TECHNISCHE UNIVERSITÄT WIEN

DISSERTATION

$\mathcal{H}$-inverses of Gram matrices

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der Technischen Wissenschaften unter der Leitung von

Univ.-Prof. Jens Markus Melenk, PhD
E101 - Institut für Analysis und Scientific Computing, TU Wien

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Dipl.-Ing. Niklas Angleitner

Diese Dissertation haben begutachtet:

1. **Univ.-Prof. Jens Markus Melenk, PhD**
   Institut für Analysis und Scientific Computing, Technische Universität Wien
2. **Univ.-Prof. Dr. Mario Bebendorf**
   Fakultät für Mathematik, Physik und Informatik, Universität Bayreuth
3. **Prof. Dr. Holger Wendland**
   Fakultät für Mathematik, Physik und Informatik, Universität Bayreuth

Wien, am 25. Mai 2022

Kurzfassung

Wir betrachten eine geeignete Klasse von Gram-Matrizen und zeigen, dass die zugehörigen Inversen hervorragend durch hierarchischen Matrizen approximiert werden können. Die Einträge einer solchen Gram-Matrix ergeben sich aus einer vorgegebenen Bilinearform auf einem geeigneten Funktionenraum sowie einer endlichen Menge von Basisfunktionen. Derartige Matrizen treten häufig im Zusammenhang mit Galerkin-Diskretisierungen von partiellen Differentialgleichungen auf, welche zur Beschreibung zahlreicher Probleme aus Physik, Technik und angewandter Mathematik verwendet werden.

Die hier relevanten Funktionenräume sind die gewohnten Sobolev-Räume ganzzahliger Ordnung und die Bilinearformen ergeben sich als Varianten der natürlichen Innenprodukte auf diesen Räumen. Eine wichtige Voraussetzung für unsere Analyse ist die Gültigkeit einer *diskreten Caccioppoli-Ungleichung*, welche wir in einem *Finite-Elemente*-Setting sowie einem *Radiale Basisfunktionen*-Setting nachweisen. Die Voraussetzungen an die Basisfunktionen, welche zur Assemblierung der Gram-Matrix verwendet werden, sind sehr allgemein gehalten und es wird insbesondere keine Lokalität gefordert. Wir setzen stattdessen eine gewisse Art von Lokalität für die zugehörige *duale* Basis voraus.

Die Fragestellung, inwiefern inverse Gram-Matrizen durch datenschwache Alternativen approximiert werden können, ist nicht neu. In mehr als zwei Jahrzehnten Forschungsarbeit wurden unterschiedliche Herangehensweisen erarbeitet (z.B., [BH03], [Bör10], [Fau15]). Das Ziel dieser Arbeit besteht darin, diese Ideen in einem abstrakteren Rahmen zu formulieren, wodurch sich ein breiteres Anwendungsspektrum ergibt. Inbesondere können wir Gitterbasierte sowie Gitter-freie Probleme auf lokal verfeinerter Gittern und Punktwolken in beliebigen Raumdimensionen behandeln. Des Weiteren sind unstetige PDE-Koeffizienten, nicht-polygonale Rechengebiete sowie nicht-lokale Basisfunktionen erlaubt.

Diese Dissertation basiert auf den Werken [AFM21a], [AFM21b] and [AFM22], welche in Zusammenarbeit mit Dr. Markus Faustmann sowie Univ.-Prof. Jens Markus Melenk, PhD im Zuge des Doktoratsstudiums des Autors an der Technischen Universität Wien angefertigt wurden.

Abstract

In this thesis, we prove that the inverse of a certain type of Gram matrix can be approximated well from the class of hierarchical matrices. The entries of a Gram matrix are determined by a bilinear form on a suitable function space and by a finite set of basis functions. Such matrices appear frequently in the context of Galerkin discretizations of partial differential equations and many related problems in physics, engineering and applied mathematics.

As for the function spaces, we are mainly concerned with the usual Sobolev spaces of integer order and the bilinear forms under consideration are variants of the inherent inner products on these spaces. An important prerequisite for the analysis is the validity of a *discrete Caccioppoli inequality*, which we derive in a *finite element* setting and a *radial basis function* setting. The assumptions on the basis functions that make up the Gram matrix are very mild and do not incorporate locality. In fact, we only require some form of locality for the corresponding *dual* basis.

The question of low-cost approximability of inverse Gram matrices is certainly not new. In more than two decades of research, different approaches have been made to answer this question (e.g., [BH03], [Bör10], [Fau15]). The goal of this work is to unify these ideas and rephrase them in a more abstract framework, which can be applied to a larger class of problems. In particular, we can treat mesh-based and mesh-less problems, locally refined meshes/point clouds in arbitrary space dimensions, rough PDE coefficients, non-polygonal computation domains and non-local basis functions.

This thesis is based on the papers [AFM21a], [AFM21b] and [AFM22], which were composed as part of the author's doctoral studies at Technische Universität Wien in collaboration with Dr. Markus Faustmann and Univ.-Prof. Jens Markus Melenk, PhD.

Danksagung

Ich danke meiner langjährigen Partnerin für ihr fortwährendes Engagement bei der Betreuung und Erziehung unserer beiden Söhne. Ohne ihrer Bereitschaft, einen erheblichen Teil der Kinderbetreuungszeit zu übernehmen, wäre die Fertigstellung dieser Arbeit deutlich erschwert worden.

Meinen Eltern danke ich für die unermüdliche und bedingungslose Unterstützung in allen Belangen des Lebens. Das Durchhaltevermögen, welches für eine so umfangreiche Abschlussarbeit unerlässlich ist, kann ich nur auf ihr Vorbild zurückführen.

Schließlich möchte ich mich bei Prof. Jens Markus Melenk und Dr. Markus Faustmann für die vielen Stunden gemeinsamer Tüftelei bedanken.

Danke!

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 25. Mai 2022

Niklas Angleitner

Contents

1 Introduction

1.1 The Galerkin method

Consider a real Hilbert space V and let $a(\cdot,\cdot) : V \times V \longrightarrow \mathbb{R}$ be a continuous, coercive, possibly non-symmetric, bilinear form (cf. D.4.2). Denote by V^* the dual space of V, i.e., the space of continuous, linear functionals on V. We are interested in the following abstract variational problem:

Problem 1.1. *Let $f \in V^*$ be given. Find $u_\infty \in V$ such that*

$$\forall v \in V : \qquad a(u_\infty, v) = f(v).$$

The well-known Lax-Milgram Lemma (e.g., [BS08, Theorem 2.7.7]) guarantees that there exists a unique solution $u_\infty \in V$ of this problem. However, the Hilbert space V is typically infinite-dimensional and the solution u_∞ cannot be computed exactly. In the *Galerkin method* (e.g., [BS08, Section 2.6]), the space V is replaced by a suitable finite-dimensional subspace $V_N \subseteq V$. The result is a discrete variational problem:

Problem 1.2. *Let $f \in (V_N)^*$ be given. Find $u \in V_N$ such that*

$$\forall v \in V_N : \qquad a(u, v) = f(v).$$

Once again, the Lax-Milgram Lemma yields existence of a unique solution $u \in V_N$. Since V_N is finite-dimensional, the discrete solution u is computable. To this end, a basis

$$\{\varphi_1, \dots, \varphi_N\} \subseteq V_N$$

must be chosen. In a typical application of the Galerkin method, there is more than one candidate for such a basis and the particular choice of the basis functions φ_n has far-reaching practical consequences. In fact, a poor choice of basis can render the method infeasible, even on modern computer hardware. Once the basis functions φ_n are elected, we make the ansatz

$$u = \sum_{n=1}^{N} \boldsymbol{c}_n \varphi_n \in V_N,$$

where $\boldsymbol{c} \in \mathbb{R}^N$ is an unknown coefficient vector. Then, introducing the *Gram matrix*[1]

$$\boldsymbol{A} := (a(\varphi_n, \varphi_m))_{m,n=1}^{N} \in \mathbb{R}^{N \times N}$$

[1] The term *Gram matrix* is often reserved for matrices arising from proper inner products (e.g., [HJ13, Thorem 7.2.10.]). Our bilinear form $a(\cdot,\cdot)$ need not be symmetric, but we will use the term anyways.

and the load vector

$$\boldsymbol{f} := (f(\varphi_m))_{m=1}^N \in \mathbb{R}^N,$$

we can rewrite P.1.2 as an equivalent linear system of equations (LSE):

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{f}.$$

The unique solvability of P.1.2 already ensures that the system matrix $\boldsymbol{A}$ is invertible. In particular, this LSE can be used to compute the unknown coefficient vector $\boldsymbol{c}$ and thus the unknown solution u to P.1.2. To do so, we have a couple of different options:

1. *Gaussian elimination with partial pivoting:* Denote by $b \in \{0, \dots, N\}$ the bandwidth[2] of $\boldsymbol{A}$. It is well-known that Gaussian elimination with partial pivoting takes $\mathcal{O}(bN)$ memory and $\mathcal{O}(b^2N)$ time in order to compute the solution of $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{f}$ (e.g., [TB97, Algorithm 21.1.] and [DER17, Chapter 8]). In particular, if $b \eqsim N^\beta$, for some $\beta \in [0, 1]$, then the memory and time requirements amount to $\mathcal{O}(N^{1+\beta})$ and $\mathcal{O}(N^{1+2\beta})$, respectively. If the problem size N becomes too large (say, hundreds of thousands or even millions), this approach might become infeasible.

   For example, consider the case where $\Omega := (0, 1)^d$, $V := H^1(\Omega)$ and $a(\cdot, \cdot) := \langle \cdot, \cdot \rangle_V$. Cutting Ω into $\mathcal{O}(N^{1/d})$ equal slices along each coordinate axis, we can construct a simplicial, uniform mesh $\mathcal{T} \subseteq \mathrm{Pow}(\Omega)$ with $N \in \mathbb{N}$ nodes (cf. D.2.60). Denote by φ_n the *hat function* corresponding to the n-th mesh node (cf. L.2.72). Then, no matter the ordering of the indices, there will always be at least two hat functions φ_n, φ_m with $\mathrm{supp}(\varphi_n) \cap \mathrm{supp}(\varphi_m) \neq \emptyset$ and $|n - m| \gtrsim N^{1-1/d}$. It follows that $b \gtrsim N^{1-1/d}$, so that the solution of $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{f}$ uses $\mathcal{O}(N^{2-1/d})$ memory and $\mathcal{O}(N^{3-2/d})$ time.

2. $\boldsymbol{PLU}$*-decomposition:* If $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{f}$ needs to be solved for multiple right-hand sides $\boldsymbol{f} \in \mathbb{R}^N$, it might be favorable to compute a $\boldsymbol{PLU}$-decomposition of $\boldsymbol{A}$ in advance. In fact, Gaussian elimination with partial pivoting generates a permutation matrix $\boldsymbol{P} \in \mathbb{R}^{N \times N}$, a unit lower triangular matrix $\boldsymbol{L} \in \mathbb{R}^{N \times N}$ and an upper triangular matrix $\boldsymbol{U} \in \mathbb{R}^{N \times N}$ such that $\boldsymbol{A} = \boldsymbol{PLU}$. Then, the computation of $\boldsymbol{c}$ can be done in two steps: First, solve $\boldsymbol{L}\tilde{\boldsymbol{c}} = \boldsymbol{P}^{-1}\boldsymbol{f}$ for $\tilde{\boldsymbol{c}} \in \mathbb{R}^N$. Second, solve $\boldsymbol{U}\boldsymbol{c} = \tilde{\boldsymbol{c}}$ for $\boldsymbol{c}$. Now, since $\boldsymbol{L}$ and $\boldsymbol{U}$ also have bandwidth $\mathcal{O}(b)$, these two systems can be solved in $\mathcal{O}(bN)$ time by forward- and backward substitution. In total, if $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{f}$ is to be solved for $k \in \mathbb{N}$ different right-hand sides, this precedure takes $\mathcal{O}(bN)$ memory and $\mathcal{O}(b^2N + kbN)$ time.

3. *(P)CG:* Consider the case where $\boldsymbol{A}$ is symmetric and positive definite (SPD):

$$\boldsymbol{A}^T = \boldsymbol{A}, \qquad \forall \boldsymbol{d} \in \mathbb{R}^N \backslash \{\boldsymbol{0}\} : \quad \langle \boldsymbol{A}\boldsymbol{d}, \boldsymbol{d} \rangle_2 > 0.$$

   The well-known *conjugate gradient method (CG)* (e.g., [Atk89, Section 8.9], [Epp13, Section 9.3.3]) generates a sequence $(\boldsymbol{c}_k)_{k \in \mathbb{N}} \subseteq \mathbb{R}^N$ of approximations $\boldsymbol{c}_k \approx \boldsymbol{c}$ in a way

[2] The bandwidth b is the smallest number such that $\boldsymbol{A}_{mn} = 0$, for all $m, n \in \{1, \dots, N\}$ with $|m - n| \geq b$.

that only requires us to perform matrix-vector-multiplications $\boldsymbol{d} \mapsto \boldsymbol{A}\boldsymbol{d}$, for some $\boldsymbol{d} \in \mathbb{R}^N$. According to [Epp13, Theorem 9.7], there holds the error bound

$$\|\boldsymbol{c} - \boldsymbol{c}_k\|_2 \leq 2\ \mathrm{cond}_2(\boldsymbol{A})^{1/2} \left(\frac{1 - \mathrm{cond}_2(\boldsymbol{A})^{-1/2}}{1 + \mathrm{cond}_2(\boldsymbol{A})^{-1/2}} \right)^{k-1} \|\boldsymbol{c} - \boldsymbol{c}_1\|_2,$$

where $\mathrm{cond}_2(\boldsymbol{A}) := \|\boldsymbol{A}\|_2 \|\boldsymbol{A}^{-1}\|_2 \geq 1$ is the spectral condition number of $\boldsymbol{A}$. Clearly, the convergence speed is determined by the magnitude of $\mathrm{cond}_2(\boldsymbol{A})$, the best case scenario being $\mathrm{cond}_2(\boldsymbol{A}) = \mathcal{O}(1)$ as $N \to \infty$. In many instances[3], however, we have $\mathrm{cond}_2(\boldsymbol{A}) = \mathcal{O}(N^\alpha)$, for some constant $\alpha > 0$. In this case, as many as $k = \mathcal{O}(\ln(\varepsilon^{-1})N^{\alpha/2})$ iteration steps are necessary to reach some prescribed tolerance $\varepsilon > 0$. Since the matrix-vector-multiplication in each step costs at least $\mathcal{O}(N)$ floating point operations, the overall complexity of $\mathcal{O}(\ln(\varepsilon^{-1})N^{1+\alpha/2})$ might still be too much to handle.

This observation is the basis for the *preconditioned conjugate gradient method (PCG)* (e.g., [GVL13, Section 11.5.2]), where the LSE $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{f}$ is replaced by the following, equivalent LSE:

$$\begin{aligned} \boldsymbol{P}^{-1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}\tilde{\boldsymbol{c}} &= \boldsymbol{P}^{-1/2}\boldsymbol{f}, \\ \boldsymbol{P}^{1/2}\boldsymbol{c} &= \tilde{\boldsymbol{c}}. \end{aligned}$$

Here, $\boldsymbol{P} \in \mathbb{R}^{N \times N}$ is a *preconditioner* matrix that should satisfy the following requirements:

a) $\boldsymbol{P}$ is SPD.

b) $\mathrm{cond}_2(\boldsymbol{A}\boldsymbol{P}) = \mathcal{O}(1)$.

c) Every LSE of the form $\boldsymbol{P}\boldsymbol{d} = \boldsymbol{g}$, where $\boldsymbol{g} \in \mathbb{R}^N$ is given and $\boldsymbol{d} \in \mathbb{R}^N$ is sought, can be solved quickly (ideally in $\mathcal{O}(N)$ time).

In theory, the PCG method is just the CG method applied to the equivalent LSE above. However, the SPD matrix $\boldsymbol{P}^{-1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}$ need not be computed explicitly. In fact, for each step of the iteration, it suffices to perform a matrix-vector-multiplication $\boldsymbol{d} \mapsto \boldsymbol{A}\boldsymbol{d}$ and solve an LSE $\boldsymbol{P}\boldsymbol{d} = \boldsymbol{g}$ (apart from a few vector additions and scalar products). The construction of good preconditioners is a vast field of research (see, e.g., [Gre97, Part II], [GVL13, Section 11.5], [TB97, Lecture 40] and the references therein).

4. *Hierarchical $\boldsymbol{L}\boldsymbol{U}$-decomposition:* The inefficiency of standard $\boldsymbol{P}\boldsymbol{L}\boldsymbol{U}$-decompositions is due to the amount of fill-in that occurs in the triangular factors $\boldsymbol{L}$ and $\boldsymbol{U}$. To overcome this problem, [Beb07] showed that so-called $\mathcal{H}$-matrices (cf. Chapter 3) can be used as approximate $\boldsymbol{L}\boldsymbol{U}$-factors for finite element stiffness matrices on quasi-uniform meshes. In fact, under the assumption that the exact inverse $\boldsymbol{A}^{-1}$ can be approximated well by an $\mathcal{H}$-matrix, Bebendorf showed that, for all $\varepsilon > 0$, $\mathcal{H}$-matrices

[3] Example: If $\boldsymbol{A}$ is the FEM stiffness matrix for the problem $(-\Delta u = f,\ u|_{\partial\Omega} = 0)$ on a uniform mesh $\mathcal{T} \subseteq \mathrm{Pow}(\Omega)$ with meshsize $h > 0$, then $\mathrm{cond}_2(\boldsymbol{A}) \eqsim h^{-2} \eqsim N^{2/d}$, according to [EG06, Theorem 4.1.].

$\boldsymbol{L}_{\mathcal{H}}$ and $\boldsymbol{U}_{\mathcal{H}}$ of block-wise rank $r \approx \ln(N)^{\alpha}\ln(\varepsilon^{-1})^{\beta}$ (for some $\alpha, \beta \geq 0$) can be constructed such that

$$\|\boldsymbol{A} - \boldsymbol{L}_{\mathcal{H}}\boldsymbol{U}_{\mathcal{H}}\|_2 \leq C\ln(N)N^{2/d}\|\boldsymbol{L}\|_2\|\boldsymbol{U}\|_2\varepsilon + \mathcal{O}(\varepsilon^2).$$

5. *Hierarchical preconditioners:* In [Beb06], the author proposed to use the $\mathcal{H}$-inversion routine (which uses efficient $\mathcal{H}$-arithmetic exclusively) to construct a preconditioner $\boldsymbol{P}_{\mathcal{H}}$ for the PCG method. Under the assumption that the algorithm produces an error of at most $\|\boldsymbol{I} - \boldsymbol{A}\boldsymbol{P}_{\mathcal{H}}\|_2 \leq \delta < 1$, it was shown that $\mathrm{cond}_2(\boldsymbol{A}\boldsymbol{P}_{\mathcal{H}}) \leq (1+\delta)/(1-\delta) = \mathcal{O}(1)$, so that the PCG method converges rapidly.

Here, we laid out a variety of strategies for the practical solution of the LSE $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{f}$. One could argue that these techniques are just different approaches to computing an approximation to the inverse matrix $\boldsymbol{A}^{-1}$. In the last two instances, [Beb07] and [Beb06], the approximation class is given by the set of $\mathcal{H}$-matrices of prescribed block-wise rank r. Then, inevitably, the following fundamental question arises:

Problem 1.3. *What are the theoretical limits for the approximation of $\boldsymbol{A}^{-1}$ from the class of data-sparse $\mathcal{H}$-matrices?*

In our main result, T.4.21, we give an upper bound for the best approximation of $\boldsymbol{A}^{-1}$ in the class of hierarchical matrices, $\mathcal{H}(\mathbb{P}^2, r)$.

1.2 Literature discussion

The literature on $\mathcal{H}$-matrices has grown substantially during the last two decades. In this overview, we focus mainly on the work that is most relevant for this dissertation. The list is by no means exhaustive.

1. The *fast multipole method (FMM)* introduced in [GR87] was named one of the "top 10 algorithms of the 20th century" in [DS00]. The authors devised a novel, groundbreaking algorithm to reduce the computational complexity of the famous *N-body problem* from $\mathcal{O}(N^2)$ to $\mathcal{O}(p^2N)$ (introducing an error $\mathcal{O}(e^{-Cp})$, for some $C > 0$).

   An instance of such an N-body problem occurs in classic celestial mechanics, where $N \in \mathbb{N}$ given planets interact with each other via gravitational forces. The trajectory $t \mapsto x_n(t) \in \mathbb{R}^3$ of the n-th planet is governed by the ordinary differential equation

   $$M_n x_n'' = \sum_{m \in \{1,\dots,N\}\setminus\{n\}} \frac{GM_nM_m}{\|x_m - x_n\|_2^3}(x_m - x_n),$$

   where $G > 0$ is a constant and where $M_n > 0$ is the mass of the n-th planet. In a typical time stepping scheme (e.g. forward Euler method), a naive implementation of this formula requires $\mathcal{O}(N^2)$ arithmetic operations to evolve the whole system one time step further.

   However, Greengard and Rokhlin found that much effort could be spared by organizing the planets into a hierarchy of groups of nearby planets. The key insight was that

the interaction between two *well separated* groups can be approximated to arbitrary accuracy by polynomials of a prescribed degree $p \in \mathbb{N}$. (Essentially a truncated Taylor series of the function $x \mapsto x/\|x\|_2^3$, which is smooth away from the origin.) Assuming that the groups have N_1 and N_2 members each, this simplification reduces the cost from $\mathcal{O}(N_1 N_2)$ to $\mathcal{O}(p(N_1 + N_2))$ and only introduces a marginal error of $\mathcal{O}(e^{-Cp})$, for some constant $C > 0$. Finally, the organization in a hierarchy facilitated a *divide and conquer* scheme which resulted in an $\mathcal{O}(p^2 N)$-algorithm.

2. In parallel, Hackbusch and Nowak developed a similar strategy known as *panel clustering method* (e.g., [HN89]). In a series of works ([Hac99], [HK00a], [HK00b], [Gra01], [Hac09]), Hackbusch, Khoromskij and Grasedyck formalized the ideas from [GR87] and [HN89] and introduced *hierarchical block partitions* $\mathbb{P}^2$ of the set of matrix indices $\{1, \dots, N\} \times \{1, \dots, N\} = \{(i, j) \,|\, i, j \in \{1, \dots, N\}\}$. The elements of $\mathbb{P}^2$ are pairs (I, J) of *clusters* $I, J \subseteq \{1, \dots, N\}$ satisfying

$$\dot{\bigcup_{(I,J)\in\mathbb{P}^2}} I \times J = \{1, \dots, N\} \times \{1, \dots, N\}.$$

In particular, given a matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, the set $\mathbb{P}^2$ induces a partition of $\boldsymbol{A}$ into a family of matrix blocks, $\{\boldsymbol{A}|_{I\times J} \in \mathbb{R}^{I\times J} \,|\, (I, J) \in \mathbb{P}^2\}$. Assuming that the (i, j)-th matrix entry $\boldsymbol{A}_{ij}$ encodes some form of interaction between two physical domains (or points) $\Omega_i, \Omega_j \subseteq \mathbb{R}^d$, it is clear that a matrix block $\boldsymbol{A}|_{I\times J}$ represents *all pairwise interactions between two groups* of domains, $\{\Omega_i \,|\, i \in I\}$ and $\{\Omega_j \,|\, j \in J\}$.

The authors constructed the partition $\mathbb{P}^2$ in an iterative manner (a tree), starting from the root $(\{1, \dots, N\}, \{1, \dots, N\})$ and splitting pairs (I, J) successively into four (or more) children $(I_1, J_1), \dots, (I_4, J_4)$ with

$$\dot{\bigcup_{l=1}^{4}} I_l \times J_l = I \times J.$$

The subdivision of a pair (I, J) stops as soon as it becomes *admissible* or *well separated*, using the terminology of [GR87]. Roughly speaking, *admissible* means that the physical sets $\bigcup_{i\in I} \Omega_i \subseteq \mathbb{R}^d$ and $\bigcup_{j\in J} \Omega_j \subseteq \mathbb{R}^d$ are small in diameter in comparison to their distance. In particular, if the interaction between these sets is described by a sufficiently smooth function, there is a good chance that the approximation mechanism of [GR87] also applies in this generalized setting.

The authors then proceeded to introduce the class of *hierarchical matrices*

$$\mathcal{H}(\mathbb{P}^2, r) \subseteq \mathbb{R}^{N\times N},$$

where $r \in \mathbb{N}$ is a prescribed rank bound on the admissible matrix blocks $\boldsymbol{A}|_{I\times J}$. Furthermore, they managed to define approximate arithmetic operations on this class including matrix-vector-multiplication, matrix-matrix-addition, matrix-matrix-multiplication and even matrix-inversion. Assuming that the partition $\mathbb{P}^2$ is constructed properly, the cost of storage and arithmetic of $\mathcal{H}$-matrices is bounded by $\mathcal{O}(r^\alpha \ln(N)^\beta N)$, for some (small) values of $\alpha, \beta \in \mathbb{N}_0$ (see, e.g., [Gra01, Chapter 5]).

As for applications of $\mathcal{H}$-matrix approximation theory, most of the early work focused on FEM- and BEM-formulations of second order elliptic PDEs as well as Fredholm integral operators.

3. In [BH03], the authors proved that the inverse $\boldsymbol{A}^{-1}$ of a FEM stiffness matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ for a second order elliptic PDE can be approximated by an $\mathcal{H}$-matrix $\boldsymbol{B} \in \mathcal{H}(\mathbb{P}^2, r)$ with the same accuracy as the FEM error $\|u_\infty - u\|_{L^2(\Omega)}$. More precisely, they showed that $\|\boldsymbol{A}^{-1} - \boldsymbol{B}\|_2 \lesssim \varepsilon_N$ under the assumption that $\|u_\infty - u\|_{L^2(\Omega)} \leq \varepsilon_N \|f\|_{L^2(\Omega)}$, for all $f \in L^2(\Omega)$. Here $u_\infty \in H^1_0(\Omega)$ is the exact solution and $u \in V_N$ is the FEM solution from a discrete ansatz space $V_N \subseteq H^1_0(\Omega)$ which is based on some mesh $\mathcal{T}_N \subseteq \mathrm{Pow}(\Omega)$. Furthermore, the authors showed that the blockwise rank r of the approximant $\boldsymbol{B}$ is bounded by $r \lesssim \ln(N)^2 \ln(\ln(N)\varepsilon_N^{-1})^{d+1}$.

   To mention a few technical details, it was assumed that $d \geq 3$, that $\Omega \subseteq \mathbb{R}^d$ is a bounded Lipschitz domain, that $a(u,v) := \langle a_1 \nabla u, \nabla v\rangle_{L^2(\Omega)}$ for some $a_1 \in L^\infty(\Omega)^{d\times d}$, that homogeneous Dirichlet boundary conditions are employed, and that the mesh $\mathcal{T}_N$ is shape regular and quasi-uniform.

   The key idea of the proof was to express the exact solution u_∞ in terms of the Green's function G for the domain Ω. Then, exploiting certain regularity properties of G, the function G was approximated by separable expansions which were then used to construct the approximant $\boldsymbol{B}$.

4. A few years later, [Bör10] improved on [BH03] by using a completely different approach. Rather than approximating Green's function, the author approximated the exact solution $u_\infty \in H^1_0(\Omega)$ directly. The key insight was that u_∞ belongs to a class of *locally harmonic functions* which satisfy some orthogonality relations on certain subsets $\omega \subseteq \Omega$. This orthogonality could then be exploited to derive a *Caccioppoli inequality* of the form $|u_\infty|_{H^1(\omega)} \lesssim \mathrm{dist}_2(\omega, \partial\omega^+)^{-1}\|u_\infty\|_{L^2(\omega^+)}$, where $\omega^+ \supseteq \omega$ is a slightly larger subset of Ω. Fitting $L \in \mathbb{N}$ concentric subsets $\omega \subseteq \omega_1 \subseteq \cdots \subseteq \omega_L \subseteq \Omega$ around some initial set $\omega \subseteq \Omega$, the author was able to approximate the exact solution $u_\infty|_\omega$ by a function $u_L \in L^2(\omega)$, producing an error $\mathcal{O}(2^{-L})$ while only using $\mathcal{O}(L^{d+1})$ degrees of freedom to do so. This procedure led to an $\mathcal{H}$-matrix approximation $\boldsymbol{B} \in \mathcal{H}(\mathbb{P}^2, r)$ of an auxiliary matrix $\boldsymbol{S} \in \mathbb{R}^{N\times N}$, which, in some sense, encoded the abstract solution operator $L^{-1} : H^{-1}(\Omega) \longrightarrow H^1_0(\Omega)$ of the underlying PDE. Citing only the case of a quasi-uniform mesh $\mathcal{T}$, the error bound was of the form $\|\boldsymbol{S} - \boldsymbol{B}\|_2 \lesssim \ln(N) N 2^{-L}$ and the blockwise rank r of $\boldsymbol{B}$ satisfied $r \lesssim L^{d+1}$.

   On the other hand, under the assumption of a shift theorem ($f \in H^{-1+\varepsilon}(\Omega) \Rightarrow u_\infty \in H^{1+\varepsilon}_0(\Omega)$), it was shown that the error between the inverse stiffness matrix $\boldsymbol{A}^{-1}$ and the auxiliary matrix $\boldsymbol{S}$ could be bounded in the form $\|\boldsymbol{A}^{-1} - \boldsymbol{S}\|_2 \lesssim N^{1-2\varepsilon/d}$ (in the case of a quasi-uniform mesh $\mathcal{T}$). The final result was

$$\|\boldsymbol{A}^{-1} - \boldsymbol{B}\|_2 \leq \|\boldsymbol{A}^{-1} - \boldsymbol{S}\|_2 + \|\boldsymbol{S} - \boldsymbol{B}\|_2 \lesssim N^{1-2\varepsilon/d} + \ln(N) N 2^{-L},$$

   so that the overall accuracy of the approximant $\boldsymbol{B}$ was again dominated by the FEM error.

5. In his dissertation [Fau15], Faustmann further improved the results from [BH03] and [Bör10]. Most importantly, he showed that $\|\boldsymbol{A}^{-1} - \boldsymbol{B}\|_2 \lesssim \varepsilon$ can be achieved for *every* $\varepsilon > 0$, independent of the FEM error. In fact, his error bounds took the form

$$\|\boldsymbol{A}^{-1} - \boldsymbol{B}\|_2 \lesssim N^{\alpha} \exp(-br^{\beta}),$$

for some constants $\alpha, \beta, b > 0$ and *arbitrary* blockwise rank $r \in \mathbb{N}$.

Here, the novelty lied in working with the discrete solution $u \in V_N$ instead of the exact solution $u_\infty \in H^1_0(\Omega)$. In fact, at no point of the derivation was the exact solution u_∞ needed in any form. Nevertheless, the proof could be seen as a *fully discrete* analogue of the one in [Bör10]. The FEM solution u is locally harmonic in a discrete sense and satisfies a discrete version of the Caccioppoli inequality. The advantage of this approach was that no auxiliary matrix $\boldsymbol{S}$ needed to be introduced so that the triangle inequality at the end of [Bör10] could be omitted. Therefore, the final result is not polluted by the FEM error so that the $\mathcal{H}$-matrix approximation indeed reaches arbitrary accuracy.

The analysis was carried out for elliptic operators $Lu := -\operatorname{div}(a_1 \cdot \nabla u) + a_2 \cdot \nabla u + a_3 u$ with rough coefficients $a_1 \in L^\infty(\Omega, \mathbb{R}^{d\times d})$, $a_2 \in L^\infty(\Omega, \mathbb{R}^d)$, $a_3 \in L^\infty(\Omega, \mathbb{R})$ combined with Dirichlet-, Neumann- and Robin boundary conditions. The thesis also covers the Navier-Lamé equation of linear elasticity as well as the boundary element formulation of the homogeneous Laplace problem. Finally, we mention that the meshes $\mathcal{T}$ were assumed to be shape-regular and quasi-uniform.

1.3 This work's contribution

Here, in this thesis, we dwell on the fully discrete approach from [Fau15] and apply it in a more abstract framework. The formulation is general enough to allow for a simultaneous treatment of mesh-based finite element problems as well as mesh-free *radial basis function (RBF)* interpolation problems. The main selling points of the abstract framework are summarized below:

1. Mesh-based and mesh-less problems (i.e., point clouds) can be treated in the same way.

2. Meshes and point clouds can be highly non-uniform (e.g., locally refined- or exponentially graded meshes).

3. All spatial dimensions $d \in \mathbb{N}$ and all integer Sobolev orders $k \in \mathbb{N}_0$ are covered.

4. The computational domain $\Omega \subseteq \mathbb{R}^d$ need *not* be a polyhedron. In fact, we only need the existence of an extension operator $E_\Omega : H^k(\Omega) \longrightarrow H^k(\mathbb{R}^d)$ in the sense of D.2.48.

5. The basis functions $\varphi_1, \ldots, \varphi_N \in V_N$ need *not* have local supports. Instead, we require that the corresponding dual basis $\lambda_1, \ldots, \lambda_N \in (V_N)^*$ is in some sense local (cf. A.4.11).

6. More emphasis is put on the crucial role of the discrete Caccioppoli inequality, since it is a key ingredient of the construction.[4]

Under these circumstances, we will show that the inverse system matrix $\boldsymbol{A}^{-1} \in \mathbb{R}^{N\times N}$ can be approximated by an $\mathcal{H}$-matrix $\boldsymbol{B}_r \in \mathcal{H}(\mathbb{P}^2, r)$, for arbitrary $r \in \mathbb{N}$. The precise error bound (cf. T.4.21) is much in the spirit of [Fau15].

Caveat: As was the case in [BH03], [Bör10] and [Fau15], our construction of the approximant $\boldsymbol{B}$ is a theoretical existence result, because it involves certain inaccessible, abstract operators. In contrast, the previously mentioned $\mathcal{H}$-inversion routine (e.g., [Hac09, Section 7.5]) produces a *concrete* $\mathcal{H}$-matrix $\tilde{\boldsymbol{B}} \in \mathcal{H}(\mathbb{P}^2, r)$ which is supposed to approximate $\boldsymbol{A}^{-1}$. To the best of our knowledge, a rigorous bound for the error $\|\boldsymbol{A}^{-1} - \tilde{\boldsymbol{B}}\|_2$ is still missing to this day. However, *if* it can be proved that this algorithm produces a Céa-type best-approximation $\tilde{\boldsymbol{B}}$ in the sense

$$\|\boldsymbol{A}^{-1} - \tilde{\boldsymbol{B}}\|_2 \lesssim \inf_{\hat{\boldsymbol{B}}\in\mathcal{H}(\mathbb{P}^2,r)} \|\boldsymbol{A}^{-1} - \hat{\boldsymbol{B}}\|_2,$$

then our existence result readily yields the convergence $\tilde{\boldsymbol{B}} \to \boldsymbol{A}^{-1}$ as $r \to \infty$.

1.4 How to read this thesis

The advanced reader is probably familiar with most of the results from Chapter 2. However, the following parts should *not* be skipped, because they contain non-standard results that are important throughout the subsequent chapters:

1. In Section 2.3, we introduce axes-parallel boxes $B \subseteq \mathbb{R}^d$ along with their *inflated* cousins $B^\delta \subseteq \mathbb{R}^d$.
2. In Section 2.4, we define the concepts of *shape regularity*, *overlap* and *spread* for families of subsets $\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$.
3. In Section 2.8.7, we prove a continuous variant of a *Caccioppoli inequality*.
4. In Section 2.9.6, we construct a *discrete* cut-off function on a simplicial mesh $\mathcal{T}$.

Then, in Chapter 3, we introduce the class of *hierarchical matrices* and show how a matrix $\boldsymbol{B} \in \mathbb{R}^{N\times N}$ can be subdivided into a family of matrix blocks $\{\boldsymbol{B}|_{I\times J} \,|\, (I, J) \in \mathbb{P}^2\}$. To this end, we construct a *block partition* $\mathbb{P}^2$ using a *geometrically balanced clustering strategy* based on axes-parallel boxes.

Chapter 4 contains the main results of this thesis, T.4.20 and T.4.21. We introduce a general set of assumptions under which these results can be derived. The proof of T.4.20 is quite intricate and thus delayed to Chapter 5.

Finally, in Chapter 6 and Chapter 7, we apply the abstract framework from Chapter 4 to a *finite element* discretization of a second-order elliptic PDE and to a *radial basis function* interpolation problem. At the end of each chapter, we demonstrate the plausibility of our theoretical analysis by means of numerical experiments.

[4] If the reader intends to apply this abstract framework to a new problem, the first thing to check should be the validity of the discrete Caccioppoli inequality and the locality of the dual basis $\lambda_1, \dots, \lambda_N$.

2 Preliminary results

2.1 Notation

1. We use the convention $\mathbb{N} := \{1, 2, 3, \dots\}$ and $\mathbb{N}_0 := \{0, 1, 2, 3, \dots\}$.

2. The cardinality of countable sets M is denoted by $\#M$.

3. $\mathrm{Pow}(M)$ is the power set of a given set M (i.e., the set of all subsets).

4. "Large" matrices and vectors are typeset in boldface letters. For example, if a PDE problem is discretized with $N \gg 1$ degrees of freedom, a linear system of equations of size $N \times N$ needs to be solved. In this context, the stiffness matrix and load vector are denoted by $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ and $\boldsymbol{f} \in \mathbb{R}^N$, respectively.

5. For all matrices $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ and all index sets $I, J \subseteq \{1, \dots, N\}$, we denote by $\boldsymbol{A}|_{I \times J} \in \mathbb{R}^{I \times J}$ the matrix block that is formed by all entries $\boldsymbol{A}_{ij}$ with $(i, j) \in I \times J$.

6. If an inequality involves a multiplicative constant $C > 0$, which does not depend on critical parameters, we use the symbols "$\lesssim$" and "$\gtrsim$". For example, we write $a \lesssim b \lesssim c \lesssim d$ instead of $a \leq C_1 b \leq C_2 c \leq C_3 d$. The notation $a \eqsim b$ is used if both $a \lesssim b$ and $a \gtrsim b$ hold true.

7. In the context of algorithmic complexity, we also use the capital Landau notation $f(N) = \mathcal{O}(g(N))$ to describe a relation of the form $f(N) \lesssim g(N)$.

8. The *Kronecker delta* is denoted by

$$\forall i, j \in \mathbb{N}_0 : \qquad \delta_{ij} := \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} .$$

   Similarly, the *characteristic function* of a set $\Omega \subseteq \mathbb{R}^d$ is given by

$$\forall x \in \mathbb{R}^d : \qquad \mathbb{I}_\Omega(x) := \begin{cases} 1 & \text{if } x \in \Omega \\ 0 & \text{if } x \notin \Omega \end{cases} .$$

9. Subsets Ω of the d-dimensional coordinate space $\mathbb{R}^d$ are frequently referred to as *physical* sets. For a subset $\Omega \subseteq \mathbb{R}^d$ to be a *domain*, we only require it to be open. In the context of a variational problem on some domain Ω, we will also use the name *computational domain* to emphasize its intended use.

10. The d-dimensional Lebesgue measure of a measurable subset $\Omega \subseteq \mathbb{R}^d$ is denoted by $\mathrm{meas}(\Omega) = \mathrm{meas}_d(\Omega) \in [0, \infty]$. For subsets $\Gamma \subseteq \partial\Omega$, we write $\mathrm{meas}_{d-1}(\Gamma)$ for the $(d-1)$-dimensional surface measure.

11. Let $\Omega \subseteq \mathbb{R}^d$ be a set. We denote by $\overline{\Omega}$ its closure, by Ω° its interior, and by $\partial\Omega := \overline{\Omega} \cap \overline{\mathbb{R}^d \setminus \Omega}$ its boundary.

12. Let $\omega, \Omega \subseteq \mathbb{R}^d$ be measurable sets. We say that ω *is compactly contained in* Ω, if there exists a compact set $K \subseteq \mathbb{R}^d$ such that $\omega \subseteq K \subseteq \Omega$. In this case, we write $\omega \Subset \Omega$.

13. Let $\Omega \subseteq \mathbb{R}^d$ be open. We say that Ω is *(path-)connected*, if, for any two points $x, y \in \Omega$, there exists a continuous function $\gamma : [0,1] \longrightarrow \mathbb{R}^d$ such that $\gamma(0) = x$, $\gamma(1) = y$ and $\gamma([0,1]) \subseteq \Omega$.

    A subset $\omega \subseteq \Omega$ is a *connected component* of Ω, if it is connected and if there exists no other connected subset $\tilde{\omega} \subseteq \Omega$ such that $\omega \subsetneq \tilde{\omega}$ (i.e., ω is maximal with respect to "$\subseteq$").

14. Let $\Omega \subseteq \mathbb{R}^d$ be open. Unless explicitly told otherwise, all function spaces on Ω are meant to be real-valued.

15. Let $\Omega \subseteq \mathbb{R}^d$ be open. The support of a function $v : \Omega \longrightarrow \mathbb{R}$ is defined as

$$\operatorname{supp}(v) := \overline{\{x \in \Omega \,|\, v(x) \neq 0\}},$$

    where the closure is taken in $\mathbb{R}^d$.

16. Let $\Omega \subseteq \mathbb{R}^d$ and let $v : \Omega \longrightarrow \mathbb{R}$ be a function. For every subset $\omega \subseteq \Omega$, we denote by $v|_\omega : \omega \longrightarrow \mathbb{R}$ the restriction of v to ω, i.e.,

$$\forall x \in \omega : \qquad (v|_\omega)(x) := v(x).$$

17. Let $\Omega \subseteq \mathbb{R}^d$ be open. Let $v : \Omega \longrightarrow \mathbb{R}$ and $w : \Omega \longrightarrow \mathbb{R}^d$ be sufficiently smooth. For every $i \in \{1, \dots, d\}$, we denote by $\partial_i v : \Omega \longrightarrow \mathbb{R}$ the partial derivative with respect to the i-th coordinate. For higher-order partial derivatives, we write, e.g., $\partial_{ij} v = \partial_i \partial_j v$ and also $\partial_i^{(k)} v = \partial_i \cdots \partial_i v$ ($k \in \mathbb{N}$ times). We set

$$\nabla v := (\partial_i v)_{i=1}^d, \qquad \Delta v := \sum_{i=1}^d \partial_{ii} v, \qquad \operatorname{div} w := \sum_{i=1}^d \partial_i w_i.$$

    Depending on the context, ∇v is either a column- or a row-vector.

18. We use the usual multi-index notation: Let $d \in \mathbb{N}$ and $\alpha = (\alpha_1, \dots, \alpha_d), \beta = (\beta_1, \dots, \beta_d) \in \mathbb{N}_0^d$. We set $\alpha \pm \beta := (\alpha_1 \pm \beta_1, \dots, \alpha_d \pm \beta_d)$ and write $\alpha \leq \beta$, if $\alpha_i \leq \beta_i$, for all $i \in \{1, \dots, d\}$. Similarly, we write $\alpha < \beta$, if $\alpha \leq \beta$ but $\alpha \neq \beta$. We set

$$|\alpha| := \alpha_1 + \cdots + \alpha_d, \qquad \alpha! := \alpha_1! \dots \alpha_d!, \qquad \binom{\alpha}{\beta} := \frac{\alpha!}{\beta!(\alpha - \beta)!}.$$

    For all $x \in \mathbb{R}^d$ and all sufficiently smooth functions $v : \Omega \longrightarrow \mathbb{R}$ ($\Omega \subseteq \mathbb{R}^d$ open), we write

$$x^\alpha := x_1^{\alpha_1} \dots x_d^{\alpha_d}, \qquad \mathrm{D}^\alpha v = \partial_1^{(\alpha_1)} \dots \partial_d^{(\alpha_d)} v.$$

Let $k \in \mathbb{N}_0$. Statements of the form "For all $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq k$, ..." are usually abbreviated as "For all $|\alpha| \leq k$, ...". Likewise, sums of the form $\sum_{\alpha \in \mathbb{N}_0^d : |\alpha| \leq k} (\dots)$ will be written as $\sum_{|\alpha| \leq k} (\dots)$.

19. Inside integrals, we frequently drop the arguments of the integrands. E.g., we abbreviate $\int_\Omega v(x) w(x) \, \mathrm{d}x$ by $\int_\Omega vw \, \mathrm{d}x$.

2.2 Norms, diameters, distances

Definition 2.1. *Let $d \in \mathbb{N}$ and $p \in [1, \infty]$. We define*

$$\forall x \in \mathbb{R}^d : \qquad \|x\|_p := \begin{cases} (\sum_{i=1}^d |x_i|^p)^{1/p} & \text{if } p \in [1, \infty) \\ \max_{i \in \{1,\dots,d\}} |x_i| & \text{if } p = \infty \end{cases}.$$

In the case $p = 2$, we define

$$\forall x, y \in \mathbb{R}^d : \qquad \langle x, y \rangle_2 := \sum_{i=1}^d x_i y_i.$$

Definition 2.2. *For all $M, N \in \mathbb{N}$, the* spectral norm *of a matrix $A \in \mathbb{R}^{M \times N}$ is denoted by*

$$\|A\|_2 := \sup_{x \in \mathbb{R}^N} \frac{\|Ax\|_2}{\|x\|_2}.$$

Definition 2.3. *For all $\Omega \subseteq \mathbb{R}^d$, we define*

$$\operatorname{diam}_2(\Omega) := \sup_{x,y \in \Omega} \|y - x\|_2 \in [0, \infty].$$

Lemma 2.4. *For all $\Omega \subseteq \mathbb{R}^d$, there hold the relations*

$$\operatorname{diam}_2(\Omega^\circ) \leq \operatorname{diam}_2(\Omega) = \operatorname{diam}_2(\overline{\Omega}).$$

Proof. The relations $\operatorname{diam}_2(\Omega^\circ) \leq \operatorname{diam}_2(\Omega) \leq \operatorname{diam}_2(\overline{\Omega})$ are trivial, because $\Omega^\circ \subseteq \Omega \subseteq \overline{\Omega}$. On the other hand, for all $n \in \mathbb{N}$ and all $x, y \in \overline{\Omega}$, we may pick points $x_n, y_n \in \Omega$ with $\|x - x_n\|_2 + \|y - y_n\|_2 \leq 1/n$, so that

$$\operatorname{diam}_2(\overline{\Omega}) \leq \sup_{x,y \in \overline{\Omega}} \|y - y_n\|_2 + \|y_n - x_n\|_2 + \|x_n - x\|_2 \leq \operatorname{diam}_2(\Omega) + \frac{1}{n} \overset{n}{\to} \operatorname{diam}_2(\Omega).$$

□

Definition 2.5. *For all subsets $\Omega_1, \Omega_2 \subseteq \mathbb{R}^d$, we set*

$$\operatorname{dist}_2(\Omega_1, \Omega_2) := \inf_{\substack{x \in \Omega_1 \\ y \in \Omega_2}} \|y - x\|_2.$$

Definition 2.6. *For all $x \in \mathbb{R}^d$ and $r \geq 0$, we define the balls*

$$\begin{aligned} \mathrm{Ball}_2(x,r) &:= \{y \in \mathbb{R}^d \,|\, \|y-x\|_2 < r\}, \\ \overline{\mathrm{Ball}}_2(x,r) &:= \{y \in \mathbb{R}^d \,|\, \|y-x\|_2 \leq r\}. \end{aligned}$$

Lemma 2.7. *1. For all $x \in \mathbb{R}^d$ and all $r \geq 0$, there holds*

$$\mathrm{meas}(\mathrm{Ball}_2(x,r)) = \mathrm{meas}(\overline{\mathrm{Ball}}_2(x,r)) = C(d) r^d,$$

where $C(d) = \pi^{d/2}(d/2)^{-1}\Gamma(d/2)^{-1}$.

2. For all $x_1, x_2 \in \mathbb{R}^d$ and $r_1, r_2 > 0$, there holds the following equivalence:

$$\mathrm{Ball}_2(x_1,r_1) \cap \mathrm{Ball}_2(x_2,r_2) = \emptyset \quad \Leftrightarrow \quad r_1 + r_2 \leq \|x_2 - x_1\|_2.$$

Proof. Ad item 1: See, e.g., [Fle77, Formula (5.46)].

Ad item 2: If $\mathrm{Ball}_2(x_1,r_1) \cap \mathrm{Ball}_2(x_2,r_2) \neq \emptyset$, then we can pick a point $x \in \mathbb{R}^d$ with $\|x - x_1\|_2 < r_1$ and $\|x - x_2\|_2 < r_2$, so that $\|x_2 - x_1\|_2 \leq \|x_2 - x\|_2 + \|x - x_1\|_2 < r_1 + r_2$. On the other hand, if $r_1 + r_2 > \|x_2 - x_1\|_2$, then the point $x := (r_2 x_1 + r_1 x_2)/(r_1 + r_2) \in \mathbb{R}^d$ lies in the intersection of $\mathrm{Ball}_2(x_1,r_1)$ and $\mathrm{Ball}_2(x_2,r_2)$:

$$\|x - x_1\|_2 = r_1 \|x_2 - x_1\|_2/(r_1 + r_2) < r_1, \qquad \|x - x_2\|_2 = r_2 \|x_2 - x_1\|_2/(r_1 + r_2) < r_2.$$

□

2.3 Axes-parallel boxes

Axes-parallel boxes play an important role throughout this work. In Chapter 3, we will use them for the purpose of *geometric clustering*, i.e., to subdivide a given point cloud $x_1, \ldots, x_N \in \mathbb{R}^d$ into multiple groups. To this end, we will need a way to split a given box into smaller ones and also to "inflate" a box by a given amount. Then, in Section 5.4, we use axes-parallel boxes again in a "partition of unity" argument on a family of overlapping boxes. Finally, in T.5.16, we work with a nested sequence of axes-parallel boxes.

Definition 2.8. *A subset $B \subseteq \mathbb{R}^d$ is called* (axes-parallel) box, *if there exist $a_i, b_i \in \mathbb{R}$, $a_i < b_i$, such that*

$$B = \bigtimes_{i=1}^{d} [a_i, b_i).$$

We denote the set of all boxes by $\mathbb{B}$.

Note that we consider half-open boxes so that we can tile the full space $\mathbb{R}^d$ without any holes or slits. For easy reference later on, we state the following trivial facts:

Lemma 2.9. *For all $B \in \mathbb{B}$, there hold the following relations:*

$$B \neq \emptyset, \qquad \mathrm{diam}_2(B) \in (0,\infty), \qquad \mathrm{meas}(B) \in (0,\infty).$$

Next, we demonstrate how to divide a given box $A \in \mathbb{B}$ into smaller ones. Here, we split A in half along each one of the coordinate axes, producing 2^d smaller boxes.

Definition 2.10. *Consider a box* $A = \bigtimes_{i=1}^d [a_i, b_i) \in \mathbb{B}$ *and let* $\Lambda := \{0,1\}^d$. *We define* subboxes $(A^{(\lambda)})_{\lambda\in\Lambda}$ *in the following way:*

$$A^{(\lambda)} := \bigtimes_{i=1}^d [(1-\lambda_i)a_i + \lambda_i(a_i+b_i)/2, (1-\lambda_i)(a_i+b_i)/2 + \lambda_i b_i) \in \mathbb{B}.$$

We denote by

$$\operatorname{sons}(A) := \{A^{(\lambda)} \mid \lambda \in \Lambda\} \subseteq \mathbb{B}$$

the sons *of the box* A.

Let us have a look at an example in $d = 2$ spatial dimensions: If $A = [0,6) \times [0,2)$, then

$$\begin{aligned} A^{((0,1))} &= [0,3) \times [1,2), & A^{((1,1))} &= [3,6) \times [1,2), \\ A^{((0,0))} &= [0,3) \times [0,1), & A^{((1,0))} &= [3,6) \times [0,1). \end{aligned}$$

Lemma 2.11. *There hold the following properties:*

1. *For all* $\lambda, \tilde\lambda \in \Lambda$ *with* $\lambda \neq \tilde\lambda$, *there holds* $A^{(\lambda)} \neq A^{(\tilde\lambda)}$. *In particular,*

$$\#\operatorname{sons}(A) = 2^d.$$

2. *Let* $A \in \mathbb{B}$. *Then the subboxes* $\operatorname{sons}(A)$ *form a partition of* A:
   - *There holds* $\emptyset \notin \operatorname{sons}(A)$.
   - *For all* $B, \tilde B \in \operatorname{sons}(A)$ *with* $B \neq \tilde B$, *there holds* $B \cap \tilde B = \emptyset$.
   - *For all* $B \in \operatorname{sons}(A)$, *there holds* $B \subseteq A$. *Furthermore,*

$$\bigcup_{B\in\operatorname{sons}(A)}{\cdot}\, B = A.$$

3. *Let* $A, \tilde A \in \mathbb{B}$ *with* $A \cap \tilde A = \emptyset$. *Then*

$$\operatorname{sons}(A) \cap \operatorname{sons}(\tilde A) = \emptyset.$$

4. *Let* $A \in \mathbb{B}$. *For all* $B \in \operatorname{sons}(A)$, *there hold the relationships*

$$\operatorname{diam}_2(B) = 2^{-1}\operatorname{diam}_2(A), \qquad \operatorname{meas}(B) = 2^{-d}\operatorname{meas}(A).$$

Proof. Ad item 1: Let $\lambda, \tilde\lambda \in \Lambda$ with $\lambda \neq \tilde\lambda$. Then it can easily be verified that the point $x := ((1-\lambda_i)a_i + \lambda_i(a_i+b_i)/2)_{i=1}^d \in \mathbb{R}^d$ lies in $A^{(\lambda)}$, but not in $A^{(\tilde\lambda)}$. Therefore, $A^{(\lambda)} \neq A^{(\tilde\lambda)}$. In particular, $\#\operatorname{sons}(A) = \#\Lambda = 2^d$.

Ad item 2: The fact that $\emptyset \notin \operatorname{sons}(A)$ follows from L.2.9. To see the disjointness, let $B, \tilde B \in \operatorname{sons}(A)$ with $B \neq \tilde B$ be given. According to item 1, there exist $\lambda, \tilde\lambda \in \Lambda$ with $\lambda \neq \tilde\lambda$, such that $B = A^{(\lambda)}$ and $\tilde B = A^{(\tilde\lambda)}$. Now, abbreviate $A = \bigtimes_{i=1}^d A_i$ and $A^{(\lambda)} = \bigtimes_{i=1}^d A_i^{(\lambda_i)}$,

where $A_i := [a_i, b_i)$ and $A_i^{(\lambda_i)} := [(1-\lambda_i)a_i + \lambda_i(a_i+b_i)/2, (1-\lambda_i)(a_i+b_i)/2 + \lambda_i b_i)$. Since $\lambda \neq \tilde{\lambda}$, one can easily check that $A_i^{(\lambda_i)} \cap A_i^{(\tilde{\lambda}_i)} = \emptyset$, for at least one $i \in \{1, \dots, d\}$. The disjointness of B and $\tilde{B}$ then follows from

$$B \cap \tilde{B} = A^{(\lambda)} \cap A^{(\tilde{\lambda})} = \left(\bigtimes_{i=1}^{d} A_i^{(\lambda_i)} \right) \cap \left(\bigtimes_{i=1}^{d} A_i^{(\tilde{\lambda}_i)} \right) = \bigtimes_{i=1}^{d} (A_i^{(\lambda_i)} \cap A_i^{(\tilde{\lambda}_i)}) = \emptyset.$$

On the other hand, since $A_i = A_i^{(0)} \cup A_i^{(1)}$, for all $i \in \{1, \dots, d\}$, we get

$$\begin{aligned} A &= \{x \in \mathbb{R}^d \,|\, x_1 \in A_1, \dots, x_d \in A_d\} \\ &= \{x \in \mathbb{R}^d \,|\, x_1 \in A_1^{(0)} \cup A_1^{(1)}, \dots, x_d \in A_d^{(0)} \cup A_d^{(1)}\} \\ &= \bigcup_{\lambda \in \Lambda} \{x \in \mathbb{R}^d \,|\, x_1 \in A_1^{(\lambda_1)}, \dots, x_d \in A_d^{(\lambda_d)}\} \\ &= \bigcup_{\lambda \in \Lambda} A^{(\lambda)} = \bigcup_{B \in \mathrm{sons}(A)} B. \end{aligned}$$

Ad item 3: Let $A, \tilde{A} \in \mathbb{B}$ with $A \cap \tilde{A} = \emptyset$. If $\mathrm{sons}(A) \cap \mathrm{sons}(\tilde{A})$ was not empty, we could pick a box $B \in \mathbb{B}$ with $B \in \mathrm{sons}(A)$ and $B \in \mathrm{sons}(\tilde{A})$. Then, item 2 yields $B \subseteq A \cap \tilde{A} = \emptyset$, which contradicts L.2.9.

Ad item 4: We only prove the first identity, since the second one is very similar. Given $B = A^{(\lambda)}$, for some $\lambda \in \Lambda$, we compute

$$\begin{aligned} \mathrm{diam}_2(B)^2 &= \sum_{i=1}^{d} \Big((1-\lambda_i)(a_i+b_i)/2 + \lambda_i b_i - (1-\lambda_i)a_i + \lambda_i(a_i+b_i)/2 \Big)^2 \\ &= \sum_{i=1}^{d} ((b_i - a_i)/2)^2 = 2^{-2}\mathrm{diam}_2(A)^2. \end{aligned}$$

This finishes the proof.

□

In D.2.10, we learned how to split a given box into smaller pieces. Next, we show how to increase the size of a box:

Definition 2.12. *Let $B = \bigtimes_{i=1}^{d} [a_i, b_i) \in \mathbb{B}$ and $\delta \geq 0$. We define the* inflated box[1]

$$B^{\delta} := \bigtimes_{i=1}^{d} [a_i - \delta, b_i + \delta) \in \mathbb{B}.$$

Note that B^{δ} is again a box. In particular, we can iterate $(B^{\delta})^{\delta} = B^{2\delta}$, $((B^{\delta})^{\delta})^{\delta} = B^{3\delta}$, et cetera. In the next lemma, we provide a short summary of the relevant properties of inflated boxes.

Lemma 2.13. *1. For all $B \in \mathbb{B}$, $\delta \geq 0$, $x \in B$ and $y \in \mathbb{R}^d$ with $\|y - x\|_2 \leq \delta$, there holds $y \in B^{\delta}$.*

[1] The notation is similar to the one from D.2.10. $B^{(\lambda)}$ is a smaller box than B and B^{δ} is a larger one.

2. *For all $B \in \mathbb{B}$, $\delta \geq 0$ and $y \in B^\delta$, there exists a point $x \in B$ such that*

$$\|y - x\|_2 \leq \sqrt{d}\delta.$$

3. *For all $B \in \mathbb{B}$ and $\delta \geq 0$, there hold the bounds*

$$\operatorname{diam}_2(B)/\sqrt{d} + 2\sqrt{d}\delta \leq \operatorname{diam}_2(B^\delta) \leq \operatorname{diam}_2(B) + 2\sqrt{d}\delta.$$

4. *For all $B_1, B_2 \in \mathbb{B}$ and $\delta_1, \delta_2 \geq 0$, there hold the bounds*

$$\operatorname{dist}_2({B_1}^{\delta_1}, {B_2}^{\delta_2}) \leq \operatorname{dist}_2(B_1, B_2) \leq \operatorname{dist}_2({B_1}^{\delta_1}, {B_2}^{\delta_2}) + \sqrt{d}(\delta_1 + \delta_2).$$

Proof. Ad item 1: Let $x \in B$ and $y \in \mathbb{R}^d$ with $\|y - x\|_2 \leq \delta$. Then $a_k \leq x_k < b_k$, for all $k \in \{1, \dots, d\}$. Since $|y_k - x_k| \leq \|y - x\|_2 \leq \delta$, we get $a_k - \delta \leq x_k - \delta \leq x_k - |y_k - x_k| \leq y_k$ and similarly $y_k < b_k + \delta$.

Ad item 2: Let $y \in B^\delta$, i.e., $a_k - \delta \leq y_k < b_k + \delta$, for all $k \in \{1, \dots, d\}$. Abbreviating $c_k := (a_k + b_k)/2$, we define a point $x \in \mathbb{R}^d$ in the following way:

$$\forall k \in \{1, \dots, d\} : \qquad x_k := \frac{b_k - c_k}{b_k + \delta - c_k}(y_k - c_k) + c_k.$$

Using the bound

$$|y_k - c_k| = \max\{y_k - c_k, c_k - y_k\} \leq \max\{b_k + \delta - c_k, c_k - a_k + \delta\} = b_k + \delta - c_k,$$

we get $|x_k - c_k| \leq |b_k - c_k| = (b_k - a_k)/2$, telling us that $x_k \in [a_k, b_k]$, for all $k \in \{1, \dots, d\}$. In fact, checking the case $y_k = b_k + \delta$ explicitly, there even holds $x_k \in [a_k, b_k)$, so that $x \in B$. Finally, we have the error bound $|y_k - x_k| = \delta|y_k - c_k|/(b_k + \delta - c_k) \leq \delta$, which readily implies $\|y - x\|_2 \leq \sqrt{d}\delta$ after summation over all k.

Ad item 3, left-hand inequality: Follows from the norm equivalence $\|\cdot\|_2 \leq \|\cdot\|_1 \leq \sqrt{d}\|\cdot\|_2$:

$$\sqrt{d}\operatorname{diam}_2(B^\delta) \geq \operatorname{diam}_1(B^\delta) = \operatorname{diam}_1(B) + 2d\delta \geq \operatorname{diam}_2(B) + 2d\delta.$$

Ad item 3, right-hand inequality: For all $y_1, y_2 \in B^\delta$, pick points $x_1, x_2 \in B$ as described in item 2. Then,

$$\operatorname{diam}_2(B^\delta) \leq \sup_{y_1, y_2 \in B^\delta} \|y_2 - x_2\|_2 + \|x_2 - x_1\|_2 + \|x_1 - y_1\|_2 \leq \operatorname{diam}_2(B) + 2\sqrt{d}\delta.$$

Ad item 4, left-hand inequality: Follows immediately from the inclusions $B_1 \subseteq {B_1}^{\delta_1}$ and $B_2 \subseteq {B_2}^{\delta_2}$.

Ad item 4, right-hand inequality: Using item 2 again, we estimate

$$\begin{aligned}\operatorname{dist}_2(B_1, B_2) &= \inf_{\substack{b_1 \in B_1 \\ b_2 \in B_2}} \|b_2 - b_1\|_2 \leq \inf_{\substack{a_1 \in {B_1}^{\delta_1} \\ a_2 \in {B_2}^{\delta_2}}} \inf_{b_2 \in B_2} \|b_2 - a_2\|_2 + \|a_2 - a_1\|_2 + \inf_{b_1 \in B_1} \|a_1 - b_1\|_2 \\ &\leq \operatorname{dist}_2({B_1}^{\delta_1}, {B_2}^{\delta_2}) + \sqrt{d}(\delta_1 + \delta_2).\end{aligned}$$

□

2.4 Shape regularity, overlap and spread

Shape regularity is a well-known concept in the literature (e.g., [Bra13, Chapter 2, Section 5], [BS08, Chapter 4] and [LB13, Section 3.1]). It is a prerequisite for almost all stability and error estimates in the context of finite element discretizations. In mesh-based methods, shape regularity is frequently accompanied by the notion of *overlap*, which somehow reflects the interaction between adjacent mesh elements. Finally, in the realm of mesh-less methods, we have to assume some sort of spatial boundedness of the degrees of freedom. To this end, we introduce *spread*.

Definition 2.14. *For every subset $\Omega \subseteq \mathbb{R}^d$, we define*

$$h_\Omega := \operatorname{diam}_2(\Omega).$$

Remark 2.15. *In this work, we are dealing mainly with two types of subsets $\Omega \subseteq \mathbb{R}^d$. On one hand, we have "small" subsets $\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$ coming from an approximation process (e.g., mesh elements $T \in \mathcal{T}$, D.2.60). On the other hand, we use "large" axes parallel boxes $B \in \mathbb{B}$ (cf. D.2.8) to subdivide the smaller sets Ω_n into multiple groups. The diameters of the sets Ω_n are typically $\mathcal{O}(N^{-\alpha})$, for some $\alpha > 0$, whereas the diameters of the boxes $B \in \mathbb{B}$ range from $\mathcal{O}(N^{-\alpha})$ up to $\mathcal{O}(1)$. As is customary in numerical analysis, we will use the character h for the diameters of the "small" sets Ω_n, but we will* not *use it for the diameters of the mostly "large" boxes B.*

Definition 2.16. *Let $\sigma_{\mathrm{shp}} \geq 1$ and $\Omega \subseteq \mathbb{R}^d$. We say*[2] *that Ω* has shape regularity *σ_{shp}, if*

$$h_\Omega \in (0, \infty)$$

and if there exists a point $x_{\mathrm{shp}} \in \Omega$ such that[3]

$$\operatorname{Ball}_2(x_{\mathrm{shp}}, (2\sigma_{\mathrm{shp}})^{-1} h_\Omega) \subseteq \Omega.$$

In this case, x_{shp} is called an incenter *of Ω. Similarly, a family of subsets $\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$ is said to* have shape regularity *σ_{shp}, if each individual set Ω_n has shape regularity σ_{shp}.*

Note that an incenter need not be unique. We summarize the relevant facts about shape regular sets:

Lemma 2.17. *Let $\Omega, \tilde{\Omega} \subseteq \mathbb{R}^d$ be given sets with shape regularity $\sigma_{\mathrm{shp}} \geq 1$. Furthermore, let $x_{\mathrm{shp}} \in \Omega$ and $\widetilde{x_{\mathrm{shp}}} \in \tilde{\Omega}$ be given incenters.*

1. *There holds*

$$x_{\mathrm{shp}} \in \operatorname{Ball}_2(x_{\mathrm{shp}}, (2\sigma_{\mathrm{shp}})^{-1} h_\Omega) \subseteq \Omega^\circ,$$

where Ω° is the interior of Ω.

[2] We will also use the phrase "Ω is σ_{shp}-shape regular" or simply "Ω is shape regular" and implicitly assume that a constant $\sigma_{\mathrm{shp}} \geq 1$ was prescribed in advance.

[3] The factor 2^{-1} guarantees that balls themselves are shape regular with $\sigma_{\mathrm{shp}} = 1$.

2. *There hold the relations*

$$\operatorname{meas}(\Omega) \leq h_\Omega^d \leq C(d)\sigma_{\mathrm{shp}}^d \operatorname{meas}(\Omega).$$

3. *If* $\Omega^\circ \cap \tilde{\Omega}^\circ = \emptyset$, *then*

$$h_\Omega + h_{\tilde{\Omega}} \leq 2\sigma_{\mathrm{shp}} \|\widetilde{x_{\mathrm{shp}}} - x_{\mathrm{shp}}\|_2.$$

Proof. Ad item 1: Since $h_\Omega > 0$, it is clear that $x_{\mathrm{shp}} \in \mathrm{Ball}_2(x_{\mathrm{shp}}, (2\sigma_{\mathrm{shp}})^{-1}h_\Omega)$. The subset $\mathrm{Ball}_2(x_{\mathrm{shp}}, (2\sigma_{\mathrm{shp}})^{-1}h_\Omega) \subseteq \Omega$ is open and thus contained in the largest open subset of Ω, which is the interior Ω°.

Ad item 2: Denote by $e_1, \dots, e_d \in \mathbb{R}^d$ the Euclidean unit vectors. For all $k \in \{1, \dots, d\}$, let $a_k := \inf_{x\in\Omega}\langle x, e_k\rangle_2$ and $b_k := \sup_{x\in\Omega}\langle x, e_k\rangle_2$. Then $B := \bigtimes_{k=1}^d [a_k, b_k]$ is an axes-parallel bounding box of Ω and we get

$$\operatorname{meas}(\Omega) \leq \operatorname{meas}(B) = \prod_{k=1}^d b_k - a_k = \prod_{k=1}^d \sup_{x,y\in\Omega} \langle x - y, e_k\rangle_2 \leq h_\Omega^d.$$

On the other hand, we have

$$C(d)\sigma_{\mathrm{shp}}^{-d} h_\Omega^d \overset{L.2.7}{=} \operatorname{meas}(\mathrm{Ball}_2(x_{\mathrm{shp}}, (2\sigma_{\mathrm{shp}})^{-1}h_\Omega)) \overset{D.2.16}{\leq} \operatorname{meas}(\Omega).$$

Ad item 3: Using step 1, we have

$$\mathrm{Ball}_2(x_{\mathrm{shp}}, (2\sigma_{\mathrm{shp}})^{-1}h_\Omega) \cap \mathrm{Ball}_2(\widetilde{x_{\mathrm{shp}}}, (2\sigma_{\mathrm{shp}})^{-1}h_{\tilde{\Omega}}) \subseteq \Omega^\circ \cap \tilde{\Omega}^\circ = \emptyset.$$

It then follows from L.2.7 that $(2\sigma_{\mathrm{shp}})^{-1}(h_\Omega + h_{\tilde{\Omega}}) \leq \|\widetilde{x_{\mathrm{shp}}} - x_{\mathrm{shp}}\|_2$.

□

When dealing with *multiple* sets $\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$, we have to account for the possibility of overlap.

Definition 2.18. *Let* $\sigma_{\mathrm{ovlp}} \geq 1$, $N \in \mathbb{N}$ *and* $\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$. *We say that the family* $\{\Omega_1, \dots, \Omega_N\}$ has overlap σ_{ovlp}, *if there holds*[4]

$$\max_{n\in\{1,\dots,N\}} \#\{m \in \{1, \dots, N\} \,|\, \Omega_m^\circ \cap \Omega_n^\circ \neq \emptyset\} \leq \sigma_{\mathrm{ovlp}}.$$

The number σ_{ovlp} allows us to quantify how many sets Ω_n can agglomerate at a given point in space. In particular, if a set Ω_n has "too many" neighbours Ω_m, their diameters cannot be arbitrarily large.

Lemma 2.19. *Let* $\sigma_{\mathrm{shp}}, \sigma_{\mathrm{ovlp}} \geq 1$. *Let* $N \in \mathbb{N}$ *and consider a family* $\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$ *of sets with shape regularity* σ_{shp} *and overlap* σ_{ovlp}. *Furthermore, for every* $n \in \{1, \dots, N\}$, *let* $x_n \in \Omega_n$ *be an incenter (cf. D.2.16). Then, for every index set* $I \subseteq \{1, \dots, N\}$ *with* $\#I > \sigma_{\mathrm{ovlp}}$, *there holds the bound*

$$\max_{n\in I} h_{\Omega_n} \leq 2\sigma_{\mathrm{shp}} \max_{m,n\in I} \|x_m - x_n\|_2.$$

[4] Note that we require the *interiors* of Ω_m and Ω_n to overlap. In particular, the intersection must have non-zero Lebesgue measure.

