Condition number estimation and reduction for the finite cell method

a study of, and solution for, conditioning problems regarding the finite cell method

de Prenter, F.

Award date:
2015

Condition number estimation and reduction for the Finite Cell Method

A study of, and solution for,
conditioning problems regarding
the Finite Cell Method

Frits de Prenter
0718450

Eindhoven University of Technology
Department of Mechanical Engineering
&
Department of Applied Mathematics

10 March 2015

A thesis presented for the degree of Master of Science

Supervisors:	**Committee:**
Clemens Verhoosel	Clemens Verhoosel
Harald van Brummelen	Harald van Brummelen
Gertjan van Zwieten	Sorin Pop

Contents

1 Introduction

Over the last decades, numerical solution methods for Partial Differential Equations (PDE's) have become essential tools in the understanding, design and optimisation of physical processes. In particular the Finite Element Method (FEM) plays an important role in present-day engineering and has proven its effectiveness on a wide range of applications. Many variations and extensions of this method have been introduced to improve its efficiency and comprehensiveness, a well-known recent example of which is Isogeometric Analysis.

Isogeometric Analysis (IGA, [1]) replaces the function space that is used in the Finite Element Method by a spline-based function space as used in Computer Aided Design (CAD). These spline-based function spaces have better approximation properties and avoid computationally expensive discretisation procedures for CAD-generated structures. The Finite Cell Method can further improve the flexibility of IGA applied to CAD-generated structures by weakly implementing essential boundary conditions on trimmed objects and coupled domains [2]. By advanced integration procedures applied in the Finite Cell Method, it enables the application of IGA on geometrically and topologically complex structures without laborious meshing procedures [3].

The Finite Cell Method (FCM, *e.g.* [2–17]) was introduced in [4] as an extension of standard Galerkin methods, in which the grid does not need to match, but simply needs to overlap the problem's domain by weakly implementing essential boundary conditions using Nitsche's method [18]. The complexity of the problem's domain is not captured by the mesh, but by the integration scheme, which splits the discretised domain into a computational part over which integration is performed and a fictitious part. FCM shows advantages over standard Galerkin methods for problems which are hard, time-consuming or even impossible to mesh, or would need frequent remeshing in the traditional way.

The Finite Cell method has been found to be prone to conditioning problems, which impedes solving the resulting linear system. Different modifications that focus on reducing the condition number have been proposed, such as basis function elimination [5] and virtual stiffness [2, 7–15]. These modifications limit the condition number, but generally not to a satisfactory level. Without a proper conditioning technique, it may be inevitable to adjust the FCM grid and apply a direct solver in order to solve the linear system resulting from FCM formulations [2, 8].

The primary research objective of this work is to analyse the origins of the conditioning problems associated with FCM, and to study the possibilities for ameliorating these problems by means of diagonal pre- and post-conditioning. An important novel contribution of this work is the derivation of an explicit relation between the condition number and the smallest basis function support for uniform grids. This relation reveals a strong dependence of the condition number on the order of the employed spline discretisation, which stipulates the need to develop a strategy to improve the conditioning of FCM.

Motivated by the aforementioned relation, in this work this improvement is established by means of basis function scaling, which is shown to be equivalent to diagonal-like pre- and post-conditioning of the system. The usage of basis function scaling in combination with Nitsche's method for imposing essential boundary conditions is studied in detail in this work. It is demonstrated that the proposed scaling strategy drastically improves the condition number for a series of problems. Convergence problems of iterative solvers related to the poor conditioning of FCM are rigorously resolved by means of the proposed scaling strategy.

Section 2 presents the variational formulation of the Finite Cell Method and analyses the conditioning in detail. An explicit relation between the condition number and the smallest basis function support for uniform grids is derived and experimentally verified. Section 3 describes the strategy to improve the conditioning by basis function scaling and tests this strategy on various examples. In Section 4 the process of solving linear systems iteratively is described in detail and the effect of improving the conditioning is demonstrated. Section 5 presents the results of several numerical examples to further study the effect of basis function scaling. Section 6 contains some concluding remarks and recommendations concerning future research topics.

2 The Finite Cell Method: An analysis of system conditioning

In this section the variational formulation of the Finite Cell Method (FCM) is presented, together with an analysis of the solvability and conditioning of the resulting system.

2.1 Variational formulation

Consider the second-order, elliptic PDE over a domain $\Omega \subset \mathbb{R}^d$

$$\begin{cases} -\mathrm{div}(A\nabla u) = f & \text{in } \Omega, \\ nA\nabla u = g & \text{on } \Gamma_n, \\ u = \overline{u} & \text{on } \Gamma_e, \end{cases} \tag{2.1}$$

with A bounded, symmetric and strongly positive. Γ_n and Γ_e represent a natural and essential boundary respectively, furthermore $\Gamma_n \cap \Gamma_e = \emptyset$, $\Gamma_n \cup \Gamma_e = \partial\Omega$ and $\Gamma_e \neq \emptyset$. Problem (2.1) for example represents either linear elasticity with A a stiffness tensor and u and f vector-valued deformations and body forces, or Laplace's problem with A the identity matrix and u and f scalar fields. Boundary-fitted Galerkin methods such as the Finite Element Method use a grid that matches Ω and can therefore impose the essential boundary condition on Γ_e in a strong manner (*i.e.* encoded into the function space) and use the weak form

$$\begin{aligned} &\text{find } u \in H^1_e(\Omega) \text{ such that:} \\ &a(v,u) = l_a(v) \quad \forall v \in H^1_0(\Omega), \end{aligned} \tag{2.2}$$

with

$$\begin{aligned} a(v,u) &= \int_\Omega \nabla v A \nabla u \mathrm{d}\Omega, \\ l_a(v) &= \int_\Omega v f \mathrm{d}\Omega + \int_{\Gamma_n} v g \mathrm{d}\Gamma, \end{aligned} \tag{2.3}$$

and where $H^1_e(\Omega) = \{u \in H^1(\Omega) | u = \overline{u} \text{ on } \Gamma_e\}$ and $H^1_0(\Omega) = \{u \in H^1(\Omega) | u = 0 \text{ on } \Gamma_e\}$. $a(\cdot,\cdot)$ is bilinear, symmetric, bounded and coercive on $H^1_0(\Omega)$ and therefore forms an inner product that induces the energy norm $\|\cdot\|_a = \sqrt{a(\cdot,\cdot)}$ that is equivalent with the H^1_0-seminorm on Ω.

The Finite Cell Method is an unfitted Galerkin method, and uses a grid that overlaps Ω such as *e.g.* a discretisation of $\Omega \cup \Omega_{\text{fict}}$ as shown in Figure 2.1. In the FCM literature (*e.g.* [2–17]) generally a physical domain (Ω or Ω_{phys}), a fictitious extension (Ω_{fict}) and a complete or embedding domain (Ω, Ω_{C} or Ω_{e}) -which is the union of the physical and fictitious domain- is defined. In this work the fictitious domain will be completely omitted and therefore the physical domain is simply referred to by Ω. As essential boundary conditions cannot be imposed in a strong manner on unmatching grids, the boundary conditions

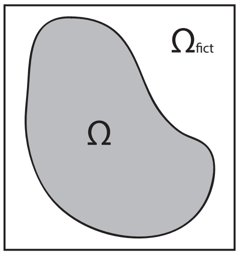


Figure 2.1: A geometrically complex domain Ω that is embedded in a rectilinear domain $\Omega \cup \Omega_{\text{fict}}$.

on Γ_e are weakly imposed using a Nitsche type boundary condition [18], giving the weak form

$$\begin{aligned}&\text{find } u \in H^1(\Omega) \text{ such that:}\\ &a(v,u) + b(v,u) = l_a(v) + l_b(v) \quad \forall v \in H^1(\Omega),\end{aligned} \tag{2.4}$$

with

$$\begin{aligned}b(v,u) &= -\int_{\Gamma_e} (vnA\nabla u + unA\nabla v)\,\mathrm{d}\Gamma,\\ l_b(v) &= -\int_{\Gamma_e} \overline{u}nA\nabla v \mathrm{d}\Gamma.\end{aligned} \tag{2.5}$$

$a(\cdot,\cdot)$ is not coercive on $H^1(\Omega)$ however, and $a(\cdot,\cdot)+b(\cdot,\cdot)$ is not even bounded or positive (and certainly not coercive) on $H^1(\Omega)$. In finite-dimensional subspaces of $H^1(\Omega)$, boundedness is not an issue however, and $a(\cdot,\cdot)+b(\cdot,\cdot)$ can be made coercive in a finite-dimensional subspace $\mathcal{V}^h(\Omega) \subset H^1(\Omega)$ by adding a penalty, $c(v,u)$, which maintains the consistency. This yields

$$a(v,u) + b(v,u) + c(v,u) = l_a(v) + l_b(v) + l_c(v) \quad \forall v \in H^1(\Omega), \tag{2.6}$$

with

$$\begin{aligned}c(v,u) &= \int_{\Gamma_e} \beta vu \mathrm{d}\Gamma,\\ l_c(v) &= \int_{\Gamma_e} \beta v\overline{u} \mathrm{d}\Gamma,\end{aligned} \tag{2.7}$$

and where the penalty factor β is either a global or an element-wise positive constant. A global approach is presented here, but a local (element-wise) approach can be derived similarly. The penalty operator $c(\cdot,\cdot)$ gets its name from the penalty method [6], where it is used to apply essential boundary conditions. Introducing a global or element-wise, computable constant C and postulating the condition that

$$\|nA\nabla v\|^2_{L^2(\Gamma_e)} \leq Ca(v,v) \quad \forall v \in \mathcal{V}^h(\Omega), \tag{2.8}$$

the combined operator $a(\cdot,\cdot)+b(\cdot,\cdot)+c(\cdot,\cdot)$ is coercive on $\mathcal{V}^h(\Omega)$ for $\beta > C$ as

$$\begin{aligned}
|b(v,v)| \leq & 2\|nA\nabla v\|_{L^2(\Gamma_e)}\|v\|_{L^2(\Gamma_e)} && \text{Cauchy-Schwarz} \\
\leq & \frac{1}{\varepsilon}\|nA\nabla v\|^2_{L^2(\Gamma_e)} + \varepsilon\|v\|^2_{L^2(\Gamma_e)} && \text{Peter-Paul} \\
\leq & \frac{C}{\varepsilon}a(v,v) + \varepsilon\|v\|^2_{L^2(\Gamma_e)} && (2.8) \\
\leq & a(v,v) + c(v,v) \quad \forall v \in \mathcal{V}^h(\Omega) && C < \varepsilon < \beta,
\end{aligned} \tag{2.9}$$

and therefore

$$\begin{aligned}
& a(v,v)+b(v,v)+c(v,v) \geq a(v,v) - |b(v,v)| + c(v,v) \\
\geq & \frac{\varepsilon - C}{\varepsilon}a(v,v) + (\beta-\varepsilon)\|v\|^2_{L^2(\Gamma_e)} \\
\geq & \delta\|v\|^2_{H^1(\Omega)} \quad \forall v \in \mathcal{V}^h(\Omega) && \text{Poincaré},
\end{aligned} \tag{2.10}$$

for some $\delta > 0$. The last inequality in (2.10) is a specific form of the Poincaré inequality as can be found in lemma B.63 in [19]. The minimal C for which (2.8) holds coincides with the largest eigenvalue λ of the generalised eigenvalue problem [5]

$$\mathbf{E}\underline{v} = \lambda\mathbf{A}\underline{v}, \tag{2.11}$$

with

$$\begin{aligned}
\mathbf{E} &= \int_{\Gamma_e} \left(nA\nabla\underline{N}\right)\left(nA\nabla\underline{N}^T\right) \mathrm{d}\Gamma, \\
\mathbf{A} &= \int_{\Omega} \nabla\underline{N}A\nabla\underline{N}^T \mathrm{d}\Omega,
\end{aligned} \tag{2.12}$$

and where $\underline{N}$ is a vector containing a basis of $\mathcal{V}^h(\Omega)$. As shown by [18], C scales with $1/\widetilde{h}$, where $\widetilde{h}$ is either the typical length scale in the element (element-wise constant) or the typical length scale in the smallest element (global constant). All examples presented here use $\beta = 2C$ as proposed by [5]. This choice does not affect the strategy to improve the conditioning of the system, however, and allows optimising the stability parameter as in [8]. For notational convenience, the combined operators are denoted as

$$\begin{aligned}
k(v,u) &= a(v,u) + b(v,u) + c(v,u) \\
&= \int_{\Omega} \nabla vA\nabla u\mathrm{d}\Omega + \int_{\Gamma_e} \big(\beta vu - (vnA\nabla u + unA\nabla v)\big)\mathrm{d}\Gamma, \\
l(v) &= l_a(v) + l_b(v) + l_c(v) \\
&= \int_{\Omega} vf\mathrm{d}\Omega + \int_{\Gamma_e} \left(\beta v\overline{u} - \overline{u}nA\nabla v\mathrm{d}\Gamma\right) + \int_{\Gamma_n} vg\mathrm{d}\Gamma.
\end{aligned} \tag{2.13}$$

As $k(\cdot,\cdot)$ is symmetric, bilinear, bounded and coercive on $\mathcal{V}^h(\Omega)$ (w.r.t. the H^1-norm) it forms an inner product that induces the FCM-norm $\|v\|_k = \sqrt{k(v,v)}$ that is equivalent with the H^1-norm on Ω.

2.2 Condition numbers

In order to define the condition number, first the Euclidean vector-norm and induced matrix-norm are defined as

$$\|\underline{x}\|_2 = \sqrt{\underline{x}^T\underline{x}}, \quad \|\mathbf{A}\|_2 = \max_{\underline{x}} \frac{\|\mathbf{A}\underline{x}\|_2}{\|\underline{x}\|_2}. \tag{2.14}$$

When solving a system of the form $\mathbf{A}\underline{x} = \underline{b}$, the condition number

$$\kappa_2(\mathbf{A}) = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2, \tag{2.15}$$

is of much interest. To begin with, the convergence speed of most iterative methods to find $\underline{x}$ is dependent on $\kappa_2(\mathbf{A})$ [20]. For example, the Conjugate Gradient method has a convergence bound given by [20]

$$\|\underline{x} - \underline{x}_i\|_{\mathbf{A}} \leq 2\left(\frac{\sqrt{\kappa_2(\mathbf{A})} - 1}{\sqrt{\kappa_2(\mathbf{A})} + 1}\right)^i \|\underline{x} - \underline{x}_0\|_{\mathbf{A}}, \tag{2.16}$$

where $\underline{x}_i$ is the approximation after i iterations and $\|\cdot\|_{\mathbf{A}}$ denotes the energy norm $\|\underline{x}\|^2_{\mathbf{A}} = \underline{x}^T\mathbf{A}\underline{x}$. Furthermore the condition number determines whether the residual is a good estimate for the error as

$$\frac{\|\underline{b} - \mathbf{A}\underline{x}_i\|_2}{\|\mathbf{A}\|_2} \leq \|\underline{x} - \underline{x}_i\|_2 \leq \kappa_2(\mathbf{A})\frac{\|\underline{b} - \mathbf{A}\underline{x}_i\|_2}{\|\mathbf{A}\|_2}. \tag{2.17}$$

Moreover, the condition number is especially important when using finite precision arithmetics due to error propagation. Assume $\mathbf{A}\underline{x} = \underline{b}$ and $\mathbf{A}\left(\underline{x} + \tilde{\underline{x}}\right) = \underline{b} + \tilde{\underline{b}}$ with $\tilde{\underline{b}}$ a perturbation of $\underline{b}$ and $\tilde{\underline{x}}$ the resulting perturbation of $\underline{x}$. Then the quotient of the relative errors is bounded by the condition number

$$\begin{aligned}\frac{\|\tilde{\underline{x}}\|_2/\|\underline{x}\|_2}{\|\tilde{\underline{b}}\|_2/\|\underline{b}\|_2} = \frac{\|\tilde{\underline{x}}\|_2\|\underline{b}\|_2}{\|\tilde{\underline{b}}\|_2\|\underline{x}\|_2} = \frac{\|\mathbf{A}^{-1}\tilde{\underline{b}}\|_2\|\mathbf{A}\underline{x}\|_2}{\|\tilde{\underline{b}}\|_2\|\underline{x}\|_2} &\leq \frac{\|\mathbf{A}^{-1}\|_2\|\tilde{\underline{b}}\|_2\|\mathbf{A}\|_2\|\underline{x}\|_2}{\|\tilde{\underline{b}}\|_2\|\underline{x}\|_2} \\ &= \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 = \kappa_2(\mathbf{A}),\end{aligned} \tag{2.18}$$

where the inequality follows directly from definition (2.14). Similarly, when there is a perturbation of $\mathbf{A}$, *i.e.* $\left(\mathbf{A} + \tilde{\mathbf{A}}\right)\left(\underline{x} + \tilde{\underline{x}}\right) = \underline{b}$, the quotient of the relative errors is also bounded by $\kappa_2(\mathbf{A})$

$$\begin{aligned}\frac{\|\tilde{\underline{x}}\|_2/\|\underline{x}\|_2}{\|\tilde{\mathbf{A}}\|_2/\|\mathbf{A}\|_2} = \frac{\|\tilde{\underline{x}}\|_2\|\mathbf{A}\|_2}{\|\tilde{\mathbf{A}}\|_2\|\underline{x}\|_2} &= \frac{\|-\left(\mathbf{A} + \tilde{\mathbf{A}}\right)^{-1}\tilde{\mathbf{A}}\underline{x}\|_2\|\mathbf{A}\|_2}{\|\tilde{\mathbf{A}}\|_2\|\underline{x}\|_2} \\ &\leq \frac{\|\left(\mathbf{A} + \tilde{\mathbf{A}}\right)^{-1}\|_2\|\tilde{\mathbf{A}}\|_2\|\underline{x}\|_2\|\mathbf{A}\|_2}{\|\tilde{\mathbf{A}}\|_2\|\underline{x}\|_2} \\ &= \|\left(\mathbf{A} + \tilde{\mathbf{A}}\right)^{-1}\|_2\|\mathbf{A}\|_2 \\ &\leq \frac{\|\mathbf{A}^{-1}\|_2\|\mathbf{A}\|_2}{1 - \|\mathbf{A}^{-1}\|_2\|\tilde{\mathbf{A}}\|_2} \approx \kappa_2(\mathbf{A}),\end{aligned} \tag{2.19}$$

where the last approximate equality holds for small $\tilde{\mathbf{A}}$. Hence, when the condition number is large, a small variation in $\mathbf{A}$ or $\underline{b}$ can result in a large error in the solution $\underline{x}$. This effect will be demonstrated in Section 4.

