{A}$ associated with the GFEM based on $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$. With a suitable ordering of the shape function of $\mathcal{S}$, the matrix $\mathbf{A}$ is of the form

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \tag{4.3}$$

where $\mathbf{A}_{ij}$ are block matrices. The matrix $\mathbf{A}_{11} = \{B(N_i, N_j)\}_{i,j\in I\setminus\{0\}}$, which is the stiffness matrix of the basic part of GFEM, is the standard $N \times N$ FE

stiffness matrix. The $(N+1)\times(N+1)$ matrix $\mathbf{A}_{22}$ is of the form $\mathbf{A}_{22} = \{B(N_i\overline{\varphi}_2^{[i]}, N_j\overline{\varphi}_2^{[j]})\}_{i,j\in I}$. Also $\mathbf{A}_{21} = \mathbf{A}_{12}^T$. For the clarity of notation, we will write $\mathbf{A}_{22} = \{(\mathbf{A}_{22})_{ij}\}_{i,j=1}^{M}$, where $M = N+1$. Note that $(\mathbf{A}_{22})_{jj}$ are associated with the vertices x_{j-1}, respectively, and the GFEM solution u_h is computed by postprocessing. We remark that, in general, M will vary based on the application and $(\mathbf{A}_{22})_{jj}$ will be associated with some vertex $x_{i(j)}$.

We first note that $\overline{\varphi}_2^{[i]}(x_j) = 0$ for $j = i-1, i, i+1$. Therefore it is easy to show that $\mathcal{S}_1$ and $\overline{\mathcal{S}}_2$ are orthogonal in the inner product $B(\cdot,\cdot)$, i.e.,

$$B(v_1, v_2) = 0, \quad \forall\ v_1 \in \mathcal{S}_1,\ v_2 \in \overline{\mathcal{S}}_2. \tag{4.4}$$

Thus it is immediate that $\mathbf{A}_{12}$ and $\mathbf{A}_{21}$ in (4.3) are "zero-matrices".

The matrix $\mathbf{A}_{11}$ is tridiagonal and is constructed by the assembly process from the element stiffness matrices $A_{11}^{(k)}$, for the element $\tau_k = [x_{k-1}, x_k]$, $k = 1, 2, \cdots, N$. the matrices $A_{11}^{(k)}$ are given by

$$A_{11}^{(k)} = \frac{1}{h}\hat{A}_{11}^{(k)}; \quad \hat{A}_{11}^{(k)} := \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},\ 2 \le k \le N, \text{ and } \hat{A}_{11}^{(1)} := [1]. \tag{4.5}$$

Similarly, the matrix $\mathbf{A}_{22}$, which is also tridiagonal, is constructed by the assembly process from the element matrices

$$A_{22}^{(k)} = \begin{bmatrix} B_{\tau_k}(N_{k-1}\overline{\varphi}_2^{[k-1]}, N_{k-1}\overline{\varphi}_2^{[k-1]}) & B_{\tau_k}(N_k\overline{\varphi}_2^{[k]}, N_{k-1}\overline{\varphi}_2^{[k-1]}) \\ B_{\tau_k}(N_{k-1}\overline{\varphi}_2^{[k-1]}, N_k\overline{\varphi}_2^{[k]}) & B_{\tau_k}(N_k\overline{\varphi}_2^{[k]}, N_k\overline{\varphi}_2^{[k]}) \end{bmatrix}, \tag{4.6}$$

for the element τ_k, $k = 1, 2, \cdots, N$, and $B_{\tau_k}(w, v) := \int_{\tau_k} au'v'\,dx$. A direct computation yields

$$A_{22}^{(k)} = h^3\hat{A}_{22}^{(k)}; \quad \hat{A}_{22}^{(k)} = \begin{bmatrix} \frac{2}{15} & \frac{1}{30} \\ \frac{1}{30} & \frac{2}{15} \end{bmatrix}.$$

It is easy to check that the matrix $\hat{A}_{22}^{(k)}$ is positive definite (the eigenvalues are $\frac{1}{10}$ and $\frac{1}{6}$) and thus

$$\frac{h^3}{10}\|y\|^2 \le y^T A_{22}^{(k)} y \le \frac{h^3}{6}\|y\|^2, \quad \forall\ y = (y_1, y_2) \in \mathbb{R}^2. \tag{4.7}$$

We now consider the diagonal matrix $\mathbf{D} = diag(\mathbf{D}_1, \mathbf{D}_2)$ with

$$\mathbf{D}_1 = diag(\mathbf{d_1}),\ \mathbf{d_1} = m_1^{-1/2}(2^{-1/2}, \cdots, 2^{-1/2}, 1)^T \in \mathbb{R}^N,$$
$$\mathbf{D}_2 = diag(\mathbf{d_2}),\ \mathbf{d_2} = m_2^{-1/2}(1, 2^{-1/2}, \cdots, 2^{-1/2}, 1)^T \in \mathbb{R}^{N+1},$$

where $m_1 = 1/h$ and $m_2 = 2h^3/15$.

We next define

$$\widehat{\mathbf{A}} := \mathbf{DAD} = \begin{bmatrix} \mathbf{D}_1\mathbf{A}_{11}\mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2\mathbf{A}_{22}\mathbf{D}_2 \end{bmatrix} = \begin{bmatrix} \widehat{\mathbf{A}}_{11} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{A}}_{22} \end{bmatrix},$$

where $\widehat{\mathbf{A}}_{11} = \mathbf{D}_1\mathbf{A}_{11}\mathbf{D}_1$ and $\widehat{\mathbf{A}}_{22} = \mathbf{D}_2\mathbf{A}_{22}\mathbf{D}_2$. Clearly $\widehat{\mathbf{A}}_{11}$ and $\widehat{\mathbf{A}}_{22}$ are $N \times N$ and $(N+1)\times(N+1)$ tri-diagonal matrices, respectively. The diagonal elements of $\widehat{\mathbf{A}}_{11}$ and $\widehat{\mathbf{A}}_{22}$ are equal to 1. Consequently, diagonal elements of $\widehat{\mathbf{A}}$ are equal to 1 and the scaled condition number $\mathfrak{K}(\mathbf{A})$ of $\mathbf{A}$ is $\kappa_2(\widehat{\mathbf{A}})$.

Proposition 4.2 *Suppose $\mathfrak{K}(\mathbf{A})$ be the scaled condition number of $\mathbf{A}$ and let $\lambda_{min}(\widehat{\mathbf{A}}_{11})$, $\lambda_{max}(\widehat{\mathbf{A}}_{11})$ be the smallest and largest eigenvalue of $\widehat{\mathbf{A}}_{11}$, respectively. Then*

$$\mathfrak{K}(\mathbf{A}_{11}) \leq \mathfrak{K}(\mathbf{A}) \leq \mathfrak{K}(\mathbf{A}_{11})\frac{\max\{1, C_2/\lambda_{max}(\widehat{\mathbf{A}}_{11})}{\min\{1, C_1/\lambda_{min}(\widehat{\mathbf{A}}_{11})\}},$$

where $C_1 = 3/4$ and $C_2 = 5/4$.

Proof: Let $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)^T \in \mathbb{R}^{2N+1}$, where $\mathbf{z}_1 \in \mathbb{R}^N$ and $\mathbf{z}_2 \in \mathbb{R}^{N+1}$. Then

$$\begin{aligned} \mathbf{z}^T\widehat{\mathbf{A}}\mathbf{z} &= (\mathbf{D}_1\mathbf{z}_1)^T\mathbf{A}_{11}(\mathbf{D}_1\mathbf{z}_1) + (\mathbf{D}_2\mathbf{z}_2)^T\mathbf{A}_{22}(\mathbf{D}_2\mathbf{z}_2) \\ &= \mathbf{z}_1^T\widehat{\mathbf{A}}_{11}\mathbf{z}_1 + \mathbf{z}_2^T\widehat{\mathbf{A}}_{22}\mathbf{z}_2. \end{aligned} \tag{4.8}$$

Let $\mathbf{z}_2 = (y_1, y_2, \cdots, y_{N+1})^T$, then $\mathbf{D}_2\mathbf{z}_2 = m_2^{-\frac{1}{2}}(y_1, 2^{-\frac{1}{2}}y_2, \cdots, 2^{-\frac{1}{2}}y_N, y_{N+1})^T$, where $m_2 = 2h^3/15$. We define $\mathbf{z}_{2,k} := (y_k, y_{k+1})^T$, and

$$\begin{aligned} &\overline{\mathbf{z}}_{2,1} := m_2^{-1/2}(y_1, 2^{-1/2}y_2)^T, \quad \overline{\mathbf{z}}_{2,N} := m_2^{-1/2}(2^{-1/2}y_N, y_{N+1})^T, \\ &\overline{\mathbf{z}}_{2,k} := (2m_2)^{-1/2}(y_k, y_{k+1})^T, \quad \text{for } k = 2, \cdots, N-1. \end{aligned}$$

Recalling that $\mathbf{A}_{22}$ could be obtained from the element matrices $A_{22}^{(i)}$ through the assembly process, using (4.7) we get,

$$\begin{aligned} \mathbf{z}_2^T\widehat{\mathbf{A}}_{22}\mathbf{z}_2 &= (\mathbf{D}_2\mathbf{z}_2)^T\mathbf{A}_{22}(\mathbf{D}_2\mathbf{z}_2) = \sum_{k=1}^{N} \overline{\mathbf{z}}_{2,k}{}^T A_{22}^{(k)}\overline{\mathbf{z}}_{2,k} \\ &\leq \frac{h^3}{6}\sum_{k=1}^{N}\|\overline{\mathbf{z}}_{2,k}\|^2 = \frac{h^3}{6}\sum_{i=1}^{N+1} m_2^{-1}y_i^2 = \frac{5}{4}\|\mathbf{z}_2\|^2. \end{aligned}$$

Similarly from (4.7), we also get

$$\frac{3}{4}\|\mathbf{z}_2\|^2 \leq \mathbf{z}_2^T\widehat{\mathbf{A}}_{22}\mathbf{z}_2,$$

and therefore from (4.8),

$$\mathbf{z}_1^T\widehat{\mathbf{A}}_{11}\mathbf{z}_1 + C_1\|\mathbf{z}_2\|^2 \leq \mathbf{z}^T\widehat{\mathbf{A}}\mathbf{z} \leq \mathbf{z}_1^T\widehat{\mathbf{A}}_{11}\mathbf{z}_1 + C_2\|\mathbf{z}_2\|^2. \tag{4.9}$$

It is clear from above that

$$\begin{aligned} \mathbf{z}^T\widehat{\mathbf{A}}\mathbf{z} &\geq \mathbf{z}_1^T\widehat{\mathbf{A}}_{11}\mathbf{z}_1 + C_1\|\mathbf{z}_2\|^2 \\ &\geq \lambda_{min}(\widehat{\mathbf{A}}_{11})\|\mathbf{z}_1\|^2 + C_1\|\mathbf{z}_2\|^2 \\ &\geq \min\{C_1, \lambda_{min}(\widehat{\mathbf{A}}_{11})\}\|\mathbf{z}\|^2, \end{aligned}$$

where $C_1 := \frac{3}{4}$. Therefore,

$$\lambda_{min}(\widehat{\mathbf{A}}) \geq \min\{C_1, \lambda_{min}(\widehat{\mathbf{A}}_{11})\} = \lambda_{min}(\widehat{\mathbf{A}}_{11}) \min\{1, C_1/\lambda_{min}(\widehat{\mathbf{A}}_{11})\}.$$

Similarly from the upper bound of (4.9), we can show that

$$\lambda_{max}(\widehat{\mathbf{A}}) \leq \lambda_{max}(\widehat{\mathbf{A}}_{11}) \max\{1, C_2/\lambda_{max}(\widehat{\mathbf{A}}_{11})\},$$

where $C_2 = \frac{5}{4}$. Thus

$$\begin{aligned} \mathfrak{K}(\mathbf{A}) = \kappa_2(\widehat{\mathbf{A}}) = \frac{\lambda_{max}(\widehat{\mathbf{A}})}{\lambda_{min}(\widehat{\mathbf{A}})} &\leq \frac{\lambda_{max}(\widehat{\mathbf{A}}_{11}) \max\{1, C_2/\lambda_{max}(\widehat{\mathbf{A}}_{11})\}}{\lambda_{min}(\widehat{\mathbf{A}}_{11}) \min\{1, C_1/\lambda_{min}(\widehat{\mathbf{A}}_{11})\}} \\ &= \mathfrak{K}(\mathbf{A}_{11}) \frac{\max\{1, C_2/\lambda_{max}(\widehat{\mathbf{A}}_{11})\}}{\min\{1, C_1/\lambda_{min}(\widehat{\mathbf{A}}_{11})\}}, \end{aligned}$$

where $\mathfrak{K}(\mathbf{A}_{11}) = \kappa_2(\widehat{\mathbf{A}}_{11})$. Thus we have the required upper bound of $\mathfrak{K}(\mathbf{A})$.

Now let $\mathbf{z}_1$ be an eigenvector of $\widehat{\mathbf{A}}_{11}$ associated with $\lambda_{max}(\widehat{\mathbf{A}}_{11})$. Also let $\mathbf{z}_2 = \mathbf{0}$. Then from (4.8), we have

$$\mathbf{z}^T \widehat{\mathbf{A}} \mathbf{z} = \lambda_{max}(\widehat{\mathbf{A}}_{11}) \|\mathbf{z}_1\|^2 = \lambda_{max}(\widehat{\mathbf{A}}_{11}) \|\mathbf{z}\|^2,$$

and therefore,

$$\lambda_{max}(\widehat{\mathbf{A}}) \geq \lambda_{max}(\widehat{\mathbf{A}}_{11}).$$

Similarly, considering $\mathbf{z}_1$ to be an eigenvector of $\widehat{\mathbf{A}}_{11}$ associated with $\lambda_{min}(\widehat{\mathbf{A}}_{11})$ and $\mathbf{z}_2 = \mathbf{0}$, we have

$$\mathbf{z}^T \widehat{\mathbf{A}} \mathbf{z} = \lambda_{min}(\widehat{\mathbf{A}}_{11}) \|\mathbf{z}_1\|^2 = \lambda_{min}(\widehat{\mathbf{A}}_{11}) \|\mathbf{z}\|^2,$$

and therefore,

$$\lambda_{min}(\widehat{\mathbf{A}}) \leq \lambda_{min}(\widehat{\mathbf{A}}_{11}).$$

Now,

$$\mathfrak{K}(\mathbf{A}) = \frac{\lambda_{max}(\widehat{\mathbf{A}})}{\lambda_{min}(\widehat{\mathbf{A}})} \geq \frac{\lambda_{max}(\widehat{\mathbf{A}}_{11})}{\lambda_{min}(\widehat{\mathbf{A}}_{11})} = \mathfrak{K}(\mathbf{A}_{11}),$$

which is the required lower bound of $\mathfrak{K}(\mathbf{A})$. □

The Proposition 4.2 establishes that $\mathfrak{K}(\mathbf{A}) \approx \mathfrak{K}(\mathbf{A}_{11})$, i.e., the scaled condition numbers of the stiffness matrices for the GFEM and the basic part of the GFEM are of same order. Thus the GFEM with $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ is indeed an SGFEM.

Remark 4.3 We note that the orthogonality of the spaces $\mathcal{S}_1$ and $\overline{\mathcal{S}}_2$ was essential in proving Proposition 4.2. This property does not hold in general. Later we will define a notion of "almost orthogonality" of $\mathcal{S}_1$ and $\overline{\mathcal{S}}_2$, which will address this issue. ∎

Remark 4.4 The inequality (4.7) played an important role in obtaining Proposition 4.2. This property depends on the functions in $\overline{V}_i$. For general approximation spaces $\overline{V}_i$, we need an assumption that will be presented later. ∎

Remark 4.5 SGFEM uses $\overline{V}_i$ as the enrichment space, which is a modification of V_i. Other modifications of V_i have been reported in different contexts. For example, the *shifting* modification, namely, $\overline{\varphi}_j^{[i]}(x) = \varphi_j^{[i]}(x) - \varphi_j^{[i]}(x_i)$, $j > 0$, is used in XFEM in the context of approximation error as well as enforcement the *Kronecker delta property* (see [23]).∎

4.2 SGFEM and its analysis:

We now present the SGFEM for (2.1), with $a \in L^\infty(\Omega)$ and $0 < \alpha \leq a(x) \leq \beta$. Moreover, suppose it is a priori known that $u(0) = 0$ (we will further comment on a priori information later). We consider the uniform mesh $\{\tau_k\}_{k\in I\setminus\{0\}}$ with the set of vertices $\mathcal{T}$, as described in Section 3. Recall that the hat function N_i and the patch ω_i is associated with each $x_i \in \mathcal{T}$. We will refer to $\{x_{i-1}, x_i, x_{i+1}\}$ as the *vertices* of ω_i; the vertices of ω_0, ω_N are $\{x_0, x_1\}$, $\{x_{N-1}, x_N\}$, respectively. Let

$$\mathcal{T}_1,\, \mathcal{T}_2 \subset \mathcal{T}; \quad \zeta_1 := \mathrm{card}(\mathcal{T}_1),\ \zeta_2 := \mathrm{card}(\mathcal{T}_2); \quad \zeta_1,\, \zeta_2 \leq N+1.$$

We define $\mathcal{S}_1 = \sum_{x_i\in\mathcal{T}_1} a_i N_i$, $a_i \in \mathbb{R}$; $\mathcal{T}_1$ will be referred to as $\mathcal{S}_1$-*relevant set of vertices*. We consider $\mathcal{T}_1 = \{x_i \in \mathcal{T} : 1 \leq i \leq N\}$ as in the example in Section 4.1. For other choices of $\mathcal{T}_1$, we refer to Remark 4.6.

For $x_i \in \mathcal{T}$, let $V_i = \mathrm{span}\{\varphi_j^{[i]}\}_{j=0}^{n_i} \subset H^1(\omega_i)$ such that there exists $\xi^i \in V_i$ satisfying $\|u - \xi^i\|_{\mathcal{E}(\overset{\circ}{\tau}_k)} \leq \epsilon_i$ for all $\tau_k \subset \overline{\omega}_i$. Clearly, $\|u - \xi^i\|_{\mathcal{E}(\omega_i)} \leq 2\epsilon_i$. We consider the modified space $\overline{V}_i = \mathrm{span}\{\overline{\varphi}_j^{[i]}\}_{j=1}^{n_i}$, where

$$\overline{\varphi}_j^{[i]} = \varphi_j^{[i]} - \mathcal{I}_{\omega_i}\varphi_j^{[i]}; \quad \mathcal{I}_{\omega_i}\varphi_j^{[i]} := \sum_{i-1\leq k\leq i+1} \varphi_j^{[i]}(x_k) N_k\big|_{\omega_i}.$$

$\mathcal{I}_{\omega_i} v$ is the piecewise linear interpolant of $v \in H^1(\omega_i)$ on the patch ω_i based on the vertices of ω_i; we adjust $\mathcal{I}_{\omega_0}$ and $\mathcal{I}_{\omega_N}$ accordingly as before. It is important to note that if for some $x_i \in \mathcal{T}$, $V_i = \{\xi \in H^1(\omega_i) : \xi|_{\tau_k} \in \mathcal{P}^1(\tau_k)$ for all $\tau_k \subset \overline{\omega}_i\}$, then $\overline{V}_i = \{0\}$. Also $\overline{\xi}^i(x_k) = 0$ with $k = i-1, i, i+1$ for all $\overline{\xi}^i \in \overline{V}_i$. We refer to a patch ω_i as *enriched* if $\overline{V}_i \neq \{0\}$. Let $\mathcal{T}_2 := \{x_i \in \mathcal{T} : \omega_i$ is enriched$\}$ and define $\overline{\mathcal{S}}_2 = \sum_{x_i\in\mathcal{T}_2} N_i\overline{V}_i$; $\mathcal{T}_2$ will be referred to as the $\overline{\mathcal{S}}_2$-*relevant set of vertices*. In Section 4.1, we chose $\mathcal{T}_2 = \mathcal{T}$. We will present examples with $\zeta_2 << N+1$ (i.e., only few patches enriched) later in the paper.

