Adaptive Lösungskonzepte – Adaptive Solution Concepts

Prof. Dr. Wolfgang Dahmen, WS 11/12

DRAFT VERSION, February 9, 2017

Contents

1 Introduction

Adaptive solution concepts aim at achieving a desired "quality" of a numerical solution (usually in terms of error or deviation bounds with brespect to a given norm) at possibly low numerical cost by exploiting information acquired during the solution process, thereby extending the scope of computability.

In a real life application context such methods draw from several disciplines such as computer science, physical modeling, and mathematics. This lecture is concerned with the relevant mathematical foundations.

The problem areas of interest can be roughly divided into the following categories:

- Capturing (mathematical) objects that are given *explicitly* in terms of measurements or observations, (imaging, machine learning).
- Capturing (mathematical) objects that are given only *implicitly* as solutions of differential or integral equations, in brief operator equations.

In each of these categories one may encounter different regimes, namely

- Low spatial dimension: classical models in continuum mechanics, up to three spatial and a time variable. Computational complexity is then essentially governed by the *regularity* (measure of smoothness in terms of differentiability or related properties) of the searched objects;
- high spatial dimension: one tries to capture functions of *many* (even several thousand) variables. Examples are: models in high-dimensional phase space (Schrödinger equation for electronic structure calculations, Fokker-Planck equations, parametric PDEs, stemming e.g. from stochastic PDEs, data-mining, etc.). The comploexity of solution concepts is now no longer be goverend by regularity alone. Resulting effects are commionly referred to as *curse of dimensionality*. It can be mitigated or avoided by using *sparsity*. This means roughly speaking that the searched object can be approximated within a desired accuracy tolerance by relatively few terms of a (possibly problem dependent) basis (or more generally dictionary).

Methodologies addressing these issues draw from several mathematical areas such as numerical analysis, partial differential equations (PDEs), harmonic analysis, functional analysis, statistical estimation. In this course, I try to bring out the basic ideas and the interconnections between the various conceptual platforms. I do not aim at providing all details but rather at working towards a common conceptual platform that helps to place the various aspects in a propere context and navigate between them.

2 The Role of Nonlinear Approximation Theory, Examples and Basics

We discuss first some basic principles from approximation theory on which the analysis and understanding of adaptive methods are based upon. We distinguish two basic paradigms:

- linear methods and
- nonlinear methods.

Adaptive methods are instances of nonlinear schemes.

2.1 A Guiding Example: Approximation by Piecewise Constants

2.1.1 Linear Methods

Setting:

$$\begin{aligned} \Omega &= (0,1) \\ \mathcal{P}_n &= \left\{ \Big[\frac{k-1}{n}, \frac{k}{n}\Big), k = 1, ..., n \right\} \qquad \text{uniform partitions} \end{aligned}$$

Approximation system:

$$\mathcal{S}_n := \left\{ g = \sum_{I \in \mathcal{P}_n} c_I \chi_I : c_I \in \mathbb{R}, I \in \mathcal{P}_n \right\} \tag{2.1.1}$$

where

$$\chi_I(x) = \begin{cases} 1, & \text{if } x \in I\,, \\ 0, & \text{else}\,. \end{cases}$$

Task: understand the error

$$e_n(f) := \inf_{g \in \mathcal{S}_n} \|f - g\|_{L_\infty(\Omega)}\ , \tag{2.1.2}$$

where

$$\|f\|_{L_\infty(\Omega)} = \sup_{x \in \Omega} |f(x)|\ .$$

Questions:

- How fast can $e_n(f)$ decay as $n \to \infty$?
- On which properties of f does the decay of $e_n(f)$ depend?

Basic concept: **Approximation classes**

$$\mathcal{A}^r\left((\mathcal{S}_n)_{n\in\mathbb{N}}, C(\Omega)\right) = \{f \in C(\Omega) : \underbrace{\sup_{n\in\mathbb{N}} n^r e_n(f)}_{:=|f|_{\mathcal{A}^r}} < \infty\} \tag{2.1.3}$$

$$\|f\|_{\mathcal{A}^r} =: \|f\|_{L_\infty(\Omega)} + |f|_{\mathcal{A}^r}$$

Exercise 2.1.1. *Show that* $\|f\|_{\mathcal{A}^r}$ *is a quasi-norm for the space* $\mathcal{A}^r\left((\mathcal{S}_n)_{n\in\mathbb{N}}, C(\Omega)\right)$.

How to read membership to $\mathcal{A}^r\left((\mathcal{S}_n)_{n\in\mathbb{N}}, C(\Omega)\right)$?

$$f \in \mathcal{A}^r\left((\mathcal{S}_n)_{n\in\mathbb{N}}, C(\Omega)\right) \quad \Rightarrow \quad \inf_{g\in\mathcal{S}_n} \|f-g\|_{L_\infty(\Omega)} \le |f|_{\mathcal{A}^r}\, n^{-r}$$

Problem:

Characterize $\mathcal{A}^r\left((\mathcal{S}_n)_{n\in\mathbb{N}}, C(\Omega)\right)$ in terms of "intrinsic" properties of f such as smoothness properties – this is a central question in approximation theory.

Hölder continuity:

$$\mathrm{Lip}\,(r, C(\Omega)) := \{f \in C(\Omega) : \exists C > 0\colon \; |f(x) - f(y)| \le C\,|x-y|^r\,, x, y \in \Omega\} \tag{2.1.4}$$

$$\begin{aligned} \|f\|_{\mathrm{Lip}(r,C(\Omega))} &:= \|f\|_{L_\infty(\Omega)} + |f|_{\mathrm{Lip}(r,C(\Omega))} \\ |f|_{\mathrm{Lip}(r,C(\Omega))} &:= \inf\left\{C : \frac{|f(x)-f(y)|}{|x-y|^r} \le C \;\forall\, x, y \in \Omega\right\} \end{aligned}$$

(Common alternative notation: $\mathrm{Lip}\,(r, C(\Omega)) = C^r(\Omega)$, $r < 1$)

Theorem 2.1.1. *One has*

$$\begin{aligned} &\mathcal{A}^r\left((\mathcal{S}_n)_{n\in\mathbb{N}}, C(\Omega)\right) = \mathrm{Lip}\,(r, C(\Omega))\,, \; 0 < r \le 1 \textit{ and} \\ &\exists\, c, C > 0\colon \quad c\,\|f\|_{\mathrm{Lip}(r,C(\Omega))} \le \|f\|_{\mathcal{A}^r} \le C\,\|f\|_{\mathrm{Lip}(r,C(\Omega))}\,. \end{aligned} \tag{2.1.5}$$

Thus smoothness is characterized by approximability.

Proof. 1) "Direct Theorem": Show $\mathrm{Lip}\,(r, C(\Omega)) \subseteq \mathcal{A}^r\left((\mathcal{S}_n)_{n\in\mathbb{N}}, C(\Omega)\right)$, i.e.,

$$\|f\|_{\mathcal{A}^r} \overset{!}{\le} C\,\|f\|_{\mathrm{Lip}(r,C(\Omega))}\,. \tag{2.1.6}$$

In other words we need to find a constant C and for each $n \in \mathbb{N}$ a $g = g_n \in \mathcal{S}_n$ such that $n^r\|f-g\|_{L_\infty(\Omega)} \le C|f|_{\mathrm{Lip}(r,C(\Omega))}$. Given $f \in \mathrm{Lip}\,(r, C(\Omega))$ let $g := \sum_{I\in\mathcal{P}_n} f(\xi_I)\chi_I \in \mathcal{S}_n$, where ξ_I is the midpoint of the interval I. Then we have

$$\begin{aligned} \|f-g\|_{L_\infty(\Omega)} &= \max_{I\in\mathcal{P}_n}\max_{x\in I} |f(x) - f(\xi_I)| \\ &\le \max_{I\in\mathcal{P}_n}\max_{x\in I} \frac{|f(x)-f(\xi_I)|}{|x-\xi_I|^r}\,|x-\xi_I|^r\,, \end{aligned}$$

and by Hölder continuity

$$\begin{aligned}&\le \max_{I\in\mathcal{P}_n}\left(\frac{|I|}{2}\right)^r |f|_{\mathrm{Lip}(r,C(\Omega))}\\ &= (2n)^{-r}\,|f|_{\mathrm{Lip}(r,C(\Omega))}\ .\end{aligned}$$

Therefore

$$e_n(f) \le n^{-r}2^{-r}\,|f|_{\mathrm{Lip}(r,C(\Omega))}$$

and

$$|f|_{\mathcal{A}^r} \le 2^{-r}\,|f|_{\mathrm{Lip}(r,C(\Omega))}\,.$$

This is (2.1.6) which yields

$$\mathrm{Lip}\,(r, C(\Omega)) \subseteq \mathcal{A}^r\left((\mathcal{S}_n)_{n\in\mathbb{N}}, C(\Omega)\right)\ .$$

2) "Inverse Theorem": Show $\mathcal{A}^r\left((\mathcal{S}_n)_{n\in\mathbb{N}}, C(\Omega)\right) \subseteq \mathrm{Lip}\,(r, C(\Omega))$ which means

$$c\,\|f\|_{\mathrm{Lip}(r,C(\Omega))} \overset{!}{\le} \|f\|_{\mathcal{A}^r}\,. \tag{2.1.7}$$

To see this, pick $n \ge 2$ and any $x, y \in \Omega$ such that $\frac{1}{n} \le |x-y| \le \frac{1}{n-1}$. Let $g \in \mathcal{S}_n$ such that $\|f-g\|_{L_\infty(\Omega)} \le |f|_{\mathcal{A}^r}\,n^{-r}$. We consider two cases.

Case a): $x, y \in I \in \mathcal{P}_n$.

$$\begin{aligned}|f(x)-f(y)| &= |f(x)-g(x)+g(x)-f(y)|\\ &= |f(x)-g(x)+g(y)-f(y)|\\ &\le |f(x)-g(x)|+|g(y)-f(y)|\\ &\le 2\|f-g\|_{L_\infty(\Omega)},\end{aligned}$$

Hence,

$$\frac{|f(x)-f(y)|}{|x-y|^r} \le 2\|f-g\|_{L_\infty(\Omega)}|x-y|^{-r} \le 2\|f-g\|_{L_\infty(\Omega)}n^{-r} \le 2|f|_{\mathcal{A}^r},$$

by our choice of g. This proves the claim in case a).

Case b): x, y do not belong to the same interval. In this case, since $n \ge 2$, they must belong to adjacent intervals. Let a be the shared endpoint of these two intervals, then

$$|f(x)-f(y)| \le |f(x)-f(a)|+|f(a)-f(y)|\ ,$$

where w.l.o.g. $x < y$. Applying case a) to (x, a) and (a, y) we obtain

$$\frac{|f(x) - f(y)|}{|x - y|^r} \leq \frac{|f(x) - f(a)|}{|x - a|^r} + \frac{|f(y) - f(a)|}{|y - a|^r} \leq 4\,|f|_{\mathcal{A}^r}\,, \tag{2.1.8}$$

and hence $|f|_{\mathrm{Lip}(r,C(\Omega))} \leq 4\,|f|_{\mathcal{A}^r}$. □

Remark 2.1.1. *This is an example of* **linear approximation***, although*

$$f \to g^* = \operatorname*{argmin}_{g \in \mathcal{S}_n} \|f - g\|_{L_\infty(\Omega)}$$

is a nonlinear mapping. We speak of linear approximation because we approximate from linear spaces given independently of f.

Remark 2.1.2. $r = 1$ *is the highest possible order achieved by approximation from* $\mathcal{S}_n$ *in the sense that*

$$e_n(f) = \mathcal{O}(n^{-r}) \textit{ for some } r > 1 \quad \Rightarrow \quad f = \textit{const.} \tag{2.1.9}$$

This is called saturation.

Proof. See Corollary 2.1.1 later below. □

2.1.2 Nonlinear Approximation

Next we discuss an example of **nonlinear approximation**. Consider

$$\Sigma_n = \{g : \exists \text{ partition } \mathcal{P} \text{ of } \Omega, \#\mathcal{P} \leq n, g = \sum_{I \in \mathcal{P}} c_I \chi_I, c_I \in \mathbb{R}\}. \tag{2.1.10}$$

Note:

$$f, g \in \Sigma_n \quad \Rightarrow \quad f + g \in \Sigma_{2n} \tag{2.1.11}$$

that is Σ_n is not a linear space. Now we are interested in

$$\sigma_n(f) := \inf_{g \in \Sigma_n} \|f - g\|_{L_\infty(\Omega)}$$

which is a *nonlinear* best approximation error. We can still define related approximation spaces

$$\begin{aligned} \mathcal{A}^r\left((\Sigma_n)_{n\in\mathbb{N}}, C(\Omega)\right) &= \{f \in C(\Omega) : \sup_{n\in\mathbb{N}} n^r \sigma_n(f) < \infty\} \\ \|f\|_{\mathcal{A}^r} := \|f\|_{L_\infty(\Omega)} + |f|_{\mathcal{A}^r}\,, &\qquad |f|_{\mathcal{A}^r} := \sup_{n\in\mathbb{N}} n^r \sigma_n(f)\,. \end{aligned} \tag{2.1.12}$$

Remark 2.1.3. *$\mathcal{A}^r\left((\Sigma_n)_{n\in\mathbb{N}}, C(\Omega)\right)$ is a linear space and $\|\cdot\|_{\mathcal{A}^r}$ is a quasi-norm, i.e., there exists a constant C such that*

$$\|f+g\|_{\mathcal{A}^r} \le C\left(\|f\|_{\mathcal{A}^r} + \|g\|_{\mathcal{A}^r}\right) . \tag{2.1.13}$$

Proof. exercise □

Problem: Again we wish to characterize $\mathcal{A}^r\left((\Sigma_n)_{n\in\mathbb{N}}, C(\Omega)\right)$. We discuss only the case $r = 1$.
Questions:

- What does nonlinearity buy us? Do we have

$$\mathcal{A}^1\left((\mathcal{S}_n)_{n\in\mathbb{N}}, C(\Omega)\right) \subsetneqq \mathcal{A}^1\left((\Sigma_n)_{n\in\mathbb{N}}, C(\Omega)\right) \ ?$$

- What is the main principle that gives us something better ?

The main tool for this is another smoothness notion

$$V(f,\Omega) := \sup_{n, 0\le x_0<\ldots<x_n\le 1} \sum_{j=1}^{n} |f(x_j) - f(x_{j-1})| \ ,$$

the **total variation** of f. The closure of $C(\Omega)$ under this metric is called $BV(\Omega)$, the space of functions with **bounded variation**.

Remark 2.1.4. *It can be shown that every BV-function can be written as a sum of two monotone functions, see [32].*

Theorem 2.1.2. *One has*

$$\mathcal{A}^1\left((\Sigma_n)_{n\in\mathbb{N}}, C(\Omega)\right) = BV(\Omega) \cap C(\Omega)\,, \tag{2.1.14}$$

$$|f|_{\mathcal{A}^1} = \frac{1}{2} V(f,\Omega)\,. \tag{2.1.15}$$

Comments: A short discussion of this result prior to the proof.

Corollary 2.1.1. *The highest nontrivial approximation rate achievable by piecewise constants on n intervals is $\mathcal{O}(n^{-1})$:*

$$\left(\sigma_n(f) \le Cn^{-r} \textit{ for some } r > 1\right) \quad \Rightarrow \quad (f = \textit{ const})$$

Exercise 2.1.2. *One has*

$$\mathrm{Lip}\left(1, C(\Omega)\right) \subsetneqq BV(\Omega) \cap C(\Omega), \tag{2.1.16}$$

i.e., the nonlinear method provides the optimal rate for a strictly larger class of functions than that for which the linear method achieves the optimal rate.

Comments:

- nonlinear aproximation does not increase the highest possible rate but
- nonlinear approximation retains the highest rate for a much larger class of functions, so it compensates lack of regularity to a certain extend.

Exercise 2.1.3. *i)* $f(x) = \sqrt{x}$, $\Omega = [0,1]$, *compare* $e_n(f), \sigma_n(f)$. *More generally:* $f(x) = x^s$, $0 < s < 1$, $f \in \mathrm{Lip}\,(r, C(\Omega))$ *if and only if* $r \le s$.

ii) Let

$$W^1(L_1(\Omega)) := \{f : \int_\Omega |f'(t)|\ dt < \infty\}.$$

Show that

$$W^1(L_1(\Omega)) \subsetneqq BV(\Omega) \tag{2.1.17}$$

and that for $f \in W^1(L_1(\Omega))$, *we have*

$$V(f,\Omega) = \int_\Omega |f'(t)|\ dt\,. \tag{2.1.18}$$

The above two examples illustrate the principal distinctions between linear and nonlinear schemes. The superior performance of nonlinear approximation hinges on an *equidistribution* principle. In the above "Direct Theorem" the variation was equidistributed. The underlying approximation method is only an idealized conceptual trick because one generally does not have the information about the target function f needed to construct such an approximation.

Proof of Theorem 2.1.2: Suppose that $f \in BV(\Omega) \cap C(\Omega)$. To show (2.1.15) observe that $V(f,\cdot)$ is set additive, that is for $[a,b] = [a,c] \cup [c,b]$ one has

$$V(f,[a,b]) = V(f,[a,c]) + V(f,[c,b])\ .$$

The key idea is to choose those intervals for piecewise polynomial approximation that **equidistribute** variation, that is given $f \in BV(\Omega) \cap C(\Omega)$ (for the direct theorem $BV \subseteq \mathcal{A}^1$), choose $0 = x_0 < \ldots < x_n = 1$ such that

$$V(f, I_j) = \frac{1}{n} V(f,\Omega), \qquad j = 1..n,\ I_j = [x_{j-1}, x_j)\,. \tag{2.1.19}$$

Clearly, the x_j's may be very irregularily distributed and their location depends on f. (Equipartition of the range, not of the domain).

Next observe that for any partition $\mathcal{P}$ the best constant approximation on $I \in \mathcal{P}$ is given by the median $m_I(f)$ of f on I.

$$m_I(f) := \frac{1}{2}\left(\max_{x\in I} f(x) + \min_{x\in I} f(x)\right)\ , \tag{2.1.20}$$

that is

$$m_I(f) = \operatorname*{argmin}_{c\in\mathbb{R}} \|f-c\|_{L_\infty(I)}, \quad \|f-m_I(f)\|_{L_\infty(I)} \leq \frac{V(f,I)}{2}.$$

Hence

$$e_{I_k}(f) = \inf_{c\in\mathbb{R}} \|f-c\|_{L_\infty(I_k)} = \|f-m_{I_k}(f)\|_{L_\infty(I_k)} \leq \frac{V(f,I_k)}{2} = \frac{1}{2}\frac{V(f,\Omega)}{n}.$$

Thus

$$\sigma_n(f) \leq \max_{k=1,..,n} e_{I_k}(f) \leq \frac{1}{2n}V(f,\Omega)$$

and thereby

$$2\,|f|_{\mathcal{A}^1} \leq V(f,\Omega) \tag{2.1.21}$$

which gives the "Direct Theorem":

$$BV(\Omega)\cap C(\Omega) \subseteq \mathcal{A}^1\left((\Sigma_n)_{n\in\mathbb{N}}, C(\Omega)\right) \text{ and } |f|_{\mathcal{A}^1} \leq \frac{1}{2}V(f,\Omega).$$

For the "Inverse Theorem" assume $f \in \mathcal{A}^1\left((\Sigma_n)_{n\in\mathbb{N}}, C(\Omega)\right)$ and show $f \in BV(\Omega)$ i.e., we wish to show that $V(f,\Omega) \overset{!}{\leq} C|f|_{\mathcal{A}^1}$. To see this, onsider an arbitrary partition $\mathcal{P} = \{[x_{j-1},x_j) = I_j, j = 1,..,n\}$, and set $g := \sum_{I\in\mathcal{P}} f(\xi_I)\chi_I$ for some $\xi_I \in I$. Observe that

$$V(g,\Omega) = \sum_{j=1}^{n} \left|g(\xi_{I_j}) - g(\xi_{I_{j-1}})\right| = \sum_{j=1}^{n} \left|f(\xi_{I_j}) - f(\xi_{I_{j-1}})\right| \leq V(f,\Omega).$$

Since

$$\left|f(\xi_{I_j}) - f(\xi_{I_{j-1}})\right| \leq |\underbrace{f(\xi_{I_j})}_{=g(\xi_{I_j})} - f(x_{j-1})| + |f(x_{j-1}) - \underbrace{f(\xi_{I_{j-1}})}_{g(\xi_{I_{j-1}})}| \leq 2\,\|f-g\|_{L_\infty(\Omega)}$$

it follows that

$$V(g,\Omega) \leq 2n\,\|f-g\|_{L_\infty(\Omega)}. \tag{2.1.22}$$

We use this to show next that also $V(f,\Omega) \leq Cn\,\|f-g\|_{L_\infty(\Omega)}$ for some constant C which would prove our claim. To that end, consider any $t_0 < t_1 < ... < t_k$ in Ω and note that

$$\begin{aligned}\sum_{j=1}^{k} |f(t_j) - f(t_{j-1})| &\leq \sum_{j=1}^{k} \left(|f(t_j) - g(t_j)| + |g(t_{j-1}) - f(t_{j-1})| + |g(t_j) - g(t_{j-1})|\right)\\ &\leq V(g,\Omega) + 2k\,\|f-g\|_{L_\infty(\Omega)}\\ &\overset{(2.1.22)}{\leq} 2(n+k)\,\|f-g\|_{L_\infty(\Omega)}.\end{aligned}$$

Taking the infimum over all $g \in \Sigma_n$ gives

$$\sum_{j=1}^{k} |f(t_j) - f(t_{j-1})| \leq 2(n+k)\sigma_n(f) = 2\left(1 + \frac{k}{n}\right) n\sigma_n(f) \overset{f\in\mathcal{A}^1}{\leq} 2\left(1 + \frac{k}{n}\right) |f|_{\mathcal{A}^1} .$$

And for the limit $n \to \infty$

$$\sum_{j=1}^{k} |f(t_j) - f(t_{j-1})| \leq 2\,|f|_{\mathcal{A}^1} .$$

Since k, t_j are arbitrary this yields $V(f, \Omega) \leq 2\,|f|_{\mathcal{A}^1}$. □

2.1.3 Towards Algorithmic Realizations

Unfortunately, finding optimal partitions by equilibrating variation is ussually practically infeasible.

Question:

> Can one find an algorithm that works for every continuous function and achieves the same rate $\mathcal{O}(n^{-1})$ if $f \in BV(\Omega) \cap C(\Omega)$ (without using this knowledge directly).
> An algorithm that realizes the best possible rate without prior knowledge of the properties of f is called **universal**.

A central idea is to *equidistribute* local errors. Recall that $m_I(f)$ is the median of f on I and that

$$m_I(f) = \underset{c\in\mathbb{R}}{\operatorname{argmin}} \|f - c\|_{L_\infty(I)} .$$

Ideal scheme:

- Pick $\epsilon > 0$ target accuracy.
- Choose partition $\mathcal{P}_\epsilon$ such that for all $I \in \mathcal{P}_\epsilon$:

$$e_I(f) = \inf_{c\in\mathbb{R}} \|f - c\|_{L_\infty(I)} = \|f - m_I(f)\|_{L_\infty(I)} = \epsilon .$$

- Set

$$g_\epsilon = \sum_{I\in\mathcal{P}} m_I(f)\chi_I \quad \Rightarrow \quad \|f - g_\epsilon\|_{L_\infty(\Omega)} = \epsilon .$$

Question: How many intervals belong to $\mathcal{P}_\epsilon$?

$$(\#\mathcal{P}_\epsilon)\cdot\epsilon = \sum_{I\in\mathcal{P}_\epsilon}\epsilon = \sum_{I\in\mathcal{P}_\epsilon} e_I(f) \overset{\text{if } f\in BV(\Omega)\cap C(\Omega)}{\leq} \sum_{I\in\mathcal{P}_\epsilon}\frac{V(f,I)}{2} = \frac{1}{2}V(f,\Omega)\,,$$

which gives

$$\epsilon = \|f-g_\epsilon\|_{L_\infty(\Omega)} \leq \frac{1}{2}(\#\mathcal{P}_\epsilon)^{-1}V(f,\Omega)\,. \tag{2.1.23}$$

Thus, we obtain

$$\#\mathcal{P}_\epsilon \leq \frac{V(f,\Omega)}{2\epsilon}. \tag{2.1.24}$$

In particular, taking $\epsilon = \frac{V(f,\Omega)}{2}n^{-1}$ leads to $(\#\mathcal{P}_\epsilon) \leq n$, i.e., the error equidistribution principle realizes the optimal rate. This latter principle works in most adaptive methods.

This ideal scheme is, of course, difficult to implement because it is hard to find $\mathcal{P}_\epsilon$. A more practical alternative uses the *greedy paradigm* to push local errors towards equilibration:

Adaptive refinements (here in the simplest prototype version):

Algorithm 2.1.1. $\Omega = (0,1)$, *fix* $\epsilon > 0$, $\mathcal{P}_{\text{good}}, \mathcal{P}_{\text{bad}}$

```
i)   Set $\mathcal{P}_{bad} = \{\Omega\}$, $\mathcal{P}_{good} = \emptyset$
ii)  if $\mathcal{P}_{bad} \neq \emptyset$, for $I \in \mathcal{P}_{bad}$
         if $e_I(f) \leq \epsilon, \mathcal{P}_{bad}\backslash\{I\} \to \mathcal{P}_{bad}, \mathcal{P}_{good}\cup\{I\} \to \mathcal{P}_{good}$
         else $(\mathcal{P}_{bad}\backslash\{I\})\cup\{I_0, I_1\} \to \mathcal{P}_{bad}, \mathcal{P}_{good} \to \mathcal{P}_{good}$
         where   $I = I_0 \cup I_1$, $|I_0| = |I_1| = \frac{|I|}{2}$
iii) if $\mathcal{P}_{bad} = \emptyset$, set $\mathcal{P}_\epsilon = \mathcal{P}_{good}$
         set $g_\epsilon = \sum_{I\in\mathcal{P}_\epsilon} m_I(f)\chi_I$
```

Algorithm 2.1.1 is an example of *tree approximation*. Hence it is more restrictive since only dyadic intervals $I = [l\,2^{-k}, (l+1)2^{-k})$, $l = 0, ..., 2^k-1$, $k \in \mathbb{N}_0$ occur. The question is: What do we lose?

For $f \in C(\Omega)$, Algorithm 2.1.1 terminates after finitely many steps.

Set

$$g_\epsilon := \sum_{I\in\mathcal{P}_\epsilon} m_I(f)\chi_I\,.$$

By construction

$$\|f-g_\epsilon\|_{L_\infty(\Omega)} \leq \epsilon\,. \tag{2.1.25}$$

How "good" this algorithm is depends on how many intervals belong to $\mathcal{P}_\epsilon$. So we are again interested in the relation

$$\epsilon \overset{?}{\leftrightarrow} \#\mathcal{P}_\epsilon$$

depending on properties of f. Remember from (2.1.24): we have shown that for $f \in BV(\Omega) \cap C(\Omega)$ the ideal scheme produces a partition $\mathcal{P}^*_\epsilon$ with $\#\mathcal{P}^*_\epsilon \leq (2\epsilon)^{-1}V(f,\Omega)$.

To estimate $\#\mathcal{P}_\epsilon$ for the output $\mathcal{P}_\epsilon$ of Algorithm 2.1.1 we we need a new smoothness measure:

Hardy-Littlewood maximal function
For $f \in L_1(\Omega)$ consider

$$Mf(x) = \sup_{I \ni x} \frac{1}{|I|} \int_{I \subseteq \Omega} |f(t)| \, dt \,. \tag{2.1.26}$$

Mapping properties (see e.g. [32, 31]):

Theorem 2.1.3. *For* $1 < p \leq \infty$*:*

$$Mf \in L_p(\Omega) \quad \Leftrightarrow \quad f \in L_p(\Omega) \,.$$

For $p = 1$*:*

$$Mf \in L_1(\Omega) \quad \Leftrightarrow \quad |f| \log(1 + |f|) \in L_1(\Omega) \,.$$

Our new smoothness condition is

$$M(f') \in L_1(\Omega) \,.$$

(This is a stronger condition than $f' \in L_1(\Omega) \quad \Leftrightarrow \quad f \in W^1(L_1(\Omega)) \subset BV(\Omega)$)

Remark 2.1.5. *For any* f *with* $Mf' \in L_1$ *one has*

$$(\#\mathcal{P}_\epsilon) \leq \max\left\{1, \epsilon^{-1} \|Mf'\|_{L_1(\Omega)}\right\} \,. \tag{2.1.27}$$

In other words, the algorithm constructs an approximation with accuracy $\mathcal{O}(n^{-1})$ *at the expense of* n *pieces whenever* $Mf' \in L_1$.

Proof. We need to count $\#\mathcal{P}_\epsilon$ and w.l.o.g. $\#\mathcal{P}_\epsilon > 1$. Key observation: If $I \in \mathcal{P}_\epsilon$ its parent $\hat{I}$ belongs to $\mathcal{P}_{bad}$, i.e., $e_{\hat{I}}(f) > \epsilon$. Hence one has

$$\epsilon < e_{\hat{I}}(f) \leq \frac{1}{2}V(f,\hat{I}) \overset{(2.1.18)}{=} \frac{|\hat{I}|}{2} \frac{1}{|\hat{I}|} \int_{\hat{I}} |f'(t)| \, dt = |I| \frac{1}{|\hat{I}|} \int_{\hat{I}} |f'(t)| \, dt \leq |I| \, Mf'(x),$$

for any $x \in I$. Let $\bar{x}$ minimize Mf' over I. Then

$$Mf'(\bar{x}) \leq \frac{1}{|I|} \int_I |Mf'(t)|\, dt,$$

so that

$$\epsilon \leq |I| Mf'(\hat{x}) \leq \int_I |Mf'(t)| dt,$$

and consequently

$$\|Mf'\|_{L_1(\Omega)} = \int_\Omega |Mf'(t)|\ dt = \sum_{I \in \mathcal{P}_\epsilon} \int_I |Mf'(t)|\ dt \geq (\#\mathcal{P}_\epsilon) \cdot \epsilon\,.$$

Thus,

$$\#\mathcal{P}_\epsilon \leq \frac{\|Mf'\|_{L_1(\Omega)}}{\epsilon}. \tag{2.1.28}$$

Moreover, for $n \in \mathbb{N}$ and $\epsilon = \|Mf'\|_{L_1(\Omega)}\, n^{-1}$ the algorithm creates a partition $\mathcal{P}_\epsilon$ of at most n intervals, giving accuracy $\|Mf'\|_{L_1}\, n^{-1}$, it provides the optimal rate for the class

$$M_1(\Omega) := \{f \in L_1(\Omega) : Mf' \in L_1(\Omega)\}.$$

. □

What is the meaning of $Mf' \in L_1$?

The theorem says that

$$|f'| \log(1 + |f'|) \in L_1(\Omega) \tag{2.1.29}$$

which is strictly (but only a little) stronger than $f' \in L_1(\Omega)$ which implies $f \in BV(\Omega) \cap C(\Omega)$. In fact, $f' \in L_p(\Omega)$, $p > 1$, implies (2.1.29). Hence the adaptive algorithm realizes the same work-accuracy balance $\epsilon \leftrightarrow n(\epsilon)$ as the "best partition" scheme under slightly stronger smoothness requirements. We shall encounter later more general results of this type.

Comments:

1. In the above abstract form Algorithm 2.1.1 is still idealized. However, as will be seen later, it is very close to practical versions arising for instance in *machine learning*, *data coarsening*, or implicitly in *PDE solvers*.

2. So far: Approximation methods were based on *localization and equidistribution*. Moreover linear and nonlinear approximability could be characterized through *regularity viz. smoothness* properties of the approximand

f. Another important class of adaptation concepts is based on *representations* of f in terms of a *basis* (or dictionary). This is particularly important in the spatially high-dimensional regime. In either case it is essential to have some understanding of the involved function spaces describing the properties of f.

2.2 A General Framework: Approximation Classes

The above examples suggest the following abstract setting:

- Suppose that $\mathbb{X}$ with norm $\|\cdot\|_{\mathbb{X}}$ is a (quasi-)Banach space.
- Let $\Sigma_n \subset \mathbb{X}, n \in \mathbb{N}_0$, be determined by n degrees of freedom/parameters s.t.

$$g \in \Sigma_n \;\Rightarrow\; cg \in \Sigma_n,\, c \in \mathbb{R}, \quad \Sigma_n + \Sigma_n \subseteq \Sigma_{an} \text{ some } a \geq 1, \tag{2.2.1}$$

for instance, piecewise constants of fixed degree on arbitrary partitions into n intervals. Let

$$\sigma_n(f)_{\mathbb{X}} := \inf_{g \in \Sigma_n} \|f - g\|_{\mathbb{X}} \tag{2.2.2}$$

denote the error of best n-term approximation.

- Consider the *approximation class*

$$\begin{aligned} \mathcal{A}^r((\Sigma_n), \mathbb{X}) &:= \{f \in X : |f|_{\mathcal{A}^r_q} < \infty\} \\ \|f\|_{\mathcal{A}^r_q} &:= \|f\|_{\mathbb{X}} + |f|_{\mathcal{A}^r_q}, \end{aligned} \tag{2.2.3}$$

where

$$|f|_{\mathcal{A}^r_q} := \begin{cases} \sup_{n \in \mathbb{N}} n^r \sigma_n(f)_{\mathbb{X}}, & q = \infty, \\ \left(\sum_{n \in \mathbb{N}} \left(n^r \sigma_n(f)_{\mathbb{X}}\right)^q \frac{1}{n}\right)^{1/q}, & 0 < q < \infty. \end{cases} \tag{2.2.4}$$

Thus, we put all $f \in \mathbb{X}$ into the same bucket whose error of best n-term approximation decays *at least* as $O(n^{-r})$. For $q < \infty$ this decay must be a little stronger to make the series converge. So q can be viewed as a "fine tuning" parameter. Its relevance will become clear later.

Since $q = \infty$ is perhaps the case of primary interest we usually write for convenience

$$\mathcal{A}^r((\Sigma_n), \mathbb{X}) = \mathcal{A}^r_\infty((\Sigma_n), \mathbb{X}),$$

as in the earlier examples. In particular, $f \in \mathcal{A}^r((\Sigma_n), \mathbb{X})$ means that for each $n \in \mathbb{N}$ we can find a $g_n \in \Sigma_n$ such that

$$\|f - g_n\|_{\mathbb{X}} \leq n^{-r} |f|_{\mathcal{A}^r}.$$

Remark 2.2.1. *Another way to read* $t \in \mathcal{A}^r((\Sigma_n), \mathbb{X})$ *is that for any given target accuracy* $\varepsilon > 0$ *it suffices to take*

$$n(\varepsilon) := \left\lceil \Big(\frac{|f|_{\mathcal{A}^r}}{\varepsilon}\Big)^{\frac{1}{r}} \right\rceil \quad \Rightarrow \quad \exists\, g \in \Sigma_{n(\varepsilon)} \ s.t.\ \|f - g\|_{\mathbb{X}} \le \varepsilon,$$

compare this with (2.1.24), (2.1.28). *Obviously the larger* r *the fewer terms are needed.*

Remark 2.2.2. *It is often convenient to work with the expression*

$$\Big(\sum_{j\in\mathbb{N}_0} \big(2^{rj}\sigma_{2^j}(f)_{\mathbb{X}}\big)^q\Big)^{1/q} \sim |f|_{\mathcal{A}^r_q} \tag{2.2.5}$$

which indeed can be shown to be equivalent.

Linear Approximation:
We talk about *linear* approximation when the Σ_n are linear spaces such as splines or finite element spaces with respect to fixed meshes

Non-linear Approximation:
Σ_n are non-linear sets (only satisfying (2.2.1)) such as rational functions, free-knot splines, adaptive mesh-refinements.

Let $\mathbb{P}_m(\Omega)$ denote the linear space of polynomials of (total) *order* m (degree $m-1$) over Ω. For a given partition $\mathcal{P}$ of Ω let

$$\mathbb{P}_m(\mathcal{P}) := \Big\{ \sum_{T\in\mathcal{P}} \chi_T P_T : P_T \in \mathbb{P}_m(T),\ T \in \mathcal{P} \Big\}$$

denote the space of piecewise polynomials of order m subordinate to the partition $\mathcal{P}$. The results in the previous section can be reinterpreted as follows.

Remark 2.2.3. *Linear approximation classes:*

$$\mathcal{A}^r((\mathbb{P}_1(\mathcal{P}_n)), C(0,1)) = \mathrm{Lip}(r, C(0,1))$$

piecewise constants on uniform partitions $\mathcal{P}_n$ *of* $(0,1)$.

Remark 2.2.4. *Non-linear approximation classes:*

$$\mathcal{A}^1((\Sigma_n), C(0,1)) = BV(0,1)$$

Σ_n = *piecewise constants with* n *arbitrary pieces, i.e.,*

$$\Sigma_n = \bigcup \{\mathbb{P}_1(\mathcal{P}) : \#(\mathcal{P}) \le n\}.$$

Comments:
In the above examples we identify approximation classes with intrinsic classical smoothness spaces. Why is this relevant? we can see from the type of smoothness whether a nonlinear or adaptive approximation performs better than simpler linear methods. Information about the smoothness of a function is, for instance, provided by *regularity theory for PDEs*. In general the understanding of the complexity and performance of adaptive schemes draws crucially on the deep interrelation between approximability and regularity.

2.3 A Primer on Function Spaces

In view of the findings in the last section, we collect next some basics on function spaces.

2.3.1 Sobolev Spaces

Classical differentiability is understood in a pointwise sense. So called weak differentiability relaxes these requirements leading to the notion of *weak derivatives*.

Let

$$\|f\|_{L_p(\Omega)} := \begin{cases} \left(\int_\Omega |f(x)|^p dx\right)^{1/p}, & 0 < p < \infty, \\ \operatorname{esssup}_{x\in\Omega}|f(x)|, & p = \infty. \end{cases}$$

Note, this is only a quasi-norm when $p < 1$. The spaces of p-integrable (equivslence classes of) functions are given by

$$L_p(\Omega) := \left\{ f \text{ measurable } : \|f\|_{L_p(\Omega)} < \infty \right\}.$$

These are (quasi-)Banach spaces (i.e., complete normed linear spaces).

Remark 2.3.1. *Assume that $\Omega \subset \mathbb{R}^d$ is a bounded domain. When $r < p$ one can use Hölder's inequality to show that*

$$\|f\|_{L_r(\Omega)} \le C\|f\|_{L_p(\Omega)}, \quad f \in L_p(\Omega), \tag{2.3.1}$$

with C depending on r, p, Ω, i.e., $\|\cdot\|_{L_r(\Omega)}$ is a weaker norm than $\|\cdot\|_{L_p(\Omega)}$. Measuring smoothness in L_r is therefore less demanding than measuring smoothness in L_p, when $p > r$. This will be used later intensely, see also Remark 2.3.2.

Consider the standard pointwise partial derivatives $\partial^\alpha := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}$, $\alpha \in \mathbb{Z}_+^d$. Here $|\alpha| = \alpha_1 + \cdots + \alpha_d$. A function $f \in L_p(\Omega)$ possesses the αth weak derivative $D^\alpha f \in L_p(\Omega)$ if

$$\int_\Omega f(x)\partial^\alpha\phi(x)dx = (-1)^{|\alpha|}\int_\Omega D^\alpha f(x)\phi(x)dx, \quad \forall\ \phi \in C_0^\infty(\Omega).$$

Exercise 2.3.1. *One has* $D^\alpha f(x) = \partial^\alpha f(x)$ *whenever* $f \in C^{|\alpha|}(\Omega)$.

Sobolev semi-norm:

$$|f|_{W^m(L^p(\Omega))} := \Big(\sum_{|\alpha|=m} \|\partial^\alpha f\|^p_{L^p(\Omega)} \Big)^{1/p}, \quad 1 \le p \le \infty;$$

Sobolev-norm:

$$\|f\|_{W^m(L_p(\Omega))} := \Big(\|f\|^p_{W^{m-1}(L_p(\Omega))} + |f|^p_{W^m(L_p(\Omega))} \Big)^{1/p}, \quad \|f\|_{W^0(L_p(\Omega))} := \|f\|_{L_p(\Omega)}$$

Assume that $\Omega \subset \mathbb{R}^d$ is open, and satisfies a uniform cone-condition (at every point x on the boundary $\partial\Omega$ of Omega one can fit a cone of a fixed opening angle and peak x fully in $\bar{\Omega}$).
Sobolev-spaces:

$$W^m(L_p(\Omega)) := \big\{ f \in L_p(\Omega) : \|f\|_{W^m(L_p(\Omega))} < \infty \big\}.$$

Remark 2.3.2. *One has the following trivial continuous embeddings:*

$$m \le n \Rightarrow W^n(L_p(\Omega)) \subseteq W^m(L_p(\Omega)), \quad p \le q \Rightarrow W^m(L_q(\Omega)) \subseteq W^m(L_p(\Omega)),$$

where the second one follows from Remark 2.3.1.

Recall that a normed linear space $(\mathbb{Y}, \|\cdot\|_{\mathbb{Y}})$ is called *continuously embedded* in $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ if

$$\|f\|_{\mathbb{X}} \le C\|f\|_{\mathbb{Y}}, \quad \forall f \in \mathbb{Y}. \tag{2.3.2}$$

An important special case is $p = 2$. In this case one obtains *Hilbert spaces*, i.e., the norms are induced by inner products. One often uses the notation:

$$H^m(\Omega) := W^m(L_2(\Omega)), \quad H^m_0(\Omega) := \mathrm{clos}_{\|\cdot\|_{H^m}} \big(C^\infty_0(\Omega) \big)$$

Exercise 2.3.2. *The function* $f(x) := |x|$ *belongs to* $H^1((-1,1))$ *but not to* $C^1((-1,1))$.

Another special case is the *Lipshitz space* discussed earlier: $W^1(L_\infty(\Omega)) = \mathrm{Lip}(1, C(\Omega))$.

2.3.2 Bounded Variation

A convenient way to define equivalent norms for the space $BV(\Omega)$ for $\Omega \subset \mathbb{R}^d$ is to use duality. To this end, for $\phi \in (L_p(\Omega))^d$ let

$$\|\phi\|_{L_p(\Omega)} = \Big\| \Big(\sum_{i=1}^d |\phi_i|^2 \Big)^{1/2} \Big\|_{L_p(\Omega)}.$$

For $1 < p \leq \infty$ the existence of the weak derivative can be characterized as follows: there exists a constant $C < \infty$ such that

$$\int_\Omega f(x)\partial^\alpha \varphi(x)dx \leq C\|\varphi\|_{L_q(\Omega)}, \quad \varphi \in C_0^\infty(\Omega), \ \frac{1}{p} + \frac{1}{q} = 1, \tag{2.3.3}$$

and in particular that $f \in W^1(L_p(\Omega))$ if and only if

$$\int_\Omega f(x)\mathrm{div}\varphi(x)dx \leq C\|\varphi\|_{L_q(\Omega)}, \quad \varphi \in (L_q(\Omega))^d. \tag{2.3.4}$$

Here we use that $L_q(\Omega)$ is the dual space of $L_p(\Omega)$. For $p = 1$, $q = \infty$ this does not characterize $W^1(L_1(\Omega))$ because $L_1(\Omega)$ is not the dual of $L_\infty(\Omega)$. Instead, in this way one characterizes the space $BV(\Omega)$ of functions of *bounded total variation*. Specifically, define

$$|f|_{BV(\Omega)} := \max\left\{ \int_\Omega f(x)\mathrm{div}\varphi(x)dx : \varphi \in \left(C_0^\infty(\Omega)\right)^d \text{ and } \|\varphi\|_{L_\infty(\Omega)} = 1 \right\}. \tag{2.3.5}$$

Remark 2.3.3. *One can show that* $f \in BV(\Omega)$ *if and only if* $\sup_{h\in\mathbb{R}^d} \frac{\|\Delta_h f\|_{L_1(\Omega_h)}}{|h|} < \infty$, *where* $\Delta_h f(x) := (x + h) - f(x)$, $\Omega_h := \{x \in \Omega : x + h \in \Omega\}$. *Thus* $BV(\Omega) = \mathrm{Lip}(1, L_1(\Omega))$.

Exercise 2.3.3. *Suppose that* E *is a domain in* Ω, $\bar{E} \subset \Omega$ *whose boundary* ∂E *has finite Hausdorff dimension* $\mathcal{H}^{d-1}(\partial E)$. *When the boundary is smooth enough one has*

$$|\chi_E|_{BV(\Omega)} = \mathcal{H}^{d-1}(\partial E).$$

This shows also that $BV(\Omega) \neq W^1(L_1(\Omega))$.

2.3.3 Besov Spaces

A series deficiency of the above smoothness classes is that they are too coarse grained. A classical example is the Trace Theorem which states that the trace opereator - a substitute for the restriction operator defined when a function is defined pointwise - maps for instance $H^1(\Omega)$ onto $H^{1/2}(\partial\Omega)$, i.e., traces of functions with one weak derivative in $L_2(\Omega)$ are only half differentiable in L_2 on $\partial\Omega$. An example of spaces with non-integer smoothness are the Lipshitz spaces (also called Hölder spaces) encountered in earlier sections where smoothness is measured in L_∞. The concept of *Besov spaces* is important for (at least) two reasons:

- it allows one to formulated non-integer degree of smoothness for all $0 < p \leq \infty$ in a unified way;
- nonlinear and adaptive approximation is goverend by Besov-regularity, not just by standard Sobolev regularity.

Basic ingredients: The perhaps most important point is to describe smoothness in a "derivative free" way. Instead, one uses so called *moduli of smoothness*:

Difference operators:

$$(\Delta_h f)(x) := f(x+h)-f(x), \quad \Delta_h^m f = \Delta_h^{m-1}(\Delta_h f), \quad \Omega_{kh} := \{x \in \Omega : x+lh \in \Omega,\ l \le k\};$$

Modulus of smoothness:

$$\omega_m(f,t,\Omega)_p := \sup_{|h|\le t} \|\Delta_h^m f\|_{L_p(\Omega_{mh})}, \quad 0 < p < \infty. \tag{2.3.6}$$

For $p = \infty$ the integration over Ω_{mh} is replaced by the supremum.

Comments and related results:

- Notice that
$$\omega_m(g,t,\Omega)_p = 0 \quad \forall\, g \in \mathbb{P}_m,$$
because mth order differences annihilate polynomials of degree $m-1$, i.e., order m, just like derivatives.

- For each fixed $f \in L_p(\Omega)$ one has $\omega_1(f,t,\Omega)_p \to 0$, $t \to 0$ which expresses continuity in L_p just like continuity in $C(\Omega)$:
$$\omega_1(f,t,\Omega)_\infty = \sup_{x\in\Omega} \sup_{|h|\le t} |f(x+h) - f(x)| \to 0, \quad t \to 0.$$

- It is easy to show (triangle inequality) that
$$\omega_n(f,t,\Omega)_p \le C\omega_k(f,t,\Omega)_p, \quad k \le n,\ C = C(k,n), \tag{2.3.7}$$
so that each fixed $f \in L_p(\Omega)$ one has $\omega_n(f,t,\Omega)_p \to 0$, $t \to 0$, holds for an fixed $n \in \mathbb{N}$, see e.g. [32].

- On a deeper level the modulus characterizes compactness in L_p in the sense that $\mathcal{F}$ is a compact subset of $L_p(\Omega)$ if and only if
$$\lim_{t\to 0} \sup_{f\in\mathcal{F}} \omega_m(f,t,\Omega)_p = 0,$$
which corresponds to Ascoli's Theorem for $p = \infty$.

Clearly, for an arbitrary $f \in L_p(\Omega)$ the convergence of $\omega_m(f,t,\Omega)_p$ may be arbitrarily slow. However, the smoother f the faster one expects the convergence to be. This suggests to describe smoothness by quantifying the decay of the modulus.

Definition 2.3.1. *Let* $s > 0$ *and* $0 < p, q \leq \infty$ *and let* $m \in \mathbb{N}$ *be an integer such that* $s < m$

$$|f|_{B^s_q(L_p(\Omega))} := \begin{cases} \left(\int_0^\infty (t^{-s}\omega_m(f,t,\Omega)_p)^q \frac{dt}{t} \right)^{1/q}, & 0 < q < \infty, \\ \sup_{t>0} t^{-s}\omega_m(f,t,\Omega)_p, & q = \infty, \end{cases}$$

Then

$$B^s_q(L_p(\Omega)) := \left\{ f \in L_p(\Omega) : |f|_{B^s_q(L_p(\Omega))} < \infty \right\}$$

and

$$\|f\|_{B^s_q(L_p(\Omega))} := \|f\|_{L_p(\Omega)} + |f|_{B^s_q(L_p(\Omega))}.$$

Remark 2.3.4. *In analogy to* (2.1.4) *one can define Lipschitz spaces for* L_p *by*

$$\operatorname{Lip}(s, L_p(\Omega)) := \{f \in L_p(\Omega) : \sup_{t>0} t^{-s}\omega_1(f,t,\Omega)_p := |f|_{\operatorname{Lip}(s,L_p(\Omega)} < \infty\}. \tag{2.3.8}$$

Obviously, one has

$$\operatorname{Lip}(s, L_p(\Omega)) = B^s_\infty(L_p(\Omega)), \quad 0 < s < 1. \tag{2.3.9}$$

One can show that

$$\begin{aligned} \operatorname{Lip}(1, L_p(\Omega)) &= W^1(L_p(\Omega)), \quad 1 < p \leq \infty \\ \operatorname{Lip}(1, L_1(\Omega)) &= BV(\Omega) \neq W^1(L_1(\Omega)). \end{aligned} \tag{2.3.10}$$

The smoothness characterized by this definition seems to depend on m, the order of the modulus, because s is limited by m through $s < m$. It is a consequence of *Marchaud's inequality* for moduli of smoothness that this is not the case in the sense that

$$\left(\int_0^\infty (t^{-s}\omega_m(f,t,\Omega)_p)^q \frac{dt}{t} \right)^{1/q} + \|f\|_{L_p(\Omega)} \sim \left(\int_0^\infty (t^{-s}\omega_k(f,t,\Omega)_p)^q \frac{dt}{t} \right)^{1/q} + \|f\|_{L_p(\Omega)}, \tag{2.3.11}$$

as long as $m, k > s$, see [32].

Remark 2.3.5. *One may wonder about the role of the additional parameter* $q \in (0, \infty]$. *It will be explained later in a bit more detail. Here it suffices to note that the strongest information is given by* s *and* p*, while* q *is used as a "fine-tuning" parameter.*

There is another important fact at the heart of analysing the Besov spaces, sometimes referred to as *Whitney's Theorem* that says

$$\inf_{g \in \mathbb{P}_m} \|f - g\|_{L_p(\Omega)} \sim \omega_m(g, \Omega)_p, \quad f \in L_p(\Omega), \tag{2.3.12}$$

with constants of equivalence depending on m, p, and Ω, where

$$\omega_m(g, \Omega)_p := \sup_{t>0} \omega_m(g, t, \Omega)_p. \tag{2.3.13}$$

(2.3.12) says that best local polynomial approximation scales like the modulus. One direction is familiar from Taylor's expansion. Thus, local polynomial approximability is also a smoothness measure. One very familiar consequence of Whitney's Theorem can be stated as follows: let D be a fixed "reference domain" (e.g. unit simplex, unit cube) and let $(D_h)_{h \geq 0}$ be a family of affine images of D of shrinking size $h = \operatorname{diam} D_h$ where the affine mappings have uniformly conditioned linear parts, then rescaling arguments show that (2.3.12) implies

$$\inf_{g \in \mathbb{P}_m} \|f - g\|_{L_p(D_h)} \leq C h^m |f|_{W^m(L_p(D_h))}, \quad f \in W^m(L_p(D_h)), \tag{2.3.14}$$

where C depends on m, p, the bound on the condition of the affine mappings, and on the reference domain D. In fact, there is a more general version

$$\inf_{g \in \mathbb{P}_m} \|f - g\|_{L_p(D_h)} \leq C h^\delta |f|_{B^s_\infty(L_r(D_h))}, \quad \delta = s - \Big(\frac{d}{r} - \frac{d}{p}\Big). \tag{2.3.15}$$

When $r = p$ this gives the order h^s of approximation corresponding to the fact that smoothness is s. When $r < p$ measuring smoothness in L_r is weaker than mesuring smoothness in L_p (see Remark 2.3.1). Thus, one can still quantify the error even when f is not smooth in L_p. It lowers the rate though by $\frac{d}{r} - \frac{d}{p} > 0$. Conversely, when $r > p$ one gains in approximation order taking advantage of the fact that f has smoothness s in a stronger metric.

Additional Background Facts: We record next some further useful facts without proof and refer to [32] for details. In particular, this concerns the relation between Sobolev and Besov spaces.

Remark 2.3.6. *The following properties hold:*

1. $0 < s < m$*:* $B^s_q(L_p(\Omega))$ interpolates *between* $L_p(\Omega)$ *and* $W^m(L_p(\Omega))$.
2. $p, q \geq 1$*:* $B^s_q(L_p(\Omega))$ *is a Banach space, otherwise only a quasi-Banach space.*
3. $p, q < 1$ *is important for* nonlinear *approximation, as shown later.*
4. $B^m_p(L_p(\Omega)) \neq W^m(L_p(\Omega))$ *for* $p \neq 2$*, but* $B^s_2(L_2(\Omega)) = H^s(\Omega)$*,* $s \in R$.
5. *Equivalent semi-norm:* $\|(a_\lambda)_{\lambda \in \mathcal{I}}\|_{\ell_p(\mathcal{I})} := \Big(\sum_{\lambda \in \mathcal{I}} |a_\lambda|^p\Big)^{1/p}$

$$|f|_{B^s_q(L_p(\Omega))} := \|(2^{sj} \omega_m(f, 2^{-j}, \Omega)_p)_{j \in \mathbb{Z}_+}\|_{\ell_q(\mathbb{Z}_+)}.$$

This follows from discretizing the integral in Definition 2.3.1. Compare this with the expression (2.2.5) *appearing in the approximation classes.*

6. *Defining*

$$W^s(L_p(\Omega)) := B^s_p(L_p(\Omega)), \quad s \notin \mathbb{N},$$

one recovers the Sobolev-Slobodezcki spaces, defined through multiple integrals. Moreover $B^s_\infty(L_\infty(\Omega))$ *are for all* $s > 0$ *the spaces referred to as Hölder spaces.*

The key word above is "interpolation". 1. says that the Besov spaces $B^s_q(L_p(\Omega))$, $k < s < k+1$, somehow fill the "gap" between $W^k(L_p(\Omega))$ and $W^{k+1}(L_p(\Omega))$. *Interpolation of Banach spaces* is a concept addressing the following objective. Suppose that $\mathbb{X}, \|\cdot\|_{\mathbb{X}}$, $\mathbb{Y}, \|\cdot\|_{\mathbb{Y}}$ are two Banach spaces where we assume that $\mathbb{Y} \subset \mathbb{X}$ in the sense of continuous embeddings. More precisely, we assume that $\mathbb{Y}$ has a semi-norm $|\cdot|_{\mathbb{Y}}$ and

$$\|\cdot\|_{\mathbb{X}} + |\cdot|_{\mathbb{Y}} \sim \|\cdot\|_{\mathbb{Y}}, \tag{2.3.16}$$

i.e., the left expression is an equivalent norm on $\mathbb{Y}$. Simple examples are $\mathbb{X} = L_p(\Omega)$, $\mathbb{Y} = W^k(L_p(\Omega))$, $|\cdot|_{\mathbb{Y}} = |\cdot|_{W^k(L_p(\Omega))}$.

Remark 2.3.7. *We are considering a somewhat specialized setting. In general one need not require an embedding of one space into the other but rather works with the spaces* $\mathbb{X} + \mathbb{Y}, \mathbb{X} \cap \mathbb{Y}$. *For our purposes the specialized scenario suffices, see [5] for the general picture.*

We wish to "fill up" the "interval" $[\mathbb{X}, \mathbb{Y}]$ by certain "intermediate" spaces $[\mathbb{X}, \mathbb{Y}]_{\theta,q}$, parametrized by $\theta \in [0, 1]$, $q \in (0, \infty]$ (as explained later) in the following way:

- Let $\mathcal{L}(\mathbb{X}, \mathbb{Y})$ denote the space of bounded linear operators from $\mathbb{X}$ to $\mathbb{Y}$ endowed with the norm

  $$\|\mathcal{B}\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})} := \sup_{v \in \mathbb{X}} \frac{\|\mathcal{B}v\|_{\mathbb{Y}}}{\|v\|_{\mathbb{X}}},$$

  and suppose that $\mathcal{B} \in \mathcal{L}(\mathbb{X}_1, \mathbb{Y}_1)$ and $\mathcal{B} \in \mathcal{L}(\mathbb{X}_2, \mathbb{Y}_2)$.

- The interpolation method that yields the interpolation spaces

  $$[\mathbb{X}_1, \mathbb{Y}_1]_{\theta,q}, \quad [\mathbb{X}_2, \mathbb{Y}_2]_{\theta,q}$$

  for fixed θ, q should ensure that

  $$\|\mathcal{B}\|_{\mathcal{L}([\mathbb{X}_1,\mathbb{Y}_1]_{\theta,q},[\mathbb{X}_2,\mathbb{Y}_2]_{\theta,q})} \leq \|\mathcal{B}\|^{1-\theta}_{\mathcal{L}(\mathbb{X}_1,\mathbb{Y}_1)} \|\mathcal{B}\|^{\theta}_{\mathcal{L}(\mathbb{X}_2,\mathbb{Y}_2)}. \tag{2.3.17}$$

  i.e., $\mathcal{B}$ is also bounded as a mapping between the interpolated spaces.

A classical application arises in deriving error estimates in the finite element method. Suppose one has constructed a projector Q_h whose range is a finite element space V_h subordinate to a mesh of mesh size h. Certain such projectors are so calles "quasi-interpolants" which are bounded mappings in $\mathcal{L}(L_2(\Omega), L_2(\Omega))$ and also in $\mathcal{L}(H^1(\Omega), H^1(\Omega))$ (L_2- and H^1-stable). Typical error estimates (NumaIV) then read

$$\|f - Q_h f\|_{L_2(\Omega)} \leq Ch|f|_{H^1(\Omega)}. \tag{2.3.18}$$

L_2-boundedness also gives $\|f - Q_h f\|_{L_2(\Omega)} \leq C\|f\|_{L_2(\Omega)} = Ch^0\|f\|_{L_2(\Omega)}$. Once one knows that $H^s(\Omega) = [L_2(\Omega), H^1(\Omega)]_{s,2}$ an application of (2.3.17) to $\mathcal{B} := I - Q_h$ yields

$$\|f - Q_h f\|_{L_2(\Omega)} \leq Ch^s|f|_{H^s(\Omega)}, \tag{2.3.19}$$

i.e., the rate s is determined by the difference of smoothness between the norms on the left and on the right.

The K-functional: The decisive tool to construct the interpolation spaces $[\mathbb{X}, \mathbb{Y}]_{\theta,q}$ is the K-functional (see [5] and the literature quoted there) introduced by J. Peetre. Under the assumption (2.3.16) it takes the form

$$K(f,t) = K(f,t;\mathbb{X},\mathbb{Y}) := \inf_{g\in\mathbb{Y}} \left\{\|f-g\|_{\mathbb{X}} + t|g|_{\mathbb{Y}}\right\}. \tag{2.3.20}$$

Thus, $K(f,t)$ measures closeness of $f \in \mathbb{X}$ to some element in the subspace $\mathbb{Y}$. When insisting of approximating f too well by an element from $\mathbb{Y}$ the second term may become large. The optimal compromise given by $K(f,t)$ is therefore also a measure of smoothness when the semi-norm $|\cdot|_{\mathbb{Y}}$ measures smoothness (as in the above examples).

Definition 2.3.2. *For $0 < q \leq \infty$, $0 < \theta < 1$, let*

$$\|f\|_{[\mathbb{X},\mathbb{Y}]_{\theta,q}} := \begin{cases} \left(\int_0^\infty (t^{-\theta}K(f,t;\mathbb{X},\mathbb{Y}))^q \frac{dt}{t}\right)^{1/q}, & 0 < q < \infty, \\ \sup_{t>0} t^{-\theta}K(f,t;\mathbb{X},\mathbb{Y}), & q = \infty. \end{cases} \tag{2.3.21}$$

and

$$[\mathbb{X},\mathbb{Y}]_{\theta,q} := \{f \in \mathbb{X} : \|f\|_{[\mathbb{X},\mathbb{Y}]_{\theta,q}} < \infty\}.$$

A glimpse at Definition 2.3.1 reveals that the interpolation norm (2.3.21) has exactly the same structure as the Besov-semi-norm. In addition there is an important result by H. Johnen and K. Scherer [39] saying that for $1 \leq p \leq \infty$ and domains Ω satisfying a uniform cone condition

$$K(f,t;L_p(\Omega), W^m(L_p(\Omega))) \sim \omega_m(f,t,\Omega)_p \tag{2.3.22}$$

with constants depending only on p, m, Ω. This explains the following remark.

Remark 2.3.8. *Besov spaces are interpolation spaces between* L_p*- and Sobolev spaces. A typical example is*

$$B^s_q(L_p(\Omega)) = [L_p(\Omega), W^m(L_p(\Omega)]_{\theta,q}, \quad \theta = s/m.$$

A further important result is that the approximation classes $\mathcal{A}^r_q$ *are interpolation spaces under the following circumstances: suppose there exists* $R > 0$ *such that the elements of* $\mathbb{Y}$ *satisfy the* Jackson inequality

$$\sigma_n(f)_{\mathbb{X}} \leq Cn^{-R}|f|_{\mathbb{Y}}, \quad n \in \mathbb{N},\ f \in \mathbb{Y}, \tag{2.3.23}$$

as well as a companion Bernstein inequality

$$|g|_{\mathbb{Y}} \leq Cn^{R}\|g\|_{\mathbb{X}}, \quad g \in \Sigma_n,\ n \in \mathbb{N}. \tag{2.3.24}$$

Then one has

$$\mathcal{A}^r_q((\Sigma_n), \mathbb{X}) = [\mathbb{X}, \mathbb{Y}]_{\theta,q}, \quad \theta = r/R. \tag{2.3.25}$$

This sometimes referred to as Jackson-Bernstein-Theory.

The so called *Re-iteration Theorem* says that interpolating between two interpolation spaces is the same as interpolating between the extreme spaces. This shows that interpolating between Besov spaces again yields Besov spaces. Likewise interpolating between approximation classes yields approximation classes.

Now the role of the parameter q becomes clearer as a means to distinguish between different possible ways of interpolation yielding the same primary smoothness s, say.

We refer to [32, 31] for more details.

Exercise 2.3.4. *Let* $\mathcal{P}_n$ *denote again the uniform partition of* $\Omega = (0, 1)$ *into* n *subintervals of equal length* $h = 1/n$ *and let* $\Sigma_n = \mathbb{P}_m(\mathcal{P}_n)$*,* $m \in \mathbb{N}$ *fixed. Determine the* R *for which the Jackson and Bernstein estimates* (2.3.23), (2.3.24) *hold when* $\mathbb{X} = L_p(\Omega)$*,* $\mathbb{Y} = W^k(L_p(\Omega))$*.*

Topography of Function Spaces: Interpolation theory helps also in deriving the following chart showing which spaces are embedded in which ones. We assume that Ω is bounded or a torus.

We summarize the embedding of Besov and Sobolev spaces on Figure 1. In this figure, a smoothness space measuring s derivatives in L_p - such as $W^s(L_p)$ or $B^s_q(L_p)$ for some $q > 0$ - is represented by the point $(1/p, s)$ in the upper-right quadrant. If $\mathbb{X}$ is a smoothness space on a domain Ω that satisfies the uniform cone condition, which is represented by the point $(1/p, t)$, then

- Spaces $\mathbb{Y}$ represented by a point in region

  $$I := \{(1/r, s) \ : \ s > t \text{ and } r > p\}$$

  embed in $\mathbb{X}$ when Ω is bounded, and this embedding is compact.

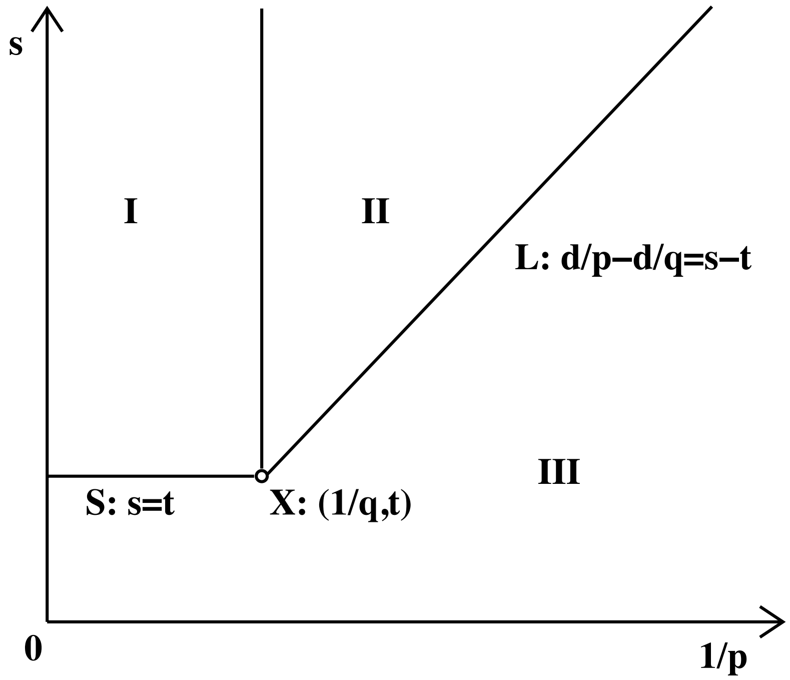


Figure 1: Graphical summary of Besov and Sobolev embeddings

- Spaces $\mathbb{Y}$ represented by a point in region

  $$\mathrm{II} := \{(1/r, s)\ :\ s > t,\ r \leq p \text{ and } s - t > d/r - d/p\}$$

  embed in $\mathbb{X}$, and this embedding is compact when Ω is bounded.