2.3 A condition-number bound for SPD-matrices

Because $k(\cdot,\cdot)$ is symmetric and coercive on $\mathcal{V}^h(\Omega)$, the system matrix $\mathbf{K}$ is symmetric and positive definite (SPD). Due to the symmetry of $\mathbf{K}$, it has real-valued eigenvalues ($\lambda_i \in \mathbb{R}$) and an orthonormal set of eigenvectors ($\|\underline{\xi}_i\|_2 = 1$ $\forall i$, $\underline{\xi}_i^T \underline{\xi}_j = 0$ for $i \neq j$). As a result of this, for every vector $\underline{v} \in \mathbb{R}^n$ there exists a unique vector $\underline{\alpha} \in \mathbb{R}^n$ such that $\underline{v} = \sum_i \alpha_i \underline{\xi}_i$ and $\|\underline{\alpha}\|_2 = \|\underline{v}\|_2$. Therefore

$$\|\mathbf{K}\|_2^2 = \max_{\|\underline{v}\|_2=1} \|\mathbf{K}\underline{v}\|_2^2 = \max_{\|\underline{\alpha}\|_2=1} \|\mathbf{K}\sum_i \alpha_i \underline{\xi}_i\|_2^2 = \max_{\|\underline{\alpha}\|_2=1} \|\sum_i \lambda_i \alpha_i \underline{\xi}_i\|_2^2, \quad (2.20)$$

and due to the orthonormality of the eigenvectors

$$\|\mathbf{K}\|_2^2 = \max_{\|\underline{\alpha}\|_2=1} \sum_i \lambda_i^2 \alpha_i^2 = |\lambda|_{\max}^2. \quad (2.21)$$

Equation (2.21) follows from creating a set of indices I such that $|\lambda_i| = |\lambda|_{\max}$ for $i \in I$ and $|\lambda_i| < |\lambda|_{\max}$ for $i \notin I$. The maximum is then attained for $\sum_{i\in I} \alpha_i^2 = 1$ and $\alpha_i = 0$ for $i \notin I$. Similarly

$$\begin{aligned} \|\mathbf{K}^{-1}\|_2^2 &= \max_{\|\underline{v}\|_2=1} \|\mathbf{K}^{-1}\underline{v}\|_2^2 = \max_{\|\underline{v}\|_2=1} \frac{1}{\|\mathbf{K}\underline{v}\|_2^2} = \max_{\|\underline{\alpha}\|_2=1} \frac{1}{\|\sum_i \lambda_i \alpha_i \underline{\xi}_i\|_2^2} \\ &= \max_{\|\underline{\alpha}\|_2=1} \frac{1}{\sum_i \lambda_i^2 \alpha_i^2} = \frac{1}{|\lambda|_{\min}^2}, \end{aligned} \quad (2.22)$$

yielding

$$\kappa_2(\mathbf{K}) = \frac{|\lambda|_{\max}}{|\lambda|_{\min}}. \quad (2.23)$$

Also, due to the symmetry of $\mathbf{K}$, the Rayleigh quotient is bounded by the eigenvalues

$$\frac{\underline{v}^T \mathbf{K} \underline{v}}{\underline{v}^T \underline{v}} = \frac{\sum_{i,j} \alpha_i \alpha_j \underline{\xi}_i^T \mathbf{K} \underline{\xi}_j}{\sum_{i,j} \alpha_i \alpha_j \underline{\xi}_i^T \underline{\xi}_j} = \frac{\sum_i \lambda_i \alpha_i^2}{\sum_i \alpha_i^2}, \quad (2.24)$$

hence

$$\lambda_{\min} \leq \frac{\underline{v}^T \mathbf{K} \underline{v}}{\underline{v}^T \underline{v}} \leq \lambda_{\max} \quad \forall \underline{v} \in \mathbb{R}^n. \quad (2.25)$$

Inequality (2.25) is sharp as $\lambda_i = \underline{\xi}_i^T \mathbf{K} \underline{\xi}_i$ and therefore it holds that,

$$\begin{aligned} \lambda_{\min} &= \min_{\underline{v}\neq\underline{0}} \frac{\underline{v}^T \mathbf{K} \underline{v}}{\underline{v}^T \underline{v}} = \min_{\|\underline{v}\|_2=1} \underline{v}^T \mathbf{K} \underline{v}, \\ \lambda_{\max} &= \max_{\underline{v}\neq\underline{0}} \frac{\underline{v}^T \mathbf{K} \underline{v}}{\underline{v}^T \underline{v}} = \max_{\|\underline{v}\|_2=1} \underline{v}^T \mathbf{K} \underline{v}. \end{aligned} \quad (2.26)$$

Due to the positive definiteness of $\mathbf{K}$

$$\begin{aligned}|\lambda|_{\min} &= \lambda_{\min},\\ |\lambda|_{\max} &= \lambda_{\max},\end{aligned} \tag{2.27}$$

and therefore

$$\kappa_2(\mathbf{K}) = \frac{|\lambda|_{\max}}{|\lambda|_{\min}} = \max_{\|\underline{u}\|_2=1, \|\underline{v}\|_2=1} \frac{\underline{u}^T\mathbf{K}\underline{u}}{\underline{v}^T\mathbf{K}\underline{v}}. \tag{2.28}$$

The maximum over the full set $\|\underline{u}\|_2 = 1$, $\|\underline{v}\|_2 = 1$ is bounded from below by the maximum over the smaller set of standard unit vectors, hence

$$\max_{\|\underline{u}\|_2=\|\underline{v}\|_2=1} \frac{\underline{u}^T\mathbf{K}\underline{u}}{\underline{v}^T\mathbf{K}\underline{v}} \geq \max_{i,j} \frac{\underline{e}_i^T\mathbf{K}\underline{e}_i}{\underline{e}_j^T\mathbf{K}\underline{e}_j} = \max_{i,j} \frac{K_{ii}}{K_{jj}}, \tag{2.29}$$

with $\underline{e}_i$ the i^{th} standard unit vector. Combining (2.28) and (2.29) yields

$$\kappa_2(\mathbf{K}) \geq \max_{i,j} \frac{K_{ii}}{K_{jj}}. \tag{2.30}$$

Inequality (2.30) is a lower bound for the condition number of SPD-matrices. The next subsection will show that (2.30) is very useful to estimate the condition number of FCM-matrices.

2.4 Application to FCM-matrices

For FCM-matrices

$$K_{ii} = k(\phi_i, \phi_i) = \|\phi_i\|_k^2, \tag{2.31}$$

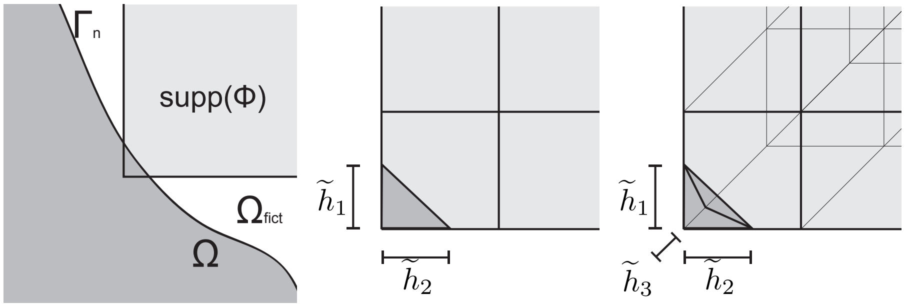
with ϕ_i the i^{th} basis function. Therefore, inequality (2.30) applied to FCM-matrices yields

$$\kappa_2(\mathbf{K}) \geq \max_{i,j} \frac{\|\phi_i\|_k^2}{\|\phi_j\|_k^2}. \tag{2.32}$$

Because basis functions whose support only slightly intersects Ω can become arbitrarily small, the condition number can grow arbitrarily large. Especially large grids have many cut-off functions and are therefore likely to have at least one function that is cut off undesirably.

To estimate the condition number of an FCM-matrix, a discretisation with B-spline basis functions as described in appendix A is considered. When an unmatching, one-dimensional, uniform grid of order p and grid size h $(= \xi_{i+1} - \xi_i)$ is cut off by the boundary between ξ_i and ξ_{i+1}, the last function is only supported on $[\xi_i, \xi_{i+1}] \cup \Omega$ and is locally proportional to $(x/h)^p$, with x a local coordinate originating in ξ_i. When a multidimensional, uniform, rectilinear grid of order p and grid size h in all directions is cut off, functions whose support only intersects Ω on (a part of) one element, as displayed in Figure 2.2, are locally proportional to

$$\phi(\underline{x}) \propto \prod_{i=1}^{d} \left(\frac{x_i}{h}\right)^p, \tag{2.33}$$



(a) Schematic representation of a support that is cut off by the boundary.

(b) A two-dimensional support that is cut off within one element.

(c) A three-dimensional support that is cut off within one element.

Figure 2.2: A basis function whose support only slightly intersects Ω and is therefore cut off.

with x_i the local coordinate in the i^{th} direction and d the number of dimensions.

When it is assumed that such a basis function is cut off by a natural boundary, then $b(\phi, \phi) = c(\phi, \phi) = 0$ and $\|\phi\|_k$ is proportional to the H_0^1-seminorm. Therefore, with $\widetilde{h}_i$ as in Figure 2.2

$$\begin{aligned}
\|\phi\|_k^2 \propto \|\phi\|_{H_0^1}^2 &= \int\limits_{\Omega} \sum_{i=1}^{d} (\partial_{x_i}\phi)^2 \,\mathrm{d}\Omega \\
&= \int\limits_{x_1=0}^{\widetilde{h}_1} \int\limits_{x_2=0}^{\widetilde{h}_2(1-x_1/\widetilde{h}_1)} \cdots \int\limits_{x_d=0}^{\widetilde{h}_d(1-\sum_{i=1}^{d-1} x_i/\widetilde{h}_i)} \sum_{i=1}^{d} (\partial_{x_i}\phi)^2 \prod_{i=1}^{d} \mathrm{d}x_i \\
&\propto \int\limits_{x_1=0}^{\widetilde{h}_1} \int\limits_{x_2=0}^{\widetilde{h}_2(1-x_1/\widetilde{h}_1)} \cdots \int\limits_{x_d=0}^{\widetilde{h}_d(1-\sum_{i=1}^{d-1} x_i/\widetilde{h}_i)} \sum_{i=1}^{d} \left(\partial_{x_i} \prod_{i=1}^{d} \left(\frac{x_i}{h}\right)^p\right)^2 \prod_{i=1}^{d} \mathrm{d}x_i \qquad (2.34) \\
&\propto \frac{1}{h^{2pd}} \int\limits_{x_1=0}^{\widetilde{h}_1} \int\limits_{x_2=0}^{\widetilde{h}_2(1-x_1/\widetilde{h}_1)} \cdots \int\limits_{x_d=0}^{\widetilde{h}_d(1-\sum_{i=1}^{d-1} x_i/\widetilde{h}_i)} \sum_{i=1}^{d} x_i^{-2} \prod_{i=1}^{d} x_i^{2p} \mathrm{d}x_i \\
&= \frac{1}{h^{2pd}} \sum_{i=1}^{d} \widetilde{h}_i^{-2} \prod_{i=1}^{d} \widetilde{h}_i^{2p+1} \int\limits_{y_1=0}^{1} \int\limits_{y_2=0}^{(1-y_1)} \cdots \int\limits_{y_d=0}^{1-\sum_{i=1}^{d-1} y_i} y_1^{-2} \prod_{i=1}^{d} y_i^{2p} \mathrm{d}y_i
\end{aligned}$$

$$\begin{aligned}&\propto \frac{1}{h^{2pd}}\sum_{i=1}^{d}\widetilde{h}_i^{-2}|\mathrm{supp}(\phi)|^{2p+1}\\&\approx \frac{1}{h^{2pd}}|\mathrm{supp}(\phi)|^{2p+1-2/d},\end{aligned}\tag{2.34}$$

where in the approximation in the last step it is assumed that $\mathcal{O}(\widetilde{h}_1) = \mathcal{O}(\widetilde{h}_2) = \cdots = \mathcal{O}(\widetilde{h}_d)$, which is the worst-case scenario as this results in the, in norm, smallest function. The norm of a basis function that is not cut off is estimated on a full element. For $p > 0$ the support is larger than one element, but for the determination of the order of magnitude it suffices to only consider a single element, hence

$$\begin{aligned}\|\phi\|_k^2 \propto \|\phi\|_{H_0^1}^2 &\propto \int\limits_{x_1=0}^{h}\int\limits_{x_2=0}^{h}\cdots\int\limits_{x_d=0}^{h}\sum_{i=1}^{d}\left(\partial_{x_i}\prod_{i=1}^{d}\left(\frac{x_i}{h}\right)^p\right)^2\prod_{i=1}^{d}\mathrm{d}x_i\\&\propto \frac{1}{h^{2pd}}\int\limits_{x=0}^{h}x^{2p-2}\mathrm{d}x\left(\int\limits_{x=0}^{h}x^{2p}\mathrm{d}x\right)^{d-1}\\&= \frac{h^{2pd+d-2}}{h^{2pd}}\int\limits_{x=0}^{1}x^{2p-2}\mathrm{d}x\left(\int\limits_{x=0}^{1}x^{2p}\mathrm{d}x\right)^{d-1}\\&\propto \frac{h^{2pd+d-2}}{h^{2pd}}.\end{aligned}\tag{2.35}$$

Combining (2.32), (2.34), (2.35) and denoting the basis function with the smallest support by $\widetilde{\phi}$ yields

$$\begin{aligned}\kappa_2(\mathbf{K}) \geq \max_{i,j}\frac{\|\phi_i\|_k^2}{\|\phi_j\|_k^2} &\propto \frac{h^{2pd+d-2}}{|\mathrm{supp}(\widetilde{\phi})|^{2p+1-2/d}}\\&= \left(\frac{h^d}{|\mathrm{supp}(\widetilde{\phi})|}\right)^{2p+1-2/d} = \eta^{-(2p+1-2/d)},\end{aligned}\tag{2.36}$$

with η the minimal volume fraction defined as

$$\eta = \min_i \eta_i = \min_i \frac{|\mathrm{supp}(\phi_i)|}{h^d},\tag{2.37}$$

where it is assumed that the smallest cut-off element is of the form displayed in Figure 2.2, which is a reasonable assumption as shown in [21]. It should be noted that for splines of order $p > 0$, an uncut function ϕ_i has a volume fraction $\eta_i > 1$ as it is supported on more than one element. Every unmatching grid has at least one function ϕ_i with $\eta_i < 1$ however, and hence $0 < \eta < 1$. The final result is

$$\kappa_2(\mathbf{K}) \geq C \propto \eta^{-(2p+1-2/d)},\tag{2.38}$$

which indicates a strong dependence on the spline order p, which is inconvenient as for accuracy considerations higher-order splines are preferred. Especially for FCM this is a major drawback, as one of the advantages of FCM is the simplicity at which higher order discretisations can be implemented. The derivation of (2.38) can also be done in a more general form for the H_0^n-seminorm, which results in

$$\kappa_2(\mathbf{K}) \geq C \propto \eta^{-(2p+1-2n/d)}, \tag{2.39}$$

and will be used to estimate the condition number for L^2-projections in the next subsection. When a function is cut off by an essential boundary, the situation is slightly different. To keep the analysis of these situations as simple as possible, it is assumed that $\widetilde{h}_1 = \widetilde{h}_2 = \cdots = \widetilde{h}_d = \widetilde{h}$, which is valid to estimate the orders of magnitude. For a small cut-off element as in Figure 2.2, $b(\phi, \phi)$ can be estimated by

$$\begin{aligned} b(\phi,\phi) \propto \phi \partial_n \phi |\Gamma| &\approx \overbrace{\left(\frac{\widetilde{h}}{h}\right)^{pd}}^{\phi} \overbrace{\frac{\widetilde{h}^{pd-1}}{h^{pd}}}^{\partial_n \phi} \overbrace{\widetilde{h}^{d-1}}^{|\Gamma|} = \frac{\widetilde{h}^{d(2p+1-2/d)}}{h^{2pd}} \\ &\approx \frac{1}{h^{2pd}} |\mathrm{supp}(\phi)|^{2p+1-2/d}, \end{aligned} \tag{2.40}$$

which is the same order of magnitude as (2.34). The stability term $c(\phi, \phi)$ can be estimated by

$$\begin{aligned} c(\phi,\phi) \propto \beta \phi^2 |\Gamma| &\approx \overbrace{\frac{1}{\widetilde{h}}}^{\beta} \overbrace{\left(\frac{\widetilde{h}}{h}\right)^{2pd}}^{\phi^2} \overbrace{\widetilde{h}^{d-1}}^{|\Gamma|} = \frac{\widetilde{h}^{d(2p+1-2/d)}}{h^{2pd}} \\ &\approx \frac{1}{h^{2pd}} |\mathrm{supp}(\phi)|^{2p+1-2/d}, \end{aligned} \tag{2.41}$$

and is therefore of the same order of magnitude as (2.34) as well, where it is used that β scales with $1/\widetilde{h}$ as mentioned in Subsection 2.1. From (2.40) and (2.41) it follows that the norm of a small cut-off function when it is cut by an essential boundary is of the same order of magnitude as when it is cut by a natural boundary. For a function that is not cut off undesirably by an essential boundary, $b(\phi, \phi)$ can be estimated by

$$b(\phi,\phi) \propto \phi \partial_n \phi |\Gamma| \approx \overbrace{1}^{\phi} \overbrace{\frac{1}{h}}^{\partial_n \phi} \overbrace{h^{d-1}}^{|\Gamma|} = h^{d-2}, \tag{2.42}$$

which is the same order of magnitude as (2.35). When an element-wise stabilisation parameter is applied, $c(\phi, \phi)$ for such a function can be estimated by

$$c(\phi,\phi) \propto \beta \phi^2 |\Gamma| \approx \overbrace{\frac{1}{h}}^{\beta} \overbrace{1}^{\phi^2} \overbrace{h^{d-1}}^{|\Gamma|} = h^{d-2}, \tag{2.43}$$

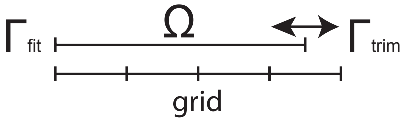


Figure 2.3: Schematic representation of the one-dimensional test case.

which is also the same order of magnitude as (2.35). Therefore, when elementwise stabilisation is applied, (2.38) also holds for essential boundary conditions on an unmatching boundary. When a global stabilisation parameter is applied, $c(\phi, \phi)$ for such a function can be estimated by

$$c(\phi,\phi) \propto \beta\phi^2|\Gamma| \approx \overbrace{\frac{1}{h\eta^{1/d}}}^{\beta}\overbrace{1}^{\phi^2}\overbrace{h^{d-1}}^{|\Gamma|} = h^{d-2}\eta^{-1/d}, \tag{2.44}$$

where it is used that $\widetilde{h} \approx h\eta^{1/d}$ as $\eta = (\widetilde{h}/h)^d$. As a result of this, there will be functions with a norm of magnitude $h^{d-2}\eta^{-1/d}$ in the system such that the application of (2.32) to a globally stabilised system yields

$$\kappa_2(\mathbf{K}) \geq C \propto \eta^{-(2p+1-2/d)}\eta^{-1/d} = \eta^{-(2p+1-1/d)}. \tag{2.45}$$

Inequality (2.45) indicates that the condition number of globally stabilised systems has a stronger dependence on η than the condition number of locally stabilised systems, because small values of η not only cause a certain function to become very small, but also cause other functions to become very large.

