# Lecture notes on defect correction

P.W. Hemker

Version: September 5, 2001

RATIO RES TEMPERAT OMNES

# Contents

# Chapter 1

# Defect Correct Processes

## 1.1 Introduction

Many problems in numerical mathematics can be cast into the form of an equation

$$Fz = y,$$

Here $z \in D$ is an unknow quantity or function; a right-hand-side $y \in \hat{D}$ and a mapping $F : D \subset E \to \hat{D} \subset \hat{E}$ are given; $E$ and $\hat{E}$ are linear spaces. The element $z \in D$ has to be found such that the equation $Fz = y$ is satisfied. Frequently we cannot or do not want to solve the above mentioned equation directly, because this would exceed our computational capabilities. On the other hand we may be able solve simpler equations that are all similar to the previous equation:

$$\tilde{F}\tilde{z} = \tilde{y},$$

for some approximation $\tilde{F} : D \to \hat{D}$ of the operator $F$ and for arbitrary $\tilde{y} \in \hat{D}$. Sometimes this yields the possibility to solve the original equation by means of an iterative process known as a DCP: Defect Correction Process [22] [3].

Defect correction processes are based on the following idea:

- let an *initial approximation* $z_0$ for the solution to the original equation be given,
- consider the *defect* $d_0 := F(z_0) - y$ of this initial approximation as a quantity which indicates to what extent the problem has (not) been solved,
- use this information in a simplified version of the problem, i.e. consider the approximate operator $\tilde{F}$, to obtain an appropriate *correction* quantity,
- apply this correction to the initial approximation to obtain a new (hopefully better) approximation.

Of course, the above process may now be repeated, using the newly obtained approximation as a new 'initial' approximation.
A few instances of the basic principle are well known. We mention a few examples:

**Example 1.1** *Newton's method*
We are interested in computing a zero of the nonlinear function $F$. The equation to be solved is then given by

$$F(z) = 0.$$

Let an initial approximation $z_0$ for the solution be given. A (hopefully convergent) sequence $z_1, z_2, \ldots$ of approximations for the solution is then generated by

$$z_{i+1} \,=\, z_i \,-\, (F'(z_i))^{-1} d_i\,, \qquad i \,=\, 0, 1, 2, \ldots,$$

where the defect $d_i$ is defined by

$$d_i \,=\, F(z_i) \,-\, 0 \,=\, F(z_i)\,.$$

Note that in each step of this iterative process a different approximation $\tilde{F}_i$ for the function $F$ is used. In the $i$-th step this approximating function $\tilde{F}_i$ is a local linearisation of $F$, given by:

$$\tilde{F}_i(z) \,=\, F(z_i) \,+\, F'(z_i)\,(z \,-\, z_i)\,,$$

where $F'(z_i)$ is defined by

$$F(z) \,=\, F(z_i) \,+\, F'(z_i)\,(z \,-\, z_i) \,+\, \mathcal{O}(\|z - z_i\|^2).$$

**Example 1.2** *Iterative refinement*
A second example of a DCP is the iterative refinement of a given approximate solution for a linear system. Usually one obtains an approximate solution $z_0$ for the linear system $Az \,=\, y$ by computing a decomposition $A \,=\, LU$ ($L$ and $U$ respectively lower and upper triangular matrices), and then solving the two triangular systems $Lw \,=\, y$, $Uz \,=\, w$ directly. The approximation $z_0$ will be contaminated by rounding errors that affect the matrix-decomposition and the solution of the triangular systems; $z_0$ can be improved by the following DCP:

$$z_{i+1} \,=\, z_i \,-\, e_i, \quad i \,=\, 0, 1, 2, \ldots,$$

with $e_i$ the solution of the system

$$LUe_i \,=\, d_i, \qquad \text{where} \quad d_i \,=\, Az_i - y.$$

## 1.2 Elementary Defect Correction Processes

We consider the equation

$$Fz \,=\, y, \tag{1.1}$$

where $F : D \subset E \to \hat{D} \subset \hat{E}$ is a bijective, continuous, generally nonlinear operator; $E$ and $\hat{E}$ are Banach spaces. The domain $D$ and the range $\hat{D}$ are closed subsets given with $F$; $\hat{D}$ contains an appropriate neighbourhood of $y$. Hence, for every $\tilde{y} \in \hat{D}$ there exists, in $D$, exactly one solution of $Fz \,=\, \tilde{y}$. The solution of the given equation (1.1) is denoted by $z^*$.

We call the problem of finding $z$ such that $Fz \,=\, \tilde{y}$ (for a given $\tilde{y} \in \hat{D}$) a *neighbouring problem*. In order to introduce the Defect Correction Processes to solve the equation (1.1), we make the following assumptions:

- We assume that the *defect*

  $$d(\tilde{z}) \,:=\, F\tilde{z} \,-\, \tilde{y}$$

  can be evaluated for approximate solutions $\tilde{z} \in D$ to all neighbouring problems $F\tilde{z} \,=\, \tilde{y}$.

- Furthermore, we assume that we can readily solve the *approximate problem*

  $$\tilde{F}z \,=\, \tilde{y}, \tag{1.2}$$

for $\tilde{y} \in \hat{D}$, i.e. we assume that we can evaluate the solution operator $\tilde{G}$ of (1.2). In other words, we assume the existence of $\tilde{G} : \hat{D} \to D$, an *approximate inverse* of $F$ such that (in some appropriate sense)

$$\tilde{G}F\,\tilde{z} \approx \tilde{z} \quad \text{for} \quad \tilde{z} \in D,$$

and

$$F\tilde{G}\,\tilde{y} \approx \tilde{y} \quad \text{for} \quad \tilde{y} \in \hat{D}.$$

Let us assume that we know some approximation $\tilde{z} \in D$ for $z^*$ and that we have computed its defect

$$d(\tilde{z}) := F\tilde{z} - y = F\tilde{z} - Fz^*.$$

In the general (nonlinear) case, there are two simple ways to use this information for the computation of a (hopefully better) approximation $\tilde{\tilde{z}}$ by means of solving problems of type (1.2):

**(A)** The error $\Delta z$ satisfies the relation:

$$\Delta z := \tilde{z} - z^* = F^{-1}(y + d(\tilde{z})) - F^{-1}y.$$

If we could compute $\Delta z$ then $\tilde{z} - \Delta z$ would yield the exact solution to our original problem. Not being able to evaluate $F^{-1}$, we use the approximate inverse $\tilde{G}$ instead and compute $\Delta\tilde{z} = \tilde{G}(y + d(\tilde{z})) - \tilde{G}y$. We then use $\Delta\tilde{z}$ as a correction for $\tilde{z}$. The new approximation $\tilde{\tilde{z}}$ is then given by:

$$\begin{aligned} \tilde{\tilde{z}} &:= \tilde{z} - \Delta\tilde{z} = \tilde{z} - [\tilde{G}(y + d(\tilde{z})) - \tilde{G}y] = \\ &= \tilde{z} - \tilde{G}F\tilde{z} + \tilde{G}y = (I - \tilde{G}F)\,\tilde{z} + \tilde{G}y. \end{aligned}$$

We notice that $\tilde{\tilde{z}}$ is a sum (in $E$) of elements in $D$.

**(B)** The second approach is dual to (A) in the sense that (B) uses the addition in $\hat{D}$, whereas (A) uses the addition in $D$. Suppose we have $l \in \hat{D}$. We then may consider $\tilde{z} = F^{-1}l$ as an approximation for $z^*$. The defect for $\tilde{z}$ is:

$$d(\tilde{z}) = F\tilde{z} - y = l - y.$$

If we take $\tilde{l} = l - d(\tilde{z})$, we obtain the exact solution $z^*$ by $z^* = F^{-1}\tilde{l}$. Because we do not want to solve (1.1) directly, we use the approximate inverse $\tilde{G}$ instead of $F^{-1}$. We then obtain:

$$\tilde{z} = \tilde{G}\tilde{l},$$

$$d(\tilde{z}) = F\tilde{G}\tilde{l} - y,$$

$$\tilde{\tilde{l}} = \tilde{l} - d(\tilde{z}) = (I - F\tilde{G})\tilde{l} + y.$$

The iterative use of these basic defect correction processes leads to the following algorithms:

$$\text{(DCPA)} \qquad \begin{cases} z_0 &= \tilde{G}y, \\ z_{i+1} &= (I - \tilde{G}F)z_i + \tilde{G}y, \end{cases}$$

and

$$\text{(DCPB)} \qquad \begin{cases} l_0 &= y, \\ l_{i+1} &= (I - F\tilde{G})\,l_i + y. \end{cases}$$

In the latter process we define approximations $z_i$ by $z_i = \tilde{G}l_i$.
**Remarks:**

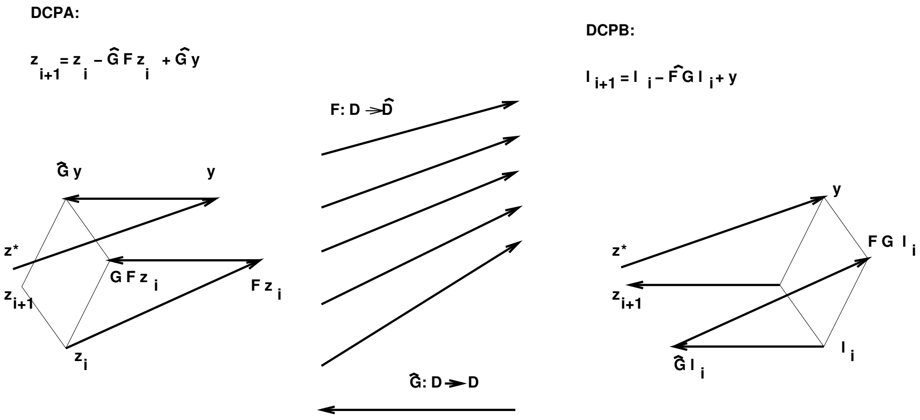


Figure 1.1: DCPA and DCPB

- In (DCPA) or (DCPB) we formulate both the iterative process and the (standard) initial approximant.
- It is essential that $\tilde{G}$ is relatively simple, i.e. it is much easier to find a solution for (1.2) than for (1.1).
- It is the existence of the approximate inverse $\tilde{G}$ which is essential, it is *not* the existence of the approximate operator $\tilde{F}$.

If and only if the mapping $\tilde{G} : \hat{D} \to D$ is injective, its left inverse $\tilde{F}$ exists[1]. Similarly the right inverse $\tilde{F} : D \to \hat{D}$ of $\tilde{G}$ exists if and only if $\tilde{G}$ is surjective[2]. In general our approximate inverse $\tilde{G}$ doesn't need to be linear and is neither necessarily injective nor surjective.

If $\tilde{G}$ is injective, its left inverse $\tilde{F}$ exists and we can write DCPB as

$$(\text{DCPB}^\star) \qquad \left\{ \begin{array}{lcl} \tilde{F}z_0 & = & y, \\ \tilde{F}z_{i+1} & = & \tilde{F}z_i - Fz_i + y, \end{array} \right.$$

or, equivalently,

$$\left\{ \begin{array}{lcl} z_0 & = & \tilde{G}y, \\ z_{i+1} & = & \tilde{G}[(\tilde{F} - F)z_i + y]. \end{array} \right.$$

In some applications, the operator $\tilde{F} - F$ is much simpler to evaluate than $F$, so that there is an advantage in using this approach.

In case of a linear operator $\tilde{G}$ we can write both DCPA and DCPB as:

$$(\text{DCPL}) \qquad \left\{ \begin{array}{lcl} z_0 & = & \tilde{G}\,y, \\ z_{i+1} & = & z_i - \tilde{G}\,(Fz_i - y) = z_i - \tilde{G}\,d(z_i). \end{array} \right.$$

**Definition 1.3** A mapping $f : X \to Y$ is called *affine* if there exists a constant element $c \in Y$ such that $f(\cdot) - c : X \to Y$ is a linear mapping.

[1]Because $\tilde{G} : \hat{D} \to D$ injective is equivalent with

$$\forall\, z \in \tilde{G}(\hat{D}) \subset D \quad \exists!\, x \in D \quad \tilde{G}x = z\,,$$

the mapping $\tilde{F} : \tilde{G}(\hat{D}) \to \hat{D}$ which maps $z$ onto the uniquely determined $x \in \hat{D}$ such that $\tilde{G}x = z$, is the left inverse of $\tilde{G}$, i.e. $\tilde{F}\tilde{G} = I_{\hat{D}}$,

[2]We then have $\tilde{G}\tilde{F} = I_D$.

This definition implies:

- $f(0) = c$,
- $f(x+y) = f(x) + f(y) - c$,
- $f(x-y+z) = f(x) - f(y) + f(z)$,

**Theorem 1.4** If $\tilde{G}$ is an affine mapping, then the sequences $\{z_i\}$ in DCPA and $\{z_i\}$ in DCPB are identical.
**Proof:** Let $\{l_i\}_{i=0,1,2,\ldots}$, and $\{z_i\}_{i=0,1,2,\ldots}$, be defined as in DCPB, then:

$$\begin{aligned}
z_0 &= \tilde{G}\,l_0 = \tilde{G}\,y\,, \\
z_{i+1} &= \tilde{G}\,l_{i+1} = \tilde{G}\,(l_i - F\tilde{G}\,l_i + y) \\
&= \tilde{G}\,l_i + \tilde{G}y - \tilde{G}\,(F\tilde{G}l_i) \\
&= z_i + \tilde{G}y - \tilde{G}Fz_i = (I - \tilde{G}F)z_i + \tilde{G}y.
\end{aligned}$$

This means that the values from this sequence $\{z_i\}$ satisfy exactly the generation rules for the sequence $\{z_i\}$ from DCPA. Hence both sequence are identical. ■

## 1.3 Convergence of the basic defect correction processes

In the following $F : D \subset E \to \hat{D} \subset \hat{E}$ is a general nonlinear operator.

**Definition 1.5** $F$ is called *bounded* if bounded subsets of $D$ are mapped onto bounded subsets in $\hat{D}$, and $F$ is called *Lipschitz* if

$$\exists\, k > 0 \;\; \forall\, x, y, \in D \;\; \|Fx - Fy\|_{\hat{E}} \le k\,\|x - y\|_E\,.$$

The *Lipschitz constant* $|||F|||$ is defined by

$$|||F||| = \sup_{x,y\in D, x\neq y} \frac{\|Fx - Fy\|_{\hat{E}}}{\|x-y\|_E}\,,$$

$F$ is called a *(strict) contraction (mapping)* if $|||F||| < 1$; $F$ is *non-expansive* if $|||F||| \le 1$.

Apparently

$$\|Fx - Fy\|_{\hat{E}} \le |||F|||_{D\subset E\to\hat{D}\subset\hat{E}}\,\|x - y\|_E,$$

and for a linear operator $F$

$$|||F||| = \|F\|\,.$$

**Definition 1.6** An iterative process $z^{(i+1)} = H(z^{(i)}, z^{(i-1)}, \ldots)$ has a *fixed point* $z^\star$ if $z^\star = H(z^\star, z^\star, \ldots)$.

Because $D$ (or $\hat{D}$) is a closed subset of a Banach space $E$ (or $\hat{E}$), the set $D$ (or $\hat{D}$) is a complete metric space. For contraction mappings of a complete metric space into itself, we have the following important theorem.

**Theorem 1.7** *(Banach's contraction principle).* Let M be a contraction mapping of a complete metric space $D$ into itself. Then M has a unique fixed point $u$ in $D$. Moreover, if $x_0$ is an arbitrary point in $D$, and $\{x_n\}$ is defined by

$$x_{n+1} = Mx_n, \quad (n = 0, 1, 2, \ldots),$$

then $\lim_{n\to\infty} x_n = u$ and

$$\|x_n - u\| \le \frac{|||M|||^n}{1-|||M|||}\, \|x_1 - x_0\| \tag{1.3}$$

**Proof:** Let $x_0 \in D$ and let $x_{n+1} = Mx_n, \quad (n = 0,1,2,...)$, and set $k = |||M|||$. Then we have

$$\|x_{r+1} - x_{s+1}\| = \|Mx_r - Mx_s\| \le k\|x_r - x_s\|$$

and hence

$$\|x_{r+1} - x_r\| \le k^r \|x_1 - x_0\|.$$

For given $p$ and $q$ with $p > q$ we have

$$\begin{aligned} \|x_p - x_q\| &\le k^q \|x_{p-q} - x_0\| \\ &\le k^q\{\|x_{p-q} - x_{p-q-1}\| + ... + \|x_1 - x_0\|\} \\ &\le k^q\{k^{p-q-1} + ... + k + 1\}\|x_1 - x_0\| \\ &\le \tfrac{k^q}{1-k}\|x_1 - x_0\|. \end{aligned} \tag{1.4}$$

The right-hand side of eq.(1.4) tends to zero as $q \to \infty$. Hence the sequence $\{x_n\}$ is Cauchy, and since $D$ is complete, $\{x_n\}$ converges to an element $u$ of $D$. As $\|x_{n+1} - Mu\| = \|Mx_n - Mu\| \le k\|x_n - u\|$ and $\|x_n - u\| \to 0$ for $n \to \infty$, we have $Mu = \lim_{n\to\infty} x_{n+1} = u$, i.e. $u$ is a fixed point of $M$.

For uniqueness, suppose $v$ is another fixed point of $M$, $v \in D$, $v \ne u$, and $v = Mv$, then $\|u - v\| = \|Mu - Mv\| \le k\|u - v\|$. This gives $(1-k)\|u - v\| \le 0$. Since $1 - k > 0$ we have $\|u - v\| = 0$, i.e. $u = v$.

To obtain eq.(1.3) we have

$$\|u - x_n\| \le \|u - x_p\| + \|x_p - x_n\| \le \|u - x_p\| + \frac{k^n}{1-k}\|x_1 - x_0\|$$

for $n < p$ by (1.4). Letting $p \to \infty$, we obtain

$$\|u - x_n\| \le \frac{k^n}{1-k}\|x_1 - x_0\|.$$

■

For many generalisations of the Banach contraction principle see [13]

### 1.3.1 Convergence of DCPA

For DCPA we have $z_{i+1} - z^* = (I - \tilde{G}F)z_i + \tilde{G}Fz^* - z^*$, hence

$$z_{i+1} - z^* = (I - \tilde{G}F)z_i - (I - \tilde{G}F)z^*.$$

**Definition 1.8** We define the *amplification operator of the error* in DCPA to be

$$M_A = I - \tilde{G}F.$$

The exact solution $z^*$ of (1.1) is a fixed point of the iteration DCPA, i.e. $z^* = (I - \tilde{G}F)z^* + \tilde{G}y$. Moreover, for any fixed point $\hat{z}$ of DCPA we have:

$$\hat{z} = (I - \tilde{G}F)\hat{z} + \tilde{G}y,$$

and hence

$$\tilde{G}F\hat{z} = \tilde{G}y = \tilde{G}Fz^*.$$

As a direct consequence of this we find the following

**Theorem 1.9** If DCPA has a fixed point $\hat{z} \in D$ with $F\hat{z} \in \hat{D}$ and if $\tilde{G}$ is injective, then $F\hat{z} = y$, i.e. then $\hat{z}$ is a solution of (1.1).

The convergence of DCPA clearly depends on the *contractivity* of the amplification operator $M_A = I - \tilde{G}F$. We formulate the following

**Theorem 1.10** Let $M_A : D \to D$ be a contraction and let $\tilde{G} : \hat{D} \to D$ be injective. Then DCPA converges to the solution $z^*$ of (1.1).

**Proof:** The DCPA iteration operator $A : D \to D$ is given by:

$$Az = (I - \tilde{G}F)z + \tilde{G}y. \tag{1.5}$$

Clearly $A$ is a contraction on (the complete metric space) $D$ as well, hence a unique fixed point $\hat{z}$ for $A$ exists and DCPA converges to this fixed point $\hat{z}$. By theorem 1.9 we know that $\hat{z}$ is a solution of (1.1). By assumption $F$ is bijective, hence $\hat{z}$ is the unique solution of (1.1). ∎

**Remark:** Even, if $\tilde{G}$ is not injective, then the solution $z^*$ of (1.1) and the fixed point $\hat{z}$ of DCPA are mapped by $\tilde{G}F$ onto the same element of $\tilde{G}(\hat{D})$, although we have not necessarily $F\hat{z} = y = Fz^*$. In other words: if $\tilde{G}$ is not injective, $\tilde{G}$ defines equivalence classes in $\hat{D}$, viz. the classes of points that are all mapped to the same point of $D$. Now $F\hat{z}$ and $Fz^*$ are elements of the same equivalence class.

### 1.3.2 Convergence of DCPB

For DCPB we have, with $\tilde{G}l^* = z^*$,

$$l_{i+1} - l^* = (I - F\tilde{G})\, l_i - (I - F\tilde{G})\, l^*.$$

**Definition 1.11** We define the *amplification operator of the residual* in DCPB to be

$$M_B = I - F\tilde{G}.$$

For any fixed point $\hat{l}$ of DCPB we have: $\hat{l} = (I - F\tilde{G})\,\hat{l} + y$, hence $F\tilde{G}\,\hat{l} = y = Fz^*$. As direct consequence of this, we find the following.

**Theorem 1.12** If DCPB has a fixed point $\hat{l} \in \hat{D}$ then $\tilde{G}\hat{l}$ is a solution of (1.1) in $\tilde{G}(\hat{D})$. Because we have assumed $F$ to be injective, we also know that $\tilde{G}\hat{l}$ is the unique solution of (1.1).

**Remarks:**

- The convergence of DCPB depends on the contractivity of the amplification operator $M_B = I - F\tilde{G}$.
- If $\tilde{G} : \hat{D} \to D$ is not surjective, it may occur that solutions $z^*$ of (1.1) have the property that $z^* \notin \tilde{G}(\hat{D})$ and hence no $\hat{l} \in \hat{D}$ exists such that $\tilde{G}\hat{l} = z^*$. In that case no fixed point $\hat{l} \in \hat{D}$ can exist. See also example 1.18 in Section 1.5.

**Theorem 1.13** Let $M_B : \hat{D} \to \hat{D}$ be a contraction on the Banach space $\hat{D}$. Then DCPB converges to an element $l^* \in \hat{D}$ such that $z^* = \tilde{G}l^*$ is the solution of (1.1).

### 1.3.3 Convergence of DCPL

For DCPL (with $\tilde{G}$ linear) we have:

$$\begin{aligned} z_{i+1} - z^* &= z_i - \tilde{G}Fz_i + \tilde{G}Fz^* - z^* \\ &= z_i - z^* - \tilde{G}(Fz_i - Fz^*) \\ &= (I - \tilde{G}F)z_i - (I - \tilde{G}F)z^*. \end{aligned}$$

If $F$ is linear as well, the amplification operator $M_L = I - \tilde{G}F$ of the error and the amplification operator $\overline{M}_L = I - F\tilde{G}$ of the residual are both linear operators and for the convergence of the iteration we can consider $\|M_L\|$, $\|\overline{M}_L\|$, $\rho(M_L)$, $\rho(\overline{M}_L)$ . (We notice that $\rho(M_L) = \rho(\overline{M}_L) = \lim_{n\to\infty}(\|M_L^n\|)^{1/n)}$ .

## 1.4 More elaborate versions of the principle

In this section we extend the idea of the defect correction process in several ways. First, we allow different approximate inverses to serve in one iteration process and we consider the process obtained when a fixed combination of approximate inverses is used repeatedly in a defect correction process. Secondly, it is possible to substitute different operators $F_i$ for $F$ during iteration. Further, we describe the iterative and the recursive application of the defect correction principle.

### 1.4.1 Non-stationary DCPs

We can use different approximate inverses in each iteration step and thus obtain non-stationary DCPs. Then the iteration steps of DCPA and DCPB read respectively

$$z_{i+1} = (I - \tilde{G}_i F)z_i + \tilde{G}_i y,$$

and

$$l_{i+1} = (I - F\tilde{G}_i)l_i + y.$$

In this way we are able to adapt the approximate inverse during the iteration and we can try to find proper sequences $\{\tilde{G}_i\}$ to accelerate the convergence of the iteration. We mention three examples: (i), (ii) and (iii).

$$(i) \qquad \tilde{G}_i = \tilde{G}(z_i).$$

Here the approximate inverse depends on the last iterand computed. This is the case e.g. in Newton's method for the solution of the non-linear equations, where $\tilde{G}(z) = (F'(z))^{-1}$, with $F'$ the Fréchet derivative of the operator $F$ in the problem (1.1). See e.g. example 1.1 .

$$(ii) \qquad \tilde{G}_i = \tilde{G}(\omega_i).$$

The approximate inverse depends on a single real parameter $\omega_i$. This is the case e.g. in non-stationary relaxation processes for the solution of linear systems. The value $\omega_i$ can be taken from a fixed sequence of values or it can be computed adaptively during the iteration process.

$$(iii) \qquad \tilde{G}_i \in \{\tilde{G}_1, \tilde{G}_2\}.$$

Here, in each iteration step the approximate inverse is chosen from a set of two (or possibly more) fixed approximate inverses. This is the case e.g. in Brakhage's and Atkinson's methods for the solution of Fredholm integral equations of the 2nd kind. (See [1] and [4]).

### 1.4.2 Combined DCPs

We now assume that $F$ is linear and consider a fixed combination of two *linear* approximate inverse operators $\tilde{G}$ and $\tilde{\tilde{G}}$. Then we consider two iteration steps in the non-stationary DCPA in which, in turn, the one and the other of the two approximate inverses is used. These two iteration steps

$$z_{i+1/2} = (I - \tilde{G}F)z_i + \tilde{G}y$$

and

$$z_{i+1} = (I - \tilde{\tilde{G}}F)z_{i+1/2} + \tilde{\tilde{G}}y$$

combine into a single iteration step of the form

$$z_{i+1} = (I - \tilde{\tilde{G}}F)(I - \tilde{G}F)z_i + (\tilde{\tilde{G}} - \tilde{\tilde{G}}F\tilde{G} + \tilde{G})y =$$

$$= (I - (\tilde{\tilde{G}} - \tilde{\tilde{G}}F\tilde{G} + \tilde{G})F)z_i + (\tilde{\tilde{G}} - \tilde{\tilde{G}}F\tilde{G} + \tilde{G})y.$$

This is again an iteration step of type (DCPA) with the approximate inverse

$$\hat{G} = \tilde{\tilde{G}} - \tilde{\tilde{G}}F\tilde{G} + \tilde{G}.$$

The amplification operator of this new process is the product of the amplification operators of the elementary processes:

$$M = I - \hat{G}F = (I - \tilde{\tilde{G}}F)(I - \tilde{G}F).$$

Thus, the combination of two different DCP-steps (with linear $F$ and $\tilde{G}$) can be seen as one "big" DCP-step.

Again assuming that $F$ and $\tilde{G}$ are linear operators, we now consider $\sigma$ applications of the same approximate inverse. Using DCPA, this can be described in matrix notation as follows:

$$\begin{pmatrix} z_{i+\sigma} \\ y \end{pmatrix} = \begin{pmatrix} I - \tilde{G}F & \tilde{G} \\ 0 & I \end{pmatrix}^{\sigma} \begin{pmatrix} z_i \\ y \end{pmatrix}$$

$$= \begin{pmatrix} (I - \tilde{G}F)^{\sigma} & \sum_{m=0}^{\sigma-1}(I - \tilde{G}F)^m\tilde{G} \\ 0 & I \end{pmatrix} \begin{pmatrix} z_i \\ y \end{pmatrix}.$$

Thus, we see that these $\sigma$ applications of the same DCPA-step lead to the following process:

$$z_{i+1} = (I - \tilde{G}F)^{\sigma}z_i + \sum_{m=0}^{\sigma-1}(I - \tilde{G}F)^m\tilde{G}y.$$

Then the relation

$$\sum_{m=0}^{\sigma-1}(I - \tilde{G}F)^m\tilde{G} = [I - (I - \hat{G}F)^{\sigma}]F^{-1},$$

allows us to consider the above process as a new DCPA with amplification operator of the error

$$M = (I - \tilde{G}F)^{\sigma} ,$$

and approximate inverse

$$\hat{G} = \sum_{m=0}^{\sigma-1}(I - \tilde{G}F)^m\tilde{G} = [I - (I - \tilde{G}F)^{\sigma}]F^{-1}. \quad (1.6)$$

### 1.4.3 Iterative application of DCPs

We will now pay attention to another possibility mentioned before, viz. the substitution of different operators $F_i$ for $F$ during iteration. This is important if we study discretised continuous problems. We consider all (discrete) $F_i$ as approximations to one (continuous) 'target' operator $F^\star$. As long as the approximate solution is not a very good approximation, i.e. in the beginning of the iteration, we take operators $\{F_i\}$ that are simple to evaluate and we will take better approximations that converge to $F^\star$ in some sense as the iteration proceeds.

If we apply this technique, we approximately solve a sequence of problems $\{P_k\}_{k=1,2,\ldots}$, of the form

$$(P_k) \qquad\qquad F_k z = y_k,$$

where we use the approximate solution of $(P_{k-1})$ as a starting value for the iteration of $(P_k)$.

One possible application is to select $\{F_i\}$ which are discrete approximations of *higher and higher order* to a continuous operator $F$. The approximate inverse $\tilde{G} = F_0^{-1}$ may be kept constant during the process.

Another example is the Mesh Continuation Method in which $\{F_i\}$ are discretisations on *finer and finer meshes* of an analytic operator $F$. When combined with a multigrid technique for the solution of the discrete problems, this is called Nested Iteration ([9]) or Full Multigrid (FMG) [8].

### 1.4.4 Recursive application of DCPs

Generally, the evaluation of the approximate inverse operator $\tilde{G}_i$ implies the solution of an equation which is (essentially) of a simpler type than the original equation. However, also this simpler equation may be of a kind that we want to solve by means of a DCP. For this we need an even simpler equation to solve, etc.. Thus, the execution of a single iteration step may imply the activation of a new (simpler to solve) DCP. In this way we can construct a recursive application of DCP's in which on the lowest level of recursion a very simple equation is to be solved. Multigrid iteration is an example of this procedure.

### 1.4.5 Generalisation for nonlinear problems: DCPN

A generalisation of DCPA, specially for nonlinear problems, can be introduced via DCPL. Let us first consider the process DCPA and suppose that $\tilde{G}$ is differentiable. Since $\tilde{G}'(\overline{y})$, $\overline{y} \in \hat{D}$, is a linear operator we can use it to obtain a linear approximation (linearisation) of the operator $\tilde{G}$. We then simply use DCPL with approximate inverse $\tilde{G}'(\overline{y})$:

$$z_{i+1} \approx z_i - \tilde{G}'(\overline{y})\,(F z_i - y).$$

For $\mu \in I\!R$, $\mu \neq 0$ and $u \in \hat{E}$ we have:

$$\tilde{G}'(\overline{y})u = \mu \tilde{G}'(\overline{y})\,\frac{u}{\mu} \approx \mu \tilde{G}(\overline{y} + \frac{u}{\mu}) - \mu \tilde{G}(\overline{y}).$$

Because, in general, the Fréchet derivative $\tilde{G}'(\overline{y})$ is not available for computation, we replace (using the above with $u = F z_i - y$) the form $\tilde{G}'(\overline{y})(F z_i - y)$ by

$$\mu \tilde{G}(\tilde{\overline{y}} + (F z_i - y)/\mu) - \mu \tilde{G}(\overline{y}).$$

We then obtain the following DCP:

$$(\text{DCPN}) \qquad \left\{ \begin{array}{lcl} z_0 & = & \tilde{G} y \\ z_{i+1} & = & z_i - \mu \tilde{G}(\overline{y} + (F z_i - y)/\mu) + \mu \tilde{G}(\overline{y}). \end{array} \right.$$

**Remarks:**

- In this new defect correction process one still has the freedom to choose the parameters $\mu$ and $\overline{y}$. For the choice $\overline{y} = y$ and $\mu = 1$, DCPN coincides with DCPA.
- For a large enough $\mu$, we may guarantee that for any defect $Fz_i - y$, the operator $\tilde{G}$ is evaluated only in a sufficiently small neighbourhood of $\overline{y}$.
- In the general case (i.e. for arbitrary values of $\mu$ and $\overline{y}$), the solution $z^*$ of (1.1) is a fixed point of DCPN. Sometimes the converse is also true:

**Theorem 1.14** Let $\hat{z}$ be a fixed point of DCPN and let $\tilde{G} : \hat{D} \to D$ be injective. Then $\hat{z}$ is the solution of (1.1).
**Proof:** Because $\hat{z}$ is a fixed point of DCPN we immediately see that

$$\tilde{G}(\overline{y} + (F\hat{z} - y)/\mu) = \tilde{G}(\overline{y}).$$

By the assumption that $\tilde{G}$ is injective, we see that $F\hat{z} - y = 0$. ■

## 1.5 Examples of defect correction processes

In Section 1.1 we already mentioned two examples of defect correction processes: Newton iteration and iterative refinement. We will return briefly to the latter example. In the case of iterative refinement for linear systems, the operator

$$F : \mathbb{R}^n \to \mathbb{R}^n$$

is a square matrix, and $z, y \in \mathbb{R}^n$ are $n$-vectors. The approximate inverse $\tilde{G}$ represents the numerical solution by means of (an approximation of ) a LU-decomposition, which has been obtained by numerical means and for which we may write

$$LU = F + E.$$

$E$ is the error in the LU-decomposition. The process of iterative refinement now reads:

$$\left\{ \begin{array}{lcl} LUz_0 & = & y \\ z_{i+1} & = & z_i - e_i \ , \quad i = 0, 1, 2, \dots . \end{array} \right.$$

with $e_i$ the solution of the system

$$LUe_i = d_i; \quad d_i = Fz_i - y.$$

Clearly, this is DCPA with approximate inverse

$$\tilde{G} = (F + E)^{-1}.$$

Because $\tilde{G}$ is linear, the above process is equivalent to a DCPB (Theorem 1.4).
We will now give a few more examples of defect correction processes.