- Spaces $\mathbb{Y}$ represented by a point on segment

  $$S := \{(1/r, s)\ :\ s = t \text{ and } r > p\}$$

  may embed in $\mathbb{X}$ when Ω is bounded, depending on the precise definition of $\mathbb{X}$ and $\mathbb{Y}$ - for example $B^s_{q_1}(Lr(\Omega)) \subset B^s_{q_2}(L_p(\Omega))$ if and only if $q_1 \leq q_2$ - and this embedding is not compact.

- Spaces $\mathbb{Y}$ represented by a point on line

  $$L := \{(1/r, s)\ :\ s > t \text{ and } s - t = d/r - d/p\}$$

  may embed in $\mathbb{X}$ depending on the precise definition of $\mathbb{X}$ and $\mathbb{Y}$ - for example $B^s_{q_1}(L_r(\Omega)) \subset B^t_{q_2}(L_p(\Omega))$ if and only if $q_1 \leq q_2$ - and this embedding is not compact.

- Spaces $\mathbb{Y}$ represented by a point in the remaining region III do not embed in $\mathbb{X}$.

The spaces $B^k_q(L_\infty(\Omega))$, $W^k(L_\infty(\Omega))$, $C^k(\Omega)$, $k \in \mathbb{N}_0$, $0 < q \le \infty$, are all associated with the point $(0,k) = (1/\infty, k)$ on the vertical coordinate axis. The spaces $B^1_q(L_1(\Omega))$, $W^1(L_1(\Omega)), BV(\Omega)$ are associated with the point $(1,1) = (1/1,1)$. When r decreases, i.e., $1/r$ increases to the right, the spaces L_r get larger, i.e., smoothness of the same order s gets weaker when r decreases. The space $B^s_r(L_r(\Omega))$ is still embedded in $L_p(\Omega)$ as long as

$$\frac{1}{r} \le \frac{s}{d} + \frac{1}{p}, \quad \frac{1}{\tau} = \frac{s}{d} + \frac{1}{p} \quad \text{being the critical embedding.} \tag{2.3.26}$$

2.4 A First Application - Back to Adaptive Piecewise Polynomial Approximation

Let $\Omega := (0,1)^d$, $\mathcal{P}_m$ the space of polynomials of total order m (degree $m-1$) on $\mathbb{R}^d$. Let

$$\Sigma_n := \bigcup \{\mathcal{P}_m(\mathcal{P}) : \#(\mathcal{P}) \le n,\ \mathcal{P} \text{ a dyadic partition}\}.$$

Here we call $\mathcal{P}$ a dyadic partition if all its cells result from a successive dyadic refinement of some "father cell" starting with the root $\{\Omega\}$. A dyadic partition of a cube means its subdivision into 2^d congruent cubes (the children).

Remark 2.4.1. *The number of degrees of freedom carried by Σ_n is not n but $n \dim \mathcal{P}_m = n\binom{m-1+d}{d}$. Since m, d are fixed the number of degrees of freedom is still uniformly proportional to n. Since all estimates involve some fixed constant we retain the simpler notation just using n.*

We shall later use that such partitions can be identified with the set of leaves of a *tree* with single root $\{\Omega\}$, see Figure 2.

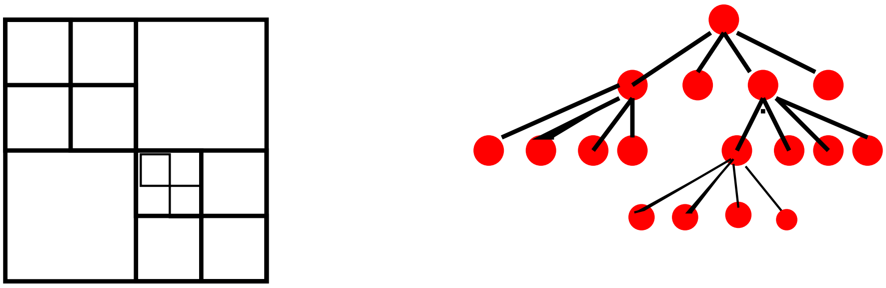


Figure 2: Dyadic refinements, tree representation

We consider the following setting:

- $(\mathcal{P}_j)_{j\in\mathbb{N}_0}$ hierarchy of uniform nested partitions; $\mathcal{T} := \bigcup_{j\in\mathbb{N}_0} \mathcal{P}_j$ partition tree with roots in $\mathcal{P}_0$
- $\mathfrak{P}_n :=$ set of all dyadic leaf-partitions with at most n cells;

- m fixed, $\Sigma_n := \bigcup\{\mathbb{P}_m(\mathcal{P}) : \mathcal{P} \in \mathfrak{P}_n\}$
- $e(f,T)_p := \inf_{g\in\mathbb{P}_m} \|f-g\|_{L_p(T)}$, $\quad \mathcal{C}(T) =$ set of children of $T \in \mathcal{T}$

The following algorithm is in principle identical to Algorithm 2.1.1, only the ingredients: norm, higher polynomial order, local error indicator, differ.

Algorithm 2.4.1.

```
1: Initialize: Set threshold $\eta > 0$, $\mathcal{B} := \{T \in \mathcal{P}_0 : e(f,T)_p > \eta\}$ set of "bad" cells, $\mathcal{G} := \{T \in \mathcal{P}_0 : e(f,T)_p \le \eta\}$ set of "good" cells
2: Output: partition $\mathcal{P}_\eta$ with $e(f,T)_p \le \eta, T \in \mathcal{P}_\eta$
3: while $\mathcal{B} \neq \emptyset$
4:     for $T \in \mathcal{B}$
           $\mathcal{B} \to (\mathcal{B}\setminus\{T\}) \cup \{T' \in \mathcal{C}(T) : e(f,T')_p > \eta\}$
           $\mathcal{G} \to \mathcal{G} \cup \{T' \in \mathcal{C}(T) : e(f,T')_p \le \eta\}$
5:     end for do
6: end while do
7: output $\mathcal{P}_\eta = \mathcal{G}$
```

Let $g_T \in \mathcal{P}_m(T)$ denote a best approximation to f on T. Then

$$g_\eta := \sum_{T\in\mathcal{P}_\eta} g_T \chi_T \in \mathcal{P}_m(\mathcal{P}_\eta)$$

is the piecewise polynomial approximant corresponding to the output partition $\mathcal{P}_\eta$.

Theorem 2.4.1. *Let* $0 < p \le \infty$, $f \in B^s_\tau(L_\tau(\Omega))$, $0 < s < m$, $\delta := s - \frac{d}{\tau} + \frac{d}{p} > 0$, $g_\eta := \underset{g\in\mathcal{P}_m(\mathcal{P}_\eta)}{\operatorname{argmin}} \|f-g\|_{L_p(\Omega)}$, *then*

$$\|f-g_\eta\|_{L_p(\Omega)} \le C(\#\mathcal{P}_\eta)^{-s/d}|f|_{B^s_\tau(L_\tau(\Omega))} \quad (C = C(p,\tau,m,\delta)) \tag{2.4.1}$$

In addition

$$\sigma_n(f)_{L_p(\Omega)} = \inf_{g\in\Sigma_n} \|f-g\|_{L_p(\Omega)} \le Cn^{-s/d}|f|_{B^s_\tau(L_\tau(\Omega))} \tag{2.4.2}$$

Hence

$$B^s_\tau(L_\tau(\Omega)) \subset \mathcal{A}^{s/d}\big((\Sigma_n), L_p(\Omega)\big), \quad \textit{when} \quad \frac{1}{\tau} < \frac{s}{d} + \frac{1}{p} \tag{2.4.3}$$

Comments 2.4.1. *1. Note that we do* not *have* $\|f-g_\eta\|_{L_p(\Omega)} \le \eta$ *when* $p < \infty$. *One only has*

$$\|f-g_\eta\|^p_{L_p(\Omega)} \le \eta^p \#\mathcal{P}_\eta.$$

2. *In contrast to Theorem 2.1.3 we have now a result for variable smoothness order. For* $s = 1$ *however Theorem 2.1.3 gives a slightly stronger result because for any* $\tau > 1$ *the condition* $f' \in L_\tau(\Omega)$ *implies* $f \in M_1(\Omega)$ *(here:* $s = 1, p = \infty, d = 1$*).*

3. *For approximation in one spatial variable (*$d = 1$*) we obtained error bounds of the form* $\lesssim n^{-s}$ *when* s *is the degree of smoothness and* n *the number of degrees of freedom. In Theorem 2.1.1 the number of degrees of freedom*n *is related to a* mesh-size $h = 1/n$*, i.e.,* $n = h^{-1}$*. For a uniform Cartesian grid of mesh-size* h *for the unit square (*$d = 2$*) we have* $n = h^{-2}$ *and in general for the unit cube in* $\mathbb{R}^d$*, the mesh-size* h *corresponds to* $n = h^{-d}$ *degrees of freedom. More generally, consider for bounded domains in* $\mathbb{R}^d$ *a* quasi-uniform *partition* $\mathcal{P}_h$*, i.e., all cells have approximately the same diameter* h *and all cells are "ball-like", i.e., the ratio of the radii of the smallest circumscribed ball and the largest inscribed ball remains uniformly bounded over all cells. Such partitions are called* shape-regular*. One then has*
$$\#\mathcal{P}_h \sim h^{-d}, \quad h \sim (\#\mathcal{P}_h)^{-1/d}. \tag{2.4.4}$$
Hence, when measuring accuracy not by mesh-size but by the number of degrees of freedom, because a mesh-size does not make sense for locally refined partitions, we expect that errors decay at best like
$$\text{error for smoothness } s \text{ and } n \text{ degrees of freedom scales at best like } n^{-s/d}. \tag{2.4.5}$$
For smoothness s *the rate* $n^{-s/d}$ *is best possible. Thus smoothness becomes less and less effective when the spatial dimension* d *increases. The error estimates* (2.4.1)*,* (2.4.2) *reflect exactly this behavior.*

4. *The theorem says that the simple adaptive algorithm realizes the optimal order for approximands in the Besov space* $B^s_\tau(L_\tau(\Omega))$*. But one can actually give examples of elements in the approximation class* $\mathcal{A}^{s/d}\big((\Sigma_n), L_p(\Omega)\big)$ *which do not belong to* $B^s_\tau(L_\tau(\Omega))$*. So there exist functions which can be approximated by elements from* Σ_n *with order* $n^{-s/d}$ *but the corresponding partition is not found by the simple greedy strategy in Algorithm 2.4.1.*

5. *We can relate Theorem 2.4.1 to the embedding diagram in Figure 1. This is illustrated below in Figure 2.4. The smaller* τ*, subject to the constraint* $\frac{1}{\tau} < \frac{s}{d} + \frac{1}{p}$*, the larger the space* $B^s_\tau(L_\tau(\Omega))$*, but the theorem is no longer valid when* $\frac{1}{\tau} = \frac{s}{d} + \frac{1}{p}$*, i.e., when* $B^s_\tau(L_\tau(\Omega))$ *is on the critical embedding line. In brief, the simple adaptive scheme provides optimal orders (although not exhausting* $\mathcal{A}^{s/d}\big((\Sigma_n), L_p(\Omega)\big)$*) up to the embedding line which is as far as one can go so that smoothness* s *still ensures embedding in* $L_p(\Omega)$*.*

Proof. (Sketch) (i) Since cells in $\mathcal{P}_\eta$ are good $\rightsquigarrow$
$$\|f - g_\eta\|_{L_p(\Omega)} \le (\#\mathcal{P}_\eta)^{1/p}\eta. \tag{2.4.6}$$

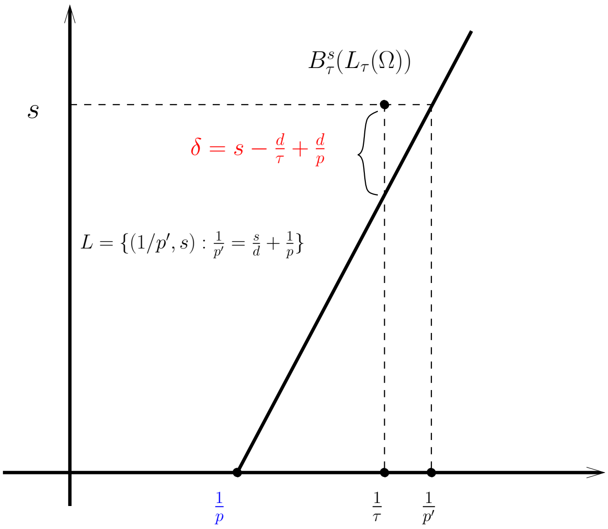

$B^s_\tau(L_\tau(\Omega))$
s
$\delta = s - \frac{d}{\tau} + \frac{d}{p}$
$L = \{(1/p', s) : \frac{1}{p'} = \frac{s}{d} + \frac{1}{p}\}$
$\frac{1}{p}$
$\frac{1}{\tau}$
$\frac{1}{p'}$

Figure 3: $B^s_\tau(L_\tau(\Omega)) \subset \mathcal{A}^{s/d}\big((\Sigma_n), L_p(\Omega)\big)$, $\frac{1}{\tau} < \frac{s}{d} + \frac{1}{p}$

(ii) count $\#\mathcal{P}_\eta$: in Theorem 2.1.3 the counting is based on estimating the local errors by the maximal function. Here, the same strategy works, estimating local errors this time by local Besov-norms. Whitney's Theorem says $\inf_{g \in \mathbb{P}_m} \|f - g\|_{L_p(\Omega)} \le C\omega_m(f, \operatorname{diam}\Omega, \Omega)_p$, (see (2.3.15)) $\rightsquigarrow$

$$\inf_{g \in \mathbb{P}_m} \|f - g\|_{L_p(T)} \le C(\operatorname{diam} T)^{s - \frac{d}{\tau} + \frac{d}{p}} |f|_{B^s_\tau(L_\tau(T))}, \tag{2.4.7}$$

recall that $s - \frac{d}{\tau} + \frac{d}{p} = \delta$, $\operatorname{diam} T \sim 2^{-j}$ when T has refinement generation j.

Let $\mathcal{P}_j$ denote all dyadic cells of generaltion/level j (mesh-size 2^{-j}). $T \in \mathcal{P}_\eta \cap \mathcal{P}_j$ $\Rightarrow e(f, P(T))_p > \eta$, $P(T)$ parent of T, $T \in \mathcal{P}_j \rightsquigarrow \operatorname{diam} T \sim 2^{-j}$, then (2.4.7) $\rightsquigarrow$

$$\eta \lesssim 2^{-j\delta} |f|_{B^s_\tau(L_\tau(P(T)))}. \tag{2.4.8}$$

Next note that the Besov-seminorm is **set-subadditive**. This means that for any partition $\mathcal{P}$ of Ω

$$\sum_{T \in \mathcal{P}} |f|^\tau_{B^s_\tau(L_\tau(T))} \lesssim |f|^\tau_{B^s_\tau(L_\tau(\Omega))},$$

compare integer order Sobolev semi-norms. This cannot directly be seen from Definition 2.3.1 because the modulus of continuity is not obviously subadditive with respect to sets. However, there is a variant, the so called *averaged modulus* where the "sup" is replaced by an average

$$\bar{\omega}_m(f, t, \Omega)_p := \left(\frac{1}{|B(0,t)|} \int_{B(0,t)} \|\Delta^m_h f\|^p_{L_p(\Omega_{mh})} dh \right)^{1/p}. \tag{2.4.9}$$

It can be shown that both variants are equivalent, i.e., each one can be bounded by a uniform constant multiple of the other one, where the constants depend only on m, p. Thus, replacing the standard modulus by the averaged one in the definition of the Besov semi-norm, yields an equivalent Besov semi-norm and one can use that the Besov-seminorm is indeed set-subadditive. Thus, summing over all level-j cells $T \in \mathcal{P}_\eta \cap \mathcal{P}_j$, (2.4.8) yields

$$\#(\mathcal{P}_\eta \cap \mathcal{P}_j)\eta^\tau \lesssim 2^{-j\tau\delta}|f|^\tau_{B^s_\tau(L_\tau(\Omega))} \quad \rightsquigarrow \quad \#(\mathcal{P}_\eta \cap \mathcal{P}_j) \lesssim 2^{-j\tau\delta}\eta^{-\tau}|f|^\tau_{B^s_\tau(L_\tau(\Omega))}. \tag{2.4.10}$$

On the other hand,

$$\#(\mathcal{P}_\eta \cap \mathcal{P}_j) \leq \#(\mathcal{P}_j) \overset{(2.4.4)}{\lesssim} 2^{jd}. \tag{2.4.11}$$

Hence

$$\#(\mathcal{P}_\eta \cap \mathcal{P}_j) \lesssim \min\left\{2^{jd}, 2^{-j\tau\delta}\eta^{-\tau}|f|^\tau_{B^s_\tau(L_\tau(\Omega))}\right\}. \tag{2.4.12}$$

Now j_0 tip over point where both bounds are essentially of equal order, i.e., $2^{j_0 d} \sim 2^{-j_0\tau\delta}\eta^{-\tau}|f|^\tau_{B^s_\tau(L_\tau(\Omega))}$. Then one obtains

$$\begin{aligned}\#(\mathcal{P}_\eta) &= \sum_{j\geq 0}\#(\mathcal{P}_\eta \cap \mathcal{P}_j) \lesssim \sum_{j\leq j_0} 2^{jd} + \sum_{j>j_0} 2^{-j\tau\delta}\eta^{-\tau}|f|^\tau_{B^s_\tau(L_\tau(\Omega))}\\ &\lesssim 2^{j_0 d} + 2^{-j_0\tau\delta}\eta^{-\tau}|f|^\tau_{B^s_\tau(L_\tau(\Omega))} \lesssim 2^{-j_0\tau\delta}\eta^{-\tau}|f|^\tau_{B^s_\tau(L_\tau(\Omega))}.\end{aligned} \tag{2.4.13}$$

Now $2^{j_0 d} \sim 2^{-j_0\tau\delta}\eta^{-\tau}|f|^\tau_{B^s_\tau(L_\tau(\Omega))}$ means that

$$2^{j_0} \sim \left(\eta^{-\tau}|f|^\tau_{B^s_\tau(L_\tau(\Omega))}\right)^{-\frac{\tau\delta}{d+\tau\delta}},$$

and therefore

$$2^{-j_0\tau\delta}\eta^{-\tau}|f|^\tau_{B^s_\tau(L_\tau(\Omega))} \lesssim \left(\eta^{-\tau}|f|^\tau_{B^s_\tau(L_\tau(\Omega))}\right)^{\frac{d}{d+\tau\delta}}. \tag{2.4.14}$$

Thus, (2.4.13) yields

$$\#(\mathcal{P}_\eta) \lesssim \left(\eta^{-\tau}|f|^\tau_{B^s_\tau(L_\tau(\Omega))}\right)^{\frac{d}{d+\tau\delta}},$$

and hence

$$\eta \lesssim |f|_{B^s_\tau(L_\tau(\Omega))}\#(\mathcal{P}_\eta)^{-\frac{d+\tau\delta}{\tau d}} = |f|_{B^s_\tau(L_\tau(\Omega))}\#(\mathcal{P}_\eta)^{-\frac{s}{d}-\frac{1}{p}}. \tag{2.4.15}$$

Insert this into (2.4.6) ⇝ (2.4.1). □

Remark 2.4.2. *The same arguments apply to other isotropic refinement methods (to be learnt later) such as shape preserving bisections like "newwest-vertex-bisection" for simplices. Then the diameter of refined cells shrinks at a slower rate ρ^{-1} for some $\rho > 1$. The result would remain the same because of the same geometric series effect.*

3 Tree-Based Algorithms

3.1 Some Basic Notions

A tree is a specific instance of a graph, namely, it is a set of nodes $\mathcal{T}$, certain pairs of which correspond to the edges of the graph. We assume that when a path of edges connects two nodes, this path is unique, which means that the graph has no cycle. In addition, one fixes an orientation for any edge between two nodes T and T', by saying that T' is a *child* of T and that T is its *parent*. The orientation must be chosen so that every node has at most one parent. A node without a parent is called a root. We denote by $\mathcal{C}(T)$ the set of all *children* of T and refer to any two elements of $\mathcal{C}(T)$ as *siblings*. We confine the subsequent discussions to trees for which each parent has at most a fixed number $M \in \mathbb{N}$ of children, i.e.

$$2 \leq \#(\mathcal{C}(T)) \leq M, \quad T \in \mathcal{T}. \tag{3.1.1}$$

In addition, we require that every node in a tree has a finite number of ancestors, and therefore every node is linked to a single root by a unique path.

So far this permits the existence of several roots. In this case one sometimes speaks of a *forest*. For simplicity we consider in what follows (mainly) the case of trees $\mathcal{T}$ with a *single* root denoted by $R(\mathcal{T})$. Most findings however carry over to forests.

Given a domain Ω and a *refinement rule*, for instance, bisecting a triangle into two triangles by splitting the longest edge, or by subdividing a cube in $\mathbb{R}^d$ into 2^d congruent cubes (dyadic subdivision) gives rise to a (geometric) tree. In fact, view Ω as the root node/cell; a consecutive refinement of a cell – viz. a node T – creates the set of children $\mathcal{C}(T)$ as new nodes. Repeating this process creates a tree $\mathcal{T}$, see Figure 2. Refining all cells ad infimum creates an infinite tree, called the *master tree* $\mathcal{T}^*$.

A cell T in a finite tree $\mathcal{T} \prec \mathcal{T}^*$ is called a *leaf* if none of its children (in $\mathcal{T}^*$) belongs to $\mathcal{T}$. A tree is called *complete* if one of the elements of $\mathcal{C}(\mathcal{T})$ belongs to $\mathcal{T}$ then $\mathcal{C}(T) \subset \mathcal{T}$.

The set of leaves of a complete tree $\mathcal{T}$

$$\mathcal{L}(\mathcal{T}) := \{T \in \mathcal{T} : \mathcal{C}(T) \cap \mathcal{T} = \emptyset\}$$

forms a *partition* of the root cell Ω. We say a partition $\mathcal{P}$ is induced by a tree $\mathcal{T}$ if $\mathcal{P} = \mathcal{L}(\mathcal{T})$. Conversely, every partition generated by a successive refinement of the root corresponds to a tree which encodes the refinement history. This is expressed by writing

$$\mathcal{P} = \mathcal{L}(\mathcal{T}), \quad \mathcal{T} = \mathcal{T}(\mathcal{P}) \text{ if } \mathcal{L}(\mathcal{T}) = \mathcal{P}.$$

Exercise 3.1.1. *Obviously* $\#(\mathcal{L}(\mathcal{T})) \leq \#(\mathcal{T})$. *Show that* $\#(\mathcal{T}) \leq 2\#(\mathcal{L}(\mathcal{T}))$.

Why considering trees? Algorithm 2.4.1 is the simplest example of a tree-based algorithm that generates a partition $\mathcal{P} = \mathcal{L}(\mathcal{T})$ where $\mathcal{T}(\mathcal{P})$ encodes its refinement history. One way to interprete Algorithm 2.4.1 is that it attempts to grow a tree which realizes a given target accuracy at the expense of a possibly small partition. More generally, one tries to find among all trees with at most n leaf nodes the one that optimizes a certain criterion. In this section we discuss this issue in a somewhat more abstract setting which covers the situation of Algorithm 2.4.1 as a special case. The quality criterion will be given in terms of error functionals.

Suppose we can associate with each node/cell $T \in \mathcal{T}^*$ an *error functional* $e(T) = e(f;T)$ (which is to represent a local approximation to f on T) satisfying the *weak subadditivity* property

$$\sum_{T' \in \mathcal{L}(\mathcal{T}')} e(T') \leq Ce(T), \quad \text{for any tree } \mathcal{T}' \text{ with } R(\mathcal{T}') = T, \tag{3.1.2}$$

where C is a fixed constant. When $C = 1$ we call the error functional *subadditive*.

Goal (ideal): Given $\varepsilon > 0$ find $\mathcal{T}_\varepsilon \subset \mathcal{T}^*$ such that

$$\mathcal{T}_\varepsilon = \underset{\mathcal{T} \prec \mathcal{T}^*}{\operatorname{argmin}}\{\#\mathcal{T} : e(\mathcal{T}) \leq \varepsilon\}, \quad \text{where} \quad e(\mathcal{T}) := \sum_{T \in \mathcal{L}(\mathcal{T})} e(T). \tag{3.1.3}$$

Example 3.1.1. 1. *Local polynomial approximation in L_p, $0 < p < \infty$:*

$$e(T) = e(f;T)_p^p := \inf_{P \in \mathcal{P}_m} \|f - P\|_{L_p(T)}^p.$$

This error functional is even subadditive ($C = 1$ in (3.1.2)) and it is used in Algorithm 2.4.1.

2. *Empirical errors in* machine learning *(regression, classification). Suppose that ρ is an unknown measure on a space $Z := X \times Y$, $d\rho(x,y) = d\rho(y|x)d\rho_X(x)$, ($d\rho_X$ is the so called marginal measure). The goal is to estimate the* **regression function**

$$f_\rho(x) := \mathbb{E}(y|x) = \int_Y y\, d\rho(y|x) \tag{3.1.4}$$

see Figure 3.1, from independent identically (with respect to ρ) distributed (i.i.d.) samples $Z_n := \{(x_i, y_i) : i = 1, \ldots, n\}$. Think of every x_i as a list of answers to a catalog of questions which are kept by a bank together with a success measure y_i obtained when giving a loan to the ith client who came up with the answers x_i. The regression function then tells the bank what the success rate of the decision would be in expectation.

Estimating f_ρ *means to construct an estimator* $\hat{f} = \hat{f}_{Z_n}$ *that minimizes the* **Risk Functional**

$$\mathcal{E}(f) := \int_Z (y - f(x))^2 d\rho,$$

which can be shown to decompose as

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) + \|f - f_\rho\|^2_{L_2(X,\rho_X)}. \tag{3.1.5}$$

Thus, minimizing the risk means to best approximate the (unknown) regression function in an L_2*-norm, in this case the* L_2*-norm with respect to the (unknown) measure* ρ*.*

A common strategy is to construct the estimator $\hat{f}$ *as a piecewise polynomial on a partition that should be chosen adaptively based on the given data (samples). Given a partition* $\mathcal{P}$ *of the domain on which* f_ρ *lives we determine for each* $T \in \mathcal{P}$ *a polynomial* P_T *determined by*

$$P_T = \underset{P \in \mathcal{P}_m}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - P(x_i))^2 \chi_T(x_i), \quad \hat{f} = \sum_{T \in \mathcal{P}} \chi_T P_T. \tag{3.1.6}$$

Notice that T *should contain sufficiently many samples to determine* P_T*.*

A natural error functional would in this case be

$$e(T) = e(f_\rho; T) := \frac{1}{n} \sum_{i=1}^{n} (y_i - P_T(x_i))^2 \chi_T(x_i), \tag{3.1.7}$$

which is again subadditive. Note that these indicators are computable. Since the output, depending on which way these indicators are used, is based on random samples, it itself is a random variable. *Therefore, the resulting estimator* $\hat{f}$ *is a random variable. In the end, one has to prove how accurate the estimator is, for instance in terms of*

$$\mathbb{E}_{\rho^n}\left(\|\hat{f} - f_\rho\|^2_{L_2(X,\rho_X)}\right).$$

This is done in mathematical/statistical learning theory.

3. *Local error estimators arise also from estimating residuals when solving certain types of PDEs. This will be discussed in detail later.*

Complexity: Consider

$$\mathfrak{T}_n := \{\mathcal{T} \subset \mathcal{T}^* : R(\mathcal{T}) = R(\mathcal{T}^*), \#(\mathcal{L}(\mathcal{T})) \leq n\} \tag{3.1.8}$$

the collection of all finite trees whose set of leaves has at most n elements. A brute force method for accomplishing (3.1.3) would be to compute $e(\mathcal{T})$ for each $\mathcal{T} \in \mathfrak{T}_n$. This would be an NP-hard problem.

Remark 3.1.1. *A complete search through* $\mathfrak{T}_n$ *has exponential cost in* n*, i.e.,* $\#(\mathfrak{T}_n) \sim a^n$ *for some* $a > 1$*.*

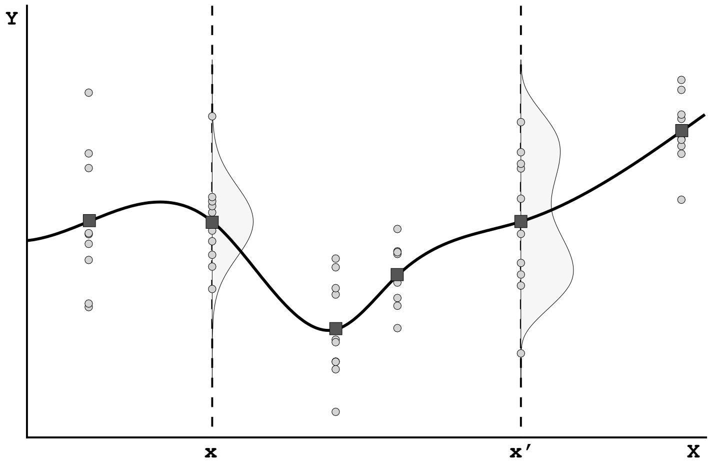


Figure 4: regression function

3.2 The Binev-DeVore Algorithm

The render the **goal** more tractable one can relax is slightly. This leads to the notion of *weak optimality*.

Definition 3.2.1. *We call a tree $\mathcal{T}$* weakly optimal *for an error function* e *associated with $\mathcal{T}^*$ if there exists a positive constant* $1 \le C^* < \infty$ *such that for any tree $\tilde{\mathcal{T}} \subset \mathcal{T}^*$ one has*

$$C^* e(\tilde{\mathcal{T}}) \le e(\mathcal{T}) \quad \Rightarrow \quad \#(\mathcal{T}) \le C^* \#(\tilde{\mathcal{T}}). \tag{3.2.1}$$

The following modification of Algorithm 2.4.1 realizes weak optimality. Specifically, consider the following *modified* error functionals:

$$\tilde{e}(T) := \begin{cases} e(T), & T = R(\mathcal{T}); \\ \left(\frac{1}{e(T)} + \frac{1}{\tilde{e}(\hat{T})}\right)^{-1}, & T \in \mathcal{C}(\hat{T}),\ \hat{T}\,\text{parent of}\,T. \end{cases} \tag{3.2.2}$$

Thus, the larger the level of T (i.e., the number of edges needed to connect T to the root $R(\mathcal{T})$) the smaller becomes the modified error functional $\tilde{e}(T)$ compared with $e(T)$.

Algorithm 3.2.1.

```
1: Initialize: ${R(\mathcal{T}^*)} \to \mathcal{T}$
2: while $e(\mathcal{T}) > \varepsilon$ do
3:     for $T \in \operatorname{argmax}_{T' \in \mathcal{L}(\mathcal{T})} \tilde{e}(T')$ set $\mathcal{T} \to \mathcal{T} \cup \mathcal{C}(T)$
4: end while
5: Output: $\mathcal{T}_\varepsilon$ with $e(\mathcal{T}_\varepsilon) \le \varepsilon$
```

The algorithm is another instance of a greedy method. It successively refines the cell with the currently largest modified error functional.

To describe the performance of Algorithm 3.2.1 it is convenient to use a slightly different complexity measure. For a given finite tree $\mathcal{T} \subset \mathcal{T}^*$ let

$$n(\mathcal{T}) := \#(\mathcal{T}) - \#(\mathcal{L}(\mathcal{T})). \tag{3.2.3}$$

By Exercise 3.1.1, we have

$$n(\mathcal{T}) \sim \#(\mathcal{L}(\mathcal{T})) \sim \#(\mathcal{T}). \tag{3.2.4}$$

Define

$$\sigma_n(e) := \min_{\mathcal{T} \subset \mathcal{T}^* : n(\mathcal{T}) \leq n} e(\mathcal{T}). \tag{3.2.5}$$

A proof of the following result can be found in [6].

Theorem 3.2.1. *Assume that* (3.1.2) *holds with* $C = 1$ *and let the modified error functionals be defined by* (3.2.2)*. Then any output* $\mathcal{T}$ *of Algorithm 3.2.1 satisfies*

$$e(\mathcal{T}) \leq \Big(\frac{n(\mathcal{T})}{n(\mathcal{T}) - k}\Big) \sigma_k(e) \quad \textit{whenever} \quad k < n(\mathcal{T}). \tag{3.2.6}$$

Thus, choosing, in particular, $k = \lfloor n(\mathcal{T})/2 \rfloor$ one sees that (3.2.1) holds with $C^* = 2$. Therefore, one obtains the following immediate consequence of Theorem 3.2.1.

Corollary 3.2.1. *Let* Σ_n *denote the set of all subtrees of the master tree* $\mathcal{T}^*$ *satisfying* $n(\mathcal{T}) \leq n$ *where* $n(\mathcal{T})$ *is defined by* (3.2.3)*. Denoting by* $\mathcal{A}^s_\infty((\Sigma_n))$ *the set of those error functions* e *associated with* $\mathcal{T}^*$ *for which* $\sigma_n(e) \leq Mn^{-s}$ *holds for some constant* $M < \infty$*. Then, for each* $n \in \mathbb{N}$ *Algorithm 3.2.1 outputs a tree* $\mathcal{T}_n$ *such that*

$$e(\mathcal{T}_n) \leq 2^{1+1/s} M n^{-s}, \quad n \in \mathbb{N}. \tag{3.2.7}$$

Hence the algoritm is rate-optimal *for any polynomial convergence rate.*

Algorithm 3.2.1 is superior to Algorithm 2.4.1 since it applies to a wider spectrum of applications and realizes the order $O(n^{-r})$ for *all* functions in a corresponding approximation class $\mathcal{A}^r$.

3.3 Optimal pruning: CART

The modified greedy algorithm discussed in the previous section *grows* trees with weakly optimal approximation properties in the sense of (3.2.1). In this section we discuss a strategy for *coarsening* a given tree in a way that one obtains a sequence of subtrees that provide even *optimal* approximations within the initial tree *without* a complete search. It is based on the concept of *optimal pruning* which goes back to Leo Breiman and Jerome H. Friedman in 1973 [8] who used it for the the generation of *Classification and Regression Trees* (CART) in the context of machine learning. In this context local error indicators are obtained by sample measurements in terms of the local least squares errors (3.1.7).

For simplicity, we continue to assume that the underlying master tree $\mathcal{T}^*$ has a single root $T_0 = R(\mathcal{T}^*)$ and that the error functionals e associated with a fixed master tree $\mathcal{T}^*$ are subadditive, i.e., satisfy (3.1.2) with $C = 1$.

Ideally, given a target accuracy $\varepsilon > 0$ and an error function e, we would like to find a tree $\mathcal{T}(\varepsilon) \subset \mathcal{T}^*$ that satisfies (3.2.1) with $C^* = 1$, i.e.,

$$e(\mathcal{T}(\varepsilon)) \leq \varepsilon, \quad e(\mathcal{T}) \leq \varepsilon \;\Rightarrow\; \#(\mathcal{T}) \geq \#(\mathcal{T}(\varepsilon)). \tag{3.3.1}$$

We discuss in this section in which sense pruning gets close to such *ideal* approximations without going through a complete tree search. The underlying method has two major stages:

- The first one is to *grow* a finite tree $\hat{\mathcal{T}}$ of $\mathcal{T}^*$ which, roughly speaking, is large enough for a given range of target accuracies.
- Given such a $\hat{\mathcal{T}}$, the second step consists in reducing $\hat{\mathcal{T}}$ to certain subtrees with the aid of a particular *pruning* strategy which turn out to be optimal in the sense of (3.3.1) at least *in* $\hat{\mathcal{T}}$.

A natural way of generating $\hat{\mathcal{T}}$ is to apply Algorithm 3.2.1. We describe now the second stage, namely the process of pruning. A tree $\mathcal{T}'$ is called a *pruned subtree* of $\mathcal{T}$ if $R(\mathcal{T}') = R(\mathcal{T})$ and if $\mathcal{T}' \subset \mathcal{T}$. We express this by writing

$$\mathcal{T}' \preceq \mathcal{T}, \quad \text{or } \mathcal{T}' \prec \mathcal{T} \text{ when } \mathcal{T}' \neq \mathcal{T}. \tag{3.3.2}$$

Moreover, we denote for $T \notin \mathcal{L}(\mathcal{T})$ by $\mathcal{B}(T;\mathcal{T})$ the *branch* of $\mathcal{T}$ rooted in T, i.e., the tree which is comprised of T and all its descendents in $\mathcal{T}$:

$$\mathcal{B}(T;\mathcal{T}) = \{T' \in \mathcal{T} : T \text{ is an ancestor of } T'\}. \tag{3.3.3}$$

Throughout the remainder of this section we work under the following

Assumption: For any $T \in \hat{\mathcal{T}} \setminus \mathcal{L}(\hat{\mathcal{T}})$ one has

$$e(T) > e(\mathcal{B}(T;\hat{\mathcal{T}})). \tag{3.3.4}$$

Property (3.3.4) means that $\hat{\mathcal{T}}$ does not contain any unnecessary nodes in the following sense. Each refinement of a non-leaf node decreases the error associated with that node. One can always ensure (3.3.4), if necessary, as follows. For any $T \in \hat{\mathcal{T}}$ which is a parent of some leaf node, check whether

$e(T) = \sum_{T' \in \mathcal{C}(T)} e(T')$. If in this case we remove the collection $\mathcal{C}(T)$ one still has $e(\hat{\mathcal{T}} \setminus \mathcal{C}(T)) = e(\hat{\mathcal{T}})$, i.e., this pruning step does not sacrifice accuarcy. It is easy to see that after at most finitely many such removals one arrives at a tree satisfying (3.3.4) but has the same accuracy as the original one.

Complexity Penalization: The key to describing the second stage of pruning is *complexity penalization*: consider for a penalty parameter $\mu \geq 0$ and a given finite subtree $\mathcal{T}$ of $\mathcal{T}^*$ with a single root $R(\mathcal{T}) = R$, the functional

$$E_\mu(\mathcal{T}) := e(\mathcal{T}) + \mu\#(\mathcal{L}(\mathcal{T})). \tag{3.3.5}$$

For the trivial tree $\{T\}$ with root T we briefly write $E_\mu(T) = E_\mu(\{T\}) = e(T) + \mu$.

Clearly, for every finite tree $\mathcal{T} \subset \mathcal{T}^*$ and any $\mu \geq 0$ there exists a tree $\mathcal{T}'$ satisfying

$$E_\mu(\mathcal{T}') = \min_{\tilde{\mathcal{T}} \preceq \mathcal{T}} E_\mu(\tilde{\mathcal{T}}). \tag{3.3.6}$$

Definition 3.3.1. *A tree $\mathcal{T}'$ satisfying* (3.3.6) *is called a μ-optimally pruned subtree of $\mathcal{T}$.*

Remark 3.3.1. *Clearly, when growing $\mathcal{T}$ the first summand $e(\mathcal{T})$ decreases but the second summand increases with the number of the leaf nodes (complexity of the corresponding partition). Therefore*

- *minimizing for a given penalty parameter μ the functional $E_\mu(\mathcal{T})$ over all finite subtrees seeks a compromize between accuracy and complexity;*
- *Increasing μ decreases the size of minimizers of E_μ. Specifically, when*

$$\mu \geq \frac{e(R(\mathcal{T})) - e(\mathcal{T})}{\#\mathcal{L}(\mathcal{T}) - 1}, \tag{3.3.7}$$

then the root $\{R(\mathcal{T})\}$ is the unique minimal μ-optimally pruned subtree of $\mathcal{T}$.

A frequently used tool is to decompose a tree into the branches of the root children. Notice that $e(\mathcal{T}) = \sum_{T' \in \mathcal{C}(R)} e(\mathcal{B}(T'; \mathcal{T}))$ and

$$\#(\mathcal{L}(\mathcal{T})) = \sum_{T' \in \mathcal{C}(R)} \#(\mathcal{L}(\mathcal{B}(T', \mathcal{T}))), \tag{3.3.8}$$

so that

$$E_\mu(\mathcal{T}) = \sum_{T' \in \mathcal{C}(R(\mathcal{T}))} E_\mu(\mathcal{B}(T'; \mathcal{T})). \tag{3.3.9}$$

The following simple fact that μ-optimality is inherited by branches is useful in what follows.

Exercise 3.3.1. *If for some $\mathcal{T}_\mu \neq \{R\}$ one has $E_\mu(\mathcal{T}_\mu) = \min_{\tilde{\mathcal{T}} \preceq \mathcal{T}} E_\mu(\tilde{\mathcal{T}})$, then for each $T' \in \mathcal{C}(R)$ the branch $\mathcal{B}(T'; \mathcal{T}_\mu)$ is a μ-optimally pruned subtree of $\mathcal{B}(T'; \mathcal{T})$.*

Uniqueness of Minimal μ-Optimally Pruned Subtrees: It is less obvious that there exists for each $\mu \geq 0$ a *unique* μ-optimally pruned subtree of *minimal* cardinality. To see this the main vehicle is the following recursion.

Lemma 3.3.1. *For any finite $\mathcal{T} \subset \mathcal{T}^*$ with root $\mathrm{R} = \mathrm{R}(\mathcal{T})$ one has*

$$\min_{\mathcal{T}' \preceq \mathcal{T}} \mathrm{E}_\mu(\mathcal{T}') = \min \Big\{ \mathrm{E}_\mu(\mathrm{R}), \sum_{\mathrm{T}' \in \mathcal{C}(\mathrm{R})} \min_{\tilde{\mathcal{T}} \preceq \mathcal{B}(\mathrm{T}';\mathcal{T})} \mathrm{E}_\mu(\tilde{\mathcal{T}}) \Big\}. \tag{3.3.10}$$

Proof: It follows from (3.3.9) that for any subtree $\tilde{\mathcal{T}}$ of $\mathcal{T}$ with root R

$$\min \Big\{ \mathrm{E}_\mu(\mathrm{R}), \sum_{\mathrm{T}' \in \mathcal{C}(\mathrm{R})} \min_{\mathcal{T}' \preceq \mathcal{B}(\mathrm{T}';\tilde{\mathcal{T}})} \mathrm{E}_\mu(\mathcal{T}') \Big\} \leq \mathrm{E}_\mu(\tilde{\mathcal{T}}). \tag{3.3.11}$$

To show the converse inequality note first that there is nothing to show when $\min_{\mathcal{T}' \preceq \mathcal{T}} \mathrm{E}_\mu(\mathcal{T}') \geq \mathrm{E}_\mu(\mathrm{R})$. Therefore suppose now that $\min_{\mathcal{T}' \preceq \mathcal{T}} \mathrm{E}_\mu(\mathcal{T}') < \mathrm{E}_\mu(\mathrm{R})$. Since E_μ trivially possesses at least one minimizer in $\mathcal{T}$ rooted in R, we can choose for each $\mathrm{T}' \in \mathcal{C}(\mathrm{R})$ a μ-optimally pruned subtree $\mathcal{T}(\mathrm{T}')$ of $\mathcal{B}(\mathrm{T}';\mathcal{T})$. Defining then

$$\tilde{\mathcal{T}} := \bigcup_{\mathrm{T}' \in \mathcal{C}(\mathrm{R})} \mathcal{T}(\mathrm{T}'), \tag{3.3.12}$$

one readily concludes that

$$\mathrm{E}_\mu(\tilde{\mathcal{T}}) = \sum_{\mathrm{T}' \in \mathcal{C}(\mathrm{R})} \min_{\mathcal{T}' \preceq \mathcal{B}(\mathrm{T}';\mathcal{T})} \mathrm{E}_\mu(\mathcal{T}'), \tag{3.3.13}$$

which completes the proof. □

The next observation asserts the uniqueness of minimizers of minimal size.

Theorem 3.3.1. *Assume that $\mu \geq 0$ and let $\mathcal{T}$ be any finite subtree of $\mathcal{T}^*$ (with single root R). Then there exists a unique minimal μ-optimally pruned subtree $\mathcal{T}_\mu(\mathcal{T})$ of $\mathcal{T}$, i.e.,*

$$\mathrm{E}_\mu(\mathcal{T}_\mu(\mathcal{T})) = \min_{\mathcal{T}' \preceq \mathcal{T}} \mathrm{E}_\mu(\mathcal{T}'), \quad \mathrm{E}_\mu(\mathcal{T}') = \mathrm{E}_\mu(\mathcal{T}_\mu(\mathcal{T})),\ \mathcal{T}' \preceq \mathcal{T} \ \Rightarrow\ \mathcal{T}_\mu(\mathcal{T}) \preceq \mathcal{T}'. \tag{3.3.14}$$

Proof: We only need to show the second relation in (3.3.14). There is nothing to show when the tree is trivial, i.e., $\mathcal{T} = \mathrm{R}(\mathcal{T})$. For any $\mathrm{T} \notin \mathcal{L}(\mathcal{T})$ and any descendant T' of T in $\mathcal{T}$ ($\mathrm{T}' \in \mathcal{B}(\mathrm{T};\mathcal{T})$) let $\ell(\mathrm{T},\mathrm{T}')$ denote the number of edges required to connect T with T'. Then define

$$\mathrm{d}(\mathrm{T},\mathcal{T}) := \max_{\mathrm{T}' \in \mathcal{L}(\mathcal{T})} \ell(\mathrm{T},\mathrm{T}')$$

denote the "maximal distance" of T from $\mathcal{L}(\mathcal{T})$ (in other words the largest refinement level in $\mathcal{T}$). Assume that we have proved for some $k \geq 0$ the assertion for any $\mathcal{T}$ with $d(R(\mathcal{T}), \mathcal{T}) \leq k$ and consider now a tree $\mathcal{T}$ with $d(R(\mathcal{T}), \mathcal{L}(\mathcal{T})) = k+1$. Suppose that there exist two different μ-optimally pruned substrees $\mathcal{T}_{\mu,i} \preceq \mathcal{T}$, $i = 1, 2$, of equal minimal cardinality $\#(\mathcal{T}_{\mu,1}) = \#(\mathcal{T}_{\mu,2})$. In particular, this implies that $E_\mu(\mathcal{T}_{\mu,i}) < E_\mu(R(\mathcal{T}))$, $i = 1, 2$. On the one hand, we then know from Lemma 3.3.1 that

$$E_\mu(\mathcal{T}_{\mu,i}) = \sum_{T' \in \mathcal{C}(R)} \min_{\mathcal{T}' \preceq \mathcal{B}(T';\mathcal{T})} E_\mu(\mathcal{T}'), \quad i = 1, 2, \tag{3.3.15}$$

while, on the other hand, (3.3.9) says that

$$E_\mu(\mathcal{T}_{\mu,i}) = \sum_{T' \in \mathcal{C}(R)} E_\mu(\mathcal{B}(T'; \mathcal{T}_{\mu,i})), \quad i = 1, 2. \tag{3.3.16}$$

Now, since $\mathcal{T}_{\mu,1} \neq \mathcal{T}_{\mu,2}$ there must exist a $T' \in \mathcal{C}(R)$ such that $\mathcal{B}(T'; \mathcal{T}_{\mu,1}) \neq \mathcal{B}(T'; \mathcal{T}_{\mu,2})$. Both must be μ-optimally prunded subtrees of $\mathcal{B}(T'; \mathcal{T})$ with minimal size because $d(T', \mathcal{B}(T'; \mathcal{T}_{\mu,i})) = k$. This is a contradiction. □

Construction of Minimal Optimally Pruned Subtrees: We proceed collecting a few further properties of μ-optimally pruned minimal subtrees. In fact, we often make use of the following immediate stability result.

Exercise 3.3.2. *Suppose that $\mathcal{T}' \preceq \mathcal{T}$. Then*

$$\mathcal{T}_\mu(\mathcal{T}) \preceq \mathcal{T}' \quad \Rightarrow \quad \mathcal{T}_\mu(\mathcal{T}) = \mathcal{T}_\mu(\mathcal{T}'), \tag{3.3.17}$$

i.e., a μ-optimally pruned minimal subtree of a given $\mathcal{T}$ stays optimal in any subtree $\mathcal{T}'$ of $\mathcal{T}$ containing it.

We now turn to an efficient way of generating μ-optimal subtrees with the aid of pruning. As indicated in Remark 3.3.1, the minimal optimally pruned trees become smaller when μ increases. In particular, for μ sufficiently large one eventually has $\mathcal{T}_\mu(\mathcal{T}) = \{R(\mathcal{T})\}$.

This observation generalizes as follows. Consider a node $T \in \mathcal{T} \setminus \mathcal{L}(\mathcal{T})$. Given μ, we can cut the branch $\mathcal{B}(T; \mathcal{T})$ without increasing error functional if and only if

$$e(T) + \mu \leq e(\mathcal{B}(T; \mathcal{T})) + \mu \# \mathcal{L}(\mathcal{B}(T; \mathcal{T})) \;\Leftrightarrow\; \mu \geq \frac{e(T) - e(\mathcal{B}(T; \mathcal{T}))}{\#(\mathcal{L}(\mathcal{B}(T; \mathcal{T}))) - 1}.$$

Thus, the tip-over value of μ for which pruning pays off is given by the function $\zeta(\cdot; \mathcal{T}) : \mathcal{T} \setminus \mathcal{L}(\mathcal{T}) \to \mathbb{R}_+$ defined by

$$\zeta(T; \mathcal{T}) := \frac{e(T) - e(\mathcal{B}(T; \mathcal{T}))}{\#(\mathcal{L}(\mathcal{B}(T; \mathcal{T}))) - 1}, \quad T \in \mathcal{T} \setminus \mathcal{L}(\mathcal{T}). \tag{3.3.18}$$

In fact, one easily verifies the following facts.

Exercise 3.3.3. *For each* $T \in \mathcal{T} \setminus \mathcal{L}(\mathcal{T})$ *one has*

$$\begin{aligned} \mu \leq \zeta(T;\mathcal{T}) \quad &\textit{if and only if} \quad E_\mu(T) \geq E_\mu(\mathcal{B}(T;\mathcal{T})), \\ \mu < \zeta(T;\mathcal{T}) \quad &\textit{if and only if} \quad E_\mu(T) > E_\mu(\mathcal{B}(T;\mathcal{T})). \end{aligned} \tag{3.3.19}$$

That is, the size of $\zeta(T;\mathcal{T})$ relative to the current value of μ, tells whether the descendants of T in $\mathcal{T}$ should be pruned or not. More precisely, let

$$\mu_1 := \min_{T \in \mathcal{T} \setminus \mathcal{L}(\mathcal{T})} \zeta(T;\mathcal{T}). \tag{3.3.20}$$

Thus, when $\mu < \mu_1$ gradually increases, as soon as $\mu = \mu_1$ there is a $T \in \mathcal{T} \setminus \mathcal{L}(\mathcal{T})$ whose branch can be cut without increasing E_{μ_1}. From (3.3.19) one can derive the following observation.

Lemma 3.3.2. *As long as* $\mu < \mu_1$, *defined in* (3.3.20), $\mathcal{T}$ *is the minimal* μ*-optimally pruned subtree of itself. For* $\mu = \mu_1$ *the tree* $\mathcal{T}$ *is still* μ_1*-optimally pruned but no longer minimal. In fact*

$$\mathcal{T}_{\mu_1}(\mathcal{T}) = \{T \in \mathcal{T} : \zeta(\hat{T};\mathcal{T}) > \mu_1 \textit{ for all ancestors } \hat{T} \textit{ of } T\} \cup \{R(\mathcal{T})\}. \tag{3.3.21}$$

Proof. **Exercise** □

Suppose now that we have fixed the initial tree $\hat{\mathcal{T}} =: \mathcal{T}_0$ with root $T_0 = R(\mathcal{T}_0)$, satisfying (3.3.4), i.e., $\mathcal{T}_0$ is the minimal 0-optimally pruned subtree of itself ($\mu = 0$), and define $\mathcal{T}_1 := \mathcal{T}_{\mu_1}(\mathcal{T}_0)$ for μ_1 defined by (3.3.20). Then, given $\mathcal{T}_k \preceq \mathcal{T}_0$ such that $\mathcal{T}_k \neq \{T_0\}$, let

$$\mu_{k+1} := \min_{T \in \mathcal{T}_k \setminus \mathcal{L}(\mathcal{T}_k)} \zeta(T;\mathcal{T}_k), \tag{3.3.22}$$

and define

$$\mathcal{T}_{k+1} := \mathcal{T}_k \setminus \bigcup \{(\mathcal{B}(T;\mathcal{T}_k) \setminus \{T\}) : T \in \mathcal{T}_k \setminus \mathcal{L}(\mathcal{T}_k),\ \zeta(T;\mathcal{T}_k) = \mu_{k+1}\}. \tag{3.3.23}$$

Thus, $\mathcal{T}_{k+1}$ is obtained from $\mathcal{T}_k$ by cutting away those branches in $\mathcal{T}_k$ whose roots minimize $\zeta(\cdot;\mathcal{T}_k)$. Since $\mathcal{T}_{k+1} \prec \mathcal{T}_k$, repeating this process must terminate at some $m := k+1$ when $\mathcal{T}_{k+1} = \{T_0\}$. The properties of the sequence of pruned subtrees

$$\{R(\mathcal{T}_0)\} = \mathcal{T}_m \prec \mathcal{T}_{m-1} \prec \cdots \prec \mathcal{T}_1 \prec \mathcal{T}_0 = \hat{\mathcal{T}} \tag{3.3.24}$$

can be summarized as follows (for a complete proof see [8] or the last paragraph below).

Theorem 3.3.2. *One has* $\mu_0 = 0 < \mu_1 < \cdots < \mu_m$ *and the trees* $\mathcal{T}_k$ *defined by* (3.3.23) *satisfy*

$$\mathcal{T}_k = \{T \in \mathcal{T}_{k-1} : \zeta(\hat{T};\mathcal{T}_{k-1}) > \mu_k \textit{ for all ancestors } \hat{T} \textit{ of } T\} \cup \{T_0\}, \tag{3.3.25}$$

i.e.,

$$\mathcal{T}_k = \mathcal{T}_{\mu_k}(\mathcal{T}_{k-1}) = \mathcal{T}_{\mu_k}(\mathcal{T}_0), \quad k = 1, \dots, m. \tag{3.3.26}$$

Moreover, one has

$$\mathcal{T}_\mu(\mathcal{T}_0) = \begin{cases} \mathcal{T}_0, & \mu < \mu_1, \\ \mathcal{T}_k, & \mu_k \leq \mu < \mu_{k+1},\ 1 \leq k < m, \\ \{T_0\}, & \mu \geq \mu_m. \end{cases} \tag{3.3.27}$$

As stated in (3.3.27), $\mathcal{T}_k = \mathcal{T}_{\mu_k}(\mathcal{T}_{k-1})$ is a μ_{k+1}-optimally pruned subtree of $\mathcal{T}_0$ but no longer its minimal one which is $\mathcal{T}_{k+1}$. Hence $E_{\mu_{k+1}}(\mathcal{T}_k) = E_{\mu_{k+1}}(\mathcal{T}_{k+1})$ which means

$$\mu_{k+1} = \frac{e(\mathcal{T}_{k+1}) - e(\mathcal{T}_k)}{\#(\mathcal{L}(\mathcal{T}_k)) - \#(\mathcal{L}(\mathcal{T}_{k+1}))}, \quad 0 < k < m. \tag{3.3.28}$$

Thus the transition threshold μ_{k+1} is the average local error of the nodes in $\mathcal{L}(\mathcal{T}_k) \setminus \mathcal{L}(\mathcal{T}_{k+1})$.

Using induction, the above findings can be reformulated as follows.

Remark 3.3.2. *Let for* $1 \leq k < m$

$$\zeta_k(T) := \begin{cases} \zeta(T; \mathcal{T}_k), & T \in \mathcal{T}_k \setminus \mathcal{L}(\mathcal{T}_k), \\ \zeta_{k-1}(T), & \textit{otherwise.} \end{cases} \tag{3.3.29}$$

Then one has for $0 \leq \mu < \infty$

$$\mathcal{T}_\mu(\mathcal{T}_0) = \{T \in \mathcal{T}_0 : \zeta_{m-1}(\hat{T}) > \mu \textit{ for all ancestors } \hat{T} \textit{ of } T\} \cup \{T_0\}. \tag{3.3.30}$$

Computational Cost: The computational cost for finding the hierarchy of trees $\mathcal{T}_k$ depends on the cost of assessing the quantities $e(T)$, $T \in \hat{\mathcal{T}}$. Assigning a cost unit to each node, a trivial lower bound for the complexity is therefore the cardinality $\#(\hat{\mathcal{T}})$ of the initial tree. Given the values $e(T)$, $T \in \hat{\mathcal{T}} = \mathcal{T}_0$, one next has to evaluate $\zeta_0(T)$, defined by (3.3.29). This requires determining the quantities $e(\mathcal{B}(T; \mathcal{T}_0))$ for $T \in \mathcal{T}_0 \setminus \mathcal{L}(\mathcal{T}_0)$. Working towards the root and using that for $T' \in \mathcal{C}(\hat{T})$ one has $e(\mathcal{B}(\hat{T}; \mathcal{T}_0)) = \sum_{T' \in \mathcal{C}(\hat{T})} e(\mathcal{B}(T'; \mathcal{T}_0)$, the total cost is again proportional to $\#(\mathcal{T}_0)$. Repeating this argument for $\mathcal{T}_k$, $0 < k \leq m$, for the total cost of computing the quantities $\zeta_k(T), T \in \mathcal{T}_k$, a crude estimate shows that the whole process takes at most the order of

$$\sum_{k=0}^{m} \#(\mathcal{T}_k) \lesssim m\#(\hat{\mathcal{T}}) \tag{3.3.31}$$

operations which is, of course, by far less than a complete search through all subtrees of $\hat{\mathcal{T}}$.

Optimality Properties: It is now straightforward to translate the properties of the trees $\mathcal{T}_k$, $k = 1, \dots, \mathcal{T}_m$, into optimality relations for the underlying approximations.

Theorem 3.3.3. *Assume that* $\mathcal{T}_0 = \hat{\mathcal{T}}$ *satisfies* (3.3.4). *Then the hierarchy of nested subtrees* $\mathcal{T}_k$, $k = 0, \dots, m$, *given by* (3.3.25) *has the following propery. The values*

$$\varepsilon_k := e(\mathcal{T}_k), \quad k = 0, \dots, m, \tag{3.3.32}$$

satisfy

$$\varepsilon_0 < \varepsilon_1 < \cdots < \varepsilon_m, \tag{3.3.33}$$

and

$$\mathcal{T} \subseteq \hat{\mathcal{T}},\ e(\mathcal{T}) \leq \varepsilon_k \quad \Rightarrow \quad \#(\mathcal{T}) \geq \#(\mathcal{T}_k), \quad k = 0, \dots, m, \tag{3.3.34}$$

i.e., the trees $\mathcal{T}_k$ *are subtrees of* $\hat{\mathcal{T}}$ *of minimal cardinality realizing accuracy* ε_k.

Proof. Since $\mathcal{T}_{k+1} \prec \mathcal{T}_k$ $k = 0, \dots, m$, it is clear that $\varepsilon_k \leq \varepsilon_{k+1}$. On the other hand, $\varepsilon_k = \varepsilon_{k+1}$ would imply

$$\begin{aligned} E_{\mu_k}(\mathcal{T}_{k+1}) &= e(\mathcal{T}_{k+1}) + \mu_k \#(\mathcal{L}(\mathcal{T}_{k+1})) \\ &= e(\mathcal{T}_k) + \mu_k \#(\mathcal{L}(\mathcal{T}_{k+1})) \\ &< e(\mathcal{T}_k) + \mu_k \#(\mathcal{L}(\mathcal{T}_k)) \\ &= E_{\mu_k}(\mathcal{T}_k), \end{aligned} \tag{3.3.35}$$

contradicting the optimality of $\mathcal{T}_k$. Concerning (3.3.34), suppose that $e(\mathcal{T}) \leq \varepsilon_k = e(\mathcal{T}_k)$ and $\#(\mathcal{T}) < \#(\mathcal{T}_k)$. Then

$$E_{\mu_k}(\mathcal{T}) = e(\mathcal{T}) + \mu_k \#(\mathcal{L}(\mathcal{T})) < e(\mathcal{T}_k) + \mu_k \#(\mathcal{L}(\mathcal{T}_k)) = E_{\mu_k}(\mathcal{T}_k) \tag{3.3.36}$$

contradicting the optimality of $\mathcal{T}_k$. □

Construction of an Initial Tree $\mathcal{T}_0$: A natural approach is then to use the tree algorithm discussed in the previous section for the construction of a $\hat{\mathcal{T}}$ satisfying $e(\hat{\mathcal{T}}) \leq \varepsilon_0$. To this end, let for $\eta > 0$, $\mathcal{T}(\eta)$ denote the smallest tree in $\mathcal{T}^*$ satisfying $e(\mathcal{T}(\eta)) \leq \eta$. From Theorem 3.2.1 we know that $\#(\mathcal{T}(\varepsilon_0/2)) \geq (\#(\hat{\mathcal{T}}))/2$.

We know that the $\mathcal{T}_k$ constructed by pruning are *optimal* within the initial tree $\mathcal{T}_0 = \hat{\mathcal{T}} = \mathcal{T}_g(\varepsilon_0)$, where $\mathcal{T}_g(\eta)$ is the smallest tree generated by the modified greedy algorithm that satisfies $e(\mathcal{T}_g(\eta)) \leq \eta$. We show next that the $\mathcal{T}_k$ remain *in essence* optimal in the whole master tree $\mathcal{T}^*$.

To that end, fix any $\eta \geq \varepsilon_0$. Then $\hat{\mathcal{T}}(\eta) := \hat{\mathcal{T}} \cap \mathcal{T}(\eta)$ satisfies

$$\eta \leq e(\hat{\mathcal{T}}(\eta)) = \sum_{T \in \mathcal{L}(\mathcal{T}(\eta)) \cap \hat{\mathcal{T}}} e(T) + \sum_{T \in \mathcal{L}(\hat{\mathcal{T}}) \cap \mathcal{T}(\eta)} e(T) \leq \eta + \varepsilon_0 \leq 2\eta. \tag{3.3.37}$$

Let $\mathcal{T}_g(2\eta)$ be the smallest tree generated by the modified greedy algorithm that satisfies $e(\mathcal{T}_g(2\eta)) \leq 2\eta$. Since $2\eta \geq \varepsilon_0$ we have $\mathcal{T}_g(2\eta) \prec \hat{\mathcal{T}}$ and, by Theorem 3.2.1, we also know that

$$\#(\mathcal{T}_g(2\eta)) \leq 2\#(\mathcal{T}(\eta)) \leq 2\#(\hat{\mathcal{T}}(\eta)) \leq 2\#(\mathcal{T}(\eta)). \tag{3.3.38}$$

Now pick $k = k(\eta)$ such that

$$\varepsilon_k \geq 2\eta > \varepsilon_{k-1}. \tag{3.3.39}$$

Then Theorem 3.3.3 and (3.3.38) yield

$$\#(\mathcal{T}_k) \leq \#(\mathcal{T}_g(2\eta)) \leq 2\#(\mathcal{T}(\eta)). \tag{3.3.40}$$

Combining (3.3.39) and (3.3.40) shows that optimal pruning provides class-optimal tree based approximations.

Proposition 3.3.1. *Assume that the initial tree $\hat{\mathcal{T}}$ used in the pruning process is generated by the modified greedy algorithm* **Mod-Greedy** *with target accuracy ε_0. Then the optimally pruned trees $\mathcal{T}_k \prec \hat{\mathcal{T}}$ are weakly optimal in $\mathcal{T}^*$. Moreover, whenever the error functional e belongs to the approximation class $\mathcal{A}^s((\Sigma_n))$, defined in Corollary 3.2.1, one has*

$$e(\mathcal{T}_k) \leq C(\#(\mathcal{T}_k))^{-s}, \quad k = 0, \dots, m, \tag{3.3.41}$$

where C depends only on e.

Proof of Theorem 3.3.2: We need some preliminaries.

Lemma 3.3.3. *The minimal μ-optimal subtrees are monotone in the following sense*

$$\mu' \geq \mu \quad \Rightarrow \quad \mathcal{T}_{\mu'}(\mathcal{T}) \preceq \mathcal{T}_\mu(\mathcal{T}). \tag{3.3.42}$$

Moreover, abbreviating $\mathcal{T}_\mu(\mathcal{T}) = \mathcal{T}_\mu$, one has

$$\mu' > \mu,\ \#(\mathcal{T}_{\mu'}) < \#(\mathcal{T}_\mu) \quad \Rightarrow \quad \mu < \frac{e(\mathcal{T}_{\mu'}) - e(\mathcal{T}_\mu)}{\#(\mathcal{L}(\mathcal{T}_\mu)) - \#(\mathcal{L}(\mathcal{T}_{\mu'}))} \leq \mu'. \tag{3.3.43}$$

Proof. When $\mathcal{T}_\mu(\mathcal{T}) = \{R(\mathcal{T})\}$ (3.3.42) is trivial $\mathcal{T}_{\mu'}(\mathcal{T})$ must be trivial as well. The general case (3.3.42) follows now from (3.3.10) by induction on $\#(\mathcal{T}_\mu(\mathcal{T}))$.