Remark 4.6 The sets $\mathcal{T}_1$, $\mathcal{T}_2 \subset \mathcal{T}$ provide a framework to address numerical treatment of many applications. Selection of both sets depends on a priori information on the problem and its solution. Selection of $\mathcal{T}_2$ will be apparent from the examples in Section 5. Suppose $\mathcal{T}_0 \subset \mathcal{T}$ contains all the vertices $x_j \in \mathcal{T}$, where it is known a priori that $u(x_j) = 0$. We choose $\mathcal{T}_1 = \mathcal{T}\setminus\mathcal{T}_0$. Typically,

$\mathcal{T}_1$ will not contain any boundary vertex with homogeneous Dirichlet condition. However $\mathcal{T}_1$ may exclude other vertices in $\mathcal{T}$ based on a priori information. For example, let $f(x) = \sum_{k=0}^{\infty} c_k \cos[2\pi(2k+1)x]$, $a(x) = 1$, and suppose it it known that $u(0) = 0$. Then $u(1/4) = u(3/4) = 0$, and the vertices $x_j \notin \mathcal{T}_1$ if $x_j = 1/4$ or $x_j = 3/4$. Thus we can accommodate many a priori information in this framework. Only for simplicity, we have considered $\mathcal{T}_0 = \{x_0\}$ in this section. ∎

We now consider a GFEM with

$$\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2 = \sum_{x_i \in \mathcal{T}_1} a_i N_i + \sum_{x_i \in \mathcal{T}_2} N_i \overline{V}_i \,. \tag{4.10}$$

Note that $v(0) = 0$ for all $v \in \mathcal{S}$. We will show that this GFEM is an SGFEM, under certain assumptions on the space $\overline{\mathcal{S}}_2$, which we will present later. We mention that $\mathcal{T}_1$ and $\mathcal{T}_2$ are called $\mathcal{S}_1$ and $\overline{\mathcal{S}}_2$ relevant vertices, respectively, since the degrees of freedom associated only with these vertices appear in the GFEM.

We first present an approximation result for the GFEM with $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$.

Theorem 4.7 *Let $u \in H^1(\Omega)$ be the solution of (2.1). Suppose for each $x_i \in \mathcal{T}_2$, there exists $\bar{\xi}^i \in \overline{V}_i$ and $C_1 > 0$, independent of i, such that*

$$\|u - \mathcal{I}_{\omega_i} u - \bar{\xi}^i\|_{L^2(\omega_i)} \leq C_1 \, diam(\omega_i) \|u - \mathcal{I}_{\omega_i} u - \bar{\xi}^i\|_{\mathcal{E}(\omega_i)},$$

and $\|u - \mathcal{I}_{\omega_i} u - \bar{\xi}^i\|_{\mathcal{E}(\omega_i)} \leq \epsilon_i$. Then there exists $v \in \mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ such that

$$\|u - v\|_{\mathcal{E}(\Omega)} \leq C \big\{ \sum_{x_i \in \mathcal{T} \setminus \mathcal{T}_2} \|u - \mathcal{I}_{\omega_i} u\|^2_{\mathcal{E}(\omega_i)} + \sum_{x_i \in \mathcal{T}_2} \epsilon_i^2 \big\}^{1/2}. \tag{4.11}$$

Proof: Let $\mathcal{I}_h u = \sum_{x_i \in \mathcal{T}} u(x_i) N_i$ be the piecewise linear interpolant of u. We note that $\mathcal{I}_h u = \mathcal{I}_{\omega_i} u$ on ω_i. Define $w := u - \mathcal{I}_h u$ and let $\overline{v} := \sum_{x_i \in \mathcal{T}_2} N_i \bar{\xi}^i \in \overline{\mathcal{S}}_2$. Then recalling that $\{N_i\}_{x_i \in \mathcal{T}}$ is a PU, we have

$$w - \overline{v} = \sum_{x_i \in \mathcal{T}} N_i w - \sum_{x_i \in \mathcal{T}_2} N_i \bar{\xi}^i = \sum_{x_i \in \mathcal{T} \setminus \mathcal{T}_2} N_i w + \sum_{x_i \in \mathcal{T}_2} N_i (w - \bar{\xi}^i).$$

Therefore

$$\|w - \overline{v}\|^2_{\mathcal{E}(\Omega)} \leq C \big[\big\| \sum_{x_i \in \mathcal{T} \setminus \mathcal{T}_2} N_i w \big\|^2_{\mathcal{E}(\Omega)} + \big\| \sum_{x_i \in \mathcal{T}_2} N_i (w - \bar{\xi}^i) \big\|^2_{\mathcal{E}(\Omega)}. \tag{4.12}$$

We first address the last term of (4.12). Using the fact that $x \in \Omega$ is in at most two patches ω_i, ω_{i+1}, we see that the sum $\sum_{x_i \in \mathcal{T}_2} [N_i (w - \bar{\xi}^i)]'$ has at most two terms for any $x \in \Omega$. Using this observation, the assumption that $\|N_i'\|_{L^\infty(\Omega)} \leq C[\text{diam}\{\omega_i\}]^{-1}$, and the hypothesis of the Theorem, we can show

that

$$\begin{aligned} \| \sum_{x_i \in \mathcal{T}_2} N_i (w - \bar{\xi}^i) \|^2_{\mathcal{E}(\Omega)} &\leq C \Big[\sum_{x_i \in \mathcal{T}_2} \frac{\|w - \bar{\xi}^i\|^2_{L^2(\omega_i)}}{\mathrm{diam}\{\omega_i\}^2} \\ &\qquad + \sum_{x_i \in \mathcal{T}_2} \|w - \bar{\xi}^i\|^2_{\mathcal{E}(\omega_i)} \Big] \\ &\leq \sum_{x_i \in \mathcal{T}_2} \|w - \bar{\xi}^i\|^2_{\mathcal{E}(\omega_i)} \leq \sum_{x_i \in \mathcal{T}_2} \epsilon_i^2. \end{aligned} \tag{4.13}$$

(We refer to the proof of Theorem 3.2 in [4] for details of the argument leading to (4.13)). Using exactly same argument and the interpolation estimate $\|w\|_{L^2(\omega_i)} = \|u - \mathcal{I}_{\omega_i} u\|_{L^2(\omega_i)} \leq Ch \|u - \mathcal{I}_{\omega_i} u\|_{\mathcal{E}(\omega_i)}$, we get

$$\| \sum_{x_i \in \mathcal{T} \setminus \mathcal{T}_2} N_i w \|^2_{\mathcal{E}(\Omega)} \leq C \sum_{x_i \in \mathcal{T} \setminus \mathcal{T}_2} \|u - \mathcal{I}_{\omega_i} u\|^2_{\mathcal{E}(\omega_i)}.$$

Therefore, from (4.12) and (4.13), we have

$$\|w - \overline{v}\|^2_{\mathcal{E}(\Omega)} \leq C \big[\sum_{x_i \in \mathcal{T} \setminus \mathcal{T}_2} \|u - \mathcal{I}_{\omega_i} u\|^2_{\mathcal{E}(\omega_i)} + \sum_{x_i \in \mathcal{T}_2} \epsilon_i^2 \big].$$

Finally, writing $w = u - \mathcal{I}_h u$ and setting $v = \mathcal{I}_h u + \overline{v} \in \mathcal{S}_1 + \overline{\mathcal{S}}_2$, we get the desired result. □

We mention that unlike in Theorem 4.1, we did not assume $\mathcal{T}_2 = \mathcal{T}$ in Theorem 4.7. We further note that $\mathcal{I}_h u$ for $u \in H^1(\Omega)$ is not defined in higher dimensions, since the point values of u, in general, do not exist in higher dimensions (in contrast to 1-d). However, using a generalized interpolant based on the average of u in a ball around the vertices x_i, the proof of the above result can be easily generalized to higher dimensions.

Remark 4.8 From the proof of Proposition 4.1, it is clear that accurate local approximation of $u - \mathcal{I}_h u$ by functions in $\overline{V}_i$ is crucial to obtain the desired result. This is the main idea of SGFEM – the spaces $\overline{V}_i$ are constructed such that the functions in $\overline{V}_i$ accurately approximate $u - \mathcal{I}_h u$ in ω_i. This is in contrast to the standard GFEM, where the functions in local approximating spaces V_i accurately approximate u in ω_i. ∎

Remark 4.9 We note that $\overline{V}_i = \{0\}$ for $x_i \in \mathcal{T} \setminus \mathcal{T}_2$. If $u \in H^1(\Omega)$ is locally smooth, namely, $u \in H^2(\omega_i)$ for $x_i \in \mathcal{T} \setminus \mathcal{T}_2$, then $\|u - \mathcal{I}_{\omega_i} u\|_{\mathcal{E}(\omega_i)} \leq Ch |u|_{H^2(\omega_i)}$ for $x_i \in \mathcal{T} \setminus \mathcal{T}_2$ and (4.11) could be written as

$$\|u - v\|_{\mathcal{E}(\Omega)} \leq C \big\{ h^2 \sum_{x_i \in \mathcal{T} \setminus \mathcal{T}_2} |u|^2_{H^2(\omega_i)} + \sum_{x_i \in \mathcal{T}_2} \epsilon_i^2 \big\}^{1/2}. \tag{4.14}$$

By incorporating the available information on the solution u in V_i, for $x_i \in \mathcal{T}_2$, we can have $\epsilon_i = O(h)$, and consequently, $\|u - v\|_{\mathcal{E}(\Omega)} = O(h)$. The set $\mathcal{T}_2$ can

be chosen adaptively with respect to a prescribed tolerance, which we do not elaborate in this paper. ∎

Remark 4.10 A rate of convergence of $O(h)$ for various problems have been reported for the Corrected XFEM (which is also a GFEM); see e.g., [22]. However, for the crack propagation problems, the enrichment spaces V_i in XFEM requires the use of a *ramp-function* to obtain the $O(h)$ rate of convergence. In contrast, the GFEM based on $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ does not require the use of a ramp-function to obtain the rate of convergence of $O(h)$.

We now address the scaled condition number of the stiffness matrix of the GFEM with $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$. For clarity of the exposition, we will present the analysis for the case when $n_i = 1$ i.e., $\overline{V}_i = \text{span}\{\overline{\varphi}_1^{[i]}\}$. The analysis for general n_i is similar.

As in the example presented in Section 4.1, the stiffness matrix $\mathbf{A}$ is of the form $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$, where $\mathbf{A}_{11} = \{B(N_i, N_j)\}_{x_i, x_j \in \mathcal{T}_1}$ is the $\zeta_1 \times \zeta_1$ stiffness matrix of the basic part of the GFEM. Let $\mathbf{D}_1$ be a diagonal matrix with $(\mathbf{D}_1)_{ii} = (\mathbf{A}_{11})_{ii}^{-1/2}$. Clearly, the diagonal elements of

$$\widehat{\mathbf{A}}_{11} := \mathbf{D}_1 \mathbf{A}_{11} \mathbf{D}_1 \tag{4.15}$$

are equal to 1.

The matrix $\mathbf{A}_{22}$ plays a central role in our analysis and depends on elements that have been enriched. We will refer to an element $\tau_k = [x_{k-1}, x_k]$ as *enriched* if (a) $x_{k-1} \in \mathcal{T}_2$ and $\overline{\varphi}_1^{[k-1]}|_{\tau_k} \not\equiv 0$, or (b) $x_k \in \mathcal{T}_2$ and $\overline{\varphi}_1^{[k]}|_{\tau_k} \not\equiv 0$. Let

$$\mathcal{K}_{enr} := \{\tau_k : \tau_k \text{ is enriched}\}.$$

The matrix $\mathbf{A}_{22}$ is constructed by the assembly process using the element stiffness matrices $A_{22}^{(k)}$ defined only on $\tau_k \in \mathcal{K}_{enr}$.

We now address the structure of the element matrices $A_{22}^{(k)}$ in detail and set up some notions and notations that will be used in the analysis. We denote the vertices of the element τ_k as $x_1^{(k)} := x_{k-1}$ and $x_2^{(k)} := x_k$; we consider only $\tau_k \in \mathcal{K}_{enr}$. The element stiffness matrix $A_{22}^{(k)}$ is of the form

$$A_{22}^{(k)} = \begin{bmatrix} b_{11}^{(k)} & b_{12}^{(k)} \\ b_{12}^{(k)} & b_{22}^{(k)} \end{bmatrix}, \tag{4.16}$$

where $b_{ij}^{(k)} = B_{\tau_k}\big(N_{k-2+i}\overline{\varphi}_1^{[k-2+i]}, N_{k-2+j}\overline{\varphi}_1^{[k-2+j]}\big)$, $1 \le i, j \le 2$.

If $b_{11}^{(k)}, b_{22}^{(k)} > 0$, then $A_{22}^{(k)}$ is 2×2 and we say that the local stiffness matrix $A_{22}^{(k)}$ is *associated with the vertices* $x_1^{(k)} = x_{k-1}$ and $x_2^{(k)} = x_k$. We define a diagonal matrix $D^{(k)} = \text{diag}\{\delta_1^{(k)}, \delta_2^{(k)}\}$ with $\delta_1^{(k)}, \delta_2^{(k)} > 0$, such that the diagonal elements of

$$\hat{A}_{22}^{(k)} := D^{(k)} A_{22}^{(k)} D^{(k)}$$

are of equal to 1 or $O(1)$, independent of h. Clearly, $\delta_1^{(k)}, \delta_2^{(k)}$ are associated with vertices x_{k-1}, x_k respectively.

On the other hand, if $b_{22}^{(k)} = 0$ (i.e., $\overline{\varphi}_1^{[k]}|_{\tau_k} \equiv 0$ and consequently $b_{12}^{(k)} = b_{21}^{(k)} = 0$) in (4.16), then the local stiffness matrix $A_{22}^{(k)} = [b_{11}^{(k)}]$ is of size 1×1 and is associated only with the vertex $x_1^{(k)} = x_{k-1}$. We define $D^{(k)} = [\delta_1^{(k)}]$, where $\delta_1^{(k)} = \{b_{11}^{(k)}\}^{-1/2}$; $\delta_1^{(k)}$ is associated with the vertex $x_1^{(k)} = x_{k-1}$. Similarly, if $b_{11}^{(k)} = 0$ in (4.16), then the local stiffness matrix $A_{22}^{(k)} = [b_{22}^{(k)}]$ is associated only with the vertex $x_2^{(k)} = x_k$. Also $D^{(k)} = [\delta_2^{(k)}]$ with $\delta_1^{(k)} = \{b_{22}^{(k)}\}^{-1/2}$ associated with the vertex $x_2^{(k)} = x_k$. Let $\varsigma^{(k)}$ be the number of vertices associated with the local stiffness matrix $A_{22}^{(k)}$. Thus the size of $A_{22}^{(k)}$ is $\varsigma^{(k)} \times \varsigma^{(k)}$; note that $\varsigma^{(k)}$ is either 1 or 2 with our assumption $n_i = 1$.

Recall that $\mathbf{A}_{22}$ is obtained by the assembly process using the element stiffness matrices $A_{22}^{(k)}$; the size of $\mathbf{A}_{22}$ is $\zeta_2 \times \zeta_2$. Let $c = (c_1, c_2, \cdots, c_{\zeta_2})$, then

$$c^T \mathbf{A}_{22} c = \sum_{\tau_k \in \mathcal{K}_{enr}} [c^{(k)}]^T A_{22}^{(k)} c^{(k)}, \tag{4.17}$$

where $c^{(k)} \in \mathbb{R}^{\varsigma^{(k)}}$. Moreover, the components of $c^{(k)}$ are also the components of c that correspond to those vertices of τ_k that are associated with $A_{22}^{(k)}$. For example, if $b_{11}^{(k)}, b_{22}^{(k)} > 0$ in $A_{22}^{(k)}$, then as mentioned before, the vertices $x_1^{(k)} = x_{k-1}$, $x_2^{(k)} = x_k$ are associated with $A_{22}^{(k)}$. Suppose the components $c_{j(k)-1}, c_{j(k)}$ of c are associated with the vertices x_{k-1}, x_k, respectively, of τ_k. Then $c^{(k)} = [c_{j(k)-1}, c_{j(k)}]^T$. Similarly, if $A_{22}^{(k)} = [b_{11}^{(k)}]$, then $A_{22}^{(k)}$ is associated with $x_1^{(k)}$ and $c^{(k)} = [c_{j(k)-1}]$ – a vector with one component. Later in our analysis, we will use (4.17) with a particular vector c and $c^{(k)}$ as defined above.

Next we note that each vertex x_i of the FE mesh is associated with a FE star – union of all elements $\tau_k \subset \overline{\omega}_i$ (equivalently, union of all elements τ_k with common vertex x_i). For $x_i \in \mathcal{T}_2$, we define $\mathcal{K}_i := \{\tau_k \in \mathcal{K}_{enr} : \tau_k \subset \overline{\omega}_i\}$. For $x_i \in \mathcal{T}_2$ and $\tau_k \in \mathcal{K}_i$, we set the index $1 \le l(i,k) \le 2$ as follows. We first note that $k \in \{i, i+1\}$. For $k = i$, we set $l(i,k) = l(i,i) = 2$ and for $k = i+1$, we set $l(i,k) = l(i,i+1) = 1$. Thus $l(i,k)$ is the index such that $x_{l(i,k)}^{(k)} = x_i$; note $x_{l(i,k)}^{(k)}$ may not be associated with $A_{22}^{(k)}$. We define

$$\mathcal{K}_i^* := \{\tau_k \in \mathcal{K}_i : x_{l(i,k)}^{(k)} \text{ is associated with } A_{22}^{(k)}\}.$$

Thus $\mathcal{K}_i^*$ is the set of $\tau_k \in \mathcal{K}_i$ such that $\overline{\varphi}_1^{[i]}|_{\tau_k} \not\equiv 0$. For $x_i \in \mathcal{T}_2$, we define

$$\Delta_i := \sum_{\tau_k \in \mathcal{K}_i^*} [\delta_{l(i,k)}^{(k)}]^{-2}, \tag{4.18}$$

which will be used later in our analysis.

Each diagonal element of $\mathbf{A}_{22}$ is associated with a vertex in $\mathcal{T}_2$. Let $(\mathbf{A}_{22})_{j_i j_i}$ be associated with $x_i \in \mathcal{T}_2$. Moreover, we note that $(\mathbf{A}_{22})_{j_i j_i} = \sum_{\tau_k \in \mathcal{K}_i^*} b_{l(i,k),l(i,k)}^{(k)}$,

where $b_{pq}^{(k)}$ was defined in (4.16). Thus $(\mathbf{A}_{22})_{j_i j_i} > 0$ for all $x_i \in \mathcal{T}_2$ (i.e., all the diagonal elements of $\mathbf{A}_{22}$ are positive). We now define the diagonal matrix $\mathbf{D}_2 = \text{diag}\{d_1, d_2, \cdots, d_{\zeta_2}\}$ with $d_j = (\mathbf{A}_{22})_{jj}^{-1/2}$, $1 \le j \le \zeta_2$. Note that $d_{j_i} = (\mathbf{A}_{22})_{j_i j_i}^{-1/2}$ is associated with $x_i \in \mathcal{T}_2$. Clearly, the diagonal elements of

$$\widehat{\mathbf{A}}_{22} := \mathbf{D}_2 \mathbf{A}_{22} \mathbf{D}_2 \tag{4.19}$$

are equal to 1. Define the diagonal matrix $\mathbf{D} := \text{diag}\{\mathbf{D}_1, \mathbf{D}_2\}$. Since the diagonal elements of $\widehat{\mathbf{A}}_{11}$, $\widehat{\mathbf{A}}_{22}$ (see (4.15), (4.19)) are equal to 1, the diagonal elements of

$$\widehat{\mathbf{A}} := \mathbf{DAD} = \begin{bmatrix} \widehat{\mathbf{A}}_{11} & \widehat{\mathbf{A}}_{12} \\ \widehat{\mathbf{A}}_{21} & \widehat{\mathbf{A}}_{22} \end{bmatrix} \tag{4.20}$$

are also equal to 1. Also $\widehat{\mathbf{A}}_{12} = \mathbf{D}_1 \mathbf{A}_{12} \mathbf{D}_2$ and $\widehat{\mathbf{A}}_{21} = \widehat{\mathbf{A}}_{12}^T$.