**Example 1.15** *Relaxation methods*
All *stationary, fully consistent iterative methods of degree one* for the solution of linear systems $Ax = b$ can be written as

$$x_{i+1} = x_i - P(Ax_i - b),$$

where $P$ is a non-singular matrix (cf. [26]). These iterative methods are defect correction processes of type $A$ with approximate inverse $\tilde{G} = P$. They are equivalent to the corresponding

DCPBs, because $P$ is linear. Many of these methods (known as relaxation methods) are often used for the solution of special sparse linear systems. We introduce the following notation:

$$A = L + D + U;$$

$L$ is a strictly lower triangular matrix $D = \text{diag}\,(A)$, a diagonal matrix, and $U$ is a strictly upper triangular matrix. Using this notation, Table 1.1 summarises some possible choices for $\tilde{F} = P^{-1}$, together with the name of the corresponding iterative method.

| $\tilde{F} = P^{-1}$ | Name of the method | | Remarks |
|---|---|---|---|
| $D$ | J | Jacobi | |
| $\omega^{-1}D$ | JOR | | $\omega > 0$ |
| $D + L$ | GS | Gauss Seidel | |
| $\omega^{-1}D + L$ | SOR | | $\omega > 0$ |
| $p^{-1}I$ | RF | Stationary Richardson | $p$ scalar |
| $P^{-1}$ | GRF | Generalized Richardson | $P$ non-singular diagonal matrix |

Table 1.1: Relaxation methods

**Example 1.16** *A nonlinear two-point boundary-value problem (Stetter [22])*
We consider the two-point boundary-value problem

$$\begin{cases} u'' - e^u = 0, & \text{on } (-1,1), \\ u(-1) = u(1) = 0. \end{cases} \tag{1.7}$$

This defines the problem

$$Fu = 0,$$

where $F : C_0^2[-1,1] \to C(-1,1)$. Note that these spaces are Banach spaces with respect to the sup-norm.

We construct an approximate problem $\tilde{F}u = y$, $y \in C(-1,1)$, replacing $e^u$ by $0.99 + 0.81u$ (i.e. a reasonable approximation if $-0.4 \le u \le 0$ ). Thus we get the following approximate problem:

$$\begin{cases} u'' - 0.81u - 0.99 = y, & \text{on}(-1,1), \\ u(-1) = u(1) = 0. \end{cases}$$

This is a linear two-point boundary-value problem and we can write its solution explicitly as

$$u(t) = \int_{-1}^{1} K(t,z)\,[0.99 + y(z)]\,dz, \tag{1.8}$$

for some suitable kernel function $K(t,z)$. For this purpose we can use Green's function for problems of the following type:

$$u'' - \gamma^2 u = 0 \quad \text{on}\,(-1,1), \quad u(-1) = u(1) = 0.$$

For this problem the kernel function reads:

$$K(t,z) = \begin{cases} C \cdot \sinh(\gamma(1+t)) \cdot \sinh(\gamma(1-z)) & t \le z, \\ K(z,t) & t \ge z. \end{cases}$$

The integral operator (1.8) defines an approximate inverse $\tilde{G}$ for the problem (1.7). With this $\tilde{G}$ we can construct a DCPA or a DCPB to find the solution of (1.7). Both processes are equivalent since $\tilde{G}$ is an affine operator.

**Example 1.17** *A singular system*
The following is an example of a DCP for a singular linear system. We consider the finite-dimensional linear system

$$Ax = b,$$

where $A$ is a singular matrix; $A$ is approximated by a nonsingular matrix $\tilde{F}$ . We denote $\tilde{G} = (\tilde{F})^{-1}$ and consider the DCPB

$$l_0 = b; \quad l_{i+1} = (I - A\tilde{G})l_i + b; \quad x_i = \tilde{G}l_i$$

or, equivalently, the DCPA

$$\begin{cases} x_0 & = \tilde{G}b, \\ x_{i+1} & = (I - \tilde{G}A)x_i + \tilde{G}b, \end{cases}$$

where $\tilde{G} = \tilde{F}^{-1}$. The iterand $x_i$ can be written as

$$x_i = \sum_{j=0}^{i} (I - \tilde{G}A)^j \, \tilde{G} \, b.$$

Let us now take

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \text{ and } \tilde{F} = \begin{pmatrix} \epsilon & 0 \\ 1 & 1 \end{pmatrix},$$

for $\epsilon \neq 0$. We then have

$$\tilde{G} = \begin{pmatrix} 1/\epsilon & 0 \\ -1/\epsilon & 1 \end{pmatrix} \text{ and } I - \tilde{G}A = \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix},$$

and also

$$(I - \tilde{G}A)^j = \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix}, \quad j = 1, 2, \ldots.$$

Hence we see that

$$\begin{aligned} x_i &= \textstyle\sum_{j=0}^{i} (I - \tilde{G}A)^j \, \tilde{G} \, b \\ &= \textstyle\sum_{j=1}^{i} \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1/\epsilon & 0 \\ -1/\epsilon & 1 \end{pmatrix} b + \begin{pmatrix} 1/\epsilon & 0 \\ -1/\epsilon & 1 \end{pmatrix} b \\ &= \tfrac{i+1}{\epsilon} \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix} b + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} b \end{aligned}$$

Clearly, the sequence $\{x_i\}$ is not converging. We also see that the sequence $\{l_i\}$ in the DCPB will not vanish, because

$$(I - A\tilde{G}) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Now we take a slightly more general $A$ and a general $\tilde{G}$:

$$A = \begin{pmatrix} 0 & 0 \\ a & 1 \end{pmatrix} \text{ and } \tilde{G} = \begin{pmatrix} p & q \\ r & s \end{pmatrix},$$

the amplification operator for the error reads

$$M_A = I - \tilde{G}A = \begin{pmatrix} 1 - aq & -q \\ -as & 1 - s \end{pmatrix}.$$

This operator $M_A$ has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 1 - s - aq$. Because of the eigenvalue 1 in the amplification operator $M_A$, it is clear that no $\tilde{G}$ can be found such that the process converges. More generally, for arbitrary singular matrices $A$ and arbitrary matrices $\tilde{G}$ we know that $\|I - A\tilde{G}\| \geq 1$ and $\|I - \tilde{G}A\| \geq 1$, because $A$ must have at least one eigenvalue equal to zero, hence $I - \tilde{G}A$ has, necessarily, at least one eigenvalue equal to one. The same holds for $I - A\tilde{G}$.
The point of this example is clear: we try to solve a singular linear system; the result is that, quite rightly, both DCPA and DCPB refuse to converge.

**Example 1.18** *Non-existence of a fixed point* $\hat{l}$
This is an example on the non-existence of a fixed point $\hat{l}$ of DCPB, whereas a fixed point $\hat{x}$ of DCPA exists. Our target problem $Fx = y$ is to find the solution of the initial value problem

$$\begin{cases} x' + \lambda x = 0\,, & \text{on } [0.1], \\ x(0) = 1, & \lambda \neq -1\,. \end{cases} \tag{1.9}$$

The approximate problem $\tilde{F}x = y$ is to find a linear function $x$ on [0,1] such that

$$\begin{cases} x'(1) + \lambda x(1) = y(1)\,, \\ x(0) = 1\,; \end{cases} \tag{1.10}$$

(i.e. we try to find an approximate solution by one single backward Euler step). The sets and spaces we consider are:

$$\begin{aligned}
E &= C^1[0,1], \\
D &= C^1_B[0,1] = \{x \in E \,|\, x(0) = 1\}, \\
\hat{E} &= C[0,1], \\
\hat{D} &= \hat{E}, \\
\tilde{G}\hat{D} &= \{t \to 1 + Mt \,|\, M \in I\!R\}, \\
F\tilde{G}\hat{D} &= \{t \to M + \lambda + \lambda M t \,|\, M \in I\!R\}.
\end{aligned}$$

Note that $\tilde{G}$ is the solution-operator of (1.10), i.e. for any $y \in \hat{D}$ the operator $\tilde{G}$ returns a function $x : t \to 1 + Mt$, $M \in I\!R$, that satisfies (1.10) for the given $y \in \hat{D}$.

In the following we will use $\underline{0}$ to denote the function which yields zero for all arguments, i.e.

$$\underline{0} : t \to 0 \quad \text{for all } t \in [0,1].$$

First we apply DCPB with $l_0 = y = \underline{0}$, to obtain

$$l_1 = l_0 - F\tilde{G}l_0 + y = \underline{0} - F\tilde{G}\underline{0} + \underline{0} = -F\tilde{G}\underline{0}.$$

We remember that evaluating $\tilde{G}\underline{0}$ means solving (1.10) with right hand side 0. Thus we obtain

$$x_0(t) = (\tilde{G}\underline{0})(t) = 1 - \frac{\lambda t}{1+\lambda},$$

hence

$$l_1(t) = -(F\tilde{G}\underline{0})(t) = -((\tilde{G}\underline{0})' + \lambda(\tilde{G}\underline{0}))(t) = -\frac{\lambda^2}{1+\lambda}(1-t),$$

with corresponding

$$x_1(t) = (\tilde{G}l_1)(t) = 1 - \frac{\lambda t}{1+\lambda}.$$

By induction we easily show that, for $n = 1, 2, ...,$ the sequence $l_n(t)$ diverges

$$l_n(t) = \frac{\lambda^2}{1+\lambda} n(t-1), \quad l_n(1) = 0,$$

$$x_n(t) = (\tilde{G} l_n)(t) = 1 + \frac{l_n(1) - \lambda}{1+\lambda} \cdot t = 1 - \frac{\lambda t}{1+\lambda}.$$

Now we apply DCPA to obtain

$$x_0(t) = (\tilde{G} y)(t) = (\tilde{G}\underline{0})(t) = 1 - \frac{\lambda t}{1+\lambda}.$$

$$(F x_0)(t) = \frac{\lambda^2}{1+\lambda}(1-t),$$

$$(\tilde{G} F x_0)(t) = 1 - \frac{\lambda}{1+\lambda} t = x_0(t),$$

hence

$$x_1 = x_0 - \tilde{G} F x_0 + x_0 = x_0.$$

Again we see $x_n = x_0$ for $n = 0, 1, 2, ...$ .

Because $\tilde{G}$ is affine, we knew beforehand that the sequences $\{x_n\}$ for DCPA and DCPB are equal (Theorem 1.4). Clearly, $\tilde{G}$ is *not* injective in this example because the fixed point $\hat{x} = x_0$ of DCPA is *not* the solution of the original problem (Theorem 1.9) except for $\lambda = 0$. Furthermore, DCPB cannot have a fixed point $\hat{l}$, otherwise, according to theorem 1.12, $\tilde{G}\hat{l}$ would be a solution of the original problem (1.9). We saw that $x_n = x_0$, $n = 0, 1, 2, ...,$ and $\tilde{G}\hat{l} = x_0$. This $x_0$ is clearly not a solution of (1.9).

# Chapter 2

# Discretisation of continuous equations

## 2.1 Introduction

Defect correction in the general sense of Chapter 1 does not refer to special properties of the spaces $E$ and $\hat{E}$ for which the equation (1.1) is defined. They may be infinite dimensional function spaces, and the mappings from Chapter 1 may be analytic. For numerical computation, however, we must resort to *finite dimensions* so that the elements of the spaces can be identified with N-tuples of real numbers. If the original problem (1.1) is in an infinite dimensional setting, we have to model it in finite dimensions. This process of modeling into finite dimensions is called *discretisation* because discrete analogues of continuous functions and operators are introduced. In such a discretisation the (continuous) functions are generally represented by their values at a finite set of points: *"the grid"*. The power of the discretisation approach is that simultaneously an infinite set of grids can be considered, which contains increasingly finer grids, i.e. grids with increasingly more gridpoints. The corresponding finite dimensional spaces of gridfunctions have higher and higher dimensions, tending to infinity as the number of gridpoints increases.

## 2.2 Approximation of spaces

As in Chapter 1 we denote the continuous problem by

$$Fz = y \, , \tag{2.1}$$

where $F : D \subset E \to \hat{D} \subset \hat{E}$ is a non-linear operator and $E$ , $\hat{E}$ are infinite dimensional Banach spaces. The discretisation of the problem (2.1) is an associated problem

$$F_h z_h = y_h \, , \tag{2.2}$$

where $F_h : D_h \subset E_h \to \hat{D}_h \subset \hat{E}_h$ and $E_h, \hat{E}_h$ are finite dimensional Banach spaces with

$$N(h) := \dim (E_h) = \dim (\hat{E}_h).$$

By different $h \in \mathcal{H}$, $\mathcal{H}$ an index set, different discretisations of the same problem are indicated. The index set $\mathcal{H}$ is such that $\lim_{h\to 0}$ is defined and $\lim_{h\to 0} N(h) = \infty$ . In order to construct a relation between the problems (2.1) and (2.2), we first define relations between $E$ and $E_h$, $h \in \mathcal{H}$, and similarly between $\hat{E}$ and $\hat{E}_h$.

**Definition 2.1** An *approximation* (or *discretisation* ) *of the space* $E$ is a set of triples

$$(E_h, P_h, R_h)_{h\in\mathcal{H}},$$

in which
$P_h : E_h \to E$ is a linear injection, called *prolongation,*
$R_h : E \to E_h$ is a linear surjection, called *restriction.*

The relation between the different spaces and mappings is summarised in the following diagram:

$$\begin{tikzcd}
\mathbf{E} \arrow[r, "\mathbf{F}"] \arrow[d, shift left, "\mathbf{R_h}"] & \bar{\mathbf{E}} \arrow[d, shift left, "\bar{\mathbf{R}}_\mathbf{h}"] \\
\mathbf{E_h} \arrow[r, "\mathbf{F_h}"'] \arrow[u, shift left, "\mathbf{P_h}"] & \bar{\mathbf{E}}_\mathbf{h} \arrow[u, shift left, "\bar{\mathbf{P}}_\mathbf{h}"]
\end{tikzcd}$$

Figure 2.1: Spaces and operators in discretisation

Given discretisations of the spaces $E$ and $\hat{E}$ by $\{E_h, P_h, R_h\}_{h\in\mathcal{H}}$ and $\{\hat{E}_h, \bar{P}_h, \bar{R}_h\}_{h\in\mathcal{H}}$ respectively, a natural way to construct discrete operators for $F : E \to \hat{E}$ is to take

$$F_h = \bar{R}_h\, F\, P_h\,.$$

For the discrete right-hand side, $y_h$, we can take

$$y_h = \bar{R}_h\, y\,,$$

to obtain the discrete problem

$$\bar{R}_h\, F\, P_h\, z_h = \bar{R}_h\, y\,.$$

This discrete operator (or problem) is called the *Galerkin discretisation* of the continuous operator (problem).

As an example, the usual Finite Element discretisations [23] [12] are Galerkin discretisations. In the most straightforward finite element methods (FEMs), the solution $z^*$ of equation 2.1 is approximated by a piecewise polynomial function over a triangularisation (or division in quadrilaterals) of the domain of definition. These functions are represented by their values at a proper set of nodal points. $R_h$ is the restriction of continuous functions over $\Omega$ to their function values at these nodal points. $P_h$ is the interpolation from the nodal values to the piecewise polynomial functions. The restriction $\bar{R}_h$ is defined by weighting with a proper set of functions

$$(\bar{R}_h y)_i \;=\; \int_\Omega y(x)\,\phi_i(x)\,dx.$$

Usually, in the finite element method the set $\{\phi_i\}$ is a basis for the same set of piecewise polynomial functions as used for $P_h$. It is typical for FEMs that the support of a $\phi_i$ is only limited to a (very) small part of $\Omega$.

Thus far we considered $F : E \to \hat{E}$, where $E$ and $\hat{E}$ are simple Banach spaces. However, it often makes sense to view $F$ as an operator between a set of different (related) Banach spaces.

**Example 2.2** *Different Banach spaces*
Consider the problem

$$-\Delta u = f \;\text{ on }\; \Omega \subset I\!R^2\,,$$

$$u = 0 \text{ on } \partial\Omega\,,$$

where $\Omega \subset I\!R^2$ is a simply connected domain with a smooth boundary $\partial\Omega$. We may consider

$$-\Delta : C_0^2(\Omega) \to C^0(\Omega)\,,$$

as well as

$$-\Delta : H_0^1(\Omega) \cap H^2(\Omega) \to H^0(\Omega)\,,$$

or

$$-\Delta : H_0^1(\Omega) \to H^{-1}(\Omega)\,.$$

(For the definition of the spaces see the Appendix.)

In this way we may consider the general operator $F$ as a mapping defined between scales of Banach spaces

$$F : E^\alpha \to \hat{E}^\alpha, \quad \alpha_0 \le \alpha \le \alpha_1.$$

Two typical examples of scales of Banach spaces are $E^\alpha = C^\alpha$, the space of continuously differentiable functions, or the scale of Sobolev spaces $E^\alpha = H^\alpha(\Omega)$. For a scale of embedded Banach spaces $E^\alpha$, $\alpha_0 \le \alpha \le \alpha_1$ we consider its discretisation

$$(E_h^\alpha, P_h, R_h)_{h\in\mathcal{H}},$$

where the prolongations $P_h$ and restrictions $R_h$ are independent of $\alpha$ and the norms on $E_h^\alpha$ are chosen such that

$$\lim_{h\to 0}\|R_h z\|_{E_h^\alpha} = \|z\|_{E^\alpha}$$

for each $\alpha \in [\alpha_0, \alpha_1]$.

**Definition 2.3** An approximation $(E_h^\alpha, P_h, R_h)_{h\in\mathcal{H}}$ of $E^\alpha$, $\alpha_0 \le \alpha \le \alpha_1$, is *convergent of order* $p$ if

$$\|P_h R_h - I\|_{E^\alpha \leftarrow E^{\alpha+p}} = \mathcal{O}(h^p) \quad (\text{ for } h \to 0)$$

for $\alpha_0 \le \alpha \le \alpha_1$.

**Remarks:**

- If $E^\alpha$ is the scale of Hilbert spaces $H^\alpha(\Omega)$, and if $\|P_h R_h - I\|_{H^0 \leftarrow H^p} = \mathcal{O}(h^p)$ for some $p$, then this is true for all integers $\tilde{p}$, $0 \le \tilde{p} \le p$. For a proof of this statement see [7]. So, with $E^\alpha = H^\alpha$, an approximation $(E_h, P_h, R_h)_{h\in\mathcal{H}}$ of $E^\alpha$ is convergent of order $p$ if $p$ is a number for which

$$\|P_h R_h - I\|_{H^0 \leftarrow H^p} = \mathcal{O}(h^p) \quad (\text{ for } h \to 0) \tag{2.3}$$

- In other texts we can find the definition of order of convergence (and later also order of consistency) defined as the largest number $p^*$, such that the condition of definition 2.3 (or definition 2.18) is satisfied for all integer $p$ with $0 \le p \le p^*$. We will call this "the highest order".

- Of course, the value of $p$ in Definition 2.3 depends on the actual $\alpha$-indexing of the scale of Banach spaces $E^\alpha$. Usually we consider only the possibilities where $E^\alpha$ are continuous function spaces, with $E^\alpha = C^\alpha(\Omega)$, or where $E^\alpha$ are Sobolev spaces, with $E^\alpha = H^\alpha(\Omega)$. Hence, usually reference is made to values of $p$ corresponding with this indexing.

In the following $E^\alpha, \hat{E}^\alpha$ are again scales of Banach spaces and

$$L : E^\alpha \to \hat{E}^\alpha$$

is a continuous *linear* mapping

**Definition 2.4** A continuous linear operator $L$ is called $\sigma$-stable for some $\sigma \geq 0$ if

$$L^{-1} : \hat{E}^\alpha \to E^{\alpha-\sigma}$$

is bounded.

**Remarks:**

- The value of $\sigma$ by this definition depends, again, on the $\alpha$-indexing of the spaces $E^\alpha$ and $\hat{E}^\alpha$
- If $L$ is 0-stable then $L$ is a bijection since $\|L\|_{\hat{E}^\alpha \leftarrow E^\alpha} < C_1$ and $\|L^{-1}\|_{E^\alpha \leftarrow \hat{E}^\alpha} < C_2$. If $L$ is 0-stable then it is $\sigma$-stable for all $\sigma \geq 0$ since from $\|L^{-1}y\|_{E^{\alpha-\sigma}} \leq \|L^{-1}y\|_{E^\alpha}$ and from $\|L^{-1}y\|_{E^\alpha} \leq C\|y\|_{\hat{E}^\alpha}$ it follows that
$$\|L^{-1}y\|_{E^{\alpha-\sigma}} \leq C\|y\|_{\hat{E}^\alpha}.$$

**Example 2.5** *Bounded and unbounded operators*
If we take $E^\alpha = H^\alpha(\mathbb{R}^d)$ and $\hat{E}^\alpha = H^{\alpha-2}(\mathbb{R}^d)$, $\alpha \geq 2$, then $I : E^\alpha \to \hat{E}^\alpha$ is bounded and has an unbounded inverse. For, if $u \in H^\alpha(\mathbb{R}^d)$, then $\|u\|_{H^{\alpha-2}(\Omega)} \leq \|u\|_{H^\alpha(\Omega)}$ and there doesn't exist a constant $C > 0$ such that

$$\|u\|_{H^\alpha(\Omega)} \leq C\|u\|_{H^{\alpha-2}(\Omega)} \quad \text{for all } u \in H^\alpha(\Omega)\,;$$

$I : E^{\alpha-2} \to \hat{E}^\alpha$ is bounded and has a bounded inverse since in this case $\|u\|_{E^{\alpha-2}} = \|u\|_{\hat{E}^\alpha}$
We see that with this choice of $E^\alpha$ and $\hat{E}^\alpha$ the mapping $I$ is 2-stable.

**Definition 2.6** The discretisation $L_h$ of $L$ is called $\sigma$-stable for some $\sigma \geq 0$ if

$$L_h^{-1} : \hat{E}_h^\alpha \to {E_h}^{\alpha-\sigma}$$

is bounded uniformly in $h$, i.e.

$$\|L_h^{-1}\|_{{E_h}^{\alpha-\sigma} \leftarrow E_h^\alpha} \leq C$$

with $C$ independent of $h$.

We now wish to introduce stability for sets of prolongations and restrictions (these are also linear mappings). However, a difficulty arises because neither prolongations nor restrictions are bijections. For this reason we will work towards a somewhat modified definition of stability for these prolongations and restrictions.

**Definition 2.7**
i) A set of prolongations $\{P_h | \; P_h : E_h^\alpha \to E^\alpha,\; h \in \mathcal{H}\}$ is called *bounded for* $\alpha = \tilde{\alpha}$ if

$$\|P_h\|_{E^{\tilde{\alpha}} \leftarrow E_h^{\tilde{\alpha}}} \leq C$$

uniformly in $h$. It is called *bounded* (without reference to some $\tilde{\alpha} \in [\alpha_0, \alpha_1]$) if the inequality holds for all $\tilde{\alpha} \in [\alpha_0, \alpha_1]$.
ii) A set of restrictions $\{R_h \,|\, R_h : E^\alpha \to E_h^\alpha,\; h \in H\}$ is called *bounded for* $\alpha = \tilde{\alpha}$ if

$$\|R_h\|_{E^{\tilde{\alpha}} \leftarrow E_h^{\tilde{\alpha}}} \leq C$$

uniformly in $h$. It is called *bounded* (without reference to some $\tilde{\alpha} \in [\alpha_0, \alpha_1]$) if the inequality holds for all $\tilde{\alpha} \in [\alpha_0, \alpha_1]$.

**Definition 2.8**
i) Since $P_h : E_h^\alpha \to E^\alpha$ is an injection, a left inverse $\hat{R}_h : E^\alpha \to E_h^\alpha$ exists such that

$$\hat{R}_h P_h = I_h \, .$$

This $\hat{R}_h$ is called the *restriction related to* $P_h$
ii) Since $R_h : E_h^\alpha$ is a surjection, right inverses $\hat{P}_h : E_h^\alpha \to E^\alpha$ exist such that

$$R_h \hat{P}_h = I_h \, .$$

Such $\hat{P}_h$ are called *prolongations related to* $R_h$.
iii) A right inverse $\hat{P}_h$, related to $R_h$, is called *optimal prolongation related to* $R_h$ if the norm

$$\|I - \hat{P}_h R_h\|$$

is minimal.

We are now in the position to define stability of prolongations and restriction in an obvious way :

**Definition 2.9**
i) A set of prolongations $\{P_h \,|\, P_h : E_h^\alpha \to E^\alpha, \; h\in\mathcal{H}\}$ is called *stable for* $\alpha = \tilde{\alpha}$ if it is bounded for $\alpha = \tilde{\alpha}$ and the set of related restrictions $\{\hat{R}_h \,|\, \hat{R}_h : E^\alpha \to E_h^\alpha, \; \hat{R}_h P_h = I_h, \; h\in\mathcal{H}\}$ is bounded for $\alpha = \tilde{\alpha}$. It is called *stable* (without reference to some $\tilde{\alpha}\in[\alpha_0, \alpha_1]$) if the above is true for all $\tilde{\alpha}\in[\alpha_0, \alpha_1]$.
ii) A set of restrictions $\{R_h \,|\, R_h : E^\alpha \to E_h^\alpha, \; h\in\mathcal{H}\}$ is called *stable for* $\alpha = \tilde{\alpha}$ if it is bounded for $\alpha = \tilde{\alpha}$ and there exists a related set of prolongations $\{\hat{P}_h \,|\, \hat{P}_h : E_h^\alpha \to E^\alpha, \; R_h \hat{P}_h = I_h, \; h\in H\}$ which is bounded for $\alpha = \tilde{\alpha}$. It is called *stable* (without reference to some $\tilde{\alpha}\in[\alpha_0, \alpha_1]$) if the above holds for all $\tilde{\alpha}\in[\alpha_0, \alpha_1]$.

**Definition 2.10** An approximation $\{E_h^\alpha, P_h, R_h\}_{h\in\mathcal{H}}$ is called *bounded (stable) for* $\alpha = \tilde{\alpha}$ if the corresponding sets of prolongations and restrictions are *bounded (stable)* for $\alpha = \tilde{\alpha}$. It is called *bounded (stable)* (without reference to some $\tilde{\alpha}\in[\alpha_0, \alpha_1]$) if the above is true for all $\tilde{\alpha}\in[\alpha_0, \alpha_1]$.

**Definition 2.11**
i) A prolongation $P_h : E_h^\alpha \to E^\alpha$ is called *bounded from below* for $\alpha = \tilde{\alpha}$ if

$$\exists C > 0 \quad \inf_{x_h \neq 0} \frac{\|P_h x_h\|_{E^{\tilde{\alpha}}}}{\|x_h\|_{E_h^{\tilde{\alpha}}}} \geq C.$$

It is called *bounded from below* (without reference to some $\tilde{\alpha}\in[\alpha_0, \alpha_1]$) if the above is true for all $\tilde{\alpha}\in[\alpha_0, \alpha_1]$
ii) A restriction $R_h : E^\alpha \to E_h^\alpha$ is called *bounded from below for* $\alpha = \tilde{\alpha}$ if

$$\exists\, C > 0 \quad \inf_{x_h \in R_h E^{\tilde{\alpha}}, \, x_h \neq 0} \quad \sup_{x \in E^{\tilde{\alpha}}, \, R_h x = x_h} \frac{\|R_h x\|_{E_h^{\tilde{\alpha}}}}{\|x\|_{E^{\tilde{\alpha}}}} \geq C.$$

It is called *bounded from below* (without reference to some $\tilde{\alpha}\in[\alpha_0, \alpha_1]$) if the above is true for all $\tilde{\alpha}\in[\alpha_0, \alpha_1]$. With the aid of the following theorems the reader will become familiar with the definitions given above.

**Theorem 2.12** Let $\hat{R}_h$ be a restriction related to the prolongation $P_h : E_h^\alpha \to E^\alpha$ then
i) if $\hat{R}_h$ is bounded, $\|\hat{R}_h\| \neq 0$, then $P_h$ is bounded from below;
ii) if $P_h$ is bounded from below, then $\hat{R}_h : P_h E_h^\alpha \subset E^\alpha \to E_h^\alpha$ is bounded.
**Proof:**
i) By assumption we have $\hat{R}_h P_h = I_h$, $\|\hat{R}_h\| \leq C$, $C > 0$ so for $x_h \in E_h^\alpha$ $\|x_h\| = \|\hat{R}_h P_h x_h\| \leq \|\hat{R}_h\| \, \|P_h x_h\| \leq C \|P_h x_h\|$ and thus $\frac{\|P_h x_h\|}{\|x_h\|} \geq C$ if $x_h \neq 0$. Therefore $P_h$ is bounded from below.
ii) Since $P_h$ is bounded from below,

$$\begin{array}{l} \exists C > 0 \ \forall x_h \in E_h^\alpha \quad \|P_h x_h\| \geq C\|x_h\| = C\|\hat{R}_h P_h x_h\| \,, \\ \exists C > 0 \ \forall x_h \in E_h^\alpha \quad \frac{\|\hat{R}_h P_h x_h\|}{\|P_h x_h\|} \leq C^{-1} \,, \\ \exists C > 0 \quad \|\hat{R}_h\|_{E_h^\alpha \leftarrow P_h E_h^\alpha \subset E^\alpha} = \sup_{x \in P_h E_h^\alpha, x \neq 0} \frac{\|\hat{R}_h x\|}{\|x\|} \leq C^{-1} \,. \end{array}$$

■

**Theorem 2.13** Let $\hat{P}_h$ be a prolongation related to the restriction $R_h : E^\alpha \to E_h^\alpha$, then
i) if $\hat{P}_h$ is bounded then $R_h$ is bounded from below;
ii) if $R_h$ is bounded from below, Domain $(\hat{P}_h) =$ Range $(R_h)$ and $\hat{P}_h R_h$ is bounded then $\hat{P}_h$ is bounded.
**Proof:**
i) By assumption we know

$$\begin{array}{l} \exists C > 0 \ \sup_{x_h \in E_h^\alpha, x_h \neq 0} \frac{\|\hat{P}_h x_h\|}{\|x_h\|} = \|\hat{P}_h\| \leq C^{-1} \,, \\ \exists C > 0 \ \inf_{x_h \in E_h^\alpha, x_h \neq 0} \frac{\|R_h \hat{P}_h x_h\|}{\|\hat{P}_h x_h\|} = \inf_{x_h \in E_h^\alpha, x_h \neq 0} \frac{\|x_h\|}{\|\hat{P}_h x_h\|} \geq C \,, \\ \exists C > 0 \ \forall x_h \in E_h^\alpha, x_h \neq 0 \ \exists x \in E^\alpha \, (\text{namely } x = \hat{P}_h x_h) \ \frac{\|R_h x\|}{\|x\|} \geq C \,. \\ \exists C > 0 \ \inf_{x_h \in R_h E^\alpha, x_h \neq 0} \sup_{x \in E^\alpha, R_h x = x_h} \frac{\|R_h x\|}{\|x\|} \geq C \,. \end{array}$$

Hence $R_h$ is bounded from below.
ii) Since Domain $(\hat{P}_h) =$ Range $(R_h) = R_h E^\alpha$ we have

$$\|\hat{P}_h\| = \sup_{x_h \in \mathrm{Dom}\,(\hat{P}_h), x_h \neq 0} \frac{\|\hat{P}_h x_h\|}{\|x_h\|} = \sup_{x_h \in R_h E^\alpha, x_h \neq 0} \frac{\|\hat{P}_h x_h\|}{\|x_h\|} \,,$$

so

$$\|\hat{P}_h\|^{-1} = \inf_{x_h \in R_h E^\alpha, x_h \neq 0} \frac{\|x_h\|}{\|\hat{P}_h x_h\|} = \inf_{x_h \in R_h E^\alpha, x_h \neq 0} \sup_{x \in E^\alpha, x_h = R_h x} \frac{\|R_h x\|}{\|\hat{P}_h R_h x\|} \,.$$

Because of the boundedness of $\hat{P}_h R_h$, there exist some $C > 0$ such that $\|\hat{P}_h R_h x\| \leq C\|x\|$ for all $x \in E$, so we have

$$\|\hat{P}_h\|^{-1} \geq \inf_{x_h \in R_h E^\alpha, x_h \neq 0} \sup_{x \in E^\alpha, x_h = R_h x} C^{-1} \frac{\|R_h x\|}{\|x\|}.$$

By virtue of the fact that $R_h$ is bounded from below we conclude that $\|\hat{P}_h\|^{-1} \geq C$ for some $C > 0$, and thus $\hat{P}_h$ is bounded. ■

## 2.3 Discretisation errors

Discretisation methods are used to approximate the solution of problem (2.1) by computation of solutions of the related problems (2.2). The difference between the solution $z^*$ of (2.1) and the solution $z_h^*$ of (2.2) can be called the error caused by the discretisation. More precisely, this kind of error is defined as a global discretisation error. On the other hand, a measure of the fact that the solution $z_h^*$ of (2.2) only satisfies the original problem (2.1) approximately can also be called discretisation error. This kind of error is called the local discretisation error[1].

**Definition 2.14** Let $z^*$ be the solution of (2.1) and let $z_h^*$ be the solution of (2.2). We now define two kinds of *global discretisation error* of a discretisation:

$$e_h^1(z^*) = z_h^* - R_h z^*,$$

$$e_h^2(z^*) = P_h z_h^* - z^*.$$

Note that $e_h^1(z^*) \in E_h^\alpha$ and $e_h^2(z^*) \in E^\alpha$ .