Moreover, by definition of μ-optimality, one has $E_{\mu'}(\mathcal{T}_{\mu'}) \leq E_{\mu'}(\mathcal{T}_\mu)$, $E_\mu(\mathcal{T}_\mu) \leq E_\mu(\mathcal{T}_{\mu'})$, and therefore

$$\begin{array}{lcl} e(\mathcal{T}_{\mu'}) + \mu'\#(\mathcal{L}(\mathcal{T}_{\mu'})) & \leq & e(\mathcal{T}_\mu) + \mu'\#(\mathcal{L}(\mathcal{T}_\mu)) \\ e(\mathcal{T}_\mu) + \mu\#(\mathcal{L}(\mathcal{T}_\mu)) & \leq & e(\mathcal{T}_{\mu'}) + \mu\#(\mathcal{L}(\mathcal{T}_{\mu'})). \end{array} \tag{3.3.44}$$

When $\mu' > \mu$ and $\#(\mathcal{T}_{\mu'}) < \#(\mathcal{T}_\mu)$, one has by the second inequality in (3.3.44) that $e(\mathcal{T}_\mu) + \mu\#(\mathcal{L}(\mathcal{T}_\mu)) < e(\mathcal{T}_{\mu'}) + \mu\#(\mathcal{L}(\mathcal{T}_\mu))$ which means $e(\mathcal{T}_\mu) < e(\mathcal{T}_{\mu'})$ so that the quotient in (3.3.43) is indeed nonzero and (3.3.44) readily yields (3.3.43) with $<$ replaced by $\leq$. Moreover, there must exist a $\tilde{\mu} \in (\mu, \mu')$ such that $\mathcal{T}_\mu$ is still $\tilde{\mu}$-optimal and hence $e(\mathcal{T}_\mu) + \tilde{\mu}\#(\mathcal{L}(\mathcal{T}_\mu)) \leq e(\mathcal{T}_{\mu'}) + \mu\#(\mathcal{L}(\mathcal{T}_\mu))$. Since $\tilde{\mu} > \mu$ the lower inequality in (3.3.43) must be strict. □

The next observation offers a useful description of the trees $\mathcal{T}_\mu(\mathcal{T})$.

Lemma 3.3.4. *If* $E_\mu(T) \geq E_\mu(\mathcal{B}(T; \mathcal{T}))$ *for all* $T \in \mathcal{T} \backslash \mathcal{L}(\mathcal{T})$*, then* $E_\mu(\mathcal{T}) = \min\{E_\mu(\tilde{\mathcal{T}}) : \tilde{\mathcal{T}} \preceq \mathcal{T}\}$*, i.e.,* $\mathcal{T}$ *is already a* μ*-optimally pruned subtree of itself, while the minimal* μ*-optimally pruned subtree is given by*

$$\mathcal{T}_\mu(\mathcal{T}) = \{T \in \mathcal{T} : E_\mu(\hat{T}) > E_\mu(\mathcal{B}(\hat{T}; \mathcal{T})) \textit{ for all ancestors } \hat{T} \textit{ of } T\} \cup \{R(\mathcal{T})\}. \quad (3.3.45)$$

Proof. The fact that, under the above assumptions, $E_\mu(\mathcal{T})$ is already minimal follows from the fact that there is no branch whose removal would strictly lower the cost-complexity E_μ. Of course, $\mathcal{T}$ may not be minimal yet.

To prove (3.3.45) denote the right hand side of (3.3.45) by $\tilde{\mathcal{T}}$ which indeed is a pruned subtree of $\mathcal{T}$. We show first that $\mathcal{T}_\mu(\mathcal{T}) \subseteq \tilde{\mathcal{T}}$. In fact, whenever $T \in \mathcal{T}_\mu(\mathcal{T}) \setminus \{R(\mathcal{T})\}$ any ancestor $\hat{T}$ of T must satisfy $E_\mu(\mathcal{B}(\hat{T}; \mathcal{T})) < E_\mu(\hat{T})$ since otherwise it could be cut away without increasing E_μ while reducing size. But since $T \in \mathcal{B}(\hat{T}; \mathcal{T})$ this contradicts the minimality of $\mathcal{T}_\mu(\mathcal{T})$. This shows that indeed $\mathcal{T}_\mu(\mathcal{T}) \preceq \tilde{\mathcal{T}}$. Conversely, suppose that $T \in \tilde{\mathcal{T}} \setminus \mathcal{T}_\mu(\mathcal{T})$. Then there must exist an ancestor $\hat{T}$ of T for which $E_\mu(\hat{T}) \leq E_\mu(\mathcal{B}(\hat{T}; \mathcal{T}))$ which contradicts the definition of $\tilde{\mathcal{T}}$. □

Note that it immediately follows from Lemma 3.3.4 that for any $T \in \mathcal{T}_{\mu_1} \setminus \mathcal{L}(\mathcal{T}_{\mu_1}(\mathcal{T}))$ the subtree $\mathcal{B}(T; \mathcal{T}_{\mu_1}(\mathcal{T}))$ is the minimal μ_1-optimally pruned subtree of $\mathcal{B}(T; \mathcal{T})$, i.e.,

$$\mathcal{T}_{\mu_1}(\mathcal{B}(T; \mathcal{T})) = \mathcal{B}(T; \mathcal{T}_{\mu_1}(\mathcal{T})), \quad (3.3.46)$$

which is used next to establish the following monotonicity property of $\zeta(\cdot; \cdot)$ with respect to the second argument.

Lemma 3.3.5. *For* ζ *defined by* (3.3.18) *and any* $T \in \mathcal{T}_{\mu_1}(\mathcal{T}) \setminus \mathcal{L}(\mathcal{T}_{\mu_1}(\mathcal{T}))$ *one has*

$$\begin{aligned} &\zeta(T; \mathcal{T}_{\mu_1}(\mathcal{T})) > \zeta(T; \mathcal{T}), \quad \textit{if } \mathcal{B}(T; \mathcal{T}_{\mu_1}(\mathcal{T})) \prec \mathcal{B}(T; \mathcal{T}), \\ &\zeta(T; \mathcal{T}_{\mu_1}(\mathcal{T})) = \zeta(T; \mathcal{T}), \quad \quad \quad \textit{else}. \end{aligned} \quad (3.3.47)$$

Proof. Let $T \in \mathcal{T}_{\mu_1}(\mathcal{T}) \setminus \mathcal{L}(\mathcal{T}_{\mu_1}(\mathcal{T}))$. The second relation in (3.3.5) holds by definition so that we may assume that $\mathcal{B}(T; \mathcal{T}_{\mu_1}(\mathcal{T}))$ is strictly contained in the branch $\mathcal{B}(T; \mathcal{T})$. As stated in (3.3.46), $\mathcal{B}(T; \mathcal{T}_{\mu_1}(\mathcal{T}))$ is the minimal μ_1-optimally pruned subtree of $\mathcal{B}(T; \mathcal{T})$. Now choose $\bar{\mu} > \mu_1$ large enough to ensure that $\{T\}$ is the smallest $\bar{\mu}$-optimally pruned subtree of $\mathcal{B}(T; \mathcal{T})$. Since for $\mu < \mu_1$, $\mathcal{B}(T; \mathcal{T})$ is the minimal μ-optimally pruned subtree of itself we infer from (3.3.43) in Lemma 3.3.3 that

$$\frac{e(T)-e(\mathcal{B}(T;\mathcal{T}_{\mu_1}(\mathcal{T})))}{\#(\mathcal{L}(\mathcal{B}(T;\mathcal{T}_{\mu_1}(\mathcal{T}))))-1} > \mu_1 \geq \frac{e(\mathcal{B}(T,\mathcal{T}_{\mu_1}(\mathcal{T})))-e(\mathcal{B}(T;\mathcal{T}))}{\#(\mathcal{L}(\mathcal{B}(T;\mathcal{T})))-\#(\mathcal{L}(\mathcal{B}(T,\mathcal{T}_{\mu_1}(\mathcal{T}))))}. \tag{3.3.48}$$

In fact, the left inequality corresponds to taking $\mu' = \bar{\mu}$, $\mu = \mu_1$ in Lemma 3.3.3, while the right inequality follows from the choice $\mu' = \mu_1, \mu = \mu$. Now (3.3.48) implies

$$\begin{aligned} e(T)-e(\mathcal{B}(T;\mathcal{T})) &= e(T)-e(\mathcal{B}(T;\mathcal{T}_{\mu_1}(\mathcal{T}))) + e(\mathcal{B}(T;\mathcal{T}_{\mu_1}(\mathcal{T}))) - e(\mathcal{B}(T;\mathcal{T})) \\ &< (e(T)-e(\mathcal{B}(T;\mathcal{T}_{\mu_1}(\mathcal{T})))) \\ &\quad \times \left(1+\frac{\#(\mathcal{L}(\mathcal{B}(T;\mathcal{T})))-\#(\mathcal{L}(\mathcal{B}(T,\mathcal{T}_{\mu_1}(\mathcal{T}))))}{\#(\mathcal{L}(\mathcal{B}(T;\mathcal{T}_{\mu_1}(\mathcal{T}))))-1}\right) \\ &= (e(T)-e(\mathcal{B}(T;\mathcal{T}_{\mu_1}(\mathcal{T})))) \left(\frac{\#(\mathcal{L}(\mathcal{B}(T;\mathcal{T}))-1}{\#(\mathcal{L}(\mathcal{B}(T;\mathcal{T}_{\mu_1}(\mathcal{T}))))-1}\right) \end{aligned} \tag{3.3.49}$$

which provides (3.3.47). □

Proof of Theorem 3.3.2: The validity of (3.3.25) for $k = 1$ has been already stated in Remark 3.3.2 for $\mathcal{T} = \mathcal{T}_0$. Suppose (3.3.25) holds for some $k \geq 1$. It follows from Lemma 3.3.5 that $\mu_{k+1} > \mu_k$. Furthermore, the right hand side of (3.3.23) in addition to T_0 consists of precisely those $T \in \mathcal{T}_k$ whose ancestors $\hat{T}$ satisfy $\zeta(\hat{T};\mathcal{T}_k) > \mu_{k+1}$ which is (3.3.25). Applying Remark 3.3.2 for $\mathcal{T} = \mathcal{T}_k$, yields $\mathcal{T}_{k+1} = \mathcal{T}_{\mu_{k+1}}$ which is the first relation in (3.3.26). The second relation follows from (3.3.42). The remaining claim is again a consequence of Lemma 3.3.4 and (3.3.19) which finishes the proof. □

4 Bases and Dictionaries

4.1 Preview and Motivation

We consider next another framework for nonlinear or adaptive approximation. The previous examples used approximations based on *partitions and localization*. In this section we consider approximations based on **representations**.

Idea: Representation of real numbers: Fix $b \in \mathbb{N}$, then for each $x \in \mathbb{R}$ there exist integers $d_j \in \{0, ..., b-1\}$ – digits – such that

$$x = \pm \sum_{j=-L}^{\infty} d_j b^{-j} .$$

So a real number x can be identified with a generally infinite sequence of integers $\mathbf{d} = (\mathbf{d_j})_{j=-L}^{\infty}$. If x is needed within some accuracy tolerance, the sequence is truncated and replaced by a finite one.

Question: Can we do this with functions as well? what are suitable bases for functions?

Candidates are the trigononemtric system, or series of orthogonal polynomials (Legendre, Tchebychev, Laguerre, etc.). However, these systems are very restrictive in the following sense. In several dimensions they work via tensor-products only on product domains (which for certain cases is important). Moreover, the expansion in these systems do not provide any local spatial information on the represented function.

An alternative general approach to function representations which is very flexibel looks as follows. Suppose $\mathbb{X}$ is a Banach space and suppose that $(P_j)_{j\in\mathbb{N}_0}$ is a sequence of (for the moment linear) mappings of $\mathbb{X}$ onto some dense hierarchy of nested linear spaces

$$V_0 \subset V_1 \subset \cdots \subset V_n \subset \cdots \mathbb{X}, \quad \overline{\bigcup_{j\in\mathbb{N}_0} V_j}^{\|\cdot\|_{\mathbb{X}}} = \mathbb{X}, \tag{4.1.1}$$

such that the $P_j f$ converge to f in some sense. Then, *formally* we have

$$f = P_0 f + \sum_{J=1}^{\infty} (P_j - P_{j-1}) f. \tag{4.1.2}$$

The terms $(P_j - P_{j-1})f$ represent "detail information" added when going from a coarser resolution V_{j-1} to the next higher resolution V_j. These details are contained in the spaces

$$W_{j-1} := \text{range}(P_j - P_{j-1}).$$

If we had a *basis* Ψ^{j-1} for each W_{j-1} and a basis Φ_0 for the coarsest space V_0, then the collection

$$\Psi := \Phi_0 \bigcup_{j=0}^{\infty} \Psi^j \tag{4.1.3}$$

is a *candidate* for a basis for all of $\mathbb{X}$.

Since we are working in an infinite dimensional space the problem is to give the above ingredients a precise meaning regarding the convergence of the above expansions. This requires some discussion of bases in Banach spaces in general.

The results obtained in this section are needed for understanding applications to image compression/encoding and adaptive methods for operator equations, especially in high dimensions.

4.2 Bases in Banach Spaces

Recall that a **Banach space** $\mathbb{X}$ is a normed linear space which is complete, i.e., Cauchy sequences have a limit in $\mathbb{X}$.

Example 4.2.1. *1. Λ a countable index set, $\ell_p(\Lambda)$: p-summable sequences:*

$$\|\mathbf{d}\|_{l_p} = \begin{cases} \left(\sum_{\lambda\in\Lambda} |d_\lambda|^p\right)^{\frac{1}{p}}, & 1 \le p < \infty \\ \sup_{\lambda\in\Lambda} |d_\lambda|, & p = \infty, \end{cases} \qquad \textit{where } \mathbf{d} = (d_\lambda)_{\lambda\in\Lambda}$$

2. p-integrable functions:

$$L_p(\Omega) = \{f \textit{ measurable} : \|f\|_{L_p(\Omega)} := \left(\int_\Omega |f(x)|^p \; dx\right)^{\frac{1}{p}} < \infty\},\ 1 \le p \le \infty;$$

or

$$L_p(\Omega, \mu) = \{f \textit{ measurable} : \|f\|_{L_p(\Omega,\mu)} := \left(\int_\Omega |f(x)|^p \; d\mu\right)^{\frac{1}{p}} < \infty\},\ 1 \le p \le \infty;$$

(for $p < 1$ these are only quasi-Banach spaces);

3. Continous functions $C(\Omega), \|\cdot\|_{L_\infty(\Omega)}$.

$\mathbb{X}$ is called *separable* if there exists a dense countable subset, that is every $f \in \mathbb{X}$ can be approximated arbitrarily well by linear combinations from this subset.

Bases: The notion basis is well understood in finite dimensional spaces.
A collection $\Psi = \{\psi_\lambda : \lambda \in \Lambda\}$, Λ countable, is a **basis** if every $f \in \mathbb{X}$ has a unique expansion

$$f = \sum_{\lambda\in\Lambda} d_\lambda(f)\psi_\lambda. \tag{4.2.1}$$

The sequence of "digits" $(d_\lambda)_{\lambda\in\Lambda} = \mathbf{d}(f)$ completely determines f. The specification of the sense in which the expansion converges gives rise to different notions of bases.

Definition 4.2.1. *Ψ is called a* **Schauder basis** *of $\mathbb{X}$ if for some ordering*

$$\Lambda = \{\lambda_k, k \in \mathbb{N}\}$$

every $f \in \mathbb{X}$ has a unique coordinate sequence

$$\mathbf{d}(f) = (d_{\lambda_k}(f))_{k\in\mathbb{N}}$$

such that

$$\left\| \sum_{k=1}^{n} d_{\lambda_k}(f)\psi_{\lambda_k} - f \right\|_{\mathbb{X}} \overset{n\to\infty}{\longrightarrow} 0.$$

For practical purposes a somewhat stronger notion of basis is important.

Definition 4.2.2. *A Schauder basis Ψ is called an* **unconditional basis**, *if there exists a constant $C < \infty$, such that if for any $\Gamma \subset \Lambda, \#\Gamma < \infty$ and any $d_\lambda, c_\lambda, \lambda \in \Gamma$ one has $|c_\lambda| \le |d_\lambda|, \lambda \in \Gamma$, then*

$$\left\| \sum_{\lambda\in\Gamma} c_\lambda\psi_\lambda \right\|_{\mathbb{X}} \le C \left\| \sum_{\lambda\in\Gamma} d_\lambda\psi_\lambda \right\|_{\mathbb{X}} . \tag{4.2.2}$$

This has important implications of practical relevance:

1. $\mathbb{X}$-norm is stable under damping coefficients (see thresholding in image processing).
2. The convergence of partial sums becomes independent of the ordering Λ, that is for any $(\Gamma_k)_{k\in\mathbb{N}}, \Gamma_k \subset \Lambda, \#\Gamma_k < \infty, \bigcup_{k\in\mathbb{N}} \Gamma_k = \Lambda$, for
$$P_\Gamma f = \sum_{\lambda\in\Gamma} d_\lambda(f)\psi_\lambda$$
one has $\|P_{\Gamma_k} f - f\|_{\mathbb{X}} \overset{k\to\infty}{\longrightarrow} 0$ and
$$\sup_{\Gamma\subset\Lambda, \#\Gamma<\infty} \|P_\Gamma\|_{\mathcal{L}(\mathbb{X},\mathbb{X})} < \infty . \tag{4.2.3}$$

Remark 4.2.1. *$L_1(\mathbb{R}^d)$ has no unconditional basis, but $L_p(\Omega), 1 < p < \infty$ all do. Wavelet bases are unconditional in that range.*

Remark 4.2.2. *One often finds an equivalent definition of unconditional basis: $\exists$ a constant C, such that for every $\Gamma \subset \Lambda, \#\Gamma < \infty$, any $d_\lambda, \lambda \in \Gamma$, every $\epsilon_\lambda \in \{\pm 1\}$*

$$\left\| \sum_{\lambda\in\Gamma} \epsilon_\lambda d_\lambda\psi_\lambda \right\|_{\mathbb{X}} \le C \left\| \sum_{\lambda\in\Gamma} d_\lambda\psi_\lambda \right\|_{\mathbb{X}} . \tag{4.2.4}$$

Hilber Spaces: A complete normed linear space $\mathcal{H}$ is called **Hilbert space** if $\|f\|_{\mathcal{H}} = \langle f, f\rangle_{\mathcal{H}}^{\frac{1}{2}}$ where $\langle \cdot, \cdot\rangle_{\mathcal{H}}$ is the scalar product on $\mathcal{H}$.

Examples for Hilbert spaces $\mathcal{H}$:

1. $\mathcal{H} = \mathbb{R}^d, \quad \|x\|_{\ell_2}^2 = x^T x = \sum_{j=1}^d x_j^2$
   $\mathcal{H} = \mathbb{C}^d, \quad \|x\|_{\ell_2}^2 = x^* x = \sum_{j=1}^d |x_j|^2$
2. $\mathcal{H} = \ell_2(\Lambda), \quad \|\mathbf{d}\|_{\ell_2(\Lambda)} = \left(\sum_{\lambda\in\Lambda} |d_\lambda|^2\right)^{\frac{1}{2}}$
   $\langle \mathbf{d}, \mathbf{g}\rangle = \sum_{\lambda\in\Lambda} d_\lambda \overline{g}_\lambda$
3. $L_2(\Omega), \quad \|f\|_{L_2(\Omega)} = \left(\int_\Omega |f(x)|^2 \, dx\right)^{\frac{1}{2}}$
   $\langle f, g\rangle = \int_\Omega f(x)\overline{g(x)}\, dx$
4. $H^s(\Omega)$ Sobolev spaces

When $\mathbb{X} = \mathcal{H}$ is a Hilbert space an important class of unconditional bases are **Riesz bases**.

Definition 4.2.3. *A collection $\Psi \subset \mathcal{H}$ is a* **Riesz basis** *if it is dense (finite linear combinations are dense in $\mathcal{H}$) and if $\exists\, 0 < c_\Psi, C_\Psi < \infty$ such that for any $\Gamma \subset \Lambda, \#\Gamma < \infty$, any $d_\lambda, \lambda \in \Gamma$ one has*

$$c_\Psi \left\|(d_\lambda)\right\|_{\ell_2(\Gamma)} \le \left\|\sum_{\lambda\in\Gamma} d_\lambda \psi_\lambda\right\|_{\mathcal{H}} \le C_\Psi \left\|(d_\lambda)\right\|_{\ell_2(\Gamma)} . \tag{4.2.5}$$

Note that (4.2.5) implies (4.2.2), (4.2.4), for $|c_\lambda| \le |d_\lambda|$ we have

$$\left\|\sum_{\lambda\in\Gamma} c_\lambda \psi_\lambda\right\|_{\mathcal{H}} \overset{(4.2.5)}{\le} C_\Psi \left\|(c_\lambda)\right\|_{\ell_2(\Gamma)} \le C_\Psi \left\|(d_\lambda)\right\|_{\ell_2(\Gamma)} \overset{(4.2.5)}{\le} \frac{C_\Psi}{c_\Psi} \left\|\sum_{\lambda\in\Gamma} d_\lambda \psi_\lambda\right\|_{\mathcal{H}} .$$

Remark 4.2.3. *If Ψ is a Riesz basis the mapping*

$$F : \mathbf{d} \in \ell_2(\Lambda) \mapsto \sum_{\lambda\in\Lambda} d_\lambda \psi_\lambda$$

is an isomorphism from $\ell_2(\Lambda)$ onto $\mathcal{H}$ and

$$\|F\|_{\mathcal{L}(\ell_2,\mathcal{H})} \le C_\Psi , \quad \left\|F^{-1}\right\|_{\mathcal{L}(\mathcal{H},\ell_2)} \le \frac{1}{c_\Psi} . \tag{4.2.6}$$

Proof. We proceed in several steps:

Steop 1: F is well defined and bounded. Obviously, the restriction of F to finitely supported sequences is linear. Now pick any $\mathbf{d} \in \ell_2(\Lambda)$, increasing sequence of finite index sets $(\Lambda_k)_{k\in\mathbb{N}}, \Lambda_k \subset \Lambda_{k+1}, \#\Lambda_k < \infty$, $\Lambda = \bigcup_{k\in\mathbb{N}} \Lambda_k$. Define $f_k := \sum_{\lambda\in\Lambda_k} d_\lambda\psi_\lambda \in \mathcal{H}$. Our claim is that $(f_k)_{k\in\mathbb{N}}$ is a Cauchy sequence. In fact

$$\begin{aligned}\|f_{k+m} - f_k\|_{\mathcal{H}} = \Big\|\sum_{\lambda\in\Lambda_{k+m}\setminus\Lambda_k} d_\lambda\psi_\lambda\Big\|_{\mathcal{H}} &\overset{(4.2.5)}{\leq} C_\Psi \left\|(d_\lambda)_{\lambda\in\Lambda_{k+m}\setminus\Lambda_k}\right\|_{\ell_2}\\ &\leq C_\Psi \left\|(d_\lambda)_{\lambda\in\Lambda\setminus\Lambda_k}\right\|_{\ell_2} \overset{k\to\infty}{\longrightarrow} 0.\end{aligned}$$

Hence there exists a limit $f = \sum_{\lambda\in\Lambda} d_\lambda\psi_\lambda \in \mathcal{H}$ and the first relation in (4.2.6) holds:

$$\begin{aligned}\Big\|\underbrace{\sum_{\lambda\in\Lambda} d_\lambda\psi_\lambda}_{=f}\Big\|_{\mathcal{H}} = \Big\|\underbrace{\sum_{\lambda\in\Lambda_k} d_\lambda\psi_\lambda}_{=f_k} + f - f_k\Big\|_{\mathcal{H}} &\leq \Big\|\sum_{\lambda\in\Lambda_k} d_\lambda\psi_\lambda\Big\|_{\mathcal{H}} + \underbrace{\|f - f_k\|_{\mathcal{H}}}_{\longrightarrow 0, k\to\infty}\\ &\leq C_\Psi \|(d_\lambda)\|_{\ell_2(\Lambda)} + \|f - f_k\|_{\mathcal{H}}\ .\end{aligned}$$

Letting k tend to infinity shows independence of the particular Cauchy sequence so that F is well defined and bounded.

Step 2: F is bijective: By denseness of Ψ, every $f \in \mathcal{H}$ must have such an expansion so that F is surjective. To verify injectivity, abbreviate $\mathbf{d}_\Gamma := (\mathbf{d}_\lambda)_{\lambda\in\Gamma}$

$$\begin{aligned}\|\mathbf{d}\|_{\ell_2} = \|\mathbf{d}_{\Lambda_k} + (\mathbf{d} - \mathbf{d}_{\Lambda_k})\|_{\ell_2} &\leq \|\mathbf{d}_{\Lambda_k}\|_{\ell_2} + \|\mathbf{d} - \mathbf{d}_{\Lambda_k}\|_{\ell_2}\\ &\leq \frac{1}{c_\Psi}\Big\|\sum_{\lambda\in\Lambda_k} d_\lambda\psi_\lambda\Big\|_{\mathcal{H}} + \underbrace{\|\mathbf{d} - \mathbf{d}_{\Lambda_k}\|_{l_2}}_{\to 0,\, k\to\infty}.\end{aligned}$$

Letting k go to ∞, yields

$$c_\Psi \|\mathbf{d}\|_{\ell_2} \leq \Big\|\sum_{\lambda\in\Lambda} d_\lambda\psi_\lambda\Big\|_{\mathcal{H}} = \|\mathsf{F}(\mathbf{d})\|_{\mathcal{H}}\ .$$

This shows injectivity and implies

$$\left\|\mathsf{F}^{-1}f\right\|_{\ell_2} \leq \frac{1}{c_\Psi}\|f\|_{\mathcal{H}}\ .$$

thus verifying (4.2.6). □

Remark 4.2.4. *For Ψ Riesz basis, every $f \in \mathcal{H}$ has a unique expansion*

$$f = \sum_{\lambda \in \Lambda} d_\lambda(f)\psi_\lambda$$

such that

$$c_\Psi \left\|(d_\lambda(f))\right\|_{\ell_2} \leq \|f\|_{\mathcal{H}} \leq C_\Psi \left\|(d_\lambda(f))\right\|_{\ell_2} . \tag{4.2.7}$$

Remark 4.2.5. *The relevance of Riesz basis is that small changes in f correspond to small changes in the "coordinates" of f, the coeffcicient sequence $(\mathbf{f}) = \left(\langle \mathbf{f}, \tilde{\psi}_\lambda\rangle_{\mathcal{H}}\right)_{\lambda \in \Lambda}$. How tightly these perturbations depend on each other depends on the ratio C_Ψ / c_Ψ of the* Riesz-constants*. That quotient is also called* **condition** *of the Riesz basis.*

Dual Riesz bases:

Proposition 4.2.1. *If Ψ is a Riesz basis for the Hilbert space $\mathcal{H}$ there exist $\tilde{\psi}_\lambda \in \mathcal{H}, \lambda \in \Lambda$ such that $d_\lambda(f) = \langle f, \tilde{\psi}_\lambda\rangle_{\mathcal{H}}$ and*

$$\langle \psi_\lambda, \tilde{\psi}_\nu\rangle_{\mathcal{H}} = \delta_{\lambda,\nu}, \quad \lambda, \nu \in \Lambda. \tag{4.2.8}$$

Moreover, $\tilde{\Psi} = \{\tilde{\psi}_\lambda, \lambda \in \Lambda\}$ is also a Riesz basis for $\mathcal{H}$.

Proof. The adjoint F^* of the mapping $F : \ell_2(\Lambda) \to \mathcal{H}$, given by $F(\mathbf{d}) = \sum_{\lambda \in \Lambda} \mathbf{d}_\lambda \psi_\lambda$, is

$$F^* : \mathcal{H}' = \mathcal{H} \to \ell_2(\Lambda)' = \ell_2(\Lambda), \quad F^*(f) = \left(\langle \psi_\lambda, f\rangle_{\mathcal{H}}\right)_{\lambda \in \Lambda} \tag{4.2.9}$$

because

$$\langle F(\mathbf{d}), \mathbf{f}\rangle_{\mathcal{H}} = \left\langle \sum_{\lambda \in \Lambda} \mathbf{d}_\lambda \psi_\lambda, \mathbf{f} \right\rangle_{\mathcal{H}} = \sum_{\lambda \in \Lambda} \mathbf{d}_\lambda \langle \psi_\lambda, \mathbf{f}\rangle_{\mathcal{H}} = \left\langle \mathbf{d}, \left(\langle \psi_\lambda, \mathbf{f}\rangle_{\mathcal{H}}\right)_{\lambda \in \Lambda} \right\rangle_{\ell_2} = \langle \mathbf{d}, \mathbf{F}^*(\mathbf{f})\rangle_{\ell_2}.$$

It is well-known that the adjoint of a linear mapping is boundedly invertible if and only if the mapping is and the norms are equal. Since $(F^*)^{-1} = (F^{-1})^*$ this means

$$\|F^*\|_{\mathcal{L}(\ell_2(\Lambda),\mathcal{H})} = \|F\|_{\mathcal{L}(\mathcal{H},\ell_2(\Lambda))}, \quad \|(F^*)^{-1}\|_{\mathcal{L}(\mathcal{H},\ell_2(\Lambda))} = \|F^{-1}\|_{\mathcal{L}(\ell_2(\Lambda),\mathcal{H}))}. \tag{4.2.10}$$

Now define for $\mathbf{e}_\lambda := (\delta_{\lambda,\nu})_{\nu \in \Lambda}$

$$\tilde{\psi}_\lambda := (F^*)^{-1}(\mathbf{e}_\lambda), \quad \lambda \in \Lambda. \tag{4.2.11}$$

Then, noting that $\psi_\lambda = F(\mathbf{e}_\lambda)$, one has

$$\langle \psi_\nu, \tilde{\psi}_\lambda\rangle_{\mathcal{H}} = \langle F(\mathbf{e}_\nu), (F^*)^{-1}(\mathbf{e}_\lambda)\rangle_{\mathcal{H}} = \langle \mathbf{e}_\nu, F^*(F^*)^{-1}\mathbf{e}_\lambda\rangle_{\ell_2} = \delta_{\nu,\lambda}, \quad \nu, \lambda \in \Lambda,$$

which is (4.2.8). The fact that $\tilde{\Psi} := \{\tilde{\psi}_\lambda : \lambda \in \Lambda\}$ is also a Riesz basis (with the same condition as Ψ) follows from (4.2.10) **(Exercise)**. □

We have used above that, by the *Riesz-Representation Theorem*, every bounded linear functional on a Hilbert space has a *representer* as an element in the Hilbert space itself. One and the same functional can have different representations. In practical situations when the Hilbert space is different from $L_2(\Omega)$, for instance, when $\mathcal{H} = H_0^1(\Omega)$ (see § 2.3.1), linear functionals are usually not represented by elements in $H_0^1(\Omega)$. Therefore, it is reasonable to introduce the space of bounded linear functionals of a Hilbert space $\mathcal{H}$ as a new object. Recall that in general, for any Banach space $\mathbb{X}$, the space of bounded linear functionals $\mathcal{L}(\mathbb{X}, \mathbb{R})$ is usually denoted by $\mathbb{X}'$ which becomes also a Banach space under the norm

$$\|z\|_{\mathbb{X}'} := \sup_{v \in \mathbb{X} \setminus \{0\}} \frac{z(v)}{\|v\|_{\mathbb{X}}}, \tag{4.2.12}$$

(compare with the definition of an operator norm). The action of $z \in \mathbb{X}'$ on $\mathbb{X}$ is often denoted by $z(v)$ but also by $\langle v, z \rangle = \langle v, z \rangle_{\mathbb{X}, \mathbb{X}'}$. Here, $\langle \cdot, \cdot \rangle$ does **not** denote an inner product but a *dual pairing* in the above sense. The latter notation especially in the context of Hilbert spaces stems from the fact that in many cases the linear functional has a "natural" representation in terms of another "pivot" space "between" $\mathcal{H}$ and $\mathcal{H}'$. The well-known classical example is $\mathcal{H} = H_0^1(\Omega)$ whose dual is denoted by $H^{-1}(\Omega) := (H_0^1(\Omega))'$. The pivot space is now $L_2(\Omega)$ with

$$H_0^1(\Omega) \subset L_2(\Omega) \subset H^{-1}(\Omega), \tag{4.2.13}$$

in the sense of compact embeddings. In fact, as shown in Numerical Analysis IV, any *function* $f \in L_2(\Omega)$ induces a *functional* $\ell \in H^{-1}(\Omega) := (H_0^1(\Omega))'$ defined by

$$\ell_f(v) := \int_\Omega v(x) f(x) dx, \quad v \in H_0^1(\Omega).$$

Exercise 4.2.1. *Determine the representation of ℓ_f as an element of $H_0^1(\Omega)$ when $H_0^1(\Omega)$ is endowed with the norm $\|v\|_{H^1(\Omega)} := \left(\|v\|_{L_2(\Omega)}^2 + \|\nabla v\|_{L_2(\Omega)}^2\right)^{1/2}$. Can one replace this norm just by $\|\nabla v\|_{L_2(\Omega)}$? If so which effect does that have on the representation of ℓ_f?.*

Exercise 4.2.2. *Let $\Psi = \{\psi_\lambda : \lambda \in \Lambda\}$ be a Riesz basis for the Hilbert space $\mathcal{H}$. Then there exists a Riesz basis $\tilde{\Psi}$ for its dual $\mathcal{H}'$ satisfying*

$$\tilde{\psi}_\nu(\psi_\lambda) = \langle \psi_\lambda, \tilde{\psi}_\lambda \rangle = \delta_{\nu,\lambda}, \quad \lambda, \nu \in \Lambda,$$

and

$$(C_\Psi)^{-1} \big\| (\langle \psi_\lambda, w \rangle)_{\lambda \in \Lambda} \big\|_{\ell_2} \leq \|w\|_{\mathcal{H}'} \leq (c_\Psi)^{-1} \big\| (\langle \psi_\lambda, w \rangle)_{\lambda \in \Lambda} \big\|_{\ell_2}, \quad w \in \mathcal{H}'. \tag{4.2.14}$$

Orthonormal Bases: There is an important special case of Riesz bases, namely: $c_\Psi = C_\Psi = 1$

$$\|(d_\lambda(f))\|_{\ell_2} = \left\| \sum_{\lambda\in\Lambda} d_\lambda(f)\psi_\lambda \right\|_{\mathcal{H}} , \tag{4.2.15}$$

that is Ψ is an **orthonormal basis,** $\psi_\lambda = \tilde{\psi}_\lambda$. In fact for $\#\Gamma < \infty, \Gamma \subset \Lambda$:

$$\|\mathbf{d}\|_{\ell_2}^2 = \left\| \sum_{\lambda\in\Gamma} d_\lambda\psi_\lambda \right\|_{\mathcal{H}}^2 = \left\langle \sum_{\lambda\in\Gamma} d_\lambda\psi_\lambda, \sum_{\lambda\in\Gamma} d_\lambda\psi_\lambda \right\rangle_{\mathcal{H}} = \sum_{\lambda,\nu\in\Gamma} d_\lambda\overline{d_\nu}\langle\psi_\lambda,\psi_\nu\rangle_{\mathcal{H}} = \mathbf{d}^*M_\Gamma\mathbf{d}$$

for all $\mathbf{d}$ with $\operatorname{supp}\mathbf{d} \subseteq \Gamma$ where $M_\Gamma = (\langle\psi_\lambda,\psi_\nu\rangle_{\mathcal{H}})_{\lambda,\nu\in\Gamma}$. This yields

$$\mathbf{d}^*\mathbf{d} = \mathbf{d}^*M_\Gamma\mathbf{d} \text{ where } M_\Gamma \text{ s.p.d.}$$

Hence all eigenvalues of M_Γ must be equal to one wich implies $M_\Gamma = \mathrm{Id}$, that is

$$\langle\psi_\lambda,\psi_\nu\rangle_{\mathcal{H}} = \delta_{\lambda,\nu}, \quad \lambda,\nu \in \Lambda, \tag{4.2.16}$$

i.e., Ψ is an orthonormal basis for $\mathcal{H}$.

Thus orthonormal bases are perfectly conditioned with condition number equal to one: the coefficient norm *equals* the function norm. Moreover, an orthonormal bases is dual to itself (when representing functionals in $\mathcal{H}$).

Remark 4.2.6. *The following examples of orthonormal bases indicate that it is generally difficult to construct practicable orthonormal systems, e.g. for more general domain geometries. Orthonormal functions are typically global (e.g. because of Gram-Schmidt). Riesz bases can be viewed as a relaxed version which will later be seen to lead to localizable basis elements. The prize for this "enhanced" practicality is to accept a somewhat larger condition of the basis, which nevertheless relates perturbations of the function to perturbations of its coefficients in a uniform way.*

Example 4.2.2. *1.* $\mathcal{H} = L_2([0,1])$, $\|f\|_{L_2([0,1])} = \left(\int_0^1 |f(t)|^2\, dt\right)^{\frac{1}{2}} =: (f,f)_{[0,1]}^{\frac{1}{2}}$: *The so called* **Haar basis** *is the simplest example of a* **wavelet basis**. *It has multilevel structure:*

$$\Psi = \{\psi_{(-1,0)}, \psi_{j,k} : k = 0,..,2^j-1, j = 0,1,2,\ldots\}$$

where

$$\psi_{-1,0}(t) = \phi(t) := \chi_{[0,1)}(t) \quad \textit{scaling function},$$

and

$$\psi_{(0,0)}(t) := \psi(t) = \phi(2t) - \phi(2t-1), \quad \psi_{j,k}(t) = 2^{\frac{j}{2}}\psi(2^jt-k).$$

with

$$\Lambda = \{(-1,0),(j,k), k = 0..2^j-1, j = 0,1,2,\ldots\}.$$

2. *Trigonometric basis:* $\mathcal{H} = L_{2,2\pi} = \{f(\cdot + 2\pi k) = f$ *a.e.*, $f \in L_2([-\pi, \pi])\}$, $\langle f, g\rangle = \int_{-\pi}^{\pi} f(x)\overline{g(x)}\,dx$

$$\Psi = \{e_k : k \in \mathbb{Z}\}, e_k(t) = \frac{1}{\sqrt{2\pi}} e^{itk} \qquad (i^2 = -1).$$

The coefficients $\langle f, e_k\rangle = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} f(t)e^{-itk}\,dt = \hat{f}(k)$ *are the classical* **Fourier coefficients** *and the projection*

$$P_{2n+1}f(t) = \sum_{|k|\le n} \hat{f}(k)e^{ikt}$$

provides the Fourier partial sums.

The Haar basis:

Exercise 4.2.3. *(a) Show that* Ψ *defined above is indeed an orthonormal system and a basis for all of* $L_2((0,1))$.

(b) Show also that the collection

$$\phi_{j,k}(t) = 2^{\frac{j}{2}}\phi(2^j t - k), \quad k = 0, \dots, 2^j - 1. \tag{4.2.17}$$

is an orthonormal basis for $\mathbb{P}_1(\mathcal{P}_{2^j})$ *where* $\mathcal{P}_{2^j}$ *is the uniform partition of* $[0,1]$ *into intervals of length* 2^{-j}.

(c) Show that

$$\Psi_{J-1} := \{\phi, \psi_{j,k} : k = 0, \dots, 2^j - 1,\ 0 \le j \le J - 1\}$$

is also an orthonormal basis for $\mathbb{P}_1(\mathcal{P}_{2^j})$.

Figure 5 illustrates the decomposition of a piecewise constant into an *average* on a coarser mesh and a *fluctuation* on the fine mesh.

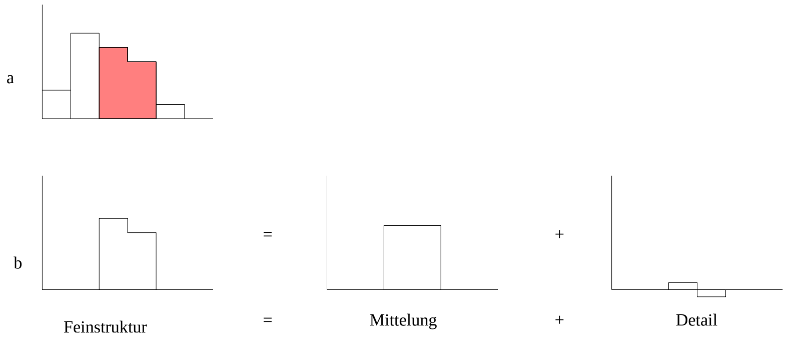


Figure 5: Decomposition of a piecewise constant

The generators of the Haar basis $\phi(t)$ and $\psi(t)$ are illustrated in Figure 6. Integration against ϕ is an average. Integration against the “oscillatory” profile ψ anihilates the constant part.

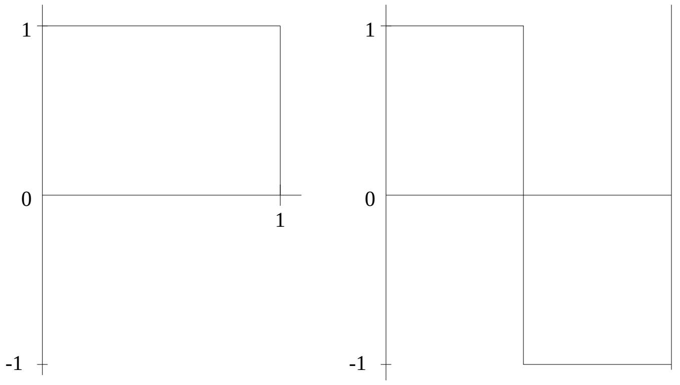


Figure 6: scaling function $\phi(t)$ "mother" wavelet $\psi(t)$

Exercise 4.2.4. *Estimate the quantity* $|\langle f, \psi_{j,k}\rangle_{[0,1]}|$ *when* $f \in W^1(L_p((k2^{-j}, (k+1)2^{-j}))$, *to see that the wavelet coefficient* $|\langle f, \psi_{j,k}\rangle_{[0,1]}|$ *is "small" when* f *is smooth on the support of* $\psi_{j,k}$. *This is heavily used in image compression.*

Consider the function

$$v_J := \sum_{k=0}^{2^J-1} p_{j,k}\phi_{j,k} \in \mathbb{P}_1(\mathcal{P}_{2^j}).$$

By Exercise 4.2.3, v_J can also be expanded in terms of the basis Ψ_{J-1}.

Exercise 4.2.5. *Derive a transformation that generates the wavelet coefficients* $d_{j,k} = d_{j,k}(v_J)$ *of* v_J. *Hint: one can proceed in a cascadic way starting from the finest level* J, *see Figure 7. and* (4.2.18)

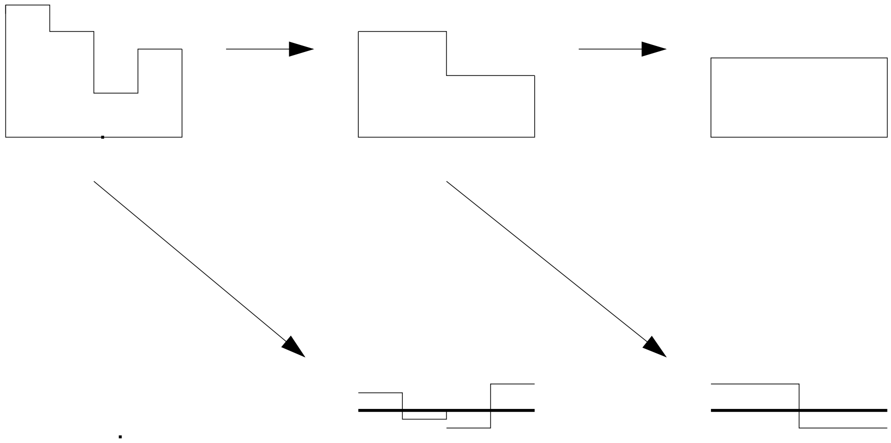


Figure 7: wavelet-transform

$$\sum_{k=0}^{2^J-1} p_{J,k}\phi_{J,k}(x) = \sum_{k=0}^{2^{J-1}-1} p_{J-1,k}\phi_{J-1,k}(x) + \sum_{k=0}^{2^{J-1}-1} d_{J-1,k}\psi_{J-1,k}(x) \qquad (4.2.18)$$

In either case, the important consequence of a Riesz basis is that approximation in $\mathcal{H}$ is equivalent to approximation in $\ell_2(\Lambda)$. We'll say more later on *wavelet-Riesz bases* need for *image compression* as well as the *adaptive solution of operator equations*.

A first example of "multiresolution-analysis": As shown above

$$P_j f := \sum_{k=0}^{2^j-1} \langle f, \phi_{j,k}\rangle_{[0,1]}\phi_{j,k} = \sum_{(l,k)\in\Lambda_{j-1}} \langle f, \psi_{l,k}\rangle_{[0,1]}\psi_{l,k} \qquad (4.2.19)$$

where

$$\Lambda_{j-1} = \{(-1,0), (l,k), k = 0..2^l - 1, l = 0,1,2,\dots,j-1\},$$

projects $L_2(\Omega)$ ($\Omega = (0,1)$) to the space

$$V_j := \mathbb{P}_1(\mathcal{P}_{2^j}) \qquad (4.2.20)$$

of piecewise constants on the uniform partition of $(0,1)$ into intervals of length 2^{-j}. Note that

$$P_J f = \sum_{j=0}^{J} (P_j - P_{j-1})f, \quad \text{where } P_{-1}f := 0. \qquad (4.2.21)$$

Since by denseness of the $\bigcup_{j\geq 0} V_j$ in $L_2(\Omega)$,

$$\lim_{j\to\infty} \|f - P_j f\|_{L_2(\Omega)} = 0, \qquad (4.2.22)$$

the *telescoping expansion*

$$f = \sum_{j=0}^{\infty} (P_j - P_{j-1})f \qquad (4.2.23)$$

converges in $L_2(\Omega)$.

Exercise 4.2.6. *Show that the "fluctuation" operators $Q_j := P_j - P_{j-1}$ are also projectors and*

$$W_{j-1} := (P_j - P_{j-1})V_j := \{(P_j - P_{j-1})f : f \in V_j\}, \qquad (4.2.24)$$

is the orthogonal complement of V_{j-1} in the refined space V_j, i.e.,

$$V_j = V_{j-1} \oplus W_{j-1}.$$

Moreover,

$$\|f\|^2_{L_2(\Omega)} = \sum_{j=0}^{\infty} \|(P_j - P_{j-1})f\|^2_{L_2(\Omega)}, \quad f \in L_2(\Omega). \qquad (4.2.25)$$

This gives the *multiscale-decomposition*

$$L_2(\Omega) = V_0 \bigoplus_{j=0}^{\infty} W_j \tag{4.2.26}$$

of $L_2(\Omega)$.

Comments 4.2.1. *The following should help interpreting the above findings.*

- (4.2.25) *indicates the importance of such telescoping expansions which, in particular, motivates their pivotal role in the next section.*
- *Note that*

  $$\Psi_j \setminus \Psi_{j-1} = \{\psi_{j-1,k} : k = 0, \dots, 2^j - 1\}$$

  is an orthonormal basis of W_j *and therefore*

  $$\|(P_j - P_{j-1})f\|^2_{L_2(\Omega)} = \sum_{k=0}^{2^{j-1}-1} \left|\langle f, \psi_{j,k}\rangle_{[0,1]}\right|^2, \tag{4.2.27}$$

  so that

  $$\|f\|^2_{L_2(\Omega)} = \sum_{(j,k)\in\Lambda} \left|\langle f, \psi_{j,k}\rangle_{[0,1]}\right|^2. \tag{4.2.28}$$

  Note that discarding small wavelet coefficients on the right hand side will cause only small quantifiable changes in the L_2*-norm of* f*. We have seen that wavelet coefficients are small when* f *is smooth on the respective support. This is the basis of image compression.*

A sketch of an application to image compression/encoding: Here is a rough sketch of the compression scheme in [19] (representing essentially the JPEG2000 standard).

- view the digital image as a *piecewise constant* function over the pixel raster (vector valued if colored image);
- for bivariate wavelets by taking tensor products of univariate wavelets (usually higher order analogs to the Haar wavelet with more vanishing moments for better compression, see Exercise 4.2.4;
- transform the piecewise constant function into its wavelet representation using the above fast transform;
- *quantize* the wavelet coefficients in a certain systematic fashion to retain a bitstream of possibly shortest length for guaranteeing an overall ℓ_2 target accuracy.

Figure 11 shows a visual comparison between an original and compressed image.

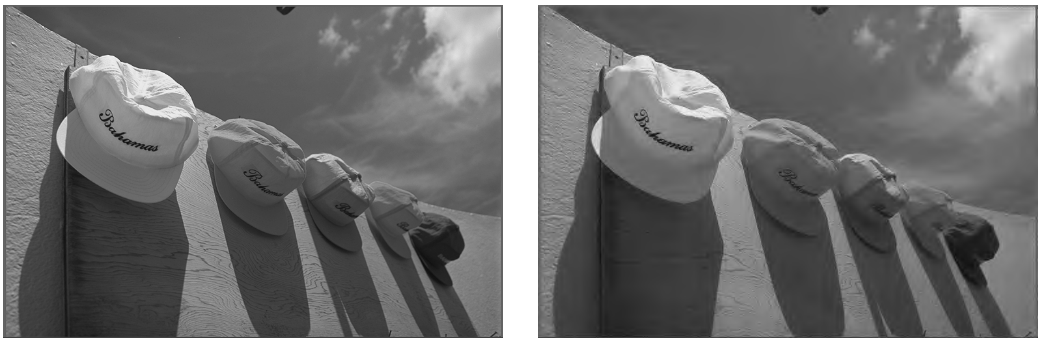
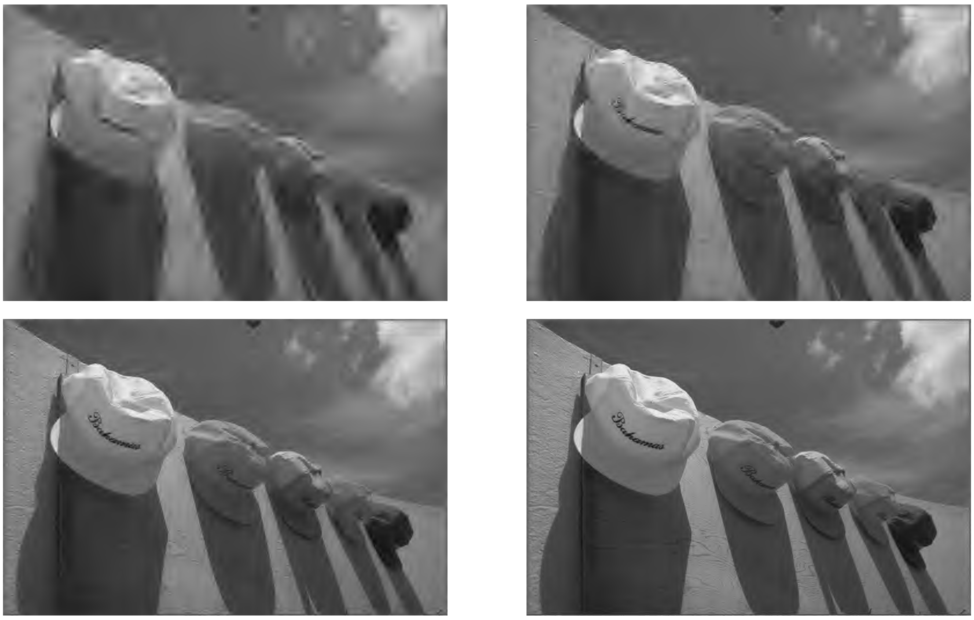


Figure 8: left: original 768 $\times$ 512 pixel 8 bit graylevel depth $\rightsquigarrow$ 384 KB naive storage allocation; right: compressed version 3.5 % of original storage

Thus, the image – a function – is decomposed into its multiscale contributions using the fact, that discarding small wavelet coefficients has only a small quantifiable effect on the image. If the image contains large smooth regions with little variation, many wavelet coefficients will be small. This is illustrated by the following sequence of compressed versions.



The decomposition of the image into portions with dyadic frequency scales is shown in Figure 9.

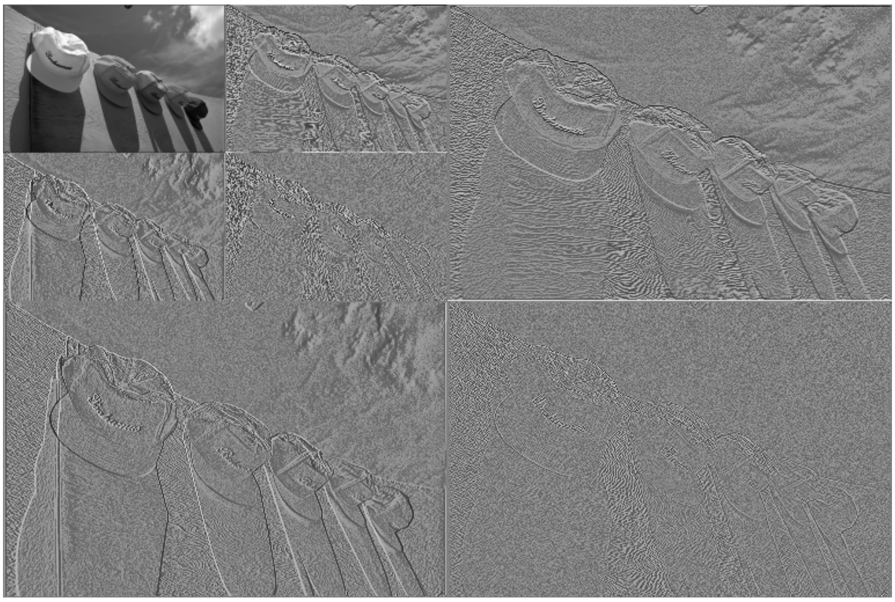


Figure 9: Multiscale decomposition of the image

Of course this compression scheme is a highly nonlinear process. The following discussions provide a basis for analyzing such schemes.

The main distinction from Fourier transforms (which provide frequency decompositions) the wavelet transform reflects decompositions into frequency ranges combined with spatial localization.

4.3 Linear versus Nonlinear Approximation

Given a basis with good stability properties the approximation of functions can be reduced to the approximation of sequences. Again, there are *linear* and *nonlinear* strategies.

Given a basis $\Psi \subset \mathbb{X}$ in a Banach space $\mathbb{X}$, consider an ordering $\Lambda = \{\lambda_k, k \in \mathbb{N}\}$. Then

$$e_n(f)_{\mathbb{X}} := \inf_{g \in \mathbb{X}_n} \|f - g\|_{\mathbb{X}}, \mathbb{X}_n = \text{span}\{\psi_{\lambda_k} : k = 1, .., n\}$$

is the error of the **best linear approximation** to f from $\mathbb{X}_n$.

Similarly let

$$\Sigma_n = \left\{ \sum_{\lambda \in \Gamma} d_\lambda \psi_\lambda : \#\Gamma \leq n, d_\lambda \in \mathbb{K} \right\} \qquad \text{(for unconditional } \Psi)$$

be the *nonlinear* set of k-term expansions and

$$\sigma_n(f)_{\mathbb{X}} := \inf_{g \in \Sigma_n} \|f - g\|_{\mathbb{X}}$$

the error of **best n-term approximation**.
Note: $\Sigma_n + \Sigma_n \subset \Sigma_{2n}$

Exercise 4.3.1. *If Ψ is unconditional then*

$$\sigma_n(f)_X \le \inf_{\#\Gamma \le n} \|f - P_\Gamma f\|_X \le C\sigma_n(f)_X \tag{4.3.1}$$

where

$$C = 1 + \sup_{\#\Gamma \le n} \|P_\Gamma\|_{X \to X}, \qquad P_\Gamma f = \sum_{\lambda \in \Gamma} d_\lambda(f)\psi_\lambda .$$

Good news: A *coordinate projection* realizes, up to a uniform constant, the error of the best n-term approximation. Such an approximation is called **near-best n-term approximation**.

Bad news: Under the above general assumptions, the search for the best $\Gamma \subset \Lambda, \#\Gamma = n$, is usually infeasible.

This latter obstruction disappears for more specific bases, for instance, for Riesz bases in a Hilbert space.

4.4 Basis Approximation in Hilbert Spaces

Let $\mathcal{H}$ be a seperable Hilbert space and $\Psi \subset \mathcal{H}$ an orthonormal basis. Then

$$P_\Gamma : f \to \sum_{\lambda \in \Gamma} \langle f, \psi_\lambda \rangle \psi_\lambda$$

is the orthogonal projection of $\mathcal{H}$ onto $\mathcal{H}_\Gamma := \operatorname{span}\{\psi_\lambda, \lambda \in \Gamma\}$ because by (4.2.16)

$$\begin{aligned}\langle f - P_\Gamma f, \psi_\nu \rangle &= \langle f, \psi_\nu \rangle - \langle P_\Gamma f, \psi_\nu \rangle \\ &= \langle f, \psi_\nu \rangle - \langle f, \psi_\nu \rangle \\ &= 0 .\end{aligned}$$

$P_\Gamma f$ realizes the unique best approximation in $\mathcal{H}_\Gamma$ because for any $g_\Gamma \in \mathcal{H}_\Gamma$ we have

$$\|f - \underbrace{g_\Gamma}_{\in \mathcal{H}_\Gamma}\|_{\mathcal{H}}^2 = \|f - \underbrace{P_\Gamma f - g_\Gamma}_{\in \mathcal{H}_\Gamma} + P_\Gamma f\|_{\mathcal{H}}^2 \overset{\text{Pythagoras}}{=} \|f - P_\Gamma f\|_{\mathcal{H}}^2 + \|P_\Gamma f - g_\Gamma\|_{\mathcal{H}}^2 \tag{4.4.1}$$

and thus

$$\|f - P_\Gamma f\|_{\mathcal{H}} = \min_{g \in \mathcal{H}} \|f - g\|_{\mathcal{H}} . \tag{4.4.2}$$

Linear approximation: For a *given* ordering $\Lambda = \{\lambda_k : k \in \mathbb{N}\}$ and associated spaces $\mathcal{H}_n := \operatorname{span}\{\psi_{\lambda_k} : 1 \leq k \leq n\}$, one has

$$e_n(f)_{\mathcal{H}} = \left(\sum_{k=n+1}^{\infty} |\langle f, \psi_{\lambda_k} \rangle|^2 \right)^{\frac{1}{2}} . \tag{4.4.3}$$

Nonlinear approximation: Since

$$\|f - P_\Gamma f\|_{\mathcal{H}} = \left(\sum_{\lambda \notin \Gamma} |\langle f, \psi_\lambda \rangle_{\mathcal{H}}|^2 \right)^{\frac{1}{2}}$$

the error is obviously minimized over all Γ, $\#\Gamma \leq n$, if Γ is comprised of the first n coefficients with *largest* absolute value. So the best set Γ^* can be read off the coefficients. Γ is not necessarily unique if several coefficients have the same absolute value. Obviously this selection depends on f and is therefore nonlinear.

This suggests to take a *greedy strategy*: Consider the **decreasing rearrangement** of coefficients.
Let Λ be ordered as $\Lambda = \{\lambda_k : k \in \mathbb{N}\}$ with

$$\underbrace{|\langle f, \psi_{\lambda_1} \rangle|}_{=d_1^*} \geq \underbrace{|\langle f, \psi_{\lambda_2} \rangle|}_{=d_2^*} \geq \ldots \geq \underbrace{|\langle f, \psi_{\lambda_k} \rangle|}_{=d_k^*} . \tag{4.4.4}$$

Let

$$\Lambda_k = \{\lambda_j\}_{j=1}^k, \qquad G_n f := \sum_{j=1}^{n} \underbrace{\langle f, \psi_{\lambda_j} \rangle}_{d_j^*} \psi_{\lambda_j} . \tag{4.4.5}$$

Then

$$\sigma_n(f)_{\mathcal{H}} = \|f - G_n f\|_{\mathcal{H}} = \left(\sum_{j=n+1}^{\infty} \left|\langle f, \psi_{\lambda_j} \rangle\right|^2 \right)^{\frac{1}{2}} = \sigma_n(\mathbf{d})_{l_2} \tag{4.4.6}$$

with

$$\mathbf{d} = \mathbf{d}(\mathbf{f}) := (\langle \mathbf{f}, \psi_\lambda \rangle)_{\lambda \in \Lambda} .$$

Exercise 4.4.1. *Assume that Ψ is a Riesz basis for $\mathcal{H}$ and let*

$$G_n f = \sum_{j=1}^{n} \langle f, \tilde{\psi}_{\lambda_j} \rangle \psi_{\lambda_j}, \qquad \left|\langle f, \tilde{\psi}_{\lambda_1} \rangle\right| \geq \left|\langle f, \tilde{\psi}_{\lambda_2} \rangle\right| \geq \ldots .$$

Then, one has

$$\sigma_n(f)_{\mathcal{H}} \leq \|f - G_n f\|_{\mathcal{H}} \leq \frac{C_\Psi}{c_\Psi} \sigma_n(f)_{\mathcal{H}}. \tag{4.4.7}$$

where

$$\frac{C_\Psi}{c_\Psi} = \textit{ condition of the Riesz basis.}$$

The operators G_n are sometimes called *greedy projectors*.

The main upshot of the above observations is that best n-term approximation in a Hilbert space is characterized by best n-term approximation in ℓ_2. To make this precise, let

$$\begin{aligned} \Sigma_n(\Psi) &:= \left\{ v = \textstyle\sum_{\lambda\in\Gamma} v_\lambda \psi_\lambda : \#\Gamma \le n \right\} \subset \mathcal{H}, \\ \Sigma_n(\Lambda) &:= \left\{ \mathbf{v} \in \ell_2(\Lambda) : \#(\operatorname{supp} \mathbf{v}) \le \mathbf{n} \right\} \subset \ell_2(\Lambda). \end{aligned} \tag{4.4.8}$$

Recall that $F : \ell_2(\Lambda) \to \mathcal{H}$ is an isomorphism if and only if Ψ is a Riesz basis.

Exercise 4.4.2. *Let Ψ be a Riesz basis for the Hilbert space $\mathcal{H}$. Then*

$$\mathbf{v} \in \mathcal{A}^r((\Sigma_\mathbf{n}(\Lambda)), \ell_2(\Lambda)) \quad \Leftrightarrow \quad \mathbf{F}(\mathbf{v}) \in \mathcal{A}^r((\Sigma_\mathbf{n}(\Psi)), \mathcal{H}), \tag{4.4.9}$$

and

$$C_\Psi^{-1} \sigma_n(f)_\mathcal{H} \le \sigma_n\big((\langle f, \tilde{\psi}_\lambda \rangle_\mathcal{H})\big)_{\ell_2(\Lambda)} \le c_\Psi^{-1} \sigma_n(f)_\mathcal{H}. \tag{4.4.10}$$

In this sense one can identify $\mathcal{A}^r((\Sigma_n(\Lambda)), \ell_2(\Lambda))$ with $\mathcal{A}^r((\Sigma_n(\Psi)), \mathcal{H})$, or more precisely $\mathcal{A}^r((\Sigma_n(\Psi), \mathcal{H}) = F\big(\mathcal{A}^r((\Sigma_n(\Lambda)), \ell_2(\Lambda))\big)$.

The last fact motivates studying approximation in sequence spaces.

4.5 Basics About Approximation in Sequence Space

Whenever one works in a Hilbert space, linear and nonlinear approximation happens in sequence spaces. In this section we discuss basic concepts in sequence spaces like best n-term approximation or thresholding. These are prerequisites for adaptive solvers to be discussed later. As we have learnt before, when dealing with adaptive approximation errors are better estimated by smoothness in different metrics which requires dealing with more general sequence spaces ℓ_p rather than only ℓ_2.

Recall that for a countable index set Λ

$$\ell_p(\Lambda) = \left\{ \mathbf{d} : \|\mathbf{d}\|_{\ell_p(\Lambda)} = \left\{ \begin{array}{ll} \left(\sum\limits_{\lambda\in\Lambda} |d_\lambda|^p \right)^{\frac{1}{p}} & , \ 0 < p < \infty \\ \sup\limits_{\lambda\in\Lambda} |d_\lambda| & , \ p = \infty \end{array} \right\} < \infty \right\}.$$

4.5.1 Linear Approximation

In the case of **linear approximation** we have

$$\begin{aligned} \Lambda &= \{\lambda_k : k \in \mathbb{N}\}, \qquad \mathbb{X}_n = \operatorname{span}\{\mathbf{d} : \operatorname{supp}\mathbf{d} \subseteq \Lambda_n\} \\ \Lambda_n &= \{\lambda_k, k = 1, .., n\} \\ e_n(\mathbf{d})_{\ell_p} &= \inf_{g \in X_n} \|\mathbf{d} - \mathbf{g}\|_{\ell_p} \end{aligned}$$

and the approximation spaces

$$\mathcal{A}^r_\infty\left((\mathbb{X}_n), l_p(\Lambda)\right) = \left\{\mathbf{d} \in \ell_p(\Lambda) : \sup_{n \geq 1} n^r e_n(\mathbf{d})_{\ell_p} =: |\mathbf{d}|_{\mathcal{A}^r} < \infty\right\},$$

or more generally

$$\mathcal{A}^r_q\left((\mathbb{X}_n), l_p(\Lambda)\right) = \left\{\mathbf{d} \in \ell_p(\Lambda) : |\mathbf{d}|_{\mathcal{A}^r_q} = \left(\sum_{n=1}^{\infty} \left(n^r e_n(\mathbf{d})_{\ell_p}\right)^q \frac{1}{n}\right)^{\frac{1}{q}} < \infty\right\}, \; 0 < q < \infty \, .$$

Exercise 4.5.1. *Show that*

$$\begin{aligned} |f|^q_{\mathcal{A}^r_q} = \sum_{n>0} (n^r \sigma_n(f)_X)^q \frac{1}{n} &\sim \sum_{j=0}^{\infty} \left(2^{rj} \sigma_{2^j}(f)_X\right)^q \\ \sum_{n>0} (n^r e_n(f)_X)^q \frac{1}{n} &\sim \sum_{j=0}^{\infty} \left(2^{rj} e_{2^j}(f)_X\right)^q \, . \end{aligned} \tag{4.5.1}$$

Goal: We wish to characterize such approximation spaces in terms of intrinsic sequence properties. We begin with the following general facts (see also Exercise 4.2.6 and Comments 4.2.1).