We will show that the GFEM with $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ is an SGFEM, under the following assumptions on the local approximation spaces $\overline{V}_i$ and the enrichment part of $\mathcal{S}$, namely $\overline{S}_2$.

Assumption 1 *The spaces $\mathcal{S}_1$ and $\overline{\mathcal{S}}_2$ are* almost orthogonal *with respect to the inner product $B(\cdot,\cdot)$, i.e., there exist constants $0 < L_1, U_1 < \infty$, independent of h, such that*

$$L_1\big\{\|v_1\|_{\mathcal{E}(\Omega)}^2 + \|v_2\|_{\mathcal{E}(\Omega)}^2\big\} \le |B(v_1+v_2, v_1+v_2)| \le U_1\big\{\|v_1\|_{\mathcal{E}(\Omega)}^2 + \|v_2\|_{\mathcal{E}(\Omega)}^2\big\},$$

for all $v_1 \in \mathcal{S}_1$ and $v_2 \in \overline{\mathcal{S}}_2$.

Assumption 2 *For $\tau_k \in \mathcal{K}_{enr}$, there exist constants $0 < L_2, U_2 < \infty$, independent of k and h such that*

$$L_2 \|[D^{(k)}]^{-1}\mathbf{x}\|^2 \le \mathbf{x}^T A_{22}^{(k)} \mathbf{x} \le U_2 \|[D^{(k)}]^{-1}\mathbf{x}\|^2, \quad \forall\ \mathbf{x} \in \mathbb{R}^{\varsigma^{(k)}},$$

where the diagonal matrices $D^{(k)}$ have been defined before.

Assumption 3 *For $x_i \in \mathcal{T}_2$, there exist constants $0 < L_3, U_3 < \infty$, independent of i and h such that*

$$L_3 \le (\mathbf{A}_{22})_{j_i j_i}^{-1} \Delta_i \le U_3,$$

where $(\mathbf{A}_{22})_{j_i j_i}$ is the diagonal element of $\mathbf{A}_{22}$ associated with x_i, and Δ_i is as defined in (4.18).

The following result is an easy consequence of Assumption 1.

Lemma 4.11 *Let $x = (\xi^T, \eta^T)^T \in \mathbb{R}^{\zeta_1+\zeta_2}$ where $\xi \in \mathbb{R}^{\zeta_1}$ and $\eta \in \mathbb{R}^{\zeta_2}$. Then there exist positive constants L_1 and U_1, independent of h, such that*

$$L_1 \left[\xi^T \mathbf{A}_{11} \xi + \eta^T \mathbf{A}_{22} \eta\right] \le x^T \mathbf{A} x \le U_1 \left[\xi^T \mathbf{A}_{11} \xi + \eta^T \mathbf{A}_{22} \eta\right],$$

where $\mathbf{A}$, $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are matrices defined before.

Proof: Let $\xi = (\xi_i)_{x_i \in \mathcal{T}_1}$ and $\eta = (\eta_i)_{x_i \in \mathcal{T}_2}$. Consider $v_1 = \sum_{x_i \in \mathcal{T}_1} \xi_i N_i \in \mathcal{S}_1$ and $v_2 = \sum_{x_i \in \mathcal{T}_2} \eta_i N_i \overline{\varphi}_1^{[i]} \in \overline{\mathcal{S}}_2$. Then $B(v_1 + v_2, v_1 + v_2) = x^T \mathbf{A} x$, $B(v_1, v_1) = \xi^T \mathbf{A}_{11} \xi$, and $B(v_2, v_2) = \eta^T \mathbf{A}_{22} \eta$. The desired result is now immediate from Assumption 1. □

Theorem 4.12 *Suppose the Assumptions 1, 2, and 3 are satisfied. Let* $\mathbf{A}$ *be the stiffness matrix of the GFEM with* $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$. *Then*

$$\frac{L_1}{U_1} \mathfrak{K}(\mathbf{A}_{11}) \leq \mathfrak{K}(\mathbf{A}) \leq \mathfrak{K}(\mathbf{A}_{11}) \frac{U_1}{L_1} \frac{\max\left\{1, U_2 U_3 / \lambda_{max}(\widehat{\mathbf{A}}_{11})\right\}}{\min\left\{1, L_2 L_3 / \lambda_{min}(\widehat{\mathbf{A}}_{11})\right\}},$$

where $\lambda_{min}(\widehat{\mathbf{A}}_{11})$, $\lambda_{max}(\widehat{\mathbf{A}})$ *are the smallest and largest eigenvalues, respectively, of the matrix* $\widehat{\mathbf{A}}_{11}$ *defined before.*

Remark 4.13 This result shows that under the Assumptions 1, 2, and 3, the scaled condition numbers of the stiffness matrices of the GFEM with $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ and the basic part of the GFEM are of the same order. Thus the GFEM with $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ is indeed an SGFEM.

Proof: Let $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)^T \in \mathbb{R}^{\zeta_1 + \zeta_2}$, where $\mathbf{z}_1 \in \mathbb{R}^{\zeta_1}$ and $\mathbf{z}_2 \in \mathbb{R}^{\zeta_2}$. Then from the definition of $\widehat{\mathbf{A}}$ (see (4.20)), we have $\mathbf{z}^T \widehat{\mathbf{A}} \mathbf{z} = \mathbf{z}^T \mathbf{D} \mathbf{A} \mathbf{D} \mathbf{z} = (\mathbf{D}\mathbf{z})^T \mathbf{A} (\mathbf{D}\mathbf{z})$, and since $\mathbf{D}\mathbf{z} = \left[(\mathbf{D}_1 \mathbf{z}_1)^T, (\mathbf{D}_2 \mathbf{z}_2)^T\right]^T$, from Lemma 4.11 we get

$$\begin{aligned} L_1 \left[(\mathbf{D}_1 \mathbf{z}_1)^T \mathbf{A}_{11} (\mathbf{D}_1 \mathbf{z}_1) + (\mathbf{D}_2 \mathbf{z}_2)^T \mathbf{A}_{22} (\mathbf{D}_2 \mathbf{z}_2)\right] &\leq \mathbf{z}^T \widehat{\mathbf{A}} \mathbf{z} \\ &\leq U_1 \left[(\mathbf{D}_1 \mathbf{z}_1)^T \mathbf{A}_{11} (\mathbf{D}_1 \mathbf{z}_1) + (\mathbf{D}_2 \mathbf{z}_2)^T \mathbf{A}_{22} (\mathbf{D}_2 \mathbf{z}_2)\right]. \end{aligned} \tag{4.21}$$

Let $\mathbf{z}_2 = (f_1, f_2, \cdots, f_{\zeta_2})^T$ and consider $\mathbf{D}_2 = \mathrm{diag}(d_1, d_2, \cdots, d_{\zeta_2})$ with $d_i = (\mathbf{A}_{22})_{ii}^{-1/2}$ as defined before. Then $\mathbf{D}_2 \mathbf{z}_2 = (d_1 f_1, d_2 f_2, \cdots, d_{\zeta_2} f_{\zeta_2})^T$. Recall that d_{j_i} is associated with $x_i \in \mathcal{T}_2$. Consequently, $d_{j_i} f_{j_i}$ is associated with $x_i \in \mathcal{T}_2$.

Consider an element $\tau_k \in \mathcal{K}_{enr}$. Following the notation given after (4.17), let $\overline{\mathbf{z}}_2^{(k)} := (\mathbf{D}_2 \mathbf{z}_2)^{(k)} \in \mathbb{R}^{\varsigma^{(k)}}$ such that the components of $\overline{\mathbf{z}}_2^{(k)}$ are the components of $\mathbf{D}_2 \mathbf{z}_2$ corresponding to the vertices of τ_k associated with $A_{22}^{(k)}$. Now from (4.17) and using Assumption 2, we have

$$(\mathbf{D}_2 \mathbf{z}_2)^T \mathbf{A}_{22} (\mathbf{D}_2 \mathbf{z}_2) = \sum_{\tau_k \in \mathcal{K}_{enr}} {\overline{\mathbf{z}}_2^k}^T A_{22}^{(k)} \overline{\mathbf{z}}_2^k \geq L_2 \sum_{\tau_k \in \mathcal{K}_{enr}} \|[D^{(k)}]^{-1} \overline{\mathbf{z}}_2^k\|^2. \tag{4.22}$$

We note that if $D^{(k)} = \mathrm{diag}\{\delta_1^{(k)}, \delta_2^{(k)}\}$, then

$$\|[D^{(k)}]^{-1} \overline{\mathbf{z}}_2^k\|^2 = [\delta_1^{(k)}]^{-2} [d_{j(k)-1}]^2 [f_{j(k)-1}]^2 + [\delta_2^{(k)}]^{-2} [d_{j(k)}]^2 [f_{j(k)}]^2,$$

where $\overline{\mathbf{z}}_2^k = [d_{j(k)-1}, f_{j(k)-1}]^T$ following the notation given after (4.17). Similarly, if $D^{(k)} = [\delta_1^{(k)}]$, then $\|[D^{(k)}]^{-1} \overline{\mathbf{z}}_2^k\|^2 = [\delta_1^{(k)}]^{-2} [d_{j(k)-1}]^2 [f_{j(k)-1}]^2$, and if $D^{(k)} = [\delta_2^{(k)}]$, then $\|[D^{(k)}]^{-1} \overline{\mathbf{z}}_2^k\|^2 = [\delta_2^{(k)}]^{-2} [d_{j(k)}]^2 [f_{j(k)}]^2$.

Now, it is important to note that if $\mathcal{J}_1 := \{d_{j(k)-1}f_{j(k)-1}, d_{j(k)}f_{j(k)}\}_{\tau_k \in \mathcal{K}_{enr}}$ (where the repeated elements appear only once) and $\mathcal{J}_2 := \{d_{j_i}f_{j_i}\}_{x_i \in \mathcal{T}_2}$, then $\mathcal{J}_1 = \mathcal{J}_2$. Thus from (4.22), we have

$$(\mathbf{D}_2\mathbf{z}_2)^T\mathbf{A}_{22}(\mathbf{D}_2\mathbf{z}_2) \geq L_2 \sum_{x_i \in \mathcal{T}_2} \Delta_i d_{j_i}^2 f_{j_i}^2 \geq L_2 L_3 \|\mathbf{z}_2\|^2, \tag{4.23}$$

where we used Assumption 3 to get the last inequality. Similarly, we can show that

$$(\mathbf{D}_2\mathbf{z}_2)^T\mathbf{A}_{22}(\mathbf{D}_2\mathbf{z}_2) \leq U_2 U_3 \|\mathbf{z}_2\|^2.$$

Therefore from (4.21) and using the definition of $\widehat{\mathbf{A}}_{11}$, we get

$$L_1 \left[\mathbf{z}_1^T \widehat{\mathbf{A}}_{11}\mathbf{z}_1 + L_2 L_3 \|\mathbf{z}_2\|^2\right] \;\leq\; \mathbf{z}^T\widehat{\mathbf{A}}\mathbf{z} \leq U_1 \left[\mathbf{z}_1^T \widehat{\mathbf{A}}_{11}\mathbf{z}_1 + U_2 U_3 \|\mathbf{z}_2\|^2\right]. \tag{4.24}$$

Now from the lower bound of $\mathbf{z}^T\widehat{\mathbf{A}}\mathbf{z}$ in (4.24), we have

$$\mathbf{z}^T\widehat{\mathbf{A}}\mathbf{z} \quad \geq \quad L_1 \left[\lambda_{min}(\widehat{\mathbf{A}}_{11})\|\mathbf{z}_1\|^2 + L_2 L_3 \|\mathbf{z}_2\|^2\right],$$

and therefore,

$$\lambda_{min}(\widehat{\mathbf{A}}) \geq L_1\, \lambda_{min}(\widehat{\mathbf{A}}_{11})\, \min\left\{1, L_2 L_3/\lambda_{min}(\widehat{\mathbf{A}}_{11})\right\}. \tag{4.25}$$

Similarly, using the upper bound of $\mathbf{z}^T\widehat{\mathbf{A}}\mathbf{z}$ in (4.24), we can show

$$\lambda_{max}(\widehat{\mathbf{A}}) \leq U_1\, \lambda_{max}(\widehat{\mathbf{A}}_{11})\, \max\left\{1, U_2 U_3/\lambda_{max}(\widehat{\mathbf{A}}_{11})\right\}. \tag{4.26}$$

Thus from (4.25) and (4.26), we have

$$\mathfrak{K}(\mathbf{A}) = \frac{\lambda_{max}(\widehat{\mathbf{A}})}{\lambda_{min}(\widehat{\mathbf{A}})} \leq \mathfrak{K}(\mathbf{A}_{11})\, \frac{U_1}{L_1}\, \frac{\max\left\{1, U_2 U_3/\lambda_{max}(\widehat{\mathbf{A}}_{11})\right\}}{\min\left\{1, L_2 L_3/\lambda_{min}(\widehat{\mathbf{A}}_{11})\right\}}, \tag{4.27}$$

which the required upper bound. The required lower bound could be obtained by following the exact arguments in Proposition 4.2 and (4.24). Thus we get the desired result. □

We mention that the notions and notations developed leading to Theorem 4.12 can also be extended to higher dimensions. An element will have n_e vertices, e.g., n_e could be 3 or 4 in 2-d. And the element stiffness matrices $A_{22}^{(k)}$ could be at most $n_e \times n_e$. The assembly argument (4.17) could be easily generalized to higher dimensions. For a given vertex x_i and an enriched element τ_k in the FE star associated with x_i, the index $l(i,k)$ will again represent the local index of the vertex $x_{l(i,k)}^{(k)}$ of τ_k that coincides with x_i, i.e., $x_{l(i,k)}^{(k)} = x_i$. The expressions for Δ_i, $(\mathbf{A}_{22})_{ii}$ and the Assumpsions 1, 2, 3, are exactly same in higher dimensions. Using these notions, the proof of Theorem 4.12 can be easily extended to higher dimensions. The approach presented here can also be extended for elasticity equations etc. We note however, the notations become a little more involved if $n_i > 1$.

Remark 4.14 We now make comments on the assumptions. The Assumption 1 is always satisfied in 1-d. Let $B_0(u,v) := \int_\Omega u'\,v'\,dx$. Since $\overline{\varphi}_j^{[i]}(x_k) = 0$ for $k = i-1, i, i+1$, it can be easily shown that $B_0(v_1, v_2) = 0$ for all $v_1 \in \mathcal{S}_1$ and $v_2 \in \overline{\mathcal{S}}_2$. Therefore,

$$\begin{aligned} &B(v_1+v_2, v_1+v_2) \geq \alpha B_0(v_1+v_2, v_1+v_2) \\ &= \alpha[B_0(v_1,v_1) + B_0(v_2,v_2)] \geq \frac{\alpha}{\beta}[\|v_1\|^2_{\mathcal{E}(\Omega)} + \|v_2\|^2_{\mathcal{E}(\Omega)}]. \end{aligned}$$

Similarly, we can show that

$$B(v_1+v_2, v_1+v_2) \leq \frac{\beta}{\alpha}[\|v_1\|^2_{\mathcal{E}(\Omega)} + \|v_2\|^2_{\mathcal{E}(\Omega)}],$$

and thus Assumption 1 is satisfied with $L_1 = \frac{\alpha}{\beta}$ and $L_2 = \frac{\beta}{\alpha}$. In higher dimensions, this assumption has to be checked.

Assumption 2 is equivalent to $L_2\|\mathbf{y}\|^2 \leq \mathbf{y}^T \hat{A}_{22}^{(k)}\mathbf{y} \leq U_2\|\mathbf{y}\|^2$ for all $\mathbf{y} \in \mathbb{R}^{\zeta^{(k)}}$. Thus $\hat{A}_{22}^{(k)}$ is uniformly positive definite in k and its eigenvalues are uniformly bounded.

It is always possible to choose the diagonal matrix $D^{(k)}$ such that Assumption 3 is satisfied. For example, it is easy to check that Assumption 3 is satisfied with $L_3 = U_3 = 1$ by choosing $D^{(k)} = \text{diag}\{\delta_1^k, \delta_2^{(k)}\}$ with $\delta_2^{(k)} = (b_{jj}^{(k)})^{-1/2}$. The Assumption 3 is trivially satisfied with $L_3 = U_3 = 1$ when $D^{(k)}$ is a 1×1 matrix. ∎

Remark 4.15 As shown in the Appendix, the implementation of the SGFEM does not require scaling the stiffness matrix, i.e., the linear system involving the stiffness matrix $\mathbf{A}$, and not scaled version $\widehat{\mathbf{A}}$, is solved. The scaling was used only to define $\mathfrak{K}(\mathbf{A})$ and to study its order through Theorem 4.12. We will show in the Appendix that $\mathfrak{K}(\mathbf{A})$ is an indicator of the loss of accuracy in the computed solution of the linear system associated with FEM, GFEM, and SGFEM. ∎

5 Applications:

In this section we will present the SGFEM, when applied to three specific applications. We will primarily address in detail the scaled condition number of the stiffness matrix of the method and show that the assumptions presented in the last section hold. The SGFEM, applied to each of these applications, will based on the uniform mesh $\{\tau_k\}_{k\in I\setminus\{0\}}$ with the set of vertices $\mathcal{T}$, defined before.

5.1 Interface Problems

Let $a(x)$ in (2.1) be a piecewise constant function and let f be smooth. We will consider two situations, namely, $a(x) = a_1(x)$ and $a(x) = a_2(x)$, where

$$a_1(x) = \begin{cases} \frac{1}{2}, & 0 \le x < b^* \\ 1, & b^* \le x \le 1 \end{cases} \quad \text{and} \quad a_2(x) = \begin{cases} 1, & 0 \le x < b_1^* \\ \frac{1}{2}, & b_1^* \le x < b_2^* \\ 1, & b_2^* \le x \le 1 \end{cases}$$

We note that the solution u of (2.1) does not belong to $H^2(\Omega)$.