**Definition 2.15** A sequence of discretisations (2.2) of the problem (2.1) is *discrete convergent of order p* , if

$$\|e_h^1(z^*)\|_{E_h^\alpha} = \mathcal{O}(h^p) \quad (\text{ for } h \to 0)$$

**Definition 2.16** A sequence of discretisation (2.2) of the problem (2.1) *is convergent of order p* , if

$$\|e_h^2(z^*)\|_{E^\alpha} = \mathcal{O}(h^p) \quad (\text{ for } h \to 0)$$

**Definition 2.17** With $z^*$ as above the *local discretisation* (or *truncation*) error of a discretisation is defined by

$$\tau_h(z^*) = F_h R_h z^* - \bar{R}_h F z^*$$

(See Figure 2.1 for a summary of the different operators).

**Definition 2.18** A sequence of discretisations (2.2) of the problem (2.1) is *consistent of order p,* if

$$\|\tau_h(z^*)\| = \mathcal{O}(h^p) \quad (\text{ for } h \to 0)$$

**Remark:** It is clear that the orders $p$, as defined in the definitions 2.15, 2.16 and 2.17 depend on the choice of the norms in $E_h^\alpha$, $E^\alpha$ or $\hat{E}_h^\alpha$ respectively.

We now give two theorems which reflect the relations between the different kinds of discretisation errors.

**Theorem 2.19** Let $(E_h, P_h, R_h)_{h\in\mathcal{H}}$ be a convergent approximation of the space $E$ of order $p$ with bounded prolongations. Suppose the solution $z^*$ of (2.1) satisfies $\|z^*\| \le C,$. Then a sequence of discretisations (2.2) has order of convergence $p$ if it has order of discrete convergence $p$.

**Proof:**

We have

$$\begin{aligned} e_h^2(z^*) = P_h z_h^* - z^* &= P_h z_h^* - P_h R_h z^* + P_h R_h z^* - z^* \\ &= P_h(z_h^* - R_h z^*) - (I - P_h R_h) z^* \\ &= P_h e_h^1(z^*) - (I - P_h R_h) z^* \,. \end{aligned}$$

By assumption $\|P_h| \le C_2$, $\|z^*\| \le C_1$, $\|e_h^1(z^*)\| = \mathcal{O}(h^p)$, and $\|I - P_h R_h\|_{E^\alpha \leftarrow E^{\alpha+p}} = \mathcal{O}(h^p)$ therefore $\|e_h^2(z^*)\| \le \|P_h\| . \|e_h^1(z^*)\| + \|z^*\| . \|I - P_h R_h\| = \mathcal{O}(h^p)$. ■

[1] There seems to be no clear explanation for the technical terms *local* and *global* associated with the meaning of these words in natural language.

**Theorem 2.20** Suppose the discrete operator $F_h$ from (2.2) is 0-stable and $y_h$ satisfies $y_h = \bar{R}_h y$. Then the discretisation (2.2) has order of discrete convergence at least $p$ if it has order of consistency $p$.

**Proof:** We have

$$\begin{aligned} e_h^1(z^*) = z_h^* - R_h z^* &= F_h^{-1} y_h - R_h z^* \\ &= F_h^{-1} \bar{R}_h y - R_h z^* \\ &= F_h^{-1} \bar{R}_h F z^* - R_h z^* \\ &= F_h^{-1} (\bar{R}_h F z^* - F_h R_h z^*) \\ &= -F_h^{-1} \tau_h(z^*). \end{aligned}$$

By assumption $\|F_h^{-1}\| \leq C$ and $\|\tau_h(z^*)\| = \mathcal{O}(h^p)$ and hence $\|e_h^1(z^*)\| \leq \|\tau_h^{-1}\|.\|F_h(z^*)\| = \mathcal{O}(h^p)$. ∎

If we consider discretisation (2.2) for which $y_h = \bar{R}_h y$, then we have

$$\begin{aligned} e_h^1(z^*) = z_h^* - R_h z^* &= F_h^{-1} y_h - R_h z^* \\ &= F_h^{-1} \bar{R}_h y - R_h z^* \\ &= (F_h^{-1} \bar{R}_h F - R_h) z^* \end{aligned}$$

and

$$\begin{aligned} e_h^2(z^*) &= P_h z_h^* - z^* \\ &= P_h F_h^{-1} y_h - F^{-1} y \\ &= (P_h F_h^{-1} \bar{R}_h - F^{-1}) y \end{aligned}$$

and as before

$$\tau_h(z^*) = (F_h R_h - \bar{R}_h F) z^*.$$

Notice that in the above definitions we considered the discretisation of a particular problem $Fz = y$, with solution $z^*$. We can also consider the discretisation of the operator $F : D \subset E \to D_h \subset E_h$ then the discretisation error are also operators:

$$\begin{aligned} e_h^1 :&\quad E^{\alpha+p} \to E_h^\alpha ; &\quad e_h^1(z) = (F_h^{-1} \bar{R}_h F - R_h)(z) , \\ \tilde{e}_h^2 :&\quad \hat{E}^{\alpha+p} \to E^\alpha ; &\quad \tilde{e}_h^2(y) = (P_h F_h^{-1} \bar{R}_h - F^{-1})(y) , \\ \tau_h :&\quad E^\alpha \to \hat{E}_h^{\alpha+p} ; &\quad \tau_h^1(z) = (F_h R_h - \bar{R}_h F)(z) . \end{aligned}$$

Generally these mapping are not defined on the entire spaces $E^\alpha$ and $\hat{E}^\alpha$ but on subsets $D \subset E^\alpha$ and $\hat{D} \subset \hat{E}^\alpha$ which depend on the class of continuous problems under consideration. For the definition of $p$, the order relation should hold uniformly for all $z^*$ in some subset $D \subset E$ of suitable functions. Usually $D$ is a subset containing sufficiently smooth functions. Based on the above, we define the *order of discrete convergence, convergence* and *consistency of a sequence of discretisations of an operator* $F$:

**Definition 2.21** The *order of discrete convergence* of a sequence of discretisations of an *operator* $F : E^{\alpha+p} \to \hat{E}^\alpha$ is $p$ if the following estimate holds:

$$\|F_h^{-1} \bar{R}_h F - R_h\|_{E_h^\alpha \leftarrow D \subset E^{\alpha+p}} = \mathcal{O}(h^p) \quad (\text{ for } h \to 0).$$

**Definition 2.22** A sequence of discretisations of an operator $F : E^{\alpha+p} \to \hat{E}^\alpha$ is *convergent of order p,* if

$$\|P_h F_h^{-1} \bar{R}_h - F^{-1}\|_{E^\alpha \leftarrow \hat{D} \subset \hat{E}^{\alpha+p}} = \mathcal{O}(h^p) \quad (\text{ for } h \to 0).$$

**Definition 2.23** A sequence of discretisations of an operator $F : E^{\alpha+p} \to \hat{E}^\alpha$ is *consistent of order p,* if

$$\|F_h R_h - \bar{R}_h F\|_{\hat{E}_h^\alpha \leftarrow D \subset E^{\alpha+p}} = \mathcal{O}(h^p) \quad (\text{ for } h \to 0).$$

**Definition 2.24**
i) Suppose the discretisation (2.2) is consistent of order $p$ then the local discretisation error admits an *asymptotic expansion in* $h$ if

$$\tau_h(z^*) = h^p \bar{R}_h D(z^*)$$

where $D(z^*) = D_0(z^*) + hD_1(z^*) + ... + h^j D_j(z^*) + \mathcal{O}(h^{j+1})$ and $D_i : E^\alpha \to \hat{E}^\alpha$ independent of $h$, $i = 1, 2, ..., j$.
ii) Suppose the discretisation (2.2) has order of discrete convergence $p$. Then the global discretisation error $e_h^1(z^*)$ admits an *asymptotic expansion in* $h$ if

$$e_h^1(z^*) = h^p R_h G(z^*)$$

where $G(z^*) = G_0(z^*) + hG_1(z^*) + ... + h^j G_j(z^*) + \mathcal{O}(h^{j+1})$ and $G_i : E^\alpha \to E^\alpha$ independent of $h$, $i = 1, 2, ..., j$ .

**Example 2.25** *Local and global error estimates*
Consider the initial value problem

$$\begin{cases} x'(t) - p(t)x(t) - q(t) = 0, & t \in [t_0, t_0 + T] \\ x(t_0) = x_0 \end{cases} \tag{2.4}$$

Suppose $p, q \in C^1[t_0, t_0 + T]$, so that the solution $x^* \in C^2[t_0, t_0 + T]$. We wish to approximate $x^*$ with the Euler-Forward method. We define an equidistant grid $t_0, t_1, ...t_{N+1}$ on $[t_0, t_0 + T]$ such that $t_i = t_0 + ih$, where $h = \frac{T}{N+1}$. The discrete problem now is given by the recursion

$$\begin{cases} x_h(t_{i+1}) - x_h(t_i) - h\,(p(t_i)x_h(t_i) + q(t_i)) = 0, & i = 0, 1, ..., N-1, \\ x_h(t_0) = x_0. \end{cases} \tag{2.5}$$

To obtain an expression for the local discretisation error we examine how well $x^*$ satisfies the discrete problem:

$$\begin{array}{l} x^*(t_{i+1}) - x^*(t_i) - h\,(p(t_i)x^*(t_i) + q(t_i)) = \\ x^*(t_i) + hx^{*\prime}(t_i) + \frac{1}{2}h^2 x^{\star\prime\prime}(\xi_i) - x^*(t_i) - h(p(t_i)x^*(t_i) + q(t_i)). \end{array}$$

Note that $x^*$ is the solution of the continuous problem and therefore $x^{*'}(t_i) = p(t_i)x^*(t_i) + q(t_i)$, so $x^*(t_{i+1}) - x^*(t_i) - h(p(t_i)\,x^*(t_i) + q(t_i)) = \frac{1}{2}h^2 x^{\star\prime\prime}(\xi_i)$, $\xi_i \in [t_i, t_{i+1}]$. As $x^* \in C^2[t_0, t_0 + T]$ there exists $C > 0$ such that $\|x^{\star\prime\prime}\|_\infty \leq C$ and it follows that the discretisation (2.5) is consistent of order 2.
Furthermore, if $x^* \in C^k[t_0, t_0 + T]$ for some $k \geq 3$, then the local discretisation error admits an asymptotic expansion in $h$ because we then have

$$x^*(t_{i+1}) = x^*(t_i) + hx^*(t_i) + ... + \frac{h^{k-1}}{(k-1)!} x^{*\,(k-1)}(t_i) + \mathcal{O}(h^{k-1}),$$

so

$$\begin{array}{l} x^*(t_{i+1}) - x^*(t_i) - h\,(p(t_i)x^*(t_i) + q(t_i)) = \\ \frac{1}{2}h^2 x^{\star\prime\prime}(t_i) + ... + \frac{h^{k-1}}{(k-1)!} x^{\star(k-1)}(t_i) + \mathcal{O}(h^{k-1}). \end{array}$$

We now return to the case that $x^* \in C^2[t_0, t_0 + T]$ and we consider the global discretisation error $e_h^1(x^*)$. If we define $e_i = [e_h^1(x^*)]_i$ , $i = 0, 1, ..., N$, then $e_0 = 0$ because $x_h(t_0) = x^*(t_0)$. Furthermore, if we define $\tau_i = [\tau_h(x^*)]_i$, then

$$x_h^*(t_{i+1}) - x_h^*(t_i) - h(p(t_i)x_h^*(t_i) + q(t_i)) = 0$$

(cf. (2.4)) and

$$x^*(t_{i+1}) - x^*(t_i) - h(p(t_i)\, x^*(t_i) + q(t_i)) = \tau_{i+1}.$$

Subtraction of these equations gives

$$e_{i+1} = (1 + hp_i)e_i + \tau_{i+1}, \quad \text{where } \ p_i = p(x_i).$$

Take $\sigma_i = 1 + hp_i$ then, for $i = 1, 2, ..., N-1$,

$$\begin{aligned} e_i \ &= \sigma_{i-1} e_{i-1} + \tau_i \\ &= \sigma_{i-1}(\sigma_{i-2} e_{i-2} + \tau_{i-1}) + \tau_i \\ &= \sigma_{i-1}(\sigma_{i-2}(\sigma_{i-3} e_{i-3} + \tau_{i-2}) + \tau_{i-1}) + \tau_i \end{aligned}$$

etc.. For constant $\sigma$ we obtain

$$\begin{aligned} e_i \ &= \sigma^i e_0 + \sigma^{i-1}\tau_1 + \sigma^{i-2}\tau_2 + ... + \sigma\tau_{i-1} + \tau_i = \\ &= \textstyle\sum_{j=1}^{i} \sigma^{i-j}\tau_j . \end{aligned} \tag{2.6}$$

Now $\sigma^k = (1 + hp)^k = e^{k \log(1+hp)} = e^{k(ph+\mathcal{O}(h^2))} = e^{kh(p+\mathcal{O}(h))}$.
Because $kh \le T$, $k = 0, 1, ..., N-1$, there exists some constant $M > 0$ such that

$$\sigma^k \ \le M, \qquad k = 0, 1, ..., N-1.$$

So

$$\begin{aligned} &|e_i| \ \le \textstyle\sum_{j=1}^{i} |\sigma^{i-j}\tau_j| \ \le \ M \sum_{j=1}^{i} |\tau_j| \\ &= \ M \textstyle\sum_{j=1}^{i} \frac{1}{2} h^2 |x^{\star\prime\prime}(\xi_j)| \ \le \ \frac{1}{2} M h^2 i \|x^{\star\prime\prime}\|_\infty \qquad (hi \le T) \\ &\le \ \tfrac{1}{2} T M h \|x^{\star\prime\prime}\|_\infty \,, \end{aligned}$$

and we have shown that the discretisation (2.5) has order of discrete convergence 1.

For showing that the global discretisation error admits an asymptotic expansion in $h$ if $x^* \in C^k[t_0, t_0+T]$, $k \ge 3$, we can substitute the asymptotic expansion for the local discretisation error for $\tau_j$ in the formula (2.6).

Estimates of the local discretisation error are often used to control and design grid adaptation algorithms. Estimates of the global discretisation error $e_h^1(x^*)$ assess the accuracy of the computed approximation.

## 2.4 Nested discretisations

In many respects it is not essential that the problem to be discretised is continuous. We can also “discretise” a finite dimensional problem (2.2) further to get a discretisation

$$F_H z_H = y_H, \quad H \in \mathcal{H}, \tag{2.7}$$

where

$$F_H : D_H \subset E_H \ \to \ \hat{D}_H \subset \hat{E}_H .$$

For the dimension $N(H)$ of $E_H$ and $\hat{E}_H$ we have

$$N(H) \ < \ N(h).$$

It is clear that if (2.2) is a discretisation of (2.1) and (2.7) a discretisation of (2.2), then (2.7) is also a discretisation of (2.1). As (2.7) is a discretisation of (2.2), there exist prolongations

$$P_{Hh} : \, E_H \to E_h \,,$$

$$\bar{P}_{Hh} : \hat{E}_H \to \hat{E}_h ,$$

and restrictions

$$R_{hH} : E_h \to E_H ,$$
$$\bar{R}_{hH} : \hat{E}_h \to \hat{E}_H .$$

so we can construct the following diagram. The operators $P_h$ and $P_{hH}$ are linear injections

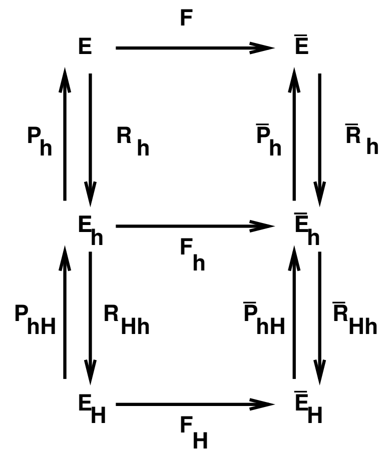


Figure 2.2: Nested discretisations

and $R_h$ and $R_{Hh}$ are linear surjections. Therefore, also $P_hP_{hH} : E_H \to E$ is a linear injection and $R_{Hh}R_h : E \to E_H$ a linear surjection.

**Definition 2.26** If $P_hP_{hH} = P_H$ and $R_{Hh}R_h = R_H$, and if $E_H$,$P_H$ and $R_H$ belong to $\{E_h, P_h, R_h\}_{h\in\mathcal{H}}$, we call $(E_h, P_h, R_h)$ and $(E_H, P_H, R_H)$ *nested discretisations* of the space $E$. To distinguish between two such related discretisations we call $(E_h, P_h, R_h)$ the *finer* and $(E_H, P_H, R_H)$ the *coarser* discretisation of the space $E$.

In this way a subset of $\{E_h, P_h, R_h\}_{h\in\mathcal{H}}$ may form a *nested sequence of discretisations* of $E$.

Discretisation properties between two nested discretisations are called the relative discretisation properties. We now will define such relative properties between discretisations $(E_h, P_h, R_h)$ and $(E_H, P_H, R_H)$.

**Definition 2.27** The spaces $E_h^\alpha$ and $E_H^\alpha$, $\alpha_0 \le \alpha \le \alpha$, have *relative order of approximation* $p$ if

$$\|P_{hH}R_{Hh} - I_h\|_{E_h^\alpha \leftarrow E_h{}^{\alpha+p}} = \mathcal{O}(H^p) \quad (\text{ for } h \to 0).$$

**Definition 2.28** The *order of relative discrete convergence* of two of discretisations of an *operator* $F : E_h^{\alpha+p} \to \hat{E}_h^\alpha$ is $p$ if the following estimate holds:

$$\|F_H^{-1}\bar{R}_{Hh}F_h - R_{Hh}\|_{E_h^\alpha \leftarrow D\subset E_h^{\alpha+p}} = \mathcal{O}(H^p) \;\; (\text{ for } H \to 0).$$

**Definition 2.29** Two discretisations of an operator $F : E^{\alpha+p} \to \hat{E}^\alpha$ are *relative convergent of order* $p$, if

$$\|P_{hH}F_H^{-1}\bar{R}_{Hh} - F_h^{-1}\|_{E_h^\alpha \leftarrow \hat{D}\subset \hat{E}_h^{\alpha+p}} = \mathcal{O}(H^p) \;\; (\text{ for } H \to 0).$$

**Definition 2.30** The *relative local discretisation (or truncation) error of two nested discretisations* is defined by

$$\tau_{Hh}(z_h^*) = F_H R_{Hh} z_h^* - \bar{R}_{Hh} F_h z_h^*.$$

**Definition 2.31** The *relative order of consistency of two nested discretisations of a problem* is a value of $p$ for which

$$\|\tau_{Hh}(z_h^*)\| = \mathcal{O}(H^p) \quad (\text{ for } h \to 0).$$

**Definition 2.32** The *relative order of consistency of two operators* $F_h \, and \, F_H$ from nested discretisations is a value of $p$ for which

$$\|F_H R_{Hh} - \bar{R}_{Hh} F_h\|_{\hat{E}^{\alpha}_H \leftarrow A \subset E_h{}^{\alpha+p}} = \mathcal{O}(H^p) \quad (\text{ for } h \to 0).$$

**Theorem 2.33** Let two operators from a sequence of nested discretisations be consistent of orders $p_1$ and $p_2$ respectively. Let the restriction $R_h$ be bounded from below and let the restriction $\bar{R}_{Hh}$ be bounded. Then the two operators are relatively consistent of order $(\min(p_1, p_2)$.
**Proof:** $R_h : E \to E_h$ is bounded from below, so there exists a constant $C_1 > 0$ such that

$$\forall x_h \in R_h E, x_h \neq 0 \;\; \exists x \in E, \; R_h x = x_h, \; \frac{\|R_h x\|}{\|x\|_E} \geq C_1.$$

If we let $T = F_H R_{Hh} - \bar{R}_{Hh} F_h : E_h \to \hat{E}_H$, then

$$\begin{aligned} \|TR_h\|_{\hat{E}_H \leftarrow E} &= \sup_{x \in E} \frac{\|TR_h x\|_{\hat{E}}H}{\|x\|_E} = \\ &= \sup_{x \in E} \frac{\|TR_h x\|_{\hat{E}_H}}{\|R_h x\|_{E_h}} \frac{\|R_h x\|_{E_h}}{\|x\|_E}, \end{aligned}$$

with $E_x = \{\tilde{x} \in E \,|\, R_h \tilde{x} = R_h x\}$ it follows that

$$\begin{aligned} \|TR_h\|_{\hat{E}_H \leftarrow E} &= \sup_{x \in E} \quad \sup_{\tilde{x} \in E_x} \frac{\|TR_h \tilde{x}\|_{\hat{E}_H}}{\|R_h \tilde{x}\|_{E^h}} \frac{\|R_h \tilde{x}\|_{E^h}}{\|\tilde{x}\|_E} \geq \\ \geq \quad &\sup_{x \in E} \quad \sup_{\tilde{x} \in E_x, \|R_h \tilde{x}\|_{E_h} \geq C_1 \|\tilde{x}\|_E} \frac{\|TR_h \tilde{x}\|_{\hat{E}_H}}{\|R_h \tilde{x}\|_{E_H}} \frac{\|R_h \tilde{x}\|_{E_h}}{\|\tilde{x}\|_E} C. \end{aligned}$$

The last inequality holds because "$R_h$ bounded from below" implies

$$E_x \supseteq \{\tilde{x} \in E_x \,|\, \|R_h \tilde{x}\|_{E_h} \geq C_1 \|\tilde{x}\|_E\} \neq \emptyset.$$

We now see that

$$\begin{aligned} \|TR_h\|_{\hat{E}_H \leftarrow E} &\geq C_1 \sup_{x \in E} \sup_{\tilde{x} \in E_x} \frac{\|TR_h \tilde{x}\|_{\hat{E}_H}}{\|R_h \tilde{x}\|_{E_h}} \\ &= \sup_{x_h \in E_h} C_1 \frac{\|Tx_h\|_{\hat{E}_H}}{\|x_h\|_{E}h} = C_1 \|T\|_{\hat{E}_H \leftarrow E_h}. \end{aligned}$$

Using this we find

$$\begin{aligned} &\|F_H R_{Hh} - \bar{R}_{Hh} F_h\|_{\hat{E}_H \leftarrow Eh} \leq C_1{}^{-1} \|F_H R_{Hh} R_h - \bar{R}_{Hh} F_h R_h\|_{\hat{E}_H \leftarrow E} = \\ &C_1{}^{-1} \|F_H R_H - \bar{R}_H F + \bar{R}_{Hh} \bar{R}_h F - R_{Hh} F_h R_h\|_{\hat{E}_H \leftarrow E} \leq \\ &C_1{}^{-1} \|F_H R_H - \bar{R}_H F\|_{\hat{E}_H \leftarrow E} + C_1{}^{-1} \|\bar{R}_{Hh}\|_{\hat{E}_H \leftarrow \hat{E}_h} \|\bar{R}_h F - F_h R_h\|_{\hat{E}_h \leftarrow E}. \end{aligned}$$

By assumption $\|\bar{R}_{Hh}\|_{\hat{E}_H \leftarrow E_h} = C_2$, $\|F_H R_H - \bar{R}_H F\|_{\hat{E}_H \leftarrow E} = \mathcal{O}(H^{p_1})$, (for $H \to 0$), and $\|F_h R_h - \bar{R}_h F\|_{\hat{E}_h \leftarrow E} = \mathcal{O}(h^{p_2})$, for $h \to 0$, so we conclude

$$\|F_H R_{Hh} - \bar{R}_{Hh} F_h\|_{\hat{E}_H \leftarrow E_h} = \mathcal{O}(H^{\min(p_1, p_2)}), \qquad (\text{ for } H \to 0)$$

■

## 2.5 Galerkin discretisation

Given discretisations $\{E_h^\alpha, P_h, R_h\}$ of $E^\alpha$ and $\{\hat{E}_h^\alpha, \bar{P}_h, \bar{R}_h\}$ of $\hat{E}^\alpha$, a natural way to construct discrete operators for $F : E^\alpha \to \hat{E}^\alpha$ is to take

$$F_h = \bar{R}_h F P_h : E_h^\alpha \to \hat{E}_h^\alpha. \tag{2.8}$$

For the discrete right hand side $y_h$ we can take

$$y_h = \bar{R}_h y_h \tag{2.9}$$

thus, we obtain the discrete problem

$$\bar{R}_h F P_h z_h = \bar{R}_h y. \tag{2.10}$$

The discrete operator (2.8) (or problem (2.10)) is called the *Galerkin discretisation* of the continuous operator $F$ (problem (2.1)).

**Example 2.34** *Finite elements. See also section (2.6)*
The usual finite element discretisations are Galerkin discretisations. For the simplest finite element methods the solution $z^*$ of the problem (2.1) is approximated by a piecewise polynomial function over a division of the domain $\Omega$ (usually a division in triangles or quadrangles). These functions are determined, and may thus be represented, by their values at a proper set of nodal points. $R_h$ is the restriction of continuous functions to their values at the nodal points. $P_h$ is the interpolation of these nodal values to the piecewise polynomial function. The restriction $\bar{R}_h$ is defined by weighting with a proper set of functions:

$$(\bar{R}_h y)_i = \int_\Omega y \cdot \phi_i \, d\Omega \,. \tag{2.11}$$

where usually we take for $\{\phi_i \,|\, i = 1, 2, ..., N(h)\}$ the same basis of piecewise polynomial functions as is taken for $P_h$. In such a finite element method a proper set of piecewise polynomial functions and a proper set of nodal points in selected, such that the approximation $\{E_h^\alpha, P_h, R_h\}_{h\in\mathcal{H}}$ of the set $E^\alpha$ has a sufficiently high order. Furthermore, the set of prolongations $\{P_h \,|\, P_h : {E_h}^\alpha \to E^\alpha, h\in\mathcal{H}\}$ should be stable.

**Theorem 2.35** Let $\{E_h^\alpha, P_h, R_h\}_{h\in\mathcal{H}}$ be a convergent approximation of the set $E^\alpha$ of order $p$. Let $\{\hat{E}^\alpha, \bar{P}_h, \bar{R}_h\}$ be an approximation of the set $\hat{E}^\alpha$ with bounded restrictions. Suppose $F : E^\alpha \to \hat{E}^\alpha$ is bounded. Then a sequence of Galerkin discretisations of the operator $F$ is consistent of order $p$.
**Proof:**

$$\begin{aligned}
&\|F_h R_h - \bar{R}_h F\|_{E_h^\alpha \leftarrow A\subset E^{\alpha+p}} = \\
&\|\bar{R}_h F P_h R_h - \bar{R}_h F\|_{\hat{E}_h^\alpha \leftarrow A\subset E^{\alpha+p}} \le \\
&\|\bar{R}_H\|_{\hat{E}_h^\alpha \leftarrow \hat{E}^\alpha} \, \|F\|_{\hat{E}^\alpha \leftarrow E^\alpha} \, \|P_h R_h - I\|_{E^\alpha \leftarrow E^{\alpha+p}} = \mathcal{O}(h^p), \quad (\text{ for } h \to 0).
\end{aligned}$$

■

Analogous to the Galerkin discretisation of a continuous operator (problem), we may consider the Galerkin discretisation of a discrete operator (problem) as follows

$$F_H = R_{Hh} F_h P_{hH} \tag{2.12}$$

and

$$R_{Hh} F_h P_{hH} z_H = R_{Hh} y_h. \tag{2.13}$$

In the case of nested discretisations for the spaces $E^\alpha$ *and* $\hat{E}^\alpha$, we obtain a *nested sequence of Galerkin discretisations* with the properties

$$F_H = R_{Hh} F_h P_{hH} = R_{Hh} R_h F P_h P_{hH} = R_H F P_H$$

and

$$y_H = \bar{R}_{Hh} y_h = \bar{R}_{Hh} \bar{R}_h y = \bar{R}_H y$$

## 2.6 A more elaborate example

Let $E = H^2(\Omega)$, $\hat{E} = H^0(\Omega)$ for some rectangle $\Omega$. Let the problem $Fu = y$ be given by

$$\begin{cases} -\Delta u = f_\Omega & \text{on} \quad \Omega\,, \\ u + \epsilon \frac{\partial u}{\partial n} = f_\Gamma & \text{on} \quad \Gamma = \partial\Omega\,. \end{cases} \tag{2.14}$$

For an arbitrary $v \in H^1(\Omega)$ we have by partial integration

$$\int_\Omega f_\Omega \, v \; d\Omega = \int_\Omega -\Delta u \, v \; d\Omega =$$

$$= \int_\Omega \nabla u \cdot \nabla v \; d\Omega - \int_\Gamma \frac{\partial u}{\partial n} v \; d\Gamma =$$

$$= \int_\Omega \nabla u \cdot \nabla v \; d\Omega - \int_\Gamma (\frac{1}{\epsilon} f_\Gamma - \frac{1}{\epsilon} u) \, v \; d\Gamma$$

Instead of solving the original problem (2.14), we solve the problem: find $u \in H^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v \; d\Omega + \frac{1}{\epsilon} \int_\Gamma u \, v \; d\Gamma = \tag{2.15}$$

$$\int_\Omega f_\Omega \, v \; d\Omega + \frac{1}{\epsilon} \int_\Gamma f_\Gamma \, v \; d\Gamma \qquad \text{for all } v \in H^1(\Omega).$$

Since $v \in H^1(\Omega)$, the bounded linear functional $f$, defined by $f(v) = \int_\Omega f_\Omega v \; d\Omega + \frac{1}{\epsilon} \int_\Gamma f_\Gamma v \; d\Gamma$ can be seen as an element of $H^{-1}(\Omega)$. The same holds for the bounded linear functional $B(u, \cdot)$, defined by

$$B(u, v) = \int_\Omega \nabla u \cdot \nabla v \; d\Omega + \frac{1}{\epsilon} \int_\Gamma u \, v \; d\Gamma.$$

The reformulated equation $B(u, v) = f(v)$ is defined for all $u \in H^1(\Omega)$ and the problem: *find* $u \in H^1(\Omega)$ *such that* $B(u, \cdot) = f(\cdot)$ *in* $H^{-1}(\Omega)$ can be considered as a generalisation of the original problem: *find* $u \in H^2(\Omega)$ *such that (2.14) holds.*

By $F(u) = B(u, \cdot)$ in $H^{-1}(\Omega)$ we introduce the operator $F : H^1(\Omega) \to H^{-1}(\Omega)$, which is a generalisation of the operator $F : H^2(\Omega) \to H^0(\Omega)$, defined by (2.14). We also may consider $F$ as an operator between the spaces $H^{1+\alpha}(\Omega)$ and $H^{-1+\alpha}(\Omega), \alpha \geq 0$ (scale of Hilbert spaces).