Theorem 4.5.1. *Let $\mathbb{X}$ be a quasi-Banach space (like $\ell_p(\Lambda), 0 < p \leq \infty$), let the sets Σ_n be* **admissible**, *that is*

$$\begin{aligned} (i) &\;\; 0 \in \Sigma_n \\ (ii) &\;\; \Sigma_n \subset \Sigma_{n+1} \\ (iii) &\;\; c\Sigma_n = \Sigma_n \\ (iv) &\;\; \Sigma_n + \Sigma_n \subset \Sigma_{an} \textit{ for some fixed } a \in \mathbb{N} \end{aligned} \tag{4.5.2}$$

(all satisfied for nested linear spaces).
Assume: $G_n : \mathbb{X} \to \Sigma_n$ *gives a* **near-best** *approximation with some constant* C_0, *that is*

$$\|f - G_n f\|_{\mathbb{X}} \leq C_0 \sigma_n(f)_{\mathbb{X}}, \qquad f \in \mathbb{X}, n \geq 1 \, . \tag{4.5.3}$$

Then, setting $g_j = g_j(f) = G_{2^j} f \in \Sigma_{2^j}$ *we have*

$$|f|_{\mathcal{A}^r_q((\Sigma_n),X)} \sim \left\| \left(2^{rj} \|f - g_j\|_X\right)_{j\in\mathbb{N}_0} \right\|_{\ell_q} \sim \left\| \left(2^{rj} \|g_{j+1} - g_j\|_{\mathbb{X}}\right)_{j\in\mathbb{N}_0} \right\|_{\ell_q} \tag{4.5.4}$$

and

$$\|f\|_{\mathcal{A}^r_q} \sim \|g_0\|_{\mathbb{X}} + \left\| \left(2^{rj} \|f - g_j\|_{\mathbb{X}}\right)_{j\in\mathbb{N}_0} \right\|_{\ell_q} \sim \|g_0\|_{\mathbb{X}} + \left\| \left(2^{rj} \|g_{j+1} - g_j\|_X\right)_{j\in\mathbb{N}_0} \right\|_{\ell_q} \tag{4.5.5}$$

with constants $C(r, q, \mathbb{X}, C_0)$.

Note that the main non-trivial step lies in the transition from the terms $\|f - g_j\|_{\mathbb{X}}$ to the terms $\|g_{j+1} - g_j\|_{\mathbb{X}}$. These latter terms provide the key for later characterizations of function spaces by wavelet coefficients and is therefore quite essential (although looking innocent).

The main tools for handling this transition are

- telescoping expansions;
- discrete Hardy inequalities.

Discrete Hardy inequalities: :

Theorem 4.5.2. First Hardy's inequality:
For any fixed number m, *let* $b_j := 2^{-mj}\sum_{l\le j} 2^{ml} a_l$. *Then, for all* $0 < q \le \infty$ *and* $s < m$, *we have*

$$\|(2^{sj} b_j)_{j\in\mathbb{Z}}\|_{\ell_q} \le C \|(2^{sj} a_j)_{j\in\mathbb{Z}}\|_{\ell_q}, \tag{4.5.6}$$

where C *depends on* s, q, *and* m. *The same result holds with* $b_j := 2^{-mj}\left(\sum_{l\le j} 2^{mpl} a_l^p\right)^{1/p}$ *for any* $p > 0$, *where* C *depends on* s, p, q, *and* m.

Second Hardy's inequality:
Let $(a_j)_{j\in\mathbb{N}}$ *be any sequence of positive numbers. Let* $b_j := \left(\sum_{l\ge j} a_l^p\right)^{\frac{1}{p}}$ *for any fixed* $0 < p < \infty$. *Then for any* $q > 0$, $s > 0$, *one has*

$$\left\|(2^{sj} b_j)_j\right\|_{\ell_q} \le C \left\|(2^{sj} a_j)_j\right\|_{\ell_q} \tag{4.5.7}$$

where $C = C(s, q, p)$.

Proof. For both inequalities, we only treat the case $p = 1$, since the generalization to $p \neq 1$ follows by applying the result established for $p = 1$ with a_j replaced by a_j^p and (q, s, m) replaced by $(q/p, sp, mp)$. We also only have to consider the case where the norm M on the right side of these inequalities is finite.

First Hardy's inequality:
When $q=\infty$, we have $a_l \leq M2^{-sl}$, and therefore

$$b_j \leq M2^{-mj}\sum_{l\leq j} 2^{(m-s)l} \leq \frac{M}{1-2^{s-m}}2^{-sj}, \tag{4.5.8}$$

which proves the result with $C=(1-2^{s-m})^{-1}$. When $1<q<\infty$, we define $t := \frac{m-s}{2}$. Using the conjugate exponent q' such that $\frac{1}{q}+\frac{1}{q'}=1$, we can write

$$\begin{aligned}
\sum_{j\in\mathbb{Z}}(2^{sj}b_j)^q &= \sum_{j\in\mathbb{Z}} 2^{(s-m)qj}\Big(\sum_{l\leq j} 2^{ml}a_l\Big)^q \\
&\leq \sum_{j\in\mathbb{Z}} 2^{(s-m)qj}\Big(\sum_{l\leq j} 2^{tq'l}\Big)^{q/q'}\Big(\sum_{l\leq j} 2^{(m-t)ql}a_l^q\Big) \\
&\lesssim \sum_{j\in\mathbb{Z}} 2^{(s-m+t)qj}\Big(\sum_{l\leq j} 2^{(m-t)ql}a_l^q\Big) \\
&= \sum_{l\in\mathbb{Z}} a_l^q 2^{(m-t)ql}\Big(\sum_{j\geq l} 2^{(s-m+t)qj}\Big) \\
&\lesssim \sum_{l\in\mathbb{Z}} 2^{sql}a_l^q,
\end{aligned}$$

where the multiplicative constants depend on s, q and m. When $q\leq 1$, we simply write

$$\begin{aligned}
\sum_{j\in\mathbb{Z}}(2^{sj}b_j)^q &= \sum_{j\in\mathbb{Z}} 2^{(s-m)qj}\Big(\sum_{l\leq j} 2^{ml}a_l\Big)^q \\
&\leq \sum_{j\in\mathbb{Z}} 2^{(s-m)qj}\sum_{l\leq j} 2^{mql}a_l^q \\
&= \sum_{l\in\mathbb{Z}} a_l^q 2^{mql}\Big(\sum_{j\geq l} 2^{(s-m)qj}\Big) \\
&\lesssim \sum_{l\in\mathbb{Z}} 2^{sql}a_l^q,
\end{aligned}$$

where the multiplicative constants depend on s, q and m.
Second Hardy's inequality: Consider first
$\underline{q=\infty}$: Let $k := \sup_j 2^{sj}a_j = \left\|(2^{sj}a_j)_j\right\|_{\ell_\infty} < \infty$. Then

$$\begin{aligned}
& a_j \leq 2^{-sj}k \\
\Rightarrow\quad & b_j = \sum_{l=j}^{\infty} a_l \leq k\sum_{l=j}^{\infty} 2^{-sl} = k(1-2^{-s})^{-1}2^{-sj} \\
\Rightarrow\quad & \left\|(2^{sj}b_j)_j\right\|_{\ell_\infty} \leq \frac{k}{1-2^{-s}} = (1-2^{-s})^{-1}\left\|(2^{sj}a_j)_j\right\|_{\ell_\infty}
\end{aligned}$$

$\underline{1 < q < \infty}$: Pick any $0 < t < s$, let $\frac{1}{q'} + \frac{1}{q} = 1$. Then

$$\begin{aligned}
\left\| \left(2^{sj} b_j\right)_j \right\|_{\ell_q}^q &= \sum_{j \geq 0} (2^{sj} b_j)^q = \sum_{j\geq 0} 2^{sjq} \left(\sum_{l \geq j} a_l \right)^q \\
&= \sum_{j\geq 0} 2^{sjq} \left(\sum_{l \geq j} 2^{-tl} (2^{tl} a_l) \right)^q \overset{\text{Hölder}}{\leq} \sum_{j \geq 0} 2^{sjq} \underbrace{\left(\sum_{l\geq j} 2^{-tlq'} \right)^{\frac{q}{q'}}}_{\leq C(t,q)\cdot 2^{-jtq}} \left(\sum_{l \geq j} 2^{tlq} a_l^q \right) \\
&\lesssim \sum_{j\geq 0} 2^{jq(s-t)} \left(\sum_{l \geq j} 2^{ltq} a_l^q \right) = \sum_{l=0}^{\infty} 2^{ltq} a_l^q \underbrace{\sum_{j=0}^{l} 2^{jq(s-t)}}_{\lesssim 2^{lq(s-t)}} \lesssim \sum_{l=0}^{\infty} (2^{ls} a_l)^q \\
&= \left\| (2^{ls} a_l)_l \right\|_{\ell_q}^q
\end{aligned}$$

$\underline{0 < q \leq 1}$:

$$\begin{aligned}
\left\| \left(2^{sj} b_j\right)_j \right\|_{\ell_q}^q &= \sum_{j=0}^{\infty} (2^{sj} b_j)^q = \sum_{j=0}^{\infty} 2^{sjq} \left(\sum_{l=j}^{\infty} a_j \right)^q \overset{q \leq 1}{\leq} \sum_{j=0}^{\infty} 2^{sjq} \sum_{l=j}^{\infty} a_l^q \\
&= \sum_{l=0}^{\infty} a_l^q \sum_{j=0}^{l} 2^{sjq} \lesssim \sum_{l=0}^{\infty} (2^{sl} a_l)^q = \left\| (2^{sl} a_l)_l \right\|_{\ell_q}^q
\end{aligned}$$

□

Proof. Back to Theorem **??**: The first equivalence in (4.5.4) is trivial because

$$\sigma_{2^j}(f)_{\mathbb{X}} \leq \|f - G_{2^j} f\|_{\mathbb{X}} \leq C_0 \sigma_{2^j}(f)_{\mathbb{X}} .$$

As for the second equivalence: Note that since X is a quasi-Banach space

$$\|g_{j+1} - g_j\|_{\mathbb{X}} = \|g_{j+1} - f + f - g_j\|_{\mathbb{X}} \lesssim C_0 \left(\sigma_{2^{j+1}}(f)_{\mathbb{X}} + \sigma_{2^j}(f)_X \right) \lesssim \sigma_{2^j}(f)_{\mathbb{X}} .$$

Using (4.5.1) this yields

$$\left(\sum_{j \geq 0} \left(2^{rj} \|g_{j+1} - g_j\|_{\mathbb{X}}\right)^q \right) \lesssim |f|_{\mathcal{A}_q^r}^q .$$

To prove the converse inequality note first (by telescoping expansion):

$$f = g_0 + \sum_{j=0}^{\infty} (g_{j+1} - g_j) , \qquad \text{the sum converges in } \mathbb{X}.$$

Now suppose that $\|u+v\|_{\mathbb{X}}^{\mu} \leq \|u\|_{\mathbb{X}}^{\mu} + \|v\|_{\mathbb{X}}^{\mu}$ for some $\mu > 0$ (for $\mathbb{X} = \ell_p, L_p, \mu = \min\{1, p\}$). Then

$$\sigma_{2^j}(f)_{\mathbb{X}}^{\mu} \leq \|f - g_j\|_{\mathbb{X}}^{\mu} = \left\| \sum_{k \geq j} (g_{k+1} - g_k) \right\|_{\mathbb{X}}^{\mu} \leq \sum_{k \geq j} \|g_{k+1} - g_k\|_{\mathbb{X}}^{\mu} .$$

Hence

$$\begin{aligned} |f|_{\mathcal{A}_q^r}^q &\overset{(4.5.1)}{\sim} \left\| \left(2^{jr} \sigma_{2^j}(f)_{\mathbb{X}}\right) \right\|_{\ell_q}^q = \sum_{j=0}^{\infty} 2^{jrq} \sigma_{2^j}(f)_{\mathbb{X}}^q \\ &\leq \sum_{j \geq 0} 2^{jrq\frac{\mu}{\mu}} \left(\sum_{k \geq j} \|g_{k+1} - g_k\|_{\mathbb{X}}^{\mu} \right)^{\frac{q}{\mu}} . \end{aligned}$$

Applying Theorem 4.5.2, (4.5.7) with $a_j = \|g_{j+1} - g_j\|_{\mathbb{X}}$, $b_j = \left(\sum_{k \geq j} a_k^{\mu}\right)^{\frac{1}{\mu}}$ yields

$$\left\| \left(2^{jr} \sigma_{2^j}(f)_{\mathbb{X}}\right)_j \right\|_{\ell_q} \lesssim \left\| \left(2^{jr} \|g_{j+1} - g_j\|_{\mathbb{X}}\right)_j \right\|_{\ell_q} ,$$

which proves (4.5.4).

As for (4.5.5) we have

$$\|g_0\|_{\mathbb{X}}^{\mu} \leq \|g_0 - f\|_{\mathbb{X}}^{\mu} + \|f\|_{\mathbb{X}}^{\mu} \leq C_0^{\mu} \sigma_1(f)_{\mathbb{X}}^{\mu} + \|f\|_{\mathbb{X}}^{\mu} \leq (1 + C_0^{\mu}) \|f\|_{\mathbb{X}}^{\mu} ,$$

and conversely

$$\|f\|_{\mathbb{X}}^{\mu} \leq \|g_0\|_{\mathbb{X}}^{\mu} + C_0^{\mu} \sigma_1(f)_{\mathbb{X}} \lesssim (1 + C_0^{\mu}) \|g_0\|_{\mathbb{X}}^{\mu} . \qquad \square$$

As a first application we take $\mathbb{X} = \ell_p(\Lambda)$ which gives a characterization of $\mathcal{A}_q^r((\mathbb{X}_n), \ell_p(\Lambda))$ $(q = p)$.

To this end, consider the *weighted sequence spaces*

$$\ell_p^r(\Lambda) = \{\mathbf{d} \in \ell_p(\Lambda) : \|\mathbf{d}\|_{\ell_p^r} = \|(k^r d_{\lambda_k})_k\|_{\ell_p} < \infty\}. \tag{4.5.9}$$

Theorem 4.5.3. *For* $r > 0$, $p > 0$ *one has*

$$\mathcal{A}_p^r((\mathbb{X}_n), \ell_p(\Lambda)) = \ell_p^r(\Lambda)$$

with equivalent norms.

Proof. ($\mathbb{X} = \ell_p(\Lambda)$, $G_n = P_n$ in Theorem 4.5.1). By (4.5.5) we have

$$\|\mathbf{d}\|_{\mathcal{A}_p^r} \sim \|P_1 \mathbf{d}\|_{\ell_p} + \left\| \left(2^{rj} \|P_{2^{j+1}} \mathbf{d} - P_{2^j} \mathbf{d}\|_{\ell_p}\right)_j \right\|_{\ell_p} = |d_{\lambda_1}| + \left\| \left(2^{rj} \left(\sum_{k=2^j+1}^{2^{j+1}} |d_{\lambda_k}|^p \right)^{\frac{1}{p}} \right)_j \right\|_{\ell_p}$$

and because $2^{rj} \leq k^r \leq 2^{r(j+1)} = 2^r 2^{rj}$

$$\sim |d_{\lambda_1}| + \left\| \left(\sum_{k=2^j+1}^{2^{j+1}} k^{rp} |d_{\lambda_k}|^p \right)^{\frac{1}{p}} \right\|_{\ell_p} = \left(\sum_{k=1}^{\infty} (k^r |d_{\lambda_k}|)^p \right)^{\frac{1}{p}} = \|\mathbf{d}\|_{\ell_p^r} \qquad \square$$

Exercise 4.5.2. *Two cases of interest:*

- *derive an analogous characterization for* $q \neq p$*;*
- *2π-periodic functions and Fourier transforms: Let* $H_{2\pi}^m = \{f \in L_{2,2\pi} : f^{(l)} \in L_{2,2\pi}, l \leq m\}$
  show $\left(\hat{f}(k)\right)_{k\in\mathbb{Z}} \in \mathcal{A}_2^m\left((\mathbb{X}_n), \ell_2(\mathbb{Z})\right)$*,* $\quad \mathbb{X}_n = \operatorname{span}\{e_k : |k| \leq n\}$

4.5.2 Nonlinear Approximation

Now take $\Sigma_n := \{\mathbf{d} \in \ell_\mathbf{p}(\Lambda) : \operatorname{supp} \mathbf{d} \leq \mathbf{n}\}$ the set of all n-*sparse* sequences:

$$\mathcal{A}_q^r\left((\Sigma_n), \ell_p(\Lambda)\right) = \left\{ \mathbf{d} \in \ell_p(\Lambda) : |\mathbf{d}|_{\mathcal{A}_q^r} := \begin{cases} \sup_n n^r \sigma_n(\mathbf{d})_{\ell_p} < \infty \\ \left(\sum_{n\geq 1} \left(n^r \sigma_n(\mathbf{d})_{\ell_p}\right)^q \frac{1}{n}\right)^{\frac{1}{q}} < \infty \end{cases} \right\}$$
$$0 < p, q \leq \infty$$
$$\|\mathbf{d}\|_{\mathcal{A}_q^r} = \|\mathbf{d}\|_{\ell_p} + |\mathbf{d}|_{\mathcal{A}_q^r} \tag{4.5.10}$$

In the following we assume that

$$\#\{\lambda : |d_\lambda| > \eta\} < \infty\,. \tag{4.5.11}$$

For $p < \infty$ this is automatically true but for $p = \infty$ this is an additional assumption. Due to this assumption the decreasing rearrangement $(d_k^*)_{k\in\mathbb{N}}$ of $\mathbf{d}$ is well-defined, i.e., there is an ordering of $\Lambda = (\lambda_k)_{k\in\mathbb{N}}$ such that $|d_{\lambda_1}| = d_1^* \geq |d_{\lambda_2}| = d_2^* \geq ....$

Rearrangements are used to define so called **Lorentz spaces**. As a first instance the space *weak* ℓ_p is defined by

$$w\ell_p(\Lambda) = \{\mathbf{d} : \sup_{k\geq 1} k^{\frac{1}{p}} d_k^* =: \|\mathbf{d}\|_{w\ell_p} < \infty\}\,. \tag{4.5.12}$$

More generally

$$\ell_{p,q}(\Lambda) := \left\{ \mathbf{d} : \left(\sum_{k=1}^{\infty} \left(k^{\frac{1}{p}} d_k^*\right)^q \frac{1}{k} \right)^{\frac{1}{q}} =: \|\mathbf{d}\|_{\ell_{p,q}} < \infty \right\}, \tag{4.5.13}$$

that is $\ell_{p,\infty}(\Lambda) = w\ell_p(\Lambda)$, $\ell_{p,p}(\Lambda) = \ell_p(\Lambda)$.

It can be seen from the definition of $\|\cdot\|_{\ell_{p,q}(\Lambda)}$ that the spaces are *interpolation spaces* (see Definition 2.3.2), i.e.,

$$\ell_{p,q}(\Lambda) = [\ell_{p_1}(\Lambda), \ell_{p_2}(\Lambda)]_{\theta,q} = \ell_{p,q}, \quad \frac{1}{p} = \frac{1-\theta}{p_1} + \frac{\theta}{p_2}. \tag{4.5.14}$$

$w\ell_p$ and $\ell_{p,q}$ are important for the understanding of nonlinear and adaptive approximation. This is explained by the following main result of this section.

Theorem 4.5.4. *For any* $r > 0$, $0 < p \leq \infty$ *one has*

$$\begin{aligned} &\mathcal{A}^r_\infty\left((\Sigma_n), \ell_p(\Lambda)\right) = w\ell_\tau(\Lambda), \qquad \frac{1}{\tau} = r + \frac{1}{p} \\ &\textit{and } \|\cdot\|_{\mathcal{A}^r_\infty} \sim \|\cdot\|_{w\ell_\tau} \end{aligned} \tag{4.5.15}$$

with constants depending on p, τ, *which blow up as* $\tau \to p$.

Moreover, for any $r > 0, 0 < q < \infty, 0 < p \leq \infty$ *one has*

$$\mathcal{A}^r_q\left((\Sigma_n), \ell_p(\Lambda)\right) = \ell_{\tau,q}(\Lambda), \qquad \frac{1}{\tau} = r + \frac{1}{p}, \tag{4.5.16}$$

and in particular

$$\mathcal{A}^r_\tau\left((\Sigma_n), \ell_p(\Lambda)\right) = \ell_{\tau,\tau}(\Lambda) = \ell_\tau(\Lambda)\,. \tag{4.5.17}$$

Comments on Theorem 4.5.4, and discussion of the spaces $w\ell_\tau(\Lambda)$: A few comments before turning to the proof of Theorem 4.5.4: The rate of nonlinear approximation is determinated by the summability index of the smoothness norm. The larger r, the smaller τ and the more concentrated is $\mathbf{d}$. The case $q = \infty$ is of particular interest.

Remark 4.5.1. *Theorem 4.5.4,* (4.5.15) *ensures the validity of the* **Jackson estimate**

$$\sigma_n(\mathbf{d})_{\ell_p} \leq c(\tau, p) n^{-r} \|\mathbf{d}\|_{w\ell_\tau} \tag{4.5.18}$$

with $\frac{1}{\tau} = r + \frac{1}{p}$, $c(p,\tau) \sim \left(\frac{\tau}{p-\tau}\right)$. *Note that the relation* $\frac{1}{\tau} = r + \frac{1}{p}$ *corresponds to the* critical embedding *in Figure 2.4, i.e.,* $\delta = 0$.

Remark 4.5.2. *In* (4.5.18) $c(\tau, p)$ *can be replaced by* 1 *if* $w\ell_\tau$ *is replaced by* ℓ_τ, *i.e.,*

$$\sigma_n(\mathbf{d})_{l_p} \leq n^{-r} \|\mathbf{d}\|_{\ell_\tau} \tag{4.5.19}$$

for $\frac{1}{\tau} = r + \frac{1}{p}$ *(with constant one). This is often called* Stechkin's inequality.

Proof. We have

$$\begin{aligned}\sigma_n(\mathbf{d})^p_{\ell_p} &= \sum_{k=n+1}^{\infty} (d_k^*)^p \le (d_n^*)^{p-\tau} \sum_{k=n+1}^{\infty} (d_k^*)^\tau \\ &\le \left(\frac{1}{n}\sum_{k=1}^{n}(d_k^*)^\tau\right)^{\frac{p-\tau}{\tau}} \left(\sum_{k=n+1}^{\infty}(d_k^*)^\tau\right) \\ &\le n^{1-\frac{p}{\tau}} \|\mathbf{d}\|^{p-\tau}_{l_\tau} \|\mathbf{d}\|^{\tau}_{l_\tau} \\ &= n^{1-\frac{p}{\tau}} \|\mathbf{d}\|^{p}_{\ell_\tau} .\end{aligned}$$

This implies

$$\sigma_n(\mathbf{d})_{\ell_p} \le n^{\frac{1}{p}-\frac{1}{\tau}} \|\mathbf{d}\|_{\ell_\tau} . \qquad \square$$

Relation between weak and strong ℓ_τ-spaces:

Remark 4.5.3. *a) One has* $\ell_p(\Lambda) \subsetneqq w\ell_p(\Lambda) \subset \ell_{p+\epsilon}(\Lambda)$ *for any* $\epsilon > 0$.

b) For any $\tau < p$ *one has* $\|\cdot\|_{\ell_p} \le c(\tau, p)\, \|\cdot\|_{w\ell_\tau}$.

Proof. a) Exercise.
b) We have

$$\begin{aligned}\|\mathbf{d}\|^p_{\ell_p} &= \sum_\lambda |d_\lambda|^p = \sum_\lambda |d_\lambda|^\tau |d_\lambda|^{p-\tau} = \sum_{k\in\mathbb{N}} \left(k^{\frac{1}{\tau}} d_k^*\right)^\tau k^{-1}(d_k^*)^{p-\tau} \\ &\le \|\mathbf{d}\|^\tau_{wl_\tau} \sum_{k\in\mathbb{N}} k^{-1}(d_k^*)^{p\frac{p-\tau}{p}} \overset{\text{Hölder}}{\le} \|\mathbf{d}\|^\tau_{wl_\tau} \left(\sum_{k\in\mathbb{N}} k^{-\frac{p}{\tau}}\right)^{\frac{\tau}{p}} \left(\sum_{k\in\mathbb{N}} (d_k^*)^p\right)^{\frac{p-\tau}{p}} \\ &\le \underbrace{c(p-\tau)}_{\text{explodes for } \tau\to p} \|\mathbf{d}\|^\tau_{wl_\tau} \|\mathbf{d}\|^{p-\tau}_{\ell_p} ,\end{aligned}$$

and hence

$$\|\mathbf{d}\|^\tau_{l_p} \le c(p-\tau)\, \|\mathbf{d}\|^\tau_{wl_\tau} \quad \Leftrightarrow \quad \|\mathbf{d}\|_{\ell_p} \le c(p,\tau)^{\frac{1}{\tau}}\, \|\mathbf{d}\|_{w\ell_\tau} .$$

$\square$

It is often convenient to employ a different way of describing the norm $\|\mathbf{d}\|_{w\ell_p}$. To this end, consider the **redistribution function**

$$\eta \mapsto \mu_{\mathbf{d}}(\eta) := \#\{\lambda : |d_\lambda| \ge \eta\} . \qquad (4.5.20)$$

Remark 4.5.4. *Let* $0 < p < \infty$. $\mathbf{d} \in w\ell_p(\Lambda)$ *if and only if there exists a constant* M *such that*

$$\mu_{\mathbf{d}}(\eta) = \#\{\lambda : |d_\lambda| \geq \eta\} \leq M\eta^{-p} \tag{4.5.21}$$

and for the smallest constant M *in* (4.5.21) *we have*

$$M = \sup_{\eta>0} \eta^p \mu_{\mathbf{d}}(\eta) = \|\mathbf{d}\|^p_{w\ell_p} . \tag{4.5.22}$$

Proof. Pick any $\eta > 0$, choose the largest $m \in \mathbb{N}$ such that $d^*_m \geq \eta$, that is from $\mu_{\mathbf{d}}(\eta) = m$ follows $\eta^p \mu_{\mathbf{d}}(\eta) \leq m(d^*_m)^p \overset{\text{def.}}{\leq} \|\mathbf{d}\|^p_{w\ell_p}$. Then, since $\eta > 0$ was arbitrary,

$$M \leq \|\mathbf{d}\|^p_{w\ell_p} .$$

Conversely: For $\epsilon > 0$ choose m such that

$$m(d^*_m)^p \geq \|\mathbf{d}\|^p_{w\ell_p} - \epsilon$$

but since $m = \mu_{\mathbf{d}}(d^*_m)$ we have

$$m(d^*_m)^p = (d^*_m)^p \mu_{\mathbf{d}}(d^*_m) \leq \sup_{\eta>0} \eta^p \mu_{\mathbf{d}}(\eta) = M,$$

and since ϵ was arbitrary

$$M \geq \|\mathbf{d}\|^p_{w\ell_p} .$$ □

Finally we record one further equivalent quantity

Remark 4.5.5. *Consider*

$$\sup_{\eta>0} \eta^p \#\{\lambda : 2\eta \geq |d_\lambda| > \eta\} =: \tilde{M} . \tag{4.5.23}$$

Then

$$\tilde{M} \leq \underbrace{\|\mathbf{d}\|^p_{w\ell_p}}_{=M} \leq (1 - 2^{-p})^{-1} \tilde{M} . \tag{4.5.24}$$

Proof. Clearly $\tilde{M} \leq M = \|\mathbf{d}\|^p_{w\ell_p}$. Conversely,

$$\begin{aligned} \#\{\lambda : |d_\lambda| > \eta\} &= \sum_{j\geq 0} \#\{\lambda : 2^{j+1}\eta \geq |d_\lambda| > 2^j\eta\} \overset{(4.5.23)}{\leq} \sum_{j\geq 0} 2^{-jp}\eta^{-p}\tilde{M} \\ &= \eta^{-p}\tilde{M} \sum_{j\geq 0} 2^{-jp} = \eta^{-p}\tilde{M} \frac{1}{1 - 2^{-p}} \end{aligned}$$ □

Proof of Theorem 4.5.4: Assume first $p < \infty$: For $\mathbf{d} \in w\ell_\tau$ show $\|bd\|_{w\ell_\tau} \lesssim \|\mathbf{d}\|_{\mathcal{A}^r_\infty}$:

$$\begin{aligned}\sigma_n(\mathbf{d})^p_{\ell_p} &= \sum_{k=n+1}^{\infty} (d^*_k)^p = \sum_{k=n+1}^{\infty} k^{\frac{-p}{\tau}} \left(k^{\frac{1}{\tau}} d^*_k\right)^p \\ &\le \|\mathbf{d}\|^p_{wl_\tau} \sum_{k=n+1}^{\infty} k^{\frac{-p}{\tau}} \le \|\mathbf{d}\|^p_{w\ell_\tau} \int_n^\infty s^{\frac{-p}{\tau}}\, ds = \|\mathbf{d}\|^p_{w\ell_\tau} \left(\frac{\tau}{p-\tau}\right) n^{1-\frac{p}{\tau}}\end{aligned}$$

$$\begin{aligned}\Rightarrow \quad & \sigma_n(\mathbf{d})_{\ell_p} \le n^{\frac{1}{p}-\frac{1}{\tau}} \left(\frac{\tau}{p-\tau}\right)^{\frac{1}{p}} \|\mathbf{d}\|_{w\ell_\tau} = n^{-r} \left(\frac{\tau}{p-\tau}\right)^{\frac{1}{p}} \|\mathbf{d}\|_{w\ell_\tau} \\ \Rightarrow \quad & |\mathbf{d}|_{\mathcal{A}^r_\infty} \le \left(\frac{\tau}{p-\tau}\right)^{\frac{1}{p}} \|\mathbf{d}\|_{w\ell_\tau} \\ \Rightarrow \quad & \mathbf{d} \in \mathcal{A}^r_\infty.\end{aligned} \tag{4.5.25}$$

Also, since $\tau < p$, we can use Remark 4.5.3 b) which says $\|\cdot\|_{\ell_p} \le c(p,\tau)\, \|\cdot\|_{w\ell_\tau}$. Since $w\ell_\tau \subseteq \mathcal{A}^r_\infty$ we conclude

$$\|\mathbf{d}\|_{\mathcal{A}^r_\infty} \lesssim \left(1 + \left(\frac{\tau}{p-\tau}\right)^{\frac{1}{p}}\right) \|\mathbf{d}\|_{w\ell_\tau}\ . \tag{4.5.26}$$

For the converse direction suppose $\mathbf{d} \in \mathcal{A}^r_\infty$. Then for any $m \in \mathbb{N}$ we have

$$\begin{aligned} m(d^*_{2m})^p &\le \sum_{k=m+1}^{2m} (d^*_k)^p \le \sum_{k=m+1}^{\infty} (d^*_k)^p = \sigma_m(\mathbf{d})^p_{\ell_p} \overset{\text{def. of } \mathbf{d}\in\mathcal{A}^r_\infty}{\le} m^{-rp}\, |\mathbf{d}|^p_{\mathcal{A}^r_\infty} \\ &= m^{1-\frac{p}{\tau}}\, |\mathbf{d}|^p_{\mathcal{A}^r_\infty} \\ \Rightarrow \quad & m^{\frac{1}{\tau}} d^*_{2m} \le |\mathbf{d}|_{\mathcal{A}^r_\infty}\ .\end{aligned}$$

So considering different cases we get

$$\begin{cases} 2^{-\frac{1}{\tau}}(2m)^{\frac{1}{\tau}} d^*_{2m} \le |\mathbf{d}|_{\mathcal{A}^r_\infty}\ , m \in \mathbb{N} \\ (2m+1)^{\frac{1}{\tau}} \frac{m^{\frac{1}{\tau}}}{(2m+1)^{\frac{1}{\tau}}} d^*_{2m+1} \le |\mathbf{d}|_{\mathcal{A}^r_\infty}\ , m \in \mathbb{N} \\ d^*_1 \le |\mathbf{d}|_{\mathcal{A}^r_\infty} \end{cases}$$

which gives us

$$\|\mathbf{d}\|_{w\ell_\tau} \le 3^{\frac{1}{\tau}}\, \|\mathbf{d}\|_{\mathcal{A}^r_\infty}\ .$$

Consider now $p = \infty$: Let $r = \frac{1}{\tau}$ and assume that $\mathbf{d} \in w\ell_\tau$

$$\sigma_n(\mathbf{d})_{\ell_\infty} = d^*_{n+1} \le n^{-\frac{1}{\tau}}\, \|\mathbf{d}\|_{w\ell_\tau} = n^{-r}\, \|\mathbf{d}\|_{w\ell_\tau}\ .$$

By the definition of $|\mathbf{d}|_{\mathcal{A}^r_\infty}$ it follows that

$$|\mathbf{d}|_{\mathcal{A}^r_\infty} \leq \|\mathbf{d}\|_{w\ell_\tau} ,$$

and consequently

$$\|\mathbf{d}\|_{\mathcal{A}^r_\infty} \leq 2 \, \|\mathbf{d}\|_{w\ell_\tau} .$$

Conversely: For $\mathbf{d} \in \mathcal{A}^r_\infty$ we have

$$n^{\frac{1}{\tau}} d^*_{n+1} = n^r d^*_{n+1} = n^r \sigma_n(\mathbf{d})_{\ell_\infty} \leq |\mathbf{d}|_{\mathcal{A}^r_\infty} .$$

Taking the supremum yields

$$\sup_{n \geq 1} n^{\frac{1}{\tau}} d^*_n \leq 2^{\frac{1}{\tau}} \, |\mathbf{d}|_{\mathcal{A}^r_\infty} ,$$

and by $d^*_1 = \|\mathbf{d}\|_{\ell_\infty}$ follows

$$\|\mathbf{d}\|_{w\ell_\tau} \leq 2^{\frac{1}{\tau}} \, \|\mathbf{d}\|_{\mathcal{A}^r_\infty} ,$$

The proof of the remaining assertion (4.5.16) follows from interpolation, using (4.5.14) and (2.3.25) in Remark 2.3.8. □

Different objectives: N-term approximation relates the total error to the number of terms needed to achieve it. It is based on sorting coefficients by size. Instead of cutting according to a desired number of terms we may also truncate the coefficient sequence according to their size. This is called "thresholding".

- So far, we prescribe the *budget* n and asked for the n-term error $\sigma_n(\mathbf{d})_{\ell_p}$.
- Now we want to control the *size* of the coefficients and ask for the error incurred by cutting off coefficients below a given *threshold* (Image Compression).

To this end consider for any $\eta > 0$

$$(T_\eta \mathbf{d})_\lambda := \begin{cases} d_\lambda & \text{if } |d_\lambda| \geq \eta \, , \\ 0 & \text{otherwise} \, . \end{cases}$$

How does this relate to $\sigma_n(\mathbf{d})_{\ell_p}$?

Theorem 4.5.5. *Let* $0 < \tau < p \le \infty$*, then* $\mathbf{d}$ *belongs to* $w\ell_\tau(\Lambda)$ *with* $\frac{1}{\tau} - \frac{1}{p} =: r$ *if and only if*

$$\underbrace{\sup_{\eta>0} \eta^{\frac{\tau}{p}-1} \|\mathbf{d} - T_\eta \mathbf{d}\|_{\ell_p}}_{=:C^*(\mathbf{d})} < \infty \tag{4.5.27}$$

and

$$C^*(\mathbf{d}) \le C \|\mathbf{d}\|_{w\ell_\tau}^{\frac{\tau}{p}} . \tag{4.5.28}$$

Let us see what this means: From (4.5.27) and (4.5.28) we obtain

$$\begin{aligned}
\|\mathbf{d} - T_\eta \mathbf{d}\|_{\ell_p} &\le C \|\mathbf{d}\|_{w\ell_\tau}^{\frac{\tau}{p}} \eta^{1-\frac{\tau}{p}} = C \|\mathbf{d}\|_{w\ell_\tau}^{\frac{\tau}{p}-1} \eta^{1-\frac{\tau}{p}} \|\mathbf{d}\|_{w\ell_\tau} \\
&= C \left(\frac{\eta}{\|\mathbf{d}\|_{w\ell_\tau}} \right)^{1-\frac{\tau}{p}} \|\mathbf{d}\|_{w\ell_\tau} \overset{(4.5.22)}{\le} C \left(\mu_{\mathbf{d}}(\eta)^{-\frac{1}{\tau}} \right)^{1-\frac{\tau}{p}} \|\mathbf{d}\|_{w\ell_\tau} \\
&= C\mu_{\mathbf{d}}(\eta)^{\frac{1}{p}-\frac{1}{\tau}} \|\mathbf{d}\|_{w\ell_\tau} .
\end{aligned}$$

Hence we obtain

$$\|\mathbf{d} - T_\eta \mathbf{d}\|_{\ell_p} \le C \left(\# \operatorname{supp} T_\eta \mathbf{d}\right)^{-r} \|\mathbf{d}\|_{w\ell_\tau}$$

which means that thresholding has best $(n = \mu_{\mathbf{d}}(\eta))$-term performance.

Proof. of Theorem 4.5.5 ($p < \infty$): Assume $\mathbf{d} \in w\ell_\tau$, show (4.5.27), (4.5.28):

$$\begin{aligned}
\|\mathbf{d} - T_\eta \mathbf{d}\|_{\ell_p}^p &= \sum_{|d_\lambda| \le \eta} |d_\lambda|^p = \sum_{j \ge 0} \sum_{2^{-(j+1)}\eta < |d_\lambda| \le 2^{-j}\eta} |d_\lambda|^p \\
&\le \sum_{j \ge 0} 2^{-jp} \eta^p \underbrace{\#\{\lambda : |d_\lambda| > 2^{-(j+1)}\eta\}}_{=\mu_{\mathbf{d}}(2^{-(j+1)}\eta)} \overset{(4.5.22)}{\le} \sum_{j \ge 0} 2^{-jp} \eta^p \|\mathbf{d}\|_{w\ell_\tau}^\tau \eta^{-\tau} 2^{\tau(j+1)} \\
&= \|\mathbf{d}\|_{w\ell_\tau}^\tau \eta^{p-\tau} 2^\tau \sum_{j \ge 0} 2^{-j(p-\tau)} \le c(p,\tau) \eta^{p-\tau} \|\mathbf{d}\|_{w\ell_\tau}^\tau
\end{aligned}$$

This proves (4.5.27) and (4.5.28).

Suppose now that $\eta^{1-\frac{\tau}{p}} \|\mathbf{d} - T_\eta \mathbf{d}\|_{\ell_p} \le C^*$. We wish to show that $\mathbf{d} \in w\ell_\tau$. To this end, consider for any $\eta > 0$

$$\eta^p \#\{\lambda : 2\eta \ge |d_\lambda| > \eta\} = \sum_{\lambda : \eta < |d_\lambda| \le 2\eta} \eta^p \le \sum_{\lambda : \eta < |d_\lambda| \le 2\eta} |d_\lambda|^p \le \sum_{\lambda : |d_\lambda| \le 2\eta} |d_\lambda|^p = \|\mathbf{d} - T_{2\eta} \mathbf{d}\|_{\ell_p}^p .$$

Hence

$$\eta^\tau \#\{\lambda : 2\eta \ge |d_\lambda| > \eta\} \le \eta^{\tau-p} \|\mathbf{d} - T_{2\eta} \mathbf{d}\|_{\ell_p}^p = 2^{-(\tau-p)} (2\eta)^{\tau-p} \|\mathbf{d} - T_{2\eta} \mathbf{d}\|_{\ell_p}^p$$

which is by assumption

$$\leq 2^{p-\tau}(C^*)^p\,.$$

By (4.5.24) we obtain

$$\|\mathbf{d}\|_{w\ell_p} \leq (1-2^{-p})^{-1}2^{p-\tau}(C^*)^p. \qquad \square$$

So far we have prescribed:

- the budget n and
- the minimal coefficient size.

There is yet another way of monitoring the best n-term approximation, namely

- prescribing the target accuracy

and choosing the smallest number of terms to achieve that error. We see next how to bound this number.

Exercise 4.5.3. *Assume that* $\mathbf{d} \in w\ell_\tau(\Lambda)$. *Define the coarsening operator*

$$\mathcal{C}_\epsilon(\mathbf{d}) = (d_k^*)_{k=1}^{n(\epsilon)}$$

where

$$n(\epsilon) = \operatorname*{argmin}_n \left\{ \left(\sum_{k>n} (d_k^*)^p \right)^{\frac{1}{p}} \leq \epsilon \right\},$$

i.e., $\mathcal{C}_\epsilon$ *coarsens* $\mathbf{d}$ *back to shortest subsequence – which is a best* $n(\epsilon)$*-term approximation – that realizes the target accuracy* ϵ. *Then one has*

$$n(\epsilon) \lesssim \epsilon^{-\frac{1}{r}} \|\mathbf{d}\|_{w\ell_\tau}^{\frac{1}{r}}, \qquad \frac{1}{\tau} = r + \frac{1}{p}\,. \tag{4.5.29}$$

Proof. Exercise. □

4.6 Wavelets and Function Spaces: a Quick Tour

So called *wavelet bases* form an important class of Riesz bases for L_2 and corresponding Sobolev bases. They also provide simple characterizations of Besov spaces. The most convenient framework for *constructing wavelet* bases is the concept of *Multiresolution Spaces* that has been already briefly indicated earlier, see (4.1.1). The envisaged characterization of function spaces through wavelet bases can be based on the characterization of function spaces through such multiresolution spaces.

The main difference from what one mostly finds in the literature is to avoid making essential use of Fourier-transforms because the concepts should eventually work on general domains rather than on the whole Euclidean space or the torus.

4.6.1 Multiresolution Spaces

In their classical form such multiresolution spaces V_j are generated by a single so called *scaling function* $\phi(x)$ defined on $\mathbb{R}$, see e.g. [30, 36]. Then defining (as in the example of the Haar wavelet) $\phi_{j,k}(x) := 2^{j/2}\phi(2^j x - k)$ and defining $V_j := \operatorname{span}\{\phi_{j,k} : j, k \in \mathbb{Z}\}$ the spaces V_j are indeed nested, $V_j \subset V_{j+1}$, if the scaling function is *refinable*, i.e., there exist coefficients a_k such that

$$\phi(x) = \sum_{k\in\mathbb{Z}} a_k \phi(2x - k), \quad x \in \mathbb{R}, \tag{4.6.1}$$

which says $V_0 \subset V_1$ and implies the general case by dilation. One also needs to quantify the *stability* of the translates $\phi(\cdot - k)$. The wavelets are then obtained (as in the case of the Haar basis) by identifying a basis $\{\psi(\cdot - k) : k \in \mathbb{Z}\}$ for

$$W_0 := V_1 \ominus V_0, \tag{4.6.2}$$

to obtain then a multi-scale basis $\psi_{j,k}$ again by dilation and translation. For the resulting collection Ψ to form a Riesz-basis for $L_2(\mathbb{R})$, say it is important how the complement in (4.6.2) is formed. An *orthogonal* complement as in the Haar-case would be perfect.

The advantage of the above framework lies in the simplicity and convenience of the pivotal role of *dilation and translation* and how this allows one to use Fourier techniques.

The disadvantage is that this works well on all of $\mathbb{R}$ (and $\mathbb{R}^d$ by tensor-products) and also on the torus by periodization but not on more general domains arising in applications.

It is therefore important to note that the framework based on dilation and translation is by no means essential. Instead the following general notion of *Multiresolution Spaces* which indeed applies to a much wider scope of practically relevant scenarios extracts the essential features driving *multi-scale analysis*, see also [23].

Definition 4.6.1. *Let $\Omega \subset \mathbb{R}^d$ be a domain. A* multiresolution approximation *is a sequence $(V_j)_{j\geq 0}$ of spaces of functions defined on Ω which satisfies the following properties.*

(i) *Nestedness: One has $V_j \subset V_{j+1}$ for all $j \geq 0$.*

(ii) *Denseness: For all $0 < p \leq \infty$, and all $f \in L_p(\Omega)$ or f uniformly continuous when $p = \infty$, there exists a sequence $(f_j)_{j\geq 0}$ with $f_j \in V_j$ such that $\lim_{j\to\infty} \|f - f_j\|_{L_p} = 0$.*

(iii) *Generating functions: For each $j \geq 0$, there exists a finite or countable family $\{\varphi_\gamma\}_{\gamma\in\Gamma_j}$ of compactly supported functions from $L_\infty(\Omega)$ with finitely overlapping supports (any $x \in \Omega$ is contained in at most finitely many supports) and such that V_j is the set of functions of the form*

$$g := \sum_{\gamma\in\Gamma_j} c_\gamma \varphi_\gamma, \tag{4.6.3}$$

where the c_γ are real numbers.

(iv) *Support properties: For each* $j \geq 0$*, there exists a family of domains* $\{S_\gamma\}_{\gamma\in\Gamma_j}$ *which is a covering of* Ω *in the sense that*

$$|\Omega \setminus \cup_{\gamma\in\Gamma_j} S_\gamma| = 0, \tag{4.6.4}$$

and such that

$$|\text{supp}(\varphi_\gamma) \setminus S_\gamma| = 0. \tag{4.6.5}$$

The family $\mathcal{F}_S := \{S_\gamma\}_{\gamma\in\Gamma}$ *with* $\Gamma := \bigcup_{j\geq 0} \Gamma_j$ *is shape preserving and satisfies*

$$0 < \inf_{j\geq 0} \inf_{\gamma\in\Gamma_j} 2^j \text{diam}(S_\gamma) \leq \sup_{j\geq 0} \sup_{\gamma\in\Gamma_j} 2^j \text{diam}(S_\gamma) < \infty, \tag{4.6.6}$$

and

$$\sup_{\gamma\in\Gamma} \#E_\gamma < \infty, \tag{4.6.7}$$

where $E_\gamma := \{\mu \in \Gamma_j : \ S_\mu \cap S_\gamma \neq \emptyset\}$ *for* $\gamma \in \Gamma_j$.

(v) *Local* L_p*-stability: For all* $0 < p \leq \infty$*, there exist constants* $0 < c_p \leq C_p$*, such that for all* $j \geq 0$*, all* $g \in V_j$*, and all* $\gamma \in \Gamma_j$*, one has*

$$c_p 2^{j(\frac{d}{2}-\frac{d}{p})} \|\mathbf{c}\|_{\ell_p(E_\gamma)} \leq \|g\|_{L_p(S_\gamma)} \leq C_p 2^{j(\frac{d}{2}-\frac{d}{p})} \|\mathbf{c}\|_{\ell_p(E_\gamma)}, \tag{4.6.8}$$

for all vectors $\mathbf{c} = (c_\mu)_{\mu\in E_\gamma}$.

Remark 4.6.1. *Note that the scaling functions* φ_γ *are normalized in* L_2*. If we replace for a* $p \neq 2$ *the function* φ_γ *by* $\varphi_\gamma / \|\varphi_\gamma\|_{L_p(\Omega)}$ *the factors* $2^{j(\frac{d}{2}-\frac{d}{p})}$ *do not occur in* (4.6.8).

Exercise 4.6.1. 1. *When* Ω *is bounded the* V_j *are finite dimensional.*

2. *The local* L_p*-stability property* (4.6.8) *implies that the functions* $(\varphi_\gamma)_{\gamma\in\Gamma_j}$ *are linearly independent*

$$\sum_{\gamma\in\Gamma_j} c_\gamma \varphi_\gamma = 0 \Rightarrow c_\gamma = 0, \quad \gamma \in \Gamma_j, \tag{4.6.9}$$

and therefore constitute a basis of V_j.

3. *Show that* (4.6.8) *implies*

$$c_p 2^{|\gamma|(\frac{d}{2}-\frac{d}{p})} \leq \|\varphi_\gamma\|_{L_p} \leq C_p 2^{|\gamma|(\frac{d}{2}-\frac{d}{p})}, \quad \gamma \in \Gamma. \tag{4.6.10}$$

as well as the global stability property

$$c_p 2^{j(\frac{d}{2}-\frac{d}{p})} \|\mathbf{c}\|_{\ell_p} \leq \| \sum_{\gamma\in\Gamma_j} c_\gamma \varphi_\gamma \|_{L_p(\Omega)} \leq C_p 2^{j(\frac{d}{2}-\frac{d}{p})} \|\mathbf{c}\|_{\ell_p}, \tag{4.6.11}$$

for all $\mathbf{c} = (c_\gamma)_{\gamma\in\Gamma_j}$.

For us the case $p = 0$ plays a pivotal role. Therefore, we assume in what follows that the φ_γ are normalized in L_2, i.e.,

$$\|\varphi_\gamma\|_{L_2} = 1, \quad \gamma \in \Gamma. \tag{4.6.12}$$

Notational convention:

$$\gamma \in \Gamma_j \quad \Leftrightarrow \quad |\gamma| = j.$$

Example: $\gamma = (j, k)$, $|\gamma| = j$ denotes the dyadic level.

The definition of $V_{j+1} \ominus V_j$ relies crucially on the existence of a *dual* or *biorthogonal multiresolution* which accompanies the given *primal* multiresolution $(V_j)_{j\in\mathbb{N}_0}$ in the following sense.

Definition 4.6.2. *Let* $(V_j)_{j\geq 0}$ *be a multiresolution approximation on a domain* $\Omega \subset \mathbb{R}^d$ *and let* $1 \leq p \leq \infty$. *We say that a family* $\{\tilde{\varphi}_\gamma\}_{\gamma\in\hat{E}\Gamma}$ *of functions from* $L_p(\Omega)$, *or from the space* $\mathcal{M}(\Omega)$ *of Radon measure in the case* $p = 1$, *is a system of* L_p*-stable* dual scaling functions *if and only if it satisfies the following properties.*

(i) *Finite overlapping and localization: for all* $\gamma \in \Gamma$,

$$|\mathrm{supp}(\tilde{\varphi}_\gamma) \setminus S_\gamma| = 0, \tag{4.6.13}$$

where the S_γ *are the sets introduced in Definition* 4.6.1.

(ii) L_p *bounds: there exists a constant* $\tilde{C}_p > 0$ *such that one has*

$$\|\tilde{\varphi}_\gamma\|_{L_p(\Omega)} \leq \tilde{C}_p 2^{|\gamma|(\frac{d}{2}-\frac{d}{p})}. \tag{4.6.14}$$

In the case $p = 1$ *the* L_1 *norm is replaced by the dual norm*

$$\|f\|_{\mathcal{M}} = \sup\{\langle f, g\rangle_{\mathcal{M},\mathcal{C}} \ : \ g \in \mathcal{C}(\bar{\Omega}),\ \|g\|_{L_\infty} = 1\}, \tag{4.6.15}$$

where $\langle\cdot,\cdot\rangle_{\mathcal{M},\mathcal{C}}$ *denotes the duality product between Radon measures and continuous functions.*

(iii) *Biorthogonality:*

$$\langle \varphi_\gamma, \tilde{\varphi}_\mu\rangle = \int_\Omega \varphi_\gamma(x)\tilde{\varphi}_\mu(x)dx = \delta_{\gamma,\mu}, \quad \gamma, \mu \in \Gamma_j, \quad j \geq 0, \tag{4.6.16}$$

where $\langle\cdot,\cdot\rangle$ *is* $\langle\cdot,\cdot\rangle_{\mathcal{M},\mathcal{C}}$ *in the case* $p = 1$, *assuming that the scaling functions* φ_γ *are continuous.*

Exercise 4.6.2. *The value of* p *is fixed in the above definition. However, it is easy to check that an* L_p*-stable dual system is also* L_r*-stable for any* $1 \leq r \leq p$.

Remark 4.6.2. *Whether such multiresolution spaces are suitable for characterizing function spaces relies on the validity of two types of inequalities:* Berntein and Jackson inequalities, *similar to the characterization of approximation spaces as interpolation spaces.*

Canonical projectors The systems of dual scaling functions induce associated projectors in a natural way that will turn out to have near-best approximation properties.

Exercise 4.6.3. *For the primal and dual scaling functions from Definitions* 4.6.1, 4.6.2 *and* $1 \le p \le \infty$ *show that*

$$P_j f := \sum_{\gamma \in \Gamma_j} \langle f, \tilde{\varphi}_\gamma \rangle \varphi_\gamma \tag{4.6.17}$$

are projectors from $L_p(\Omega)$ *onto* V_j.

Remark 4.6.3. *Lagrange interpolation and orthogonal projections are special cases of projectors. The projectors* (4.6.17) *are special in that the the functions* $\tilde{\varphi}_\gamma$ *inducing the dual functionals are also refinable. If for* $p = 2$ (4.6.16) *holds for* $\tilde{\varphi}_\gamma = \varphi_\gamma$*, the* P_j *is the orthogonal projector onto* V_j.

Introducing the *influence* domains

$$R_\gamma := \cup\{S_\mu \ : \ \mu \in E_\gamma\}, \tag{4.6.18}$$

we observe that the value of $P_j f$ in S_γ depends only on the value of f in R_γ. The projectors P_j are locally stable, as shown by the following result.

Assumption 4.6.1. *The collection of sets* $\{S_\gamma\}_{\gamma \in \Gamma}, \{R_\gamma\}_{\Gamma \in \Gamma}$, $\Gamma := \bigcup_{j \in \mathbb{N}_0} \Gamma_j$*, are uniformly shape regular.*

The first observation is that these projectors are (globally and locally) stable.

Theorem 4.6.1. *Let* $(V_j)_{j \ge 0}$ *be a multiresolution approximation on a domain* $\Omega \subset \mathbb{R}^d$ *and let* $1 \le p \le \infty$*. Assume that the dual system* $\{\tilde{\varphi}_\gamma\}_{\gamma \in \Gamma}$ *is* $L_{p'}$*-stable with* p' *the conjugate exponent of* p*. Then one has for all* $f \in L_p(\Omega)$*, if* $p < \infty$*, or* $f \in C(\bar{\Omega})$*, if* $p = \infty$*,*

$$\|P_j f\|_{L_p(\Omega)} \le C\|f\|_{L_p(\Omega)}, \quad j \ge 0, \tag{4.6.19}$$

where C *depends on the stability constants* C_p *and* $\tilde{C}_{p'}$*, and on the maximal number of supports the dual scaling functions overlapping any given* $x \in \Omega$*. One also has*

$$\|P_j f\|_{L_p(S_\gamma)} \le C\|f\|_{L^p(R_\gamma)}, \quad |\gamma| = j, \ \ j \ge 0. \tag{4.6.20}$$

Proof. We know from (4.6.11) that

$$\|P_j f\|_{L^p(\Omega)} \le C_p 2^{j(\frac{d}{2} - \frac{d}{p})} \|\mathbf{c}\|_{\ell_p} \tag{4.6.21}$$

with $\mathbf{c} = (\langle f, \tilde{\varphi}_\gamma \rangle)_{\gamma \in \Gamma_j}$. Using Hölder's inequality, we obtain.

$$|\langle f, \tilde{\varphi}_\gamma \rangle| \le \|f\|_{L_p(\tilde{S}_\gamma)} \|\tilde{\varphi}_\gamma\|_{L_{p'}} \le \|f\|_{L_p(\tilde{S}_\gamma)} \tilde{C}_{p'} 2^{j(\frac{d}{2} - \frac{d}{p'})}. \tag{4.6.22}$$

Combining these two estimates and using the finite overlapping properties of the dual scaling functions we obtain (4.6.19). Applying (4.6.19) to $f\chi_{R_\gamma}$ gives (4.6.20) since $P_j(f\chi_{R_\gamma}) = P_j f$ on S_γ. □

An immediate consequence of this stability is the so called *Lebesgue-type inequality.*

Exercise 4.6.4. *Under the assumptions of Theorem 4.6.1 there exists for each* p*,* $1 \leq p \leq \infty$ *(with* $C(\bar{\Omega})$ *in place of* $L_\infty(\Omega)$*) a constant* $\bar{C}_p$ *such that*

$$\|f - P_j f\|_{L_p(\Omega)} \leq \bar{C}_p \inf_{g \in V_j} \|f - g\|_{L_p(\Omega)}, \quad j \in \mathbb{N}_0. \tag{4.6.23}$$

Give a bound for $\bar{C}_p$*. That is, the* P_j *provide near-best (linear) approximations from the spaces* V_j*.*

We can now apply Theorem 4.5.1 with the following specifications:

- $\mathbb{X} = L_p(\Omega)$, $\Omega \subset \mathbb{R}^d$ a bounded domain;
- $g_j = P_j f$, $j \in \mathbb{N}_0$.

Corollary 4.6.1. *With the above specifications we have for* $1 \leq p \leq \infty$*,* $0 < q < \infty$*,* $s > 0$

$$\|f\|_{\mathcal{A}_q^{s/d}((\Sigma_n);L_p(\Omega))} \sim \|P_0 f\|_{L_p(\Omega)} + \Big(\sum_{j=1}^{\infty} 2^{jsq} \|(P_{j+1} - P_j)f\|^q_{L_p(\Omega)} \Big)^{1/q} \tag{4.6.24}$$

and

$$\|f\|_{\mathcal{A}_\infty^{s/d}((\Sigma_n);L_p(\Omega))} \sim \|P_0 f\|_{L_p(\Omega)} + \sup_{j \in \mathbb{N}} 2^{js} \|(P_{j+1} - P_j)f\|_{L_p(\Omega)}. \tag{4.6.25}$$

Proof. Since Ω is bounded the cardinality of the supports S_γ, $\gamma \in \Gamma_j$ scales (because of finite overlap (4.6.7)) like $\#\Gamma_j \sim 2^{dj}$. By (4.5.3) and (4.6.23) we have

$$\|f - P_j f\|_{L_p(\Omega)} \leq C_0 \sigma_{a2^{dj}}(f)_p.$$

Taking $\Sigma_n := V_j$ for $a2^{dj} = \dim V_j \leq n < a2^{d(j+1)} = a2^d 2^{dj} = \dim V_{j+1}$, $r = s/d$, the claims (4.6.24), (4.6.25) forllow from Theorem 4.5.1. □

Thus, the approximation spaces associated with the multiresolution $(V_j)_{j \in \mathbb{N}_0}$ is characterized as expected by the decay of the fluctuations $\|(P_{j+1} - P_j)f\|_{L_p(\Omega)}$ between successive resolution levels.

The next question is whether the approximation spaces are related to classical function spaces. This turns out to be true provided that the multiresolution satisfies certain *Bernstein- and Jackson-type inequalities.*

Jackson-Inequality: The projectors P_j discussed above provide near-best approximations but how rapidly the approximation errors tend to zero requires more properties of the multiresolution $(V_j)_{j \in \mathbb{N}_0}$. Specifically, this dependes solely on the *order of exactness* of multiresolution $(V_j)_{j \geq 0}$.

Definition 4.6.3. *The multiresolution* $(V_j)_{j \geq 0}$ *is said to have* polynomial exactness *of order* m*, if and only if*

$$\mathbb{P}_m \subset V_j, \tag{4.6.26}$$

for all $j \geq 0$*.*

We can then establish a *Jackson-type-inequality.*

Theorem 4.6.2. *Let* $(V_j)_{j\geq 0}$ *be a multiresolution approximation on a domain* $\Omega \subset \mathbb{R}^d$ *such that the spaces* V_j *have polynomial exactness of order* m. *Then for all* $1 \leq p \leq \infty$,

$$\inf_{g\in V_j} \|f-g\|_{L^p(\Omega)} \leq C\omega_m(f, 2^{-j})_p, \quad f \in L^p(\Omega), \tag{4.6.27}$$

for all $j \geq 0$, *where the constant* C *depends on* m, p, *the constants in* (4.6.8) *and* (4.6.11), *the supremum in* (4.6.7), *and on the uniform shape parameter for the sets* $S_\gamma, R_\gamma, \gamma \in \Gamma$.

Proof. Asuume that $p < \infty$ (the case $p = \infty$ is even simpler). Then for any $\gamma \in \Gamma_j$ one has for any polynomial g over S_γ

$$\begin{aligned}\|f-P_jf\|^p_{L_p(S_\gamma)} &\lesssim \|f-g\|^p_{L_p(S_\gamma)} + \|g-P_jf\|^p_{L_p(S_\gamma)}\\ &\overset{(4.6.26)}{=} \|f-g\|^p_{L_p(S_\gamma)} + \|P_j(g-f)\|^p_{L_p(S_\gamma)}\\ &\overset{(4.6.20)}{\leq} (1+C)\|f-g\|^p_{L_p(\hat{S}_\gamma)},\end{aligned}$$

where

$$\hat{S}_\gamma := \bigcup\{S_\mu : \mu \in \Gamma_j, |S_\gamma \cap S_\mu| > 0\}.$$

By (4.6.7) and (4.6.6) one has $\operatorname{diam} \hat{S}_\gamma \lesssim 2^{-j}$. By Whitney's Theorem (see (2.3.12) and (2.3.13)), we have

$$\|f-P_jf\|^p_{L_p(S_\gamma)} \lesssim \omega_m(f, \hat{S}_\gamma)_p \lesssim \omega_m(f, 2^{-j}, \hat{S}_\gamma)_p, \tag{4.6.28}$$

uniformly in j. Now use that only a uniformly bounded finite number of the $\hat{S}_\gamma$ overlap at a given location and the fact that the modulus of smoothness is equivalent to its averaged version (2.4.9), which is subadditive, summing over the local esitmates $\gamma \in \Gamma_j$ yields (4.6.27). □

Corollary 4.6.2. *If* (4.6.26) *holds one has for* $0 < s < m$, m *from* (4.6.26)

$$B^s_q(L_p(\Omega)) \subseteq \mathcal{A}^{s/d}_q((\Sigma_n); L_p(\Omega)) \tag{4.6.29}$$

in the sense of a continuous embedding.

Proof. Noting that

$$\|(P_{j+1}-P_j)f\|_{L_p(\Omega)} \leq \|P_{j+1}f - f\|_{L_p(\Omega)} + \|f-P_jf\|_{L_p(\Omega)} \lesssim \omega_m(f, 2^{-j}, \Omega)_p,$$

and substituting these bounds into the right hand side of (4.6.24), (4.6.25) and using Remark 2.3.6, 5., shows that

$$\|f\|_{\mathcal{A}^{s/d}_q} \lesssim \|f\|_{B^s_q(L_p(\Omega))}, \tag{4.6.30}$$

which confirms the claim. □

Thus polynomial exactness of order m implies that Besov spaces of smoothness $s < m$ are contained in the approximation spaces.

We will see next that one even has *equality* in (4.6.29) provided that the multiresolution $(V_j)_{j\in\mathbb{N}_0}$ satisfies in addition to the Jackson-type estimates also certain *Bernstein-type estimates.*

Bernstein type inequality:

Definition 4.6.4. *The multiresolution* $(V_j)_{j\geq 0}$ *(respectively the generating scaling functions) is said to have* smoothness of order s_p *in* $L_p(\Omega)$ *if and only if there exists an integer* $\bar{m} \geq 1$ *and a constant* C *such that*

$$\|\Delta_h^{\bar{m}}\varphi_\gamma\|_{L_p(\Omega_{nh})} \leq C\min\{1, 2^{|\gamma|}|h|\}^{s_p} 2^{|\gamma|(\frac{d}{2}-\frac{d}{p})}, \quad h \in \mathbb{R}^d, \quad \gamma \in \Gamma. \tag{4.6.31}$$

Exercise 4.6.5. *Let* $p = 2$, $\varphi_{j,k}(x) := 2^{j/2}\chi_{[0,1)}(2^j x - k)$. *What is the largest* s_2 *in this case?*

We begin with a Bernstein type inequality which controls "difference" quotions in L_p by the L_p norm for functions in V_j (similar to the inverse estimates encountered earlier).

Theorem 4.6.3. *Let* $(V_j)_{j\geq 0}$ *be a multiresolution approximation on a domain* $\Omega \subset \mathbb{R}^d$ *and let* $0 < p \leq \infty$. *If the primal scaling functions satisfy the smoothness condition* (4.6.31) *for some* $s_p > 0$ *and* $\bar{m} > s_p$, *then*

$$\omega_{\bar{m}}(g, t)_p \leq C\Big(\min\{1, 2^j t\}\Big)^{s_p} \|g\|_{L_p(\Omega)}, \quad g \in V_j, \tag{4.6.32}$$

for all $t > 0$ *and* $j \geq 0$, *where* C *depends on* $\bar{m}$, d, p, *on the supremum in* (4.6.7), *and on the constants in* (4.6.31) *and* (4.6.11).