We first consider $a(x) = a_1(x)$. We consider the set $\mathcal{T}_2 \subset \mathcal{T}$ as before. There exists an m such that $b^* \in \overset{o}{\tau}_{m+1} = (x_m, x_{m+1})$ and therefore, $b^* \in \omega_m \cap \omega_{m+1}$. For $x_i \in \mathcal{T}$, we consider $V_i = \text{span}\{1, \varphi_1^{[i]} = \int_{x_{i-1}}^{x} (1/a_1(t))dt\}$. Clearly, for $i \neq m, m+1$, we have $V_i = \text{span}\{1, (x - x_{i-1})\}$. Therefore recalling that $\overline{V}_i = \text{span}\{\overline{\varphi}_1^{[i]}\}$, where $\overline{\varphi}_1^{[i]} = \varphi_1^{[i]} - \mathcal{I}_{\omega_i}\varphi_1^{[i]}$, we get $\overline{V}_i = \{0\}$ for $i \neq m, m+1$. We set $\mathcal{T}_2 = \{x_m, x_{m+1}\} \subset \mathcal{T}$ and from the definition of $\overline{S}_2$, we have

$$\overline{S}_2 = \sum_{x_i \in \mathcal{T}_2} N_i \overline{V}_i = N_m \overline{V}_m + N_{m+1} \overline{V}_{m+1}\,.$$

We further note that $\varphi_1^{[m]}$ is linear on τ_m and therefore, $\overline{\varphi}_1^{[m]}|_{\tau_m} = 0$. Similarly, $\overline{\varphi}_1^{[m+1]}|_{\tau_{m+2}} = 0$. Therefore τ_{m+1} is the only enriched element, i.e., $\mathcal{K}_{enr} = \{\tau_{m+1}\}$, and $\mathbf{A}_{22} = A_{22}^{(m+1)}$. Also, we can easily show that $\overline{\varphi}_1^{[m]}|_{\tau_{m+1}} = \overline{\varphi}_1^{[m+1]}|_{\tau_{m+1}}$. Let $b^* = x_m + \beta h$ with $0 < \beta < 1$. Then from a direct computation, we have

$$A_{22}^{(m+1)} = \begin{bmatrix} h\beta(1-\beta)^2(\frac{3}{2} + \beta - 2\beta^2)/3 & h\beta^2(1-\beta)^2(1+4\beta)/6 \\ h\beta^2(1-\beta)^2(1+4\beta)/6 & h\beta^2(1-\beta)(1+2\beta^2)/3 \end{bmatrix}. \tag{5.1}$$

Clearly, $A_{22}^{(m+1)}$ is associated with the vertices x_m, x_{m+1}. We choose the diagonal matrix $D^{(m+1)} = \text{diag}\{\delta_1^{(m+1)}, \delta_2^{(m+1)}\}$, where

$$\delta_1^{(m+1)} = h^{-1/2}\beta^{-1/2}(1-\beta)^{-1}, \quad \delta_2^{(m+1)} = h^{-1/2}\beta^{-1}(1-\beta)^{-1/2}. \tag{5.2}$$

Then

$$\begin{aligned} \hat{A}_{22}^{(m+1)} &= D^{(m+1)} A_{22}^{(m+1)} D^{(m+1)} \\ &= \begin{bmatrix} (\frac{3}{2} + \beta - 2\beta^2)/3 & \beta^{1/2}(1-\beta)^{1/2}(1+4\beta)/6 \\ \beta^{1/2}(1-\beta)^{1/2}(1+4\beta)/6 & (1+2\beta^2)/3 \end{bmatrix}. \end{aligned} \tag{5.3}$$

The diagonal elements of $\hat{A}_{22}^{(m+1)}$ are $O(1)$ for all $0 < \beta < 1$. Also the eigenvalues of $\hat{A}_{22}^{(m+1)}$ are $\lambda_1 = (2-\beta)/6$ and $\lambda_2 = (1+\beta)/2$. Therefore, recalling Remark 4.14, we have

$$\frac{1}{6}\|[D^{(m+1)}]^{-1}\mathbf{x}\|^2 \le \mathbf{x}^T A_{22}^{(m+1)} \mathbf{x} \le \|[D^{(m+1)}]^{-1}\mathbf{x}\|^2, \quad \forall\, \mathbf{x} \in \mathbf{R}^2, \tag{5.4}$$

and hence, Assumption 2 is satisfied with $L_2 = \frac{1}{6}$ and $U_2 = 1$.

We set $\mathbf{D}_2 = \text{diag}\{d_1, d_2\}$ with $d_i = (\mathbf{A}_{22})_{ii}^{-1/2}$. Clearly, the diagonal elements of $\widehat{\mathbf{A}}_{22} = \mathbf{D}_2\mathbf{A}_{22}\mathbf{D}_2$ are equal to 1. Recall that $\mathcal{T}_2 = \{x_m, x_{m+1}\}$ and $\mathcal{K}_m = \mathcal{K}_{m+1} = \{\tau_{m+1}\}$. Therefore, $l(m, m+1) = 1$ and $l(m+1, m+1) = 2$, where the index $l(i,k)$ for $x_i \in \mathcal{T}_2$ and $\tau_k \in \mathcal{K}_i$ was defined just before (4.18). We also have $\mathcal{K}_m^* = \mathcal{K}_{m+1}^* = \{\tau_{m+1}\}$. Therefore from (4.18), we have $\Delta_m = [\delta_1^{(m+1)}]^{-2}$ and $\Delta_{m+1} = [\delta_2^{(m+1)}]^{-2}$. Also the vertices $x_m, x_{m+1} \in \mathcal{T}_2$ are associated with the diagonal elements $(\mathbf{A}_{22})_{j_m j_m}, (\mathbf{A}_{22})_{j_{m+1} j_{m+1}}$, respectively, of $\mathbf{A}_{22}$, where $j_m = 1, j_{m+1} = 2$. It is easy to check that

$$1 < (\mathbf{A}_{22})_{11}^{-1}\Delta_m,\ (\mathbf{A}_{22})_{22}^{-1}\Delta_{m+1} \leq 6$$

and the Assumption 3 is satisfied with $L_3 = 1$ and $U_3 = 6$.

We have shown in Remark 4.14 that the Assumption 1 is always satisfied in 1-d; in this case $L_1 = \frac{1}{2}$ and $U_1 = 2$. Therefore, from Theorem 4.12, we have that $\mathfrak{K}(\mathbf{A}) = \mathcal{O}(h^{-2})$, and thus the GFEM with $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ is indeed an SGFEM. We further note that Assumptions 1, 2, 3 are satisfied for any $0 < \beta < 1$, i.e., the constants L_1, U_1, L_2, U_2, L_3 and U_3 are independent of β. Therefore $\mathfrak{K}(\mathbf{A}) = \mathcal{O}(h^{-2})$ even when $\beta \approx 0$ or $\beta \approx 1$, i.e., when the point of discontinuity b^* of $a_1(x)$ is close to the one of the vertices x_i (see also Remark 5.1).

We next consider the (2.1) with $a(x) = a_2(x)$. We again choose $V_i = \text{span}\{1, \varphi_1^{[i]} = \int_{x_{i-1}}^{x} (1/a_2(t))dt\}$. If the points of discontinuity b_1^*, b_2^* of $a_2(x)$ are separated, e.g., $b_1^* \in \overset{o}{\tau}_l$ and $b_2^* \in \overset{o}{\tau}_{l^*}$ with $|l - l^*| \geq 2$, then we can again show that the GFEM with $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ is an SGFEM, based on the arguments given above.

Suppose there is an m such that $b_1^* \in \overset{o}{\tau}_m$ and $b_2^* \in \overset{o}{\tau}_{m+1}$. Moreover, suppose $b_1^* = x_{m-1} + h/2$ and $b_2^* = x_m + \beta h$ with $0 < \beta < 1$. Note that b_1^* is away from the vertices, whereas, b_2^* could be close to either x_m ($\beta \approx 0$) or x_{m+1} ($\beta \approx 1$). As before, let $\overline{V}_i = \text{span}\{\overline{\varphi}_1^{[i]}\}$; clearly, $\overline{V}_i = \{0\}$ for $i \neq m-1, m, m+1$. Therefore $\mathcal{T}_2 = \{x_{m-1}, x_m, x_{m+1}\}$ and

$$\overline{S}_2 = \sum_{i=m-1,m,m+1} N_i \overline{V}_i\,.$$

We further note that $\overline{\varphi}_1^{[m-1]}\big|_{\tau_{m-1}} = \overline{\varphi}_1^{[m+1]}\big|_{\tau_{m+2}} = 0$. Also it can be shown that $\overline{\varphi}_1^{[m-1]}\big|_{\tau_m} = \overline{\varphi}_1^{[m]}\big|_{\tau_m}$ and $\overline{\varphi}_1^{[m]}\big|_{\tau_{m+1}} = \overline{\varphi}_1^{[m+1]}\big|_{\tau_{m+1}}$. Therefore $\mathcal{K}_{enr} = \{\tau_m, \tau_{m+1}\}$ (i.e., τ_m, τ_{m+1} are the only enriched elements), and hence $\mathbf{A}_{22}$ is assembled from local stiffness matrices $A_{22}^{(m)}$ and $A_{22}^{(m+1)}$.

From direct computation, we get

$$A_{22}^{(m)} = \begin{bmatrix} \frac{h}{16} & \frac{h}{32} \\ \frac{h}{32} & \frac{h}{32} \end{bmatrix},$$

and it is associated with the vertices x_{m-1} and x_m. The matrix $A_{22}^{(m+1)}$ is same as in (5.1) and is associated with x_m and x_{m+1}. We choose $D^{(m)} =$

$\text{diag}(\delta_1^{(m)},\ \delta_2^{(m)})$ with $\delta_1^{(m)} = \delta_2^{(m)} = h^{-1/2}$ and $D^{(m+1)} = \text{diag}(\delta_1^{(m+1)},\ \delta_2^{(m+2)})$ with $\delta_1^{(m+1)}$, $\delta_2^{(m+2)}$ as given in (5.2). Then

$$\hat{A}_{22}^{(m)} := D^{(m)} A_{22}^{(m)} D^{(m)} = \begin{bmatrix} \frac{1}{16} & \frac{1}{32} \\ \frac{1}{32} & \frac{1}{16} \end{bmatrix}.$$

Clearly the diagonal elements of $\hat{A}_{22}^{(m)}$ are $O(1)$. The eigenvalues of $\hat{A}_{22}^{(m)}$ are $\lambda_1 = 1/32$ and $\lambda_2 = 3/32$ and therefore (recall Remark 4.14),

$$\frac{1}{32}\|[D^{(m)}]^{-1}\mathbf{x}\|^2 \leq \mathbf{x}^T A_{22}^{(m)} \mathbf{x} \leq \frac{3}{32}\|[D^{(m)}]^{-1}\mathbf{x}\|^2. \tag{5.5}$$

Next, the matrix $\hat{A}_{22}^{(m+1)} := D^{(m+1)} A_{22}^{(m+1)} D^{(m+1)}$ is same as the matrix given in (5.3). The diagonal elements of $\hat{A}_{22}^{(m+1)}$ are $O(1)$ and its eigenvalues are $\lambda_1 = (2-\beta)/6$ and $\lambda_2 = (1+\beta)/2$. Therefore

$$\frac{1}{6}\|[D^{(m+1)}]^{-1}\mathbf{x}\|^2 \leq \mathbf{x}^T A_{22}^{(m+1)} \mathbf{x} \leq \|[D^{(m+1)}]^{-1}\mathbf{x}\|^2, \quad \forall\ \mathbf{x} \in \mathbf{R}^2.$$

Thus the above inequality together with (5.5) implies that Assumption 2 is satisfied with $L_2 = \frac{1}{32}$ and $U_2 = 1$ for all $0 < \beta < 1$.

The matrix $\mathbf{A}_{22}$ is assembled from the matrices $A_{22}^{(m)}$, $A_{22}^{(m+1)}$ and is given by

$$\mathbf{A}_{22} = \begin{bmatrix} \frac{h}{16} & \frac{h}{32} & 0 \\ \frac{h}{16} & \frac{h}{32} + \frac{h\beta(1-\beta)^2}{3}\left(\frac{3}{2} + \beta - 2\beta^2\right) & \frac{h\beta^2(1-\beta)^2}{3}\left(\frac{1}{2} + 2\beta\right) \\ 0 & \frac{h\beta^2(1-\beta)^2}{3}\left(\frac{1}{2} + 2\beta\right) & \frac{h\beta^2(1-\beta)}{3}\left(1 + 2\beta^2\right) \end{bmatrix}.$$

We choose $\mathbf{D}_2 = \text{diag}(d_1,\ d_2,\ d_3)$ with $d_i = (\mathbf{A}_{22})_{ii}^{-1/2}$. Clearly the diagonal elements of $\widehat{\mathbf{A}}_{22} := \mathbf{D}_2 \mathbf{A}_{22} \mathbf{D}_2$ are equal to 1. Consider the vertex $x_m \in \mathcal{T}_2$. Then $\mathcal{K}_m = \{\tau_m, \tau_{m+1}\}$ and $l(m,m) = 2$, $l(m,m+1) = 1$. Also in this case, $\mathcal{K}_i^* = \mathcal{K}_i$. Therefore, from (4.18), we have $\Delta_m = [\delta_2^{(m)}]^{-2} + [\delta_1^{(m+1)}]^{-2}$. Similarly, we can show that $\Delta_{m-1} = [\delta_1^{(m)}]^{-2}$ and $\Delta_{m+1} = [\delta_2^{(m+1)}]^{-2}$. We also note that the vertices $x_{m-1}, x_m, x_{m+1} \in \mathcal{T}_2$ are associated with the diagonal elements $(\mathbf{A}_{22})_{j_{m-1}j_{m-1}}, (\mathbf{A}_{22})_{j_m j_m}, (\mathbf{A}_{22})_{j_{m+1}j_{m+1}}$, respectively, of $\mathbf{A}_{22}$, where $j_{m-1} = 1, j_m = 2, j_{m+1} = 3$. An easy calculation yields

$$1 \leq (\mathbf{A}_{22})_{11}^{-1}\Delta_{m-1},\ (\mathbf{A}_{22})_{22}^{-1}\Delta_m,\ (\mathbf{A}_{22})_{33}^{-1}\Delta_{m+1} \leq 16.$$

Thus Assumption 3 is satisfied with $L_3 = 1$ and $U_3 = 16$ for all $0 < \beta < 1$. We have shown before that Assumption 1 is always satisfied in 1-d. Therefore from Theorem 4.12, we infer that $\mathfrak{K}(\mathbf{A}) = \mathcal{O}(h^{-2})$; the result is true even when $\beta \approx 0$ or $\beta \approx 1$. Thus the GFEM with $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ is indeed an SGFEM.

We remark that for $a(x) = a_1(x)$ or $a(x) = a_2(x)$, we can show that there exists $\bar{\xi}^i \in \overline{V}_i$ such that $\|u - \mathcal{I}_{\omega_i} u - \bar{\xi}^i\|_{\mathcal{E}(\omega_i)} = O(h)$ for each $x_i \in \mathcal{T}_2$. Thus using the standard interpolation estimates and using Theorem 4.7, we have $\|u - u_h\|_{\mathcal{E}(\Omega)} = O(h)$, where u_h is the SGFEM solution.

Remark 5.1 Note that $A_{22}^{(m+1)}$ and thus $\mathbf{A}_{22}$, $\mathbf{A}$ degenerate as $\beta \to 0$ or $\beta \to 1$. Let ϵ_0 be small, say, $\epsilon_0 = 10^{-14}$. We adjust the implementation when $\beta \le \epsilon_0$ or $1-\beta \le \epsilon_0$ by setting $\beta = \epsilon_0$ or $1-\beta = \epsilon_0$, respectively. We emphasize that $\mathfrak{K}(\mathbf{A})$ is bounded independently of β. ∎

5.2 Problems with singular solutions

Let $a(x) = 1$ in (2.1) and suppose $f(x)$ be such that the solution u of (2.1)-(2.2) is of the form $u = x^\alpha + u_0$, where $\frac{1}{2} < \alpha < \frac{3}{2}$, $\alpha \neq 1$, and u_0 is smooth with $u_0(0) = 0$. Clearly $u \notin H^2(\Omega)$. Let $0 < D < 1$ and set $\Omega_l := (0, D)$, $\Omega_r := (D, 1)$. Then $u \in H^2(\Omega_r)$ and $|u|_{H^2(\Omega_r)} \le C[\,|x^\alpha|_{H^2(\Omega_r)} + |u_0|_{H^2(\Omega_r)}]$. Clearly, $|u|_{H^2(\Omega_r)}$ depends on D and is extremely large for $D \approx 0$.

We consider $\mathcal{T}_1 \subset \mathcal{T}$ as before. Let $\mathcal{T}_2 := \{x_i \in \mathcal{T} : \omega_i \cap \Omega_l \neq \emptyset\}$, where the patches ω_i have been defined before. Clearly, $x_0, x_1 \in \mathcal{T}_2$. Let $k^* \in I$ be the largest index such that $x_i \in \mathcal{T}_2$ for $0 \le i \le k^* - 1$. For $x_i \in \mathcal{T}_2$, let

$$V_i = \text{span}\{1, \varphi_1^{[i]} = (x - x_i), \varphi_2^{[i]} = x^\alpha|_{\omega_i}\},$$

and for $x_i \in \mathcal{T}\backslash\mathcal{T}_2$, let $V_i = \text{span}\{1, \varphi_1^{[i]} = (x - x_i)\}$. Clearly, $\overline{V}_i = \{0\}$ for $x_i \in \mathcal{T}\backslash\mathcal{T}_2$. For $x_i \in \mathcal{T}_2$, we have

$$\overline{V}_i = \text{span}\{\overline{\varphi}_2^{[i]} = \sigma_i\} \neq 0,$$

where $\sigma_i := (x^\alpha - \mathcal{I}_{\omega_i} x^\alpha)|_{\omega_i}$; recall $\mathcal{I}_{\omega_i} x^\alpha$ is the piecewise linear polynomial that interpolates x^α at the vertices $\{x_{i-1}, x_i, x_{i+1}\}$ of ω_i for $i \neq 0$, and $\mathcal{I}_{\omega_0} x^\alpha$ interpolates x^α at $\{x_0, x_1\}$. For an element $\tau_k \subset \overline{\omega}_i$ (with $x_i \in \mathcal{T}_2$), we define $\sigma^{(k)} := (x^\alpha - I_k x^\alpha)|_{\tau_k}$, where $I_k x^\alpha \in \mathcal{P}^1(\tau_k)$ interpolates x^α at x_{k-1}, x_k. Clearly, $\mathcal{I}_{\omega_i} x^\alpha = I_k x^\alpha$ on $\tau_k \subset \overline{\omega}_i$. It is also clear that $[\sigma^{(k)}]' \not\equiv 0$ on $\tau_k \subset \overline{\omega}_i$.

We define $\overline{\mathcal{S}}_2 = \sum_{i=0}^{k^*-1} N_i \overline{V}_i$ and we consider the GFEM based $\mathcal{S} = \mathcal{S}_1 + \overline{S}_2$. We first address the convergence of the GFEM solution u_h. It is easy to show that for $x_i \in \mathcal{T}_2$, there exists $\bar{\xi}_i \in V_i$ such that

$$\|u - \mathcal{I}_{\omega_i} u - \bar{\xi}_i\|_{\mathcal{E}(\omega_i)} \le Ch|u_0|_{H^2(\omega_i)}\,.$$

Also for $x_i \in \mathcal{T}\backslash\mathcal{T}_2$, from standard interpolation result we have

$$\|u - \mathcal{I}_{\omega_i} u\|_{\mathcal{E}(\omega_i)} \le Ch|u|_{H^2(\omega_i)} \le Ch[|x^\alpha|_{H^2(\omega_i)} + |u_0|_{H^2(\omega_i)}].$$

Therefore, from Theorem 4.7, there exists $v \in \mathcal{S}_1 + \overline{\mathcal{S}}_2$ such that

$$\begin{aligned}\|u - v\|_{\mathcal{E}(\Omega)} &\le Ch\Big[\sum_{x_i \in \mathcal{T}_2} |u_0|^2_{H^2(\omega_i)} + \sum_{x_i \in \mathcal{T}\backslash\mathcal{T}_2} \{|x^\alpha|^2_{H^2(\omega_i)} + |u_0|^2_{H^2(\omega_i)}\}\Big]^{1/2} \\ &\le Ch\big[|u_0|^2_{H^2(\Omega)} + |x^\alpha|^2_{H^2(\Omega_r)}\big]^{1/2}.\end{aligned}$$

Thus we have $\|u - u_h\|_{\mathcal{E}(\Omega)} \le Ch$, where u_h is the GFEM solution; note that C depends on $|x^\alpha|_{H^2(\Omega_r)}$ and thus on D.

We note that Ω_l is independent of h. However, if $D = h^\gamma$, $\gamma < 1$ (i.e., $|\Omega_l| = h^\gamma$), then one can show that $\|u - u_h\|_{\mathcal{E}(\Omega)} = O(h^{1-\gamma})$. Thus if we enrich only a fixed number of patches in the neighborhood of the singularity, we loose the optimal order of convergence.