Next we want to discretise the problem given by (2.15). We divide the region $\Omega$ into triangles as is demonstrated in Figure 2.3. The nodes $N_{ij}^h$ of the triangulation are the vertices of the triangles. With each node $N_{ij}^h$ we associate the piecewise linear function $\phi_{ij}^h$ defined by

$$\phi_{ij}^h(N_{kl}) = \begin{cases} 1 & \text{if } (i,j) = (k,l) \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.16}$$

See Figure 2.4. The function $u$ will be approximated by

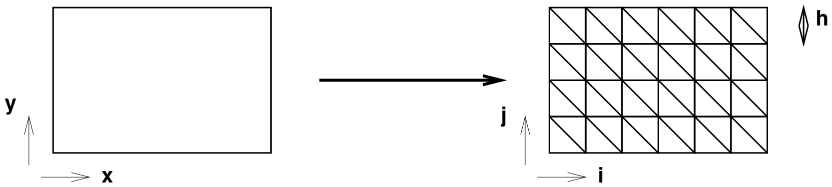


Figure 2.3: A triangularisation of a domain

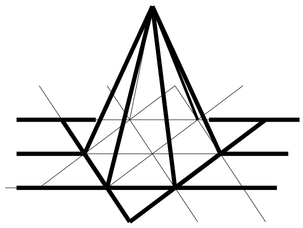

Figure 2.4: A piecewise linear basis function

$$u_h = \sum_{i,j} a_{ij}^h \, \phi_{ij}^h .$$

Notice that $u_h(N_{ij}^h) = a_{ij}^h$, where the constants $a_{ij}^h$ are determined by

$$\int_\Omega \nabla u_h \cdot \nabla v \; d\Omega + \frac{1}{\epsilon} \int_\Gamma u_h \, v \; d\Gamma = \int_\Omega f_\Omega v \, d\Omega + \frac{1}{\epsilon} \int_\Gamma f_\Gamma \, v \; d\Gamma \qquad \text{for all } v \in \text{Span}\,\{\phi_{kl}^h\}, \tag{2.17}$$

or

$$\int_\Omega \nabla u_h \cdot \nabla \phi_{kl}^h \; d\Omega + \frac{1}{\epsilon} \int_\Gamma u_h \phi_{kl}^h \; d\Gamma$$
$$= \int_\Omega f_\Omega \phi_{kl}^h d\Omega + \frac{1}{\epsilon} \int_\Gamma f_\Gamma \phi_{kl}^h \; d\Gamma, \qquad \text{for all } k,l,$$

or

$$\sum_{i,j} a_{ij}^h \{ \int_\Omega \nabla \phi_{ij}^h \cdot \nabla \phi_{kl}^h \; d\Omega + \frac{1}{\epsilon} \int_\Gamma \phi_{ij}^h \, \phi_{kl}^h \; d\Gamma \, \} \; = \tag{2.18}$$
$$= \int_\Omega f_\Omega \phi_{kl}^h \; d\Omega + \frac{1}{\epsilon} \int_\Gamma f_\Gamma \phi_{kl}^h \; d\Gamma \qquad \text{for all } \; k,l.$$

We define $E_h$ as the set of function values at the nodes $N_{ij}^h$, i.e. $E_h = I\!R^{N(h)}$.
$P_h : E_h \to E$ is the piecewise linear interpolation over the triangles,
$R_h : E \to E_h$ is the restriction of a function to its nodal function values, and
$\bar{R}_h : \hat{E} \to \hat{E}_h = I\!R^{N(h)}$ is defined by

$$(\bar{R}_h y)_{ij} = \int_\Omega y_\Omega \, \phi_{ij}^h \, d\Omega \; + \; \frac{1}{\epsilon} \int_\Gamma y_\Gamma \, \phi_{ij}^h \, d\Gamma \, ,$$

and $a_h \in I\!R^{N(h)}$ is the vector with coordinates $a_{ij}^h$. Now we may write (2.18) as the linear system

$$F_h a_h = y_h, \tag{2.19}$$

where

$$F_h = \bar{R}_h F P_h, \tag{2.20}$$

and

$$y_h = \bar{R}_h f. \tag{2.21}$$

We also can divide the region $\Omega$ into larger triangles to obtain *a coarser discretisation*

$$F_H a_H = y_H$$

If we take a division as shown in Figure 2.5, and the corresponding piecewise linear functions $\phi_{ij}^H$ are defined in the same way as we previously defined the functions $\phi_{ij}^h$, then

$$\text{Span}\,\{\phi_{ij}^H\} \subset \text{Span}\,\{\phi_{ij}^h\}$$

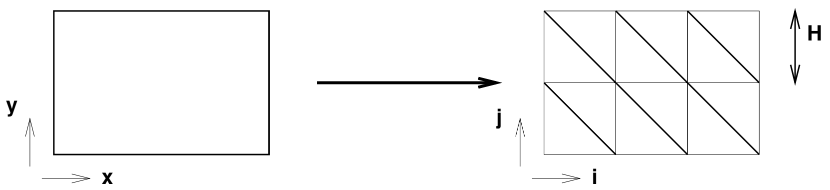


Figure 2.5: A coarser triangulation

The space $E_H$ and the operators $P_H$,$R_H$ and $\bar{R}_H$ are defined completely analogous to the space $E_h$ and the operators $P_h, R_h$ and $\bar{R}_h$. We now want to examine the operators $R_{Hh}$,$P_{Hh}$ and $\bar{R}_{Hh}$. We notice that we have constructed the coarse discretisation such that one coarse triangle can be seen as a union of triangles from the finer discretisation.
In Figure 2.6 a part of the fine and a part of the coarse division of $\Omega$ into triangles is shown in one picture. The thin lines indicate edges that belong to the fine division but don't belong to the coarse. The thick lines indicate edges that belong to both divisions. It is clear that

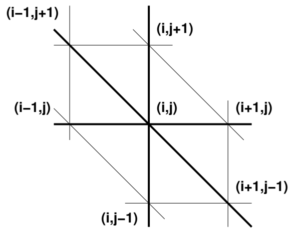


Figure 2.6: Neighbouring nodal points in a triangulation

the operator $R_{Hh}$ converts an element of $E_h(= I\!R^{N(h)})$ into an element of $E_H(= I\!R^{N(H)})$ by taking the function values at the nodes of the coarse triangulation. This is denoted as

$$R_{Hh} \approx \begin{bmatrix} 0 & 0 & \\ 0 & 1 & 0 \\ & 0 & 0 \end{bmatrix}_h .$$

Furthermore we have

$$\phi_{i,j}^H = \phi_{i,j}^h + 1/2(\phi_{i,j-1}^h + \phi_{i+1,j-1}^h + \phi_{i-1,j}^h + \phi_{i+1,j}^h + \phi_{i-1,j+1}^h + \phi_{i,j+1}^h), \tag{2.22}$$

so we have

$$P_{hH} \approx \begin{bmatrix} 1/2 & 1/2 & \\ 1/2 & 1 & 1/2 \\ & 1/2 & 1/2 \end{bmatrix}_h .$$

If we combine (2.22) with the definitions of the operators $\bar{R}_H$ and $\bar{R}_h$ it becomes clear that a similar relation for these operators holds:

$$\begin{array}{lllllllll} (\bar{R}_H y)_{ij} & = & (\bar{R}_h y)_{ij} & + & 1/2( & (\bar{R}_h y)_{i,j-1} & + & (\bar{R}_h y)_{i+1,j-1} & + & (\bar{R}_h y)_{i-1,j} & + \\ & & & & & (\bar{R}_h y)_{i+1,j} & + & (\bar{R}_h y)_{i-1,j+1} & + & (\bar{R}_h y)_{i,j+1}\ ) & \end{array}$$

So we have

$$\bar{R}_{Hh} \approx \begin{bmatrix} 1/2 & 1/2 & \\ 1/2 & 1 & 1/2 \\ & 1/2 & 1/2 \end{bmatrix}_h$$

with the operators $R_{Hh}$, $P_{hH}$ and $\bar{R}_{Hh}$ defined as indicated above, it follows that

$$P_H = P_h P_{hH}\ ;\ R_H = R_{Hh} R_h\ ;\ \bar{R}_H = \bar{R}_{Hh} \bar{R}_h\ .$$

In this way a *nested sequence of Galerkin discretisations* can be constructed.

If piecewise polynomials of degree $k$ are used for the interpolation $P_h$ on triangles, then we have the following error estimate (cf. [18])

$$\|P_h a_h - u\|_{L^2(\Omega)} \leq C h^{k+1} \|u\|_{H^{k+1}(\Omega)}. \tag{2.23}$$

From this error estimate we now derive the relative order of convergence for the fine and coarse discretisations of the operator $F$ (cf. definition 2.22). In the following derivation the stability of $P_h$ and $\bar{R}_h$ is used. (which will not be proved here.)

For the remainder of this section we take $E = \hat{E} = L^2(\Omega)$. We have to prove

$$\|F_h^{-1} - P_{hH} F_H^{-1} \bar{R}_{Hh}\|_{E_h \leftarrow B \subset \hat{E}_h} \leq C H^p \quad \text{for some } C \text{ and } p.$$

Take an arbitrary $y_h \in \hat{E}_h$. Since $\bar{R}_h$ is surjective, there exists some $y \in \hat{E}$ such that $y_h = \bar{R}_h y$. Let $y_H = R_{Hh} y_h$. Assume $u$ is the solution of $Fu = y$; $a^h$ is the solution of $F_h a_h = y_h$, (thus $u_h = P_h a_h$); $a^H$ is the solution of $F_H a_H = y_H$ and $u_H = P_H a_H$. Then

$$\begin{array}{l} \|F_h^{-1} y_h - P_{hH} F_H^{-1} \bar{R}_{Hh} y_h\|_{E_h} \\ = \|a_h - P_{hH} F_H^{-1} y_H\|_{E_h} \\ = \|a_h - P_{hH} a_H\|_{E_h} \\ = \|\hat{R}_h (P_h a_h - u + u - P_h P_{hH} a_H)\|_{E_h} \\ = \|\hat{R}_h (u_h - u + u - u_H)\|_{E_h} \\ \leq \|\hat{R}_h\| \left(\|u_h - u\|_{L^2(\Omega)} + \|u_H - u\|_{L^2(\Omega)}\right) \quad \text{(by(2.23))} \\ \leq \|\hat{R}_h\| (C_1 h^2 + C_2 H^2) \|y\|_{L^2(\Omega)} \\ = \|\hat{R}_h\| (C_1 h^2 + C_2 H^2) \|\hat{\bar{P}}_h y_h\|_{L^2(\Omega)} \\ \leq \|\hat{R}_h\|\, \|\hat{\bar{P}}_h\| C_3 H^2 \|y_h\|_{L^2(\Omega)} \\ = C_4 H^2 \|y_h\|_{L^2(\Omega)} \end{array}$$

since $\hat{R}_h$ and $\hat{\bar{P}}_h$ are bounded because of the stability of $R_h$ and $\bar{R}_h$.
It now follows that $\|F_h^{-1} - P_{hH} F_H^{-1} \bar{R}_{Hh}\|_{E_h \leftarrow B \subset \hat{E}_h} \leq C_4 H^2$, so the fine and coarse discretisations of the operator $F$ have relative convergence order 2.

Starting from the same triangulation as given in Figure 2.5, it is also possible to extend the number of nodes in the triangulation as shown in Figure 2.7

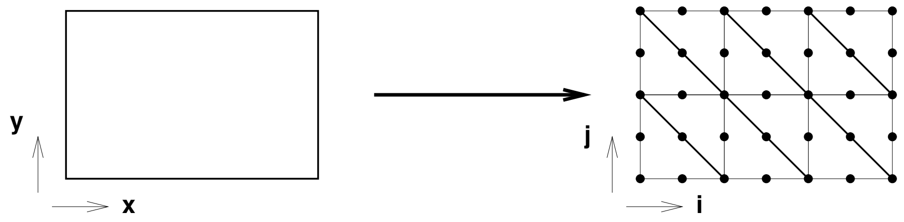


Figure 2.7: Nodal points for a piecewise quadratic approximation

Now we can associate with each node $N_{ij}$ a piecewise parabolic function $\phi_{ij}^h$ with the property (2.16). (Piecewise parabolic: i.e. a $C^0(\Omega)$-function that is a quadratic polynomial on each triangle). Again $u$ is approximated by $u_h = \Sigma_{ij}\, a_{ij}^h \phi_{ij}^h$, but now $P_h : E_h \to E$ is the piecewise parabolic interpolation over the triangles and $R_h, \bar{R}_h$ and $a_h$ remain unchanged, taking notice of the fact that $\phi_{ij}^h$ in the definition of $\bar{R}_h$ now denotes a parabolic function.

Also in this case we can define a coarser discretisation where $P_H$ and $\bar{R}_H$ are defined by piecewise parabolics. Thus again, a nested set of Galerkin discretisations can be obtained. Now the error estimate (2.23) holds with $k = 2$, so with the use of the same $h$ we have a better approximation when piecewise parabolic functions are used (provided that the solution is smooth enough). However, the computational effort necessary to find the approximation $u_h$ of $u$ is much greater when we use the piecewise parabolic functions.

The reason for this greater computational effort is the following. With both the piecewise linear and parabolic functions, the discretisation leads to a linear system $F_h a_h = y_h$ where the matrix $F_h$ is $N(h) \times N(h)$, but this matrix contains only a relatively small number of non-zero elements. From equation (2.18) (from which the elements of $F_h$ are derived) we see that the element of $F_h$ corresponding to $((i,j),(k,l))$ is zero if $\text{Support}(\phi_{ij}^h) \bigcap \text{Support}(\phi_{kl}^h)$ has measure zero. When we consider figure 2.8 it becomes clear that if $N_{kl}$ is an interior node of the triangulation, corresponding to the piecewise linear functions, the equation (2.18) corresponding to this node contains at most 7 variables, whereas this number is 9 or 19 in the triangulation corresponding to the piecewise parabolic functions. So the construction and the solution of the linear system $F_h a_h = y_h$ will be much more time consuming when piecewise parabolic functions are used.

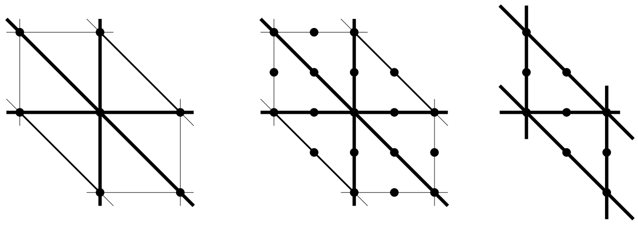

Figure 2.8: Related nodal points for piecewise parabolic approximations

# Chapter 3

# Defect correction and discretisation

## 3.1 Introduction

In the foregoing chapters we discussed defect correction and discretisation. It will now be shown how these techniques can be combined to approximate efficiently the solution of a continuous problem. The basic principle is that less accurate discretisations accelerate the solution of more accurate discretisations. More precisely

1. the solution of a lower order discretisation may accelerate the solution of a higher order discretisation, or

2. the solution of a coarser discretisation may accelerate the solution of a finer discretisation.

The first case will be considered in the next section, the latter will be treated in more detail in Section 3.3.

## 3.2 A fundamental theorem

**Example 3.1** *An accurate and an inaccurate discretisation*
We consider a two-dimensional second-order linear elliptic boundary value problem (e.g. problem (2.14) and two discretisations of it, both obtained by using the finite element method (cf. Section 2.6). The first discretisation with piecewise linear functions on a triangularisation, is denoted by

$$F_h z_h = y_h \,, \quad F_h : E_h \to \hat{E}_h. \tag{3.1}$$

The other discretisation, with piecewise quadratics on the same triangularisation, is denoted by

$$F_h^* z_h = y_h^* \,, \quad F_h^* : E_h^* \to \hat{E}_h^* \,. \tag{3.2}$$

Because we use basicly the same triangularisation for both discretisations, we can identify $E_h$, $E_h^*$, $\hat{E}_h$, and $\hat{E}_h^*$ all with $\mathbb{R}^{N(h)}$, and the matrices $F_h$ and $F_h^*$ have the same dimension.

If the solution of the boundary value problem is smooth, the operator $F_h^*$ will yield a more accurate approximation than $F_h$. From this point of view, it seems advantageous to solve (3.2) to obtain an approximation of the solution of the continuous problem. However, as was shown in section 2.6, $F_h^*$ has a more complex structure than $F_h$. Therefore we prefer to solve (3.1) and we shall use its solution as the initial value in a defect correction process for (3.2) as follows

$$\begin{cases} F_h z_h^{(1)} = y_h \\ F_h z_h^{(i+1)} = (F_h - F_h^*) z_h^{(i)} + y_h^* \quad i = 1, 2, \ldots . \end{cases} \tag{3.3}$$

A stationary point for the iteration (3.3) is a solution for (3.2). Later we shall see that in many cases one or a few of these iteration steps may be enough to obtain a sufficiently accurate approximation to the solution of the original problem. An approximation not less accurate than the one that will be obtained by direct solution of (3.2).
We will now place the above example in a more general context. We consider the continuous problem

$$Lu = f\,, \tag{3.4}$$

where $L : E^\alpha \to \hat{E}^\alpha$ is a linear operator, and $\{E^\alpha |\, \alpha_0 \leq \alpha \leq \alpha_1\}$, $\{\hat{E}^\alpha | \alpha_0 \leq \alpha \leq \alpha_1\}$ are scales of Banach spaces (cf. Appendix). The solution of (3.4) is called $\hat{u}$. We consider a less accurate discretisation of (3.4):

$$L_h u_h = \bar{R}_h f \tag{3.5}$$

and a more accurate discretisation of (3.4):

$$L_h^* u_h = \bar{R}_h^* f \tag{3.6}$$

where $L_h, L_h^* : E_h^\alpha \to \hat{E}_h^\alpha$ are linear operators, and $\{E_h^\alpha | \alpha_0 \leq \alpha \leq \alpha_1\}$, $\{\hat{E}_h^\alpha | \alpha_0 \leq \alpha \leq \alpha_1\}$ are scales of Banach spaces corresponding with $\{E^\alpha | \alpha_0 \leq \alpha \leq \alpha_1\}$ and $\{\hat{E}^\alpha | \alpha_0 \leq \alpha \leq \alpha_1\}$ respectively. As in the example, the defect correction process

$$\begin{cases} L_h u_h^{(1)} = \bar{R}_h f\,, \\ L_h u_h^{(i+1)} = (L_h - L_h^*) u_h^{(i)} + \bar{R}_h^* f\,, \quad i = 1, 2 \ldots\,, \end{cases} \tag{3.7}$$

is used to approximate the solution of (3.6). When the defect correction process (3.7) converges slowly (or does not converge) the use of (3.7) will be less efficient than the immediate solution of (3.6). The following theorem will show under what conditions the use of (3.7) might be preferable.

**Theorem 3.2** Let the problem (3.4) be discretised by (3.5) and (3.6), and let

1. $L_h$ be $\sigma$-stable for some $\sigma \geq 0$,
2. $L_h$ be consistent of order $\tilde{p}$, for all $\tilde{p} \in [0, p]$,
3. $L_h^*$ be consistent of order $\tilde{p}^*$, for all $\tilde{p}^* \in [0, p^*]$, $p^* > p$,
4. $L_h$ and $L_h^*$ be relatively consistent of order $p$.

Then the iterands in (3.7) satisfy the error estimate

$$\|u_h^{(i)} - R_h \hat{u}\|_{E_h^\omega} \leq C h^{\min(p^*, ip)} \|\hat{u}\|_{E^{\omega + \beta_i}}\,, \tag{3.8}$$

with $\beta_i = \min(\sigma + p^*, i(\sigma + p))$ and $C$ independent of $h$ and further $\omega, i$ such that $\omega, \omega + i(p + \sigma) \in [\alpha_0, \alpha_1]$.
**Proof:** We will first put the conditions 1) - 4) in a more explicit form:

1. $\exists C_1$ independent of $h : \|L_h^{-1}\|_{E_h^{\alpha-\sigma} \leftarrow \hat{E}_h^\alpha} \leq C_1$

2. $p$ is a real number for which

$$\|\bar{R}_h L - L_h R_h\|_{\hat{E}_h^\alpha \leftarrow E^{\alpha+\tilde{p}}} \leq C_2 h^{\tilde{p}}$$

for all $0 \leq \tilde{p} \leq p$ and some $C_2$ independent of $h$. This implies that

$$\|\bar{R}_h L\hat{u} - L_h R_h \hat{u}\|_{\hat{E}_h^\alpha} = \|\bar{R}_h f - L_h R_h \hat{u}\|_{\hat{E}_h^\alpha}, \leq C_2 h^{\tilde{p}} \|\hat{u}\|_{E^{\alpha+\tilde{p}}} \tag{3.9}$$

for all $0 \leq \tilde{p} \leq p$.

3. $p^*$ is a real number for which

$$\|\bar{R}_h^* L - L_h^* R_h\|_{\hat{E}_h^\alpha \leftarrow E^{\alpha+\tilde{p}}} \leq C_3 h^{\tilde{p}}$$

for all $0 \leq \tilde{p} \leq p$ and some $C_3$ independent of $h$. This implies that

$$\|\bar{R}_h^* L\hat{u} - L_h^* R_h \hat{u}\|_{\hat{E}_h^\alpha} = \|\bar{R}_h^* f - L_h^* R_h^* \hat{u}\|_{\hat{E}_h^\alpha} \leq C_3 h^{\tilde{p}} \|\hat{u}\|_{E^{\alpha+\tilde{p}}}$$

for all $0 \leq \tilde{p} \leq p^*$

4. 

$$\|L_h - L_h^*\|_{\hat{E}_h^\alpha \leftarrow E_h^{\alpha+\tilde{p}}} \leq C_4 h^{\tilde{p}}$$

for all $0 \leq \tilde{p} \leq p$ and some $C_4$ independent of $h$.

Now we are in the position to prove the theorem by induction:

$$u_h^{(1)} - R_h \hat{u} = L_h^{-1} \bar{R}_h f - R_h \hat{u} = L_h^{-1} [\bar{R}_h L - L_h R_h] \hat{u} \,.$$

Hence

$$\|u_h^{(1)} - R_h \hat{u}\|_{E_h^\omega} \leq \|L_h^{-1}\|_{E_h^\omega \leftarrow \hat{E}_h^{\omega+\sigma}} \|\bar{R}_h L - L_h R_h\|_{\hat{E}_h^{\omega+\sigma} \leftarrow E^{\omega+\sigma+p}} \|\hat{u}\|_{E^{\omega+\sigma+p}} \,.$$

With the use of 1) $(\alpha = \omega + \sigma)$ and 2) $(\alpha = \omega + \sigma)$ we find

$$\|u_h^{(1)} - R_h \hat{u}\|_{E_h^\omega} \leq C_1 C_2 h^p \|\hat{u}\|_{E^{\omega+\sigma+p}} \,.$$

Because (3.6) is a discretisation of higher order than (3.5) we have $p^* \geq p$ and it follows that

$$\|u_h^{(1)} - R_h \hat{u}\|_{E_h^\omega} \leq C h^{\min(p^*,p)} \|\hat{u}\|_{E^{\omega+\min(\sigma+p^*,\sigma+p)}} \,,$$

and thus the theorem is proved for $i = 1$.

Now, suppose that the theorem is proved for some $i \geq 1$. We will prove it for $i + 1$:

$$\begin{aligned} u_h^{(i+1)} - R_h \hat{u} &= L_h^{-1} [(L_h - L_h^*) u_h^{(i)} + \bar{R}_h^* f] - R_h \hat{u} \\ &= u_h^{(i)} - R_h \hat{u} - L_h^{-1} [L_h^* u_h^{(i)} - \bar{R}_h^* L \hat{u}] \\ &= [I - L_h^{-1} L_h^*] \, [u_h^{(i)} - R_h \hat{u}] - L_h^{-1} [L_h^* R_h - \bar{R}_h^* L] \hat{u} \\ &= L_h^{-1} [L_h - L_h^*] \, [u_h^{(i)} - R_h \hat{u}] - L_h^{-1} [L_h^* R_h - \bar{R}_h^* L] \hat{u} \,. \end{aligned}$$

It now follows that

$$\begin{aligned} \|u_h^{(i+1)} - R_h \hat{u}\|_{E_h^\omega} &\leq \|L_h^{-1} [L_h - L_h^*] \, [u_h^{(i)} - R_h \hat{u}]\|_{E_h^\omega} + \|L_h^{-1} [L_h^* R_h - \bar{R}_h^* L] \hat{u}\|_{E_h^\omega} \\ &\leq \|L_h^{-1}\|_{E_h^\omega \leftarrow \hat{E}_h^{\omega+\sigma}} \|L_h - L_h^*\|_{\hat{E}_h^{\omega+\sigma} \leftarrow E_h^{\omega+\sigma+p}} \|u_h^{(i)} - R_h \hat{u}\|_{E_h^{\omega+\sigma+p}} + \end{aligned}$$

$$\|L_h^{-1}\|_{E_h^\omega \leftarrow \hat{E}_h^{\omega+\sigma}} \, \|L_h^* R_h - \bar{R}_h^* L\|_{\hat{E}_h^\omega+\sigma \leftarrow E^{\omega+\sigma+p^*}} \, \|\hat{u}\|_{E^{\omega+\sigma+p^*}} \, .$$

Using 1) ($\alpha = \omega + \sigma$), 3) ($\alpha = \omega + \sigma$) and 4) ($\alpha = \omega + \sigma$) we find

$$\|u_h^{(i+1)} - R_h \hat{u}\|_{E_h^\omega} \ \leq \ C_1 C_4 h^p \|u_h^{(i)} - R_h \hat{u}\|_{E_h^{\omega+\sigma+p}} + C_1 C_3 h^{p^*} \|\hat{u}\|_{E^{\omega+\sigma+p^*}} \, .$$

Application of the induction hypothesis leads to

$$\begin{aligned} \|u_h^{(i+1)} - R_h \hat{u}\|_{E_h^\omega} \ &\leq \ C_1 C_4 h^p C_5 h^{\min(p^*, ip)} \|\hat{u}\|_{E^{\omega+\sigma+p+\beta_i}} + C_1 C_3 h^{p^*} \|\hat{u}\|_{E^{\omega+\sigma+p^*}} \ \leq \\ &\leq \ C_1 C_4 C_5 h^{\min(p^*, (i+1)p)} \|\hat{u}\|_{E^{\omega+\beta_{i+1}}} + C_1 C_3 h^{\min(p^*, (i+1)p)} \|\hat{u}\|_{E^{\omega+\beta_{i+1}}} \\ &\leq \ C h^{\min(p^*, (i+1)p)} \|\hat{u}\|_{E^{\omega+\beta_{i+1}}} \, . \end{aligned}$$

In the one but last inequality we used the fact that

$$\| \cdot \|_\alpha \leq \| \cdot \|_{\alpha+\eta} \ \text{ for } \ \eta \geq 0 \, .$$

■

**Remarks:**

- The theorem requires no stability of $L_h^*$.
- If $\bar{R}_h = \bar{R}_h^*$ and the set of restrictions $\{R_h | h \in \mathcal{H}\}$ is stable, then the requirement 4) from the theorem follows from the requirements 2) and 3):

$$\begin{aligned} \|L_h - L_h^*\|_{\hat{E}_h^\alpha \leftarrow E_h^{\alpha+\tilde{p}}} &= \|(L_h - L_h^*) R_h \hat{P}_h\|_{\hat{E}_h^\alpha \leftarrow E_h^{\alpha+\tilde{p}}} \\ &\leq \ \|L_h R_h - L_h^* R_h\|_{\hat{E}_h^\alpha \leftarrow E^{\alpha+\tilde{p}}} \, \|\hat{P}_h\|_{E^{\alpha+\tilde{p}} \leftarrow E_h^{\alpha+\tilde{p}}} \\ &\leq \ C_6 \|L_h R_h - \bar{R}_h L + \bar{R}_h L - L_h^* R_h\|_{\hat{E}_h^\alpha \leftarrow E^{\alpha+\tilde{p}}} \\ &\leq \ C_6 \{ \|L_h R_h - \bar{R}_h L\|_{\hat{E}_h^\alpha \leftarrow E^{\alpha+\tilde{p}}} + \|\bar{R}_h L - L_h^* R_h\|_{\hat{E}_h^\alpha \leftarrow E^{\alpha+\tilde{p}}} \} \\ &\leq \ C_6 \{ C_2 h^{\tilde{p}} + C_3 h^{\tilde{p}} \} \ \leq \ C h^{\tilde{p}} \, , \ \text{for all } \ 0 \leq \tilde{p} \leq p \, . \end{aligned}$$

## 3.3 DCP with an approximate inverse of deficient rank

**Example 3.3** *A fine and a coarser discretisation*
Here we show how a coarser discretisation can be used to accelerate the solution process for a finer discretisation. Again we take as a model problem the two-dimensional second-order linear elliptic boundary value problem. We are interested in a coarse and a fine finite-element discretisation of this problem with piecewise linear functions on a triangularisation. Again the finer discretisation is denoted by

$$F_h z_h = y_h \, , \tag{3.10}$$

and the coarser by

$$F_H z_H = y_H \, . \tag{3.11}$$

It is assumed that the two discretisations are nested (e.g. $H = 2h$), i.e. that there exist a prolongation $P_{hH}$ and a restriction $\bar{R}_{Hh}$ with property

$$P_H = P_h P_{hH} \, , \ \bar{R}_H = \bar{R}_{Hh} \bar{R}_h \, .$$

It is clear that the solution of (3.10) is much more expensive to compute than the solution of (3.11), because the size of matrix for $F_h$ is larger than that for $F_H$. For this reason we are inclined to use the solution of (3.11) as an auxiliary problem in a defect-correction process for the solution of (3.10). It seems natural to use $P_{hH}F_H^{-1}\bar{R}_{Hh}$ as an approximate inverse for $F_h$. This results in the *coarse grid correction process*

$$\begin{cases} z_h^{(1)} & = P_{hH}z_H \\ z_h^{(i+1)} & = z_h^{(i)} - P_{hH}F_H^{-1}\bar{R}_{Hh}(F_h z_h^{(i)} - y_h) \end{cases} \tag{3.12}$$

Notice that the operator $P_{hH}F_H^{-1}\bar{R}_{Hh}$ is not full rank: $\text{Rank}(P_{hH}F_H^{-1}\bar{R}_{Hh}) = N(H) < N(h)$. The following figure may illustrate this

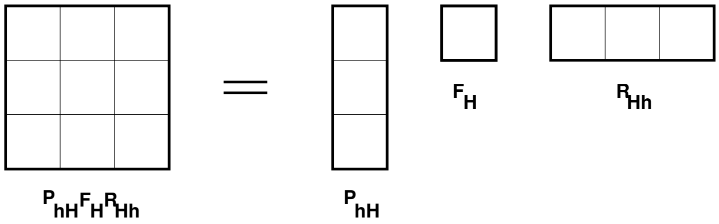


Figure 3.1: Rank deficiency.

From now on the above example will be generalised. The more general form of the iteration step in (3.12) is

$$z^{(i+1)} = z^{(i)} - \tilde{G}(Fz^{(i)} - y)\,, \tag{3.13}$$

where

$$\begin{array}{l} \tilde{G} = PS\bar{R} : \hat{E}_h \to E_h \ \text{ is not full rank ,} \\ S \ : \ \hat{E_H} \to E_H \ \text{ is full rank ,} \\ \bar{R} \ : \ \hat{E}_h \to \hat{E}_H \ \text{ is a restriction and ,} \\ P \ : \ E_H \to E_h \ \text{ is a prolongation. ,} \\ F \ : \ E_h \to \hat{E}_h \ \text{ is full rank .} \end{array}$$

**Definition 3.4** Let $S, \bar{R}, P$ and $F$ be as above. Then we call the operator $E : E_H \to \hat{E}_H$, defined by

$$E = S^{-1} - \bar{R}FP\,,$$

the *deviation from the Galerkin approximation* corresponding to (3.13).

Notice that with the use of Galerkin approximation $S = (RFP)^{-1}$, we have $E = 0$.

**Definition 3.5** Let $\hat{R}$ be a restriction related to $P$, i.e. its left inverse, $\hat{R} = (P^TP)^{-1}P^T$. Let $e$ be the error in the approximation $z^{(i)}$, i.e. $e = z^{(i)} - F^{-1}y$. Then we define $e_s$, em the smooth part of the error, and $e_u$, the em unsmooth or em rapidly varying part of the error by

$$e_s = P\hat{R}e; \quad e_u = (I - P\hat{R})e\,.$$

**Remark:** The operator $P\hat{R}$ represents the projection of $E_h$ on Range$(P) \subset E_h$. The smooth part of the error is the part of the error lying in Range$(P)$, the remaining part, $e_u$, lying in its complement Range$(P)^T$.

We now want to examine the effect of the amplification operator of the error, $M = I - \tilde{G}F$, on an error $e = e_s + e_u$. We examine $e_s$ and $e_u$ separately. Since $e_s \in$ Range$(P)$, we have $e_s = P\hat{R}e_s$, and

$$\begin{aligned} Me_s = \quad & MP\hat{R}e_s = (I - PS\bar{R}F)P\hat{R}e_s = P(I - S\bar{R}FP)\hat{R}e_s \\ & = P(I - S(S^{-1} - E))\hat{R}e_s = PSE\hat{R}\,e_s \in \text{Range}(P), \text{ and} \\ Me_u = \quad & (I - PS\bar{R}F)e_u = e_u - PS\bar{R}Fe_u. \end{aligned}$$

So we have the situation as shown in figure 3.2. We can follow the same procedure for the residual.

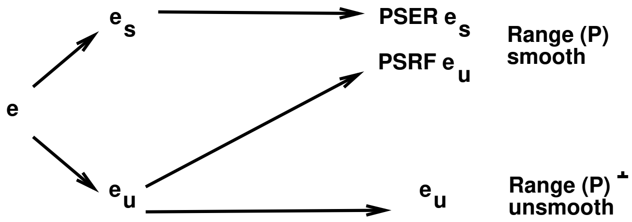


Figure 3.2: Effect of coarse grid correction on the error.

**Definition 3.6** Let $\hat{\bar{P}}$ be a prolongation related to $\bar{R}$, i.e. a right inverse, e.g. $\hat{\bar{P}} = \bar{R}^T(RR^T)^{-1}$. Let $r$ be the residual for some approximation $z^{(i)}$, i.e. $r = Fz^{(i)} - y$. Then we define $r_s$, *the smooth part of the residual,* and $r_u$, the *unsmooth* or *rapidly varying part of the residual* by

$$r_s = \hat{\bar{P}}\bar{R}r; \quad r_u = (I - \hat{\bar{P}}\bar{R})r.$$

**Remark:** The operator $\hat{\bar{P}}\bar{R}$ represents the projection of $\hat{E}_h$ on the complement of Kernel$(\bar{R})$ and $r_u \in$ Kernel$(\bar{R})$.

The effect of the amplification operator of the residual, $\overline{M} = I - F\tilde{G}$, on $r = r_s + r_u$, can now be derived from its effect on $r_s$ and $r_u$:

$$\begin{aligned} \overline{M}r_s \quad & = (I - FPS\bar{R})r_s \\ & = (I - \hat{\bar{P}}\bar{R}FPS\bar{R} + \hat{\bar{P}}\bar{R}FPS\bar{R} - FPS\bar{R})r_s \\ & = (I - \hat{\bar{P}}(S^{-1} - E)S\bar{R} + (\hat{\bar{P}}\bar{R} - I)FPS\bar{R})r_s \\ & = (I + \hat{\bar{P}}ES\bar{R} - \hat{\bar{P}}\bar{R} - (I - \hat{\bar{P}}\bar{R})FPS\bar{R})r_s \\ & = \hat{\bar{P}}ES\bar{R}r_s - (I - \hat{\bar{P}}\bar{R})F\tilde{G}r_s\,, \end{aligned}$$

where $(I - \hat{\bar{P}}\bar{R})F\tilde{G}r_s \in$Kernel$(\bar{R})$, and $\hat{\bar{P}}ES\bar{R}r_s \in$Kernel$(\bar{R})^T$, the complement of Kernel$(R)$, because $\hat{\bar{P}}\bar{R}(\hat{\bar{P}}ES\bar{R}r_s) = \hat{\bar{P}}ES\bar{R}r_s$ .
Furthermore, because $r_u \in$Kernel$(\bar{R})$, we have

$$\overline{M}r_u = (I - FPS\bar{R})r_u = r_u,$$

We summarise this in Figure 3.3.