Proof. We combine (4.6.31) with (4.6.10) applied to $u_\gamma = \Delta_h^{\bar{m}}\varphi_\gamma$. This yields, for any sequence $\mathbf{c} = (c_\gamma)_{\gamma\in\Gamma_j}$ and $g = \sum_{\gamma\in\Gamma_j} c_\gamma\varphi_\gamma \in V_j$,

$$\|\Delta_h^{\bar{m}} g\|_{L_p(\Omega)} = \|\sum_{\gamma\in\Gamma_j} c_\gamma \Delta_h^{\bar{m}}\varphi_\gamma\|_{L_p(\Omega)} \leq C\min\{1, 2^{|\gamma|}|h|\}^{s_p} 2^{|\gamma|(\frac{d}{2}-\frac{d}{p})}\|\mathbf{c}\|_{\ell_p}, \tag{4.6.33}$$

where C depends on $\bar{m}$, d, p, the supremum in (4.6.7), and the constant in (4.6.31). It follows that

$$\omega_{\bar{m}}(g, t)_p \leq C\Big(\min\{1, 2^j t\}\Big)^{s_p} 2^{j(\frac{d}{2}-\frac{d}{p})}\|\mathbf{c}\|_{\ell_p}, \tag{4.6.34}$$

which, combined with the lower bound in (4.6.11) yields (4.6.32). □

Remark 4.6.4. *One easily derives from the definition of the Besov-seminorms (see Definition 2.3.1) the Bernstein type inequality*

$$|g|_{B_\infty^{s_p}(L_p(\Omega))} \leq C 2^{s_p j}\|g\|_{L_p(\Omega)}, \quad g \in V_j. \tag{4.6.35}$$

Theorem 4.6.4. *Let* $(V_j)_{j\geq 0}$ *be a multiresolution approximation on a domain* $\Omega \subset \mathbb{R}^d$ *and let* $0 < q \leq \infty$, $1 \leq p \leq \infty$. *Assume that the spaces* V_j *have polynomial exactness of order* m, *and that the primal scaling functions satisfy the smoothness condition* (4.6.31) *for some* $s_p > 0$ *and* $\bar{m} > s_p$. *If in addition* $0 < s < \min\{s_p, m\}$, *one has*

$$B_q^s(L_p(\Omega)) = \mathcal{A}_q^{s/d}((V_j); L_p(\Omega)), \tag{4.6.36}$$

with equivalent quasi-norms. The embedding and equivalence constants depend on p, q, s, s_p, m, *and on the constants in* (4.6.32) *and* (4.6.27). *In particular, one has for the above range of parameters*

$$\|f\|_{B^s_q(L_p(\Omega))} \sim \|P_0 f\|_{L_p(\Omega)} + \Big(\sum_{j=1}^{\infty} 2^{jsq} \|(P_{j+1} - P_j) f\|^q_{L_p(\Omega)}\Big)^{1/q}, \tag{4.6.37}$$

with the usuaql interpretation when $q = \infty$.

Proof. In view of Corollaries 4.6.1, 4.6.2 and (4.6.30), we need only to show

$$\|f\|_{B^s_q(L_p(\Omega))} \lesssim \|f\|_{\mathcal{A}^{s/d}_q}. \tag{4.6.38}$$

To see this,

$$\begin{aligned} \omega_{\bar{m}}(f, 2^{-j})_p &\le \omega_{\bar{m}}(f - P_j f, 2^{-j})_p + \omega_{\bar{m}}(P_0 f, 2^{-j})_p + \textstyle\sum_{l=1}^{j} \omega_{\bar{m}}((P_l - P_{l-1}) f, 2^{-j})_p \\ &\lesssim \|f\|_{L_p(\Omega)} + 2^{-js_p} \textstyle\sum_{l=0}^{j} 2^{ls_p} \sigma_{a2^{dl}}(f)_{L_p(\Omega)}, \end{aligned}$$

where the second inequality uses the Bernstein type inequality (4.6.32). By application of the first Hardy's inequality (4.5.6), we conclude that for any $0 < s < s_p$,

$$\|(2^{sj} \omega_{\bar{m}}(f, 2^{-j})_p)_{j \ge 0}\|_{\ell_q} \lesssim \|f\|_{L_p(\Omega)} + |f|_{\mathcal{A}^{s/d}_q}, \tag{4.6.39}$$

which gives the continuous embedding (4.6.30). □

Remark 4.6.5. *Analogous statements can be proved also for* $0 < p < 1$ *but with considerably more work.*

4.6.2 Wavelet-Based Characterization

The spaces

$$W_j := (P_{j+1} - P_j) V_{j+1} \tag{4.6.40}$$

clearly satisfy

$$V_{j+1} = V_j + W_j.$$

suppose that we have found for each level j a basis

$$\Psi^j := \{\psi_\lambda : \lambda \in \Lambda_j\}$$

which is L_p-stable in the sense of (4.6.11), i.e.,

$$2^{j\left(\frac{d}{2} - \frac{d}{p}\right)} \|(d_\lambda)_{\lambda \in \Lambda_j}\|_{\ell_p(\Lambda_j)} \sim \Big\| \sum_{\lambda \in \Lambda_j} d_\lambda \psi_\lambda \Big\|_{L_p(\Omega)}, \quad w \in W_j, \tag{4.6.41}$$

with uniform constants. As before this means that we use the normalization

$$\|\psi_\lambda\|_{L_2(\Omega)} = 1, \quad \lambda \in \Lambda. \tag{4.6.42}$$

Then, given f each fluctuation $(P_{j+1} - P_j)f \in W_j$ has a unique representation

$$(P_{j+1} - P_j)f = \sum_{\lambda \in \Lambda_j} d_\lambda(f)\psi_\lambda. \tag{4.6.43}$$

Inserting this in (4.6.37) yields the norm equivalence

$$\|f\|_{B^s_q(L_p(\Omega))} \sim \|P_0 f\|_{L_p(\Omega)} + \Big(\sum_{j=1}^{\infty} 2^{jq\left(s + \frac{d}{2} - \frac{d}{p}\right)} \|(d_\lambda(f))_{\lambda \in \Lambda_j}\|^q_{\ell_p(\Lambda_j)} \Big)^{1/q}, \tag{4.6.44}$$

characterizing elements in $B^s_q(L_p(\Omega))$ by the fact that the coefficient sequences $(d_\lambda(f))_{\lambda \in \Lambda}$ of its elements belong to a certain weighted sequence space.

Remark 4.6.6. *We highlight two important cases:*

1. *Let* $p = q = 2$*: then* (4.6.44) *simplifies to*

$$\|f\|_{H^s(\Omega)} \sim \|f\|_{B^2_2(L_2(\Omega))} \sim |d_0(f)| + \Big(\sum_{\lambda \in \Lambda} 2^{2s|\lambda|} |d_\lambda(f)|^2 \Big)^{1/2}, \tag{4.6.45}$$

where we have used that $B^s_2(L_2) = H^s$, $s \geq 0$ *with equivalent norms, see Remark 2.3.6, 4.*

2. *Analogous statements hold when* $(V_j)_{j \in \mathbb{N}_0}$ *is a multiresolution for some closed subspace* H^s *of* $H^s(\Omega)$ *such as* $H^s_0(\Omega)$ *defined as the closure of* $C^\infty_0(\Omega)$ *with respect to* $\| \cdot \|_{H^s(\Omega)}$.

3. *Let*

$$\frac{1}{\tau} = \frac{s}{d} + \frac{1}{2}. \tag{4.6.46}$$

Then (4.6.44) *simplifies to*

$$\|f\|_{B^s_\tau(L_\tau(\Omega))} \sim \|\mathbf{d}(\mathbf{f})\|_{\ell_\tau(\Lambda)}, \tag{4.6.47}$$

i.e., the Besov spaces on the Sobolev-embedding line for L_2 *are equivalent to* ℓ_τ*-sequence spaces.*

4. *When normalizing the* ψ_λ *in* L_p *instead of* L_2 *the analogous relation to* (4.6.47) *holds with 2 replaced by* p.

4.6.3 Riesz Bases for L_2 and H^s

We emphasize that all these characterizations are valid for $s > 0$. A natural (and as will be seen important) question is: what happens when $s = 0$. Specifically, under which circumstances does (4.6.43) lead to a Riesz-basis for $L_2(\Omega)$?

Remark 4.6.7. *A first simple affirmative answer of the last question is the case* $\tilde{\varphi}_\gamma = \varphi_\gamma$*, i.e., the dual scaling functions equals the primal one which means that the* P_j *are* orthogonal *projectors. Then, one easily checks that* $P_{j+1} - P_j$ *are also orthogonal projectors, so that in this case*

$$\|f\|^2_{L_2(\Omega)} = \sum_{j=0}^{\infty} \|(P_j - P_{j-1})f\|^2_{L_2(\Omega)}.$$

The Haar basis is the simplest example for this situation.

As indicated before orthonormality is hard to realize practically basis functions which simultaneously are orthonormal, compactly supported and have a higher non-trivial smoothness.

For the construction of non-orthonormal bases which are Riesz-bases a so called dual multiresolution plays a crucial role. The reason lies in the following facts:

Exercise 4.6.6. *Let* $V_j \subset \mathcal{H}$ *be a nested sequence of closed subspaces with associated projectors* $Q_j : \mathcal{H} \to V_j$*. Show that the following properties are equivalent:*

1. *The projectors commute*
$$Q_j Q_k = Q_k Q_j, \quad j, k \in \mathbb{N}_0. \tag{4.6.48}$$

2. *The operators* $Q_{j+1} - Q_j$ *are also projectors.*

3. *The ranges* $\tilde{V}_j :=$ range Q_j^* *are also nested*
$$\tilde{V}_j \subset \tilde{V}_{j+1}, \quad j \in \mathbb{N}_0, \tag{4.6.49}$$
where Q_j^* *are the adjoints of* Q_j*.*

Exercise 4.6.7. *As an immediate consequence, if one of the above properties hold one has*

$$(Q_{j+1} - Q_j)\big((Q_{k+1} - Q_k)f\big) = ((Q_{j+1} - Q_j)f)\delta_{j,k}, \quad j, k \in \mathbb{N}_0, \tag{4.6.50}$$

and therefore

$$\begin{aligned}\langle (Q_{j+1} - Q_j)f, (Q^*_{l+1} - Q^*_l)f)\rangle_{\mathcal{H}} &= \langle (Q_{j+1} - Q_j)f, (Q^*_{j+1} - Q^*_j)f)\rangle_{\mathcal{H}}\delta_{j,l}, \quad j, l \in \mathbb{N}_0 \\ &= \langle (Q_{j+1} - Q_j)f, f)\rangle_{\mathcal{H}}\delta_{j,l}.\end{aligned} \tag{4.6.51}$$

Hence (with $Q_{-1} = 0$)

$$\begin{aligned}\|f\|^2_{\mathcal{H}} &= \Big\langle \sum_{j\in\mathbb{N}_0} (Q_j - Q_{j-1})f, \sum_{l\in\mathbb{N}_0} (Q^*_l - Q^*_{l-1})f \Big\rangle \\ &= \sum_{j\in\mathbb{N}_0} \sum_{j\in\mathbb{N}_0} \langle (Q_j - Q_{j-1})f, (Q^*_j - Q^*_{j-1})f \rangle \\ &\leq \sum_{j\in\mathbb{N}_0} \|(Q_j - Q_{j-1})f\|_{\mathcal{H}} \|(Q^*_j - Q^*_{j-1})f\|_{\mathcal{H}} \\ &\leq \Big(\sum_{j\in\mathbb{N}_0} \|(Q_j - Q_{j-1})f\|^2_{\mathcal{H}} \Big)^{1/2} \Big(\sum_{j\in\mathbb{N}_0} \|(Q^*_j - Q^*_{j-1})f\|^2_{\mathcal{H}} \Big)^{1/2} \\ &= \big\| \big(\|(Q_j - Q_{j-1})f\|_{\mathcal{H}} \big)_{j\in\mathbb{N}_0} \big\|_{\ell_2} \big\| \big(\|(Q^*_j - Q^*_{j-1})f\|_{\mathcal{H}} \big)_{j\in\mathbb{N}_0} \big\|_{\ell_2}.\end{aligned} \tag{4.6.52}$$

Thus if we knew that either

$$\left\|\left(\|(Q_j - Q_{j-1})f\|_{\mathcal{H}}\right)_{j\in\mathbb{N}_0}\right\|_{\ell_2} \lesssim \|f\|_H \text{ or } \left\|\left(\|(Q_j^* - Q_{j-1}^*)f\|_{\mathcal{H}}\right)_{j\in\mathbb{N}_0}\right\|_{\ell_2} \lesssim \|f\|_{\mathcal{H}}, \tag{4.6.53}$$

we would conclude that

$$\|f\|_{\mathcal{H}} \sim \left\|\left(\|(Q_j - Q_{j-1})f\|_{\mathcal{H}}\right)_{j\in\mathbb{N}_0}\right\|_{\ell_2}, \quad f \in \mathcal{H}. \tag{4.6.54}$$

Unfortunately, just based on the properties 1. to 3. above, one cannot conclude (4.6.53). However, under additional information on the ranges V_j and $\tilde{V}_j$ of Q_j, Q_j^*, respectively, one can prove (4.6.54), see [22].

To relate these observations to the previous section, note that (4.6.51) can be viewed as a *biorthogonality relation*.

Remark 4.6.8. *Let* $p = 2$. *By* (4.6.16)*, the adjoint of* $P_j f = \sum_{\gamma\in\Gamma_j}\langle f, \tilde{\varphi}_\gamma\rangle\varphi_\gamma$ *is given by*

$$P_j^* f = \sum_{\gamma\in\Gamma_j}\langle f, \varphi_\gamma\rangle\tilde{\varphi}_\gamma. \tag{4.6.55}$$

Let

$$\tilde{V}_j := \text{range } P_j^* = \text{span}\{\tilde{\varphi}_\gamma : \gamma \in \Gamma_j\}. \tag{4.6.56}$$

If the spaces $\tilde{V}_j$ *are* **also** nested

$$\tilde{V}_j \subset \tilde{V}_{j+1} \tag{4.6.57}$$

then $(\tilde{V}_j)_{j\in\mathbb{N}_0}$ *satisfies all conditions of a multiresolution and is called* dual multiresolution*. Clearly,* (4.6.57) *is equivalent to the refinability of the* $\tilde{\varphi}_\gamma$*, i.e., one can write for* $\gamma \in \Gamma_j$

$$\tilde{\varphi}_\gamma = \sum_{\nu\in\Gamma_{j+1}} a_{\gamma,\nu}\tilde{\varphi}_\nu \tag{4.6.58}$$

for some $(a_{\gamma,\nu})_{\nu\in\Gamma_\nu} \in \ell_2(\Gamma_{j+1})$.

Remark 4.6.9. *When* (4.6.57) *(or equivalentyl* (4.6.58)*) hold all the statements in Exercise 4.6.6 apply to the projectors* $Q_j = P_j$. *Thus, defining*

$$W_j := (P_{j+1} - P_j)V_{j+1}, \quad \tilde{W}_j := (P_{j+1}^* - P_j^*)\tilde{V}_{j+1}, \tag{4.6.59}$$

(4.6.50) *implies the* biorthogonality conditions

$$W_j \perp \tilde{W}_k, \quad j \neq k, \quad \tilde{W}_j \perp V_k, \quad k \leq j. \tag{4.6.60}$$

It should be noted though that, given φ_γ, $\gamma \in \Gamma_j$, it is relatively easy to construct a stable dual system in the sense of (4.6.16). It is not easy at all to ensure that the dual system is also refinable, i.e., satisfies (4.6.58), see [30].

Theorem 4.6.5. *Assume that* $(V_j)_{j\in\mathbb{N}_0}, (\tilde{V}_j)_{j\in\mathbb{N}_0}$ *is a pair of dual multiresolutions for* $L_2(\Omega)$ *which have exactness orders* $m, \tilde{m}$*, respectively. Moreover assume that both multiresolutions have smoothness* (4.6.31) *of orders* $0 < s_2, \tilde{s}_2$*, respectively. Then, the following norm-equivalences hold:*

1. *For each* $0 \le s < \min\{s_2, m, \tilde{m}\}$ *there exist constants* $c_s \le C_s$ *such that (for* $\langle f, g\rangle := \int_\Omega fg dx$, $P_{-1} := 0$*)*

$$c_s \sum_{j=0}^{\infty} 2^{2sj} \|(P_j - P_{j-1})f\|^2_{L_2(\Omega)} \le \|f\|^2_{H^s(\Omega)} \le C_s \sum_{j=0}^{\infty} 2^{2sj} \|(P_j - P_{j-1})f\|^2_{L_2(\Omega)}, \quad (4.6.61)$$

   uniformly for $f \in H^s(\Omega)$.

2. *For each* $0 \le s < \min\{\tilde{s}_2, \tilde{m}, \bar{\tilde{m}}\}$ *there exist constants* $\tilde{c}_s \le \tilde{C}_s$ *such that*

$$\tilde{c}_s \sum_{j=0}^{\infty} 2^{2sj} \|(P^*_j - P^*_{j-1})f\|^2_{L_2(\Omega)} \le \|f\|^2_{H^s(\Omega)} \le \tilde{C}_s \sum_{j=0}^{\infty} 2^{2sj} \|(P^*_j - P^*_{j-1})f\|^2_{L_2(\Omega)}, \quad (4.6.62)$$

   uniformly for $f \in H^s(\Omega)$*, where* C_s/c_s, $\tilde{C}_s/\tilde{c}_s$ *tend to infinity when* $s \to s_2, \tilde{s}_2$*, respectively.*

In particular, one has (for $s = 0$*)*

$$\Big(\sum_{j=0}^{\infty} \|(P^*_j - P^*_{j-1})f\|^2_{L_2(\Omega)}\Big)^{1/2} \sim \Big(\sum_{j=0}^{\infty} \|(P_j - P_{j-1})f\|^2_{L_2(\Omega)}\Big)^{1/2} \sim \|f\|_{L_2(\Omega)}. \quad (4.6.63)$$

Proof. (Sketch) For $s > 0$ (4.6.61), (4.6.62) immediately follow from Theorem 4.6.4. To prove (4.6.63) we need the following lemma.

Lemma 4.6.1. *For* $s < \min\{s_2, \tilde{s}_2\}$ *one has*

$$\|g\|_{(H^s(\Omega))'} \sim \Big(\sum_{j=0}^{\infty} 2^{-2js} \|(P^*_j - P^*_{j-1})g\|^2_{L_2(\Omega)}\Big)^{1/2} \sim \Big(\sum_{j=0}^{\infty} 2^{-2js} \|(P_j - P_{j-1})g\|^2_{L_2(\Omega)}\Big)^{1/2}, \quad (4.6.64)$$

where the constants depend on the constants in (4.6.61), (4.6.62).

Proof of Lemma 4.6.1: Let us abbreviate

$$D_j f := (P_j - P_{j-1})f, \quad D^*_j f := (P^*_j - P^*_{j-1})f, \quad j \in \mathbb{N}_0.$$

Then

$$\begin{aligned}
|\langle f, g\rangle| &= \Big|\Big\langle \sum_{j=0}^{\infty} D_j f, \sum_{k=0}^{\infty} D^*_k g\Big\rangle\Big| \overset{(4.6.50)}{=} \Big|\sum_{j=0}^{\infty} \langle D_j f, D^*_j g\rangle\Big| \\
&= \Big|\sum_{j=0}^{\infty} \langle 2^{sj} D_j f, 2^{-sj} D^*_j g\rangle\Big| \\
&\le \sum_{j=0}^{\infty} 2^{sj} \|D_j f\|_{L_2(\Omega)} 2^{-sj} \|D^*_j g\|_{L_2(\Omega)} \\
&\le \Big(\sum_{j=0}^{\infty} 2^{2sj} \|D_j f\|^2_{L_2(\Omega)}\Big)^{1/2} \Big(\sum_{j=0}^{\infty} 2^{2sj} \|D^*_j g\|^2_{L_2(\Omega)}\Big)^{1/2}. \quad (4.6.65)
\end{aligned}$$

This shows that, in view of (4.6.61),

$$\|g\|_{(H^s(\Omega))'} \lesssim \Big(\sum_{j=0}^{\infty} 2^{-2sj} \|D_j^* g\|_{L_2(\Omega)}^2 \Big)^{1/2}. \tag{4.6.66}$$

Interchaning the roles of D_j and D_j^*, and using (4.6.62), shows that for $s < s_2, \tilde{s}_2$ also

$$\|g\|_{(H^s(\Omega))'} \lesssim \Big(\sum_{j=0}^{\infty} 2^{-2sj} \|D_j g\|_{L_2(\Omega)}^2 \Big)^{1/2}. \tag{4.6.67}$$

To prove a converse etsimate, note first that since $\varphi_\gamma, \tilde{\varphi}_\gamma \in H^s(\Omega)$, $\gamma \in \Gamma$. Thus, given $g \in (H^s(\Omega))'$, since $D_j g, D_j^* g$ are well defined, we consider

$$f_g := \sum_{j=0}^{\infty} 2^{-2sj} D_j g,$$

and claim that $f_g \in H^s(\Omega)$. In fact, again by (4.6.50) we have $D_k f_g = 2^{-2sk} D_k g$ and therefore

$$\begin{aligned} \|f_g\|_{H^s(\Omega)}^2 &\overset{(4.6.61)}{\sim} \sum_{k=0}^{\infty} 2^{2sk} \|D_k f_g\|_{L_2(\Omega)}^2 \\ &= \sum_{k=0}^{\infty} 2^{2sk} 2^{-4sk} \|D_k g\|_{L_2(\Omega)}^2 \\ &= \sum_{k=0}^{\infty} 2^{-2sk} \|D_k g\|_{L_2(\Omega)}^2. \end{aligned} \tag{4.6.68}$$

Therefore

$$\begin{aligned} \|g\|_{(H^s(\Omega))'} &= \sup_{f \in H^s(\Omega)} \frac{\langle f, g\rangle}{\|f\|_{H^s(\Omega)}} \geq \frac{\langle f_g, g\rangle}{\|f_g\|_{H^s(\Omega)}} = \frac{\Big\langle \sum_{j=0}^{\infty} 2^{-2sj} D_j g, g \Big\rangle}{\|f_g\|_{H^s(\Omega)}} \\ &= \frac{\sum_{j=0}^{\infty} 2^{-2sj} \|D_j g\|_{L_2(\Omega)}^2}{\|f_g\|_{H^s(\Omega)}} \\ &\overset{(4.6.68)}{\sim} \Big(\sum_{j=0}^{\infty} 2^{-2sj} \|D_j g\|_{L_2(\Omega)}^2 \Big)^{1/2}, \end{aligned} \tag{4.6.69}$$

which finishes the proof. □

We return to the proof of Theorem 4.6.5. It is known from interpolation theory ([5]) that

$$[(H^s(\Omega))', H^s(\Omega)]_{\frac{1}{2},2} = L_2(\Omega). \tag{4.6.70}$$

Likewise for the weighted sequence spaces ℓ_2^r (see also (4.5.9))

$$\|\mathbf{d}\|_{\ell_2^s} := \Big(\sum_{\mathbf{j=0}}^{\infty} \mathbf{2^{2sj}|d_j|^2}\Big)^{\mathbf{1/2}},$$

one has

$$[\ell_2^{-s}, \ell_2^s]_{\frac{1}{2},2} = \ell_2. \tag{4.6.71}$$

By the norm equivalences (4.6.61), (4.6.62) and (4.6.64), the claim (4.6.63) follows. □

The construction of wavelet-type Riesz basis consists now in constructing for each level j, j ≥ 0, bases

$$\Psi^j := \{\psi_\lambda : \lambda \in \Lambda^j\}, \quad \tilde{\Psi}^j := \{\tilde{\psi}_\lambda : \lambda \in \Lambda^j\},$$

for the complement spaces $W_j, \tilde{W}_j$ where

$$\Lambda^j := \{\lambda \in \Lambda : |\lambda| = j\}, \quad j \geq 0,$$

with the following properties:

$$\bar{c}\|\mathbf{d}_{\Lambda^j}\|_{\ell_2(\Lambda^j)} \leq \Big\|\sum_{\lambda\in\Lambda^j} \mathbf{d}_\lambda\psi_\lambda\Big\|_{\mathbf{L_2}(\Omega)} \leq \bar{\mathbf{C}}\|\mathbf{d}_{\Lambda^j}\|_{\ell_2(\Lambda^j)}, \tag{4.6.72}$$

(and analogously for $\tilde{\Psi}^j$,) as well as

$$\langle\psi_\lambda, \tilde{\psi}_\nu\rangle = \delta_{\lambda,\nu}. \tag{4.6.73}$$

Moreover, setting

$$\psi_\gamma = \varphi_\gamma, \ \tilde{\psi}_\gamma := \tilde{\varphi}_\gamma, \quad \gamma \in \Gamma_0 := \Lambda_{-1}, \tag{4.6.74}$$

(4.6.73) holds by definition. Moreover, the following facts follow immediately from (4.6.60).

Proposition 4.6.1. *The collections*

$$\Psi := \bigcup_{j\geq -1} \Psi^j, \quad \tilde{\Psi} := \bigcup_{j\geq -1} \tilde{\Psi}^j,$$

are biorthogonal Riesz bases *for* $L_2(\Omega)$.

Exercise 4.6.8. *Show that under the assumptions of Theorem 4.6.5*

$$\Psi_s := \{\psi_\lambda/\|\psi_\lambda\|_{H^s(\Omega)} : \lambda \in \Gamma_0 \cup \Lambda\} \tag{4.6.75}$$

is a Riesz basis for $H^s(\Omega)$ *(respectively some closed subspace of it) for*

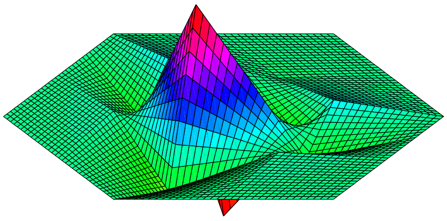
$$-\min\{\tilde{s}_2, \tilde{m}, \bar{\tilde{m}}\} < s < \min\{s_2, m, \bar{m}, \}.$$

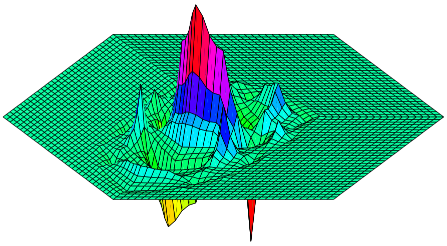
The statement remains true when $\psi_\lambda/\|\psi_\lambda\|_{H^s(\Omega)}$ *is replaced by* $2^{-s|\lambda|}\psi_\lambda$.

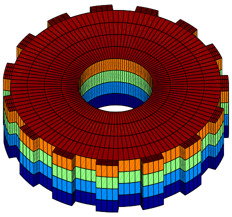
Comments 4.6.1. *1. Compare the pair $\Psi, \tilde{\Psi}$ with Exercise 4.2.2.*

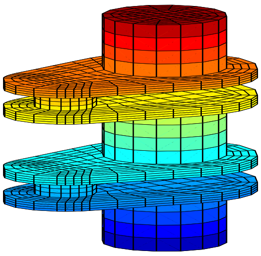
2. *Constructing a Riesz for $L_2(\Omega)$ is more "difficult" than for $H^s(\Omega)$, $s > 0$. The latter requires controling only the primal multiresolution.*

3. *There are several concrete constructions of wavelet Riesz bases on realistic domains, see [29] for multiresolutions generated by standard hierarchies of finite element spaces;*
   [24, 27] for wavelets on bounded intervals incorporating boundary conditions;
   [10, 11, 25] for wavelets on bounded domains using isoparametric mappings to hyper rectangles (see the figures below);
   [26] for general compact manifolds.
   (See [14, 12] for further reading).









5 Adaptive Methods: General Background

5.1 Preliminary Remarks

Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and a vector $\mathbf{b} \in \mathbb{R}^n$, *solvability* of a linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{5.1.1}$$

means in precise terms that for *any* $\mathbf{b} \in \mathbb{R}^n$ there exists a *unique* $\mathbf{x} \in \mathbb{R}^n$ satisfying (5.1.1). Another way to say this is that $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^n$ is a *bijection*, where the domain and the range of the "operator" $\mathbf{A}$ are the same.

Formally, a PDE such as the *Poisson equation* with homogeneous Dirichlet boundary conditions

$$-\Delta u = f \quad \text{in } \Omega, \quad u|_{\partial\Omega} = 0, \tag{5.1.2}$$

may also be regarded as such a linear operator equation. However, the meaning of solvability is now much less clear. It doesn't make any sense to say "for *any* right hand side" f which is now an element of some *infinitely dimensional* space $\mathbb{F}$, say. So a first question is what is a "suitable" choice for $\mathbb{F}$ – giving rise to the next question: what is the meaning of "suitable"?

Of course, one can avoid these uncomfortable questions by jumping directly to a discretization of (5.1.2) and worrying only about the invertibility of the resulting matrix. However, as we know, this is only seemingly a solution because what happens when the underlying mesh size tends to zero? For that reason one formulates the concept of *well-posedness*.

5.2 Well-Posed Problems

We concentrate first on *linear operator equations*: given a linear operator F and a right hand side f find u such that

$$Fu = f, \tag{5.2.1}$$

where we had $F = -\Delta$ in (5.1.2). Obviously, there is important information missing, namely for f in which normed linear space $\mathbb{F}$? Moreover, even if we provide such an $\mathbb{F}$ for which a unique solution u exists, any computational approach will only approximate u based on perturbed information of f (e.g. due to round-off). Therefore, the problem is only meaningful if we can quantify the effect of perturbing the data on the variation of the solution. The latter can only be quantified if we also choose a norm for measuring exactness of the solution, which means we also prescribe a normed linear space $\mathbb{X}$ where we seek the solution.

Definition 5.2.1. *The problem* (5.2.1) *is called well-posed, if for each* $f \in \mathbb{F}$ *there esists a unique* $u \in \mathbb{X}$ *satisfying* (5.2.1) *depending continuously on* f.

Recall that for linerar operators continuity is equivalent to boundedness, so continuous dependence of $u \in \mathbb{X}$ on the data $f \in \mathbb{F}$ just means that

$$\|u\|_{\mathbb{X}} \leq C\|f\|_{\mathbb{F}} \tag{5.2.2}$$

holds for some constant C independent of $f \in \mathbb{F}$. An equivalent way to say this is that

$$F \in \mathcal{L}(\mathbb{X}, \mathbb{F}), \quad F^{-1} \text{ exists in } \mathcal{L}(\mathbb{F}, \mathbb{X}). \tag{5.2.3}$$

Remark 5.2.1. *A first part of the problem is to actually identify a pair of spaces* $\mathbb{X}, \mathbb{F}$ *for which* infinite dimensional *problem is well-posed.*

Remark 5.2.2. *One cannot expect to develop numerical methods without accounting for the choice of* $\mathbb{X}, \mathbb{F}$*. This is particularly crucial when developing* adaptive *methods.*

Note that in general the spaces $\mathbb{X}$ and $\mathbb{F}$ are *different*. More precisely, this is the case when the operator F has *nonzero order* such as the Laplace operator lowering smoothness by the order two. (There are for instance integral operators of the second kind which map cartain spaces $\mathbb{X}$ onto themselves but for differential operators one has $\mathbb{X} \neq \mathbb{F}$ to guarantee well-posedness. This is in contrast to the finite dimensional case (5.1.1). Nevertheless this shows by the fact that for decreasing meshsizes the stiffness matrices for (5.1.2) are known to grow increasingly *ill-conditioned* which is the prize for ignoring the different metrics on $\mathbb{X}$ and $\mathbb{F}$ in a well-posed infinite dimensional formulation.

Remark 5.2.3. *The particular relevance of well-posedness for* adaptivity *can be seen as follows: suppose* $u_H \in \mathbb{X}_H \subset \mathbb{X}$ *is an approximation to the solution* u *of* $Fu = f$*. An adaptive method would try to determine next a* refined *trial space* $\mathbb{X}_h \supset \mathbb{X}_H$ *that reduces the error at the expense of possibly few additional degrees of freedom. Unfortunately, since* u *is unknown, one cannot extract corresponding information directly from the error* $u - u_H$*. However,* u_H *being now given, the* residual

$$f - Fu_H = F(u - u_H)$$

is, in principle, known.

To exploit residuals for error estimation note the following: boundedness and bounded invertibility (5.2.2) give

$$\|F(u_H) - f\|_{\mathbb{F}} = \|F(u_H - u)\|_{\mathbb{F}} \leq \|F\|_{\mathcal{L}(\mathbb{X},\mathbb{F})} \|u_H - u\|_{\mathbb{X}}$$

which implies

$$\begin{aligned} \|F\|^{-1}_{\mathcal{L}(\mathbb{U},\mathbb{F})} \|f - F(w_H)\|_{\mathbb{F}} \leq \|u - u_H\|_{\mathbb{X}} &= \left\|F^{-1}(F(u - u_H))\right\|_{\mathbb{X}} \\ &\leq \left\|F^{-1}\right\|_{\mathcal{L}(\mathbb{F},\mathbb{U})} \|F(u_H) - f\|_{\mathbb{F}} , \end{aligned}$$

and hence

$$(\|F\|_{\mathcal{L}(\mathbb{X},\mathbb{F})})^{-1}\|f - Fu_H\|_{\mathbb{F}} \le \|u - u_H\|_{\mathbb{X}} \le \|F^{-1}\|_{\mathcal{L}(\mathbb{F},\mathbb{X})}\|f - Fu_H\|_{\mathbb{F}}. \tag{5.2.4}$$

Thus, the smaller the *condition number*

$$\kappa_{\mathbb{X},\mathbb{F}}(F) := \|F\|_{\mathcal{L}(\mathbb{X},\mathbb{F})}\|F^{-1}\|_{\mathcal{L}(\mathbb{F},\mathbb{X})} \tag{5.2.5}$$

of $F \in \mathcal{L}(\mathbb{X}, \mathbb{F})$ is, the tighter is the error $\|u - u_H\|_{\mathbb{X}}$ estimated by the residual $\|f - Fu_H\|_{\mathbb{F}}$ in $\mathbb{F}$. Again, an improper choice of $\mathbb{X}, \mathbb{F}$ may entail large or infinite condition numbers.

Remark 5.2.4. *Essentially all rigorously founded adaptive methods are based on* a posteriori *bounds for the residuals $\|f - Fu_H\|_{\mathbb{F}}$. Deriving such bounds may be nontrivial and the subsequent discussions address scenarios where this can be done. More precisely, we will discuss two formally different frameworks which, however, use in essence this same paradigm.*

5.3 Stable Variational Formulations

Remark 5.3.1. *For the residual to estimate the error well it is important that $\kappa_{\mathbb{X},\mathbb{F}}(F)$ be small – quantified well-posedness. In this case we talk about* stability, *although it is usually hard to prescribe a concrete bound for $\kappa_{\mathbb{X},\mathbb{F}}(F)$ to mathematically* define *stability. The notion of stability is therefore more of a guideline.*

We address now the issue of identifying "good pairs" of spaces $\mathbb{X}, \mathbb{F}$ in the above sense. A very powerful strategy for this task is to contrive a *variational formulation* of a (linear or nonlinear) operator vequation

$$F(u) = f \tag{5.3.1}$$

To explain this, assume first, as before, that F is a linear operator. Look for a Hilbert space $\mathbb{U}$ (which is to host the solution – "trial space") and a possibly different Hilbert space $\mathbb{V}$ ("test space") so that for all $w \in \mathbb{U}$, $F(v)$ is a bounded linear functional on $\mathbb{V}$, i.e., $F(v) \in \mathbb{V}'$, the normed dual of $\mathbb{V}$. Hence the *dual pairing*

$$\langle F(w), v\rangle = (F(w))(v)$$

is well defined for any $w \in \mathbb{U}, v \in \mathbb{V}$.

Remark 5.3.2. $\langle z, v\rangle = z(v)$ *does not stand for a scalar product but for the action of a functional $z \in \mathbb{V}'$ on $v \in \mathbb{V}$. This notation is motivated by the fact that in many cases (not always!) this action can be expressed by an ℓ_2-inner product. The scalar products on $\mathbb{U}, \mathbb{V}$ are denoted by $\langle\cdot,\cdot\rangle_{\mathbb{U}}, \langle\cdot,\cdot\rangle_{\mathbb{V}}$.*

Recall that $\mathbb{V}'$ is endowed with the norm $\|z\|_{\mathbb{V}'} := \sup_{v\in\mathbb{V}} \frac{\langle z,v\rangle}{\|v\|_{\mathbb{V}}}$ so that

$$\|F(v)\|_{\mathbb{V}'} = \sup_{v\in\mathbb{V}} \frac{\langle F(w), v\rangle}{\|v\|_{\mathbb{V}}} < \infty\,.$$

In particular,

$$A(v,w) := \langle F(w), v\rangle, \qquad (w,v) \in \mathbb{U}\times\mathbb{V}, \tag{5.3.2}$$

is then a well-defined bilinear form on $\mathbb{U}\times\mathbb{V}$. In this case $F \in \mathcal{L}(\mathbb{U},\mathbb{V}')$, i.e., the weak formulation of the problem suggests taking $\mathbb{F} = \mathbb{V}'$.

One can also look at it the other way around, taking the bilinear form as the starting point. Then $F : F\mathbb{U} \to \mathbb{V}'$ is induced by (5.3.2) by the Riesz-Representation Theorem.

Definition 5.3.1. *The problem: Given* $f \in \mathbb{V}'$, *find* $u \in \mathbb{U}$ *so that*

$$A(u,v) = \langle f, v\rangle \qquad \forall\, v \in \mathbb{V}, \tag{5.3.3}$$

is called a $(\mathbb{U},\mathbb{V})$-**stable variational formulation** *of* (5.3.1) *if* F, *defined by* (5.3.2), *is an isomorphism from* $\mathbb{U}$ *onto* $\mathbb{V}'$. *In this case* (5.2.5) *takes the form*

$$\kappa_{\mathbb{U},\mathbb{V}'}(F) := \|F\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')}\, \|F^{-1}\|_{\mathcal{L}(\mathbb{V}',\mathbb{U})} < \infty\,. \tag{5.3.4}$$

Clearly, (5.2.2) becomes now

$$\|u\|_{\mathbb{U}} = \|F^{-1}f\|_{\mathbb{U}} \le \|F^{-1}\|_{\mathcal{L}(\mathbb{V}',\mathbb{U})}\|f\|_{\mathbb{V}'}, \tag{5.3.5}$$

i.e., $C = \|F^{-1}\|_{\mathcal{L}(\mathbb{V}',\mathbb{U})}$ in (5.2.2).

5.4 Babuška-Nečas-Theory

There is a well known criterion for stability, due to Babuška-Nečas. (see Numa IV).

Theorem 5.4.1. (5.3.4) *holds if the following is true:*

(i) **continuity** *of* $A(\cdot,\cdot) : \mathbb{U}\times\mathbb{V} \to \mathbb{R}$: $\exists\, C_a < \infty$ *such that*

$$|A(w,v)| \le C_a\, \|w\|_{\mathbb{U}}\, \|v\|_{\mathbb{V}}\,, \qquad (w,v) \in \mathbb{U}\times\mathbb{V}\,. \tag{5.4.1}$$

(ii) **inf-sup-condition**: $\exists\, c_\alpha > 0$ *such that*

$$\inf_{v\in\mathbb{V}} \sup_{w\in\mathbb{U}} \frac{A(w,v)}{\|w\|_{\mathbb{U}}\, \|v\|_{\mathbb{V}}} \ge c_\alpha\,. \tag{5.4.2}$$

(iii) For all $w \in \mathbb{U}$ there exists $v \in \mathbb{V}$ such that $A(w, v) \neq 0$.

Corollary 5.4.1. *If the bilinear form $A(\cdot,\cdot)$ is continuous on $\mathbb{U} \times \mathbb{U}$, i.e.,* (5.4.1) *holds for $\mathbb{V} = \mathbb{U}$, and if $A(\cdot,\cdot)$ is in addition* coercive *on $\mathbb{U}$, i.e.,*

$$A(w, w) \geq c_\alpha \|w\|_{\mathbb{U}}^2, \quad w \in \mathbb{U}, \tag{5.4.3}$$

then the operator F induced by $A(\cdot,\cdot)$ through $\langle Fw, v\rangle = A(w, v)$, $w, v \in \mathbb{U}$, satisfies (5.4.7)*,* (5.4.8) *and hence* (5.4.9)*. Such problems are briefly called* $\mathbb{U}$-elliptic.

To illustrate these facts let us consider again the simplest possible example.

Example 5.4.1. *Return to the Poisson problem* (5.1.2)

$$\begin{aligned} F(u) = -\Delta u &= f \quad \text{in } \Omega\,, \\ u &= 0 \quad \text{on } \partial\Omega\,. \end{aligned} \tag{5.4.4}$$

Then

$$a(v, w) = \int_\Omega (-\Delta v) w \, dx = \int_\Omega \nabla v \cdot \nabla w \, dx \qquad \forall\, v, w \in C_0^\infty(\Omega)$$

is (by closure) well defined on $H_0^1(\Omega) \times H_0^1(\Omega)$. So here $\mathbb{U} = \mathbb{V} = H_0^1(\Omega)$, where

$$\begin{aligned} H_0^1(\Omega) &= \overline{C_0^\infty(\Omega)}^{H^1}\,, \\ \|v\|_{H^1(\Omega)}^2 &= \|v\|_{L_2(\Omega)}^2 + \|\nabla v\|_{L_2(\Omega)}^2\,. \end{aligned}$$

By Cauchy-Schwarz the bilinear form $a(\cdot,\cdot)$ is continuous,

$$|a(v, w)| \leq \|v\|_{H^1}\,\|w\|_{H^1}\,. \qquad (\rightsquigarrow C_a = 1) \tag{5.4.5}$$

Moreover, by the Poincaré inequality there exists $\alpha > 0$ such that

$$a(v, v) \geq c_\alpha\,\|v\|_{H^1}^2\,. \tag{5.4.6}$$

The estimates (5.4.5) *and* (5.4.6) *imply conditions (i), (ii) and (iii) in Theorem 5.4.1.*

To see that Theorem 5.4.1 implies (5.3.4) note the boundedness of F as a mapping into $\mathbb{V}'$ follows from the boundedness of the bilinear form (5.4.1):

$$\|F(w)\|_{\mathbb{V}'} = \sup_{v \in \mathbb{V}} \frac{\langle F(w), v\rangle}{\|v\|_{\mathbb{V}}} = \sup_{v \in \mathbb{V}} \frac{A(w, v)}{\|v\|_{\mathbb{V}}} \overset{(5.4.1)}{\leq} \sup_{v \in \mathbb{V}} \frac{C_a\,\|w\|_{\mathbb{U}}\,\|v\|_{\mathbb{U}}}{\|v\|_{\mathbb{V}}} = C_a\,\|w\|_{\mathbb{U}}\,,$$

which yields

$$\|F\|_{\mathcal{L}(\mathbb{U},\mathbb{V}')} \leq C_a\,. \tag{5.4.7}$$

On the other hand, by (5.4.2), we have

$$\|F(w)\|_{\mathbb{V}'} = \sup_{v\in\mathbb{V}} \frac{\langle F(w), v\rangle}{\|v\|_{\mathbb{V}}} \geq c_\alpha \|w\|_{\mathbb{U}}.$$

This implies

$$\|F^{-1}\|_{\mathcal{L}(\mathbb{V}',\mathbb{U})} \leq \frac{1}{c_\alpha}, \tag{5.4.8}$$

which gives indeed

$$\kappa_{\mathbb{U},\mathbb{V}'}(F) \leq \frac{C_a}{c_\alpha}. \tag{5.4.9}$$

Note that (5.2.4) becomes

$$\|F\|^{-1}_{\mathcal{L}(\mathbb{U},\mathbb{V}')} \|f - F(u_H)\|_{\mathbb{V}'} \leq \|u - u_H\|_{\mathbb{U}} \leq \left\|F^{-1}\right\|_{\mathcal{L}(\mathbb{V}',\mathbb{U})} \|f - F(u_H)\|_{\mathbb{V}'}. \tag{5.4.10}$$

Remark 5.4.1. *The spaces $\mathbb{F} = \mathbb{V}'$ hosting the data are in such a variational framework typically* dual *spaces. Dual norms are usually not easy to evaluate. Thus, an adaptive scheme based on using a posteriori information in terms of* residuals in such a dual norm, *has to address this issue and we shall learn how to deal with this in at least two different ways.*

5.5 Nonlinear Problems

Idea: A problem is called stable if its linearizations are stable:
Find again X, Y such that $F : \mathbb{U} \to \mathbb{V}'$ is well defined by

$$\langle F(u), v\rangle = \langle f, v\rangle, \qquad v \in \mathbb{V}. \tag{5.5.1}$$

Motivation: $R : \mathbb{R} \to \mathbb{R}$, search for x with $R(x) = 0$. Newton's method converges well if the tangent has a nonvanishing slope, i.e., the Jacobian is nonsingular.

The same here: Consider

$$R(u) := F(u) - f.$$

What does it mean to say: The "derivative" does not vanish near the solution?

Consider the **Fréchet derivative** of R at v defined by

$$\begin{aligned}\langle DR(v)w, z\rangle &:= \lim_{t\to 0} \frac{1}{t}\langle R(v + tw) - R(v), z\rangle \\ &= \lim_{t\to 0} \frac{1}{t}\langle F(v + tw) - F(v), z\rangle \\ &= \langle DF(v)w, z\rangle \qquad \forall z \in \mathbb{V}.\end{aligned}$$

The Fréchet derivative is (for each v) a linear mapping from $\mathbb{U}$ to $\mathbb{V}'$. We call (5.5.1) **stable** in the neighborhood $\mathcal{N}$ of a solution u if $DF(w) : \mathbb{U} \to \mathbb{V}'$ is an isomorphism for each $w \in \mathcal{N}$, i.e., for each $w \in \mathcal{N}$ there exist c_w, C_w such that

$$c_w \|z\|_{\mathbb{U}} \leq \|DF(w)z\|_{\mathbb{V}'} \leq C_w \|z\|_{\mathbb{U}} \qquad \forall z \in \mathbb{U}\,. \tag{5.5.2}$$

5.6 Examples of Stable Variational Formulations

We collect the relevant facts without proof.

5.6.1 Saddle Point Problems

Poisson's equation gives rise to a coercive bilinear form. Its weak formulation can be understood as first order conditions for the minimization of the "energy" functional

$$J(v) := \frac{1}{2}a(v,u) - \langle f, v\rangle\,.$$

If such a minimization problem is subjected to **constraints** one obtains a **saddle point problem**, losing coercivity! In the context of constrained optimization saddle point problems arise under the name of *Karush-Kuhn-Tucker conditions*.

The weak formulation of saddle point problems: given Hilbert spaces $\mathbb{X}, \mathbb{M}$ and bilinear forms

$$a(\cdot,\cdot) : \mathbb{X} \times \mathbb{X} \to \mathbb{R}, \quad b(\cdot,\cdot) : \mathbb{X} \times \mathbb{M} \to \mathbb{R},$$

with data $[f, g] \in \mathbb{X}' \times \mathbb{M}'$, find $[u, p] \in \mathbb{X} \times \mathbb{M} =: \mathbb{U}$ such that

$$\begin{array}{lcll} a(u,v) + b(v,p) & = & \langle f, v\rangle & \forall v \in \mathbb{X}\,, \\ b(u,q) & = & \langle g, q\rangle & \forall q \in \mathbb{M}\,, \end{array} \tag{5.6.1}$$

which is equivalent to

$$\hat{A}([u,p],[v,q]) = \langle f, v\rangle + \langle g, q\rangle \qquad \forall [v,q] \in \mathbb{U} = \mathbb{X} \times \mathbb{M} = \mathbb{V}, \tag{5.6.1a}$$

where

$$\hat{A}(\cdot,\cdot) : (\mathbb{X} \times \mathbb{M}) \times (\mathbb{X} \times \mathbb{M}) \to \mathbb{R}\,,$$
$$\hat{A}([u,p],[v,q]) := a(u,v) + b(v,p) + b(u,q)\,.$$

Defining the linear operators $A : \mathbb{X} \to \mathbb{X}'$, $B : \mathbb{X} \to \mathbb{M}'$, $B^* : \mathbb{M} \to \mathbb{X}'$ by

$$\langle Av, w\rangle := a(v,w), \qquad \langle Bv, q\rangle := b(v,q) = \langle v, B^* q\rangle$$

(5.6.1) can be written as

$$F([u,p]) = \begin{pmatrix} A & B^* \\ B & 0 \end{pmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} =: \tilde{f} \in \mathbb{X}' \times \mathbb{M}' = \mathbb{U}'\,. \tag{5.6.1b}$$

Note that (5.6.1a) is a non-coercive, indefinite problem.

Theorem 5.6.1 ([9, 7])**.** (5.6.1) *has a unique solution* $[u,p]$ *if there exist constants* $C_a, C_b < \infty$, $\alpha, \beta > 0$, *such that*

$$|a(v,w)| \le C_a \|v\|_H \|w\|_{\mathbb{X}}, \quad |b(v,q)| \le C_b \|v\|_{\mathbb{X}} \|q\|_{\mathbb{M}} \qquad \forall v,w \in \mathbb{X}, q \in \mathbb{M}, \tag{5.6.2}$$

$$|a(v,v)| \ge c_\alpha \|v\|_{\mathbb{X}}^2 \qquad \forall v \in \mathbb{Y} = \{v \in \mathbb{X} : b(v,q) = 0 \ \forall q \in \mathbb{M}\} = \mathrm{Ker}(B), \tag{5.6.3}$$

$$\inf_{q\in\mathbb{M}} \sup_{v\in\mathbb{X}} \frac{b(v,q)}{\|v\|_{\mathbb{X}} \|q\|_{\mathbb{M}}} \ge c_\beta . \tag{5.6.4}$$

Moreover (5.6.1) *is stable, i.e.,*

$$\tilde{c} \, \|[v,q]\|_{\mathbb{U}} \le \|F([v,q])\|_{\mathbb{U}'} \le \tilde{C} \, \|[v,q]\|_{\mathbb{U}} \qquad \forall [v,q] \in \mathbb{U} = \mathbb{X} \times \mathbb{M} \tag{5.6.5}$$

where $\|[v,q]\|_{\mathbb{U}}^2 = \|v\|_{\mathbb{X}}^2 + \|q\|_{\mathbb{M}}^2$, $\tilde{c} = \tilde{c}(C_a, C_b, c_\alpha, c_\beta)$, $\tilde{C} = \tilde{C}(C_a, C_b, c_\alpha, c_\beta)$ *and* $\kappa_{\mathbb{U},\mathbb{U}'}(F) \le \frac{\tilde{C}}{\tilde{c}}$.

In fact, one can show that (5.6.2)–(5.6.4) implies the validity of the assumptions of Theorem 5.4.1 (see for example [7]).

Example 5.6.1. Stokes system*:*

$$\begin{aligned} -\Delta u + \nabla p &= f && \textit{in } \Omega \\ \operatorname{div} u &= 0 && \textit{in } \Omega \\ u &= 0 && \textit{on } \partial\Omega \end{aligned}$$

$$\begin{aligned} \rightsquigarrow \qquad & a(u,v) = \int_\Omega \nabla u : \nabla v \, dx, \quad b(v,q) = -\int \operatorname{div} v \; q \, dx \\ & \mathbb{X} = H_0^1(\Omega)^d, \quad \mathbb{M} = L_{2,0}(\Omega) = \{v \in L_2(\Omega) : \int_\Omega v \, dx = 0\} \\ & \mathbb{Y} = \{v \in H_0^1(\Omega)^d : \operatorname{div} v = 0 \textit{ weakly}\} \end{aligned}$$

One can show that the onditions of Theorem 5.6.1 are satisfied.

Example 5.6.2. Fictitious domain method *Let* $\Omega \subset \square$ *where* Ω *is complicated or moves and* $\square$ *is simple.*
Question: *How to deal with "moving" boundary conditions?*
Idea: *Find* $[u,p] \in \mathbb{U} := H^1(\square) \times H^{-\frac{1}{2}}(\Gamma := \partial\Omega)$ *such that*

$$\begin{aligned} \underbrace{\langle \nabla u, \nabla v \rangle}_{a(u,v)} + \underbrace{\langle \gamma v, p \rangle_\Gamma}_{b(v,q)} &= \langle f, v \rangle && \forall v \in H^1(\square), \\ \langle \gamma u, q \rangle_\Gamma &= \langle g, q \rangle_\Gamma && \forall q \in H^{-\frac{1}{2}}(\Gamma), \end{aligned} \tag{5.6.6}$$

where $\gamma : H^1(\Box) \to H^{\frac{1}{2}}(\Gamma)$ *is the trace map. It is essentially a consequence to the Trace Theorem, that a classical solution of* (5.6.6) *solves*

$$\begin{aligned} -\Delta u &= f \qquad \textit{in } \Omega \\ u &= g \qquad \textit{on } \partial\Omega \end{aligned}$$

and that (5.6.6) *satisfies the assumptions of Theorem 5.6.1.*

Example 5.6.3. Mixed formulation

$$\begin{aligned} -\operatorname{div}(a\nabla u) &= f \qquad \textit{in } \Omega \\ u &= 0 \qquad \textit{on } \partial\Omega \end{aligned}$$

Write this as a first order system:

$$\begin{aligned} a^{-1} q &= -\nabla u \\ q &:= -a\nabla u \end{aligned}$$

This gives rise to the variational formulation: Find $[q, u] \in \mathbb{U} := L_2(\Omega)^d \times H^1_0(\Omega)$ *such that*

$$\begin{aligned} \langle a^{-1}q, r\rangle + \langle r, \nabla u\rangle &= 0 \qquad \forall\, r \in L_2(\Omega)^d\,, \\ -\langle q, \nabla v\rangle &= \langle f, v\rangle \quad \forall\, v \in H^1_0(\Omega)\,. \end{aligned} \tag{5.6.7}$$

Here

$$\begin{aligned} a(q, r) &= \int_\Omega a^{-1} q \cdot r \, dx, \qquad & b(u, r) &= \int_\Omega r \cdot \nabla u \, dx \\ \mathbb{X} &= L_2(\Omega)^d, & \mathbb{M} &= H^1_0(\Omega) \end{aligned}$$

We claim that (5.6.7) *also satisfies the conditions in Theorem 5.6.1, see for example [7].*

5.6.2 Quasilinear Equations

Consider the nonlinear equation

$$\begin{aligned} -\Delta u + u^3 &= f \qquad \text{in } \Omega\,, \\ u &= 0 \qquad \text{on } \partial\Omega\,. \end{aligned} \tag{5.6.8}$$

Claim: For $d \leq 3$, $\mathbb{U} = \mathbb{V} = H^1_0(\Omega)$ is still appropriate. Show first that $F(v) \in \left(H^1_0(\Omega)\right)' = H^{-1}(\Omega)$ for $v \in H^1_0(\Omega)$ (so that $\langle F(v), w\rangle$ is defined for all $w \in H^1_0(\Omega)$).

To this end, note

$$\langle F(v), w\rangle = \langle \nabla v, \nabla w\rangle + \langle v^3, w\rangle\,.$$

We want to show that

$$|\langle F(v), w\rangle| \leq C(v)\, \|w\|_{H^1}\ .$$

Clearly $|\langle\nabla u, \nabla w\rangle| \leq \|v\|_{H^1} \|w\|_{H^1}$. Furthermore

$$\begin{aligned} \left|\langle v^3, w\rangle\right| &\leq \int_\Omega |v|^3 |w| \, dx = \int_\Omega |v|^2 |vw| \, dx \\ &\leq \left(\int_\Omega |v|^4 \, dx\right)^{\frac{1}{2}} \left(\int_\Omega |v|^2 |w|^2 \, dx\right)^{\frac{1}{2}} \\ &\leq \|v\|^2_{L_4(\Omega)} \|v\|_{L_4(\Omega)} \|w\|_{L_4(\Omega)} = \|v\|^3_{L_4(\Omega)} \|w\|_{L_4(\Omega)} \,. \end{aligned} \tag{5.6.9}$$

Here we invoke a Sobolev embedding theorem (see Figure 1). For $d \leq 3$ one has

$$H^1_0(\Omega) \subseteq L_4(\Omega)\,, \tag{5.6.10}$$

$$\text{that is } \|v\|_{L_4(\Omega)} \leq C \|v\|_{H^1(\Omega)} \, \forall v \in H^1_0(\Omega)\,.$$

Combining (5.6.9) with (5.6.10) yields

$$\left|\langle v^3, w\rangle\right| \leq C \|v\|^3_{H^1(\Omega)} \|w\|_{H^1(\Omega)} \,.$$

This gives

$$|\langle F(v), w\rangle| \leq \left(\|v\|_{H^1(\Omega)} + C \|v\|^3_{H^1(\Omega)}\right) \|w\|_{H^1(\Omega)} \tag{5.6.11}$$

which implies

$$F(v) \in H^{-1}(\Omega)\,.$$

To show stability, determine $DF(v)$: For $z \in H^1_0(\Omega)$

$$\begin{aligned} \langle F(v+tw) - F(v), z\rangle &= \langle \nabla(v+tw), \nabla z\rangle - \langle \nabla v, \nabla z\rangle + \langle (v+tw)^3, z\rangle - \langle v^3, z\rangle \\ &= t\langle \nabla w, \nabla z\rangle + 3t\langle v^2 w, z\rangle + 3t^2\langle v w^2, z\rangle + t^3\langle w^3, z\rangle \,. \end{aligned}$$

Hence

$$\frac{1}{t}\langle F(v+tw) - F(v), z\rangle \overset{t\to 0}{\to} \langle \nabla w, \nabla z\rangle + 3\langle v^2 w, z\rangle\,,$$

and thus

$$\langle DF(v)w, z\rangle = \langle \nabla w, \nabla z\rangle + 3\langle v^2 w, z\rangle\,. \tag{5.6.12}$$

Question: Is $DF(v)w \in H^{-1}(\Omega)$?
By the same argument as before we conclude that

$$|\langle DF(v)w, z\rangle| \leq \|w\|_{H^1(\Omega)} \|z\|_{H^1(\Omega)} + 3C \|v\|^2_{H^1(\Omega)} \|w\|_{H^1(\Omega)} \|z\|_{H^1(\Omega)} \,,$$

and therefore

$$\|DF(v)w\|_{H^{-1}(\Omega)} = \sup_{z \in H^1_0(\Omega)} \frac{\langle DF(v)w, z\rangle}{\|z\|_{H^1(\Omega)}} \leq \left(1 + 3C \|v\|^2_{H^1(\Omega)}\right) \|w\|_{H^1(\Omega)} \,, \tag{5.6.13}$$

that is, $DF(v) : H_0^1(\Omega) \to H^{-1}(\Omega)$ is bounded.
Moreover,

$$\langle DF(v)w, w\rangle = |w|^2_{H^1(\Omega)} + 3\int_\Omega v^2 w^2 \, dx \geq |w|^2_{H^1(\Omega)} \overset{\text{Poincaré}}{\geq} C_\Omega \|w\|^2_{H^1(\Omega)} ,$$

and thus

$$\|DF(v)w\|_{H^{-1}(\Omega)} = \sup_{z\in H_0^1(\Omega)} \frac{\langle DF(v)w, z\rangle}{\|z\|_{H^1(\Omega)}} \geq \frac{\langle DF(v)w, w\rangle}{\|w\|_{H^1(\Omega)}} \geq C_\Omega \|w\|_{H^1(\Omega)} .$$

Hence (5.5.2) holds with $c_v = C_\Omega$, $C_v = \left(1 + 3C \|v\|^2_{H^1(\Omega)}\right)$.

5.6.3 Boundary Integral Equations

For details on this topic see [40]. Motivation: Exterior domain problems.
Solve

$$\begin{aligned} -\Delta U &= 0 && \text{in } \Omega^c = \mathbb{R}^3\backslash\Omega \\ U &= f && \text{on } \partial\Omega =: \Gamma \\ U(x) &\to 0, && |x| \to \infty \end{aligned} \tag{5.6.14}$$

Obstruction: The domain is unbounded.

Idea: Transform the problem into an equivalent one living on Γ (**dimension reduction**).
Let

$$\mathcal{E}(x, y) := \frac{1}{4\pi |x - y|}$$

where $|x - y|$ is the Euclidean norm and define

$$(\mathcal{V}v)(x) := \int_\Gamma \mathcal{E}(x, y)v(y) \, d\Gamma_y, \qquad x \in \Gamma .$$

(singular integral operator / single layer potential operator)
Find u such that

$$\mathcal{V}u = f \qquad \text{on } \Gamma , \tag{5.6.15}$$

(integral equation of the first kind) – note that f lives on Γ. Then one can show that

$$U(x) := \int_\Gamma \mathcal{E}(x, y)u(y) \, d\Gamma_y, \qquad x \in \Omega^c$$

solves (5.6.14). Here

$$\begin{aligned} &a(u, v) = \langle \mathcal{V}u, v\rangle_\Gamma, \\ &\mathbb{U} = \mathbb{V} = H^{-\frac{1}{2}}(\Gamma), \quad \mathbb{U}' = \left(H^{-\frac{1}{2}}(\Gamma)\right)' = H^{\frac{1}{2}}(\Gamma) , \end{aligned}$$

so this is an operator of order -1.

Alternative: Rewrite (5.6.14) as integral equation of the second kind (double layer potential). Define

$$\begin{aligned}(\mathcal{K}v)(x) &:= \int_\Gamma \frac{\partial}{\partial n_y}\mathcal{E}(x,y)v(y)\,d\Gamma_y \\ &= \int_\Gamma \frac{1}{4\pi}\frac{n_y^T(x-y)}{|x-y|^3}v(y)\,d\Gamma_y, \qquad x \in \Gamma.\end{aligned}$$

Consider

$$\mathcal{L}u := \left(\frac{1}{2}+\mathcal{K}\right)u = f \qquad \text{on } \Gamma.$$

Then

$$U(x) := \int_\Gamma \mathcal{K}(x,y)u(y)\,d\Gamma_y, \qquad x \in \Omega^c$$

solves (5.6.14).

Here one has

$$A(u,v) = \langle u, \mathcal{L}v\rangle_\Gamma, \quad L_2(\Gamma) = \mathbb{U} = \mathbb{U}' = \mathbb{V}' = \mathbb{V}.$$

One can show that

$$\mathcal{L} : L_2(\Gamma) \to L_2(\Gamma)$$

is an isomorphism. $\mathcal{L}$ has order zero, hence linear systems are uniformly well conditioned independent of the discretization.

5.6.4 High-Dimensional Elliptic Problems

Parametric PDEs: are PDEs involving coefficients which depend on additional parameters. Such parameters could describe shapes and typically arise in design or optimization problems where one seeks "optimal parameters". Another important problem source concerns problems where coefficients exhibit such a fine scale structure that a direct numerical resolution is too costly. For instance, in porous media flow exact permeability (diffusion) coefficients in Darcy flow to represent the porosity are out of reach. One possibility is to view such coefficients as *random fields*. Expanding such random fields in suitable bases (Karhunen-Loewe expansion) yields coefficients that depend in principle even on infinitely many parameters. This is a central issue in "Uncertainty Quantification". A typical model problem reads as follows (see [2]):

Let $\mathcal{I} = \{1,\ldots,d\}$ or $\mathcal{I} = \mathbb{N}$ in the finite or infinite dimensional case, a given index set and consider the *affinely parametrized* diffusion problem of the form

$$F(y)u := -\mathrm{div}\big(a(y)\nabla u\big) = f, \quad a(y) := \bar{a} + \sum_{j\geq 1} y_j\theta_j, \tag{5.6.16}$$

with homogeneous Dirichlet boundary conditions, posed in the weak sense on a spatial domain Ω. Such models arise, for instance, when parametrizing random diffusion coefficient fields modeling the permeability in porous media flow. We require that the expansion of the diffusion coefficient $a(y)$ satisfies the *uniform ellipticity assumption*

$$\sum_{j\geq 1} |\theta_j(x)| \leq \bar{a}(x) - \underline{\alpha}, \quad x \in \Omega, \tag{5.6.17}$$

for some $\underline{\alpha} > 0$. This implies that for $\mathbb{X} = H_0^1(\Omega)$, there exist $0 < r \leq R < \infty$ such that

$$\langle F(y)v, w\rangle \leq R\|v\|_{\mathbb{X}}\|w\|_{\mathbb{X}} \quad \text{and} \quad \langle F(y)v, v\rangle \geq r\|v\|_V^2, \quad v, w \in \mathbb{X},\ y \in \mathcal{Y}, \tag{5.6.18}$$

which implies in particular that $F(y)$ is boundedly invertible uniformly in $y \in \mathcal{Y}$, with

$$\|F(y)\|_{\mathcal{L}(\mathbb{X}',\mathbb{X})} \leq r^{-1}, \quad y \in \mathcal{Y}. \tag{5.6.19}$$

Moreover, writing briefly $L^2(\mathcal{Y}) = L^2(\mathcal{Y}; d\mu)$ for a given probability measure μ on $\mathcal{Y}$, we can define the operator $F : \mathbb{U} := \mathbb{X}\times L^2(\mathcal{Y}) = L^2(\mathcal{Y}; V) \to \mathbb{U}' = \mathbb{V}'\times L^2(\mathcal{Y})$ by

$$a(v, w) := \langle Fv, w\rangle_{\mathbb{U}} = \int_{\mathcal{Y}}\int_{\Omega} \langle F(y)v, w\rangle dx d\mu(y), \quad v, w \in \mathbb{U}, \tag{5.6.20}$$

The problem

$$a(u, v) = f(v), \quad v \in \mathbb{U}, \tag{5.6.21}$$

is again well-posed over $\mathbb{U}$.

The particular challenge of this type of problems lies in the fact that the solutions u are now functions of the spatial variables $x \in \Omega$ and also of the (possibly infinitely many) parametric variables $y \in \mathcal{Y}$, see [2].