Proof. Let $n \in I$. Then there must exist an index $m \in I$ such that $\Omega_m^\circ \cap \Omega_n^\circ = \emptyset$, because otherwise we would get the contradiction

$$\sigma_{\text{ovlp}} < \#I = \#\{\tilde{m} \in I \,|\, \Omega_{\tilde{m}}^\circ \cap \Omega_n^\circ \neq \emptyset\} \leq \#\{\tilde{m} \in \{1,\dots,N\} \,|\, \Omega_{\tilde{m}}^\circ \cap \Omega_n^\circ \neq \emptyset\} \overset{D.2.18}{\leq} \sigma_{\text{ovlp}}.$$

We obtain

$$\text{Ball}_2(x_m, (2\sigma_{\text{shp}})^{-1} h_{\Omega_m}) \cap \text{Ball}_2(x_n, (2\sigma_{\text{shp}})^{-1} h_{\Omega_n}) \overset{L.2.17}{\subseteq} \Omega_m^\circ \cap \Omega_n^\circ = \emptyset,$$

which in turn implies $(2\sigma_{\text{shp}})^{-1}(h_{\Omega_m} + h_{\Omega_n}) \leq \|x_m - x_n\|_2$ (cf. L.2.7). Then,

$$h_{\Omega_n} \leq 2\sigma_{\text{shp}} \|x_m - x_n\|_2 \leq 2\sigma_{\text{shp}} \max_{\tilde{m},\tilde{n} \in I} \|x_{\tilde{m}} - x_{\tilde{n}}\|_2.$$

Taking the maximum over all $n \in I$, the desired bound follows.

□

Questions of overlap also arise when dealing with integrals. If a function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is integrated over subsets $\Omega_1, \Omega_2 \subseteq \mathbb{R}^d$ with $\Omega_1 \cap \Omega_2 \neq \emptyset$, then we have $\int_{\Omega_1 \cup \Omega_2} f \,\mathrm{d}x \neq \int_{\Omega_1} f \,\mathrm{d}x + \int_{\Omega_2} f \,\mathrm{d}x$ in general, because the contribution $\int_{\Omega_1 \cap \Omega_2} f \,\mathrm{d}x$ is counted twice on the right-hand side. To measure the discrepancy between the two, we can make use of the quantity σ_{ovlp} again.

Lemma 2.20. *Let $\Omega \subseteq \mathbb{R}^d$ be open, $k \in \mathbb{N}_0$, $p \in [1, \infty)$ and $f \in W^{k,p}(\Omega)$ (cf. D.2.37). Let $\sigma_{\text{ovlp}} \geq 1$, $N \in \mathbb{N}$ and consider a family $\Omega_1, \dots, \Omega_N \subseteq \Omega$ of sets with overlap σ_{ovlp}. Then, for all index sets $I \subseteq \{1,\dots,N\}$, there hold the bounds*

$$\|f\|_{W^{k,p}(\Omega_I)}^p \leq \sum_{n \in I} \|f\|_{W^{k,p}(\Omega_n)}^p \leq \sigma_{\text{ovlp}} \|f\|_{W^{k,p}(\Omega_I)}^p,$$

where $\Omega_I := \bigcup_{n \in I} \Omega_n \subseteq \Omega$.

Proof. For every subset $\omega \subseteq \Omega$, denote by $\mathbb{I}_\omega \in L^\infty(\Omega)$ its characteristic function (cf. Section 2.1). Since

$$\|f\|_{W^{k,p}(\omega)}^p = \sum_{|\alpha| \leq k} \|\mathrm{D}^\alpha f\|_{L^p(\omega)}^p = \int_\Omega \mathbb{I}_\omega(x) \sum_{|\alpha| \leq k} |(\mathrm{D}^\alpha f)(x)|^p \,\mathrm{d}x,$$

it suffices to show that, for almost all $x \in \Omega$,

$$\mathbb{I}_{\Omega_I}(x) \leq \sum_{n \in I} \mathbb{I}_{\Omega_n}(x) \leq \sigma_{\text{ovlp}} \mathbb{I}_{\Omega_I}(x).$$

The left-hand inequality follows readily from the identity $\Omega_I = \bigcup_{n \in I} \Omega_n$ and the fact that $\mathbb{I}_{\omega_1 \cup \omega_2} \leq \mathbb{I}_{\omega_1} + \mathbb{I}_{\omega_2}$, for all subsets $\omega_1, \omega_2 \subseteq \Omega$. To see the right-hand inequality, let $x \in \Omega \backslash M$, where $M := \bigcup_{n \in I} \partial\Omega_n$ satisfies $\text{meas}(M) = 0$. (The set M is defined such that $x \in \Omega_n$, if and only if $x \in \Omega_n^\circ$.) In the case $x \notin \Omega_I$, both sides of the inequality become zero.

In the remaining case $x \in \Omega_I$, we can find an index $n_0 \in I$ such that $x \in \Omega_{n_0} \setminus M \subseteq \Omega_{n_0}^\circ$. In particular,

$$\sum_{n\in I} \mathbb{I}_{\Omega_n}(x) = \#\{n \in I \,|\, x \in \Omega_n^\circ\} \leq \#\{n \in I \,|\, \Omega_n^\circ \cap \Omega_{n_0}^\circ \neq \emptyset\} \overset{D.2.18}{\leq} \sigma_{\text{ovlp}} = \sigma_{\text{ovlp}} \mathbb{I}_{\Omega_I}(x).$$

This finishes the proof.

□

We finish this section with the definition of *spread*. For mesh-based methods, spread is not an issue, since the computation domains is typically assumed to be bounded anyways. However, in the context of mesh-less methods, we are usually working with point clouds $x_1, \dots, x_N \in \mathbb{R}^d$ (or tiny bubbles $\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$) that need not be associated with any underlying domain. In order to rule out the case of individual points wandering off too far from the cloud's center, we assume a uniform bound on the cloud diameter. (Note that this does *not* rule out a scenario where the cloud as a whole wanders off to infinity.)

Definition 2.21. *Let $\sigma_{\text{sprd}} \geq 1$, $N \in \mathbb{N}$ and $\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$. We say that the family $\{\Omega_1, \dots, \Omega_N\}$* has spread σ_{sprd}, *if*

$$\operatorname{diam}_2\left(\bigcup_{n=1}^N \Omega_n\right) \leq \sigma_{\text{sprd}}.$$

Clearly, if the diameter of the family $\Omega_1, \dots, \Omega_N$ is bounded by σ_{sprd}, we can wrap it in an axes-parallel box $B \in \mathbb{B}$ (cf. D.2.8) with side length σ_{sprd}.

Lemma 2.22. *Let $\sigma_{\text{sprd}} \geq 1$, $N \in \mathbb{N}$ and $\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$ be a family of sets with spread σ_{sprd}. Then, there exists a box $B \in \mathbb{B}$ with the following properties:*

$$\Omega_1, \dots, \Omega_n \subseteq \overline{B}, \qquad \operatorname{diam}_2(B) = \sqrt{d}\sigma_{\text{sprd}}, \qquad \operatorname{meas}(B) = \sigma_{\text{sprd}}^d.$$

Proof. Abbreviate $\Omega := \bigcup_{n=1}^N \Omega_n \subseteq \mathbb{R}^d$. Using the Euclidean unit vectors $e_1, \dots, e_d \in \mathbb{R}^d$, we introduce the quantities

$$a_i := \inf_{x\in\Omega} \langle x, e_i\rangle_2, \qquad b_i := \sup_{x\in\Omega} \langle x, e_i\rangle_2, \qquad c_i := (a_i + b_i)/2.$$

and the box

$$B := \bigtimes_{i=1}^d [c_i - \sigma_{\text{sprd}}/2, c_i + \sigma_{\text{sprd}}/2) \in \mathbb{B}.$$

Note that

$$b_i - a_i = \sup_{x\in\Omega} \langle x, e_i\rangle_2 - \inf_{y\in\Omega} \langle y, e_i\rangle_2 = \sup_{x,y\in\Omega} \langle x - y, e_i\rangle_2 \leq \operatorname{diam}_2(\Omega) \overset{D.2.21}{\leq} \sigma_{\text{sprd}}.$$

In particular, for all $x \in \Omega$ and $i \in \{1, \dots, d\}$, we get

$$\langle x, e_i\rangle_2 \leq b_i = (a_i + b_i)/2 + (b_i - a_i)/2 \leq c_i + \sigma_{\text{sprd}}/2$$

and similarly $\langle x, e_i\rangle_2 \geq c_i - \sigma_{\text{sprd}}/2$. Therefore, we have the inclusions $\Omega_n \subseteq \Omega \subseteq \overline{B}$. Finally, we compute

$$\begin{aligned} \operatorname{diam}_2(B)^2 &= \sum_{k=1}^{d} ((c_i + \sigma_{\text{sprd}}/2) - (c_i - \sigma_{\text{sprd}}/2))^2 &= d\sigma_{\text{sprd}}^2, \\ \operatorname{meas}(B) &= \prod_{k=1}^{d} ((c_i + \sigma_{\text{sprd}}/2) - (c_i - \sigma_{\text{sprd}}/2)) &= \sigma_{\text{sprd}}^d. \end{aligned}$$

□

2.5 Affine transformations

Definition 2.23. *A function $F : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is called* affine (transformation)*, if there exist a regular matrix $A \in \mathbb{R}^{d\times d}$ and a vector $a \in \mathbb{R}^d$, such that*

$$\forall x \in \mathbb{R}^d : \qquad F(x) = Ax + a.$$

If $A = Q$, for some orthogonal matrix[5] $Q \in \mathbb{R}^{d\times d}$ with $\det(Q) = 1$, then we call F a rigid body transformation.

Note that the regularity of A is part of the definition of affinity. Furthermore, note that an affine transformation F is smooth and that there holds $\nabla F \equiv A$. In particular, for all $x \in \mathbb{R}^d$, we have $F(x) = (\nabla F)x + a$. The stability properties of affine transformations are tightly connected with the notion of shape regularity from D.2.16.

Lemma 2.24. *Let $F : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be affine.*

1. *Then F is bijective and its inverse $F^{-1} : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is again affine with*
$$\nabla(F^{-1}) = (\nabla F)^{-1}.$$

2. *Let $\hat{\Omega}, \Omega \subseteq \mathbb{R}^d$ be such that $F(\hat{\Omega}) = \Omega$. If $\hat{\Omega}$ is $\widehat{\sigma_{\text{shp}}}$-shape regular, for some $\widehat{\sigma_{\text{shp}}} \geq 1$, then Ω is σ_{shp}-shape regular, where*
$$\sigma_{\text{shp}} := \widehat{\sigma_{\text{shp}}} \|\nabla(F^{-1})\|_2 h_\Omega h_{\hat{\Omega}}^{-1} \geq 1.$$

3. *Let $\hat{\Omega}, \Omega \subseteq \mathbb{R}^d$ be such that $F(\hat{\Omega}) = \Omega$. Suppose that $\hat{\Omega}$ is $\widehat{\sigma_{\text{shp}}}$-shape regular and that Ω is σ_{shp}-shape regular, for some $\widehat{\sigma_{\text{shp}}}, \sigma_{\text{shp}} \geq 1$. Then, there hold the following bounds:*
$$\begin{aligned} C(d)^{-1}\sigma_{\text{shp}}^{-d} h_\Omega^d h_{\hat{\Omega}}^{-d} &\leq |\det \nabla F| \leq C(d)\widehat{\sigma_{\text{shp}}}^{\,d} h_\Omega^d h_{\hat{\Omega}}^{-d}, \\ h_\Omega h_{\hat{\Omega}}^{-1} &\leq \|\nabla F\|_2 \leq \widehat{\sigma_{\text{shp}}} h_\Omega h_{\hat{\Omega}}^{-1}. \end{aligned}$$

Proof. Ad item 1: If $F(x) = Ax + a$, then $F^{-1}(y) = A^{-1}y - A^{-1}a$.

Ad item 2: Write $F(x) = Ax + a$. According to D.2.16, we have $h_{\hat{\Omega}} \in (0, \infty)$. Therefore,

$$h_\Omega = \sup_{x,y\in\Omega} \|y - x\|_2 = \sup_{\hat{x},\hat{y}\in\hat{\Omega}} \|F(\hat{y}) - F(\hat{x})\|_2 = \sup_{\hat{x},\hat{y}\in\hat{\Omega}} \|A(\hat{y} - \hat{x})\|_2 \leq \|A\|_2 h_{\hat{\Omega}} < \infty.$$

[5] Recall that a regular matrix $Q \in \mathbb{R}^{d\times d}$ is orthogonal, if there holds $Q^{-1} = Q^T$. In this case $|\det(Q)| = 1$.

Similarly, $0 < h_{\hat{\Omega}} \leq \|A^{-1}\|_2 h_\Omega$, so that $h_\Omega > 0$ as well.

Next, owing to D.2.16 again, we may pick a point $\widehat{x_{\mathrm{shp}}} \in \hat{\Omega}$ such that

$$\mathrm{Ball}_2(\widehat{x_{\mathrm{shp}}}, (2\widehat{\sigma_{\mathrm{shp}}})^{-1} h_{\hat{\Omega}}) \subseteq \hat{\Omega}.$$

Set $\sigma_{\mathrm{shp}} := \widehat{\sigma_{\mathrm{shp}}} \|A^{-1}\|_2 h_\Omega h_{\hat{\Omega}}^{-1}$ and $x_{\mathrm{shp}} := F(\widehat{x_{\mathrm{shp}}}) \in \Omega$. We want to show that

$$\mathrm{Ball}_2(x_{\mathrm{shp}}, (2\sigma_{\mathrm{shp}})^{-1} h_\Omega) \subseteq \Omega.$$

To this end let $x \in \mathrm{Ball}_2(x_{\mathrm{shp}}, (2\sigma_{\mathrm{shp}})^{-1} h_\Omega)$. Then, its pre-image $\hat{x} := F^{-1}(x)$ satisfies

$$\begin{aligned} \|\hat{x} - \widehat{x_{\mathrm{shp}}}\|_2 &= \|F^{-1}(x) - F^{-1}(x_{\mathrm{shp}})\|_2 = \|A^{-1}(x - x_{\mathrm{shp}})\|_2 \leq \|A^{-1}\|_2 \|x - x_{\mathrm{shp}}\|_2 \\ &\leq \|A^{-1}\|_2 (2\sigma_{\mathrm{shp}})^{-1} h_\Omega = (2\widehat{\sigma_{\mathrm{shp}}})^{-1} h_{\hat{\Omega}}. \end{aligned}$$

We obtain

$$\hat{x} \in \mathrm{Ball}_2(\widehat{x_{\mathrm{shp}}}, (2\widehat{\sigma_{\mathrm{shp}}})^{-1} h_{\hat{\Omega}}) \subseteq \hat{\Omega}$$

and ultimately $x = F(\hat{x}) \in F(\hat{\Omega}) = \Omega$.

Ad item 3: Write $F(x) = Ax + a$. To bound the determinant, we use the well-known identity $|\det A| = \mathrm{meas}(\Omega)\mathrm{meas}(\hat{\Omega})^{-1}$ (e.g., [Rud87, Theorem 2.20, Lemma 2.23]):

$$C(d)^{-1} \sigma_{\mathrm{shp}}^{-d} h_\Omega^d h_{\hat{\Omega}}^{-d} \overset{L.2.17}{\leq} \frac{\mathrm{meas}(\Omega)}{\mathrm{meas}(\hat{\Omega})} = |\det A| = \frac{\mathrm{meas}(\Omega)}{\mathrm{meas}(\hat{\Omega})} \overset{L.2.17}{\leq} C(d) \widehat{\sigma_{\mathrm{shp}}}^d h_\Omega^d h_{\hat{\Omega}}^{-d}.$$

The lower bound $h_\Omega h_{\hat{\Omega}}^{-1} \leq \|A\|_2$ was already shown in step 2. Finally, D.2.16 allows us to pick a point $\widehat{x_{\mathrm{shp}}} \in \hat{\Omega}$ such $\overline{\mathrm{Ball}}_2(\widehat{x_{\mathrm{shp}}}, \hat{r}) \subseteq \hat{\Omega}$, for all $\hat{r} \in (0, (2\widehat{\sigma_{\mathrm{shp}}})^{-1} h_{\hat{\Omega}})$. We compute

$$\begin{aligned} \|A\|_2 &= \sup_{\|\xi\|_2 = 1} \|A\xi\|_2 = (2\hat{r})^{-1} \sup_{\|\xi\|_2 = 1} \|F(\widehat{x_{\mathrm{shp}}} + \hat{r}\xi) - F(\widehat{x_{\mathrm{shp}}} - \hat{r}\xi)\|_2 \\ &\leq (2\hat{r})^{-1} \sup_{\hat{x}, \hat{y} \in \hat{\Omega}} \|F(\hat{y}) - F(\hat{x})\|_2 = (2\hat{r})^{-1} h_\Omega. \end{aligned}$$

Sending $\hat{r} \to (2\widehat{\sigma_{\mathrm{shp}}})^{-1} h_{\hat{\Omega}}$, the upper bound $\|A\|_2 \leq \widehat{\sigma_{\mathrm{shp}}} h_\Omega h_{\hat{\Omega}}^{-1}$ follows.

□

Note that we can get analogous bounds for $|\det \nabla(F^{-1})|$ and $\|\nabla(F^{-1})\|_2$ by reversing the roles of $\hat{\Omega}$ and Ω.

2.6 Some function spaces

Definition 2.25. *Let $\Omega \subseteq \mathbb{R}^d$ be open and let $p \in \mathbb{N}_0$. We define the space of* polynomials of degree p,

$$\mathbb{P}^p(\Omega) := \left\{ v : \Omega \longrightarrow \mathbb{R} \;\middle|\; \exists (c_\alpha)_{|\alpha| \leq p} \subseteq \mathbb{R} : \forall x \in \Omega : \; v(x) = \sum_{|\alpha| \leq p} c_\alpha x^\alpha \right\}.$$

Furthermore, we set

$$\mathbb{P}^{-1}(\Omega) := \{ v : \Omega \longrightarrow \mathbb{R} \,|\, v \equiv 0 \}.$$

We mention that the dimension of this space is given by

$$\dim(\mathbb{P}^p(\Omega)) = \binom{d+p}{d} = \frac{(d+p)!}{d!p!} = \mathcal{O}(p^d),$$

if $p \in \mathbb{N}_0$, and $\dim(\mathbb{P}^{-1}(\Omega)) = 0$ (see, e.g., [EG04, Section 1.2.3]).

Note that a function $v \in \mathbb{P}^p(\Omega)$ is represented by the same coefficient set $(c_\alpha)_{|\alpha|\leq p}$ on all of Ω, even if Ω is not connected. In order to allow for distinct coefficient sets on distinct connected components of Ω, we introduce the following, slightly larger space:

Definition 2.26. *Let $\Omega \subseteq \mathbb{R}^d$ be open and let $p \in \mathbb{N}_0 \cup \{-1\}$. We define the space*

$$\mathbb{P}^p_{\text{conn}}(\Omega) := \{v : \Omega \longrightarrow \mathbb{R} \,|\, \forall \text{ connected components } \omega \subseteq \Omega : \; v|_\omega \in \mathbb{P}^p(\omega)\}.$$

For the next definition, we remind the reader of the notion of uniform continuity: Let $\Omega \subseteq \mathbb{R}^d$ be an open set. A function $v : \Omega \longrightarrow \mathbb{R}$ is said to be *continuous on* Ω, if the following statement is true:

$$\forall x \in \Omega : \forall \varepsilon > 0 : \exists \delta > 0 : \qquad \sup_{\substack{y \in \Omega: \\ \|y-x\|_2 \leq \delta}} |v(y) - v(x)| \leq \varepsilon.$$

On the other hand, the function v is *uniformly continuous on* Ω, if

$$\forall \varepsilon > 0 : \exists \delta > 0 : \qquad \sup_{\substack{x,y \in \Omega: \\ \|y-x\|_2 \leq \delta}} |v(y) - v(x)| \leq \varepsilon.$$

In this case, v can be extended to a continuous function $v : \overline{\Omega} \longrightarrow \mathbb{R}$ (e.g., [AF03, Section 1.28]). If, additionally, v is bounded, then its extension to $\overline{\Omega}$ is bounded as well.

Definition 2.27. *Let $\Omega \subseteq \mathbb{R}^d$ be open and $k \in \mathbb{N}_0 \cup \{\infty\}$. We set*[6]

$$\begin{aligned} C^k(\Omega) &:= \{v : \Omega \longrightarrow \mathbb{R} \,|\, v \text{ is } k \text{ times continuously differentiable}\}, \\ C^k(\overline{\Omega}) &:= \{v \in C^k(\Omega) \,|\, \forall |\alpha| \leq k : \mathrm{D}^\alpha v \text{ unif. cont. and bounded}\}, \\ C^k_0(\Omega) &:= \{v \in C^k(\Omega) \,|\, \operatorname{supp}(v) \Subset \Omega\}. \end{aligned}$$

The functions $v \in C^\infty_0(\Omega)$ are called test functions.

2.7 Lebesgue spaces

Lebesgue spaces are a core pillar of the theory of partial differential equations. More details can be found, e.g., in [Fle77, Section 5.13], [Rud87, Chapter 3], [AF03, Chapter 2], [Bre11, Chapter 4] or [Leo17, Appendix B.7.].

[6] See Section 2.1 for the definition of the relation "$A \Subset B$".

Definition 2.28. *Let $I \subseteq \mathbb{N}$ and $p \in [1,\infty]$. We define the* sequence space

$$l^p(I) := \{v = (v_i)_{i\in I} \in \mathbb{R}^I \mid \|v\|_{l^p(I)} < \infty\},$$

where

$$\|v\|_{l^p(I)} := \begin{cases} (\sum_{i\in I} |v_i|^p)^{1/p} & \text{if } p \in [1,\infty) \\ \sup_{i\in I} |v_i| & \text{if } p = \infty \end{cases}.$$

In the case $p = 2$, we set

$$\langle v, w\rangle_{l^2(I)} := \sum_{i\in I} v_i w_i.$$

If the set I has the form $I = \{1,\dots,N\}$, for some $N \in \mathbb{N}$, then we abbreviate $l^p(N) := l^p(\{1,\dots,N\})$.

Lemma 2.29. *Let $I \subseteq \mathbb{N}$ and let $p,q,r \in [1,\infty]$ with $1/p + 1/q = 1/r$. Then, for all $v \in l^p(I)$ and $w \in L^q(I)$, there holds the* Hölder inequality

$$\|vw\|_{l^r(I)} \le \|v\|_{l^p(I)} \|w\|_{l^q(I)}.$$

Proof. See, e.g., [AF03, Corollary 2.5]. □

In the special case $p = q = 2$ and $r = \infty$, Hölder's inequality is also known as the Cauchy-Schwarz inequality.

Definition 2.30. *Let $\Omega \subseteq \mathbb{R}^d$ be a measurable set and $p \in [1,\infty]$. We define the* Lebesgue space

$$L^p(\Omega) := \{v : \Omega \longrightarrow \mathbb{R} \mid v \text{ is measurable}, \ \|v\|_{L^p(\Omega)} < \infty\},$$

where

$$\|v\|_{L^p(\Omega)} := \begin{cases} (\int_\Omega |v(x)|^p \,\mathrm{d}x)^{1/p} & \text{if } p \in [1,\infty) \\ \operatorname{ess\,sup}_{x\in\Omega} |v(x)| & \text{if } p = \infty \end{cases}.$$

In the case $p = 2$, we set

$$\langle v, w\rangle_{L^2(\Omega)} := \int_\Omega v(x) w(x) \,\mathrm{d}x.$$

Lemma 2.31. *Let $\Omega \subseteq \mathbb{R}^d$ be measurable and $p,q,r \in [1,\infty]$ with $1/p + 1/q = 1/r$. Then, for all $v \in L^p(\Omega)$ and $w \in L^q(\Omega)$, there hold $vw \in L^r(\Omega)$ and the* Hölder inequality

$$\|vw\|_{L^r(\Omega)} \le \|v\|_{L^p(\Omega)} \|w\|_{L^q(\Omega)}.$$

Proof. See, e.g., [AF03, Corollary 2.5]. □

An immediate consequence of Hölder's inequality is the following result:

Lemma 2.32. *Let $\Omega \subseteq \mathbb{R}^d$ be measurable with $\operatorname{meas}(\Omega) < \infty$. Then, for all $p,\tilde{p} \in [1,\infty]$ with $\tilde{p} \le p$, there holds*

$$L^p(\Omega) \subseteq L^{\tilde{p}}(\Omega).$$

Finally, we need to know what happens to the space $L^p(\Omega)$ if we perform an affine coordinate transformation (cf. D.2.23).

Lemma 2.33. *Let $F : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be affine and let $\hat{\Omega}, \Omega \subseteq \mathbb{R}^d$ be such that $F(\hat{\Omega}) = \Omega$. Let $p \in [1, \infty]$ and $v : \Omega \longrightarrow \mathbb{R}$ be a measurable function. Then there holds the following equivalence[7]:*

$$v \in L^p(\Omega) \quad \Leftrightarrow \quad v \circ F \in L^p(\hat{\Omega}).$$

Furthermore, there holds the relation[8]

$$|\det \nabla F|^{1/p} \|v \circ F\|_{L^p(\hat{\Omega})} = \|v\|_{L^p(\Omega)}.$$

Proof. The case $p = \infty$ is trivial and the case $p \in [1, \infty)$ follows from the well-known (e.g., [Rud87, Theorem 7.26]) transformation rule for integrals,

$$\int_{\Omega} v \, \mathrm{d}x = \int_{\hat{\Omega}} v \circ F \cdot |\det \nabla F| \, \mathrm{d}x.$$

□

2.8 Sobolev spaces

The literature on Sobolev spaces is vast and we can only name a few: [Gri85, Chapter 1], [McL00, Chapter 3], [GT01, Chapter 7], [EG04, Appendix B.3], [BS08, Chapter 1], [Eva10, Section 5.2.], [Bre11, Chapter 8] or [Neč12, Chapter 2]. Whole books dedicated to the subject include, among others, [Maz85], [AF03] and [Leo17].

2.8.1 Definition

A fundamental concept in the theory of Sobolev spaces is the notion of *weak derivatives*. Given an open set $\Omega \subseteq \mathbb{R}^d$, a function $v \in C^k(\Omega)$, a test function $w \in C_0^\infty(\Omega)$ and a multi-index $\alpha \in \mathbb{N}_0$, it is well known that there holds the formula of partial integration:

$$\int_{\Omega} v(\mathrm{D}^\alpha w) \, \mathrm{d}x = (-1)^{|\alpha|} \int_{\Omega} (\mathrm{D}^\alpha v) w \, \mathrm{d}x.$$

However, the formula also makes sense if we only require[9] $v, \mathrm{D}^\alpha v \in L^p(\Omega)$, for some $p \in [1, \infty]$. This observation allows us to generalize the notion of differentiability significantly:

[7] More precisely, we should write $v \circ (F|_{\hat{\Omega}}) \in L^p(\hat{\Omega})$.
[8] In the case $p = \infty$, the convention $1/\infty := 0$ is used.
[9] In fact, it suffices to require $v, \mathrm{D}^\alpha v \in L^1_{\mathrm{loc}}(\Omega)$, where $L^1_{\mathrm{loc}}(\Omega) := \{v : \Omega \longrightarrow \mathbb{R} \,|\, \forall \omega \Subset \Omega : \ v|_\omega \in L^1(\omega)\}$ denotes the space of *locally integrable functions*. However, the simpler setting of $L^p(\Omega)$ is sufficient for the purpose of the present work.

Definition 2.34. *Let $\Omega \subseteq \mathbb{R}^d$ be an open set, $p \in [1,\infty]$, $v \in L^p(\Omega)$ and $\alpha \in \mathbb{N}_0^d$. We say that v has an* α-th weak derivative, *if there exists a function $v_\alpha \in L^p(\Omega)$ such that, for all $w \in C_0^\infty(\Omega)$, there holds the identity*

$$\int_\Omega v(\mathrm{D}^\alpha w)\,\mathrm{d}x = (-1)^{|\alpha|}\int_\Omega v_\alpha w\,\mathrm{d}x.$$

As we shall see, there can be at most one such function v_α. To this end, we need the following well-known result, which goes by the name of *Fundamental Lemma of Calculus of Variations* or *du Bois-Reymond Lemma*:

Lemma 2.35. *Let $\Omega \subseteq \mathbb{R}^d$ be open, $k \in \mathbb{N}_0$ and $p \in [1,\infty]$. Let $v \in L^p(\Omega)$ be such that*

$$\int_\Omega v(\mathrm{D}^\alpha w)\,\mathrm{d}x = 0,$$

for all $w \in C_0^\infty(\Omega)$ and all $|\alpha| = k$. Then, there holds[10] *$v \in \mathbb{P}^{k-1}_{\mathrm{conn}}(\Omega)$.*

Proof. Proofs of the cases $k \in \{0,1\}$ can be found, e.g., in [AF03, Lemma 3.31] and [GH96, Chapter 1, Section 2., Subsection 2.3., Lemma 4], respectively. □

We quickly summarize the basic properties of weak derivatives.

Lemma 2.36. *Let $\Omega \subseteq \mathbb{R}^d$ be open, $p \in [1,\infty]$ and $\alpha, \beta \in \mathbb{N}_0^d$.*

1. Uniqueness: *If $u \in L^p(\Omega)$ has an α-th weak derivative $u_\alpha \in L^p(\Omega)$, then u_α is unique. In particular, the following notation is justified:*
$$\mathrm{D}^\alpha u := u_\alpha \in L^p(\Omega).$$
2. Commutativity: *Suppose $u \in L^p(\Omega)$ has an α-th and a β-th weak derivative $\mathrm{D}^\alpha u$, $\mathrm{D}^\beta u \in L^p(\Omega)$. If either one of the weak derivatives $\mathrm{D}^{\alpha+\beta}u, \mathrm{D}^\alpha(\mathrm{D}^\beta v), \mathrm{D}^\beta(\mathrm{D}^\alpha u) \in L^p(\Omega)$ exists, then all of them exist and coincide.*
3. Linearity: *If $u, v \in L^p(\Omega)$ have α-th weak derivatives $\mathrm{D}^\alpha u, \mathrm{D}^\alpha v \in L^p(\Omega)$, then, for all $a, b \in \mathbb{R}$, the function $au + bv \in L^p(\Omega)$ has an α-th weak derivative given by $\mathrm{D}^\alpha(au + bv) = a(\mathrm{D}^\alpha u) + b(\mathrm{D}^\alpha v) \in L^p(\Omega)$.*
4. Restrictions: *Consider an open subset $\omega \subseteq \Omega$. If $u \in L^p(\Omega)$ has an α-th weak derivative $\mathrm{D}^\alpha u \in L^p(\Omega)$, then the restriction $u|_\omega \in L^p(\omega)$ has an α-th weak derivative given by $\mathrm{D}^\alpha(u|_\omega) = (\mathrm{D}^\alpha u)|_\omega \in L^p(\omega)$.*

Proof. To see item 1, let $v_\alpha, \widetilde{v_\alpha} \in L^p(\Omega)$ be such that $\int_\Omega v(\mathrm{D}^\alpha w)\,\mathrm{d}x = (-1)^{|\alpha|}\int_\Omega v_\alpha w\,\mathrm{d}x$ and $\int_\Omega v(\mathrm{D}^\alpha w)\,\mathrm{d}x = (-1)^{|\alpha|}\int_\Omega \widetilde{v_\alpha} w\,\mathrm{d}x$, for all $w \in C_0^\infty(\Omega)$. Then $\int_\Omega (v_\alpha - \widetilde{v_\alpha})w\,\mathrm{d}x = 0$, for all $w \in C_0^\infty(\Omega)$, so that $v_\alpha = \widetilde{v_\alpha}$, by L.2.35. The remaining items are straightforward computations. □

[10] Recall from D.2.25 and D.2.26 that $\mathbb{P}^{-1}_{\mathrm{conn}}(\Omega) = \{0\}$.

Definition 2.37. *Let $\Omega \subseteq \mathbb{R}^d$ be open, $k \in \mathbb{N}_0$ and $p \in [1, \infty]$. We define the* Sobolev space

$$W^{k,p}(\Omega) := \{v \in L^p(\Omega) \,|\, \forall |\alpha| \leq k : \exists\, \mathrm{D}^\alpha v \in L^p(\Omega)\}.$$

The space $W^{k,p}(\Omega)$ is equipped with the norm

$$\|v\|_{W^{k,p}(\Omega)} := \begin{cases} (\sum_{|\alpha|\leq k} \|\mathrm{D}^\alpha v\|^p_{L^p(\Omega)})^{1/p} & \text{if } p \in [1,\infty) \\ \max_{|\alpha|\leq k} \|\mathrm{D}^\alpha v\|_{L^\infty(\Omega)} & \text{if } p = \infty \end{cases}.$$

Furthermore, for all $l \in \{0, \ldots, k\}$, we define the seminorm

$$|v|_{W^{l,p}(\Omega)} := \begin{cases} (\sum_{|\alpha|=l} \|\mathrm{D}^\alpha v\|^p_{L^p(\Omega)})^{1/p} & \text{if } p \in [1,\infty) \\ \max_{|\alpha|=l} \|\mathrm{D}^\alpha v\|_{L^\infty(\Omega)} & \text{if } p = \infty \end{cases}.$$

In the case $p = 2$, we write $H^k(\Omega) := W^{k,2}(\Omega)$ and set

$$\langle v, w\rangle_{H^k(\Omega)} := \sum_{|\alpha|\leq k} \langle \mathrm{D}^\alpha v, \mathrm{D}^\alpha w\rangle_{L^2(\Omega)}.$$

Remark 2.38. *For general subsets $\Omega \subseteq \mathbb{R}^d$, which are not necessarily open, we adopt the convention from [Cia78, Remark 2.1.3.] and abbreviate*

$$W^{k,p}(\Omega) := W^{k,p}(\Omega^\circ).$$

From a functional analytic point of view, the spaces $W^{k,p}(\Omega)$ and $H^k(\Omega)$ have a nice structure.

Lemma 2.39. *Let $\Omega \subseteq \mathbb{R}^d$ be open, $k \in \mathbb{N}_0$ and $p \in [1, \infty]$. The space $W^{k,p}(\Omega)$ is a Banach space. Furthermore, $H^k(\Omega)$ is a Hilbert space.*

Proof. See, e.g., [Eva10, Section 5.2., Theorem 2]. □

The concept of "weak differentiability" from D.2.34 generalizes "classic differentiability". In fact, if a function $v \in C^k(\Omega)$ is such that all of its *classic* derivatives $\mathrm{D}^\alpha v \in C^0(\Omega)$, $|\alpha| \leq k$, lie in $L^p(\Omega)$ (for some $p \in [1, \infty]$), then all *weak* derivatives $\mathrm{D}^\alpha v \in L^p(\Omega)$, $|\alpha| \leq k$, exist and coincide with the corresponding classic ones.

In contrast to possible other forms[11] of "generalized differentiability", weak derivatives still allow us to infer "$v \equiv \text{const}$" from "$\nabla v = 0$". (As for classic derivatives, a proof of this fact can be found, e.g., in [Fol02, Theorem 2.42].) To deal with the issue of non-connected sets Ω, we remind the reader of the space $\mathbb{P}^p_{\mathrm{conn}}(\Omega)$ from D.2.26.

Lemma 2.40. *Let $\Omega \subseteq \mathbb{R}^d$ be an open set. Furthermore, let $k \in \mathbb{N}_0$ and $p \in [1, \infty]$. Then, for all $v \in W^{k,p}(\Omega)$, there holds the following equivalence:*

$$|v|_{W^{k,p}(\Omega)} = 0 \qquad \Leftrightarrow \qquad v \in \mathbb{P}^{k-1}_{\mathrm{conn}}(\Omega).$$

Proof. The direction "$\Leftarrow$" is trivial and the direction "$\Rightarrow$" follows immediately from L.2.35. □

[11] One could, for example, define "generalized differentiability" as differentiability *almost* everywhere in Ω.

Many more properties of classic derivatives still hold true for weak derivatives. The next theorem is a handy tool in these types of proofs.

Theorem 2.41. *Let $k \in \mathbb{N}_0$ and $p \in [1, \infty)$.*

1. *Let $\Omega \subseteq \mathbb{R}^d$ be open. Then, for every $v \in W^{k,p}(\Omega)$, there exists a sequence*
$$(v_n)_{n\in\mathbb{N}} \subseteq W^{k,p}(\Omega) \cap C^\infty(\Omega),$$
such that[12] $\|v - v_n\|_{W^{k,p}(\Omega)} \xrightarrow{n} 0$.

2. *Let $\Omega \subseteq \mathbb{R}^d$ be an open, bounded set with a Lipschitz boundary. Then, for every $v \in W^{k,p}(\Omega)$, there exists a sequence*
$$(v_n)_{n\in\mathbb{N}} \subseteq W^{k,p}(\Omega) \cap C^\infty(\overline{\Omega}),$$
such that[13] $\|v - v_n\|_{W^{k,p}(\Omega)} \xrightarrow{n} 0$.

3. *Let $\Omega \subseteq \mathbb{R}^d$ be open. Then, for every $v \in W^{k,p}(\Omega)$, there exists a sequence*
$$(v_n)_{n\in\mathbb{N}} \subseteq C_0^\infty(\mathbb{R}^d),$$
such that $\|v - v_n\|_{L^p(\Omega)} \xrightarrow{n} 0$ and $\|v - v_n\|_{W^{k,p}(\omega)} \xrightarrow{n} 0$, for all $\omega \Subset \Omega$.

4. *For every $v \in W^{k,p}(\mathbb{R}^d)$, there exists a sequence*
$$(v_n)_{n\in\mathbb{N}} \subseteq C_0^\infty(\mathbb{R}^d),$$
such that $\|v - v_n\|_{W^{k,p}(\mathbb{R}^d)} \xrightarrow{n} 0$.

Proof. Item 1 is attributed to [MS64]. Item 2 can easily be derived from item 1 via an extension operator $E_\Omega : W^{k,p}(\Omega) \longrightarrow W^{k,p}(\mathbb{R}^d)$ (see D.2.48 below). Item 3 can be found in [Mv98, Chapter 5, Section 4, Theorem 1] or [Bre11, Theorem 9.2]. As for item 4, see [Leo17, Remark 11.26.].

□

2.8.2 Product rule

T.2.41 can be used, for example, to derive Leibniz' product rule for weak derivatives.

Lemma 2.42. *Let $\Omega \subseteq \mathbb{R}^d$ be open, $k \in \mathbb{N}_0$ and $p, q, r \in [1, \infty]$ with $1/p + 1/q = 1/r$. Let $u \in W^{k,p}(\Omega)$ and $v \in W^{k,q}(\Omega)$. Then $uv \in W^{k,r}(\Omega)$ and, for all $|\alpha| \leq k$, there holds the Leibniz formula*
$$\mathrm{D}^\alpha(uv) = \sum_{\beta\leq\alpha} \binom{\alpha}{\beta} (\mathrm{D}^\beta u)(\mathrm{D}^{\alpha-\beta} v) \in L^r(\Omega).$$
In particular, for all $l \in \{0, \dots, k\}$, there holds the bound
$$|uv|_{W^{l,r}(\Omega)} \leq C(d, \Omega, k, p, q, r) \sum_{j=0}^{l} |u|_{W^{j,p}(\Omega)} |v|_{W^{l-j,q}(\Omega)}.$$

[12] In other words, $W^{k,p}(\Omega) \cap C^\infty(\Omega)$ is a dense subspace of $W^{k,p}(\Omega)$.
[13] In other words, $W^{k,p}(\Omega) \cap C^\infty(\overline{\Omega})$ is a dense subspace of $W^{k,p}(\Omega)$.

Proof. We only prove the case $k = 1$. Let $u \in W^{1,p}(\Omega)$ and $v \in W^{1,q}(\Omega)$. Using Hölder's inequality (cf. L.2.31), we get, for all $i \in \{1, \dots, d\}$,

$$\begin{array}{rcll} \|uv\|_{L^r(\Omega)} & \leq & \|u\|_{L^p(\Omega)}\|v\|_{L^q(\Omega)} & < \infty, \\ \|(\partial_i u)v + u(\partial_i v)\|_{L^r(\Omega)} & \leq & |u|_{W^{1,p}(\Omega)}\|v\|_{L^q(\Omega)} + \|u\|_{L^p(\Omega)}|v|_{W^{1,q}(\Omega)} & < \infty. \end{array}$$

It remains to prove that $(\partial_i u)v + u(\partial_i v) \in L^r(\Omega)$ is indeed the i-th weak partial derivative of $uv \in L^r(\Omega)$ in the sense of D.2.34. To this end, we introduce the quantities

$$\tilde{p} := \begin{cases} p/r & \text{if } r < \infty \\ 2 & \text{if } r = \infty \end{cases}, \qquad \tilde{q} := \begin{cases} q/r & \text{if } r < \infty \\ 2 & \text{if } r = \infty \end{cases}.$$

Keeping in mind that $p, q, r \in [1, \infty]$ and that $1/p + 1/q = 1/r$, the relevant properties of $\tilde{p}$ and $\tilde{q}$ are the following:

$$\tilde{p}, \tilde{q} \in [1, \infty], \qquad 1/\tilde{p} + 1/\tilde{q} = 1, \qquad \tilde{p} \leq p, \qquad \tilde{q} \leq q, \qquad (\tilde{p} < \infty \text{ or } \tilde{q} < \infty).$$

Now, let $w \in C_0^\infty(\Omega)$ be given. Since $\operatorname{supp}(w) \Subset \Omega$ (cf. D.2.27), we can find an open, bounded set $\omega \subseteq \Omega$ such that $\operatorname{supp}(w) \subseteq \omega$. Using $\tilde{p} \leq p$ and $\tilde{q} \leq q$, we know from L.2.32 that

$$u|_\omega \in W^{1,p}(\omega) \subseteq W^{1,\tilde{p}}(\omega), \qquad v|_\omega \in W^{1,q}(\omega) \subseteq W^{1,\tilde{q}}(\omega).$$

W.l.o.g., let us assume that $\tilde{q} < \infty$ (recall that $\tilde{p} < \infty$ or $\tilde{q} < \infty$ or both). Then we may apply T.2.41 to the space $W^{1,\tilde{q}}(\omega)$ and obtain a sequence $(v_n)_{n\in\mathbb{N}} \subseteq W^{1,\tilde{q}}(\omega) \cap C^\infty(\omega)$ such that $\|v - v_n\|_{W^{1,\tilde{q}}(\omega)} \overset{n}{\to} 0$. In particular, using Hölder's inequality from L.2.31 for the conjugate exponents $\tilde{p}$ and $\tilde{q}$, we have

$$\begin{array}{rcll} |\int_\omega uv(\partial_i w)\,\mathrm{d}x - \int_\omega uv_n(\partial_i w)\,\mathrm{d}x| & \leq & \|u\|_{L^{\tilde{p}}(\omega)}\|v - v_n\|_{L^{\tilde{q}}(\omega)}|w|_{W^{1,\infty}(\omega)} & \overset{n}{\to} 0, \\ |\int_\omega u(\partial_i v)w\,\mathrm{d}x - \int_\omega u(\partial_i v_n)w\,\mathrm{d}x| & \leq & \|u\|_{L^{\tilde{p}}(\omega)}|v - v_n|_{W^{1,\tilde{q}}(\omega)}\|w\|_{L^\infty(\omega)} & \overset{n}{\to} 0, \\ |\int_\omega (\partial_i u)vw\,\mathrm{d}x - \int_\omega (\partial_i u)v_n w\,\mathrm{d}x| & \leq & |u|_{W^{1,\tilde{p}}(\omega)}\|v - v_n\|_{L^{\tilde{q}}(\omega)}\|w\|_{L^\infty(\omega)} & \overset{n}{\to} 0. \end{array}$$

Note that the integral $\int_\omega (\partial_i u)v_n w\,\mathrm{d}x$ on the third line is susceptible for partial integration (i.e., D.2.34), since $v_n w \in C_0^\infty(\omega)$. The integrand then becomes $-u\partial_i(v_n w)$ and we can apply the classic Leibniz rule $\partial_i(v_n w) = (\partial_i v_n)w + v_n(\partial_i w)$. Putting everything together, we compute

$$\int_\Omega ((\partial_i u)v + u(\partial_i v))w\,\mathrm{d}x = \int_\omega (\partial_i u)vw + u(\partial_i v)w\,\mathrm{d}x \overset{n}{\leftarrow} \int_\omega (\partial_i u)v_n w + u(\partial_i v_n)w\,\mathrm{d}x$$

$$\overset{\text{p.i.}}{=} \int_\omega -u\partial_i(v_n w) + u(\partial_i v_n)w\,\mathrm{d}x \overset{\text{Lbz.}}{=} -\int_\omega uv_n(\partial_i w)\,\mathrm{d}x \overset{n}{\to} -\int_\omega uv(\partial_i w)\,\mathrm{d}x = -\int_\Omega uv(\partial_i w)\,\mathrm{d}x.$$

This concludes the proof. □

2.8.3 Chain rule

Next, we develop the chain rule for the weak derivatives. The chain rule is often used in the context of a *scaling argument*, where stability/error bounds on a family of subsets $\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$ are reduced to stability/error bounds on a single *reference element* $\hat{\Omega} \subseteq \mathbb{R}^d$. Similar results on the chain rule can be found, e.g., in [AF03, Theorem 3.41], [EG04, Lemma 1.101.], [Bre11, Proposition 9.6] or [Bra13, Lemma 6.6].

For the next lemma, we remind the reader of D.2.23 and D.2.16, where we defined affine transformations and shape regular sets.

Lemma 2.43. *Let $F : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be affine and let $\hat{\Omega}, \Omega \subseteq \mathbb{R}^d$ be sets with $F(\hat{\Omega}) = \Omega$. Suppose that $\hat{\Omega}$ has shape regularity $\widehat{\sigma_{\mathrm{shp}}} \geq 1$ and that Ω has shape regularity $\sigma_{\mathrm{shp}} \geq 1$. Furthermore, let $k \in \mathbb{N}_0$ and $p \in [1, \infty]$. Then, for every measurable function $v : \Omega \longrightarrow \mathbb{R}$, there holds the following equivalence[14]:*

$$v \in W^{k,p}(\Omega) \quad \Leftrightarrow \quad v \circ F \in W^{k,p}(\hat{\Omega}).$$

In this case, for all $l \in \{0, \dots, k\}$ and all $j_1, \dots, j_l \in \{1, \dots, d\}$, we have[15]

$$\partial_{j_1} \cdots \partial_{j_l} (v \circ F) = \sum_{i_1, \dots, i_l = 1}^{d} (\partial_{i_1} \cdots \partial_{i_l} v) \circ F \cdot (\nabla F)_{i_1 j_1} \cdots (\nabla F)_{i_l j_l} \in L^p(\hat{\Omega}).$$

Furthermore, there exists a constant $C = C(d, k, p, \hat{\Omega}, \widehat{\sigma_{\mathrm{shp}}}, \sigma_{\mathrm{shp}}) \geq 1$, such that, for all $l \in \{0, \dots, k\}$,

$$C^{-1} h_\Omega^l |v|_{W^{l,p}(\Omega)} \leq h_\Omega^{d/p} |v \circ F|_{W^{l,p}(\hat{\Omega})} \leq C h_\Omega^l |v|_{W^{l,p}(\Omega)}.$$

Proof. We only prove the case $k = 1$. Abbreviate $A := \nabla F \in \mathbb{R}^{d \times d}$ and let $v \in W^{1,p}(\Omega)$ and $j \in \{1, \dots, d\}$. Using L.2.33, a straightforward computation proves that $v \circ F \in L^p(\hat{\Omega})$ as well as $\sum_{i=1}^d (\partial_i v) \circ F \cdot A_{ij} \in L^p(\hat{\Omega})$ (very similar to the stability bounds below).

We argue that $\sum_{i=1}^d (\partial_i v) \circ F \cdot A_{ij}$ is indeed the j-th weak partial derivative of $v \circ F$ in the sense of D.2.34. To this end, let $\hat{w} \in C_0^\infty(\hat{\Omega})$ be given. Since F is affine, there holds $\hat{w} \circ F^{-1} \in C_0^\infty(\Omega)$. Furthermore, using the chain rule for classic derivatives, there holds the following (pointwise) identity on Ω:

$$\begin{aligned} \sum_{i=1}^d \partial_i (\hat{w} \circ F^{-1}) A_{ij} &= \sum_{i=1}^d \sum_{m=1}^d (\partial_m \hat{w}) \circ F^{-1} \cdot (A^{-1})_{mi} A_{ij} \\ &= \sum_{m=1}^d (\partial_m \hat{w}) \circ F^{-1} \cdot (A^{-1} A)_{mj} \qquad = \quad (\partial_j \hat{w}) \circ F^{-1}. \end{aligned}$$

[14] More precisely, we should write $v \circ (F|_{\hat{\Omega}}) \in W^{k,p}(\hat{\Omega})$.

[15] Alternatively, one could use Faa di Bruno's formula for $\mathrm{D}^\alpha(v \circ F)$, for all multi-indices $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq k$. See, e.g., [CS96] for an explicit formula in the multivariate setting.

With this identity, we compute

$$\int_{\hat\Omega}\Big(\sum_{i=1}^d(\partial_i v)\circ F\cdot A_{ij}\Big)\hat w\,\mathrm{d}x=\sum_{i=1}^d A_{ij}\int_{\hat\Omega}((\partial_i v)\circ F)\hat w\,\mathrm{d}x$$
$$=\sum_{i=1}^d A_{ij}|\det A|^{-1}\int_\Omega(\partial_i v)(\hat w\circ F^{-1})\,\mathrm{d}x\overset{D.2.34}{=}-\sum_{i=1}^d A_{ij}|\det A|^{-1}\int_\Omega v\partial_i(\hat w\circ F^{-1})\,\mathrm{d}x$$
$$=-|\det A|^{-1}\int_\Omega v((\partial_j\hat w)\circ F^{-1})\,\mathrm{d}x=-\int_{\hat\Omega}(v\circ F)(\partial_j\hat w)\,\mathrm{d}x.$$

Since $\hat w\in C_0^\infty(\hat\Omega)$ was arbitrary, it follows that $\partial_j(v\circ F)=\sum_{i=1}^d(\partial_i v)\circ F\cdot A_{ij}$, as required. Finally, to see the stability bounds, we compute (with an implicit constant $C(d,p)$)

$$(\sigma_{\mathrm{shp}}^{-1}h_\Omega h_{\hat\Omega}^{-1})^{d/p}\|v\circ F\|_{L^p(\hat\Omega)}\overset{L.2.24}{\lesssim}|\det\nabla F|^{1/p}\|v\circ F\|_{L^p(\hat\Omega)}\overset{L.2.33}{=}\|v\|_{L^p(\Omega)}.$$

Similarly, using the bound

$$\max_{i,j\in\{1,\dots,d\}}|A_{ij}|=\max_{i,j\in\{1,\dots,d\}}|\langle Ae_j,e_i\rangle_2|\le\|A\|_2,$$

we get, for all $j\in\{1,\dots,d\}$,

$$(\sigma_{\mathrm{shp}}^{-1}h_\Omega h_{\hat\Omega}^{-1})^{d/p}\|\partial_j(v\circ F)\|_{L^p(\hat\Omega)}=(\sigma_{\mathrm{shp}}^{-1}h_\Omega h_{\hat\Omega}^{-1})^{d/p}\Big\|\sum_{i=1}^d(\partial_i v)\circ F\cdot A_{ij}\Big\|_{L^p(\hat\Omega)}$$
$$\le(\sigma_{\mathrm{shp}}^{-1}h_\Omega h_{\hat\Omega}^{-1})^{d/p}\sum_{i=1}^d\|(\partial_i v)\circ F\|_{L^p(\hat\Omega)}\|A\|_2\lesssim\sum_{i=1}^d\|\partial_i v\|_{L^p(\Omega)}\|A\|_2^{L.2.24}\lesssim\widehat{\sigma_{\mathrm{shp}}}h_\Omega h_{\hat\Omega}^{-1}|v|_{W^{1,p}(\Omega)}.$$

In summary, for all $l\in\{0,1\}$, we have

$$(\sigma_{\mathrm{shp}}^{-1}h_\Omega h_{\hat\Omega}^{-1})^{d/p}|v\circ F|_{W^{l,p}(\hat\Omega)}\le C(d,p,l)(\widehat{\sigma_{\mathrm{shp}}}h_\Omega h_{\hat\Omega}^{-1})^l|v|_{W^{l,p}(\Omega)}.$$

This concludes the first case where $v\in W^{1,p}(\Omega)$ was asserted. To see the reverse direction, suppose that $v:\Omega\longrightarrow\mathbb{R}$ is such that $\hat v:=v\circ F\in W^{1,p}(\hat\Omega)$. Since F^{-1} is again a regular affine transformation (cf. L.2.24), we can apply the previous results to the function $\hat v\circ F^{-1}$ and get $v=\hat v\circ F^{-1}\in W^{1,p}(\Omega)$. Reversing the roles of $\hat\Omega$ and Ω, it follows that, for all $l\in\{0,1\}$,

$$\begin{aligned}(\widehat{\sigma_{\mathrm{shp}}}^{-1}h_{\hat\Omega}h_\Omega^{-1})^{d/p}|v|_{W^{l,p}(\Omega)}&=(\widehat{\sigma_{\mathrm{shp}}}^{-1}h_{\hat\Omega}h_\Omega^{-1})^{d/p}|\hat v\circ F^{-1}|_{W^{l,p}(\Omega)}\\&\le C(d,p,l)(\sigma_{\mathrm{shp}}h_{\hat\Omega}h_\Omega^{-1})^l|\hat v|_{W^{l,p}(\hat\Omega)}\\&=C(d,p,l)(\sigma_{\mathrm{shp}}h_{\hat\Omega}h_\Omega^{-1})^l|v\circ F|_{W^{l,p}(\hat\Omega)}.\end{aligned}$$

This finishes the proof.

□

2.8.4 Lipschitz boundaries, traces and extensions

Many important results in the theory of Sobolev spaces require some regularity assumptions on the boundary of the computation domain $\Omega \subseteq \mathbb{R}^d$. The concept of a *locally Lipschitz-continuous boundary* (or simply *Lipschitz boundary*) turned out to be a particularly fruitful one (e.g., [Cia78, Section 1.2.], [McL00, Pages 89-96], [AF03, Chapter 4] or [Neč12, Section 1.1.3]). Here, we restrict the presentation to the case of open, bounded subsets $\Omega \subseteq \mathbb{R}^d$, but the notion of a Lipschitz boundary also exists for open, unbounded sets (e.g., [AF03, Chapter 4]).

Definition 2.44. *Let $\delta > 0$. A function $\gamma : (-\delta, \delta)^{d-1} \longrightarrow \mathbb{R}$ is called* Lipschitz continuous*, if*

$$\sup_{x,\tilde{x} \in (-\delta,\delta)^{d-1}} \frac{|\gamma(x) - \gamma(\tilde{x})|}{\|x - \tilde{x}\|_2} < \infty.$$

For the next definition, we remind the reader of D.2.23, where we introduced rigid body transformations.

Definition 2.45. *Let $\Omega \subseteq \mathbb{R}^d$ be open and bounded. We say that $\Omega \subseteq \mathbb{R}^d$ has a* Lipschitz boundary*, if there exist $\varepsilon, \delta > 0$, $L \in \mathbb{N}$, rigid body transformations $F_1, \ldots, F_L : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ and Lipschitz continuous functions[16] $\gamma_1, \ldots, \gamma_L : (-\delta, \delta)^{d-1} \longrightarrow \mathbb{R}$, such that*

$$\begin{aligned}
\bigcup_{l=1}^{L} \{F_l(x,y) \,|\, x \in (-\delta,\delta)^{d-1},\ y \in \gamma_l(x) + (-\varepsilon, 0)\} &\subseteq \Omega, \\
\bigcup_{l=1}^{L} \{F_l(x,y) \,|\, x \in (-\delta,\delta)^{d-1},\ y = \gamma_l(x)\} &= \partial\Omega, \\
\bigcup_{l=1}^{L} \{F_l(x,y) \,|\, x \in (-\delta,\delta)^{d-1},\ y \in \gamma_l(x) + (0, \varepsilon)\} &\subseteq \mathbb{R}^d \backslash \overline{\Omega}.
\end{aligned}$$

Lipschitz boundaries allow us to define boundary values for functions $v \in W^{1,p}(\Omega)$ in the form of a *trace operator* $(\cdot)|_\Gamma : W^{1,p}(\Omega) \longrightarrow L^p(\Gamma)$. For the definition of the space $L^p(\Gamma)$, $\Gamma \subseteq \partial\Omega$, we refer the reader to [Neč12, Section 2.4]. More details on the trace operator can be found, e.g., in [Gri85, Section 1.5] or [McL00, Pages 100-106].

Lemma 2.46. *Let $\Omega \subseteq \mathbb{R}^d$ be an open, bounded set with a Lipschitz boundary and let $\Gamma \subseteq \partial\Omega$ be a part of the boundary with $\mathrm{meas}_{d-1}(\Gamma) > 0$. Furthermore, let $p \in [1, \infty)$. Then, there exists a linear* trace operator

$$(\cdot)|_\Gamma : W^{1,p}(\Omega) \longrightarrow L^p(\Gamma)$$

with the following properties:

1. *For all $v \in W^{1,p}(\Omega) \cap C^\infty(\overline{\Omega})$ and $x \in \Gamma$, there holds $(v|_\Gamma)(x) = v(x)$.*
2. *For all $v \in W^{1,p}(\Omega)$, there holds the stability bound[17]*

$$\|v\|_{L^p(\Gamma)} \leq C(d, \Omega, \Gamma, p) \|v\|_{W^{1,p}(\Omega)}.$$

[16] In the case $d = 1$, the functions γ_l have to be replaced with constant values $\gamma_l \in \mathbb{R}$.
[17] Here and in the sequel we abbreviate $\|v|_\Gamma\|_{L^p(\Gamma)}$ by $\|v\|_{L^p(\Gamma)}$.

Proof. See, e.g., [Neč12, Section 2.4] or [Leo17, Chapter 18]. □

The trace operator allows us to define the space $H_0^1(\Omega)$, which will make a short appearance in Chapter 6.

Definition 2.47. *Let $\Omega \subseteq \mathbb{R}^d$ be an open, bounded set with a Lipschitz boundary and let $\Gamma \subseteq \partial\Omega$ be a part of the boundary with $\mathrm{meas}_{d-1}(\Gamma) > 0$. Furthermore, let $p \in [1, \infty)$. We define the space*

$$W_0^{1,p}(\Omega, \Gamma) := \{v \in W^{1,p}(\Omega) \,|\, v|_\Gamma = 0\}.$$

In the case $\Gamma = \partial\Omega$, we abbreviate $W_0^{1,p}(\Omega) := W_0^{1,p}(\Omega, \partial\Omega)$. Furthermore, if $p = 2$, we set $H_0^1(\Omega, \Gamma) := W_0^{1,2}(\Omega, \Gamma)$ and again $H_0^1(\Omega) := H_0^1(\Omega, \partial\Omega)$.

Another important aspect of Lipschitz boundaries is the fact that they allow us to extend a function $v \in W^{k,p}(\Omega)$ to a function $\tilde{v} \in W^{k,p}(\mathbb{R}^d)$.

Definition 2.48. *Let $\Omega \subseteq \mathbb{R}^d$ be an open set, $k \in \mathbb{N}_0$ and $p \in [1, \infty]$. We say that Ω is a* $W^{k,p}$-extension domain, *if there exists a linear operator*

$$E_\Omega : W^{k,p}(\Omega) \longrightarrow W^{k,p}(\mathbb{R}^d)$$

such that, for all $v \in W^{k,p}(\Omega)$ and all $l \in \{0, \dots, k\}$,

$$(E_\Omega v)|_\Omega = v, \qquad \|E_\Omega v\|_{W^{l,p}(\mathbb{R}^d)} \leq C(d, \Omega, k, p)\|v\|_{W^{l,p}(\Omega)}.$$

More details on extension domains can be found, e.g., in [Leo17, Section 13.1.].

Lemma 2.49. *1. The set $\Omega = \mathbb{R}^d$ itself is a $W^{k,p}$-extension domain.*

2. Let $\Omega \subseteq \mathbb{R}^d$ be an open, bounded[18] set with a Lipschitz boundary. Then, Ω is a $W^{k,p}$-extension domain.

Proof. Ad item 1: Take E_Ω as the identity operator from $W^{k,p}(\mathbb{R}^d)$ to $W^{k,p}(\mathbb{R}^d)$.
Ad item 2: See [Ste70, Chapter 6, Section 3]. □

2.8.5 Embedding theorems

Next, we briefly discuss embedding theorems. While the literature on this subject is extensive (e.g., [AF03], [BS08], [Bre11]), we only need the following two results. The first one is known as a *Sobolev embedding theorem* and the second one as a *Rellich-Kondrachov embedding theorem*.

Theorem 2.50. *Let $\Omega \subseteq \mathbb{R}^d$ be an open, bounded[19] set with a Lipschitz boundary. Furthermore, let $k \in \mathbb{N}$ be such that $k > d/2$. Then, there holds the continuous embedding*

$$H^k(\Omega) \subseteq C^0(\overline{\Omega}).$$

[18] The assumption of boundedness can be dropped, if the notion of Lipschitz boundaries for unbounded domains from [AF03, Chapter 4] is used.

[19] Once again, the assumption of boundedness can be dropped, if Lipschitz boundaries are defined as in [AF03, Chapter 4].

In other words, for all $v \in H^k(\Omega)$, there holds[20] *$v \in C^0(\overline{\Omega})$ and*

$$\|v\|_{C^0(\overline{\Omega})} \leq C(d, \Omega, k)\|v\|_{H^k(\Omega)}.$$

Proof. See, e.g., [AF03, Theorem 4.12]. □

Theorem 2.51. *Let $\Omega \subseteq \mathbb{R}^d$ be an open, bounded*[21] *set with a Lipschitz boundary. Furthermore, let $k \in \mathbb{N}$. Then, the there holds the compact embedding*

$$H^k(\Omega) \subseteq H^{k-1}(\Omega).$$

In other words, for every bounded sequence $(v_n)_{n\in\mathbb{N}} \subseteq H^k(\Omega)$, there exist a subsequence $(v_{n(i)})_{i\in\mathbb{N}}$ and a function $v \in H^{k-1}(\Omega)$, such that $\|v - v_{n(i)}\|_{H^{k-1}(\Omega)} \xrightarrow{i} 0$.

Proof. See, e.g., [AF03, Theorem 6.3]. □

2.8.6 Poincaré inequality

An important consequence of the Rellich-Kondrachov embedding theorem is the so-called *Poincaré inequality*, of which there exist several variants. For the purpose of the present work, we need the following version:

Lemma 2.52. *Let $\Omega \subseteq \mathbb{R}^d$ be an open, bounded set with a Lipschitz boundary and $k \in \mathbb{N}_0$. Let $(Z, \|\cdot\|_Z)$ be a normed space and $\iota_Z : H^k(\Omega) \longrightarrow Z$ be a linear operator with the following properties:*

1. *For all $v \in H^k(\Omega)$, there holds $\|\iota_Z v\|_Z \leq C(d, \Omega, k, Z, \iota_Z)\|v\|_{H^k(\Omega)}$.*
2. *For all $v \in \mathbb{P}^{k-1}_{\mathrm{conn}}(\Omega)$ with $\iota_Z v = 0$, there holds $v = 0$.*

Then, there holds the following Poincaré type inequality*:*

$$\forall v \in H^k(\Omega): \qquad \|v\|_{H^k(\Omega)} \leq C(d, \Omega, k, Z, \iota_Z)(|v|_{H^k(\Omega)} + \|\iota_Z v\|_Z).$$

Proof. The case $k = 0$ is trivial, so let us assume that $k \geq 1$. If the Poincaré type inequality were not true, then we could find a sequence $(v_n)_{n\in\mathbb{N}} \subseteq H^k(\Omega)$ with $\|v_n\|_{H^k(\Omega)} = 1$ (after possible normalization) and

$$|v_n|_{H^k(\Omega)} + \|\iota_Z v_n\|_Z < \frac{1}{n}\|v_n\|_{H^k(\Omega)} = \frac{1}{n} \xrightarrow{n} 0.$$

According to T.2.51, there exist a function $v \in H^{k-1}(\Omega)$ and a subsequence of $(v_n)_{n\in\mathbb{N}}$, denoted by $(v_n)_{n\in\mathbb{N}}$ again, such that $\|v - v_n\|_{H^{k-1}(\Omega)} \xrightarrow{n} 0$. Then, in view of the bound

$$\|v_n - v_m\|_{H^k(\Omega)} \lesssim |v_n|_{H^k(\Omega)} + |v_m|_{H^k(\Omega)} + \|v_n - v\|_{H^{k-1}(\Omega)} + \|v - v_m\|_{H^{k-1}(\Omega)},$$

[20] More precisely: We can pick a continuous representative from the equivalence class v.
[21] Here, in contrast to T.2.50, the assumption of boundedness *cannot* be omitted.

it follows that $(v_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $H^k(\Omega)$. Since $H^k(\Omega)$ is complete (cf. L.2.39), there exists a function $w \in H^k(\Omega)$ such that $\|w - v_n\|_{H^k(\Omega)} \xrightarrow{n} 0$. From the inequality

$$\|v - w\|_{H^{k-1}(\Omega)} \leq \|v - v_n\|_{H^{k-1}(\Omega)} + \|v_n - w\|_{H^k(\Omega)} \xrightarrow{n} 0$$

we get that $v = w \in H^k(\Omega)$ and that $\|v - v_n\|_{H^k(\Omega)} = \|w - v_n\|_{H^k(\Omega)} \xrightarrow{n} 0$. Therefore,

$$\begin{aligned} |v|_{H^k(\Omega)} + \|\iota_Z v\|_Z &\leq |v - v_n|_{H^k(\Omega)} + |v_n|_{H^k(\Omega)} + \|\iota_Z(v - v_n)\|_Z + \|\iota_Z v_n\|_Z \\ &\lesssim |v_n|_{H^k(\Omega)} + \|v - v_n\|_{H^k(\Omega)} + \|\iota_Z v_n\|_Z \xrightarrow{n} 0. \end{aligned}$$

We obtain $|v|_{H^k(\Omega)} = 0$ and $\iota_Z v = 0$. Using L.2.40 and the assumption on the operator ι_Z, we obtain $v \in \mathbb{P}^{k-1}_{\text{conn}}(\Omega)$ and ultimately $v = 0$. This produces the contradiction

$$1 = \lim_{n\to\infty} \|v_n\|_{H^k(\Omega)} = \|v\|_{H^k(\Omega)} = 0.$$

□

One of the most famous variants of the Poincaré inequality (e.g., [Leo17, Theorem 13.19]) reads as follows:

Corollary 2.53. *Let $\Omega \subseteq \mathbb{R}^d$ be an open, bounded set with Lipschitz boundary and let $H^1_0(\Omega)$ be defined as in D.2.47. Then, there holds the following bound:*

$$\forall v \in H^1_0(\Omega): \qquad \|v\|_{H^1(\Omega)} \leq C(d,\Omega)|v|_{H^1(\Omega)}.$$

Proof. Apply L.2.52 to the space $Z := L^2(\partial\Omega)$ and the trace operator $\iota_Z := (\cdot)|_{\partial\Omega}$ from L.2.46. □

2.8.7 Inverse- and Caccioppoli inequality

The Poincaré type inequality from L.2.52 allows us to bound *lower* order derivatives of a function $v \in H^k(\Omega)$ by *higher* order derivatives (plus an additional term that accounts for polynomials). On the other hand, bounding *higher* order derivatives by *lower* order ones is certainly not possible for *arbitrary* functions $v \in H^k(\Omega)$. To see this, consider the sequence $v_n := \sin(2\pi n\,\cdot)$ on $\Omega := (0,1) \subseteq \mathbb{R}$. Then $\|v_n\|_{L^2(\Omega)} = 1/\sqrt{2}$ and $|v_n|_{H^1(\Omega)} = \sqrt{2}\pi n$, so that the inequality $|v_n|_{H^1(\Omega)} \leq C\|v_n\|_{L^2(\Omega)}$ *cannot* hold true for a constant $C > 0$ independent of n.

If, however, we have additional information about $v \in H^k(\Omega)$, then such a bound is indeed possible. Below, we present two instances of such inequalities. The first one is concerned with polynomials and goes by the name of *inverse-* or *Markov inequality*.

Lemma 2.54. *Let $\Omega \subseteq \mathbb{R}^d$ be an open, bounded and convex set. Let $k, l \in \mathbb{N}_0$ with $l \leq k$ and $p \in \mathbb{N}_0$. Then, for all $v \in \mathbb{P}^p(\Omega)$, there holds the* inverse inequality

$$p^{-2k}|v|_{H^k(\Omega)} \leq C(d,\Omega,k)p^{-2l}|v|_{H^l(\Omega)}.$$

Proof. The case $(k = 1, l = 0)$ can be found, e.g., in [Dit92]. The general case follows easily by induction on k. □

The second instance is known as a *Caccioppoli inequality* or *inverse Poincaré* inequality (see, e.g., [Gia83, Proposition 2.1] and [Eva10, Section 6.3.] for similar bounds). We provide a short proof, because *discrete* Caccioppoli inequalities will play an important role later on (cf. A.4.19) and it might be helpful to the reader to see how such an inequality can be proved in a continuous setting.

Lemma 2.55. *Let $\Omega \subseteq \mathbb{R}^d$ be an open, bounded set with a Lipschitz boundary. Let $D \subseteq \mathbb{R}^d$ and consider a function $f \in L^2(\Omega)$ with $\operatorname{supp}(f) \subseteq \Omega \cap D$. Denote by $u \in H^1_0(\Omega)$ the weak solution of $(-\Delta u = f,\ u|_{\partial\Omega} = 0)$, i.e.,*

$$\forall v \in H^1_0(\Omega): \qquad \langle \nabla u, \nabla v\rangle_{L^2(\Omega)} = \langle f, v\rangle_{L^2(\Omega)}.$$

Furthermore, let $B \in \mathbb{B}$ be a box (D.2.8) and let $\delta > 0$ be such that the inflated box $B^\delta \in \mathbb{B}$ (D.2.12) satisfies $B^\delta \cap D = \emptyset$. Then, there holds the Caccioppoli type inequality

$$\delta |u|_{H^1(\Omega\cap B)} \leq C(d,\Omega)\|u\|_{L^2(\Omega\cap B^\delta)}.$$

Proof. The key element of the proof is a *cut-off function* κ with the following properties (cf. L.5.3):

$$\kappa \in C^\infty_0(\mathbb{R}^d), \quad \operatorname{supp}(\kappa) \subseteq B^\delta, \quad \kappa|_B \equiv 1, \quad 0 \leq \kappa \leq 1, \quad |\kappa|_{W^{1,\infty}(\mathbb{R}^d)} \lesssim \delta^{-1}.$$

Now, the fact that the product $v := \kappa^2 u$ lies in $H^1_0(\Omega)$ allows us to use v as a test function in the variational equation that defines u. On the one hand, since $\operatorname{supp}(f) \subseteq \Omega \cap D$ and $\operatorname{supp}(v) \subseteq \Omega \cap B^\delta$, we know that

$$\langle \nabla u, \nabla v\rangle_{L^2(\Omega)} = \langle f, v\rangle_{L^2(\Omega)} = \langle f, v\rangle_{L^2(\Omega\cap D\cap B^\delta)} = 0.$$

On the other hand, we can expand $\nabla v = \nabla(\kappa^2 u) = 2\kappa u \nabla\kappa + \kappa^2 \nabla u$ to find that

$$0 = \langle \nabla u, \nabla v\rangle_{L^2(\Omega)} = 2\langle \kappa\nabla u, u\nabla\kappa\rangle_{L^2(\Omega)} + \|\kappa\nabla u\|^2_{L^2(\Omega)}.$$

Solving for $\|\kappa\nabla u\|^2_{L^2(\Omega)}$ and applying the Cauchy-Schwarz inequality, we get

$$\begin{aligned}\|\kappa\nabla u\|^2_{L^2(\Omega)} &= -2\langle \kappa\nabla u, u\nabla\kappa\rangle_{L^2(\Omega)} \\ &\leq 2\|\kappa\nabla u\|_{L^2(\Omega)}\|u\nabla\kappa\|_{L^2(\Omega)} \lesssim \delta^{-1}\|\kappa\nabla u\|_{L^2(\Omega)}\|u\|_{L^2(\Omega\cap B^\delta)},\end{aligned}$$

which ultimately results in the desired bound:

$$|u|_{H^1(\Omega\cap B)} = \|\kappa\nabla u\|_{L^2(\Omega\cap B)} \leq \|\kappa\nabla u\|_{L^2(\Omega)} \leq \delta^{-1}\|u\|_{L^2(\Omega\cap B^\delta)}.$$

□

We mention that this proof works for other geometric shapes than boxes as well (e.g., balls). The only requirement for the sets B and B^δ is that a cut-off function κ with all of the mentioned properties can be constructed.