Suppose we have a smoothing operator $A$, which suppresses unsmooth components and leaves smooth components (more or less) unchanged. If we apply this operator before performing a coarse-grid correction, the result will be a much smaller error if the deviation from the Galerkin approximation is small. Figure 3.4 illustrates this point. The reader may easily check

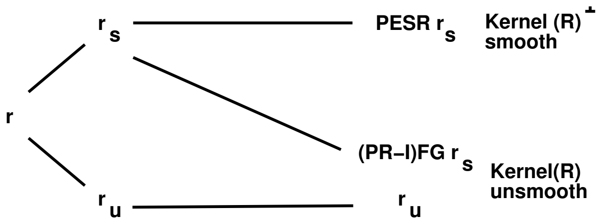


Figure 3.3: Effect of coarse grid correction on the residual.

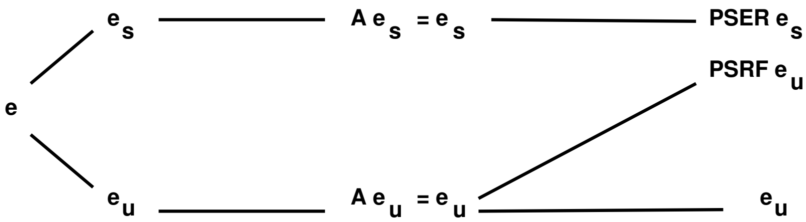


Figure 3.4: Effect of smoothing and coarse grid correction on the error.

that applying the above process to the residual $r$ does not lead to a guaranteed smaller residual. However, if we perform a coarse-grid correction before applying the smoothing operator $A$, the result will be a much smaller residual, whereas the error may not improve. The first statement is illustrated in Figure 3.5.

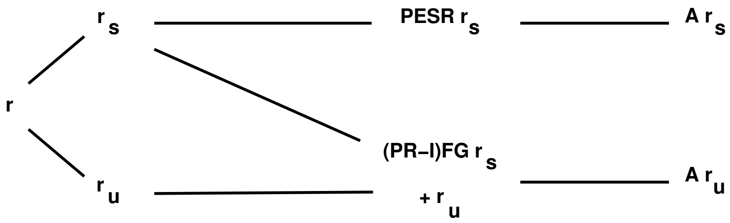


Figure 3.5: Effect of coarse grid correction and smoothing on the residual.

These considerations show that the duality between error and residual is reflected in the importance of relaxation before or after the application of the coarse grid correction. For reduction of the residual *relaxation after,* for reduction of the error *relaxation before* the coarse grid correction is required.

# Chapter 4

# Example of a two-grid algorithm

## 4.1 Introduction

Before we introduce multigrid algorithms, we first give a simple example of a two-grid algorithm. This is a pre-version of a multigrid algorithm. The example shows a simple two-point boundary value problem and a simple iterative solution method for which the behaviour of the iterative process can be analyzed exactly. In this example the main features of a multigrid algorithm are already visible. We consider the two-point boundary value problem

$$-u''(x) = f(x)\,, \quad x \in \Omega = (0,1), \tag{4.1}$$

$$u(0) = u(1) = 0\,.$$

Below, this problem will be discretised and its discretisation will be solved by means of the two-grid algorithm.

## 4.2 Discretisation

We provide $\bar{\Omega}$ with a uniform mesh $\bar{\Omega}_h = \{x_i | x_i = i/N;\ i = 0, 1, ..., N\}$, where $h$ equals $1/N$, as is shown in Figure 4.1.

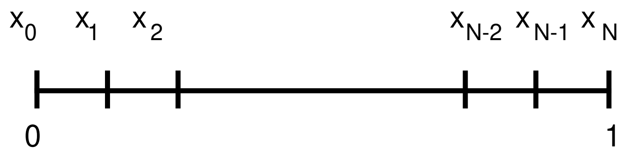


Figure 4.1: The one-dimensional mesh

The problem (4.1) is discretised by a finite difference method (FDM), using the mesh $\Omega_h$. (The reader may check for himself that the use of the finite element method (FEM) leads to exactly the same discrete problem, when piecewise linear test and trial functions are used). Writing $u(x_i) =: u_i$, we have

$$u''(x_i) = \frac{1}{h^2}(u_{i-1} - 2u_i + u_{i+1}) + \mathcal{O}(h^2), \quad i = 1, 2, ..., N-1.$$

This can be verified with the aid of Taylor-expansions. Keeping this in mind the following discretisation of (4.1) is obvious:

$$\frac{1}{h^2}(-u_{i-1}+2u_i-u_{i+1}) = f_i\,, \quad i=1,2,...,N-1. \tag{4.2}$$

Implementation of the boundary conditions, $u_0 = u_N = 0$, yields the discrete linear system

$$L_h u_h = f_h\,, \tag{4.3}$$

with the $(N-1)\times(N-1)$ matrix

$$L_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & & \ddots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix},$$

and the $(N-1)$ vectors

$$u_h = (u_1, u_2, ..., u_{N-1})^T.$$
$$f_h = (f_1, f_2, ..., f_{N-1})^T.$$

## 4.3 The smoothing process

We consider the damped Jacobi-relaxation method for the iterative solution of (4.3). One iteration step in this process reads

$$u_h^{(i+1)} = u_h^{(i)} - \omega D_h^{-1}(L_h u_h^{(i)} - f_h), \tag{4.4}$$

where $D_h$ is the main diagonal of $L_h$ and $\omega \in I\!R$ is the damping parameter. By $\hat{u}_h$ we denote the solution of (4.3); then the error in the $i$-th iterand is

$$e_h^{(i)} = u_h^{(i)} - \hat{u}_h.$$

and

$$e_h^{(i+1)} = M_h^{DJM} e_h^{(i)},$$

where $M_h^{DJM}$ is the amplification-operator of the error for the damped Jacobi-method, which reads

$$M_h^{DJM} = I - \omega D_h^{-1} L_h, \tag{4.5}$$

as can be seen easily from (4.4). The eigensystem of the matrix

$$L_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & & \ddots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix},$$

is $\{\lambda_m, \phi_m\}_{m=1,2...N-1}$, with eigenvalues

$$\lambda_m = \frac{2-2\,\cos(m\pi h)}{h^2} = \frac{4\,\sin^2(m\pi h/2)}{h^2}\,, \tag{4.6}$$

and eigenfunctions

$$\phi_m = (\sin(m\pi h),\, \sin(2m\pi h),\, ...,\, \sin((N-1)m\pi h))^T\,. \tag{4.7}$$

Noticing that $\phi_m$ must be considered as a discrete function on the mesh $\Omega_h = \{x_i | x_i = i/N;\, i = 1, 2, ..., N-1\}$, the following terminology is natural. An eigenfunction $\phi_m$ is said to have a low (high) frequency if $m$ is small (large). From (4.5) and (4.6) it is clear that $M_h^{DJM}$ has the same eigenfunctions $\phi_m$ as $L_h$ and eigenvalues

$$\lambda(M_h^{DJM})_m = 1 - \frac{\omega}{2} 4 \sin^2(m\pi h/2) = \cos^2(m\pi h/2) + (1-2\omega)\,\sin^2(m\pi h/2). \tag{4.8}$$

The index $m$ takes the values $1 \le m \le N-1$ and hence the arguments

$$0 < \frac{\pi h}{2} \le \frac{m\pi h}{2} \le \frac{(N-1)\pi h}{2} < \frac{\pi}{2}.$$

In order to get monotonous decreasing $\lambda(M_h^{DJM})_m$ for increasing $m$, we select $\omega = 1/2$; we then find

$$\lambda(M_h^{DJM})_m = \cos^2(m\pi h/2).$$

We see that eigenfunctions with a low frequency are damped slowly by $M_h^{DJM}$, whereas eigenfunctions having a high frequency are damped efficiently. After a few iterations with the damped Jacobi method ($\omega = 1/2$), the rapidly varying component in the error will have almost vanished. However, the slowly varying components will hardly be affected: the error has not become much smaller. The total error has become much smoother. Application of one step of the process (4.4) is therefore also called a *smoothing step.*

## 4.4 The coarse grid correction

We just saw how to approximate $\hat{u}_h$ with a smooth error. Now we will show how to decrease the size of the error when it is smooth. Let $\tilde{u}_h$ be the approximation with a smooth error of the discrete solution $\hat{u}_h$. Then the error $\tilde{e}_h$ given by $\tilde{e}_h = \tilde{u}_h - \hat{u}_h$ satisfies the equation

$$L_h\tilde{e}_h = L_h\tilde{u}_h - L_h\hat{u}_h = L_h\tilde{u}_h - f_h = d_h = -r_h, \tag{4.9}$$

$r_h$ is the is the residual of $\tilde{u}_h$; $d_h$ is the defect of $\tilde{u}_h$. If $\tilde{e}_h$ is a smooth function, we are well able to represent it well on a coarser grid. To find it we may solve the discrete problem on a coarser grid e.g.

$$L_H\tilde{e}_H = -r_H = d_H\,,\; H = 2h, \tag{4.10}$$

where (4.10) is a discretisation of (4.9). We want that $\tilde{e}_H$ is a good coarse-grid approximation of $\tilde{e}_h$. For this we need

- a restriction $\bar{R}_{Hh}$,
- a coarse-grid operator $L_H$,
- a prolongation $P_{hH}$,

as is shown in Figure 4.2. For $L_H$ we take an operator similar to $L_h$, only with a mesh $H = 2h$ instead of $h$. Thus, $L_H$ is a $(N/2-1)\times(N/2-1)$ matrix. (We assume that $N$ is an even integer.) A simple prolongation, $P_{hH}$, is found by linear interpolation. This operator $P_{hH}$ is defined by $u_h = P_{hH}u_H$ such that

$$\begin{cases} u_h(x_i) &=& u_H(x_i) & \text{if } x_i \in \Omega_H, \\ u_h(x_i) &=& (u_H(x_i+h) + u_H(x_i-h))/2 & \text{if } x_i \notin \Omega_H, \end{cases}$$

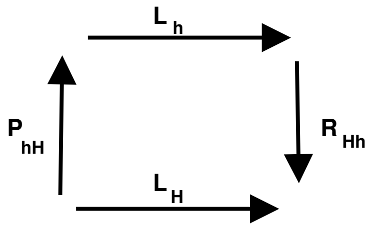


Figure 4.2: The Galerkin relation between coarse and fine grid operator.

for all $x_i = \Omega_h$. A simple restriction, $\bar{R}_{Hh}$, is found by *injection.* This restriction is defined by $f_H = \bar{R}_{Hh} f_h$, with $f_h \in \hat{E}_h$ and $f_H \in \hat{E}_H$ such that

$$f_H(x_i) = f_h(x_i), \text{ for all } x_i \in \Omega_H .$$

Another possible *restriction* is the *weighed restriction.* defined by $f_H = \bar{R}^*_{Hh} f_h$, with $f_h \in \hat{E}_h$ and $f_H \in \hat{E}_H$ such that

$$f_H(x_i) = (f_h(x_i + h) + 2 f_h(x_i) + f_h(x_i - h))/4,$$

for all $x_i \in \Omega_H$. The operators $L_H, P_{hH}, \bar{R}_{Hh}$ and $\bar{R}^*_{Hh}$ can be described explicitly by their matrices

$$L_H = \frac{1}{H^2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & & \ddots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}, \text{ an } (N/2-1) \times (N/2-1) \text{ matrix,} \tag{4.11}$$

$$P_{hH} = \begin{pmatrix} 1/2 & & & & \\ 1 & & & & \\ 1/2 & 1/2 & & & \\ & 1 & & & \\ & 1/2 & 1/2 & & \\ & & 1 & & \\ & & 1/2 & 1/2 & \\ & & & \vdots & \\ & & & 1/2 & 1/2 \\ & & & & 1 \\ & & & & 1/2 \end{pmatrix}, \text{ an } (N-1) \times (N/2-1) \text{ matrix,.} \tag{4.12}$$

$$\bar{R}_{Hh} = \begin{pmatrix} 0 & 1 & 0 & & & & & & \\ & & 0 & 1 & 0 & & & & \\ & & & & 0 & 1 & 0 & & \\ & & & & & & \ddots & & \\ & & & & & & 0 & 1 & 0 \end{pmatrix}, \text{an } (N/2-1) \times (N-1) \text{ matrix, and} \tag{4.13}$$

$$\bar{R}^*_{Hh} = \begin{pmatrix} 1/4 & 1/2 & 1/4 & & & & \\ & & 1/4 & 1/2 & 1/4 & & \\ & & & & \ddots & & \\ & & & & 1/4 & 1/2 & 1/4 \end{pmatrix}, \text{ also an } (N/2-1) \times (N-1) \text{ matrix .} \tag{4.14}$$

The vectors $u_h$ and $u_H$ have the form

$$u_h = (u_h(h), u_h(2h), ..., u_h((N-1)h))^T,$$

and

$$u_H = (u_H(H), u_H(2H), ..., u_H((N/2-1)H))^T.$$

The coarse discretisation (4.10) is used to improve the approximate solution $\tilde{u}_h$ as follows:

- first we compute $\tilde{e}_H$ from either $L_H\tilde{e}_H = d_H = \bar{R}_{Hh}d_h$ or $L_H\tilde{e}_H = d_H = \bar{R}^*_{Hh}d_h$,
- then we subtract the interpolated error from the approximate solution:

$$\hat{\tilde{u}}_h := \tilde{u}_h - P_{hH}\tilde{e}_H.$$

This correction of the approximate solution is the *coarse grid correction step :*

$$\hat{\tilde{u}}_h := \tilde{u}_h - P_{hH}L_H^{-1}\bar{R}_{Hh}(L_h\tilde{u}_h - f_h). \tag{4.15}$$

The amplification operator of the error in a coarse grid correction step is clearly either

$$M_h^{CGC} = I - P_{hH}L_H^{-1}\bar{R}_{Hh}L_h$$

or

$$M_h^{CGC*} = I - P_{hH}L_H^{-1}\bar{R}^*_{Hh}L_h\,.$$

## 4.5 The two level algorithm

The Two Level Algorithm (TLA), which is a preversion of the Multi Level Algorithm, is an iterative process for the solution of (4.3), in which each step consists of

- a number of $p$ smoothing steps,
- a coarse grid correction step,
- a number of $q$ smoothing steps.

In the order to see what the effect is of a TLA step on the error in an approximate solution, we decompose the error into eigenfunction components

$$e_h = \sum_m \alpha_m\phi_m,$$

and we consider the effect on a single eigenfunction. The amplification operator of the error in one step of the two level algorithm will be denoted by $M_h^{TLA}$. For $\phi_m$ as given (4.7) we shall compute

$$M_h^{TLA}\phi_m = (M_h^{DJM})^q(M_h^{CGC})(M_h^{DJM})^p\phi_m \text{ and}$$
$$M_h^{TLA*}\phi_m = (M_h^{DJM})^q(M_h^{CGC*})(M_h^{DJM})^p\phi_m.$$

We will first consider the two level algorithm in which the simple restriction $\bar{R}_{Hh}$ is used in the coarse-grid correction step. To this end we first compute

$$M_h^{CGC}\phi_m = \phi_m - P_{hH}L_H^{-1}\bar{R}_{Hh}L_h\phi_m$$

$$= \phi_m - \frac{4}{h^2}\sin^2(m\pi h/2)P_{hH}L_H^{-1}\bar{R}_{Hh}\phi_m.$$

$$P_{hH} L_H^{-1} \bar{R}_{Hh} \begin{pmatrix} \sin(\pi m h) \\ \sin(2\pi m h) \\ \sin(3\pi m h) \\ \vdots \\ \sin((N-1)\pi m h) \end{pmatrix} = P_{hH} L_H^{-1} \begin{pmatrix} \sin(2\pi m h) \\ \sin(4\pi m h) \\ \sin(6\pi m h) \\ \vdots \\ \sin((N-2)\pi m h) \end{pmatrix}.$$

Since $2h = H$, and

$$(\sin(\pi m H), \sin(2\pi m H), \sin(3\pi m H), \ldots, \sin((N/2-1)\pi m H))^T$$

is an eigenvector of $H^2 L_H$, with eigenvalue $4\sin^2(m\pi H/2)$, the last expression equals

$$\frac{H^2}{4\sin^2(m\pi H/2)} P_{hH} \begin{pmatrix} \sin(\pi m H) \\ \sin(2\pi m H) \\ \sin(3\pi m H) \\ \vdots \\ \sin((N/2-1)\pi m H) \end{pmatrix} =$$

$$= \frac{H^2}{4\sin^2(m\pi H/2)} \cdot \frac{1}{2} \begin{pmatrix} \sin(0) + \sin(\pi m H) \\ 2\ \sin(\pi m H) \\ \sin(\pi m H) + \sin(2\pi m H) \\ 2\ \sin(2\pi m H) \\ \vdots \\ 2\ \sin((N/2-1)\pi m H) \\ \sin((N/2-1)\pi m H) + \sin((N/2)\pi m H) \end{pmatrix} =$$

$$= \frac{h^2}{\sin^2(m\pi h)} \cdot \frac{1}{2} \begin{pmatrix} 2.\sin(\pi m h)\cos(\pi m h) \\ 2.\sin(2\pi m h) \\ 2.\sin(3\pi m h)\cos(\pi m h) \\ 2.\sin(4\pi m h) \\ 2.\sin(5\pi m h)\cos(\pi m h) \\ \vdots \\ 2.\sin((N-2)\pi m h) \\ 2.\sin((N-1)\pi m h)\cos(\pi m h) \end{pmatrix}$$

the last equality being valid because

$$\sin\alpha + \sin\beta = 2\ \sin(\frac{\alpha+\beta}{2})\ \cos(\frac{\alpha-\beta}{2}).$$

If we write $C_m := \cos(\pi m h/2)$ and $S_m := \sin(\pi m h/2)$ and take $a = C_m^2$, $b = -S_m^2$, we obtain

$$\begin{aligned} a - b &= 1\,, \\ a + b &= \cos(m\pi h)\,. \end{aligned}$$

If we furthermore notice that

$$\sin(k\pi(N-m)h) = \sin(k\pi N_h - k\pi m h) = \sin(k\pi - k\pi m h) =$$

$$(-1)^k \sin(-k\pi m h) = (-1)^{k+1}\sin(k\pi m h)$$

then it is clear that

$$\frac{h^2}{\sin^2(m\pi h)} \begin{pmatrix} \sin(\pi mh)\cos(\pi mh) \\ \sin(2\pi mh) \\ \sin(3\pi mh)\cos(\pi mh) \\ \sin(4\pi mh) \\ \vdots \\ \sin((N-2)\pi mh) \\ \sin((N-1)\pi mh)\cos(\pi mh) \end{pmatrix}$$

$$\frac{h^2}{\sin^2(m\pi h)} \left[ a \begin{pmatrix} \sin(\pi mh) \\ \sin(2\pi mh) \\ \sin(3\pi mh) \\ \sin(4\pi mh) \\ \vdots \\ \sin((N-1)\pi mh) \end{pmatrix} + b \begin{pmatrix} +\sin(\pi mh) \\ -\sin(2\pi mh) \\ +\sin(3\pi mh) \\ -\sin(4\pi mh) \\ \vdots \\ +\sin((N-1)\pi mh) \end{pmatrix} \right] =$$

$$= \frac{h^2(a\phi_m + b\phi_{N-m})}{\sin^2(\pi mh)},$$

so we have because $\sin^2(\pi mh) = 4S_m^2 C_m^2$

$$P_{hH} L_H^{-1} \bar{R}_{Hh} \phi_m = \frac{h^2(C_m^2\phi_m - S_m^2\phi_{N-m})}{4S_m^2 C_m^2} \tag{4.16}$$

and

$$M_h^{CGC}\phi_m = \phi_m - \frac{1}{C_m^2}(C_m^2\phi_m - S_m^2\phi_{N-m}) = \frac{S_m^2}{C_m^2}\phi_{N-m}\,.$$

Using (4.8) with $\omega = 1/2$ this results in

$$\begin{aligned} M_h^{TLA}\phi_m &= (M_h^{DJM})^q (M_h^{CGC})(M_h^{DJM})^p \phi_m \\ &= C_m^{2p}(M_h^{DJM})^q (M_h^{CGC})\phi_m \\ &= C_m^{2p-2} S_m^2 (M_h^{DJM})^q \phi_{N-m} \\ &= C_{N-m}^{2q} S_m^2 C_m^{2p-2} \phi_{N-m}\,. \end{aligned}$$

Because $C_{N-m} = S_m$ and $S_{N-m} = C_m$ we have

$$M_h^{TLA}\phi_m = S_m^{2q+2} C_m^{2p-2}\phi_{N-m}\,,$$

and

$$M_h^{TLA}\phi_{N-m} = C_m^{2q+2} S_m^{2p-2}\phi_m\,.$$

Thus we see that

$$\begin{aligned} M_h^{TLA}(\alpha\,\phi_m + \beta\,\phi_{N-m}) &= \beta\, C_m^{2q+2} S_m^{2p-2}\phi_m + \alpha S_m^{2q+2} C_m^{2p-2}\phi_{N-m} \\ &= \alpha'\,\phi_m + \beta'\,\phi_{N-m}, \end{aligned}$$

which can be denoted in matrix notation by

$$\begin{pmatrix} \alpha' \\ \beta' \end{pmatrix} = \begin{pmatrix} 0 & C_m^{2q+2}\, S_m^{2p-2} \\ S_m^{2q+2}\, C_m^{2p-2} & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

and we write therefore

$$_mM_h^{TLA} = \begin{pmatrix} 0 & C_m^{2q+2} S_m^{2p-2} \\ S_m^{2q+2} C_m^{2p-2} & 0 \end{pmatrix}.$$

Analogously we find for the amplification operator of the residual of theTLA-algorithm

$$ {}_m\overline{M}_h^{TLA} = \begin{pmatrix} 0 & S_m^{2p} C_m^{2q} \\ C_m^{2p} S_m^{2q} & 0 \end{pmatrix} . $$

The matrices ${}_m M_h^{TLA}$ and ${}_m\overline{M}_h^{TLA}$ have eigenvalues $\lambda_{1,2} = \pm (C_m S_m)^{p+q}$, thus

$$ \rho(M_h^{TLA}) = \rho(\overline{M}_h^{TLA}) \ \le\ 2^{-(p+q)} . $$

We now will determine the spectral norm of $M_h^{TLA}$ in order to know by what factor the norm of the error decreases by a single TLA iteration step. To this end it is convenient to have the following lemma available.

**Lemma 4.1**

$$ \max_{x\in[0,1]} x^a (1-x)^b = \frac{a^a b^b}{(a+b)^{a+b}} =: f(a,b) $$

**Proof:** The proofs follows by straight forward computation ∎

We will use the following equality

$$ \|M_h^{TLA}\| = \max_{m=1,2,\ldots,N/2} \sqrt{\max_{1,2} |\lambda_{1,2}(({}_m M_h^{TLA})^T ({}_m M_h^{TLA}))|} . $$

We easily compute

$$ ({}_m M_h^{TLA})^T ({}_m M_h^{TLA}) = \begin{pmatrix} S_m^{4q+4} C_m^{4p-4} & 0 \\ 0 & C_m^{4q+4} S_m^{4p-4} \end{pmatrix} ; $$

Hence,

$$ \begin{aligned} \|M_h^{TLA}\| &= \max_m \{ S_m^{2q+2} C_m^{2p-2}, C_m^{2q+2} S_m^{2p-2} \} \\ &\le \max_{t\in[0,1]} t^{q+1}(1-t)^{p-1} \\ &= f(q+1,p-1) = \frac{(q+1)^{q+1}(p-1)^{p-1}}{(p+q)^{p+q}} . \end{aligned} $$

And we conclude

if $p=0$ then $\lim_{h\to 0} \|M_h^{TLA}\| \to \infty$ ,
if $p=1$ then $\lim_{h\to 0} \|M_h^{TLA}\| = 1$ ,
if $p>1$ then $\lim_{h\to 0} \|M_h^{TLA}\| \le f(q+1,p-1) \ <\ 1$ .

Now the same can be done for $\overline{M}_h^{TLA}$ :

$$ \|\overline{M}_h^{TLA}\| = \max_{m=1,2,\ldots,N/2} \sqrt{\max_{1,2} |\, \lambda_{1,2}(({}_m\overline{M}_h^{TLA})^T ({}_m\overline{M}_h^{TLA}))\, |} . $$

Since

$$ ({}_m\overline{M}_h^{TLA})^T ({}_m\overline{M}_h^{TLA}) = \begin{pmatrix} S_m^{4q} C_m^{4p} & 0 \\ 0 & C_m^{4q} S_m^{4p} \end{pmatrix} $$

it follows that

$$ \begin{aligned} \|\overline{M}_h^{TLA}\| &= \max_m \{ S_m^{2q} C_m^{2p}, C_m^{2q} S_m^{2p} \} \\ &\le \max_{t\in[0,1]} t^q (1-t)^p \\ &= f(q,p) = \frac{q^q\, p^p}{(p+q)^{p+q}} . \end{aligned} $$

And we conclude

$$\begin{array}{lll} \text{if } p=0 & \text{then} & \lim_{h\to 0} \|\bar{M}_h^{TLA}\| = 1 \ , \\ \text{if } q=0 & \text{then} & \lim_{h\to 0} \|\bar{M}_h^{TLA}\| = 1 \ , \\ \text{if } p,q \neq 0 & \text{then} & \lim_{h\to 0} \|\bar{M}_h^{TLA}\| \le f(q,p) \ < \ 1 \ . \end{array}$$

In summary

$$\rho(M_h^{TLA}) = \rho(\overline{M}_h^{TLA}) \ \le \ 2^{-(p+q)}$$

$$\begin{array}{lll} \|M_h^{TLA}\| \to \infty & \text{for } h \to 0 & \text{if } p=0 \\ \|M_h^{TLA}\| \to 1 & & \text{if } p=1 \\ \|M_h^{TLA}\| \le f(q+1,p-1) < 1 & & \text{if } p>1 \end{array}$$

$$\begin{array}{lll} \|\bar{M}_h^{TLA}\| \to 1 & \text{for } h \to 0 & \text{if } p=0 \\ \|\bar{M}_h^{TLA}\| \to 1 & & \text{if } q=0 \\ \|\bar{M}_h^{TLA}\| \le f(q,p) < 1 & & \text{if } p,q \neq 0 \end{array}$$

If we apply the TLA with $\bar{R}^*_{Hh}$ instead of $\bar{R}_{Hh}$ we get

$$P_{hH} L_H^{-1} \bar{R}^*_{Hh} \phi_m =$$
$$= P_{hH} L_H^{-1} \tfrac{1}{4} \cdot \begin{pmatrix} \sin(m\pi h) + 2\sin(2m\pi h) + \sin(3m\pi h) \\ \sin(3m\pi h) + 2\sin(4m\pi h) + \sin(5m\pi h) \\ \vdots \\ \sin((2k-1)m\pi h) + 2\sin(2km\pi h) + \sin((2k+1)m\pi h) \\ \vdots \\ \sin((N-3)m\pi h) + 2\sin((N-2)m\pi h) + \sin((N-1)m\pi h) \end{pmatrix}$$

Using

$$\sin((2k-1)\alpha) + \sin(2k\alpha) + \sin(2k\alpha) + \sin((2k+1)\alpha)$$
$$= \ 2\ \sin(\frac{1}{2}(4k-1)\alpha)\cos(\alpha/2) + 2\sin((4k+1)\alpha/2)\cos(\alpha/2)$$
$$= \ 4\ \sin(2k\alpha)\cos^2(\alpha/2)\,,$$

it is not difficult to see that the above expression equals

$$\begin{aligned} & P_{hH} L_H^{-1} \tfrac{1}{4} \begin{pmatrix} 4\cos^2(m\pi h/2)\sin(2m\pi h) \\ 4\cos^2(m\pi h/2)\sin(4m\pi h) \\ \vdots \\ 4\cos^2(m\pi h/2)\sin(2(N/2-1)m\pi h) \end{pmatrix} = \\ & = \ C_m^2 P_{hH} L_H^{-1} \begin{pmatrix} \sin(m\pi H) \\ \sin(2m\pi H) \\ \vdots \\ \sin((N/2-1)m\pi H) \end{pmatrix} = \\ & = \ \frac{h^2(C_m^2\phi_m - S_m^2\phi_{N-m})}{4S_m^2} \end{aligned} \tag{4.17}$$

(cf. the computation which led to (4.16)). Thus we find

$$M_h^{CGC*}\phi_m = \phi_m - \frac{h^2(C_m^2\phi_m - S_m^2\phi_{N-m})}{4S_m^2} \cdot \frac{4S_m^2}{h^2} =$$

$$= S_m^2 \phi_m + S_m^2 \phi_{N-m} \, ,$$

and

$$M_h^{CGC*} \phi_{N-m} = C_m^2 \phi_m + C_m^2 \phi_{N-m} \, .$$

This can be written in matrix form again as ${}_mM_h^{CGC*} = \begin{pmatrix} S_m^2 & C_m^2 \\ S_m^2 & C_m^2 \end{pmatrix}$ because

$$M_h^{CGC*}(\alpha\phi_m + \beta\phi_{N-m}) = (\alpha S_m^2 + \beta C_m^2)\phi_m + (\alpha S_m^2 + \beta C_m^2)\phi_{N-m}$$

$$= \alpha'\phi_m + \beta'\phi_{N-m} \, ,$$

with

$$\begin{pmatrix} \alpha' \\ \beta' \end{pmatrix} = {}_mM_h^{CGC*} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} .$$

Using the fact that ${}_mM_h^{DJM} = \begin{pmatrix} C_m^2 & 0 \\ 0 & S_m^2 \end{pmatrix}$ we find

$${}_mM_h^{TLA*} = (\,{}_mM_h^{DJM})^q (\,{}_mM_h^{CGC*})(\,{}_mM_h^{DJM})^p$$

$$= \begin{pmatrix} C_m^{2p+2q} S_m^2 & C_m^{2q+2} S_m^{2p} \\ C_m^{2p} S_m^{2q+2} & C_m^2 S_m^{2p+2q} \end{pmatrix}$$

Similarly we find for the amplification operator of the residual

$$\begin{pmatrix} \alpha' \\ \beta' \end{pmatrix} = {}_m\overline{M}_h^{CGC*} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} S_m^2 & S_m^2 \\ C_m^2 & C_m^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} .$$

The effect of the completeTLA-algorithm on low- and high-frequency components of the residual is now described by

$${}_m\overline{M}_h^{TLA*} = \begin{pmatrix} C_m^2 & 0 \\ 0 & S_m^2 \end{pmatrix}^q \begin{pmatrix} S_m^2 & S_m^2 \\ C_m^2 & C_m^2 \end{pmatrix} \begin{pmatrix} C_m^2 & 0 \\ 0 & S_m^2 \end{pmatrix}^p$$

$$= \begin{pmatrix} S_m^2 C_m^{2p+2q} & C_m^{2q} S_m^{2p+2} \\ S_m^{2q} C_m^{2p+2} & C_m^2 S_m^{2p+2q} \end{pmatrix} .$$

The eigenvalues of both ${}_mM_h^{TLA*}$ and ${}_m\overline{M}_h^{TLA*}$ are 0 and $S_m^2 C_m^{2p+2q} + C_m^2 S_m^{2p+2q}$. Therefore we have for the spectral radius of $M_h^{TLA*}$ and $\overline{M}_h^{TLA*}$,

$$\rho(M_h^{TLA*}) = \rho(\overline{M}_h^{TLA*}) = \max_m |S_m^2 C_m^{2p+2q} + C_m^2 S_m^{2p+2q}|$$

$$\leq \max_{t\in[0,1]} (t(1-t)^{p+q} + (1-t)t^{p+q}).$$

Using Lemma 4.1 we find that

$$\rho(M_h^{TLA*}) = \rho(\overline{M}_h^{TLA*}) \leq 2\, f(1, p+q)$$

and thus

$$\rho(M_h^{TLA*}) = \rho(\overline{M}_h^{TLA*}) \begin{array}{ll} \leq 1 & \text{if } p+q = 0 (\text{ without the lemma}), \\ \leq 1/2 & \text{if } p+q = 1, \\ \leq \frac{8}{27} & \text{if } p+q = 2. \end{array}$$

To compute $\rho(M_h^{TLA*}) = \rho(\overline{M}_h^{TLA*})$ for large $p+q$ we first notice that for large $n$

$$\max_{t\in[0,1]}((1-t)t^n) = f(1,n) = \frac{n^n}{(n+1)^{n+1}} = \frac{1}{n}(1-\frac{1}{n+1})^{n+1} \approx \frac{1}{ne},$$

which shows that

$$\rho(M_h^{TLA*}) = \rho(\overline{M}_h^{TLA*}) \approx \frac{1}{(p+q)e} \text{ for } (p+q)\to\infty .$$

This describes the convergence rate of the two level algorithm for large $p+q$.