2.5 Results

To verify the scaling relations (2.38), (2.39) and (2.45), several numerical experiments were done. In one dimension, Laplace's problem on the domain $\Omega = (0, 1 - h + \eta h)$ was discretised by a uniform grid with $h = 1/4$ and $p \in \{1, 2, 3\}$. A schematic representation of this test case is displayed in Figure 2.3. At Γ_{fit}, the grid matched the domain and an essential boundary condition was applied in a strong manner. At Γ_{trim}, the grid did not match the domain, and in the first experiment a natural boundary condition was applied and in the second experiment a globally stabilised essential boundary condition was applied in a weak manner. It should be noted that stabilisation in one dimension is global by definition, as the size of the boundary is independent of the volume fraction.

The resulting condition numbers from these discretisations are plotted against η in Figure 2.4 for the natural boundary condition, and in Figure 2.5 for the essential boundary condition, in which also the penalty factors β are displayed. In these figures, the red circles depict the condition numbers and the white circles

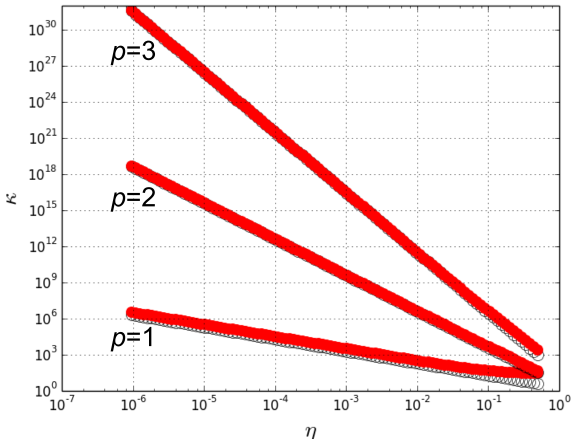
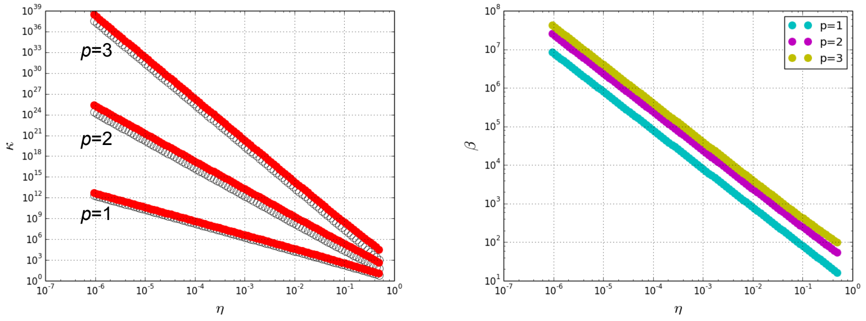


Figure 2.4: Condition numbers against η for the one-dimensional test case with a natural boundary condition along the trimmed boundary for $p \in \{1, 2, 3\}$.

depict the lower bounds for the condition numbers from (2.30). In Figure 2.4, a slope of -1 $(= -(2p+1-2/d) = -(2+1-2))$ is visible for $p = 1$, a slope of -3 for $p = 2$ and a slope of -5 for $p = 3$, which is in agreement with (2.38) for a natural boundary condition. In Figure 2.5a, a slope of -2 is visible for $p = 1$, a slope of -4 for $p = 2$ and a slope of -6 for $p = 3$, which is in agreement with (2.45) for a globally stabilised essential boundary condition, which the stabilisation in one dimension is by definition. The penalty factors β in Figure 2.5b form a slope of -1, which is explained by $\beta \propto 1/\widetilde{h} \propto 1/\eta$ in one dimension.

In two dimensions, an L^2-projection was done and Laplace's problem was solved on a half ring with inner radius 1/2 and outer radius 1. A schematic representation of this test case is displayed in Figure 2.6. A uniform, rectilinear grid with $h = 1/4$ was used with $p \in \{0, 1, 2, 3\}$ for the L^2-projection and $p \in \{1, 2, 3\}$ for Laplace's problem. The grid matched the domain along the straight edges of the half ring (Γ_{fit}) and was unmatching at the inner and outer radius of the ring (Γ_{trim}). By shifting the grid along Γ_{fit} with steps of $h/2000$, 1001 different discretisations were generated. For Laplace's problem, an essential boundary condition was applied in a strong manner along Γ_{fit} and a natural boundary condition, an element-wise stabilised essential boundary condition and a globally stabilised essential boundary condition were applied along Γ_{trim} in a weak manner.

The resulting condition numbers (and for the globally stabilised case also the penalty factors β) are plotted against η in Figure 2.7 for the L^2-projection and in Figures 2.8, 2.9 and 2.10 for Laplace's problem with a natural boundary condition, an element-wise stabilised essential boundary condition and a globally



(a) Condition numbers.

(b) Penalty factors.

Figure 2.5: Condition numbers and penalty factors against η for the one-dimensional test case with an essential boundary condition along the trimmed boundary for $p \in \{1, 2, 3\}$.

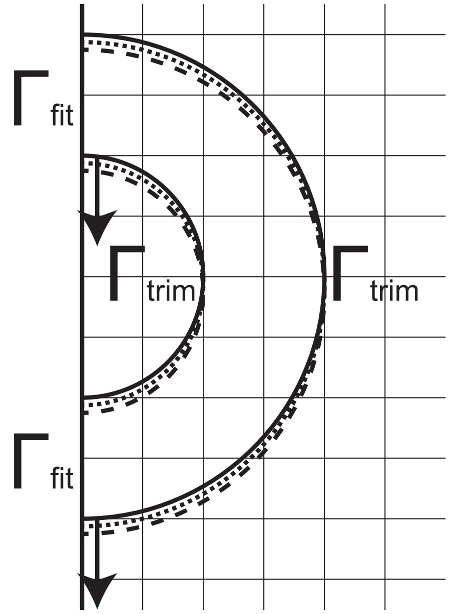


Figure 2.6: Schematic representation of the two-dimensional test case.

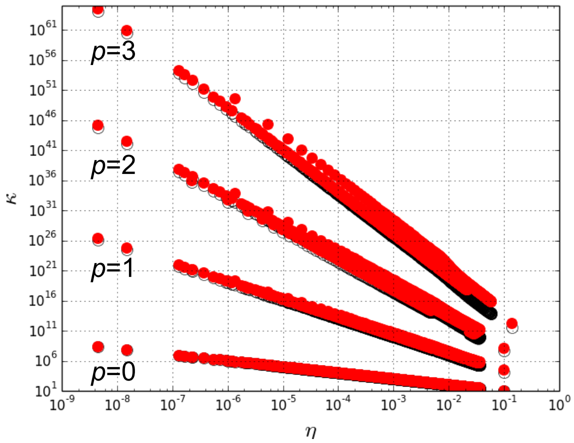


Figure 2.7: Condition numbers against η for the two-dimensional test case with an L^2-projection for $p \in \{0, 1, 2, 3\}$.

stabilised essential boundary condition along Γ_{trim}, respectively. In these figures the red circles depict the condition numbers and the white circles depict the lower bounds for the condition numbers from (2.30). In Figure 2.7, a slope of -1 is visible for $p = 0$, a slope of -3 for $p = 1$, a slope of -5 for $p = 2$ and a slope of -7 for $p = 3$, which is in agreement with (2.39) for an L^2-projection. In Figures 2.8 and 2.9, a slope of -2 is visible for $p = 1$, a slope of -4 for $p = 2$ and a slope of -6 for $p = 3$, which is in agreement with (2.38) for a natural boundary condition or an element-wise stabilised essential boundary condition. In Figure 2.10a, a slope of $-2\frac{1}{2}$ is visible for $p = 1$, a slope of $-4\frac{1}{2}$ for $p = 2$ and a slope of $-6\frac{1}{2}$ for $p = 3$, which is in agreement with (2.45) for a globally stabilised essential boundary condition. As predicted by inequality (2.45), global stabilisation results in a steeper slope of the condition number than local stabilisation. The penalty factors β in Figure 2.10b form a slope of $-\frac{1}{2}$, which is explained by $\beta \propto 1/\widetilde{h} \propto 1/\eta^{1/d} = \eta^{-1/2}$ in two dimensions.

In three dimensions, Laplace's problem is solved on a three-dimensional extension of the two-dimensional domain; a hemisphere with a centred hemispherical exclusion with inner radius 1/2 and outer radius 1. A schematic representation of this test case is displayed in Figure 2.11. A uniform, rectilinear grid with $h = 1/4$ was used with $p \in \{1, 2, 3\}$. The grid matched the domain along the flat base of the hemisphere (Γ_{fit}) and was unmatching at the inner and outer radius (Γ_{trim}). By shifting the grid along Γ_{fit} in 10 steps of $h/20$ in both directions along the base of the hemisphere, 66 different discretisations were generated (due to symmetry). An essential boundary condition was applied in a strong manner along Γ_{fit} and a natural boundary condition and an element-wise stabilised essential boundary condition were applied along Γ_{trim} in a weak manner.

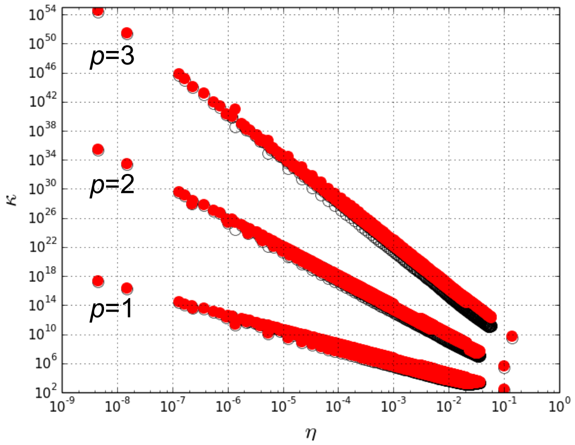


Figure 2.8: Condition numbers against η for the two-dimensional test case with a natural boundary condition along the trimmed boundary for $p \in \{1, 2, 3\}$.

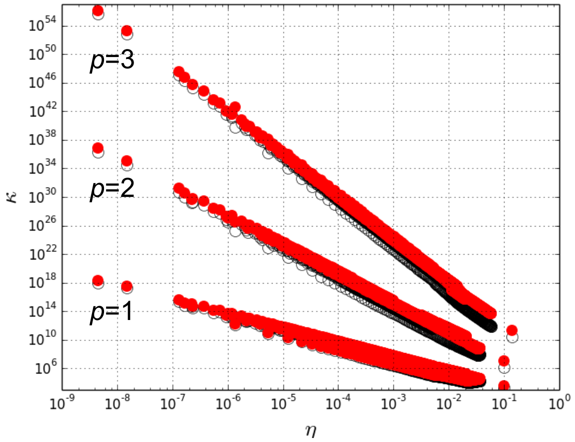
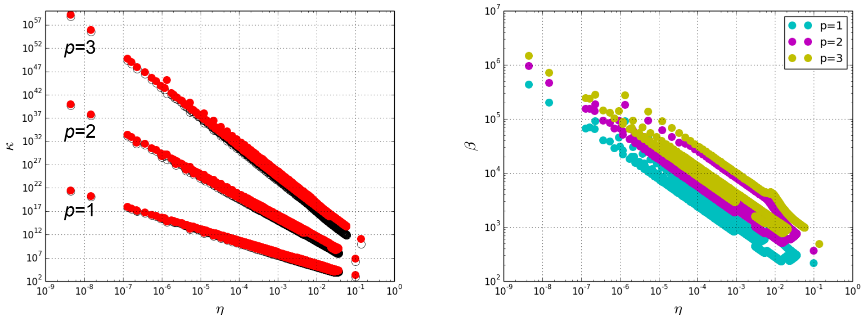


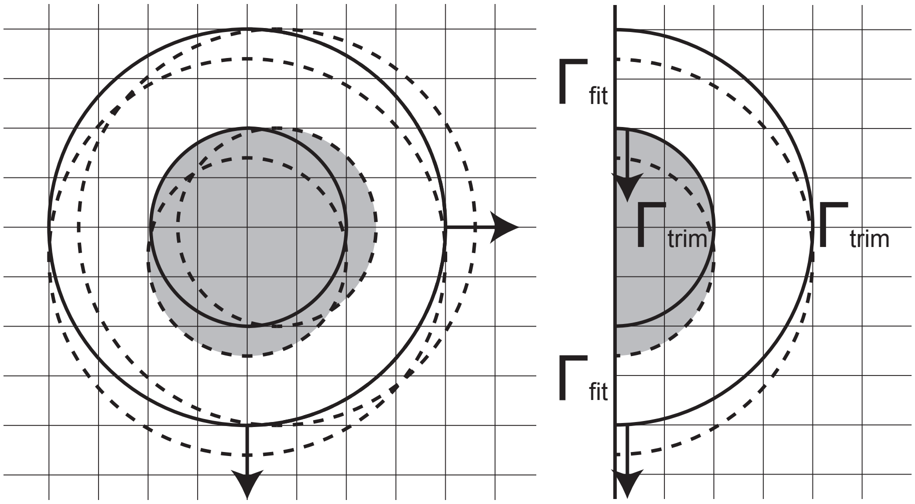
Figure 2.9: Condition numbers against η for the two-dimensional test case with an element-wise stabilised essential boundary condition along the trimmed boundary for $p \in \{1, 2, 3\}$.



(a) Condition numbers.

(b) Penalty factors.

Figure 2.10: Condition numbers and penalty factors against η for the two-dimensional test case with a globally stabilised essential boundary condition along the trimmed boundary for $p \in \{1, 2, 3\}$.



(a) Left view.

(b) Front and top view.

Figure 2.11: Schematic representation of the three-dimensional test case.

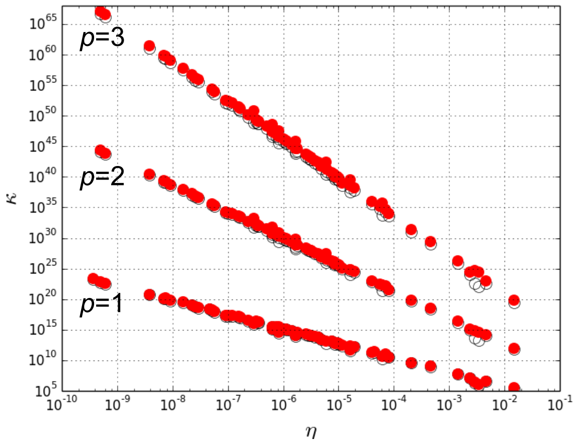


Figure 2.12: Condition numbers against η for the three-dimensional test case with a natural boundary condition along the trimmed boundary for $p \in \{1, 2, 3\}$.

The resulting condition numbers are plotted against η in Figures 2.12 and 2.13 for the natural boundary condition and the element-wise stabilised essential boundary condition along Γ_{trim}, respectively. The red circles depict the condition numbers and the white circles depict the lower bounds for the condition numbers from (2.30). In both figures, a slope of -2 is visible for $p = 1$, a slope of -4 for $p = 2$ and a slope of -6 for $p = 3$, which is in agreement with (2.38) for a natural boundary condition or an element-wise stabilised essential boundary condition.

Experiments with an L^2-projection in one and three dimensions and with a globally stabilised essential boundary condition along Γ_{trim} in three dimensions have also been found to yield the theoretically predicted proportionality rates from (2.38), (2.39) and (2.45), but are not presented here for the sake of brevity.

2.6 Existing conditioners and solvers

Currently there are different approaches to improve or overcome the conditioning problems associated with FCM. The authors of [5] have eliminated basis functions with a volume fraction smaller than a certain volume fraction ε (which was set to 10^{-6}). Following (2.38), solving a two-dimensional Poisson's problem with cubic splines then still allows condition numbers up to 10^{24}. To significantly reduce the condition number with this strategy, ε would have to be set much larger, which reduces the quality of the used function space. In many articles (*e.g.* [2, 7–15]), the norm of cut-off functions is increased by virtual stiffness,

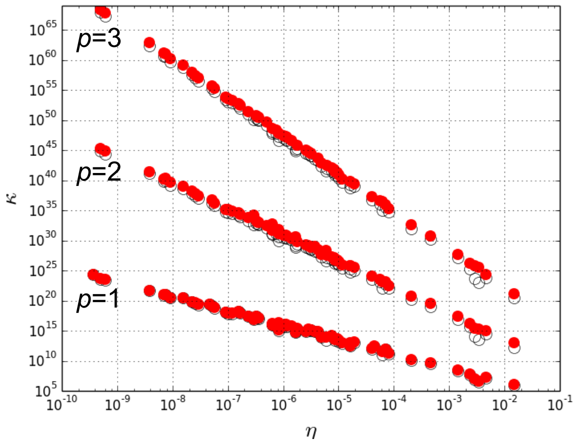
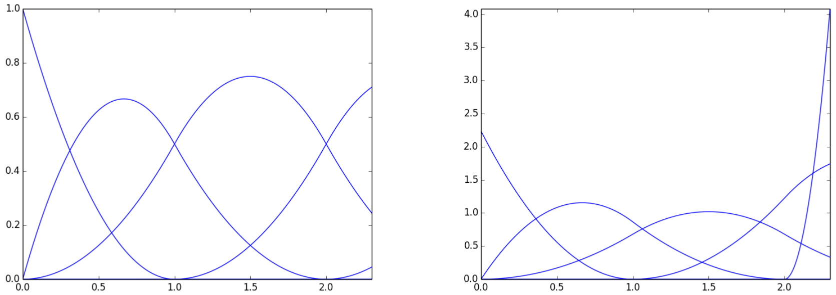


Figure 2.13: Condition numbers against η for the three-dimensional test case with an element-wise stabilised essential boundary condition along the trimmed boundary for $p \in \{1, 2, 3\}$.

which implies that $a(\cdot,\cdot)$ multiplied by a small parameter α -usually in the range of $(10^{-8} - 10^{-14})$ for higher-order splines [2]- is also integrated over the fictitious domain. Although this approach has been shown to be feasible in many situations, from a mathematical perspective this is not consistent with the original problem and decreases the accuracy with increasing α [11]. Furthermore, following (2.30), this still allows condition numbers of up to α^{-1}. Therefore this forces an inconvenient compromise between conditioning and accuracy to be made. Due to the lack of proper conditioning techniques, the authors of [8] resort to direct solvers which perform better than iterative methods for the poorly conditioned systems encountered in FCM. As discussed in Subsection 2.2 however, direct solvers applied to systems with large condition numbers can still give inaccurate results through error propagation. Furthermore, direct solvers generally require $\mathcal{O}(n^3)$ flops and $\mathcal{O}(n^2)$ memory storage, opposed to $\mathcal{O}(n^2)$ flops and $\mathcal{O}(n)$ memory storage that are typically required for iterative solvers for an $n \times n$ system [22–24]. Therefore iterative solvers are preferred for large-scale simulations. As all these strategies have clear drawbacks, a new strategy to improve the conditioning of FCM systems is proposed in the next section.