2.8.8 Deny-Lions Lemma

The error bound in the following corollary goes by the name of *Deny-Lions Lemma* (e.g., [Cia78, Theorem 3.1.1.]) and can also be seen as a variant of the *Bramble-Hilbert Lemma* ([BH71, Theorem 1, Theorem 2]).

Corollary 2.56. *Let $\Omega \subseteq \mathbb{R}^d$ be an open, bounded, connected set with a Lipschitz boundary and $k \in \mathbb{N}_0$. Furthermore, denote by*

$$J : H^k(\Omega) \longrightarrow \mathbb{P}^{k-1}(\Omega)$$

the orthogonal projection onto the closed subspace $\mathbb{P}^{k-1}(\Omega) \subseteq H^k(\Omega)$. Then, for all $v \in H^k(\Omega)$, there holds the following error bound:

$$\|v - Jv\|_{H^k(\Omega)} = \inf_{w \in \mathbb{P}^{k-1}(\Omega)} \|v - w\|_{H^k(\Omega)} \leq C(d, \Omega, k)|v|_{H^k(\Omega)}.$$

Proof. We apply L.2.52 to the normed space $Z := H^k(\Omega)$ and the linear operator $\iota_Z := J$. Since J is an orthogonal projection, there holds $\|\iota_Z v\|_Z \leq \|v\|_{H^k(\Omega)}$, for all $v \in H^k(\Omega)$. Furthermore, for all $v \in \mathbb{P}^{k-1}(\Omega)$ with $\iota_Z v = 0$, the projection property of J yields $v = Jv = \iota_Z v = 0$. Now L.2.52 tells us that, for all $w \in H^k(\Omega)$,

$$\|w\|_{H^k(\Omega)} \leq C(d, \Omega, k)(|w|_{H^k(\Omega)} + \|Jw\|_{H^k(\Omega)}).$$

Finally, given $v \in H^k(\Omega)$, we plug in $w := v - Jv \in H^k(\Omega)$ and find that

$$\|v - Jv\|_{H^k(\Omega)} \lesssim |v - Jv|_{H^k(\Omega)} + \|J(v - Jv)\|_{H^k(\Omega)} = |v|_{H^k(\Omega)}.$$

□

2.9 Meshes

In this section, we introduce the basic concepts regarding *(simplicial) meshes*. We will need these results later on in Chapter 6, where we apply the abstract framework from Chapter 4 to a mesh-based *finite element problem*. The introduction of a mesh on a given computational domain $\Omega \subseteq \mathbb{R}^d$ is often the first step in the analysis of the finite element method for the discretization of partial differential equations. For further reading, see, e.g., [Cia78, Chapter 2], [EG04, Section 1.3], [BS08, Chapter 3], [Bra13, Chapter 2, Section 5] or [LB13, Chapter 3].

2.9.1 Simplices

Definition 2.57. *Let $d \in \mathbb{N}$ and $k \in \{0, \dots, d\}$. A subset $S \subseteq \mathbb{R}^d$ is called* k-simplex, *if there exist points $N_0, \dots, N_k \in \mathbb{R}^d$ such that the vectors $\{N_1 - N_0, \dots, N_k - N_0\} \subseteq \mathbb{R}^d$ are linearly independent and such that*

$$S = \left\{ N_0 + \sum_{i=1}^{k} t_i(N_i - N_0) \,\middle|\, t_1, \dots, t_k \geq 0,\ \sum_{i=1}^{k} t_i \leq 1 \right\}.$$

The points $\mathcal{N}(S) := \{N_0, \dots, N_k\}$ are called nodes of S. *In the case $k = d$, we drop the prefix "d-" and call S a* simplex.

Note that S is closed, i.e., it contains its boundary. Furthermore, we mention that there is some ambiguity in the set of nodes $\mathcal{N}(S)$ in that a reordering of the nodes produces the same physical set $S \subseteq \mathbb{R}^d$. In $d = 3$ space dimensions, a 3-simplex is a tetrahedron, a 2-simplex is a triangle, a 1-simplex is a line segment and a 0-simplex is a point (all of which are subsets/elements of $\mathbb{R}^3$).

Lemma 2.58. *Let $T \subseteq \mathbb{R}^d$ be a simplex with shape regularity $\sigma_{\mathrm{shp}} \geq 1$ (cf. D.2.16). Then, there hold the relations*

$$(2\sigma_{\mathrm{shp}}^{-1})h_T \leq \min_{\substack{M,N\in\mathcal{N}(T):\\ M\neq N}} \|M-N\|_2 \leq \max_{M,N\in\mathcal{N}(T)} \|M-N\|_2 = h_T.$$

Proof. Let $\mathcal{N}(T) = \{N_0, \dots, N_d\}$ and assume w.l.o.g. that $\min_{M\neq N} \|M-N\|_2 = \|N_d - N_0\|_2$. Consider the $(d-1)$-dimensional, parallel hyperplanes

$$\begin{aligned} \Gamma_0 &:= N_0 + \operatorname{span}\{N_i - N_0 \,|\, i \in \{1,\dots,d-1\}\}, \\ \Gamma_d &:= N_d + \operatorname{span}\{N_i - N_0 \,|\, i \in \{1,\dots,d-1\}\} \end{aligned}$$

and denote the enclosed slice by

$$\Omega := \{(1-t)x + ty \,|\, x \in \Gamma_0, y \in \Gamma_d, t \in [0,1]\} \subseteq \mathbb{R}^d.$$

Since $\mathrm{Ball}_2(x_{\mathrm{shp}}, (2\sigma_{\mathrm{shp}})^{-1}h_T) \subseteq T \subseteq \Omega$ (cf. D.2.16), we have

$$(2\sigma_{\mathrm{shp}})^{-1}h_T \leq \operatorname{dist}_2(\Gamma_0,\Gamma_d) \leq \|N_d - N_0\|_2 = \min_{\substack{M,N\in\mathcal{N}(T):\\ M\neq N}} \|M-N\|_2.$$

The other relations being trivial, it remains to show that $h_T \leq \max_{M,N} \|M-N\|_2$. To this end, consider arbitrary points $x, y \in T$. We expand $x = \sum_{i=0}^d s_i N_i$ and $y = \sum_{j=0}^d t_j N_j$ with coefficients $t_0, \dots, t_d, s_0, \dots, s_d \in [0,1]$ satisfying $\sum_{i=0}^d s_i = \sum_{j=0}^d t_j = 1$. Then,

$$h_T = \sup_{s_i,t_j} \left\| \sum_{i=0}^d s_i N_i - \sum_{j=0}^d t_j N_j \right\|_2 = \sup_{s_i,t_k} \left\| \sum_{i,j=0}^d s_i t_j (N_i - N_j) \right\|_2 \leq \max_{N,N\in\mathcal{N}(T)} \|M-N\|_2.$$

□

One particular simplex will play an important role in the sequel.

Definition 2.59. *Let $\hat{N}_0 := 0 \in \mathbb{R}^d$ and, for all $i \in \{1,\dots,d\}$, let $\hat{N}_i := e_i \in \mathbb{R}^d$, the i-th Euclidean unit vector. The corresponding simplex $\hat{T} \subseteq \mathbb{R}^d$ is called* reference element.

2.9.2 Simplicial meshes

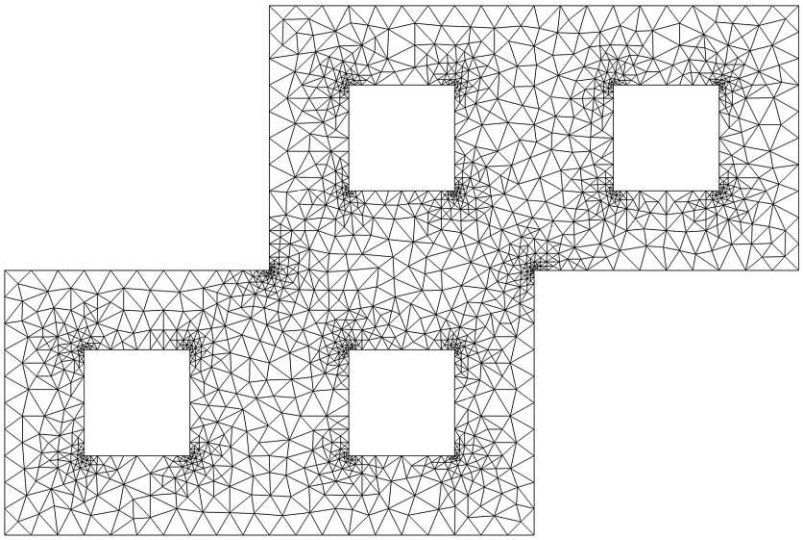

Figure 2.1: A simplicial mesh in 2D.

Definition 2.60. *Let $\Omega \subseteq \mathbb{R}^d$ be an open, bounded set with Lipschitz boundary (cf. D.2.45). Furthermore, let $\sigma_{\mathrm{shp}}, \sigma_{\mathrm{lqu}} \geq 1$ be given constants. A system $\mathcal{T} \subseteq \mathrm{Pow}(\overline{\Omega})$ is called* (simplicial) mesh on Ω, *if the following conditions are satisfied:*

1. *The set $\mathcal{T}$ is finite.*
2. *Every $T \in \mathcal{T}$ is a simplex in the sense of D.2.57.*
3. *Every $T \in \mathcal{T}$ has shape regularity σ_{shp} (cf. D.2.16).*
4. *For all $T, \tilde{T} \in \mathcal{T}$ with $T \neq \tilde{T}$ and $T \cap \tilde{T} \neq \emptyset$, there exists a $k \in \{0, \dots, d-1\}$, such that $T \cap \tilde{T}$ is a k-simplex with $\mathcal{N}(T \cap \tilde{T}) \subseteq \mathcal{N}(T) \cap \mathcal{N}(\tilde{T})$.*
5. *There holds*
$$\overline{\Omega} = \bigcup_{T \in \mathcal{T}} T.$$
6. *For all $T, \tilde{T} \in \mathcal{T}$ with $\mathcal{N}(T) \cap \mathcal{N}(\tilde{T}) \neq \emptyset$, there holds the bound*
$$h_{\tilde{T}} \leq \sigma_{\mathrm{lqu}} h_T.$$

Remark 2.61. *Item 6 says that $\mathcal{T}$ is* locally quasi-uniform, *i.e., simplices $T, \tilde{T} \in \mathcal{T}$ sharing a common node N have comparable diameters. This assumption is automatically fulfilled in $d \geq 2$ space dimensions. To see this, one can exploit the Lipschitz property of $\partial\Omega$ to construct a "fan" of elements $T_1, \dots, T_L \in \mathcal{T}$ with*

$$T_1 = T, \qquad T_L = \tilde{T}, \qquad N \in \mathcal{N}(T_l), \qquad \#(\mathcal{N}(T_l) \cap \mathcal{N}(T_{l+1})) \geq 2.$$

Since T_l and T_{l+1} have a common edge, we obtain the relation $h_{T_{l+1}} \leq C(\sigma_{\mathrm{shp}})h_{T_l}$ from L.2.58. Solving the recursion, it follows that $h_{\tilde{T}} \leq C(\sigma_{\mathrm{shp}})^{L-1}h_T$. Finally, since all T_l have N as a common node, the upcoming L.2.64 yields $L \leq C(d, \sigma_{\mathrm{shp}})$ as well.

Definition 2.62. *Let $\Omega \subseteq \mathbb{R}^d$ be an open, bounded set with Lipschitz boundary. We say that Ω is a* polyhedron, *if there exists a mesh $\mathcal{T} \subseteq \mathrm{Pow}(\overline{\Omega})$ on Ω.*

Note that, according to D.2.57, every $T \in \mathcal{T}$ is a closed subset of $\mathbb{R}^d$. In order to remove the ambiguity in the orientation of the simplices, we henceforth assume that, for every $T \in \mathcal{T}$, an ordering of the nodes $\mathcal{N}(T)$ has been fixed in advance. Before we go on, we introduce a few names:

Definition 2.63. 1. *The members T of a mesh $\mathcal{T}$ are called* (mesh) elements. *In analogy to D.3.6, a subset $\mathcal{B} \subseteq \mathcal{T}$ is called* (mesh) cluster.

2. *The* (mesh) nodes *are given by $\mathcal{N} := \bigcup_{T\in\mathcal{T}} \mathcal{N}(T)$.*

3. *For every $T \in \mathcal{T}$, we fix an incenter $x_T \in T$ (cf. D.2.16).*

4. *For every physical subset $B \subseteq \mathbb{R}^d$, we define the* patch[22]

$$\mathcal{T}(B) := \{T \in \mathcal{T} \,|\, T \cap B \neq \emptyset\}.$$

In particular, $\mathcal{T}(T)$ is the patch of a mesh element $T \in \mathcal{T}$, $\mathcal{T}(x_0) := \mathcal{T}(\{x_0\})$ is the patch of a point $x_0 \in \mathbb{R}^d$ and $\mathcal{T}(N)$ is the patch of a node $N \in \mathcal{N}$.

5. *For all $\mathcal{B} \subseteq \mathcal{T}$, we set*

$$h_{\mathcal{B}} := h_{\max,\mathcal{B}} := \max_{T\in\mathcal{B}} h_T, \qquad h_{\min,\mathcal{B}} := \min_{T\in\mathcal{B}} h_T.$$

Note that, given an element $T \in \mathcal{T}$ and a node $N \in \mathcal{N}$, there holds the equivalence

$$N \in \mathcal{N}(T) \quad \Leftrightarrow \quad T \in \mathcal{T}(N).$$

In the next lemma, we show that the number of mesh elements in any given node/element patch is uniformly bounded.

Lemma 2.64. *Let $\Omega \subseteq \mathbb{R}^d$ be a polyhedron and $\mathcal{T} \subseteq \mathrm{Pow}(\overline{\Omega})$ be a mesh.*

1. *For all $T, \tilde{T} \in \mathcal{T}$ with $T \neq \tilde{T}$, there holds*

$$T \cap \tilde{T} \subseteq (\partial T) \cap (\partial \tilde{T}).$$

In particular, $T^\circ \cap \tilde{T}^\circ = \emptyset$.

2. *For all $x_0 \in \mathbb{R}^d$ and $T_0 \in \mathcal{T}$, there hold the bounds*

$$\begin{aligned} \#\mathcal{T}(x_0) &\leq C(d, \sigma_{\mathrm{shp}}), \\ \#\mathcal{T}(T_0) &\leq C(d, \sigma_{\mathrm{shp}}). \end{aligned}$$

[22] Note that $\mathcal{T}(B) \subseteq \mathcal{T}$ is just a collection of mesh elements, whereas $\bigcup \mathcal{T}(B) \subseteq \mathbb{R}^d$ is the corresponding physical set.

Proof. Ad item 1: Let $T, \tilde{T} \in \mathcal{T}$ with $T \neq \tilde{T}$. If $T \cap \tilde{T} = \emptyset$, then the statement becomes trivial. On the other hand, if $T \cap \tilde{T} \neq \emptyset$, then we know from D.2.60 that $S := T \cap \tilde{T}$ is a k-simplex, for some $k \in \{0, \dots, d-1\}$, and that $\mathcal{N}(S) \subseteq \mathcal{N}(T) \cap \mathcal{N}(\hat{T})$. But then S must be a subset of a hyperplane going through one of T's faces so that $S \subseteq \partial T$. The same holds true for $\tilde{T}$, which ultimately leads to $S \subseteq (\partial T) \cap (\partial \tilde{T})$.

Ad item 2: The idea is to shrink the patch elements $\mathcal{T}(x_0)$ towards the common point x_0 so that they all have the same size. To this end, let

$$h_{\min} := \min_{T \in \mathcal{T}(x_0)} h_T.$$

Now, consider an element $T \in \mathcal{T}(x_0)$. We define $\tilde{T} := F_T(T) \subseteq \mathbb{R}^d$, where F_T is the following affine transformation (cf. D.2.23):

$$\forall x \in \mathbb{R}^d : \qquad F_T(x) := \frac{h_{\min}}{h_T}(x - x_0) + x_0 = \frac{h_{\min}}{h_T} x + \left(1 - \frac{h_{\min}}{h_T}\right) x_0.$$

The mapping F_T shrinks the original element T towards the point x_0, which itself lies both in T and $\tilde{T}$. Since $h_{\min}/h_T \in [0,1]$, the shrunk element $\tilde{T}$ consists of convex combinations of points from T. But T is convex, so that there must hold $\tilde{T} \subseteq T$. As for the diameter of $\tilde{T}$, we have

$$\sigma_{\mathrm{shp}}^{-1} h_{\min} = \sigma_{\mathrm{shp}}^{-1} \|\nabla F_T\|_2 h_T \overset{L.2.24}{\leq} h_{\tilde{T}} \overset{L.2.24}{\leq} \|\nabla F_T\|_2 h_T = h_{\min}.$$

Furthermore, according to L.2.24, $\tilde{T}$ is $\widetilde{\sigma_{\mathrm{shp}}}$-shape regular, where

$$\widetilde{\sigma_{\mathrm{shp}}} := \sigma_{\mathrm{shp}} \|\nabla(F_T^{-1})\|_2 h_{\tilde{T}} h_T^{-1} = \sigma_{\mathrm{shp}} h_{\tilde{T}} h_{\min}^{-1} \leq \sigma_{\mathrm{shp}}.$$

Now, since $x_0 \in \tilde{T}$, there holds the following chain of inclusions:

$$B_T := \mathrm{Ball}_2(\widetilde{x_{\mathrm{shp}}}, (2\widetilde{\sigma_{\mathrm{shp}}})^{-1} h_{\tilde{T}}) \overset{L.2.17}{\subseteq} \tilde{T}^\circ \subseteq \tilde{T} \subseteq \overline{\mathrm{Ball}}_2(x_0, h_{\tilde{T}}) \subseteq \overline{\mathrm{Ball}}_2(x_0, h_{\min}) =: B.$$

In particular, we have

$$\sigma_{\mathrm{shp}}^{-2d} h_{\min}^d \leq (\widetilde{\sigma_{\mathrm{shp}}}^{-1} h_{\tilde{T}})^d \overset{L.2.7}{\lesssim} \mathrm{meas}(B_T), \qquad \mathrm{meas}(B) \overset{L.2.7}{\lesssim} h_{\min}^d.$$

Note that the balls $\{B_T \,|\, T \in \mathcal{T}(x_0)\}$ are pairwise disjoint, because the supersets $T^\circ \supseteq \tilde{T}^\circ \supseteq B_T$ are pairwise disjoint (cf. item 1). Putting everything together, we obtain

$$\sigma_{\mathrm{shp}}^{-2d} h_{\min}^d \#\mathcal{T}(x_0) \lesssim \sum_{T \in \mathcal{T}(x_0)} \mathrm{meas}(B_T) = \mathrm{meas}\left(\dot{\bigcup_{T \in \mathcal{T}(x_0)}} B_T \right) \leq \mathrm{meas}(B) \lesssim h_{\min}^d.$$

Dividing by $h_{\min}^d$, we obtain the desired bound $\#\mathcal{T}(x_0) \leq C(d, \sigma_{\mathrm{shp}})$. This finishes the case of a single point $x_0 \in \mathbb{R}^d$.

Finally, let $T_0 \in \mathcal{T}$. For every $T \in \mathcal{T}(T_0)$, we know from D.2.60 that T_0 and T share at least one common node. Therefore, using the previously established bound for single points,

$$\#\mathcal{T}(T_0) \leq \sum_{N \in \mathcal{N}(T_0)} \#\mathcal{T}(N) \leq C(d, \sigma_{\mathrm{shp}}).$$

□

2.9.3 The affine reference mappings

Many aspects of the finite element method are easier to handle on the simplex $\hat{T} \subseteq \mathbb{R}^d$ from D.2.59. Since every mesh element $T \in \mathcal{T}$ is a simplex as well, it is not surprising that we can find an affine transformation (cf. D.2.23) between $\hat{T}$ and T.

Definition 2.65. *For every mesh element $T \in \mathcal{T}$ with nodes $\mathcal{N}(T) = \{N_0, \dots, N_d\}$, we define the affine transformation*

$$F_T : \left\{ \begin{array}{ccc} \mathbb{R}^d & \longrightarrow & \mathbb{R}^d \\ x & \longmapsto & N_0 + \sum_{i=1}^d x_i (N_i - N_0) \end{array} \right. .$$

Note that F_T can also be interpreted as a function $F_T : \hat{T} \longrightarrow T$ and we will frequently do so in the sequel.

What D.2.65 tells us is that we can think of mesh elements $T \in \mathcal{T}$ having a particularly "nice" form. Suppose, for example, that we want to prove some inequality on one of the elements $T \in \mathcal{T}$. More often than not, there will be implicit constants $C > 0$ involved, which depend on the integration domain, i.e., the element T itself. Without any further information about the shape of T, this could be problematic. In order to avoid this pitfall, we can proceed in three steps: First, we use F_T to transform the goal inequality to the reference element $\hat{T}$. Second, we prove the inequality on $\hat{T}$ and inherit a constant $C(\hat{T}) > 0$, which need not bother us. Third, we transform the inequality back to the mesh element T itself. Note that transforming back and forth between $\hat{T}$ and T poses no problem, because the stability properties of F_T are controlled by the shape regularity constant σ_{shp} (cf. L.2.24 and L.2.43).

An instance of such an argument can be found in the next lemma, which states that the behaviour of a polynomial of degree 1 is determined by its values on the element's nodes.

Lemma 2.66. *For all $T \in \mathcal{T}$ and all $v \in \mathbb{P}^1(T)$, there hold the following bounds:*

$$\begin{aligned} \|v\|_{L^\infty(T)} &\leq C(d, \sigma_{\mathrm{shp}}) \Big(\min_{x \in T} |v(x)| + h_T |v|_{W^{1,\infty}(T)} \Big), \\ h_T |v|_{W^{1,\infty}(T)} &\leq C(d, \sigma_{\mathrm{shp}}) \max_{M,N \in \mathcal{N}(T)} |v(M) - v(N)|. \end{aligned}$$

Proof. Denote by $\hat{T} \subseteq \mathbb{R}^d$ the reference element from D.2.59 and recall that its nodes are given by $\mathcal{N}(\hat{T}) = \{0, e_1, \dots, e_d\}$, where $e_i \in \mathbb{R}^d$ is the i-th Euclidean unit vector.

Consider a polynomial $w \in \mathbb{P}^1(\hat{T})$. Since $|w(\cdot)|$ is continuous on the compact set $\hat{T}$, we can pick a point $\hat{x}_0 \in \hat{T}$ such that $|w(\hat{x}_0)| = \min_{\hat{x} \in \hat{T}} |w(\hat{x})|$. Then, using a Taylor expansion of w around $\hat{x}_0$, we get

$$\|w\|_{L^\infty(\hat{T})} \leq |w(\hat{x}_0)| + d |w|_{W^{1,\infty}(\hat{T})} \mathrm{diam}_2(\hat{T}) \lesssim \min_{\hat{x} \in \hat{T}} |w(\hat{x})| + |w|_{W^{1,\infty}(\hat{T})}.$$

On the other hand, a Taylor expansion around 0 tells us that $\partial_i w \equiv w(e_i) - w(0)$, for all $i \in \{1, \dots, d\}$. In particular,

$$|w|_{W^{1,\infty}(\hat{T})} = \max_{i \in \{1,\dots,d\}} |w(e_i) - w(0)| \leq \max_{\hat{M},\hat{N} \in \mathcal{N}(\hat{T})} |w(\hat{M}) - w(\hat{N})|.$$

Now, denote by $F_T : \hat{T} \longrightarrow T$ the affine transformation from D.2.65. Then, for all $v \in \mathbb{P}^1(T)$,

$$\begin{aligned} \|v\|_{L^\infty(T)} &\overset{L.2.43}{\lesssim} \|v \circ F_T\|_{L^\infty(\hat{T})} \lesssim \min_{\hat{x}\in\hat{T}} |(v \circ F_T)(\hat{x})| + |v \circ F_T|_{W^{1,\infty}(\hat{T})} \\ &\overset{L.2.43}{\lesssim} \min_{x\in T} |v(x)| + h_T |v|_{W^{1,\infty}(T)}. \end{aligned}$$

Furthermore, since F_T maps the nodes of $\hat{T}$ to the nodes of T,

$$\begin{aligned} h_T |v|_{W^{1,\infty}(T)} &\overset{L.2.43}{\lesssim} |v \circ F_T|_{W^{1,\infty}(\hat{T})} \leq \max_{\hat{M},\hat{N}\in\mathcal{N}(\hat{T})} |(v \circ F_T)(\hat{M}) - (v \circ F_T)(\hat{N})| \\ &= \max_{M,N\in\mathcal{N}(T)} |v(M) - v(N)|. \end{aligned}$$

This concludes the proof.

□

Lemma 2.67. *Let $k, l \in \mathbb{N}_0$ with $l \leq k$ and $p \in \mathbb{N}_0$. For all $T \in \mathcal{T}$ and all $v \in \mathbb{P}^p(T)$, there holds the* inverse inequality

$$p^{-2k} h_T^k |v|_{H^k(T)} \leq C(d, k, \sigma_{\mathrm{shp}}) p^{-2l} h_T^l |v|_{H^l(T)}.$$

Proof. Since $\hat{T}$ is bounded and convex, we know from L.2.54 that, for all $w \in \mathbb{P}^p(\hat{T})$, there holds

$$p^{-2k} |w|_{H^k(\hat{T})} \lesssim p^{-2l} |w|_{H^l(\hat{T})}.$$

Denote by $F_T : \hat{T} \longrightarrow T$ the affine element transformation from D.2.65 and let $v \in \mathbb{P}^p(T)$. Since $v \circ F_T \in \mathbb{P}^p(\hat{T})$, we get

$$p^{-2k} h_T^k |v|_{H^k(T)} \overset{L.2.43}{\lesssim} p^{-2k} h_T^{d/2} |v \circ F_T|_{H^k(\hat{T})} \lesssim p^{-2l} h_T^{d/2} |v \circ F_T|_{H^l(\hat{T})} \overset{L.2.43}{\lesssim} p^{-2l} h_T^l |v|_{H^l(T)}.$$

□

2.9.4 Mesh refinement

In D.2.60, we required that mesh elements sharing a common node have comparable diameters, but we made no assumptions about the *global* distribution of element diameters. In particular, heavily non-uniform meshes such as exponentially graded- or locally refined ones are allowed (cf. D.6.15).

Remark 2.68. *Given a polyhedron $\Omega \subseteq \mathbb{R}^d$, a shape regular, possibly non-uniform family of meshes $(\mathcal{T}_l)_{l\in N}$ can be constructed from an arbitrary initial mesh $\mathcal{T}_1$ via a procedure called* adaptive mesh refinement. *Assuming that $\mathcal{T}_l$ is already defined, the construction of $\mathcal{T}_{l+1}$ is done in four steps:*

1. *For every element $T \in \mathcal{T}_l$, a so-called* error estimator $\eta_T \in [0, \infty)$ *is computed.*

2. *From the error estimators* $(\eta_T)_{T\in\mathcal{T}_l}$*, a subset* $\mathcal{M}_l \subseteq \mathcal{T}_l$ *of* marked elements *is determined. (Typically the elements with the largest error estimators.)*

3. *The marked elements* $T \in \mathcal{M}_l$ *are subdivided into smaller elements.*

4. *If needed, elements which are "close" to marked elements also get subdivided. (Might be necessary to ensures the adjacency condition from D.2.60.)*

The error estimators η_T *can be used to specify regions of* Ω *where the current mesh* $\mathcal{T}_l$ *is too coarse. The choice of a specific type of error estimator usually depends on the application the user has in mind (e.g., [CFPP14] or [EG04, Chapter 10]). If the element subdivision in step* 3 *is done via* bisection *([Ste08]), then the mesh family* $(\mathcal{T}_l)_{l\in\mathbb{N}}$ *is indeed shape regular in the sense of D.2.60.*

We close this short section with a rigorous proof of the seemingly trivial fact that, given a number $\delta > 0$ and a mesh element $T \in \mathcal{T}$, we can split T into a family of smaller simplices of diameter $\eqsim \delta$.

Lemma 2.69. *Let* $\Omega \subseteq \mathbb{R}^d$ *be a polyhedron,* $\mathcal{T} \subseteq \mathrm{Pow}(\overline{\Omega})$ *be a mesh and* $C_0 \geq 1$ *be a given constant. Let* $T \in \mathcal{T}$ *and let* $\delta > 0$ *be such that* $\delta < C_0 h_T$*. Then, there exists a family* $\mathcal{S} \subseteq \mathrm{Pow}(\mathbb{R}^d)$ *of simplices* $S \subseteq \mathbb{R}^d$ *with the following properties:*

1. *There holds* $\bigcup_{S\in\mathcal{S}} S = T$*.*

2. *Every* $S \in \mathcal{S}$ *is* $\widetilde{\sigma_{\mathrm{shp}}}$*-shape regular, where* $\widetilde{\sigma_{\mathrm{shp}}} = C(d)\sigma_{\mathrm{shp}}$*.*

3. *There hold the bounds* $h_{\max,\mathcal{S}} \leq \delta \leq C(d)C_0 h_{\min,\mathcal{S}}$*.*

Proof. Denote by $\hat{T} \subseteq \mathbb{R}^d$ the reference element from D.2.59 and by $F : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ the affine element transformation from D.2.65 which maps $F(\hat{T}) = T$. We know from L.2.24 that there exists a constant $C_1 := C(d) \geq 1$, such that

$$C_1^{-1} h_T \leq \|\nabla F\|_2 \leq C_1 h_T, \qquad C_1^{-1} h_T^{-1} \leq \|\nabla(F^{-1})\|_2 \leq C_1 \sigma_{\mathrm{shp}} h_T^{-1}.$$

Now, let $M \in \mathbb{N}$ be a free parameter. According to [EG00], the simplex $\hat{T}$ can be subdivided into a family $\hat{\mathcal{S}} \subseteq \mathrm{Pow}(\hat{T})$ of simplices $\hat{S} \subseteq \mathbb{R}^d$ with the following properties (for some $C_2 := C(d) \geq 1$):

1. There holds $\bigcup_{\hat{S}\in\hat{\mathcal{S}}} \hat{S} = \hat{T}$.

2. Every $\hat{S} \in \hat{\mathcal{S}}$ is C_2-shape regular.

3. For every $\hat{S} \in \hat{\mathcal{S}}$, there hold the bounds $C_2^{-1} M^{-1} \leq h_{\hat{S}} \leq C_2 M^{-1}$.

We define the corresponding family of simplices on T:

$$\mathcal{S} := \{F(\hat{S}) \,|\, \hat{S} \in \hat{\mathcal{S}}\} \subseteq \mathrm{Pow}(T).$$

Clearly, $\bigcup_{S\in\mathcal{S}} S = T$. Furthermore, for every $S = F(\hat{S}) \in \mathcal{S}$, we know from L.2.24 that S is σ-shape regular, where

$$\sigma := C_2 \|\nabla(F^{-1})\|_2 h_S h_{\hat{S}}^{-1} \leq C_2 \|\nabla(F^{-1})\|_2 \|\nabla F\|_2 \leq C_2 (C_1 \sigma_{\mathrm{shp}} h_T^{-1})(C_1 h_T) = C_1^2 C_2 \sigma_{\mathrm{shp}}.$$

Finally, let us derive the bounds for $h_{\max,\mathcal{S}}$ and $h_{\min,\mathcal{S}}$: L. 2.24 tells us that $h_S \leq h_{\hat{S}} \|\nabla F\|_2 \leq C_2 h_S$, which readily implies

$$C_1^{-1} C_2^{-2} h_T M^{-1} \leq h_S \leq C_1 C_2 h_T M^{-1}.$$

Now, using the ceiling function $\lceil \cdot \rceil$, we choose

$$M := \lceil C_1 C_2 h_T \delta^{-1} \rceil \in \mathbb{N}.$$

Since we assumed $\delta < C_0 h_T$, we have

$$\begin{array}{rcccccl} M & = & \lceil C_1 C_2 h_T \delta^{-1} \rceil & \geq & C_1 C_2 h_T \delta^{-1}, & & \\ M & = & \lceil C_1 C_2 h_T \delta^{-1} \rceil & \leq & C_1 C_2 h_T \delta^{-1} + 1 & < & (C_1 C_2 + C_0) h_T \delta^{-1}, \end{array}$$

which leads us to the bounds

$$\begin{array}{rcccccl} h_{\max,\mathcal{S}} & = & \max\limits_{S \in \mathcal{S}} h_S & \leq & C_1 C_2 h_T M^{-1} & \leq & \delta, \\ h_{\min,\mathcal{S}} & = & \min\limits_{S \in \mathcal{S}} h_S & \geq & C_1^{-1} C_2^{-2} h_T M^{-1} & \geq & C_1^{-1} C_2^{-2} (C_1 C_2 + C_0)^{-1} \delta. \end{array}$$

This concludes the proof. □

2.9.5 Spline spaces

Next, we introduce the well-known *spline spaces*. To this end, we remind the reader of D.2.25, D.2.37 and D.2.47, where we defined polynomial and Sobolev spaces.

Definition 2.70. *Let $\Omega \subseteq \mathbb{R}^d$ be a polyhedron and $\mathcal{T} \subseteq \mathrm{Pow}(\overline{\Omega})$ be a mesh. For all $p \in \mathbb{N}$, we define the* spline spaces

$$\begin{array}{rcl} \mathbb{S}^{p,1}(\mathcal{T}) & := & \{v \in H^1(\Omega) \,|\, \forall T \in \mathcal{T} : v \circ F_T \in \mathbb{P}^p(\hat{T})\}, \\ \mathbb{S}_0^{p,1}(\mathcal{T}) & := & \{v \in H_0^1(\Omega) \,|\, \forall T \in \mathcal{T} : v \circ F_T \in \mathbb{P}^p(\hat{T})\}. \end{array}$$

Similarly, for all $p \in \mathbb{N}_0$, we set

$$\mathbb{S}^{p,0}(\mathcal{T}) := \{v \in L^2(\Omega) \,|\, \forall T \in \mathcal{T} : v \circ F_T \in \mathbb{P}^p(\hat{T})\}.$$

Note that, trivially, $\mathbb{S}_0^{p,1}(\mathcal{T}) \subseteq \mathbb{S}^{p,1}(\mathcal{T}) \subseteq \mathbb{S}^{p,0}(\mathcal{T})$. We will refer to functions $v \in \mathbb{S}^{p,0}(\mathcal{T})$ as being *discrete*.

Given a simplex $T \subseteq \mathbb{R}^d$ and a polynomial $v \in \mathbb{P}^p(T)$, the support of v can either be empty or almost all of T. Since there are no other possibilities in between, it makes sense to introduce a slightly different notion of supports for discrete functions:

Definition 2.71. *Let $\Omega \subseteq \mathbb{R}^d$ be a polyhedron and $\mathcal{T} \subseteq \mathrm{Pow}(\overline{\Omega})$ be a mesh. For all $p \in \mathbb{N}_0$ and $v \in \mathbb{S}^{p,0}(\mathcal{T})$, we set*

$$\mathrm{supp}_{\mathcal{T}}(v) := \{T \in \mathcal{T} \,|\, v|_T \not\equiv 0\}.$$

Note that $\operatorname{supp}_{\mathcal{T}}(v)$ is a set of mesh elements, rather than a physical set. In particular, we have $\operatorname{supp}_{\mathcal{T}}(v) \subseteq \mathcal{T}$ and $\bigcup \operatorname{supp}_{\mathcal{T}}(v) \subseteq \mathbb{R}^d$. (In comparison, the usual support is a subset $\operatorname{supp}(v) \subseteq \mathbb{R}^d$.)

One of the most widely used ansatz spaces in real world FEM applications is $\mathbb{S}^{1,1}(\mathcal{T})$, i.e., the lowest-order case $p = 1$. A natural choice for a set of basis functions is associated with the mesh nodes $\mathcal{N}$.

Lemma 2.72. *Let $\Omega \subseteq \mathbb{R}^d$ be a polyhedron and let $\mathcal{T} \subseteq \operatorname{Pow}(\overline{\Omega})$ be a mesh with nodes $\mathcal{N}$. There exists a system of* hat functions

$$\{\psi_N \,|\, N \in \mathcal{N}\} \subseteq \mathbb{S}^{1,1}(\mathcal{T})$$

with the following properties:

1. *$\{\psi_N \,|\, N \in \mathcal{N}\}$ is a basis of $\mathbb{S}^{1,1}(\mathcal{T})$ and $\{\psi_N \,|\, N \in \mathcal{N} \backslash \partial\Omega\}$ is a basis of $\mathbb{S}^{1,1}_0(\mathcal{T})$.*
2. *For all $M, N \in \mathcal{N}$, there holds $\psi_N(M) = \delta_{NM}$.*
3. *For all $N \in \mathcal{N}$, there holds $\operatorname{supp}_{\mathcal{T}}(\psi_N) \subseteq \mathcal{T}(N)$.*
4. *For all $N \in \mathcal{N}$, there holds $0 \leq \psi_N \leq 1$.*
5. *There holds $\sum_{N \in \mathcal{N}} \psi_N \equiv 1$ on all of Ω.*

Proof. See, e.g., [EG04, Section 1.1.2] for the case $d = 1$ or [LB13, Section 3.2.2] for the case $d = 2$. □

2.9.6 Discrete cut-off functions

In Section 2.8.7, we already hinted that Caccioppoli type inequalities will play an important role later on. During the proof of L.2.55, we then used a *smooth cut-off function* $\kappa \in C^\infty(\Omega)$ for the derivation of a Caccioppoli inequality in the continuous problem setting. This result will serve as a blueprint for the proof of our main result, T.4.21. However, T.4.21 is derived in a fully discrete setting (cf. A.4.19), meaning we only work with the solution of the discrete problem, rather than the continuous one.

In Chapter 6, we then verify the assumption from A.4.19 and prove a *discrete* version of the Caccioppoli in the context of a finite element discretization. Therefore, it is not surprising that we also need a *discrete* version of the cut-off function κ.

Let us quickly recapitulate the relevant properties of the function κ from L.2.55: We had

$$\kappa \in C^\infty(\Omega), \qquad \kappa|_{\Omega \cap B} \equiv 1, \qquad \operatorname{supp}(\kappa) \subseteq \Omega \cap B^\delta,$$

where $B^\delta \in \mathbb{B}$ is a slightly inflated version of the box $B \in \mathbb{B}$ (cf. D.2.8 and D.2.12). A first attempt on a discrete counterpart would obviously be κ's nodal interpolant $\tilde{\kappa} \in \mathbb{S}^{1,1}(\mathcal{T})$. The problem with this approach is that $\tilde{\kappa}$ might have a prohibitively large support. At the very least, the interpolation process adds *one* layer of mesh elements T to the support of κ (which encompasses at least $\Omega \cap B$). This is not problematic in the vicinity of elements $T \in \mathcal{T}$ with "small" diameters $h_T \lesssim \delta$. However, close to elements $T \in \mathcal{T}$ with "large"

diameters $h_T \gtrsim \delta$, even a single layer of additional elements increases the support by h_T, which we cannot control in terms of δ. In particular, even by tweaking δ, there is no hope of achieving $\operatorname{supp}(\tilde{\kappa}) \subseteq \Omega \cap B^\delta$. Here, the reason for failure is that the initial cut-off function κ is coupled to the shape of the box $B \subseteq \mathbb{R}^d$ instead of the (possibly highly irregular) set

$$\bigcup \{T \in \mathcal{T}(B) \,|\, h_T \lesssim \delta\} \subseteq \mathbb{R}^d.$$

In this section, we construct a discrete cut-off function $\tilde{\kappa} \in \mathbb{S}^{1,1}(\mathcal{T})$ that addresses this very problem (cf. L.2.77). First, we need a discrete analogue for the inflated boxes $B^\delta \in \mathbb{B}$ from D.2.12. To this end, recall from D.2.63 that, for every mesh element $T \in \mathcal{T}$, we fixed an incenter $x_T \in T$ in the sense of D.2.16. Furthermore, recall that $h_{\mathcal{B}}$ is the maximal element diameter in a subset $\mathcal{B} \subseteq \mathcal{T}$.

Definition 2.73. *Let $\Omega \subseteq \mathbb{R}^d$ be a polyhedron and $\mathcal{T} \subseteq \operatorname{Pow}(\overline{\Omega})$ be a mesh. For all $\mathcal{B} \subseteq \mathcal{T}$ and all $\delta \geq 0$, we define the* inflated cluster

$$\mathcal{B}^\delta := \{T \in \mathcal{T} \,|\, \exists \tilde{T} \in \mathcal{B} : \|x_T - x_{\tilde{T}}\|_2 \leq \delta\}.$$

Lemma 2.74. *1. For all $\mathcal{B} \subseteq \mathcal{T}$ and $\delta \geq 0$, there holds the bound*

$$h_{\mathcal{B}^\delta} \leq \max\{h_{\mathcal{B}}, 2\sigma_{\mathrm{shp}}\delta\}.$$

2. Let $B \in \mathbb{B}$ be a box, $\mathcal{B} \subseteq \mathcal{T}(B)$ and $\delta > 0$ with $3h_{\mathcal{B}} \leq \delta$. Let $\varepsilon := (6\sigma_{\mathrm{shp}})^{-1}\delta > 0$. Then, there holds the inclusion

$$\bigcup \mathcal{B}^\varepsilon \subseteq \Omega \cap B^\delta.$$

Proof. Ad item 1: Let $T \in \mathcal{B}^\delta$. By definition of $\mathcal{B}^\delta$, there exists an element $\tilde{T} \in \mathcal{B}$ such that $\|x_T - x_{\tilde{T}}\|_2 \leq \delta$. In the case $T = \tilde{T}$, we have $h_T = h_{\tilde{T}} \leq h_{\mathcal{B}}$. On the other hand, if $T \neq \tilde{T}$, then L.2.64 implies $\tilde{T}^\circ \cap T^\circ = \emptyset$ and we find that

$$h_T \overset{\text{L.2.17}}{\leq} 2\sigma_{\mathrm{shp}}\|x_T - x_{\tilde{T}}\|_2 \leq 2\sigma_{\mathrm{shp}}\delta.$$

Ad item 2: Consider an element $T \in \mathcal{B}^\varepsilon$. Note that, using item 1, we have

$$h_T \leq \max\{h_{\mathcal{B}}, 2\sigma_{\mathrm{shp}}\varepsilon\} \overset{\text{Def.}\varepsilon}{=} \max\{h_{\mathcal{B}}, \delta/3\} = \delta/3.$$

Now, according to D.2.73, there exists an element $\tilde{T} \in \mathcal{B}$ such that $\|x_T - x_{\tilde{T}}\|_2 \leq \varepsilon$. Note that $h_{\tilde{T}} \leq h_{\mathcal{B}} \leq \delta/3$, according to the assumption $3h_{\mathcal{B}} \leq \delta$. Since $\tilde{T} \in \mathcal{B} \subseteq \mathcal{T}(B)$, we can pick a point $x \in \tilde{T} \cap B$ and find that, for all $y \in T$,

$$\|y - x\|_2 \leq \|y - x_T\|_2 + \|x_T - x_{\tilde{T}}\|_2 + \|x_{\tilde{T}} - x\|_2 \leq h_T + \varepsilon + h_{\tilde{T}} \leq \delta/3 + \delta/3 + \delta/3 = \delta.$$

Due to L.2.13, this implies $y \in B^\delta$. Taking the union over all $T \in \mathcal{B}^\varepsilon$ and $y \in T$, the desired result follows.

□

Now we know how to inflate a set of mesh elements $\mathcal{B} \subseteq \mathcal{T}$ by a prescribed amount $\delta > 0$. Note that this inflation process is isotropic (uniform in all directions), does not care for the *number* of added "element layers" and also ignores the shape of the underlying domain Ω. In fact, $\mathcal{B}^\delta$ might include mesh elements $T \in \mathcal{T}$ that lie on "the other side" of a gap/hole in Ω. The geodesic distance (i.e., the shortest path *inside* Ω) between the elements $\tilde{T} \in \mathcal{B}$ and the elements $T \in \mathcal{B}^\delta$ might be much longer than the beeline δ. The idea of geodesic distances is the topic of the next definition.

Definition 2.75. *Let $\Omega \subseteq \mathbb{R}^d$ be a polyhedron and $\mathcal{T} \subseteq \operatorname{Pow}(\overline{\Omega})$ be a mesh with nodes $\mathcal{N}$.*

1. *A set $\mathcal{K} = \{N_1, \dots, N_L\} \subseteq \mathcal{N}$ is called* node chain, *if, for every $l \in \{1, \dots, L-1\}$, there exists an element $T_l \in \mathcal{T}$ such that*
$$N_l, N_{l+1} \in \mathcal{N}(T_l).$$

2. *For every node chain $\mathcal{K} = \{N_1, \dots, N_L\} \subseteq \mathcal{N}$ with $L = 1$, we set $|\mathcal{K}| := 0$. If $L \geq 2$, then*
$$|\mathcal{K}| := \sum_{l=1}^{L-1} \|N_{l+1} - N_l\|_2.$$
We refer to $|\mathcal{K}|$ as the length *of the node chain.*

3. *Let $N, M \in \mathcal{N}$. If there exists a node chain $\mathcal{K} \subseteq \mathcal{N}$ with $N, M \in \mathcal{K}$, then we define*
$$\operatorname{dist}_{\mathcal{N}}(N, M) := \min\{|\mathcal{K}| \,|\, \mathcal{K} \subseteq \mathcal{N} \text{ node chain with } N, M \in \mathcal{K}\}.$$
If no such node chain exists, then $\operatorname{dist}_{\mathcal{N}}(N, M) := \infty$. For subsets $\mathcal{M} \subseteq \mathcal{N}$, we set
$$\operatorname{dist}_{\mathcal{N}}(N, \mathcal{M}) := \inf_{M \in \mathcal{M}} \operatorname{dist}_{\mathcal{N}}(N, M) \quad (\in [0, \infty]).$$
We call $\operatorname{dist}_{\mathcal{N}}(\cdot, \cdot)$ the geodesic node distance.

We collect the relevant properties of node chains and the geodesic node distance.

Lemma 2.76. *Let $\Omega \subseteq \mathbb{R}^d$ be a polyhedron and $\mathcal{T} \subseteq \operatorname{Pow}(\overline{\Omega})$ be a mesh with nodes $\mathcal{N}$.*

1. *The function $\operatorname{dist}_{\mathcal{N}}(\cdot, \cdot)$ defines a metric on $\mathcal{N}$ (with values in $[0, \infty]$).*

2. *For all $T \in \mathcal{T}$ and all $\mathcal{M} \subseteq \mathcal{N}$, there holds*
$$\max_{N_1, N_2 \in \mathcal{N}(T)} |\operatorname{dist}_{\mathcal{N}}(N_1, \mathcal{M}) - \operatorname{dist}_{\mathcal{N}}(N_2, \mathcal{M})| \leq h_T.$$

3. *Denote by $\sigma_{\mathrm{lqu}} \geq 1$ the constant from D.2.60 and let $\mathcal{K} = \{N_1, \dots, N_L\} \subseteq \mathcal{N}$ be a chain. Then, for all $T_{\mathrm{start}}, T_{\mathrm{end}} \in \mathcal{T}$ with $N_1 \in \mathcal{N}(T_{\mathrm{start}})$ and $N_L \in \mathcal{N}(T_{\mathrm{end}})$, there holds the bound*
$$h_{T_{\mathrm{end}}} \leq \sigma_{\mathrm{lqu}} \max\{h_{T_{\mathrm{start}}}, 2\sigma_{\mathrm{shp}}|\mathcal{K}|\}.$$

Proof. Ad item 1: We only prove the triangle inequality: Let $N_1, N_2, N_3 \in \mathcal{N}$ and assume that $\operatorname{dist}_{\mathcal{N}}(N_1, N_2)$ and $\operatorname{dist}_{\mathcal{N}}(N_2, N_3)$ are both finite (otherwise, the triangle inequality is trivially fulfilled). Then we can find minimal node chains $\mathcal{K}_{12}, \mathcal{K}_{23} \subseteq \mathcal{N}$ connecting N_1 with N_2 and N_2 with N_3, respectively. We can then easily form a node chain from N_1 to N_3 by $\mathcal{K}_{13} := \mathcal{K}_{12} \cup \mathcal{K}_{23}$. Since N_2 is part of both node chains, we get

$$\operatorname{dist}_{\mathcal{N}}(N_1, N_3) \leq |\mathcal{K}_{13}| \leq |\mathcal{K}_{12}| + |\mathcal{K}_{23}| = \operatorname{dist}_{\mathcal{N}}(N_1, N_2) + \operatorname{dist}_{\mathcal{N}}(N_2, N_3).$$

Ad item 2: Let $T \in \mathcal{T}$ and $N_1, N_2 \in \mathcal{N}(T)$. Since $\mathcal{K} := \{N_1, N_2\}$ *is* a node chain with $N_1, N_2 \in \mathcal{K}$, there holds

$$\operatorname{dist}_{\mathcal{N}}(N_1, N_2) \leq |\mathcal{K}| = \|N_2 - N_1\|_2 \leq h_T.$$

In particular, for all $\mathcal{M} \subseteq \mathcal{N}$,

$$\begin{aligned} \operatorname{dist}_{\mathcal{N}}(N_1, \mathcal{M}) &= \inf_{M \in \mathcal{M}} \operatorname{dist}_{\mathcal{N}}(N_1, M) \\ &\leq \inf_{M \in \mathcal{M}} \operatorname{dist}_{\mathcal{N}}(N_1, N_2) + \operatorname{dist}_{\mathcal{N}}(N_2, M) \leq \operatorname{dist}_{\mathcal{N}}(N_2, \mathcal{M}) + h_T. \end{aligned}$$

An analogous bound holds true for reversed roles of N_1 and N_2. The asserted bound then follows readily.

Ad item 3: In the case $L = 1$, we have $N_1 \in \mathcal{N}(T_{\mathrm{start}}) \cap \mathcal{N}(T_{\mathrm{end}})$, proving that this set is not empty. Then, D.2.60 implies

$$h_{T_{\mathrm{end}}} \leq \sigma_{\mathrm{lqu}} h_{T_{\mathrm{start}}}.$$

In the remaining case $L \geq 2$, there exists an element $T_{L-1} \in \mathcal{T}$ such that $N_{L-1}, N_L \in \mathcal{N}(T_{L-1})$. It follows that $N_L \in \mathcal{N}(T_{L-1}) \cap \mathcal{N}(T_{\mathrm{end}})$, so that

$$h_{T_{\mathrm{end}}} \overset{D.2.60}{\leq} \sigma_{\mathrm{lqu}} h_{T_{L-1}} \overset{L.2.58}{\leq} 2\sigma_{\mathrm{shp}}\sigma_{\mathrm{lqu}} \|N_L - N_{L-1}\|_2 \leq 2\sigma_{\mathrm{shp}}\sigma_{\mathrm{lqu}} |\mathcal{K}|.$$

□

We close this section with the promised discrete cut-off function. The construction is similar to [AFM21a, Lemma 3.18] and makes use of the geodesic node distance $\operatorname{dist}_{\mathcal{N}}(\cdot, \cdot)$.

Lemma 2.77. *Let $\Omega \subseteq \mathbb{R}^d$ be a polyhedron and $\mathcal{T} \subseteq \operatorname{Pow}(\overline{\Omega})$ be a mesh. Denote by $\sigma_{\mathrm{shp}}, \sigma_{\mathrm{lqu}} \geq 1$ the constants from D.2.60. Let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ be such that $\delta \geq 4\sigma_{\mathrm{lqu}} h_{\mathcal{B}}$. Then, there exists a* discrete cut-off function

$$\kappa_{\mathcal{B}}^{\delta} \in \mathbb{S}^{1,1}(\mathcal{T})$$

with the following properties:

1. *There holds the inclusion $\operatorname{supp}_{\mathcal{T}}(\kappa_{\mathcal{B}}^{\delta}) \subseteq \mathcal{B}^{\delta}$.*
2. *There holds $\kappa_{\mathcal{B}}^{\delta}|_{\mathcal{B}} \equiv 1$ and $0 \leq \kappa_{\mathcal{B}}^{\delta} \leq 1$.*
3. *For every $l \in \{0, 1\}$, there holds the stability bound*

$$|\kappa_{\mathcal{B}}^{\delta}|_{W^{l,\infty}(\Omega)} \leq C(d, \sigma_{\mathrm{shp}}, \sigma_{\mathrm{lqu}}) \delta^{-l}.$$

Proof. Denote by $\mathcal{N}$ the nodes of $\mathcal{T}$ and by $\{\psi_N \,|\, N \in \mathcal{N}\} \subseteq \mathbb{S}^{1,1}(\mathcal{T})$ the basis of hat functions from L.2.72. We introduce a subset of nodes $\mathcal{M} \subseteq \mathcal{N}$ and a parameter $\alpha > 0$:

$$\mathcal{M} := \bigcup_{T\in\mathcal{B}} \mathcal{N}(T), \qquad \alpha := (8\sigma_{\mathrm{shp}}\sigma_{\mathrm{lqu}})^{-1} > 0.$$

Now, let

$$\kappa_{\mathcal{B}}^{\delta} := \kappa := \sum_{N\in\mathcal{N}} \kappa_N \psi_N \in \mathbb{S}^{1,1}(\mathcal{T}),$$

where the nodal values κ_N are defined as

$$\kappa_N := \max\left\{0, 1 - \frac{\mathrm{dist}_{\mathcal{N}}(N,\mathcal{M})}{\alpha\delta}\right\} \in [0,1].$$

The idea is that the node values κ_N fall off with a constant slope δ^{-1} along the node chains that make up the shortest connection between any given node $N \in \mathcal{N}$ and the set $\mathcal{M}$.

Ad item 1: Let $T \in \mathrm{supp}_{\mathcal{T}}(\kappa)$, i.e., $\kappa|_T \not\equiv 0$ (cf. D.2.71). Since the polynomial $\kappa|_T \in \mathbb{P}^1(T)$ is uniqely determined by the values $\{\kappa_N \,|\, N \in \mathcal{N}(T)\}$, there must exist a node $N \in \mathcal{N}(T)$ with $\kappa_N \neq 0$, i.e., $\mathrm{dist}_{\mathcal{N}}(N,\mathcal{M}) < \alpha\delta$. Since $\mathrm{dist}_{\mathcal{N}}(N,\mathcal{M})$ is the length of the shortest chain from N to $\mathcal{M} = \bigcup_{\tilde{T}\in\mathcal{B}} \mathcal{N}(\tilde{T})$, we can find an element $T_{\mathrm{start}} \in \mathcal{B}$ and a chain $\mathcal{K} = \{N_1, \dots, N_L\}$ such that

$$N_1 \in \mathcal{N}(T_{\mathrm{start}}), \qquad N_L = N \in \mathcal{N}(T), \qquad |\mathcal{K}| = \mathrm{dist}_{\mathcal{N}}(N,\mathcal{M}) < \alpha\delta.$$

Exploiting both the definition of the parameter α and the assumption $\delta \geq 4\sigma_{\mathrm{lqu}} h_{\mathcal{B}}$, we get the following bound for the Euclidean distance between the incenters $x_{T_{\mathrm{start}}} \in T_{\mathrm{start}}$ and $x_T \in T$:

$$\begin{aligned}\|x_{T_{\mathrm{start}}} - x_T\|_2 &\leq \|x_{T_{\mathrm{start}}} - N_1\|_2 + |\mathcal{K}| + \|N_L - x_T\|_2 \leq h_{T_{\mathrm{start}}} + |\mathcal{K}| + h_T \\ &\overset{L.2.76}{\leq} h_{T_{\mathrm{start}}} + |\mathcal{K}| + \sigma_{\mathrm{lqu}} \max\{h_{T_{\mathrm{start}}}, 2\sigma_{\mathrm{shp}}|\mathcal{K}|\} \leq 2\sigma_{\mathrm{lqu}} h_{\mathcal{B}} + 4\sigma_{\mathrm{shp}}\sigma_{\mathrm{lqu}}\alpha\delta \leq \delta/2 + \delta/2 = \delta.\end{aligned}$$

According to D.2.73, this proves $T \in \mathcal{B}^{\delta}$ and ultimately $\mathrm{supp}_{\mathcal{T}}(\kappa) \subseteq \mathcal{B}^{\delta}$.

Ad item 2: Let $T \in \mathcal{B}$. For all $N \in \mathcal{N}(T)$, there holds $N \in \mathcal{M}$, so that $\mathrm{dist}_{\mathcal{N}}(N,\mathcal{M}) = 0$ and $\kappa_N = 1$. For the remaining nodes $N \in \mathcal{N}\backslash\mathcal{N}(T)$, the support properties of the hat functions ψ_N guarantee $\psi_N|_T \equiv 0$ (cf. L.2.72). Thus,

$$\kappa|_T \overset{L.2.72}{=} \sum_{N\in\mathcal{N}(T)} \kappa_N(\psi_N|_T) + \sum_{N\in\mathcal{N}\backslash\mathcal{N}(T)} \kappa_N(\psi_N|_T) = \Big(\sum_{N\in\mathcal{N}} \psi_N\Big)\Big|_T \overset{L.2.72}{\equiv} 1.$$

Furthermore, since $0 \leq \psi_N \leq 1$ and $\kappa_N \in [0,1]$, we have

$$0 \leq \sum_{N\in\mathcal{N}} \kappa_N\psi_N = \kappa, \qquad \kappa = \sum_{N\in\mathcal{N}} \kappa_N\psi_N \leq \sum_{N\in\mathcal{N}} \psi_N = 1.$$

Ad item 3: The fact that $0 \leq \kappa \leq 1$ immediately gives $\|\kappa\|_{L^\infty(\Omega)} \leq 1$. In order to get a bound for $|\kappa|_{W^{1,\infty}(\Omega)}$, let $T \in \mathcal{T}$ be given. Using the identity $\kappa(N) = \kappa_N$, for all $N \in \mathcal{N}$, and the bound $|\max\{0,t\} - \max\{0,s\}| \leq |t-s|$, for all $t, s \in \mathbb{R}$, we compute

$$\begin{aligned} h_T |\kappa|_{W^{1,\infty}(T)} &\overset{L.\,2.66}{\lesssim} \max_{N_1,N_2 \in \mathcal{N}(T)} |\kappa_{N_1} - \kappa_{N_2}| \\ &\leq (\alpha\delta)^{-1} \max_{N_1,N_2 \in \mathcal{N}(T)} |\mathrm{dist}_{\mathcal{N}}(N_1,\mathcal{M}) - \mathrm{dist}_{\mathcal{N}}(N_2,\mathcal{M})| \\ &\overset{L.\,2.76}{\leq} (\alpha\delta)^{-1} h_T \overset{\mathrm{Def.}\alpha}{\lesssim} \delta^{-1} h_T. \end{aligned}$$

Dividing by h_T and taking the maximum over all $T \in \mathcal{T}$, we obtain $|\kappa|_{W^{1,\infty}(\Omega)} \leq C(d, \sigma_{\mathrm{shp}}, \sigma_{\mathrm{lqu}})\delta^{-1}$.

□

3 Hierarchical matrices

This chapter shall serve as an introduction to the theory of *hierarchical matrices* ($\mathcal{H}$-matrices in short) and contains everything the reader needs to know in preparation for the main result of this thesis, T.4.21. Most importantly, we develop the *block partition* $\mathbb{P}^2$ that we need for the definition of $\mathcal{H}(\mathbb{P}^2, r)$, the class of $\mathcal{H}$-matrices. For more introductory material on $\mathcal{H}$-matrices, we suggest the dissertation [Gra01] and the book [Hac09].

3.1 Motivation

Before we start with the rigorous construction of the block partition $\mathbb{P}^2$, we want to articulate the ideas presented in Section 1.2 in more detail. Denote by

$$\boldsymbol{A} := (a(\varphi_n, \varphi_m))_{m,n=1}^{N} \in \mathbb{R}^{N\times N}$$

the Gram matrix from Section 1.1, i.e., $a(\cdot,\cdot)$ is a bilinear form on some suitable Hilbert space V, $V_N \subseteq V$ is a finite-dimensional subspace and $\{\varphi_1, \dots, \varphi_N\} \subseteq V_N$ is a basis. To be more specific, assume that V is a function space on some computational domain $\Omega \subseteq \mathbb{R}^d$ and that the sets

$$\Omega_n := \operatorname{supp}(\varphi_n) \subseteq \mathbb{R}^d$$

are small, e.g., tiny balls or mesh elements.

We consider two index sets $I, J \subseteq \{1, \dots, N\}$ (so-called *clusters*) and the corresponding physical domains

$$\Omega_I := \bigcup_{i\in I} \Omega_i \subseteq \mathbb{R}^d, \qquad \Omega_J := \bigcup_{j\in J} \Omega_j \subseteq \mathbb{R}^d.$$

Suppose that Ω_I and Ω_J are *well separated* in the following sense: There exist a constant $\sigma_{\text{adm}} > 0$ and boxes $B_I, B_J \in \mathbb{B}$ (cf. D.2.8) such that $\Omega_I \subseteq B_I$, $\Omega_J \subseteq B_J$ and such that the following *admissibility condition* is satisfied:

$$\max\{\operatorname{diam}_2(B_I), \operatorname{diam}_2(B_J)\} \leq \sigma_{\text{adm}} \operatorname{dist}_2(B_I, B_J).$$

Clearly, the interaction between the groups $\{\varphi_i \,|\, i \in I\}$ and $\{\varphi_j \,|\, j \in J\}$ is somehow encoded in the matrix block $\boldsymbol{A}|_{I\times J} \in \mathbb{R}^{I\times J}$. If the physical law behind the bilinear form $a(\cdot,\cdot)$ is governed by a "well-behaved" kernel function (e.g., *asymptotically smooth* as in [Hac09, Section 4.2.4]), then it is safe to assume that the interdependence of the groups $\{\varphi_i \,|\, i \in I\}$ and $\{\varphi_j \,|\, j \in J\}$ can be modeled using fewer bits of information than expected.

In a naive implementation, the memory requirements to store the (possibly fully populated) matrix block $\boldsymbol{A}|_{I\times J} \in \mathbb{R}^{I\times J}$ amount to $\#I\#J$. However, if its singular values decay

rapidly, we can reduce the memory footprint considerably without losing too much information. To this end, recall (e.g., [Str80, Section 6.3]) that any matrix $\boldsymbol{B} \in \mathbb{R}^{I\times J}$ can be written in the form of a *singular values decomposition (SVD)*

$$\boldsymbol{B} = \boldsymbol{X}\boldsymbol{Y}^T \in \mathbb{R}^{I\times J},$$

where $\boldsymbol{X} \in \mathbb{R}^{I\times J}$ is the product of an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{I\times I}$ with a diagonal matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{I\times J}$ and where $\boldsymbol{Y} \in \mathbb{R}^{J\times J}$ is an orthogonal matrix. The diagonal entries of $\boldsymbol{\Sigma}$ are given by the *singular values* $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ of the original matrix $\boldsymbol{B}$. Then, given a *rank bound* $r \in \mathbb{N}$ with $r \ll \min\{\#I, \#J\}$, we can assemble a *truncated singular values decomposition*

$$\boldsymbol{B}_r := \boldsymbol{X}_r\boldsymbol{Y}_r^T \in \mathbb{R}^{I\times J},$$

where $\boldsymbol{X}_r \in \mathbb{R}^{I\times r}$ and $\boldsymbol{Y}_r \in \mathbb{R}^{J\times r}$ only contain the first r columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Note that $\boldsymbol{B}_r$ has the same number of rows and columns as $\boldsymbol{B}$, but the individual matrix entries come from much shorter sums (J_r only contains the first r members of J):

$$\forall (i,j) \in I \times J : \qquad \boldsymbol{B}_{ij} = \sum_{k\in J} \boldsymbol{X}_{ik}\boldsymbol{Y}_{jk}, \qquad (\boldsymbol{B}_r)_{ij} = \sum_{k\in J_r} \boldsymbol{X}_{ik}\boldsymbol{Y}_{jk}.$$

The matrices $\boldsymbol{X}_r$ and $\boldsymbol{Y}_r$ can be regarded as an efficient representation of $\boldsymbol{B}_r$, provided that we store them as separate entities and refrain from carrying out the implied multiplication. While the matrix $\boldsymbol{B}_r$ would need $\#I\#J$ bits of memory, the cumulative cost of storing $\boldsymbol{X}_r$ and $\boldsymbol{Y}_r$ as two separate matrices only amounts to $r(\#I + \#J)$. The truncation error between $\boldsymbol{B}$ and $\boldsymbol{B}_r$ is given by (e.g., [TB97, Theorem 5.8.])

$$\|\boldsymbol{B} - \boldsymbol{B}_r\|_2 = \sigma_{r+1},$$

which should quickly tend to zero (as $r \to \infty$) if the initial matrix $\boldsymbol{B}$ has a "low information content". In particular, we can choose a small value[1] for the rank bound r, so that the reduction in memory cost from $\#I\#J$ to $r(\#I + \#J)$ is indeed significant.

Finally, we mention that the representation of $\boldsymbol{B}_r$ via $\boldsymbol{X}_r$ and $\boldsymbol{Y}_r$ can even be used to perform matrix-vector-multiplications efficiently: Given some input vector $\boldsymbol{c} \in \mathbb{R}^J$, we first compute $\tilde{\boldsymbol{c}} := \boldsymbol{Y}_r^T\boldsymbol{c}$ and then $\boldsymbol{X}_r\tilde{\boldsymbol{c}}$, which produces $\boldsymbol{B}_r\boldsymbol{c} = \boldsymbol{X}_r\boldsymbol{Y}_r^T\boldsymbol{c}$. With this simple two-step procedure, the total work load reduces to $\mathcal{O}(r(\#I + \#J))$, which is again much better than $\mathcal{O}(\#I\#J)$ for the naive matrix-vector-product $\boldsymbol{B}_r\boldsymbol{c}$.

Figure 3.1 shall serve as a visualization of the difference between storing the matrix $\boldsymbol{B}_r$ explicitly versus storing its constituents $\boldsymbol{X}_r$ and $\boldsymbol{Y}_r$ separately. The left-hand matrix $\boldsymbol{B}_r \in \mathbb{R}^{12\times 10}$ needs $12 \cdot 10 = 120$ units of memory, whereas the right-hand matrices $\boldsymbol{X}_r \in \mathbb{R}^{12\times 2}$ and $\boldsymbol{Y}_r \in \mathbb{R}^{10\times 2}$ only need $12 \cdot 2 + 10 \cdot 2 = 44$ units. Although $\boldsymbol{B}_r$ is an almost fully populated matrix here, its information content is less than 37%.

[1] Typical real world applications require r to be chosen on the order of $\mathcal{O}(\ln(N))$ or even $\mathcal{O}(1)$.

$$\begin{bmatrix} -4 & 9 & -6 & -1 & 0 & -6 & 8 & -4 & -6 & -2 \\ 8 & -17 & 12 & 5 & 2 & 10 & -14 & 6 & 8 & 8 \\ -6 & 10 & -9 & -12 & -7 & -2 & 5 & 1 & 5 & -17 \\ -4 & 10 & -6 & 2 & 2 & -8 & 10 & -6 & -10 & 2 \\ -6 & 15 & -9 & 3 & 3 & -12 & 15 & -9 & -15 & 3 \\ 6 & -9 & 9 & 15 & 9 & 0 & -3 & -3 & -9 & 21 \\ -4 & 7 & -6 & -7 & -4 & -2 & 4 & 0 & 2 & -10 \\ 2 & -8 & 3 & -10 & -7 & 10 & -11 & 9 & 17 & -13 \\ 8 & -12 & 12 & 20 & 12 & 0 & -4 & -4 & -12 & 28 \\ 2 & -8 & 3 & -10 & -7 & 10 & -11 & 9 & 17 & -13 \\ 6 & -14 & 9 & 0 & -1 & 10 & -13 & 7 & 11 & 1 \\ 0 & -1 & 0 & -3 & -2 & 2 & -2 & 2 & 4 & -4 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -4 & 1 \\ 3 & 2 \\ 2 & -2 \\ 3 & -3 \\ -3 & -3 \\ 2 & 1 \\ -1 & 4 \\ -4 & -4 \\ -1 & 4 \\ -3 & 2 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} -2 & 0 \\ 4 & -1 \\ -3 & 0 \\ -2 & -3 \\ -1 & -2 \\ -2 & 2 \\ 3 & -2 \\ -1 & 2 \\ -1 & 4 \\ -3 & -4 \end{bmatrix}^T$$

Figure 3.1: Representing a (12×10)-matrix by a (12×2)- and a (10×2)-matrix.

Now let us return to the Gram matrix $\boldsymbol{A} = (a(\varphi_n, \varphi_m))_{m,n=1}^N$ from the beginning of this section. Truncated SVDs promise good compression rates for blocks $\boldsymbol{A}|_{I\times J} \in \mathbb{R}^{I\times J}$, whose domains $\Omega_I, \Omega_J \subseteq \mathbb{R}^d$ are well separated (i.e., included in admissible boxes). However, there will also be matrix blocks $\boldsymbol{A}|_{I\times J}$, whose domains Ω_I and Ω_J are *not* well-separated (e.g., on the diagonal of $\boldsymbol{A}$, where $I \cap J \neq \emptyset$ and thus $\Omega_I \cap \Omega_J \neq \emptyset$). In these cases, we need to make sure that $\min\{\#I, \#J\}$ is smaller than some predefined threshold $\sigma_{\text{small}} \geq 1$. Then, we can simply store the full matrix block $\boldsymbol{A}|_{I\times J}$ as is, which results in a memory cost of

$$\#I\#J = \min\{\#I, \#J\} \cdot \max\{\#I, \#J\} \leq \sigma_{\text{small}}(\#I + \#J).$$

We are now left with the following task:

Problem 3.1. *Construct a partition of the full matrix index set* $\{1, \dots, N\} \times \{1, \dots, N\}$ *into blocks* $I \times J$*, such that each block satisfies one or both of the following conditions:*

1. *There holds* $\min\{\#I, \#J\} \leq \sigma_{\text{small}}$.
2. *The physical sets* $\Omega_I, \Omega_J \subseteq \mathbb{R}^d$ *are well separated.*

In the remainder of this chapter, we will use the *adaptive, geometrically balanced clustering strategy* from [GHLB04] to construct such a partition. Before we proceed, a few remarks are necessary.