In order to see what is the effect of a *single* TLA iteration step on the residual we have to study $\|\overline{M}_h^{TLA*}\|$, the spectral norm of the amplification operator of the residual.

$$\|\overline{M}_h^{TLA*}\| = \max_{m=1,2,\dots,N/2} \sqrt{\max_{1,2} |\, \lambda_{1,2}((\,_m\overline{M}_h^{TLA*})^T(\,_m\overline{M}_h^{TLA*}))\,|}$$

$$= \max_m \sqrt{S_m^4 C_m^{4p+4q} + S_m^{4q}C_m^{4p+4} + C_m^{4q}S_m^{4p+4} + C_m^4 S_m^{4p+4q}}$$

$$= \max_m \sqrt{(S_m^4 C_m^{4q} + C_m^4 S_m^{4q})(C_m^{4p} + S_m^{4p})}.$$

Hence,

$$\limsup_{h\to 0} \|\overline{M}_h^{TLA*}\| \approx$$

$$\max_{0\le t\le 1} \sqrt{t^2(1-t)^{2q} + (1-t)^2 t^{2q}} \cdot \sqrt{(1-t)^{2p} + t^{2p}}.$$

If $p=q=0$ we find

$$\lim_{h\to 0} \|\overline{M}_h^{TLA*}\| \approx \max_{0\le t\le 1} \sqrt{t^2 + (1-t)^2}\,\sqrt{2} = \sqrt{2}.$$

If $q=0, p\neq 0$

$$\lim_{h\to 0} \|\overline{M}_h^{TLA*}\| =$$

$$\max_{0\le t\le 1} \sqrt{t^2+(1-t)^2}\,\sqrt{(1-t)^{2p}+t^{2p}} = 1, \quad (t=0).$$

If $p=0, q\neq 0$

$$\lim_{h\to 0} \|\overline{M}_h^{TLA*}\| = \sqrt{2} \max_{0\le t\le 1} \sqrt{t^2(1-t)^{2q} + (1-t)^2 t^{2q}},$$

and by lemma 4.1 this yields $1/2$ for $q=1$, and for $q\to\infty$ this yields

$$\sqrt{2} \max_{0\le t\le 1} \sqrt{t^2(1-t)^{2q}} = \sqrt{2} \max_{t\in[0,1]} (t(1-t)^q) =$$

$$= \sqrt{2}\,\frac{q^q}{(q+1)^{q+1}} \to \frac{\sqrt{2}}{qe}.$$

If $p,q>0$

$$\|\overline{M}_h^{TLA*}\| \approx \max_m \sqrt{S_m^4 C_m^{4q} C_m^{4p}} = \max_m (S_m^2 C_m^{2p+2q}),$$

$$\lim_{h\to 0} \|\overline{M}_h^{TLA*}\| = \max_{0\le t\le 1} t(1-t)^{p+q} = \frac{1}{p+q}\left(\frac{p+q}{p+q+1}\right)^{p+q+1} \to \frac{1}{e(p+q)}.$$

for $p+q\to\infty$.

Summarizing we see

$$
\begin{array}{lll}
\rho(M_h^{TLA*}) = \rho(\overline{M}_h^{TLA*}) & \to 1 \ \text{ for } h \to 0, & \text{if } p+q=0, \\
 & \to \frac{1}{2} & \text{if } p+q=1, \\
 & \to \frac{8}{27} & \text{if } p+q=2, \\
 & \to \frac{1}{e(p+q)} & \text{for } (p+q) \to \infty \cdot \\
 & & \\
\|\overline{M}_h^{TLA*}\| & \to \sqrt{2} \text{ for } h \to 0, & \text{if } p=0, q=0 \\
 & \to 1 & \text{if } p \neq 0, q=0, \\
 & \to \frac{1}{2} & \text{if } p=0, q=1, \\
 & \to \frac{\sqrt{2}}{qe} & \text{if } p=0, q \to \infty, \\
 & \to \frac{1}{e(p+q)} & \text{if } p \neq 0 \, (p+q) \to \infty.
\end{array}
$$

$\|M_h^{TLA*}\|$ yields the same values, with $p$ and $q$ interchanged.

**Remark:** In [9] we can see that the difference between the above values for finite values of $h$ and for $h \to 0$ is very small ($\mathcal{O}(h^2)$).

We conclude that relaxations *after* coarse grid corrections are of use for a small norm of the amplification operator of the *residual,* whereas relaxations *before* yield a small norm of the operator of the *error.* We noticed this already in a more theoretical context in chapter 3.

# Chapter 5

# Multigrid algorithms

## 5.1 Introduction

In the framework of defect correction processes multigrid algorithms are easy to explain. For this purpose we consider a continuous problem 2.1 and two discretisations of this problem on grids with meshwidth $h$ and $H$:

$$F_h z_h = y_h \quad \text{and} \quad F_H z_H = y_H,\ H > h. \tag{5.1}$$

The operators $F_h : D_h \subset E_h \to \hat{D}_h \subset \hat{E}_h$ and $F_H : D_H \subset E_H \to \hat{D}_H \subset \hat{E}_H$ are mappings between discrete spaces $E_h$, $\hat{E}_h$, $E_H$ and $\hat{E}_H$. Further a linear injection $P_{hH} : E_H \to E_h$ (*prolongation* or *interpolation*) and a linear surjection $R_{Hh} : \hat{E}_h \to \hat{E}_H$ (*restriction)* are given. A multigrid algorithm for the approximate solution of a discretised problem $F_h z_h = y_h$ is an iterative process in which one iteration cycle consists of

- $p$ (pre-) relaxation steps ($p \in I\!N$),
- a coarse grid correction step,
- $q$ (post-) relaxation steps ($q \in I\!N$).

The relaxation steps are defect correction steps as e.g. damped Jacobi (JOR), Gauss-Seidel (GS), symmetric Gauss-Seidel (SGS), incomplete LU-decomposition iteration (ILU) etc. Their main purpose is to reduce the unsmooth part of the error (see Section 3.3). The remaining, smooth error can be represented well on a coarser grid by means of some restriction.

The coarse grid correction step is a defect correction step of type DCPA where the approximate inverse $\tilde{G}_h$ is given by

$$\tilde{G}_h = P_{hH} F_H^{-1} \bar{R}_{Hh} \tag{5.2}$$

(see Figure 5.1). The use of the approximate inverse (5.2) implies that we solve the defect equation on a coarse grid, with the help of a coarse discretisation of our problem. This is only meaningful if (part of) the error before the defect correction step is representable on the coarse grid.

Thus far we have only described an algorithm using two grids. If we do not solve the defect equation on a coarser grid directly, but if we approximate its solution by application of a few iteration steps of the same algorithm on the coarser level, we obtain a recursive process where we have to solve a discretised problem directly only on the very coarsest grid. The resulting process is a true multigrid-algorithm. One complete iteration step in a multigrid process is called a multigrid *cycle.*

## 5.2 Two-level algorithms

Again we consider the two discretisations (5.1). Their relation via restrictions and prolongations is shown in Figure 5.1. We present the algorithms in the form of ALGOL-like programmes.

$$\begin{tikzcd} D_H \arrow[r, "F_h"] \arrow[d, shift right, "R_{Hh}"'] & \hat{D}_H \arrow[d, "\bar{R}_{Hh}"] \\ D_H \arrow[u, shift right, "P_{hH}"'] \arrow[r, "F_H"] & \hat{D}_H \end{tikzcd}$$

Figure 5.1: Relations between operators and spaces in a two-grid algorithm

First we describe two auxiliary procedures:

- **proc** solve = (**operator** $F_h$, **ref vector** $z_h$, **vector** $y_h$) **void:**
  This procedure uses the operator $F_h$ and the right-hand-side $y_h$ to solve (approximately) the equation
  $$F_h z_h = y_h \, .$$
  On entry, $z_h$ should contain an appropriate initial value for the (possibly) iterative solution process. On exit $z_h$ contains a (better approximate) solution of the problem $F_h z_h = y_h$.

- **proc** relax = (**operator** $F_h$, **ref vector** $z_h$, **vector** $y_h$) **void:**
  This procedure performs one iteration step of a suitable relaxation method for the equation $F_h z_h = y_h$.

Now we explain the essential coarse grid correction step. Given an approximation $z_h$ to the true solution $z_h^*$ of our discrete problem, we consider the residual

$$r_h = y_h - F_h z_h.$$

With the error $e_h = z_h - z_h^*$, we have

$$F_h(z_h - e_h) = y_h = r_h + F_h z_h. \tag{5.3}$$

For a linear operator $F_h$ this reduces to

$$F_h e_h = -r_h. \tag{5.4}$$

Instead of solving equation (5.3) directly, we compute the solution of a similar equation on a coarse grid

$$F_H w_H = \bar{R}_{Hh} r_h + F_H R_{Hh} z_h \tag{5.5}$$

and then use $P_{hH}(w_H - R_{Hh} z_h)$ as an approximation for the error $-e_h$, and hence as a correction quantity in the DCP. For a linear operator $F_h$ the above reduces to

$$F_H e_H = -\bar{R}_{Hh} r_h,$$

which gives the correction quantity $P_{hH} e_H$ . Now we describe one iteration step of the nonlinear two-level algorithm:

**proc** TGM = (**int** $p, q$, **ref vector** $z_h$, $y_h$) **void**:
**begin**
  **to** $p$ **do** relax $(F_h, z_h, y_h)$ ;
  **vector** $r_H := \bar{R}_{Hh}(y_h - F_h z_h) + F_H R_{Hh} z_h$;
  **vector** $w_H := R_{Hh} z_h$;
  solve $(F_H, w_H, r_H)$;
  $z_h := z_h + P_{hH}(w_H - R_{Hh} z_h)$;
  **to** $q$ **do** relax$(F_h, z_h, y_h)$;
**end**;

For a linear operator $F_h$ we can use the following simplified version:

**proc** TGML = (**int** $p, q$, **ref vector** $z_h, y_h$) **void**:
**begin**
  **to** $p$ **do** relax $(F_h, z_h, y_h)$;
  **vector** $r_H := \bar{R}_{Hh}(y_h - F_h z_h)$;
  **vector** $w_H := 0$;
  solve $(F_H, w_H, r_H)$;
  $z_h := z_h + P_{hH} w_H$;
  **to** $q$ **do** relax $(F_h, z_h, y_h)$;
**end**;

**Remark:** The coarse-grid correction in the two-level algorithm can easily be seen as an example of a DCPN as treated in Section 1.4.5. Therefore we take in (DCPN): $F := F_h$; $z := u_h$; $\tilde{G} := P_{hH} F_H^{-1} \overline{R}_{Hh}$ and $\overline{y} := \overline{P}_{hH} F_H R_{Hh} u_h^{(i)}$ with $\overline{P}_{hH}$ such that $\overline{R}_{Hh} \overline{P}_{hH} = I_H$. Then we find $\tilde{G} y = P_{hH} R_{Hh} u_h^{(i)}$ and with $u_H^{\mathrm{old}} = R_{Hh} u_h^{(i)}$ and $u_H^{\mathrm{new}}$ the solution of

$$F_H u_H^{\mathrm{new}} = F_H u_H^{\mathrm{old}} - \overline{R}_{Hh}(F_h u_h^{(i)} - y_h)/\mu$$

the (DCPN) reads as

$$u_h^{(i+1)} = u_h^{(i)} + \mu P_{hH}(u_H^{\mathrm{new}} - u_H^{\mathrm{old}}) \ .$$

## 5.3 Multi-level algorithms

Consider a sequence of grids with meshwidths $h_i$, $h_{i-1} > h_i$, $i = 0, 1, 2, \ldots$. Often one uses $h_{i-1} = 2h_i$. We now describe one cycle of a multi-level algorithm to solve the problem $F_{h_i} z_{h_i} = r_{h_i}$. The algorithm uses a sequence of approximate solutions $z = [z_{h_i}, z_{h_{i-1}}, \ldots, z_{h_0}]$ and a sequence of right-hand-sides $r = [r_{h_i}, r_{h_{i-1}}, \ldots, r_{h_0}]$. At entrance these data are given only for the finest grid.

**proc** MGM = (**int** $i, p, q, \sigma$, **ref vector** $z, r$) **void**:
**begin**
  **operator** $F_h = F_{h_i}, F_H = F_{h_{i-1}}$;
  **vector** $z_h = z_{h_i}$, $y_h = r_{h_i}$;
  **to** $p$ **do** relax $(F_h, z_h, y_h)$;
  **vector** $r_H = r_{h_{i-1}} := \bar{R}_{Hh}(y_h - F_h z_h) + F_H R_{Hh} z_h$;
  **vector** $w_H = z_{h_{i-1}} := R_{Hh} z_h$;
  **if** $i = 0$
    **then** solve $(F_H, w_H, r_H)$

```
        else to σ do MGM (i − 1, p, q, σ, z, r)
    fi;
    z_h := z_h + P_hH (w_H − R_Hh z_h);
    to q do relax (F_h, z_h, y_h); end;
```

For a linear operator $F_h$ we have the following simplified version:

```
proc MGML = (int i, p, q, σ, ref [ ] vector z, r ) void:
begin
    operator F_h = F_{h_i}, F_H = F_{h_{i-1}} ;
    vector z_h = z_{h_i}, y_h = r_{h_i} ;
    to p do relax (F_h, z_h, y_h);
    vector r_H = r_{h_{i-1}} := R̄_Hh (y_h − F_h z_h);
    vector w_H = z_{h_{i-1}} := 0;
    if i = 0
        then solve (F_H, w_H, r_H)
        else to σ do MGML(i − 1, p, q, σ, z, r)
    fi;
    z_h := z_h + P_hH w_H;
    to q do relax (F_h, z_h, y_h);
end;
```

Multigrid methods based upon MGM, respectively MGML, are also know by the names FAS resp. CS. These names stand for *Full Approximation Storage scheme* and *Correction Storage scheme* respectively. As the names already indicate, with CS we only compute corrections on the coarser grid, whereas with FAS we compute approximate solutions on all levels. When the MGML-algorithm has converged we have $F_h z_h = y_h$; when the MGM-algorithm has converged we have in addition $z_{h_{i-1}} = R_{h_{i-1}h_i} z_{h_i}$, i.e. on the coarse grids we find the restriction of the solution on the fine grid.

**Theorem 5.1** Consider an application of MGM where $h_i/h_{i+1} = H/h$ is constant for all $i = 0, 1, 2, \ldots$. Let $d$ be the dimension of the grid, i.e. the number of space dimensions. If $\sigma < (H/h)^d$ then the total amount of work in a multigrid cycle is proportional to the amount of the work on the fine grid.

**Proof:** Let $W$ be the amount of computational work needed to perform relaxations, operator evaluations, restriction and prolongation on the finest grid. On every next coarser grid the number of nodal points is reduced by a factor $(h/H)^d$. Hence the amount of work on the coarser grid is reduced by the same factor. If we consider an infinite number of grids, the total amount of work is given by

$$W_{\mathrm{tot}} \leq \sum_{n=0}^{\infty} [\sigma (h/H)^d]^n W = \frac{W}{1 - \sigma (h/H)^d}.$$

The above series converges if and only if $\sigma < (H/h)^d$. This implies that $W_{\mathrm{tot}}$ is proportional to $W$ if $\sigma$ is sufficiently small. ■

In the above multigrid algorithms (MGM, MGML) the fixed numbers $p, q$ and $\sigma$ determine the *strategy* of the algorithm. Other multigrid algorithms may terminate iterations sooner or later, depending on the convergence or other conditions that can be checked during the computation. Multigrid algorithms that make use of this possibility have an *adaptive strategy;* algorithms where the iterations are controlled only by the fixed numbers $p, q$ and $\sigma$ have a *fixed strategy.* MG-cycles with $\sigma = 1$ are called V-cycles, those with $\sigma = 2$ are called W-cycles. V-cycles that have either $p = 0$ or $q = 0$ are called *sawtooth cycles.*

In the following figures we show for some fixed strategies how is switched between the different levels of discretisation. We see that -essentially- most relaxation sweeps are performed on the coarser levels. In all diagrams the number of the levels is 4, the coarsest level is denoted by 0. Segments between tick-marks on a level $> 0$ denote the execution of a relaxation step on this level; a segment on level 0 denotes the direct solution on the coarsest level. Let

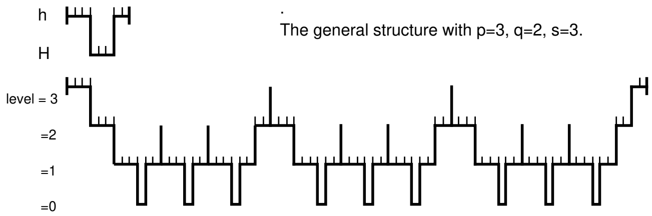


Figure 5.2: A general multigrid cycle with $p = 3, q = 2, \sigma = 3$

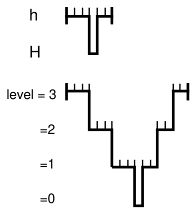


Figure 5.3: A V-cycle with $p = 3, q = 2$.

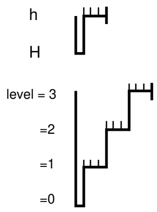


Figure 5.4: A saw-tooth cycle with $p = 0, q = 3, \sigma = 1$.

$M_h^{REL}$ and $\overline{M}_h^{REL}$ denote the amplification operators of the relaxation for error and residual respectively. For TGML the amplification operator for the error is given by

$$M_h^{TGML,p,q} = (M_h^{REL})^q (I_h \; - \; P_{hH} F_H^{-1} \bar{R}_{Hh} F_h)(M_h^{REL})^p$$

$$= (M_h^{REL})^q (F_h^{-1} \; - \; P_{hH} F_H^{-1} \bar{R}_{Hh})(\overline{M}_h^{REL})^p F_h .$$

We denote the amplification operator of a multi-level iteration step (MGML) on the $h$-level of discretisation by $M_h^{MGML,p,q,\sigma}$, or $M_h^{MGML}$ for short. The same amplification operator on the next coarser level we denote by $M_H^{MGML,p,q,\sigma}$ or $M_H^{MGML}$. In the multigrid cycle the approximate inverse is not given by

$$P_{hH} F_H^{-1} \bar{R}_{Hh} ,$$

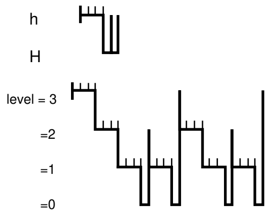


Figure 5.5: A W-cycle with $p = 3, q = 0, \sigma = 2$

because $F_H^{-1}$ is approximated by application of $\sigma$ steps of a DCP. The amplification operator of this DCP is given by $M_H^{MGML}$. Hence the approximate inverse of the $\sigma$ iteration steps together is given by

$$(I_H \, - \, (M_H^{MGML})^\sigma) F_H^{-1}$$

(see Section 1.4.2). Consequently, the amplification operator of the coarse grid correction inMGML is

$$I_h \, - \, P_{hH}(I_H \, - \, (M_H^{MGML})^\sigma) F_H^{-1} \bar{R}_{Hh} F_h$$

and we have

$$\begin{aligned} M_h^{MGML,p,q,\sigma} &= (M_h^{REL})^q (I_h \, - \, P_{hH}(I_H \, - \, (M_H^{MGML})^\sigma) F_H^{-1} \bar{R}_{Hh} F_h)(M_h^{REL})^p \\ &= M_h^{TGML} \, + \, (M_h^{REL})^q P_{hH} (M_H^{MGML})^\sigma F_H^{-1} \bar{R}_{Hh} F_h (M_h^{REL})^p \, . \end{aligned} \tag{5.6}$$

## 5.4 Full multigrid method (FMG)

The multigrid cycles we described in the previous section yield iterative improvement of a solution on a fine grid and therefore they need some initial estimate of the solution on this finest grid to start with. One possible algorithm is to obtain the initial estimate by interpolation from a solution on the next coarser grid, which has previously been calculated by a similar algorithm. Using this algorithm we start solving the problem on the coarsest grid. An algorithm of this type is called a *Nested Iteration* process:

**proc** nested iteration = (**int** $l$,[ ] **int** $i$, **ref** [ ] **vector** $z, y$) **void**:
**begin**
  solve $(F_0, z[0], y[0])$; { – sufficiently accurate – }
  **for** $k$ **to** $l$
  **begin**
    $z[k] := P^*_{k,k-1} z[k-1]$;
    **for** $m$ **to** $i[k]$ iteration $(k, z[k], y[k])$;
  **end**
**end**

**Remarks:**

- $l + 1$ is the number of levels available; the coarsest level is denoted by 0.
- $y[k]$ is the right-hand-side for the equation to be solved on level $k$.
- $i[k]$ denotes the number of iteration steps needed on level $k$.

- $P^*_{k,k-1}$ is an interpolation from level $k-1$ to level $k$. This operator is not necessarily the same as the prolongation operator used in the the multigrid cycles; it is usually more accurate.

The procedure "iteration" represents one step of a suitable iterative solution process. In the multigrid context we will use MGM (FAS) or MGML (CS) to replace "iteration". In this case we call the resulting method a *full multigrid method* (FMG). In an FMG-algorithm the coarse grids have a double function:

- providing the iterative process with an initial estimate;
- speeding up the process on a finer grid.

A typical FMG algorithm with one V-cycle ($p=\nu_1, q=\nu_2, \sigma=1$) per coarse grid, is shown in figure 5.4.1. (i.e. we have $i[k]=1$ for all $k$).

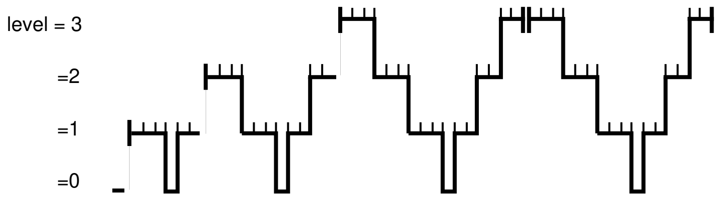


Figure 5.6: A FMG starting phase and an additional V-cycle

The strength of the FMG method is the combination of the sufficiently accurate initial estimate obtained by interpolation from the coarse grids, together with the convergence rate of the MG cycling procedure that is independent of the meshwidth. This combination makes that finally a strategy is found by which the solution of the discrete system on the finest grid is obtained with an accuracy that is of the same order of magnitude as the truncation error for this finest grid, by an amount of work that is only direct proportional to the number of degrees of freedom in the finest grid. Thus, the solution of the continuous problem is approximated up to truncation error accuracy by an amount of work that is proportional to the number of points in the finest grid. This is shown in the following theorem.

**Theorem 5.2** If we consider a Nested Iteration process with a sequence of discretisations $F_k z_k = y_k$, $k=0,1,2,\dots$ ,

$$F_k = F_{h_k}, \quad z_k = z_{h_k}, \quad y_k = y_{h_k}, \qquad h_0 > h_1 > h_2 > \dots , \tag{5.7}$$

such that (i) $h_{k-1}/h_k \le C_1$ for all $k$, (ii) the discretisations $F_k z_k = y_k$ and $F_{k-1}z_{k-1} = y_{k-1}$ are relatively convergent of order $p$, i.e.

$$\|P^*_{k,k-1} z_{k-1} - z_k\| \le C_0 h^p_{k-1} , \tag{5.8}$$

and (iii) the convergence factor of the iterative procedure "iteration" is independent of $h$, i.e.

$$\|z_k^{(j+1)} - z_k\| \le C_2 \|z_k^{(j)} - z_k\| ,$$

with $C_2$ independent of $h$. Then, *with* $i[k]=i$ *independent of* $k$, the result of the nested iteration procedure $\tilde{z}_k = z_k^{(i)}$ satisfies

$$\|\tilde{z}_k - z_k\|_{E_k} \le \frac{C_1^p C_2^i}{1 - C_1^p C_2^i \|P\|} C_0 h_k^p , \tag{5.9}$$

where $\|P\| = \sup_k \|P^*_{k,k-1}\|$ and it is assumed that $C_1^p C_2^i \|P\| \le 1$.
**Proof:** For each $k \ge 0$ we know

$$\|z_k^{(i)} - z_k\| \le C_2^i \, \|z_k^{(0)} - z_k\| \, .$$

In the nested iteration algorithm it is required that $z_0$ is computed sufficiently accurate. We require

$$\|\tilde{z}_0 - z_0\| \le C_1^p \; C_2^i \; C_0 h_0^p,$$

i.e. we require that truncation error accuracy has been attained. Then the statement of the theorem is satisfied for $k = 0$. For an arbitrary $k > 0$ we use induction. We assume that the theorem is satisfied for $k-1$, then

$$\begin{aligned} \|z_k^{(0)} - z_k\| &= \|P^*_{k,k-1} z_{k-1}^{(i)} - z_k\| \\ &\le \|P^*_{k,k-1}\| \; \|z_{k-1}^{(i)} - z_{k-1}\| + \|P^*_{k,k-1} z_{k-1} - z_k\| \\ &\le \|P\| \; \|z_{k-1}^{(i)} - z_{k-1}\| + C_0 h_{k-1}^p \end{aligned}$$

so that

$$\begin{aligned} \|z_k^{(i)} - z_k\| &\le C_2^i \; \|P\| \; \|z_{k-1}^{(i)} - z_{k-1}\| + C_2^i C_1^p C_0 h_k^p \\ &\le C_2^i C_1^p C_0 h_k^p + C_2^i \; \|P\| \; \{ C_2^i C_1^p C_0 h_{k-1}^p + C_2^i \; \|P\| \; \{...\}\} \\ &= C_0 C_1^p C_2^i \, [h_k^p + C_2^i \|P\| \{ h_{k-1}^p + C_2^i \|P\| \{...\}\}] \\ &= C_0 C_1^p C_2^i h_k^p \, [1 + C_2^i \|P\| \{ C_1^p + C_2^i \|P\| \{...\}\}] \\ &= C_0 C_1^p C_2^i h_k^p \, [1 + C_2^i \|P\| C_1^p + (C_2^i \|P\| C_1^p)^2 + ...] \\ &= \frac{C_1^p C_2^i}{1 - C_1^p C_2^i \|P\|} \, C_0 h_k^p \, . \end{aligned}$$

■

**Remark:** Notice that in the error estimate we recognise: $C_0 h_k^p$ the truncation error on level $k$, and $C_2$ the convergence factor of the iteration cycle. Usually, the mesh ratio $C_1 = 2$ and

$$\|P\| = \sup_H \|P_{hH}\| = \sup_H \sup_{x_H \in E_H} \frac{\|P_{hH} x_H\|_{E_h}}{\|x_H\|_{E_H}} = 1 \, .$$

**Remark:** We notice that our definition of relative convergence (Definition 2.29) was in terms of

$$\|P_{hH} L_H^{-1} \bar{R}_{Hh} - L_h^{-1}\| \le C H^p \, .$$

We recognise the equivalence with the assumption in the above theorem by

$$\begin{aligned} & \|(P_{hH} L_H^{-1} \bar{R}_{Hh} - L_h^{-1}) y_h\| \le C H^p \|y_h\| \\ \iff \; & \|P_{hH} L_H^{-1} \bar{R}_{Hh} y_h - L_h^{-1} y_h\| \le C H^p \|y_h\| \\ \iff \; & \|P_{hH} z_H - z_h\| \le \; C \; H^p \|y_h\| \\ \iff \; & \|P_{hH} z_H - z_h\| \le \; C \; H^p \frac{\|\bar{R}_H y\|}{\|y\|} \|y\| \end{aligned}$$

If $\bar{R}_H$ is bounded and stable then $c_1 \|y\| \le \|R_H y\| \le c_2 \|y\|$ and we recognise the equivalence immediately. We can derive the relative convergence of order $p$ between two discretisations also from the convergence of both. This is shown in the following theorem.

**Theorem 5.3** Consider the continuous equation $Fz = y$ and let the sequence of discretisations $F_h z_h = y_h$, $h \in \mathcal{H}$ have an order of convergence $p$. If there exists a nested sequence of stable prolongations

$$\{ P_h \mid P_h : E_h \to E, \; h \in \mathcal{H} \}$$

such that $P_h P_{hH} = P_H$ for $h, H \in \mathcal{H}$, $H > h$, then the discretisations $F_h z_h = y_h$ and $F_H z_H = y_H$ are relatively convergent of order $p$.
**Proof:**

$$\begin{aligned} \|P_{hH} z_H - z_h\| &\leq \|\hat{\hat{R}}_h P_h (P_{hH} z_H - z_h)\| \\ &\leq \|\hat{\hat{R}}_h\| \; \|P_h P_{hH} z_H - P_h z_h\| \\ &\leq \|\hat{\hat{R}}_h\| \; (\|P_H z_H - z\| + \|z - P_h z_h\|) \\ &\leq C(CH^p + Ch^p) \quad \leq CH^p \,. \end{aligned}$$

■

Associated with a nested iteration is a particular type of multigrid iteration cycle. this iteration cycle consists of a of consecutive V- or W-cycles on a sequence of finer and finer grids. This is similar to the inital phase of a FMG process. This iteration cycle is called an *F-cycle*. It can be considered as the repeated execution of the FMG starting phase, first applied to the solution and later on the corrections to the solution.

# Chapter 6

# Convergence proofs for multigrid methods

## 6.1 The structure of a TGM convergence proof (Hackbusch)

In order to construct a framework in which convergence of several multigrid algorithms can be proved we first give sufficient conditions for convergence of a two level algorithm (cf. [9]).

**Definition 6.1** Let $G_h$ be the approximate inverse related to some relaxation process for the solution of the equation $L_h u_h = f_h$. The relaxation process is said to have a *proper smoothing property of order* $\alpha$ if

$$\|L_h(I - G_h L_h)^\nu\| \leq C_0(\nu) h^{-\alpha} \tag{6.1}$$

with $C_0(\nu)$ independent of $h$, $C_0(\nu) \to 0$ as $\nu \to \infty$ . (This definition is a slightly simplified version of the one given in ([9])

**Remark:** Noticing that we have for the amplification operator of the error and residual of the relaxation respectively

$$M_h^{REL} = I_h - G_h L_h,$$

and

$$\overline{M}_h^{REL} = I_h - L_h G_h,$$

the property (6.1) can also be written as

$$\|L_h (M_h^{REL})^\nu\| \leq C_0(\nu) h^{-\alpha} \tag{6.2}$$

or

$$\|(\overline{M}_h^{REL})^\nu L_h\| \leq C_0(\nu) h^{-\alpha}. \tag{6.3}$$

**Example 6.2** *The model problem*
In Chapter 4 we saw for the example operator $L_h$ together with the damped Jacobi method as a relaxation method

$$\|L_h(I - G_h L_h)^\nu\| = \frac{4}{h^2} \sup_m \left| \begin{pmatrix} S_m^2 & 0 \\ 0 & C_m^2 \end{pmatrix} \begin{pmatrix} C_m^2 & 0 \\ 0 & S_m^2 \end{pmatrix}^\nu \right| =$$

$$= \frac{4}{h^2} \sup_m \left| \begin{pmatrix} C_m^{2\nu} S_m^2 & 0 \\ 0 & S_m^{2\nu} C_m^2 \end{pmatrix} \right| \le$$
$$\le \frac{4}{h^2} \max_{t \in [0,1]} t(1-t)^\nu \approx \frac{4}{h^2} \cdot \frac{1}{\nu e} \quad \text{for} \quad \nu \to \infty \,.$$

Hence, for this operator $L_h$, damped Jacobi has the proper smoothing property of order 2.

**Theorem 6.3** In a two level algorithm, if

- (i) the operators $L_h$ and $L_H$ are relatively convergent of order $\alpha$,
- (ii) the relaxation process for $L_h$ satisfies a proper smoothing property of order $\alpha$,
- (iii) the ratio $H/h$ is bounded, and
- (iv) $\|M_h^{REL}\| \le C$ uniformly in $h$,

then, for the error-amplification operator of a two-level cycle with $p$ pre- and $q$ post-relaxation steps, $M_h^{TLA,p,q}$, we have the estimate

$$\|M_h^{TLA,p,q}\| \le C \cdot C_0(p),$$

where $C$ is independent of $h$ and $C_0(p) \to 0$ as $p \to \infty$.
**Proof:**

$$\begin{aligned} M_h^{TLA,p,q} &= (M_h^{REL})^q \, M_h^{CGC} (M_h^{REL})^p \\ &= (M_h^{REL})^q \, (I_h - P_{hH} L_H^{-1} \bar{R}_{Hh} L_h)(M_h^{REL})^p \\ &= (M_h^{REL})^q \, (L_h^{-1} - P_{hH} L_H^{-1} \bar{R}_{Hh}) L_h (M_h^{REL})^p \end{aligned}$$

so

$$\|M_h^{TLA,p,q}\| \quad \le \quad \|M_h^{REL}\|^q \|L_h^{-1} - P_{hH} L_H^{-1} \bar{R}_{Hh}\| \; \|L_h (M_h^{REL})^p\| \,.$$

$$\begin{array}{ll} (i) \implies & \|L_h^{-1} - P_{hH} L_H^{-1} R_{Hh}\| \le C \cdot H^\alpha \,, \\ (ii) \implies & \|L_h (M_h^{REL})^p\| \le C_0(p) \cdot h^{-\alpha} \,, \\ (iv) \implies & \|M_h^{REL}\|^q \le C. \end{array}$$

Thus, by (iii) it directly follows that

$$\|M_h^{TLA,p,q}\| \le C \cdot (H/h)^\alpha \cdot C_0(p) \le C \cdot C_0(p)$$

■

**Theorem 6.4** If the conditions (i), (ii) and (iii) of Theorem 6.3 are satisfied, and instead of (iv) we have

$$\|\overline{M}_h^{REL}\| \le C \quad \text{uniformly for all} \quad h \,,$$

then for the residual-amplification operator $\overline{M}_h^{TLA,p,q}$ of the two-level algorithm with $p$ pre- and $q$ post-relaxation steps, we have

$$\|\overline{M}_h^{TLA,p,q}\| \le C \, C_0(q),$$

where $C$ is independent of $h$ and $C_0(q) \to 0$ as $q \to \infty$.
**Proof:** Completely analogous to the proof of Theorem 6.3. ■

**Remark:** Notice that the estimate for $\|M_h^{TLA,p,q}\|$ does not depend on $q$ whereas the estimate for $\|\overline{M}_h^{TLA,p,q}\|$ does not depend on $p$.