We now address the scaled condition number of the stiffness matrix $\mathbf{A}$ of the GFEM. We note that the matrix $\mathbf{A}_{22}$ is assembled from element stiffness matrices $A_{22}^{(k)}$ for the element τ_k, where τ_k is enriched. We note that the set of enriched elements is given by $\mathcal{K}_{enr} := \{\tau_k \in \{\tau_l\}_{l\in I\setminus\{0\}} : x_k \in \mathcal{T}_2\}$. We further note that if $\tau_k \in \mathcal{K}_{enr}$, then $\tau_j \in \mathcal{K}_{enr}$ for $1 \le j \le k$. Also from the definition of k^*, it is clear that $\tau_{k^*} \in \mathcal{K}_{enr}$ and $\tau_j \notin \mathcal{K}_{enr}$ for $j \ge k^*+1$. Now, for $\tau_k \in \mathcal{K}_{enr}$, $k \ne k^*$, the matrices $A_{22}^{(k)}$ are of the form $A_{22}^{(k)} = \{b_{lm}^{(k)}\}_{l,m=1}^2$; the entries $b_{lm}^{(k)}$ are as given by

$$b_{lm}^{(k)} = \int_{\tau_k} (N_{k-2+l}\sigma)'(N_{k-2+m}\sigma)'\,dx\,.$$

Also, since $x_{k^*} \notin \mathcal{T}_2$, we have $A_{22}^{(k^*)} = [b_{11}^{(k^*)}]$ (an 1×1 matrix), where $b_{11}^{(k^*)}$ is given by the above expression.

Lemma 5.2 *The entries of the matrix $A_{22}^{(k)}$ are as follows:*

$$b_{11}^{(k)} = \int_{\tau_k} N_{k-1}^2 {\sigma'}^2\,dx, \quad b_{22}^{(k)} = \int_{\tau_k} N_k^2 {\sigma'}^2\,dx,$$
$$b_{12}^{(k)} = b_{21}^{(k)} = \int_{\tau_k} N_{k-1}N_k {\sigma'}^2\,dx\,.$$

The proof is easy and we do not present it here. □

It is clear from above that for $\tau_k \in \mathcal{K}_{enr}$ and $k \ne k^*$, the diagonal elements $b_{11}^{(k)}, b_{22}^{(k)} > 0$ and therefore $A_{22}^{(k)}$ is associated with x_{k-1}, x_k. Also $b_{11}^{(k^*)} > 0$ and thus $A_{22}^{(k^*)}$ is associated with x_{k^*-1}. A simple observation yields that the size of $\mathbf{A}_{22}$ is $k^* \times k^*$.

Let $\tau_k \in \mathcal{K}_{enr}$ and set $x_{k-1/2} := (k-\frac{1}{2})h$; $x_{k-1/2}$ is the mid-point of τ_k. We define

$$G_k = \left|[x^\alpha]''(x_{k-1/2})\right| = |\alpha(\alpha-1)(k-\tfrac{1}{2})^{\alpha-2}h^{\alpha-2}|.$$

Note that for $1 \le j \le k^*-1$, $\tau_{j+1} \in \mathcal{K}_{enr}$ implies $\tau_j \in \mathcal{K}_{enr}$, and we have

$$1 \le \frac{G_j}{G_{j+1}} = \Big(\frac{j+\frac{1}{2}}{j-\frac{1}{2}}\Big)^{2-\alpha} \le 3^{2-\alpha}. \tag{5.6}$$

We now obtain a few results, which will be used to establish that the GFEM with $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ is an SGFEM.

Lemma 5.3 *For $x_k \in \mathcal{K}_{enr}$, there exist positive constants C_1^*, C_2^*, independent of k and h but may depend on α, such that*

$$C_1^* h^{3/2} \le \frac{\|[\sigma^{(k)}]'\|_{L^2(\tau_k)}}{G_k} \le C_2^* h^{3/2}.$$

Proof: (a) First let $2 \le k \le k^*$ and let $g(x) = [\sigma^{(k)}]'(x)$ for $x \in \tau_k$. Then

$$\begin{aligned} \max|g'(x)| &= \max_{x\in\tau_k}|[\sigma^{(k)}]''(x)| = |\alpha(\alpha-1)|x_{k-1}^{\alpha-2} \\ &= |\alpha(\alpha-1)|(k-\tfrac{1}{2})^{\alpha-2}h^{\alpha-2}\Big(\frac{k-1}{k-\frac{1}{2}}\Big)^{\alpha-2} \\ &= G_k\Big(\frac{k-\frac{1}{2}}{k-1}\Big)^{2-\alpha} \le G_k\Big(\frac{3}{2}\Big)^{2-\alpha} := M_k. \end{aligned} \tag{5.7}$$

Similarly,

$$\min|g'(x)| = G_k\Big(\frac{k-\frac{1}{2}}{k}\Big)^{2-\alpha} \ge G_k\Big(\frac{3}{4}\Big)^{2-\alpha} := m_k. \tag{5.8}$$

We next note that g' does not change sign in τ_k and thus g is monotonic in τ_k. Also, since $\int_{\tau_k} g\,dx = \sigma^{(k)}\big|_{x_{k-1}}^{x_k} = 0$, there exists a unique $x_k^* = x_{k-1} + \zeta h \in \tau_k$ with $0 < \zeta < 1$ such that $g(x_k^*) = 0$ and x_k^* is characterized by

$$\int_{x_{k-1}}^{x_k^*} |g|\,dx = \int_{x_k^*}^{x_k} |g|\,dx\,. \tag{5.9}$$

We now obtain bounds on ζ, independent of k. Since $g(x_k^*) = 0$, it is clear from the mean value theorem, (5.7), and (5.8) that

$$\begin{aligned} m_k|x - x_k^*| &\le \min|g'||x - x_k^*| \le |g(x)| \\ &\le \max|g'||x - x_k^*| \le M_k|x - x_k^*|, \quad x \in \tau_k\,. \end{aligned} \tag{5.10}$$

Consequently,

$$m_k\,\frac{\zeta^2h^2}{2} \le \int_{x_{k-1}}^{x_k^*} |g|\,dx \le M_k\frac{\zeta^2h^2}{2} \tag{5.11}$$

and

$$m_k\,\frac{(1-\zeta)^2h^2}{2} \le \int_{x_k^*}^{x_k} |g|\,dx \le M_k\frac{(1-\zeta)^2h^2}{2}\,. \tag{5.12}$$

Now from (5.9), (5.11), and (5.12), we have

$$m_k\,\frac{\zeta^2h^2}{2} \le \int_{x_{k-1}}^{x_k^*} |g|\,dx = \int_{x_k^*}^{x_k} |g|\,dx \le M_k\frac{(1-\zeta)^2h^2}{2}\,,$$

and thus

$$\zeta \le \frac{\sqrt{M_k}}{\sqrt{M_k}+\sqrt{m_k}} = \frac{(3/2)^{(2-\alpha)/2}}{(3/2)^{(2-\alpha)/2}+(3/4)^{(2-\alpha)/2}}\,,$$

where we used the definition of m_k and M_k given in (5.8) and (5.7) respectively.

Using a similar argument we obtain a lower bound of ζ; we summarize the bounds of ζ as

$$\begin{aligned} &\zeta_l \le \zeta \le \zeta_r, \quad \text{where} \\ &\zeta_l = \frac{(3/4)^{(2-\alpha)/2}}{(3/2)^{(2-\alpha)/2}+(3/4)^{(2-\alpha)/2}} \\ \text{and} \qquad &\zeta_r = \frac{(3/2)^{(2-\alpha)/2}}{(3/2)^{(2-\alpha)/2}+(3/4)^{(2-\alpha)/2}}\,. \end{aligned} \tag{5.13}$$

Finally, from (5.10) and using the definition of M_k (given in (5.7)), we get

$$\begin{aligned}\int_{\tau_k} |g|^2\,dx &= \int_{x_{k-1}}^{x_k^*} |g|^2\,dx + \int_{x_k^*}^{x_k} |g|^2\,dx \\ &\le M_k^2 \int_{x_{k-1}}^{x_k^*} (x_k^* - x)^2\,dx + M_k^2 \int_{x_k^*}^{x_k} (x - x_k^*)^2\,dx \\ &= \frac{M_k^2 h^3}{3}[\zeta^3 + (1-\zeta)^3] \\ &\le \frac{G_k^2 (3/2)^{2(2-\alpha)} h^3}{3}[\zeta_r^3 + (1-\zeta_l)^3],\end{aligned}$$

and similarly, we have

$$\begin{aligned}\int_{\tau_k} |g|^2\,dx &\ge \frac{m_k^2 h^3}{3}[\zeta^3 + (1-\zeta)^3] \\ &\ge \frac{G_k^2 (3/4)^{2(2-\alpha)} h^3}{3}[\zeta_l^3 + (1-\zeta_r)^3].\end{aligned}$$

Thus

$$\bar{C}_1^* h^{3/2} \le \frac{\|[\sigma^{(k)}]'\|_{L^2(\tau_k)}}{G_k} \le \bar{C}_2^* h^{3/2}, \quad \text{for } 2 \le i \le N,$$

where $\bar{C}_1^* = (3/4)^{2-\alpha}\sqrt{[\zeta_l^3 + (1-\zeta_r)^3]/3}$ and $\bar{C}_2^* = (3/2)^{2-\alpha}\sqrt{[\zeta_r^3 + (1-\zeta_l)^3]/3}$.

(b) We now consider $k = 1$. We note that on $\tau_1 = (0, h)$,

$$[\sigma^{(1)}]'(x) = \alpha x^{\alpha-1} - h^{\alpha-1}.$$

By a direct computation, we get

$$\int_0^h |[\sigma^{(1)}]'|^2\,dx = \frac{(\alpha-1)^2 h^{2\alpha-1}}{2\alpha-1}\,.$$

Therefore,

$$\frac{\int_0^h |[\sigma^{(1)}]'|^2\,dx}{G_1^2} = \frac{(\alpha-1)^2 h^{2\alpha-1}}{(2\alpha-1)\alpha^2(\alpha-1)^2 h^{2\alpha-4} 2^{4-2\alpha}} = \frac{h^3}{(2\alpha-1)\,\alpha^2\,2^{4-2\alpha}} := \bar{C}^*.$$

Hence we get the desired result with $C_1^* = \min(\bar{C}_1^*, \bar{C}^*)$ and $C_2^* = \max(\bar{C}_2^*, \bar{C}^*)$. □

Lemma 5.4 *Suppose $\tau_k \in \mathcal{K}_{enr}$ and let $l_k(x)$ be a linear function, defined on τ_k, such that $l_k(x_{k-1}) = y_1$ and $l_k(x_k) = y_2$. Then there exists a positive constant C_3^*, independent of k and h but may depend on α, such that*

$$\|[\sigma^{(k)}]' l_k\|_{L^2(\tau_k)} \ge C_3^* G_k h^{3/2} (y_1^2 + y_2^2)^{1/2}.$$

Proof: (a) Let $2 \le k \le k^*$ and define $g(x) = [\sigma^{(k)}]'(x)$ for $x \in \tau_k$. On τ_k, we have seen in the proof of Lemma 5.3 that $g(x_k^*) = 0$ where $x_k^* = x_{k-1} + \zeta h$ and $0 < \zeta_l \le \zeta \le \zeta_r < 1$. We have also seen that

$$m_k|x - x_k^*| \le |g(x)| \le M_k|x - x_k^*|, \quad \forall\ x \in \tau_k\,, \tag{5.14}$$

where

$$m_k = G_k(3/4)^{2-\alpha}, \quad M_k = G_k(3/2)^{2-\alpha}.$$

Let

$$\bar{x}_k \equiv x_{k-1} + \zeta_r h + \frac{1-\zeta_r}{2}\, h.$$

Then

$$|\bar{x}_k - x_k^*| = |x_{k-1} + \zeta_r h + \frac{1-\zeta_r}{2}\, h - x_{k-1} - \zeta h| \ge \frac{1-\zeta_r}{2}\, h. \tag{5.15}$$

Also from the definition of $\bar{x}_k$, it is clear that $g(x) \neq 0$ in $(\bar{x}_k, x_k)$ and thus from (5.14) and (5.15), we have

$$|g(x)| \ge m_k|\bar{x}_k - x_k^*| \ge \frac{1-\zeta_r}{2}\, m_k h. \tag{5.16}$$

Therefore,

$$\int_{x_{k-1}}^{x_k} |g|^2|l_k|^2\, dx \ge \int_{\bar{x}_k}^{x_k} |g|^2|l_k|^2\, dx \ge m_k^2 h^2 \frac{(1-\zeta_r)^2}{4} \int_{\bar{x}_k}^{x_k} |l_k|^2\, dx\,. \tag{5.17}$$

We make the change of variable $y = \frac{x-\bar{x}_k}{h}\frac{2}{1-\zeta_r}$. Then

$$F(y_1, y_2) := \frac{\int_{\bar{x}_k}^{x_i} |l_k|^2\, dx}{y_1^2 + y_2^2} = \frac{(1-\zeta_r)h \int_0^1 |\tilde{l}(y)|^2\, dy}{2(y_1^2 + y_2^2)}\,,$$

where

$$\tilde{l}(y) = l_k(\bar{x}_k + \frac{(1-\zeta_r)hy}{2}) = y_1\, \frac{1-\zeta_r}{2}(1-y) + y_2\, (\frac{1+\zeta_r}{2} + \frac{1-\zeta_r}{2}y).$$

Thus $F(y_1, y_2)$ is independent of k. We next note that $F(y_1, y_2)$ is a continuous function and $F(\beta y_1, \beta y_2) = F(y_1, y_2)$. It is well known that the minimum of $F(y_1, y_2)$ is attained on the compact set $y_1^2 + y_2^2 = 1$. Hence there is a constant C_{min}, independent of k but may depend on ζ_r, such that

$$0 < C_{min}^2 \frac{(1-\zeta_r)h}{2} \le F(y_1, y_2) = \frac{\int_{\bar{x}_k}^{x_k} |l|^2\, dx}{y_1^2 + y_2^2}\,.$$

Thus from (5.17), we have

$$\begin{aligned} \int_{x_{k-1}}^{x_k} |g|^2|l_k|^2\, dx &\ge C_{min}^2 m_k^2 h^3 \frac{(1-\zeta_r)^3}{8}(y_1^2 + y_2^2) \\ &= B_1^{*2} G_k^2 h^3 (y_1^2 + y_2^2), \end{aligned}$$

where $B_1^{*2} = (3/4)^{2(2-\alpha)} C_{min}^2 (1-\zeta_r)^3/8$.

(b) We now consider $k = 1$. On $\tau_1 = (0, h)$, we have $g(x) = \alpha x^{\alpha-1} - h^{\alpha-1}$. It is easy to see that $g(x_1^*) = 0$, where $x_1^* = \zeta h$ with $\zeta = \zeta(\alpha) = \alpha^{1/1-\alpha}$. Since $\zeta(\alpha)$ is increasing for $\frac{1}{2} < \alpha < \frac{3}{2}$ (with ζ redefined for $\alpha = 1$), we have $\zeta \le \zeta^* \equiv \zeta(3/2) = (2/3)^2$.

Set $\bar{x}_1 = \zeta^* h + (1-\zeta^*)h/2$. Since $|g(x)|$ is increasing in (x_1^*, h), we have $|g(x)| \ge \bar{g}_{min} \equiv |g(\bar{x}_1)|$ on $(\bar{x}_1, h)$. Therefore,

$$\int_0^h |g|^2 |l_1|^2 \, dx > \int_{\bar{x}_1}^h |g|^2 |l_1|^2 \, dx \ge \bar{g}_{min}^2 \int_{\bar{x}_1}^h |l_1|^2 \, dx$$
$$= \frac{\bar{g}_{min}^2 G_1^2}{\alpha^2 |\alpha-1|^2 h^{2(\alpha-2)}/2^{2(\alpha-2)}} \int_{\bar{x}_1}^h |l_1|^2 \, dx = C^2 G_1^2 h^2 \int_{\bar{x}_1}^h |l_1|^2 \, dx \,,$$

where

$$C^2 = \frac{2^{4-2\alpha} \bar{g}_{min}^2}{\alpha^2(\alpha-1)^2 h^{2(\alpha-1)}} = \frac{2^{4-2\alpha}[\alpha \frac{(1+\zeta^*)}{2} - 1]^2}{\alpha^2(\alpha-1)^2} \,.$$

As before, we can show that

$$\int_{\bar{x}_1}^h |l_1|^2 \, dx \ge C_{min} \frac{(1-\zeta^*)h}{2} (y_1^2 + y_2^2)^{1/2},$$

and therefore,

$$\int_0^h |g|^2 |l_1|^2 \, dx \ge B_2^{*2} G_1^2 h^3 (y_1^2 + y_2^2)^{1/2},$$

where $B_2^{*2} = C^2 C_{min}(1-\zeta^*)h/2$. Finally, defining $C_3^* = \min(B_1^*, B_2^*)$ and recalling that $g = [\sigma^{(k)}]'$, we get the desired result. □

Now, for $k \ne k^*$, consider the diagonal matrix $D^{(k)} = \mathrm{diag}(\delta_1^{(k)}, \delta_2^{(k)})$ with $\delta_1^{(k)} = \delta_2^{(k)} = G_k^{-1} h^{-3/2}$ and set $\hat{A}_{22}^{(k)} = D^{(k)} A_{22}^{(k)} D^{(k)}$. The diagonal elements of $\hat{A}_{22}^{(k)}$ (see Lemma 5.2) are

$$\bar{b}_{11}^{(k)} = \frac{1}{G_k^2 h^3} \int_{\tau_k} N_{k-1}^2 \sigma'^2 \, dx \,, \quad \bar{b}_{22}^{(k)} = \frac{1}{G_k^2 h^3} \int_{\tau_k} N_k^2 \sigma'^2 \, dx \,.$$

Using Lemmas 5.4 and 5.3, it is clear that

$$C_3^* \le \bar{b}_{11}^{(k)} \le \frac{1}{G_k^2 h^3} \int_{\tau_k} \sigma'^2 \, dx \le C_2^*,$$

where C_2^*, C_3^* are independent of k and h. Similarly,

$$C_3^* \le \bar{b}_{22}^{(k)} \le \frac{1}{G_k^2 h^3} \int_{\tau_k} \sigma'^2 \, dx \le C_2^*.$$

We let $D^{(k^*)} = [\delta_1^{(k^*)}]$ with $\delta_1^{(k^*)} = G_{k^*}^{-1} h^{-3/2}$. Using similar arguments we show the $C_3^* \le \bar{b}_{11}^{(k^*)} \le C_2^*$, where $\hat{A}_{22}^{(k^*)} = D^{(k^*)} A_{22}^{(k^*)} D^{(k^*)} = [\bar{b}_{11}^{(k^*)}]$. Thus the diagonal elements of $\hat{A}_{22}^{(k)}$ are $O(1)$ for all $\tau_k \in \mathcal{K}_{enr}$.

We next show that the element matrices $A_{22}^{(k)}$ satisfy the Assumption 2.