Remark 3.2. *In our discussion of truncated SVDs, we argued that the* pair *of matrices* $(\boldsymbol{X}_r, \boldsymbol{Y}_r)$ *can be seen as an efficient representation of* $\boldsymbol{B}_r$*. Continuing this theme, the elements of the block partition* $\mathbb{P}^2$ *will be* pairs (I, J) *of clusters* $I, J \subseteq \{1, \dots, N\}$*, as opposed to cartesian products* $I \times J$*. We hope that this slight misnomer is no cause for confusion.*

Remark 3.3. *For the ease of presentation, the motivation was formulated in terms of the Gram matrix* $\boldsymbol{A}$ *itself. However, the ultimate goal of this thesis is an approximation*

result for its inverse $\boldsymbol{A}^{-1}$. The heuristic still applies, though, because our assumptions on the bilinear form $a(\cdot,\cdot)$ and the basis functions φ_i in Chapter 4 will guarantee that $\boldsymbol{A}^{-1}$ is again a Gram matrix (cf. L.4.15).

Remark 3.4. *A common compression technique for large matrices $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is exploiting sparsity. If most entries of $\boldsymbol{A}$ are zero, then it might be cheaper to store the positions and values of the non-zero-entries in three short lists and use these to represent $\boldsymbol{A}$. We want to emphasize, however, that the cost efficiency of $\mathcal{H}$-matrices* does not *come from sparsity. In fact, an $\mathcal{H}$-matrix $\boldsymbol{A}$ might very well be fully populated. The only requirement for $\boldsymbol{A}$ to be an $\mathcal{H}$-matrix is that the "admissible" subblocks $\boldsymbol{A}|_{I\times J}$ must have a small rank, so that a cheap representation via matrix pairs $(\boldsymbol{X}_r, \boldsymbol{Y}_r)$ is possible.*

3.2 The characteristic sets Ω_n

Definition 3.5. *Let $\sigma_{\mathrm{shp}}, \sigma_{\mathrm{ovlp}}, \sigma_{\mathrm{sprd}} \geq 1$ and $N \in \mathbb{N}$. We consider a family of subsets*

$$\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$$

with shape regularity σ_{shp}, overlap σ_{ovlp} and spread σ_{sprd} (cf. D.2.16, D.2.18, D.2.21). The sets Ω_n are called characteristic sets.

Definition 3.6.

1. *A subset $I \subseteq \{1,\dots,N\}$ is called* cluster.
2. *For all clusters $I \subseteq \{1,\dots,N\}$, we define*

$$\Omega_I := \bigcup_{n\in I} \Omega_n \subseteq \mathbb{R}^d.$$

3. *For every $n \in \{1,\dots,N\}$, we fix an incenter $x_n \in \Omega_n$ (cf. D.2.16). The points $x_1,\dots,x_N$ are called* characteristic points.
4. *For all $n \in \{1,\dots,N\}$ and $I \subseteq \{1,\dots,N\}$, we set (cf. D.2.14)*

$$h_n := h_{\Omega_n}, \quad h_I := h_{\max,I} := \max_{n\in I} h_n, \quad h_{\min,I} := \min_{n\in I} h_n, \quad h_{\min} := \min_{n\in\{1,\dots,N\}} h_n.$$

The upcoming clustering algorithm mainly deals with the characteristic points $x_1,\dots,x_N$ and the quantities σ_{shp}, σ_{ovlp}, σ_{sprd}. The characteristic sets $\Omega_1,\dots,\Omega_N$ only play a minor role.

Definition 3.7. *Denote by $x_1,\dots,x_N \in \mathbb{R}^d$ the characteristic points from D.3.6. For every subset $B \subseteq \mathbb{R}^d$, we define the corresponding cluster*

$$\iota(B) := \{n \in \{1,\dots,N\} \,|\, x_n \in B\}.$$

3.3 The box tree $\mathbb{T}$

In this section, we use a *geometric clustering algorithm* to construct a *box tree* $\mathbb{T}$:

1. The nodes of $\mathbb{T}$ are axes-parallel boxes $B \in \mathbb{B}$ (cf. D.2.8).
2. The root is a large box containing all of the characteristic points $x_1, \dots, x_N$.
3. Using the splitting method $\mathrm{sons}(\cdot) : \mathbb{B} \longrightarrow \mathrm{Pow}(\mathbb{B})$ from D.2.10, the initial box is successively split into smaller boxes.
4. The splitting stops when the number of characteristic points x_n inside the current box falls below a predefined threshold.
5. The boxes at the leaves of the tree form a partition of the root box. In particular, they can be used to partition the points $x_1, \dots, x_N$ (and thus the abstract index set $\{1, \dots, N\}$).

To get the clustering algorithm going, we need a box to start with. The assumption about the family $\{\Omega_1, \dots, \Omega_N\}$ having spread σ_{sprd} allows us to utilize a previous result.

Definition 3.8. *Denote by $B_{\mathrm{start}} \in \mathbb{B}$ the box from L.2.22.*

Lemma 3.9. *There hold the following properties:*

$$\begin{aligned} \Omega_1, \dots, \Omega_N &\subseteq \overline{B_{\mathrm{start}}}, & \iota(B_{\mathrm{start}}) &= \{1, \dots, N\}, \\ \mathrm{diam}_2(B_{\mathrm{start}}) &= \sqrt{d}\sigma_{\mathrm{sprd}}, & \mathrm{meas}(B_{\mathrm{start}}) &= \sigma_{\mathrm{sprd}}^d. \end{aligned}$$

Proof. See L.2.22. □

Next, we introduce the threshold for the stopping criterion. We will refer to this number as being a *clustering parameter*.

Definition 3.10. *Denote by $\sigma_{\mathrm{ovlp}} \geq 1$ the quantity from D.3.5. Let $\sigma_{\mathrm{small}} \geq 1$ be a number that satisfies*

$$\sigma_{\mathrm{ovlp}} \leq \sigma_{\mathrm{small}}.$$

For a given system of characteristic sets $\Omega_1, \dots, \Omega_N$, it might be impossible to determine the precise value of σ_{ovlp}. However, if the sets Ω_n are constructed algorithmically, a theoretical upper bound to σ_{ovlp} may be available (e.g., L.6.8). D.3.10 then says that σ_{small} must be chosen larger than this theorical bound.

We encode the stopping criterion of the algorithm in form of a subset $\mathbb{B}_{\mathrm{stop}} \subseteq \mathbb{B}$. If a branch of the tree $\mathbb{T}$ reaches $\mathbb{B}_{\mathrm{stop}}$, it won't grow any further.

Definition 3.11. *Let $\sigma_{\mathrm{small}} \geq 1$ be the clustering parameter from D.3.10. A box $B \in \mathbb{B}$ is called* small, *if there holds*

$$\#\iota(B) \leq \sigma_{\mathrm{small}}.$$

We set

$$\mathbb{B}_{\mathrm{stop}} := \{B \in \mathbb{B} \mid B \text{ is small}\} \subseteq \mathbb{B}.$$

The boxes $B \in \mathbb{B}$ that have not yet reached the stopping set $\mathbb{B}_{\text{stop}}$ must contain more than σ_{small} characteristic points x_n. Since we assumed $\sigma_{\text{small}} \geq \sigma_{\text{ovlp}}$ in D.3.10, such a box B then also contains more than σ_{ovlp} characteristic points. According to L.2.19, this puts an upper limit on the diameters h_n of the characteristic sets Ω_n:

Lemma 3.12. *Denote by $\sigma_{\text{shp}} \geq 1$ the quantity from D.3.5. Then, for every box $B \in \mathbb{B}\backslash\mathbb{B}_{\text{stop}}$, there holds the bound*

$$h_{\iota(B)} \leq 2\sigma_{\text{shp}}\text{diam}_2(B).$$

Proof. Since $B \notin \mathbb{B}_{\text{stop}}$, we know from D.3.11 and D.3.10 that $\#\iota(B) > \sigma_{\text{small}} \geq \sigma_{\text{ovlp}}$. Then, using the fact that $x_n \in B$, for all $n \in \iota(B)$ (cf. D.3.7), we get

$$h_{\iota(B)} \overset{D.3.6}{=} \max_{n\in\iota(B)} h_{\Omega_n} \overset{L.2.19}{\leq} 2\sigma_{\text{shp}} \max_{m,n\in\iota(B)} \|x_m - x_n\|_2 \leq 2\sigma_{\text{shp}}\text{diam}_2(B).$$

□

Now that we know how to start, split and stop, we can construct the individual levels of the box tree $\mathbb{T}$:

Definition 3.13. *Let $B_{\text{start}} \in \mathbb{B}$ be the box from D.3.8, let $\text{sons}(\cdot) : \mathbb{B} \longrightarrow \text{Pow}(\mathbb{B})$ be the splitting procedure defined in D.2.10 and denote by $\sigma_{\text{small}} \geq 1$ the clustering parameter from D.3.10. Furthermore, let $\mathbb{B}_{\text{stop}} \subseteq \mathbb{B}$ be defined as in D.3.11. We define a sequence $(\mathbb{T}_l)_{l\in\mathbb{N}}$ of subsets $\mathbb{T}_l \subseteq \mathbb{B}$ in a recursive manner:*

$$\begin{aligned} & \mathbb{T}_1 := \{B_{\text{start}}\}, \\ \forall l \geq 2 : \quad & \mathbb{T}_l := \{B \,|\, A \in \mathbb{T}_{l-1}\backslash\mathbb{B}_{\text{stop}},\ B \in \text{sons}(A)\}. \end{aligned}$$

Remark 3.14. *This recursive definition can easily be converted to an actual computer program. Assuming the level $\mathbb{T}_{l-1} \subseteq \mathbb{B}$ has already been computed, we simply iterate over all $A \in \mathbb{T}_{l-1}$ and determine the corresponding index cluster $\iota(A) \subseteq \{1, \ldots, N\}$ through a series of "point-in-box" checks (cf. D.3.7). If the final score $\#\iota(A)$ exceeds σ_{small}, then A is split up and produces 2^d sons on the next level $\mathbb{T}_l$.*

The number of "point-in-box" checks can be reduced significantly, if the boxes A are stored along with their associated clusters $\iota(A)$. In this case, the tree nodes are the pairs $(A, \iota(A))$ and we split A and $\iota(A)$ simultaneously. The clusters $\iota(B)$ of the children $B \in \text{sons}(A)$ can then be determined from $\iota(A)$ and we don't need to go through all N points $x_1, \ldots, x_N$ afresh.

In fact, we could also work with an alternate definition of $\mathbb{T}$ where every node is a pair (B, I) of a box $B \in \mathbb{B}$ and a cluster $I \subseteq \{1, \ldots, N\}$. However, the upcoming results really only depend on the properties of the boxes B and not so much on the properties of the index sets I. Therefore, we proceed with the original definition of $\mathbb{T}$, which is far less cumbersome anyways.

We summarize the most important properties of the sequence $(\mathbb{T}_l)_{l\in\mathbb{N}}$:

Lemma 3.15. *1. Let $l \in \mathbb{N}$. For all $B \in \mathbb{T}_l$, there holds $B \subseteq B_{\mathrm{start}}$. Furthermore, for all $B, \tilde{B} \in \mathbb{T}_l$ with $B \neq \tilde{B}$, there holds $B \cap \tilde{B} = \emptyset$. In particular,*

$$\dot{\bigcup_{B \in \mathbb{T}_l}} B \subseteq B_{\mathrm{start}}.$$

2. For all $l \in \mathbb{N}$ and all $B \in \mathbb{T}_l$, there hold the following identities:

$$\mathrm{diam}_2(B) = 2\sqrt{d}\sigma_{\mathrm{sprd}} 2^{-l}, \qquad \mathrm{meas}(B) = (2\sigma_{\mathrm{sprd}})^d 2^{-dl}.$$

3. For all $l, k \in \mathbb{N}$ with $l \neq k$, there holds $\mathbb{T}_l \cap \mathbb{T}_k = \emptyset$.

Proof. Ad item 1: We only prove the statement about disjointness. Induction basis $l = 1$: Trivial, since $\mathbb{T}_1$ only contains one element. Induction step $l - 1 \mapsto l$: Let $B, \tilde{B} \in \mathbb{T}_l$ with $B \neq \tilde{B}$ be given. By definition of $\mathbb{T}_l$, there exist $A, \tilde{A} \in \mathbb{T}_{l-1} \backslash \mathbb{B}_{\mathrm{stop}}$ such that $B \in \mathrm{sons}(A)$ and $\tilde{B} \in \mathrm{sons}(\tilde{A})$. If $A = \tilde{A}$, then $B, \tilde{B} \in \mathrm{sons}(A)$ and L.2.11 yields $B \cap \tilde{B} = \emptyset$. On the other hand, if $A \neq \tilde{A}$, then $B \cap \tilde{B} \subseteq A \cap \tilde{A} = \emptyset$ by the induction hypothesis.

Ad item 2: Follows easily from L.2.11 and L.3.9 by induction on l.

Ad item 3: If $\mathbb{T}_l \cap \mathbb{T}_k \neq \emptyset$, then we can find a box B with $B \in \mathbb{T}_l$ and $B \in \mathbb{T}_k$. Then, using item 2, we get

$$l = -\log_2(2^{-l}) = -\log_2((2\sqrt{d}\sigma_{\mathrm{sprd}})^{-1}\mathrm{diam}_2(B)) = -\log_2(2^{-k}) = k.$$

□

Next, let us demonstrate that the sequence $(\mathbb{T}_l)_{l \in \mathbb{N}}$ must halt at some point. The determining factors are the quantities $\sigma_{\mathrm{shp}}, \sigma_{\mathrm{sprd}} \geq 1$ and $h_{\min} > 0$ from D.3.5 and D.3.6.

Lemma 3.16. *There exists an $l \in \mathbb{N}$ such that $\mathbb{T}_l \subseteq \mathbb{B}_{\mathrm{stop}}$. The minimizer*

$$L := \min\{l \in \mathbb{N} \,|\, \mathbb{T}_l \subseteq \mathbb{B}_{\mathrm{stop}}\}$$

has the following properties:

1. There holds $\mathbb{T}_L \neq \emptyset$ and $\mathbb{T}_{L+1} = \mathbb{T}_{L+2} = \cdots = \emptyset$.

2. There holds the bound

$$L \leq \log_2(8\sqrt{d}\sigma_{\mathrm{shp}}\sigma_{\mathrm{sprd}}h_{\min}^{-1}).$$

Proof. Ad item 1: If there existed a sequence $(B_l)_{l \in \mathbb{N}}$ such that $B_l \in \mathbb{T}_l \backslash \mathbb{B}_{\mathrm{stop}}$, for all $l \in \mathbb{N}$, then we would get the following contradiction:

$$0 < h_{\min} \leq h_{\iota(B_l)} \overset{L.3.12}{\leq} 2\sigma_{\mathrm{shp}}\mathrm{diam}_2(B_l) \overset{L.3.15}{=} 4\sqrt{d}\sigma_{\mathrm{shp}}\sigma_{\mathrm{sprd}}2^{-l} \overset{l}{\to} 0.$$

Now, denote by $L \in \mathbb{N}$ the minimal value such that $\mathbb{T}_L \subseteq \mathbb{B}_{\mathrm{stop}}$. If $L = 1$, then trivially $\mathbb{T}_L = \{B_{\mathrm{start}}\} \neq \emptyset$. If $L \geq 2$, then $\mathbb{T}_{L-1} \not\subseteq \mathbb{B}_{\mathrm{stop}}$, due to the minimality assumption. It follows that $\mathbb{B}_{\mathrm{stop}} \backslash \mathbb{T}_{L-1} \neq \emptyset$ and ultimately $\mathbb{T}_L \neq \emptyset$. As for the subsequent levels, the

relation $\mathbb{T}_L \backslash \mathbb{B}_{\text{stop}} = \emptyset$ immediately implies $\mathbb{T}_{L+1} = \emptyset$. Then, inductively, $\mathbb{T}_l = \emptyset$, for all $l \geq L+1$.

Ad item 2: In the case $L \geq 2$, the minimality of L tells us that there must exist a $B \in \mathbb{T}_{L-1}$ with $B \notin \mathbb{B}_{\text{stop}}$. Using the same estimates as in the previous step, we find that $h_{\min} \leq 4\sqrt{d}\sigma_{\text{shp}}\sigma_{\text{sprd}}2^{-(L-1)}$. This can easily be rearranged into the equivalent form $L \leq \log_2(8\sqrt{d}\sigma_{\text{shp}}\sigma_{\text{sprd}}h_{\min}^{-1})$. In the remaining case $L = 1$, the bound

$$h_{\min} \overset{D.3.6}{=} \min_{n\in\{1,\dots,N\}} \operatorname{diam}_2(\Omega_n) \overset{L.3.9}{\leq} \operatorname{diam}_2(B_{\text{start}}) \overset{L.3.9}{=} \sqrt{d}\sigma_{\text{sprd}} \leq 4\sqrt{d}\sigma_{\text{shp}}\sigma_{\text{sprd}}$$

readily yields $L = \log_2(2) \leq \log_2(8\sqrt{d}\sigma_{\text{shp}}\sigma_{\text{sprd}}h_{\min}^{-1})$ as well.

□

The previous lemma tells us that the sequence $(\mathbb{T}_l)_{l\in\mathbb{N}}$ terminates after roughly $\ln(h_{\min}^{-1})$ steps. The non-trivial levels constitute the *box tree* $\mathbb{T}$ to be defined next. Keep in mind that the levels $\mathbb{T}_l$ are pairwise disjoint, i.e., a box $B \in \mathbb{B}$ cannot occur on more than one level (cf. L.3.15).

Definition 3.17. *Let $(\mathbb{T}_l)_{l\in\mathbb{N}}$ be the sequence from D.3.13 and $L \in \mathbb{N}$ be the value from L.3.16. We define the* box tree

$$\mathbb{T} := \bigcupdot_{l=1}^{L} \mathbb{T}_l \subseteq \mathbb{B}$$

and set $\operatorname{depth}(\mathbb{T}) := L$. *The sets $\mathbb{T}_l$ are called* levels of $\mathbb{T}$. *Furthermore, we say that $\mathbb{T}$* is based on the clustering parameters B_{start}, $\operatorname{sons}(\cdot)$ and σ_{small} *(D.3.8, D.2.10, D.3.10).*

For the remainder of this section, we are concerned with the computational complexity of storing and assembling the box tree $\mathbb{T}$.

Lemma 3.18. *Let $B_{\text{start}} \in \mathbb{B}$ be the box from D.3.8. Furthermore, let $f : \operatorname{Pow}(\mathbb{R}^d) \longrightarrow [0,\infty)$ be an additive function, i.e., for all $M, \tilde{M} \subseteq \mathbb{R}^d$ with $M \cap \tilde{M} = \emptyset$, there holds*

$$f(M \cup \tilde{M}) = f(M) + f(\tilde{M}).$$

Then, there holds the bound

$$\sum_{B\in\mathbb{T}} f(B) \leq \operatorname{depth}(\mathbb{T}) f(B_{\text{start}}).$$

Proof. Every non-negative, additive function is monotone, i.e., for all $M, \tilde{M} \subseteq \mathbb{R}^d$ with $M \subseteq \tilde{M}$, there holds

$$f(M) \leq f(M) + f(\tilde{M}\backslash M) = f(M \cup (\tilde{M}\backslash M)) = f(\tilde{M}).$$

Now, abbreviate $L := \operatorname{depth}(\mathbb{T})$. Recall from L.3.15 that, for each $l \in \mathbb{N}$, the boxes $B \in \mathbb{T}_l$ are pairwise disjoint and that $\bigcupdot_{B\in\mathbb{T}_l} B \subseteq B_{\text{start}}$. Therefore,

$$\sum_{B\in\mathbb{T}} f(B) \overset{D.3.17}{\leq} \sum_{l=1}^{L} \sum_{B\in\mathbb{T}_l} f(B) = \sum_{l=1}^{L} f\Big(\bigcupdot_{B\in\mathbb{T}_l} B \Big) \leq \sum_{l=1}^{L} f(B_{\text{start}}) = L f(B_{\text{start}}).$$

□

L.3.18 allows us to find an upper bound for $\#\mathbb{T}$ in terms of the quantity $\sigma_{\text{small}} \geq 1$ from D.3.10, the tree depth from D.3.17 and the number of characteristic points.

Lemma 3.19. *There holds the bound*

$$\#\mathbb{T} \leq C(d)\sigma_{\text{small}}^{-1}\text{depth}(\mathbb{T})N.$$

Proof. Abbreviating $L := \text{depth}(\mathbb{T})$, we compute

$$\#\mathbb{T} \leq \sum_{l=1}^{L} \#\mathbb{T}_l \overset{D.3.13}{\lesssim} \sum_{l=2}^{L} \#(\mathbb{T}_{l-1}\backslash\mathbb{B}_{\text{stop}}) \overset{L.3.15}{=} \#\left(\left(\dot{\bigcup_{l=2}^{L}} \mathbb{T}_{l-1}\right)\backslash\mathbb{B}_{\text{stop}}\right) \leq \#(\mathbb{T}\backslash\mathbb{B}_{\text{stop}})$$

$$= \sum_{B\in\mathbb{T}\backslash\mathbb{B}_{\text{stop}}} 1 \overset{D.3.11}{<} \sigma_{\text{small}}^{-1} \sum_{B\in\mathbb{T}} \#\iota(B) \overset{L.3.18}{\leq} \sigma_{\text{small}}^{-1}\text{depth}(\mathbb{T})\#\iota(B_{\text{start}}) \overset{L.3.9}{=} \sigma_{\text{small}}^{-1}\text{depth}(\mathbb{T})N.$$

This concludes the proof. □

Remark 3.20. *In R.3.14, we argued that one should consider storing the pairs $(B, \iota(B))$ instead of just B. In this case, the total memory usage and assembly time of $\mathbb{T}$ amounts to $\mathcal{O}(\sum_{B\in\mathbb{T}} \#\iota(B))$. Looking at the proof of L.3.19 again, we actually showed that*

$$\sum_{B\in\mathbb{T}} \#\iota(B) \leq \text{depth}(\mathbb{T})N.$$

As for the memory usage alone, we can do even better: After splitting a node $(A, \iota(A))$ into its sons $(B, \iota(B))$, we can relabel the characteristic points $x_1, \ldots, x_N$ in a way such that $\iota(B)$ is a contiguous subset of $\mathbb{N}$. Regardless of its cardinality, such a cluster can be represented by 2 numbers alone, namely its minimum and its maximum. Then, the cost of storing $\mathbb{T}$ is again on the order of $\mathcal{O}(\#\mathbb{T})$.

3.4 The product box tree $\mathbb{T}^2$

The leaves of the box tree $\mathbb{T}$ can be used to partition the index set $\{1, \ldots, N\}$ into a family of clusters $I \subseteq \{1, \ldots, N\}$. Next, we introduce the *product box tree* $\mathbb{T}^2$ that will allow us to divide the set $\{1, \ldots, N\} \times \{1, \ldots, N\}$ into a *block partition*. One might be tempted to use the tensor product tree $\mathbb{T} \times \mathbb{T} = \{(B_1, B_2) \,|\, B_1, B_2 \in \mathbb{T}\}$ for this task, but, according to L.3.19, this structure contains up to $\#(\mathbb{T} \times \mathbb{T}) = \#\mathbb{T}\#\mathbb{T} \lesssim \sigma_{\text{small}}^{-2}\text{depth}(\mathbb{T})^2N^2$ members, which is prohibitively large. Instead, we construct a substructure $\mathbb{T}^2 \subseteq \mathbb{T} \times \mathbb{T}$ that contains just $\mathcal{O}(\#\mathbb{T})$ elements.

The construction of the product box tree $\mathbb{T}^2$ is very similar to the one of $\mathbb{T}$:

1. The nodes of $\mathbb{T}^2$ are pairs (B_1, B_2) of axes-parallel boxes $B_1, B_2 \in \mathbb{B}$.
2. The root is just the pair $(B_{\text{start}}, B_{\text{start}})$.
3. Splitting a pair $(A_1, A_2) \in \mathbb{B}^2$ means splitting A_1 and A_2 individually (cf. D.2.10) and forming all pairs (B_1, B_2) of boxes $B_1 \in \text{sons}(A_1)$ and $B_2 \in \text{sons}(A_2)$.

4. The splitting of a pair (B_1, B_2) stops as soon as it is deemed *small* or *admissible* (cf. D.3.22).

5. The leaves of the tree can be used to form a partition of $\{1, \dots, N\} \times \{1, \dots, N\}$.

The stopping criterion is the biggest conceptual novelty, because it involves the spatial distance *between* the boxes B_1 and B_2 that make up the pair (B_1, B_2). This kind of coupling is the reason why $\mathbb{T}^2$ contains significantly fewer elements than the full tensor product $\mathbb{T} \times \mathbb{T}$. We begin by laying out the playing field for the algorithm.

Definition 3.21. *We denote the set of pairs of boxes by* $\mathbb{B}^2 := \mathbb{B} \times \mathbb{B}$.

We already know what the root of the tree $\mathbb{T}^2$ will be and we also know how to split pairs of boxes. It remains to define a rigorous stopping criterion.

Definition 3.22. *Let* $\sigma_{\text{small}} \geq 1$ *be the clustering parameter from D.3.10. Let* $\tilde{\sigma}_{\text{adm}} > 0$ *be a another clustering parameter. A pair of boxes* $(B_1, B_2) \in \mathbb{B}^2$ *is called* small*, if*

$$\min\{\#\iota(B_1), \#\iota(B_2)\} \leq \sigma_{\text{small}}.$$

Similarly, the pair (B_1, B_2) *is called* admissible*, if*

$$\max\{\operatorname{diam}_2(B_1), \operatorname{diam}_2(B_2)\} \leq \tilde{\sigma}_{\text{adm}} \operatorname{dist}_2(B_1, B_2).$$

We set

$$\mathbb{B}^2_{\text{stop}} := \{(B_1, B_2) \in \mathbb{B}^2 \mid (B_1, B_2) \text{ is small or admissible}\}.$$

Note that a pair (B_1, B_2) is small, iff at least one of its components is small in the sense of D.3.11. Furthermore, we emphasize that $\mathbb{B}^2_{\text{stop}}$ is *not* the same as $\mathbb{B}_{\text{stop}} \times \mathbb{B}_{\text{stop}}$ (see L.3.24 below).

Remark 3.23. *In the literature on* $\mathcal{H}$*-matrices, the admissibility of a pair* $(B_1, B_2) \in \mathbb{B}^2$ *is sometimes phrased in terms of the* minimum *of the diameters, i.e., via the relation*

$$\min\{\operatorname{diam}_2(B_1), \operatorname{diam}_2(B_2)\} \leq \tilde{\sigma}_{\text{adm}} \operatorname{dist}_2(B_1, B_2).$$

However, as we shall see later in L.3.26, if a pair (B_1, B_2) *is an element of the product cluster tree* $\mathbb{T}^2$*, then its components* B_1 *and* B_2 *lie on the same level of the box tree* $\mathbb{T}$*. L.3.15 then implies that their diameters must coincide, so that, trivially,*

$$\min\{\operatorname{diam}_2(B_1), \operatorname{diam}_2(B_2)\} = \max\{\operatorname{diam}_2(B_1), \operatorname{diam}_2(B_2)\}.$$

Hence, in the case of geometrically balanced clustering, the two notions of admissibility are equivalent.

Lemma 3.24. *There hold the following inclusions:*

$$\begin{aligned} \mathbb{B}_{\text{stop}} \times \mathbb{B}_{\text{stop}} \subseteq (\mathbb{B}_{\text{stop}} \times \mathbb{B}) \cup (\mathbb{B} \times \mathbb{B}_{\text{stop}}) &\subseteq \mathbb{B}^2_{\text{stop}}, \\ \{B \in \mathbb{B} \mid (B, B) \in \mathbb{B}^2_{\text{stop}}\} &\subseteq \mathbb{B}_{\text{stop}}. \end{aligned}$$

Proof. We only prove the bottom line. Let $B \in \mathbb{B}$ be such that $(B, B) \in \mathbb{B}^2_{\text{stop}}$. According to D.3.22, the pair (B, B) is small or admissible. However, it cannot be admissible, since otherwise

$$0 \overset{L.2.9}{<} \operatorname{diam}_2(B) = \max\{\operatorname{diam}_2(B), \operatorname{diam}_2(B)\} \leq \tilde{\sigma}_{\text{adm}} \operatorname{dist}_2(B, B) = 0.$$

Therefore, (B, B) must be small, so that $\#\iota(B) = \min\{\#\iota(B), \#\iota(B)\} \leq \sigma_{\text{small}}$. According to D.3.11, this means $B \in \mathbb{B}_{\text{stop}}$. □

The levels of the product box tree $\mathbb{T}^2$ are again defined recursively:

Definition 3.25. *Let $B_{\text{start}} \in \mathbb{B}$ be the box from D.3.8, let $\operatorname{sons}(\cdot) : \mathbb{B} \longrightarrow \operatorname{Pow}(\mathbb{B})$ be the splitting procedure from D.2.10 and denote by $\sigma_{\text{small}} \geq 1$ and $\tilde{\sigma}_{\text{adm}} > 0$ the clustering parameters from D.3.10 and D.3.22. Furthermore, let $\mathbb{B}^2_{\text{stop}} \subseteq \mathbb{B}^2$ be defined as in D.3.22. We define a sequence $(\mathbb{T}^2_l)_{l \in \mathbb{N}}$ of subsets $\mathbb{T}^2_l \subseteq \mathbb{B}^2$ in a recursive manner:*

$$\begin{aligned} & \mathbb{T}^2_1 &:=\ & \{(B_{\text{start}}, B_{\text{start}})\}, \\ \forall l \geq 2 : \quad & \mathbb{T}^2_l &:=\ & \{(B_1, B_2) \,|\, (A_1, A_2) \in \mathbb{T}^2_{l-1} \backslash \mathbb{B}^2_{\text{stop}},\ B_1 \in \operatorname{sons}(A_1),\ B_2 \in \operatorname{sons}(A_2)\}. \end{aligned}$$

Once again, $\mathbb{T}^2_l$ is *not* the same as $\mathbb{T}_l \times \mathbb{T}_l$. The reason being that a pair $(A_1, A_2) \in \mathbb{T}^2_{l-1}$ can be admissible without A_1 or A_2 being small. Such a pair will not produce any sons in $\mathbb{T}^2_l$, whereas A_1 and A_2 *do* have sons in $\mathbb{T}_l$.

Lemma 3.26. *1. For all $l \in \mathbb{N}$, there hold the inclusions*

$$\{(B, B) \,|\, B \in \mathbb{T}_l\} \ \subseteq\ \mathbb{T}^2_l \ \subseteq\ \mathbb{T}_l \times \mathbb{T}_l.$$

2. For all $l, k \in \mathbb{N}$ with $l \neq k$, there holds $\mathbb{T}^2_l \cap \mathbb{T}^2_k = \emptyset$.

Proof. Item 1, left-hand inclusion: The case $l = 1$ is trivial. To see the induction step $l - 1 \mapsto l$, let $B \in \mathbb{T}_l$ be given. According to the definition of $\mathbb{T}_l$ (cf. D.3.13), there exists a box $A \in \mathbb{T}_{l-1} \backslash \mathbb{B}_{\text{stop}}$ such that $B \in \operatorname{sons}(A)$. The induction hypothesis implies $(A, A) \in \mathbb{T}^2_{l-1}$ and L.3.24 yields $(A, A) \in \mathbb{T}^2_{l-1} \backslash \mathbb{B}^2_{\text{stop}}$. Then, by D.3.25, $(B, B) \in \mathbb{T}^2_l$.

Item 1, right-hand inclusion: The case $l = 1$ is trivial. To see the induction step $l - 1 \mapsto l$, let $(B_1, B_2) \in \mathbb{T}^2_l$. By definition of $\mathbb{T}^2_l$, there exists a pair $(A_1, A_2) \in \mathbb{T}^2_{l-1} \backslash \mathbb{B}^2_{\text{stop}}$ such that $B_1 \in \operatorname{sons}(A_1)$ and $B_2 \in \operatorname{sons}(A_2)$. Using L.3.24 and the induction hypothesis, we have

$$(A_1, A_2) \in \mathbb{T}^2_{l-1} \backslash \mathbb{B}^2_{\text{stop}} \subseteq (\mathbb{T}_{l-1} \times \mathbb{T}_{l-1}) \backslash (\mathbb{B}_{\text{stop}} \times \mathbb{B}_{\text{stop}}) = (\mathbb{T}_{l-1} \backslash \mathbb{B}_{\text{stop}}) \times (\mathbb{T}_{l-1} \backslash \mathbb{B}_{\text{stop}}),$$

so that $A_1, A_2 \in \mathbb{T}_{l-1} \backslash \mathbb{B}_{\text{stop}}$. According to D.3.13, this implies $(B_1, B_2) \in \mathbb{T}_l \times \mathbb{T}_l$.

Ad item 2: Let $l, k \in \mathbb{N}$ with $l \neq k$. Then, using item 1,

$$\mathbb{T}^2_l \cap \mathbb{T}^2_k \subseteq (\mathbb{T}_l \times \mathbb{T}_l) \cap (\mathbb{T}_k \times \mathbb{T}_k) = (\mathbb{T}_l \cap \mathbb{T}_k) \times (\mathbb{T}_l \cap \mathbb{T}_k) \overset{L.3.15}{=} \emptyset \times \emptyset = \emptyset.$$

□

The next lemma establishes the fact that the trees $\mathbb{T}$ and $\mathbb{T}^2$ have the same depth.

Lemma 3.27. *There exists an $l \in \mathbb{N}$ such that $\mathbb{T}_l^2 \subseteq \mathbb{B}_{\text{stop}}^2$. The minimizer*

$$L := \min\{l \in \mathbb{N} \,|\, \mathbb{T}_l^2 \subseteq \mathbb{B}_{\text{stop}}^2\}$$

has the following properties:

1. *There holds $\mathbb{T}_L^2 \neq \emptyset$ and $\mathbb{T}_{L+1}^2 = \mathbb{T}_{L+2}^2 = \cdots = \emptyset$.*
2. *There holds $L = \operatorname{depth}(\mathbb{T})$, where $\operatorname{depth}(\mathbb{T})$ is defined as in D.3.17.*

Proof. Abbreviate $M := \operatorname{depth}(\mathbb{T}) = \min\{l \in \mathbb{N} \,|\, \mathbb{T}_l \subseteq \mathbb{B}_{\text{stop}}\}$. The relation

$$\mathbb{T}_M^2 \overset{L.3.26}{\subseteq} \mathbb{T}_M \times \mathbb{T}_M \subseteq \mathbb{B}_{\text{stop}} \times \mathbb{B}_{\text{stop}} \overset{L.3.24}{\subseteq} \mathbb{B}_{\text{stop}}^2$$

proves that there exists an $l \in \mathbb{N}$ such that $\mathbb{T}_l^2 \subseteq \mathbb{B}_{\text{stop}}^2$. Now, denote by $L \in \mathbb{N}$ the minimal value such that $\mathbb{T}_L^2 \subseteq \mathbb{B}_{\text{stop}}^2$. The proof of item 1 is completely analogous to the one in L.3.16, so let us go straight to item 2: On one hand, the minimality of L immediately yields $L \leq M$. On the other hand, if $L < M$ were true, then $\mathbb{T}_M^2 = \emptyset$ by item 1. However, since $\mathbb{T}_M \neq \emptyset$ by L.3.16, we would end up with the following contradiction:

$$\emptyset \neq \{(B, B) \,|\, B \in \mathbb{T}_M\} \overset{L.3.26}{\subseteq} \mathbb{T}_M^2 = \emptyset.$$

□

As was the case in D.3.17, we are only interested in the non-trivial levels $\mathbb{T}_l^2$. Furthermore, recall from L.3.26 that they are pairwise disjoint, meaning that a box pair (B_1, B_2) cannot occur on more than one level.

Definition 3.28. *Let $(\mathbb{T}_l^2)_{l \in \mathbb{N}}$ be the sequence from D.3.25 and $L \in \mathbb{N}$ be the value from L.3.27. We define the* product box tree

$$\mathbb{T}^2 := \bigcupdot_{l=1}^{L} \mathbb{T}_l^2 \subseteq \mathbb{B}^2$$

and set $\operatorname{depth}(\mathbb{T}^2) := L$. The sets $\mathbb{T}_l^2$ are called levels of $\mathbb{T}^2$. *Furthermore, we say that $\mathbb{T}^2$* is based on the clustering parameters B_{start}, $\operatorname{sons}(\cdot)$, σ_{small} and $\tilde{\sigma}_{\text{adm}}$ *(D.3.8, D.2.10, D.3.10, D.3.22).*

In the remainder of this section, we will see why $\mathbb{T}^2$ is indeed much smaller than $\mathbb{T} \times \mathbb{T}$.

Definition 3.29. *We define the* sparsity constant

$$\sigma_{\text{sparse}}(\mathbb{T}^2) := \max\left\{ \max_{B \in \mathbb{T}} \#\{(B_1, B_2) \in \mathbb{T}^2 \,|\, B_1 = B\}, \quad \max_{B \in \mathbb{T}} \#\{(B_1, B_2) \in \mathbb{T}^2 \,|\, B_2 = B\} \right\}.$$

The next lemma shows that the value of $\sigma_{\text{sparse}}(\mathbb{T}^2)$ is indeed uniformly bounded.

Lemma 3.30. *There holds the uniform bound*

$$\sigma_{\text{sparse}}(\mathbb{T}^2) \leq C(d)(2 + \tilde{\sigma}_{\text{adm}}^{-1})^d.$$

Proof. We focus on finding a bound for the left-hand maximum in D.3.29 and suggest that the right-hand maximum can be treated analogously. First, recalling the inclusion $\mathbb{T}_l^2 \subseteq \mathbb{T}_l \times \mathbb{T}_l$ from L.3.26, we simplify[2]

$$\max_{B\in\mathbb{T}} \#\{(B_1,B_2)\in\mathbb{T}^2 \mid B_1 = B\} = \max_{l\in\mathbb{N}} \max_{B\in\mathbb{T}_l} \#\{(B_1,B_2)\in\mathbb{T}_l^2 \mid B_1 = B\}.$$

Now, for all $l\in\mathbb{N}$ and all $B\in\mathbb{T}_l$, let us abbreviate

$$\mathbb{T}_l^2(B) := \{(B_1,B_2)\in\mathbb{T}_l^2 \mid B_1 = B\} \subseteq \mathbb{T}_l^2.$$

Clearly, if we can find a constant $C\geq 1$ such that $\#(\mathbb{T}_l^2(B)) \leq C$, then also $\sigma_{\text{sparse}}(\mathbb{T}^2)\leq C$.

Let us first have a look at the non-trivial case $l\geq 2$: Given a pair $(B_1,B_2)\in\mathbb{T}_l^2(B)$, we know from D.3.25 that there exists a pair $(A_1,A_2)\in\mathbb{T}_{l-1}^2\backslash\mathbb{B}_{\text{stop}}^2$ such that $B_1\in\operatorname{sons}(A_1)$ and $B_2\in\operatorname{sons}(A_2)$. First, according to L.2.11, we have $B_1\subseteq A_1$ and $B_2\subseteq A_2$. Second, since $(A_1,A_2)\in\mathbb{T}_{l-1}^2\subseteq\mathbb{T}_{l-1}\times\mathbb{T}_{l-1}$, we know from L.3.15 that

$$\operatorname{diam}_2(A_1) = 4\sqrt{d}\sigma_{\text{sprd}}2^{-l}, \qquad \operatorname{diam}_2(A_2) = 4\sqrt{d}\sigma_{\text{sprd}}2^{-l}.$$

And third, since $(A_1,A_2)\in\mathbb{B}^2\backslash\mathbb{B}_{\text{stop}}^2$, we know from D.3.21 that the pair (A_1,A_2) is neither small nor admissible. In particular,

$$\max\{\operatorname{diam}_2(A_1),\operatorname{diam}_2(A_2)\} > \tilde{\sigma}_{\text{adm}}\operatorname{dist}_2(A_1,A_2).$$

Now, fix a point $b\in B$ and let $a_1\in\overline{A_1}$, $a_2\in\overline{A_2}$ be such that $\|a_2-a_1\|_2 = \operatorname{dist}_2(A_1,A_2)$. Then, since $b\in B = B_1\subseteq A_1$,

$$\begin{aligned}
\sup_{b_2\in B_2}\|b_2-b\|_2 &\leq \sup_{b_2\in B_2}\|b_2-a_2\|_2 + \|a_2-a_1\|_2 + \|a_1-b\|_2\\
&\leq \operatorname{diam}_2(A_2) + \operatorname{dist}_2(A_1,A_2) + \operatorname{diam}_2(A_1)\\
&\leq \operatorname{diam}_2(A_2) + \tilde{\sigma}_{\text{adm}}^{-1}\max\{\operatorname{diam}_2(A_1),\operatorname{diam}_2(A_2)\} + \operatorname{diam}_2(A_1)\\
&= 4\sqrt{d}(2+\tilde{\sigma}_{\text{adm}}^{-1})\sigma_{\text{sprd}}2^{-l}.
\end{aligned}$$

Abbreviating $C_0 := 4\sqrt{d}(2+\tilde{\sigma}_{\text{adm}}^{-1})\sigma_{\text{sprd}}$, we just proved that

$$\bigcup_{(B_1,B_2)\in\mathbb{T}_l^2(B)} B_2 \subseteq \overline{\operatorname{Ball}}_2(b, C_0 2^{-l}).$$

Furthermore, this union must be disjoint. To see this, consider two pairs (B_1,B_2), $(\tilde{B}_1,\tilde{B}_2)\in\mathbb{T}_l^2(B)$ with $(B_1,B_2)\neq(\tilde{B}_1,\tilde{B}_2)$. According to L.3.26, we have $B_1,B_2,\tilde{B}_1,\tilde{B}_2\in\mathbb{T}_l$. Now, since $B_1 = B = \tilde{B}_1$, it must be the case that $B_2\neq\tilde{B}_2$. Then, L.3.15 already implies $B_2\cap\tilde{B}_2 = \emptyset$.

Next, we compute

$$\begin{aligned}
C(d)C_0^d 2^{-dl} &\overset{L.2.7}{=} \operatorname{meas}(\overline{\operatorname{Ball}}_2(b,C_0 2^{-l})) \geq \operatorname{meas}\Bigg(\bigcup_{(B_1,B_2)\in\mathbb{T}_l^2(B)} B_2\Bigg)\\
&= \sum_{(B_1,B_2)\in\mathbb{T}_l^2(B)} \operatorname{meas}(B_2) \overset{L.3.15}{=} (2\sigma_{\text{sprd}})^d 2^{-dl}\#(\mathbb{T}_l^2(B)).
\end{aligned}$$

[2] In other words, given a box $B\in\mathbb{T}$, it suffices to look for "partners" B_2 on the same tree level.

Solving for $\#(\mathbb{T}_l^2(B))$, we obtain the desired bound:

$$\#(\mathbb{T}_l^2(B)) \leq \frac{C(d)C_0^d 2^{-dl}}{(2\sigma_{\text{sprd}})^d 2^{-dl}} \leq C(d)(2+\tilde{\sigma}_{\text{adm}}^{-1})^d.$$

This concludes the case $l \geq 2$. Finally, in the case $l = 1$, we have $\mathbb{T}_1 = \{B_{\text{start}}\}$, so that

$$\#(\mathbb{T}_1^2(B_{\text{start}})) = \#\{(B_{\text{start}}, B_{\text{start}})\} = 1 \leq C(d)(2+\tilde{\sigma}_{\text{adm}}^{-1})^d.$$

□

The sparsity constant allows us to bound sums over $\mathbb{T}^2$ by sums over $\mathbb{T}$.

Lemma 3.31. *Let $f_1, f_2 : \mathbb{B} \longrightarrow [0,\infty)$ be given functions. Then, there holds the bound*

$$\sum_{(B_1,B_2)\in\mathbb{T}^2} f_1(B_1) + f_2(B_2) \leq \sigma_{\text{sparse}}(\mathbb{T}^2) \sum_{B\in\mathbb{T}} f_1(B) + f_2(B).$$

Proof. Using the definition of the sparsity constant $\sigma_{\text{sparse}}(\mathbb{T}^2)$ from D.3.29, we compute

$$\begin{aligned}\sum_{(B_1,B_2)\in\mathbb{T}^2} f_1(B_1) + f_2(B_2) &\leq \sum_{B\in\mathbb{T}} \left(\sum_{\substack{(B_1,B_2)\in\mathbb{T}^2:\\ B_1=B}} f_1(B_1) + \sum_{\substack{(B_1,B_2)\in\mathbb{T}^2:\\ B_2=B}} f_2(B_2) \right) \\ &= \sum_{B\in\mathbb{T}} \left(\#\{(B_1,B_2)\in\mathbb{T}^2 \,|\, B_1 = B\} f_1(B) + \#\{(B_1,B_2)\in\mathbb{T}^2 \,|\, B_2 = B\} f_2(B) \right) \\ &\leq \sigma_{\text{sparse}}(\mathbb{T}^2) \sum_{B\in\mathbb{T}} f_1(B) + f_2(B).\end{aligned}$$

□

As an immediate consequence, we get the following result:

Corollary 3.32. *There holds the bound*

$$\#(\mathbb{T}^2) \leq \sigma_{\text{sparse}}(\mathbb{T}^2)\#\mathbb{T}.$$

Proof. We apply L.3.31 to the functions $f_1(B) := f_2(B) := 1/2$. Then,

$$\#(\mathbb{T}^2) = \sum_{(B_1,B_2)\in\mathbb{T}^2} 1 \leq \sigma_{\text{sparse}}(\mathbb{T}^2) \sum_{B\in\mathbb{T}} 1 = \sigma_{\text{sparse}}(\mathbb{T}^2)\#\mathbb{T}.$$

□

Remark 3.33. *The time needed to compute the product box tree $\mathbb{T}^2$ is proportional to its memory footprint. The recursion in D.3.25 requires us to go through all pairs $(A_1, A_2) \in \mathbb{T}_{l-1}^2$ and check whether $(A_1, A_2) \in \mathbb{B}_{\text{stop}}^2$. According to D.3.22, we need two checks for smallness in the sense of D.3.11 and one check for admissibility. The information about smallness is already available in the box tree $\mathbb{T}$ (i.e., a trivial lookup) and admissibility of boxes can easily be checked in $\mathcal{O}(1)$ time.*

3.5 The leaves $\mathbb{T}^2_{\text{stop}}$

Recall from R.3.2 that the goal of this chapter is to construct a family $\mathbb{P}^2$ of cluster pairs (I, J) such that $\dot{\bigcup}_{(I,J)\in\mathbb{P}^2} I \times J = \{1, \dots, N\} \times \{1, \dots, N\}$. The product box tree $\mathbb{T}^2$ contains pairs of boxes that lie on top of each other, meaning that pairs of characteristic points (x_i, x_j) will end up on more than one level of $\mathbb{T}^2$. To get a proper partition of $\{1, \dots, N\} \times \{1, \dots, N\}$, we may only take the box pairs from the *leaves* of $\mathbb{T}^2$.

Definition 3.34. *We define the* leaves *of* $\mathbb{T}^2$ *by*

$$\mathbb{T}^2_{\text{stop}} := \mathbb{T}^2 \cap \mathbb{B}^2_{\text{stop}}.$$

Lemma 3.35. *In the sense of a disjoint union, there holds*

$$\dot{\bigcup_{(B_1,B_2)\in\mathbb{T}^2_{\text{stop}}}} B_1 \times B_2 = B_{\text{start}} \times B_{\text{start}}.$$

Proof. First, we prove pairwise disjointness: Consider two pairs $(B_1, B_2), (\tilde{B}_1, \tilde{B}_2) \in \mathbb{T}^2_{\text{stop}}$ with $(B_1, B_2) \neq (\tilde{B}_1, \tilde{B}_2)$ and let $l, \tilde{l} \in \mathbb{N}$ be such that $(B_1, B_2) \in \mathbb{T}^2_l$ and $(\tilde{B}_1, \tilde{B}_2) \in \mathbb{T}^2_{\tilde{l}}$. Let us check the case $l < \tilde{l}$ first: Backtracking the predecessors of $(\tilde{B}_1, \tilde{B}_2)$, we can find a pair $(\tilde{A}_1, \tilde{A}_2) \in \mathbb{T}^2_l \backslash \mathbb{B}^2_{\text{stop}}$ such that $\tilde{B}_1 \subseteq \tilde{A}_1$ and $\tilde{B}_2 \subseteq \tilde{A}_2$ (cf. L. 2.11). Note that $B_1, B_2, \tilde{A}_1, \tilde{A}_2 \in \mathbb{T}_l$ by L.3.26. There must hold $(B_1, B_2) \neq (\tilde{A}_1, \tilde{A}_2)$, because $(B_1, B_2) \in \mathbb{B}^2_{\text{stop}}$, whereas $(\tilde{A}_1, \tilde{A}_2) \notin \mathbb{B}^2_{\text{stop}}$. If $B_1 \neq \tilde{A}_1$, then already $B_1 \cap \tilde{A}_1 = \emptyset$, according to L.3.15. Similarly, if $B_2 \neq \tilde{A}_2$, then $B_2 \cap \tilde{A}_2 = \emptyset$. Either way, it follows that

$$(B_1 \times B_2) \cap (\tilde{B}_1 \times \tilde{B}_2) \subseteq (B_1 \times B_2) \cap (\tilde{A}_1 \times \tilde{A}_2) = (B_1 \cap \tilde{A}_1) \times (B_2 \cap \tilde{A}_2) = \emptyset.$$

This concludes the proof of pairwise disjointness in the case $l < \tilde{l}$. Due to symmetry, the case $l > \tilde{l}$ is completely analogous. Finally, in the case $l = \tilde{l}$, the pair $(\tilde{B}_1, \tilde{B}_2)$ itself plays the role of $(\tilde{A}_1, \tilde{A}_2)$.

It remains to show that the sets $\{B_1 \times B_2 \,|\, (B_1, B_2) \in \mathbb{T}^2_{\text{stop}}\}$ make up all of $B_{\text{start}} \times B_{\text{start}}$. To this end, we introduce subsets $M_l, M_{l,\text{stop}} \subseteq \mathbb{R}^d \times \mathbb{R}^d$, $l \in \{1, \dots, L\}$, $L := \text{depth}(\mathbb{T}^2)$, in the following way:

$$M_l := \bigcup_{(A_1,A_2)\in\mathbb{T}^2_l} A_1 \times A_2, \qquad M_{l,\text{stop}} := \bigcup_{(A_1,A_2)\in\mathbb{T}^2_l\cap\mathbb{B}^2_{\text{stop}}} A_1 \times A_2.$$

For all $l \in \mathbb{N}$, $l \geq 2$, we have the following recursion:

$$\begin{aligned}
M_{l-1} &= \bigcup_{(A_1,A_2)\in\mathbb{T}^2_{l-1}} A_1 \times A_2 \\
&\overset{L.2.11}{=} \bigcup_{\substack{(A_1,A_2)\in\\ \mathbb{T}^2_{l-1}\cap\mathbb{B}^2_{\text{stop}}}} A_1 \times A_2 \quad \cup \bigcup_{\substack{(A_1,A_2)\in\\ \mathbb{T}^2_{l-1}\setminus\mathbb{B}^2_{\text{stop}}}} \left(\bigcup_{B_1\in\text{sons}(A_1)} B_1 \right) \times \left(\bigcup_{B_2\in\text{sons}(A_2)} B_2 \right) \\
&= \bigcup_{\substack{(A_1,A_2)\in\\ \mathbb{T}^2_{l-1}\cap\mathbb{B}^2_{\text{stop}}}} A_1 \times A_2 \quad \cup \bigcup_{\substack{(A_1,A_2)\in\\ \mathbb{T}^2_{l-1}\setminus\mathbb{B}^2_{\text{stop}}}} \bigcup_{\substack{B_1\in\text{sons}(A_1),\\ B_2\in\text{sons}(A_2)}} B_1 \times B_2 \\
&\overset{D.3.25}{=} \bigcup_{\substack{(A_1,A_2)\in\\ \mathbb{T}^2_{l-1}\cap\mathbb{B}^2_{\text{stop}}}} A_1 \times A_2 \quad \cup \bigcup_{(B_1,B_2)\in\mathbb{T}^2_l} B_1 \times B_2 \\
&= M_{l-1,\text{stop}} \cup M_l.
\end{aligned}$$

Since $\mathbb{T}^2_1 = \{(B_{\text{start}}, B_{\text{start}})\}$, the first element of the recursion reads $M_1 = B_{\text{start}} \times B_{\text{start}}$. On the other hand, the last element satisfies $M_{L,\text{stop}} = M_L$, because in L.3.27 we defined L such that $\mathbb{T}^2_L \subseteq \mathbb{B}^2_{\text{stop}}$. Then,

$$B_{\text{start}} \times B_{\text{start}} = M_1 = M_{1,\text{stop}} \cup M_2 = \cdots = \bigcup_{l=1}^{L} M_{l,\text{stop}} = \bigcup_{(B_1,B_2)\in\mathbb{T}^2_{\text{stop}}} B_1 \times B_2,$$

which concludes the proof.

□

3.6 The block partition $\mathbb{P}^2$

We are finally in the position to define the block partition $\mathbb{P}^2$. Recall from D.3.7 that $\iota(B) = \{n \in \{1,\ldots,N\} \,|\, x_n \in B\}$ is the index cluster that belongs to a given physical set $B \subseteq \mathbb{R}^d$. We mention that some of the box pairs $(B_1, B_2) \in \mathbb{T}^2_{\text{stop}}$ might be "empty" in the sense that they contain none of the characteristic points $x_1, \ldots, x_N$.

Definition 3.36. *Denote by $\mathbb{T}^2_{\text{stop}}$ the leaves of the product box tree as defined in D.3.34. We define the* block partition

$$\mathbb{P}^2 := \{(\iota(B_1), \iota(B_2)) \,|\, (B_1, B_2) \in \mathbb{T}^2_{\text{stop}} \text{ with } \iota(B_1) \neq \emptyset,\ \iota(B_2) \neq \emptyset\}.$$

Lemma 3.37. *In the sense of a disjoint union, there holds*

$$\dot{\bigcup_{(I_1,I_2)\in\mathbb{P}^2}} I_1 \times I_2 = \{1,\ldots,N\} \times \{1,\ldots,N\}.$$

Proof. We start with pairwise disjointness: Let $(I_1, I_2), (\tilde{I}_1, \tilde{I}_2) \in \mathbb{P}^2$ with $(I_1, I_2) \neq (\tilde{I}_1, \tilde{I}_2)$. By definition of $\mathbb{P}^2$, there exist $(B_1, B_2), (\tilde{B}_1, \tilde{B}_2) \in \mathbb{T}^2_{\text{stop}}$ such that $(I_1, I_2) = (\iota(B_1), \iota(B_2))$

and $(\tilde{I}_1, \tilde{I}_2) = (\iota(\tilde{B}_1), \iota(\tilde{B}_2))$. Note that there must hold $(B_1, B_2) \neq (\tilde{B}_1, \tilde{B}_2)$, because otherwise (I_1, I_2) and $(\tilde{I}_1, \tilde{I}_2)$ would coincide. It then follows from L.3.35 that $(B_1 \times B_2) \cap (\tilde{B}_1 \times \tilde{B}_2) = \emptyset$. Therefore,

$$\begin{aligned}(I_1 \times I_2) \cap (\tilde{I}_1 \times \tilde{I}_2) &= \{(m,n) \,|\, x_m \in B_1, x_n \in B_2\} \cap \{(m,n) \,|\, x_m \in \tilde{B}_1, x_n \in \tilde{B}_2\} \\ &= \{(m,n) \,|\, (x_m, x_n) \in (B_1 \times B_2) \cap (\tilde{B}_1 \times \tilde{B}_2)\} = \emptyset.\end{aligned}$$

Finally, we have

$$\bigcup_{\substack{(I_1,I_2)\\ \in \mathbb{P}^2}} I_1 \times I_2 = \bigcup_{\substack{(B_1,B_2)\\ \in \mathbb{T}^2_{\mathrm{stop}}}} \{(m,n) \,|\, x_m \in B_1, x_n \in B_2\} = \left\{(m,n) \,\middle|\, (x_m, x_n) \in \bigcup_{\substack{(B_1,B_2)\\ \in \mathbb{T}^2_{\mathrm{stop}}}} B_1 \times B_2\right\}$$
$$\overset{\text{L.3.35}}{=} \{(m,n) \,|\, (x_m, x_n) \in B_{\mathrm{start}} \times B_{\mathrm{start}}\} \overset{\text{L.3.9}}{=} \{1, \dots, N\} \times \{1, \dots, N\}.$$

This finishes the proof.

□

The previous lemma tells us that any given matrix $\boldsymbol{B} \in \mathbb{R}^{N \times N}$ can be represented by a family of matrix blocks, i.e.,

$$\boldsymbol{B} \quad \leftrightarrow \quad \{\boldsymbol{B}|_{I_1 \times I_2} \,|\, (I_1, I_2) \in \mathbb{P}^2\}.$$

The clustering algorithm in D. 3.13 generates a hierarchy of boxes $B \in \mathbb{B}$ based on the positions of the characteristic points $x_1, \dots, x_N \in \mathbb{R}^d$ from D.3.6 alone. But there is no guarantee that the corresponding characteristic sets $\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$ from D.3.5 are fully contained in these boxes. We can inflate the boxes slightly (cf., D. 2.12) to rectify this inconvenience, but then the inflated box pairs do not satisfy the original admissibility condition from D.3.22 any more. However, by tuning the clustering parameter $\tilde{\sigma}_{\mathrm{adm}} > 0$ appropriately, we can regain admissibility with respect to a different admissibility parameter $\sigma_{\mathrm{adm}} > 0$ (the one we actually care about).

Lemma 3.38. *Denote by $\sigma_{\mathrm{shp}}, \sigma_{\mathrm{sprd}} \geq 1$ and $h_{\min} > 0$ the quantities from D. 3.5 and D.3.6. Let $\sigma_{\mathrm{adm}} > 0$ be a given number (yet another clustering parameter). Suppose that the product box tree $\mathbb{T}^2$ from D.3.28 is based on the clustering parameters B_{start}, $\mathrm{sons}(\cdot)$, σ_{small} and $\tilde{\sigma}_{\mathrm{adm}}$ (D.3.8, D.2.10, D.3.10, D.3.22), where $\tilde{\sigma}_{\mathrm{adm}}$ is chosen as*

$$\tilde{\sigma}_{\mathrm{adm}} := \frac{\sigma_{\mathrm{adm}}}{(1 + 4\sqrt{d}\sigma_{\mathrm{shp}})(1 + \sigma_{\mathrm{adm}})} > 0.$$

Then, for every pair $(I_1, I_2) \in \mathbb{P}^2$, there holds at least one of the following statements:

1. *There holds*
$$\min\{\#I_1, \#I_2\} \leq \sigma_{\mathrm{small}}.$$

2. There exist boxes $A_1, A_2 \in \mathbb{B}$ such that[3]

$$\begin{aligned} \Omega_{I_1} &\subseteq A_1, \\ \Omega_{I_2} &\subseteq A_2, \\ h_{\min} &\leq \min\{\operatorname{diam}_2(A_1), \operatorname{diam}_2(A_2)\}, \\ \max\{\operatorname{diam}_2(A_1), \operatorname{diam}_2(A_2)\} &\leq \sigma_{\mathrm{adm}} \operatorname{dist}_2(A_1, A_2), \\ \operatorname{dist}_2(A_1, A_2) &\leq \sqrt{d}\sigma_{\mathrm{sprd}}. \end{aligned}$$

Proof. Let $(I_1, I_2) \in \mathbb{P}^2$ be given. According to D.3.36, there exists a pair $(B_1, B_2) \in \mathbb{T}^2_{\mathrm{stop}} = \mathbb{T}^2 \cap \mathbb{B}^2_{\mathrm{stop}}$ (cf. D.3.34) such that $(I_1, I_2) = (\iota(B_1), \iota(B_2))$. In particular, by definition of $\mathbb{B}^2_{\mathrm{stop}}$ (cf. D.3.22), at least one of the following conditions is satis´fied:

$$\begin{aligned} \min\{\#I_1, \#I_2\} &\leq \sigma_{\mathrm{small}}, \\ \max\{\operatorname{diam}_2(B_1), \operatorname{diam}_2(B_2)\} &\leq \tilde{\sigma}_{\mathrm{adm}} \operatorname{dist}_2(B_1, B_2). \end{aligned}$$

If the first condition is satisfied, then we already have what we want. It remains to check the case where the first condition is violated and the second one is valid. Using the quantities $h_{I_1} = \max_{n \in I_1} h_n$ and h_{I_2} from D.3.6, we define the inflated boxes (cf. D.2.12)

$$A_1 := {B_1}^{h_{I_1}} \in \mathbb{B}, \qquad A_2 := {B_2}^{h_{I_2}} \in \mathbb{B}.$$

We will show that the boxes A_1 and A_2 have all of the desired properties. To this end, let $y \in \Omega_{I_1}$. Then there exists an index $n \in I_1$ such that $y \in \Omega_n$. Note that the characteristic point $x_n \in \Omega_n$ (cf. D.3.6) satisfies $x_n \in B_1$ by definition of $I_1 = \iota(B_1)$ (cf. D.3.7). Then, since $\|y - x_n\|_2 \leq \operatorname{diam}_2(\Omega_n) = h_n \leq h_{I_1}$, L.2.13 tells us that $y \in {B_1}^{h_{I_1}} = A_1$. Using a similar argument for I_2, we obtain the inclusions

$$\Omega_{I_1} \subseteq A_1, \qquad \Omega_{I_2} \subseteq A_2.$$

Next, since we are currently in the case $\min\{\#I_1, \#I_2\} > \sigma_{\mathrm{small}}$, we know from D.3.11 that $B_1, B_2 \in \mathbb{B} \backslash \mathbb{B}_{\mathrm{stop}}$. According to L.3.12, it follows that

$$h_{I_1} = h_{\iota(B_1)} \leq 2\sigma_{\mathrm{shp}} \operatorname{diam}_2(B_1), \qquad h_{I_2} \leq 2\sigma_{\mathrm{shp}} \operatorname{diam}_2(B_2).$$

Then, abbreviating $\gamma := 1 + 4\sqrt{d}\sigma_{\mathrm{shp}}$, we compute

$$\begin{aligned} &\max\{\operatorname{diam}_2(B_1), \operatorname{diam}_2(B_2)\} \\ &\quad\leq \tilde{\sigma}_{\mathrm{adm}} \operatorname{dist}_2(B_1, B_2) \\ &\quad\overset{L.2.13}{\leq} \tilde{\sigma}_{\mathrm{adm}} \operatorname{dist}_2(A_1, A_2) + \sqrt{d}\tilde{\sigma}_{\mathrm{adm}}(h_{I_1} + h_{I_2}) \\ &\quad\leq \tilde{\sigma}_{\mathrm{adm}} \operatorname{dist}_2(A_1, A_2) + 2\sqrt{d}\sigma_{\mathrm{shp}}\tilde{\sigma}_{\mathrm{adm}}(\operatorname{diam}_2(B_1) + \operatorname{diam}_2(B_2)) \\ &\quad\leq \tilde{\sigma}_{\mathrm{adm}} \operatorname{dist}_2(A_1, A_2) + \gamma\tilde{\sigma}_{\mathrm{adm}} \max\{\operatorname{diam}_2(B_1), \operatorname{diam}_2(B_2)\}. \end{aligned}$$

[3] Lines 3 and 5 are merely a byproduct and we mention them only for easier reference later on. The reader should focus his attention on lines 1, 2 and 4.

Since $\gamma\tilde{\sigma}_{\mathrm{adm}} = \sigma_{\mathrm{adm}}/(1+\sigma_{\mathrm{adm}}) < 1$, we can absorb the last term in the left-hand side of the overall inequality. It follows that

$$\begin{aligned}
\max\{\mathrm{diam}_2(A_1), \mathrm{diam}_2(A_2)\} &\overset{L.2.13}{\leq} \max\{\mathrm{diam}_2(B_1) + 2\sqrt{d}h_{I_1}, \mathrm{diam}_2(B_2) + 2\sqrt{d}h_{I_2}\} \\
&\leq \gamma \max\{\mathrm{diam}_2(B_1), \mathrm{diam}_2(B_2)\} \\
&\leq \frac{\gamma\tilde{\sigma}_{\mathrm{adm}}}{1-\gamma\tilde{\sigma}_{\mathrm{adm}}}\mathrm{dist}_2(A_1, A_2) \\
&\overset{\mathrm{Def.}\tilde{\sigma}_{\mathrm{adm}}}{=} \sigma_{\mathrm{adm}}\mathrm{dist}_2(A_1, A_2).
\end{aligned}$$

The lower bound for the diameters of A_1 and A_2 can be seen as follows:

$$\begin{aligned}
h_{\min} &\overset{D.3.6}{\leq} \min\{h_{I_1}, h_{I_2}\} \\
&\overset{L.2.13}{\leq} \min\{\mathrm{diam}_2(B_1{}^{h_{I_1}}), \mathrm{diam}_2(B_1{}^{h_{I_1}})\} = \min\{\mathrm{diam}_2(A_1), \mathrm{diam}_2(A_2)\}.
\end{aligned}$$

It remains to prove the upper bound for the distance between A_1 and A_2. From D.3.34, L.3.26 and L.3.15 we know that $B_1, B_2 \subseteq B_{\mathrm{start}}$. Therefore,

$$\mathrm{dist}_2(A_1, A_2) = \mathrm{dist}_2(B_1{}^{h_{I_1}}, B_2{}^{h_{I_2}}) \overset{L.2.13}{\leq} \mathrm{dist}_2(B_1, B_2) \leq \mathrm{diam}_2(B_{\mathrm{start}}) \overset{L.3.9}{=} \sqrt{d}\sigma_{\mathrm{sprd}}.$$

This finishes the proof.

□

L.3.38 tells us that the cluster pairs $(I_1, I_2) \in \mathbb{P}^2$ can be categorized into two different groups.

Definition 3.39. *We define*

$$\mathbb{P}^2_{\mathrm{small}} := \{(I_1, I_2) \in \mathbb{P}^2 \mid \min\{\#I_1, \#I_2\} \leq \sigma_{\mathrm{small}}\}, \qquad \mathbb{P}^2_{\mathrm{adm}} := \mathbb{P}^2 \backslash \mathbb{P}^2_{\mathrm{small}}.$$

The next lemma will come in handy when we estimate the memory requirements for an arbitrary $\mathcal{H}$-matrix (cf. L.3.44).

Lemma 3.40. *There holds the bound*

$$\sum_{(I_1,I_2)\in\mathbb{P}^2} \#I_1 + \#I_2 \leq 2\sigma_{\mathrm{sparse}}(\mathbb{T}^2)\mathrm{depth}(\mathbb{T})N.$$

Proof. We compute

$$\begin{aligned}
\sum_{(I_1,I_2)\in\mathbb{P}^2} \#I_1 + \#I_2 &\overset{D.3.36}{\leq} \sum_{(B_1,B_2)\in\mathbb{T}^2_{\mathrm{stop}}} \#\iota(B_1) + \#\iota(B_2) \overset{D.3.34}{\leq} \sum_{(B_1,B_2)\in\mathbb{T}^2} \#\iota(B_1) + \#\iota(B_2) \\
&\overset{L.3.31}{\leq} 2\sigma_{\mathrm{sparse}}(\mathbb{T}^2) \sum_{B\in\mathbb{T}} \#\iota(B) \overset{L.3.18}{\leq} 2\sigma_{\mathrm{sparse}}(\mathbb{T}^2)\mathrm{depth}(\mathbb{T})\#\iota(B_{\mathrm{start}}) \\
&\overset{L.3.9}{=} 2\sigma_{\mathrm{sparse}}(\mathbb{T}^2)\mathrm{depth}(\mathbb{T})N.
\end{aligned}$$

□

Finally, we need to relate the operator norm of a full matrix $\boldsymbol{B} \in \mathbb{R}^{N\times N}$ to the norms of its subblocks $\boldsymbol{B}|_{I_1\times I_2}$ as determined by $\mathbb{P}^2$.