(a) Unscaled basis functions.

(b) Basis functions scaled to $\|\phi\|_{L^2} = 1$.

Figure 3.1: Different bases for the same finite-dimensional function space.

3 Improving the conditioning by means of basis function scaling

3.1 Basis function scaling

As shown in subsections 2.3, 2.4 and 2.5, the conditioning problems associated with the Finite Cell Method are caused by the large differences between the magnitude of the norms of separate basis functions. To improve the conditioning, it is proposed to scale the basis functions such that all basis functions have the same norm. Doing this does not change the finite-dimensional function space but just changes the basis that is used for it. Therefore the approximated solution is unaffected. An example with $\|\phi\|_{L^2} = 1$ for all basis functions is displayed in Figure 3.1.

As a result of the basis function scaling (from here on simply referred to as scaling), inequality (2.32) yields $\kappa_2 \geq 1$ as a lower bound for the condition number. It should be noted that this strategy only decreases the lower bound for the condition number from inequality (2.38) and does not guarantee to reduce the condition number itself. Also, singular systems cannot become regular after scaling the basis functions. In Subsection 3.2 and Section 5, the effectiveness of scaling will be studied numerically.

The scaled system of equations can be formed by pre- and post-conditioning the original system of equations

$$\mathbf{K}\underline{u} = \underline{l}, \tag{3.1}$$

with the diagonal matrix

$$\mathbf{D} = \begin{bmatrix} \frac{1}{\|\phi_1\|_k} & & \\ & \ddots & \\ & & \frac{1}{\|\phi_n\|_k} \end{bmatrix}, \tag{3.2}$$

such that the scaled system of equations becomes

$$\begin{aligned} \mathbf{DKD}\underline{y} &= \mathbf{D}\underline{l}, \\ \underline{u} &= \mathbf{D}\underline{y}. \end{aligned} \tag{3.3}$$

The scaled matrix has the value 1 on all of its diagonals and is diagonally dominant because of the Cauchy-Schwarz inequality

$$\mathbf{DKD} = \begin{cases} (DKD)_{i,i} = \frac{k(\phi_i,\phi_i)}{\|\phi_i\|_k^2} = \frac{\|\phi_i\|_k^2}{\|\phi_i\|_k^2} = 1, \\ (DKD)_{i,j} = \frac{k(\phi_i,\phi_j)}{\|\phi_i\|_k\|\phi_j\|_k} \leq \frac{\|\phi_i\|_k\|\phi_j\|_k}{\|\phi_i\|_k\|\phi_j\|_k} = 1 \quad (i \neq j). \end{cases} \tag{3.4}$$

System (3.3) shows that scaling is the same as using a diagonal pre- and post-conditioner. It is known from literature that diagonal pre- and post-conditioners do not guarantee to reduce the condition number [25, 26]. In [27] it is mentioned that especially for the Euclidean norm, optimality of a diagonal pre- and post-conditioner is usually not easily verified. However, Theorem 4.3 in [27] does prove that

$$\kappa_2(\mathbf{DKD}) \leq q\kappa_2(\widetilde{\mathbf{D}}\mathbf{K}\widetilde{\mathbf{D}}), \tag{3.5}$$

with $\mathbf{D}$ the diagonal matrix as in (3.2), $\widetilde{\mathbf{D}}$ the diagonal matrix that optimally scales $\mathbf{K}$ and q the maximal number of nonzero elements in a row or column of $\mathbf{K}$. For rectilinear grids, which are applied in all experiments in this work, $q = (p+1)^d$ when a scalar equation is solved and $q = \widetilde{d}(p+1)^d$ when a vector equation is solved, with p the spline order and d and $\widetilde{d}$ the number of dimensions of the domain and the vector-field respectively. Because this bounds the condition number by a known constant multiplied by the optimally scaled condition number, scaling with $\mathbf{D}$ is at least quasi-optimal.

It is noted that scaling with the FCM-norm is not the only option. As $b(\phi, \phi)$ can be smaller than zero, it is possible that a function with a large influence on the solution is small in norm. It can therefore be argued that it might be better to scale with $\sqrt{a(\phi,\phi) + |b(\phi,\phi)| + c(\phi,\phi)}$ or $\sqrt{a(\phi,\phi) + c(\phi,\phi)}$, which are equivalent as $|b(\phi,\phi)| \leq a(\phi,\phi) + c(\phi,\phi)$ (see (2.9)), such that

$$\begin{aligned} a(\phi,\phi) + c(\phi,\phi) &\leq a(\phi,\phi) + |b(\phi,\phi)| + c(\phi,\phi) \\ &\leq 2\left(a(\phi,\phi) + c(\phi,\phi)\right). \end{aligned} \tag{3.6}$$

Defining diagonal matrices $\mathbf{D}_{a|b|c}$, $\mathbf{D}_{ac}$ and $\mathbf{D}_{*}$ as

$$\begin{aligned}
\mathbf{D}_{a|b|c} &= \begin{cases} \left(D_{a|b|c}\right)_{i,i} = \frac{1}{\sqrt{a(\phi_i,\phi_i)+|b(\phi_i,\phi_i)|+c(\phi_i,\phi_i)}}, \\ \left(D_{a|b|c}\right)_{i,j} = 0 \quad (i \neq j), \end{cases} \\
\mathbf{D}_{ac} &= \begin{cases} \left(D_{ac}\right)_{i,i} = \frac{1}{\sqrt{a(\phi_i,\phi_i)+c(\phi_i,\phi_i)}}, \\ \left(D_{ac}\right)_{i,j} = 0 \quad (i \neq j), \end{cases} \\
\mathbf{D}_{*} &= \begin{cases} \left(D_{*}\right)_{i,i} = \frac{\sqrt{a(\phi_i,\phi_i)+c(\phi_i,\phi_i)}}{\sqrt{a(\phi_i,\phi_i)+|b(\phi_i,\phi_i)|+c(\phi_i,\phi_i)}}, \\ \left(D_{*}\right)_{i,j} = 0 \quad (i \neq j), \end{cases} \quad = \mathbf{D}_{a|b|c}\mathbf{D}_{ac}^{-1},
\end{aligned} \tag{3.7}$$

it holds that $\mathbf{D}_{a|b|c} = \mathbf{D}_{*}\mathbf{D}_{ac}$ and $\mathbf{D}_{ac} = \mathbf{D}_{*}^{-1}\mathbf{D}_{a|b|c}$ and furthermore

$$\begin{aligned}
\|\mathbf{D}_{*}\|_2 &= \max_{i\in\{1,\ldots,n\}} \frac{\sqrt{a(\phi_i,\phi_i)+c(\phi_i,\phi_i)}}{\sqrt{a(\phi_i,\phi_i)+|b(\phi_i,\phi_i)|+c(\phi_i,\phi_i)}} \leq 1, \\
\|\mathbf{D}_{*}^{-1}\|_2 &= \max_{i\in\{1,\ldots,n\}} \frac{\sqrt{a(\phi_i,\phi_i)+|b(\phi_i,\phi_i)|+c(\phi_i,\phi_i)}}{\sqrt{a(\phi_i,\phi_i)+c(\phi_i,\phi_i)}} \leq \sqrt{2},
\end{aligned} \tag{3.8}$$

such that $\kappa_2(\mathbf{D}_{*}) \leq \sqrt{2}$. Equivalence follows from

$$\begin{aligned}
\kappa_2(\mathbf{D}_{a|b|c}\mathbf{K}\mathbf{D}_{a|b|c}) &= \kappa_2(\mathbf{D}_{*}\mathbf{D}_{ac}\mathbf{K}\mathbf{D}_{ac}\mathbf{D}_{*}) \\
&\leq \kappa_2(\mathbf{D}_{*})^2\kappa_2(\mathbf{D}_{ac}\mathbf{K}\mathbf{D}_{ac}) \leq 2\kappa_2(\mathbf{D}_{ac}\mathbf{K}\mathbf{D}_{ac}), \\
\kappa_2(\mathbf{D}_{ac}\mathbf{K}\mathbf{D}_{ac}) &= \kappa_2(\mathbf{D}_{*}^{-1}\mathbf{D}_{a|b|c}\mathbf{K}\mathbf{D}_{a|b|c}\mathbf{D}_{*}^{-1}) \\
&\leq \kappa_2(\mathbf{D}_{*})^2\kappa_2(\mathbf{D}_{a|b|c}\mathbf{K}\mathbf{D}_{a|b|c}) \leq 2\kappa_2(\mathbf{D}_{a|b|c}\mathbf{K}\mathbf{D}_{a|b|c}),
\end{aligned} \tag{3.9}$$

where it is used that, for induced norms

$$\begin{aligned}
\kappa(\mathbf{AB}) &= \|\mathbf{AB}\|\left\|\left(\mathbf{AB}\right)^{-1}\right\| = \|\mathbf{AB}\|\|\mathbf{B}^{-1}\mathbf{A}^{-1}\| \\
&\leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\|\mathbf{B}\|\|\mathbf{B}^{-1}\| = \kappa(\mathbf{A})\kappa(\mathbf{B}).
\end{aligned} \tag{3.10}$$

For large β, it can even be shown that scaling with $\|\phi\|_k$ is equivalent to scaling with $\sqrt{a(\phi,\phi)+c(\phi,\phi)}$. When the penalty factor $\beta \geq \gamma C$ everywhere on the domain for some constant $\gamma > 1$, (2.9) can be rewritten as

$$|b(\phi,\phi)| \leq \frac{C}{\varepsilon}a(\phi,\phi) + \frac{\varepsilon}{\gamma C}c(\phi,\phi) = \frac{1}{\sqrt{\gamma}}\left(a(\phi,\phi)+c(\phi,\phi)\right), \tag{3.11}$$

for $\varepsilon = \sqrt{\gamma}C$ (note that this implies $C < \varepsilon < \beta$). Therefore

$$\begin{aligned}
\frac{\sqrt{\gamma}-1}{\sqrt{\gamma}}\left(a(\phi,\phi)+c(\phi,\phi)\right) &\leq a(\phi,\phi)+b(\phi,\phi)+c(\phi,\phi) \\
&\leq \frac{\sqrt{\gamma}+1}{\sqrt{\gamma}}\left(a(\phi,\phi)+c(\phi,\phi)\right),
\end{aligned} \tag{3.12}$$

such that by a similar derivation

$$\frac{\sqrt{\gamma}-1}{\sqrt{\gamma}+1}\kappa_2(\mathbf{DKD}) \leq \kappa_2(\mathbf{D}_{ac}\mathbf{K}\mathbf{D}_{ac}) \leq \frac{\sqrt{\gamma}+1}{\sqrt{\gamma}-1}\kappa_2(\mathbf{DKD}). \tag{3.13}$$

As in this work all experiments were done with $\beta = 2C$, scaling with alternative scaling factors did not yield any significant differences in the condition number compared to standard diagonal scaling. Because of that, we propose to scale with the FCM-norm as this is the simplest option, the only option that guarantees diagonal dominance and because of the mathematical evidence that this bounds the condition number by q times the optimally scaled condition number. When the value of β is reduced however, it may be profitable to scale with a different factor.

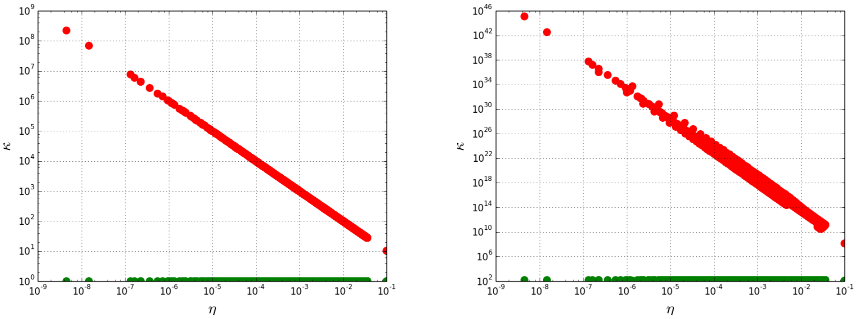
Because, for β large enough, the scaling procedure is independent of β, improving the condition number through scaling can be applied parallel with optimising β for accuracy, as proposed in [8]. Besides optimising β for accuracy, one could also study the effect of varying the value of β on the condition number, especially in combination with scaling. Studying these effects and optimising β for conditioning and accuracy in combination with scaling is beyond the scope of this work.

3.2 Results

Basis function scaling has been applied to the test cases from Section 2.5. The scaling significantly decreases the condition number and, except for systems with a globally stabilised essential boundary condition, has been observed to result in a condition number that is practically independent of η. Figure 3.2 displays the scaled and unscaled condition numbers for the L^2-projection in two dimensions for $p = 0$ and $p = 2$. In all figures in this section, the green circles represent the scaled condition numbers and the red circles represent the unscaled condition numbers. For $p = 0$, $\mathbf{K}$ is diagonal, such that after scaling it becomes the identity matrix with $\kappa_2(\mathbf{DKD}) = 1$, which is visible in Figure 3.2a. This is clearly optimally scaled, as also follows from (3.5).

Figures 3.3 and 3.4 display the scaled and unscaled condition numbers for Laplace's problem in two dimensions with a natural boundary condition and an element-wise stabilised essential boundary condition for $p = 2$ and $p = 3$. Figure 3.5 displays the scaled and unscaled condition numbers for Laplace's problem in three dimensions with a natural boundary condition and an element-wise stabilised essential boundary condition for $p = 2$. It is visible that for all experiments the condition numbers are significantly reduced and are practically independent of η.

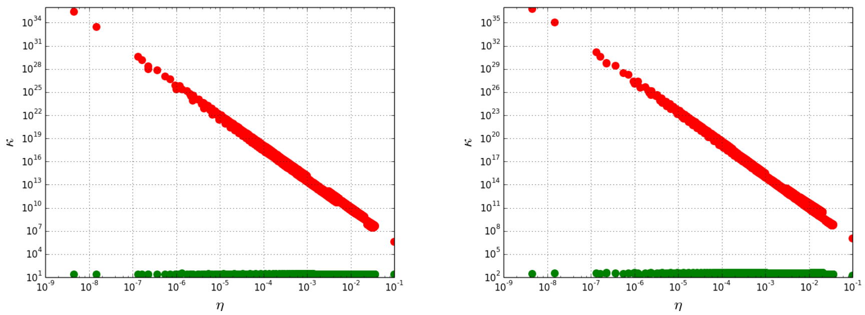
Figure 3.6 displays the scaled and unscaled condition numbers for Laplace's problem in two and three dimensions with a globally stabilised essential boundary condition for $p = 2$. It is visible that for all experiments the condition



(a) $p = 0$.

(b) $p = 2$.

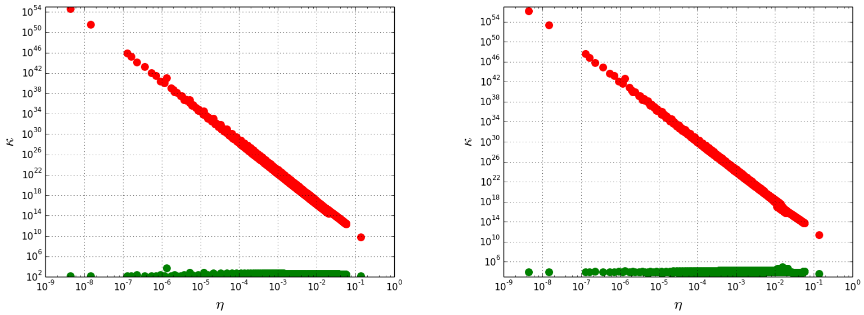
Figure 3.2: Unscaled (red) and scaled (green) condition numbers against η for the two-dimensional L^2-projection.



(a) Natural boundary condition.

(b) Element-wise stabilised essential boundary condition.

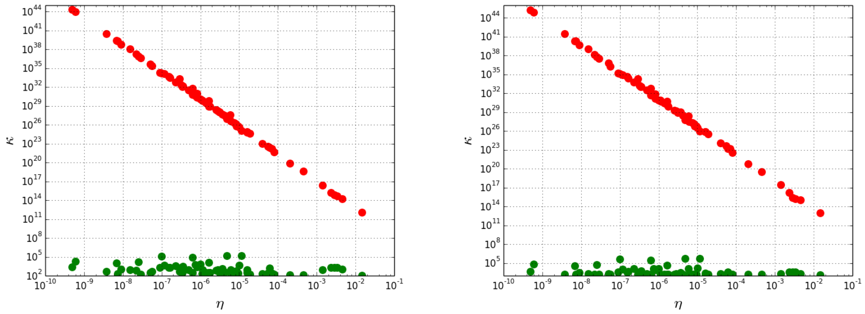
Figure 3.3: Unscaled (red) and scaled (green) condition numbers against η for the two-dimensional test case for $p = 2$.



(a) Natural boundary condition.

(b) Element-wise stabilised essential boundary condition.

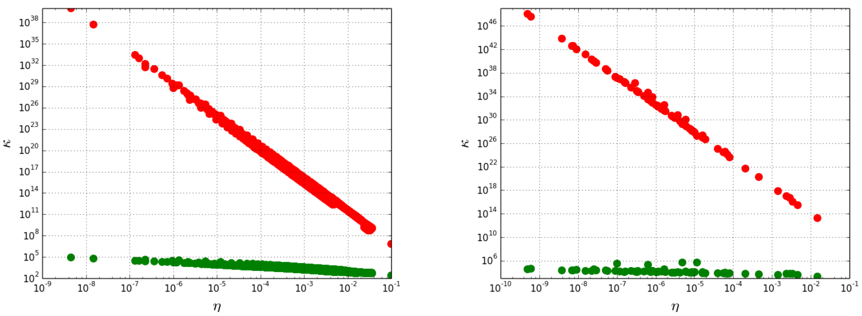
Figure 3.4: Unscaled (red) and scaled (green) condition numbers against η for the two-dimensional test case for $p = 3$.



(a) Natural boundary condition.

(b) Element-wise stabilised essential boundary condition.

Figure 3.5: Unscaled (red) and scaled (green) condition numbers against η for the three-dimensional test case for $p = 2$.



(a) Two-dimensional.

(b) Three-dimensional.