## 6.2 The structure of a MGM convergence proof (Hackbusch)

Whereas in the two-level algorithm we have to solve a coarse-grid problem in each iteration step, in the multi-level algorithm we solve this problem approximately, by the application of $\sigma$ iteration steps of the same multi-level algorithm on the coarse level. In other words: a multi-level algorithm is a two-level algorithm in which the application of $L_H^{-1}$ is replaced by $\sigma$ multi-level algorithm sweeps on the $H$-level; i.e. $L_H^{-1}$ is replaced by $[I-(M_H^{MGM})^\sigma]L_H^{-1}$, cf. Section 1.4. We will now give sufficient conditions for the convergence of a multi-level algorithm (cf. ([9]).

In the following we denote the error amplification operator of a multi-level iteration step on the $h$-level of discretisation by $M_h^{MLA}$, or - if we want to specify the strategy - we denote it by $M_h^{MLA,p,q,\sigma}$. The same operator on the next coarser level is denoted by $M_H^{MLA}$.

**Theorem 6.5** Suppose the conditions (i), (ii) and (iii) of Theorem 6.3 are satisfied and in addition

- (v) $P_{hH}$ is bounded and bounded from below, uniformly in $h$,
- (vi) $(M_h^{REL})^p$ and $(M_h^{REL})^q$ are bounded uniformly in $h$,
- (vii) $p$ is sufficiently large,

then
$$\|M_h^{MLA,p,q,\sigma}\| \le \|M_h^{TLA,p,q}\| + C\|M_H^{MLA}\|^\sigma .$$

**Proof:**
$$\begin{aligned}
&M_h^{MLA,p,q,\sigma}\\
&= (M_h^{REL})^q(I-P_{hH}(I-(M_H^{MLA})^\sigma)L_H^{-1}\bar R_{Hh}L_h)(M_h^{REL})^p\\
&= M_h^{TLA,p,q} + (M_h^{REL})^q P_{hH}(M_H^{MLA})^\sigma L_H^{-1}\bar R_{Hh}L_h(M_h^{REL})^p
\end{aligned}$$

so
$$\begin{aligned}
\|M_h^{MLA,p,q,\sigma}\| \le\ & \|M_h^{TLA,p,q}\| + \\
& + \|(M_h^{REL})^q\|\cdot\|P_{hH}\|\cdot\|M_H^{MLA}\|^\sigma\cdot\|L_H^{-1}\bar R_{Hh}L_h(M_h^{REL})^p\| .
\end{aligned} \tag{6.4}$$

Further
$$\begin{aligned}
- \quad & P_{hH}L_H^{-1}\bar R_{Hh}L_h(M_h^{REL})^p\\
&= (L_h^{-1} - P_{hH}L_H^{-1}\bar R_{Hh} - L_h^{-1})L_h(M_h^{REL})^p\\
&= (L_h^{-1} - P_{hH}L_H^{-1}\bar R_{Hh})L_h(M_h^{REL})^p - (M_h^{REL})^p\\
&= M_h^{TLA,p,0} - (M_h^{REL})^p.
\end{aligned}$$

Since $P_{hH}$ is bounded from below uniformly in $h$, according to (v), some $C>0$ exists, independent of $h$, such that
$$\|L_H^{-1}\bar R_{Hh}L_h(M_h^{REL})^p\| \le C\|M_h^{TLA,p,0} - (M_h^{REL})^p\|$$

Substituting this in (6.4) gives
$$\begin{aligned}
& \|M_h^{MLA,p,q,\sigma}\|\\
\le\ & \|M_h^{TLA,p,q}\| + C\|(M_h^{REL})^q\|\cdot\|P_{hH}\|\cdot\|M_H^{MLA}\|^\sigma\cdot\|M_h^{TLA,p,0} - (M_h^{REL})^p\|\\
\le\ & \|M_h^{TLA,p,q}\| + C\|(M_h^{REL})^q\|\cdot\|P_{hH}\|\cdot\|M_H^{MLA}\|^\sigma(\|M_h^{TLA,p,0}\| + \|(M_h^{REL})^p\|)
\end{aligned}$$

$$
\begin{array}{lll}
(i),(ii),(iii) & \Longrightarrow & \|M_h^{TLA,p,0}\| \le C\cdot C_0(p); \quad \text{cf. proof of Theorem6.3}\\
(v) & \Longrightarrow & \|P_{hH}\| \le C,\\
(vi) & \Longrightarrow & \|(M_h^{REL})^q\| \le C, \quad \|(M_h^{REL})^p\| \le C\\
(vii) & \Longrightarrow & C_0(p) \le C.
\end{array}
$$

Now it directly follows that

$$\|M_h^{MLA,p,q,\sigma}\| \le \|M_h^{TLA,p,q}\| + C\|M_H^{MLA}\|^\sigma.$$

■

We can obtain a similar result for the amplification operator of the residual:

**Theorem 6.6** Suppose the conditions (i), (ii) and (iii) of Theorem 6.3 are satisfied and in addition

- (viii) $\bar{R}_{Hh}$ is bounded and bounded from below uniformly in $h$,
- (ix) $(\overline{M}_h^{REL})^p$ and $(\overline{M}_h^{REL})^q$ are bounded uniformly in $h$,
- (x) $q$ is sufficiently large.

Then

$$\|\overline{M}_h^{MLA,p,q,\sigma}\| \le \|\overline{M}_h^{TLA,p,q}\| + C\|\overline{M}_H^{MLA}\|^\sigma.$$

**Proof:** The proof is similar to the proof of Theorem 6.5. ■

From the Theorem 6.3 and 6.4 in combination with the hypotheses of Theorem 6.5 and Theorem 6.6 we may immediately conclude that

$$\|M_h^{TLA,p,q}\| \le C\cdot C_0(p),$$

and

$$\|\overline{M}_h^{TLA,p,q}\| \le C\cdot C_0(q).$$

We now denote the sequence of discretisations used in the multilevel algorithm by

$$L_{h_k}u_{h_k} = f_{h_k}, \quad k = 0,1,2,...$$

with $|h_0| > |h_1| > |h_2| > ...$, or briefly by

$$L_k u_k = f_k, \quad k = 0,1,2,...$$

Here $k$ is the level of the discretisation. We define

$$\kappa_k = \|M_k^{TLA}\|, \quad k = 0,1,2,... \tag{6.5}$$

and

$$\tilde{\kappa}_k = \|M_k^{MLA}\|, \quad k = 0,1,2,... \tag{6.6}$$

If the conditions from Theorem 6.5 hold for all levels of the discretisation, then we can derive the following recursive relation for $\tilde{\kappa}_k$

$$\begin{cases} \tilde{\kappa}_1 = \kappa_1\,, \\ \tilde{\kappa}_k \le \kappa_k + C\tilde{\kappa}_{k-1}^\sigma, \end{cases} \tag{6.7}$$

where $\kappa_k \le C\cdot C_0(p)$ independent of $k$. It can be proved that for $C_0(p)$ small enough and $\sigma$ large enough (e.g. $\sigma = 2$), we have $\tilde{\kappa}_k \le C < 1$. A similar recursion can be obtained for $\|\overline{M}_k^{MLA}\|$.

For a concrete multi-level algorithm used for solving e.g. a boundary value problem, convergence can be proved with the help of the above theorems. Here we will not go into the details of verifying the hypotheses of the theorems.

## 6.3 Tools for a MGM convergence proof

In the following sections we will give a MGM convergence proof following Bank and Douglas [2]. In this section we introduce several basic notions necessary for this proof. Let $\mathcal{H}$ be a Hilbert space, $a(\cdot,\cdot)$: $\mathcal{H}\times\mathcal{H}\to I\!R$ be a continuous, symmetric, positive definite, bilinear form, and $g(\cdot):\mathcal{H}\to I\!R$ be a continuous linear functional. The *continuous problem* to be solved is now defined as:

Find $u\in\mathcal{H}$ such that

$$a(u,v)=g(v)\ \text{ for all } v\in\mathcal{H}. \tag{6.8}$$

We will look for approximate solutions in nested finite dimensional spaces:

$$M_1\subseteq M_2\subseteq M_3\subseteq ...\subseteq M_j\subset\mathcal{H},$$

$$\dim\ (M_k)=N_k.$$

The *discrete problem* now is formulated as:

Find $z\in M_j$ such that

$$a(z,v)=g(v)\ \text{ for all } v\in M_j. \tag{6.9}$$

Let the set $\{\omega_k|k=1,2,...,N_j\}$ form a basis for $M_j$, then (6.9) can be formulated as: find $z$ such that

$$a(z,\omega_k)=g(\omega_k)\quad\text{for}\quad k=1,2,...,N_j; \tag{6.10}$$

or equivalently

$$A\vec{z}=\vec{g} \tag{6.11}$$

where $A$ is the $N_j\times N_j$ matrix with $A_{i,k}=a(\omega_i,\omega_k)$; $\vec{g}$ is the $N_j$-vector with $g_i=g(\omega_i)$, and $\vec{z}$ is the unknown $N_j$-vector with the coefficients of $z$. The problem (6.11) will be solved by a multigrid method.

#### Relaxation (smoothing)

We now consider the smoothing procedure (the relaxation). Let $b_j(\cdot,\cdot):M_j\times M_j\to I\!R$ be some approximation to $a(\cdot,\cdot)$, the so called *smoother* associated with $M_j$, such that $b_j$ like $a$ is a continuous, symmetric, positive definite bilinear form. The corresponding approximation for the matrix $A$ is $B$ with

$$B_{ik}=b_j(\omega_i,\omega_k).$$

We furthermore assume that

$$\sup_{v\in M_j}\frac{a(v,v)}{b_j(v,v)}\le 1\,, \tag{6.12}$$

or equivalently

$$\sup_{\vec{v}\in I\!R^{N_j}}\frac{\vec{v}^TA\vec{v}}{\vec{v}^TB\vec{v}}\le 1. \tag{6.13}$$

**Example 6.7** *The model problem*

In our example from Chapter 4 we had

$$A=L_h=\frac{1}{h^2}\begin{pmatrix}2&-1&&\\-1&2&-1&\\&&\ddots&-1\\&&-1&2\end{pmatrix}$$

and this matrix was approximated by

$$B = \frac{1}{\omega} D_h = \frac{2}{\omega h^2} I,$$

for some damping factor $\omega$. We already know that $A$ has eigenvalues $\frac{4}{h^2} S_m^2$ and $B$ has only one eigenvalue, namely $\frac{2}{\omega h^2}$. Therefore,

$$\sup_{\vec{v} \in \mathbb{R}^{N_j}} \frac{\vec{v}^T A v}{\vec{v}^T B v} \leq \max_m 2\omega S_m^2 .$$

Thus, (6.12) is satisfied when $\omega \leq 1/2$.

**Definition 6.8** Let $S$ be a positive definite symmetric matrix with eigenvalue decomposition $S = \Phi \Lambda \Phi^{-1}$; $\Phi$ is the matrix with orthonormal eigenvectors of $S$ such that $\Phi\Phi^T = I$ and $\Lambda$ is the diagonal matrix with the eigenvalues of $S$.

We define a function $f$ on $S$ as follows:

$$f(S) = \Phi f(\Lambda) \Phi^{-1} .$$

Let us now consider the generalised eigenvalue problem

$$A\Psi = B\Psi\Lambda, \tag{6.14}$$

with $\Lambda$ (diagonal) and $\Psi$ unknown, $A$ and $B$ as before. If we take $\Phi$ such that

$$\Psi = B^{-1/2}\Phi ,$$

then (6.14) can be written as the symmmetric positive definite eigenvalue problem

$$(B^{-1/2} A B^{-1/2})\Phi = \Phi\Lambda. \tag{6.15}$$

It is clear that there exists an orthonormal system $\{\phi_i\}$ of eigenvectors of $B^{-1/2}AB^{-1/2}$ and a set of eigenvalues $\{\lambda_i\}$ with $\lambda_i > 0$, and thus $\Psi$ and $\Lambda$ can be determined. We have the following useful relations:

**Lemma 6.9**

$$\Psi^T A \Psi = \Lambda\, , \tag{6.16}$$

$$\Psi B \Psi = I\, , \tag{6.17}$$

$$0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_{N_j} \leq 1\, . \tag{6.18}$$

**Proof:**

$$\begin{aligned} a) \quad \Psi^T A \Psi &= (B^{-1/2}\Phi)^T A (B^{-1/2}\Phi) \\ &= \Phi^T B^{-1/2} A B^{-1/2} \Phi \\ &= \Phi^T \Phi \Lambda \\ &= \Lambda\, . \end{aligned}$$

$$\begin{aligned} b) \quad \Psi^T B \Psi &= (B^{-1/2}\Phi)^T B (B^{-1/2}\Phi) \\ &= \Phi^T B^{-1/2} B B^{-1/2} \Phi) \\ &= \Phi^T \Phi \\ &= I\, . \end{aligned}$$

c) Let $\psi_i$ be the $i$-th column of $\Psi$. hen, because of (6.16) and the fact that $A$ is positive definite, we have

$$0 \,<\, \lambda_i = \psi_i^T A \psi_i \,.$$

Using (6.13) and (6.17) we obtain

$$0 \,<\, \lambda_i = \psi_i^T A \psi_i \,\leq\, \psi_i^T B \psi_i = 1.$$

Arranging the $\lambda_i$'s gives

$$0 \,<\, \lambda_1 \,\leq\, \lambda_2 \,\leq \ldots. \leq\, \lambda_{N_j} \leq 1.$$

■

**Definition 6.10** On $M_j$ we define a norm $\|\cdot\|_\theta$, $\theta \in I\!R$, by

$$\begin{aligned} \|v\|_\theta^2 \;&= \vec{v}^T B^{1/2} (B^{-1/2} A B^{-1/2})^\theta B^{1/2} \vec{v} \\ &= \vec{v}^T B^{1/2} \Phi \Lambda^\theta \Phi^{-1} B^{1/2} \vec{v} \\ &= \vec{c}^T \Lambda^\theta \vec{c} \end{aligned} \tag{6.19}$$

with $\vec{v} = \Psi \vec{c}$ and $(\vec{v})_i$ such that $v = \sum_{i=1}^{N_j} v_i \omega_i$.

**Remark:** It can be easily verified (use $\lambda_i > 0$), that the above definition really defines a norm.

**Example 6.11** *The model problem*
We analyse the vector $\vec{c}$ in case of the example from Chapter 4. Again we take $\omega = 1/2$.

$$B = \frac{2}{\omega h^2} I = \frac{4}{h^2} I \quad \text{and hence} \quad B^{1/2} = \frac{2}{h} I \,,$$

$$\vec{c} = \Psi^{-1} \vec{v} = \Phi^T B^{1/2} \vec{v} = \frac{2}{h} \Phi^T \vec{v}. \tag{6.20}$$

If we denote the $i$-th column of $\Phi$ by $\phi_i$ then we find for the $i$-th component of $\vec{c}$:

$$c_i = \frac{2}{h} < \phi_i, \vec{v} > .$$

Thus, in this case the $i$-th component of the vector $c$ is a measure of the size of the component in the direction $\phi_i$ of $\vec{v}$.

As a consequence of the preceding definition we have the following

**Lemma 6.12**

$$\|v\|_0^2 = \vec{v}^T B \vec{v} \tag{6.21}$$

$$\|v\|_1^2 = \vec{v}^T A \vec{v} \tag{6.22}$$

$$|a(v,w)| \,\leq\, \|v\|_{1-\alpha} \|w\|_{1+\alpha}, \qquad \alpha \in I\!R. \tag{6.23}$$

**Proof:** (6.21) and (6.22) are trivial consequences of Definition 6.10. We will therefore prove only (6.23). If we write $v = \Sigma_i v_i \omega_i$, $w = \Sigma_i w_i \omega_i$ and denote the vectors with the coefficients $v_i$ and $w_i$ as usual by $\vec{v}$ respectively $\vec{w}$, we find

$$\begin{aligned} a(v,w) \;&= \vec{v}^T A \vec{w} \\ &= \vec{v}^T B^{1/2} \Phi \Lambda \Phi^{-1} B^{1/2} \vec{w}, \end{aligned}$$

the last equality following from (6.15).

If we write

$$\vec{c} = \Phi^{-1} B^{1/2} \vec{v} \quad (\text{or} \;\; \vec{v} = \Psi \vec{c}\,),$$
$$\vec{d} = \Phi^{-1} B^{1/2} \vec{w} \quad (\text{or} \;\; \vec{w} = \Psi \vec{d}\,),$$

we find

$$a(v,w) = \vec{c}^T \Lambda \vec{d} = \sum_i c_i \lambda_i d_i = \sum_i (c_i \lambda_i^{(1-\alpha)/2})\,(d_i \lambda_i^{(1+\alpha)/2}),$$

and with the use of the Cauchy-Schwartz inequality:

$$\begin{array}{rl} |a(v,w)| & \leq \sqrt{\sum_i (c_i \lambda_i^{(1-\alpha)/2})^2} \;\; \sqrt{\sum_i (d_i \lambda_i^{(1+\alpha)/2})^2} \\ & = \sqrt{\sum_i c_i \lambda_i^{1-\alpha} c_i} \;\; \sqrt{\sum_i d_i \lambda_i^{1+\alpha} d_i} \\ & = \sqrt{\vec{c} \Lambda^{1-\alpha} \vec{c}} \; \sqrt{\vec{d}^T \Lambda^{1+\alpha} \vec{d}} \\ & = \|v\|_{1-\alpha} \; \|w\|_{1+\alpha}. \end{array}$$

■

## 6.4 The effect of the smoothing operator

When solving problem (6.11) by a multigrid method, a smoothing operator will be applied several times. An approximate inverse $B^{-1}$ of $A$ is at our disposal. The smoothing operator is therefore defined as

$$S = I - B^{-1} A = B^{-1}(B - A). \tag{6.24}$$

We now want to analyse what happens to an error $e \in M_j$ when $S$ is applied.

**Theorem 6.13** Let $e \in M_j$, $\vec{c^e}$ such that $\vec{e} = \Psi \vec{c^e}$, (cf. equation 6.20) or equivalently: $\vec{c^e}$ such that $e = \sum_i c_i^e \, \psi_i$. With $S$ as in (6.24) and $\Lambda$ as in the preceding section we have

$$Se = \sum_i c_i^e (1 - \lambda_i)\, \psi_i. \tag{6.25}$$

**Proof:** Let $\vec{c}^S e$ be such that $S\vec{e} = \Psi \vec{c}^{Se}$, then

$$\begin{array}{rl} \vec{c}^{Se} & = \Psi^{-1} S \vec{e} \\ & = \Phi^{-1} B^{1/2} S \vec{e} \\ & = \Phi^{-1} B^{1/2} B^{-1} (B - A) \vec{e} \\ & = \Phi^{-1} B^{1/2} \vec{e} - \Phi^{-1} B^{-1/2} A \vec{e} \\ & = \Psi^{-1} \vec{e} - \Phi^{-1} B^{-1/2} A \Psi \vec{c^e} \\ & = \vec{c^e} - \Phi^{-1} B^{-1/2} A B^{-1/2} \Phi \vec{c^e} \\ & = [I - \Phi^{-1} B^{-1/2} A B^{-1/2} \Phi] \vec{c^e} \\ & = [I - \Lambda] \vec{c^e}\,. \end{array}$$

So $c_i^{Se} = (1 - \lambda_i) c_i^e$ and hence

$$Se = \sum_i c_i^{Se}\, \psi_i = \sum_i c_i^e (1 - \lambda_i)\, \psi_i\,.$$

■

As in Chapter 4, we see that components having high frequencies (large $\lambda_i$) are damped very well, whereas components having low frequencies (small $\lambda_i$) remain almost unchanged.

As a consequence of Theorem 6.13 we find

**Theorem 6.14** With $f$ as in lemma 4.1, $\alpha \in \mathbb{R}$, $p$ an integer we have

$$\|S^p e\|_\theta \leq f(p, \alpha/2) \ \ \|e\|_{\theta-\alpha}. \tag{6.26}$$

**Proof:** It follows from Theorem 6.13 that

$$S^p e = \sum_i c_i^e \, (1-\lambda_i)^p \, \psi_i \, ,$$

and thus

$$\vec{c}^{S^p e} = (I - \Lambda)^p \, \vec{c}^e \, .$$

By definition

$$\begin{aligned} \|S^p e\|_\theta^2 &= \vec{c}^{S^p e} \Lambda^\theta (\vec{c}^{S^p e})^T \\ &= \textstyle\sum_i (c_i^e)^2 (1-\lambda_i)^{2p} \lambda_i^\theta \\ &= \textstyle\sum_i (c_i^e)^2 \lambda_i^{\theta-\alpha} (1-\lambda_i)^{2p} \lambda_i^\alpha . \end{aligned}$$

By (6.18) $0 < \lambda_i \leq 1$ and therefore

$$\|S^p e\|_\theta^2 \leq \max_{0 \leq \lambda \leq 1} [(1-\lambda)^{2p} \lambda^\alpha] \sum_i (c_i^e)^2 \lambda_i^{\theta-\alpha} = f(2p, \alpha) \, \| \, e\|_{\theta-\alpha}^2 \, ,$$

or

$$\|S^p e\|_\theta \leq f(p, \alpha/2) \, \|e\|_{\theta-\alpha}.$$

■

**Example 6.15** *The model problem*

We now want to apply the above theory to our example from Chapter 4. Again we take $\omega = 1/2$.

$$A = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}, \quad B = \frac{4}{h^2} I$$

and hence

$$B^{-1/2} A B^{-1/2} = \frac{1}{4} \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & & \ddots & \\ & & -1 & 2 \end{pmatrix}$$

with eigenvalues $\lambda_m = S_m^2$ and corresponding eigenvectors

$$\phi_m = \begin{pmatrix} \sin(m\pi h) \\ \sin(2m\pi h) \\ \dots \\ \sin(N_j m\pi h) \end{pmatrix},$$

$$\Phi = (\phi_1, \phi_2, ..., \phi_{N_j}).$$

We have $\Phi = B^{1/2}\Psi$ and hence the following relation for the columns of $\Phi$ and $\Psi$ is valid:

$$\phi_m = \frac{2}{h}\psi_m .$$

Now we are able to compute $\|v\|_\theta$ and $\|S^p v\|_\theta$ for $v \in M_j$. Let $v = \Sigma_m \alpha_m \phi_m \in M_j$, then

$$v = \sum_m \alpha_m \frac{2}{h}\psi_m ,$$

and therefore (cf. Definition 6.10)

$$\begin{aligned}
\|v\|_\theta^2 &= \textstyle\sum_m (\alpha_m \frac{2}{h})^2 \lambda_m^\theta \\
&= \textstyle\sum_m \alpha_m^2 \frac{4}{h^2} S_m^{2\theta} . \\
\|v\|_\theta &= \textstyle\frac{2}{h}\sqrt{\sum_m \alpha_m^2 S_m^{2\theta}} .
\end{aligned}$$

In Chapter 4 we already observed that $S\phi_m = C_m^2 \phi_m$ and hence

$$\begin{aligned}
S^p v = \;& S^p \textstyle\sum_m \alpha_m \phi_m \\
&= \textstyle\sum_m \alpha_m S^p \phi_m \\
&= \textstyle\sum_m \alpha_m C_m^{2p} \phi_m \\
&= \textstyle\sum_m \alpha_m C_m^{2p} \cdot \frac{2}{h}\psi_m ,
\end{aligned}$$

thus

$$\begin{aligned}
\|S^p v\|_\theta^2 = \;& \textstyle\sum_m (\alpha_m C_m^{2p} \cdot \frac{2}{h})^2 \lambda_m^\theta \\
&= \textstyle\sum_m \alpha_m^2 C_m^{4p} \frac{4}{h^2} S_m^{2\theta} , \\
\|S^p v\|_\theta &= \textstyle\frac{2}{h}\sqrt{\sum_m \alpha_m^2 C_m^{4p} S_m^{2\theta}} .
\end{aligned}$$

## 6.5 The effect of the coarse grid correction

We will denote the amplification operator of the error associated with the coarse grid correction by

$$C : M_j \to M_j .$$

In a coarse grid correction, a smooth error $v^{\mathrm{old}} \in M_j$ is approximated in $M_{j-1}$. This approximation is called $v^{corr}$ and hence

$$v^{\mathrm{new}} - v^{\mathrm{old}} = -v^{corr} \in M_{j-1}$$

or

$$Cv^{\mathrm{old}} - v^{\mathrm{old}} \in M_{j-1} .$$

Because we consider a nested sequence of Galerkin discretisations (6.9), the result of the coarse grid correction is that the new error has no component left in $M_{j-1}$. Thus we have the following two properties of the coarse grid operator $C : M_j \to M_j$

$$Cv - v \in M_{j-1} \quad \text{for all} \quad v \in M_j . \tag{6.27}$$

$$a(Cv, w) = 0 \quad \text{for all} \quad w \in M_{j-1},\ v \in M_j . \tag{6.28}$$

**Remark:** Property (6.28) is also written as $Cv \in M_{j-1}^{\perp}$.

**Lemma 6.16** If C satisfies (6.27) and (6.28) then

$$C^2 = C.$$

**Proof:** If $v \in M_j$ then $Cv \in M_j$, and hence by (6.27)

$$CCv - Cv \in M_{j-1}.$$

Using (6.28) with $v = Cv - v$, $w = CCv - Cv$ we find

$$a(CCv - Cv, CCv - Cv) = 0,$$

or by (6.22)

$$\|CCv - Cv\|_{\theta=1} = 0.$$

Thus $CCv = Cv$ for all $v \in M_j$. ∎

From now on we will use the following notation

$$|||\cdot||| = \|\cdot\|_{\theta=1}\,.$$

**Lemma 6.17** Let C satisfy (6.27) and (6.28). If $\tilde{C}$ is such that

$$\tilde{C}v - v \in M_{j-1} \quad \text{for all} \quad v \in M_j,$$

and

$$|||\tilde{C}v - Cv||| \; < \; \gamma|||Cv - v||| \;\text{for some}\, \gamma \; > \; 0,$$

then

$$|||\tilde{C}v \; + \; (\gamma - 1)Cv||| \; \leq \; \gamma|||v|||$$

**Proof:** For all $v \in M_j$

$$\begin{aligned} &|||\tilde{C}v \; + \; (\gamma-1)Cv|||^2 = \\ &= a(\tilde{C}v + (\gamma-1)Cv\,,\, \tilde{C}v + (\gamma-1)Cv) = \\ &= a(\tilde{C}v - Cv\,,\, \tilde{C}v - Cv) + 2\gamma\, a(\tilde{C}v - Cv, Cv) + \gamma^2 a(Cv, Cv) = \\ &= |||\tilde{C}v - Cv|||^2 \; + \; \gamma^2|||Cv|||^2. \end{aligned}$$

The last equality being valid because

$$\tilde{C}v - Cv = (\tilde{C}v - v) - (Cv - v) \in M_{j-1}$$

and hence

$$a(\tilde{C}v - Cv, Cv) = 0.$$

Now the assumption that

$$|||\tilde{C}v - Cv||| \; < \; \gamma|||Cv - v|||$$

yields

$$|||\tilde{C}v + (\gamma-1)Cv|||^2 \; \leq \; \gamma^2(|||Cv - v|||^2 + |||Cv|||^2) \; = \; \gamma^2|||v|||^2.$$

(Note that $a(Cv - v, Cv) = 0$ and thus we may use Pythagoras' rule.) Taking the square root on both sides we obtain the desired result. ∎

From now on we will assume that the following hypothesis is satisfied:

**Hypothesis 6.18** For given $\alpha > 0$, and $\theta \in I\!R$, there exists some constant $\kappa \geq 1$ such that for all $v \in M_j \bigcap M_{j-1}^{\perp}$

$$\|v\|_{\theta-\alpha}^2 \leq \kappa^{\alpha}\|v\|_{\theta}^2 \quad \text{uniform for all } j. \tag{6.29}$$

**Remark:** If $\kappa$ can be chosen such that $\frac{1}{\kappa} \leq \lambda_i \leq 1$ for all $i$, uniform in $j$, then hypothesis 6.18 holds for this $\kappa$. For if we take $\vec{v} = \Psi\vec{c}$, then we have the following equivalences

$$\|v\|_{\theta-\alpha} \leq \kappa^{\alpha/2}\|v\|_{\theta} \iff$$

$$\|v\|_{\theta-\alpha}^2 \leq \kappa^{\alpha}\|v\|_{\theta}^2 \iff$$

$$\sum_i c_i^2\lambda_i^{\theta-\alpha} \leq \kappa^{\alpha}\sum_i c_i^2\lambda_i^{\theta}.$$

Hence, if $\frac{1}{\kappa} \leq \lambda_i \leq 1$ then $\kappa^{\alpha} \geq \lambda_i^{-\alpha}$ and (6.29) is valid. Notice that in practice there is often no uniform positive lower bound for $\lambda_i$ because $\lambda_i$ depends on $j$.

If for given $\theta$ and $\alpha$, a $\kappa$ exists such that (6.29) holds, then

$$\kappa \geq \sup_j \sup_{v \in M_j \cap M_{j-1}^{\perp}} \frac{\|v\|_{\theta-\alpha}^{2/\alpha}}{\|v\|_{\theta}^{2/\alpha}} \tag{6.30}$$

and vice versa; if for given $\theta$ and $\alpha$ the supremum (6.30) exists, then (6.29) holds with $\kappa$ as in (6.30).

**Lemma 6.19** If hypothesis 6.18 is true for $\alpha_0$ and $\theta_0$ then it is true for $\alpha$ and $\theta_0$ with $0 < \alpha \leq \alpha_0$.
**Proof:** This proof is left as an exercise. ∎

**Example 6.20** *The model problem*
For our example in Chapter 4 we will determine some $\kappa$ for $\theta = \alpha = 1$. To do this we have to know first which functions $v$ are in $M_j \cap M_{j-1}^{\perp}$. In Chapter 4 we saw that for the amplification operator of the error, $C = M_h^{CGC*}$ corresponding to the coarse grid correction with $\bar{R}_{Hh}^*$ instead of $\bar{R}_{Hh}$

$$C(\phi_m) = S_m^2\phi_m + S_m^2\phi_{N-m},$$

$$C(\phi_{N-m}) = C_m^2\phi_m + C_m^2\phi_{N-m},$$

hence

$$C(\phi_m + \phi_{N-m}) = \phi_m + \phi_{N-m},$$

$$C(C_m^2\phi_m - S_m^2\phi_{N-m}) = 0.$$

Using (6.27) and (6.28) we find

$$M_{j-1} \supseteq \text{Span}\,\{C_m^2\phi_m - S_m^2\phi_{N-m}\}$$

and

$$M_{j-1}^{\perp} \supseteq \text{Span}\,\{\phi_m + \phi_{N-m}\}.$$

Counting the dimensions of the spaces it follows that

$$M_{j-1} = \text{Span}\,\{C_m^2\phi_m - S_m^2\phi_{N-m}\}$$

and

$$M_{j-1}^{\perp} = \text{Span}\,\{\phi_m + \phi_{N-m}\}.$$

Now

$$\kappa = \sup_{v\in M_j\cap M_{j-1}^\perp} \frac{\|v\|_{\theta-\alpha}^{2/\alpha}}{\|v\|_\theta^{2/\alpha}} = \sup_{v\in M_{j-1}^\perp} \frac{\|v\|_0^2}{\|v\|_1^2}\,.$$

Let $v = \sum_m \alpha_m\phi_m$, then $\|v\|_\theta^2 = \frac{4}{h}\sum_m \alpha_m^2 S_m^{2\theta}$ (cf. example 6.15) and hence

$$\kappa = \sup_{v\in M_{j-1}^\perp} \frac{\frac{4}{h}\sum_m \alpha_m^2}{\frac{4}{h}\sum_m \alpha_m^2 S_m^2} = \sup_{v\in M_{j-1}^\perp} \frac{\sum_m \alpha_m^2}{\sum_m \alpha_m^2 S_m^2} =$$

$$= \sup_{v\in M_{j-1}^\perp} \frac{\sum_{m=1}^{N/2}(\alpha_m^2 S_m^2 + \alpha_{N-m}^2)}{\sum_{m=1}^{N/2}(\alpha_m^2 S_m^2 + \alpha_{N-m}^2 S_{N-m}^2)}\,,$$

the last equality being valid because $N$ is even (assumption from Chapter 4).