Lemma 3.41. *For all $\boldsymbol{B} \in \mathbb{R}^{N\times N}$, there holds the bound*

$$\|\boldsymbol{B}\|_2 \leq \sigma_{\mathrm{sparse}}(\mathbb{T}^2)\mathrm{depth}(\mathbb{T}) \max_{(I_1,I_2)\in\mathbb{P}^2} \|\boldsymbol{B}|_{I_1\times I_2}\|_2.$$

Proof. Let $\boldsymbol{x} \in \mathbb{R}^N$ and set $\boldsymbol{y} := \boldsymbol{B}\boldsymbol{x} \in \mathbb{R}^N$. We compute

$$\begin{aligned}
\|\boldsymbol{B}\boldsymbol{x}\|_2^2 &= \langle \boldsymbol{y}, \boldsymbol{B}\boldsymbol{x}\rangle_2 = \sum_{i,j=1}^{N} \boldsymbol{y}_i \boldsymbol{B}_{ij} \boldsymbol{x}_j \overset{L.3.37}{=} \sum_{(I_1,I_2)\in\mathbb{P}^2} \sum_{(i,j)\in I_1\times I_2} \boldsymbol{y}_i \boldsymbol{B}_{ij} \boldsymbol{x}_j \\
&= \sum_{(I_1,I_2)\in\mathbb{P}^2} \langle \boldsymbol{y}|_{I_1}, \boldsymbol{B}|_{I_1\times I_2}\boldsymbol{x}|_{I_2}\rangle_{l^2(I_1)} \leq \sum_{(I_1,I_2)\in\mathbb{P}^2} \|\boldsymbol{y}\|_{l^2(I_1)} \|\boldsymbol{B}|_{I_1\times I_2}\|_2 \|\boldsymbol{x}\|_{l^2(I_2)} \\
&\leq \left(\max_{(I_1,I_2)\in\mathbb{P}^2} \|\boldsymbol{B}|_{I_1\times I_2}\|_2 \right) \sum_{(I_1,I_2)\in\mathbb{P}^2} \|\boldsymbol{y}\|_{l^2(I_1)} \|\boldsymbol{x}\|_{l^2(I_2)} \\
&\leq \left(\max_{(I_1,I_2)\in\mathbb{P}^2} \|\boldsymbol{B}|_{I_1\times I_2}\|_2 \right) \left(\sum_{(I_1,I_2)\in\mathbb{P}^2} \|\boldsymbol{y}\|_{l^2(I_1)}^2 \right)^{1/2} \left(\sum_{(I_1,I_2)\in\mathbb{P}^2} \|\boldsymbol{x}\|_{l^2(I_2)}^2 \right)^{1/2}.
\end{aligned}$$

The term in the middle can be treated as follows:

$$\begin{aligned}
\sum_{(I_1,I_2)\in\mathbb{P}^2} \|\boldsymbol{y}\|_{l^2(I_1)}^2 &\overset{D.3.36}{\leq} \sum_{(B_1,B_2)\in\mathbb{T}^2_{\mathrm{stop}}} \|\boldsymbol{y}\|_{l^2(\iota(B_1))}^2 \overset{D.3.34}{\leq} \sum_{(B_1,B_2)\in\mathbb{T}^2} \|\boldsymbol{y}\|_{l^2(\iota(B_1))}^2 \\
&\overset{L.3.31}{\leq} \sigma_{\mathrm{sparse}}(\mathbb{T}^2) \sum_{B\in\mathbb{T}} \|\boldsymbol{y}\|_{l^2(\iota(B))}^2 \overset{L.3.18}{\leq} \sigma_{\mathrm{sparse}}(\mathbb{T}^2)\mathrm{depth}(\mathbb{T}) \|\boldsymbol{y}\|_{l^2(\iota(B_{\mathrm{start}}))}^2 \\
&\overset{L.3.9}{=} \sigma_{\mathrm{sparse}}(\mathbb{T}^2)\mathrm{depth}(\mathbb{T}) \|\boldsymbol{y}\|_2^2.
\end{aligned}$$

With an analogous bound for $\sum_{(I_1,I_2)\in\mathbb{P}^2} \|\boldsymbol{x}\|_{l^2(I_2)}^2$, we get

$$\|\boldsymbol{B}\boldsymbol{x}\|_2^2 \leq \sigma_{\mathrm{sparse}}(\mathbb{T}^2)\mathrm{depth}(\mathbb{T}) \left(\max_{(I_1,I_2)\in\mathbb{P}^2} \|\boldsymbol{B}|_{I_1\times I_2}\|_2 \right) \|\boldsymbol{y}\|_2 \|\boldsymbol{x}\|_2.$$

Finally, dividing by $\|\boldsymbol{y}\|_2 = \|\boldsymbol{B}\boldsymbol{x}\|_2$ and taking the supremum over all $\boldsymbol{x} \in \mathbb{R}^N$, the alleged inequality follows.

□

This finishes our construction of the block partition $\mathbb{P}^2$. We collect all of our findings in a corollary.

Corollary 3.42. *Let $\sigma_{\mathrm{shp}}, \sigma_{\mathrm{ovlp}}, \sigma_{\mathrm{sprd}} \geq 1$ and $N \in \mathbb{N}$. Consider a family of subsets*

$$\Omega_1, \ldots, \Omega_N \subseteq \mathbb{R}^d$$

with shape regularity σ_{shp}, overlap σ_{ovlp} and spread σ_{sprd} (cf. D.2.16, D.2.18, D.2.21). Denote by $h_{\min} > 0$ the minimal element diameter as defined in D.3.6. Finally, let $\sigma_{\mathrm{small}} \geq 1$ be a number with $\sigma_{\mathrm{ovlp}} \leq \sigma_{\mathrm{small}}$ and let $\sigma_{\mathrm{adm}} > 0$. Then, there exist sets $\mathbb{P}^2$, $\mathbb{P}^2_{\mathrm{small}}$, $\mathbb{P}^2_{\mathrm{adm}}$ with the following properties:

1. *There holds* $\mathbb{P}^2 = \mathbb{P}^2_{\mathrm{small}} \cup \mathbb{P}^2_{\mathrm{adm}}$.
2. *The elements of* $\mathbb{P}^2$ *are tuples* (I, J) *of clusters* $I, J \subseteq \{1, \dots, N\}$.
3. *There holds*
$$\bigcup_{(I,J)\in\mathbb{P}^2} I \times J = \{1, \dots, N\} \times \{1, \dots, N\}.$$
4. *For every* $(I, J) \in \mathbb{P}^2_{\mathrm{small}}$, *there holds*
$$\min\{\#I, \#J\} \leq \sigma_{\mathrm{small}}.$$
5. *For every* $(I, J) \in \mathbb{P}^2_{\mathrm{adm}}$, *there exist axes-parallel boxes* $B, D \in \mathbb{B}$ *with the following properties*[4]*:*
$$\begin{aligned} \Omega_I &\subseteq B, \\ \Omega_J &\subseteq D, \\ h_{\min} &\leq \min\{\mathrm{diam}_2(B), \mathrm{diam}_2(D)\}, \\ \max\{\mathrm{diam}_2(B), \mathrm{diam}_2(D)\} &\leq \sigma_{\mathrm{adm}} \mathrm{dist}_2(B, D), \\ \mathrm{dist}_2(B, D) &\leq \sqrt{d}\sigma_{\mathrm{sprd}}. \end{aligned}$$
6. *There holds the bound*
$$\sum_{(I,J)\in\mathbb{P}^2} \#I + \#J \leq C(d, \sigma_{\mathrm{shp}}, \sigma_{\mathrm{sprd}}, \sigma_{\mathrm{adm}}) \ln(h_{\min}^{-1}) N.$$
7. *For all* $\boldsymbol{B} \in \mathbb{R}^{N\times N}$, *there holds the bound*
$$\|\boldsymbol{B}\|_2 \leq C(d, \sigma_{\mathrm{shp}}, \sigma_{\mathrm{sprd}}, \sigma_{\mathrm{adm}}) \ln(h_{\min}^{-1}) \max_{(I,J)\in\mathbb{P}^2} \|\boldsymbol{B}|_{I\times J}\|_2.$$

Proof. We obviously pick the systems from D.3.36 and D.3.39. Items 1, 2, 4 are trivial, item 3 was proved in L.3.37 and item 5 follows from the dichotomy in L.3.38. Finally, to see items 6 and 7, recall that

$$\tilde{\sigma}_{\mathrm{adm}} \overset{L.3.38}{=} \frac{\sigma_{\mathrm{adm}}}{(1 + 4\sqrt{d}\sigma_{\mathrm{shp}})(1 + \sigma_{\mathrm{adm}})} = C(d, \sigma_{\mathrm{shp}}, \sigma_{\mathrm{adm}}),$$

so that

$$\sigma_{\mathrm{sparse}}(\mathbb{T}^2) \overset{L.3.30}{\leq} C(d)(2 + \tilde{\sigma}_{\mathrm{adm}}^{-1})^d \leq C(d, \sigma_{\mathrm{shp}}, \sigma_{\mathrm{adm}}).$$

Furthermore,

$$\mathrm{depth}(\mathbb{T}) \overset{L.3.16}{\leq} \log_2(8\sqrt{d}\sigma_{\mathrm{shp}}\sigma_{\mathrm{sprd}}h_{\min}^{-1}) \leq C(d, \sigma_{\mathrm{shp}}, \sigma_{\mathrm{sprd}}) \ln(h_{\min}^{-1}).$$

[4] Recall from D.3.6 that $\Omega_I := \bigcup_{n\in I} \Omega_n \subseteq \mathbb{R}^d$, for every cluster $I \subseteq \{1, \dots, N\}$.

We conclude that

$$\sum_{(I,J)\in\mathbb{P}^2} \#I + \#J \overset{L.3.40}{\leq} 2\sigma_{\mathrm{sparse}}(\mathbb{T}^2)\mathrm{depth}(\mathbb{T})N \leq C(d,\sigma_{\mathrm{shp}},\sigma_{\mathrm{sprd}},\sigma_{\mathrm{adm}})\ln(h_{\min}^{-1})N$$

and that

$$\begin{aligned}\forall \boldsymbol{B}\in\mathbb{R}^{N\times N}: \qquad \|\boldsymbol{B}\|_2 &\overset{L.3.41}{\leq} \sigma_{\mathrm{sparse}}(\mathbb{T}^2)\mathrm{depth}(\mathbb{T}) \max_{(I,J)\in\mathbb{P}^2} \|\boldsymbol{B}|_{I\times J}\|_2 \\ &\leq C(d,\sigma_{\mathrm{shp}},\sigma_{\mathrm{sprd}},\sigma_{\mathrm{adm}})\ln(h_{\min}^{-1}) \max_{(I,J)\in\mathbb{P}^2} \|\boldsymbol{B}|_{I\times J}\|_2.\end{aligned}$$

This finishes the proof. □

3.7 The class of hierarchical matrices $\mathcal{H}(\mathbb{P}^2, r)$

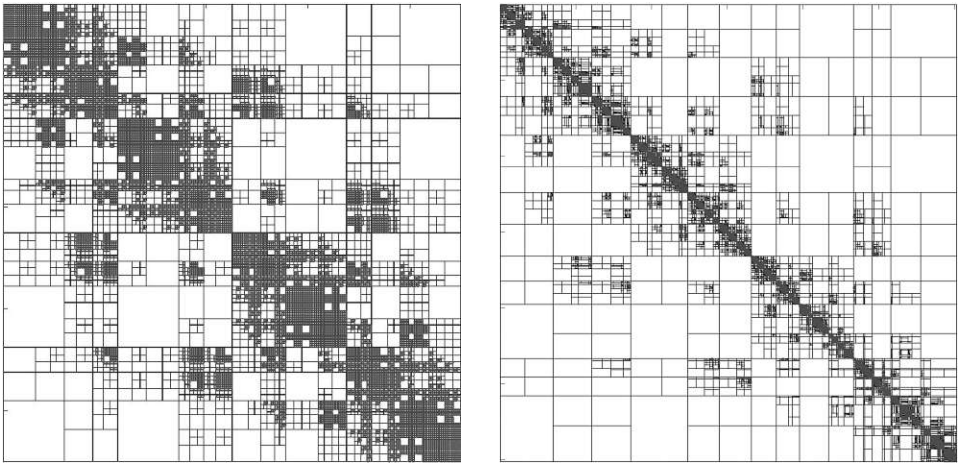

Figure 3.2: Two typical $\mathcal{H}$-matrices.

Once the block partition $\mathbb{P}^2 = \mathbb{P}^2_{\mathrm{small}} \cup \mathbb{P}^2_{\mathrm{adm}}$ is available, the corresponding class of hierarchical matrices is easy to describe.

Definition 3.43. *Let $\mathbb{P}^2$, $\mathbb{P}^2_{\mathrm{small}}$, $\mathbb{P}^2_{\mathrm{adm}}$ be defined as in C.3.42. Furthermore, let $r\in\mathbb{N}$ be a given rank bound. We define the class[5] of* $\mathcal{H}$-matrices *by*

$$\mathcal{H}(\mathbb{P}^2,r) := \{\boldsymbol{B}\in\mathbb{R}^{N\times N} \,|\, \forall (I,J)\in\mathbb{P}^2_{\mathrm{adm}} : \mathrm{rank}(\boldsymbol{B}|_{I\times J})\leq r\}.$$

We end this chapter with a note on the memory requirements for an arbitrary $\mathcal{H}$-matrix $\boldsymbol{B}\in\mathcal{H}(\mathbb{P}^2,r)$. Since the rank of an *admissible* block $\boldsymbol{B}|_{I\times J}$ is bounded by r, we know from

[5] Note that $\mathcal{H}(\mathbb{P}^2,r)$ is *not* a vector space, because the sum of two matrices with rank r has rank $2r$, in general.

Section 3.1 that there exist matrices $\boldsymbol{X}_r \in \mathbb{R}^{I\times r}$ and $\boldsymbol{Y}_r \in \mathbb{R}^{J\times r}$ such that $\boldsymbol{B}|_{I\times J} = \boldsymbol{X}_r \boldsymbol{Y}_r^T$. The pair $(\boldsymbol{X}_r, \boldsymbol{Y}_r)$ then takes up $r(\#I + \#J)$ units of memory. The *small* blocks $\boldsymbol{B}|_{I\times J}$, on the other hand, can simply be stored in full. The total memory requirements for $\boldsymbol{B}$ then add up to

$$\sum_{(I,J)\in\mathbb{P}^2_{\text{small}}} \#I\#J + \sum_{(I,J)\in\mathbb{P}^2_{\text{adm}}} r(\#I + \#J).$$

This quantity can be bounded as follows:

Lemma 3.44. *For all $r \in \mathbb{N}$, there holds the bound*

$$\sum_{(I,J)\in\mathbb{P}^2_{\text{small}}} \#I\#J + \sum_{(I,J)\in\mathbb{P}^2_{\text{adm}}} r(\#I + \#J) \leq C(d, \sigma_{\text{shp}}, \sigma_{\text{sprd}}, \sigma_{\text{adm}})(\sigma_{\text{small}} + r)\ln(h_{\min}^{-1})N.$$

Proof. For all $(I, J) \in \mathbb{P}^2_{\text{small}}$, we know from C.3.42 that

$$\#I\#J = \min\{\#I, \#J\}\max\{\#I, \#J\} \leq \sigma_{\text{small}}(\#I + \#J).$$

Then, using item 6 from C.3.42,

$$\begin{aligned}\sum_{(I,J)\in\mathbb{P}^2_{\text{small}}} \#I\#J + \sum_{(I,J)\in\mathbb{P}^2_{\text{adm}}} r(\#I + \#J) &\leq (\sigma_{\text{small}} + r)\sum_{(I,J)\in\mathbb{P}^2} \#I + \#J \\ &\leq C(d, \sigma_{\text{shp}}, \sigma_{\text{sprd}}, \sigma_{\text{adm}})(\sigma_{\text{small}} + r)\ln(h_{\min}^{-1})N.\end{aligned}$$

□

4 The main results

This chapter contains the main results of this thesis. We present an abstract framework for the approximation of inverse Gram matrices from the class of $\mathcal{H}$-matrices discussed in Chapter 3.

4.1 The discrete model problem

As usual, the first step of a rigorous analysis is to fix the functional analytic setting. Here, we use the Sobolev space $H^k(\Omega)$ from D.2.37 as the ambient space.

Definition 4.1. *Let $d, k \in \mathbb{N}$ and let $\Omega \subseteq \mathbb{R}^d$ be an H^k-extension domain (cf. D.2.48). Let*

$$V \subseteq H^k(\Omega)$$

be a closed subspace with the following property:

$$\forall \kappa \in C_0^\infty(\mathbb{R}^d) : \forall v \in V : \qquad (\kappa|_\Omega)v \in V.$$

Recall from D.2.37 that the natural inner product and norm on $H^k(\Omega)$ are given by

$$\langle v, w\rangle_{H^k(\Omega)} = \sum_{|\alpha|\leq k} \langle \mathrm{D}^\alpha v, \mathrm{D}^\alpha w\rangle_{L^2(\Omega)}, \qquad \|v\|_{H^k(\Omega)} = \left(\sum_{|\alpha|\leq k} \|\mathrm{D}^\alpha v\|^2_{L^2(\Omega)}\right)^{1/2}.$$

Apart from $\langle\cdot,\cdot\rangle_{H^k(\Omega)}$, we need another bilinear form on V, which need not be symmetric:

Definition 4.2. *Let $a(\cdot,\cdot) : V \times V \longrightarrow \mathbb{R}$ be a continuous, coercive bilinear form, i.e., there exists a constant $\sigma_{\mathrm{coco}} \geq 1$ such that, for all $u, v \in V$,*

$$|a(u,v)| \leq \sigma_{\mathrm{coco}} \|u\|_{H^k(\Omega)} \|v\|_{H^k(\Omega)}, \qquad \sigma_{\mathrm{coco}}^{-1} \|v\|^2_{H^k(\Omega)} \leq a(v,v).$$

Next, following Section 1.1, we introduce a discrete ansatz space.

Definition 4.3. *Let $N \in \mathbb{N}$ and let*

$$V_N \subseteq V$$

be a finite-dimensional subspace with $\dim(V_N) = N$. Furthermore, let

$$\{\varphi_1, \ldots, \varphi_N\} \subseteq V_N$$

be a basis of this space. The corresponding coordinate mapping is denoted by

$$\Phi : \left\{ \begin{array}{ccc} \mathbb{R}^N & \longrightarrow & V_N \\ \boldsymbol{c} & \longmapsto & \sum_{n=1}^N \boldsymbol{c}_n \varphi_n \end{array} \right. .$$

Note that Φ must be bijective, because the functions φ_n form a basis of V_N.

We also need the dual space of V_N. Since V_N is finite-dimensional, every linear functional on V_N is already continuous (with respect to any norm) so that the algebraic- and the topological dual space coincide.

Definition 4.4. *We define the dual space*

$$V_N^* := (V_N)^* := \{f : V_N \longrightarrow \mathbb{R} \,|\, f \text{ linear}\}.$$

For all $f \in V_N^*$ and $v \in V_N$, we set

$$\langle f, v\rangle_* := f(v), \qquad \|f\|_* := \sup_{v \in V_N} \frac{|\langle f, v\rangle_*|}{\|v\|_{H^k(\Omega)}}.$$

Our assumptions on the bilinear form $a(\cdot,\cdot)$ ensure that P.1.2 is a well-posed problem.

Lemma 4.5. *Let $f \in V_N^*$. Then, there exists a unique function $u \in V_N$ that satisfies the following discrete variational problem:*

$$\forall v \in V_N : \qquad a(u, v) = \langle f, v\rangle_*.$$

Furthermore, there holds the stability bound

$$\|u\|_{H^k(\Omega)} \leq \sigma_{\text{coco}} \|f\|_*.$$

Proof. See, e.g., [BS08, Theorem 2.7.7, Remark 2.7.11] (Lax-Milgram Lemma). □

Definition 4.6. *The linear operator*

$$S_N : V_N^* \longrightarrow V_N$$

that maps a right-hand side $f \in V_N^*$ to the corresponding discrete solution $S_N f := u \in V_N$ is called discrete solution operator.

Remark 4.7. *The continuous problem P.1.1 only serves as a reference. The upcoming analysis is based solely on the properties of the discrete problem P.1.2.*

Before we go to the next section, let us quickly lay out which applications we have in mind:

1. In Chapter 6, we will look at a finite element discretization of a second-order elliptic PDE with homogeneous Dirichlet boundary conditions. There, the ambient space is $V := H_0^1(\Omega)$ and $a(\cdot,\cdot)$ is the usual bilinear form for a second-order elliptic differential operator. The discrete ansatz space V_N is the spline space $\mathbb{S}_0^{p,1}(\mathcal{T})$ on a mesh $\mathcal{T}$ (cf. D.2.60 and D.2.70).

2. In Chapter 7, we are interested in a *radial basis function* interpolation problem, where a function of the form $u := \sum_{n=1}^N \boldsymbol{c}_n \varphi(\cdot - x_n)$ is used to interpolate given target values $\boldsymbol{f} \in \mathbb{R}^N$ on a set of predefined interpolation points $x_1, \ldots, x_N \in \mathbb{R}^d$. The correct ambient space for this problem is $V = H^k(\mathbb{R}^d)$, for some $k > d/2$, and the bilinear form $a(\cdot,\cdot)$ is a variant of the natural inner product on $H^k(\mathbb{R}^d)$. The ansatz space V_N is the span of the translates $\varphi(\cdot - x_n)$.

4.2 The Gram matrix $\boldsymbol{A}$

As discussed in Section 1.1, the discrete variational problem P.1.2 can be written as a linear system of equations, which is governed by the following matrix:

Definition 4.8. *We define the* Gram matrix

$$\boldsymbol{A} := (a(\varphi_n, \varphi_m))_{m,n=1}^N \in \mathbb{R}^{N\times N}.$$

The next lemma establishes a few basic properties of this matrix. Recall that $\Phi : \mathbb{R}^N \longrightarrow V_N$ is the coordinate mapping from D.4.3.

Lemma 4.9. *1. For all $\boldsymbol{c}, \boldsymbol{d} \in \mathbb{R}^N$, there holds the identity*

$$\langle \boldsymbol{A}\boldsymbol{c}, \boldsymbol{d}\rangle_2 = a(\Phi\boldsymbol{c}, \Phi\boldsymbol{d}).$$

2. The matrix $\boldsymbol{A}$ is positive definite[1], i.e.,

$$\forall \boldsymbol{c} \in \mathbb{R}^N\backslash\{\boldsymbol{0}\} : \qquad \langle \boldsymbol{A}\boldsymbol{c}, \boldsymbol{c}\rangle_2 > 0.$$

3. The matrix $\boldsymbol{A}$ is invertible.

Proof. Ad item 1: For all $\boldsymbol{c}, \boldsymbol{d} \in \mathbb{R}^N$, we have

$$\langle \boldsymbol{A}\boldsymbol{c}, \boldsymbol{d}\rangle_2 \overset{D.4.8}{=} \sum_{m,n=1}^N a(\varphi_n, \varphi_m)\boldsymbol{c}_n\boldsymbol{d}_m = a\Big(\sum_{n=1}^N \boldsymbol{c}_n\varphi_n, \sum_{m=1}^N \boldsymbol{d}_m\varphi_m\Big) \overset{D.4.3}{=} a(\Phi\boldsymbol{c}, \Phi\boldsymbol{d}).$$

Ad item 2: Let $\boldsymbol{c} \in \mathbb{R}^N\backslash\{\boldsymbol{0}\}$. Using the identity from item 1 and exploiting the bijectivity of the coordinate mapping Φ, we compute

$$\langle \boldsymbol{A}\boldsymbol{c}, \boldsymbol{c}\rangle_2 = a(\Phi\boldsymbol{c}, \Phi\boldsymbol{c}) \overset{D.4.2}{\geq} \sigma_{\text{coco}}^{-1}\|\Phi\boldsymbol{c}\|_{H^k(\Omega)}^2 > 0.$$

Ad item 3: Since $\boldsymbol{A}$ is a square matrix, it suffices to show injectivity. In fact, for every $\boldsymbol{c} \in \mathbb{R}^N$ with $\boldsymbol{c} \neq \boldsymbol{0}$, the relation $\langle \boldsymbol{A}\boldsymbol{c}, \boldsymbol{c}\rangle_2 > 0$ from item 2 already implies $\boldsymbol{A}\boldsymbol{c} \neq \boldsymbol{0}$. □

4.3 The dual basis $\lambda_1, \ldots, \lambda_N$

Our main result, T.4.21, is a statement about the approximability of the inverse $\boldsymbol{A}^{-1} \in \mathbb{R}^{N\times N}$ from the class of hierarchical matrices. The proof is based on an explicit formula for $\boldsymbol{A}^{-1}$ in terms of the discrete solution operator $S_N : V_N^* \longrightarrow V_N$ from D.4.6. If we want to express the action of $\boldsymbol{A}^{-1}$ on a given vector $\boldsymbol{f} \in \mathbb{R}^N$, we first need to convert $\boldsymbol{f}$ to a linear functional $f \in V_N^*$, which can be plugged into S_N. Since V_N is finite-dimensional, we have $\dim(V_N^*) = \dim(V_N) = N$ (e.g., [Axl15, Lemma 3.95]) and we may pick a basis with N elements:

[1] Note that $\boldsymbol{A}$ need not be symmetric, though.

Definition 4.10. *Denote by $\{\varphi_1, \dots, \varphi_N\} \subseteq V_N$ the basis functions from D.4.3. We define the* dual basis *$\{\lambda_1, \dots, \lambda_N\} \subseteq V_N^*$ via the following conditions:*

$$\forall n, m \in \{1, \dots, N\} : \qquad \langle \lambda_n, \varphi_m \rangle_* = \delta_{nm}.$$

Note that the functionals λ_n indeed form a basis of the dual space V_N^*, so that the name dual *basis* is justified. Furthermore, we mention that the dual basis is unique. The following locality assumption is essential for the subsequent analysis.

Assumption 4.11. *There exists a family of subsets*

$$\Omega_1, \dots, \Omega_N \subseteq \overline{\Omega}$$

with the following properties:

1. *There exist numbers $k_0 \in \{0, \dots, k\}$ and $\sigma_{\text{stab}} \geq 1$ such that, for all $n \in \{1, \dots, N\}$ and all $v \in V_N$,*
$$|\langle \lambda_n, v \rangle_*| \leq \sigma_{\text{stab}} \|v\|_{H^{k_0}(\Omega_n)}.$$
2. *There exist numbers $\sigma_{\text{shp}}, \sigma_{\text{ovlp}}, \sigma_{\text{sprd}} \geq 1$ such that the sets $\Omega_1, \dots, \Omega_N$ have shape regularity σ_{shp}, overlap σ_{ovlp} and spread σ_{sprd} (D.2.16, D.2.18, D.2.21).*

The second part of the assumption allows us to apply the results from Chapter 3. We adopt the name *characteristic sets* from D.3.5 and also some notation from D.3.6:

Definition 4.12. *For all $I \subseteq \{1, \dots, N\}$ and all $n \in \{1, \dots, N\}$, we define*

$$\Omega_I := \bigcup_{n \in I} \Omega_n \subseteq \mathbb{R}^d, \qquad h_{\Omega_n} := \operatorname{diam}_2(\Omega_n), \qquad h_{\min} := \min_{n \in \{1, \dots, N\}} h_{\Omega_n}.$$

The dual basis comes with its own set of coordinate mappings:

Definition 4.13. *We define the operators*

$$\Lambda : \begin{cases} \mathbb{R}^N & \longrightarrow & V_N^* \\ \boldsymbol{f} & \longmapsto & \sum_{n=1}^N \boldsymbol{f}_n \lambda_n \end{cases}, \qquad \Lambda^T : \begin{cases} V_N & \longrightarrow & \mathbb{R}^N \\ v & \longmapsto & (\langle \lambda_n, v \rangle_*)_{n=1}^N \end{cases}.$$

We summarize the essential properties of Λ and Λ^T and their connection to the coordinate mapping $\Phi : \mathbb{R}^N \longrightarrow V_N$ from D.4.3.

Lemma 4.14. 1. *The operators Λ and Λ^T are transposed in the following sense:*

$$\forall \boldsymbol{f} \in \mathbb{R}^N : \forall v \in V_N : \qquad \langle \Lambda \boldsymbol{f}, v \rangle_* = \langle \boldsymbol{f}, \Lambda^T v \rangle_2.$$

2. *The operators Φ and Λ are dual in the following sense:*

$$\forall \boldsymbol{f}, \boldsymbol{c} \in \mathbb{R}^N : \qquad \langle \Lambda \boldsymbol{f}, \Phi \boldsymbol{c} \rangle_* = \langle \boldsymbol{f}, \boldsymbol{c} \rangle_2.$$

3. *There holds $\Lambda^T = \Phi^{-1}$. In other words,*

$$\forall \boldsymbol{c} \in \mathbb{R}^N : \qquad \Lambda^T \Phi \boldsymbol{c} = \boldsymbol{c}, \qquad \forall v \in V_N : \qquad \Phi \Lambda^T v = v.$$

4. *Denote by* $k_0 \in \{0, \dots, k\}$ *and* $\sigma_{\mathrm{stab}}, \sigma_{\mathrm{ovlp}} \geq 1$ *the quantities from A.4.11. Then, for all* $\boldsymbol{f} \in \mathbb{R}^N$, $I \subseteq \{1, \dots, N\}$ *and* $v \in V_N$, *there hold the following relations*[2]*:*

$$\begin{aligned} \|\Lambda \boldsymbol{f}\|_* &\leq \sigma_{\mathrm{stab}} \sigma_{\mathrm{ovlp}}^{1/2} \|\boldsymbol{f}\|_2, \\ \|\Lambda^T v\|_{l^2(I)} &\leq \sigma_{\mathrm{stab}} \sigma_{\mathrm{ovlp}}^{1/2} \|v\|_{H^{k_0}(\Omega_I)}. \end{aligned}$$

Proof. Ad item 1: For all $\boldsymbol{f} \in \mathbb{R}^N$ and $v \in V_N$, we have

$$\langle \Lambda \boldsymbol{f}, v \rangle_* = \left\langle \sum_{n=1}^N \boldsymbol{f}_n \lambda_n, v \right\rangle_* = \sum_{n=1}^N \boldsymbol{f}_n \langle \lambda_n, v \rangle_* = \langle \boldsymbol{f}, \Lambda^T v \rangle_2.$$

Ad item 2: For all $\boldsymbol{f}, \boldsymbol{c} \in \mathbb{R}^N$, there holds

$$\langle \Lambda \boldsymbol{f}, \Phi \boldsymbol{c} \rangle_* = \left\langle \sum_{n=1}^N \boldsymbol{f}_n \lambda_n, \sum_{m=1}^N \boldsymbol{c}_m \varphi_m \right\rangle_* = \sum_{n,m=1}^N \langle \lambda_n, \varphi_m \rangle_* \boldsymbol{f}_n \boldsymbol{c}_m \overset{D.4.10}{=} \sum_{n=1}^N \boldsymbol{f}_n \boldsymbol{c}_n = \langle \boldsymbol{f}, \boldsymbol{c} \rangle_2.$$

Ad item 3: Let $\boldsymbol{c} \in \mathbb{R}^N$. Then, abbreviating $\boldsymbol{f} := \boldsymbol{c} - \Lambda^T \Phi \boldsymbol{c} \in \mathbb{R}^N$ and using the identities from items 1 and 2, we obtain

$$\|\boldsymbol{c} - \Lambda^T \Phi \boldsymbol{c}\|_2^2 = \langle \boldsymbol{f}, \boldsymbol{c} - \Lambda^T \Phi \boldsymbol{c} \rangle_2 = \langle \boldsymbol{f}, \boldsymbol{c} \rangle_2 - \langle \boldsymbol{f}, \Lambda^T \Phi \boldsymbol{c} \rangle_2 = \langle \boldsymbol{f}, \boldsymbol{c} \rangle_2 - \langle \Lambda \boldsymbol{f}, \Phi \boldsymbol{c} \rangle_* = 0.$$

Next, let $v \in V_N$. Since Φ is bijective, we may introduce the coefficient vector $\boldsymbol{c} := \Phi^{-1} v \in \mathbb{R}^N$. Using the previous identity, we get

$$\Phi \Lambda^T v = \Phi(\Lambda^T \Phi \boldsymbol{c}) = \Phi \boldsymbol{c} = v.$$

Ad item 4: Let $\boldsymbol{f} \in \mathbb{R}^N$. To get an estimate for $\|\Lambda \boldsymbol{f}\|_*$, we compute, for arbitrary $v \in V_N$,

$$\begin{aligned} |\langle \Lambda \boldsymbol{f}, v \rangle_*| &= \left| \left\langle \sum_{n=1}^N \boldsymbol{f}_n \lambda_n, v \right\rangle_* \right| = \left| \sum_{n=1}^N \boldsymbol{f}_n \langle \lambda_n, v \rangle_* \right| \\ &\leq \sum_{n=1}^N |\boldsymbol{f}_n| |\langle \lambda_n, v \rangle_*| \overset{A.4.11}{\leq} \sigma_{\mathrm{stab}} \sum_{n=1}^N |\boldsymbol{f}_n| \|v\|_{H^{k_0}(\Omega_n)} \\ &\overset{\text{C.S.}}{\leq} \sigma_{\mathrm{stab}} \|\boldsymbol{f}\|_2 \left(\sum_{n=1}^N \|v\|_{H^{k_0}(\Omega_n)}^2 \right)^{1/2} \overset{L.2.20}{\leq} \sigma_{\mathrm{stab}} \sigma_{\mathrm{ovlp}}^{1/2} \|\boldsymbol{f}\|_2 \|v\|_{H^{k_0}(\Omega)}. \end{aligned}$$

Since $k_0 \leq k$, we can plug in $\|v\|_{H^{k_0}(\Omega)} \leq \|v\|_{H^k(\Omega)}$ and obtain

$$\|\Lambda \boldsymbol{f}\|_* \overset{D.4.4}{=} \sup_{v \in V_N} \frac{|\langle \Lambda \boldsymbol{f}, v \rangle_*|}{\|v\|_{H^k(\Omega)}} \leq \sigma_{\mathrm{stab}} \sigma_{\mathrm{ovlp}}^{1/2} \|\boldsymbol{f}\|_2.$$

Finally, for all $I \subseteq \{1, \dots, N\}$ and $v \in V_N$, we have

$$\|\Lambda^T v\|_{l^2(I)}^2 = \sum_{n \in I} \langle \lambda_n, v \rangle_*^2 \overset{A.4.11}{\leq} \sigma_{\mathrm{stab}}^2 \sum_{n \in I} \|v\|_{H^{k_0}(\Omega_n)}^2 \overset{L.2.20}{\leq} \sigma_{\mathrm{stab}}^2 \sigma_{\mathrm{ovlp}} \|v\|_{H^{k_0}(\Omega_I)}^2.$$

This concludes the proof. □

[2] The norm $\|\cdot\|_*$ was defined in D.4.4 and the sets Ω_I were introduced in D.4.12.

4.4 The inverse matrix A^{-1}

Next, we derive the promised representation formula for the inverse Gram matrix $\boldsymbol{A}^{-1}$ in terms of the discrete solution operator S_N from D.4.6 and the coordinate mappings Λ, Λ^T from D.4.13. The mapping properties of these operators can be visualized as follows:

$$\mathbb{R}^N \xrightarrow{\Lambda} V_N^* \xrightarrow{S_N} V_N \xrightarrow{\Lambda^T} \mathbb{R}^N.$$

As we shall see now, $\boldsymbol{A}^{-1}$ is the matrix that represents the composition of these three operators. Furthermore, we show that $\boldsymbol{A}^{-1}$ is again a Gram matrix.

Lemma 4.15. *Denote by $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ the Gram matrix from D.4.8. Then, there holds the following identity:*

$$\forall \boldsymbol{f} \in \mathbb{R}^N : \qquad \boldsymbol{A}^{-1}\boldsymbol{f} = \Lambda^T S_N \Lambda \boldsymbol{f}.$$

In particular, we have

$$\forall m, n \in \{1, \dots, N\} : \qquad (\boldsymbol{A}^{-1})_{mn} = \langle \lambda_m, S_N \lambda_n \rangle_*.$$

Proof. Let $\boldsymbol{f} \in \mathbb{R}^N$. Setting $v := S_N \Lambda \boldsymbol{f} \in V_N$, we know from L.4.14 that $\Phi \Lambda^T v = v$. Then, for all $\boldsymbol{c} \in \mathbb{R}^N$, we compute

$$\langle \boldsymbol{A}\Lambda^T v, \boldsymbol{c}\rangle_2 \overset{L.4.9}{=} a(\Phi\Lambda^T v, \Phi\boldsymbol{c}) = a(v, \Phi\boldsymbol{c}) = a(S_N\Lambda\boldsymbol{f}, \Phi\boldsymbol{c}) \overset{D.4.6}{=} \langle \Lambda\boldsymbol{f}, \Phi\boldsymbol{c}\rangle_* \overset{L.4.14}{=} \langle \boldsymbol{f}, \boldsymbol{c}\rangle_2.$$

Since $\boldsymbol{c} \in \mathbb{R}^N$ was arbitrary, we get $\boldsymbol{A}\Lambda^T v = \boldsymbol{f}$ and ultimately $\boldsymbol{A}^{-1}\boldsymbol{f} = \Lambda^T v = \Lambda^T S_N \Lambda \boldsymbol{f}$. Finally, for all $m, n \in \{1, \dots, N\}$, using the Euclidean unit vectors $\boldsymbol{e}_m, \boldsymbol{e}_n \in \mathbb{R}^N$,

$$(\boldsymbol{A}^{-1})_{mn} = \langle \boldsymbol{e}_m, \boldsymbol{A}^{-1}\boldsymbol{e}_n\rangle_2 = \langle \boldsymbol{e}_m, \Lambda^T S_N \Lambda \boldsymbol{e}_n\rangle_2 \overset{L.4.14}{=} \langle \Lambda\boldsymbol{e}_m, S_N\Lambda\boldsymbol{e}_n\rangle_* = \langle \lambda_m, S_N\lambda_n\rangle_*.$$

This finishes the proof. □

4.5 The discrete Caccioppoli inequality

In L.2.55, we saw an example of a continuous variational problem whose solutions satisfy a so-called *Caccioppoli inequality*, i.e., a bound of a strong norm on a small set by a weaker norm on a slightly larger set. Here, we require a discrete version of such an inequality.

Definition 4.16. *The support of a vector $\boldsymbol{f} \in \mathbb{R}^N$ is defined by*

$$\operatorname{supp}(\boldsymbol{f}) := \{n \in \{1, \dots, N\} \,|\, \boldsymbol{f}_n \neq 0\}.$$

In analogy to D.3.7, we make the following definition:

Definition 4.17. *For every subset $D \subseteq \mathbb{R}^d$, we define the corresponding cluster*

$$\iota(D) := \{n \in \{1, \dots, N\} \,|\, \Omega_n \subseteq D\}.$$

Note that D.4.17 is slightly different from D.3.7, because, in order for an index $n \in \{1, \dots, N\}$ to lie in $\iota(D)$, the whole set Ω_n must be included in D. From this point onward, we only need the mapping $\iota : \mathbb{B} \longrightarrow \mathrm{Pow}\{1, \dots, N\}$ as defined in D.4.17.

Definition 4.18. *Denote by $S_N : V_N^* \longrightarrow V_N$ and $\Lambda : \mathbb{R}^N \longrightarrow V_N^*$ the operators from D.4.6 and D.4.13. For every subset $D \subseteq \mathbb{R}^d$, we define the subspace*

$$V_{\mathrm{sol}}(D) := \{S_N \Lambda \boldsymbol{f} \mid \boldsymbol{f} \in \mathbb{R}^N \text{ with } \mathrm{supp}(\boldsymbol{f}) \subseteq \iota(D)\} \subseteq V_N.$$

In particular, every function $u \in V_{\mathrm{sol}}(D)$ can be written in the following form (for some $\boldsymbol{f}_n \in \mathbb{R}$):

$$u = S_N\Big(\sum_{n \in \iota(D)} \boldsymbol{f}_n \lambda_n\Big).$$

Next, we remind the reader of D.2.8 and D.2.12, where we defined axes-parallel boxes $B \in \mathbb{B}$ and their inflated cousins $B^\delta \in \mathbb{B}$. We make the following assumption:

Assumption 4.19. *There exists a constant $\sigma_{\mathrm{Cacc}} \geq 1$, such that, for all $D \in \mathbb{B}$ and all $u \in V_{\mathrm{sol}}(D)$, the following statement is true: For all $B \in \mathbb{B}$ and all radii $\delta > 0$ satisfying $B^\delta \cap D = \emptyset$, there holds the* discrete Caccioppoli inequality

$$\delta^k |u|_{H^k(\Omega \cap B)} \leq \sigma_{\mathrm{Cacc}} \sum_{l=0}^{k-1} \delta^l |u|_{H^l(\Omega \cap B^\delta)}.$$

Figure 4.1 shall serve as a visual guide for the situation in A.4.19.

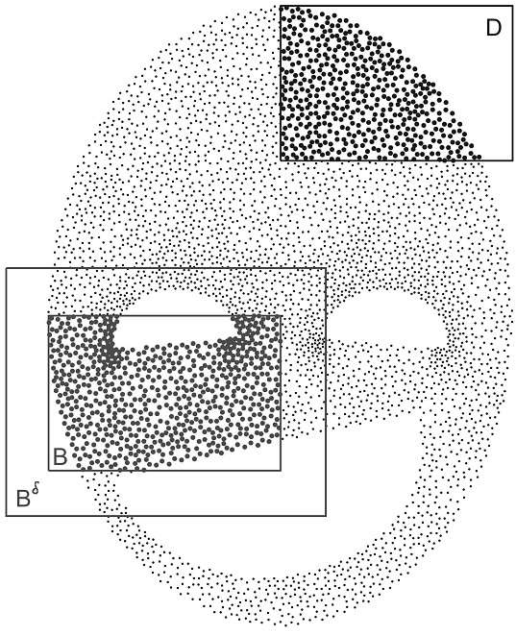

Figure 4.1: An example of boxes $B, D \in \mathbb{B}$ and $\delta > 0$ with $B^\delta \cap D = \emptyset$.

4.6 The main results

We are finally in the position to formulate the main results of this thesis. The bulk of the work lies in the construction of an approximation operator $Q^r_{B,D} : V_{\mathrm{sol}}(D) \longrightarrow V_{\mathrm{sol}}(D)$ for functions u lying in the solution space $V_{\mathrm{sol}}(D)$ from D.4.18. We only state the theorem here (cf. T.4.20) and defer its proof to Chapter 5. Then, in T.4.21, we use this operator $Q^r_{B,D}$ to construct approximations $\boldsymbol{B}^r_{I,J} \in \mathbb{R}^{I\times J}$ of the admissible blocks $\boldsymbol{A}^{-1}|_{I\times J}$ of the inverse Gram matrix $\boldsymbol{A}^{-1}$. Stitching these matrix blocks $\boldsymbol{B}^r_{I,J}$ together then produces an approximant $\boldsymbol{B}_r \approx \boldsymbol{A}^{-1}$ in the class of $\mathcal{H}$-matrices, $\mathcal{H}(\mathbb{P}^2, r)$.

We summarize the objects and assumptions that are relevant for the formulation of T.4.20 and T.4.21:

1. $d, k \in \mathbb{N}$, $\Omega \subseteq \mathbb{R}^d$ is an H^k-extension domain (cf. D.2.48) and $V \subseteq H^k(\Omega)$ is the subspace from D.4.1.
2. $a(\cdot,\cdot) : V \times V \longrightarrow \mathbb{R}$ is the bilinear form from D.4.2. The constant $\sigma_{\mathrm{coco}} \geq 1$ describes both continuity and coercivity.
3. $N \in \mathbb{N}$ and $V_N \subseteq V$ is an N-dimensional subspace. The system $\{\varphi_1, \dots, \varphi_N\} \subseteq V_N$ constitutes a basis (cf. D.4.3).
4. $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is the Gram matrix from D.4.8.
5. $\mathbb{B}$ is the set of axes-parallel boxes in $\mathbb{R}^d$ (cf. D.2.8).
6. It is assumed that the dual basis $\{\lambda_1, \dots, \lambda_N\} \subseteq V_N^*$ is *local* in the sense of A.4.11. The numbers $k_0 \in \{0, \dots, k\}$ and $\sigma_{\mathrm{stab}} \geq 1$ govern the stability bound and the quantities $\sigma_{\mathrm{shp}}, \sigma_{\mathrm{ovlp}}, \sigma_{\mathrm{sprd}} \geq 1$ describe shape regularity, overlap and spread of the characteristic sets $\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$ (cf. D.2.16, D.2.18, D.2.21).
7. It is assumed that there holds a discrete Caccioppoli inequality on the subspaces $V_{\mathrm{sol}}(D) \subseteq V_N$ from D.4.18, where $D \in \mathbb{B}$. The respective constant is called $\sigma_{\mathrm{Cacc}} > 0$ (cf. A.4.19).
8. $\mathbb{P}^2 = \mathbb{P}^2_{\mathrm{small}} \cup \mathbb{P}^2_{\mathrm{adm}}$ is the block partition from C.3.42, based on the characteristic sets $\Omega_n \subseteq \mathbb{R}^d$ from A.4.11. The quantity $h_{\min} > 0$ is the minimal diameter of the sets Ω_n (cf. D.3.6) and the numbers $\sigma_{\mathrm{small}} \geq 1$ and $\sigma_{\mathrm{adm}} > 0$ are the clustering parameters from C.3.42 (recall that $\sigma_{\mathrm{small}} \geq \sigma_{\mathrm{ovlp}}$ was required). The symbol $\mathcal{H}(\mathbb{P}^2, r) \subseteq \mathbb{R}^{N\times N}$ denotes the class of hierarchical matrices based on the block partition $\mathbb{P}^2$ and the rank bound $r \in \mathbb{N}$ (cf. D.3.43).

Now let us start with our first main result:

Theorem 4.20. *Consider two boxes $B, D \in \mathbb{B}$ that satisfy the following bounds:*

$$h_{\min} \leq \mathrm{diam}_2(B) \leq \sigma_{\mathrm{adm}} \mathrm{dist}_2(B, D) \leq \sigma_{\mathrm{adm}} \sqrt{d} \sigma_{\mathrm{sprd}}.$$

Then, for every $r \in \mathbb{N}$, there exists a linear operator

$$Q^r_{B,D} : V_{\mathrm{sol}}(D) \longrightarrow V_{\mathrm{sol}}(D)$$

with the following properties:

1. *There holds the rank bound*
$$\operatorname{rank}(Q^r_{B,D}) \leq r.$$
2. *There exist numbers* $C_0 \geq 1$ *and* $\sigma_{\mathrm{exp}} > 0$ *of the form*
$$C_0 = C(d, k, \Omega, \sigma_{\mathrm{sprd}}, \sigma_{\mathrm{adm}}), \qquad \sigma_{\mathrm{exp}} = C(d, k, \Omega, \sigma_{\mathrm{sprd}}, \sigma_{\mathrm{adm}})^{-1}\sigma_{\mathrm{Cacc}}^{-1},$$
such that, for all $m \in \{0, \ldots, k\}$ *and* $u \in V_{\mathrm{sol}}(D)$*, the following error bound is satisfied:*
$$\|u - Q^r_{B,D}u\|_{H^m(\Omega\cap B)} \leq C_0\sigma_{\mathrm{Cacc}}^m h_{\mathrm{min}}^{-m} \exp(-\sigma_{\mathrm{exp}} r^{1/(d+1)})\|u\|_{H^k(\Omega)}.$$

Proof. The proof will be given in the next chapter. □

T.4.20 tells us that the number of degrees of freedom to describe a function $u \in V_{\mathrm{sol}}(D)$ can be greatly reduced without losing too much information. In our second main result, we use this property to construct an $\mathcal{H}$-matrix approximation of the inverse Gram matrix.

Theorem 4.21. *For every* $r \in \mathbb{N}$*, there exists an* $\mathcal{H}$*-matrix*
$$\boldsymbol{B}_r \in \mathcal{H}(\mathbb{P}^2, r)$$
with the following properties:

1. *The memory requirements to store* $\boldsymbol{B}_r$ *can be bounded by*
$$C(d, \sigma_{\mathrm{shp}}, \sigma_{\mathrm{sprd}}, \sigma_{\mathrm{adm}})(\sigma_{\mathrm{small}} + r)\ln(h_{\mathrm{min}}^{-1})N.$$
2. *There exist numbers* $C_0 \geq 1$ *and* $\sigma_{\mathrm{exp}} > 0$ *of the form*
$$C_0 = C(d, k, \Omega, \sigma_{\mathrm{coco}}, \sigma_{\mathrm{shp}}, \sigma_{\mathrm{sprd}}, \sigma_{\mathrm{adm}}), \qquad \sigma_{\mathrm{exp}} = C(d, k, \Omega, \sigma_{\mathrm{sprd}}, \sigma_{\mathrm{adm}})^{-1}\sigma_{\mathrm{Cacc}}^{-1}$$
such that the following error bound is satisfied:
$$\|\boldsymbol{A}^{-1} - \boldsymbol{B}_r\|_2 \leq C_0\sigma_{\mathrm{stab}}^2\sigma_{\mathrm{ovlp}}\sigma_{\mathrm{Cacc}}^{k_0}\ln(h_{\mathrm{min}}^{-1})h_{\mathrm{min}}^{-k_0}\exp(-\sigma_{\mathrm{exp}} r^{1/(d+1)}).$$

Proof. Let $r \in \mathbb{N}$. We construct the matrix $\boldsymbol{B}_r$ in a block-wise manner. For all $(I, J) \in \mathbb{P}^2_{\mathrm{small}}$, we simply use the matrix $\boldsymbol{A}^{-1}$ itself:
$$\boldsymbol{B}_r|_{I\times J} := \boldsymbol{A}^{-1}|_{I\times J}.$$
Now let $(I, J) \in \mathbb{P}^2_{\mathrm{adm}}$. According to C.3.42, we can find boxes $B, D \in \mathbb{B}$ such that
$$\Omega_I \subseteq B, \qquad \Omega_J \subseteq D, \qquad h_{\mathrm{min}} \leq \operatorname{diam}_2(B) \leq \sigma_{\mathrm{adm}}\operatorname{dist}_2(B, D) \leq \sigma_{\mathrm{adm}}\sqrt{d}\sigma_{\mathrm{sprd}},$$
where $\Omega_I = \bigcup_{n\in I}\Omega_n \subseteq \overline{\Omega}$ and $\Omega_J = \bigcup_{n\in J}\Omega_n \subseteq \overline{\Omega}$ (cf. D.4.12). Denote by $E_J : \mathbb{R}^J \longrightarrow \mathbb{R}^N$ the trivial extension by zeros and by $R_I : \mathbb{R}^N \longrightarrow \mathbb{R}^I$ the restriction of a vector $\boldsymbol{c} \in \mathbb{R}^N$ to the entries $(\boldsymbol{c}_i)_{i\in I}$. Furthermore, denote by $S_N : V_N^* \longrightarrow V_N$, $\Lambda : \mathbb{R}^N \longrightarrow V_N^*$ and

$\Lambda^T : V_N \longrightarrow \mathbb{R}^N$ the operators from D.4.6 and D.4.13, respectively. Then, using the representation formula from L.4.15, we have the following identity:

$$\forall \boldsymbol{f} \in \mathbb{R}^J : \qquad \boldsymbol{A}^{-1}|_{I\times J}\boldsymbol{f} = R_I \Lambda^T S_N \Lambda E_J \boldsymbol{f}.$$

Note that $\operatorname{supp}(E_J\boldsymbol{f}) \subseteq J \subseteq \iota(D)$ (cf. D.4.17) so that $S_N\Lambda E_J\boldsymbol{f} \in V_{\mathrm{sol}}(D)$, according to D.4.18. In particular, we may plug this function into the operator $Q^r_{B,D} : V_{\mathrm{sol}}(D) \longrightarrow V_{\mathrm{sol}}(D)$ from T.4.20. Now, let $\boldsymbol{B}^r_{I,J} \in \mathbb{R}^{I\times J}$ be the matrix that represents the linear operator $R_I\Lambda^T Q^r_{B,D} S_N \Lambda E_J$, i.e.,

$$\forall \boldsymbol{f} \in \mathbb{R}^J : \qquad \boldsymbol{B}^r_{I,J}\boldsymbol{f} = R_I \Lambda^T Q^r_{B,D} S_N \Lambda E_J \boldsymbol{f}.$$

Clearly,

$$\operatorname{rank}(\boldsymbol{B}^r_{I,J}) \leq \operatorname{rank}(Q^r_{B,D}) \overset{T.4.20}{\leq} r.$$

As for the error bound, we get, for every $\boldsymbol{f} \in \mathbb{R}^J$:

$$\begin{aligned}
\|\boldsymbol{A}^{-1}|_{I\times J}\boldsymbol{f} - \boldsymbol{B}^r_{I,J}\boldsymbol{f}\|_{l^2(I)} &= \|\Lambda^T(\mathrm{id} - Q^r_{B,D})S_N\Lambda E_J\boldsymbol{f}\|_{l^2(I)} \\
&\overset{L.4.14}{\lesssim} \sigma_{\mathrm{stab}}\sigma_{\mathrm{ovlp}}^{1/2}\|(\mathrm{id} - Q^r_{B,D})S_N\Lambda E_J\boldsymbol{f}\|_{H^{k_0}(\Omega_I)} \\
&\overset{T.4.20}{\lesssim} \sigma_{\mathrm{stab}}\sigma_{\mathrm{ovlp}}^{1/2}\sigma_{\mathrm{Cacc}}^{k_0}h_{\min}^{-k_0}\exp(-\sigma_{\exp}r^{1/(d+1)})\|S_N\Lambda E_J\boldsymbol{f}\|_{H^k(\Omega)} \\
&\overset{L.4.5}{\lesssim} \sigma_{\mathrm{stab}}\sigma_{\mathrm{ovlp}}^{1/2}\sigma_{\mathrm{Cacc}}^{k_0}h_{\min}^{-k_0}\exp(-\sigma_{\exp}r^{1/(d+1)})\|\Lambda E_J\boldsymbol{f}\|_* \\
&\overset{L.4.14}{\leq} \sigma_{\mathrm{stab}}^2\sigma_{\mathrm{ovlp}}\sigma_{\mathrm{Cacc}}^{k_0}h_{\min}^{-k_0}\exp(-\sigma_{\exp}r^{1/(d+1)})\|\boldsymbol{f}\|_{l^2(J)}.
\end{aligned}$$

Now, set

$$\boldsymbol{B}_r|_{I\times J} := \boldsymbol{B}^r_{I,J}.$$

Then the definition of the matrix $\boldsymbol{B}_r \in \mathbb{R}^{N\times N}$ is complete. Since the ranks of the admissible blocks of $\boldsymbol{B}_r$ are bounded by r, it is clear that $\boldsymbol{B}_r \in \mathcal{H}(\mathbb{P}^2, r)$ (cf. D.3.43). The global error bound can be seen as follows:

$$\begin{aligned}
\|\boldsymbol{A}^{-1} - \boldsymbol{B}_r\|_2 &\overset{C.3.42}{\lesssim} \ln(h_{\min}^{-1}) \max_{(I,J)\in\mathbb{P}^2_{\mathrm{adm}}} \|\boldsymbol{A}^{-1}|_{I\times J} - \boldsymbol{B}^r_{I,J}\|_2 \\
&\lesssim \sigma_{\mathrm{stab}}^2\sigma_{\mathrm{ovlp}}\sigma_{\mathrm{Cacc}}^{k_0}\ln(h_{\min}^{-1})h_{\min}^{-k_0}\exp(-\sigma_{\exp}r^{1/(d+1)}).
\end{aligned}$$

This completes the proof of item 2. Item 1 is taken from L.3.44.

□

5 Construction of the operator $Q^r_{B,D}$

5.1 Overview

In this chapter, we provide the proof of T. 4.20, i.e., we construct the operator $Q^r_{B,D} : V_{\mathrm{sol}}(D) \longrightarrow V_{\mathrm{sol}}(D)$. First, let us give a rough overview:

1. Let $B, D \in \mathbb{B}$ be admissible boxes, let $r \in \mathbb{N}$ and let $u \in V_{\mathrm{sol}}(D)$. Our goal is to construct a "low-dimensional" function $\tilde{u} \in V_{\mathrm{sol}}(D)$ such that the error $\|u-\tilde{u}\|_{H^k(\Omega\cap B)}$ is small.

2. Since B is "far away" from D, we can inflate it $L \in \mathbb{N}$ times by a tiny amount $\delta > 0$ before hitting D (cf. Figure 5.1). This procedure generates a sequence of nested boxes:
$$B = B^{\delta\cdot 0} \subseteq B^{\delta\cdot 1} \subseteq \cdots \subseteq B^{\delta l} \subseteq \cdots \subseteq B^{\delta L} \subseteq \mathbb{R}^d \setminus D.$$

3. The hardest part of the proof is to construct a so-called[1] *single-step coarsening operator* $Q^\delta_{\tilde{B},D} : V_{\mathrm{sol}}(D) \longrightarrow V_{\mathrm{sol}}(D)$, where $\tilde{B}$ is any one of the boxes $B^{\delta l}$. This operator has rank $\mathcal{O}(L^d)$ and satisfies an error bound of the form
$$\|u - Q^\delta_{\tilde{B},D} u\|_{\Omega\cap\tilde{B},k,H} \leq \frac{1}{2}\|u\|_{\Omega\cap\tilde{B}^\delta,k,H},$$
where $H = \mathcal{O}(\delta)$ and where $\|\cdot\|_{\Omega\cap\tilde{B},k,H}$ is a certain H-weighted H^k-norm. We then combine L instances of this operator (one for each box) into a *multi-step coarsening operator* $Q^{\delta,L}_{B,D} : V_{\mathrm{sol}}(D) \longrightarrow V_{\mathrm{sol}}(D)$. This operator has rank $\mathcal{O}(L^{d+1})$ and the error recursion reduces roughly to
$$\|u - Q^{\delta,L}_{B,D} u\|_{H^k(\Omega\cap B)} \lesssim 2^{-L}\|u\|_{H^k(\Omega)}.$$

4. Finally, the operator $Q^r_{B,D}$ is just $Q^{\delta,L}_{B,D}$ for a certain choice of the parameters δ and L, where, roughly,
$$L \eqsim r^{1/(d+1)}, \qquad \delta \eqsim \mathrm{dist}_2(B,D)L^{-1}.$$

Let us fill in a few more details regarding the single-step coarsening operator, since its design is quite complicated.

1. Let $B, D \in \mathbb{B}$ and $\delta > 0$ be such that $B^\delta \cap D = \emptyset$. Given $u \in V_{\mathrm{harm}}(D)$, we want to find a "low-dimensional" function $\tilde{u} \in V_{\mathrm{sol}}(D)$ such that $\|u - Q^\delta_{B,D} u\|_{\Omega\cap B,k,H}$ is small.

[1] The name shall reflect the fact that a comparatively *coarse* mesh-like structure is used for approximation.

2. Since we know almost nothing about the shape of the set $\Omega \cap B$, it might be difficult to find a good approximant using standard approximation techniques with polynomials. We circumvent this problem by first extending u to a global function $E_\Omega u \in H^k(\mathbb{R}^d)$ and then looking for an approximant $\tilde{u} \in H^k(\mathbb{R}^d)$ such that $\|E_\Omega u - \tilde{u}\|_{B,k,H}$ is small. The extension of u is possible due to the assumption of Ω being an H^k-extension domain (cf. D.4.1).

3. The approximation step itself is done by a *low-rank approximation operator* $J^H : H^k(\mathbb{R}^d) \longrightarrow H^k(\mathbb{R}^d)$. We use a *partition of unity method*, i.e., we subdivide $\mathbb{R}^d$ into a family of congruent, overlapping boxes $T \in \mathbb{B}$ of side length $\mathcal{O}(H)$, where $H > 0$ is a free parameter. Along with the boxes T comes a family $(g_T)_T \subseteq C_0^\infty(\mathbb{R}^d)$ of bump functions which sum to 1 at every point $x \in \mathbb{R}^d$. Given $v \in H^k(\mathbb{R}^d)$, we pick, for each box T, a polynomial $v_T \in \mathbb{P}^{k-1}(\mathbb{R}^d)$ such that $\sum_{l=0}^k H^l |v - v_T|_{H^l(T)}$ is small. We then set $J^H v := \sum_T v_T g_T$ and derive a global error bound on $\mathbb{R}^d$ in terms of H.

4. Then, we restrict the output of $J^H E_\Omega u$ from $\mathbb{R}^d$ to Ω. We obtain an object in $H^k(\Omega)$, which in turn needs to be mapped to the subspace $V_{\mathrm{sol}}(D)$.

5. Now comes the part where things get complicated. Recall that our goal is to achieve an error bound of the form

$$\|u - \tilde{u}\|_{\Omega\cap B,k,H} \leq \frac{1}{2}\|u\|_{\Omega\cap B^\delta,k,H},$$

   i.e., the integration domain on the right-hand side must not exceed $\Omega \cap B^\delta$. Since J^H is a global operator acting on the full space $\mathbb{R}^d$, we need to squeeze in *two* smooth cut-off functions $\kappa \in C_0^\infty(\mathbb{R}^d)$ (with $\kappa|_B \equiv 1$ and $\mathrm{supp}(\kappa) \subseteq B^{\delta/2}$) in the right places[2]:

   a) The first κ is applied even *before* the extension from Ω to $\mathbb{R}^d$ is done. One of the last steps of the error estimation is the stability bound $\|E_\Omega v\|_{H^k(\mathbb{R}^d)} \lesssim \|v\|_{H^k(\Omega)}$ and we need the cut-off function to reduce the remainder to $\|v\|_{H^k(\Omega\cap B^{\delta/2})}$.

   b) The second κ is applied right *after* the extension from Ω to $\mathbb{R}^d$. The problem is that the extension operator E_Ω is not local, meaning that $\mathrm{supp}(E_\Omega(\kappa u)) \subseteq \mathbb{R}^d$ might be much larger than $\mathrm{supp}(\kappa u) \subseteq \Omega \cap B^{\delta/2}$. As a remedy, we simply multiply $E_\Omega(\kappa u)$ with κ again.

6. At this point, our approximant looks like

$$\tilde{u} = (J^H \kappa E_\Omega(\kappa u))|_\Omega \in H^k(\Omega)$$

   and we have to find a way to turn $\tilde{u}$ into an element of $V_{\mathrm{sol}}(D)$ again. It is tempting to use the orthogonal projection $P : H^k(\Omega) \longrightarrow V_{\mathrm{sol}}(D)$ for this purpose, but this approach interferes with the cut-off functions κ. The problem is that $(\kappa E_\Omega(\kappa u))|_\Omega$ need not lie in the space $V_{\mathrm{sol}}(D)$ again and we would have to balance the approximation properties of the operators J^H and P (which seems impossible).

   The trick is to introduce a slightly larger space $V_{\mathrm{sol}}(B, D) \supseteq V_{\mathrm{sol}}(D)$ that *is* closed under multiplication with the cut-off function κ. Then, we can use the orthogonal

[2] Strictly speaking, we have $\kappa \in C^\infty(\overline{\Omega})$ in the first instance and $\kappa \in C_0^\infty(\mathbb{R}^d)$ in the second instance.

projection $P^H_{B,D} : H^k(\Omega) \longrightarrow V_{\mathrm{sol}}(B,D)$ (with respect to a certain H-weighted H^k-norm) and we don't get irreconcilable error terms. The last step is to use a "minimum norm extension" operator $E_{B,D} : V_{\mathrm{sol}}(B,D) \longrightarrow V_{\mathrm{sol}}(D)$, so that the final output of the single-step coarsening operator $Q^\delta_{B,D}$ indeed lies in $V_{\mathrm{sol}}(D)$.

7. Using the error bound of the operator J^H and the stability bounds of all the other operators, we then end up with the following estimate:

$$\|u - Q^\delta_{B,D}u\|_{\Omega\cap B,k,H} \lesssim (H/\delta)^k \sum_{l=0}^{k} \delta^l |u|_{H^l(\Omega\cap B^{\delta/2})}.$$

At this point, the discrete Caccioppoli inequality from A.4.19 enters the picture and we can kill the k-th summand in the right-hand sum (in exchange for a slightly larger intergration domain). Finally, choosing $H = \mathcal{O}(\delta)$ correctly, we can produce the promised prefactor $1/2$ for the error bound in the H-weighted norms.

5.2 Weighted Sobolev norms

In D.2.37, we introduced the Sobolev spaces $H^k(\Omega)$ along with the following quantities:

$$\langle v,w\rangle_{H^k(\Omega)} := \sum_{|\alpha|\le k} \langle \mathrm{D}^\alpha v, \mathrm{D}^\alpha w\rangle_{L^2(\Omega)}, \qquad |v|_{H^l(\Omega)} := \Big(\sum_{|\alpha|=l} \|\mathrm{D}^\alpha v\|^2_{L^2(\Omega)}\Big)^{1/2},$$
$$\|v\|_{H^k(\Omega)} := \Big(\sum_{l=0}^{k} |v|^2_{H^l(\Omega)}\Big)^{1/2}.$$

In this chapter, we frequently use weighted variants thereof.

Definition 5.1. *Let $\Omega \subseteq \mathbb{R}^d$ be open, $k \in \mathbb{N}_0$ and $\varepsilon > 0$. For all $v, w \in H^k(\Omega)$, we set*

$$\langle v,w\rangle_{\Omega,k,\varepsilon} := \sum_{l=0}^{k} \varepsilon^{2l} \sum_{|\alpha|=l} \langle \mathrm{D}^\alpha v, \mathrm{D}^\alpha w\rangle_{L^2(\Omega)}, \qquad |v|_{\Omega,k,\varepsilon} := \varepsilon^k |v|_{H^l(\Omega)},$$
$$\|v\|_{\Omega,k,\varepsilon} := \Big(\sum_{l=0}^{k} \varepsilon^{2l} |v|^2_{H^l(\Omega)}\Big)^{1/2}.$$

Lemma 5.2. *Let $\Omega \subseteq \mathbb{R}^d$ be open, $k \in \mathbb{N}_0$ and $\varepsilon, \delta > 0$. For all $v \in H^k(\Omega)$, there hold the following inequalities:*

$$\begin{aligned} k^{-1/2} \sum_{l=0}^{k} \varepsilon^l |v|^l_{H^l(\Omega)} &\le \|v\|_{\Omega,k,\varepsilon} \le \sum_{l=0}^{k} \varepsilon^l |v|^l_{H^l(\Omega)}, \\ \min\{1,\varepsilon\}^k \|v\|_{H^k(\Omega)} &\le \|v\|_{\Omega,k,\varepsilon} \le \max\{1,\varepsilon\}^k \|v\|_{H^k(\Omega)}, \\ \min\{1,\varepsilon/\delta\}^k \|v\|_{\Omega,k,\delta} &\le \|v\|_{\Omega,k,\varepsilon} \le \max\{1,\varepsilon/\delta\}^k \|v\|_{\Omega,k,\delta}. \end{aligned}$$

In particular, the norms $\|\cdot\|_{H^k(\Omega)}$ and $\|\cdot\|_{\Omega,k,\varepsilon}$ are equivalent.

Proof. The first line follows from the norm equivalence $k^{-1/2}\|\cdot\|_1 \leq \|\cdot\|_2 \leq \|\cdot\|_1$ on $\mathbb{R}^k$. As for the second and third line, we only show the right-hand inequality:

$$\|v\|^2_{\Omega,k,\varepsilon} = \sum_{l=0}^{k} \varepsilon^{2l}|v|^2_{H^l(\Omega)} \leq \left(\max_{l\in\{0,\dots,k\}} \varepsilon^{2l}\right)\|v\|^2_{H^k(\Omega)} = \max\{1,\varepsilon\}^{2k}\|v\|^2_{H^k(\Omega)}.$$

Similarly,

$$\|v\|^2_{\Omega,k,\varepsilon} = \sum_{l=0}^{k} \varepsilon^{2l}|v|^2_{H^l(\Omega)} = \sum_{l=0}^{k} (\varepsilon/\delta)^{2l}\delta^{2l}|v|^2_{H^l(\Omega)} \leq \max\{1,\varepsilon/\delta\}^{2k}\|v\|^2_{\Omega,k,\delta}.$$

□

5.3 The cut-off operator $K^\delta_{\omega,B}$

In this short section, we introduce the *cut-off operator* $K^\delta_{\omega,B} : H^k(\omega) \longrightarrow H^k(\omega)$, which multiplies any given input $v \in H^k(\omega)$ with a fixed, smooth cut-off function $\kappa^\delta_B \in C^\infty_0(\mathbb{R}^d)$. The subscript ω is necessary, because we will need two separate instances of this operator, one on the H^k-extension domain $\omega = \Omega$ from D.4.1 and one on the full space $\omega = \mathbb{R}^d$ (cf. Section 5.1).

We remind the reader of D.2.8 and D.2.12, where we defined axes-parallel boxes $B \in \mathbb{B}$ and their inflated relatives $B^\delta \in \mathbb{B}$.

Lemma 5.3. *Let $B \in \mathbb{B}$ and $\delta > 0$. Then, there exists a* smooth cut-off function

$$\kappa^\delta_B \in C^\infty_0(\mathbb{R}^d)$$

with the following properties:

1. *There holds the inclusion* $\operatorname{supp}(\kappa^\delta_B) \subseteq B^\delta$.
2. *There holds* $\kappa^\delta_B|_B \equiv 1$ *and* $0 \leq \kappa^\delta_B \leq 1$.
3. *For every $l \in \mathbb{N}_0$, there holds the stability bound* $|\kappa^\delta_B|_{W^{l,\infty}(\mathbb{R}^d)} \leq C(d,l)\delta^{-l}$.

Proof. See [Hör90, Theorem 1.4.1.]

□

It is convenient to wrap the action of multiplying a given function with this cut-off function into an operator.

Definition 5.4. *Let $\omega \subseteq \mathbb{R}^d$ be open, $B \in \mathbb{B}$ and $\delta > 0$. Denote by $\kappa^\delta_B \in C^\infty_0(\mathbb{R}^d)$ the smooth cut-off function from L.5.3. We define the* cut-off operator[3]

$$K^\delta_{\omega,B} : \left\{ \begin{array}{ccc} H^k(\omega) & \longrightarrow & H^k(\omega) \\ v & \longmapsto & \kappa^\delta_B v \end{array} \right. .$$

[3] More precisely, we should write $(\kappa^\delta_B|_\omega)v$ instead of $\kappa^\delta_B v$.

Let us quickly summarize the defining properties of this operator:

Lemma 5.5. *1. For all $v \in H^k(\omega)$, there holds $\operatorname{supp}(K^\delta_{\omega,B} v) \subseteq \omega \cap B^\delta$.*

2. For all $v \in H^k(\omega)$, there holds $(K^\delta_{\omega,B} v)|_{\omega\cap B} = v|_{\omega\cap B}$.

3. For all $v \in H^k(\omega)$, there holds the stability bound (cf. D.5.1)

$$\|K^\delta_{\omega,B} v\|_{\omega,k,\delta} \leq C(d,k)\|v\|_{\omega\cap B^\delta,k,\delta}.$$

Proof. Items 1 and 2 follow immediately from L.5.3. To see item 3, we compute

$$\|K^\delta_{\omega,B} v\|_{\omega,k,\delta} \overset{L.5.2}{\leq} \sum_{l=0}^{k} \delta^l |K^\delta_{\omega,B} v|_{H^l(\omega)} = \sum_{l=0}^{k} \delta^l |\kappa^\delta_B v|_{H^l(\omega\cap B^\delta)}$$
$$\overset{L.2.42}{\lesssim} \sum_{l=0}^{k} \delta^l \sum_{j=0}^{l} |\kappa^\delta_B|_{W^{l-j,\infty}(\mathbb{R}^d)} |v|_{H^j(\omega\cap B^\delta)} \overset{L.5.3}{\lesssim} \sum_{j=0}^{k} \delta^j |v|_{H^j(\omega\cap B^\delta)} \overset{L.5.2}{\lesssim} \|v\|_{\omega\cap B^\delta,k,\delta}.$$

□

5.4 The low-rank approximation operator J^H

This operator is based on a *partition of unity method*, which is a well-known concept in general approximation theory (see, e.g., [Hör90, Section 1.4.] or [BM97]). The basic idea is to construct a family $\mathcal{T}$ of overlapping boxes $T \in \mathbb{B}$ that covers the full space $\mathbb{R}^d$. Furthermore, we need a corresponding family of bump functions $g_T \in C^\infty_0(\mathbb{R}^d)$ that sums to 1 at each individual point $x \in \mathbb{R}^d$.