Proposition 5.5 *For $\tau_k \in \mathcal{K}_{enr}$, the matrices $A_{22}^{(k)}$ satisfies Assumption 2.*

Proof: Suppose $k \neq k^*$ and let $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$. Then, using Lemma 5.2, we have

$$\begin{aligned}\mathbf{x}^T A_{22}^{(k)} \mathbf{x} &= b_{11}^{(k)} x_1^2 + 2b_{12}^{(k)} x_1 x_2 + b_{22}^{(k)} x_2^2 \\ &= \int_{\tau_k} [x_1^2 N_{k-1}^2 + 2x_1 x_2 N_{k-1} N_k + x_2^2 N_k^2][\sigma^{(k)}]'^2 \, dx \\ &= \int_{\tau_k} [x_1 N_{k-1} + x_2 N_k]^2 [\sigma^{(k)}]'^2 \, dx \leq 2(x_1^2 + x_2^2) \int_{\tau_k} [\sigma^{(k)}]'^2 \, dx \,,\end{aligned}$$

and using Lemma 5.3, we have

$$\mathbf{x}^T A_{22}^{(k)} \mathbf{x} \leq C_2^* h^3 G_k^2 \|\mathbf{x}\|^2 = C_2^* \|[D^{(k)}]^{-1} \mathbf{x}\|^2.$$

Next from Lemma 5.4, it is immediate that

$$\begin{aligned}\mathbf{x}^T A_{22}^{(k)} \mathbf{x} &= \int_{\tau_k} [x_1 N_{k-1} + x_2 N_k]^2 [\sigma^{(k)}]'^2 \, dx \\ &\geq C_1^* G_k^2 h^3 (y_1^2 + y_2^2) = C_1^* \|[D^{(k)}]^{-1} \mathbf{x}\|^2.\end{aligned}$$

Similar bounds fox $A_{22}^{(k^*)} x^2$ for all $x \in \mathbb{R}$ also hold. Thus Assumption 2 is satisfied with $L_2 = C_1^*$ and $U_2 = C_2^*$. □

Next, recalling that $\mathbf{A}_{22}$ is $k^* \times k^*$, we choose the diagonal matrix $\mathbf{D}_2 = \text{diag}(d_1, d_2, \cdots, d_{k^*})$ with $d_j = (\mathbf{A}_{22})_{jj}^{-1/2}$. Clearly, the diagonal elements of $\widehat{\mathbf{A}}_{22} = \mathbf{D}_2 \mathbf{A}_{22} \mathbf{D}_2$ are equal to 1. Note that $(\mathbf{A}_{22})_{jj}$, $1 \leq j \leq k^*$, is associated with the vertex $x_{j-1} \in \mathcal{T}_2$. Also, $(\mathbf{A}_{22})_{11} = b_{11}^{(0)}$ and $(\mathbf{A}_{22})_{jj} = b_{22}^{(j-1)} + b_{11}^{(j)}$ for $2 \leq j \leq k^*$.

Now, for $x_i \in \mathcal{T}_2$ and $i \neq 0, k^* - 1$ (recall $x_i \notin \mathcal{T}_2$ for $k^* \leq i \leq N$), we have $\mathcal{K}_i = \{\tau_i, \tau_{i+1}\}$, $l(i,i) = 2$, $l(i, i+1) = 1$ and $\mathcal{K}_i^* = \mathcal{K}_i$. Also $\mathcal{K}_0 = \{\tau_1\}$, $l(0,1) = 1$, $\mathcal{K}_0^* = \mathcal{K}_0$ and $\mathcal{K}_{k^*-1} = \{\tau_{k^*}\}$, $l(k^* - 1, k^*) = 1$, $\mathcal{K}_{k^*-1}^* = \mathcal{K}_{k^*-1}$. Therefore, from (4.18), we have $\Delta_i = [\delta_2^{(i)}]^{-2} + [\delta_1^{(i+1)}]^{-2}$ for $x_i \in \mathcal{T}_2$ and $i \neq 1, k^* - 1$; also $\Delta_0 = [\delta_1^{(1)}]^{-2}$ and $\Delta_{k^*-1} = [\delta_1^{(k^*)}]^{-2}$.

We now show that Assumption 3 is satisfied.

Proposition 5.6 *Let $(\mathbf{A}_{22})_{jj}$, $1 \leq j \leq k^*$, be the diagonal elements of $\mathbf{A}_{22}$ and consider Δ_i for $x_i \in \mathcal{T}_2$, defined above. Then Assumption 3 is satisfied.*

Proof: Let $x_i \in \mathcal{T}_2$ and $i \neq 0, k^* - 1$; x_i is associated with $(\mathbf{A}_{22})_{j_i j_i}$, where $j_i = i + 1$. Therefore using the definition of $(\mathbf{A}_{22})_{i+1,i+1}$, $\delta_1^{(i+1)}$, and $\delta_2^{(i)}$, we have

$$(\mathbf{A}_{22})_{j_i j_i}^{-1} \Delta_i = (\mathbf{A}_{22})_{i+1,i+1}^{-1} \Delta_i = \frac{G_i^2 h^3}{b_{22}^{(i)} + b_{11}^{(i+1)}} + \frac{G_{i+1}^2 h^3}{b_{22}^{(i)} + b_{11}^{(i+1)}} \,. \tag{5.18}$$

Now using Lemmas 5.4, 5.3, and (5.6), it is immediate that

$$\begin{aligned} C_3^{*2} &\leq \frac{b_{22}^{(i)}}{G_i^2 h^3} \leq C_2^{*2}, \\ \frac{C_3^{*2}}{3^{2(2-\alpha)}} \leq \frac{b_{11}^{(i+1)}}{G_{i+1}^2 h^3}\frac{G_{i+1}^2}{G_i^2} &= \frac{b_{11}^{(i+1)}}{G_{i+1}^2 h^3} \leq C_2^{*2}, \end{aligned}$$

and therefore,

$$\frac{1}{2C_2^{*2}} \leq \frac{G_i^2 h^3}{b_{22}^{(i)} + b_{11}^{(i+1)}} \leq \frac{3^{2(2-\alpha)}}{C_3^{*2}(1+3^{2(2-\alpha)})}. \tag{5.19}$$

Similarly, we get

$$\frac{1}{C_2^{*2}(1+3^{2(2-\alpha)})} \leq \frac{G_{i+1}^2 h^3}{b_{22}^{(i)} + b_{11}^{(i+1)}} \leq \frac{1}{2C_2^{*2}},$$

and combining (5.18),(5.19), we infer that there exist constants L_3, U_3, such that

$$L_3 \leq (\mathbf{A}_{22})^{-1}_{j_i j_i}\Delta_i \leq U_3,$$

where

$$\begin{aligned} L_3 &= \frac{1}{2C_2^{*2}} + \frac{1}{C_2^{*2}(1+3^{2(2-\alpha)})}, \\ U_3 &= \frac{1}{2C_3^{*2}} + \frac{3^{2(2-\alpha)}}{C_3^{*2}(1+3^{2(2-\alpha)})}. \end{aligned}$$

Thus Assumption 3 hold for $x_i \in \mathcal{T}_2$, $i \neq 0, k^*-1$. The proofs for $x_i \in \mathcal{T}_2$, $i = 0, k^*-1$ are simpler and we do not include them here, □

Based on Propositions 5.5, 5.6, it is clear that Assumptions 2 and 3 are satisfied. Assumption 1 always hold in 1-d. Thus from Theorem 4.12, $\mathfrak{K}(\mathbf{A}) = O(h^{-2})$ and the GFEM with $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ is an SFEM.

5.3 Problems with discontinuous solutions

We now address a problem, which is different from (2.1). Let $\Omega = (0,1)$ and set $\Omega_l = (0,c)$ and $\Omega_r = (c,1)$, where $0 < c < 1$ is fixed. Consider

$$H(\Omega) := \Big\{ v \in L_2(\Omega) : v(0) = v(1) = 0, \int_{\Omega_l} v'^2 dx < \infty, \text{ and } \int_{\Omega_l} v'^2 dx < \infty \Big\}.$$

Then $(H(\Omega), \|\cdot\|_H)$ is a Hilbert space, where

$$\|v\|_H^2 := |v|^2_{H^1(\Omega_l)} + |v|^2_{H^1(\Omega_r)}.$$

We note that $H_0^1(\Omega) \subset H(\Omega)$ and functions in $H(\Omega)$ may have jump discontinuity at $x = c$.

For $f \in L_2(\Omega)$, we consider the problem

$$u \in H(\Omega),\ B(u,v) = F(v), \quad \forall\ v \in H(\Omega), \tag{5.20}$$

where

$$B(u,v) := \int_{\Omega_l} u'v'\,dx + \int_{\Omega_r} u'v'\,dx \text{ and } F(v) := \int_{\Omega} fv\,dx.$$

The bilinear form $B(\cdot,\cdot)$ is coercive and bounded in $H(\Omega)$. Also $F(\cdot)$ is a bounded linear functional on $H(\Omega)$. Thus the problem (5.20) has a unique solution.

If f and the solution $u \in H(\Omega)$ of (5.20) are smooth in Ω_l and Ω_r, then u is the solution of the boundary value problem

$$\begin{aligned} &-u'' = f \text{ on } \Omega_l, \quad -u'' = f \text{ on } \Omega_r, \\ &u(0) = u(1) = 0 \text{ and } u'(c^-) = u'(c^+) = 0. \end{aligned}$$

This problem mimics the problem with a crack in higher dimensions, where the solution is discontinuous along the crack line away from the crack-tip.

We now give a characterization of the solution of (5.20). We will use the Heaviside function

$$H_c(x) = \begin{cases} 1, & 0 \le x < c; \\ -1, & c \le x \le 1. \end{cases} \tag{5.21}$$

Lemma 5.7 *Suppose $u \in H(\Omega)$ such that $u'(c^-) = u'(c^+) = 0$ and*

$$\int_{\Omega_l} (u'')^2 dx < \infty, \quad \int_{\Omega_l} (u'')^2 dx < \infty.$$

Then

$$u(x) = s(x) + \tilde{u}(x),$$

where s is a step function with discontinuity at $x = c$ and $\tilde{u} \in H^2(\Omega)$

Proof: We first note that $u(c^-)$ and $u(c^+)$ are well defined. We define

$$\tilde{u} = u - \frac{u(c^-) - u(c^+)}{2}\,[H_c - 1], \tag{5.22}$$

where $H_c(x)$ is given in (5.21). It is easy to check that

$$\begin{aligned} &\tilde{u}\big|_{\Omega_l} = u\big|_{\Omega_l},\ \tilde{u}(c^-) = u(c^-), \\ &\tilde{u}'\big|_{\Omega_l} = u'\big|_{\Omega_l}, \text{ and } \tilde{u}'(c^-) = u'(c^-). \end{aligned}$$

Similarly,

$$\begin{aligned} &\tilde{u}\big|_{\Omega_r} = u\big|_{\Omega_r} + [u(c^-) - u(c^+)],\ \tilde{u}(c^+) = u(c^-), \\ &\tilde{u}'\big|_{\Omega_r} = u'\big|_{\Omega_r}, \text{ and } \tilde{u}'(c^+) = u'(c^+). \end{aligned}$$

Note that $\tilde{u}(c^-) = \tilde{u}(c^+) = u(c^-)$. We define $\tilde{u}(c) = u(c^-)$ so that $\tilde{u}$ is continuous at $x = c$ and thus is continuous on Ω. Also since $u'(c^-) = u'(c^+) = 0$, it is clear from above that $\tilde{u}'(c^-) = \tilde{u}'(c^+) = 0$. We define $\tilde{u}'(c) = 0$ so that $\tilde{u}'$ is continuous at $x = c$, and consequently $\tilde{u}'$ is continuous in Ω. Moreover,

$$\int_\Omega (\tilde{u}'')^2 dx = \int_{\Omega_l} (\tilde{u}'')^2 dx + \int_{\Omega_r} (\tilde{u}'')^2 dx = \int_{\Omega_l} (u'')^2 dx + \int_{\Omega_r} (u'')^2 dx \leq \infty.$$

Thus $\tilde{u} \in H^2(\Omega)$ and considering $s = \frac{u(c^-)-u(c^+)}{2}\,[H_c - 1]$ in (5.22), we get the desired result. □

Suppose $c \notin \mathcal{T}$, i.e., c is not a vertex of the mesh. Therefore, there exists an m such that $c \in \tau_{m+1}$ and hence, $c \in \omega_m \cap \omega_{m+1}$. We assume that $m \neq 1, N$; this is always achieved for h small. Since $u(0) = u(1) = 0$, we consider $\mathcal{T}_1 = \mathcal{T}\backslash\{x_0, x_N\}$ (see Remark 4.6).

For $1 \leq i \leq N-1$, we consider $V_i = \text{span}\{1, \varphi_1^{[i]} = (x-x_i), \varphi_2^{[i]} = H_c(x)\}$ and we set $V_0 = \text{span}\{\varphi_1^{[0]} = (x-x_0)\}$ and $V_N = \text{span}\{\varphi_1^{[N]} = (x-x_N)\}$. Note that $V_i \in H(\omega_i)$ for $i \in I$ (i.e., for $x_i \in \mathcal{T}$). Clearly, $\overline{V}_i = \{0\}$ for $i \in I\backslash\{m, m+1\}$. We set $\mathcal{T}_2 = \{x_m, x_{m+1}\} \subset \mathcal{T}$ and define

$$\overline{\mathcal{S}}_2 = N_m \overline{V}^1_m + N_{m+1} \overline{V}^1_{m+1}.$$

We consider the GFEM with $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$.

Since $\varphi_2^{[m]} = H_c$ is constant in τ_m, we have $\overline{\varphi}_2^{[m]}|_{\tau_m} = 0$. Similarly, $\overline{\varphi}_2^{[m+1]}|_{\tau_{m+2}} = 0$. Therefore $\mathcal{K}_{enr} = \{\tau_{m+1}\}$. Moreover, the functions $\overline{\varphi}_2^{[m]}$, $\overline{\varphi}_2^{[m+1]}$ are discontinuous at $x = c$, their values are zero at $x = x_m, x_{m+1}$, and $\overline{\varphi}_2^{[m]}|_{\tau_{m+1}} = \overline{\varphi}_2^{[m+1]}|_{\tau_{m+1}}$.

We assume that f is such that solution $u \in H(\Omega)$ of (5.20) satisfies the assumptions of Lemma 5.7 and $u = s + \tilde{u}$, where s is a step-function with a discontinuity at $x = c$ and $\tilde{u} \in H^2(\Omega)$. We now address the convergence of the GFEM solution. We first note that Theorem 4.7 hold for $u \in H(\Omega)$ with $\mathcal{E}(\Omega), \mathcal{E}(\omega_i)$ replaced by $H(\Omega), H(\omega_i)$ and with $\overline{V}_i \in H(\omega_i)$. Now for $x_i \in \mathcal{T}\backslash\mathcal{T}_2$, we have $u \in H^2(\omega_i)$ and from the standard interpolation result

$$\|u - \mathcal{I}_{\omega_i} u\|_{H(\omega_i)} = \|u - \mathcal{I}_{\omega_i} u\|_{\mathcal{E}(\omega_i)} \leq Ch|u|_{H^2(\omega_i)} = Ch|\tilde{u}|_{H^2(\omega_i)}. \tag{5.23}$$

For $x_m \in \mathcal{T}_2$, it is easy to show that there exists $\bar{\xi}^m \in \overline{V}_m$ such that $u - \mathcal{I}_{\omega_m} u - \bar{\xi}^m = \tilde{u} - \mathcal{I}_{\omega_m} \tilde{u}$ on ω_m. Therefore, $\|u - \mathcal{I}_{\omega_m} u - \bar{\xi}^m\|_{L^2(\omega_m)} \leq C|\omega_m|\, \|u - \mathcal{I}_{\omega_m} u - \bar{\xi}^m\|_{\mathcal{E}(\omega_m)}$, and from the standard interpolation result, we have

$$\|u - \mathcal{I}_{\omega_m} u - \bar{\xi}^m\|_{H(\omega_m)} = \|\tilde{u} - \mathcal{I}_{\omega_m} \tilde{u}\|_{H^1(\omega_m)} \leq Ch|\tilde{u}|_{H^2(\omega_m)}. \tag{5.24}$$

Similarly, there exists $\bar{\xi}_{m+1} \in \overline{V}_{m+1}$ such that

$$\|u - \mathcal{I}_{\omega_{m+1}} u - \bar{\xi}_{m+1}\|_{H(\omega_{m+1})} \leq Ch|\tilde{u}|_{H^2(\omega_{m+1})}.$$

Therefore combining (5.23), (5.24), (5.3) and using the Theorem 4.7 with modifications as mentioned above, we infer that there exists $v \in \mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ such that

$$\|u - v\|_{H(\Omega)} \leq Ch|\tilde{u}|_{H^2(\Omega)}.$$

Therefore, $\|u - u_h\|_{H(\Omega)} = O(h)$, where u_h is the GFEM solution.

We next address the scaled condition number of the stiffness matrix of the GFEM. Since τ_{m+1} is the only element in $\mathcal{K}_{enr}$, we have $\mathbf{A}_{22} = A_{22}^{(m+1)}$.

Let $c = x_m + \beta h$ with $0 < \beta < 1$. A direct computation yields that

$$A_{22}^{(m+1)} = \frac{4}{h}\begin{bmatrix} (4 - 9\beta + 6\beta^2)/3 & -\frac{1}{3} + 2\beta - 2\beta^2 \\ -\frac{1}{3} + 2\beta - 2\beta^2 & (1 - 3\beta + 6\beta^2)/3 \end{bmatrix}.$$

Thus $A_{22}^{(m+1)}$ is associated with vertices x_m and x_{m+1}. We choose the diagonal matrix $D^{(m+1)} = \text{diag}\{\delta_1^{(m+1)}, \delta_2^{(m+1)}\}$ with $\delta_1^{(m+1)} = \delta_2^{(m+1)} = h^{1/2}/2$. Then

$$\hat{A}_{22}^{(m+1)} = D^{(m+1)} A_{22}^{(m+1)} D^{(m+1)} = \begin{bmatrix} (4 - 9\beta + 6\beta^2)/3 & -\frac{1}{3} + 2\beta - 2\beta^2 \\ -\frac{1}{3} + 2\beta - 2\beta^2 & (1 - 3\beta + 6\beta^2)/3 \end{bmatrix}.$$

The diagonal elements of $\hat{A}_{22}^{(m+1)}$ are $O(1)$ for any $0 < \beta < 1$. The eigenvalues of $\hat{A}_{22}^{(m+1)}$ are $\lambda_1 = \frac{5}{6} - 2\beta + 2\beta^2 - T$ and $\lambda_2 = \frac{5}{6} - 2\beta + 2\beta^2 + T$, where $T = \frac{1}{6}\sqrt{13 - 84\beta + 228\beta^2 - 288\beta^3 + 144\beta^4}$ (obtained from *MAPLE*). It can be shown that

$$\frac{1}{6} \le \lambda_1 \le \frac{5}{6} - \frac{\sqrt{13}}{6}, \quad \frac{1}{2} \le \lambda_2 \le \frac{5}{6} + \frac{\sqrt{13}}{6}.$$

Thus, as before, we have

$$\frac{1}{6}\|[D^{(m+1)}]^{-1}\mathbf{x}\|^2 \le \mathbf{x}^T A_{22}^{(m+1)}\mathbf{x} \le \Big(\frac{5}{6} + \frac{\sqrt{13}}{6}\Big)\|[D^{(m+1)}]^{-1}\mathbf{x}\|^2, \quad \forall\, \mathbf{x} \in \mathbb{R}^2,$$

and the Assumption 2 is satisfied with $L_2 = \frac{1}{6}$ and $U_2 = \frac{5}{6} + \frac{\sqrt{13}}{6}$.

We choose $\mathbf{D}_2 = \text{diag}\{d_1, d_2\}$ with $d_i = (\mathbf{A}_{22})_{ii}^{-1/2}$. Clearly, the diagonal elements of $\widehat{\mathbf{A}}_{22} = \mathbf{D}_2\mathbf{A}_{22}\mathbf{D}_2$ are equal to 1. As in the first example (i.e., when $a(x) = a_1(x)$) in Section 5.1, we have $\mathcal{T}_2 = \{x_m, x_{m+1}\}$ and $\mathcal{K}_m = \mathcal{K}_{m+1} = \{\tau_{m+1}\}$. Therefore, $l(m, m+1) = 1$ and $l(m+1, m+1) = 2$. Also $\mathcal{K}_m^* = \mathcal{K}_{m+1}^* = \{\tau_{m+1}\}$ and therefore from (4.18), we have $\Delta_m = [\delta_1^{(m+1)}]^{-2}$ and $\Delta_{m+1} = [\delta_2^{(m+1)}]^{-2}$. Also $x_m, x_{m+1} \in \mathcal{T}_2$ are associated with $(\mathbf{A}_{22})_{y_m y_m}, (\mathbf{A}_{22})_{y_{m+1} y_{m+1}}$, respectively, where $y_m = 1$, and $y_{m+1} = 2$. It is easy to check that

$$\frac{3}{4} \le (\mathbf{A}_{22})_{11}^{-1}\Delta_m,\ (\mathbf{A}_{22})_{22}^{-1}\Delta_{m+1} \le \frac{24}{5}.$$

Thus Assumption 3 is satisfied with $L_3 = 3/4$ and $U_3 = 24/5$. Therefore from Theorem 4.12, we have $\mathfrak{K}(\mathbf{A}) = O(h^{-2})$, where $\mathbf{A}$ is the stiffness matrix, for all $0 < \beta < 1$.