5.6.5 Space-Time Variational Formulation of Parabolic Problems

Model problem: $u = u(x, t)$

$$\begin{aligned} \partial_t u &= \Delta u && \text{in } D_T := \Omega \times [0, T] \\ u &= 0 && \text{on } \partial\Omega \times [0, T] \\ u(\cdot, 0) &= u_0 && \text{on } \Omega \end{aligned} \tag{5.6.22}$$

which gives rise to the variational formulation

$$\int_{D_T} \partial_t u\, v - \Delta u\, v \, dx\, dt = 0 \qquad \forall v \in \text{test space } \mathbb{V}\,.$$

Question: What is the right test space $\mathbb{V}$? Integration by parts yields

$$\int_\Omega u(x,T)\,v(x,T)\,dx - \int_\Omega u_0(x)\,v(x,0)\,dx - \int_{D_T} u\,\partial_t v\,dx\,dt + \int_{D_T} \nabla_x u(x,t)\cdot\nabla_x v(x,t)\,dx\,dt = 0\,.$$

Suppose $v(\cdot,T) = 0$, then it follows that

$$\int_0^T \int_\Omega \left(\nabla_x u(x,t)\cdot\nabla_x v(x,t) - u(x,t)\,\partial_t v(x,t)\right)\,dx\,dt - \int_\Omega u_0(x)\,v(x,0)\,dx = 0\,.$$

So appropriate spaces are

$$\begin{aligned} &\mathbb{V} := H^1_{0,\Gamma_T}(D_T) \qquad \text{where } \Gamma_T = \partial\Omega\times[0,T)\cup\Omega\times\{T\} \\ &\mathbb{U} := L_2([0,T), H^1_0(\Omega)) \end{aligned}$$

with the norms

$$\|u\|^2_{\mathbb{U}} = \int_0^T \|u(\cdot,t)\|^2_{H^1}\,dt, \qquad \|v\|_{\mathbb{V}} = \|v\|_{H^1(D_T)}\,.$$

Then (5.6.1) takes the form

$$A(u,v) = l(v), \qquad v\in H^1_{0,\Gamma_T}(D_T) = \mathbb{V}\,, \tag{5.6.23}$$

where

$$l(v) = \int_\Omega u_0(x)\,v(x,0)\,dx, \qquad A(u,v) = \int_{D_T} \nabla_x u\cdot\nabla_x v - u\,\partial_t v\,dx\,dt\,. \tag{5.6.24}$$

One can show continuity of $A(\cdot,\cdot)$ by using Cauchy-Schwarz:

$$|A(v,w)| \le C\,\|v\|_{\mathbb{U}}\,\|w\|_{\mathbb{V}}\,,$$

inf-sup stability will be shown later.

Comments 5.6.1.

a) *One has to verify that* $l\in\mathbb{V}'$. *This requires that the trace of* v *on* $\Omega\times\{0\}$ *is well-defined as a function in* $L_2(\Omega)$. *Then*

$$\begin{aligned} |l(v)| = \left|\int_\Omega u_0 v\,dx\right| &\le \|u_0\|_{L_2(\Omega)}\,\|v(\cdot,0)\|_{L_2(\Omega)} \\ &\le \|u_0\|_{L_2(\Omega)}\,\|v(\cdot,0)\|_{H^{\frac12}_{00}(\Omega)} \le C\,\|u_0\|_{L_2(\Omega)}\,\|v\|_{H^1(D_T)}\,. \end{aligned}$$

b) *Dirichlet condition:* $u(x,0) = u_0(x)$ *is a natural boundary condition.*

Remember:

$$H_0^1(\Omega) := \overline{C_0^\infty}^{H^1}\,,$$
$$H_0^s(\Omega) := \overline{C_0^\infty}^{H^s}\,.$$

An alternative to define spaces with zero trace is

$$H_{00}^s(\Omega) := \left\{v \in H^s(\Omega) : \chi_\Omega v \in H^s(\mathbb{R}^d)\right\}$$

where $\chi_\Omega v$ extends v to 0 outside of Ω.

$$H_{00}^s(\Omega) \neq H_0^s(\Omega) \qquad \text{for } s \in \frac{1}{2} + \mathbb{N}_0\,.$$

5.6.6 What About Transport?

Now regard the convection-diffusion equation

$$\begin{aligned} -\epsilon\Delta u + b\cdot\nabla u + c\,u &= f \qquad &&\text{in } \Omega\,,\\ u &= 0 &&\text{on } \partial\Omega\,. \end{aligned} \tag{5.6.25}$$

Assume

$$\begin{aligned} c(x) - \frac{1}{2}\operatorname{div} b(x) \geq 0, \qquad & x \in \Omega\,,\\ \|c\|_{L_\infty} \leq C_0, \qquad & \|b\|_{L_\infty} \leq C_1\,. \end{aligned}$$

Choose again $\mathbb{U} = \mathbb{V} = H_0^1(\Omega)$. Then (5.6.25) has a unique weak solution in $H_0^1(\Omega)$,

$$A(u,v) = \int_\Omega \epsilon\nabla u\cdot\nabla v + (b\cdot\nabla u)v + c\,u\,v\,dx, \qquad \mathbb{U} = H_0^1(\Omega) = \mathbb{V}\,.$$

One can show

$$A(v,v) = \int_\Omega \epsilon\,|\nabla v|^2 + \left(c - \frac{1}{2}\operatorname{div} b\right) v^2\,dx \geq \epsilon C_\Omega\,\|v\|_{H^1(\Omega)}^2$$

as well as continuity

$$|A(u,v)| \leq \underbrace{C\,\|b\|_{L_\infty}}_{C_a} \|u\|_{H^1}\,\|v\|_{H^1}\,.$$

But for $\|b\|_{L_\infty} \gg \epsilon$ (convection – dominated case) one has

$$\kappa_{H_0^1(\Omega),H_0^1(\Omega)}(A) \leq \frac{C\,\|b\|_{L_\infty}}{C_\Omega\epsilon} \gg 1\,.$$

So in theory this is stable but it is not well conditioned; the constants matter! Here other spaces are needed. In such cases one will end up with $\mathbb{U} \neq \mathbb{V}$.

5.7 A Road Map for Adaptive Wavelet Schemes

In this section we discuss a first adaptive paradigm that can be applied to all examples discussed before. In all cases one proceeds along the following steps. However, we concentrate on symmetric formulations, i.e., $\mathbb{U} = \mathbb{V}$.

(I) Verify that

$$F(u) = f \tag{5.7.1}$$

gives rise to a stable variational formulation in the following sense: find a Hilbert space $\mathbb{U}$ such that for each $f \in \mathbb{U}'$ the problem

$$\langle F(u), v\rangle = \langle f, v\rangle \qquad v \in \mathbb{U}$$

has a unique solution $u \in \mathbb{U}$, depending continuously on the data $f \in \mathbb{U}'$. In other words, F should have the *mapping properties*

$$\kappa_{\mathbb{U},\mathbb{U}'}(F) \le \kappa \text{ or } \kappa_{\mathbb{U},\mathbb{U}'}(DF(w)) \le \kappa,\ w \in \mathcal{N}(u), \text{ when } F \text{ is nonlinear.}$$

(II) Find a Riesz basis $\Psi = \{\psi_\lambda : \lambda \in \Lambda\} \subset \mathbb{U}$ for $\mathbb{U}$ and write **(I)** as an *equivalent* infinite system (transformation step)

$$\mathbf{F}(\mathbf{u}) = \mathbf{f} \tag{5.7.2}$$

where the unknown $u \in \mathbb{U}$ is expressed as

$$u = \sum_{\lambda\in\Lambda} u_\lambda \psi_\lambda, \qquad \mathbf{u} = (u_\lambda)_{\lambda\in\Lambda}$$

and $\mathbf{F}(\cdot)$ is the corresponding (nonlinear) operator in wavelet representation.

(III) Show that the stability of the variational formulation combined with the Riesz-basis property imply that (5.7.2) is *well-conditioned* in $\ell_2(\Lambda)$, that is

$$D\mathbf{F}(\mathbf{w}) : \ell_2(\Lambda) \to \ell_2(\Lambda)$$

is an isomorphism for each $\mathbf{w} \in \ell_2(\Lambda)$.
(*Note: The transformation gains us something because: Before we had $\mathbb{U} \to \mathbb{U}'$, for example $H^1 \to H^{-1}$, so this changed smoothness which results in a bad condition; you have to precondition*)

(IV) Contrive an "idealized" iteration

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \mathbf{C}_n(\mathbf{f} - \mathbf{F}(\mathbf{u}^n)) \tag{5.7.3}$$

on $\ell_2(\Lambda)$ that converges with a guaranteed error reduction per step

$$\left\|\mathbf{u}^{n+1} - \mathbf{u}\right\|_{l_2} \le \rho \left\|\mathbf{u}^n - \mathbf{u}\right\|_{l_2} \tag{5.7.4}$$

for some $\rho \in (0, 1)$. $\mathbf{C}_n$ is a preconditioner.
(*Note: So far everything is infinite-dimensional and hence not computable*)

(V) Develop "computable" versions of (5.7.3) by applying **F** only *approximately* within suitable error tolerances.

(VI) Prove that the perturbed scheme converges with "optimal complexity". Recall that for a given Riesz basis Ψ the error of best n-term approximation is given by

$$\sigma_n(u)_{\mathbb{U}} := \min_{\#\Gamma \le n; d_\lambda \in \mathbb{R}} \left\| u - \sum_{\lambda \in \Gamma} d_\lambda \psi_\lambda \right\|_{\mathbb{U}},$$

and behaves like $\sigma_n(\mathbf{u})_{\ell_2(\Lambda)}$ when $\mathbf{u}$ is the array of expansion coefficients of u, see Exercise 4.4.2).

Dream-Theorem. *If the solution* u *of* (5.7.1) *belongs to*

$$\mathcal{A}^s_\infty((\Sigma_n), \mathbb{U}) = \{v \in \mathbb{U} : \sup_{n \in \mathbb{N}} n^s \sigma_n(v)_{\mathbb{U}} = |v|_{\mathcal{A}^s_\infty} < \infty\},$$

then for any target accuracy $\epsilon > 0$ *the scheme outputs a finite sequence* $\mathbf{u}(\epsilon)$ *such that* $u(\epsilon) = \sum_{\lambda \in \Lambda} u^\epsilon_\lambda \psi_\lambda$ *satisfies*

$$\|u - u(\epsilon)\|_{\mathbb{U}} \le \epsilon, \qquad \#\operatorname{supp}\{\mathbf{u}(\epsilon)\} = \#(u^\epsilon_\lambda)_{\lambda \in \Lambda_\epsilon} \lesssim \epsilon^{-\frac{1}{s}} |u|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} \tag{5.7.5}$$

and

$$\#\text{operations} \lesssim \epsilon^{-\frac{1}{s}} |u|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} .$$

Note that $\sup n^s \sigma_n = c < \infty$ means $\sigma_n \lesssim n^{-s}$. In order to obtain an error of order ϵ as in (5.7.5), in general $n \sim \epsilon^{-\frac{1}{s}}$ coefficients are required. In this sense, the above is the best one can expect.

5.8 Transformation to Equivalent Problem in $\ell_2(\Lambda)$ (II), (III)

We consider first linear problems. Suppose we have found a Hilbert space $\mathbb{U}$ such that the problem: find $u \in \mathbb{U}$ such that

$$A(u, v) = \langle f, v \rangle \qquad \forall v \in \mathbb{U}, \tag{5.8.1}$$

for the linear operator equation

$$Au = f$$

(that is $A(u, v) = \langle Au, v \rangle$) is stable, i.e.,

$$\kappa_{\mathbb{U},\mathbb{U}'}(A) < \infty.$$

This means that there are constants c_a, C_a such that the mapping property (MP) holds:

$$c_a \|v\|_{\mathbb{U}} \leq \|Av\|_{\mathbb{U}'} \leq C_a \|v\|_{\mathbb{U}}, \qquad v \in \mathbb{U}, \tag{MP}$$

and that $\Psi = \{\psi_\lambda : \lambda \in \Lambda\} \subset \mathbb{U}$ is a Riesz basis for $\mathbb{U}$, i.e., there exist constants c_Ψ, C_Ψ such that the norm equivalence (NE) holds:

$$c_\Psi \|\mathbf{v}\|_{\ell_2} \leq \left\| \sum_{\lambda\in\Lambda} v_\lambda \psi_\lambda \right\|_{\mathbb{U}} \leq C_\Psi \|\mathbf{v}\|_{\ell_2}, \tag{NE}$$

where $\mathbf{v} = (v_\lambda)_{\lambda\in\Lambda} \in \ell_2(\Lambda)$.

Theorem 5.8.1. *Let* $\mathbf{A} := (A(\psi_\nu, \psi_\lambda))_{\lambda,\nu\in\Lambda}$, $\mathbf{f} = (\langle f, \psi_\lambda\rangle)_{\lambda\in\Lambda}$. *Then* (5.8.1) *is equivalent to*

$$\mathbf{Au} = \mathbf{f}, \tag{5.8.2}$$

and

$$u = \sum_{\lambda\in\Lambda} u_\lambda \psi_\lambda$$

solves (5.8.1). *Moreover, when the mapping property (MP) and the norm equivalence (NE) hold then* (5.8.2) *is* $\ell_2(\Lambda)$*−stable, that is* $\mathbf{A} : \ell_2(\Lambda) \to \ell_2(\Lambda)$ *is an isomorphism. More precisely*

$$c_a c_\Psi^2 \|\mathbf{v}\|_{\ell_2} \leq \|\mathbf{Av}\|_{\ell_2} \leq C_a C_\Psi^2 \|\mathbf{v}\|_{\ell_2}, \qquad \mathbf{v} \in \ell_2(\Lambda). \tag{5.8.3}$$

Remark 5.8.1. *While* $A : \mathbb{U} \to \mathbb{U}'$ *typically changes regularity, the representation* $\mathbf{A}$ *does not, i.e.,* (5.8.2) *is well-conditioned as a mapping from a space onto itself, i.e., for the same metric in domain and range.*

Proof. We know $u \in \mathbb{U}$ has a unique representation $u = \sum_{\lambda\in\Lambda} u_\lambda \psi_\lambda$ with the sequence $\mathbf{u} = (u_\lambda)_{\lambda\in\Lambda}$. To prove the first part of the assertion, since Ψ is a basis, we have

$$\begin{aligned}
& A(u, v) = \langle f, v\rangle \qquad \forall v \in \mathbb{U} \\
\Leftrightarrow\ & A(u, \psi_\lambda) = \langle f, \psi_\lambda\rangle \qquad \forall \lambda \in \Lambda \\
\Leftrightarrow\ & A\left(\sum_{\nu\in\Lambda} u_\nu \psi_\nu, \psi_\lambda\right) = \langle f, \psi_\lambda\rangle \qquad \forall \lambda \in \Lambda \\
\Leftrightarrow\ & \sum_{\nu\in\Lambda} A(\psi_\nu, \psi_\lambda) u_\lambda = \langle f, \psi_\lambda\rangle \qquad \forall \lambda \in \Lambda \\
\Leftrightarrow\ & (\mathbf{Au})_\lambda = \langle f, \psi_\lambda\rangle \qquad \forall \lambda \in \Lambda \\
\Leftrightarrow\ & \mathbf{Au} = \mathbf{f}
\end{aligned}$$

which proves (5.8.2).

To establish stability (5.8.3), we apply the result from Exercise 4.2.2 in Section 4.2, namely:

$$\frac{1}{C_\Psi}\|\mathbf{w}\|_{\ell_2} \le \Big\|\underbrace{\sum_{\lambda\in\Lambda} w_\lambda\tilde{\psi}_\lambda}_{=:w}\Big\|_{\mathbb{U}'} \le \frac{1}{c_\Psi}\|\mathbf{w}\|_{\ell_2}\ , \qquad \text{(NE')}$$

where $w_\lambda = \langle w, \psi_\lambda\rangle$ (since $\langle\psi_\lambda,\tilde{\psi}_\nu\rangle = \delta_{\lambda\nu}$) and hence

$$\frac{1}{C_\Psi}\|(\langle w,\psi_\lambda\rangle)_{\lambda\in\Lambda}\|_{\ell_2} \le \|w\|_{\mathbb{U}'} \le \frac{1}{c_\Psi}\|(\langle w,\psi_\lambda\rangle)_{\lambda\in\Lambda}\|_{\ell_2}\ .$$

Now

$$\begin{aligned}\|\mathbf{v}\|_{\ell_2} &\overset{\text{(NE)}}{\le} c_\Psi^{-1}\Big\|\underbrace{\sum_{\lambda\in\Lambda} v_\lambda\psi_\lambda}_{=:v}\Big\|_{\mathbb{U}} \overset{\text{(MP)}}{\le} c_\Psi^{-1}c_a^{-1}\|Av\|_{\mathbb{U}'}\\ &\overset{\text{(NE')}}{\le} c_a^{-1}c_\Psi^{-2}\|(\langle Av,\psi_\lambda\rangle)_{\lambda\in\Lambda}\|_{\ell_2} = c_a^{-1}c_\Psi^{-2}\|\mathbf{Av}\|_{\ell_2}\end{aligned}$$

yields the left inequality of (5.8.3).

For the upper bound recall: $Av = \sum_{\lambda\in\Lambda}\langle Av,\psi_\lambda\rangle\tilde{\psi}_\lambda$. Since by (MP),

$$\|Av\|_{\mathbb{U}'} \le C_a\|v\|_{\mathbb{U}}\ ,$$

one concludes

$$\begin{aligned}\|\mathbf{Au}\|_{\ell_2} = \|(\langle Av,\psi_\lambda\rangle)_{\lambda\in\Lambda}\|_{\ell_2} &\overset{\text{(NE')}}{\le} C_\Psi\|Av\|_{\mathbb{U}'}\\ &\overset{\text{(MP)}}{\le} C_\Psi C_a\|v\|_{\mathbb{U}} \overset{\text{(NE)}}{\le} C_\Psi^2 C_a\|\mathbf{v}\|_{\ell_2}\ ,\end{aligned} \qquad \square$$

which completes the proof.

Remark 5.8.2. *Under the hypotheses of Theorem 5.8.1 one has*

$$c_a c_\Psi^2\|\mathbf{u}-\mathbf{v}\|_{\ell_2} \le \|\mathbf{f}-\mathbf{Av}\|_{\ell_2} \le \mathbf{C_a C_\Psi^2}\|\mathbf{u}-\mathbf{v}\|_{\ell_2}, \quad \mathbf{v}\in\ell_2(\Lambda), \qquad (5.8.4)$$

i.e., errors are equivalent to residuals, both in the same *ℓ_2-metric. The bounds are the better the smaller $\frac{C_a C_\Psi^2}{c_a c_\Psi^2}$.*

Exercise 5.8.1.

a) *Suppose* (5.8.1) *is the variational formulation of the Poisson problem with $\mathbb{U} = H_0^1(\Omega)$ ($-\Delta u = f$ in $\Omega, u = 0$ on $\partial\Omega$). Let $\mathbf{A}_h$ denote the standard FE stiffness matrix for a Galerkin discretization with respect to a finite element space on a quasi-uniform mesh with mesh size h. How does $\kappa_2(\mathbf{A}_h)$ grow with decreasing mesh size h?*

b) *Suppose you have a Riesz basis* Ψ *for* $H^1_0(\Omega)$. *For any* $\Lambda_h \subset \Lambda, \#\Lambda_h = N$, *let* $\Psi_{\Lambda_h} = \{\psi_\lambda : \lambda \in \Lambda_h\} \subset H^1_0(\Omega)$. *Let* $\mathbb{U}_{\Lambda_h} = \mathrm{span}\{\psi_\lambda : \lambda \in \Lambda_h\} \subset \mathbb{U}$ *and use this as a trial space for the Galerkin scheme. Let* $A_h^\Psi = (a(\psi_\nu, \psi_\lambda))_{\lambda,\nu\in\Lambda_h} \in \mathbb{R}^{N\times N}$ *be the corresponding stiffness matrix. Show that*

$$\kappa_2(A_h^\Psi) \leq \frac{C_a C_\Psi^2}{c_a c_\Psi^2}. \tag{5.8.5}$$

Exercise 5.8.2. *Suppose one employs Galerkin schemes for a* $\mathbb{U}$*-elliptic problem based on a multiresolution hierarchy* $(\mathbb{U}_n)_{n\in\mathbb{N}}$ *of trial spaces. Suppose further that* Ψ *is a Riesz basis for* $\mathbb{U}$ *and that the trial spaces in the multiresolution hierarchy are spanned by subsets of* Ψ. *Describe a preconditioner for the Galerkin stiffness matrices* $\mathbf{A}_n$ *with respect to the standard scaling bases for the spaces* $\mathbb{U}_n$, *so that the Euclidean condition numbers remain uniformly bounded.*

The same principles work also when $F(u) = f$, with nonlinear F. Then

$$\langle F(u), v\rangle = \langle f, v\rangle, \qquad v \in \mathbb{U} \tag{5.8.6}$$

is equivalent to

$$\mathbf{F}(\mathbf{u}) = \mathbf{f} \tag{5.8.7}$$

where $\mathbf{F}(\mathbf{u}) = (\langle F(u), \psi_\lambda\rangle)_{\lambda\in\Lambda}$.

Exercise 5.8.3.

1. *If* (5.8.6) *is stable, that is* (5.3.4) *holds, then one obtains*

$$c_w c_\Psi^2 \|\mathbf{v}\|_{\ell_2} \leq \|D\mathbf{F}(\mathbf{w})\mathbf{v}\|_{\ell_2} \leq C_w C_\Psi^2 \|\mathbf{v}\|_{\ell_2} \qquad \forall \mathbf{v} \in \ell_2(\Lambda) \tag{5.8.8}$$

   for all w *in the stability region.*

2. *Identify* $D\mathbf{F}(\mathbf{w})\mathbf{v}$ *and specify this for* $F(v) = v^3$.

5.9 Idealized Iteration (IV)

The idea is to solve the infinite system (5.8.2) iteratively. Since $\mathbf{A}$ is well-conditioned one expects that such iterations reduce the error in each step by a fixed factor. Such an iteration serv es only as a conceptual starting point. At this point it is *idealized* since the involved infinite arrays cannot be computed (exactly).

Goal: Contrive an iterative scheme realizing a guaranteed fixed error reduction per step:

$$\|\mathbf{u}^{k+1} - \mathbf{u}\|_{l_2} \leq \rho \|\mathbf{u}^k - \mathbf{u}\|_{l_2}, \qquad k = 0, 1, 2, ... \tag{5.9.1}$$

for some fixed $\rho \in (0,1)$. We'll see that this is possible because of (5.8.3). We will discuss several example classes. The starting point is a *fixed point* ansatz

$$\mathbf{u} = \mathbf{u} + \mathbf{C}(\mathbf{f} - \mathbf{F}(\mathbf{u}))$$

which holds trivially for any isomorphism $\mathbf{C} : \ell_2 \to \ell_2$. This gives

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \mathbf{C}_n(\mathbf{f} - \mathbf{F}(\mathbf{u}^n)), \qquad n = 0, 1, 2, \dots. \tag{5.9.2}$$

Here the matrices $\mathbf{C}_n$ (that could but do not have to depend on $n, \mathbf{u}^n$) play the role of a *preconditioner*. For a given problem we have to find $\mathbf{C}_n$ such that (5.9.1) holds.

We discuss a few examples:

$\mathbb{U}$-elliptic problems: Let $\mathbf{F}(\mathbf{u}) = \mathbf{A}\mathbf{u}$ where $A(\cdot,\cdot)$ is symmetric and $\mathbb{U}$-elliptic (for example Poisson problem)

Claim: $\mathbf{C}_n = \mathbf{C} = \alpha\mathbf{I}$ for some suitable $\alpha > 0$ (**relaxation parameter**) satisfies (5.9.1) (Richardson scheme, simplified gradient descent).

To see this: $\mathbf{A}$ is by definition symmetric and, as a Gramian, positive definite. More precisely, as shown earlier

$$\|\mathbf{A}\| := \|\mathbf{A}\|_{\mathcal{L}(\ell_2,\ell_2)} = \sup_{\mathbf{v}\in l_2} \frac{\|\mathbf{A}\mathbf{v}\|_{l_2}}{\|\mathbf{v}\|_{l_2}} \overset{(5.8.3)}{\leq} C_a C_\Psi^2 =: C_\mathbf{A}, \tag{5.9.3}$$

and, again by (5.8.3),

$$\left\|\mathbf{A}^{-1}\right\| \leq c_\mathbf{A}^{-1}, \quad c_\mathbf{A} := c_a c_\Psi^2.$$

Then (5.9.2) reads:

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \alpha(\mathbf{f} - \mathbf{A}\mathbf{u}^n) = (\mathbf{I} - \alpha\mathbf{A})\mathbf{u}^n + \alpha\mathbf{f} =: \Phi(\mathbf{u}^n)$$

and

$$\begin{aligned} \Phi(v) - \Phi(w) &= \|(\mathbf{I} - \alpha\mathbf{A})(\mathbf{v} - \mathbf{w})\|_{l_2} \\ &\leq \underbrace{\|\mathbf{I} - \alpha\mathbf{A}\|}_{=:\rho} \|\mathbf{v} - \mathbf{w}\|_{l_2} \end{aligned}$$

So we need to show $\rho < 1$. Thus we have to find α such that the spectral radius fulfills $\sigma(\mathbf{I} - \alpha\mathbf{A}) < 1$. We know that $\lambda_{min}(\mathbf{A}) \geq c_\mathbf{A}$ and $\lambda_{max}(\mathbf{A}) \leq C_\mathbf{A}$. Consequently $\sigma(\mathbf{I} - \alpha\mathbf{A}) \subset [1 - \alpha C_\mathbf{A}, 1 - \alpha c_\mathbf{A}]$. Hence we need

$$\begin{aligned} & 1 - \alpha C_\mathbf{A} \geq -\rho \quad \text{and} \quad 1 - \alpha c_\mathbf{A} \leq \rho \\ \Leftrightarrow \quad & \alpha C_\mathbf{A} - 1 \leq \rho \quad \text{and} \quad 1 - \alpha c_\mathbf{A} \leq \rho \\ \Leftrightarrow \quad & \frac{1-\rho}{c_\mathbf{A}} \leq \alpha \leq \frac{1+\rho}{C_\mathbf{A}}. \end{aligned}$$

α is optimal if

$$\frac{1-\rho}{c_{\mathbf{A}}} = \alpha = \frac{1+\rho}{C_{\mathbf{A}}}$$

which is equivalent to

$$\rho = \frac{C_{\mathbf{A}} - c_{\mathbf{A}}}{C_{\mathbf{A}} + c_{\mathbf{A}}} = \frac{\frac{C_{\mathbf{A}}}{c_{\mathbf{A}}} - 1}{\frac{C_{\mathbf{A}}}{c_{\mathbf{A}}} + 1} \geq \frac{\kappa_2(\mathbf{A}) - 1}{\kappa_2(\mathbf{A}) + 1} \tag{5.9.4}$$

and

$$\alpha = \frac{1+\rho}{C_{\mathbf{A}}} = \ldots = \frac{2}{C_{\mathbf{A}} + c_{\mathbf{A}}} = \frac{2}{C_a C_\Psi^2 + c_a c_\Psi^2} \tag{5.9.5}$$

Indefinite problems: Let $\mathbf{F}(\mathbf{u}) = \mathbf{A}\mathbf{u}$, suppose the operator A satisfies the mapping property (MP) but is not positive definite (example: Stokes). In this case $\mathbf{I} - \alpha\mathbf{A}$ has eigenvalues > 1 and hence is no contraction, but $\mathbf{A} : \ell_2 \to \ell_2$ is still an isomorphism satisfying

$$c_a c_\Psi^2 \|\mathbf{v}\|_{\ell_2} = c_A \|\mathbf{v}\|_{\ell_2} \leq \|\mathbf{A}\mathbf{v}\|_{\ell_2} \leq C_A \|\mathbf{v}\|_{\ell_2} = C_a C_\Psi^2 \|\mathbf{v}\|_{\ell_2} \ ,$$

and

$$\mathbf{A}\mathbf{u} = \mathbf{f} \quad \Leftrightarrow \quad \underbrace{\mathbf{A}^{\mathsf{T}}\mathbf{A}}_{=:\mathbf{B}}\mathbf{u} = \underbrace{\mathbf{A}^{\mathsf{T}}\mathbf{f}}_{=:\tilde{\mathbf{f}}} \ . \tag{5.9.6}$$

This is still well-conditioned because

$$\begin{aligned} \|\mathbf{B}\mathbf{v}\|_{\ell_2} &= \sup_{\mathbf{w}} \frac{\mathbf{w}^{\mathsf{T}}\mathbf{B}\mathbf{v}}{\|\mathbf{w}\|_{\ell_2}} \geq \frac{\mathbf{v}^t\mathbf{B}\mathbf{v}}{\|\mathbf{v}\|_{\ell_2}} = \frac{\mathbf{v}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{A}\mathbf{v}}{\|\mathbf{v}\|_{\ell_2}} \\ &= \|\mathbf{A}\mathbf{v}\|_{\ell_2}^2 \|\mathbf{v}\|_{\ell_2} \geq c_A^2 \frac{\|\mathbf{v}\|_{\ell_2}^2}{\|\mathbf{v}\|_{\ell_2}} = c_A^2 \|\mathbf{v}\|_{\ell_2} \ . \end{aligned}$$

Conversely

$$\|\mathbf{B}\mathbf{v}\|_{\ell_2} = \|\mathbf{A}^{\mathsf{T}}\mathbf{A}\mathbf{v}\|_{\ell_2} \leq \underbrace{\|\mathbf{A}^{\mathsf{T}}\|_{\mathcal{L}(\ell_2,\ell_2)}}_{\leq C_A} \underbrace{\|\mathbf{A}\mathbf{v}\|_{\ell_2}}_{\overset{\text{MP of A}}{\leq} C_A\|\mathbf{v}\|_{\ell_2}} \leq C_A^2 \|\mathbf{v}\|_{\ell_2} \ .$$

So one still gets stability of $\mathbf{B} = \mathbf{A}^{\mathsf{T}}\mathbf{A}$, however with the condition squared. Now apply case (i) to

$$\begin{aligned} \mathbf{u}^{n+1} = \mathbf{u}^n + &\underbrace{\underbrace{\tilde{\alpha}\mathbf{A}^{\mathsf{T}}}_{=\mathbf{C}_n}(\mathbf{f} - \mathbf{A}\mathbf{u}^n)} \\ &= \tilde{\alpha}(\mathbf{A}^{\mathsf{T}}\mathbf{f} - \mathbf{A}^{\mathsf{T}}\mathbf{A}\mathbf{u}^n) \\ &= \tilde{\alpha}(\tilde{\mathbf{f}} - \mathbf{B}\mathbf{u}^n) \end{aligned}$$

with a suitable $\tilde{\alpha}$,

$$\tilde{\alpha} = \frac{2}{C_A^2 + c_A^2}\,.$$

Although squaring works in principle the quantitative performance gets worse.

Saddle point problems: Sometimes one can do better than "squaring" the problem. Consider the saddle point problem

$$\begin{aligned} a(u,v) + b(v,p) &= \langle f, v\rangle \qquad \forall\, v \in \mathbb{X}\,, \\ b(u,q) &= \langle g, q\rangle \qquad \forall\, q \in \mathbb{M}\,. \end{aligned} \tag{5.9.7}$$

Assume that (5.6.2) and (5.6.4) hold (continuity and inf-sup-condition = MP). Furthermore assume the somewhat stronger condition

$$a(v,v) \geq c_a \|v\|_{\mathbb{X}}^2\,, \qquad v \in \mathbb{X}\,. \tag{5.9.8}$$

(Remember: Originally we had this for $v \in \operatorname{Ker}(b)$, this would be technically more complicated but also sufficient. This can be found in [20]). Then

$$\begin{aligned} &A : \mathbb{X} \to \mathbb{X}'\,, \langle Au, v\rangle := a(u,v) \text{ is an isomorphism on } \mathbb{X} \to \mathbb{X}'\,, \\ &B : \mathbb{X} \to \mathbb{M}'\,, \langle Bu, q\rangle := b(u,q)\,. \end{aligned}$$

Recall here $\mathbb{U} = \mathbb{X} \times \mathbb{M}$.

Now suppose $\Psi_{\mathbb{X}} = \{\psi_\lambda^{\mathbb{X}} \in \Lambda_{\mathbb{X}}\}$, $\Psi_{\mathbb{M}} = \{\psi_\nu^{\mathbb{M}} : \nu \in \Lambda_{\mathbb{M}}\}$ are Riesz bases for $\mathbb{X}$ and $\mathbb{M}$, respectively. As before

$$\begin{aligned} u &= \sum_{\lambda \in \Lambda_{\mathbb{X}}} u_\lambda \psi_\lambda^{\mathbb{X}}\,, \qquad \mathbf{u} = (u_\lambda)_{\lambda \in \Lambda_{\mathbb{X}}} \in \ell_2(\Lambda_{\mathbb{X}})\,, \\ p &= \sum_{\lambda \in \Lambda_{\mathbb{M}}} p_\lambda \psi_\lambda^{\mathbb{M}}\,, \qquad \mathbf{p} = (p_\lambda)_{\lambda \in \Lambda_{\mathbb{M}}} \in \ell_2(\Lambda_{\mathbb{M}})\,, \end{aligned}$$

and hence (5.9.7) is by Theorem 5.8.1 equivalent to

$$\begin{pmatrix} \mathbf{A} & \mathbf{B}^T \\ \mathbf{B} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}. \tag{5.9.9}$$

Instead of squaring we use block elimination:
The first line in (5.9.9) gives

$$\mathbf{A}\mathbf{u} + \mathbf{B}^T\mathbf{p} = \mathbf{f} \quad \Leftrightarrow \quad \mathbf{u} = \mathbf{A}^{-1}\mathbf{f} - \mathbf{A}^{-1}\mathbf{B}^T\mathbf{p}\,. \tag{5.9.10}$$

Applying this to the scond line in (5.9.9) yields

$$\mathbf{g} = \mathbf{B}\mathbf{u} = \mathbf{B}\mathbf{A}^{-1}\mathbf{f} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T\mathbf{p} \quad \Rightarrow \quad \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T\mathbf{p} = \mathbf{B}\mathbf{A}^{-1}\mathbf{f} - \mathbf{g}\,.$$

Hence (5.9.9) is equivalent to

$$\mathbf{S}\mathbf{p} = \tilde{\mathbf{f}}\,, \tag{5.9.10a}$$

where $\mathbf{S} := \mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T$ is the Schur complement and $\tilde{\mathbf{f}} := \mathbf{B}\mathbf{A}^{-1}\mathbf{f} - \mathbf{g}$.

Lemma 5.9.1. *There exist constants* c_S, C_S *(depending on* c_a, c_b, C_a, C_b*) such that*

$$c_S \, \|\mathbf{q}\|_{\ell_2(\Lambda_{\mathbb{M}})} \leq \|\mathbf{Sq}\|_{\ell_2(\Lambda_{\mathbb{M}})} \leq C_S \, \|\mathbf{q}\|_{\ell_2(\Lambda_{\mathbb{M}})} . \quad \mathbf{q} \in \ell_2(\Lambda_{\mathbb{M}}). \tag{5.9.11}$$

Proof. Obviously $\mathbf{S}$ is symmetric and bounded because $\mathbf{A}^{-1}$ and $\mathbf{B}$ are bounded. Moreover,

$$\begin{aligned} \mathbf{q}^T\mathbf{Sq} = \mathbf{q}^T\mathbf{BA}^{-1}\mathbf{B}^T\mathbf{q} &= (\mathbf{B}^T\mathbf{q})^T\mathbf{A}^{-1}(\mathbf{B}^T\mathbf{q}) \geq C_{\mathbf{A}}^{-1}(\mathbf{B}^T\mathbf{q})^T(\mathbf{B}^T\mathbf{q}) \\ &\geq C_{\mathbf{A}}^{-1}(\mathbf{q}^T\mathbf{q}) \min_{\mathbf{q}'} \Big(\frac{\|\mathbf{B}^T\mathbf{q}'\|_{\ell_2(\Lambda_{\mathbb{M}})}}{\|\mathbf{q}'\|_{\ell_2(\Lambda_{\mathbb{M}})}} \Big)^2 . \end{aligned} \tag{5.9.12}$$

Since for $q' = \sum_{\lambda\in\Lambda_{\mathbb{M}}} q'_\lambda \psi_\lambda^{\mathbb{M}}, w = \sum_{\lambda\in\Lambda_{\mathbb{X}}} w_\lambda \psi_\lambda^{\mathbb{X}}$ one has

$$\begin{aligned} \|\mathbf{B}^T\mathbf{q}'\|_{\ell_2(\Lambda_{\mathbb{M}})} &= \sup_{\mathbf{w}} \frac{(\mathbf{q}')^T\mathbf{Bw}}{\|\mathbf{w}\|_{\ell_2(\Lambda_{\mathbb{X}})}} \geq \sup_{w} \frac{b(w, q')}{c_{\psi^{\mathbb{X}}}^{-1}\|w\|_{\mathbb{X}}} \geq c_b c_{\psi^{\mathbb{X}}} \|q'\|_{\mathbb{M}} \\ &\geq c_b c_{\psi^{\mathbb{X}}} c_{\psi^{\mathbb{M}}} \|\mathbf{q}'\|_{\ell_2(\Lambda_{\mathbb{M}})}, \end{aligned} \tag{5.9.13}$$

where we have used the inf-sup condition (5.6.4) in the second but last step. Inserting this into (5.9.12), yields

$$\mathbf{q}^T\mathbf{Sq} \geq C_{\mathbf{A}}^{-1} c_b c_{\psi^{\mathbb{X}}} c_{\psi^{\mathbb{M}}} (\mathbf{q}^T\mathbf{q}),$$

which yields

$$\|\mathbf{S}^{-1}\| \leq \frac{C_{\mathbf{A}}}{c_b c_{\psi^{\mathbb{X}}} c_{\psi^{\mathbb{M}}}}. \tag{5.9.14}$$

Hence, the system (5.9.10a) can be solved with the aid of the simple Richardson iteration ($\mathbf{C}_n = \alpha\mathbf{I}$ in (5.9.2)) with a suitable damping factor $\alpha > 0$ □

How to make use of (5.9.10a) ?

Uzawa iteration: In principle, the iteration

$$\mathbf{p}^{k+1} = \mathbf{p}^k + \alpha(\tilde{\mathbf{f}} - \mathbf{Sp}^k)$$

converges for suitable α with fixed $\rho < 1$.

Problem: There is an inverse in $\mathbf{S}$. Hence an explicit (approximate) application of $\mathbf{S}$ has to be avoided. To that end, go back to the fixed point iteration

$$\begin{aligned} \mathbf{p} &= \mathbf{p} + \mathbf{C}\, \underbrace{(\underbrace{\mathbf{BA}^{-1}\mathbf{f} - \mathbf{g}}_{=\tilde{\mathbf{f}}} - \underbrace{\mathbf{BA}^{-1}\mathbf{B}^T\mathbf{p}}_{=\mathbf{Sp}})}_{\substack{=\mathbf{BA}^{-1}(\mathbf{f}-\mathbf{B}^T\mathbf{p}) - \mathbf{g} \\ \overset{(5.9.10)}{=} \mathbf{Bu} - \mathbf{g}}} \\ &= \mathbf{p} + \mathbf{C}(\mathbf{Bu} - \mathbf{g}) \end{aligned}$$

This leads to the so called **Uzawa iteration**:

Algorithm 5.9.1.

1. Choose an initial guess $\mathbf{p}^0$.
2. Solve $\mathbf{u}^0 = \mathbf{A}^{-1}(\mathbf{f} - \mathbf{B}^\mathsf{T}\mathbf{p}^0)$, i.e., solve $\mathbf{A}\mathbf{u}^0 = \mathbf{f} - \mathbf{B}^\mathsf{T}\mathbf{p}^0$.
3. Given $\mathbf{p}^k, \mathbf{u}^k$, set
$$\mathbf{p}^{k+1} = \mathbf{p}^k + \mathbf{C}(\mathbf{B}\mathbf{u}^k - \mathbf{g})\,. \tag{5.9.15}$$
4. Solve $\mathbf{A}\mathbf{u}^{k+1} = \mathbf{f} - \mathbf{B}^\mathsf{T}\mathbf{p}^{k+1}$.

Comments 5.9.1.

- $\mathbf{S}$ *is not needed explicitly.*
- *Each iteration step requires an elliptic solve*
$$\mathbf{A}\mathbf{u}^k = \left(\mathbf{f} - \mathbf{B}^\mathsf{T}\mathbf{p}^k\right)\,,$$
but at each stage in a practical realization of an iterative solver you need only few steps to update the approximation to $\mathbf{u}$.
- *Then just update* $\mathbf{p}^k$ *to* $\mathbf{p}^{k+1}$

Error reduction:

$$\begin{aligned}\mathbf{p}^{k+1} - \mathbf{p} &= \mathbf{p}^k + \mathbf{C}(\mathbf{B}\mathbf{u}^k - \mathbf{g}) - (\mathbf{p} + \mathbf{C}\,\underbrace{(\mathbf{B}\mathbf{u} - \mathbf{g})}_{=0})\\ &= \mathbf{p}^k - \mathbf{p} + \mathbf{C}\mathbf{B}(\mathbf{u}^k - \mathbf{u})\,.\end{aligned} \tag{5.9.16}$$

Now, using (5.9.10) and step 4. in Algorithm 5.9.1, we obtain

$$\begin{aligned}& \mathbf{A}(\mathbf{u} - \mathbf{u}^k) = \mathbf{f} - \mathbf{B}^\mathsf{T}\mathbf{p} - \mathbf{f} + \mathbf{B}^\mathsf{T}\mathbf{p}^k = \mathbf{B}^\mathsf{T}(\mathbf{p}^k - \mathbf{p})\\ \Rightarrow\quad & (\mathbf{u} - \mathbf{u}^k) = \mathbf{A}^{-1}\mathbf{B}^\mathsf{T}(\mathbf{p}^k - \mathbf{p})\\ \overset{(5.9.16)}{\Rightarrow}\quad & \mathbf{p}^{k+1} - \mathbf{p} = \mathbf{p}^k - \mathbf{p} + \mathbf{C}\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^\mathsf{T}(\mathbf{p} - \mathbf{p}^k) = (\mathbf{I} - \mathbf{C}\mathbf{S})(\mathbf{p}^k - \mathbf{p})\,.\end{aligned}$$

So we have a contraction for suitable $\mathbf{C}$.

We summarize the above findings as follows.

Remark 5.9.1. *In all the above cases of the linear problem*

$$\mathbf{F}(\mathbf{u}) = \mathbf{f}$$

(of elliptic, indefinite, or saddle point type) we have identified a preconditioner $\mathbf{C}$ *such that the iteration*

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \mathbf{C}(\mathbf{f} - \mathbf{F}(\mathbf{u}^k)), \quad k \in \mathbb{N}_0,$$

(where $\mathbf{C}$ *may depend on* $\mathbf{u}^k$*) gives rise to a fixed error reduction in each step by a factor* $\rho < 1$*. In fact, we have shown in all cases that*

$$\|\mathbf{I} - \mathbf{CF}\| \le \rho, \quad k \in \mathbb{N}_0, \tag{5.9.17}$$

where $\|\cdot\|$ *is the spectral norm. In terms of the residual* $\mathbf{R}(\mathbf{v}) = \mathbf{f} - \mathbf{F}(\mathbf{v})$*, this can be equivalently expressed as*

$$\|\mathbf{v} - \mathbf{w} + \mathbf{C}(\mathbf{R}(\mathbf{v}) - \mathbf{R}(\mathbf{w}))\|_{\ell_2} \le \rho \|\mathbf{v} - \mathbf{w}\|_{\ell_2} \tag{5.9.18}$$

Semilinear elliptic problems: Recall:

$$\begin{aligned} F(u) := -\Delta u + u^3 &= f \qquad \text{in } \Omega\,, \\ u &= 0 \qquad \text{on } \partial\Omega\,. \end{aligned}$$

General format:

$$\langle F(u), v\rangle = a(u,v) + \langle G(u), v\rangle\,,$$

where $a(\cdot,\cdot)$ is $\mathbb{U}$-elliptic (here $\mathbb{U} = H_0^1(\Omega)$) and $G : \mathbb{U} \to \mathbb{U}'$ with the following properties:

(P1)

$$\|G(v) - G(w)\|_{\mathbb{U}'} \le C_G(\max\{\|v\|_{\mathbb{U}}, \|w\|_{\mathbb{U}}\})\, \|v - w\|_{\mathbb{U}}\ , \qquad v, w \in \mathbb{U}$$

(For fixed x, $t = v(x), s = w(x)$) we have $G(t) - G(s) = G'(\xi)(t-s)$ for some $\xi \in [t,s]$, $C_G(t)$ increases for increasing t). This latter property is generalized as follows:

(P2) G is **monotone**, i.e.,

$$\langle v - w, G(v) - G(w)\rangle \ge 0 \qquad \forall v, w \in \mathbb{U}\,.$$

In the example this follows immediately from the structure of $G(t) = t^3$,

$$\int_\Omega (v(x) - w(x))(G(v(x)) - G(w(x)))\,dx \overset{\text{for fixed } x}{\rightsquigarrow} (t-s)(t^3 - s^3) \ge 0\,.$$

One can show that (5.5.2) (stability) follows from (P1) and (P2) with $C_w = (C_a + C_G(\|w\|_{\mathbb{U}}))$ and $c_w = C_\Omega$ (Poincaré constant).

As before, transform the problem into

$$\mathbf{F}(\mathbf{u}) = \mathbf{f}$$

with

$$\mathbf{F}(\mathbf{u}) = \mathbf{A}\mathbf{u} + \mathbf{G}(\mathbf{u}), \quad \mathbf{A} = (a(\psi_\nu, \psi_\lambda))_{\lambda,\nu\in\Lambda},$$

and

$$\mathbf{G}(\mathbf{v}) = (\langle G(v), \psi_\lambda \rangle)_{\lambda \in \Lambda} .$$

So

$$\langle F(u), v \rangle = \langle f, v \rangle , \qquad v \in \mathbb{U},$$

is equivalent to

$$\mathbf{R}(\mathbf{u}) := \mathbf{f} - \mathbf{A}\mathbf{u} - \mathbf{G}(\mathbf{u}) = 0 . \tag{5.9.19}$$

Remark 5.9.2. *One has (monotonicity in Riesz coordinates)*

$$(\mathbf{v} - \mathbf{w})^{\mathsf{T}}(\mathbf{G}(\mathbf{v}) - \mathbf{G}(\mathbf{w})) \geq 0 , \qquad \mathbf{v}, \mathbf{w} \in \ell_2(\Lambda) .$$

Proof. Expand the arguments in $\mathbb{U}$ (using the dual Riesz basis Ψ)

$$v - w = \sum_{\lambda \in \Lambda} (v_\lambda - w_\lambda)\psi_\lambda,$$

and the images in the range $\mathbb{U}'$ (using $\tilde{\Psi}$)

$$G(v) - G(w) = \sum_{\lambda \in \Lambda} \langle G(v) - G(w), \psi_\lambda \rangle \tilde{\psi}_\lambda.$$

Substituting these expansions, yields

$$\begin{aligned} 0 &\overset{(P2)}{\leq} \langle v - w, G(v) - G(w) \rangle \\ &= \sum_{\lambda, \nu \in \Lambda} (v_\lambda - w_\lambda) \underbrace{\langle \psi_\lambda, \tilde{\psi}_\nu \rangle}_{\delta_{\lambda\nu}} (\mathbf{G}(\mathbf{v}) - \mathbf{G}(\mathbf{w}))_\nu \\ &= \sum_{\lambda \in \Lambda} (v_\lambda - w_\lambda) \, (\mathbf{G}(\mathbf{v}) - \mathbf{G}(\mathbf{w}))_\lambda \\ &= (\mathbf{v} - \mathbf{w})^{\mathsf{T}} \, (\mathbf{G}(\mathbf{v}) - \mathbf{G}(\mathbf{w})) , \end{aligned}$$

which confirms the monotonicity of the nonlinear operator $\mathbf{G}$ on $\ell_2(\Lambda)$. □

Remark 5.9.3. *Note that to compute the coefficient arrays $\mathbf{G}(\mathbf{v}) = (\langle \mathbf{G}(\mathbf{v}), \psi_\lambda \rangle)_{\lambda \in \Lambda}$ one only needs the primal basis functions ψ_λ. The dual basis functions $\tilde{\psi}_\lambda$ are never used in computations, only in the analysis.*

Remark 5.9.4. *The concept of monotonicity for nonlinear operators generalizes coercivity of linear operators. (convince yourself).*

Now study the simplest iteration for (5.9.19) (Richardson)

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \alpha \mathbf{R}(\mathbf{u}^n) , \qquad n = 0, 1, 2, ... \tag{5.9.20}$$

with suitable α.

Because of the nonlinearity we need first a priori bounds on the solution and a neighbourhood on which linearizations are well-conditioned. For this purpose we need several preparations.

Exercise 5.9.1. *The monotonicity of* G *implies positive semi-definiteness of the linearization* $\mathrm{D}\mathbf{G}(\mathbf{w})$ *for each* $\mathbf{w}$*, i.e.,*

$$\mathbf{v}^{\mathrm{T}}\mathbf{DG}(\mathbf{w})\mathbf{v} \geq \mathbf{0}, \quad \mathbf{v} \in \ell_2(\Lambda).$$

Exercise 5.9.2. *The operator* $-\mathbf{R}$ *is strictly monotone, i.e.,*

$$\langle w - v, \mathrm{R}(v) - \mathrm{R}(w)\rangle \geq c_a \|w - v\|_{\mathbb{U}}^2, \quad w, v \in \mathbb{U}. \tag{5.9.21}$$

Next we carry some properties on the the function space side over to the $\ell_2(\Lambda)$ side.

(a) The analog of (P1) in Riesz coordinates is

$$\begin{aligned}
\|\mathbf{G}(\mathbf{v}) - \mathbf{G}(\mathbf{w})\|_{\ell_2} &\overset{\text{def.}}{=} \|(\langle \mathrm{G}(v) - \mathrm{G}(w), \psi_\lambda\rangle)_{\lambda\in\Lambda}\|_{\ell_2} \\
&\overset{\mathrm{NE}'}{\leq} C_\Psi \|\mathrm{G}(v) - \mathrm{G}(w)\|_{\mathbb{U}'} \\
&\overset{\text{(P1)}}{\leq} C_\Psi C_G(\max\{\|v\|_X, \|w\|_X\}) \|v - w\|_{\mathbb{U}} \\
&\leq \underbrace{C_\Psi^2 C_G(\max\{\|v\|_{\mathbb{U}}, \|w\|_{\mathbb{U}}\})}_{=:\hat{C}(\max\{\|\mathbf{v}\|_{\ell_2}, \|\mathbf{w}\|_{\ell_2}\})} \|\mathbf{v} - \mathbf{w}\|_{\ell_2} .
\end{aligned}$$

Thus

$$\|\mathbf{G}(\mathbf{v}) - \mathbf{G}(\mathbf{w})\|_{\ell_2} \leq \hat{C}(\max\{\|\mathbf{v}\|_{\ell_2}, \|\mathbf{w}\|_{\ell_2}\}) \|\mathbf{v} - \mathbf{w}\|_{\ell_2} . \tag{5.9.22}$$

(b) Bound $\|\mathbf{u}\|_{\ell_2}$:

$$\begin{aligned}
\|\mathbf{u} - \mathbf{v}\|_{\ell_2} \|\mathbf{R}(\mathbf{v})\|_{\ell_2} &\overset{\text{Cauchy-Schwarz}}{\geq} (\mathbf{u} - \mathbf{v})^{\mathrm{T}}(\mathbf{R}(\mathbf{v}) - \underbrace{\mathbf{R}(\mathbf{u})}_{=0}) \\
&= (\mathbf{u} - \mathbf{v})^{\mathrm{T}}(\mathbf{A}(\mathbf{u} - \mathbf{v}) + \mathbf{G}(\mathbf{u}) - \mathbf{G}(\mathbf{v})) \\
&\overset{\text{Remark 5.9.2}}{\geq} (\mathbf{u} - \mathbf{v})^{\mathrm{T}}\mathbf{A}(\mathbf{u} - \mathbf{v}) \\
&\overset{\text{Rayleigh quotient}}{\geq} c_{\mathbf{A}} \|\mathbf{u} - \mathbf{v}\|_{\ell_2}^2 ,
\end{aligned}$$

where $c_{\mathbf{A}} = c_a c_\Psi^2$. Hence

$$\|\mathbf{u} - \mathbf{v}\|_{\ell_2} \leq c_{\mathbf{A}}^{-1} \|\mathbf{R}(\mathbf{v})\|_{\ell_2} . \tag{5.9.23}$$

In particular for $\mathbf{v} = \mathbf{0}$ we have

$$\|\mathbf{u}\|_{\ell_2} \leq c_{\mathbf{A}}^{-1} \|\mathbf{R}(\mathbf{0})\|_{\ell_2} = c_{\mathbf{A}}^{-1} \|\mathbf{f}\|_{\ell_2} . \tag{5.9.24}$$

Thus, defining

$$B_\delta(\mathbf{u}) = \{\mathbf{v} : \|\mathbf{v}-\mathbf{u}\|_{\ell_2} \le \delta\},$$

we conclude that

$$\begin{aligned}\|\mathbf{v}\|_{\ell_2} &\le \|\mathbf{u}\|_{\ell_2} + c_{\mathbf{A}}^{-1}\|\mathbf{R}(\mathbf{v})\|_{\ell_2}\\ &\le c_{\mathbf{A}}^{-1}\big(\|\mathbf{f}\|_{\ell_2} + \|\mathbf{R}(\mathbf{v})\|_{\ell_2}\big)\\ &\le c_{\mathbf{A}}^{-1}\big(\|\mathbf{f}\|_{\ell_2} + \max_{\tilde{\mathbf{v}}\in B_\delta(\mathbf{u})}\|\mathbf{R}(\tilde{\mathbf{v}})\|_{\ell_2}\big). \end{aligned} \tag{5.9.25}$$

(c) Bound for $\hat{C}(\max\{\|\mathbf{v}\|_{\ell_2}, \|\mathbf{w}\|_{\ell_2}\})$ over a neighborhood of $\mathbf{u}$:

Remark 5.9.5. *Given any $\delta > 0$, let*

$$\zeta(\delta) := c_A^{-1}\left(\|\mathbf{f}\|_{\ell_2} + \max_{\mathbf{v}\in B_\delta(\mathbf{u})}\|\mathbf{R}(\mathbf{v})\|_{\ell_2}\right). \tag{5.9.26}$$

Then we infer from (5.9.25) *that*

$$\max_{\mathbf{v},\mathbf{w}\in B_\delta(\mathbf{u})} \hat{C}(\max\{\|\mathbf{v}\|_{\ell_2}, \|\mathbf{w}\|_{\ell_2}\}) \le \hat{C}(\zeta(\delta)) =: C^*(\delta) \tag{5.9.27}$$

(d) Main preparation for error reduction: Recall:

$$\begin{aligned}\mathbf{u}^{n+1} &= \mathbf{u}^n + \alpha\mathbf{R}(\mathbf{u}^n)\\ \mathbf{u}^{n+1} - \mathbf{u} &= \mathbf{u}^n - \mathbf{u} + \underbrace{\alpha(\mathbf{R}(\mathbf{u}^n) - \mathbf{R}(\mathbf{u}))}_{\substack{\text{for error reduction:}\\ \text{estimate this term}}}\ .\end{aligned}$$

Hence, we need to consider quantities of the form $\mathbf{w} - \mathbf{v} + \alpha(\mathbf{R}(\mathbf{w}) - \mathbf{R}(\mathbf{v})$, compare with the linear case (5.9.18) in Remark 5.9.1.

Remark 5.9.6. *One has*

$$\mathbf{v} - \mathbf{w} + \mathbf{C}(\mathbf{R}(\mathbf{v}) - \mathbf{R}(\mathbf{w})) = (\mathbf{I} - \mathbf{C}\mathbf{M}(\mathbf{v},\mathbf{w}))(\mathbf{v}-\mathbf{w}), \tag{5.9.28}$$

with

$$\mathbf{M}(\mathbf{v},\mathbf{w}) = \int_0^1 \mathbf{A} + D\mathbf{G}(\mathbf{w} + s(\mathbf{v}-\mathbf{w}))\,ds\,. \tag{5.9.29}$$

($D\mathbf{G}$ *= linearization* $\hat{=}$ *derivative).*

Proof. Consider $\mathbf{g}(t) := \mathbf{G}(\mathbf{w} + t(\mathbf{v} - \mathbf{w}))$. Then

$$\mathbf{g}(1) = \mathbf{G}(\mathbf{v}), \quad \mathbf{g}(0) = \mathbf{G}(\mathbf{w})$$

which gives

$$\mathbf{G}(\mathbf{v}) - \mathbf{G}(\mathbf{w}) = \mathbf{g}(1) - \mathbf{g}(0) = \int_0^1 \mathbf{g}'(s)\,ds = \int_0^1 D\mathbf{G}(\mathbf{w} + s(\mathbf{v} - \mathbf{w}))(\mathbf{v} - \mathbf{w})\,ds$$

and therefore

$$\mathbf{R}(\mathbf{v}) - \mathbf{R}(\mathbf{w}) = -\left(\mathbf{A} + \int_0^1 D\mathbf{G}(\mathbf{w} + s(\mathbf{v} - \mathbf{w}))\,ds\right)(\mathbf{v} - \mathbf{w}),$$

confirming the claim. □

Next we analyze the mapping properties of $\mathbf{M}(\mathbf{v}, \mathbf{w})$ and show first coercivity. In fact, by monotonicity of $\mathbf{G}$,

$$\begin{aligned}(\mathbf{v} - \mathbf{w})^T\mathbf{M}(\mathbf{v}, \mathbf{w})(\mathbf{v} - \mathbf{w}) &\geq \lambda_{min}(\mathbf{A})\,\|\mathbf{v} - \mathbf{w}\|_{\ell_2}^2 + (\mathbf{v} - \mathbf{w})^T(\mathbf{G}(\mathbf{v}) - \mathbf{G}(\mathbf{w})) \\ &\geq \lambda_{min}(\mathbf{A})\,\|\mathbf{v} - \mathbf{w}\|_{\ell_2}^2 \\ &\geq c_A\,\|\mathbf{v} - \mathbf{w}\|_{\ell_2}^2\,. \qquad (5.9.30)\end{aligned}$$

As for an upper bound, we have

$$\begin{aligned}\|\mathbf{M}(\mathbf{v}, \mathbf{w})(\mathbf{v} - \mathbf{w})\|_{\ell_2} &= \|\mathbf{A}(\mathbf{v} - \mathbf{w}) + \mathbf{G}(\mathbf{v}) - \mathbf{G}(\mathbf{w})\|_{\ell_2} \\ &\overset{(5.9.22)}{\leq} (C_{\mathbf{A}} + \hat{C}(\max\{\|\mathbf{v}\|_{\ell_2}, \|\mathbf{w}\|_{\ell_2}\}))\,\|\mathbf{v} - \mathbf{w}\|_{\ell_2} \\ &\overset{\text{for } \mathbf{v},\mathbf{w} \in B_\delta(\mathbf{u})}{\leq} (C_{\mathbf{A}} + C^*(\delta))\,\|\mathbf{v} - \mathbf{w}\|_{\ell_2}\,. \qquad (5.9.31)\end{aligned}$$

Now fix the initial guess $\mathbf{u}^0$ (for example $\mathbf{u}^0 = \mathbf{0}$). Then (5.9.23) yields

$$\|\mathbf{u} - \mathbf{u}^0\|_{\ell_2} \leq c_{\mathbf{A}}^{-1}\|\mathbf{R}(\mathbf{u}^0)\|_{\ell_2} =: \delta_0\,, \qquad (5.9.32)$$

and hence

$$\mathbf{u}^0 \in B_{\delta_0}(\mathbf{u})\,.$$

Now we can establish error reduction in the idealized iteration (5.9.20).

Proposition 5.9.1. *For*

$$\begin{aligned}&0 < \alpha < \frac{2c_{\mathbf{A}}}{(C_{\mathbf{A}} + C^*(\delta_0))^2} \\ &\rho = \rho(\alpha) := \left(1 - 2\alpha c_{\mathbf{A}} + \alpha^2(C_{\mathbf{A}} + C^*(\delta))^2\right)^{\frac{1}{2}} < 1\end{aligned} \qquad (5.9.33)$$

the iteration

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \alpha\mathbf{R}(\mathbf{u}^n)$$

satisfies (5.9.1), *i.e.,* $\|\mathbf{u}^{k+1} - \mathbf{u}\|_{\ell_2} \leq \rho\|\mathbf{u}^k - \underline{u}\|_{\ell_2}$, *with* ρ *from* (5.9.33).

Proof. By (5.9.28)

$$\mathbf{u}^{n+1} - \mathbf{u} = \mathbf{u}^n - \mathbf{u} + \alpha(\mathbf{R}(\mathbf{u}^n) - \underbrace{\mathbf{R}(\mathbf{u})}_{=0}) = (\mathbf{I} - \alpha\mathbf{M}(\mathbf{u}^n, \mathbf{u}))\,(\mathbf{u}^n - \mathbf{u})\,.$$

Thus by (5.9.30) and (5.9.31)

$$\begin{aligned}
\|\mathbf{u}^1 - \mathbf{u}\|_{\ell_2}^2 &= \left\|\left(\mathbf{I} - \alpha\mathbf{M}(\mathbf{u}^0, \mathbf{u})\right)(\mathbf{u}^0 - \mathbf{u})\right\|_{\ell_2}^2 \\
&= (\mathbf{u}^0 - \mathbf{u})^T \left(\mathbf{I} - \alpha\mathbf{M}(\mathbf{u}^0, \mathbf{u})\right)^T \left(\mathbf{I} - \alpha\mathbf{M}(\mathbf{u}^0, \mathbf{u})\right)(\mathbf{u}^0 - \mathbf{u}) \\
&= \|\mathbf{u}^0 - \mathbf{u}\|_{\ell_2}^2 - 2\alpha(\mathbf{u}^0 - \mathbf{u})^T\mathbf{M}(\mathbf{u}^0, \mathbf{u})(\mathbf{u}^0 - \mathbf{u}) + \alpha^2 \left\|\mathbf{M}(\mathbf{u}^0, \mathbf{u})(\mathbf{u}^0 - \mathbf{u})\right\|_{\ell_2}^2 \\
&\leq \|\mathbf{u}^0 - \mathbf{u}\|_{\ell_2}^2 - 2\alpha c_{\mathbf{A}}\|\mathbf{u}^0 - \mathbf{u}\|_{\ell_2}^2 + \alpha^2(C_{\mathbf{A}} + C^*(\delta))^2\|\mathbf{u}^0 - \mathbf{u}\|_{\ell_2}^2 \\
&= \left(1 - 2\alpha c_{\mathbf{A}} + \alpha^2(C_{\mathbf{A}} + C^*(\delta))^2\right)\|\mathbf{u}^0 - \mathbf{u}\|_{\ell_2}^2 \\
&= \rho^2\|\mathbf{u}^0 - \mathbf{u}\|_{\ell_2}^2\,.
\end{aligned}$$

Hence

$$\mathbf{u}^1 \in B_{\delta_0}(\mathbf{u})\,.$$

The assertion follows by repetition. □

Corollary 5.9.1. *For any* $\mathbf{v}, \mathbf{w} \in B_{\delta_0}(\mathbf{u})$ *one has*

$$\|\mathbf{v} - \mathbf{w} + \alpha(\mathbf{R}(\mathbf{v}) - \mathbf{R}(\mathbf{w}))\|_{\ell_2} \leq \rho\,\|\mathbf{v} - \mathbf{w}\|_{\ell_2} \tag{5.9.34}$$

for ρ *from* (5.9.33)*, see Remark 5.9.1 for the linear case.*

Proof. Follows from Remark 5.9.6 and Proposition 5.9.1. □

Comments 5.9.2. $\mathbf{C} = \alpha\mathbf{I}$ *was the simplest choice that works for a monotone nonlinearity* G*. Further choices are*

$$\begin{aligned}
&\mathbf{C}_n := \alpha D\mathbf{F}(\overline{\mathbf{u}})^T \\
\textit{or } &\mathbf{C}_n := \alpha D\mathbf{F}(\mathbf{u}^n)^T \qquad \textit{(Gauß-Newton)} \\
\textit{or } &\mathbf{C}_n := D\mathbf{F}(\mathbf{u}^n)^{-1} \qquad \textit{(Newton's Method)}
\end{aligned}$$

for any fixed $\overline{\mathbf{u}}$ *and suitable* α*.*

Remark 5.9.7. *In all those cases (for suitable* $B_{\delta_0}(\mathbf{u})$*) one can find* $0 < \beta < \infty$ *and* $\rho = \rho(\delta_0) < 1$ *such that*

$$\|\mathbf{u} - \mathbf{v}\|_{\ell_2} \overset{(5.9.23)}{\leq} \beta\,\|\mathbf{C}\mathbf{R}(\mathbf{v})\|_{\ell_2}\,, \qquad \mathbf{v} \in B_{\delta_0}(\mathbf{u}) \tag{5.9.35}$$

and

$$\|\mathbf{v} - \mathbf{w} + \mathbf{C}(\mathbf{R}(\mathbf{v}) - \mathbf{R}(\mathbf{w}))\|_{\ell_2} \leq \rho\,\|\mathbf{v} - \mathbf{w}\|_{\ell_2}\,, \tag{5.9.36}$$

see Remark 5.9.1 for the linear case.

For details see [17].