Remark 5.6. *In this section, we use much the same notation as for simplicial meshes in Section 2.9, because the concepts are so similar. However, we emphasize that the symbol $\mathcal{T}$ now denotes a family of overlapping, axes-parallel boxes rather than a simplicial[4] mesh in the sense of D.2.60. To make the distinction clearer, we will call the members $T \in \mathcal{T}$* cells *(as opposed to* elements*).*

Definition 5.7. *Let $H > 0$.*

1. We define the reference cell

$$\hat{T} := [-1/4, 5/4)^d \in \mathbb{B}.$$

2. We define the family

$$\mathcal{T} := \{H(\hat{T} + m) \,|\, m \in \mathbb{Z}^d\}.$$

3. For every $T = H(\hat{T} + m) \in \mathcal{T}$, we define the following affine transformation (cf. D.2.23):

$$\forall x \in \mathbb{R}^d : \qquad F_T(x) := H(x + m).$$

[4] If anything, $\mathcal{T}$ would be a tensor product mesh, but the fact that the "elements" are overlapping prevent it from being an actual tensor product mesh.

4. For every physical subset $B \subseteq \mathbb{R}^d$, we define its patch

$$\mathcal{T}(B) := \{T \in \mathcal{T} \,|\, T \cap B \neq \emptyset\}.$$

Note that the elements $T \in \mathcal{T}$ are overlapping and that the transformations F_T are set up such that $F_T(\hat{T}) = T$.

Lemma 5.8. *1. For every $B \in \mathbb{B}$, there holds the bound*

$$\#\mathcal{T}(B) \leq C(d)(1 + \operatorname{diam}_2(B)/H)^d.$$

2. For all $\mathcal{S} \subseteq \mathcal{T}$, all $l \in \{0, \dots, k\}$ and all $v \in H^k(\mathbb{R}^d)$, there hold the bounds

$$|v|^2_{H^l(\bigcup \mathcal{S})} \leq \sum_{S \in \mathcal{S}} |v|^2_{H^l(S)} \leq C(d)|v|^2_{H^l(\bigcup \mathcal{S})}.$$

Proof. Ad item 1: For every $T \in \mathcal{T}(B)$, we can pick a point $x \in T \cap B$. Then, for all $y \in T$, we have $\|y - x\|_2 \leq \operatorname{diam}_2(T) = (3\sqrt{d}/2)H$ and L.2.13 implies

$$\bigcup_{T \in \mathcal{T}(B)} T \subseteq B^{(3\sqrt{d}/2)H},$$

where the right-hand side is an inflated box (cf. D.2.12). On the other hand, since the subsets $\{F_T([0,1)^d) \,|\, T \in \mathcal{T}(B)\}$ are pairwise disjoint, we have

$$H^d \cdot \#\mathcal{T}(B) = \sum_{T \in \mathcal{T}(B)} \operatorname{meas}(F_T([0,1)^d)) = \operatorname{meas}\Big(\bigcup_{T \in \mathcal{T}(B)} F_T([0,1)^d) \Big) \leq \operatorname{meas}\Big(\bigcup_{T \in \mathcal{T}(B)} T \Big)$$
$$\leq \operatorname{meas}(B^{(3\sqrt{d}/2)H}) \leq \operatorname{diam}_2(B^{(3\sqrt{d}/2)H})^d \overset{\text{L.2.13}}{\leq} C(d)(\operatorname{diam}_2(B) + H)^d.$$

Ad item 2: This can be proved in the same way as L.2.20. Here, according to item 1, the "overlap factor" is bounded by a constant $C(d) > 0$.

□

Next, we construct the bump functions that correspond to the individual cells $T \in \mathcal{T}$.

Lemma 5.9. *There exists a system of functions*

$$\{g_T \,|\, T \in \mathcal{T}\} \subseteq C^\infty_0(\mathbb{R}^d)$$

with the following properties:

1. For all $T \in \mathcal{T}$, there holds $\operatorname{supp}(g_T) \subseteq T$.

2. For all $T \in \mathcal{T}$ and $l \in \mathbb{N}_0$, there holds $|g_T|_{W^{l,\infty}(\mathbb{R}^d)} \leq C(d,l)H^{-l}$.

3. There holds the following identity:

$$\forall x \in \mathbb{R}^d : \qquad \sum_{T \in \mathcal{T}} g_T(x) = 1.$$

Proof. Applying L.5.3 to the situation $\Omega = \mathbb{R}^d$, $B = [0,1)^d \in \mathbb{B}$ and $\delta = 1/4$, we may pick a function $\kappa \in C_0^\infty(\mathbb{R}^d)$ with $\operatorname{supp}(\kappa) \subseteq \hat{T}$, $\kappa|_{[0,1)^d} \equiv 1$ and $0 \leq \kappa \leq 1$. The function $\tilde{\kappa} := \sum_{m\in\mathbb{Z}^d} \kappa(\cdot - m)$ then satisfies $\tilde{\kappa} \in C^\infty(\mathbb{R}^d)$, because the sum only ranges over a finite number of terms on each bounded subset of $\mathbb{R}^d$ (cf. L.5.8). Furthermore, there holds $\tilde{\kappa} \geq 1$, because for every $x \in \mathbb{R}^d$, there exists at least one $m(x) \in \mathbb{Z}^d$ such that $x - m(x) \in [0,1)^d$. Now it is not difficult to see that the function $\hat{g} := \kappa/\tilde{\kappa}$ has the following properties:

$$\operatorname{supp}(\hat{g}) \subseteq \hat{T}, \qquad \hat{g} \geq 0, \qquad \forall x \in \mathbb{R}^d : \quad \sum_{m\in\mathbb{Z}^d} \hat{g}(x-m) = 1.$$

Next, for every $T \in \mathcal{T}$, we use the affine transformation $F_T : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ from D.5.7 to define the function

$$g_T := \hat{g} \circ F_T^{-1}.$$

Clearly, $g_T \in C_0^\infty(\mathbb{R}^d)$ with $\operatorname{supp}(g_T) \subseteq T$. Furthermore, for all $l \in \mathbb{N}_0$, we have the stability bound

$$|g_T|_{W^{l,\infty}(\mathbb{R}^d)} = |\hat{g} \circ F_T^{-1}|_{W^{l,\infty}(T)} \overset{L.2.43}{\lesssim} h_T^{-l} |\hat{g}|_{W^{l,\infty}(\hat{T})} \lesssim H^{-l}.$$

Finally, for all $x \in \mathbb{R}^d$, we compute

$$\sum_{T\in\mathcal{T}} g_T(x) = \sum_{m\in\mathbb{Z}^d} \hat{g}(x/H - m) = 1.$$

This concludes the proof. □

Now we have everything we need to build the operator J^H. The basic idea was already presented in Section 5.1.

Lemma 5.10. *Let $H > 0$. Then, there exists a* low-rank approximation operator

$$J^H : H^k(\mathbb{R}^d) \longrightarrow H^k(\mathbb{R}^d)$$

with the following properties:

1. *For every box $B \in \mathbb{B}$, there holds the following local rank bound:*

$$\dim\{J^H v \,|\, v \in H^k(\mathbb{R}^d) \text{ with } \operatorname{supp}(v) \subseteq B\} \leq C(d,k)(1 + \operatorname{diam}_2(B)/H)^d.$$

2. *For all $v \in H^k(\mathbb{R}^d)$, there holds the following global error bound:*

$$\|v - J^H v\|_{\mathbb{R}^d,k,H} \leq C(d,k)|v|_{\mathbb{R}^d,k,H}.$$

Proof. Let $\hat{T} \subseteq \mathbb{R}^d$, $\mathcal{T} \subseteq \operatorname{Pow}(\mathbb{R}^d)$ and $F_T : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be defined as in D.5.7. Denote by $\{g_T \,|\, T \in \mathcal{T}\}$ the corresponding smooth partition of unity from L.5.9. Furthermore, let $\hat{J} : H^k(\hat{T}) \longrightarrow \mathbb{P}^{k-1}(\hat{T})$ be the orthogonal projection onto the closed subspace $\mathbb{P}^{k-1}(\hat{T}) \subseteq H^k(\hat{T})$.

Now, given a function $v \in H^k(\mathbb{R}^d)$, we introduce an approximating polynomial via

$$v_T := \hat{J}(v \circ F_T) \circ F_T^{-1} \in \mathbb{P}^{k-1}(\mathbb{R}^d).$$

(Note that we implicitly restricted $v \circ F_T$ from $\mathbb{R}^d$ to $\hat{T}$ and extended the polynomial $\hat{J}(v \circ F_T)$ from $\hat{T}$ to $\mathbb{R}^d$.) The asserted linear operator is defined as follows:

$$J^H : \left\{ \begin{array}{ccc} H^k(\mathbb{R}^d) & \longrightarrow & C^\infty(\mathbb{R}^d) \\ v & \longmapsto & \sum_{T\in\mathcal{T}} v_T g_T \end{array} \right. .$$

We mention that, in a neighbourhood around any given point $x \in \mathbb{R}^d$, the number of non-zero summands is finite, so that $J^H v$ is indeed a smooth function.

In order to derive the asserted rank bound, let $B \in \mathbb{B}$ be a given box. For every function $v \in H^k(\mathbb{R}^d)$ with $\operatorname{supp}(v) \subseteq B$ and every $T \in \mathcal{T}\backslash\mathcal{T}(B)$, there holds $(v \circ F_T)|_{\hat{T}} \equiv 0$, so that $v_T = 0$. Therefore,

$$\dim\{J^H v \,|\, v \in H^k(\mathbb{R}^d),\ \operatorname{supp}(v) \subseteq B\} = \dim\left\{ \sum_{T\in\mathcal{T}(B)} v_T g_T \,\middle|\, v \in H^k(\mathbb{R}^d),\ \operatorname{supp}(v) \subseteq B \right\}$$

$$\leq \dim\left\{ \sum_{T\in\mathcal{T}(B)} w_T g_T \,\middle|\, w_T \in \mathbb{P}^{k-1}(\mathbb{R}^d) \right\} \leq \dim(\mathbb{P}^{k-1}(\mathbb{R}^d))\#\mathcal{T}(B) \overset{L.5.8}{\lesssim} (1 + \operatorname{diam}_2(B)/H)^d.$$

Finally, in order to prove the error bound, let $v \in H^k(\mathbb{R}^d)$ be given. For every element $T \in \mathcal{T}$, the scaling argument L.2.43 and the Deny-Lions lemma[5] C.2.56 yield

$$\sum_{l=0}^{k} H^l |v - v_T|_{H^l(T)} \lesssim H^{d/2} \|v \circ F_T - \hat{J}(v \circ F_T)\|_{H^k(\hat{T})} \lesssim H^{d/2} |v \circ F_T|_{H^k(\hat{T})} \lesssim H^k |v|_{H^k(T)}.$$

Since the functions g_S sum to one (L.5.9), we have $v = \sum_{S\in\mathcal{T}} v g_S$. Then, for every $T \in \mathcal{T}$, we obtain the following error bound:

$$\sum_{l=0}^{k} H^l |v - J^H v|_{H^l(T)} = \sum_{l=0}^{k} H^l \Big| \sum_{S\in\mathcal{T}(T)} (v - v_S) g_S \Big|_{H^l(T)}$$

$$\leq \sum_{l=0}^{k} H^l \sum_{S\in\mathcal{T}(T)} |(v - v_S) g_S|_{H^l(T\cap S)} \overset{L.2.42}{\lesssim} \sum_{l=0}^{k} H^l \sum_{S\in\mathcal{T}(T)} \sum_{j=0}^{l} |v - v_S|_{H^j(S)} |g_S|_{W^{l-j,\infty}(\mathbb{R}^d)}$$

$$\overset{L.5.9}{\lesssim} \sum_{S\in\mathcal{T}(T)} \sum_{j=0}^{k} H^j |v - v_S|_{H^j(S)} \lesssim H^k \sum_{S\in\mathcal{T}(T)} |v|_{H^k(S)}.$$

Summing the squares over all $T \in \mathcal{T}$ and applying the bounds from L.5.8, the global error bound follows. In particular, using a simple triangle inequality, we also get the global stability bound $\|J^H v\|_{H^k(\mathbb{R}^d)} \leq C(d, H)\|v\|_{H^k(\mathbb{R}^d)} < \infty$, which proves that $J^H : H^k(\mathbb{R}^d) \longrightarrow H^k(\mathbb{R}^d)$. This finishes the proof.

□

[5] Note that $\hat{T}^\circ = (-1/4, 5/4)^d$ is open, bounded, connected and has a Lipschitz boundary.

5.5 The orthogonal projection $P^H_{B,D}$

We remind the reader of the spaces from D.4.18:

$$V_{\mathrm{sol}}(D) = \{S_N \Lambda \boldsymbol{f} \,|\, \boldsymbol{f} \in \mathbb{R}^N \text{ with } \mathrm{supp}(\boldsymbol{f}) \subseteq \iota(D)\} \subseteq V_N.$$

As mentioned in Section 5.1, these spaces have the following drawback: Let $u \in V_{\mathrm{sol}}(D)$ and let $B \in \mathbb{B}$ be a second box. If we modify u outside of B, the result need not lie in $V_{\mathrm{sol}}(D)$ again. For example, if we multiply u with a cut-off function κ that satisfies $\kappa|_B \equiv 1$, then $\kappa u \notin V_{\mathrm{sol}}(D)$, in general. To rectify this problem, we introduce a superspace $V_{\mathrm{sol}}(B, D) \supseteq V_{\mathrm{sol}}(D)$, where this is indeed the case.

Definition 5.11. *Let $B, D \in \mathbb{B}$ be given boxes. We set*

$$V_{\mathrm{sol}}(B, D) := \{u \in V \,|\, \exists \tilde{u} \in V_{\mathrm{sol}}(D) \text{ such that } \tilde{u}|_{\Omega \cap B} = u|_{\Omega \cap B}\}.$$

Note that $V_{\mathrm{sol}}(B, D)$ consists of global functions $u : \Omega \longrightarrow \mathbb{R}$ that merely happen to have a special structure on the subset $\Omega \cap B$. On the remaining part $\Omega \backslash B$, nothing is assumed about u.

Lemma 5.12. *Let $B, D \in \mathbb{B}$.*

1. *There hold the inclusions $V_{\mathrm{sol}}(D) \subseteq V_{\mathrm{sol}}(B, D) \subseteq V \subseteq H^k(\Omega)$.*
2. *For every $\delta > 0$, there holds $V_{\mathrm{sol}}(B^\delta, D) \subseteq V_{\mathrm{sol}}(B, D)$.*
3. *Let $u \in V_{\mathrm{sol}}(B, D)$. For every $v \in V$ with $v|_{\Omega \cap B} = u|_{\Omega \cap B}$, there holds $v \in V_{\mathrm{sol}}(B, D)$.*
4. *The subspace $V_{\mathrm{sol}}(B, D) \subseteq H^k(\Omega)$ is closed.*

Proof. We only prove the statement about closedness: Consider the subspace

$$Z := \{\tilde{u}|_{\Omega \cap B} \,|\, \tilde{u} \in V_{\mathrm{sol}}(D)\} \subseteq H^k(\Omega \cap B).$$

Since $\dim(Z) \leq \dim(V_N) = N < \infty$, we know that Z is a closed subspace of $H^k(\Omega \cap B)$. Note that, for any given function $u \in H^k(\Omega)$, there holds the following equivalence:

$$v \in V_{\mathrm{sol}}(B, D) \qquad \Leftrightarrow \qquad (u \in V \quad \wedge \quad u|_{\Omega \cap B} \in Z).$$

Now, let $(u_n)_{n \in \mathbb{N}} \subseteq V_{\mathrm{sol}}(B, D)$ and $u \in H^k(\Omega)$ with $\|u - u_n\|_{H^k(\Omega)} \xrightarrow{n} 0$. In particular, for every $n \in \mathbb{N}$, we know that $u_n \in V$ and that $u_n|_{\Omega \cap B} \in Z$. Since $V \subseteq H^k(\Omega)$ is closed (cf. D.4.1), we infer $u \in V$. On the other hand, the trivial bound $\|u - u_n\|_{H^k(\Omega \cap B)} \leq \|u - u_n\|_{H^k(\Omega)} \xrightarrow{n} 0$ and the closedness of Z yield $u|_{\Omega \cap B} \in Z$. According to the equivalence above, this means $u \in V_{\mathrm{sol}}(B, D)$.

□

We finish this section with a projection from $H^k(\Omega)$ to the subspace $V_{\mathrm{sol}}(B, D) \subseteq H^k(\Omega)$.

Lemma 5.13. *Let $B, D \in \mathbb{B}$ be given boxes and $H > 0$ be a given parameter. There exists a linear operator*

$$P^H_{B,D} : H^k(\Omega) \longrightarrow V_{\mathrm{sol}}(B, D)$$

with the following properties:

1. Projection: *For every $u \in V_{\mathrm{sol}}(B, D)$, there holds $P^H_{B,D} u = u$.*
2. Stability: *For all $u \in H^k(\Omega)$, there holds the stability bound*

$$\|P^H_{B,D} u\|_{\Omega,k,H} \leq \|u\|_{\Omega,k,H}.$$

Proof. From L.5.2, we know that the norms $\|\cdot\|_{H^k(\Omega)}$ and $\|\cdot\|_{\Omega,k,H}$ are equivalent (with constants depending on H). Furthermore, L.5.12 tells us that $V_{\mathrm{sol}}(B, D) \subseteq H^k(\Omega)$ is a closed subspace with respect to $\|\cdot\|_{H^k(\Omega)}$ and thus also with respect to $\|\cdot\|_{\Omega,k,H}$. In particular, the orthogonal projection $P^H_{B,D} : H^k(\Omega) \longrightarrow V_{\mathrm{sol}}(B, D)$ with respect to the H-weighted inner product $\langle\cdot,\cdot\rangle_{\Omega,k,H}$ from D.5.1 is well-defined. Item 1 is self-explanatory and item 2 follows easily from the fact that $P^H_{B,D}$ is the orthogonal projection. □

5.6 The minimum-norm extension operator $E_{B,D}$

Let $B, D \in \mathbb{B}$ be given boxes and consider an element $u \in V_{\mathrm{sol}}(B, D)$ (cf. D.5.11). By definition of this space, there exists at least one function $\tilde{u} \in V_{\mathrm{sol}}(D)$ such that $\tilde{u}|_{\Omega\cap B} = u|_{\Omega\cap B}$. However, such an extension[6] $\tilde{u}$ need not be unique. In fact, for every $\tilde{u}_0 \in V_{\mathrm{sol}}(D)$ with $\tilde{u}_0|_{\Omega\cap B} = 0$, the sum $\tilde{u}+\tilde{u}_0$ is a viable extension of u as well. Immediately, the question arises of how to make a meaningful choice from this affine subspace. For our purposes, the minimum-norm extension is sufficient[7]:

Lemma 5.14. *Let $B, D \in \mathbb{B}$ be given boxes. There exists a linear operator*

$$E_{B,D} : V_{\mathrm{sol}}(B, D) \longrightarrow V_{\mathrm{sol}}(D)$$

with the following properties:

1. *For all $u \in V_{\mathrm{sol}}(B, D)$, there holds $(E_{B,D} u)|_{\Omega\cap B} = u|_{\Omega\cap B}$.*
2. *For all $u \in V_{\mathrm{sol}}(B, D)$, there holds the bound*

$$\|E_{B,D} u\|_{H^k(\Omega)} \leq \inf_{\substack{\tilde{u}\in V_{\mathrm{sol}}(D):\\ \tilde{u}|_{\Omega\cap B} = u|_{\Omega\cap B}}} \|\tilde{u}\|_{H^k(\Omega)}.$$

[6] Usually, an *extension* of a function $f : M \longrightarrow \mathbb{R}$ is a function $\tilde{f} : \tilde{M} \longrightarrow \mathbb{R}$ which is defined on a larger set $\tilde{M} \supseteq M$ and which satisfies $\tilde{f}|_M = f$. Here, we somewhat abuse the term *extension*, because u and $\tilde{u}$ are already defined on the same set (namely Ω).

[7] In fact, we only need the process of choosing an extension to be a linear operator. We state the stability bound for the sake of completeness, but we won't actually need it later on.

Proof. Consider the subspace

$$Z := \{u \in V_{\mathrm{sol}}(D) \,|\, u|_{\Omega\cap B} = 0\} \subseteq H^k(\Omega).$$

Then Z is finite-dimensional and thus a closed subspace of $H^k(\Omega)$. In particular, we may introduce the orthogonal projection $P : H^k(\Omega) \longrightarrow Z$. Recall that, for every $\tilde{u} \in H^k(\Omega)$, the image $P\tilde{u} \in Z$ is characterized by the following variational equation:

$$\forall z \in Z : \qquad \langle P\tilde{u}, z\rangle_{H^k(\Omega)} = \langle \tilde{u}, z\rangle_{H^k(\Omega)}.$$

Now, for every given $u \in V_{\mathrm{sol}}(B, D)$, we pick a function $\tilde{u} \in V_{\mathrm{sol}}(D)$ with $\tilde{u}|_{\Omega\cap B} = u|_{\Omega\cap B}$ and set

$$E_{B,D}u := \tilde{u} - P\tilde{u} \in V_{\mathrm{sol}}(D).$$

First, let us check that this definition is independent of the choice of $\tilde{u}$. In fact, if $\overline{u} \in V_{\mathrm{sol}}(D)$ is another function with $\overline{u}|_{\Omega\cap B} = u|_{\Omega\cap B}$, then the error $\overline{u} - \tilde{u}$ lies in Z. Since P is a projection onto Z, we get $P(\overline{u} - \tilde{u}) = \overline{u} - \tilde{u}$. It follows that $\tilde{u}$ and $\overline{u}$ indeed produce the same result:

$$\overline{u} - P\overline{u} = (\overline{u} - \tilde{u}) - P(\overline{u} - \tilde{u}) + (\tilde{u} - P\tilde{u}) = \tilde{u} - P\tilde{u}.$$

Now that the mapping $E_{B,D}$ is well-defined, let us derive its main properties: Its linearity follows from $V_{\mathrm{sol}}(D)$ being a vector space and the linearity of P. To see item 1, let $u \in V_{\mathrm{sol}}(B, D)$ and pick $\tilde{u} \in V_{\mathrm{sol}}(D)$ with $\tilde{u}|_{\Omega\cap B} = u|_{\Omega\cap B}$. Since $P\tilde{u} \in Z$ vanishes on $\Omega \cap B$, we find that

$$(E_{B,D}u)|_{\Omega\cap B} = \tilde{u}|_{\Omega\cap B} - (P\tilde{u})|_{\Omega\cap B} = \tilde{u}|_{\Omega\cap B} = u|_{\Omega\cap B}.$$

Finally, to see the stability bound, let $u \in V_{\mathrm{sol}}(B, D)$ and $\overline{u} \in V_{\mathrm{sol}}(D)$ with $\overline{u}|_{\Omega\cap B} = u|_{\Omega\cap B}$. Since P is an orthogonal projection, we obtain

$$\begin{aligned} \|E_{B,D}u\|_{H^k(\Omega)} &= \|\overline{u} - P\overline{u}\|_{H^k(\Omega)} = \inf_{z\in Z} \|\overline{u} - z\|_{H^k(\Omega)} \\ &= \inf_{\substack{\tilde{u}\in V_{\mathrm{sol}}(D):\\ \tilde{u}|_{\Omega\cap B} = \overline{u}|_{\Omega\cap B}}} \|\tilde{u}\|_{H^k(\Omega)} = \inf_{\substack{\tilde{u}\in V_{\mathrm{sol}}(D):\\ \tilde{u}|_{\Omega\cap B} = u|_{\Omega\cap B}}} \|\tilde{u}\|_{H^k(\Omega)}. \end{aligned}$$

This concludes the proof. □

5.7 The single-step coarsening operator $Q^\delta_{B,D}$

This section contains the most complicated part in the derivation of our main result, T.4.21. We combine all of the operators from the previous sections in this chapter (and the extension operator from D.2.48) and construct the *single-step coarsening operator*.

Theorem 5.15. *Denote by $\sigma_{\mathrm{sprd}} > 0$ and $\sigma_{\mathrm{Cacc}} \geq 1$ the constants from A.4.11 and A.4.19. Let $B, D \in \mathbb{B}$ be given boxes and $\delta > 0$ be a free parameter with $\delta \leq \sigma_{\mathrm{sprd}}$ and $B^\delta \cap D = \emptyset$. Then, there exists a linear* single-step coarsening operator

$$Q^\delta_{B,D} : V_{\mathrm{sol}}(D) \longrightarrow V_{\mathrm{sol}}(D)$$

with the following properties:

1. *There holds the rank bound*

$$\operatorname{rank}(Q^\delta_{B,D}) \leq C(d,k,\Omega,\sigma_{\mathrm{sprd}})\sigma^d_{\mathrm{Cacc}}(1+\operatorname{diam}_2(B)/\delta)^d.$$

2. *There exists a number $H>0$ of the form*

$$H = \frac{\delta}{C(d,k,\Omega,\sigma_{\mathrm{sprd}})\sigma_{\mathrm{Cacc}}}$$

such that, for all $u \in V_{\mathrm{sol}}(D)$, there holds[8]

$$\|u - Q^\delta_{B,D}u\|_{\Omega\cap B,k,H} \leq \frac{1}{2}\|u\|_{\Omega\cap B^\delta,k,H}.$$

Proof. The alleged operator $Q^\delta_{B,D}$ is composed of seven other operators:

1. Denote by $K^{\delta/2}_{\Omega,B} : H^k(\Omega) \longrightarrow H^k(\Omega)$ the cut-off operator from D.5.4, applied to the set Ω, the box B and the parameter $\delta/2$. Similarly, let $K^{\delta/2}_{\mathbb{R}^d,B} : H^k(\mathbb{R}^d) \longrightarrow H^k(\mathbb{R}^d)$ be the cut-off operator corresponding to the set $\mathbb{R}^d$, the box B and the parameter $\delta/2$.
2. Denote by $E_\Omega : H^k(\Omega) \longrightarrow H^k(\mathbb{R}^d)$ the Sobolev extension operator from D. 2.48 (recall from D.4.1 that $\Omega \subseteq \mathbb{R}^d$ is assumed to be an H^k-extension domain).
3. Let $H > 0$ and denote by $J^H : H^k(\mathbb{R}^d) \longrightarrow H^k(\mathbb{R}^d)$ the low-rank approximation operator from L.5.10. The precise value of H will be chosen during the proof.
4. Let $R_\Omega : H^k(\mathbb{R}^d) \longrightarrow H^k(\Omega)$ be the restriction operator, i.e., $R_\Omega v := v|_\Omega$.
5. Denote by $P^H_{B,D} : H^k(\Omega) \longrightarrow V_{\mathrm{sol}}(B,D)$ the projection from L.5.13 with respect to the parameter $H > 0$ from before.
6. Let $E_{B,D} : V_{\mathrm{sol}}(B,D) \longrightarrow V_{\mathrm{sol}}(D)$ be the minimum-norm extension operator from L.5.14.

The mapping properties of these operators are summarized in the following schematic:

$$V_{\mathrm{sol}}(D) \overset{D.4.18}{\subseteq} V \overset{D.4.1}{\subseteq} H^k(\Omega) \xrightarrow{K^{\delta/2}_{\Omega,B}} H^k(\Omega) \xrightarrow{E_\Omega} H^k(\mathbb{R}^d) \xrightarrow{K^{\delta/2}_{\mathbb{R}^d,B}} H^k(\mathbb{R}^d) \dots$$
$$\dots \xrightarrow{J^H} H^k(\mathbb{R}^d) \xrightarrow{R_\Omega} H^k(\Omega) \xrightarrow{P^H_{B,D}} V_{\mathrm{sol}}(B,D) \xrightarrow{E_{B,D}} V_{\mathrm{sol}}(D).$$

Now, define

$$Q^\delta_{B,D} := E_{B,D}\, P^H_{B,D}\, R_\Omega\, J^H\, K^{\delta/2}_{\mathbb{R}^d,B}\, E_\Omega\, K^{\delta/2}_{\Omega,B} : V_{\mathrm{sol}}(D) \longrightarrow V_{\mathrm{sol}}(D).$$

[8] In fact, we may even write $k-1$ instead of k on the right-hand side.

We begin our analysis with the error bound, since it tells us how to choose the free parameter $H > 0$. To this end, let $u \in V_{\mathrm{sol}}(D)$ be given. In order to bound the error $u - Q^\delta_{B,D}u$ on $\Omega \cap B$, we first have to find similar expressions for $u|_{\Omega\cap B}$ and $(Q^\delta_{B,D}u)|_{\Omega\cap B}$.

On one hand, using the definition of $Q^\delta_{B,D}$, we have

$$\begin{aligned}(Q^\delta_{B,D}u)|_{\Omega\cap B} &= (E_{B,D}\, P^H_{B,D}\, R_\Omega\, J^H\, K^{\delta/2}_{\mathbb{R}^d,B}\, E_\Omega\, K^{\delta/2}_{\Omega,B}\, u)|_{\Omega\cap B} \\ &\overset{L.5.14}{=} (P^H_{B,D}\, R_\Omega\, J^H\, K^{\delta/2}_{\mathbb{R}^d,B}\, E_\Omega\, K^{\delta/2}_{\Omega,B}\, u)|_{\Omega\cap B}.\end{aligned}$$

On the other hand, dropping the operators $P^H_{B,D}$ and J^H from the remaining expression, let us introduce the auxiliary function

$$v := R_\Omega\, K^{\delta/2}_{\mathbb{R}^d,B}\, E_\Omega\, K^{\delta/2}_{\Omega,B}\, u \quad \in H^k(\Omega).$$

Recall from D.5.4 that the cut-off operators $K^{\delta/2}_{\mathbb{R}^d,B}$ and $K^{\delta/2}_{\Omega,B}$ realize the multiplication with the smooth cut-off function $\kappa := \kappa^{\delta/2}_B \in C^\infty_0(\mathbb{R}^d)$ from L.5.3. In particular, we can write v in the following way:

$$v = R_\Omega\, K^{\delta/2}_{\mathbb{R}^d,B}\, E_\Omega\, K^{\delta/2}_{\Omega,B}\, u = (R_\Omega\kappa)\cdot(R_\Omega E_\Omega K^{\delta/2}_{\Omega,B}u) \overset{D.2.48}{=} (\kappa^2|_\Omega)u.$$

From this representation, it follows that $v \in V$, because $u \in V_{\mathrm{sol}}(D) \subseteq V_{\mathrm{sol}}(B,D) \subseteq V$ and V is closed under multiplication with test functions (cf. D.4.1). In fact, L.5.12 and the identity

$$v|_{\Omega\cap B} = (\kappa^2|_{\Omega\cap B})(u|_{\Omega\cap B}) \overset{L.5.3}{=} u|_{\Omega\cap B}$$

even yield $v \in V_{\mathrm{sol}}(B,D)$. It follows that $P^H_{B,D}v = v$, because $P^H_{B,D}$ is a projection onto the space $V_{\mathrm{sol}}(B,D)$. We obtain the following representation of u on $\Omega \cap B$:

$$u|_{\Omega\cap B} = v|_{\Omega\cap B} = (P^H_{B,D}v)|_{\Omega\cap B} = (P^H_{B,D}\, R_\Omega\, K^{\delta/2}_{\mathbb{R}^d,B}\, E_\Omega\, K^{\delta/2}_{\Omega,B}\, u)|_{\Omega\cap B}.$$

Note that the expressions for $u|_{\Omega\cap B}$ and $(Q^\delta_{B,D}u)|_{\Omega\cap B}$ only differ by J^H. In particular, in order to estimate $(u - Q^\delta_{B,D}u)|_{\Omega\cap B}$, it suffices to have an *error* bound for J^H and *stability* bounds for all the remaining operators. As for the extension operator E_Ω, we will need the following bound, which holds true for all $v \in H^k(\Omega)$ and makes use of the assumption $\delta \le \sigma_{\mathrm{sprd}}$:

$$\|E_\Omega v\|_{\mathbb{R}^d,k,\delta} \overset{L.5.2}{\le} \sum_{l=0}^k \delta^l |E_\Omega v|_{H^l(\Omega)} \overset{D.2.48}{\lesssim} \sum_{l=0}^k \delta^l \|v\|_{H^l(\Omega)} \lesssim \sum_{l=0}^k \delta^l |v|_{H^l(\Omega)} \overset{L.5.2}{\lesssim} \|v\|_{\Omega,k,\delta}.$$

Furthermore, the assumption $B^\delta \cap D = \emptyset$ allows us to apply the discrete Caccioppoli inequality from A.4.19 to the function $u \in V_{\mathrm{sol}}(D)$, the box $B^{\delta/2}$ and the parameter $\delta/2$. Expressed in terms of the weighted norms from D.5.1, we get

$$\|u\|_{\Omega\cap B^{\delta/2},k,\delta} \le C(k)\sigma_{\mathrm{Cacc}}\|u\|_{\Omega\cap B^\delta,k-1,\delta}.$$

Finally, we have everything we need to bound the error $(u - Q^\delta_{B,D}u)|_{\Omega\cap B}$. We start with the H-weighted norm (= natural choice for J^H), switch to the δ-weighted norm in

between (= natural choice for cut-off operators and Caccioppoli inequality) and finish with the H-weighted norm again:

$$\begin{aligned}
\|u - Q^\delta_{B,D}u\|_{\Omega\cap B,k,H} &= \|P^H_{B,D}\, R_\Omega\,(\mathrm{id} - J^H)\, K^{\delta/2}_{\mathbb{R}^d,B}\, E_\Omega\, K^{\delta/2}_{\Omega,B} u\|_{\Omega\cap B,k,H}\\
&\le \|P^H_{B,D}\, R_\Omega\,(\mathrm{id} - J^H)\, K^{\delta/2}_{\mathbb{R}^d,B}\, E_\Omega\, K^{\delta/2}_{\Omega,B} u\|_{\Omega,k,H}\\
&\overset{L.5.13}{\lesssim} \|R_\Omega\,(\mathrm{id} - J^H)\, K^{\delta/2}_{\mathbb{R}^d,B}\, E_\Omega\, K^{\delta/2}_{\Omega,B} u\|_{\Omega,k,H}\\
&\le \|(\mathrm{id} - J^H)\, K^{\delta/2}_{\mathbb{R}^d,B}\, E_\Omega\, K^{\delta/2}_{\Omega,B} u\|_{\mathbb{R}^d,k,H}\\
&\overset{L.5.10}{\lesssim} |K^{\delta/2}_{\mathbb{R}^d,B}\, E_\Omega\, K^{\delta/2}_{\Omega,B} u|_{\mathbb{R}^d,k,H}\\
&\overset{D.5.1}{=} (H/\delta)^k |K^{\delta/2}_{\mathbb{R}^d,B}\, E_\Omega\, K^{\delta/2}_{\Omega,B} u|_{\mathbb{R}^d,k,\delta}\\
&\overset{L.5.5}{\lesssim} (H/\delta)^k \|E_\Omega\, K^{\delta/2}_{\Omega,B} u\|_{\mathbb{R}^d,k,\delta}\\
&\lesssim (H/\delta)^k \|K^{\delta/2}_{\Omega,B} u\|_{\Omega,k,\delta}\\
&\overset{L.5.5}{\lesssim} (H/\delta)^k \|u\|_{\Omega\cap B^{\delta/2},k,\delta}\\
&\lesssim \sigma_{\mathrm{Cacc}} (H/\delta)^k \|u\|_{\Omega\cap B^{\delta},k-1,\delta}\\
&\overset{L.5.2}{\lesssim} (\sigma_{\mathrm{Cacc}}/2)(H/\delta)^k \max\{1,\delta/H\}^{k-1} \|u\|_{\Omega\cap B^{\delta},k-1,H}.
\end{aligned}$$

Now, denote by $C_0 := C(d,k,\Omega,\sigma_{\mathrm{sprd}}) \ge 1$ the implicit cumulative constant. Then, the choice

$$H := \frac{\delta}{C_0 \sigma_{\mathrm{Cacc}}} > 0$$

guarantees that

$$C_0(\sigma_{\mathrm{Cacc}}/2)(H/\delta)^k \max\{1,\delta/H\}^{k-1} = \frac{C_0\sigma_{\mathrm{Cacc}}}{2(C_0\sigma_{\mathrm{Cacc}})^k} \max\{1, C_0\sigma_{\mathrm{Cacc}}\}^{k-1} = \frac{1}{2}$$

and the asserted error bound follows:

$$\|u - Q^\delta_{B,D}u\|_{\Omega\cap B,k,H} \le \frac{1}{2}\|u\|_{\Omega\cap B^\delta,k-1,H} \le \frac{1}{2}\|u\|_{\Omega\cap B^\delta,k,H}.$$

Finally, we turn our attention to the rank bound. We compute

$$\begin{aligned}
\mathrm{rank}(Q^\delta_{B,D}) &= \dim\{E_{B,D}\, P^H_{B,D}\, R_\Omega\, J^H\, K^{\delta/2}_{\mathbb{R}^d,B}\, E_\Omega\, K^{\delta/2}_{\Omega,B}\, u \,|\, u \in V_{\mathrm{sol}}(D)\}\\
&\overset{L.5.5}{\le} \dim\{J^H v \,|\, v \in H^k(\mathbb{R}^d) \text{ with } \mathrm{supp}(v) \subseteq B^{\delta/2}\} \overset{L.5.10}{\lesssim} (1 + \mathrm{diam}_2(B^{\delta/2})/H)^d\\
&\overset{L.2.13}{\lesssim} (1 + \mathrm{diam}_2(B)/H + \delta/H)^d \overset{\mathrm{Def.}H}{\lesssim} \sigma^d_{\mathrm{Cacc}}(1 + \mathrm{diam}_2(B)/\delta)^d.
\end{aligned}$$

This concludes the proof.

□

5.8 The multi-step coarsening operator $Q^{\delta,L}_{B,D}$

The hardest part now lies behind us and we can proceed with the plan from Section 5.1. The next step is to combine $L \in \mathbb{N}$ instances of the single-step coarsening operator $Q^\delta_{B,D}$ to a *multi-step coarsening operator* $Q^{\delta,L}_{B,D}$.

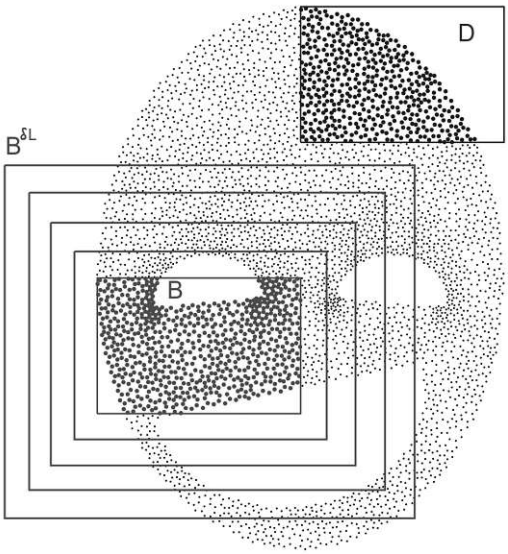

Figure 5.1: A nested sequence of inflated boxes, $B \subseteq B^\delta \subseteq \cdots \subseteq B^{\delta L}$.

Theorem 5.16. *Denote by $\sigma_{\mathrm{sprd}} > 0$ and $\sigma_{\mathrm{Cacc}} \geq 1$ the constants from A.4.11 and A.4.19, respectively. Let $B, D \in \mathbb{B}$ be given boxes and let $\delta > 0$, $L \in \mathbb{N}$ be free parameters with $\delta \leq \sigma_{\mathrm{sprd}}$ and $B^{\delta L} \cap D = \emptyset$. Then, there exists a linear* multi-step coarsening operator

$$Q^{\delta,L}_{B,D} : V_{\mathrm{sol}}(D) \longrightarrow V_{\mathrm{sol}}(D)$$

with the following properties:

1. *There holds the rank bound*

$$\mathrm{rank}(Q^{\delta,L}_{B,D}) \leq C(d,k,\Omega,\sigma_{\mathrm{sprd}})\sigma^d_{\mathrm{Cacc}}(L + \mathrm{diam}_2(B)/\delta)^{d+1}.$$

2. *For all $m \in \{0,\dots,k\}$ and all $u \in V_{\mathrm{sol}}(D)$, there holds the error bound*

$$\|u - Q^{\delta,L}_{B,D}u\|_{H^m(\Omega\cap B)} \leq C(d,k,\Omega,\sigma_{\mathrm{sprd}})(\sigma_{\mathrm{Cacc}}/\delta)^m 2^{-L}\|u\|_{H^k(\Omega\cap B^{\delta L})}.$$

Proof. Consider the following sequence of nested, inflated boxes (cf. Figure 5.1):

$$B = B^{\delta\cdot 0} \subseteq B^{\delta\cdot 1} \subseteq \cdots \subseteq B^{\delta l} \subseteq \cdots \subseteq B^{\delta L}.$$

Clearly, for each $l \in \{0,\dots,L-1\}$, there holds $(B^{\delta l})^\delta \cap D = B^{\delta(l+1)} \cap D \subseteq B^{\delta L} \cap D = \emptyset$. In particular, we may apply T.5.15 to the boxes $B^{\delta l}$, D and the parameter δ. For the corresponding single-step coarsening operators, we abbreviate

$$Q_l := Q^\delta_{B^{\delta l},D} : V_{\mathrm{sol}}(D) \longrightarrow V_{\mathrm{sol}}(D).$$

The definition of $Q^{\delta,L}_{B,D}$ is such that the subsequent error analysis becomes very simple:

$$\forall u \in V_{\mathrm{sol}}(D): \qquad Q^{\delta,L}_{B,D}u := u - (\mathrm{id} - Q_0)\circ\cdots\circ(\mathrm{id} - Q_{L-1})(u) \in V_{\mathrm{sol}}(D).$$

Denote by $H := C(d,k,\Omega,\sigma_{\mathrm{sprd}})^{-1}\sigma_{\mathrm{Cacc}}^{-1}\delta$ the number from the error bound in T.5.15. Then,

$$\begin{aligned}\|u - Q^{\delta,L}_{B,D}u\|_{\Omega\cap B,k,H} &= \|(\mathrm{id}-Q_0)\circ\cdots\circ(\mathrm{id}-Q_{L-1})(u)\|_{\Omega\cap B^{\delta\cdot 0},k,H}\\ &\le 2^{-1}\|(\mathrm{id}-Q_1)\circ\cdots\circ(\mathrm{id}-Q_{L-1})(u)\|_{\Omega\cap B^{\delta\cdot 1},k,H}\\ &\le 2^{-2}\|(\mathrm{id}-Q_2)\circ\cdots\circ(\mathrm{id}-Q_{L-1})(u)\|_{\Omega\cap B^{\delta\cdot 2},k,H}\\ &\vdots\\ &\le 2^{-L}\|u\|_{\Omega\cap B^{\delta L},k,H}.\end{aligned}$$

Using the norm equivalences from L.5.2, we also get an estimate in the standard H^m-norms, $m\in\{0,\dots,k\}$:

$$\|u - Q^{\delta,L}_{B,D}u\|_{H^m(\Omega\cap B)} \le \frac{\max\{1,H\}^k}{\min\{1,H\}^m}2^{-L}\|u\|_{H^k(\Omega\cap B^{\delta L})}.$$

Since $H\le\delta\le\sigma_{\mathrm{sprd}}$, we can plug in $\max\{1,H\}\le\sigma_{\mathrm{sprd}}$ and $\min\{1,H\}\ge H/\sigma_{\mathrm{sprd}}$, so that

$$\|u - Q^{\delta,L}_{B,D}u\|_{H^m(\Omega\cap B)} \le \sigma_{\mathrm{sprd}}^{k+m}H^{-m}2^{-L}\|u\|_{H^k(\Omega\cap B^{\delta L})} \overset{\text{Def.}H}{\lesssim} (\sigma_{\mathrm{Cacc}}/\delta)^m 2^{-L}\|u\|_{H^k(\Omega\cap B^{\delta L})}.$$

Finally, let us derive the rank bound. By induction on L, one can easily derive the following alternative representation of the operator $Q^{\delta,L}_{B,D}$:

$$Q^{\delta,L}_{B,D}u = \sum_{l=0}^{L-1} Q_l(\mathrm{id}-Q_{l+1})\circ\cdots\circ(\mathrm{id}-Q_{L-1})u.$$

(Note that the $(L-1)$-th summand is just $Q_{L-1}u$.) From this identity, we get

$$\begin{aligned}\mathrm{rank}(Q^{\delta,L}_{B,D}) &\le \sum_{l=0}^{L-1}\mathrm{rank}(Q_l) \overset{T.5.15}{\lesssim} \sigma_{\mathrm{Cacc}}^d\sum_{l=0}^{L-1}(1+\mathrm{diam}_2(B^{\delta l})/\delta)^d\\ &\overset{L.2.13}{\lesssim} \sigma_{\mathrm{Cacc}}^d\sum_{l=0}^{L-1}(1+\mathrm{diam}_2(B)/\delta+l)^d \le \sigma_{\mathrm{Cacc}}^d(L+\mathrm{diam}_2(B)/\delta)^{d+1}.\end{aligned}$$

This concludes the proof. □

5.9 The approximation operator $Q^r_{B,D}$

In this section, we complete the proof of T.4.20. We remind the reader of the relevant objects from the statement of T.4.20:

1. Denote by $\sigma_{\mathrm{sprd}}\ge 1$ and $\sigma_{\mathrm{Cacc}}\ge 1$ the constants from A.4.11 and A.4.19, respectively.

2. Let $h_{\min} > 0$ and $\sigma_{\mathrm{adm}} > 0$ be given numbers[9].
3. Let $B, D \in \mathbb{B}$ be given boxes that satisfy the following bounds:
$$h_{\min} \leq \operatorname{diam}_2(B) \leq \sigma_{\mathrm{adm}} \operatorname{dist}_2(B, D) \leq \sigma_{\mathrm{adm}} \sqrt{d} \sigma_{\mathrm{sprd}}.$$
4. Let $r \in \mathbb{N}$.

Proof of T.4.4.20. We employ the operator $Q^{\delta,L}_{B,D}$ from T.5.16 for specific values of $\delta > 0$ and $L \in \mathbb{N}$. We leave L unspecified for the moment and fix δ in relation to L:
$$\delta := \frac{\operatorname{dist}_2(B, D)}{2\sqrt{d}L} > 0.$$
The asserted bounds on B and D guarantee that
$$\delta = \frac{\operatorname{dist}_2(B, D)}{2\sqrt{d}L} \leq \frac{\sqrt{d}\sigma_{\mathrm{sprd}}}{2\sqrt{d}L} \leq \sigma_{\mathrm{sprd}}$$
and also the disjointness of $B^{\delta L}$ and D:
$$\operatorname{dist}_2(B^{\delta L}, D) \overset{L.2.2.13}{\geq} \operatorname{dist}_2(B, D) - \sqrt{d}\delta L \overset{\text{Def.}\delta}{=} \operatorname{dist}_2(B, D)/2 \geq h_{\min}/(2\sigma_{\mathrm{adm}}) > 0.$$
Therefore, the assumptions of T.5.16 are fulfilled and we may use the operator
$$Q^r_{B,D} := Q^{\delta,L}_{B,D} : V_{\mathrm{sol}}(D) \longrightarrow V_{\mathrm{sol}}(D).$$
As for the rank, we have
$$\begin{aligned}\operatorname{rank}(Q^r_{B,D}) &\overset{T.5.16}{\lesssim} \sigma^d_{\mathrm{Cacc}}(L + \operatorname{diam}_2(B)/\delta)^{d+1} \overset{\text{Def.}\delta}{=} \sigma^d_{\mathrm{Cacc}}\left(L + \frac{\operatorname{diam}_2(B)}{\operatorname{dist}_2(B, D)} 2\sqrt{d}L\right)^{d+1} \\ &\lesssim \sigma^d_{\mathrm{Cacc}}((1 + \sigma_{\mathrm{adm}})L)^{d+1} \leq (\sigma_{\mathrm{Cacc}}(1 + \sigma_{\mathrm{adm}})L)^{d+1}.\end{aligned}$$
Now, denote the implicit cumulative constant by $C_0 := C(d, k, \Omega, \sigma_{\mathrm{sprd}}) \geq 1$. Then, the choice ($\lfloor\cdot\rfloor$ is the floor function)
$$L := \left\lfloor \frac{(r/C_0)^{1/(d+1)}}{\sigma_{\mathrm{Cacc}}(1 + \sigma_{\mathrm{adm}})} \right\rfloor \in \mathbb{N}$$
leads to the desired rank bound:
$$\operatorname{rank}(Q^r_{B,D}) \overset{\text{Def.}C_0}{\leq} C_0(\sigma_{\mathrm{Cacc}}(1 + \sigma_{\mathrm{adm}})L)^{d+1} \overset{\text{Def.}L}{\leq} r.$$
Finally, let us derive the error bound. Let $m \in \{0, \ldots, k\}$ and $u \in V_{\mathrm{sol}}(D)$. Using the assumptions on B and D once again, we get
$$\begin{aligned}\|u - Q^r_{B,D}u\|_{H^m(\Omega\cap B)} &\overset{T.5.16}{\lesssim} (\sigma_{\mathrm{Cacc}}/\delta)^m 2^{-L}\|u\|_{H^k(\Omega)} \overset{\text{Def.}\delta}{=} \left(\frac{\sigma_{\mathrm{Cacc}} 2\sqrt{d}L}{\operatorname{dist}_2(B, D)}\right)^m 2^{-L}\|u\|_{H^k(\Omega)} \\ &\lesssim (\sigma_{\mathrm{Cacc}}\sigma_{\mathrm{adm}}/h_{\min})^m L^k 2^{-L}\|u\|_{H^k(\Omega)}.\end{aligned}$$

[9] In the statement of T.4.20, these numbers have a specific meaning. Here, we only need them to be positive.

Sacrificing a small fraction of 2^{-L}, we can get rid of the distracting factor L^k. We use the following elementary relations:

$$\varepsilon := \ln(2) - 1/2 \approx 0.19, \qquad \sup_{t\in[0,\infty)} \frac{t^k}{e^{\varepsilon t}} = \left(\frac{k}{e\varepsilon}\right)^k \lesssim 1, \qquad L \geq \frac{(r/C_0)^{1/(d+1)}}{2\sigma_{\mathrm{Cacc}}(1+\sigma_{\mathrm{adm}})}.$$

Then,

$$\frac{L^k}{2^L} = \frac{L^k}{e^{\varepsilon L}} \exp(-L/2) \lesssim \exp(-L/2) \leq \exp\left(-\frac{(r/C_0)^{1/(d+1)}}{4\sigma_{\mathrm{Cacc}}(1+\sigma_{\mathrm{adm}})}\right).$$

Finally, by setting

$$\sigma_{\mathrm{exp}} := (4\sigma_{\mathrm{Cacc}}(1+\sigma_{\mathrm{adm}})C_0^{1/(d+1)})^{-1} > 0,$$

we obtain the overall bound

$$\|u - Q^r_{B,D}u\|_{H^m(\Omega\cap B)} \lesssim (\sigma_{\mathrm{Cacc}}\sigma_{\mathrm{adm}}/h_{\mathrm{min}})^m \exp(-\sigma_{\mathrm{exp}} r^{1/(d+1)})\|u\|_{H^k(\Omega)}.$$

This finishes the proof.

□

6 An application in a FEM setting

The *finite element method (FEM)* is a well-established (e.g., [Cia78], [EG04], [BS08], [Bra13], [LB13]) approximation scheme for partial differential equations. Among the numerous advantages over other numerical schemes (e.g., finite differences method) is the fact that it allows for a rigourous mathematical analysis. Here, we look at the analyst's favourite example of a second order elliptic PDE with homogeneous Dirichlet boundary conditions. We apply our main result, T.4.21, to the inverse stiffness matrix $\boldsymbol{A}^{-1}$.

For the reader's convenience, we summarize the necessary steps:

1. Choose the space V from D.4.1 and the bilinear form $a(\cdot,\cdot)$ from D.4.2. Find an upper bound for the quantity $\sigma_{\mathrm{coco}} \geq 1$.
2. Choose the ansatz space $V_N \subseteq V$ and the basis $\{\varphi_1, \dots, \varphi_N\} \subseteq V_N$ from D.4.3.
3. Prove that the dual basis $\{\lambda_1, \dots, \lambda_N\} \subseteq V_N^*$ satisfies a *local* stability bound as required by A.4.11 and determine the corresponding values of $k_0 \in \{0, \dots, k\}$ and $\sigma_{\mathrm{stab}} > 0$. Determine the shape-regularity-, overlap- and spread factors σ_{shp}, σ_{ovlp}, σ_{sprd} of the characteristic sets $\Omega_1, \dots, \Omega_N \subseteq \overline{\Omega}$. Furthermore, estimate the number $h_{\min} > 0$ from D.4.12.
4. Choose the clustering parameter $\sigma_{\mathrm{small}} \geq 1$ from C. 3.42 large enough to ensure $\sigma_{\mathrm{ovlp}} \leq \sigma_{\mathrm{small}}$.
5. Prove the discrete Caccioppoli inequality from A.4.19.

6.1 An elliptic PDE

Let $d \in \{1, 2, 3\}$ and $\Omega \subseteq \mathbb{R}^d$ be a polyhedron (cf. D. 2.62). Furthermore, let $a_1 \in L^\infty(\Omega, \mathbb{R}^{d\times d})$, $a_2 \in L^\infty(\Omega, \mathbb{R}^d)$ and $a_3 \in L^\infty(\Omega, \mathbb{R})$ be given coefficient functions and $f_\Omega \in L^2(\Omega)$ be a given right-hand side. We seek a weak solution $u \in H_0^1(\Omega)$ to the following elliptic PDE[1]:

$$\begin{aligned} -\mathrm{div}(a_1 \nabla u) + a_2 \cdot \nabla u + a_3 u &= f_\Omega \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

We assume that a_1 is pointwise symmetric and that there exist constants $\alpha_1 > 0$ and $\alpha_2, \alpha_3 \geq 0$ such that, for all $x \in \Omega$ and $y \in \mathbb{R}^d$, the following relations are satisfied:

$$\begin{aligned} \alpha_1^{-1}\|y\|_2^2 &\leq \langle a_1(x)y, y\rangle_2, & \|a_1(x)\|_2 &\leq \alpha_1, \\ \|a_2(x)\|_2 &\leq \alpha_2, & |a_3(x)| &\leq \alpha_3, \\ 2C_P^2\alpha_1(\alpha_2+\alpha_3) &\leq 1, \end{aligned}$$

[1] Note that $a_1\nabla u$ is a matrix-vector product, $a_2 \cdot \nabla u$ is a dot product and $a_3 u$ is just a multiplication.

Here, $C_P := C(d, \Omega) \geq 0$ is the constant in the Poincaré inequality $\|\cdot\|_{H^1(\Omega)} \leq C_P |\cdot|_{H^1(\Omega)}$ on $H^1_0(\Omega)$ (cf. C.2.53).

6.2 The space V

First, we fix the functional analytic setting for this problem.

Definition 6.1. *Let* $k := 1$ *and*

$$V := H^1_0(\Omega).$$

For all $u, v \in V$, *set*[2]

$$a(u, v) := \langle a_1 \nabla u, \nabla v \rangle_{L^2(\Omega)} + \langle a_2 \cdot \nabla u, v \rangle_{L^2(\Omega)} + \langle a_3 u, v \rangle_{L^2(\Omega)}.$$

Lemma 6.2. *1. The set* Ω *is an* H^1*-extension domain and the space* V *satisfies the requirements from D.4.1.*

2. The bilinear form $a(\cdot, \cdot)$ *is continuous and coercive in the sense of D. 4.2 with a constant*

$$\sigma_{\text{coco}} = C(d, \Omega, \alpha_1, \alpha_2, \alpha_3).$$

Proof. Item 1: According to D.2.62 and L.2.49, the polyhedron Ω is a H^1-extension domain. Furthermore, V is a closed subspace of $H^1(\Omega)$ and, for all $\kappa \in C^\infty_0(\mathbb{R}^d)$ and $v \in V$, there holds $(\kappa|_\Omega)v \in V$.

Item 2: For all $u, v \in V$, we have

$$\begin{aligned} |a(u, v)| &\leq \int_\Omega |\langle a_1 \nabla u, \nabla v \rangle_2| + |(a_2 \cdot \nabla u) v| + |a_3 u v| \, \mathrm{d}x \\ &\leq \int_\Omega \alpha_1 \|\nabla u\|_2 \|\nabla v\|_2 + \alpha_2 \|\nabla u\|_2 |v| + \alpha_3 |u||v| \, \mathrm{d}x \lesssim \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}. \end{aligned}$$

On the other hand, using the Cauchy-Schwarz- and the Poincaré inequality,

$$\alpha_1^{-1} |u|^2_{H^1(\Omega)} \leq \int_\Omega \langle a_1 \nabla u, \nabla u \rangle_2 \, \mathrm{d}x = a(u, u) - \int_\Omega \langle a_2, \nabla u \rangle_2 u + a_3 u^2 \, \mathrm{d}x$$

$$\leq a(u, u) + \int_\Omega \alpha_2 \|\nabla u\|_2 |u| + \alpha_3 u^2 \, \mathrm{d}x \leq a(u, u) + \alpha_2 |u|_{H^1(\Omega)} \|u\|_{L^2(\Omega)} + \alpha_3 \|u\|^2_{L^2(\Omega)}$$

$$\leq a(u, u) + C_P^2 (\alpha_2 + \alpha_3) |u|^2_{H^1(\Omega)} \leq a(u, u) + (2\alpha_1)^{-1} |u|^2_{H^1(\Omega)}.$$

If we subtract the last term from both sides (and apply the Poincaré inequality again) we end up with the asserted coercivity bound:

$$(2C_P^2 \alpha_1)^{-1} \|u\|^2_{H^1(\Omega)} \leq (2\alpha_1)^{-1} |u|^2_{H^1(\Omega)} \leq a(u, u).$$

□

[2] In this chapter, we use the abbreviation $\langle F, G \rangle_{L^2(\Omega)} := \int_\Omega \langle F, G \rangle_2 \, \mathrm{d}x$ for vector-valued functions $F, G : \Omega \longrightarrow \mathbb{R}^d$.

6.3 The space V_N

Next, we fix the discrete ansatz space V_N and the basis functions φ_n. For this purpose, we use the spline spaces $\mathbb{S}_0^{p,1}(\mathcal{T})$ from D.2.70. Our assumptions on the basis functions $\{\varphi_1, \dots, \varphi_N\}$ reflect a common pattern in the construction of finite element bases: First, a polynomial basis $\{\hat{\varphi}_1, \dots, \hat{\varphi}_L\}$ on the reference element $\hat{T}$ is chosen (the *shape functions*). Then, these polynomials are transferred to the individual mesh elements $T \in \mathcal{T}$ via the affine transformations $F_T : \hat{T} \longrightarrow T$ (cf. D.2.65). Finally, these *element shape functions* are glued together along the element interfaces to construct the global basis functions φ_n. We mention that the classic *hat functions* from L.2.72 ($p = 1$) fall into this category along with the more general *Lagrange elements* ($p \geq 1$).

Definition 6.3. *Let $\mathcal{T} \subseteq \mathrm{Pow}(\mathbb{R}^d)$ be a mesh (cf. D.2.60) on the polyhedron Ω. Furthermore, let $p \in \mathbb{N}$, $L := \binom{p+d}{d}$ and let $\{\hat{\varphi}_1, \dots, \hat{\varphi}_L\} \subseteq \mathbb{P}^p(\hat{T})$ be a basis. We set $N := \dim(\mathbb{S}_0^{p,1}(\mathcal{T}))$ and consider the ansatz space*

$$V_N := \mathbb{S}_0^{p,1}(\mathcal{T}) \subseteq V.$$

A basis

$$\{\varphi_1, \dots, \varphi_N\} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T})$$

is called FEM basis*, if the following conditions are satisfied:*

1. Local supports: *For every $n \in \{1, \dots, N\}$, there exists a* characteristic element *$T_n \in \mathcal{T}$ such that*[3]
$$T_n \in \mathrm{supp}_{\mathcal{T}}(\varphi_n) \subseteq \mathcal{T}(T_n).$$

2. Simple structure: *For every $n \in \{1, \dots, N\}$ and every $T \in \mathrm{supp}_{\mathcal{T}}(\varphi_n)$, there exists an index $l(n, T) \in \{1, \dots, L\}$ such that*
$$\varphi_n|_T = \hat{\varphi}_{l(n,T)} \circ F_T^{-1}.$$

3. Local distinctness: *For all $m, n \in \{1, \dots, N\}$ with $m \neq n$ and all $T \in \mathrm{supp}_{\mathcal{T}}(\varphi_n) \cap \mathrm{supp}_{\mathcal{T}}(\varphi_m)$, there holds*
$$l(n, T) \neq l(m, T).$$

4. Stability: *There exists a constant $\gamma \geq 0$, such that, for all $c \in \mathbb{R}^L$, there holds the bound*
$$\|c\|_2 \leq C(d)p^\gamma \left\| \sum_{l=1}^{L} c_l \hat{\varphi}_l \right\|_{L^2(\hat{T})}.$$

Example 6.4. *See [FMPR15, Lemma 4.4.] for an example of shape functions $\{\hat{\varphi}_1, \dots, \hat{\varphi}_L\}$ in $d = 2$ space dimensions with $\gamma = 3$.*

In the sequel, we assume that a FEM basis $\{\varphi_1, \dots, \varphi_N\} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T})$ is given and that the characteristic elements $T_n \in \mathcal{T}$ from item 1 are kept fixed.

[3] Recall that $\mathrm{supp}_{\mathcal{T}}(v) \subseteq \mathcal{T}$ is the discrete support of a discrete function $v \in \mathbb{S}^{q,0}(\mathcal{T})$ (cf. D.2.71) and that $\mathcal{T}(T_n)$ denotes the elements touching T_n (cf. D.2.63).

6.4 A weak formulation

As usual, a weak formulation of the model problem from Section 6.1 is derived by multiplication with $v \in V_N$ and integration over Ω. Using the bilinear form $a(\cdot,\cdot)$ from D.6.1, the left-hand side of the equation can be expressed as $a(u,v)$. The right-hand side can be written in the form of a linear functional $f \in V_N^*$:

$$\forall v \in V_N : \qquad \langle f, v\rangle_* := \langle f_\Omega, v\rangle_{L^2(\Omega)}.$$

In particular, the model problem from Section 6.1 falls into the problem class described in P.1.2 and L.4.5: Find $u \in V_N$ such that

$$\forall v \in V_N : \qquad a(u,v) = \langle f, v\rangle_*.$$

6.5 The dual basis $\lambda_1, \dots, \lambda_N$

In this section, we derive a *local* stability bound for the dual basis $\lambda_1, \dots, \lambda_N \in V_N^*$ from D.4.10. Furthermore, we determine the values of the quantities σ_{shp}, σ_{ovlp} and σ_{sprd} from A.4.11.

Definition 6.5. *Denote by $T_1, \dots, T_N \in \mathcal{T}$ the characteristic mesh elements from D.6.3. For all $n \in \{1, \dots, N\}$, we set*

$$\Omega_n := T_n.$$

Lemma 6.6. *Denote by $h_{\min,\mathcal{T}} > 0$ and $h_{\min} > 0$ the quantities from D.2.63 and D.4.12, respectively. There holds the relationship*

$$h_{\min} \geq h_{\min,\mathcal{T}}.$$

Proof. We compute

$$h_{\min} \overset{D.4.12}{=} \min_{n\in\{1,\dots,N\}} h_{\Omega_n} = \min_{n\in\{1,\dots,N\}} h_{T_n} \geq \min_{T\in\mathcal{T}} h_T \overset{D.2.63}{=} h_{\min,\mathcal{T}}.$$

□

Next, we derive the local stability bound. The trick is to find a representation of the n-th dual functional $\lambda_n \in V_N^*$ in terms of a "density" $\mu_n \in L^2(\Omega)$.

Lemma 6.7. *Denote by $\sigma_{\text{shp}} \geq 1$ the shape regularity constant of the mesh $\mathcal{T}$ (cf. D.2.60) and let $\gamma \geq 0$ be defined as in D.6.3. Denote by $\{\lambda_1, \dots, \lambda_N\} \subseteq V_N^*$ the dual basis (cf. D.4.10). Then, for all $v \in V_N$, there holds the following local stability bound:*

$$|\langle \lambda_n, v\rangle_*| \leq C(d, \sigma_{\text{shp}}) p^\gamma h_{\min,\mathcal{T}}^{-d/2} \|v\|_{L^2(\Omega_n)}.$$

Proof. Denote by $F_T : \hat{T} \longrightarrow T$ the affine transformations from D.2.65. Furthermore, denote by $L \in \mathbb{N}$, $\{\hat{\varphi}_1, \dots, \hat{\varphi}_L\} \subseteq \mathbb{P}^p(\hat{T})$, $T_n \in \mathcal{T}$ and $l(n,T) \in \{1, \dots, L\}$ the objects from D.6.3.

Let $\{\hat{\mu}_1, \dots, \hat{\mu}_L\} \subseteq \mathbb{P}^p(\hat{T})$ be the basis of *dual polynomials* which is uniquely determined by the following conditions:

$$\forall k, l \in \{1, \dots, L\} : \qquad \langle \hat{\mu}_k, \hat{\varphi}_l \rangle_{L^2(\hat{T})} = \delta_{kl}.$$

Now, for each $n \in \{1, \dots, N\}$, we define a (discontinuous) function $\mu_n \in \mathbb{S}^{p,0}(\mathcal{T})$ in a piecewise manner: For every $T \in \mathcal{T} \backslash \{T_n\}$, we set $\mu_n|_T := 0$, whereas

$$\mu_n|_{T_n} := |\det(\nabla F_{T_n})|^{-1} \cdot (\hat{\mu}_{l(n,T_n)} \circ F_{T_n}^{-1}).$$

We use the function μ_n as a density to define a linear functional $\tilde{\lambda}_n \in V_N^*$:

$$\forall v \in V_N : \qquad \langle \tilde{\lambda}_n, v \rangle_* := \langle \mu_n, v \rangle_{L^2(\Omega)}.$$

If we can show that $\langle \tilde{\lambda}_n, \varphi_m \rangle_* = \delta_{nm}$, for all $n, m \in \{1, \dots, N\}$, then already $\tilde{\lambda}_n = \lambda_n$ by the uniqueness of the dual basis (cf. D.4.10). To this end, let $n, m \in \{1, \dots, N\}$. First, consider the case where $T_n \notin \operatorname{supp}_{\mathcal{T}}(\varphi_m)$. Then, the fact that $T_n \in \operatorname{supp}_{\mathcal{T}}(\varphi_n)$ from D.6.3 implies that there must hold $m \neq n$. It follows that $\delta_{nm} = 0$ and thus

$$\langle \lambda_n, \varphi_m \rangle_* = \langle \mu_n, \varphi_m \rangle_{L^2(T_n \cap \operatorname{supp}(\varphi_m))} = 0 = \delta_{nm}.$$

In the remaining case $T_n \in \operatorname{supp}_{\mathcal{T}}(\varphi_m)$, it follows from the structure assumption in D.6.3 that there exists an index $l(m, T_n) \in \{1, \dots, L\}$ such that

$$\varphi_m|_{T_n} = \hat{\varphi}_{l(m,T_n)} \circ F_{T_n}^{-1}.$$

We argue that there holds the identity $\delta_{l(n,T_n)l(m,T_n)} = \delta_{nm}$: If $n = m$, then both sides yield the value 1. On the other hand, if $n \neq m$, then the fact that $T_n \in \operatorname{supp}_{\mathcal{T}}(\varphi_n) \cap \operatorname{supp}_{\mathcal{T}}(\varphi_m)$ and the asserted local distinctness from D.6.3 yield $l(n, T_n) \neq l(m, T_n)$, so that both sides of the equation become 0. Now, using the definition of μ_n, we obtain

$$\begin{aligned} \langle \lambda_n, \varphi_m \rangle_* &= \langle \mu_n, \varphi_m \rangle_{L^2(T_n)} = |\det(\nabla F_{T_n})|^{-1} \langle \hat{\mu}_{l(n,T_n)} \circ F_{T_n}^{-1}, \hat{\varphi}_{l(m,T_n)} \circ F_{T_n}^{-1} \rangle_{L^2(T_n)} \\ &= \langle \hat{\mu}_{l(n,T_n)}, \hat{\varphi}_{l(m,T_n)} \rangle_{L^2(\hat{T})} = \delta_{l(n,T_n)l(m,T_n)} = \delta_{nm}. \end{aligned}$$

It follows that, indeed $\tilde{\lambda}_n = \lambda_n$. Then, for all $v \in V_N$, we get

$$|\langle \lambda_n, v \rangle_*| = |\langle \tilde{\lambda}_n, v \rangle_*| = |\langle \mu_n, v \rangle_{L^2(T_n)}| \leq \|\mu_n\|_{L^2(T_n)} \|v\|_{L^2(T_n)}.$$

It remains to bound the norm of μ_n. To this end, we first need a bound for the dual polynomials $\hat{\mu}_1, \dots, \hat{\mu}_L$. Expanding the k-th polynomial in the form $\hat{\mu}_k = \sum_{l=1}^L c_l \hat{\varphi}_l$ (for some $c \in \mathbb{R}^L$), we get

$$\|\hat{\mu}_k\|_{L^2(\hat{T})}^2 = \left\langle \hat{\mu}_k, \sum_{l=1}^L c_l \hat{\varphi}_l \right\rangle_{L^2(\hat{T})} = \sum_{l=1}^L c_l \langle \hat{\mu}_k, \hat{\varphi}_l \rangle_{L^2(\hat{T})} = c_k \leq \|c\|_2 \overset{D.6.3}{\lesssim} p^\gamma \|\hat{\mu}_k\|_{L^2(\hat{T})}.$$

Then,

$$\begin{aligned} \|\mu_n\|_{L^2(T_n)} &= |\det(\nabla F_{T_n})|^{-1} \|\hat{\mu}_{l(n,T_n)} \circ F_{T_n}^{-1}\|_{L^2(T_n)} \\ &\overset{L.2.33}{=} |\det(\nabla F_{T_n})|^{-1/2} \|\hat{\mu}_{l(n,T_n)}\|_{L^2(\hat{T})} \overset{L.2.24}{\lesssim} p^\gamma h_{T_n}^{-d/2} \leq p^\gamma h_{\min,\mathcal{T}}^{-d/2}. \end{aligned}$$

This concludes the proof.

□

In the last part of this section, we determine the values of the quantities $\sigma_{\text{shp}}, \sigma_{\text{ovlp}}, \sigma_{\text{sprd}}$ as required by A.4.11. For the definition of these numbers, see D.2.16, D.2.18 and D.2.21.