Figure 3.6: Unscaled (red) and scaled (green) condition numbers against η for the test case with a globally stabilised essential boundary condition along the trimmed boundary for $p = 2$.

numbers are significantly reduced, but certainly not independent of η. It is noticed that the slope in the scaled condition numbers is equal to the difference in the slope of the unscaled condition numbers of the globally stabilised essential boundary condition and the slope of the unscaled condition numbers of the natural and element-wise stabilised essential boundary condition. It is anticipated that this behaviour is due to the large stabilisation factor β for small η, which causes large differences in magnitude between the bulk and boundary terms. As mentioned in Section 2, β scales with $1/\widetilde{h} \approx \eta^{-1/d}$, which is exactly the slope in the scaled condition numbers. Verification of this hypothesis is a topic of further study.

4 Solving FCM systems using iterative solvers

Systems of equations can be solved by either direct or iterative methods. As mentioned in Section 2.6, iterative solvers are preferred as they generally only require $\mathcal{O}(n^2)$ flops and $\mathcal{O}(n)$ memory storage, opposed to direct solvers that typically require $\mathcal{O}(n^3)$ flops and $\mathcal{O}(n^2)$ memory storage for an $n \times n$ system [22–24]. However, the convergence speed of most iterative solvers depends on the condition number, hence they perform poorly on badly conditioned systems. The Conjugate Gradient (CG) method to solve symmetric positive definite matrices is one of the best known iterative solvers and has proven to be very effective [20, 28]. Because of that, in this work we will study the effectiveness of the CG method for solving FCM systems.

4.1 Conjugate Gradient method

For a symmetric positive definite (SPD) matrix $\mathbf{A}$, the bilinear operator

$$\left(\underline{x}, \underline{y}\right)_{\mathbf{A}} = \underline{x}^T \mathbf{A} \underline{y}, \tag{4.1}$$

is symmetric, bounded and coercive on $L^2\left(\mathbb{R}^n\right)$ and therefore forms an inner product that induces the norm $\|\cdot\|_{\mathbf{A}}$ for which it holds that

$$\lambda_{min} \|\underline{x}\|_2^2 \leq \|\underline{x}\|_{\mathbf{A}}^2 \leq \lambda_{max} \|\underline{x}\|_2^2, \tag{4.2}$$

with λ_{min} and λ_{max} the minimal and maximal eigenvalue of $\mathbf{A}$ respectively. It should be noted that when $\mathbf{A}$ is an FCM-matrix, it holds that $(\cdot,\cdot)_{\mathbf{A}} = k(*,*)$ and $\|\cdot\|_{\mathbf{A}} = \|*\|_k$. The Conjugate Gradient method (CG) solves SPD systems and is based on orthogonality with respect to $(\cdot,\cdot)_{\mathbf{A}}$. Suppose the system

$$\mathbf{A}\underline{x} = \underline{b}, \tag{4.3}$$

is to be solved iteratively using CG with the approximation and residual after i iterations denoted by $\underline{x}_i$ and $\underline{r}_i = \underline{b} - \mathbf{A}\underline{x}_i$ respectively. In every iteration a vector is created and added to a growing set of mutually perpendicular (w.r.t. $(\cdot,\cdot)_{\mathbf{A}}$) vectors $\{\underline{p}_1, \ldots, \underline{p}_i\}$, where $\underline{p}_i$ is created in the i^{th} iteration. The i^{th} approximation $\underline{x}_i$ is then the orthogonal projection of the solution onto the span of all vectors created so far $\langle \underline{p}_1, \ldots, \underline{p}_i \rangle$. Because $\underline{p}_i$ is orthogonal to $\langle \underline{p}_1, \ldots, \underline{p}_{i-1} \rangle$, this is the same as orthogonally projecting the solution onto $\underline{x}_{i-1} + \langle \underline{p}_i \rangle$. This reduces the computational cost significantly and causes every subsequent iteration to be equally expensive, which is the main reason for why CG is so efficient when compared to other projection methods where the computational cost is not linear with the rank of the projection space. To minimise the $\|\cdot\|_{\mathbf{A}}$-norm (and with that the FCM-norm) of the next approximation, the i^{th} projection direction $\underline{p}_i$ is based on the direction of the steepest descend

$$\nabla \|\underline{x} - \underline{x}_{i-1}\|_{\mathbf{A}}^2 \propto \mathbf{A}\left(\underline{x} - \underline{x}_{i-1}\right) = \underline{r}_{i-1}, \tag{4.4}$$

and then orthogonalised to $\langle \underline{p}_1, \dots, \underline{p}_{i-1} \rangle$. The vector $\underline{x} - \underline{x}_{i-1}$ is already orthogonal to $\langle \underline{p}_1, \dots, \underline{p}_{i-1} \rangle$ however, as x_{i-1} is the orthogonal projection of x onto $\langle \underline{p}_1, \dots, \underline{p}_{i-1} \rangle$

$$\underline{x} - \underline{x}_{i-1} \perp_{(\cdot,\cdot)_{\mathbf{A}}} \langle \underline{p}_1, \dots, \underline{p}_{i-1} \rangle. \tag{4.5}$$

Also

$$\begin{aligned}
\langle \underline{p}_1, \dots, \underline{p}_{i-1} \rangle &= \langle \underline{p}_1, \dots, \underline{p}_{i-2}, \underline{r}_{i-2} \rangle \\
&= \langle \underline{p}_1, \dots, \underline{p}_{i-2}, \underline{r}_{i-3} + \gamma \mathbf{A} \underline{p}_{i-2} \rangle && \text{for some } \gamma \\
&= \langle \underline{p}_1, \dots, \underline{p}_{i-2}, \mathbf{A} \underline{p}_{i-2} \rangle && \underline{r}_{i-3} \in \langle \underline{p}_1, \dots, \underline{p}_{i-2} \rangle \\
&= \langle \underline{p}_1, \mathbf{A} \underline{p}_1, \dots, \mathbf{A}^{i-2} \underline{p}_1 \rangle && \text{induction.}
\end{aligned} \tag{4.6}$$

Combining (4.5) and (4.6) yields

$$\begin{aligned}
\underline{x} - \underline{x}_{i-1} \perp_{(\cdot,\cdot)_{\mathbf{A}}} & \langle \underline{p}_1, \dots, \underline{p}_{i-1} \rangle \\
&= \langle \underline{p}_1, \mathbf{A} \underline{p}_1, \dots, \mathbf{A}^{i-2} \underline{p}_1 \rangle \\
&\supset \mathbf{A} \langle \underline{p}_1, \mathbf{A} \underline{p}_1, \dots, \mathbf{A}^{i-3} \underline{p}_1 \rangle \\
&= \mathbf{A} \langle \underline{p}_1, \dots, \underline{p}_{i-2} \rangle,
\end{aligned} \tag{4.7}$$

and because of the symmetry of $\mathbf{A}$

$$\underline{r}_{i-1} = \mathbf{A} \left(\underline{x} - \underline{x}_{i-1} \right) \perp_{(\cdot,\cdot)_{\mathbf{A}}} \langle \underline{p}_1, \dots, \underline{p}_{i-2} \rangle, \tag{4.8}$$

such that $\underline{r}_{i-1}$ only has to be orthogonalised to $\underline{p}_{i-1}$ to get $\underline{p}_i$, hence

$$\underline{p}_i = \underline{r}_{i-1} + \beta_{i-1} \underline{p}_{i-1}, \tag{4.9}$$

with

$$\begin{aligned}
\left(\underline{p}_i, \underline{p}_{i-1} \right)_{\mathbf{A}} &= \left(\underline{r}_{i-1} + \beta_{i-1} \underline{p}_{i-1}, \underline{p}_{i-1} \right)_{\mathbf{A}} = 0, \\
\Rightarrow \beta_{i-1} &= -\frac{\left(\underline{r}_{i-1}, \underline{p}_{i-1} \right)_{\mathbf{A}}}{\left(\underline{p}_{i-1}, \underline{p}_{i-1} \right)_{\mathbf{A}}} = -\frac{\left(\underline{r}_{i-1}, \underline{p}_{i-1} \right)_{\mathbf{A}}}{\left(\underline{r}_{i-2}, \underline{p}_{i-1} \right)_{\mathbf{A}}} && \underline{p}_{i-2} \perp_{(\cdot,\cdot)_{\mathbf{A}}} \underline{p}_{i-1} \\
&= -\frac{\left(\underline{r}_{i-1}, (\underline{x} - \underline{x}_{i-1}) - (\underline{x} - \underline{x}_{i-2}) \right)_{\mathbf{A}}}{\left(\underline{r}_{i-2}, (\underline{x} - \underline{x}_{i-1}) - (\underline{x} - \underline{x}_{i-2}) \right)_{\mathbf{A}}} \\
&= -\frac{\left(\underline{x} - \underline{x}_{i-1}, \underline{r}_{i-1} - \underline{r}_{i-2} \right)_{\mathbf{A}}}{\left(\underline{r}_{i-2}, (\underline{x} - \underline{x}_{i-1}) - (\underline{x} - \underline{x}_{i-2}) \right)_{\mathbf{A}}} && \mathbf{A} = \mathbf{A}^T \\
&= \frac{\left(\underline{x} - \underline{x}_{i-1}, \underline{r}_{i-1} \right)_{\mathbf{A}}}{\left(\underline{r}_{i-2}, \underline{x} - \underline{x}_{i-2} \right)_{\mathbf{A}}} = \frac{\underline{r}_{i-1}^T \underline{r}_{i-1}}{\underline{r}_{i-2}^T \underline{r}_{i-2}} && \underline{x} - \underline{x}_{i-1} \perp_{(\cdot,\cdot)_{\mathbf{A}}} \underline{r}_{i-2}.
\end{aligned} \tag{4.10}$$

The vector $\underline{x}$ is then orthogonally projected onto $\underline{x}_{i-1} + \langle \underline{p}_i \rangle$ such that

$$\underline{x}_i = \underline{x}_{i-1} + \alpha_i \underline{p}_i, \tag{4.11}$$

with

$$\begin{aligned}
\Big(\underline{x}-\underline{x}_i,\underline{p}_i\Big)_{\mathbf{A}} &= \Big(\underline{x}-\underline{x}_{i-1}-\alpha_i\underline{p}_i,\underline{p}_i\Big)_{\mathbf{A}} = 0,\\
\Rightarrow \alpha_i &= \frac{\Big(\underline{x}-\underline{x}_{i-1},\underline{p}_i\Big)_{\mathbf{A}}}{\Big(\underline{p}_i,\underline{p}_i\Big)_{\mathbf{A}}}\\
&= \frac{\Big(\underline{x}-\underline{x}_{i-1},\underline{r}_{i-1}+\beta_{i-1}\underline{p}_{i-1}\Big)_{\mathbf{A}}}{\Big(\underline{p}_i,\underline{p}_i\Big)_{\mathbf{A}}}\\
&= \frac{\big(\underline{x}-\underline{x}_{i-1},\underline{r}_{i-1}\big)_{\mathbf{A}}}{\Big(\underline{p}_i,\underline{p}_i\Big)_{\mathbf{A}}} = \frac{\underline{r}_{i-1}^T\underline{r}_{i-1}}{\underline{p}_i^T\mathbf{A}\underline{p}_i} \quad \underline{x}-\underline{x}_{i-1} \perp_{(\cdot,\cdot)_{\mathbf{A}}} \underline{p}_{i-1}.
\end{aligned} \tag{4.12}$$

After initialisation, CG reduces to the following algorithm.

```
**while** $\|\underline{r}\|_2 >$ tol **do**
    $\beta \leftarrow (\underline{r}^T\underline{r})/(\underline{s}^T\underline{s})$
    $\underline{p} \leftarrow \underline{r} + \beta\underline{p}$
    $\alpha \leftarrow (\underline{r}^T\underline{r})/(\underline{p}^T\mathbf{A}\underline{p})$
    $\underline{x} \leftarrow \underline{x} + \alpha\underline{p}$
    $\underline{s} \leftarrow \underline{r}$
    $\underline{r} \leftarrow \underline{b} - \mathbf{A}\underline{x}$
**end while**
```

4.2 Convergence properties of CG

In exact arithmetics, when $\underline{b}$ is built up of m eigenspaces of $\mathbf{A}$, CG converges in m iterations as in that case $\underline{x} \in \langle \underline{p}_1, \ldots, \underline{p}_m \rangle$ such that $\underline{x}_m = \underline{x}$. For the system $\mathbf{A}\underline{x} = \underline{b}$ with $\mathbf{A}$ an $n \times n$ matrix, $m \leq n$ and therefore CG ideally converges in at most n steps. In finite precision arithmetics, the orthogonality between distant projection vectors is lost however. Especially with large condition numbers, the ratio between the magnitudes of the separate eigenspaces present in the sequence $\mathbf{A}(\underline{x}-\underline{x}_0), \mathbf{A}^2(\underline{x}-\underline{x}_0), \ldots$, which is what the projection space is built-up of, rapidly grows larger than the machine precision. Because the method projects w.r.t. the $\|\cdot\|_{\mathbf{A}}$-norm, the method is still absolutely converging with respect to this norm (and with that to the FCM-norm), but generally requires more than m steps. As the $\|\cdot\|_{\mathbf{A}}$-norm is equivalent with the residual norm, the residual converges as well, but is not guaranteed to decrease in every subsequent iteration. As mentioned in Section 2.2, a convergence bound is given in [20]

$$\|\underline{x}-\underline{x}_i\|_{\mathbf{A}} \leq 2\left(\frac{\sqrt{\kappa_2(\mathbf{A})}-1}{\sqrt{\kappa_2(\mathbf{A})}+1}\right)^i \|\underline{x}-\underline{x}_0\|_{\mathbf{A}}, \tag{4.13}$$

which also indicates that CG converges faster for well-conditioned systems.

It should be noted that when solving a system, $\|\underline{x} - \underline{x}_i\|_{\mathbf{A}}$ is unknown and only the residual can be calculated. The equivalence between the residual and $\|\underline{x} - \underline{x}_i\|_{\mathbf{A}}$ depends on the condition number

$$\frac{\|\underline{b} - \mathbf{A}\underline{y}\|_2^2}{\|\mathbf{A}\|_2} \leq \|\underline{x} - \underline{x}_i\|_{\mathbf{A}}^2 \leq \kappa_2(\mathbf{A}) \frac{\|\underline{b} - \mathbf{A}\underline{y}\|_2^2}{\|\mathbf{A}\|_2}. \tag{4.14}$$

Therefore, for better-conditioned systems the residual is a better estimate for the $\| \cdot \|_{\mathbf{A}}$-norm, which is usually the quantity of interest as this is equal to the FCM-norm and equivalent to the $H^1(\Omega)$-norm.

4.3 Results

CG has been applied to the systems resulting from the two-dimensional experiment in Subsection 2.5, with $p = 2$ with a natural boundary condition and an element-wise stabilised essential boundary condition applied along the unmatching boundary. The solution that was approximated is $u = \arctan(y/x)$. Figures 4.1 and 4.3 show the convergence behaviour for systems that have large unscaled condition numbers and Figures 4.2 and 4.4 show the convergence behaviour for systems that have relatively small unscaled condition numbers. In these figures, it is visible that convergence is slow and that the $\| \cdot \|_{\mathbf{A}}$-norm and the residual norm are very loosely equivalent for large condition numbers. The iterations were stopped when the residual first became smaller than 10^{-12}, which supposedly indicated the system had converged. However, the figures show that for large condition numbers the system was clearly not yet converged in the $\| \cdot \|_{\mathbf{A}}$-norm. Figures 4.5 and 4.6 display the average convergence behaviour and standard deviation for 10000 systems in which the grid was vertically shifted along Γ_{fit} by a random variable taken from a uniform distribution in the range $[0, 1)$ multiplied by the grid size h. In these figures it is visible that not only the convergence is slower and the equivalence between the $\| \cdot \|_{\mathbf{A}}$-norm and the residual norm is looser for unscaled systems, but that also the variation is much larger, causing uncertainty in the quality of the solution of unscaled systems.

Solving systems very accurately is only relevant if the discretisation error is small as well. For the grid with $h = 1/4$ that is used here, depending on the shift, $\|u - u^h\|_k \approx 2 \cdot 10^{-2}$ for the natural boundary condition and $\|u - u^h\|_k \approx 3 \cdot 10^{-2}$ for the element-wise stabilised essential boundary condition, with u denoting the analytical solution and u^h the (fully converged) numerical solution. Therefore solving the linear system up to $\|\underline{x} - \underline{x}_i\|_{\mathbf{A}} \approx 10^{-5}$ for the natural boundary condition and $\|\underline{x} - \underline{x}_i\|_{\mathbf{A}} \approx 10^{-4}$ for the element-wise stabilised essential boundary condition -as is done for the unscaled system- would suffice and scaling is not required. When the grid is refined, however, the discretisation error decreases and at a certain moment the inaccuracy of the linear solver becomes the leading order error. As the inaccuracy of the linear solver is much larger for the unscaled system than for the scaled system, the scaled system can be solved up to a higher precision than the unscaled system. This is illustrated in Figure 4.7

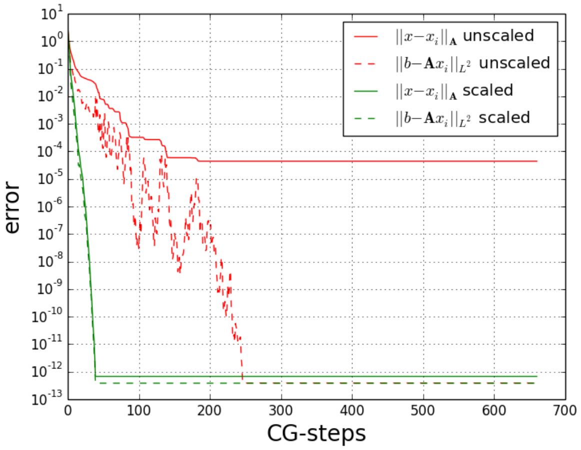


Figure 4.1: Convergence of CG for the test case with a natural boundary condition along the unmatching boundary. The grid is shifted $0.26h$, resulting in a condition number of $1.2 \cdot 10^{20}$ for the unscaled matrix and 22 for the scaled matrix.

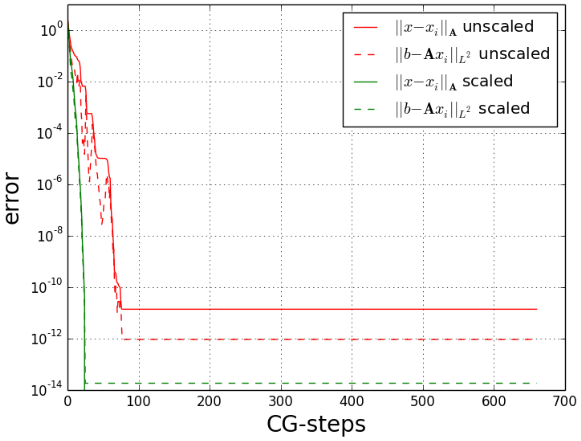


Figure 4.2: Convergence of CG for the test case with a natural boundary condition along the unmatching boundary. The grid is not shifted, resulting in a condition number of $3.9 \cdot 10^{7}$ for the unscaled matrix and 22 for the scaled matrix.