5.10 A Convergent Perturbation (V)

Given a variational problem with (MP) and a Riesz basis for the energy space (NE), we have seen that one can always find an iteration in $\ell_2(\Lambda)$ that converges with a fixed error reduction ρ per step. But even in the simplest version with $\mathbf{C}_n = \alpha\mathbf{I}$ such an iteration cannot be implemented numerically because even when $\mathbf{u}^n$ has finite support, $\mathbf{R}(\mathbf{u}^n) = \mathbf{f} - \mathbf{F}(\mathbf{u}^n)$ has in general infinite support: The idealized iteration

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \mathbf{C_n}(\mathbf{f} - \mathbf{F}(\mathbf{u}^n)) = \mathbf{u}^n + \mathbf{R}(\mathbf{f}, \mathbf{F}, \mathbf{C}_n, \mathbf{u}^n) \tag{5.10.1}$$

is not computable because

- $\mathbf{f}$ is in general infinitely supported
- even if $\#(\operatorname{supp}\mathbf{u}^n) < \infty$, $\mathbf{F}(\mathbf{u}^n)$ has in general infinite support, the $\langle F(u^n), \psi_\lambda\rangle$ might all be different from 0.

Idea: Do not work with the full load vector $\mathbf{f}$ and with the exact $\mathbf{F}(\mathbf{u})$ but only with finitely supported approximations so that each iteration has finite support. This is a perturbation of the ideal iteration.

1. *Question:* How to dynamically perturb the ideal iteration so as to still converge to $\mathbf{u}$ with "optimal" complexity (dream-theorem)?
2. *Question:* How to realize these perturbations?

To answer the first question we first completely ignore how to determine such approximations and assume that appropriate routines are available. Instead we first wish to see only which tolerances are sufficient in each iteration.

Assumption 5.10.1. *We assume at this point that we have two routines available,*

- COARSE$[\mathbf{v}, \eta] \to \mathbf{v}_\eta$ *such that*

$$\|\mathbf{v} - \mathbf{v}_\eta\|_{\ell_2} \le \eta\,, \quad \#\mathbf{v}_\eta < \infty\,. \tag{5.10.2}$$

- RES$[\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{v}, \eta] \to \mathbf{r}_\eta$ *such that*

$$\|\mathbf{R}(\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{v}) - \mathbf{r}_\eta\|_{\ell_2} \le \eta\,, \tag{5.10.3}$$

where $\mathbf{R}(\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{v}) = \mathbf{C}(\mathbf{f} - \mathbf{F}(\mathbf{v}))$.

- *We have an initial guess* $\mathbf{u}^0$ *with a bound*

$$\|\mathbf{u} - \mathbf{u}^0\|_{\ell_2} \le \delta_0\,, \tag{5.10.4}$$

see (5.9.23).

Then we consider the following scheme:

Algorithm 5.10.1.

```
SOLVE$[\mathbf{F}, \mathbf{f}, \mathbf{u}^0, \epsilon] \to \mathbf{u}(\epsilon) \in \ell_2(\Lambda)$

(i) Choose $M^*$, b such that $\frac{1+b}{M^*} \leq \frac{1}{2}$, $\overline{\rho} \in (0,1)$,
    set $\overline{\mathbf{u}} = \mathbf{u}^0$, $\delta = \delta_0$ (see (5.10.4))
    $\{w_k\}_{k=0}^{\infty}$ such that $\sum_{k=0}^{\infty} w_k = 1$.

(ii) If $\delta \leq \epsilon$ stop and output $\mathbf{u}(\epsilon) = \overline{\mathbf{u}}$;
     else set $\mathbf{v}^0 := \overline{\mathbf{u}}$, $k = 0$.
    (ii.1) Set $\eta_k := w_k \overline{\rho}^k \delta$ and
           compute $\mathbf{r}^k :=$ RES$[\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{v}^k, \eta_k]$.
    (ii.2) If
           $\beta \left(\eta_k + \|\mathbf{r}^k\|_{\ell_2}\right) \leq \frac{\delta}{M^*}$     (5.10.5)
           ($\beta$ from (5.9.35))
           set $\tilde{\mathbf{v}} := \mathbf{v}^k$ and go to (iii),
           else: Set $\mathbf{v}^{k+1} = \mathbf{v}^k + \mathbf{r}^k$, set $k+1 \to k$, go to (ii.1).

(iii) COARSE$[\tilde{\mathbf{v}}, \frac{b\delta}{M^*}] \to \overline{\mathbf{u}}$, $\frac{\delta}{2} \to \delta$, go to (ii).
```

Proposition 5.10.1. *For any* $\epsilon > 0$ *the scheme* SOLVE$[\mathbf{F}, \mathbf{f}, \mathbf{u}^0, \epsilon]$ *terminates after finitely many steps and outputs a finitely supprted sequence* $\mathbf{u}(\epsilon) \in \ell_2(\Lambda)$ *satisfying*

$$\|\mathbf{u} - \mathbf{u}(\epsilon)\|_{\ell_2} \leq \epsilon \,. \tag{5.10.6}$$

Proof. When entering (ii) for the jth time, let us denote the input $\overline{\mathbf{u}}$ by $\overline{\mathbf{u}}^j$. Moreover let $\mathbf{u}^k := \mathbf{u}^k(\overline{\mathbf{u}})$ be the exact iterates for $\mathbf{u}^{k+1} = \mathbf{u}^k + \mathbf{C}(\mathbf{f} - \mathbf{F}(\mathbf{u}^k))$, that is

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \mathbf{R}(\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{u}^k)$$

with initial guess $\mathbf{u}^0 = \overline{\mathbf{u}}$. Suppose we have shown that $\left\|\mathbf{u} - \overline{\mathbf{u}}^j\right\|_{\ell_2} \leq \epsilon_j = 2^{-j}\delta_0$ (induction assumption, which is true for $j = 0$). We wish to show now that the output $\bar{\mathbf{u}} = \bar{\mathbf{u}}^{j+1}$ of step (iii) in Algorithm 5.10.1 satisfies $\|\mathbf{u} - \bar{\mathbf{u}}^{j+1}\|_{\ell_2} \leq \epsilon_{j+1} = 2^{-j-1}\delta_0$.

We compare first the perturbed iterates $\mathbf{v}^k$ in (ii.2) with the exact iterates $\mathbf{u}^k$

$$\begin{aligned}\mathbf{v}^{k+1} - \mathbf{u}^{k+1} &= \mathbf{v}^k - \mathbf{u}^k + (\mathbf{r}^k - \mathbf{R}(\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{u}^k)) \\ &= \mathbf{v}^k - \mathbf{u}^k + \underbrace{(\mathbf{r}^k - \mathbf{R}(\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{v}^k))}_{\|\cdot\| \leq \eta_k} + \mathbf{R}(\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{v}^k) - \mathbf{R}(\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{u}^k)\,.\end{aligned}$$

By (5.10.3)

$$\|\mathbf{v}^{k+1} - \mathbf{u}^{k+1}\|_{\ell_2} \leq \eta_k + \|\mathbf{v}^k - \mathbf{u}^k + \underbrace{\mathbf{R}(\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{v}^k) - \mathbf{R}(\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{u}^k)}_{\overset{(5.9.36)}{=} \mathbf{C}(\mathbf{R}(\mathbf{v}^k) - \mathbf{R}(\mathbf{u}^k))}\|_{\ell_2} \,. \tag{5.10.7}$$

When **F** is linear we can use Remark 5.9.1. When **F** is the above nonlinear operator we would like to invoke Corollary 5.9.1, (5.9.34) or (5.9.36). An additional slight complication is that we need to show that the iterates remain in a sufficiently small neighborhood of the solution. To this end, let us assume that ρ is the error reduction on the ideal scheme belonging to the neighborhood $B_{2\delta_0}(\mathbf{u})$. So in order to invoke (5.9.36) we need to show that $\mathbf{v}^{k+1}$ stays in $B_{2\delta_0}(\mathbf{u})$. To see this we know by assumption

$$\begin{aligned} & \|\mathbf{u} - \overline{\mathbf{u}}^j\|_{\ell_2} \leq \epsilon_j = 2^{-j}\delta_0 \\ & \|\mathbf{u} - \mathbf{u}^k(\overline{\mathbf{u}}^j)\|_{\ell_2} \leq \rho^k \|\mathbf{u} - \overline{\mathbf{u}}^j\|_{\ell_2} \leq \rho^k \epsilon_j \leq \delta_0 \\ \Rightarrow \quad & \mathbf{u}^k(\overline{\mathbf{u}}^j) \in B_{\delta_0}(\mathbf{u}) \,. \end{aligned} \tag{5.10.8}$$

Also we have $\mathbf{v}^0 = \overline{\mathbf{u}}^j \in B_{\delta_0}(\mathbf{u})$. Assuming that $\mathbf{v}^k \in B_{2\delta_0}(\mathbf{u})$, we can apply (5.10.7) and (5.9.36) (or (5.9.34)) to obtain

$$\|\mathbf{v}^{k+1} - \mathbf{u}^{k+1}\|_{\ell_2} \leq \eta_k + \rho\|\mathbf{v}^k - \mathbf{u}^k\|_{\ell_2} \leq \eta_k + \rho\left(\eta_{k-1} + \rho\|\mathbf{v}^{k-1} - \mathbf{u}^{k-1}\|_{\ell_2}\right) \,,$$

and repeating this argument gives

$$\|\mathbf{v}^{k+1} - \mathbf{u}^{k+1}\|_{\ell_2} \leq \sum_{l=0}^{k} \eta_{k-l}\rho^l = \sum_{l=0}^{k} w_{k-l}\overline{\rho}^{k-l}\delta\rho^l \,.$$

Taking $\hat{\rho} := \max\{\rho, \overline{\rho}\} \in (0, 1)$, we have (since $\delta = 2^{-j}\delta_0 = \epsilon_j$)

$$\|\mathbf{v}^{k+1} - \mathbf{u}^{k+1}\|_{\ell_2} \leq \epsilon_j \hat{\rho}^k \tag{5.10.9}$$

and also, by (5.10.8),

$$\|\mathbf{v}^{k+1} - \mathbf{u}\|_{\ell_2} \leq \|\mathbf{v}^{k+1} - \mathbf{u}^{k+1}\|_{\ell_2} + \|\mathbf{u}^{k+1} - \mathbf{u}\|_{\ell_2} \leq \epsilon_j\hat{\rho}^k + \epsilon_j\rho^{k+1} \leq 2\epsilon_j\hat{\rho}^k \leq 2\delta_0 \,.$$

Hence $\mathbf{v}^{k+1} \in B_{2\delta_0}(\mathbf{u})$. The rest follows by induction ($\epsilon_j \leq \epsilon_0 = \delta_0$).

To see when (5.10.5) is satisfied (i.e., the residual has been reduced enough), note $\mathbf{v}^{k+1} = \mathbf{v}^k + \mathbf{r}^k$. Hence

$$\begin{aligned} & \|\mathbf{r}^k\|_{\ell_2} \leq \|\mathbf{v}^{k+1} - \mathbf{u}\|_{\ell_2} + \|\mathbf{v}^k - \mathbf{u}\|_{\ell_2} \leq 2\epsilon_j\hat{\rho}^k + 2\epsilon_j\hat{\rho}^{k-1} \leq 4\epsilon_j\hat{\rho}^{k-1} \qquad (5.10.10) \\ \Rightarrow \quad & \beta\left(\eta_k + \|\mathbf{r}^k\|_{\ell_2}\right) \leq \beta\epsilon_j(w_k\hat{\rho}^k + 4\hat{\rho}^{k-1}) = \beta\epsilon_j\hat{\rho}^{k-1}(w_k\hat{\rho} + 4) \,. \end{aligned}$$

Hence, setting

$$K := \operatorname*{argmin}_k \left\{ \beta(w_k\hat{\rho} + 4)\hat{\rho}^{k-1} \leq \frac{1}{M^*} \right\} \tag{5.10.11}$$

we obtain that (5.10.5) holds after <u>at most</u> K steps.

Next let us see how accurate $\mathbf{v}^k$ is after branching to (iii). We have

$$\begin{aligned}\|\mathbf{u}-\mathbf{v}^k\|_{\ell_2} &= \|\mathbf{u}-\tilde{\mathbf{v}}\|_{\ell_2} \overset{(5.9.35)}{\leq} \beta\|\mathbf{CR}(\mathbf{v}^k)\|_{\ell_2} \leq \beta\left(\|\mathbf{CR}(\mathbf{v}^k)-\mathbf{r}^k\|_{\ell_2}+\|\mathbf{r}^k\|_{\ell_2}\right)\\ &\leq \beta\left(\eta_k+\|\mathbf{r}^k\|_{\ell_2}\right) \leq \frac{\delta}{M^*} = \frac{\epsilon_j}{M^*}\,.\end{aligned}$$

Therefore, by (5.10.2), we obtain for $\bar{\mathbf{u}} = \bar{\mathbf{u}}^{j+1}$

$$\begin{aligned}\|\mathbf{u}-\overline{\mathbf{u}}\|_{\ell_2} \leq \|\mathbf{u}-\tilde{\mathbf{v}}\|_{\ell_2}+\|\tilde{\mathbf{v}}-\overline{\mathbf{u}}\|_{\ell_2} &\leq \frac{\delta}{M^*}+\frac{b\delta}{M^*}\\ &= \frac{1+b}{M^*}\delta \leq \frac{\delta}{2} = \frac{\epsilon_j}{2} = \epsilon_{j+1} = 2^{-j-1}\delta_0\,,\end{aligned}$$

which advances the induction.

Hence, we have shown that one has for all j

$$\|\mathbf{u}-\overline{\mathbf{u}}^j\|_{\ell_2} \leq \epsilon_j = 2^{-j}\delta_0\,. \tag{5.10.12}$$

Thus the target accuracy ϵ is reached after $\lceil|\log_2\epsilon|+\log_2\delta_0\rceil$ steps. □

5.11 Complexity (VI)

It remains to see what the computational complexity of computing the sequences $\overline{\mathbf{u}}^j$, and eventually $\mathbf{u}(\epsilon)$, is. There are two issues:

(a) What is the role of COARSE?

(b) The computational complexity of SOLVE is determined by the cost of computing the approximate residuals $\mathbf{r}^k$ in step (ii.1). So what are the conditions on the realization of RES?

We first discuss (a):
One has to distinguish the linear and nonlinear case. In the latter case the output of COARSE has to satisfy certain constraints on the structure of the support. This is not necessary when $F = A$ is linear. Therefore we first consider the linear case where we have no constraints on $\operatorname{supp}\mathbf{v}_\eta$ in (5.10.2). Hence COARSE can be realized by the operator $\mathcal{C}_\eta$ from Exercise 4.5.3 based on thresholding, that is for $\#\operatorname{supp}\mathbf{v} < \infty$,

$$\text{COARSE}[\mathbf{v},\eta] \to \mathbf{v}_\eta := \mathcal{C}_\eta\mathbf{v}. \tag{5.11.1}$$

More precisely, sort entries by decreasing size of absolute values; starting with the smallest entries discard them as long as the sum of their squares is at most η^2

$$\mathbf{v} \rightsquigarrow (v_1^*,\ldots,v_N^*)\,, \qquad |v_i^*| \geq |v_{i+1}^*|\;.$$

Pick the largest l such that $(v_N^*)^2 + ... + (v_{N-l}^*)^2 \le \eta^2$. Then

$$\mathbf{v}_\eta = \left(v_j^*\right)_{j=1}^{N-l-1} .$$

Obviously, $\mathbf{v}_\eta$ satisfies

$$\begin{aligned} & \|\mathbf{v} - \mathbf{v}_\eta\|_{\ell_2} \le \eta \\ \text{and} \quad & \|\mathbf{v} - \mathbf{w}\|_{\ell_2} \le \eta \;\Rightarrow\; \#\operatorname{supp}\mathbf{w} \ge \#\operatorname{supp}\mathbf{v}_\eta . \end{aligned} \tag{5.11.2}$$

Theorem 5.11.1 (Coarsening Lemma). *Let* $\mathbf{u}, \mathbf{v} \in \ell_p(\Lambda)$ *so that*

$$\|\mathbf{u} - \mathbf{v}\|_{\ell_p} \le \delta . \tag{5.11.3}$$

Fix some $a > 0$ *and set* $\eta := (1+a)\delta$. *Then*

$$\|\mathbf{u} - \mathcal{C}_\eta \mathbf{v}\|_{\ell_p} \le (2+a)\delta = \eta + \delta \le 2\eta . \tag{5.11.4}$$

Moreover, whenever $\mathbf{u} \in w\ell_\tau \, (= \mathcal{A}^r_\infty)\, , \frac{1}{\tau} = r + \frac{1}{p}$*, we have*

$$\#(\operatorname{supp}\mathcal{C}_\eta\mathbf{v}) \le 2a^{-\frac{1}{r}}\delta^{-\frac{1}{r}} \|\mathbf{u}\|^{\frac{1}{r}}_{\mathcal{A}^r_\infty} \le 2\left(\frac{2+a}{a}\right)^{\frac{1}{r}} \|\mathbf{u}\|^{\frac{1}{r}}_{\mathcal{A}^r_\infty} \|\mathbf{u} - \mathcal{C}_\eta\mathbf{v}\|^{-\frac{1}{r}}_{\ell_p} \tag{5.11.5}$$

and

$$\|\mathcal{C}_\eta\mathbf{v}\|_{\mathcal{A}^r_\infty} \le C\,\|\mathbf{u}\|_{\mathcal{A}^r_\infty} , \qquad C = C(r,p,a) , \tag{5.11.6}$$

or equivalently

$$\|\mathcal{C}_\eta\mathbf{v}\|_{w\ell_\tau} \le C\,\|\mathbf{u}\|_{w\ell_\tau} , \qquad C = C(r,p,a) . \tag{5.11.7}$$

$$\left(\begin{array}{l} \text{Recall: Optimal rate of best n-term approximation} \approx \delta^{-\frac{1}{r}} \|\mathbf{u}\|^{\frac{1}{r}}_{w\ell_\tau} ;\ \text{Exercise 4.5.5} \\ \qquad \sigma_{n(\epsilon)}(\mathbf{u})_{\ell_p} \sim \epsilon \quad \Rightarrow \quad |\mathbf{u}|_{\mathcal{A}^r_\infty} = \sup_{n>0} n^r \sigma_n(\mathbf{u})_{\ell_p} \ge n(\epsilon)^r \sigma_{n(\epsilon)}(\mathbf{u})_{\ell_p} \sim n(\epsilon)^r \epsilon \\ \qquad\qquad\qquad\qquad \Rightarrow \quad n(\epsilon) \le \epsilon^{-\frac{1}{r}} |\mathbf{u}|^{\frac{1}{r}}_{\mathcal{A}^r_\infty} \end{array} \right)$$

Proof. (5.11.4) by triangle inequality.

To prove (5.11.5), let

$$n = n(a\delta) \tag{5.11.8}$$

be the smallest n such that $\sigma_n(\mathbf{u})_{\ell_p} \le a\delta$. By definition $\sigma_{n(a\delta)}(\mathbf{u})_{\ell_p} \le a\delta$ and $\sigma_{n(a\delta)-1}(\mathbf{u})_{\ell_p} > a\delta$. Hence

$$\begin{aligned} & |\mathbf{u}|_{\mathcal{A}^r_\infty} = \sup_{n>0} n^r \sigma_n(\mathbf{u})_{\ell_p} \ge (n(a\delta)-1)^r \sigma_{n(a\delta)-1}(\mathbf{u})_{\ell_p} \ge (n(a\delta)-1)^r a\delta \\ \Rightarrow \quad & (a\delta)^{-\frac{1}{r}} |\mathbf{u}|^{\frac{1}{r}}_{\mathcal{A}^r_\infty} \ge (n(a\delta)-1) \\ \Leftrightarrow \quad & n(a\delta) \le (a\delta)^{-\frac{1}{r}} |\mathbf{u}|^{\frac{1}{r}}_{\mathcal{A}^r_\infty} + 1 . \end{aligned}$$

Note

$$n(a\delta) > 0 \quad \Leftrightarrow \quad a\delta < \|\mathbf{u}\|_{\ell_p} \quad \Leftrightarrow \quad 1 < \left(\frac{\|\mathbf{u}\|_{\ell_p}}{a\delta}\right)^{\frac{1}{r}} .$$

Employing $a^\mu + b^\mu \leq c(\mu)(a+b)^\mu$ for $\mu = \frac{1}{r}$, yields

$$n(a\delta) \leq (a\delta)^{-\frac{1}{r}} \left(|\mathbf{u}|_{\mathcal{A}^r_\infty}^{\frac{1}{r}} + \|\mathbf{u}\|_{\ell_p}^{\frac{1}{r}}\right) \leq (a\delta)^{-\frac{1}{r}} C(r) \left(|\mathbf{u}|_{\mathcal{A}^r_\infty} + \|\mathbf{u}\|_{\ell_p}\right)^{\frac{1}{r}} ,$$

with

$$C(r) \leq \begin{cases} 1, & r < 1 \Leftrightarrow \frac{1}{r} > 1 \\ 2, & r > 1 \rightarrow \frac{1}{2}\left(a^\mu + b^\mu\right) \leq \left(\frac{a+b}{2}\right)^\mu \text{by concavity} \end{cases} .$$

Then

$$n(a\delta) \leq 2(a\delta)^{-\frac{1}{r}} \|\mathbf{u}\|_{\mathcal{A}^r_\infty}^{\frac{1}{r}} . \tag{5.11.9}$$

Denoting by $\Lambda(a\delta) = \operatorname{supp} \mathcal{C}_{a\delta}\mathbf{u}$ one has $\mathcal{C}_{a\delta}\mathbf{u} = P_{\Lambda(a\delta)}\mathbf{u}$, i.e., $\mathcal{C}_{a\delta}$ is just the projection operator onto $\Lambda(a\delta)$. Hence

$$\begin{aligned} \left\|\mathbf{v} - P_{\Lambda(a\delta)}\mathbf{v}\right\|_{\ell_p} &= \left\|(I - P_{\Lambda(a\delta)})(\mathbf{v} - \mathbf{u}) + (I - P_{\Lambda(a\delta)})\mathbf{u}\right\|_{\ell_p} \\ &\leq \left\|(I - P_{\Lambda(a\delta)})(\mathbf{v} - \mathbf{u})\right\|_{\ell_p} + \underbrace{\left\|(I - P_{\Lambda(a\delta)})\mathbf{u}\right\|_{\ell_p}}_{\sigma_{n(a\delta)}(\mathbf{u})_{\ell_p} \leq a\delta} \\ &\leq \underbrace{\|\mathbf{v} - \mathbf{u}\|_{\ell_p}}_{\leq \delta \text{ by assumption}} + a\delta \\ &\leq (1+a)\delta = \eta . \end{aligned}$$

So $P_{\Lambda(a\delta)}\mathbf{v}$ realizes $\left\|\mathbf{v} - P_{\Lambda(a\delta)}\mathbf{v}\right\|_{\ell_p} \leq \eta$. By (5.11.2), one obtains

$$n(a\delta) = \# \operatorname{supp} P_{\Lambda(a\delta)}\mathbf{v} \geq \# \operatorname{supp} \mathcal{C}_\eta \mathbf{v} ,$$

and by (5.11.9)

$$\# \operatorname{supp} \mathcal{C}_\eta \mathbf{v} \leq 2(a\delta)^{-\frac{1}{r}} \|\mathbf{u}\|_{\mathcal{A}^r_\infty} , \tag{5.11.10}$$

which is the first part of (5.11.5).

By (5.11.4)

$$\begin{aligned} \eta + \delta = (2+a)\delta = \left(\frac{2+a}{a}\right) a\delta \geq \|\mathbf{u} - \mathcal{C}_\eta \mathbf{v}\|_{\ell_p} \\ \Rightarrow \quad \frac{a}{2+a} \|\mathbf{u} - \mathcal{C}_\eta \mathbf{v}\|_{\ell_p} \leq a\delta , \end{aligned}$$

which is the second part of (5.11.5).

It remains to prove the stability estimate (5.11.7). To this end, note that (5.11.5) holds with a uniform constant for all $a \geq a_0$ as long as $a_0 > 0$ is fixed. Therefore consider

$$a = a_l = 2^l, \eta_l := (1 + 2^l)\delta, \quad \mathbf{v}^\mathbf{l} := \mathcal{C}_{\eta_\mathbf{l}} \mathbf{v}^\mathbf{l}, \quad \mathbf{l} = \mathbf{0}, \ldots, |\log_\mathbf{2}(\|\mathbf{v}\|_{\ell_2}/\delta)|.$$

Then we know from (5.11.5) that

$$\#(\operatorname{supp} \mathbf{v}^\mathbf{l}) \leq \left\lceil \mathbf{2}\delta^{-\mathbf{1/r}} \mathbf{2}^{-\mathbf{1/r}} \|\mathbf{u}\|_{\mathcal{A}^\mathbf{r}_\infty}^{\mathbf{1/r}} \right\rceil =: \mathbf{m}_\mathbf{l}. \tag{5.11.11}$$

To prove (5.11.7) it suffices to show that

$$|\mathcal{C}_\eta \mathbf{v}|_{\mathcal{A}^r_\infty} = \sup_{n \in \mathbb{N}} n^r \sigma_n(\mathcal{C}_\eta \mathbf{v})_{\ell_2} \lesssim \|\mathbf{u}\|_{\mathcal{A}^r_\infty}. \tag{5.11.12}$$

Since $\#(\operatorname{supp} \mathcal{C}_\eta \mathbf{v}| l e m_0$ we have $\sigma_n(\mathcal{C}_\eta \mathbf{v})_{\ell_2} = 0$ for $n > m_0$ so that we only need to consider $n = m_l$ for l in the above range. Then

$$\begin{aligned} m_l^r \sigma_{m_l}(\mathcal{C}_\eta \mathbf{v})_{\ell_2} &\leq m_l^r \|\mathcal{C}_\eta \mathbf{v} - \mathcal{C}_{\eta_l} \mathbf{v}\|_{\ell_2} \\ &\leq m_l^r \|\mathbf{v} - \mathcal{C}_{\eta_l} \mathbf{v}\|_{\ell_2} \leq m_l^r \eta_l \\ &\leq 2^r (1 + 2^{-l} \|\mathbf{u}\|_{\mathcal{A}^r_| \mathtt{infty}}, \end{aligned}$$

which completes the proof. □

Corollary 5.11.1. *In summary, we have shown that if* COARSE$[\mathbf{v}, \eta] \to \mathbf{v}_\eta$ *is realized by* $\mathcal{C}_\eta$ *and if we know that* $\|\mathbf{u} - \mathbf{v}\|_{\ell_2} \leq \delta$*, then for* $\eta = (1 + a)\delta$*,* $a > 0$ *fixed,* $\mathbf{v}_\eta :=$ COARSE$[\mathbf{v}, \eta]$ *satisfies*

$$\begin{aligned} &\|\mathbf{u} - \mathbf{v}_\eta\|_{\ell_2} \leq (2 + a)\delta \\ &\#(\operatorname{supp} \mathbf{v}_\eta) \lesssim \eta^{-\frac{1}{r}} \|\mathbf{u}\|_{w\ell_\tau}^{\frac{1}{r}}, \qquad \|\mathbf{v}_\eta\|_{w\ell_\tau} \lesssim \|\mathbf{u}\|_{w\ell_\tau} \end{aligned} \tag{5.11.13}$$

where $\frac{1}{\tau} = r + \frac{1}{2}$*.*

Proof. Follows from Theorem 5.11.1 for $p = 2$. □

The relevance of Theorem 5.11.1 can be explained as follows: Suppose one has an approximation $\mathbf{v}$ of some unknown function $\mathbf{u}$ and suppose one knows that $\|\mathbf{u} - \mathbf{v}\|_{\ell_2} \leq \delta$ (in Algorithm 5.10.1 this is $\mathbf{v} = \mathbf{v}^\mathbf{k}$). Then coarsening (the given sequence) $\mathbf{v}$ with a somewhat coarser accuracy yields an approximation to $\mathbf{u}$ which is optimal in the $\mathcal{A}^r_\infty$-class sense. The input to step (iii) of SOLVE plays exactly this role.

The next observation is that when the computation of $\mathbf{r}^k$, or more generally, of RES$[\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{v}, \eta]$, has a comparable complexity, then one obtains the dream theorem. So we proceed as follows:

1. Formulate certain sparsity conditions on the routine RES;
2. show that under these conditions one indeed obtains the dream theorem;
3. then, for a given variational problem realize the routine RES with the desired properties.

As for (i), we require that RES so that is has the following properties analogous to the routine COARSE.

Definition 5.11.1. *The routine* RES$[\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{v}, \eta]$ *is called* **s*-computable** *if for any finitely supported input* $\mathbf{v}$ *the output*

$$\mathbf{r}_\eta = \mathrm{RES}[\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{v}, \eta]$$

satisfies for any $0 < s < s^*$ *and* $\mathbf{u} \in \mathcal{A}^s_\infty((\Sigma_n), \ell_2(\Lambda)) \left(= w\ell_\tau(\Lambda), \frac{1}{\tau} = s + \frac{1}{2}\right)$

$$\begin{aligned} \|\mathbf{r}_\eta\|_{\mathcal{A}^s_\infty} &\leq C \left(\|\mathbf{v}\|_{\mathcal{A}^s_\infty} + \|\mathbf{u}\|_{\mathcal{A}^s_\infty}\right) \\ \#(\operatorname{supp} \mathbf{r}_\eta) &\leq C\eta^{-\frac{1}{s}} \left(\|\mathbf{v}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + \|\mathbf{u}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty}\right) \\ \#\,\mathrm{flops}(\mathbf{r}_\eta) &\leq C\left(\#(\operatorname{supp} \mathbf{v}) + \#(\operatorname{supp} \mathbf{r}_\eta)\right) \end{aligned} \tag{5.11.14}$$

for $\mathbf{F}$ *linear and*

$$\begin{aligned} \|\mathbf{r}_\eta\|_{\mathcal{A}^s_\infty} &\leq C \left(\|\mathbf{v}\|_{\mathcal{A}^s_\infty} + \|\mathbf{u}\|_{\mathcal{A}^s_\infty} + 1\right) \\ \#(\operatorname{supp} \mathbf{r}_\eta) &\leq C\eta^{-\frac{1}{s}} \left(\|\mathbf{v}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + \|\mathbf{u}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + 1\right) \\ \#\,\mathrm{flops}(\mathbf{r}_\eta) &\leq C\left(\#(\operatorname{supp} \mathbf{v}) + \#(\operatorname{supp} \mathbf{r}_\eta)\right) \end{aligned} \tag{5.11.14a}$$

for $\mathbf{F}$ *nonlinear for some* C *independent of* $\mathbf{u}, \mathbf{v}$ *and depending on* s *only as* $s \to s^*$.

Theorem 5.11.2. *Given the system*

$$\mathbf{F}(\mathbf{u}) = \mathbf{f},$$

assume that (5.9.35), (5.9.36) *hold and* COARSE *satisfies* (5.11.13). *If in addition* RES *is* s^**-computable, then the following is true:*
The output $\mathbf{u}(\epsilon)$ *produced by* SOLVE$[\mathbf{f}, \mathbf{F}, \mathbf{C}, \mathbf{u}^0, \epsilon]$ *satisfies for each* $\epsilon > 0$

$$\|\mathbf{u} - \mathbf{u}(\epsilon)\|_{\ell_2} \leq \epsilon$$

and

$$\#(\operatorname{supp} \mathbf{u}(\epsilon)) \leq C\epsilon^{-\frac{1}{s}} \left(\|\mathbf{u}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + 1\right), \qquad \|\mathbf{u}(\epsilon)\|_{\mathcal{A}^s_\infty} \leq C\,\|\mathbf{u}\|_{\mathcal{A}^s_\infty} \tag{5.11.15}$$

whenever $\mathbf{u} \in \mathcal{A}^s_\infty$ *for some* $0 < s < s^*$.
Moreover

$$\#\,\mathrm{flops}\,(\mathbf{u}(\epsilon)) \lesssim \epsilon^{-\frac{1}{s}}\, \|\mathbf{u}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} \sim n(\epsilon)$$

where $n(\epsilon)$ *is minimal such that* $\sigma_{n(\epsilon)}(\mathbf{u})_{\ell_2} \leq \epsilon$.

Proof. The first part follows directly from Proposition 5.10.1. Now consider the rest of the assertion.

Bounding sparsity norms and supports: We have already shown that after at most K inner iterations in step (ii) one branches to (iii) COARSE. Again denote by j the number of times an approximation enters (ii). First we assume that our initial guess has been already subjected to COARSE ending up with the bound δ_0. Therefore we can assume by (5.11.13) that $\overline{\mathbf{u}} = \overline{\mathbf{u}}^0$ satisfies, for some $0 < s < s^*$, $j = 0$,

$$\|\overline{\mathbf{u}}^j\|_{\mathcal{A}^s_\infty} \lesssim \|\mathbf{u}\|_{\mathcal{A}^s_\infty} , \qquad \# \operatorname{supp} \overline{\mathbf{u}}^j \lesssim \epsilon_j^{-\frac{1}{s}} \|\mathbf{u}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} . \tag{5.11.16}$$

Idea: Show by induction that this remains true for a fixed uniform constant for $j+1$.

So suppose (5.11.16) holds for j. Since RES is s^*-computable we conclude that $\mathbf{r}^0 = \text{RES}[\mathbf{f}, \mathbf{F}, \mathbf{C}, \overline{\mathbf{u}}^j, \eta_0]$ satisfies

$$\begin{aligned}
\|\mathbf{r}^0\|_{\mathcal{A}^s_\infty} &\leq C \left(\|\overline{\mathbf{u}}^j\|_{\mathcal{A}^s_\infty} + \|\mathbf{u}\|_{\mathcal{A}^s_\infty} + 1 \right) \\
\#(\operatorname{supp} \mathbf{r}^0) &\leq C \eta_0^{-\frac{1}{s}} \left(\|\overline{\mathbf{u}}^j\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + \|\mathbf{u}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + 1 \right) \\
\# \operatorname{flops}(\mathbf{r}^n) &\leq C (\# \operatorname{supp} \overline{\mathbf{u}}^j + \#(\operatorname{supp} \mathbf{r}^0)) \\
&\leq C \eta_0^{-\frac{1}{s}} \left(\|\overline{\mathbf{u}}^j\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + \|\mathbf{u}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + 1 \right) .
\end{aligned}$$

Notice $\eta_0 \sim \epsilon_j$, by (5.11.16) $\left\|\overline{\mathbf{u}}^j\right\|_{\mathcal{A}^s_\infty} \lesssim \|\mathbf{u}\|_{\mathcal{A}^s_\infty}$. Hence

$$\begin{aligned}
\|\mathbf{r}^0\|_{\mathcal{A}^s_\infty} &\leq C \left(\|\mathbf{u}\|_{\mathcal{A}^s_\infty} + 1 \right) \\
\#(\operatorname{supp} \mathbf{r}^0) &\leq C \epsilon_j^{-\frac{1}{s}} \left(\|\mathbf{u}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + 1 \right) \\
\# \operatorname{flops}(\mathbf{r}^0) &\leq C \epsilon_j^{-\frac{1}{s}} \left(\|\mathbf{u}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + 1 \right) .
\end{aligned}$$

So $\mathbf{v}^1 := \overline{\mathbf{u}}^j + \mathbf{r}^0$ has an analogous bound (that is satisfies (5.11.16)).

Repeating this argument, shows that $\tilde{\mathbf{v}} = \mathbf{v}^k$ $(k \leq K)$ satisfies (5.10.5) and

$$\begin{aligned}
\|\tilde{\mathbf{v}}\|_{\mathcal{A}^s_\infty} &\leq C_K \left(\|\mathbf{u}\|_{\mathcal{A}^s_\infty} + 1 \right) \\
\#(\operatorname{supp} \tilde{\mathbf{v}}) &\leq C_K \epsilon_j^{-\frac{1}{s}} \left(\|\mathbf{u}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + 1 \right) \\
\# \operatorname{flops}(\tilde{\mathbf{v}}) &\leq C_K \epsilon_j^{-\frac{1}{s}} \left(\|\mathbf{u}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + 1 \right) ,
\end{aligned} \tag{5.11.17}$$

i.e., the total computational work in (ii) stays proportional to $\epsilon_j^{-\frac{1}{s}} \left(\|\mathbf{u}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + 1 \right)$.

Remark 5.11.1. *The constant C_K could grow in each step of the inner loop (ii) in Algorithm 5.10.1. However, we already know that the inner cycle has at most a fixed uniform number K of steps. Thus, when exiting to step (ii) the coarsening step* COARSE *resets the constant according to the coarsening lemma. This ensures a uniform control of all constants.*

Now when $\tilde{\mathbf{v}}$ enters (iii) we know

$$\|\mathbf{u} - \tilde{\mathbf{v}}\|_{\ell_2} \leq \frac{\epsilon_j}{M^*}.$$

By (5.11.13) in Corollary 5.11.1 we know that for $b > 1$ and $\frac{b+1}{M^*} \leq \frac{1}{2}$ the next input for (ii)

$$\text{COARSE}\left[\tilde{\mathbf{v}}, \frac{b\epsilon_j}{M^*}\right] =: \overline{\mathbf{u}}^{j+1}$$

satisfies

$$\|\overline{\mathbf{u}}^{j+1}\|_{\mathcal{A}^s_\infty} \leq C\,\|\mathbf{u}\|_{\mathcal{A}^s_\infty}$$

$$\left.\begin{array}{r}\#\,\text{supp}\,\overline{\mathbf{u}}^{j+1} \\ \#\,\text{flops}(\overline{\mathbf{u}}^{j+1})\end{array}\right\} \leq C\epsilon_{j+1}^{-\frac{1}{s}}\left(\|\mathbf{u}\|_{\mathcal{A}^{\frac{1}{s}}_\infty} + 1\right),$$

where C is independent of K and C_K but depends only on s, as s tends to $0, s^*$. This advances the induction over j.

Bounding the computational work: It remains to estimate the total computational work. To that end, recall: $\epsilon_j = 2^{-j}\delta_0$ and let j_0 be the smallest integer so that $\epsilon_{j_0} = 2^{-j_0}\delta_0 \leq \epsilon$. Then $2\,2^{-j_0}\delta_0 > \epsilon$ and

$$\text{work in } j^{\text{th}} \text{ block (ii)} \;=:\; w_j \;\leq\; \underbrace{C_K\left(\|\mathbf{u}\|^{\frac{1}{s}}_{\mathcal{A}^s_\infty} + 1\right)}_{=:C(\mathbf{u})}\epsilon_j^{-\frac{1}{s}}$$

Thus

$$\begin{aligned}\text{total work } &\leq \sum_{j=0}^{j_0} w_j \leq \sum_{j=0}^{j_0} C(\mathbf{u})\epsilon_j^{-\frac{1}{s}} \\ &= C(\mathbf{u})\sum_{j=0}^{j_0}(2^{-j}\delta_0)^{-\frac{1}{s}} = C(\mathbf{u})\delta_0^{-\frac{1}{s}} \quad \underbrace{\sum_{j=0}^{j_0} 2^{\frac{j}{s}}}_{=\frac{2^{\frac{j_0+1}{s}}-1}{2^{\frac{1}{s}}-1}\sim 2^{\frac{j_0}{s}}} \\ &\sim C(\mathbf{u})\left(\delta_0 2^{-j_0}\right)^{-\frac{1}{s}} \sim C(\mathbf{u})\epsilon_{j_0}^{-\frac{1}{s}} \sim C(\mathbf{u})\epsilon^{-\frac{1}{s}}\end{aligned}$$

which yields (5.11.15). □

Comments 5.11.1. *(i) Thus, under Assumption 5.10.1 and under the assumption that the routine* RES *is* s^**-computable for some* $s^* = s^*(\Psi)$*, Algorithm 5.10.1 is class-optimal for* $s < s^*$*. This means when the solution* u *belongs to* $\mathcal{A}^s_\infty((\Sigma_n), \mathbb{U})$ *(meaning that the coefficient sequence* $\mathbf{u}$ *belongs to* $\mathcal{A}^s_\infty((\Sigma_n), \ell_2(\Lambda))$*, see Exercise 4.4.2) then the algorithm outputs for any given target accuracy* ϵ *an approximate solution at class-optimal complexity* $\sim \epsilon^{-1/s}$*.*

(ii) One can derive from Remark 4.6.6, 3. that for $\mathbb{U} = H^1(\Omega)$ *(or* $H^1_0(\Omega)$*) and*

$$\frac{1}{\tau} = \frac{t}{d} + \frac{1}{2}$$

one has (with $t = ds$*)*

$$\|u\|_{B^{1+ds}_\tau(L_\tau))} \sim \|\mathbf{u}\|_{\ell_\tau(\Lambda)}, \tag{5.11.18}$$

since now the base space is H^1 *and not* L_2*. We also know from Theorem 4.5.4, Remark 4.5.1, and Exercise 4.5.3, that*

$$\mathcal{A}^s_\infty((\Sigma_n), \mathbb{U}) \hat{=} \mathcal{A}^s_\infty((\Sigma_n), \ell_2) \hat{=} w\ell_\tau(\Lambda) \supset \ell_\tau(\Lambda),$$

i.e., whenever the solution u *belongs to the Besov space* $B^{1+t}_\tau(L_\tau))$ *it is approximated with accuracy* ϵ *with optimal rate* $\epsilon^{-1/s}$*. It has been shown in [21] that the solutions to elliptic boundary value problems on Lipshitz domains have in general higher Besov regularity than Sobolev regularity, which means that an adaptive method realizes the same target accuracy at a strictly better rate.*

(iii) Now it remains to construct for a given problem routines RES *which are indeed* s^**-computable for possibly high* s^**. This will be done in the next section.*

6 Realization of RES

6.1 Overview:

Every realization of RES hinges on the following ingredients

(i) Approximation of data $\mathbf{f} = (\langle f, \psi_\lambda \rangle)_{\lambda \in \Lambda}$

(ii) Application of preconditioner $\mathbf{C}$

(iii) Application of $\mathbf{F}$.

(i): The point of view taken here is that the data are *known completely* (being provided by the user), and $\mathbf{f}$ is computed in a preprocessing step whose complexity is not counted for the performance of SOLVE!
Why? There is no way to estimate in general the complexity of approximating any given function without any a priori knowledge about it. Therefore the preprocessing has typically the following structure:

Fix any final target tolerance $\hat{\epsilon}$ such that for some finite index set $\Lambda_f(f, \hat{\epsilon}) \subset \Lambda$, $\widehat{\mathbf{f}} := \mathbf{f}|_{\Lambda_f}$ satisfies

$$\|\mathbf{f} - \widehat{\mathbf{f}}\|_{\ell_2} \leq \hat{\epsilon} \tag{6.1.1}$$

The claim is that this implies a certain accuracy of the output of SOLVE$[\widehat{\mathbf{f}}, \mathbf{F}, \mathbf{C}, \epsilon]$.

To see this, denote by $\hat{\mathbf{u}}$ the exact solution of

$$\mathbf{F}(\hat{\mathbf{u}}) = \widehat{\mathbf{f}} \qquad (\text{as opposed to } \mathbf{F}(\mathbf{u}) = \mathbf{f})\,.$$

Then, by (5.9.35)

$$\|\mathbf{u} - \hat{\mathbf{u}}\|_{\ell_2} \leq \beta \, \|\mathbf{C}\mathbf{R}(\hat{\mathbf{u}})\|_{\ell_2} \,, \qquad \hat{\mathbf{u}} \in B_{\delta_0}(\mathbf{u})\,.$$

Now

$$\mathbf{R}(\hat{\mathbf{u}}) = \mathbf{f} - \mathbf{F}(\hat{\mathbf{u}}) = \mathbf{f} - \widehat{\mathbf{f}},$$

so that

$$\begin{aligned} \|\mathbf{u} - \hat{\mathbf{u}}\|_{\ell_2} &\leq \beta \, \|\mathbf{C}\|_{\mathcal{L}(\ell_2,\ell_2)} \left\|\mathbf{f} - \widehat{\mathbf{f}}\right\|_{\ell_2} \\ &\leq \beta \, \|\mathbf{C}\|_{\mathcal{L}(\ell_2,\ell_2)} \, \hat{\epsilon}\,. \end{aligned}$$

Therefore, whenever seeking only a target accuracy $\epsilon \geq 2\beta \, \|\mathbf{C}\|_{\mathcal{L}(\ell_2,\ell_2)} \, \hat{\epsilon} =: 2\hat{\epsilon}_0$, say, one can work with the perturbed data, for $\eta \geq 2\hat{\epsilon}_0$. So one could use a routine

$$\text{DATA}[\mathbf{f}, \eta] \to \mathbf{f}_\eta \text{ such that } \|\mathbf{f} - \mathbf{f}_\eta\|_{\ell_2} \leq \eta \tag{6.1.2}$$

to be realized as follows

$$\mathbf{f}_\eta := \text{COARSE}[\widehat{\mathbf{f}}, \eta - \hat{\epsilon}]\,.$$

Then

$$\|\mathbf{f} - \mathbf{f}_\eta\|_{\ell_2} \leq \underbrace{\|\mathbf{f} - \widehat{\mathbf{f}}\|_{\ell_2}}_{\leq \hat{\epsilon}} + \underbrace{\|\widehat{\mathbf{f}} - \mathbf{f}_\eta\|_{\ell_2}}_{\leq \eta - \hat{\epsilon}} \leq \eta\,.$$

(ii): The application of $\mathbf{C}$ depends on its particular structure. In general it requires approximate application of an infinite matrix. This is treated under (iii).

For now we consider the simplest case

$$\mathbf{C} = \alpha \mathbf{I}$$

for some $\alpha > 0$.

(iii): the *application of the operator in Riesz coordinates* is a crucial ingredient which we discuss now in a little more detail.

6.2 Near-sparsity of operator representations

We show next that the representation

$$\mathbf{F}(\mathbf{u}) = (\langle f(u), \psi_\lambda\rangle)_{\lambda\in\Lambda}$$

is for a wide range of operators F and certain *multilevel-type* Riesz bases Ψ *nearly sparse* in the sense that many entries of $\mathbf{F}(\mathbf{u})$ are so small that they can be replaced by zero without affecting the mapping properties of $\mathbf{F}$.

The key properties of the Riesz bases for which the sparsity effect holds are the following:

- $\Psi \subset \mathbb{U}$ satisfies the *norm equivalencees* (NE) for the underlying trial space $\mathbb{U}$, see Section 4.6.3.
- *Multilevel-locality* of the basis functions $\psi_\lambda \in \Psi$, i.e.,

  $$\operatorname{diam}(S_\lambda) \sim 2^{-|\lambda|}, \lambda \in \Lambda, \tag{LOC}$$

  uniformly in λ. Here the indices λ encode different types of information, namely the dyadic refinement level $|\lambda|$, the spatial location $x(\lambda)$ (e.g. the center of gravity of the support S_λ), and the "type" $\mathbf{e}(\lambda)$ of the function ψ_λ (typically needed for spatial dimension $d > 1$).
- *Cancellation property:* Suppose that $\mathbb{U}$ is a closed subspace of $H^t(\Omega)$. Then inner products with the ψ_λ annihilate smooth parts. More precisely, there exists an integer $\tilde{m} \in \mathbb{N}$ such that for $1 \le p \le \infty$

  $$|\langle v, \psi_\lambda\rangle| \lesssim 2^{-|\lambda|\left(\tilde{m}+t+\frac{d}{2}-\frac{d}{p}\right)} |v|_{W^{\tilde{m}}(L_p(S_\lambda))}, \quad v \in W^{\tilde{m}}(L_p(\Omega)). \tag{CP}$$

Here are some comments:

(NE): We have seen that the relevant Hilbert spaces are Sobolev spaces, or closed subspaces of Sobolev spaces (e.g. determined by homogeneous boundary conditions) or by Cartesian products of such spaces. Thus, it suffices to consider the case where the (infinite dimensional) trial space $\mathbb{U}$ is such a (closed subspace of a) Sobolev space $H^t(\Omega)$ (whose intersection with $L_2(\Omega)$ is dense in $L_2(\Omega)$). Furthermore, recall from Exercise 4.6.8 that a wavelet-type Riesz-basis for $\mathbb{U}$ can then be obtained by rescaling an L_2-Riesz Ψ°, basis provided that $\Psi^\circ \subset \mathbb{U}$. More precisely, we have shown in Theorem 4.6.5 the following:

Remark 6.2.1. *The collections*

$$\Psi^s = \{\psi_\lambda^s : \lambda \in \Lambda\}, \quad \psi_\lambda^s = \frac{\psi_\lambda^\circ}{\|\psi_\lambda^\circ\|_{H^s(\Omega)}} \sim 2^{-s|\lambda|}\psi_\lambda^\circ, \tag{6.2.1}$$

form Riesz bases for the whole scale of spaces $\mathbb{U} \cap H^s(\Omega)$, $-\tilde{s}_2 < s < s_2$, *including* $\mathbb{U}$ *when* $t \in (-\tilde{s}_2, s_2)$.

Exercise 6.2.1. *Show that properly rescaled Haar basis functions form a Riesz basis for* $H^s(0,1)$*, for* $-\frac{1}{2} < s < \frac{1}{2}$*.*

(LOC): Regarding the structure of the multilevel index sets recall the Haar basis on $\Omega = (0,1)$:

$$\Psi := \{\psi_\lambda\,, \lambda \in \Lambda\}\,, \Lambda = \{(-1,0),(j,k) : k = 0,...,2^j-1, j = 0,1,2,...\}$$

where

$$\begin{aligned}\phi(x) &:= \chi_{[0,1]}(x) =: \phi_{0,0}(x) =: \psi_{-1,0}\\ \phi_{j,k}(x) &:= 2^{\frac{j}{2}}\phi\left(2^j x - k\right)\,, \qquad \|\phi_{j,k}\|_{L_2(0,1)} = 1\,, j \in \mathbb{N}_0\,, k = 0,..,2^j-1\\ \psi_{0,0}(x) &:= \phi(2t) - \phi(2t-1)\\ \psi_{j,k}(x) &:= 2^{\frac{j}{2}}\psi(2^j x - k)\,.\end{aligned}$$

Here we have $\lambda = (j,k)$, $|\lambda| = j, x(\lambda) = k$.

Analogously, one can construct a Haar basis on $\Omega = (0,1)^2$, using tensor products

$$\phi_{j,(k_1,k_2)}(x,y) := \phi_{j,k_1}(x)\phi_{j,k_2}(y), \quad V_j = \text{span}\{\phi_{j,k} : 0 \le k_1, k_2 \le 2^j-1\}$$

To span orthogonal complements W_j such that $V_{j+1} = V_j \oplus W_j$, observe

$$\underbrace{\dim V_{j+1}}_{=2^{2(j+1)}} = \underbrace{\dim V_j}_{=2^{2j}} + \dim W_j \quad \Rightarrow \qquad\qquad \dim W_j = 3\cdot 2^{2j}\,.$$

Hence we need three types of wavelets

$$\begin{aligned}\psi_{j,(1,0),\mathbf{k}}(x,y) &= \psi_{j,k_1}(x)\phi_{j,k_2}(y)\\ \psi_{j,(0,1),\mathbf{k}}(x,y) &= \phi_{j,k_1}(x)\psi_{j,k_2}(y)\\ \psi_{j,(1,1),\mathbf{k}}(x,y) &= \psi_{j,k_1}(x)\psi_{j,k_2}(y)\,.\end{aligned}$$

So we get the index structure

$$\Lambda = \{\lambda = (j,\mathbf{e},\mathbf{k}) : \mathbf{e} \in \{0,1\}^d, \mathbf{k} \in \{0,...,2^j-1\}^d, j = -1,0,1,...\}\,.$$

Again $|\lambda| = j$ is the level of λ, $\mathbf{e}$ denotes the *type*, and $\mathbf{k}$ the *spatial shift*.

Clearly we have

$$\text{diam}(\text{supp}\,\psi_\lambda) = \sqrt{2\,2^{-2|\lambda|}} = \sqrt{2}\,2^{-|\lambda|} \sim 2^{-|\lambda|}\,.$$

This should suffice at this point to illustrate the index structure and multilevel-locality (LOC). Since the Haar functions are discontinuous they will not form a Riesz basis for smoother spaces such as $H^1(\Omega)$.

To obtain Riesz bases for spaces $H^s(\Omega)$, $s \geq 1/2$, one needs smoother basis functions. The following considerations refer to the type of multilevel Riesz bases discussed in Section 4.6.3. For concrete constructions of such bases for realistic domains, see e.g. [10, 11, 24, 25, 26, 27, 29]. Here we only need to assume in the sequel the following key properties shared by all these constructions (see also the discussion of the Haar basis in Section 4.2):

(CP): The Haar basis functions have been shown to satisfy the cancellation property (CP) for $\tilde{m} = 1$ which was derived using a vanishing moment, see Exercise 4.2.4. This can be generalized as follows.

Remark 6.2.2. *The Riesz bases constructed in Section 4.6.3 rely on a pair of dual multiresolution sequences and associated dual pairs of bases $\Psi, \tilde{\Psi}$, see Remark 4.6.8. Moreover, the two multiresolution sequences are required to have some polynomial exactness $m, \tilde{m}$, respectively, see Definition 4.6.3,* (4.6.26). *In this case the primal basis functions ψ_λ have* vanishing moments *of order $\tilde{m}$, i.e.,*

$$\langle v, \psi_\lambda \rangle = 0, \quad \forall v \in \mathbb{P}_{\tilde{m}},\ |\lambda| \geq 1. \tag{VM}$$

Proof. Since any polynomial $g \in \mathbb{P}_{\tilde{m}}(S_\lambda)$ can be represented as a linear combination of the scaling functions generating the dual multiresolution space $\tilde{V}_{|\lambda|}$ and since by Remark 4.6.9, the ψ_λ are orthogonal to $\tilde{V}_{|\lambda|}$ which yields (VM). □

Proposition 6.2.1. *Assume that the dual multiresolution approximation $(\tilde{V}_j)_{j\in\mathbb{N}_0}$ corresponding to the dual Riesz basis has exactness order $\tilde{m}$. and that primal basis Ψ is a Riesz basis for (a closed subspace $\mathbb{U}$ of) $H^t(\Omega) = B^t_2(L_2(\Omega))$. Then the cancellation properties (CP) of order $\tilde{m}$ hold:*

$$|\langle v, \psi_\lambda \rangle| \lesssim 2^{-|\lambda|\left(\tilde{m}+t+\frac{d}{2}-\frac{d}{p}\right)} |v|_{W^{\tilde{m}}(L_p(S_\lambda))}, \quad v \in W^{\tilde{m}}(L_p(\Omega)),\ 1 \leq p \leq \infty. \tag{6.2.2}$$

Proof. By Remark 6.2.2, one has vanishing moments of order $\tilde{m}$ (VM). Hence,

$$\begin{aligned} |\langle v, \psi_\lambda \rangle| \quad &= \quad \inf_{p\in\mathbb{P}_{\tilde{m}-1}} |\langle v - P, \psi_\lambda \rangle| \\ &\overset{\text{Hölder}}{\leq} \quad \inf_{p\in\mathbb{P}_{\tilde{m}-1}} \|v - P\|_{L_p(S_\lambda)} \|\psi_\lambda\|_{L_q(S_\lambda)}\,, \end{aligned} \tag{6.2.3}$$

where $\frac{1}{p} + \frac{1}{q} = 1$. Now we use the fact that the ψ_λ are obtained by rescaling an L_2-Riesz basis $\Psi^\circ = \{\psi^\circ_\lambda\}_{\lambda\in\Lambda}$, i.e., $\psi_\lambda = 2^{-t|\lambda|}\psi^\circ_\lambda \sim \psi^\circ_\lambda / \|\psi^\circ_\lambda\|_{H^t(\Omega)}$ (Exercise 4.6.8). From (4.6.41) and (4.6.42) we know that

$$\|\psi^\circ_\lambda\|_{L_q(S_\lambda)} \sim 2^{|\lambda|\left(\frac{d}{q}-\frac{d}{2}\right)} = 2^{|\lambda|\left(\frac{d}{2}-\frac{d}{p}\right)} \;\Rightarrow\; \|\psi_\lambda\|_{L_q(S_\lambda)} \sim 2^{|\lambda|\left(\frac{d}{q}-\frac{d}{2}-t\right)}. \tag{6.2.4}$$

Next we invoke Whitney's Theorem (Deny-Lions) (3.1.6)

$$\inf_{P\in\Pi_{\tilde{m}-1}} \|v - P\|_{L_p(S_\lambda)} \leq C(p, d, \tilde{m})\, \operatorname{diam}(S_\lambda)^{\tilde{m}}\, |v|_{W^{\tilde{m}}(L_p(S_\lambda))}\,. \tag{6.2.5}$$

The claim follows from combining (6.2.4), (6.2.5) and (LOC) with (6.2.3). □

We begin with the simplest case of a linear operator

$$\mathbf{F}(\mathbf{u}) = \mathbf{A}\mathbf{u} \tag{6.2.6}$$

where

$$\mathbf{A} = (a(\psi_\nu, \psi_\lambda))_{\lambda,\nu\in\Lambda}.$$

Assumption 6.2.1.

1. *We assume in what follows that* $F : H^t \to (H^t)' =: H^{-t}$ *is an isomorphism, where* H^t *is a closed subspace of* $H^t(\Omega)$*, determined e.g. by homogeneous boundary conditions. This covers all operators discussed in Section 5.4, in particular also the global boundary integral operators in Section 5.6.3.*

2. Ψ *is a Riesz basis for* $\mathbb{U} = H^t$ *satisfying (NE), (LOC), and (CP). More specifically,* Ψ *is of multilevel-type as discussed in Theorem 4.6.5, i.e., rescaled versions of* Ψ *are still Riesz bases for the spaces* $H^s(\Omega) \cap H^t$ *for some* s *in a neighborhood of* t*, see* (6.2.4).

Key-estimate: *Under the above assumptions on* Ψ *the following type of decay estimates hold: There exist* $\sigma > \frac{d}{2}$, $\beta > d$ *such that*

$$|\langle\psi_\lambda, F\psi_\mu\rangle| \lesssim \frac{2^{-\sigma\left||\lambda|-|\mu|\right|}}{\left(1 + 2^{\min\{|\mu|,|\lambda|\}}\mathrm{dist}\,(S_\lambda, S_\mu)\right)^\beta}, \qquad \lambda, \mu \in \Lambda. \tag{6.2.7}$$

where σ *and* β *depend on the smoothness and order of (CP), respectively. Such estimates have actually been established in many different contexts, see e.g. [4, 42, 37, 38, 41, 23]*

Remark 6.2.3. *There are two types of decay-effects. The numerator in* (6.2.7) *shows that entries of* $\mathbf{A}_{\lambda,\nu}$ *are the smaller the larger the level-distance of the basis functions. The algebraic decay in the denominator is relevant when the operator* F *is global, i.e.,* $a(\psi_\lambda, \psi_{\lambda'})$ *does not necessarily vanish when* $S_\lambda \cap S_{\lambda'} = \emptyset$*, for instance, when* F *is a boundary integral operator. This is illustrated in Figure 10.*

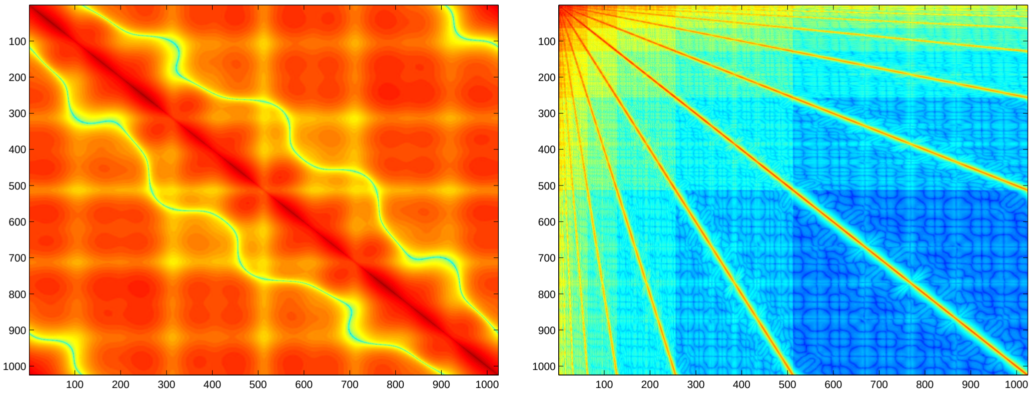


Figure 10: Left: scaling function (FE) representation of the single layer potential operator, right: wavelet representation (the darker the blue the smaller the entries

Discussion: We indicate next how estimates of the above type come about.

We consider first the regime of mutually overlapping basis functions:

$$|S_\lambda \cap S_\mu| > 0. \tag{6.2.8}$$

We wish to show that $|a_{\lambda,\mu}|$ gets small when the level difference $||\lambda| - |\mu||$ increases. The following simple argument applies to differential as well as integral operators and uses only their mapping properties. To be specific, suppose that F has the following additional continuity properties, namely that there exists a positive r such that

$$\|Fv\|_{H^{-t+s}(\Omega)} \lesssim \|v\|_{H^{t+s}(\Omega)}, \quad v \in H^{t+s}(\Omega) \cap \mathbb{U},\ 0 \le |s| < r, \tag{6.2.9}$$

which holds for the examples in the previous sections. Without loss of generality let $|\lambda| > |\mu|$.

Remark 6.2.4. *For instance, when $t = 1$ piecewise linear wavelets have such access regularity, namely $\psi_\lambda \in H^{\frac{3}{2}-\delta}(\Omega)$ for every positive δ, i.e., $r = 1/2$ in this case.*

Then, since $F\psi_\mu$ is a continuous linear functional on $H^{t-s}(\Omega)$ we derive from (6.2.9)

$$|\langle \psi_\lambda, F\psi_\mu \rangle| \le \|\psi_\lambda\|_{H^{t-s}(\Omega)} \|F\psi_\mu\|_{H^{-t+s}(\Omega)} \lesssim \|\psi_\mu\|_{H^{t+s}(\Omega)} \|\psi_\lambda\|_{H^{t-s}(\Omega)}. \tag{6.2.10}$$

Whenever

$$s < r, \quad t + s < s_2, \quad t - s > -\tilde{s}_2, \tag{6.2.11}$$

we can invoke the norm equivalences from Theorem 4.6.5 and (6.2.4) to conclude that

$$|\langle \psi_\lambda, F\psi_\mu \rangle| \lesssim 2^{-s\left||\lambda|-|\mu|\right|}, \quad \lambda, \mu \in \Lambda. \tag{6.2.12}$$

Hence, entries with overlapping wavelets still decay with increasing level difference and the strength of this decay depends essentially on the smoothness of the wavelets.

This is in general only a crude estimate and to illustrate the basic mechanism we consider next the case

$$Fu = -\mathrm{div}(a\nabla u),$$

where a is a possibly variable but smooth diffusion coefficient. Note that when ψ_λ has vanishing moments of order $\tilde{m}$ (see (VM)) its gradient $\nabla\psi_\lambda$ has even vanishing moments of order $\tilde{m}+1$ so that for $|\lambda| \geq |\mu|$, say,

$$\begin{array}{rcl} |\langle\psi_\lambda, F\psi_\mu\rangle| & = & |\langle a\nabla\psi_\mu, \nabla\psi_\lambda\rangle| \\ & \leq & \inf_{g\in\mathbb{P}_{\tilde{m}+1}} \|a\nabla\psi_\mu - g\|_{L^2(S_{\psi_\lambda})}\|\nabla\psi_\lambda\|_{L^2(\Omega)} \\ & \lesssim & 2^{-\zeta|\lambda|}|a\nabla\psi_\mu|_{H^\zeta(\Omega)}, \end{array}$$

provided that $a\nabla\psi_\mu \in H^\zeta(\Omega)$. When both a and $\nabla\psi_\lambda$ have uniformly bounded H^ζ-norms we obtain

$$|a\nabla\psi_\mu|_{H^\zeta(\Omega)} \lesssim 2^{\zeta|\mu|}2^{-\frac{d}{2}\left||\mu|-|\lambda|\right|}, \qquad (6.2.13)$$

to arrive at

$$|\langle a\nabla\psi_\mu, \nabla\psi_\lambda\rangle| \lesssim 2^{-(\zeta+\frac{d}{2})\left||\mu|-|\lambda|\right|}, \quad \lambda, \mu \in \Lambda, \qquad (6.2.14)$$

which is again of the form (6.2.12) for $s = \zeta + \frac{d}{2}$, see (6.2.7).