Lemma 6.8. *Denote by $\sigma_{\text{shp}} \geq 1$ the shape-regularity constant of the mesh $\mathcal{T}$ (cf. D.2.60). The characteristic sets $\Omega_1, \ldots, \Omega_N \subseteq \overline{\Omega}$ from D.6.5 have shape-regularity σ_{shp}, overlap σ_{ovlp} and spread σ_{sprd}, where*

$$\sigma_{\text{ovlp}} = \binom{d+p}{d}, \qquad \sigma_{\text{sprd}} = C(\Omega).$$

Proof. Since the sets Ω_n are mesh elements, they are clearly shape-regular with respect to the shape factor σ_{shp} of the mesh $\mathcal{T}$. The fact that the polyhedron Ω is by definition bounded (cf. D.2.62) immediately yields a uniformly bounded spread factor σ_{sprd}:

$$\operatorname{diam}_2\Big(\bigcup_{n=1}^{N} \Omega_n\Big) \subseteq \operatorname{diam}_2\Big(\bigcup_{T\in\mathcal{T}} T\Big) = \operatorname{diam}_2(\Omega) < \infty.$$

The overlap factor σ_{ovlp} requires a bit more work: According to D.2.18, we need to find an upper bound for the quantity

$$\max_{n\in\{1,\ldots,N\}} \#\{m \in \{1,\ldots,N\} \,|\, \Omega_m^\circ \cap \Omega_n^\circ \neq \emptyset\}.$$

To this end, recall from L.2.64 that distinct mesh elements can only intersect at their boundaries. In particular, the condition $\Omega_m^\circ \cap \Omega_n^\circ \neq \emptyset$ is satisfied, if and only if $T_m = T_n$. Therefore, it suffices to determine a bound for the quantity $\max_{T\in\mathcal{T}} \#ms(T)$, where

$$ms(T) := \{m \in \{1,\ldots,N\} \,|\, T_m = T\}.$$

Recall from the proof of L.6.7 that the dual functionals $\lambda_1, \ldots, \lambda_N \in V_N^*$ can be represented by densities $\mu_1, \ldots, \mu_N \in \mathbb{S}^{p,0}(\mathcal{T})$. Since the functionals λ_n are linearly independent, it is not difficult to see that the densities μ_n are linearly independent as well (as functions on all of Ω). Then, given an arbitrary element $T \in \mathcal{T}$, the restrictions

$$\{\mu_m|_T \,|\, m \in ms(T)\} \subseteq \mathbb{P}^p(T)$$

are again linearly independent (as functions on T). To see this, consider coefficients $(c_m)_{m\in ms(T)} \subseteq \mathbb{R}$ such that $\sum_{m\in ms(T)} c_m(\mu_m|_T) \equiv 0$ on T. Since $\operatorname{supp}(\mu_m) = T_m = T$, for all $m \in ms(T)$, we also have $\sum_{m\in ms(T)} c_m\mu_m \equiv 0$ on $\Omega\backslash T$. Combining both, we get $\sum_{m\in ms(T)} c_m\mu_m \equiv 0$ on all of Ω, so that necessarily $c_m = 0$, for all $m \in ms(T)$. Now that the linear independence of the polynomials $\{\mu_m|_T \,|\, m \in ms(T)\}$ is settled, we obtain

$$\#ms(T) = \dim(\operatorname{span}\{\mu_m|_T \,|\, m \in ms(T)\}) \leq \dim(\mathbb{P}^p(T)) = \binom{d+p}{d}.$$

This finishes the proof. □

6.6 The discrete Caccioppoli inequality

Now we come to the hardest part of this chapter. In this section, we show that the discrete Caccioppoli inequality from A.4.19 is satisfied for some constant

$$\sigma_{\mathrm{Cacc}} = C(d, \sigma_{\mathrm{shp}}, \sigma_{\mathrm{lqu}})p^{(9-d)/2},$$

where $\sigma_{\mathrm{shp}}, \sigma_{\mathrm{lqu}} \geq 1$ are the quantities from D.2.60. Let us quickly give an outline of the proof:

1. Let $D \in \mathbb{B}$ be an axes-parallel box (cf. D.2.8) and consider a function $u \in V_{\mathrm{sol}}(D)$. Furthermore, let $B \in \mathbb{B}$ and $\delta > 0$ be such that $B^\delta \cap D = \emptyset$, where $B^\delta \in \mathbb{B}$ is the inflated box (cf. D.2.12). Our goal is to show that there exists a number $\sigma_{\mathrm{Cacc}} \geq 1$ such that
$$\delta |u|_{H^1(\Omega \cap B)} \leq \sigma_{\mathrm{Cacc}} \|u\|_{L^2(\Omega \cap B^\delta)}.$$

2. To this end, we split the patch elements $\mathcal{T}(B) = \{T \in \mathcal{T} \,|\, T \cap B \neq \emptyset\}$ (cf. D.2.63) into two groups, based on the relative size of the element diameter h_T and the value of δ:
$$\begin{aligned} \mathcal{B}_{\mathrm{small}} &:= \{T \in \mathcal{T}(B) \,|\, 24\sigma_{\mathrm{shp}}\sigma_{\mathrm{lqu}} h_T \leq \delta\}, \\ \mathcal{B}_{\mathrm{large}} &:= \{T \in \mathcal{T}(B) \,|\, 24\sigma_{\mathrm{shp}}\sigma_{\mathrm{lqu}} h_T > \delta\}. \end{aligned}$$
Clearly,
$$\delta^2 |u|^2_{H^1(\Omega \cap B)} = \delta^2 \sum_{T \in \mathcal{T}(B)} |u|^2_{H^1(T \cap B)} = \delta^2 \sum_{T \in \mathcal{B}_{\mathrm{small}}} |u|^2_{H^1(T \cap B)} + \delta^2 \sum_{T \in \mathcal{B}_{\mathrm{large}}} |u|^2_{H^1(T \cap B)},$$
and we need to find bounds for both sums.

3. The large elements $T \in \mathcal{B}_{\mathrm{large}}$ with $T \subseteq B^\delta$ are easy to handle, because $\delta \lesssim h_T$, and the inverse inequality from L.2.54 already gives us
$$\delta |u|_{H^1(T \cap B)} \leq \delta |u|_{H^1(T)} \lesssim h_T |u|_{H^1(T)} \overset{L.2.54}{\lesssim} \|u\|_{L^2(T)} = \|u\|_{L^2(T \cap B^\delta)}.$$
However, there might be large elements $T \in \mathcal{B}_{\mathrm{large}}$ which are *not* fully contained in the inflated box B^δ. As a remedy, we first break T into smaller pieces $S \subseteq T$ (cf. L.2.69). Then, we apply the previous argument only to those pieces S which cover $T \cap B$ and lie inside of $T \cap B^\delta$.

4. While the large elements can be treated individually, the small elements $\mathcal{B}_{\mathrm{small}}$ need to be treated as a group. The proof is a fully discrete version of the continuous Caccioppoli inequality from L.2.55. There, we used a suitable *smooth* cut-off function $\kappa \in C^\infty(\Omega)$ and exploited the orthogonality $a(u, \kappa^2 u) = 0$, which relies on the fact that $\kappa^2 u \in H^1_0(\Omega)$ may be plugged into the variational equation that defines the function $u \in H^1_0(\Omega)$.

   Here, in the discrete setting, the fact that $\max\{h_T \,|\, T \in \mathcal{B}_{\mathrm{small}}\} \lesssim \delta$ allows us to employ a suitable *discrete* cut-off function $\kappa \in \mathbb{S}^{1,1}(\mathcal{T})$ (cf. L.2.77), which lives inside

a small neighbourhood of $\mathcal{B}_{\text{small}}$. However, the product $\kappa^2 u \in \mathbb{S}_0^{p+2,1}(\mathcal{T})$ is not an element of the FEM space $V_N = \mathbb{S}_0^{p,1}(\mathcal{T})$ and we cannot use it as a test function directly. Therefore, we take an interpolation operator

$$J_{\mathcal{T}}^p : H^2_{\text{pw}}(\mathcal{T}) \cap H^1_0(\Omega) \longrightarrow \mathbb{S}_0^{p,1}(\mathcal{T})$$

and use $J_{\mathcal{T}}^p(\kappa^2 u) \in \mathbb{S}_0^{p,1}(\mathcal{T})$ instead. As it turns out, the induced error $\kappa^2 u - J_{\mathcal{T}}^p(\kappa^2 u)$ does not pose a problem for the validity of the discrete Caccioppoli inequality.

We start with the large elements $\mathcal{B}_{\text{large}}$:

Lemma 6.9. *Let $u \in \mathbb{S}^{p,0}(\mathcal{T})$, $B \in \mathbb{B}$ and $\delta > 0$. Then, for all $T \in \mathcal{T}(B)$ with $24\sigma_{\text{shp}}\sigma_{\text{lqu}} h_T > \delta$, there holds the inequality*

$$\delta |u|_{H^1(T\cap B)} \leq C(d, \sigma_{\text{shp}}, \sigma_{\text{lqu}}) p^2 \|u\|_{L^2(T\cap B^\delta)}.$$

Proof. Let $T \in \mathcal{T}(B)$ with $24\sigma_{\text{shp}}\sigma_{\text{lqu}} h_T > \delta$. According to L.2.69, we can find a family $\mathcal{S} \subseteq \text{Pow}(\mathbb{R}^d)$ of simplices $S \subseteq \mathbb{R}^d$ with the following properties:

1. There holds $\bigcup_{S\in\mathcal{S}} S = T$.
2. Every $S \in \mathcal{T}$ is $\widetilde{\sigma_{\text{shp}}}$-shape regular, where $\widetilde{\sigma_{\text{shp}}} = C(d)\sigma_{\text{shp}}$.
3. There hold the bounds $h_{\max,\mathcal{S}} \leq \delta \leq C(d, \sigma_{\text{shp}}, \sigma_{\text{lqu}}) h_{\min,\mathcal{S}}$.

We only need the subsimplices that touch the set B,

$$\mathcal{S}(B) := \{S \in \mathcal{S} \,|\, S \cap B \neq \emptyset\}.$$

First, note that

$$T \cap B = \left(\bigcup_{S\in\mathcal{S}} S\right) \cap B = \bigcup_{S\in\mathcal{S}} (S \cap B) = \bigcup_{S\in\mathcal{S}(B)} (S \cap B) \subseteq \bigcup_{S\in\mathcal{S}(B)} S.$$

On the other hand, for every $S \in \mathcal{S}(B)$, we may pick a point $x_0 \in S \cap B$ and find the following relation:

$$\forall x \in S: \qquad \|x - x_0\|_2 \leq \text{diam}_2(S) = h_S \leq h_{\max,\mathcal{S}} \leq \delta.$$

Due to L.2.13, this implies $S \subseteq B^\delta$. In summary, we have

$$T \cap B \;\subseteq\; \bigcup_{S\in\mathcal{S}(B)} S \;\subseteq\; T \cap B^\delta.$$

Finally, using the inverse inequality from L.2.67, we compute

$$\begin{aligned} \delta^2 |u|^2_{H^1(T\cap B)} \quad &\lesssim \quad h^2_{\min,\mathcal{S}} |u|^2_{H^1(T\cap B)} \leq \sum_{S\in\mathcal{S}(B)} h_S^2 |u|^2_{H^1(S)} \\ &\overset{L.2.67}{\lesssim} \sum_{S\in\mathcal{S}(B)} p^4 \|u\|^2_{L^2(S)} \leq p^4 \|u\|^2_{L^2(T\cap B^\delta)}. \end{aligned}$$

Note that the constant from the inverse inequality depends on $\widetilde{\sigma_{\text{shp}}}$, which is bounded by $C(d)\sigma_{\text{shp}}$. This concludes the proof. □

Now that the large elements are taken care of, we turn our attention to the small elements $\mathcal{B}_{\text{small}}$. To this end, we introduce the interpolation operator $J^p_{\mathcal{T}}$ that was already mentioned in the overview paragraph at the beginning of this section.

Lemma 6.10. *Suppose $d \in \{1,2,3\}$. There exists a linear operator*[4]

$$J^p_{\mathcal{T}} : H^2_{\text{pw}}(\mathcal{T}) \longrightarrow \mathbb{S}^{p,0}(\mathcal{T})$$

with the following properties:

1. *For all $v \in H^2_{\text{pw}}(\mathcal{T}) \cap H^1_0(\Omega)$, there holds $J^p_{\mathcal{T}} v \in \mathbb{S}^{p,1}_0(\mathcal{T})$.*
2. *For all $q \in \mathbb{N}_0$ and all $v \in \mathbb{S}^{q,0}(\mathcal{T})$, there holds*[5]
$$\operatorname{supp}_{\mathcal{T}}(J^p_{\mathcal{T}} v) \subseteq \operatorname{supp}_{\mathcal{T}}(v).$$
3. *For all $v \in H^2_{\text{pw}}(\mathcal{T})$ and all $T \in \mathcal{T}$, there holds the error bound*
$$\sum_{l=0}^{1} h_T^l |v - J^p_{\mathcal{T}} v|_{H^l(T)} \leq C(d,\sigma_{\text{shp}}) p^{(1-d)/2} \inf_{w \in \mathbb{P}^p(T)} \sum_{l=0}^{2} h_T^l |v-w|_{H^l(T)}.$$

Proof. Denote by $\hat{T} \subseteq \mathbb{R}^d$ the reference element (cf. D. 2.59). In [MR20], under the assumption $d \in \{1,2,3\}$, an operator

$$\hat{J}^p : H^{(d+1)/2}(\hat{T}) \longrightarrow \mathbb{P}^p(\hat{T})$$

with the following properties was constructed:

1. For every $v \in H^{(d+1)/2}(\hat{T})$, every $k \in \{0,\dots,d-1\}$ and every k-simplex $\hat{\Gamma} \subseteq \mathbb{R}^d$ with $\mathcal{N}(\hat{\Gamma}) \subseteq \mathcal{N}(\hat{T})$ (cf. D.2.57), the value of $(\hat{J}^p v)|_{\hat{\Gamma}}$ is uniquely determined by $v|_{\hat{\Gamma}}$.
2. For every $v \in H^{(d+1)/2}(\hat{T})$, there holds the error bound
$$\|v - \hat{J}^p v\|_{H^1(\hat{T})} \leq C(d) p^{(1-d)/2} \inf_{w \in \mathbb{P}^p(\hat{T})} \|v-w\|_{H^{(d+1)/2}(\hat{T})}.$$

Now, for every $v \in H^2_{\text{pw}}(\mathcal{T})$, the image $J^p_{\mathcal{T}} v \in \mathbb{S}^{p,0}(\mathcal{T})$ is defined in a piecewise manner via the element transformations $F_T : \hat{T} \longrightarrow T$ from D.2.65:

$$(J^p_{\mathcal{T}} v)|_T := \hat{J}^p (v \circ F_T) \circ F_T^{-1}.$$

The preservation of global continuity and homogeneous boundary values follows from the first property of $\hat{J}^p$. The preservation of discrete supports is clear from the piecewise

[4] The space $H^k_{\text{pw}}(\mathcal{T})$ consists of all functions $v \in L^2(\Omega)$ with $v|_T \in H^k(T)$, for all $T \in \mathcal{T}$.
[5] Discrete supports $\operatorname{supp}_{\mathcal{T}}(\cdot)$ were defined in D.2.71.

definition of $J_{\mathcal{T}}^p$. Finally, let us derive the error bound. Let $v \in H^2_{\mathrm{pw}}(\mathcal{T})$ and $T \in \mathcal{T}$. Denote by $\hat{v} := v \circ F_T \in H^2(\hat{T})$ the corresponding pull-back. Then,

$$\sum_{l=0}^{1} h_T^l |v - J_{\mathcal{T}}^p v|_{H^l(T)} \overset{\text{L.2.43}}{\lesssim} h_T^{d/2} \|\hat{v} - \hat{J}^p \hat{v}\|_{H^1(\hat{T})} \lesssim p^{(1-d)/2} h_T^{d/2} \inf_{\hat{w} \in \mathbb{P}^p(\hat{T})} \|\hat{v} - \hat{w}\|_{H^2(\hat{T})}$$
$$\lesssim p^{(1-d)/2} \inf_{\hat{w} \in \mathbb{P}^p(\hat{T})} \sum_{l=0}^{2} h_T^{d/2} |\hat{v} - \hat{w}|_{H^l(\hat{T})} \overset{\text{L.2.43}}{\lesssim} p^{(1-d)/2} \inf_{w \in \mathbb{P}^p(T)} \sum_{l=0}^{2} h_T^l |v - w|_{H^l(T)}.$$

This concludes the proof.

□

At this point, we remind the reader of D.2.63 and D.2.73, where we defined the maximal element diameter $h_{\mathcal{B}}$ of a mesh cluster $\mathcal{B} \subseteq \mathcal{T}$ and also the inflated mesh cluster $\mathcal{B}^\delta$, $\delta > 0$. Now we have everything we need to treat the small elements $\mathcal{B}_{\mathrm{small}}$. The following result can be seen as a discrete version of L.2.55.

Theorem 6.11. *Let $\mathcal{B} \subseteq \mathcal{T}$ and $\delta > 0$ be such that $4\sigma_{\mathrm{lqu}} h_{\mathcal{B}} \leq \delta \lesssim 1$. Let $u \in \mathbb{S}_0^{p,1}(\mathcal{T})$ be such that, for all $v \in \mathbb{S}_0^{p,1}(\mathcal{T})$ with[6] $\operatorname{supp}_{\mathcal{T}}(v) \subseteq \mathcal{B}^\delta$, there holds*

$$a(u, v) = 0.$$

Then, there holds the inequality

$$\delta |u|_{H^1(\mathcal{B})} \leq C(d, \sigma_{\mathrm{shp}}, \sigma_{\mathrm{lqu}}) p^{(9-d)/2} \|u\|_{L^2(\mathcal{B}^\delta)}.$$

Proof. Since we assumed $\delta \geq 4\sigma_{\mathrm{lqu}} h_{\mathcal{B}}$, we know from L.2.77 that there exists a discrete cut-off function $\kappa := \kappa_{\mathcal{B}}^\delta \in \mathbb{S}^{1,1}(\mathcal{T})$ with the following properties:

$$\operatorname{supp}_{\mathcal{T}}(\kappa) \subseteq \mathcal{B}^\delta, \qquad \kappa|_{\mathcal{B}} \equiv 1, \qquad 0 \leq \kappa \leq 1, \qquad \forall l \in \{0, 1\} : \quad |\kappa|_{W^{l,\infty}(\Omega)} \lesssim \delta^{-l}.$$

Denote by $J_{\mathcal{T}}^p : H^2_{\mathrm{pw}}(\mathcal{T}) \longrightarrow \mathbb{S}^{p,0}(\mathcal{T})$ the approximation operator from L.6.10. Since the product $\kappa^2 u$ lies in $\mathbb{S}_0^{p+2,1}(\mathcal{T})$, we know that

$$v := J_{\mathcal{T}}^p(\kappa^2 u) \in \mathbb{S}_0^{p,1}(\mathcal{T}).$$

Furthermore, we have

$$\operatorname{supp}_{\mathcal{T}}(v) = \operatorname{supp}_{\mathcal{T}}(J_{\mathcal{T}}^p(\kappa^2 u)) \overset{\text{L.6.10}}{\subseteq} \operatorname{supp}_{\mathcal{T}}(\kappa^2 u) \subseteq \operatorname{supp}_{\mathcal{T}}(\kappa) \subseteq \mathcal{B}^\delta.$$

In particular, the assumption on u tells us that $a(u, v) = 0$. Then, using an element-wise stability bound of $a(\cdot, \cdot)$ (similar to L.6.2), we compute

$$a(u, \kappa^2 u) = a(u, \kappa^2 u - v) = a(u, (\mathrm{id} - J_{\mathcal{T}}^p)(\kappa^2 u)) \lesssim \sum_{T \in \mathcal{B}^\delta} \|u\|_{H^1(T)} \|(\mathrm{id} - J_{\mathcal{T}}^p)(\kappa^2 u)\|_{H^1(T)}.$$

[6] For the definition of $\mathcal{B}^\delta$, see D.2.73.

In order to bound the error term, we choose a point $x_0 \in T$ such that $|\kappa(x_0)| = \min_{x\in T} |\kappa(x)|$. Since $\kappa \pm \kappa(x_0) \in \mathbb{P}^1(T)$, L.2.66 provides us with the following bounds:

$$\begin{aligned}
\|\kappa + \kappa(x_0)\|_{L^\infty(T)} &\lesssim \min_{x\in T} |\kappa(x) + \kappa(x_0)| + h_T|\kappa + \kappa(x_0)|_{W^{1,\infty}(T)} \lesssim |\kappa(x_0)| + h_T\delta^{-1},\\
\|\kappa - \kappa(x_0)\|_{L^\infty(T)} &\lesssim \min_{x\in T} |\kappa(x) - \kappa(x_0)| + h_T|\kappa - \kappa(x_0)|_{W^{1,\infty}(T)} \lesssim h_T\delta^{-1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|(\mathrm{id} - J^p_{\mathcal{T}})(\kappa^2 u)\|_{H^1(T)} &\overset{L.6.10}{\lesssim} p^{(1-d)/2} h_T^{-1} \inf_{w\in\mathbb{P}^p(T)} \sum_{l=0}^{2} h_T^l |\kappa^2 u - w|_{H^l(T)}\\
&\overset{L.2.67}{\lesssim} p^{(9-d)/2} h_T^{-1} \inf_{w\in\mathbb{P}^p(T)} \|\kappa^2 u - w\|_{L^2(T)}\\
&\leq p^{(9-d)/2} h_T^{-1} \|\kappa^2 u - \kappa(x_0)^2 u\|_{L^2(T)}\\
&\leq p^{(9-d)/2} h_T^{-1} \|\kappa + \kappa(x_0)\|_{L^\infty(T)} \|\kappa - \kappa(x_0)\|_{L^\infty(T)} \|u\|_{L^2(T)}\\
&\lesssim p^{(9-d)/2} \delta^{-1} (|\kappa(x_0)| + h_T\delta^{-1}) \|u\|_{L^2(T)}.
\end{aligned}$$

Recall that the element-wise stability bound for $a(u, \kappa^2 u)$ requires us to multiply both sides with $\|u\|_{H^1(T)} \eqsim \|u\|_{L^2(T)} + |u|_{H^1(T)}$. On the right-hand side, the following two terms emerge:

$$\begin{aligned}
(|\kappa(x_0)| + h_T\delta^{-1}) \|u\|^2_{L^2(T)} &\lesssim \delta^{-1} \|u\|^2_{L^2(T)},\\
(|\kappa(x_0)| + h_T\delta^{-1}) |u|_{H^1(T)} \|u\|_{L^2(T)} &\overset{L.2.67}{\lesssim} \|\kappa\nabla u\|_{L^2(T)} \|u\|_{L^2(T)} + p^2\delta^{-1} \|u\|^2_{L^2(T)}.
\end{aligned}$$

We summarize our findings:

$$\begin{aligned}
a(u, \kappa^2 u) &\lesssim \sum_{T\in\mathcal{B}^\delta} \|u\|_{H^1(T)} \|(\mathrm{id} - J^p_{\mathcal{T}})(\kappa^2 u)\|_{H^1(T)}\\
&\lesssim p^{(9-d)/2} \delta^{-1} \sum_{T\in\mathcal{B}^\delta} (|\kappa(x_0)| + h_T\delta^{-1}) (\|u\|^2_{L^2(T)} + |u|_{H^1(T)} \|u\|_{L^2(T)})\\
&\lesssim p^{(9-d)/2} \delta^{-1} \sum_{T\in\mathcal{B}^\delta} p^2\delta^{-1} \|u\|^2_{L^2(T)} + \|\kappa\nabla u\|_{L^2(T)} \|u\|_{L^2(T)}\\
&\overset{\text{C.S.}}{\lesssim} p^{(13-d)/2} \delta^{-2} \|u\|^2_{L^2(\mathcal{B}^\delta)} + p^{(9-d)/2} \delta^{-1} \|\kappa\nabla u\|_{L^2(\Omega)} \|u\|_{L^2(\mathcal{B}^\delta)}.
\end{aligned}$$

On the other hand, we can use the definition of $a(\cdot,\cdot)$ from D.6.1 to expand the term $a(u, \kappa^2 u)$ explicitly. One of the summands is amenable to the coercivity of the PDE coefficient a_1. In particular, using Hölder's inequality and Young's inequality with a free parameter $\varepsilon > 0$, we get

$$\begin{aligned}
\|\kappa\nabla u\|^2_{L^2(\Omega)} &\lesssim \langle a_1 \kappa\nabla u, \kappa\nabla u\rangle_{L^2(\Omega)}\\
&= a(u, \kappa^2 u) - 2\langle a_1\kappa\nabla u, u\nabla\kappa\rangle_{L^2(\Omega)} - \langle a_2 \cdot \nabla u, \kappa^2 u\rangle_{L^2(\Omega)} - \langle a_3 u, \kappa^2 u\rangle_{L^2(\Omega)}\\
&\lesssim p^{(13-d)/2} \delta^{-2} \|u\|^2_{L^2(\mathcal{B}^\delta)} + p^{(9-d)/2} \delta^{-1} \|\kappa\nabla u\|_{L^2(\Omega)} \|u\|_{L^2(\mathcal{B}^\delta)}\\
&\lesssim \varepsilon^{-1} p^{9-d} \delta^{-2} \|u\|^2_{L^2(\mathcal{B}^\delta)} + \varepsilon \|\kappa\nabla u\|^2_{L^2(\Omega)}.
\end{aligned}$$

Since the Young parameter ε can be chosen arbitrarily small, we may absorb the last summand in the left-hand side of the overall inequality. Finally,

$$\delta|u|_{H^1(\mathcal{B})} = \delta\|\kappa\nabla u\|_{L^2(\mathcal{B})} \leq \delta\|\kappa\nabla u\|_{L^2(\Omega)} \lesssim p^{(9-d)/2}\|u\|_{L^2(\mathcal{B}^\delta)}.$$

This concludes the proof.

□

Remark 6.12. *During the proof of T.6.11, the interpolation operator $J^p_{\mathcal{T}} : H^2_{\mathrm{pw}}(\mathcal{T}) \longrightarrow \mathbb{S}^{p,0}(\mathcal{T})$ from L.6.10 was used to turn the function $\kappa^2 u$ into an element of the ansatz space $V_N = \mathbb{S}^{p,1}_0(\mathcal{T})$. However, since $\kappa^2 u$ lies in the discrete space $\mathbb{S}^{p+2,1}_0(\mathcal{T})$ anyways, we could achieve the same result with a similar operator $\tilde{J}^p_{\mathcal{T}} : \mathbb{S}^{p+2,0}(\mathcal{T}) \longrightarrow \mathbb{S}^{p,0}(\mathcal{T})$, which we constructed in an earlier work (cf. [AFM22, Lemma 3.9.]). The advantage of $\tilde{J}^p_{\mathcal{T}}$ over $J^p_{\mathcal{T}}$ is that its proof was carried out in an arbitrary space dimension $d \in \mathbb{N}$, whereas [MR20] assumed $d \in \{1,2,3\}$. In compensation, $\tilde{J}^p_{\mathcal{T}}$ would produce worse powers of p, because [AFM22, Lemma 3.9.] was derived in the "wrong" norms. L.6.10 is the only reason why this chapter is limited to space dimensions $d \in \{1,2,3\}$.*

We finish this section with the complete proof of the discrete Caccioppoli inequality from A.4.19. As discussed in the overview paragraph of this section, the trick is to consider "small" and "large" elements separately.

Corollary 6.13. *Denote by $\sigma_{\mathrm{shp}}, \sigma_{\mathrm{lqu}} \geq 1$ the mesh related quantities from D.2.60. Then, A.4.19 is satisfied with a constant*

$$\sigma_{\mathrm{Cacc}} = C(d, \sigma_{\mathrm{shp}}, \sigma_{\mathrm{lqu}})p^{(9-d)/2}.$$

Proof. Let $B, D \in \mathbb{B}$ and $\delta > 0$ be such that $B^\delta \cap D = \emptyset$. Furthermore, let $u \in V_{\mathrm{sol}}(D)$. We divide the patch

$$\mathcal{T}(B) = \{T \in \mathcal{T} \,|\, T \cap B \neq \emptyset\}$$

into the groups

$$\begin{aligned} \mathcal{B}_{\mathrm{small}} &:= \{T \in \mathcal{T}(B) \,|\, 24\sigma_{\mathrm{shp}}\sigma_{\mathrm{lqu}}h_T \leq \delta\}, \\ \mathcal{B}_{\mathrm{large}} &:= \{T \in \mathcal{T}(B) \,|\, 24\sigma_{\mathrm{shp}}\sigma_{\mathrm{lqu}}h_T > \delta\} \end{aligned}$$

and start estimating:

$$\delta^2|u|^2_{H^1(\Omega\cap B)} = \delta^2 \sum_{T\in\mathcal{T}(B)} |u|^2_{H^1(T\cap B)} \leq \delta^2|u|^2_{H^1(\mathcal{B}_{\mathrm{small}})} + \sum_{T\in\mathcal{B}_{\mathrm{large}}} \delta^2|u|^2_{H^1(T\cap B)}.$$

First, we treat the individual large elements with L.6.9:

$$\sum_{T\in\mathcal{B}_{\mathrm{large}}} \delta^2|u|^2_{H^1(T\cap B)} \overset{L.6.9}{\lesssim} \sum_{T\in\mathcal{B}_{\mathrm{large}}} p^4\|u\|^2_{L^2(T\cap B^\delta)} \leq \sum_{T\in\mathcal{T}} p^4\|u\|^2_{L^2(T\cap B^\delta)} = p^4\|u\|^2_{L^2(\Omega\cap B^\delta)}.$$

It remains to take care of the small elements, which we now abbreviate by $\mathcal{B} := \mathcal{B}_{\mathrm{small}}$. We want to apply T.6.11 to the set $\mathcal{B}$, the parameter $\varepsilon := (6\sigma_{\mathrm{shp}})^{-1}\delta > 0$ and the function u. First, we check that ε is indeed a viable choice:

$$4\sigma_{\mathrm{lqu}}h_{\mathcal{B}} \leq \frac{4\sigma_{\mathrm{lqu}}\delta}{24\sigma_{\mathrm{shp}}\sigma_{\mathrm{lqu}}} = \varepsilon.$$

Next, we need to verify the assumption on the function u. To this end, consider a test function $v \in \mathbb{S}_0^{p,1}(\mathcal{T})$ with $\operatorname{supp}_{\mathcal{T}}(v) \subseteq \mathcal{B}^\varepsilon$. Since $\mathcal{B} \subseteq \mathcal{T}(B)$ and since $3h_{\mathcal{B}} \leq 24\sigma_{\mathrm{shp}}\sigma_{\mathrm{lqu}}h_{\mathcal{B}} \leq \delta$, we have

$$\operatorname{supp}(v) \subseteq \bigcup \operatorname{supp}_{\mathcal{T}}(v) \subseteq \bigcup \mathcal{B}^\varepsilon \overset{L.2.74}{\subseteq} \Omega \cap B^\delta.$$

In particular, for all $n \in \iota(D)$ (cf. D.4.17), there holds[7]

$$|\langle \lambda_n, v\rangle_*| \overset{L.6.7}{\lesssim} \|v\|_{L^2(\Omega_n)} \overset{n\in\iota(D)}{\leq} \|v\|_{L^2(D)} \overset{B^\delta\cap D=\emptyset}{=} 0.$$

According to D.4.18, we can write u in the form[8] $u = S_N\Lambda \boldsymbol{f}$ (for some $\boldsymbol{f} \in \mathbb{R}^N$ with $\operatorname{supp}(\boldsymbol{f}) \subseteq \iota(D)$), so that

$$a(u,v) = a(S_N\Lambda\boldsymbol{f}, v) \overset{D.4.6}{=} \langle \Lambda\boldsymbol{f}, v\rangle_* \overset{D.4.13}{=} \left\langle \sum_{n=1}^N \boldsymbol{f}_n\lambda_n, v \right\rangle_* = \sum_{n\in\iota(D)} \boldsymbol{f}_n\langle\lambda_n, v\rangle_* = 0.$$

We may then apply T.6.11 and obtain the following bound:

$$\delta^2|u|^2_{H^1(\mathcal{B}_{\mathrm{small}})} \eqsim \varepsilon^2|u|^2_{H^1(\mathcal{B})} \overset{T.6.11}{\lesssim} p^{9-d}\|u\|^2_{L^2(\mathcal{B}^\varepsilon)} \leq p^{9-d}\|u\|^2_{L^2(\Omega\cap B^\delta)}.$$

Note that $p^{9-d} \geq p^4$, because $d \in \{1,2,3\}$. Finally, we combine the estimates for both groups:

$$\delta^2|u|^2_{H^1(\Omega\cap B)} \lesssim (p^{9-d} + p^4)\|u\|^2_{L^2(\Omega\cap B^\delta)} \lesssim p^{9-d}\|u\|^2_{L^2(\Omega\cap B^\delta)}.$$

This concludes the proof.

□

6.7 A corollary

Now that we completed all the necessary steps for the application of T.4.21, we summarize our findings in a corollary:

Corollary 6.14. *Let $p \in \mathbb{N}$. Denote by $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ the Gram matrix that corresponds to the bilinear form $a(\cdot,\cdot)$ from D. 6.1 and the FEM basis $\{\varphi_1, \dots, \varphi_N\} \subseteq \mathbb{S}_0^{p,1}(\mathcal{T})$ from D.6.3. Assume that the block partition $\mathbb{P}^2$ from C.3.42 is constructed using the parameter $\sigma_{\mathrm{small}} := \binom{d+p}{d}$. Then, for every $r \in \mathbb{N}$, there exists an $\mathcal{H}$-matrix*

$$\boldsymbol{B}_r \in \mathcal{H}(\mathbb{P}^2, r)$$

with the following properties:

1. *The memory requirements to store $\boldsymbol{B}_r$ can be bounded by*

$$C(d, \Omega, \sigma_{\mathrm{shp}}, \sigma_{\mathrm{adm}})(p^d + r)\ln(h^{-1}_{\min,\mathcal{T}})N.$$

[7] The implicit constant is of the form $C(d, \sigma_{\mathrm{shp}}, p, \gamma, h_{\min,\mathcal{T}})$.
[8] The operators $S_N : V_N^* \longrightarrow V_N$ and $\Lambda : \mathbb{R}^d \longrightarrow V_N^*$ were defined in D.4.6 and D.4.13, respectively.

2. *There exist numbers* $C_0 \geq 1$ *and* $\sigma_{\text{exp}} > 0$ *of the form*[9]

$$\begin{aligned} C_0 &= C(d, \Omega, \alpha_1, \alpha_2, \alpha_3, \sigma_{\text{shp}}, \sigma_{\text{lqu}}, \sigma_{\text{adm}}), \\ \sigma_{\text{exp}} &= C(d, \Omega, \sigma_{\text{shp}}, \sigma_{\text{lqu}}, \sigma_{\text{adm}})^{-1} p^{-(9-d)/2} \end{aligned}$$

such that the following error bound is satisfied:

$$\|\boldsymbol{A}^{-1} - \boldsymbol{B}_r\|_2 \leq C_0 p^{d+2\gamma} \ln(h_{\text{min},\mathcal{T}}^{-1}) h_{\text{min},\mathcal{T}}^{-d} \exp(-\sigma_{\text{exp}} r^{1/(d+1)}).$$

Proof. Over the course of L.6.2, L.6.6, L.6.7, L.6.8 and C.6.13, we derived the following relations:

$$\sigma_{\text{coco}} = C(d, \Omega, \alpha_1, \alpha_2, \alpha_3), \qquad k_0 = 0, \qquad \sigma_{\text{stab}} = C(d, \sigma_{\text{shp}}) p^{\gamma} h_{\text{min},\mathcal{T}}^{-d/2},$$
$$\sigma_{\text{ovlp}} = \binom{d+p}{d}, \qquad \sigma_{\text{sprd}} = C(\Omega), \qquad h_{\text{min}} \geq h_{\text{min},\mathcal{T}}, \qquad \sigma_{\text{Cacc}} = C(d, \sigma_{\text{shp}}, \sigma_{\text{lqu}}) p^{(9-d)/2}.$$

Ad item 1: The bound on the storage complexity from T.4.21 now reads

$$C(d, \sigma_{\text{shp}}, \sigma_{\text{sprd}}, \sigma_{\text{adm}})(\sigma_{\text{small}} + r) \ln(h_{\text{min}}^{-1}) N \leq C(d, \Omega, \sigma_{\text{shp}}, \sigma_{\text{adm}})(p^d + r) \ln(h_{\text{min},\mathcal{T}}^{-1}) N.$$

Ad item 2: The numbers C_0 and σ_{exp} from T.4.21 become

$$C_0 \overset{T.4.21}{=} C(d, k, \Omega, \sigma_{\text{coco}}, \sigma_{\text{shp}}, \sigma_{\text{sprd}}, \sigma_{\text{adm}}) = C(d, \Omega, \alpha_1, \alpha_2, \alpha_3, \sigma_{\text{shp}}, \sigma_{\text{adm}})$$

and

$$\sigma_{\text{exp}} \overset{T.4.21}{=} C(d, k, \Omega, \sigma_{\text{sprd}}, \sigma_{\text{adm}})^{-1} \sigma_{\text{Cacc}}^{-1} = C(d, \Omega, \sigma_{\text{shp}}, \sigma_{\text{lqu}}, \sigma_{\text{adm}})^{-1} p^{-(9-d)/2}.$$

Finally, the prefactor in the error bound from T.4.21 turns into

$$\begin{aligned} \sigma_{\text{stab}}^2 \sigma_{\text{ovlp}} \sigma_{\text{Cacc}}^{k_0} \ln(h_{\text{min}}^{-1}) h_{\text{min}}^{-k_0} &\lesssim (p^{\gamma} h_{\text{min},\mathcal{T}}^{-d/2})^2 p^d \ln(h_{\text{min},\mathcal{T}}^{-1}) \\ &= p^{d+2\gamma} \ln(h_{\text{min},\mathcal{T}}^{-1}) h_{\text{min},\mathcal{T}}^{-d}. \end{aligned}$$

□

Note that C.6.14 holds true for *any* simplicial mesh $\mathcal{T}$ as defined in D.2.60. However, in order to get a useful complexity bound, we need to make an assumption about the relationship between $h_{\text{min},\mathcal{T}}$ and N. In fact, in the extreme case $h_{\text{min},\mathcal{T}} \eqsim e^{-N}$, the complexity bound reads $\mathcal{O}(N^2)$ and the result becomes useless. For a large class of meshes, the dependence of h_{min} on N is of algebraic nature and we get satisfactory results.

Definition 6.15. *Let* $\Omega \subseteq \mathbb{R}^d$ *be a polyhedron (D.2.62) and let* $\mathcal{T} \subseteq \text{Pow}(\overline{\Omega})$ *be a mesh. Let* $\Gamma \subseteq \overline{\Omega}$ *be a set with*[10] $T^\circ \cap \Gamma = \emptyset$*, for all* $T \in \mathcal{T}$*. Furthermore, let* $H > 0$ *and* $\sigma_{\text{grade}} \in [1, \infty]$*. We say that* $\mathcal{T}$ has grading σ_{grade}*, if there exists a constant* $C := C(d, \Omega, \Gamma) \geq 1$*, such that the following relation is satisfied*[11]*:*

$$\forall T \in \mathcal{T}: \qquad C^{-1} h_T \leq \text{dist}_2(x_T, \Gamma)^{1-1/\sigma_{\text{grade}}} H \leq C h_T.$$

The case $\sigma_{\text{grade}} = 1$ *is called* uniform *grading, the case* $\sigma_{\text{grade}} \in (1, \infty)$ *is an* algebraic *grading and the case* $\sigma_{\text{grade}} = \infty$ *is an* exponential *grading.*

[9] The constants $\alpha_1, \alpha_2, \alpha_3$ were introduced in the initial problem statement, Section 6.1.
[10] In other words, Γ is a subset of the mesh's skeleton.
[11] Recall from D.2.63 that, for every mesh element $T \in \mathcal{T}$, we fixed an incenter $x_T \in T$.

We mention that the notion of "gradedness" can easily be formulated for general points clouds $x_1, \dots, x_N \in \mathbb{R}^d$ as well. Later, in Section 7.10, we will look at several examples of graded point clouds.

Lemma 6.16. *Denote by $\Gamma \subseteq \overline{\Omega}$, $H > 0$ and $\sigma_{\mathrm{grade}} \in [1, \infty]$ the quantities from D.6.15. If $\sigma_{\mathrm{grade}} \in [1, \infty)$, then there exist constants $C(d, \Omega, \sigma_{\mathrm{shp}}, \Gamma, \sigma_{\mathrm{grade}}) \geq 1$ such that*

$$(\#\mathcal{T})^{-\sigma_{\mathrm{grade}}/d} \lesssim H^{\sigma_{\mathrm{grade}}} \lesssim h_{\min,\mathcal{T}} \leq h_{\mathcal{T}} \lesssim H.$$

Proof. Since $1 - 1/\sigma_{\mathrm{grade}} \geq 0$, we can estimate

$$h_{\mathcal{T}} = \max_{T \in \mathcal{T}} h_T \overset{D.6.15}{\lesssim} \max_{T \in \mathcal{T}} \operatorname{dist}_2(x_T, \Gamma)^{1-1/\sigma_{\mathrm{grade}}} H \leq \operatorname{diam}_2(\Omega)^{1-1/\sigma_{\mathrm{grade}}} H \lesssim H.$$

On the other hand, for every $T \in \mathcal{T}$, we have

$$\operatorname{Ball}_2(x_T, (2\sigma_{\mathrm{shp}})^{-1} h_T) \cap \Gamma \overset{L.2.17}{\subseteq} T^\circ \cap \Gamma \overset{D.6.15}{=} \emptyset.$$

This implies $\operatorname{dist}_2(x_T, \Gamma) \geq (2\sigma_{\mathrm{shp}})^{-1} h_T$ and we get

$$h_T \overset{D.6.15}{\gtrsim} \operatorname{dist}_2(x_T, \Gamma)^{1-1/\sigma_{\mathrm{grade}}} H \gtrsim h_T^{1-1/\sigma_{\mathrm{grade}}} H.$$

Since it was assumed that $\sigma_{\mathrm{grade}} < \infty$, we can easily solve for h_T and obtain the relation $h_T \gtrsim H^{\sigma_{\mathrm{grade}}}$. Taking the minimum over all $T \in \mathcal{T}$, it follows that $h_{\min,\mathcal{T}} \gtrsim H^{\sigma_{\mathrm{grade}}}$.

Finally, we estimate

$$1 \lesssim \operatorname{meas}(\Omega) = \sum_{T \in \mathcal{T}} \operatorname{meas}(T) \overset{L.2.17}{\leq} \sum_{T \in \mathcal{T}} h_T^d \leq (\#\mathcal{T}) h_{\mathcal{T}}^d \lesssim (\#\mathcal{T}) H^d,$$

and deduce $(\#\mathcal{T})^{-1/d} \lesssim H$.

□

Uniformly- or algebraically graded meshes indeed yield good complexity bounds.

Corollary 6.17. *If $\mathcal{T}$ is a mesh with grading $\sigma_{\mathrm{grade}} \in [1, \infty)$, then C.6.14 holds verbatim with a complexity bound of the form*

$$\mathcal{O}((p^d + r) \ln(N) N).$$

Proof. Follows immediately from L.6.16 and the crude bound

$$\#\mathcal{T} \leq p^d \#\mathcal{T} \eqsim \dim(\mathbb{S}_0^{p,1}(\mathcal{T})) \overset{D.6.3}{=} N.$$

□

If $\mathcal{T}$ is exponentially graded, i.e., $\sigma_{\mathrm{grade}} = \infty$, then we cannot say anything about the relationship between $h_{\min}$ and N. In Figure 6.1 (third image from the left), we show an example of exponential grading towards an edge of the unit square in $\mathbb{R}^2$. In this case, the relationship is of the form $h_{\min,\mathcal{T}} \eqsim N^{-1}$. However, in the fourth image, the exponential grading steers towards a corner and the relationship reads $h_{\min,\mathcal{T}} \eqsim 2^{-N}$.

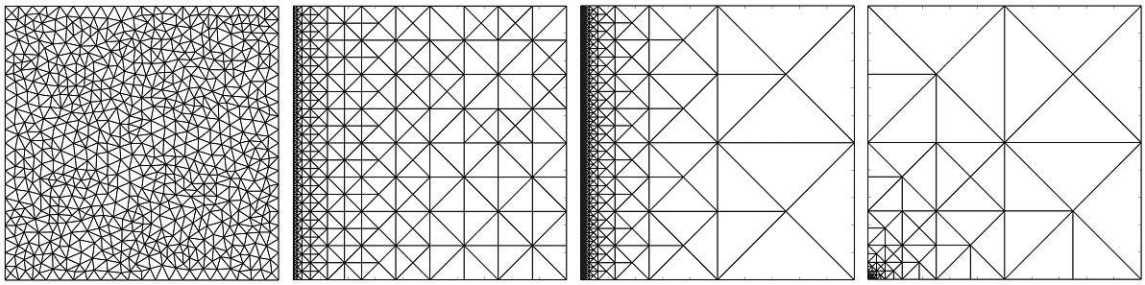

Figure 6.1: From left to right: Uniform, algebraically graded towards edge, exponentially graded towards edge, exponentially graded towards corner. The complexity bound from C.6.17 is satisfied in the first, second and third case, but *not* in the last one.

We close this section with a remark on exponentially graded meshes.

Remark 6.18. *One possible application of exponentially graded meshes can be found in the context of the* boundary concentrated FEM*, e.g., [KM02] and [KM03]. This method is similar to the boundary element method (BEM), in that most mesh element lie on the boundary of Ω. However, we mention that C.6.14 is not directly applicable to this method, because [KM03] replaces the (constant-degree) spline spaces $\mathbb{S}_0^{p,1}(\mathcal{T})$ from D.2.70 with varying-degree spline spaces $\mathbb{S}_0^{\boldsymbol{p},1}(\mathcal{T})$, $\boldsymbol{p} = \{p_T \,|\, T \in \mathcal{T}\}$.*

6.8 Numerical examples

In this subsection, we illustrate the validity of C.6.14 by means of several numerical examples. The examples are taken from [AFM21a, Section 4] and [AFM22, Section 5].

6.8.1 Algebraic grading

For the geometry we choose the *L-shape* $\Omega := ((0,1) \times (0,1)) \backslash ([1/2,1] \times [1/2,1]) \subseteq \mathbb{R}^2$ in two space dimensions. The PDE coefficients for the model problem from Section 6.1 are given by $a_1(x) = \left(\begin{smallmatrix} 10 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, $a_2(x) := \left(\begin{smallmatrix} 10x_2 \\ 0 \end{smallmatrix}\right)$ and $a_3(x) := 1$. The mesh $\mathcal{T}$ has grading $\sigma_{\mathrm{grade}} = 5$ towards $\Gamma := \{(1/2, 1/2)\}$ and the coarse mesh width is given by $H := 0.0095$. We use the lowest-order spline space $\mathbb{S}_0^{1,1}(\mathcal{T})$ from D.2.70 and the basis of hat-functions $\{\varphi_1, \dots, \varphi_N\} \subseteq \mathbb{S}_0^{1,1}(\mathcal{T})$ (cf. L.2.72). The block partition $\mathbb{P}^2$ is then constructed using the clustering strategy from Chapter 3 and the clustering parameters are set to $\sigma_{\mathrm{adm}} := 2$ and $\sigma_{\mathrm{small}} := 25$ (cf. C.3.42).

Since the approximant $\boldsymbol{B}_r \in \mathcal{H}(\mathbb{P}^2, r)$ from C.6.14 is of theoretical nature, we revert to the truncated SVDs from Section 3.1. First, we compute the exact inverse $\boldsymbol{A}^{-1} \in \mathbb{R}^{N\times N}$ explicitly. Then, for every admissible block $\boldsymbol{A}^{-1}|_{I\times J}$, we compute the first $r \in \{1, \dots, 50\}$ singular values and end up with the following *computable* error bound:

$$\|\boldsymbol{A}^{-1} - \boldsymbol{B}_r\|_2 \overset{C.3.42}{\lesssim} \ln(h_{\min}^{-1}) \max_{(I,J)\in\mathbb{P}^2_{\text{adm}}} \sigma_{r+1}(\boldsymbol{A}^{-1}|_{I\times J}).$$

The numerical example is implemented in `MATLAB` ([MAT]). For the inversion of the full matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ we use `MATLAB`'s built-in procedure `inv(...)`. For the SVDs we use `svds(...)`. Recall that an exact matrix inversion needs $\mathcal{O}(N^2)$ memory and $\mathcal{O}(N^3)$ time to compute, which effectively restricts the maximal feasible problem size to $N \approx 70.000$ on our machine.

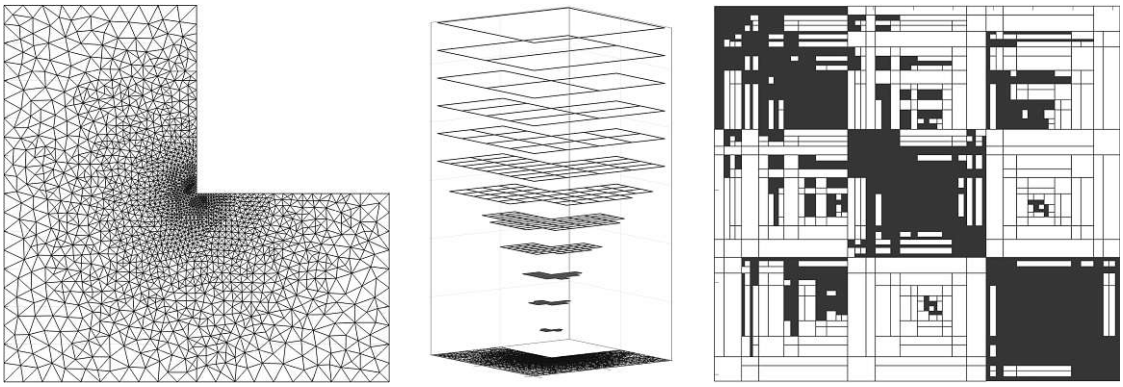

Figure 6.2: The mesh $\mathcal{T}$, the box tree $\mathbb{T}$ (cf. D.3.17) and the block partition $\mathbb{P}^2$ (cf. C.3.42) for $N \approx 2.000$ degrees of freedom.

In Figure 6.2, we choose $N \approx 2.000$ degrees of freedom. The elements are graded towards the reentrant corner with a grading exponent $\sigma_{\text{grade}} = 5$. The cluster tree $\mathbb{T}$ is clearly deeper near the grading center. The block partition $\mathbb{P}^2$ uses sorted indices internally. Only a few admissible blocks are far away from the diagonal, lots of small blocks agglomerate along the diagonal. The sparsity pattern becomes more pronounced as $N \to \infty$.

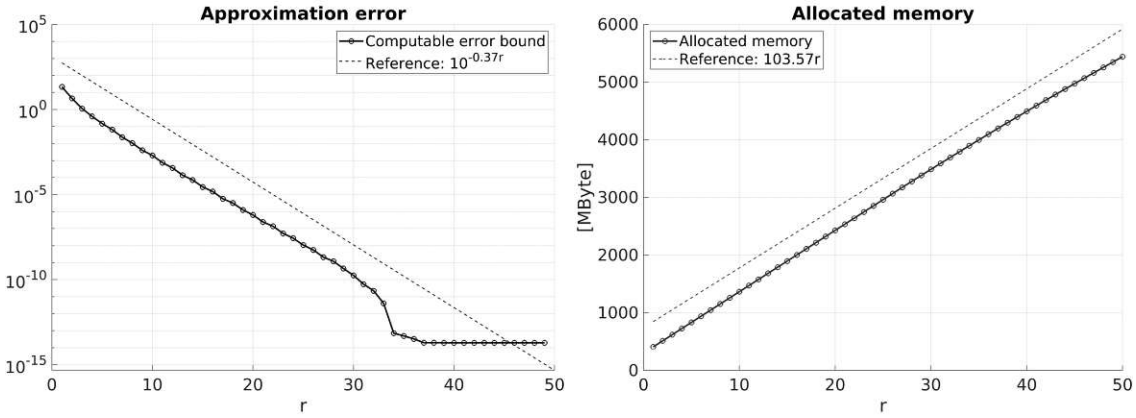

Figure 6.3: Approximation error and memory allocation for $N \approx 72.000$ degrees of freedom.

In Figure 6.3, we choose $N \approx 72.000$ degrees of freedom. The computable error bound from above (for $r \in \{1, \dots, 50\}$) is depicted on a linear abscissa and a logarithmic ordinate. The values are below a straight line with slope -0.37 indicating an *exponential decay* $\text{error}(r) \lesssim 10^{-0.37r}$. This is even better than the theoretical bound from C.6.14 and might be attributable to the fact that block-wise truncated SVDs produce the best possible $\mathcal{H}$-matrix approximant, whereas C.6.14 produces *some* $\mathcal{H}$-matrix approximant. The allocated memory in MBytes is plotted on a linear abscissa and a linear ordinate. The values are below a straight line with slope 103.57 indicating a *polynomial growth* $\text{memory}(r) \lesssim r$. Choosing a rank bound $r = 37$, for example, gives an approximation error $\approx 10^{-14}$ and uses ≈ 4.2 GByte memory. In comparison, the full matrix $\boldsymbol{A}^{-1}$ takes ≈ 41.4 GByte memory.

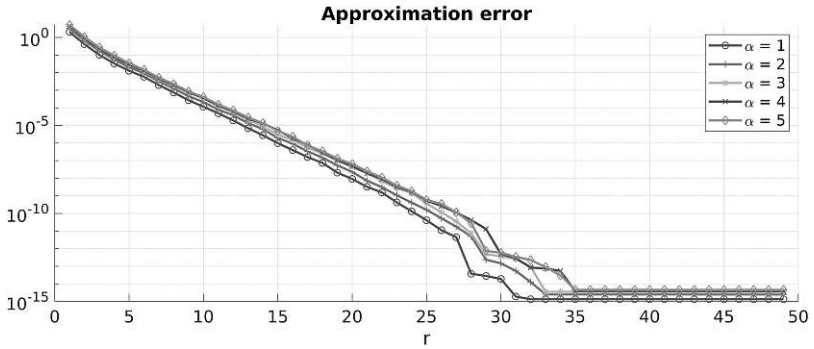

Figure 6.4: Comparison of approximation errors for different grading parameters, $\sigma_{\text{grade}} \in \{1, 2, 3, 4, 5\}$. The number of degrees of freedom is kept constant at roughly $N \approx 17.500$ throughout all five runs.

Finally, in Figure 6.4, we choose $N \approx 17.500$ degrees of freedom and multiple grading exponents in the range $\{1, 2, 3, 4, 5\}$. The case $\sigma_{\text{grade}} = 1$ corresponds to a uniform mesh, whereas $\sigma_{\text{grade}} = 5$ is "heavily" graded. Again, the computable error bound from above is

shown on a linear abscissa and a logarithmic ordinate. The convergence speed seems to be largely independent of the grading exponent σ_{grade}.

6.8.2 Exponential grading

Next, we take a look at an example on the unit square $\Omega := (0,1) \times (0,1) \subseteq \mathbb{R}^2$. The mesh $\mathcal{T}$ is exponentially graded towards the left edge ($\Gamma := \{0\} \times [0,1]$, $H := 0.25$, $\sigma_{\text{grade}} = \infty$ in D.6.15). To get a computable error bound, we proceed as in Section 6.8.1, i.e., $\boldsymbol{A}^{-1}$ is computed exactly and we use block-wise truncated SVDs.

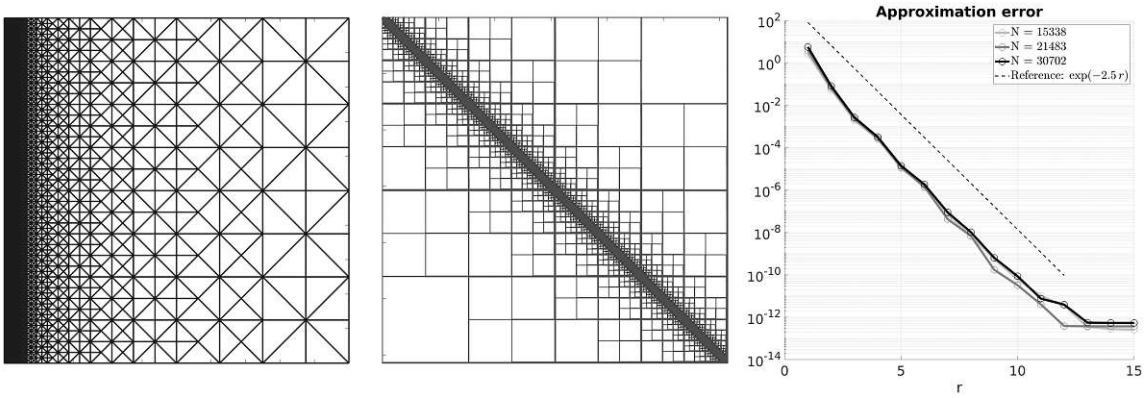

Figure 6.5: Left: The mesh $\mathcal{T}$. Center: The block partition $\mathbb{P}$. Right: Empirical approximation errors.

The right-hand image in Figure 6.5 depicts a comparison between three different problem sizes of roughly $N \approx 15.000$, $N \approx 21.500$ and $N \approx 31.000$ degrees of freedom. The error appears to decline at a rate of $\exp(-2.5r)$, which is again much better than our theoretical prediction $\exp(-\sigma_{\text{exp}} r^{1/3})$ from C.6.14.

6.8.3 Some $\mathcal{H}$-arithmetic[12]

In this final example, we use the same mesh $\mathcal{T}$ as in Section 6.8.2, but we increase the polynomial degree to $p \in \{5, 6\}$. We employ a combination of the finite element code `NGSolve` from [NGS] (which is capable of higher order polynomials) and the $\mathcal{H}$-matrix library `H2Lib` from [H2L]. Both libraries are coupled using a code which was previously used in [EMM+21]. We use the polynomial degrees $p = 5$ and $p = 6$, which lead to problem sizes of $N \approx 5.800$ and $N \approx 17.000$, respectively. This time, an $\mathcal{H}$-matrix approximant $\boldsymbol{B}_r$ is computed via an $\mathcal{H}$-Cholesky decomposition $\boldsymbol{A} \approx \boldsymbol{L}_{\mathcal{H}} \boldsymbol{L}_{\mathcal{H}}^T$ and a subsequent inversion thereof. We then use the error measure

$$\frac{\|\boldsymbol{A}^{-1} - (\boldsymbol{L}_{\mathcal{H}} \boldsymbol{L}_{\mathcal{H}}^T)^{-1}\|_2}{\|\boldsymbol{A}^{-1}\|_2} \leq \|\boldsymbol{I} - (\boldsymbol{L}_{\mathcal{H}} \boldsymbol{L}_{\mathcal{H}}^T)^{-1} \boldsymbol{A}\|_2,$$

[12] This experiment was performed by Dr. Markus Faustmann, a co-author of [AFM22].

which does not involve an explicit inversion of $\boldsymbol{A}^{-1}$. Figure 6.6 shows exponential convergence of this error measure.

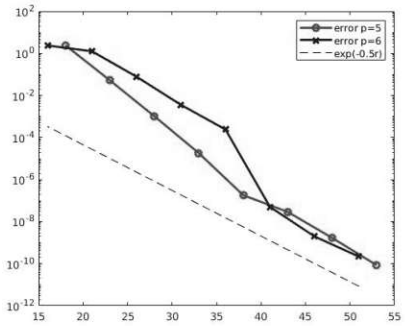

Figure 6.6: Exponential convergence of $\mathcal{H}$-matrix approximations for $p \in \{5, 6\}$.

7 An application in an RBF setting

A major drawback of mesh-based approximation schemes (such as the finite element method from the previous chapter) is the necessity for good mesh-generation algorithms. To alleviate this problem, a variety of so-called *mesh-less methods* has been developed over a period of more than four decades. In this chapter, we take a look at a *radial basis function (RBF)* interpolation problem and show how this seemingly unrelated problem can be expressed in the framework of our main result, T.4.21.

Inspiration for this part of the thesis was mainly taken from the book [Wen05], which provides a comprehensive introduction to the theory of radial basis functions. However, we mention that [Wen05] uses mostly Fourier transformation techniques, whereas the present text focuses more on the variational aspects of the theory. For an overview of general meshless methods, we refer to [DO95] and also [WQ19, Chapter 1].

The structure of this chapter is identical to the one of Chapter 6. We formulate the basic model problem, find the appropriate functional analytic setting, verify the assumptions from Section 4.6 and develop bounds for the quantities σ_{coco}, k_0, σ_{stab}, σ_{shp}, σ_{ovlp}, σ_{sprd} and $h_{\min}$. Finally, we summarize our findings in a corollary of T.4.21.

7.1 An interpolation problem

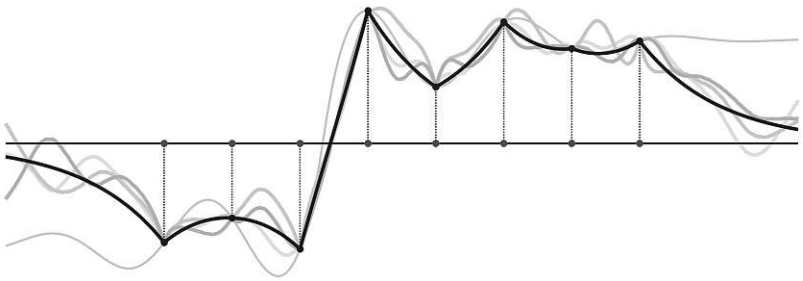

Figure 7.1: An interpolation problem in 1D.

Let $d \in \mathbb{N}$ and $N \in \mathbb{N}$. We consider a family of functions $\{\varphi_1, \dots, \varphi_N\} \subseteq C^0(\mathbb{R}^d)$ of the form

$$\varphi_n := \varphi(\cdot - x_n),$$

where $\varphi \in C^0(\mathbb{R}^d)$ is fixed and where $x_1, \dots, x_N \in \mathbb{R}^d$ are given (pairwise distinct) *interpolation points*. Furthermore, let $\boldsymbol{f} \in \mathbb{R}^N$ be a given vector of target values. We seek a

solution $u \in \operatorname{span}\{\varphi_1, \dots, \varphi_N\}$ of the following *interpolation problem*:

$$\forall m \in \{1, \dots, N\} : \qquad u(x_m) = \boldsymbol{f}_m.$$

7.2 The radial basis function φ

If we want to apply the results from Chapter 4, we first have to find a weak formulation of the interpolation problem. To this end, we need to express the point evaluations of u as integrals. This can be achieved via Fourier transformation techniques. Here, we use the following definition of Fourier transforms: For all $f \in L^1(\mathbb{R}^d)$ and all $y \in \mathbb{R}^d$,

$$\widehat{f}(y) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x) e^{-i\langle y,x\rangle_2} \,\mathrm{d}x, \qquad \check{f}(y) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x) e^{i\langle y,x\rangle_2} \,\mathrm{d}x.$$

We also use the symbols $\mathcal{F}f := \widehat{f}$ and $\mathcal{F}^{-1}f := \check{f}$. For basic properties of Fourier transforms, we refer the reader to, e.g., [Yos80, Chapter 6] or [Eva10, Section 4.3.].

Definition 7.1. *Let $k \in \mathbb{N}$ with $k > d/2$ and let $b \in (0, \infty)$. We consider the following* radial basis function[1]

$$\forall x \in \mathbb{R}^d : \qquad \varphi(x) := \frac{(4\pi)^{-d/2}}{\Gamma(k)} \int_0^\infty t^{k-d/2-1} e^{-b^2 t} e^{-\|x\|_2^2/(4t)} \,\mathrm{d}t.$$

Note that φ is indeed a radial function, i.e., $\varphi(x)$ only depends on $\|x\|_2$. We mention that φ can also be written in the form

$$\varphi(x) = \frac{(2\pi)^{-d/2}}{2^{k-1}\Gamma(k)} \left(\frac{\|x\|}{b}\right)^{k-d/2} K_{k-d/2}(b\|x\|),$$

which goes by the name of *Matérn function*, *Sobolev spline* or *Bessel potential* in the literature (e.g., [AS61]). Here,

$$K_\nu(r) := \int_0^\infty e^{-r\cosh(s)} \cosh(\nu s) \,\mathrm{d}s = \int_{-\infty}^\infty e^{-r(e^s+e^{-s})/2} e^{\nu s}/2 \,\mathrm{d}s$$

is the well-known *modified Bessel function of the second kind*. The formula for $\varphi(x)$ can be derived by plugging in $\nu = k - d/2$ and $r = b\|x\|$ and then substituting $s = \ln(2bt/\|x\|)$. Furthermore, in the case where $d \in \{1, 3, 5, \dots\}$, there holds the explicit representation

$$\varphi(x) = \frac{(4\pi)^{(1-d)/2}}{\Gamma(k)(2b)^{2L+1}} \sum_{l=0}^{L} \frac{(2L-l)!}{l!(L-l)!} (2b\|x\|)^l e^{-b\|x\|}, \qquad L := k - d/2 - 1/2 \in \mathbb{N}_0.$$

[1] Here, $\Gamma(k) := \int_0^\infty t^{k-1} e^{-t} \,\mathrm{d}t$ is the Gamma function.

This follows easily from the known identity

$$K_{L+1/2}(r) = \sum_{l=0}^{L} \frac{\pi^{1/2}(L+l)!}{l!(L-l)!} \frac{e^{-r}}{(2r)^{l+1/2}},$$

which can be found, e.g., in [GR07, Page 925].

The relevant properties of the function φ are summarized in the next lemma, which is taken from [AFM21b, Lemma 2.10.]:

Lemma 7.2. *1. For all $x_0 \in \mathbb{R}^d$, there holds $\varphi(\cdot - x_0) \in H^k(\mathbb{R}^d)$. In particular, $\varphi \in H^k(\mathbb{R}^d)$.*

2. The function φ is a fundamental solution of the differential operator

$$\mathrm{D}^{2k} := (b^2 - \Delta)^k = \sum_{l=0}^{k} \binom{k}{l} b^{2(k-l)} (-\Delta)^l.$$

More precisely, there holds the following identity:

$$\forall x_0 \in \mathbb{R}^d : \forall v \in C_0^\infty(\mathbb{R}^d) : \qquad \int_{\mathbb{R}^d} \varphi(x - x_0)(\mathrm{D}^{2k} v)(x) \,\mathrm{d}x = v(x_0).$$

Proof. Ad item 1: Using the substitution $t = s/b^2$ and the assumption $k > d/2$, we can check that the integral defining $\varphi(x)$ is indeed well-defined:

$$\begin{aligned} \int_0^\infty |t^{k-d/2-1} e^{-b^2 t} e^{-\|x\|^2/(4t)}| \,\mathrm{d}t &\leq \int_0^\infty t^{k-d/2-1} e^{-b^2 t} \,\mathrm{d}t \\ &= b^{d-2k} \int_0^\infty s^{k-d/2-1} e^{-s} \,\mathrm{d}s = \frac{\Gamma(k - d/2)}{b^{2k-d}} < \infty. \end{aligned}$$

In fact, using Fubini's Theorem for non-negative integrands and the transformations $x = \sqrt{t} y$ and $t = s/b^2$, we find that $\varphi \in L^1(\mathbb{R}^d)$:

$$\begin{aligned} \|\varphi\|_{L^1(\mathbb{R}^d)} &\asymp \int_0^\infty t^{k-d/2-1} e^{-b^2 t} \int_{\mathbb{R}^d} e^{-\|x\|^2/(4t)} \,\mathrm{d}x \,\mathrm{d}t \\ &= \left(\int_0^\infty t^{k-1} e^{-b^2 t} \,\mathrm{d}t \right) \left(\int_{\mathbb{R}^d} e^{-\|y\|^2/4} \,\mathrm{d}y \right) = \frac{C(d)\Gamma(k)}{b^{2k}} < \infty. \end{aligned}$$

Next, we compute the Fourier transform of φ. Recall (e.g., [Yos80, Chapter 6]) that the Gauß kernel $e^{-\|\cdot\|^2/2}$ is a fixpoint of the Fourier transform and that there holds the relation

$\mathcal{F}(e^{-\|\cdot\|^2/(4t)})(y) = (2t)^{d/2}e^{-t\|y\|^2}$ for all $t > 0$ and $y \in \mathbb{R}^d$. Substituting $t = s/(b^2 + \|y\|^2)$, we obtain the following expression:

$$\begin{aligned}\widehat{\varphi}(y) &= \frac{(4\pi)^{-d/2}}{\Gamma(k)} \int_0^\infty t^{k-d/2-1} e^{-b^2 t} \mathcal{F}(e^{-\|\cdot\|^2/(4t)})(y)\,\mathrm{d}t \\ &= \frac{(2\pi)^{-d/2}}{\Gamma(k)} \int_0^\infty t^{k-1} e^{-(b^2+\|y\|^2)t}\,\mathrm{d}t = \frac{(2\pi)^{-d/2}}{(b^2+\|y\|^2)^k}.\end{aligned}$$

Using polar coordinates, a straightforward computation shows that $\widehat{\varphi} \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ and we know from theory that then also $\varphi \in L^2(\mathbb{R}^d)$ (e.g., [Eva10, Section 4.3.1.]). A similar computation reveals $(1 + \|\cdot\|^2)^{k/2}\widehat{\varphi} \in L^2(\mathbb{R}^d)$ and it follows that even $\varphi \in H^k(\mathbb{R}^d)$ (e.g., [Eva10, Section 5.8.5.]). Finally, a simple integral transformation also yields $\varphi(\cdot - x_0) \in H^k(\mathbb{R}^d)$, for all $x_0 \in \mathbb{R}^d$.

Ad item 2: Let $x_0 \in \mathbb{R}^d$ and $v \in C_0^\infty(\mathbb{R}^d)$. Using standard Fourier manipulation rules and the explicit formula for $\widehat{\varphi}$, we compute

$$\begin{aligned}&\int_{\mathbb{R}^d} \varphi(x - x_0)(\mathrm{D}^{2k}v)(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \varphi(x - x_0)\mathcal{F}((b^2 + \|\cdot\|^2)^k \check{v})(x)\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} \widehat{\varphi}(y)e^{-i\langle x_0, y\rangle}(b^2 + \|y\|^2)^k \check{v}(y)\,\mathrm{d}y = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \check{v}(y)e^{-i\langle x_0, y\rangle}\,\mathrm{d}y = v(x_0).\end{aligned}$$

This concludes the proof.

□

7.3 The space V

L.7.2 provides a hint about the "correct" function space setting for the interpolation problem from Section 7.1.

Definition 7.3. *Denote by $k \in \mathbb{N}$ and $b \in (0, \infty)$ the parameters from D.7.1. Let $\Omega := \mathbb{R}^d$. We consider the* native space

$$V := H^k(\mathbb{R}^d)$$

and equip it with the following bilinear form:

$$\forall u, v \in V: \qquad a(u, v) := \sum_{l=0}^{k} \binom{k}{l} b^{2(k-l)} \sum_{|\alpha|=l} \frac{l!}{\alpha!} \langle \mathrm{D}^\alpha u, \mathrm{D}^\alpha v\rangle_{L^2(\mathbb{R}^d)}.$$

Note that V also carries its natural inner product and norm from D.2.37, $\langle\cdot,\cdot\rangle_{H^k(\mathbb{R}^d)}$ and $\|\cdot\|_{H^k(\mathbb{R}^d)}$.