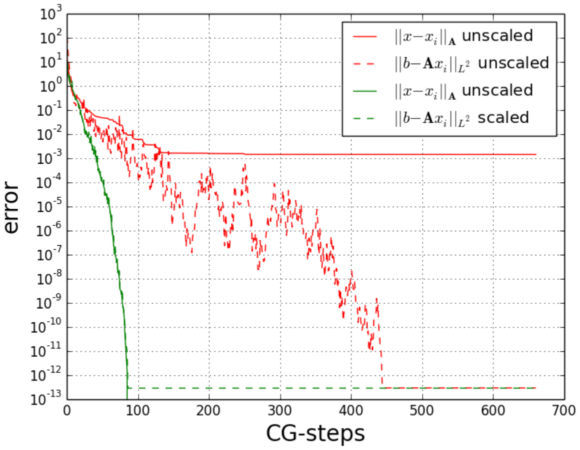


Figure 4.3: Convergence of CG for the test case with an essential boundary condition on the trimmed boundary. The grid is shifted $0.26h$, resulting in a condition number of $5.8 \cdot 10^{21}$ for the unscaled matrix and 373 for the scaled matrix.

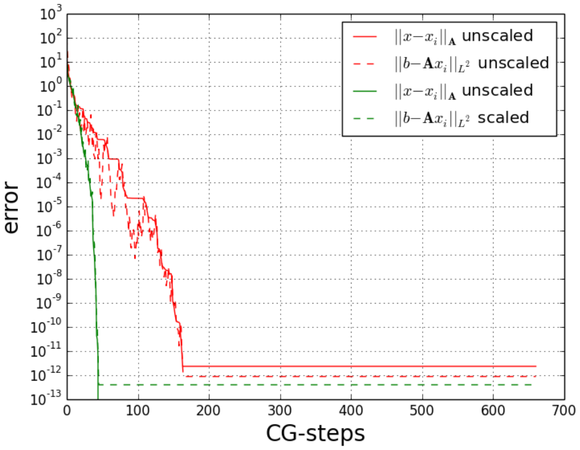
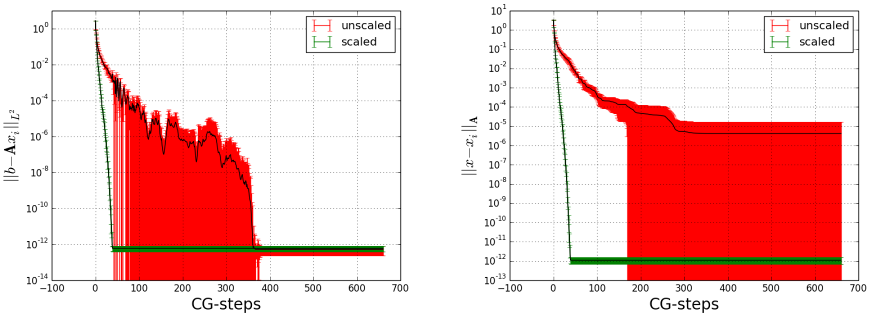


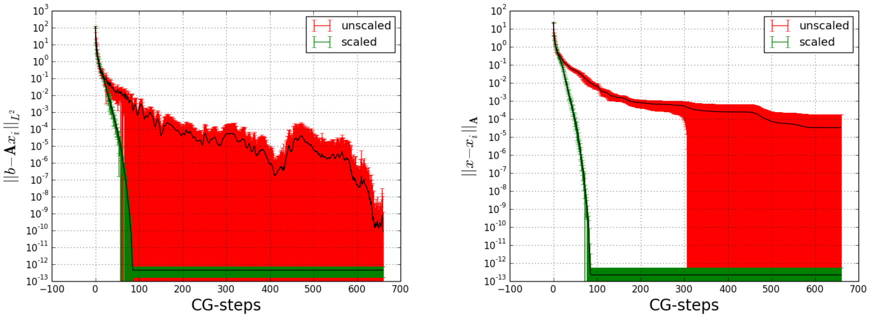
Figure 4.4: Convergence of CG for the test case with an essential boundary condition on the trimmed boundary. The grid is not shifted, resulting in a condition number of $1.0 \cdot 10^{7}$ for the unscaled matrix and 174 for the scaled matrix.



(a) $\|\underline{b} - \mathbf{A}\underline{x}_i\|_2$

(b) $\|\underline{x} - \underline{x}_i\|_{\mathbf{A}}$

Figure 4.5: Average convergence of CG for 10000 randomly shifted test cases with a natural boundary condition on the trimmed boundary.



(a) $\|\underline{b} - \mathbf{A}\underline{x}_i\|_2$

(b) $\|\underline{x} - \underline{x}_i\|_{\mathbf{A}}$

Figure 4.6: Average convergence of CG for 10000 randomly shifted test cases with an essential boundary condition on the trimmed boundary.

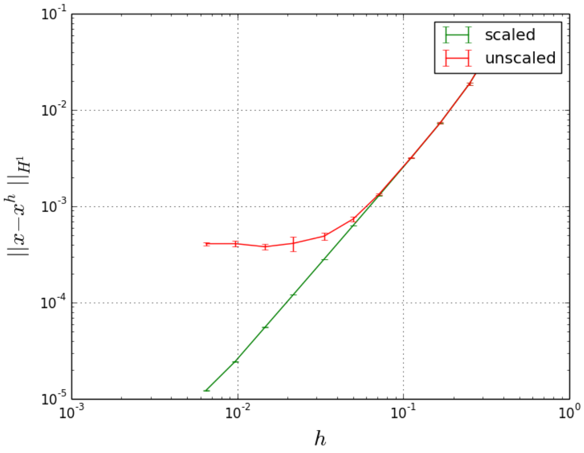


Figure 4.7: Convergence plot under grid refinement, where the iterative solver is stopped when the residual error first becomes smaller than 10^{-7} for the natural boundary condition.

for the natural boundary condition and in Figure 4.8 for the element-wise stabilised essential boundary condition. These figures display the average error and standard deviation of 10 random shifts in the H^1-norm, as the FCM-norm $\|\cdot\|_k$ is grid-dependent. It should be mentioned that with a tolerance of 10^{-12} as in the previous experiment, one needs an extremely fine grid to show this effect. Therefore the tolerance in this experiment is set to 10^{-7}. The straight line of slope 2 ($= p$) in Figure 4.7, depicting the discretisation error before the inaccuracy of the linear solver becomes dominant, is in agreement with Cea's lemma [22]. The straight line of the same slope in Figure 4.8 indicates the same convergence rate, but there is no mathematical basis for this as the used operator is grid-dependent and therefore Cea's lemma does not apply here. The slight divergence of the scaled system for the smallest grids can be explained by the dependence of the scaled condition number $\kappa_2(\mathbf{DKD})$ on the grid size h, as demonstrated in Subsection 5.2.2.

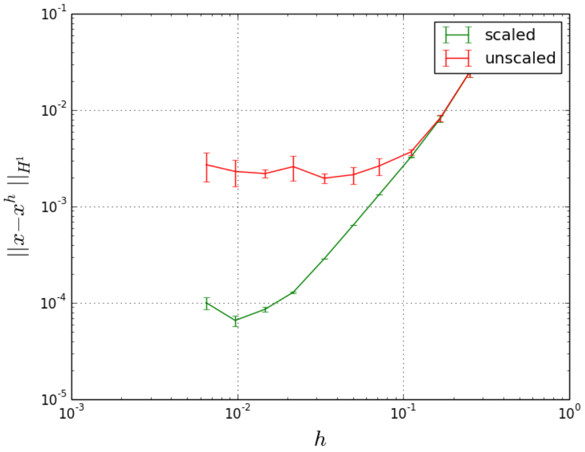
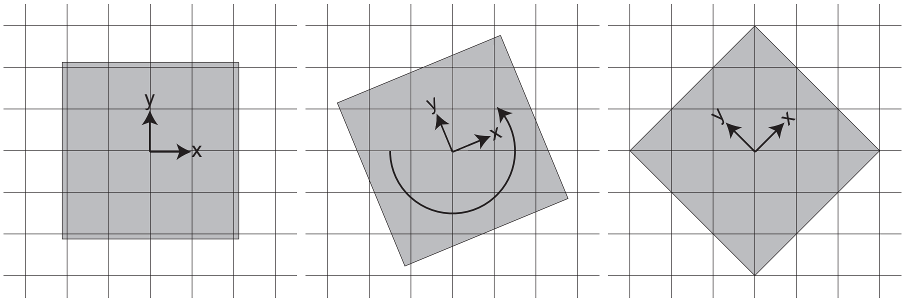


Figure 4.8: Convergence plot under grid refinement, where the iterative solver is stopped when the residual error first becomes smaller than 10^{-7} for the elementwise stabilised essential boundary condition.



(a) Original position. (b) Intermediate position. (c) Final position.

Figure 5.1: Schematic representation of the setup with the centre of the square located at the corner of an element.

5 Numerical examples

In this section, Laplace's problem and a linear elasticity problem are solved on grids at, relative to the domain, different positions and orientations to further study the effect of basis function scaling on the condition number of the Finite Cell Method. Also, the effect of grid refinement in combination with scaling on the condition number is studied.

5.1 Laplace's problem on a rotated domain

In the first example, Laplace's problem is solved on a squared domain of length 1.06. This domain is placed on a uniform, rectilinear grid with grid size $h = 1/4$ and spline order $p = 2$, with the centre of the square located at the corner of an element. The square is then rotated over 45° in 1000 steps, to create 1001 different discretisations. Essential boundary conditions are stabilised element-wise. A schematic representation of the setup is displayed in Figure 5.1.

The problem that is solved is

$$\begin{cases} -\Delta u = \cos(x) - \cos(y) & \text{in } \Omega, \\ u = \cos(x) - \cos(y) & \text{on } \partial\Omega, \end{cases} \tag{5.1}$$

for $\Omega = (-0.53, 0.53)^2$, with analytical solution

$$u = \cos(x) - \cos(y), \tag{5.2}$$

which is displayed in Figure 5.2. The discretisation errors and the condition numbers for the different discretisations are displayed in Figure 5.3 and Figure 5.4 respectively.

In Figure 5.3a, the quality of the approximation is practically independent of the smallest volume fraction η and in Figure 5.3b it is visible that the quality

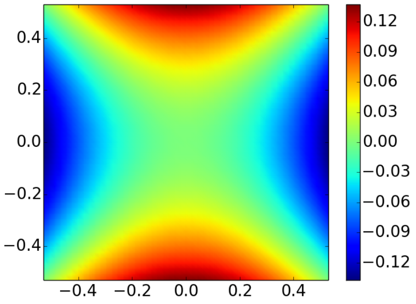
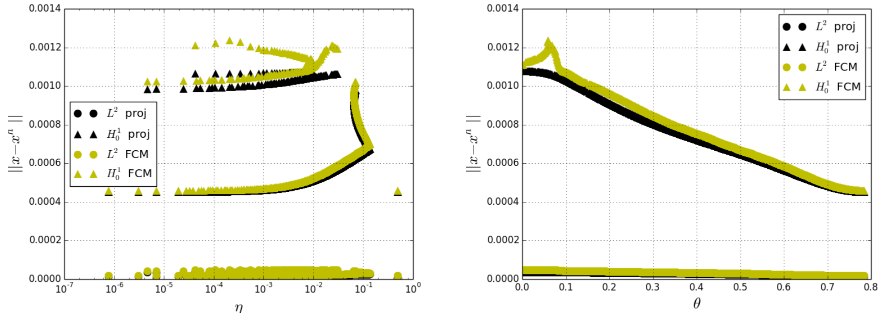


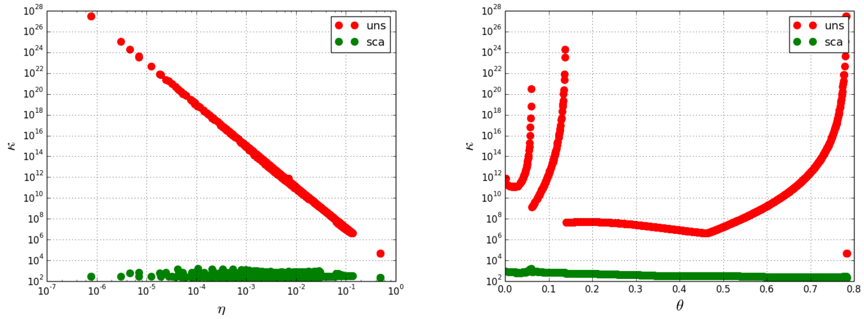
Figure 5.2: Analytical solution of Problem (5.1).



(a) Plotted against η.

(b) Plotted against θ.

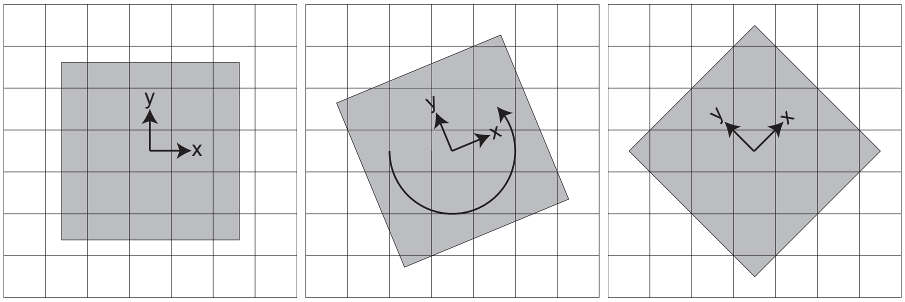
Figure 5.3: Discretisation errors in the H_0^1- and L^2-norm for the domain with the centre of the square located at the corner of an element. The yellow markers display the discretisation errors of the FCM-solution and the black markers display the discretisation errors of projections onto the used function space.



(a) Plotted against η.

(b) Plotted against θ.

Figure 5.4: Unscaled (red) and scaled (green) condition numbers for the domain with the centre of the square located at the corner of an element.



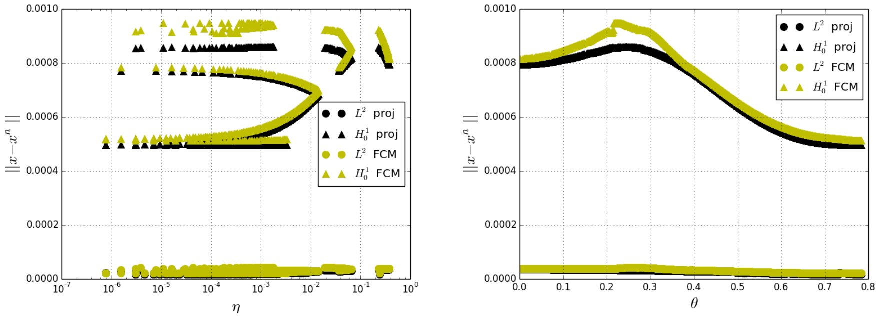
(a) Original position. (b) Intermediate position. (c) Final position.

Figure 5.5: Schematic representation of the setup with the centre of the square located at the centre of an element.

of the approximation has a slight dependence on θ. This dependence is a consequence of the fact that the discrete trial space is orientation dependent. This dependence is demonstrated by the black markers that display the discretisation error of the best possible approximation which is obtained by projecting the analytical solution onto the function space in H^1_0 and L^2 sense. In Figure 5.4a, the unscaled condition number shows a similar dependence on η as derived in Subsection 2.4 and observed in Subsection 2.5. The scaled condition numbers show no dependence on η, as was also observed in Subsection 3.2.

The same experiment was conducted with the centre of the square located at the centre of an element. A schematic representation of this setup is displayed in Figure 5.5. The discretisation errors are visible in Figure 5.6 and the condition numbers for the different discretisations are displayed in Figure 5.7.

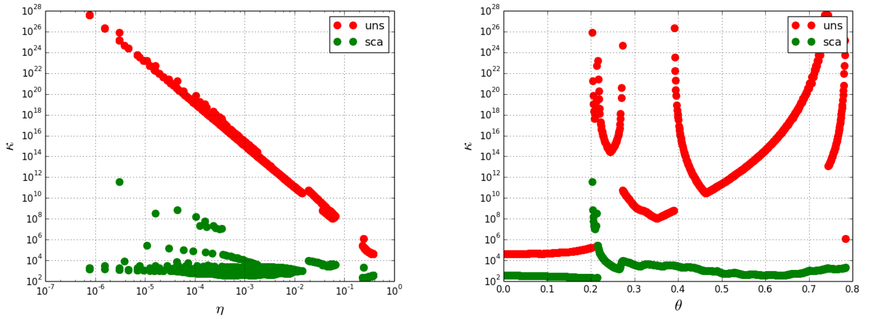
In Figure 5.6a, it is observed that the quality of the approximation is practically independent of η again and in Figure 5.6b a similar slight dependence of the quality of the approximation on θ as in Figure 5.3b is visible. In Figure 5.7a, the scaled condition number shows a very different behaviour as was observed in Figure 5.4a and Subsection 3.2 however. The scaled condition numbers are still significantly lower than the unscaled condition numbers, but for certain values of θ the scaled condition numbers are much larger than the scaled condition numbers in Figure 5.4 and additionally they are observed to be dependent on η. This behaviour has been studied by comparing the discretisations for which the scaled condition number was relatively large to discretisations with small values of η and small scaled condition numbers. It was observed that condition numbers of discretisations with small values of η caused by cut-off elements as in Figure 5.8a -which were analysed in Subsection 2.4 and present in Subsection 3.2- are reduced by scaling to significantly lower values than the condition numbers of discretisations with small values of η caused by cut-off elements as in Figure 5.8b. Also the scaled condition numbers of discretisations with small



(a) Plotted against η.

(b) Plotted against θ.

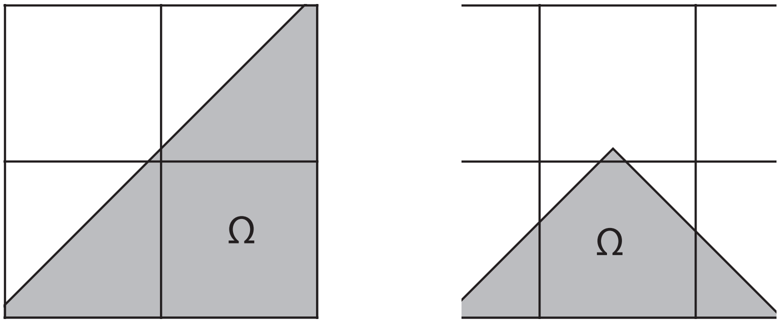
Figure 5.6: Discretisation errors in the H_0^1- and L^2-norm for the domain with the centre of the square located at the centre of an element. The yellow markers display the discretisation errors of the FCM-solution and the black markers display the discretisation errors of the projections onto the used function space.