When F is a differential operator as above, and hence local, $|S_\lambda \cap S_\mu| = 0$ already implies $a_{\lambda,\mu} = 0$. This is no longer true, of course, when F is an integral operator. We illustrate the key mechanisms for operators of the form

$$(Fu)(x) = \int_\Gamma K(x, y)u(y)d\Gamma_y, \qquad (6.2.15)$$

where the kernel K is smooth except on its diagonal and satisfies

$$|\partial_x^\alpha\partial_y^\beta K(x, y)| \lesssim \mathrm{dist}\,(x, y)^{-(d+2t+|\alpha|+|\beta|)}. \qquad (6.2.16)$$

This covers, in particular, the the cases considered inSection 5.6.3. Accordingly we assume again that $F : \mathbb{U} \to \mathbb{U}'$ is an isomorphism and $\mathbb{U}$ is a closed subspace of $H^t(\Gamma)$. To see how the cancellation properties play in this case note first that

$$\langle\psi_\lambda, F\psi_\mu\rangle = \langle K, \psi_\lambda \otimes \psi_\mu\rangle, \qquad (6.2.17)$$

which suggests to apply estimates like (6.2.2) separately with respect to x and y. To be a bit more precise let $P_{\lambda,\tilde{m}}$ be an L_∞-bounded projector of $L_\infty(S_\lambda)$ to $\mathbb{P}_{\tilde{m}}|_{S_\lambda}$ and denote by I_x, I_y the identity operator with respect to the variables

x, y, respectively. Writing $(P_{\lambda,\tilde{m}} \otimes I_y K)(x,y) := (P_{\lambda,\tilde{m}} K(\cdot, y))(x)$, we consider the Boolean sum

$$(P_{\lambda,\tilde{m}} \oplus P_{\mu,\tilde{m}})K := (P_{\lambda,\tilde{m}} \otimes I_y)K + (I_x \otimes P_{\mu,\tilde{m}})K - (P_{\lambda,\tilde{m}} \otimes P_{\mu,\tilde{m}})K. \qquad (6.2.18)$$

One readily checks that, on the one hand,

$$K - (P_{\lambda,\tilde{m}} \oplus P_{\mu,\tilde{m}})K = (I_x - P_{\lambda,\tilde{m}}) \otimes (I_y - P_{\mu,\tilde{m}})K, \qquad (6.2.19)$$

i.e., errors multiply, while, on the other hand, one still has by (VM)

$$\langle (P_{\lambda,\tilde{m}} \oplus P_{\mu,\tilde{m}})K, \psi_\lambda \otimes \psi_\mu \rangle = 0. \qquad (6.2.20)$$

Since therefore $|\langle \psi_\lambda, F\psi_\mu \rangle| = |\langle K - (P_{\lambda,\tilde{m}} \oplus P_{\mu,\tilde{m}})K, \psi_\lambda \otimes \psi_\mu \rangle|$, we can invoke the estimates from (6.2.3), (6.2.5) for $q = 1, p = \infty$ for both variables separately to conclude on account of (6.2.16) that

$$|\langle \psi_\lambda, F\psi_\mu \rangle| \lesssim \frac{2^{-(|\lambda|+|\mu|)\left(t+\tilde{m}+\frac{d}{2}\right)}}{\operatorname{dist}(S_\lambda, S_\mu)^{d+2t+2\tilde{m}}} = \frac{2^{-\left||\lambda|-|\mu|\right|\left(t+\tilde{m}+\frac{d}{2}\right)}}{\left(2^{\min\{|\mu|,|\lambda|\}}\operatorname{dist}(S_\lambda, S_\mu)\right)^{d+2t+2\tilde{m}}}. \qquad (6.2.21)$$

The rightmost reformulation in (6.2.21) shows an algebraic decay in terms of the scaled distance between the two wavelet supports measured by multiples of the larger support diameter.

In summary, (6.2.12) and (6.2.21) indicate that the typical bound for the entries $\langle \psi_\lambda, F\psi_\mu \rangle$ of the wavelet representation has the claimed form (6.2.7)

$$|\langle \psi_\lambda, F\psi_\mu \rangle| \lesssim \frac{2^{-\sigma\left||\lambda|-|\mu|\right|}}{\left(1 + 2^{\min\{|\mu|,|\lambda|\}}\operatorname{dist}(S_\lambda, S_\mu)\right)^{\beta}}, \quad \text{for some } \sigma > \frac{d}{2}, \ \beta > d, \quad \lambda, \mu \in \Lambda.$$

where σ and β depend on the smoothness and order of vanishing moments of the wavelets, respectively.

There are still infinitely many pairs of wavelets satisfying (6.2.8). Obviously, it is desirable to have σ in (6.2.7) as large as possible. Unfortunately, the above argument requires a global smoothness of the wavelets to ensure a strong decay with respect to level differences. Using structural properties of F these estimates can be refined for certain wavelet families. For instance, when F is a partial differential operator with constant coefficients, such as the Laplacian, and when the wavelets are piecewise polynomials there are more and more indices $|\lambda| > |\mu|$ for which S_λ is such that $\psi_\mu|_{S_\lambda} \in \mathbb{P}_m$, i.e., ψ_λ "sees" only a polynomial part of $F\psi_\mu$ so that, by (VM), $\langle \psi_\lambda, F\psi_\mu \rangle = 0$. This observation can be extended to more generality.

Remark 6.2.5. *In the above form the estimates* (6.2.7) *result from what is sometimes called "first compression". A so called "second compression" was introduced and analyzed by Reinhold Schneider resulting in significantly improved estimates for* $|\langle \psi_\lambda, F\psi_\mu \rangle|$

for a wide range of pseudo-differential operators and spline wavelet bases for overlapping supports S_λ, S_μ, see e.g. [38, 28]. Best possible values for σ and β for spline wavelets and many relevant differential and singular integral operators have been determined in [34, 35].

6.3 Compressible matrices

Let us denote by

$$\mathcal{M}_{\sigma,\beta} := \left\{ \mathbf{B} \in \mathbb{R}^{\Lambda\times\Lambda} : |b_{\lambda,\mu}| \lesssim 2^{-\sigma\left||\lambda|-|\mu|\right|}\left(1 + 2^{\min\{|\mu|,|\lambda|\}}\mathrm{dist}\,(S_\lambda, S_\mu)\right)^{-\beta} \right\} \qquad (6.3.1)$$

the set of matrices satisfying (6.2.7), assuming always that as above

$$\sigma > \frac{d}{2}, \quad \beta > d. \qquad (6.3.2)$$

Some important properties of this class are conveniently derived with the aid of the following simple version of the Schur-Lemma.

Lemma 6.3.1. *Let $\mathbf{B} = (b_{\lambda,\lambda'})_{\lambda,\lambda'\in\Lambda} \in \mathbb{R}^{\Lambda\times\Lambda}$. If there exists a sequence $(\omega_\lambda)_{\lambda\in\Lambda}$ with positive entries and a finite C such that*

$$\sum_{\lambda'\in\Lambda} |b_{\lambda,\lambda'}|\omega_{\lambda'} \leq C\omega_\lambda, \quad \sum_{\lambda\in\Lambda} |b_{\lambda,\lambda'}|\omega_\lambda \leq C\omega'_\lambda, \quad \lambda,\lambda' \in \Lambda, \qquad (6.3.3)$$

then one has

$$\|\mathbf{B}\| \leq C \qquad (6.3.4)$$

where C is the constant from (6.3.3) ($\|\cdot\| := \|\cdot\|_{\mathcal{L}(\ell_2,\ell_2)}$).

Proof: Let $\mathbf{D}$ denote the diagonal matrix with entries $d_{\lambda,\lambda'} = \omega_\lambda\delta_{\lambda,\lambda'}$, $\lambda,\lambda' \in \Lambda$. Then (6.3.3) just says that $\|\mathbf{D}\mathbf{B}\mathbf{D}^{-1}\|_{\mathcal{L}(\ell_\infty,\ell_\infty)} \leq C$ and $\|\mathbf{D}\mathbf{B}\mathbf{D}^{-1}\|_{\mathcal{L}(\ell_1,\ell_1)} \leq C$, whence $\|\mathbf{B}\| = \|\mathbf{D}\mathbf{B}\mathbf{D}^{-1}\| \leq C$, by interpolation. □

For convenience we abbreviate

$$d(\lambda,\lambda') = 2^{\min\{|\lambda|,|\lambda'|\}}\mathrm{dist}\,(S_\lambda, S_{\lambda'}). \qquad (6.3.5)$$

Thus, $d(\lambda,\lambda')$ expresses the distance with the diameter of the lower level basis function as the unit.

Observe first that, under the assumptions (6.3.2) all elements of $\mathcal{M}_{\sigma,\beta}$ are bounded in $\ell^2(\Lambda)$.

Proposition 6.3.1. *Every $\mathbf{B} \in \mathcal{M}_{\sigma,\beta}$ defines a bounded operator on $\ell_2(\Lambda)$.*

Proof. We apply Schur's lemma with the weights $\omega_\lambda = 2^{-|\lambda|d/2}$, $\lambda \in \Lambda$. To establish the first inequality in (6.3.3), let $\lambda \in \Lambda$ and let $|\lambda| = j$. Then, using boundedness of the domain and the fact that $\beta > d$ one arrives at the estimate

$$\sum_{|\lambda'|=j'} (1 + d(\lambda, \lambda'))^{-\beta} \lesssim 2^{d\max\{0,j'-j\}}.$$

With this estimate we obtain for the summation in space,

$$\begin{aligned} \omega_\lambda^{-1} \sum_{\lambda' \in \Lambda} \omega_{\lambda'} |b_{\lambda,\lambda'}| &\lesssim 2^{d|\lambda|/2} \sum_{j' \geq 0} 2^{-dj'/2} 2^{-\sigma|j-j'|} \sum_{|\lambda'|=j'} (1 + d(\lambda, \lambda'))^{-\beta} \\ &\lesssim \sum_{j' \geq j} 2^{-d(j'-j)/2} 2^{-\sigma(j'-j)} 2^{d(j'-j)} + \sum_{0 \leq j' < j} 2^{-d(j'-j)/2} 2^{\sigma(j'-j)} \\ &\lesssim \sum_{l \geq 0} 2^{-(\sigma - d/2)l} < \infty. \end{aligned}$$

A symmetric argument confirms the second estimate in (6.3.3) proving that $\mathbf{B}$ is bounded. □

The perhaps most important consequence of the decay properties (6.3.1) is a systematic way of *sparsitying* the matrices in $\mathcal{M}_{\sigma,\beta}$ while controling the error in the spectral norm. The first step is a truncation in scale, defining for any given $J \in \mathbb{N}$ the scale-compressed matrix $\tilde{\mathbf{B}}_J := (\tilde{b}_{\lambda,\lambda'})_{\lambda,\lambda' \in \nabla}$ by

$$\tilde{b}_{\lambda,\lambda'} := \begin{cases} b_{\lambda,\lambda'}, & ||\lambda| - |\lambda'|| \leq J/d, \\ 0, & \text{else.} \end{cases} \qquad (6.3.6)$$

The second step is a truncation in space provided by the new matrix $\mathbf{B}_J := (b'_{\lambda,\lambda'})_{\lambda,\lambda' \in \nabla}$ where

$$b'_{\lambda,\lambda'} := \begin{cases} \tilde{b}_{\lambda,\lambda'}, & d(\lambda, \lambda') \leq 2^{J/d - ||\lambda| - |\lambda'||} \gamma(||\lambda| - |\lambda'||), \\ 0, & \text{else,} \end{cases} \qquad (6.3.7)$$

Here $\gamma(n)$ is also a polynomially decreasing sequence such that $\sum_n \gamma(n)^d < \infty$. Specifically, we take $\gamma(n) := (1 + n)^{-2/d}$.

Proposition 6.3.2. *Assume that* $\mathbf{B} \in \mathcal{M}_{\sigma,\beta}$, *so that* (6.3.2) *is valid and let*

$$s^* := \min\left\{\frac{\sigma}{d} - \frac{1}{2}, \frac{\beta}{d} - 1\right\}. \qquad (6.3.8)$$

Then, given any $s < s^*$, *the matrices* $\mathbf{B}_J$, *defined by* (6.3.6) *and* (6.3.7) *satisfy*

$$N_J := \#\{\textit{nonzero entries in the rows/columns of } \mathbf{B}_J\} \leq C 2^J, \qquad (6.3.9)$$

and

$$\|\mathbf{B} - \mathbf{B}_J\| \leq C2^{-Js}, \quad J \in \mathbb{N}, \tag{6.3.10}$$

where the constant C *depends on* s *but is independent of* J*. Moreover this result also holds for* $s = s^*$ *provided* $\sigma - d/2 \neq \beta - d$.

Proof. We employ Lemma (6.3.1) to estimate first $\|\mathbf{B} - \tilde{\mathbf{B}}_J\|$. To that end, we fix $J > 0$ choose the same weights $\omega_\lambda = 2^{-|\lambda|d/2}$ as in the proof of Proposition 6.3.1. Then, as before, we obtain for any $\lambda \in \Lambda$ and $|\lambda| = j$

$$\begin{aligned}
\omega_\lambda^{-1} \sum_{\lambda'} \omega_{\lambda'} |b_{\lambda,\lambda'} - \tilde{b}_{\lambda,\lambda'}| &= \omega_\lambda^{-1} \sum_{\{\lambda' \,:\, |j - |\lambda'|| > J/d\}} \omega_{\lambda'} |b_{\lambda,\lambda'}| \\
&\lesssim \sum_{l > J/d} 2^{-(\sigma - d/2)l} \\
&\lesssim 2^{-(\sigma - d/2)J/d} \lesssim 2^{-Js},
\end{aligned}$$

which shows that

$$\|\mathbf{B} - \tilde{\mathbf{B}}_J\| \lesssim 2^{-Js}, \tag{6.3.11}$$

holds uniformly in J.

It remains to bound $\|\tilde{\mathbf{B}}_J - \mathbf{B}_j\|$ to take the spatial truncation into account. Note first, that we can immediately estimate the maximal number N_J of non-zero entries in each row and column of $\mathbf{B}_J$ by

$$N_J \lesssim \sum_{l=0}^{\lceil J/d \rceil} [2^{J/d - l} \gamma(l)]^d 2^{ld} \lesssim 2^J, \tag{6.3.12}$$

which confirms (6.3.9). In view of (6.3.11), it remains only to prove that $\|\mathbf{B}_J - \tilde{\mathbf{B}}_J\| \lesssim 2^{-Js}$. In order to estimate the spectral norm $\|\mathbf{B}_J - \tilde{\mathbf{B}}_J\|$, we use again the Schur lemma with the same weights. For each j' and $\lambda \in \Lambda$, we have

$$\sum_{\{\lambda' \,:\, d(\lambda,\lambda') > R\}} (1 + d(\lambda, \lambda'))^{-\beta} \lesssim R^{-\beta + d} 2^{d \max\{0, |\lambda'| - |\lambda|\}},$$

Therefore, for any $\lambda \in \Lambda$,

$$\begin{aligned}
\omega_\lambda^{-1} \sum_{\lambda'} \omega_{\lambda'} |b'_{\lambda,\lambda'} - \tilde{b}_{\lambda,\lambda'}| &\lesssim \sum_{l=0}^{[J/d]} 2^{-(\sigma - d/2)l} [2^{J/d - l} \gamma(l)]^{-(\beta - d)} \\
&= 2^{-sJ} [2^{-J(\beta - d - ds)/d} \sum_{l=0}^{J} 2^{[(\beta - d) - (\sigma - d/2)]l} \gamma(l)^{-(\beta - d)}].
\end{aligned}$$

In the case where $(\beta - d) < (\sigma - d/2)$ (resp.$(\beta - d) > (\sigma - d/2)$), the factor on the right of 2^{-sJ} is bounded by $C2^{-J(\beta - d - ds)/d}$ (resp. $C2^{-J(\sigma - d/2 - ds)/d}$) with C a constant independent of J and λ. Thus, when $\beta - d \neq \sigma - d/2$, we obtain the desired estimate of $\|\mathbf{B}_J - \tilde{\mathbf{B}}_J\|$ for all $s \leq s^*$. On the other hand, when $\beta - d = \sigma - d/2$, this factor is still bounded by a fixed constant provided $s < s^*$. □

As shown in the next section the estimates (6.3.9) and (6.3.10) form the basis for an efficient approximate application of **B**. To that end, a slightly different balance between accuracy and sparsity, based on the following simple observation, turns out to be more convenient.

Lemma 6.3.2. *Assume that* $\mathbf{B} \in \mathcal{M}_{\sigma,\beta}$ *and* s^* *is defined by* (6.3.8). *Then for every* $s < s^*$ *there exists summable sequences of positive numbers* $(\alpha_n)_{n\geq 0}$, $(\beta_n)_{n\geq 0}$ *and matrices* $\mathbf{B}_j$, $j \geq 0$ *such that*

$$\|\mathbf{B} - \mathbf{B}_j\| \leq \beta_j 2^{-sj}, \quad j \geq 0, \tag{6.3.13}$$

and

$$\#\{\textit{entries per row/column of } \mathbf{B}_j\} \leq \alpha_j 2^j, \quad j \geq 0. \tag{6.3.14}$$

In fact, for any $a, b < 1$, according to Proposition 6.3.2, $\lesssim 2^{aJ} = 2^{(a-1)J}2^J$ entries per row and column suffice, to provide accuracy $\lesssim 2^{-saJ} = 2^{sa(b-1)J}2^{-sabJ}$. Clearly, choosing a, b sufficiently close to one and s sufficiently close to s^* we may stlll bring $\tilde{s} = abs$ as close to s^* as we wish while the entries $\alpha_j := 2^{(a-1)j}$, $\beta_j = 2^{sa(b-1)j}$ still exhibit exponential decay.

Definition 6.3.1. *The matrix* **B** *is called* s^**-compressible if there exists a sequence* $(\mathbf{B}_j)_{j\geq 0}$ *of matrices as well as summable sequences* $(\alpha_j)_{j\geq 0}$, $(\beta_j)_{j\geq 0}$ *satisfying* (6.3.13) *and* (6.3.14). *Moreover, we define*

$$\|\mathbf{B}\|_* := \min\max\left\{\|(\alpha_j)_{j\geq 0}\|_{\ell_1(\mathbb{N}_0)}, \|(\beta_j)_{j\geq 0}\|_{\ell_1(\mathbb{N}_0)}\right\}. \tag{6.3.15}$$

Denoting by $\mathcal{C}_{s^*}$ the set of s^*-compressible matrices, we have shown that $\mathcal{M}_{\sigma,\beta} \subset \mathcal{C}_{s^*}$ when s^* is defined by (6.3.8).

6.4 Efficient application of compressible matrices

The envisaged approximate application of an s^*-compressible matrix **A** to a sequence $\mathbf{v} \in \ell_2(\Lambda)$ is *adaptive,* hence nonlinear, in that it combines *apriori* knowledge about the matrix with *a posteriori* information about the input sequence. This latter information is used through the following decomposition of **v**. Let $\mathbf{v}_{2^j}$ denote a best 2^j-term approximation to **v** and the sections

$$\mathbf{v}_{[j]} := \mathbf{v}_{2^j} - \mathbf{v}_{2^{j-1}}, \quad \mathbf{j} \geq \mathbf{0}, \quad \mathbf{v}_{2^{-1}} := \mathbf{0}. \tag{6.4.1}$$

Obviously, $\mathbf{v}_{[j]}$ is comprized of the 2^{j-1}st to 2^jth largest (in modulus) entries of **v** and

$$\|\mathbf{v}_{[j]}\|_{\ell_2} \leq \sigma_{2^j}(\mathbf{v}) + \sigma_{2^{j-1}}(\mathbf{v}) \leq (\mathbf{1} + \mathbf{2^s})\mathbf{2}^{-\mathbf{js}}|\mathbf{v}|_{\mathcal{A}^s}, \quad \mathbf{j} \geq \mathbf{0}, \tag{6.4.2}$$

where here and below

$$\mathcal{A}^s = \mathcal{A}^s_\infty((\Sigma_n), \ell_2(\Lambda)), \quad \text{with } \Sigma_n := \{\mathbf{v} \in \ell_2(\Lambda) : \|\mathbf{v}\|_{\ell_0(\Lambda)} \le \mathbf{n}\}, \tag{6.4.3}$$

where

$$\|\mathbf{v}\|_{\ell_0(\Lambda)} = \#(\mathrm{supp}\,(\mathbf{v})), \tag{6.4.4}$$

is, of course, not really a norm.

The idea is now to intertwine the telescopic expansions

$$\mathbf{A} = \sum_{j\ge 0} \mathbf{A}_j - \mathbf{A}_{j-1}, \ (\mathbf{A}_{-1} := 0), \qquad \mathbf{v} = \sum_{j\ge \mathbf{0}} \mathbf{v}_{[j]}, \tag{6.4.5}$$

to generate for each $J \in \mathbb{N}$ and approximation

$$\mathbf{w}_J := \sum_{j=0}^{J} \mathbf{A}_{J-j}\mathbf{v}_{[j]}. \tag{6.4.6}$$

Since

$$\mathbf{A}\mathbf{v} - \mathbf{w}_\mathbf{J} = \sum_{\mathbf{j}=\mathbf{0}}^{\mathbf{J}} (\mathbf{A} - \mathbf{A}_{\mathbf{J}-\mathbf{j}})\mathbf{v}_{[\mathbf{j}]} + \mathbf{A}(\mathbf{v} - \mathbf{v}_{\mathbf{2}^\mathbf{J}}) \tag{6.4.7}$$

one immediately arrives at the error estimate

$$\|\mathbf{A}\mathbf{v} - \mathbf{w}_\mathbf{J}\|_{\ell_2} \le \sum_{\mathbf{j}=\mathbf{0}}^{\mathbf{J}} \beta_{\mathbf{J}-\mathbf{j}} \mathbf{2}^{-\mathbf{s}(\mathbf{J}-\mathbf{j})} \|\mathbf{v}_{[\mathbf{j}]}\|_{\ell_2} + \|\mathbf{A}\|\sigma_{\mathbf{2}^\mathbf{J}}(\mathbf{v}). \tag{6.4.8}$$

Obviously, the right hand side decreases when J increases. Specifically, when $\mathbf{v}$ has a finite support one can compute the quantities $\|\mathbf{v}_{[j]}\|_{\ell_2}$ and find for any given target accuracy $\eta > 0$ the smallest $J = J(\eta)$ such that

$$\|\mathbf{A}\|_* \sum_{j=0}^{J} 2^{-s(J-j)} \|\mathbf{v}_{[j]}\|_{\ell_2} + \|\mathbf{A}\|\sigma_{\mathbf{2}^J}(\mathbf{v}) \le \eta. \tag{6.4.9}$$

We have thus described a computable routine

APPLY$[\mathbf{A}, \mathbf{v}, \eta] \to \mathbf{w}_\eta$

with the following property: for every finitely supported input $\mathbf{v}$ and any given target accuracy $\eta > 0$ the output $\mathbf{w}_\eta$ is given by $\mathbf{w}_\eta := \mathbf{w}_{J(\eta)}$, defined in (6.4.6), and satisfies

$$\|\mathbf{A}\mathbf{v} - \mathbf{w}_\eta\|_{\ell_2} \le \eta. \tag{6.4.10}$$

Remark 6.4.1. *The computation of the quantities* $\|\mathbf{v}_{[j]}\|$ *requires sorting the entries of* $\mathbf{v}$ *which scales like* $\#(\operatorname{supp}\mathbf{v})\log\big(\#(\operatorname{supp}\mathbf{v})\big)$. *replacing exact sorting by quasi-sorting, i.e., collecting all entries of* $\mathbf{v}$ *such that* $\alpha^j \leq |v_\lambda| < \alpha^{j-1}$ *for some fixed* $\alpha \in]0,1[$, *in some arbitrary order, amounts to a cost that stays proportional to* $\#(\operatorname{supp}\mathbf{v})$ *at the expense of an additional constant factor in the error bound* (6.4.8). *In all complexity estimates to come we tacitly assume the application of such a strategy to attribute linear cost to sorting tasks. Moreover, we assume that each entry in* $\mathbf{A}$ *can be computet by a uniformly bounded number of operations. This is, for instance, the case when the coefficients in the underlying PDE are constant or polynomial.*

Under this latter provision the properties of the rountine APPLY$[\mathbf{A},\mathbf{v},\eta]$ can be stated as follows.

Proposition 6.4.1. *Assume that* $\mathbf{A} \in \mathcal{C}_{s^*}$. *Then the following statements hold:*

1. *If* $\mathbf{v} \in \mathcal{A}^{\mathbf{s}}$ *and* $s < s^*$ *we have*
$$\#\operatorname{supp}\mathbf{w}_\eta \lesssim \eta^{-1/s}\|\mathbf{v}\|_{\mathcal{A}^s}^{\mathbf{1/s}}, \quad \eta > \mathbf{0}. \tag{6.4.11}$$

2. *When* $\mathbf{v}$ *has finite support the number of operations required to compute* $\mathbf{w}_\eta$ *is bounded by*
$$\mathbf{ops}(\mathbf{w}_\eta) \lesssim \#\operatorname{supp}\mathbf{v} + \eta^{-\mathbf{1/s}}\|\mathbf{v}\|_{\mathcal{A}^s}^{\mathbf{1/s}}. \tag{6.4.12}$$

3. *The output* $\mathbf{w}_\eta = $ APPLY$[\mathbf{A},\mathbf{v},\eta]$ *is stable in* $\mathcal{A}^s$, *i.e.,*
$$|\mathbf{w}_\eta|_{\mathcal{A}^s} \lesssim |\mathbf{v}|_{\mathcal{A}^s}, \quad \eta > \mathbf{0}, \tag{6.4.13}$$
with constants independent of $\eta > 0$. *and* $\mathbf{v}$.

4. *For every* $s < s^*$ *the matrix* $\mathbf{A}$ *takes* $\mathcal{A}^s$ *boundedly into itself.*

Proof. In view of (6.3.14) we infer from (6.4.6) that for any $J \in \mathbb{N}$
$$\#\operatorname{supp}\mathbf{w}_J \leq \sum_{j=0}^{J} \alpha_j 2^{J-j}2^j \leq \|\mathbf{A}\|_* 2^J. \tag{6.4.14}$$

On the other hand, invoking (6.4.2), one has
$$\begin{aligned}\|\mathbf{A}\mathbf{v} - \mathbf{w}_\mathbf{J}\|_{\ell_2} &\leq (1+2^s)\sum_{k\leq J}\beta_{J-k}2^{-s(j-k)}2^{-ks}|\mathbf{v}|_{\mathcal{A}^s} + \|\mathbf{A}\|\mathbf{2}^{-\mathbf{J}\mathbf{s}}|\mathbf{v}|_{\mathcal{A}^s} \\ &\leq ((1+2^s)\|\mathbf{A}\|_* + \|\mathbf{A}\|)2^{-Js}|\mathbf{v}|_{\mathcal{A}^s}. \end{aligned}\tag{6.4.15}$$

Thus, denoting by J' the smallest integer for which $((1+2^s)\|\mathbf{A}\|_* + \|\mathbf{A}\|)2^{-J's}|\mathbf{v}|_{\mathcal{A}^s} \leq \eta$ we conclude that $((1+2^s)\|\mathbf{A}\|_* + \|\mathbf{A}\|)2^{-J's}|\mathbf{v}|_{\mathcal{A}^s} > \mathbf{2}^{-\mathbf{s}}\eta$. Since $J(\eta) \leq J'$ we also have
$$((1+2^s)\|\mathbf{A}\|_* + \|\mathbf{A}\|)2^{-J(\eta)s}|\mathbf{v}|_{\mathcal{A}^s} > \mathbf{2}^{-\mathbf{s}}\eta.$$

From this and invoking (6.4.14), we deduce that

$$(\#\mathrm{supp}\,\mathbf{w}_\eta)^s \le 2^{J(\eta)s} < 2^s((1+2^s)\|\mathbf{A}\|_* + \|\mathbf{A}\|)|\mathbf{v}|_{\mathcal{A}^s}\eta^{-1}. \tag{6.4.16}$$

which gives (6.4.11).

Under the provision spelled out in Remark 6.4.1 the bound (6.4.12) is then an immediate consequence.

To confirm (6.4.13) it suffices to exhibit for each $j \in \mathbb{N}$ a $\tilde{\mathbf{w}}_j$ with $\#\mathrm{supp}\,\tilde{\mathbf{w}}_j \le C2^j$ such that $\|\mathbf{w}_\eta - \tilde{\mathbf{w}}_j\|_{\ell_2} \le C2^{-sj}|\mathbf{v}|_{\mathcal{A}^s}$. In view of (6.4.14), it suffices to consider $j \le J(\eta)$. Of course, natural candidates are $\tilde{\mathbf{w}}_j := \mathbf{w}_j$, $j \le J(\eta)$, defined by (6.4.6). The claim (6.4.13) follows now immediately from (6.4.14) and (6.4.15).

Taking J arbitrary, the same ergument shows that $|\mathbf{A}\mathbf{v}|_{\mathcal{A}^s} \lesssim |\mathbf{v}|_{\mathcal{A}^s}$ which completes the proof. □

Proposition 6.4.1 and Corollary 5.11.1 imply the following fact.

Corollary 6.4.1. *For operator representations* $\mathbf{A}$ *of the class considered above, and* $\mathbf{C} = \alpha\mathbf{I}$*, the routine*

RES $[\mathbf{f}, \mathbf{A}, \mathbf{v}, \eta] \to \mathbf{r}_\eta$, *defined by*

$$\alpha\Big(\text{COARSE}\,[\hat{\mathbf{f}}, \alpha^{-1}c\eta] - \text{APPLY}\,[\mathbf{A}, \mathbf{v}, \eta/\mathbf{2}]\Big), \tag{6.4.17}$$

with preprocessed right hand side $\hat{\mathbf{f}}$ *according to* (6.1.1) *(and* $\hat{\epsilon}$ *small enough), and* c *chosen so that*

$$c \le \frac{1}{2} - \frac{\hat{\epsilon}}{\eta},$$

is s^**-computable whenever* $\mathbf{A}$ *is* s^**-compressible. Hence the complexity bounds from Theorem 5.11.2 hold for Algorithm 5.10.1 applied to* $\mathbf{A}\mathbf{u} = \mathbf{f}$.

6.5 A Numerical Example – the Stokes Problem

Consider the Stokes system in Example 5.6.1. The following results are obtained with the Uzawa technique, see Algorithm 5.9.1. This works as follows:

- In this case the residual evaluation RES refers to the Schur complement problem (5.9.10a).
- The application of $\mathbf{F} = \mathbf{S}$ involves now an approximate elliptic solve in step 2. of Algorithm 5.9.1, and the update in step 3. requiring an approximate application of $\mathbf{B}$.
- The fact that the resulting residual realization is s^*-computable follows from Corollary 6.4.1 above, applied to the elliptic subproblem (also using Proposition 6.4.1), combined with the properties of the routine APPLY from Proposition 6.4.1 for the matrix $\mathbf{B}$.

Recall from Algorithgm 5.9.1 the ideal iteration

$$\mathbf{p}^{k+1} = \mathbf{p}^{k} + \alpha(\mathbf{B}(\mathbf{u}^{k} - \mathbf{g}), \quad \text{where} \quad \mathbf{A}\mathbf{u}^{k} = \mathbf{f} - \mathbf{B}^{T}\mathbf{p}^{k}. \tag{6.5.1}$$

The residual approximation can then be realized as follows:

RES$[\mathbf{A}, \mathbf{B}, \mathbf{f}, \mathbf{g}, \mathbf{p}^{k}, \mathbf{u}^{k}, \eta] \to \mathbf{r}_{\eta}$

1. choose $\eta_1, \ldots, \eta_5$ such that
$$\eta_4 + \|\mathbf{B}\|(\eta_3 + \|\mathbf{A}^{-1}\|(\eta_1 + \eta_2)) + \eta_5 \le \alpha^{-1}\eta. \tag{6.5.2}$$
2. compute
$$\mathbf{w} := \text{COARSE}[\mathbf{f}, \eta_1] - \text{APPLY}[\mathbf{B}^{T}, \mathbf{p}^{k}, \eta_2]$$
3. Define SOLVE$[\mathbf{A}, \mathbf{w}, \eta]$ as in Algorithm 5.10.1 and compute
$$\bar{u} := \text{SOLVE}[\mathbf{A}, \mathbf{w}, \eta_3]$$
4. compute
$$\mathbf{r}_{\eta} := \alpha\big\{ \text{APPLY}[\mathbf{B}, \bar{\mathbf{u}}, \eta_4] - \text{COARSE}[\mathbf{g}, \eta_5]\big\}$$

Exercise 6.5.1. *Show that*

$$\| \text{RES}[\mathbf{A}, \mathbf{B}, \mathbf{f}, \mathbf{g}, \mathbf{p}^{k}, \mathbf{u}^{k}, \eta] - \alpha(\mathbf{B}(\mathbf{u}^{k} - \mathbf{g})\|_{\ell_2} \le \eta.$$

Exercise 6.5.2. *Using the properties of* COARSE, APPLY *and the results on Algorithm 5.10.1, and noting that all the accuracy tolerances* η_i *are proportional to* η, *show that* RES *is* s^**-computable for the compressibility limits of* $\mathbf{A}$ *and* $\mathbf{B}$.

This gives the following result, see also [20].

Corollary 6.5.1. *The complexity bounds from Theorem 5.11.2 hold for the above Uzawa-variant of Algorithm 5.10.1 applied to the Stokes problem, see also [20].*

The theoretical results are illustrated by some numerical experiments, see Tabel 1. Velocity and pressure components are depicted in Figure 11. Both show a singularity at the reentrant corner.

$$\rho_{\mathbf{x}} := \frac{\|\mathbf{x} - \mathbf{x}_{\Lambda}\|_{\ell_2}}{\|\mathbf{x} - \mathbf{x}_{\#\Lambda}\|_{\ell_2}}, \quad r_{\mathbf{x}} := \frac{\|\mathbf{x} - \mathbf{x}_{\Lambda}\|_{\ell_2}}{\|\mathbf{x}\|_{\ell_2}},$$

Table 2 illustrates an important point. It is well-known from Brezzi's theory that (as a saddle-point problem is indefinite) its finite dimensional Galerkin formulation is not automatically stable. The finite dimensional trial spaces spaces for velocity and pressure have to satisfy the LBB-condition (Ladyzhenskaya Babuska Brezzi condition), see [9]. In the following test the wavelet bases for velocity and pressure are chosen in such a way that the linear spans of fixed finite truncations would generate trial spaces that **violate** the LBB-condition. As shown this does *not* affect the convergence of the adaptive scheme. In this sense **adaptivity stabilizes**.

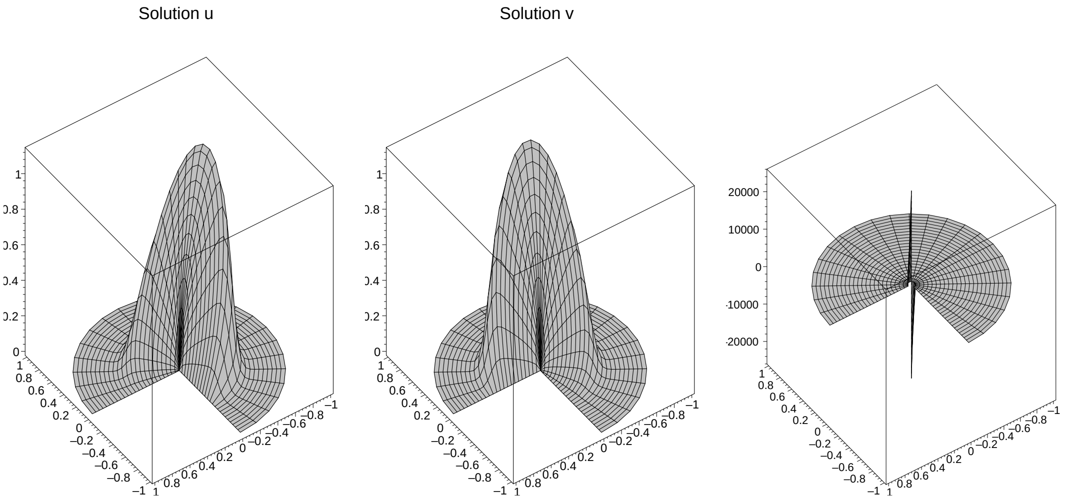


Figure 11: Exact solution for the first example. Velocity components (left and middle) and pressure (right). The pressure functions exhibits a strong singularity

It	$\#\Lambda_u$	$\rho_{\mathbf{u}}$	$r_{\mathbf{u}}$	$\#\Lambda_v$	$\rho_{\mathbf{v}}$	$r_{\mathbf{v}}$	$\#\Lambda_p$	$\rho_{\mathbf{p}}$	$r_{\mathbf{p}}$
1	33	1.04	0.6838	34	1.04	0.6744	768	130.35	1.0024
2	84	1.26	0.3427	83	1.24	0.3447	768	130.40	1.0028
3	193	1.32	0.1530	184	1.31	0.1541	768	15.37	0.5234
4	446	1.29	0.0821	450	1.29	0.0897	929	4.15	0.2218
5	1070	1.27	0.0434	1065	1.27	0.0456	1211	2.58	0.1034

Table 1: Results for the first example. Sizes of the adaptive approximations, ratio to best N-term approximation and relative error.

Exercise 6.5.3. *Explain why this does not entail any contradiction to existing theory.*

6.6 Adaptive Application of Nonlinear Operators

When the operator is nonlinear the scheme RES requires a substitute for the adaptive matrix application. Roughly speaking, given a finitely supported $\mathbf{v} \in \ell_2(\Lambda)$, produce a sparse accuracy controled approximation to $\mathbf{G}(\mathbf{v})$. It turns out that in the nonlinear case parsity of the input with nonzero entries in *arbitrary* positions cannot be preserved.

Subsets of Λ with Tree-Structure: Roughly speaking when an entry of

$$\mathbf{G}(\mathbf{u}) = (\langle G(u), \psi_\lambda \rangle)_{\lambda \in \Lambda}$$

It	$\#\Lambda_u$	ρ_u	r_u	$\#\Lambda_v$	ρ_v	r_v	$\#\Lambda_p$	ρ_p	r_p
3	5	1.00	0.7586	5	1.00	0.7588	243	2.23810	0.1196
4	20	1.13	0.4064	24	1.45	0.3979	262	2.08107	0.0612
5	61	1.47	0.2107	77	1.79	0.2107	324	2.72102	0.0339
6	178	1.33	0.1060	198	1.52	0.1306	396	2.81079	0.0209
7	294	1.19	0.0533	286	1.46	0.0744	674	2.21371	0.0108
8	478	1.25	0.0271	531	1.46	0.0362	899	1.83271	0.0071

Table 2: Results with piecewise linear trial functions for velocity and pressure - LBB condition is violated

say $\langle G(u), \psi_\lambda \rangle$, is large, available estimates do not allow one to recognize entries $\langle G(u), \psi_\mu \rangle$ with $|\mu| \le \lambda$ and $|S_\lambda \cap S_\mu| > 0$, as small.

Sparsity statements about such entries can only be established when the index sets under consideration are somewhat constraints. An appropriate constraint turns out to be *tree-structure*.

Definition 6.6.1. *A subset $\Gamma \subset \Lambda$ has tree-structure if*

$$\lambda \in \Gamma \quad \Rightarrow \quad \{\mu \in \Lambda : |\mu| \le |\lambda|,\ S_\mu \cap S_\lambda| > 0\} \subseteq \Gamma. \tag{6.6.1}$$

Figure 12 illustrates such an index set.

Next, the standard coarsening operator $\mathcal{C}_\eta$ needs to be repolaced by a *tree-coarsening* operator

$$\mathcal{C}_{\eta,\mathcal{T}} : \mathbf{v} \longrightarrow \mathbf{v}_\eta$$

which should have the following properties:

- $\mathbf{v}_\eta$ has tree structure
- $\exists C^*$ such that
$$\#(\operatorname{supp} \mathbf{v}_\eta) \le C^* \# \mathcal{T}\Big(\mathbf{v}, \frac{\eta}{C^*}\Big),$$
where $\mathcal{T}(\mathbf{v}, \epsilon)$ is the best tree-approximation with accuracy ϵ:
$$\big\|\mathbf{v} - \mathbf{v}|_{\mathcal{T}(\mathbf{v},\epsilon)}\big\|_{l_2} \le \epsilon,$$
- One has the error estimate
$$\|\mathbf{v} - \mathbf{v}_\eta\|_{l_2} \le \eta.$$

This is a version of the Coarsening Lemma (Corollary 5.11.1) based on the Tree Algorithm 3.2.1, see [17]. This requires providing some *local error functionals* (recall $S_\lambda := \operatorname{supp} \psi_\lambda$)

$$e_\lambda = e_\lambda(\mathbf{v}) := \sum_{\lambda' \in \Lambda, |\mathbf{S}_{\lambda'} \cap \mathbf{S}_\lambda| > \mathbf{0}, |\lambda'| \ge |\lambda|} \mathbf{v}_{\lambda'}^2, \tag{6.6.2}$$

representing the "energy" of $\mathbf{v}$ "over" S_λ. One can show that for $\mathbb{U} \subseteq H^1(\Omega)$

$$e_\lambda(\mathbf{v}) \lesssim \inf_{\mathbf{P}\in\mathbb{P}_{\tilde{\mathbf{m}}}} \|\mathbf{v}-\mathbf{P}\|^{\mathbf{2}}_{\mathbf{H^1}(\tilde{\mathbf{S}}_\lambda)}, \quad \tilde{\mathbf{S}}_\lambda = \bigcup\{\mathbf{S}_{\lambda'} : |\lambda'| = |\lambda|, \, |\mathbf{S}_\lambda \cap \mathbf{S}_{\lambda'}| > \mathbf{0}\},$$

where $\tilde{m}$ is the order of cancellation properties (CP) (respectively, vanishing moments (VM)). One can than verify all the above properties.

Remark 6.6.1. *Tree-structured index sets can be viewed as analogs to locally refined partitions.*

Adaptive Evaluation of Nonlinear Operators: As a substitute for APPLY$[\mathbf{A}, \cdot, \cdot]$ we need an adaptive *nonlinear* approximate evaluation scheme:

EVAL$[\mathbf{F}, \mathbf{v}, \eta] \rightarrow \mathbf{w}_\eta$ such that $\mathbf{v}_\eta$ has tree-structure and satisfies

$$\|\mathbf{F}(\mathbf{v}) - \mathbf{w}_\eta\|_{\ell_2} \le \eta. \tag{6.6.3}$$

Main Goal: construct such a scheme that is in addition s^*-computable and gives rise a residual approximation RES satisfying the properties in Definition 5.11.1.

A central ingredient is: given $\mathbf{v}$, predict and compute the significant entries of $\mathbf{F}(\mathbf{v})$. A typical result for nonlinear operators $\mathbf{F}$ with a fixed power growth reads as follwos.

Theorem 6.6.1. *There exists a positive number $\gamma > d/2$ depending on $\mathbf{F}$, the underlying energy space $\mathbb{U}$, the spatial dimension d, the cancellation properties (CP) of Ψ such that*

$$|\langle \psi_\lambda, F(\mathbf{v})\rangle| = |\mathbf{F}(\mathbf{v})_\lambda| \lesssim \sup_{\mathbf{S}_\mu \cap \mathbf{S}_\lambda \neq \emptyset} |\mathbf{v}_\mu| \mathbf{2}^{-\gamma(|\lambda|-|\mu|)}. \tag{6.6.4}$$

For details and proofs, see [18].

Remark 6.6.2. *This result shows why a tree-structure is relevant. If $|\mu| > |\lambda|$ the bound becomes meaningless. It provides best information when λ is a leaf of the tree or close to a leave, see also Figure 13 for an illustration.*

The estimates (6.6.4) can be used to predict for a given input $\mathbf{v}$ and a given accuracy tolerance $\eta > 0$ a tree-structured index set $\Gamma_\eta \subset \Lambda$ with the following properties.

Theorem 6.6.2. *For $\mathbf{v}$ and $\eta > 0$ and Γ_η as above, one has*

$$\|\mathbf{F}(\mathbf{v}) - \mathbf{F}(\mathbf{v})|_{\Gamma_\eta}\|_{\ell_2} \le \eta. \tag{6.6.5}$$

Moreover, let $s^* := \frac{2\gamma-d}{2d}$. Then, for $\mathbf{v} \in \mathcal{A}^s((\Sigma)_{\mathbf{n}}, \ell_2(\Lambda))$ and $0 < s < s^*$ one has

$$\#(\Gamma_\eta) \lesssim \eta^{-1/s}\|\mathbf{v}\|^{\mathbf{1/s}}_{\mathcal{A}^s} + \#\Lambda_{\mathbf{0}}, \quad \|\mathbf{F}(\mathbf{v})\|_{\mathcal{A}^s} \lesssim \mathbf{1} + \|\mathbf{v}\|_{\mathcal{A}^s}. \tag{6.6.6}$$

with constants independent of $\mathbf{v}$.

One can then construct EVAL according to (6.6.3) such that the resulting residual approximation scheme RES is indeed s^*-computable for nonlinearities covered by the above results. As a consequence the assertions in Theorem 5.11.2 are valid for all such cases.

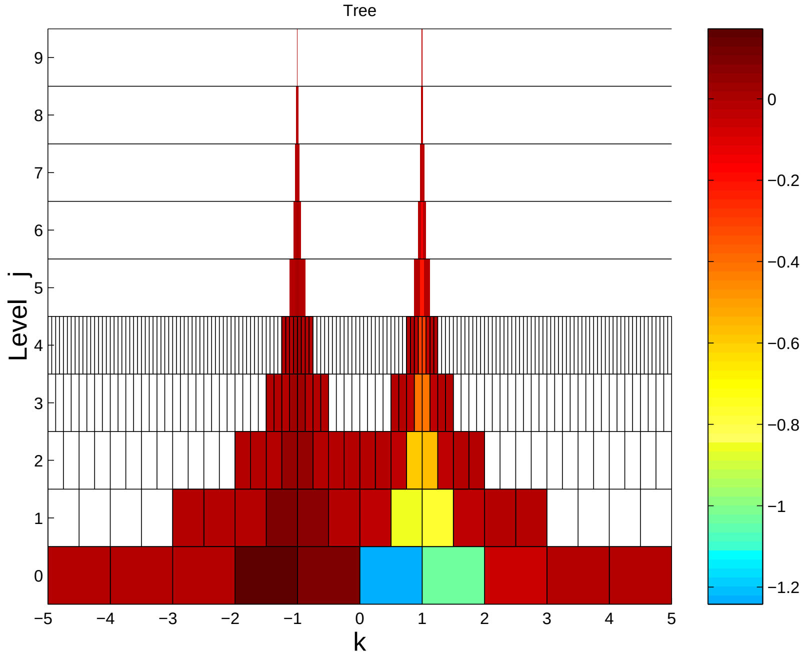


Figure 12: Iinput sequence with active coefficients under tree-structure constraints.

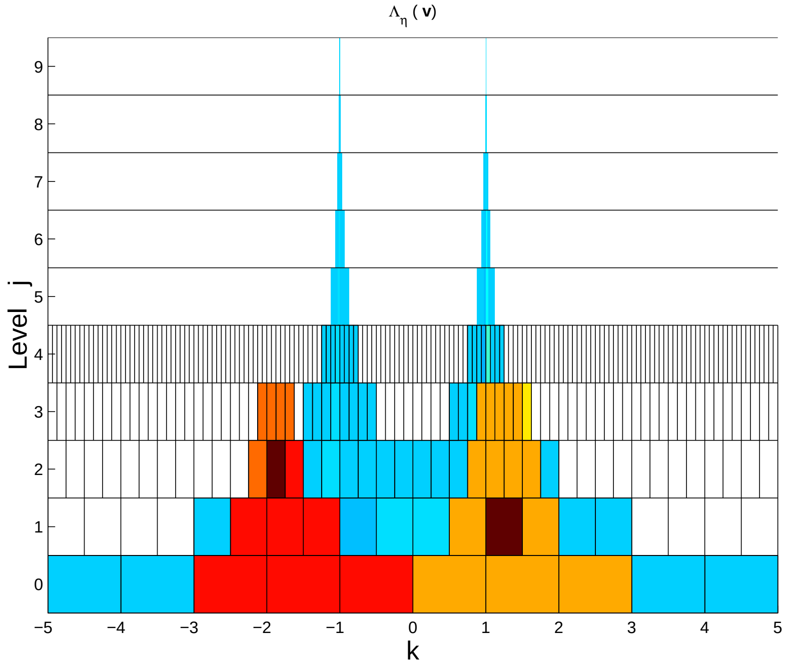


Figure 13: Resulting significant coefficients after application of a nonlinear mapping.

6.7 Can Coarsening be Avoided?

We finally discuss a version of Algorithm 5.10.1 where the use of COARSE is avoided, see [33] for all details.

This is done for the restricted problem class of symmetric positive definite operator representations $\mathbf{A} = \mathbf{A}^T$. On easily shows that $|||\mathbf{v}||| := \left(\mathbf{v}^\mathbf{T}\mathbf{A}\mathbf{v}\right)^{1/2}$ is an equivalent norm on $\ell_2(\Lambda)$ and

$$\|\mathbf{A}^{-1}\|^{-1/2}\|\mathbf{v}\|_{\ell_2} \leq |||\mathbf{v}||| \leq \|\mathbf{A}\|^{1/2}\|\mathbf{v}\|_{\ell_2}, \quad \mathbf{v} \in \ell_2(\Lambda),$$

(cf. the energy norm for elliptic problems). For simplicity we abbreviate in what follows

$$\|\cdot\|_{\ell_2(\Lambda)} =: \|\cdot\|.$$

The basic point of view is slightly different. Rather than thinking of a simple Richardson iteration it was first proposed in [15] to successively solve *defect problems*. It can be reinterpreted as a version of the general iteration

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \mathbf{C}(\mathbf{f} - \mathbf{A}\mathbf{u}^n), \quad n = 0, 1, 2, \ldots, \tag{6.7.1}$$

for the following particular "procedural" preconditioner $\mathbf{C}$. Suppose that $\mathbf{v}$ is a current approximation with finite support, find a possibly small "grown" index set $\Gamma \supset \operatorname{supp}\mathbf{v}$ which carries the "bulk" of the residual $\mathbf{f} - \mathbf{A}\mathbf{v}$, i.e., for some fixed $\theta \in (0, 1]$

$$\|P_\Gamma(\mathbf{f} - \mathbf{A}\mathbf{v})\| \geq \theta\|\mathbf{f} - \mathbf{A}\mathbf{v}\|, \tag{6.7.2}$$

where $P_\Gamma\mathbf{w}$ replaces all entries of $\mathbf{w}$ outside Γ by zero. Then one can show (using in particular the energy norm $|||\cdot|||$) that (see [33, Lemma 1.2])

$$\|\mathbf{u}_\Gamma - \mathbf{v}\| \leq (\mathbf{1} - \kappa(\mathbf{A})^{-1}\theta^2)\|\mathbf{u} - \mathbf{v}\|, \tag{6.7.3}$$

where $\mathbf{u}_\Gamma$ is the Galerkin solution in $\ell_2(\Gamma)$

$$P_\Gamma(\mathbf{A}\mathbf{u}_\Gamma) = P_\Gamma\mathbf{f}.$$

That means that the increased subspace $\ell_2(\Gamma) \subset \ell_2(\Lambda)$ provides a Galerkin solution with strict error decay. This is sometimes referred to as *saturation* property.

In terms of (6.7.1), this can be interpreted as follows:

1. given $\mathbf{u}^n$, find $\Gamma_{n+1} \supset \operatorname{supp}\mathbf{u}^n =: \Gamma_n$ such that (6.7.2) holds;
2. define the finite submatrix

$$\mathbf{A}_\Gamma := (\mathbf{A}_{\lambda,\lambda'})_{\lambda,\lambda' \in \Gamma}$$

and solve the Galerkin problem

$$\mathbf{A}_\Gamma\mathbf{w} = P_\Gamma\mathbf{f} \quad \Leftrightarrow \quad (\mathbf{A}\mathbf{w})^T\mathbf{v} = \mathbf{f}^\mathbf{T}\mathbf{v}, \quad \mathbf{v} \in \ell_2(\Gamma); \tag{6.7.4}$$

3. set

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \mathbf{w} = \mathbf{u}^{n+1} + \mathbf{A}_\Gamma^{-1}P_\Gamma(\mathbf{f} - \mathbf{A}\mathbf{u}^n). \tag{6.7.5}$$

Thus, the preconditioner $\mathbf{C} = \mathbf{C}_n = \mathbf{A}_\Gamma^{-1} P_\Gamma$ corresponds to the solution of a defect problem on the fixed subspace $\ell_2(\Gamma)$.

The reason why such "support growing" strategy could result in an optimal complexity is given by the following Lemma from [33].

Lemma 6.7.1. *Let* $\theta \in (0, \kappa(\mathbf{A})^{-1/2})$, $\mathbf{v} \in \ell_2(\Lambda)$ *and assume that* $\mathbf{u} \in \mathcal{A}^s = \mathcal{A}^s_\infty((\Sigma_n), \ell_2(\Lambda) = w\ell_\tau(\Lambda)$, $\frac{1}{\tau} = s + \frac{1}{2}$. *Then the smallest set* $\Gamma \supseteq \operatorname{supp} \mathbf{v}$ *with* (6.7.2)

$$\|P_\Gamma(\mathbf{f} - \mathbf{A}\mathbf{v})\| \geq \theta \|\mathbf{f} - \mathbf{A}\mathbf{v}\|$$

satisfies

$$\#(\Gamma \setminus \operatorname{supp} \mathbf{v}) \lesssim \|\mathbf{f} - \mathbf{A}\mathbf{v}\|^{-1/s} |\mathbf{u}|^{1/s}_{\mathcal{A}^s}. \tag{6.7.6}$$

This says that the growth-complexity scales with the optimal rate of the current error $\|\mathbf{u} - \mathbf{v}\| \sim \|\mathbf{f} - \mathbf{A}\mathbf{v}\|$ whenever $\mathbf{u} \in \mathcal{A}^s$.

In practice, one cannot determine the optimal growth-set Γ_{n+1} because the residual is in general infinitely supported. The idea is now to find a near-optimal growth-set using approximations to $\mathbf{f}$ (coarsening a prprocessed array) and an approximate application of $\mathbf{A}$ to the given approximation, using e.g. the APPLY routine. A successively grown support set is accepted if either the approximate residual is below a desired threshold or one has found some bulk of it.

GROW$[\mathbf{v}, \bar{\xi}, \eta] \to [\Gamma, \xi]$ according to the following steps:

```
1. choose constants $0 < \alpha < \omega$, $\frac{\alpha+\omega}{1-\omega} < \kappa(\mathbf{A})^{-1/2}$, set $\zeta := \frac{2\omega\bar{\xi}}{1-\omega}$
2. do $\zeta/2 \to \zeta$, $\mathbf{r} = $ COARSE$[\mathbf{f}, \zeta/2] - $ APPLY$[\mathbf{A}, \mathbf{v}, \zeta/\mathbf{2}] = $ RES$[\mathbf{A}, \mathbf{f}, \mathbf{v}, \zeta]$
   until $\xi := \|\mathbf{r}\| + \zeta \leq \eta$ or $\zeta \leq \omega\|\mathbf{r}\|$
3. if $\xi > \eta$
     determine a set $\Gamma \supset \operatorname{supp} \mathbf{v}$ with (up to a fixed constant factor) minimal cardinality such that $\|P_\Gamma \mathbf{r}\| \geq \alpha\|\mathbf{r}\|$
   else $\Gamma = \emptyset$
   endif
```

Obviously, the output depends on the choice of the parameter $\bar{\xi}$ that should ideally lie within or is not far from the interval $\frac{1-\omega}{1+\omega}\|\mathbf{f} - \mathbf{A}\mathbf{v}\|, \|\mathbf{f} - \mathbf{A}\mathbf{v}\|$, for a more detailed discussion, see [33, §2]).

One can then show that the estimate ξ for the current residual scales in the right way, namely

$$\frac{\alpha - \omega}{1 + \omega}\xi \leq \|P_\Gamma(\mathbf{f} - \mathbf{A}\mathbf{v})\|, \quad \#(\Gamma \setminus \operatorname{supp} \mathbf{v}) \lesssim \xi^{-1/s} |\mathbf{u}|^{1/s}_{\mathcal{A}^s}. \tag{6.7.7}$$

see [33, Theorem 2.4].

The next ingerdient is the approximate solution of the defect problem $\mathbf{A}\mathbf{w} = \mathbf{f} - \mathbf{A}\mathbf{v}$ projected to $\ell_2(\Gamma)$.

GALSOLVE$[\Gamma, \mathbf{f}_\Gamma, \mathbf{v}, \delta, \eta] \to \mathbf{v}_\Gamma$

1. set $\mathbf{f}_\Gamma = P_\Gamma \mathbf{f}$, δ should satisfy $\delta \geq \|\mathbf{f}_\Gamma - \mathbf{A}\mathbf{v}\|$
2. Fix a system matrix for Γ: choose J such that the compressed matrix $\mathbf{A}_J$ satisfies
$$\sigma := \|\mathbf{A} - \mathbf{A}_J\| \|\mathbf{A}^{-1}\| \leq \frac{\eta}{3\eta + 3\delta}$$
set $\mathbf{B} := P_\Gamma \frac{1}{2}(\mathbf{A}_J + \mathbf{A}_J^T)|_{\ell_2(\Gamma)}$
set $\mathbf{r}_0 := \mathbf{f}_\Gamma - P_\Gamma\big(\text{APPLY}[\mathbf{A}, \mathbf{v}, \eta/\mathbf{3}]\big)$
3. approximately solve $\mathbf{B}\mathbf{w} = \mathbf{r}_0$, i.e., find $\bar{\mathbf{w}}$ such that $\|\mathbf{r}_0 - \mathbf{B}\bar{\mathbf{w}}\| \leq \frac{\eta}{3}$ (e.g. using conjugate gradients)
set $\mathbf{v}_\Gamma = \mathbf{v} + \bar{\mathbf{w}}$

With these prerequisites one can formulate the following adaptive scheme.

Algorithm 6.7.1.

1. *Fix target accuracy $\epsilon > 0$ and an estimate ξ_{-1} for $\|\mathbf{f}\|$; take the parameters α, ω as in* GROW *and choose*
$$\gamma \in \Big(0, \frac{1}{6}\kappa(\mathbf{A})^{-1/2}\frac{\alpha - \omega}{1 + \omega}\Big), \quad \beta > 0$$
set $k = 0$, $\mathbf{v}^k = \mathbf{0}$
2. *while for* $[\Gamma_{k+1}, \xi_k] :=$ GROW$[\mathbf{v}^k, \beta\xi_{\mathbf{k-1}}, \epsilon]$, $\xi_\mathbf{k} > \epsilon$, *do*
```
    $\mathbf{f}^{k+1} := P_{\Gamma_{k+1}}\big(\text{COARSE}[\mathbf{f}, \gamma\xi_k]\big)$
    $\mathbf{v}^{\mathbf{k+1}} := \text{GALSOLVE}[\Gamma_{\mathbf{k+1}}, \mathbf{f}^{\mathbf{k+1}}, \mathbf{v}^\mathbf{k}, (\mathbf{1} + \gamma)\xi_\mathbf{k}, \gamma\xi_\mathbf{k}]$
    $k + 1 \to k$
```

Theorem 6.7.1. [33] *Algorithm 6.7.1 terminates with an approximation $\bar{\mathbf{v}}$ to $\mathbf{u}$ such that $\|\mathbf{f} - \mathbf{A}\bar{\mathbf{v}}\| \leq \epsilon$ (hence $\|\mathbf{u} - \bar{\mathbf{v}}\| \leq \|\mathbf{A}^{-1}\|\epsilon$). Moreover, when $\mathbf{u} \in \mathcal{A}^s$ for some $s < s^*$ (the compressibility limit of $\mathbf{A}$) then one has*
$$\#(\text{supp}\,\bar{\mathbf{v}}) \lesssim \epsilon^{-1/s}|\mathbf{u}|_{\mathcal{A}^s}^{1/s}$$
and the number of operations remains bounded by an absolute constant multiple of $\epsilon^{-1/s}|\mathbf{u}|_{\mathcal{A}^s}^{1/s}$.

This result can be extended to semi-linear problems but ne needs very fine stimates for the involved constants. As mentioned earlier, a currently hot context where such methods in coefficient space are very relevant are high dimensional or parameter dependent PDEs.

References

[1] R.A. Adams, *Sobolev Spaces*, Academic Press, 1978.

[2] M. Bachmayr , A. Cohen, W. Dahmen, Parametric PDEs: Sparse or low-rank approximations? IGPM Preprint #454, RWTH Aachen, July 14, 2016. http://arxiv.org/abs/1607.04444 [math.NA].

[3] A. Barron, A. Cohen, W. Dahmen, R. DeVore, Approximation and learning by greedy algorithms, Annals of Statistics, 3(No 1)(2008), 64–94.

[4] G. Beylkin, R. R. Coifman, and V. Rokhlin, Fast wavelet transforms and numerical algorithms I, Comm. Pure and Appl. Math., **44** (1991), 141–183.

[5] J. Bergh, J. Löfström, *Interpolation Spaces, An Introduction*, Springer, 1976.

[6] P. Binev, R. DeVore Fast Computation in Adaptive Tree Approximation, Numer. Math. 97 (2004), 193–217.

[7] D. Braess, *Finite Elemente*, 3rd edition, Springer, 2003.

[8] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Chapman & Hall/CRC, 1993.

[9] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, 1991.

[10] A. Canuto, A. Tabacco, K. Urban, The wavelet element method, part I: Construction and analysis, Appl. Comp. Harm. Anal., 6(1999), 1–52.

[11] A. Canuto, A. Tabacco, K. Urban, The wavelet element method, part II: Realization and additional features, Appl. Comp. Harm. Anal.

[12] J.M. Carnicer, W. Dahmen and J.M. Peña, Local decomposition of refinable spaces, Appl. Comp. Harm. Anal. 3 (1996), 127-153.

[13] A. Cohen, Wavelet methods in numerical analysis, in the Handbook of Numerical Analysis, vol. VII, P.-G. Ciarlet et J.-L. Lions eds., Elsevier, Amsterdam, 2000.

[14] A. Cohen, Numerical Analysis of Wavelet Methods, Series in Mathematics and its Applications, Elsevier, Amsterdam (2003)

[15] A. Cohen, W. Dahmen, R. DeVore, Adaptive wavelet methods for elliptic operator equations – Convergence rates, Math. Comp. 70 (2001), 27–75.

[16] A. Cohen, W. Dahmen, R. DeVore, Adaptive wavelet methods II - Beyond the elliptic case, Foundations of Computational Mathematics, 2 (2002), 203–245.

[17] A. Cohen, W. Dahmen, R. DeVore, Adaptive Wavelet Schemes for Nonlinear Variational Problems, SIAM J. Numer. Anal., (5)41(2003), 1785–1823.

[18] A. Cohen, W. Dahmen, R. DeVore, Sparse evaluation of compositions of functions using multiscale expansions, SIAM J. Math. Anal., 35 (2003), 279–303.

[19] A. Cohen, W. Dahmen, I. Daubechies, R. DeVore, Tree approximation and optimal encoding, Applied and Computational Harmonic Analysis, 11 (2001), 192–226.

[20] S. Dahlke, W. Dahmen, K. Urban, Adaptive wavelet methods for saddle point problems – Convergence rates, SIAM J. Numer. Anal., 40 (No. 4) (2002), 1230–1262.

[21] S. Dahlke, R. DeVore, Besov regularity for elliptic boundary value problems, Comm. Partial Differential Equations, 22 (1997), 1–16.

[22] W. Dahmen, Stability of multiscale transformations, Journal of Fourier Analysis and Applications, 2 (1996), 341-361.

[23] W. Dahmen, Wavelet and Multiscale Methods for Operator Equations, Acta Numerica, Cambridge University Press, 6(1997), 55–228.

[24] W. Dahmen, A. Kunoth, K. Urban, Biorthogonal spline-wavelets on the interval – Stability and moment conditions, Applied and Computational Harmonic Analysis, 6 (1999), 132–196.

[25] W. Dahmen, R. Schneider, Composite wavelet bases for operator equations, Math. Comp., 68 (1999), 1533-1567.

[26] W. Dahmen, R. Schneider, Wavelets on manifolds I. Construction and domain decomposition, SIAM Journal on Mathematical Analysis, 31 (1999), 184-230.

[27] W. Dahmen, R. Schneider, Wavelets with complementary boundary conditions – Function spaces on the cube, Results in Mathematics, 34 (1998), 255–293.

[28] W. Dahmen, H. Harbrecht, R. Schneider, Compression Techniques for Boundary Integral Equations – Optimal Complexity Estimates, SIAM J. Numer. Anal., 43:2251–2271 (2006).

[29] W. Dahmen, R. Stevenson, Element-by element construction of wavelets satisfying stability and moment conditions, SIAM Journal on Numerical Analysis, 37 (1999), 319–325.

[30] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Math. 61, SIAM, Philadelphia, 1992.

[31] R. DeVore, Nonlinear approximation, Acta Numerica, **7**, Cambridge University Press, 1998, 51-150.

[32] R. DeVore and G.G. Lorentz, *Constructive Approximation*, vol. 303, Springer Grundlehren, Springer, Berlin-Heidelberg, 1993.

[33] T. Gantumur, H. Harbrecht, R.P. Stevenson, An optimal adaptive wavelet method without coarsening of the iterands, Math. Comp., 76 (2007), 615–629.

[34] T. Gantumur, R.P. Stevenson, Computation of differential operators in wavelet coordinates, Math. Comp., 75(2006), 697–709 .

[35] T. Gantumur, R.P. Stevenson, Computation of singular integral operators in wavelet coordinates, Computing, 76(2006), 77–107.

[36] S. Jaffard, Wavelet methods for fast resolution of elliptic equations, SIAM J. Numer. Anal. 29 (1992), 965–986.

[37] T. von Petersdorff, and C. Schwab, Fully discrete multiscale Galerkin BEM, in: *Multiscale Wavelet Methods for PDEs*, W. Dahmen, A. Kurdila, and P. Oswald (eds.), Academic Press, San Diego, 1997, 287–346.

[38] R. Schneider, *Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur Lösung großer vollbesetzter Gleichungssysteme*, B.G. Teubner, Stuttgart, 1998.

[39] H. Johnen, K. Scherer, On the equivalence of the K-functional and moduli of continuity and some applications, Lecture Notes Math., 571 (1976), pp. 119–140

[40] S. Sauter, C. Schwab, Randelementmethoden: Analyse, Numerik und Implementierung schneller Algorithmen, B.G. Teubner, Stuttgart, 2005.

[41] R. P. Stevenson, On the compressibility of operators in wavelet coordinates, SIAM J. Math. Anal., 35(2004), 1110–1132.

[42] P. Tchamitchian, Wavelets, Functions, and Operators, in: *Wavelets: Theory and Applications*, G. Erlebacher, M.Y. Hussaini, and L. Jameson (eds.), ICASE/LaRC Series in Computational Science and Engineering, Oxford University Press, 1996, 83–181.