Lemma 7.4. *1. The space V satisfies the requirements from D.4.1.*

2. *The bilinear form $a(\cdot,\cdot)$ defines an inner product on V and the induced norm $\|\cdot\|_a$ is equivalent to the natural norm $\|\cdot\|_{H^k(\mathbb{R}^d)}$:*

$$\forall v \in V: \qquad C(d,k,b)^{-1}\|v\|_{H^k(\mathbb{R}^d)} \leq \|v\|_a \leq C(d,k,b)\|v\|_{H^k(\mathbb{R}^d)}.$$

In particular, $a(\cdot,\cdot)$ is continuous and coercive in the sense of D.4.2 with a constant

$$\sigma_{\text{coco}} = C(d,k,b).$$

Additionally, for all $u, v \in V$, there holds the local *stability bound*

$$|a(u,v)| \leq C(d,k,b)\|u\|_{H^k(\operatorname{supp}(v))}\|v\|_{H^k(\operatorname{supp}(u))}.$$

3. *There holds the inclusion $V \subseteq C^0(\mathbb{R}^d)$ along with the following stability bound:*

$$\forall v \in V: \qquad \|v\|_{C^0(\mathbb{R}^d)} \leq C(d,k)\|v\|_{H^k(\mathbb{R}^d)}.$$

In particular, every $v \in V$ has well-defined point values.

4. *The subspace $C_0^\infty(\mathbb{R}^d) \subseteq V$ is dense with respect to $\|\cdot\|_{H^k(\mathbb{R}^d)}$ (and also with respect to $\|\cdot\|_a$).*

5. *Denote by φ the radial basis function from D.7.1. Then, for all $x_0 \in \mathbb{R}^d$, there holds $\varphi(\cdot - x_0) \in V$. In particular, $\varphi \in V$.*

6. *There holds the following* reproducing kernel formula[2]

$$\forall v \in V: \forall x_0 \in \mathbb{R}^d: \qquad a(v, \varphi(\cdot - x_0)) = v(x_0).$$

Proof. Item 1 is trivial and items 3, 4 and 5 follows from T.2.50, T.2.41 and L.7.2, respectively.

Ad item 2: The norm equivalence $\|\cdot\|_{H^k(\mathbb{R}^d)} \eqsim \|\cdot\|_a$ follows immediately from the definition of the natural inner product $\langle\cdot,\cdot\rangle_{H^k(\mathbb{R}^d)}$ in D.2.37 and the fact that $\binom{k}{l} b^{2(k-l)} l!/\alpha! \eqsim 1$. In particular, $a(\cdot,\cdot)$ is strictly positive definite and defines an inner product on V. To see the local stability bound, we compute, for all $u, v \in V$,

$$|a(u,v)| \lesssim \sum_{|\alpha|\leq k} |\langle \mathrm{D}^\alpha u, \mathrm{D}^\alpha v\rangle_{L^2(\Omega)}| \lesssim \|u\|_{H^k(\operatorname{supp}(v))}\|v\|_{H^k(\operatorname{supp}(u))}.$$

Ad item 6: Let $v \in C_0^\infty(\Omega)$, $x_0 \in \mathbb{R}^d$ and bbreviate $\varphi_0 := \varphi(\cdot - x_0) \in V$. Using successive

[2] In particular, the triple $(V, a(\cdot,\cdot), \varphi(\cdot - \cdot))$ constitutes a so-called *reproducing kernel Hilbert space*. See, e.g., [Wen05, Section 10.1] for more details on this matter.

partial integrations, we have,

$$\begin{aligned} a(v,\varphi_0) &= \sum_{l=0}^{k} \binom{k}{l} b^{2(k-l)} \sum_{|\alpha|=l} \frac{l!}{\alpha!} \langle \mathrm{D}^\alpha v, \mathrm{D}^\alpha \varphi_0 \rangle_{L^2(\mathbb{R}^d)} \\ &= \sum_{l=0}^{k} (-1)^l \binom{k}{l} b^{2(k-l)} \left\langle \sum_{|\alpha|=l} \frac{l!}{\alpha!} \mathrm{D}^{2\alpha} v, \varphi_0 \right\rangle_{L^2(\mathbb{R}^d)} \\ &= \sum_{l=0}^{k} \binom{k}{l} b^{2(k-l)} \langle (-\Delta)^l v, \varphi_0 \rangle_{L^2(\mathbb{R}^d)} \overset{\text{L.7.2}}{=} \langle \mathrm{D}^{2k} v, \varphi_0 \rangle_{L^2(\mathbb{R}^d)} \\ &= \int_{\mathbb{R}^d} (\mathrm{D}^{2k} v)(x) \varphi(x - x_0) \, \mathrm{d}x \overset{\text{L.7.2}}{=} v(x_0). \end{aligned}$$

Now, consider a general $v \in V$. Since $C_0^\infty(\mathbb{R}^d) \subseteq V$ is dense (see item 3), we can find a sequence $(v_n)_{n\in\mathbb{N}} \subseteq C_0^\infty(\mathbb{R}^d)$ with $\|v - v_n\|_{H^k(\mathbb{R}^d)} \overset{n}{\to} 0$. In particular, by the previous argument, $a(v_n, \varphi(\cdot - x_0)) = v_n(x_0)$, so that

$$\begin{aligned} |a(v,\varphi_0) - v(x_0)| &\leq |a(v,\varphi_0) - a(v_n,\varphi_0)| + |v_n(x_0) - v(x_0)| \\ &\leq \|v - v_n\|_a \|\varphi_0\|_a + \|v_n - v\|_{C^0(\mathbb{R}^d)} \\ &\overset{\text{Items 1,2}}{\lesssim} (\|\varphi_0\|_a + 1)\|v - v_n\|_{H^k(\mathbb{R}^d)} \overset{n}{\to} 0. \end{aligned}$$

The proof is then complete. □

7.4 The space V_N

Next, we need to fix the discrete ansatz space $V_N \subseteq V$ from D.4.3 along with a basis $\{\varphi_1, \dots, \varphi_N\} \subseteq V_N$. This time, in contrast with the FEM setting (cf. D.6.3), we pick the basis functions φ_n *first* and the space V_N *last*.

Definition 7.5. *Let $\varphi \in V$ be defined as in D.7.1. Let $N \in \mathbb{N}$ and denote by $x_1, \dots, x_N \in \mathbb{R}^d$ the interpolation points from Section 7.1. For all $n \in \{1, \dots, N\}$, we set*

$$\varphi_n := \varphi(\cdot - x_n) \in V.$$

Furthermore, let

$$V_N := \operatorname{span}\{\varphi_1, \dots, \varphi_N\} \subseteq V.$$

Lemma 7.6. *The system $\{\varphi_1, \dots, \varphi_N\} \subseteq V$ is linearly independent. In particular, there holds $\dim(V_N) = N$.*

Proof. In the upcoming L.7.12, we will construct a family of linear functionals $\tilde{\lambda}_1, \dots, \tilde{\lambda}_N \in V_N^*$ with $\langle \tilde{\lambda}_n, \varphi_m \rangle_* = \delta_{nm}$. In particular, if $\boldsymbol{c} \in \mathbb{R}^N$ is such that $\sum_{m=1}^N \boldsymbol{c}_m \varphi_m \equiv 0$, then, for all $n \in \{1, \dots, N\}$,

$$0 = \langle \tilde{\lambda}_n, 0 \rangle_* = \left\langle \tilde{\lambda}_n, \sum_{m=1}^N \boldsymbol{c}_m \varphi_m \right\rangle_* = \sum_{m=1}^N \boldsymbol{c}_m \langle \tilde{\lambda}_n, \varphi_m \rangle_* = \boldsymbol{c}_n.$$

□

7.5 A weak formulation

The reproducing kernel formula from L.7.4 allows us to write the interpolation problem from Section 7.1 in weak form. Recall that, given target values $\boldsymbol{f} \in \mathbb{R}^N$, we want to find a function $u \in \operatorname{span}\{\varphi_1, \dots, \varphi_N\} = V_N$ such that $u(x_m) = \boldsymbol{f}_m$, for all $m \in \{1, \dots, N\}$. The point values of u can be expressed in the form $u(x_m) = a(u, \varphi_m)$. To encode the right-hand side in a linear functional $f \in V_N^*$, we use the dual basis $\lambda_1, \dots, \lambda_N \in V_N^*$ from D.4.10 and the coordinate mapping $\Lambda : \mathbb{R}^N \longrightarrow V_N^*$ from D.4.13:

$$f := \Lambda \boldsymbol{f} = \sum_{n=1}^{N} \boldsymbol{f}_n \lambda_n \in V_N^*.$$

Indeed, for all $m \in \{1, \dots, N\}$, we have

$$\langle f, \varphi_m \rangle_* = \sum_{n=1}^{N} \boldsymbol{f}_n \langle \lambda_n, \varphi_m \rangle_* = \boldsymbol{f}_m.$$

In particular, the interpolation problem from Section 7.1 fits into the setting from P.1.2 and L.4.5: Find $u \in V_N$ such that

$$\forall v \in V_N : \qquad a(u, v) = \langle f, v \rangle_*.$$

However, note that there is no "continuous" equivalent on the full space V (comparable to P.1.1), because the original interpolation problem from Section 7.1 already has a discrete character.

7.6 The dual basis $\lambda_1, \dots, \lambda_N$

Following the same steps as in Chapter 6, our next goal is to find a local stability bound for the dual basis $\{\lambda_1, \dots, \lambda_N\} \subseteq V_N^*$.

Definition 7.7. *Denote by $x_1, \dots, x_N \in \mathbb{R}^d$ the interpolation points from Section 7.1. We define the* separation distance

$$h_{\text{sep}} := \min_{\substack{m,n \in \{1,\dots,N\} \\ m \neq n}} \|x_n - x_m\|_2 > 0.$$

Definition 7.8. *For all $n \in \{1, \dots, N\}$, we set*

$$\Omega_n := \overline{\text{Ball}}_2(x_n, h_{\text{sep}}/2).$$

The value of h_{sep} is chosen such that these balls are essentially pairwise disjoint.

Lemma 7.9. *1. For all $m, n \in \{1, \dots, N\}$ with $m \neq n$, there holds*

$$\Omega_m^\circ \cap \Omega_n^\circ = \emptyset.$$

2. The quantity $h_{\min}$ from D.4.12 satisfies

$$h_{\min} = h_{\text{sep}}.$$

Proof. Ad item 1: Follows from L.2.7 and the fact that, for all $m, n \in \{1, \dots, N\}$ with $m \neq n$, there holds

$$h_{\text{sep}}/2 + h_{\text{sep}}/2 = h_{\text{sep}} \overset{D.7.7}{=} \min_{\substack{\tilde{m},\tilde{n} \in \{1,\dots,N\} \\ \tilde{m} \neq \tilde{n}}} \|x_{\tilde{n}} - x_{\tilde{m}}\|_2 \leq \|x_n - x_m\|_2.$$

Ad item 2: We compute

$$h_{\min} \overset{D.4.12}{=} \min_{n \in \{1,\dots,N\}} h_{\Omega_n} = \min_{n \in \{1,\dots,N\}} \operatorname{diam}_2(\overline{\operatorname{Ball}}_2(x_n, h_{\text{sep}}/2)) = h_{\text{sep}}.$$

□

Up until this point, we have not made any assumptions about the interpolation points $x_1, \dots, x_N$ that would allows us to make an inference about the spread factor σ_{sprd} from D.2.21.

Assumption 7.10. *We assume that there exists a number $\sigma_{\text{sprd}} \geq 1$, independent of N, such that*

$$2 \cdot \max_{m,n \in \{1,\dots,N\}} \|x_m - x_n\|_2 \leq \sigma_{\text{sprd}}.$$

The name "σ_{sprd}" for the constant in A.7.10 is not a coincidence.

Lemma 7.11. *Denote by $\sigma_{\text{sprd}} \geq 1$ the constant from A. 7.10. The characteristic sets $\Omega_1, \dots, \Omega_N \subseteq \mathbb{R}^d$ from D. 7.8 have shape-regularity σ_{shp}, overlap σ_{ovlp} and spread σ_{sprd}, where*

$$\sigma_{\text{shp}} = \sigma_{\text{ovlp}} = 1.$$

Proof. Since the sets $\Omega_n = \overline{\operatorname{Ball}}_2(x_n, h_{\text{sep}}/2)$ are balls with pairwise disjoint interiors (cf. L.7.9), it is clear that $\sigma_{\text{shp}} = \sigma_{\text{ovlp}} = 1$. To compute the spread, let $m, n \in \{1, \dots, N\}$, $x \in \Omega_m$ and $y \in \Omega_n$. Then,

$$\begin{aligned} \|y - x\|_2 &\leq \|y - x_n\|_2 + \|x_n - x_m\|_2 + \|x_m - x\|_2 \leq h_{\text{sep}}/2 + \|x_n - x_m\|_2 + h_{\text{sep}}/2 \\ &\overset{D.7.7}{=} \|x_n - x_m\|_2 + \min_{\substack{\tilde{m},\tilde{n} \in \{1,\dots,N\} \\ \tilde{m} \neq \tilde{n}}} \|x_{\tilde{n}} - x_{\tilde{m}}\|_2 \leq 2\|x_n - x_m\|_2 \overset{A.7.10}{\leq} \sigma_{\text{sprd}} \end{aligned}$$

and ultimately.

$$\operatorname{diam}_2\Big(\bigcup_{n=1}^{N} \Omega_n\Big) \leq \sup_{m,n \in \{1,\dots,N\}} \sup_{\substack{x \in \Omega_m \\ y \in \Omega_n}} \|y - x\|_2 \leq \sigma_{\text{sprd}}.$$

According to D.2.21, this implies the sets $\Omega_1, \dots, \Omega_N$ having spread σ_{sprd}.

□

It remains to derive a local stability bound for the dual functionals λ_n. Analogous to L.6.7, the trick is to express λ_n in terms of a suitable density function $\mu_n : \Omega \longrightarrow \mathbb{R}$.

Lemma 7.12. *Denote by h_{sep} and σ_{sprd} the quantities from D.7.7 and A.7.10. Furthermore, denote by $\{\lambda_1, \dots, \lambda_N\} \subseteq V_N^*$ the dual basis (cf. D.4.10). Then, for all $v \in V_N$, there holds the local stability bound*

$$|\langle \lambda_n, v \rangle_*| \leq C(d, k, b, \sigma_{\text{sprd}}) h_{\text{sep}}^{d/2-k} \|v\|_{H^k(\Omega_n)}.$$

Proof. Consider the following bump function:

$$\begin{aligned} \forall \|x\|_2 < 1 : \qquad \mu(x) &:= e \exp(-\tfrac{1}{1-\|x\|_2^2}), \\ \forall \|x\|_2 \geq 1 : \qquad \mu(x) &:= 0. \end{aligned}$$

It is well known (e.g., [AF03, Lemma 2.28]), that

$$\mu \in C_0^\infty(\mathbb{R}^d), \qquad \operatorname{supp}(\mu) = \overline{\text{Ball}}_2(0,1), \qquad \mu(0) = 1, \qquad 0 \leq \mu \leq 1.$$

Then, for every $n \in \{1, \dots, N\}$, we use the function

$$\mu_n := \mu(2h_{\text{sep}}^{-1}(\cdot - x_n)) \in C_0^\infty(\mathbb{R}^d) \subseteq V$$

as a density to define a linear functional $\tilde{\lambda}_n \in V_N^*$:

$$\forall v \in V_N : \qquad \langle \tilde{\lambda}_n, v \rangle_* := a(\mu_n, v).$$

Now, if we can show that $\langle \tilde{\lambda}_n, \varphi_m \rangle_* = \delta_{nm}$, then already $\tilde{\lambda}_n = \lambda_n$ by the uniqueness of the dual basis (cf. D.4.10). To this end, let $m \in \{1, \dots, N\}$. If $m = n$, then we have

$$\langle \tilde{\lambda}_n, \varphi_m \rangle_* = \langle \tilde{\lambda}_n, \varphi_n \rangle_* = a(\mu_n, \varphi_n) \overset{L.7.4}{=} \mu_n(x_n) = \mu(0) = 1 = \delta_{nm}.$$

On the other hand, if $m \neq n$, then L.7.9 tells us that $x_m \notin \Omega_n = \operatorname{supp}(\mu_n)$, so that

$$\langle \tilde{\lambda}_n, \varphi_m \rangle_* = a(\mu_n, \varphi_m) \overset{L.7.4}{=} \mu_n(x_m) = 0 = \delta_{nm}.$$

It follows that, indeed, $\tilde{\lambda}_n = \lambda_n$. In particular, for all $v \in V_N$, we have

$$|\langle \lambda_n, v \rangle_*| = |\langle \tilde{\lambda}_n, v \rangle_*| = |a(\mu_n, v)| \overset{L.7.4}{\lesssim} \|\mu_n\|_{H^k(\mathbb{R}^d)} \|v\|_{H^k(\Omega_n)}.$$

Finally, using the trivial relation $h_{\text{sep}} \leq \sigma_{\text{sprd}}/2$, the norm of μ_n can be bounded as follows:

$$\|\mu_n\|_{H^k(\mathbb{R}^d)} \eqsim \sum_{l=0}^{k} |\mu(2h_{\text{sep}}^{-1}(\cdot - x_n))|_{H^l(\mathbb{R}^d)} \lesssim \sum_{l=0}^{k} h_{\text{sep}}^{d/2-l} |\mu|_{H^l(\mathbb{R}^d)} \lesssim h_{\text{sep}}^{d/2-k}.$$

The proof is now complete. □

7.7 The discrete Caccioppoli inequality

In this section, we prove the discrete Caccioppoli inequality from A.4.19. Once again, the derivation is based on a suitable cut-off function.

Lemma 7.13. *Denote by $b \in (0,\infty)$ the parameter from D.7.1. Then, A.4.19 is satisfied with a constant*

$$\sigma_{\mathrm{Cacc}} = C(d,k,b).$$

Proof. Let $D \in \mathbb{B}$ and $u \in V_{\mathrm{sol}}(D)$ (D.4.18). Furthermore, let $B \in \mathbb{B}$ and $\delta > 0$ be such that $B^\delta \cap D = \emptyset$, where $B^\delta \in \mathbb{B}$ is the inflated box (cf. D.2.12). Our goal is to show that there exists a constant $\sigma_{\mathrm{Cacc}} \geq 1$ such that

$$\delta^k |u|_{H^k(B)} \leq \sigma_{\mathrm{Cacc}} \sum_{l=0}^{k-1} \delta^l |u|_{H^l(B^\delta)}.$$

To this end, we use the smooth cut-off function $\kappa := \kappa^\delta_{\mathbb{R}^d,B} \in C_0^\infty(\mathbb{R}^d)$ from L.5.3:

$$\mathrm{supp}(\kappa) \subseteq B^\delta, \qquad \kappa|_B \equiv 1, \qquad 0 \leq \kappa \leq 1, \qquad \forall l \in \mathbb{N}_0 : |\kappa|_{W^{l,\infty}(\mathbb{R}^d)} \lesssim \delta^{-l}.$$

Clearly, for all $m \in \{1,\dots,N\}$ with $x_m \notin B^\delta$, we have $\kappa(x_m) = 0$. On the other hand, consider an index $m \in \{1,\dots,N\}$ with $x_m \in B^\delta$. Then there must hold $m \notin \iota(D)$ (cf. D.4.17), because otherwise we would get the contradiction $x_m \in B^\delta \cap \Omega_m \subseteq B^\delta \cap D = \emptyset$. According to D.4.18, we can write u in the form[3] $u = S_N \Lambda \boldsymbol{f}$ (for some $\boldsymbol{f} \in \mathbb{R}^N$ with $\mathrm{supp}(\boldsymbol{f}) \subseteq \iota(D)$), so that

$$\begin{aligned} u(x_m) &\overset{L.7.4}{=} a(u,\varphi_m) = a(S_N\Lambda\boldsymbol{f},\varphi_m) \overset{D.4.6}{=} \langle \Lambda\boldsymbol{f},\varphi_m\rangle_* \overset{D.4.13}{=} \left\langle \sum_{n=1}^N \boldsymbol{f}_n \lambda_n, \varphi_m \right\rangle_* \\ &= \sum_{n=1}^N \boldsymbol{f}_n \langle \lambda_n,\varphi_m\rangle_* \overset{D.4.10}{=} \sum_{n=1}^N \boldsymbol{f}_n \delta_{nm} = \sum_{n\in\iota(D)} \boldsymbol{f}_n \delta_{nm} \overset{m\notin\iota(D)}{=} 0. \end{aligned}$$

In other words, the product $\kappa^2 u \in V$ vanishes at *all* interpolation points $x_1,\dots,x_N$. Now, since $u \in V_{\mathrm{sol}}(D) \subseteq V_N$, we can expand it in the form $u = \sum_{n=1}^N \boldsymbol{c}_n \varphi_n$ with certain coefficients $\boldsymbol{c}_n \in \mathbb{R}$. It follows that

$$a(\kappa^2 u, u) = a\left(\kappa^2 u, \sum_{n=1}^N \boldsymbol{c}_n \varphi_n\right) = \sum_{n=1}^N \boldsymbol{c}_n a(\kappa^2 u, \varphi_n) \overset{L.7.4}{=} \sum_{n=1}^N \boldsymbol{c}_n \kappa(x_n)^2 u(x_n) = 0.$$

On the other hand, using the definition of $a(\cdot,\cdot)$ from D.7.3 as well as Leibniz' product rule, we can expand the term $a(\kappa^2 u, u)$ explicitly (with $c_\alpha := \binom{k}{l} b^{2(k-l)} \frac{l!}{\alpha!} > 0$ and $c_{\alpha,\beta} := c_\alpha \binom{\alpha}{\beta} > 0$):

$$0 = a(\kappa^2 u, u) = \sum_{|\alpha|\leq k} c_\alpha \langle \mathrm{D}^\alpha(\kappa^2 u), \mathrm{D}^\alpha u\rangle_{L^2(\mathbb{R}^d)} = \sum_{|\alpha|\leq k} \sum_{\beta\leq\alpha} c_{\alpha,\beta} \langle \mathrm{D}^{\alpha-\beta}(\kappa^2)\mathrm{D}^\beta u, \mathrm{D}^\alpha u\rangle_{L^2(\mathbb{R}^d)}.$$

[3] The operators $S_N : V_N^* \longrightarrow V_N$ and $\Lambda : \mathbb{R}^N \longrightarrow V_N^*$ were defined in D.4.6 and D.4.13, respectively.

We transfer the summands with $\beta < \alpha$ to the other side of the equality and obtain the following expression:

$$\sum_{|\alpha|\le k} \|\kappa \mathrm{D}^\alpha u\|^2_{L^2(\mathbb{R}^d)} \lesssim \sum_{|\alpha|\le k} c_\alpha \|\kappa \mathrm{D}^\alpha u\|^2_{L^2(\mathbb{R}^d)} = -\sum_{|\alpha|\le k}\sum_{\beta<\alpha} c_{\alpha,\beta} \langle \mathrm{D}^{\alpha-\beta}(\kappa^2)\mathrm{D}^\beta u, \mathrm{D}^\alpha u\rangle_{L^2(\mathbb{R}^d)}$$

$$\lesssim \sum_{|\alpha|\le k} \Bigg[\sum_{i=1}^{d} |\langle \partial_i(\kappa^2)\mathrm{D}^{\alpha-e_i} u, \mathrm{D}^\alpha u\rangle_{L^2(\mathbb{R}^d)}| + \sum_{\substack{\beta<\alpha,\\ |\beta|\le|\alpha|-2}} |\langle \mathrm{D}^{\alpha-\beta}(\kappa^2)\mathrm{D}^\beta u, \mathrm{D}^\alpha u\rangle_{L^2(\mathbb{R}^d)}| \Bigg].$$

For the summands in the first sum, we use Young's inequality (with variable $\varepsilon > 0$):

$$\begin{aligned} |\langle \partial_i(\kappa^2)\mathrm{D}^{\alpha-e_i} u, \mathrm{D}^\alpha u\rangle_{L^2(\mathbb{R}^d)}| &= 2|\langle (\partial_i\kappa)\mathrm{D}^{\alpha-e_i} u, \kappa\mathrm{D}^\alpha u\rangle_{L^2(B^\delta)}| \\ &\lesssim \|\partial_i\kappa\|_{L^\infty(\mathbb{R}^d)} \|\mathrm{D}^{\alpha-e_i} u\|_{L^2(B^\delta)} \|\kappa\mathrm{D}^\alpha u\|_{L^2(\mathbb{R}^d)} \\ &\lesssim \delta^{-1}|u|_{H^{|\alpha|-1}(B^\delta)} \|\kappa\mathrm{D}^\alpha u\|_{L^2(\mathbb{R}^d)} \\ &\lesssim \varepsilon^{-1}\delta^{-2}|u|^2_{H^{|\alpha|-1}(B^\delta)} + \varepsilon\|\kappa\mathrm{D}^\alpha u\|^2_{L^2(\mathbb{R}^d)}. \end{aligned}$$

Note that, by choosing ε sufficiently small, we can absorb the $\mathcal{O}(\varepsilon)$-term in the left-hand side of the overall inequality.

For the summands in the second sum, we can pick an index $i \in \{1, \dots, d\}$ with $\alpha_i \ge 1$ (in the case $\alpha = 0$, the sum is empty anyways). Then, we perform partial integration with respect to the i-th coordinate:

$$\begin{aligned} &|\langle \mathrm{D}^{\alpha-\beta}(\kappa^2)\mathrm{D}^\beta u, \mathrm{D}^\alpha u\rangle_{L^2(\mathbb{R}^d)}| \\ &\quad= |\langle \mathrm{D}^{\alpha-\beta+e_i}(\kappa^2)\mathrm{D}^\beta u + \mathrm{D}^{\alpha-\beta}(\kappa^2)\mathrm{D}^{\beta+e_i} u, \mathrm{D}^{\alpha-e_i} u\rangle_{L^2(B^\delta)}| \\ &\quad\le (\|\mathrm{D}^{\alpha-\beta+e_i}(\kappa^2)\|_{L^\infty(\mathbb{R}^d)} \|\mathrm{D}^\beta u\|_{L^2(B^\delta)} \\ &\qquad\qquad + \|\mathrm{D}^{\alpha-\beta}(\kappa^2)\|_{L^\infty(\mathbb{R}^d)} \|\mathrm{D}^{\beta+e_i} u\|_{L^2(B^\delta)}) \|\mathrm{D}^{\alpha-e_i} u\|_{L^2(B^\delta)} \\ &\quad\lesssim (\delta^{-|\alpha|+|\beta|-1}|u|_{H^{|\beta|}(B^\delta)} + \delta^{-|\alpha|+|\beta|}|u|_{H^{|\beta|+1}(B^\delta)}) |u|_{H^{|\alpha|-1}(B^\delta)} \\ &\quad= \delta^{-2|\alpha|}(\delta^{|\beta|}|u|_{H^{|\beta|}(B^\delta)} + \delta^{|\beta|+1}|u|_{H^{|\beta|+1}(B^\delta)})(\delta^{|\alpha|-1}|u|_{H^{|\alpha|-1}(B^\delta)}) \\ &\quad\lesssim \delta^{-2|\alpha|} \sum_{i=0}^{|\alpha|-1} \delta^{2i}|u|^2_{H^i(B^\delta)}. \end{aligned}$$

Finally, exploiting $\kappa|_B \equiv 1$, we put everything together:

$$\delta^{2k}|u|^2_{H^k(B)} \lesssim \delta^{2k} \sum_{|\alpha|\le k} \|\kappa\mathrm{D}^\alpha u\|^2_{L^2(\mathbb{R}^d)} \lesssim \sum_{|\alpha|\le k} \delta^{2(k-|\alpha|)} \sum_{i=0}^{|\alpha|-1} \delta^{2i}|u|^2_{H^i(B^\delta)} \lesssim \sum_{i=0}^{k-1} \delta^{2i}|u|^2_{H^i(B^\delta)}.$$

This concludes the proof.

□

7.8 A corollary

All of the prerequisites for T.4.21 have now been checked and we may collect the fruits of our labour.

Corollary 7.14. *Denote by $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ the Gram matrix that corresponds to the bilinear form $a(\cdot,\cdot)$ from D.7.3 and the radial basis functions $\{\varphi_1,\dots,\varphi_N\} \subseteq V$ from D.7.5. Furthermore, denote by $h_{\mathrm{sep}} > 0$ and $\sigma_{\mathrm{sprd}} \geq 1$ the quantities from D.7.7 and A.7.10, respectively. Then, for every $r \in \mathbb{N}$, there exists an $\mathcal{H}$-matrix*

$$\boldsymbol{B}_r \in \mathcal{H}(\mathbb{P}^2, r)$$

with the following properties:

1. *The memory requirements to store $\boldsymbol{B}_r$ can be bounded by*

$$C(d,\sigma_{\mathrm{sprd}},\sigma_{\mathrm{adm}})(\sigma_{\mathrm{small}}+r)\ln(h_{\mathrm{sep}}^{-1})N.$$

2. *There exist numbers $C_0 \geq 1$ and $\sigma_{\mathrm{exp}} > 0$ of the form*

$$C_0 = C(d,k,b,\sigma_{\mathrm{sprd}},\sigma_{\mathrm{adm}}), \qquad \sigma_{\mathrm{exp}} = C(d,k,b,\sigma_{\mathrm{sprd}},\sigma_{\mathrm{adm}})^{-1},$$

such that the following error bound is satisfied:

$$\|\boldsymbol{A}^{-1} - \boldsymbol{B}_r\|_2 \leq C_0 \ln(h_{\mathrm{sep}}^{-1}) h_{\mathrm{sep}}^{d-3k} \exp(-\sigma_{\mathrm{exp}} r^{1/(d+1)}).$$

Proof. We collect the relations from L.7.4, L.7.9, L.7.12, L.7.11 and L.7.13:

$$\sigma_{\mathrm{coco}} = C(d,k,b), \qquad k_0 = k, \qquad \sigma_{\mathrm{stab}} = C(d,k,b,\sigma_{\mathrm{sprd}}) h_{\mathrm{sep}}^{d/2-k},$$
$$\sigma_{\mathrm{shp}} = 1, \qquad \sigma_{\mathrm{ovlp}} = 1, \qquad h_{\mathrm{min}} = h_{\mathrm{sep}}, \qquad \sigma_{\mathrm{Cacc}} = C(d,k,b).$$

Ad item 1: The bound on the memory complexity from T.4.21 becomes

$$C(d,\sigma_{\mathrm{shp}},\sigma_{\mathrm{sprd}},\sigma_{\mathrm{adm}})(\sigma_{\mathrm{small}}+r)\ln(h_{\mathrm{min}}^{-1})N = C(d,\sigma_{\mathrm{sprd}},\sigma_{\mathrm{adm}})(\sigma_{\mathrm{small}}+r)\ln(h_{\mathrm{sep}}^{-1})N.$$

Ad item 2: The numbers C_0 and σ_{exp} from T.4.21 read

$$C_0 \overset{T.4.21}{=} C(d,k,\Omega,\sigma_{\mathrm{coco}},\sigma_{\mathrm{shp}},\sigma_{\mathrm{sprd}},\sigma_{\mathrm{adm}}) = C(d,k,b,\sigma_{\mathrm{sprd}},\sigma_{\mathrm{adm}})$$

and

$$\sigma_{\mathrm{exp}} \overset{T.4.21}{=} C(d,k,\Omega,\sigma_{\mathrm{sprd}},\sigma_{\mathrm{adm}})^{-1}\sigma_{\mathrm{Cacc}}^{-1} = C(d,k,b,\sigma_{\mathrm{sprd}},\sigma_{\mathrm{adm}})^{-1}.$$

Finally, the prefactor of the error bound from T.4.21 turns out as

$$\sigma_{\mathrm{stab}}^2\sigma_{\mathrm{ovlp}}\sigma_{\mathrm{Cacc}}^{k_0}\ln(h_{\mathrm{min}}^{-1})h_{\mathrm{min}}^{-k_0} \lesssim (h_{\mathrm{sep}}^{d/2-k})^2\ln(h_{\mathrm{sep}}^{-1})h_{\mathrm{sep}}^{-k} = \ln(h_{\mathrm{sep}}^{-1})h_{\mathrm{sep}}^{d-3k}.$$

□

Note that C.7.14 holds true for *any* distribution of interpolation points $x_1,\dots,x_N \in \mathbb{R}^d$ (as long as the points are pairwise distinct). Using a similar notion of "gradedness" as in D.6.15, we can easily achieve the bound from C.6.17 again.

7.9 The case of a semi-definite bilinear form

The radial basis function φ from D.7.1 is very special in that its native space V (cf. D.7.3) is a Sobolev space with a proper norm $\|\cdot\|_a$. In a previous work ([AFM21b]), we also looked at the case of *thin-plate splines*, in which the bilinear form $a(\cdot,\cdot)$ is merely positive *semi*-definite. The final approximability result is very similar to T.4.21, but we had to take a few sidesteps to get there. The functional analytic setting from [AFM21b] is quite different to the one from Chapter 4, the most important difference being that the basis functions φ_n *do not* lie in the space V individually. However, for certain coefficient vectors $\boldsymbol{c} \in \mathbb{R}^N$, the linear combinations $\sum_{n=1}^N \boldsymbol{c}_n \varphi_n$ *do* lie in V. This quirk leads to the surprising non-identity

$$a\left(v, \sum_{n=1}^N \boldsymbol{c}_n \varphi_n\right) \neq \sum_{n=1}^N \boldsymbol{c}_n a(v, \varphi_n), \quad v \in V,$$

because the right-hand side is not even well-defined. In this section, we merely present the challenges of the thin-plate spline case and highlight the differences between both theories. As for proofs, we refer the interested reader to the original work [AFM21b].

For the interpolation with thin-plate splines we need a different native space:

Definition 7.15. *Let $k \in \mathbb{N}$ with $k > d/2$. We define the* Beppo-Levi space[4]

$$V := \mathrm{BL}^k(\mathbb{R}^d) := \{v \in L^1_{\mathrm{loc}}(\mathbb{R}^d) \,|\, \forall |\alpha| = k : \mathrm{D}^\alpha v \in L^2(\mathbb{R}^d)\}.$$

For all $u, v \in V$, we set

$$a(u,v) := \sum_{|\alpha|=k} \frac{k!}{\alpha!} \langle \mathrm{D}^\alpha u, \mathrm{D}^\alpha v \rangle_{L^2(\mathbb{R}^d)}, \qquad |v|_a := \sqrt{a(v,v)}.$$

Furthermore, we define the space

$$P := \mathbb{P}^{k-1}(\mathbb{R}^d) \subseteq V.$$

The function $a(\cdot,\cdot)$ defines a symmetric, positive semi-definite bilinear form on V and $|\cdot|_a$ defines a seminorm with kernel P. Furthermore, for every $v \in V$ and every open, bounded set $\omega \subseteq \mathbb{R}^d$, there holds $v \in H^k(\omega) \subseteq C^0(\omega)$.

Definition 7.16. *Denote by $x_1, \ldots, x_N \in \mathbb{R}^d$ the interpolation points from Section 7.1. We define the corresponding* evaluation operator

$$E_N : \left\{ \begin{array}{ccc} C^0(\mathbb{R}^d) & \longrightarrow & \mathbb{R}^N \\ v & \longmapsto & (v(x_n))_{n=1}^N \end{array} \right. .$$

The operator E_N is a convenient way to describe interpolation conditions:

$$(\forall m \in \{1, \ldots, N\} : u(x_m) = f(x_m)) \qquad \Leftrightarrow \qquad E_N u = E_N f.$$

[4] A common alternative name is *homogeneous Sobolev space*. For more details on theses spaces, see, e.g., [DL95], [SSN98] or [Wen05, Section 10.5].

In Section 7.5, the initial interpolation problem from Section 7.1 was rephrased as a variational problem over the discrete space $V_N = \operatorname{span}\{\varphi_1, \dots, \varphi_N\}$. This change of perspective was crucial, because it put us right into the framework of Chapter 4 (see L.4.5). In the case of thin-plate splines, the underlying variational problem looks somewhat different. Most importantly, we are now dealing with the infinite-dimensional space

$$V_0 := \{v \in V \mid E_N v = \mathbf{0}\}$$

instead of the N-dimensional space V_N.

Problem 7.17. *Let $f \in V$. Find $u \in V$, such that*

$$E_N u = E_N f, \qquad (\forall v \in V_0 : \quad a(v,u) = 0).$$

We mention that the orthogonality side conditions ensure that the solution u minimizes the seminorm $|\cdot|_a$ over the affine space $\{\tilde{u} \in V \mid E_N \tilde{u} = E_N f\}$. Furthermore, looking at the proof of L.7.13 again, we can see that these orthogonality conditions are precisely what we need to derive the discrete Caccioppoli inequality (note that $\kappa^2 u \in V_0$). P.7.17 is indeed uniquely solvable and we can express u in the form $u = u_0 + f$, where $u_0 \in V_0$ solves $(\forall v \in V_0 : a(v,u_0) = -a(v,f))$.

At first glance, the solution u looks like an infinite-dimensional object. However, it can be shown that u has the form

$$u = \sum_{n=1}^{N} \boldsymbol{c}_n \varphi_n + \sum_{l=1}^{L} \boldsymbol{d}_l \pi_l,$$

for certain $\boldsymbol{c}_n, \boldsymbol{d}_l \in \mathbb{R}$, $\varphi_n \in C^0(\mathbb{R}^d)$ and basis polynomials $\pi_1, \dots, \pi_L \in P$. This is the point where the thin-plate splines enter the stage.

Definition 7.18. *We define the* thin-plate spline

$$\begin{array}{lll} (d \in \{1,3,5,\dots\}) & \varphi(x) & := \quad C_1 \|x\|^{2k-d}, \\ (d \in \{2,4,6,\dots\}) & \varphi(x) & := \quad C_2 \|x\|^{2k-d} \ln \|x\|, \end{array}$$

where

$$C_1 := \frac{\Gamma(d/2-k)}{4^k \pi^{d/2} (k-1)!}, \qquad C_2 := \frac{(-1)^{k+(d-2)/2}}{2^{2k-1} \pi^{d/2} (k-1)! (k-d/2)!}.$$

Furthermore, for all $n \in \{1, \dots, N\}$, we set

$$\varphi_n := \varphi(\cdot - x_n).$$

We already hinted at the beginning of this section that $\varphi \notin V$, in general. In fact, in the simplest case $d = k = 1$, we have $\varphi' \eqsim |\cdot|' = \operatorname{sgn} \notin L^2(\mathbb{R})$.

Definition 7.19. *We define the following set of coefficient vectors*[5]*:*

$$\boldsymbol{C} := \{\boldsymbol{c} \in \mathbb{R}^N \mid \forall p \in P : \langle \boldsymbol{c}, E_N p \rangle_2 = 0\}.$$

[5] Up until this point, we used boldface letters only for matrices and vectors. Here, we use $\boldsymbol{C}$ for a *set* of vectors.

The relevance of the set $\boldsymbol{C}$ should become clear in the next lemma:

Lemma 7.20. *1. For all $x_0 \in \mathbb{R}^d$, there holds $\varphi(\cdot - x_0) \in C^0(\mathbb{R}^d)$. In particular, $\varphi \in C^0(\mathbb{R}^d)$.*

2. The function φ is a fundamental solution of the differential operator (cf. L.7.2)

$$\mathrm{D}^{2k} := (-\Delta)^k.$$

3. For all $\boldsymbol{c} \in \boldsymbol{C}$, there holds

$$\sum_{n=1}^{N} \boldsymbol{c}_n \varphi_n \in V.$$

Proof. See [AFM21b, Lemma 2.11.]. □

Now, in order to get the representation of the solution u of P.7.17 in terms of the translates φ_n and the polynomials π_l, we introduce the following matrices:

Definition 7.21. *We define*

$$\boldsymbol{A} := (\varphi_n(x_m))_{m,n=1}^{N} \in \mathbb{R}^{N\times N}, \qquad \boldsymbol{B} := (\pi_l(x_n))_{l=1,n=1}^{L,N} \in \mathbb{R}^{L\times N}.$$

Note that $\boldsymbol{A}$ is not a Gram matrix in the sense of D.4.8, because there is no bilinear form involved. The connection between the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and the solution u of P.7.17 is described in the following lemma:

Lemma 7.22. *Let $f \in V$ and denote by $(\boldsymbol{c}, \boldsymbol{d}) \in \mathbb{R}^N \times \mathbb{R}^L$ the unique solution of the following saddle point system:*

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B}^T \\ \boldsymbol{B} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{pmatrix} = \begin{pmatrix} E_N f \\ \boldsymbol{0} \end{pmatrix}.$$

Then, the solution $u \in V$ of P.7.17 is given by

$$u = \sum_{n=1}^{N} \boldsymbol{c}_n \varphi_n + \sum_{l=1}^{L} \boldsymbol{d}_l \pi_l \in V.$$

Proof. See [AFM21b, Lemma 3.12.]. □

Note that the first line of the system encodes the interpolation conditions $E_N u = E_N f$. In fact, the matrices $\boldsymbol{A}$ and $\boldsymbol{B}^T$ can be written in the form $\boldsymbol{A} = (E_N\varphi_1 | \dots | E_N\varphi_N)$ and $\boldsymbol{B}^T = (E_N\pi_1 | \dots | E_N\pi_L)$, so that

$$\boldsymbol{A}\boldsymbol{c} + \boldsymbol{B}^T\boldsymbol{d} = \sum_{n=1}^{N} \boldsymbol{c}_n E_N\varphi_n + \sum_{l=1}^{L} \boldsymbol{d}_l E_N\pi_l = E_N\left(\sum_{n=1}^{N} \boldsymbol{c}_n\varphi_n + \sum_{l=1}^{L} \boldsymbol{d}_l\pi_l\right) = E_N u.$$

The second line of the system reads $\boldsymbol{B}\boldsymbol{c} = \boldsymbol{0}$, which is equivalent to the condition $\boldsymbol{c} \in \boldsymbol{C}$.

The final approximation result concerns only the upper-left block of the inverse system matrix.

Theorem 7.23. *Write the inverse of the system matrix from L.7.22 in the form*

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B}^T \\ \boldsymbol{B} & \boldsymbol{0} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{S_{11}} & \boldsymbol{S_{12}} \\ \boldsymbol{S_{21}} & \boldsymbol{S_{22}} \end{pmatrix},$$

where $\boldsymbol{S_{11}} \in \mathbb{R}^{N\times N}$, $\boldsymbol{S_{21}} \in \mathbb{R}^{L\times N}$, $\boldsymbol{S_{12}} \in \mathbb{R}^{N\times L}$ *and* $\boldsymbol{S_{22}} \in \mathbb{R}^{L\times L}$. *Then, for every* $r \in \mathbb{N}$, *there exists an* $\mathcal{H}$*-matrix*

$$\boldsymbol{S}_r \in \mathcal{H}(\mathbb{P}^2, r)$$

with the following properties:

1. *The memory requirements to store* $\boldsymbol{S}_r$ *can be bounded by*

$$C(d, \sigma_{\text{sprd}}, \sigma_{\text{adm}})(\sigma_{\text{small}} + r) \ln(h_{\text{sep}}^{-1})N.$$

2. *There exist numbers* $C_0 \geq 1$ *and* $\sigma_{\text{exp}} > 0$ *of the form*

$$C_0 = C(d, k, b, \sigma_{\text{sprd}}, \sigma_{\text{adm}}), \qquad \sigma_{\text{exp}} = C(d, k, b, \sigma_{\text{sprd}}, \sigma_{\text{adm}})^{-1},$$

such that the following error bound is satisfied:

$$\|\boldsymbol{S_{11}} - \boldsymbol{S}_r\|_2 \leq C_0 \ln(h_{\text{sep}}^{-1}) h_{\text{sep}}^{d-3k} \exp(-\sigma_{\text{exp}} r^{1/(d+1)}).$$

Proof. The complexity bound is the same as in C.7.14. The error bound is essentially taken from [AFM21b, Theorem 2.18.], but *without* plugging in the assumption $1 \leq CN^{\sigma_{\text{card}}} h_{\text{sep}}^d$ from [AFM21b, Definition 2.2.]. □

7.10 Numerical examples

In this section, which is taken from [AFM21b, Section 4], we present some numerical examples to demonstrate the plausibility of C.7.14 and T.7.23. The experiments are performed in `MATLAB` ([MAT]) and `H2Lib` ([H2L]).

7.10.1 TU Wien logo

In our first example, the domain of interest is the *TU Wien* logo, which consists of the unit square in $\mathbb{R}^2$ and a series of holes in the shape of letters. We place roughly $N \approx 30.000$ interpolation points inside the logo and perform some algebraic grading (cf. D.6.15, $\sigma_{\text{grade}} = 2$) at the convex corners. As for the radial basis function, we use the thin-plate spline $\varphi(x) = \|x\|^2 \ln \|x\|$ from D.7.18 with $k = 2$.

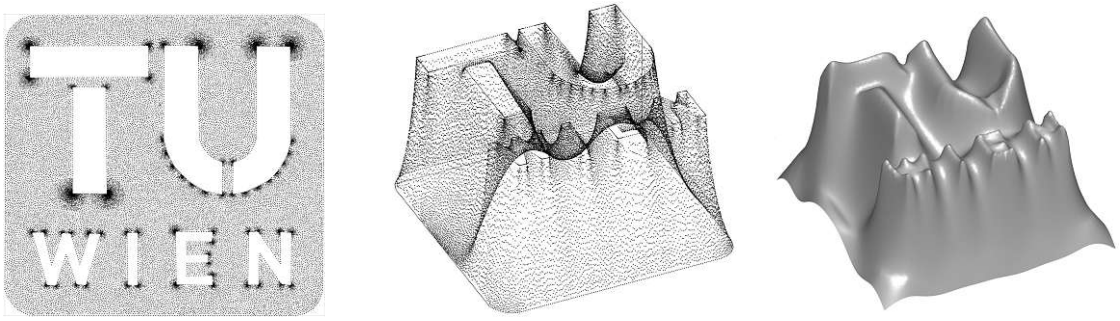

Figure 7.2: Interpolation of smooth data on a non-uniform point distribution.

The left image in Figure 7.2 shows the positions x_n of the interpolation points and the one in the middle depicts the pairs $(x_n, \boldsymbol{f}_n)$. Here, the target values $\boldsymbol{f}_n$ come from a smooth indicator function of the letters. On the right-hand side, the solution $u \in V_N$ of the interpolation problem P.7.17 is rendered.

7.10.2 A uniform grid in 2D

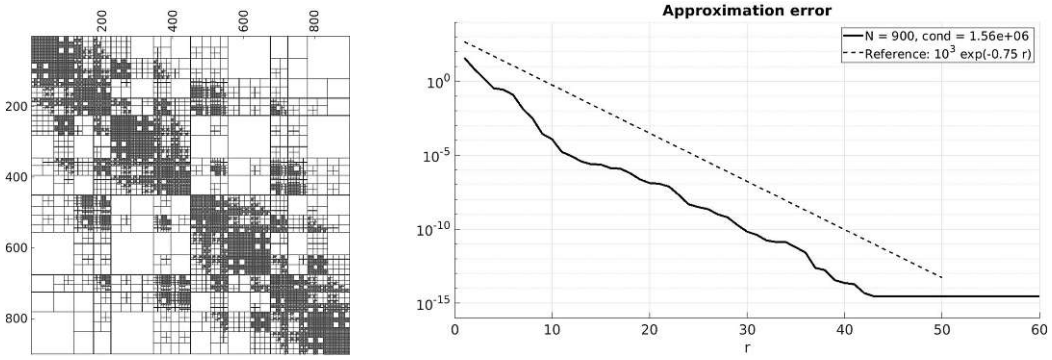

Figure 7.3: A typical hierarchical block partition and a typical error plot in 2D.

Figure 7.3 shows the results of a problem in space dimension $d = 2$. The $N = 900$ interpolation points x_n produce a regular 30×30 grid in the unit square $[0,1] \times [0,1] \subseteq \mathbb{R}^2$ (i.e., the case $\sigma_{\text{grade}} = 1$ in D.6.15). Once again, the thin plate-spline $\varphi(x) = \|x\|_2^2 \ln \|x\|_2$ with $k = 2$ is employed. In the left image, we can see a typical block partition $\mathbb{P}^2$ in the sense of C.3.42. The somewhat fractal pattern of small and admissible cluster blocks arises from the fact that we order the interpolation points in a row-wise fashion, i.e., $x_1 = (0, 0/29)$, $x_{31} = (0, 1/29)$, $x_{61} = (0, 2/29)$, et cetera.

The right-hand image is empirical evidence that the error bound in T.7.7.23 is correct. To generate this plot, we use the same strategy as in Section 6.8.1, i.e., the interpolation matrix $\left(\begin{smallmatrix} \boldsymbol{A} & \boldsymbol{B}^T \\ \boldsymbol{B} & \boldsymbol{0} \end{smallmatrix}\right)$ is inverted exactly and we compute block-wise truncated SVDs for the main block $\boldsymbol{S_{11}}$ (cf. T.7.7.23). The semi-logarithmic error plot depicts the computable error bound

from Section 6.8.1 along with a dashed reference line. The apparent similarity suggests a relation of the form $\|\boldsymbol{S_{11}} - \boldsymbol{S}_r\|_2 \lesssim C(N)\exp(-\sigma_{\text{exp}} r)$, which is again better than our theoretical prediction $C(N)\exp(-\sigma_{\text{exp}} r^{1/3})$.

On a side note, we mention that the standard 16-digit precision arithmetic in `MATLAB` is not enough to generate a conclusive error plot. As is well-established in the literature (e.g., [Wen05, Chapter 12]), the condition number of the interpolation matrix $\left(\begin{smallmatrix} \boldsymbol{A} & \boldsymbol{B}^T \\ \boldsymbol{B} & \boldsymbol{0} \end{smallmatrix}\right)$ scales very poorly with respect to the separation distance h_{sep} introduced in D.7.7. To overcome this fundamental problem, we use `MATLAB`'s *variable-precision arithmetic* `vpa(...)` with 32 digits. This brute-force approach allows us to carry out the explicit matrix inversion with sufficient accuracy.

7.10.3 A uniform grid in 3D

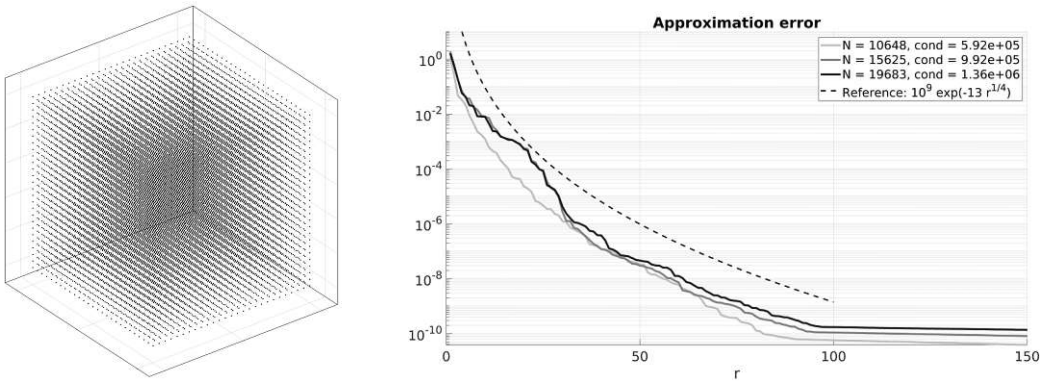

Figure 7.4: A comparison of different problem sizes N for a uniform 3D grid.

The next example, Figure 7.4, covers the case $d = 3$ and a uniform point distribution in the unit cube $[0,1] \times [0,1] \times [0,1] \subseteq \mathbb{R}^3$, visualized in the left image. This time, we use the Bessel potential $\varphi(x) = e^{-\|x\|_2}$ from D.7.1 as the basis function (with $k = 2$ and $b = 1$). The error plot shows a comparison between $N \approx 10.000$, $N \approx 15.000$ and $N \approx 20.000$ interpolation points, as well as a reference curve of the form $r \mapsto C\exp(-\sigma_{\text{exp}} r^{1/4})$. In accordance with C.7.14, the empirical decay rate seems to be independent of the problem size N.

7.10.4 An algebraically graded grid in 3D

In Figure 7.5, we investigate the influence of the grading parameter σ_{grade} from D.6.15 on the error decay rate in $d = 3$ space dimensions. Once again, we use the Bessel potential $\varphi(x) = e^{-\|x\|_2}$ as the basis function (D.7.1, $k = 2$, $b = 1$). The error plot compares the cases $\sigma_{\text{card}} \in \{1, 2, 3\}$, where $\sigma_{\text{card}} = 1$ is a uniform grid and $\sigma_{\text{card}} = 3$ is "strongly" graded towards the origin $0 \in \mathbb{R}^3$. The problem size $N \approx 10.000$ is held constant throughout all three runs. The plot suggests that the constant σ_{exp} from the error bound in C.7.14 is independent of the grading parameter σ_{card}.

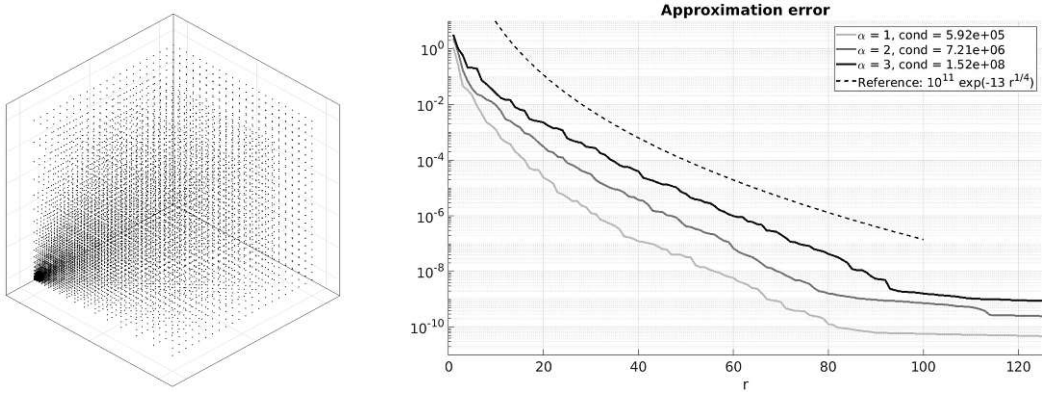

Figure 7.5: Experimenting with an algebraically graded grid in 3D.

7.10.5 Some $\mathcal{H}$-arithmetic[6]

Previous numerical results have established that $\mathcal{H}$-matrix arithmetic is a viable tool for solving RBF interpolation problems (e.g., [LM17], [LBW19], [LBW20]). In the following, we use the library `H2Lib` ([H2L]) for this purpose.

Here, we look at the thin-plate splines $\varphi(x) = \|x\|_2^2 \ln \|x\|_2$ in 2D and the Bessel potential $\varphi(x) = e^{-\|x\|_2}$ in 3D. The bessel potentials are treated as in Section 6.8.3, i.e., we compute an $\mathcal{H}$-Cholesky factorization $\boldsymbol{A} \approx \boldsymbol{A}_{\mathcal{H}} \approx \boldsymbol{L}_{\mathcal{H}} \boldsymbol{L}_{\mathcal{H}}^T$ and invert it. (The intermediate matrix $\boldsymbol{A}_{\mathcal{H}}$ is necessary, because $\boldsymbol{A}$ is fully populated.) The $\mathcal{H}$-Cholesky factorization is discussed, e.g., in [Beb07].

In the case of the thin-plate splines, the saddle point structure of the interpolation matrix makes this approach infeasible. Instead, we follow the approach from [BLB12] and [LBW19] and employ the augmented Lagrangian $\boldsymbol{A} + \gamma \boldsymbol{B}^T \boldsymbol{B}$ ($\gamma > 0$), which is SPD and therefore amenable to an $\mathcal{H}$-matrix inversion.

[6] This experiment was performed by Dr. Markus Faustmann, a co-author of [AFM21b].

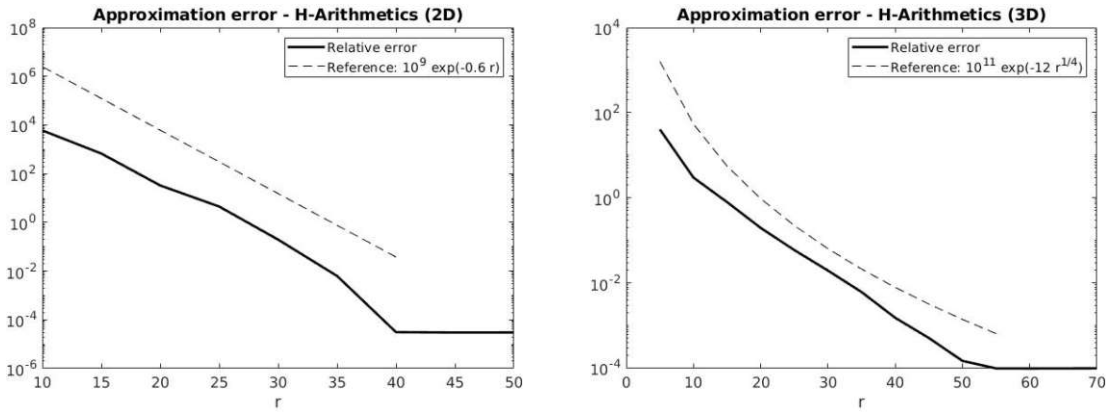

Figure 7.6: Experiment using $\mathcal{H}$-arithmetics to approximate the inverse system matrix. Left: 2D thin-plate splines. Right: 3D Bessel potential.

In Figure 7.6, we plot the error measure from Section 6.8.3 for the relative error. In the 2D case, we work with $N = 10.000$ interpolation points on the unit square in a uniform grid (i.e. $\sigma_{\text{grade}} = 1$ in D.6.15). The parameter in the definition of the augmented Lagrangian is set to $\gamma := 1$. In the 3D case, we take $N \approx 4.100$ uniformly distributed points in the unit cube. Once again, we observe exponential convergence as predicted by C.7.14 and T.7.23. However, we mention that the error flattens out before we arrive at the level of machine precision, which is most likely attributable to the initial approximation $\boldsymbol{A} \approx \boldsymbol{A}_{\mathcal{H}}$ by interpolation.

Bibliography

[AF03] Robert A. Adams and John J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[AFM21a] N. Angleitner, M. Faustmann, and J.M. Melenk. Approximating inverse FEM matrices on non-uniform meshes with $\mathcal{H}$-matrices. *Calcolo*, 58(3):Paper No. 31, 36, 2021.

[AFM21b] N. Angleitner, M. Faustmann, and J.M. Melenk. $\mathcal{H}$-inverses for RBF interpolation. `https://arxiv.org/abs/2109.05763`. 2021.

[AFM22] N. Angleitner, M. Faustmann, and J.M. Melenk. Exponential meshes and $\mathcal{H}$-matrices. `https://arxiv.org/abs/2203.09925`. 2022.

[AS61] N. Aronszajn and K.T. Smith. Theory of Bessel potentials. I. *Ann. Inst. Fourier (Grenoble)*, 11:385–475, 1961.

[Atk89] Kendall E. Atkinson. *An introduction to numerical analysis*. John Wiley & Sons, Inc., New York, second edition, 1989.

[Axl15] Sheldon Axler. *Linear algebra done right*. Undergraduate Texts in Mathematics. Springer, Cham, third edition, 2015.

[Beb06] Mario Bebendorf. Approximate inverse preconditioning of finite element discretizations of elliptic operators with nonsmooth coefficients. *SIAM J. Matrix Anal. Appl.*, 27(4):909–929, 2006.

[Beb07] Mario Bebendorf. Why finite element discretizations can be factored by triangular hierarchical matrices. *SIAM J. Numer. Anal.*, 45(4):1472–1494, 2007.

[BH03] Mario Bebendorf and Wolfgang Hackbusch. Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients. *Numer. Math.*, 95(1):1–28, 2003.

[BH71] J. H. Bramble and S. R. Hilbert. Bounds for a class of linear functionals with applications to Hermite interpolation. *Numer. Math.*, 16:362–369, 1970/71.

[BLB12] Steffen Börm and Sabine Le Borne. $\mathcal{H}$-LU factorization in preconditioners for augmented Lagrangian and grad-div stabilized saddle point systems. *Internat. J. Numer. Methods Fluids*, 68(1):83–98, 2012.

[BM97] I. Babuška and J. M. Melenk. The partition of unity method. *Internat. J. Numer. Methods Engrg.*, 40(4):727–758, 1997.

[Bör10] Steffen Börm. Approximation of solution operators of elliptic partial differential equations by $\mathcal{H}$- and $\mathcal{H}^2$-matrices. *Numer. Math.*, 115(2):165–193, 2010.

[Bra13] Dietrich Braess. *Finite Elemente : Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer-Lehrbuch Masterclass. Springer Berlin Heidelberg, Berlin, Heidelberg, 5. aufl. 2013. edition, 2013.

[Bre11] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[BS08] Susanne C. Brenner and L. Ridgway Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[CFPP14] C. Carstensen, M. Feischl, M. Page, and D. Praetorius. Axioms of adaptivity. *Comput. Math. Appl.*, 67(6):1195–1253, 2014.

[Cia78] P.G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.

[CS96] G. M. Constantine and T. H. Savits. A multivariate Faà di Bruno formula with applications. *Trans. Amer. Math. Soc.*, 348(2):503–520, 1996.

[DER17] I. S. Duff, A. M. Erisman, and J. K. Reid. *Direct methods for sparse matrices*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, second edition, 2017.

[Dit92] Z. Ditzian. Multivariate Bernstein and Markov inequalities. *J. Approx. Theory*, 70(3):273–283, 1992.

[DL95] J. Deny and J.L. Lions. Les espaces du type de Beppo Levi. *Ann. Inst. Fourier (Grenoble)*, 5:305–370 (1955), 195.

[DO95] C Armando Duarte and JT Oden. *A review of some meshless methods to solve partial differential equations*. Texas Institute for Computational and Applied Mathematics Austin, TX, 1995.

[DS00] J. Dongarra and F. Sullivan. Guest editors introduction to the top 10 algorithms. *Computing in Science and Engineering*, 2(1):22–23, 2000.

[EG00] H. Edelsbrunner and D. R. Grayson. Edgewise subdivision of a simplex. volume 24, pages 707–719. 2000. ACM Symposium on Computational Geometry (Miami, FL, 1999).

[EG04] Alexandre Ern and Jean-Luc Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[EG06] Alexandre Ern and Jean-Luc Guermond. Evaluation of the condition number in linear systems arising in finite element approximations. *M2AN Math. Model. Numer. Anal.*, 40(1):29–48, 2006.

[EMM+21] Christoph Erath, Lorenzo Mascotto, Jens Markus Melenk, Ilaria Perugia, and Alexander Rieder. Mortar coupling of *hp*-discontinuous galerkin and boundary element methods for the helmholtz equation, 2021.

[Epp13] James F. Epperson. *An introduction to numerical methods and analysis*. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2013.

[Eva10] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[Fau15] Markus Faustmann. *Approximation inverser Finite Elemente- und Randelementematrizen mittels hierarchischen Matrizen*. PhD thesis, Technische Universität Wien, 2015.

[Fle77] Wendell Fleming. *Functions of several variables*. Undergraduate Texts in Mathematics. Springer-Verlag, New York-Heidelberg, second edition, 1977.

[FMPR15] T. Führer, J. M. Melenk, D. Praetorius, and A. Rieder. Optimal additive Schwarz methods for the *hp*-BEM: the hypersingular integral operator in 3D on locally refined meshes. *Comput. Math. Appl.*, 70(7):1583–1605, 2015.

[Fol02] G.B. Folland. *Advanced Calculus*. Featured Titles for Advanced Calculus Series. Prentice Hall, 2002.

[GH96] Mariano Giaquinta and Stefan Hildebrandt. *Calculus of variations. II*, volume 311 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1996. The Hamiltonian formalism.

[GHLB04] L. Grasedyck, W. Hackbusch, and S. Le Borne. Adaptive geometrically balanced clustering of $\mathcal{H}$-matrices. *Computing*, 73(1):1–23, 2004.

[Gia83] Mariano Giaquinta. *Multiple integrals in the calculus of variations and nonlinear elliptic systems*, volume 105 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, NJ, 1983.

[GR87] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comput. Phys.*, 73(2):325–348, 1987.

[GR07] I.S. Gradshteyn and I.M. Ryzhik. *Table of integrals, series, and products*. Elsevier/Academic Press, Amsterdam, seventh edition, 2007. Translated from the Russian, Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger, With one CD-ROM (Windows, Macintosh and UNIX).

[Gra01] Lars Grasedyck. *Theorie und Anwendungen Hierarchischer Matrizen*. PhD thesis, Christian-Albrechts-Universität zu Kiel, 2001.

[Gre97] Anne Greenbaum. *Iterative methods for solving linear systems*, volume 17 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[Gri85] P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985.

[GT01] David Gilbarg and Neil S. Trudinger. *Elliptic partial differential equations of second order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

[GVL13] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

[H2L] H2Lib. Available at `http://www.h2lib.org/`.

[Hac99] W. Hackbusch. A sparse matrix arithmetic based on $\mathcal{H}$-matrices. I. Introduction to $\mathcal{H}$-matrices. *Computing*, 62(2):89–108, 1999.

[Hac09] Wolfgang Hackbusch. *Hierarchische Matrizen: Algorithmen und Analysis*. Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2009.

[HJ13] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[HK00a] W. Hackbusch and B. N. Khoromskij. A sparse $\mathcal{H}$-matrix arithmetic: general complexity estimates. volume 125, pages 479–501. 2000. Numerical analysis 2000, Vol. VI, Ordinary differential equations and integral equations.

[HK00b] W. Hackbusch and B. N. Khoromskij. A sparse $\mathcal{H}$-matrix arithmetic. II. Application to multi-dimensional problems. *Computing*, 64(1):21–47, 2000.

[HN89] W. Hackbusch and Z. P. Nowak. On the fast matrix multiplication in the boundary element method by panel clustering. *Numer. Math.*, 54(4):463–491, 1989.

[Hör90] Lars Hörmander. *The analysis of linear partial differential operators. I*, volume 256 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 1990. Distribution theory and Fourier analysis.

[KM02] B. N. Khoromskij and J. M. Melenk. An efficient direct solver for the boundary concentrated FEM in 2D. *Computing*, 69(2):91–117, 2002.

[KM03] B. N. Khoromskij and J. M. Melenk. Boundary concentrated finite element methods. *SIAM J. Numer. Anal.*, 41(1):1–36, 2003.

[LB13] Mats G. Larson and Fredrik Bengzon. *The finite element method: theory, implementation, and applications*, volume 10 of *Texts in Computational Science and Engineering*. Springer, Heidelberg, 2013.

[LBW19] Sabine Le Borne and Michael Wende. Iterative solution of saddle-point systems from radial basis function (RBF) interpolation. *SIAM J. Sci. Comput.*, 41(3):A1706–A1732, 2019.

[LBW20] Sabine Le Borne and Michael Wende. Multilevel interpolation of scattered data using $\mathcal{H}$-matrices. *Numer. Algorithms*, 85(4):1175–1193, 2020.

[Leo17] Giovanni Leoni. *A first course in Sobolev spaces*, volume 181 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2017.

[LM17] M. Löhndorf and J. M. Melenk. On thin plate spline interpolation. In *Spectral and high order methods for partial differential equations—ICOSAHOM 2016*, volume 119 of *Lect. Notes Comput. Sci. Eng.*, pages 451–466. Springer, Cham, 2017.

[MAT] MATLAB. Available at `https://mathworks.com/`.

[Maz85] Vladimir G. Maz'ja. *Sobolev spaces*. Springer Series in Soviet Mathematics. Springer-Verlag, Berlin, 1985. Translated from the Russian by T. O. Shaposhnikova.

[McL00] William McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, Cambridge, 2000.

[MR20] J. M. Melenk and C. Rojik. On commuting p-version projection-based interpolation on tetrahedra. *Math. Comp.*, 89(321):45–87, 2020.

[MS64] Norman G. Meyers and James Serrin. $H = W$. *Proc. Nat. Acad. Sci. U.S.A.*, 51:1055–1056, 1964.

[Mv98] Dragiša Mitrović and Darko Žubrinić. *Fundamentals of applied functional analysis*, volume 91 of *Pitman Monographs and Surveys in Pure and Applied Mathematics*. Longman, Harlow, 1998. Distributions—Sobolev spaces—nonlinear elliptic equations.

[Neč12] Jindřich Nečas. *Direct methods in the theory of elliptic equations*. Springer Monographs in Mathematics. Springer, Heidelberg, 2012. Translated from the 1967 French original by Gerard Tronel and Alois Kufner, Editorial coordination and preface by Šárka Nečasová and a contribution by Christian G. Simader.

[NGS] NGSolve. Available at `https://ngsolve.org/`.

[Rud87] Walter Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987.

[SSN98] H. Sohr and M. Specovius-Neugebauer. The Stokes problem for exterior domains in homogeneous Sobolev spaces. In *Theory of the Navier-Stokes equations*, volume 47 of *Ser. Adv. Math. Appl. Sci.*, pages 185–205. World Sci. Publ., River Edge, NJ, 1998.

[Ste70] Elias M. Stein. *Singular integrals and differentiability properties of functions*. Princeton Mathematical Series, No. 30. Princeton University Press, Princeton, N.J., 1970.

[Ste08] Rob Stevenson. The completion of locally refined simplicial partitions created by bisection. *Math. Comp.*, 77(261):227–241, 2008.

[Str80] Gilbert Strang. *Linear algebra and its applications*. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, second edition, 1980.

[TB97] Lloyd N. Trefethen and David Bau, III. *Numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[Wen05] Holger Wendland. *Scattered data approximation*, volume 17 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2005.

[WQ19] Hui Wang and Qing-Hua Qin. *Methods of fundamental solutions in solid mechanics*. Elsevier, 2019.

[Yos80] K. Yoshida. *Functional Analysis*. Classics in mathematics / Springer. World Publishing Company, 1980.

Lebenslauf

Persönliche Daten

Name:	Niklas Angleitner
Geburtsdatum:	[redacted]
Geburtsort:	[redacted]
Nationalität:	[redacted]
Email:	niklas.angleitner@tuwien.ac.at

Ausbildung

12.2015 - 02.2022:	**Doktoratsstudium**, Technische Mathematik, TU Wien.
01.2013 - 11.2015:	**Masterstudium**, Technische Mathematik, TU Wien.
03.2009 - 01.2013:	**Bachelorstudium**, Technische Mathematik, TU Wien.
09.2008 - 02.2009:	**Grundwehrdienst**, Hörsching.
06.2008:	**Matura** in Deutsch, Englisch, Mathematik und Biologie.
2000 - 2008:	**Schule**, BG/BRG Anton-Bruckner-Straße, Wels.

Wissenschaftliche Publikationen

1. N. Angleitner, M. Faustmann, and J.M. Melenk: *Approximating inverse FEM matrices on non-uniform meshes with $\mathcal{H}$-matrices*, Calcolo 58, 2021.
2. N. Angleitner, M. Faustmann, and J.M. Melenk: *$\mathcal{H}$-inverses for RBF interpolation*, `https://arxiv.org/abs/2109.05763`, 2021.
3. N. Angleitner, M. Faustmann, and J.M. Melenk: *Exponential meshes and $\mathcal{H}$-matrices*, `https://arxiv.org/abs/2203.09925`, 2022.

Wien, am 25. Mai 2022