(a) Plotted against η.

(b) Plotted against θ.

Figure 5.7: Unscaled (red) and scaled (green) condition numbers for the domain with the centre of the square located at the centre of an element.



(a) Small cut-off element for which only one function is only supported on the very small domain.

(b) Small cut-off element for which more than one functions are only supported on the very small domain.

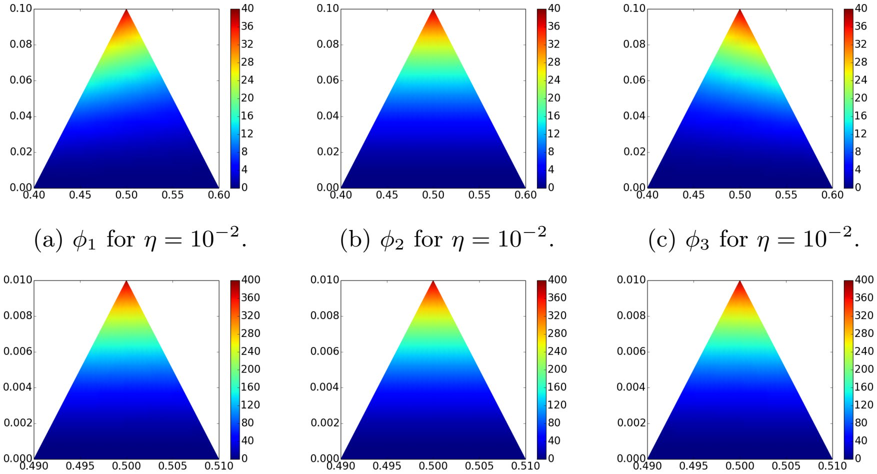
Figure 5.8: Schematic representation of two differently cut-off elements that may cause a small volume fraction η.

cut-off elements as in Figure 5.8a were practically independent of η, as opposed to condition numbers of discretisations with small cut-off elements as in Figure 5.8b, that showed a strong dependence on η.

This behaviour can be explained by the number of basis functions that are only supported on a small cut-off element. For small cut-off elements as in Figure 5.8a, only one basis function is only supported on this small cut-off element, and after scaling this results in a condition number independent of η. (For Lagrange-elements of order $p > 1$ this does not hold, as with these elements there are more than one functions that are only supported on this small cut-off element, such that a situation similar to Figure 5.8b occurs.) For small cut-off elements of the form displayed in Figure 5.8b, $p+1$ basis functions are only supported on the small cut-off element. On the intersection of their support and Ω, these functions are much more dependent on the vertical coordinate than on the horizontal coordinate however, with vertical and horizontal coordinates relative to the orientation in Figure 5.8b. Figure 5.9 shows for $p = 2$ that for the relatively large value of $\eta = 10^{-2}$ the basis functions only show a very slight dependence on the horizontal coordinate, and that for $\eta = 10^{-4}$ the functions are virtually independent of the horizontal coordinate. This causes these functions to become almost linearly-dependent, from here on referred to as quasi linearly-dependent, which is not repaired by scaling and therefore the condition number is reduced significantly less.

To verify wether the quasi linear-dependence is what limits the performance of the basis function scaling, a linear-dependentness index

$$\chi = \frac{1+\alpha}{1-\alpha}, \tag{5.3}$$



(a) ϕ_1 for $\eta = 10^{-2}$.

(b) ϕ_2 for $\eta = 10^{-2}$.

(c) ϕ_3 for $\eta = 10^{-2}$.

(d) ϕ_1 for $\eta = 10^{-4}$.

(e) ϕ_2 for $\eta = 10^{-4}$.

(f) ϕ_3 for $\eta = 10^{-4}$.

Figure 5.9: The functions that are only supported on a small cut-off element as in Figure 5.8b for $p = 2$ with $\eta \in \{10^{-2}, 10^{-4}\}$, scaled with the L^2-norm. The axes are relative to the grid size h, such that the full element is $(0, 1)^2$.

is introduced, with

$$\alpha = \max_{i,j} \frac{\left|(\phi_i, \phi_j)_k\right|}{\|\phi_i\|_k \|\phi_j\|_k}. \tag{5.4}$$

From the Cauchy-Schwarz inequality it is known that $0 \leq \alpha \leq 1$, with $\alpha = 0$ indicating that all functions are orthogonal and $\alpha = 1$ indicating that there is a pair of linearly-dependent functions. The linear-dependentness index χ is formulated as in (5.3) because this would be the condition number of a scaled system in which, except for the two functions that yield the maximum in (5.4), all other functions are orthogonal. Therefore $\alpha = 0$ yields $\chi = 1$ as a scaled system of orthogonal functions gives the identity matrix and $\alpha = 1$ yields $\chi = \infty$ as a system with a set of linearly-dependent functions is singular. The scaled condition numbers from Figure 5.7b are plotted together with χ in Figure 5.10. In this figure χ shows a similar behaviour and peaks at the same value of θ as the scaled condition number, which strongly indicates that it is the quasi linear-dependence that limits the performance of the basis function scaling. Furthermore it is visible that χ is smaller than the scaled condition numbers, which is because the functions that yield the maximum in (5.4) are not orthogonal to all other functions as assumed in (5.3).

A possible resolution to this problem is the orthonormalisation of these functions by a Gram-Schmidt procedure, as displayed in Figure 5.11 for the functions from Figure 5.9. This is a simple operation as only a very limited number of functions is involved and the inner products between these are known. Also it

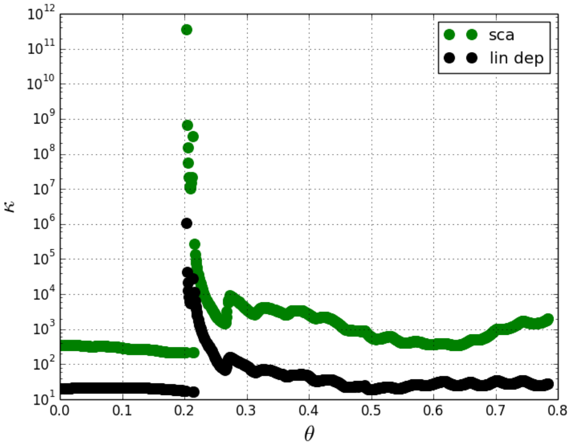


Figure 5.10: Scaled condition numbers (green) and linear-dependentness index χ (black) plotted against θ.

can be implemented using a similar pre- and post-conditioner, which in this case is not diagonal but very sparse lower triangular, such that the system becomes

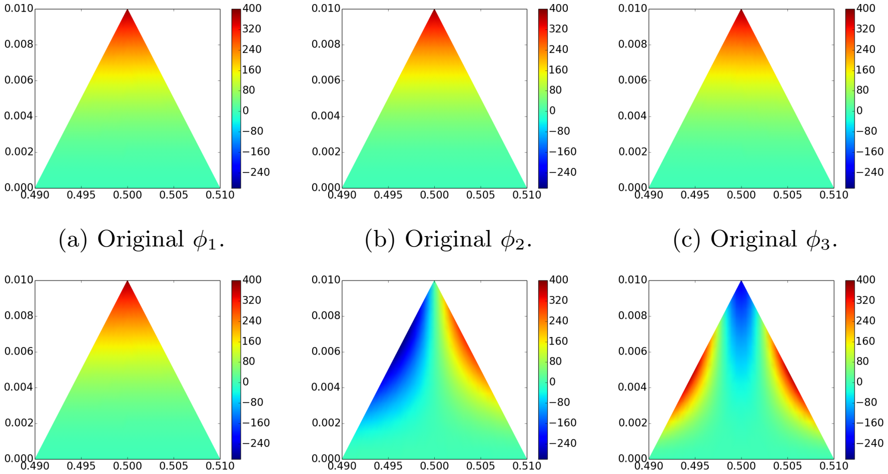
$$\begin{aligned} \mathbf{DKD}^T \underline{y} &= \mathbf{D}\underline{l}, \\ \underline{u} &= \mathbf{D}^T \underline{y}. \end{aligned} \tag{5.5}$$

Because the support of the orthonormalised functions is unchanged, the sparsity pattern of the conditioned system does not change either, such that an iteration in the solving procedure is equally expensive.

The orthonormalisation was tested by applying the Gram-Schmidt procedure to the functions with $\chi > 10$. In Figure 5.12, the resulting condition numbers are plotted for Laplace's problem on the rotating domain with the centre of the square located at the centre of an element. This shows a very promising result, but as this only involves one example, more research is required to verify the effectiveness of this method and to study at what value of χ the orthonormalisation should be applied. Furthermore there are also other possible resolutions to the problem of the quasi linearly-dependent functions, such as replacing these functions by an averaged function -which will have a negligible effect on the accuracy as only a fraction of the domain is involved- and locally refining the grid.

5.2 Linear elastic deformation of a plate with a circular exclusion

In the second example a linear elasticity problem is solved on a square of size 1 with a circular exclusion of radius $R = 1/4$ at the corner. A schematic repre-



(a) Original ϕ_1.

(b) Original ϕ_2.

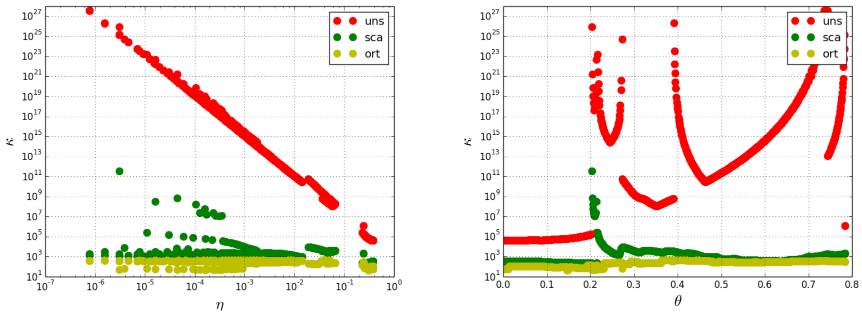
(c) Original ϕ_3.

(d) Orthogonalised $\tilde{\phi}_1$.

(e) Orthogonalised $\tilde{\phi}_2$.

(f) Orthogonalised $\tilde{\phi}_3$.

Figure 5.11: The original and orthonormalised (w.r.t. the L^2-norm) functions that are only supported on a small cut-off element as in Figure 5.8b for $p = 2$ with $\eta = 10^{-4}$, scaled with the L^2-norm. The axes are relative to the grid size h, such that the full element is $(0, 1)^2$.



(a) Plotted against η.

(b) Plotted against θ.

Figure 5.12: Unscaled (red), scaled (green) and scaled *and* orthonormalised (yellow) condition numbers for the domain with the centre of the square located at the centre of an element.

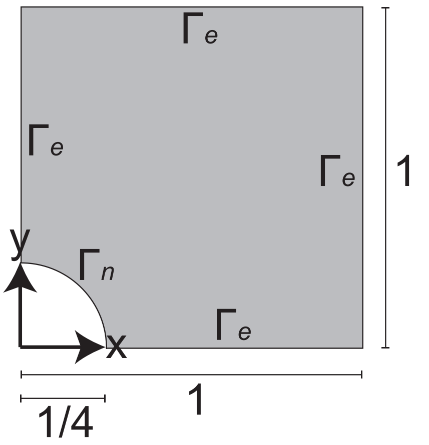


Figure 5.13: Schematic representation of the domain with a circular exclusion.

sentation of the domain is displayed in Figure 5.13.

An isotropic elasticity model is applied using Lamé's first parameter λ and Lamé's second parameter or shear modulus μ. The problem that is solved is set such that the solution coincides with the analytical solution for an infinite plate with a circular exclusion of radius R at the origin under axial tension T in the x direction

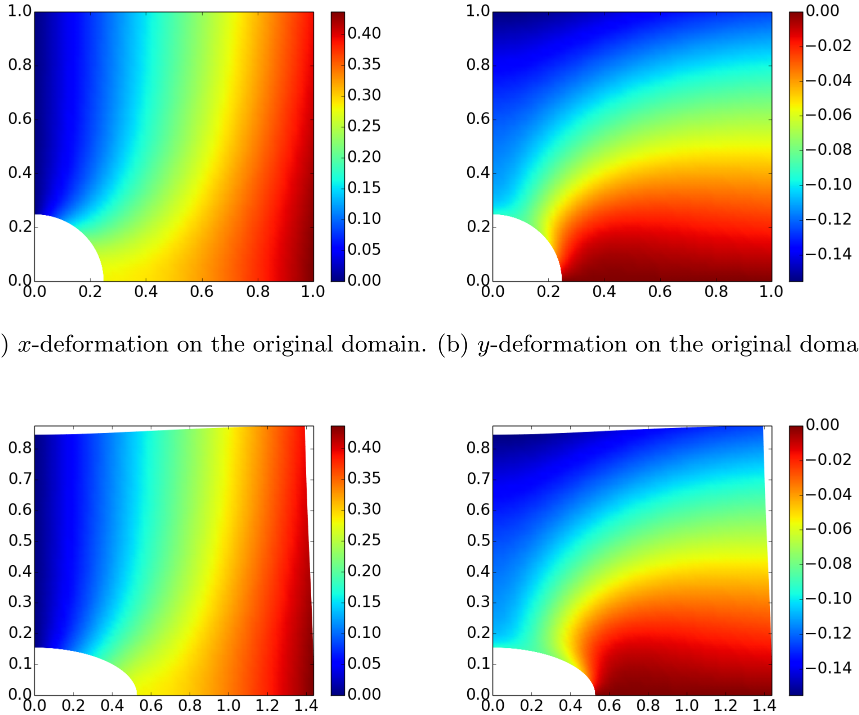
$$\begin{aligned} u_x &= \frac{Tx}{\mu}\left(\frac{(2\mu+\lambda)r^2+(\mu-\lambda)R^2}{4(\mu+\lambda)r^2}+\frac{\frac{3}{4}R^4+x^2R^2}{r^4}-\frac{x^2R^4}{r^6}\right), \\ u_y &= \frac{Ty}{\mu}\left(\frac{-\lambda r^2+(\mu+3\lambda)R^2}{4(\mu+\lambda)r^2}-\frac{\frac{3}{4}R^4+y^2R^2}{r^4}+\frac{y^2R^4}{r^6}\right), \end{aligned} \tag{5.6}$$

with $r=\sqrt{x^2+y^2}$ denoting the distance from the origin. The analytical solution for $T=1$, $\lambda=1$ and $\mu=1$, which are used in this example, is displayed in Figure 5.14.

The problem is stated as

$$\left\{ \begin{aligned} -\mathrm{div}(\sigma(u)) &= 0 \quad \text{in } \Omega, \\ n\sigma(u) &= 0 \quad \text{on } \Gamma_n, \\ u &= \overline{u} \quad \text{on } \Gamma_e, \end{aligned} \right. \tag{5.7}$$

with $\Omega=(0,1)^2\backslash B_{0,R}$, $\Gamma_n=\partial B_{0,R}\cap(0,1)^2$ and $\Gamma_e=\partial(0,1)^2\backslash B_{0,R}$ as in Figure 5.15, with $\overline{u}$ as in (5.6) and with $\sigma(u)=\left(\lambda\mathbf{II}+2\mu{}^4\mathbf{I}^s\right):\nabla u$ where $\mathbf{I}$ and ${}^4\mathbf{I}^s$ denote the second-order identity tensor and symmetric fourth-order identity tensor respectively.



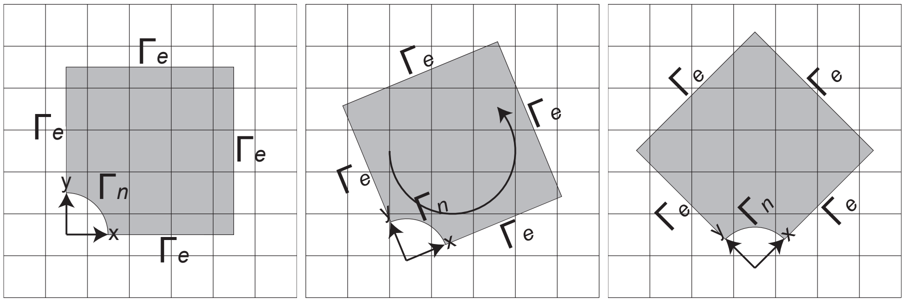
(a) x-deformation on the original domain.

(b) y-deformation on the original domain.

(c) x-deformation on the deformed domain.

(d) y-deformation on the deformed domain.

Figure 5.14: Analytical solution.



(a) Original position. (b) Intermediate position. (c) Final position.

Figure 5.15: Schematic representation of the domain with a circular exclusion rotated on the grid.

5.2.1 Grid rotation

Similar to the previous example, the domain is placed on a uniform, rectilinear grid with grid size $h = 1/4$ and spline order $p = 2$, with the centre of the square located at the centre of an element. The square is then rotated over 45° in 1000 steps, to create 1001 different discretisations. A schematic representation of the setup is displayed in Figure 5.15.

Essential boundary conditions are stabilised element-wise by separate stabilisation terms for the first and second Lamé parameter

$$\begin{aligned} c_1(u,v) &= \int_{\Gamma_e} \beta_1 (n\cdot u)(n\cdot v)\mathrm{d}\Gamma, \\ c_2(u,v) &= \int_{\Gamma_e} \beta_2 u\cdot v \mathrm{d}\Gamma, \end{aligned} \tag{5.8}$$

with

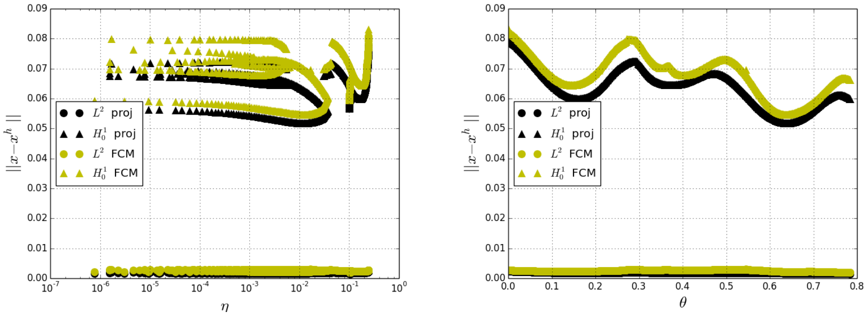
$$\begin{aligned} \beta_1 &= 2\lambda C_1, \\ \beta_2 &= 4\mu C_2, \end{aligned} \tag{5.9}$$

for C_1 and C_2 satisfying

$$\begin{aligned} \int_{\Gamma_e} \mathrm{tr}(\nabla v)^2 \mathrm{d}\Gamma &\leq \int_{\Omega} C_1 \mathrm{tr}(\nabla v)^2 \mathrm{d}\Omega, \\ \int_{\Gamma_e} (n\nabla^s v)^2 \mathrm{d}\Gamma &\leq \int_{\Omega} C_2 (\nabla^s v)^2 \mathrm{d}\Omega. \end{aligned} \tag{5.10}$$

Details about these stabilisation parameters can be found in [8]. The discretisation errors and the condition numbers for the different discretisations are displayed in Figure 5.16 and Figure 5.17 respectively.