6 Conclusion

The GFEM uses special enrichment functions, based on the available (or extracted) information on the unknown solution of the underlying variational

problem. The use of special enrichment functions gives rise to the excellent convergence properties of the GFEM. In fact, for a given problem, it is possible to choose several classes of enrichment functions such that the GFEM, employing each of these enrichment classes, will yield excellent convergence properties. However, GFEM employing some of these classes of enrichments could be ill-conditioned, i.e., there could be severe loss of accuracy in the computed solution of the linear system associated with the GFEM. The loss of accuracy could be much more than that experienced in a standard FEM. In this paper, we have presented and analyzed a modification of the GFEM – the stable GFEM (SGFEM), which does not have the problem with severe loss of accuracy. SGFEM has all the advantages of the GFEM and is also very robust with respect to the parameters of the enrichments (e.g., the parameter β in Sections 5.1 and 5.3). The loss of accuracy is characterized by the scaled condition number and is expressed through Hypothesis H, which was validated based on various examples.

The abstract framework developed in this paper has been applied to a one-dimensional problem for the clarity of exposition. This framework could also be applied to higher dimensional problems, which will be reported in a forthcoming paper.

Acknowledgement: We thank Professor C. Armando Duarte of the Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, and Professor John E. Osborn of the Department of Mathematics, University of Maryland at College Park, for fruitful discussions on certain aspects of this paper. We also thank Mr. Karl Schulz of PECOS, ICES, University of Texas at Austin, for his help on the use of *superLU* and *MUMPS* on the *Lonestar system*, Texas Advanced Computing Center, and also for various illuminating discussions and interpretation of the results, presented in this paper.

7 Appendix

Validation and Verification (V & V) is a fairly new field and is still in its developing stage ([1, 42, 7, 37]). Suppose a mathematical model of some "Reality" (e.g., a physical, chemical or biological system or process), formulated for a particular goal or purpose, is given. The objective of V & V is to assess whether the predictions based on the computed solution of a mathematical model are reliable enough so that they could be the basis for certain decisions related to the goal.

Validation is the process of building confidence on the mathematical model ([1, 42, 37]). The process is of course is constrained by the cost, available time, and skills, as explicitly underlined in [45]. It is based on a set of properly selected problems and their mathematical models for which experimental data is available. These problems are called validation problems and they are chosen with varying level of complexity; more complex problems are closer to the "Reality". Of course, obtaining the experimental data for the validation problems with increasing complexity is increasingly costly. The prediction based on the

computed solution of these problems is then compared with the experimental data. The assessment of the difference is based on a specified tolerance and a suitably selected metric (could be more than one) relative to the specific goal. If the measure of the difference is larger than the tolerance for any validation problem, the mathematical model is rejected. If none of the validation models are rejected, then one could have confidence that the mathematical model realistically describes the "Reality", with respect to the goal, beyond the scope of the chosen validation problems. The level of confidence will be based on the tolerance as well as the number and the selection of the validation problems. We mention that the set of the validation problems is finite, their selection has a large subjective component, and a philosophical question about the justification of the confidence in the mathematical model could certainly be raised (see [28]).

Numerical algorithms and their properties obtained from the mathematical analysis are always based on various assumptions that are not satisfied when the algorithm is implemented on a computer. For example, infinite precision arithmetic is often assumed while describing a numerical algorithm or stating an inference about the algorithm obtained from the analysis. However, this assumption is always violated by a computer working with finite precision arithmetic. The output from the computer implementation of the algorithm may also depend, for example, on the package in which the algorithm have been implemented, the compiler, the processor, the computing platform with single or multiple processors, among other factors. Consequently, the output may vary even when the same outcome is predicted by the mathematical analysis for two different algorithms. For example, suppose the solution of the linear system $Ax = b$ is sought using algorithms of the form $P_i^T A P_i z = P_i^T b$, $x = P_i z$, where P_i is a permutation matrix for $i = 1, 2$. Both the algorithms should yield that solution x, however, the computed solutions could be different (see Problems 1a,b, below). Thus the implementation of a numerical algorithm in a computer is analogous to a "Reality"; the goal is to obtain a particular quantity of interest for a particular purpose (related to a decision). The mathematical model of this "Reality" is the inference obtained from the mathematical analysis, or other statements based on the inference, about obtaining the quantity of interest from the algorithm. Therefore, the process of validation of the inference has to be performed to have confidence in the inference or a statement based on the inference.

We have briefly formulated the following hypothesis in the Introduction. Let $Ax = b$, $x, b \in \mathbb{R}^n$ be a linear system, where the $n \times n$ matrix A belongs to a class of sparse matrices that include the stiffness matrices associated with FEM, GFEM, or SGFEM. Let $\hat{x}$ be the computed solution of the linear system, obtained from an elimination method, e.g., some variant of Gaussian elimination. Moreover, $\hat{x}$ is computed in finite precision arithmetic with machine precision ϵ. Let $H = DAD$ where D is a diagonal matrix with $D_{ii} = A_{ii}^{-1/2}$; clearly, $H_{ii} = 1$. Recall the scaled condition number $\mathfrak{K}(A)$ of A is given by $\mathfrak{K}(A) := \kappa_2(H)$, where $\kappa_2(H) = \|H\|_2 \|H^{-1}\|_2$ is the condition number of H based on the $\|\cdot\|_2$ vector norm. Also recall $\eta := \|x - \hat{x}\|_2 / \|x\|_2$.

Hypothesis H: For n, not small,

$$\eta \approx Cn^{\beta}\mathfrak{K}(A)\epsilon;\ \beta \approx 0, \tag{7.1}$$

where $\hat{x}$ has been computed in an computing environment satisfying the IEEE standard for floating point arithmetic (with the guard digit), there is no overflow or underflow during the computation of $\hat{x}$, and C, β do not depend on n as well as other factors mentioned before.

The $\approx$ in (7.1) means there exist $0 < \bar{C}_1, \bar{C}_2$ and $0 < \bar{\beta}$ small, such that $\eta = Cn^{\beta}\mathfrak{K}(A)\epsilon$ with $\bar{C}_1 \leq C \leq \bar{C}_2$ and $|\beta| \leq \bar{\beta}$. Also this hypothesis addresses the range N for which not (almost) all digits of accuracy is lost (see Problem 3a).

Hypothesis H is based on certain mathematical inferences (results), which we will discuss later. The validation of (7.1) with respect to the tolerance $\tau = \{\tau_1, \tau_2\}$ means that $\bar{C}_2/\bar{C}_1 \leq \tau_1$ and $\bar{\beta} \leq \tau_2$. Note that τ_2 is primary and should be small for confidence in (7.1), however τ_1 could be allowed to be larger. The set of validation problems consists of stiffness matrices of FEM, GFEM, SGFEM, and other similar matrices, e.g., arising in finite difference method, applied to solve various linear elliptic variational problems of increasing complexity. For confidence in the Hypothesis, we require that (7.1) is not rejected for any of the validation problems relative to the given tolerance τ. We note that it is possible to select a tolerance such that the hypothesis is not rejected, however, the tolerance have to be admissible (e.g., reasonably small) for the decision making process. In our case, the decision will be whether to accept the SGFEM over the standard GFEM. We note that the class of matrices for which the hypothesis will be validated is not precisely defined, similar to a class of complex physical or engineering problem.

We now give a theoretical rationale for (7.1). There is a lot of literature available on the accuracy of the computed solutions of the linear system $Ax = b$. We particularly mention the classic [51] and a modern book [26] with an excellent survey of the theoretical results in the area. Typically, the loss of accuracy in the numerical solution due to round-offs is analyzed by the backward error analysis. This analysis shows that the computed solution is the exact solution of a perturbed linear system, and it provides estimates of the perturbations in terms of the data of the linear system. A bound on the loss of accuracy in the computed solution, measured by η defined before, is then obtained using the perturbation estimates.

It is well known from a standard perturbation argument that for a full matrix A,

$$\eta \leq f(n)\kappa_2(A)\epsilon, \tag{7.2}$$

where ϵ is the machine precision and $f(n)$ depends on the algorithm used to solve $Ax = b$ (see e.g., [24, 27]). In Hypothesis H, we hypothesize that $\kappa_2(A)$ is replaced by $\mathfrak{K}(A)$. We also hypothesize that $\bar{C}_1 n^{\beta} \leq f(n) \leq \bar{C}_3 n^{\beta}$ and $|\beta| \leq \bar{\beta}$, where $\bar{C}_1$, $\bar{C}_2$, and $\bar{\beta}$ are as defined before. It is important to note that in the

mathematical literature, only an upper bound of η is available; in contrast, the Hypothesis H addresses both the upper and lower bounds of η.

Consider the linear system $DADz = Db$, where D is a diagonal matrix with $D_{ii} = 2^{g_i}$ in the rage of the floating point system. Clearly $x = Dz$. We now cite the following old result of F. L. Bauer ([8]):

Theorem 7.1 *Let $\hat{x}$, $\hat{z}$ be the computed solutions of the linear systems $Ax = b$ and $DADz = Db$, respectively, obtained from an elimination method with no pivoting. Furthermore, we assume that there is no overflow or underflow in the computation of $\hat{x}$, $\hat{z}$. Then all the digits of $\hat{x}$ and $\hat{x}_D$ are same, where $\hat{x}_D = D\hat{z}$.*

We note that the result of the above Theorem is not true if the diagonal elements of D are not binary. However in that situation, the quantities $\|x-\hat{x}\|$, $\|x-\hat{x}_D\|$, and $\|\hat{x} - \hat{x}_D\|$ are of the same order.

We next note that it is possible to find a diagonal matrix such that $\kappa_2(DAD) \leq \kappa_2(A)$. For example, for $\mu > 0$, let $A = \begin{bmatrix} 1 & 1 \\ 0 & \mu \end{bmatrix}$. Then $\kappa_2(A) = 1/\mu$ is large for μ small. Let $\mu = \chi 2^{-d}$, where $1/2 < \chi < 2$. Consider $D = \begin{bmatrix} 1 & 0 \\ 0 & 2^{d/2} \end{bmatrix}$. Then $DAD = \begin{bmatrix} 1 & 2^{d/2} \\ 0 & \chi \end{bmatrix}$, and $1 \leq \kappa_2(DAD) < 2$ and consequently, $\kappa_2(DAD) < \kappa_2(A)$ for μ small. Let D^* be the diagonal matrix such that $\kappa_2(D^*AD^*) = \min_D \kappa_2(DAD)$ (minimum over all diagonal matrices D with binary diagonal elements), then from the above theorem and (7.2), we have

$$\eta \leq f(n)\,\kappa_2(D^*AD^*)\epsilon \leq f(n)\,\kappa_2(A)\epsilon.$$

Thus $\kappa_2(D^*AD^*)$ provides more accurate information about η than $\kappa_2(A)$. But in general, it is not easy to find either D^* or $\kappa_2(D^*AD^*)$. In Hypothesis H, we used $\mathfrak{K}(A) = \kappa_2(DAD)$, where D is a diagonal matrix with $D_{ii} = A_{ii}^{-1/2}$ and D_{ii} may not be binary. We note, however, that not using a binary only influences $\bar{C}_1$, $\bar{C}_2$ by factors of $1/2$ and 2 respectively. We also mention that in the literature ([14, 26]), an upper bound of the form (7.2) for η is available with $f(n) = Cn^2$ and $\kappa_2(A)$ replaced by $\mathfrak{K}(A)$ for symmetric positive definite linear systems solved by Cholesky decomposition. In Hypothesis H, we used $f(n) = Cn^\beta$, $\beta \approx 0$, based on our computational experience.

We now consider a set of validation problems, whose exact solution (experimental data) is known. The solution to these problems will be computed on various computers using double precision, i.e., with 16 digits of accuracy.

Problem 1: We consider approximating the solution $u(x) = x$ of the problem $-u''(x) = 0,\ x \in (0,1), u(0) = 0, u(1) = 1$, by the FEM using piecewise linear finite elements.

Problem 1a: We use the FE mesh vertices $x_i = ih$ for $i = 0, 1, \cdots, N$ and $h = 1/N$. The FE solution is same as the exact solution u of the problem. Let the associated linear systems be $A^{(1)}x^{(1)} = b^{(1)}$. The exact solution vector $x^{(1)}$ is known, namely, $x_i^{(1)} = ih$,

$i = 1, 2, \cdots, N$. We will solve the linear system by the standard LU decomposition algorithm for sparse matrices without partial pivoting.

Problem 1b: We use the mesh vertices $x_i = (N-i+1)h$, $i = 0, 1, \cdots, N$. The FE solution is same as the exact solution u and let the associated linear system be $A^{(2)}x^{(2)} = b^{(2)}$; it is known that $x_i^{(2)} = (N - i + 1)h$, $i = 1, \cdots, N$. Note that the elements of $x^{(2)}$ are the permuted elements of $x^{(1)}$ and thus $\|x^{(1)}\|_2 = \|x^{(2)}\|_2$. We will solve the linear system by the same algorithm as Problem 1a.

The computations are performed on a Dell Latitude PC with INTEL CORE(TM)2 CPU, 1.20GHZ.

Problem 2: We approximate the solution $u(x) = 1$ of the problem $-u''(x) = 0,\ x \in (0,1), u(0) = 1, u(1) = 1$, by the piecewise linear FEM based on the mesh vertices as in Problem 1a. Let the associated linear system be $Ax = b$. It is clear that the exact solution is given by $x_i = 1$, $i = 1, 2, \cdots, N$.

Problem 2a: The linear system is solved by a sparse matrix direct solver *superLU* [29] on a single processor.

Problem 2b: The linear system is solved by a sparse matrix direct solver *MUMPS* [2] on a single processor.

Problem 2c: The linear system is solved by *MUMPS*, using parallel computation, on 128 processors.

The computations were performed on the Lonestar system at Texas Advanced Computing Center. Lonestar is a Linux based cluster comprised of 1888 compute nodes connected via high speed quad-data rate infiniband, with each compute node containing two hex-core socket (INTEL Xeon 5680 processors) for an aggregate system size of 22656 cores. Each core runs at a peak of 3.33GHZ.

Problem 3: We consider approximating the solution $u(x) = x^2$ of the problem $-u''(x) = -2,\ x \in (0,1),\ u(0) = 0, u'(1) = 2$, by the GFEM based on $\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$ (see (3.4)). We use $n_i = 1$ and $\varphi_1^{[i]}(x) = x^2$, $i = 0, 1, \cdots, N$. We order the shape functions as $N_0\varphi_1^{[0]}$, N_1, $N_1\varphi_1^{[1]}$, $N_2, \cdots$, N_N, $N_N\varphi_1^{[N]}$ and suppose the associated stiffness matrix is $Ax = b$, where A is of the order $2N + 1$. The GFEM solution is same as the exact solution u. It is easy to see that $x_{2i+1} = 1$, $i = 0, 2, \cdot, N$ and $x_{2i} = 0$, $i = 1, 2, \cdots, N$. The linear system is solved by the same algorithm and on the same platform as in Problem 1a.

Problem 4: We consider approximating the solution of the same problem in Problem 3 by the SGFEM based on $\mathcal{S} = \mathcal{S}_1 + \overline{\mathcal{S}}_2$ (see (4.10)) with $n_i = 1$, $\mathcal{T}_2 = \mathcal{T}$, and $\overline{\varphi}_1^{[i]} = x^2 - \mathcal{I}_{\omega_i}x^2$. We order the shape functions as $N_0\overline{\varphi}_1^{[0]}$, N_1, $N_1\overline{\varphi}_1^{[1]}$, $N_2, \cdots$, N_N, $N_N\overline{\varphi}_1^{[N]}$ and suppose the associated stiffness matrix is $Ax = b$, where A is of the order $2N + 1$. The GFEM solution is same

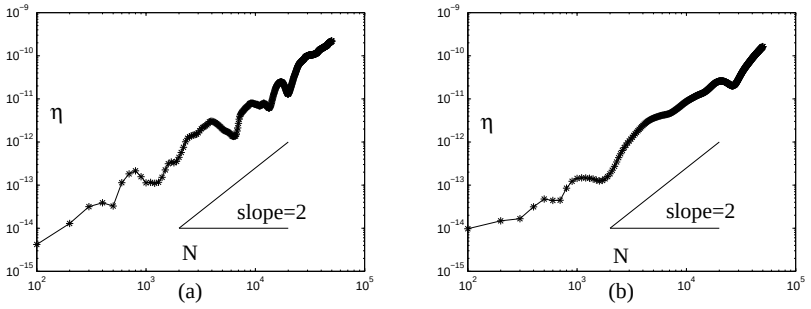

Figure 1: Log-log plots of $\eta^{(k)} = \|x^{(k)} - \hat{x}^{(k)}\|_2/\|x^{(k)}\|_2$ where $\hat{x}^{(k)}$ is the computed solution of $A^{(k)}x^{(k)} = b^{(k)}$, $k = 1, 2$, associated with FEM with vertices $x_i = ih$ and $x_i = (N - i)h$, $i = 0, 1, \cdots, N$, respectively. $\eta^{(1)}$, $\eta^{(2)}$ have been computed and presented in (a) and (b), respectively, for $N = 100, 200, \cdots, 50000$

.

as the exact solution u and it is easy to see that $x_{2i+1} = 1$, $i = 0, 2, \cdot, N$ and $x_{2i} = (ih)^2$, $i = 1, 2, \cdots, N$. The linear system is solved by the same method and on the same platform as in Problem 1a.

We will now validate Hypothesis H based on the validation problems described above. We will consider the tolerance $\tau = (\tau_1, \tau_2)$, with $\tau_1 = 400$ and $\tau_2 = 0$.

Let $\hat{x}^{(1)}$ and $\hat{x}^{(2)}$ be the computed solutions of the linear systems $A^{(1)}x^{(1)} = b^{(1)}$ and $A^{(2)}x^{(2)} = b^{(2)}$ of **Problem 1a** and **Problem 1b**, respectively. It can be shown that for large N, $\mathfrak{K}(A^{(1)}) = \mathfrak{K}(A^{(2)}) \approx 0.4\,N^2$. We have computed and presented the log-log plots of the relative errors $\eta^{(k)} = \|x^{(k)} - \hat{x}^{(k)}\|_2/\|x^{(k)}\|_2$, $k = 1, 2$, with respect to $N = 100, 200, \cdots, 50000$ in Figure 1. We have observed that $\bar{C}_1^{(k)}[0.4N^2] \leq \eta^{(k)} \leq \bar{C}_2^{(k)}[0.4N^2]$ for $k = 1, 2$ with $\bar{C}_2/\bar{C}_1 \leq 120 < \tau_1$ (note $\tau_2 = 0$). Thus we do not reject Hypothesis H. Note that we did not reject the hypothesis based only on the subset of meshes with the values of N, mentioned above. Moreover, it is interesting to note that the plots of $\eta^{(1)}$ and $\eta^{(2)}$ are quite different. Thus the computed solution is affected by changing the order of the FE mesh vertices, in spite of the fact that $\|x^{(1)}\|_2 = \|x^{(2)}\|_2$.