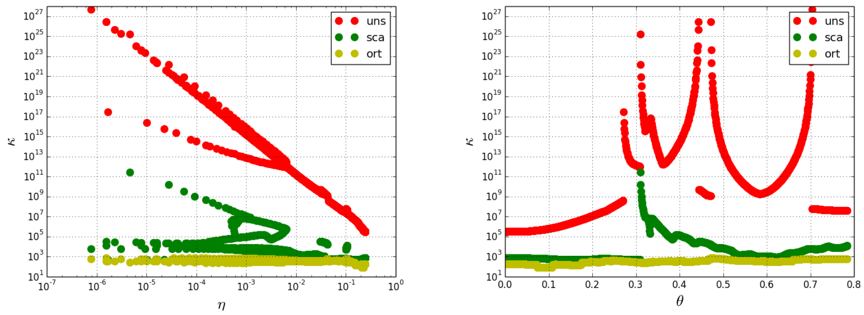
In Figure 5.16a it is observed that the quality of the approximation is practically independent of η and in Figure 5.16b a slight dependence on θ can be



(a) Plotted against η.

(b) Plotted against θ.

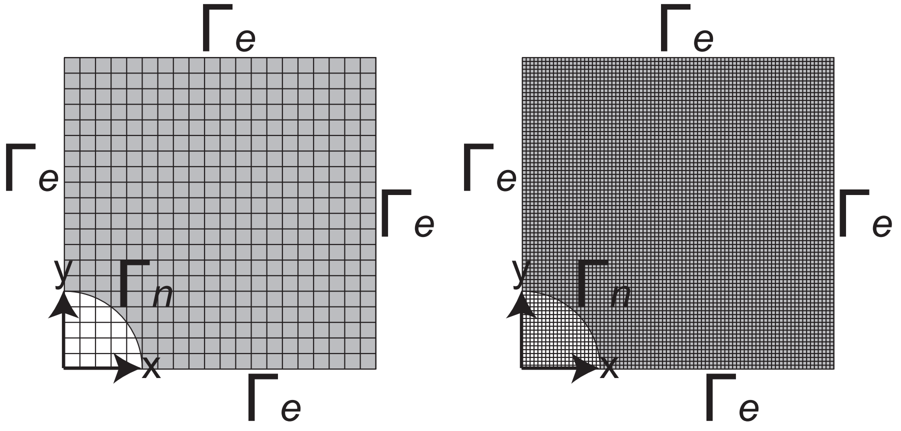
Figure 5.16: Discretisation errors in the H_0^1- and L^2-norm. The yellow markers display the discretisation errors of the FCM-solution and the black markers display the discretisation errors of projections onto the used function space.



(a) Plotted against η.

(b) Plotted against θ.

Figure 5.17: Unscaled (red), scaled (green) and scaled *and* orthonormalised (yellow) condition numbers.



(a) Grid with $h = 1/20$.

(b) Grid with $h = 1/80$.

Figure 5.18: Schematic representation of uniform grid refinement on the domain with a circular exclusion.

noticed, as was also observed in the example with Laplace's problem. Figure 5.17 shows that for certain values of θ, the condition numbers behave as derived in Subsection 2.4 and observed in Subsection 2.5 and Subsection 3.2, but for certain other values of θ a different behaviour is observed. For these other values of θ, the unscaled condition numbers show a dependence on η in a different slope than predicted by (2.38) and the scaled condition numbers are relatively large and also show a dependence on η. Studying the discretisations for which this unexpected behaviour occurred showed that at these values of θ small cut-off elements as in Figure 5.8b were present. Therefore this behaviour can be explained in a similar fashion as the effect observed in Figure 5.7 and is again resolved by the orthonormalisation procedure, which was applied for functions with a value of $\chi > 10$.

5.2.2 Grid refinement

Problem (5.7) was also solved on uniform, rectilinear grids of different grid sizes that match the essential boundaries, but do not match the natural boundary, with a spline order of $p = 2$. The essential boundary conditions in this example are applied in a strong manner, such that $b(\cdot, \cdot) = c(\cdot, \cdot) = 0$. A schematic representation of the setup is displayed in Figure 5.18. It should be noted that cut-off elements of the shape as displayed in Figure 5.8b are not possible with these discretisations, and therefore the orthonormalisation procedure is not applied here. The behaviour of the unscaled and scaled condition number as well as the discretisation error for the different grid sizes was studied. The discretisation errors and the condition numbers for the different discretisations are displayed in Figure 5.19 and Figure 5.20 respectively.

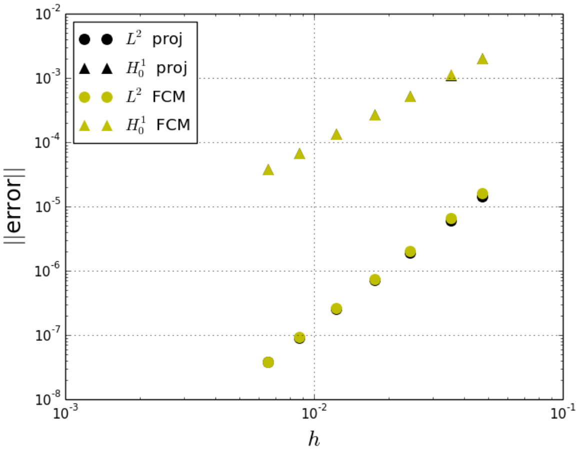
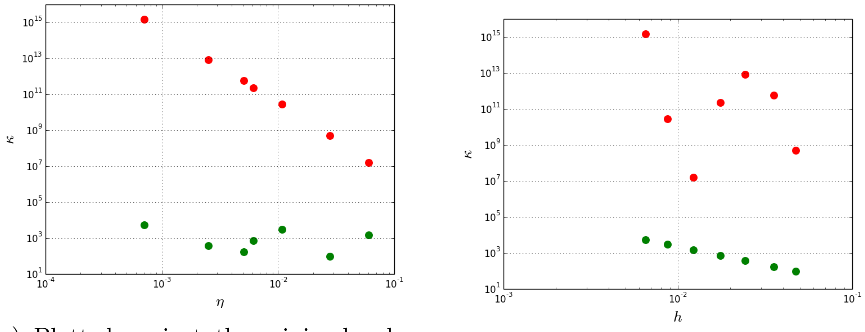


Figure 5.19: Discretisation errors in the H_0^1- and L^2-norm under grid refinement. The yellow markers display the discretisation errors of the FCM-solution and the black markers display the discretisation errors of projections onto the used function space.



(a) Plotted against the minimal volume fraction η.

(b) Plotted against the grid size h.

Figure 5.20: Unscaled (red) and scaled (green) condition numbers.

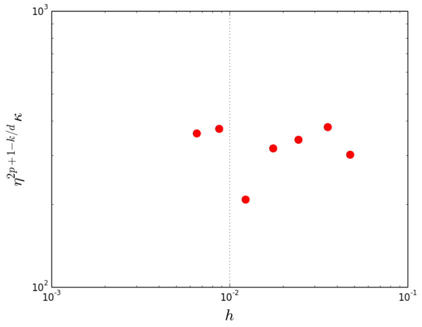


Figure 5.21: Unscaled $\eta^{2p+1-k/d}\kappa_2(\mathbf{K})$ plotted against h.

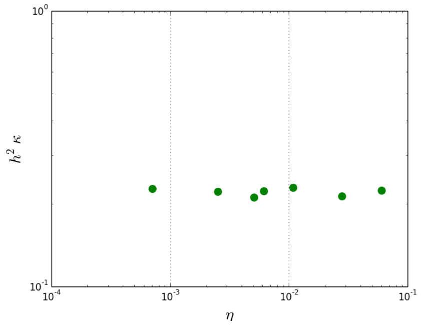


Figure 5.22: Scaled $h^2\kappa_2(\mathbf{DKD})$ plotted against η.

The straight lines of slope 2 ($= p$) for the H_0^1-error and slope 3 ($= p+1$) for the L^2-error in Figure 5.19 are in agreement with Cea's lemma [22], which applies here as $b(\cdot,\cdot) = c(\cdot,\cdot) = 0$ such that the FCM operator is independent of the grid. It should be noted that reducing the condition number by scaling is important to attain these levels of accuracy, as was demonstrated in Figures 4.7 and 4.8. Figure 5.20a shows that the unscaled condition numbers behave as derived in Subsection 2.4 and observed in Subsection 2.5. In Figure 5.20b it is visible that the scaled condition numbers are proportional to $1/h^2$, and thus are observed to exhibit the same behaviour as the condition number for standard finite elements for second-order elliptic PDE's that is also proportional to $1/h^2$, caused by the eigenvalues of the periodic functions with the longest and the shortest wavelength captured by the grid [22]. The dependence of the unscaled condition number on η is displayed more clearly in Figure 5.21, in which $\kappa_2(\mathbf{K})\eta^{2p+1-2/d}$ is plotted against h to show the relation $\kappa_2(\mathbf{K}) \approx C\eta^{-(2p+1-2/d)}$ for some constant C as predicted by (2.38). Figure 5.22 shows $\kappa_2(\mathbf{DKD}) \approx C/h^2$ for some different constant C by plotting $\kappa_2(\mathbf{DKD})h^2$ against η, which depicts a constant value.

6 Conclusion and recommendations

On the basis of the results presented in this work, it is concluded that the conditioning problems that the Finite Cell Method is prone to can be resolved by diagonal pre- and post-conditioning. From the estimate of the condition number derived in Subsection 2.4 and the experimental verification in Subsection 2.5, we can conclude that very small functions -which can occur when elements are cut- are the main cause of the conditioning problems associated with FCM. The proposed method to improve the conditioning by diagonal scaling has rigorously reduced all condition numbers in Subsection 3.2 and Section 5 and has drastically improved the convergence speed and reliability of the iterative solver in Subsection 4.3. Scaled systems do not show the discretisation-order dependent sensitivity of the condition number to the orientation of the grid. Therefore the simplicity at which higher order discretisations can be implemented in the Finite Cell Method can be exploited further.

FCM conditioning can be further improved by resolving the problem of quasi linearly-dependent functions encountered in Subsection 5.1. The proposed method of orthonormalisation shows very promising preliminary results, but more research is required to verify the effectiveness of this procedure.

A major challenge lies in generalising the scaling procedure for mixed methods, such that it can be applied to problems in the field of computational fluid dynamics (CFD). Applying FCM to CFD -where it may even show more advantages than in solid mechanics because of moving domains that generally require remeshing- will increase the range of applicability of the method. Also the simulation of flows in porous media can to a great extend be simplified by the application of the Finite Cell Method.

References

[1] J.A. Cottrell, T.J.R. Hughes, and Y. Basilevs. *Isogeometric Analysis.* Wiley, 1st edition, 2009.

[2] M. Ruess, D. Schillinger, A.I. Özkan, and E. Rank. Weak coupling for isogeometric analysis of non-matching and trimmed multi-patch geometries. *Computational methods in applied mechanics and engineering*, 269:46–71, 2014.

[3] C.V. Verhoosel, G.J. van Zwieten, B. van Rietbergen, and R. de Borst. Image-based goal-oriented adaptive isogeometric analysis with application to the micro-mechanical modeling of trabecular bone. *Computer Methods in Applied Mechanics and Engineering*, 284:138–164, 2015.

[4] J. Parvizian, A. Düster, and E. Rank. Finite cell method – h- and p-extension for embedded domain problems in solid mechanics. *Computational Mechanics*, 41:121–133, 2007.

[5] A. Embar, J. Dolbow, and I. Harari. Imposing dirichlet boundary conditions with nitshce's method and spline based finite elements. *International Journal for Numerical Methods in Engineering*, 83:877–898, 2010.

[6] S. Fernández-Méndez and A. Huerta. Imposing essential boundary conditions in mesh-free methods. *Computational methods in applied mechanics and engineering*, 193:1257–1275, 2004.

[7] M. Ruess, D. Tal, N. Trabelsi, Z. Yosibash, and E. Rank. The finite cell method for bone simulations: verification and validation. *Biomechanics and Modeling in Mechanobiology*, 11:425–437, 2012.

[8] M. Ruess, D. Schillinger, Y. Bazilevs, V. Varduhn, and E. Rank. Weakly enforced essential boundary conditions for nurbs-embedded and trimmed nurbs geometries on the basis of the finite cell methodod. *International Journal for Numerical Methods in Engineering*, 95:811–846, 2013.

[9] E. Rank, S. Kollmannsberger, C. Sorger, and A. Düster. Shell finite cell method: A high order fictious domain approach for thin-walled structures. *Computational methods in applied mechanics and engineering*, 200:3200–3209, 2011.

[10] E. Rank, M. Ruess, S. Kollmannsberger, D. Schillinger, and A. Düster. Geometric modeling, isogeometric analysis and the finite cell method. *Computational methods in applied mechanics and engineering*, 249-252:104–115, 2012.

[11] M. Dauge, A. Düster, and E. Rank. Theoretical and numerical investigation of the finite cell method. 2013.

[12] D. Schillinger, M. Ruess, N. Zander, Y. Bazilevs, A. Düster, and E. Rank. Small and large deformation analysis with the *p*- and b-spline versions of the finite cell method. *Computational Mechanics*, 50:445–478, 2012.

[13] D. Schillinger, A. Düster, and E. Rank. The *hp-d*-adaptive finite cell method for geometrically nonlinear problems of solid mechanics. *International Journal for Numerical Methods in Engineering*, 89:1171–1202, 2012.

[14] M. Joulaian and A. Düster. Local enrichment of finite cell method for problems with material interfaces. *Computational Mechanics*, 52:741–762, 2013.

[15] Z. Yang, M. Ruess, S. Kollmannsberger, A. Düster, and E. Rank. An efficient integration technique for the voxel-based finite cell method. *International Journal for Numerical Methods in Engineering*, 91:457–471, 2012.

[16] A. Düster, J. Parvizian, Z. Yang, and E. Rank. The finite cell method for three-dimensional problems of solid mechanics. *Computational methods in applied mechanics and engineering*, 197:3768–3782, 2008.

[17] A. Düster, H.G. Sehlhorst, and E. Rank. Numerical homogenization of heterogeneous and cellular materials utilizing the finite cell method. *Computational Mechanics*, 50:413–431, 2012.

[18] J. Nitsche. Über ein variations zur lösung von dirichlet-problemen bei verwendung von teilräumen die keinen randbedingungen unterworfen sind. *Abh. Math. Se. Univ.*, 36:9–15, 1970.

[19] A. Ern and J.L. Guermond. *Theory and Practice of Finite Elements.* Springer Science & Business Media, 2004.

[20] Y. Saad. *Iterative Methods for Sparse Linear Systems.* SIAM, 2nd edition, 2003.

[21] G.M. Nielson and B. Hamann. The asymptotic decider: resolving the ambiguity in marching cubes. *Proceedings of the 2nd conference on Visualization'91. IEEE Computer Society Press*, 1991.

[22] C. Johnson. *Numerical Solution of Partial Differential Equations by the Finite Element Method.* Cambridge U. Press, 1987.

[23] R. Barrett et al. *Templates for the solution of linear systems: building blocks for iterative methods*, volume 43. SIAM, 1994.

[24] O. Axelsson. *Iterative solution methods.* Cambridge University Press, 1996.

[25] F.L. Bauer. Optimally scaled matrices. *Numerische Mathematik*, 5(1):73–87, 1963.

[26] R.D. Braatz and M. Morari. Minimizing the euclidean condition number. *SIAM Journal on Control and Optimization*, 32(6):1763–1768, 1994.

[27] A. van der Sluis. Condition numbers and equilibration of matrices. *Numerische Mathematik*, 14(1):14–23, 1969.

[28] O.C. Zienkiewics, R.L. Taylor, and J.Z. Zhu. *The Finite Element Method: Its Basis and Fundamentals.* Elsevier, 6th edition, 2005.

[29] M.G. Cox. The numerical evaluation of b-splines. *IMA Journal of Applied Mathematics*, 10(2):134–149, 1972.

[30] C. de Boor. On calculating with Bs-splines. *Journal of Approximation Theory*, 6(1):50–62, 1972.

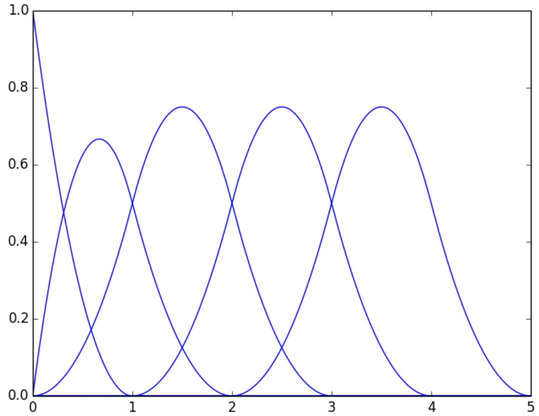


Figure A.1: An example of B-splines of order $p = 2$ with a knot vector $\Xi = \{0, 0, 0, 1, 2, 3, 4, 5\}$.

A B-spline functions

All basis functions used in this work are B-splines [1]. Starting with a non-decreasing knot vector $\Xi = \{\xi_1, \cdots, \xi_{n+p+1}\}$ containing the grid-points, B-splines of order p are constructed by the Cox-de Boor recursion formula [29, 30]. For $p = 0$ the i^{th} function is

$$N_{i,0}(x) = \begin{cases} 1 & \text{if } \xi_i \leq x < \xi_{i+1}, \\ 0 & \text{otherwise}, \end{cases} \tag{A.1}$$

and for $p > 0$ the i^{th} function is

$$N_{i,p} = \frac{x - \xi_i}{\xi_{i+p} - \xi_i} N_{i,p-1}(x) + \frac{\xi_{i+p+1} - x}{\xi_{i+p+1} - \xi_{i+1}} N_{i+1,p-1}(x). \tag{A.2}$$

B-splines are $(p + 1 - m)^{\text{th}}$-order continuous through a knot where m denotes the multiplicity (*i.e.* the amount of times that the same knot is repeated). The first knot in Figure A.1 has multiplicity $p + 1$ and therefore is discontinuous there, such that on the left side an essential boundary can be imposed in a strong manner. For this reason, matching grids always begin and end with multiplicity $p + 1$. Unmatching grids should be extended far enough beyond Ω such that all elements that intersect Ω have equally many degrees of freedom. In Figure A.1 this would imply that $\Omega = [0, c]$ with $c \leq 3$. Multidimensional B-splines are constructed by the vector product of knot vectors in different dimensions. Because of the simple grids that are applied for the Finite Cell Method, the physical domain is chosen to coincide with the parameter domain. This is not the general case for the application of B-splines however.