In **Problem 2**, we solve the linear system $Ax = b$ using two different software *superLU* and *MUMPS*; we also implement *MUMPS* on multiple processors. Let $\hat{x}^{(a)}$, $\hat{x}^{(b)}$, $\hat{x}^{(c)}$ be the computed solutions of Problems 2a, 2b, and 2c, respectively. These solutions were computed for $10 \leq N \leq 10^7$, with 90 values of N in the range $[10, 10^2)$, with 400 values of N in the range $[10^2, 10^3)$, and 360 values of N in the range $[10^i, 10^{i+1})$, $i = 3, 4, 5, 6$, and with $N = 10^7$. We presented the log-log plots of $\eta^{(k)} = \|x - \hat{x}^{(k)}\|_2/\|x\|_2$, $k = a, b$, for the values of N given before, in Figures 2a and 2b respectively. We observed that for $N \geq 100$, $\bar{C}_2/\bar{C}_1 \leq 200 < \tau_1$ for both the problems. Thus we do not reject the Hypothesis H for $N \geq 100$. Note that for $N \leq 100$, Figures 2a and b suggest that $\beta \approx -1$. It is also clear from Figure 2b that the implementation of the

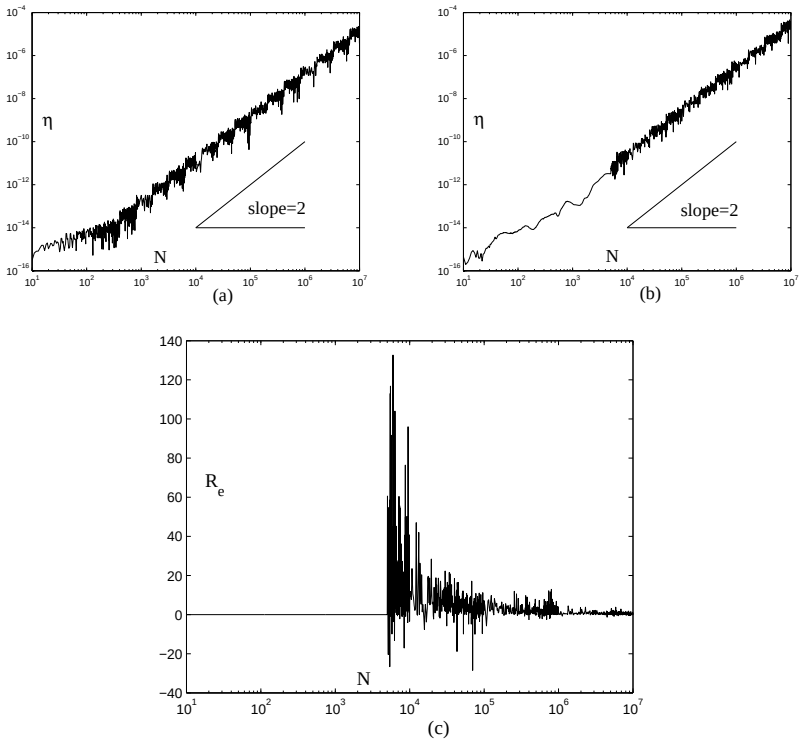

Figure 2: (a) Log-log plot of $\eta^{(a)} = \|x - \hat{x}^{(a)}\|_2/\|x\|_2$ with respect to N, where $\hat{x}^{(a)}$ is the computed solution of Problem 2a (using *superLU*). (b) Log-log plot of $\eta^{(b)} = \|x - \hat{x}^{(b)}\|_2/\|x\|_2$ with respect to N, where $\hat{x}^{(b)}$ is the computed solution of Problem 2b (using *MUMPS*). (c) Semi-log plot of $100 * (\eta^{(c)} - \eta^{(b)})/\eta^{(b)}$ with respect to N, where $\eta^{(c)} = \|x - \hat{x}^{(c)}\|_2/\|x\|_2$ and $\hat{x}^{(c)}$ is the computed solution of Problem 1c (using *MUMPS* with 128 processors). Proportionally distributed 1931 values of N in the interval $[10, 10^7]$ are used in all the figures.

algorithm in *MUMPS* changes drastically for $N > 5 \times 10^3$; this is not the case with *superLU*, as seen in Figure 2a. Thus the computed solution depends on the software package, as mentioned before. For Problem 2c, we did not display the log-log plot of $\eta^{(c)} = \|x - \hat{x}^{(c)}\|_2/\|x\|_2$ as it would be very similar to the plot of $\eta^{(b)}$ in Figure 2b. However, we computed $R_e \equiv 100(\eta^{(c)} - \eta^{(b)})/\eta^{(b)}$ – the "signed relative difference percent" — and presented the semi-log plot of R_e in Figure 2c for the same values of N, given before. For $N \leq 5 \times 10^3$, we see that $R_e \approx 0$ and values of R_e starts to oscillate for $N > 5 \times 10^3$. This indicates that the implementation in MUMPS changes drastically. Figure 2c also suggests that $\eta^{(c)}$ is larger than $\eta^{(b)}$ for most values on N, and $\eta^{(c)}$ gets closer to $\eta^{(b)}$ as N increases.

Let $\hat{x}$ be the solution of the linear system $Ax = b$ of **Problem 3**. We have $\mathfrak{K}(A) = O(N^4)$ (see Section 3.1). The log-log plot of $\eta = \|x - \hat{x}\|_2/\|x\|_2$

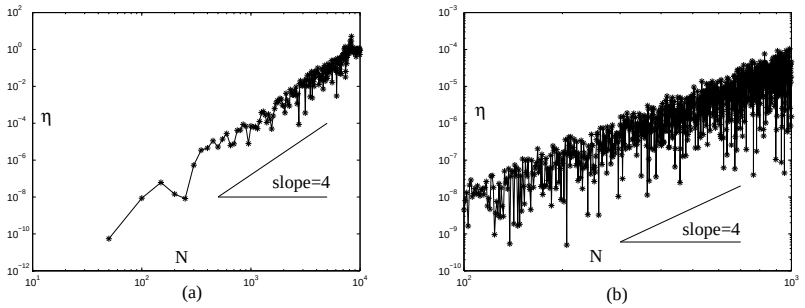

Figure 3: Plots of $\eta = \|x - \hat{x}\|_2/\|x\|_2$ where $\hat{x}$ is the computed solution of the linear system $Ax = b$ of Problem 3, associated with the GFEM with vertices $x_i = ih$, $i = 0, 1, \cdots, N$. In (a), we used $N = 50, 100, 150, \cdots, 10000$, and in (b), we used every value of N in the interval $[100, 1000]$ to show the detail.

with respect to $N = 50, 100, 150, \cdots, 10000$ have been presented in Figure 3a. In Figure 3b, we show the details in the range $100 \leq N \leq 1000$, where we have presented the log-log plot of η for every value of N in this range. Based on both these data (i.e., the values of η for every value of N in the range $100 \leq N \leq 1000$ and for $N = 1050, 1100, 1150, \cdots, 10000$),we have observed that $\bar{C}_1 N^4 \leq \eta \leq \bar{C}_2 N^4$ with $\bar{C}_2/\bar{C}_1 \leq 340 < \tau_1$ (note $\tau_2 = 0$). Thus we do not reject the Hypothesis H, again based on the subset of meshes with the values of N mentioned above. It is important to note that in Problem 3a, all the digits of accuracy were lost for $N \geq 9000$, and thus the Hypothesis H does not address the value of $N \geq 9000$. We also computed η for every value of N in the range $9000 \leq N \leq 11000$; η was of the order 1 and oscillated around 1.

Let $\hat{x}$ be the computed solution of the linear system $Ax = b$ of **Problem 4**. We have shown in this paper that $\mathfrak{K}(A) = O(h^2)$. We have presented the log-log plot of $\eta = \|x - \hat{x}\|_2/\|x\|_2$, with respect to $N = 50, 100, 150, \cdots, 10000$ in Figure 4a, and for every value of N in the range $100 \leq N \leq 1000$ in Figure 4b. Based on both these data (i.e., the values of η for every value of N in the range $100 \leq N \leq 1000$ and $N = 1050, 1100, 1150, \cdots, 10000$), we observed that $\bar{C}_1 N^2 \leq \eta \leq \bar{C}_2 N^2$ with $\bar{C}_2/\bar{C}_1 \leq 240 < \tau_1$ (note $\tau_2 = 0$). Thus we do not reject the Hypothesis H (based on meshes with these values of N).

Thus we did not reject the Hypothesis H for any validation problems with respect to the tolerance $\tau_1 = 400$ and $\tau_2 = 0$. But we would reject the Hypothesis H if we choose $\tau_1 = 300$, since $\bar{C}_2/\bar{C}_1 \leq 340 \not\leq \tau_1$ in Problem 3. However, if the values of η for every value of N in the range $[100, 1000]$ were not available (see Figure 3b), then we will have $\bar{C}_2/\bar{C}_1 \leq 250 < \tau_1$, and we thus we would not reject Hypothesis H. Hence validation depends on the values of N, i.e., on the number of validation problems considered, since each value of N (in each of Problems 1, 2, 3, and 4) constitutes a separate validation problem. But as mentioned before, the choice of the tolerance depends on the type of decision related to the goal. For example in this paper, we have to decide whether to

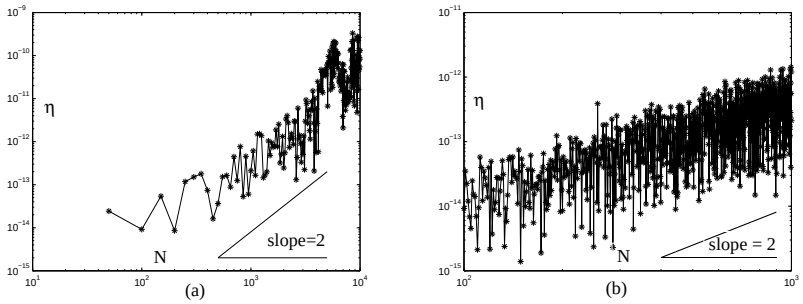

Figure 4: Plots of $\eta = \|x - \hat{x}\|_2/\|x\|_2$ where $\hat{x}$ is the computed solution of the linear system $Ax = b$ of Problem 4, associated with the SGFEM with vertices $x_i = ih$, $i = 0, 1, \cdots, N$. In (a), we used $N = 50, 100, 150, \cdots, 10000$, and in (b), we used every value of N in the interval $[100, 1000]$ to show the detail.

accept SGFEM over the standard GFEM. In this case, we may allow τ_1 to be bigger; in fact if $\tau_1 = 500$, we still accept SGFEM over GFEM since the value of η for GFEM will be much larger than the η of SGFEM for large N.

We summarize by stating that

> (a) we have confidence in Hypothesis H, based on the chosen validation problems (Problems 1–4). We underline that we have also considered other 2- and 3-dimensional validation problems for the Hypothesis H, which we do not present in this paper. We will present a more substantial validation of Hypothesis H in a future publication.
>
> (b) Because of our confidence in Hypothesis H, we prefer the use of SGFEM over GFEM, since linear system of SGFEM is less prone to the loss of accuracy than the linear system of the GFEM, when solved using an elimination method.

Remark 7.2 As mentioned before, all the computations presented here were performed with 10^{-16} accuracy. However, all the figures, presented above, indicate that that the apparent accuracy is about 10^{-18}. This is likely the effect of various cancelations. ∎

References

[1] American Society of Mechanical Engineers, New York. *ASME guide for Verification and Validation in Computational Solid Mechanics*, 2006. V&V 10.

[2] P. R. Amestoy, I. S. Duff, J. Koster, and J. Y. L'Excellent. A fully asynchronous multifrontal solver using distributed dynamic scheduling. *SIMAX*, 23:15–41, 2001.

[3] I. Babuśka and R. Lipton. Optimal local approximation spaces for generalized finite element methods with application to multiscale problems. Technical Report 10-12, ICES, University of Texas at Austin, 2010.

[4] I. Babuška, U. Banerjee, and J. Osborn. Generalized finite element methods: Main ideas, results, and perspective. *International Journal of Computational Methods*, 1(1):1–37, 2004.

[5] I. Babuška, G. Caloz, and J. Osborn. Special finite element methods for a class of second order elliptic problems with rough coefficients. *SIAM J. Numer. Anal.*, 31:945–981, 1994.

[6] I. Babuška and J. M. Melenk. The partition of unity finite element method. *Int. J. Numer. Meth. Engng.*, 40:727–758, 1997.

[7] I. Babuška and J. T. Oden. Verification and Validation in computational engineering and science: basic concepts. *Comput. Methods Appl. Mech. Engrg.*, 193:4057–4066, 2004.

[8] F. L. Bauer. Optimal scaling of matrices and the importance of the minimal condition. In C. M. Popplewell, editor, *Information Processing 62*, IFIP Congress 1962, pages 198–201, Amsterdam, 1963. North-Holland.

[9] T. Belytschko and T. Black. Elastic crack growth in finite elements with minimal remeshing. *Int. J. Numer. Meth. Engng.*, 45:601–620, 1999.

[10] S. E. Benzley. Representation of singularities with isoparametric finite elements. *Int. J. Numer. Meth. Engng.*, 8:537–545, 1974.

[11] H. Blum and M. Dobrowoski. On finite element methods for elliptic equations on domains with corners. *Computing*, 28:53–63, 1982.

[12] E. Byskov. The calculation of stress intensity factors using finite element with cracked element. *Int. J. Fract. Mech.*, 6:159–167, 1970.

[13] C. Daux, N. Moes, J. Dolbow, N. Sukumar, and T. Belytschko. Arbitrary branched and intersecting cracks with extended finite element method. *Int. J. Numer. Meth. Engng.*, 48:1741–1760, 2000.

[14] J. Demmel. On floating point error in cholesky. Technical Report CS-89-87, Dept. of Computer Science, Univ. of Tennessee, 1989.

[15] J. Dolbow, N. Moës, and T. Belytschko. Modeling fracture in Mindlin-Reissner plates with the extended finite element method. *J. Solids Struct.*, 37:7161–7183, 2000.

[16] J. E. Dolbow. *An Extended Finite Element Method with Discontinuous Enrichment for Applied Mechanics*. PhD thesis, Northwestern University, 1999.

[17] C. A. Duarte and J. T. Oden. An h-p adaptive method using clouds. *Comput. Methods Appl. Mech. Engrg.*, 139:237–262, 1996.

[18] C. A. Duarte and J. T. Oden. H-p Clouds – An *h-p* Meshless Method. *Numer. Methods Partial Differential Equations*, 12:673–705, 1996.

[19] P. Esser, J. Grande, and A. Reusken. An extended finite element method applied to levitated droplet problems. *Int. J. Numer. Meth. Engng*, 84:757–773, 2010.

[20] M. Farsad, F. J. Vernerey, and H. S. Park. An extended finite element/level set method to study surface effects on the mechanical behavior and properties of nanomaterials. *Int. J. Numer. Meth. Engng*, 84:1466–1489, 2010.

[21] G. Fix, S. Gulati, and G. I. Wakoff. On the use of singular functions with finite element approximations. *J. Comp. Phys.*, 13:209–228, 1973.

[22] T.-P. Fries. A corrected XFEM approximation without problems in blending elements. *Int. J. Numer. Meth. Engng.*, 75:503–532, 2008.

[23] T-P Fries and T. Belytschko. The extended/generalized finite element method: An overview of the method and its applications. *Int. J. Numer. Meth. Engng.*, 84:253–304, 2010.

[24] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, USA, 1996.

[25] M. Griebel and M. A. Schweitzer. A Particle-Partition of Unity method – Part VI: Adaptivity. In M. Griebel and M. A. Schweitzer, editors, *Meshfree Methods for Partial Differential Equations III*, Lecture Notes on Computer Science and Engineering, Vol. 26, pages 121–148. Springer, 2006.

[26] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, 2002.

[27] D. Kincaid and W. Cheney. *Numerical Analysis; Mathematics of Scientific Computing*. American Mathematical Society, 2002.

[28] G. B. Kleindorfer, L. O'Neill, and R. Ganeshan. Validation in similation: various positions in the philosophy of science. *Management Science*, 44:1087–1099, 1998.

[29] X. S. Li and J. W. Demmel. SuperLU-DIST: A scalable distributed-memory sparse direct solver for unsymmetric linear systems. *ACM Trans, Mathematical Software*, 29:110–140, 2003.

[30] C. Lu and B. Shanker. Generalized finite element method for vector electromafnetic problems. *IEEE Transactions on Antennas and Propagation*, 55:1369–1381, 2007.

[31] A. M. Matache, I. Babuśka, and C. Schwab. Generalized p-FEM in homogenization. *Numer. Math.*, 86, 2000.

[32] J. M. Melenk. *On Generalized Finite Element Methods*. PhD thesis, University of Maryland, 1995.

[33] J. M. Melenk and I. Babuška. The partition of unity finite element method: Basic theory and application. *Comput. Methods Appl. Mech. Engrg.*, 139:289–314, 1996.

[34] A. Menk and S. P. A. Bordas. A robust preconditioning technique for the extended finite element method. *Int. J. Meth. Engng.*, 85:1609–1632, 2011.

[35] N. Moes, J. Dolbow, and T. Belytschko. A finite element method for crack without remeshing. *Int. J. Numer. Meth. Engrg.*, 46:131–150, 1999.

[36] A. Nouy and A Clément. eXtended stochastic finite element method for the numerical simulation of heterogegeous materials with random material interfaces. *Int. J. Numer. Meth. Engng*, 83:1312–1344, 2010.

[37] W. L. Oberkampf and Ch. J. Roy. *Verification and Validation in Scientific Computing*. Cambridge University Press, New York, 2010.

[38] J. T. Oden and C. A. M. Duarte. Clouds, Cracks and FEMs. In B. Daya Reddy, editor, *Recent Developments in Computational and Applied Mechanics*, 1997.

[39] J. T. Oden, C. A. M. Duarte, and O. C. Zienkiewicz. A new cloud-based hp finite element method. *Comput. Methods Appl. Mech. Engrg.*, 153:117–126, 1998.

[40] P. O'Hara, C. A. Duarte, and T. Eason. Generalized finite element analysis for three-dimensional problems exhibiting sharp thermal gradients. *Comput. Methods Appl. Mech. Engrg.*, 198:1857–1871, 2009.

[41] A. K. Rao, I. S. Raju, and A. V. K. Murthy. A powerful hybrid method in finite element analysis. *Int. J. Numer. Meth. Engng,*, 3:389–403, 1971.

[42] P. J. Roache. *Fundamentals of Verification and Validation*. Hermosa Publisher, Albuquerque, NM, 2009.

[43] M. A. Schweitzer. *A Parallel Multilevel Partition of Unity Method for Elliptic Partial Differential Equations*. Springer, 2003. Lecture Notes in Computational Science, vol. 29.

[44] A. Simone, C. A. Duarte, and E. Van der Giessen. A generalized finite element method for polycrystals with discontinuous grain boundaries. *Int. J. Numer. Meth. Engng*, 67:1122–1145, 2006.

[45] Simulation Interoperability Standards Organization, Orlando, FL. *Guide for generic methodology for Verificatin and Validation (V&V) and acceptance of models, simulations, and data*, 2007.

[46] G. Strang and G. Fix. *An Analysis of the Finite Element Method*. Wellesley-Cambridge, 2008. 2nd. edition.

[47] T. Strouboulis, I. Babuška, and K. Copps. The design and analysis of the generalized finite element method. *Comput. Methods Appl. Mech. Engrg.*, 181:43–69, 2000.

[48] T. Strouboulis, K. Copps, and I. Babuška. The generalized finite element method: an example of its implementation and illustration of its performance. *Int. J. Numer. Meth. Engng.*, 47:1401–1417, 2000.

[49] T. Strouboulis, K. Copps, and I. Babuška. The generalized finite element method. *Comput. Methods Appl. Mech. Engrg.*, 190:4081–4193, 2001.

[50] N. Sukumar, N. Moes, B. Moran, and T. Belytschko. Extended finite element method for for three dimensional crack modelling. *Int. J. Numer. Meth. Engrg.*, 48(11):1549–1570, 2000.

[51] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. The Oxford University Press, 1988.