Since $M_{j-1}^\perp = \text{Span}\,\{\phi_m + \phi_{N-m}\}$ we have $\alpha_m = \alpha_{N-m}$ for $v = \Sigma_m\alpha_m\phi_m \in M_{j-1}^\perp$. Hence

$$\kappa = \sup_{(\alpha_1,\ldots,\alpha_{N/2})} \frac{\sum_{m=1}^{N/2}(\alpha_m^2 + \alpha_m^2)}{\sum_{m=1}^{N/2}(\alpha_m^2 S_m^2 + \alpha_m^2 C_m^2)} = \sup_{(\alpha_1,\ldots,\alpha_{N/2})} \frac{\sum_{m=1}^{N/2} 2\alpha_m^2}{\sum_{m=1}^{N/2}\alpha_m^2} = 2\,.$$

**Lemma 6.21** The validity of our hypothesis has the following consequences

$$\|Cv\|_1 \le \kappa^{\alpha/2}\|v\|_{1+\alpha} \tag{6.31}$$

$$\|CS^p v\|_1 \le \kappa^{\alpha/2} f(p,\alpha/2)\,\|v\|_1 \tag{6.32}$$

$$\|S^q Cv\|_1 \le \kappa^{\alpha/2} f(q,\alpha/2)\,\|Cv\|_1 \tag{6.33}$$

$$\|S^q Cv\|_{1+\alpha} \le \kappa^{\alpha/2} f(q,\alpha/2)\,\|v\|_{1+\alpha} \tag{6.34}$$

**Proof:**
(6.31): $\|Cv\|_1^2 = a(Cv,Cv) = a(Cv,Cv-v+v) = a(Cv,Cv-v) + a(Cv,v) = a(Cv,v)$. The last equality follows from (6.27) and (6.28). Using (6.23) and (6.29) ($\theta = 1$) yields

$$a(Cv,v) \le \|Cv\|_{1-\alpha}\|v\|_{1+\alpha} \le \kappa^{\alpha/2}\|Cv\|_1\|v\|_{1+\alpha}\,,$$

or summarising

$$\|Cv\|_1^2 \le \kappa^{\alpha/2}\|Cv\|_1\|v\|_{1+\alpha}.$$

Division by $\|Cv\|_1$ gives the desired result.
(6.32): From (6.31) follows

$$\|CS^p v\|_1 \le \kappa^{\alpha/2}\|S^p v\|_{1+\alpha}\,.$$

Now Theorem 6.14 gives

$$\|CS^p v\|_1 \le \kappa^{\alpha/2} f(p,\alpha/2)\|v\|_1.$$

(6.33): The successive use of Theorem 6.14 and our hypothesis with $\theta = 1$ yields

$$\|S^q Cv\|_1 \le f(q,\alpha/2)\|Cv\|_{1-\alpha} \le k^{\alpha/2} f(q,\alpha/2)\|Cv\|_1.$$

(6.34): Now we use Theorem 6.14 and (6.31)

$$\|S^q Cv\|_{1+\alpha} \le f(q,\alpha/2)\|Cv\|_1 \le \kappa^{\alpha/2} f(q,\alpha/2)\|v\|_{1+\alpha}.$$

■

## 6.6 The convergence proof for the TGM

In a two-level algorithm the error-amplification operator is $S^qCS^p$. For this operator we have the following

**Theorem 6.22** If the hypothesis (6.29) is true then

$$|||S^qCS^pv||| \ \leq \ \gamma|||v||| \text{ for all } v\in M_j, \tag{6.35}$$

where

$$\gamma = \inf_\alpha \ f(q,\alpha/2)\, f(p,\alpha/2)\, \kappa^\alpha,$$

the infimum taken over those $\alpha$ for which (6.29) is true.
**Proof:** The use of (6.33) and (6.34) results in

$$\begin{aligned} |||S^qCS^pv||| &\leq \kappa^{\alpha/2} f(q,\alpha/2)\, |||CS^pv||| \\ &\leq \kappa^{\alpha/2} f(q,\alpha/2)\kappa^{\alpha/2} f(p,\alpha/2)\, |||v||| \\ &\leq \gamma|||v|||\,. \end{aligned}$$

■

In the remainder of this chapter we restrict ourselves to $p = q$ and we write $m = p + q$. As a result of this restriction we have for $\gamma$ as in Theorem 6.22:

$$\gamma = \inf_\alpha\, [f(p,\alpha/2)]^2\, \kappa^\alpha = \inf_\alpha \ f(2p,\alpha)\, \kappa^\alpha = \inf_\alpha \ f(m,\alpha)\, \kappa^\alpha \cdot$$

**Lemma 6.23** If the hypothesis is true for $\kappa$ such that $\alpha = m/(\kappa-1)$, then the convergence factor $\gamma$ of the two level algorithm equals $((\kappa-1)/\kappa)^m$.
**Proof:** We compute $\gamma = \inf_\alpha \ f(m,\alpha)\, \kappa^\alpha$.

$$f(m,\alpha)\kappa^\alpha = \frac{\kappa^\alpha m^m \alpha^\alpha}{(m+\alpha)^{m+\alpha}}$$

$$\inf_\alpha \ \kappa^\alpha f(m,\alpha) = \exp\left(\inf_\alpha \ln \frac{\kappa^\alpha m^m \alpha^\alpha}{(m+\alpha)^{m+\alpha}}\right) =$$

$$\exp(\inf_\alpha\, (\alpha\, \ln \kappa + m\, \ln m + \alpha\, \ln \alpha - (m+\alpha)\, \ln(m+\alpha)))\,.$$

Now the infimum will be determined by computing the zero of the derivative of

$$g(\alpha) = \alpha \ln \kappa + m \ln m + \alpha \ln \alpha - (m+\alpha) \ln(m+\alpha).$$

$$g'(\alpha) = \ln \kappa + \ln \alpha + 1 - \ln(m+\alpha) - 1 = \ln \frac{\kappa\alpha}{m+\alpha}\cdot$$

$$g'(\alpha) = 0 \iff \frac{\kappa\alpha}{m+\alpha} = 1 \iff \alpha = \frac{m}{\kappa-1}\cdot$$

$$g''(\alpha) = \frac{1}{\alpha} - \frac{1}{m+\alpha} = \frac{m}{\alpha(\alpha+m)} \iff g''(\alpha) > 0 \ \text{ for } \ \alpha > 0\,.$$

Hence, $g$ takes its infimum at $\alpha = m/(\kappa-1)$.

We assumed the hypothesis being true for $k$ such that $\alpha = m/(k-1)$. Since the infimum has to be taken over those $\alpha$ for which (6.29) holds, we have

$$\gamma \ = \ \inf_\alpha \ f(m,\alpha)\kappa^\alpha = \frac{\left(\frac{\kappa^m}{\kappa-1}\right)^{\frac{m}{\kappa-1}} m^m}{\left(m + \frac{m}{\kappa-1}\right)^m + \frac{m}{\kappa-1}} = \left(\frac{\kappa-1}{\kappa}\right)^m.$$

■

## 6.7 The convergence proof for the W-cycle and the V-cycle

In the following theorem we introduce the operator $\tilde{C}$, an approximation of the coarse grid correction operator $C$. The condition on $\tilde{C}$ is such that $\tilde{C}$ may represent a W-cycle approximation (or a V-cycle approximation) to the coarse grid correction.

**Theorem 6.24** Suppose that the hypothesis (6.29) is true and that $\gamma$ in (6.35) satisfies $\gamma < \frac{1}{2}$. Let $\tilde{C}$ be such that

$$|||\tilde{C}v - Cv||| \leq \tilde{\gamma}^2 |||Cv - v||| \text{ for all } v \in M_j,$$

where

$$\frac{\gamma}{1-\gamma} \leq \tilde{\gamma} < 1.$$

Then

$$|||S^{m/2}\tilde{C}S^{m/2}v||| \leq \tilde{\gamma}|||v|||.$$

**Proof:**

$$\begin{aligned} |||S^{m/2}\tilde{C}S^{m/2}v||| &= |||S^{m/2}(1-\tilde{\gamma}^2)CS^{m/2}v + S^{m/2}(\tilde{C} + (\tilde{\gamma}^2 - 1)C)S^{m/2}v||| \\ &\leq (1-\tilde{\gamma}^2)\,|||S^{m/2}CS^{m/2}v||| + |||S^{m/2}(\tilde{C} + (\tilde{\gamma}^2-1)C)S^{m/2}v|||\,. \end{aligned} \tag{6.36}$$

Using Theorem 6.22 we have

$$|||S^{m/2}CS^{m/2}v||| \leq \gamma|||v|||\,.$$

Furthermore if we take $\alpha = 0$, $\theta = 1$ in Theorem 6.14 it follows that

$$|||S^{m/2}(\tilde{C} - (\tilde{\gamma}^2 - 1)C)S^{m/2}v||| \leq |||(\tilde{C} - (\tilde{\gamma}^2-1)C)S^{m/2}v|||\,.$$

In addition, application of Lemma 6.17 yields

$$|||(\tilde{C} - (\tilde{\gamma}^2 - 1)C)S^{m/2}v||| \leq \tilde{\gamma}^2|||S^{m/2}v|||\,,$$

and with Theorem 6.14 we find

$$\tilde{\gamma}^2|||S^{m/2}v||| \leq \tilde{\alpha}^2|||v|||\,.$$

Substitution of all this in (6.36) gives

$$|||S^{m/2}\tilde{C}S^{m/2}v||| \leq [(1-\tilde{\gamma}^2)\gamma + \tilde{\gamma}^2]\;|||v|||\,.$$

We assumed

$$\begin{gathered} \frac{\gamma}{1-\gamma} \leq \tilde{\gamma} < 1 \quad \Longleftrightarrow \\ \gamma \leq \tilde{\gamma} - \gamma\tilde{\gamma} \quad \Longleftrightarrow \\ \gamma(1+\tilde{\gamma})(1-\tilde{\gamma}) \leq \tilde{\gamma}(1-\tilde{\gamma}) \quad \Longleftrightarrow \\ \gamma - \gamma\tilde{\gamma}^2 \leq \tilde{\gamma} - \tilde{\gamma}^2 \quad \Longleftrightarrow \\ \tilde{\gamma}^2 + (1-\tilde{\gamma}^2)\gamma \leq \tilde{\gamma}\,. \end{gathered}$$

Hence

$$|||S^{m/2}\tilde{C}S^{m/2}v||| \leq \tilde{\gamma}|||v|||\,.$$

■

The reader may have noticed that it is not essential for $\tilde{C}$ to come from a W-cycle and that could as well come from a V-cycle. There is however a reason to consider the above as a convergence proof for the W-cycle. In a W-cycle the approximation $\tilde{C}$ of $C$ consists of a double application of an approximation $\hat{C}$ of $C$; $\hat{C}$ coming from a V-cycle. If we now assume that the quality of $\tilde{C}$ as an approximation of $C$ can be expressed as:

$$|||\hat{C}v - Cv||| \le \tilde{\gamma}|||Cv - v|||,$$

then the quality of $\tilde{C}$ as an approximation of $C$ is given by

$$\begin{aligned} |||\tilde{C}v - Cv||| &= |||\hat{C}\hat{C}v - Cv||| \\ &\le \tilde{\gamma}|||\hat{C}v - Cv||| \\ &\le \tilde{\gamma}^2|||v - Cv|||, \end{aligned}$$

which is exactly the main-hypothesis of the previous theorem.

In the following theorem we will make use of the function $\phi$ which is defined as follows:

**Definition 6.25** Let $v = \sum_i c_i\psi_i \in M_j$ and let $S : M_j \to M_j$ be some smoothing operator. Then we define the function

$$\phi : I\!R \to I\!R$$

by

$$S^{m/2}v = \sum_i c_i\,\phi(\lambda_i)\,\psi_i\,.$$

**Example 6.26** *Jacobi relaxation*
By (6.25) we see that for Jacobi-relaxation we have

$$\phi(\lambda) = (1-\lambda)^{m/2}\,.$$

**Theorem 6.27** Suppose that the hypothesis (6.29) holds. Let $\gamma$ be as in (6.35). Let $\tilde{C}$ be the approximation of the coarse grid correction operator $C$ satisfying

$$|||\tilde{C}v - Cv||| \le \hat{\gamma}|||Cv - v||| \text{ for all } v\in M_j\,,$$

where

$$\hat{\gamma} = \frac{1}{1+\beta}, \qquad \beta = \inf_{\lambda\in(0,1]} \frac{\phi^{-2}(\lambda) - 1}{\kappa^\alpha\lambda^\alpha}\,.$$

Then $|||S^{m/2}\tilde{C}S^{m/2}v||| \le \hat{\gamma}|||v|||$ for all $v\in M_j$.
**Proof:** Let first $w\in M_j$ be arbitrary. Write $\bar{v} = S^{m/2}v$, $\bar{w} = S^{m/2}w$.

For $T : M_j \to M_j$ a symmetric operator, we have ( $a(\cdot,\cdot)$ is symmetric)

$$a(Tu, v) = a(u, Tv).$$

Using this with $T = S^{m/2}$ we find

$$\begin{aligned} a(S^{m/2}\tilde{C}S^{m/2}v, w) &= a(\tilde{C}S^{m/2}v, S^{m/2}w) \\ &= a(\tilde{C}\bar{v}, \bar{w}) \\ &= a(C\bar{v} - \hat{\gamma}C\bar{v} + \tilde{C}\bar{v} - C\bar{v} + \hat{\gamma}C\bar{v}, \bar{w}) \\ &= (1-\hat{\gamma})a(C\bar{v}, \bar{w}) + a((\tilde{C} + (\hat{\gamma}-1))\bar{v}, \bar{w}) \\ &= (1-\hat{\gamma})a(C\bar{v}, C\bar{w}) + a((\tilde{C} + (\hat{\gamma}-1))\bar{v}, \bar{w})\,. \end{aligned}$$

The last equality because $C\bar{w} - \bar{w} \in M_{j-1}$ and hence

$$a(C\bar{v}, C\bar{w} - \bar{w}) = 0.$$

Now using (6.23) with $\alpha = 0$ gives

$$|a(S^{m/2}\tilde{C}S^{m/2}v, w)| \;\leq\; (1-\hat{\gamma})\; |||C\bar{v}||| \cdot |||C\bar{w}||| + |||\tilde{C}\bar{v} + (\hat{\gamma}-1)C\bar{v}|||\; |||\bar{w}|||\,,$$

or using lemma 6.17

$$|a(S^{m/2}\tilde{C}S^{m/2}v, w)| \leq$$

$$(1-\hat{\gamma})|||C\bar{v}||| \cdot |||C\bar{w}||| + \hat{\gamma}|||\bar{v}||| \cdot |||\bar{w}|||.$$

Now, with

$$\hat{v} = \begin{pmatrix} \sqrt{1-\hat{\gamma}}|||C\bar{v}||| \\ \sqrt{\hat{\gamma}}|||\bar{v}||| \end{pmatrix}, \quad \hat{w} = \begin{pmatrix} \sqrt{1-\hat{\gamma}}|||C\bar{w}||| \\ \sqrt{\hat{\gamma}}|||\bar{w}||| \end{pmatrix}$$

Cauchy-Schwartz

$$| < \hat{v}, \hat{w} > | \leq \|\hat{v}\|_{\mathrm{eucl}} \|\hat{w}\|_{\mathrm{eucl}}$$

yields

$$|a(S^{m/2}\tilde{C}S^{m/2}v, w)| \leq$$

$$\sqrt{(1-\hat{\gamma})|||C\bar{v}|||^2 + \hat{\gamma}|||\bar{v}|||^2} \cdot \sqrt{(1-\hat{\gamma})|||C\bar{w}|||^2 + \hat{\gamma}|||\bar{w}|||^2}.$$

Now we will investigate

$$(1-\hat{\gamma})\; |||C\bar{v}|||^2 + \hat{\gamma}\; |||\bar{v}|||^2.$$

Using (6.31) we find

$$(1-\hat{\gamma})\; |||C\bar{v}|||^2 + \hat{\gamma}\; |||\bar{v}|||^2 \leq (1-\hat{\gamma})\kappa^\alpha \|\bar{v}\|^2_{1+\alpha} + \hat{\gamma}\; |||\bar{v}|||^2\,.$$

Hence $S^{m/2}v = \sum_i c_i^v\, \phi(\lambda_i)\psi_i$, and

$$(1-\hat{\gamma})\; |||C\bar{v}|||^2 + \hat{\gamma}\; |||\bar{v}|||^2 \leq$$

$$\leq\; (1-\hat{\gamma})\kappa^\alpha \sum_i c_i^2 \phi^2(\lambda_i)\lambda_i^{1+\alpha} + \hat{\gamma} \sum_i c_i^2 \phi^2(\lambda_i)\lambda_i \leq$$

$$\leq\; \max_{\lambda\in[0,1]} [(1-\hat{\gamma})\kappa^\alpha \lambda^\alpha \phi^2(\lambda) + \hat{\gamma}\phi^2(\lambda)] \sum_i c_i^2 \lambda_i =$$

$$=\; \max_{\lambda\in[0,1]} [\phi^2(\lambda)((1-\hat{\gamma})\kappa^\alpha\lambda^\alpha + \hat{\gamma})] \cdot |||v|||^2.$$

Returning to the mainline of the proof, substituting the above result, we find

$$|a(S^{m/2}\tilde{C}S^{m/2}v, w| \leq |||v||| \cdot |||w||| \cdot \max_{\lambda\in[0,1]} [\phi^2(\lambda)((1-\hat{\gamma})\kappa^\alpha\lambda^\alpha + \hat{\gamma}].$$

Taking $w = S^{m/2}\tilde{C}S^{m/2}v$ and knowing that

$$a(S^{m/2}\tilde{C}S^{m/2}v, S^{m/2}\tilde{C}S^{m/2}v) = |||S^{m/2}\tilde{C}S^{m/2}v|||^2,$$

division by $|||S^{m/2}\tilde{C}S^{m/2}v|||$ yields

$$|||S^{m/2}\tilde{C}S^{m/2}v||| \leq \max_{\lambda\in[0,1]} [\phi^2(\lambda)((1-\hat{\gamma})\kappa^\alpha\lambda^\alpha + \hat{\gamma})]\; |||v|||\,. \tag{6.37}$$

Let $\beta = \inf_{\lambda\in[0,1]} \frac{\phi^{-2}(\lambda)-1}{\kappa^\alpha\lambda^\alpha} \Longrightarrow \beta\kappa^\alpha\lambda^\alpha + 1 \le \phi^{-2}(\lambda)$ for all $\lambda\in[0,1] \Longrightarrow \phi^2(\lambda)(\beta\kappa^\alpha\lambda^\alpha+1) \le 1$ for all $\lambda\in[0,1]$. Now $\hat\gamma = \frac{1}{1+\beta} \Longrightarrow \beta = \frac{1-\hat\gamma}{\hat\gamma}$. Substitution of $\frac{1-\hat\gamma}{\hat\gamma}$ for $\beta$ gives

$$\phi^2(\lambda)((\frac{1-\hat\gamma}{\hat\gamma})\kappa^\alpha\lambda^\alpha + 1) \le 1 \text{ for all } \lambda\in[0,1] \Longrightarrow$$

$$\phi^2(\lambda)((1-\hat\gamma)\kappa^\alpha\lambda^\alpha + \hat\gamma) \le \hat\gamma \text{ for all } \lambda\in[0,1] \Longrightarrow$$

$$\max_{\lambda\in[0,1]} [\phi^2(\lambda)((1-\hat\gamma)\kappa^\alpha\lambda^\alpha + \hat\gamma)] \le \hat\gamma\,.$$

Now the proof follows by substitution of this inequality in (6.37). ∎

**Example 6.28** *Jacobi relaxation*
In this example we want to compute $\hat\gamma$ when Jacobi-relaxation is used. From example 6.26 we know $\phi(\lambda) = (1-\lambda)^{m/2}$.

Hence

$$\beta = \inf_{\lambda\in[0,1]} \frac{(1-\lambda)^{-m}-1}{\kappa^\alpha\lambda^\alpha} = \frac{1}{\kappa^\alpha} \inf_{\lambda\in[0,1]} \frac{(1-\lambda)^{-m}-1}{\lambda^\alpha}\,.$$

For simplicity we assume $\alpha\in(0,1]$.

Define

$$f(\lambda) = \frac{(1-\lambda)^{-m}-1}{\lambda^\alpha} \quad \text{on} \quad (0,1)\,,$$

then

$$f'(\lambda) = \frac{(1-\lambda)^{m-1}}{\lambda^{\alpha+1}} \cdot (m\lambda - \alpha(1-\lambda) + \alpha(1-\lambda)^{m+1}).$$

We see that $f'(\lambda) \ge 0$ on (0,1) if

$$g(\lambda) = m\lambda - \alpha(1-\lambda) + \alpha(1-\lambda)^{m+1} \ge 0 \quad \text{on} \quad (0,1)\,,$$

and

$$\begin{aligned} g'(\lambda) &= m + \alpha - (m+1)\alpha(1-\lambda)^m \\ &\ge m + \alpha - (m+1)\alpha \ge 0 \ \text{ since } \ \alpha\in(0,1]. \end{aligned}$$

Further $g(0) = 0$ and hence $g(\lambda) \ge 0$ on $[0,1]$. We now conclude that $f'(\lambda) \ge 0$ on [0,1]. Using this we find

$$\beta = \frac{1}{\kappa^\alpha} \lim_{\lambda\to 0} \frac{(1-\lambda)^{-m}-1}{\lambda^\alpha} = \begin{cases} \frac{m}{\kappa}, & \alpha = 1 \qquad \text{(Taylor)} \\ 0, & 0 < \alpha < 1 \end{cases}$$

Hence

$$\hat\gamma = \frac{1}{1+\beta} = \begin{cases} \frac{\kappa}{\kappa+m}, & \text{if } \alpha = 1, \\ 1, & \text{if } 0 < \alpha < 1. \end{cases}$$

# Chapter 7

# Local mode analysis

## 7.1 Fourier transforms of continuous functions

In this section we collect well-known results with respect to Fourier transforms of functions that are defined (almost everywhere) on domains in the real $n$-dimensional space. All results mentioned in this section can be found in general texts as e.g. [14], [17], [19], [20].

Let $u \in L^2(\mathbb{R}^n)$, then its Fourier transform $\widehat{u}$ is defined by

$$\widehat{u}(y) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-ixy}\, u(x)\, dx \,. \tag{7.1}$$

Furthermore, a back-transformation formula is available:

$$\tilde{u}(x) = (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{+ixy}\, \widehat{u}(y)\, dy \,, \tag{7.2}$$

such that $\tilde{u}(x) = u(x)$ almost everywhere on $\mathbb{R}^n$. Moreover

$$\widehat{u} \in L^2(\mathbb{R}^n) \quad \text{and} \quad \|u\|_{L^2(\mathbb{R}^n)} = \|\widehat{u}\|_{L^2(\mathbb{R}^n)} \,.$$

We say that the Fourier transformation is a norm-invariant bijection $L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$.

The above definition of a Fourier transformation can be generalised to more general functions than $L^2(\mathbb{R}^n)$-functions. The same definition applies to the set of "tempered distributions" (see e.g. [20] ); in this case -again- the back transformation is available.

We see that the Fourier transform (FT) of a function defined on $\mathbb{R}^n$ is a function defined on $\mathbb{R}^n$ itself. A Fourier transformation can also be defined for a finite set of equally spaced data. In this case the FT of a set of $N$ data (the Finite Fourier transform) is again a set of $N$ coefficients (see e.g. [11]).

The FT of a periodic function (or, what is the same, the FT of a function defined on a torus) is a countable infinite set of coefficients. Analogously, in the following sections we shall introduce the Fourier transformation on an infinite set of equally spaced data. In this case the FT of such a "gridfunction" will be periodic function (a function defined on a torus).

**Definition 7.1** Let $h = (h_1, ..., h_n) \in \mathbb{R}^n$ be given, then the *h-periodisation* of a function $u : \mathbb{R}^n \to \mathbb{C}$ is defined by

$$\tilde{u}(x) = \sum_{k \in \mathbb{Z}^n} u(x - kh),$$

where $kh = (k_1 h_1, ..., k_n h_n)$.

We notice that $\tilde{u}(x)$ is a periodic function on $\mathbb{R}^n$ with period $h$; it is completely defined by a mapping $[0,h) \to \mathbb{C}$, where $[0,h)$ is defined by

$$[0,h) = [0,h_1) \times [0,h_2) \times ... \times [0,h_n)\,.$$

The FT of a function $\tilde{u}(x)$ defined on the torus $[0,h)$ is (cf. [14]) a sequence $\{c_k\}_{k\in\mathbb{Z}^n}$ defined by

$$c_k = \frac{1}{h^n(2\pi)^n}\int_0^h e^{-2\pi ikx/h}\tilde{u}(x)dx. \tag{7.3}$$

Also the Fourier transform on $[0,h)$ has its back-transformation. From this we see that the knowledge of $\widehat{u}(y)$ at only certain equally spaced points is enough to restore a periodisation of the original function $u$, whereas the complete definition of $\widehat{u}(y)$ (almost everywhere on $\mathbb{R}^n$) is needed to find the function $u(x)$ itself.

## 7.2 Gridfunctions

For a fixed "mesh" $h = (h_1, ..., h_n)$ with $h_i > 0$, $i = 1, 2, ..., n$, the regular infinite $n$-dimensional grid $\mathbb{Z}^n_h$ is defined by

$$\mathbb{Z}^n_h = \{jh | j\in\mathbb{Z}^n\}\,.$$

Further we introduce the notation

$$j/h = (j_1/h_1, ..., j_n/h_n),$$
$$h^n = h_1 \cdot h_2 \cdot ... \cdot h_n,$$
$$T^n_h = (2\pi/h]^n = (-\pi/h_1, \pi/h_1] \times ... \times (-\pi/h_n, \pi/h_n].$$

We call $T^n_h$ an $n$-dimensional torus.

**Definition 7.2** A complex or a real *gridfunction* is a mapping $\mathbb{Z}^n_h \to \mathbb{C}^d$, respectively $\mathbb{Z}^n_h \to \mathbb{R}^d$, where $d$ is the dimension of the range of the mapping.

**Remark:** We will restrict ourselves to the scalar real or complex gridfunction $\mathbb{Z}^n_h \to \mathbb{R}$ or $\mathbb{Z}^n_h \to \mathbb{C}$ and, unless stated otherwise, we shall use the word gridfunction for this kind of gridfunction exclusively. It is immediate that, with the usual addition and scalar multiplication, the set of all gridfunctions is a vector space. This vector space we denote by

$$l_h(\mathbb{Z}^n_h)$$

or, shortly, by $l_h$. For any $p \geq 1$ or $p = \infty$ the space $l_h$ can be provided with a norm

$$\|u_h\|_p = (h^n \sum_{j\in\mathbb{Z}^n} |u_h(jh)|^p)^{1/p}, \quad 1 \leq p < \infty, \tag{7.4}$$

or

$$\|u_h\|_\infty = \ \sup_{j\in\mathbb{Z}^n} |u_h(jh)|\,. \tag{7.5}$$

For a fixed $p$, $1 \leq p \leq \infty$, all gridfunctions with a finite norm $\|\cdot\|_p$ form a subspace of $l_h(\mathbb{Z}^n_h)$, which is denoted by

$$l^p_h(\mathbb{Z}^n_h)\,.$$

It is obvious that for any $p$, $1 \leq p \leq \infty$, $l^p_h(\mathbb{Z}^n_h)$ is a Banach space (cf. [25] p.35). Moreover, for $p = 2$ we know that $l^2_h(\mathbb{Z}^n_h)$ is a Hilbert space with the inner product

$$< u_h, v_h >_h = h^n \sum_{j\in\mathbb{Z}^n} u_h(jh)\overline{v_h(jh)}\,. \tag{7.6}$$

## 7.3 The Fourier transform of a gridfunction

Let $u_h : \mathbb{Z}_h^n \to \mathbb{C}$ be a gridfunction. We give the following

**Definition 7.3** The Fourier transform $\widehat{u_h}\in L^2(T_h^n)$ of $u_h\in l_h^2(\mathbb{Z}_h^n)$ is a function $T_h^n \to \mathbb{C}$, defined by

$$\widehat{u_h}(\omega) = (\frac{h}{\sqrt{2\pi}})^n \sum_{j\in\mathbb{Z}^n} e^{-ijh\omega}\, u_h(jh)\,. \tag{7.7}$$

The inverse transformation is given by

$$u_h(jh) = (\frac{1}{\sqrt{2\pi}})^n \int_{\omega\in T_h^n} e^{+ijh\omega}\, \widehat{u_h}(\omega)\, d\omega\,. \tag{7.8}$$

**Remarks:**

- We denote the Fourier transform also by

$$\widehat{u_h} \quad = \quad \mathbf{FT}(u_h).$$

- $\widehat{u_h}$ can also be considered as a $[2\pi/h]^n$-periodic function $\widehat{u_h} : \mathbb{R}^n \to \mathbb{C}$.
- By the *Parseval identity* we have

$$\|u_h\|_2 = \|\widehat{u_h}\|_{L^2(T_h^n)}\,.$$

  Hence the Fourier transformation operator $\mathbf{FT}$: $l_h^2 \to L^2(T_h^n)$ is a unitary operator.
- With the identity $\int_{T_h^n} e^{ih\omega(j-k)}d\omega = (\frac{2\pi}{h})^n\delta_{jk}$ we easily verify the back-transformation formula.

In the back-transformation formula (7.8) we see that any gridfunction $u_h\in l_h^2$ can be considered as a linear combination of gridfunctions $e_{h,\omega}$ of the form

$$e_{h,\omega}(jh) = e^{ijh\omega}, \quad \omega\in T_h^n,\, j\in\mathbb{Z}^n, \tag{7.9}$$

i.e. a periodic gridfunction with period $[\frac{2\pi}{h\omega}]^n$; such $e_{h,\omega}$ is called a *simple mode* . The parameter $\omega$ is called the *frequency* of this mode and $\widehat{u_h}(\omega)$ is the *amplitude* of the component (mode) with frequency $\omega$ in $u_h$.

Because $e_{h,\omega} \equiv e_{h,\omega+2\pi k/h}$, for all $k\in\mathbb{Z}^n$, a frequency $\omega$ cannot be distinguished from a frequency $\omega + 2\pi k/h$. This phenomenon is called *aliasing* .

It is clear from definition 7.3 that the range of frequencies that can be represented on a fine grid (small $h$) is larger than the range of those which can be represented on a coarser grid (large $h$). This explains why smooth gridfunctions (i.e. gridfunctions with relatively small amplitudes for the high frequencies) can be much better represented on a coarser grid than less smooth gridfunctions.

**Remarks:**

- For functions $u_h$ defined on a bounded domain $\Omega_h$, $\Omega_h = [-Nh, Nh]^n \bigcap \mathbb{Z}_h^n$, with periodic boundary conditions, a similar definition can be given for the Fourier transform $\widehat{u_h}$. The only difference is that the set of frequencies, $T_h^n$, available for such gridfunction $u_h$ is discrete:

$$T_h^n = \{(\pi k_1/Nh, ..., \pi k_n/Nh) \mid -N < k_j < N,\ j = 1, ..., n\} \tag{7.10}$$

- Functions on a bounded domain with homogeneous Dirichlet or Neumann boundary conditions can be considered as a special case of problems with periodic boundary conditions. In this case the elementary components in these functions are all of the form

$$e_{h,\omega} \pm e_{h,\omega} = e^{ijh\omega} \pm e^{-ijh\omega}\,.$$

  and the Fourier analysis method gives an exact description of symmetric operators $B_h$ : $l_h(\Omega_h) \to l_h(\Omega_h)$ because for symmetric operators we have $(B_h)_{i,k} = (B_h)_{k,i}$ and hence $\widehat{B_h}(\omega) = \widehat{B_h}(-\omega)$. (See Section 7.5 for a definition of $\widehat{B_h}$).

## 7.4 The relation between FTs of a function restricted to different grids

In this section we will present two theorems. One gives the relation between the FT of a continuous function defined on $\mathbb{R}^n$ and the FT of its restriction to the grid $\mathbb{Z}_h^n$. The other gives the relation between the FT of a gridfunction and the FT of its canonical restriction to a coarser grid.

**Theorem 7.4** Let $u \in L^2(\mathbb{R}^n)$ be a continuous function with FT $\widehat{u}$. Its restriction to the grid $\mathbb{Z}_h^n$ is defined by

$$u_h(jh) = (R_h^0 u)(jh) = u(jh),\; j \in \mathbb{Z}^n\,. \tag{7.11}$$

Then we have the following relation between $\widehat{u}$ and $\widehat{u_h}$:

$$\widehat{u_h}(\omega) = \sum_{k \in \mathbb{Z}^n} \widehat{u}(\omega + 2\pi k/h), \tag{7.12}$$

i.e. $\widehat{u_h}$ is the $[2\pi/h]^n$-periodisation of $\widehat{u}$ (def. 7.1).
**Proof:**

$$\begin{aligned}
\widehat{u_h}(\omega) &= (\tfrac{h}{\sqrt{2\pi}})^n \textstyle\sum_j e^{-ijh\omega} u(jh) \\
&= (\tfrac{h}{2\pi})^n \textstyle\sum_j e^{-ijh\omega} \sum_k \int_{-\pi/h}^{\pi/h} e^{ijh(y+2\pi k/h)}\, \widehat{u}(y + 2\pi k/h) dy \\
&= (\tfrac{h}{2\pi})^n \textstyle\sum_j e^{-ijh\omega} \int_{-\pi/h}^{\pi/h} e^{ijhy} \sum_k \widehat{u}(y + 2\pi k/h) dy
\end{aligned}$$

Using (7.7) and (7.8) we see that this equals

$$\sum_{k \in \mathbb{Z}^n} \widehat{u}(\omega + 2\pi k/h)\,.$$

■

**Definition 7.5** Let $u_h \in l_h(\mathbb{Z}_h^n)$ then its *canonical q-restriction* $R_q^0 u_h, (q \in \mathbb{Z}^n)$, is the gridfunction $u_H$ defined on $\mathbb{Z}_H^n = \mathbb{Z}_{qh}^n$, defined by

$$(R_q^0 u_h)(jH) = u_H(jH) = u_h(jqh)\,. \tag{7.13}$$

Sometimes we write $R_{Hh}^0$, with $H = qh$, instead of $R_q^0$.

**Lemma 7.6** If $u_h \in l_h^p(\mathbb{Z}_h^n)$ then $R_q^0 u_h \in l_H^p(\mathbb{Z}_H^n)$ with $H = qh$.
**Proof:**

$$\begin{aligned}
\|u_H\|^p_{l_H^p(\mathbb{Z}_H^n)} &= q^n h^n \textstyle\sum_{j \in \mathbb{Z}^n} |u_h(jqh)|^p \\
&\le q^n h^n \textstyle\sum_{j \in \mathbb{Z}^n} |u_h(jh)|^p \\
&\le q^n \|u_h\|^p_{l_h^p(\mathbb{Z}_h^n)} < \infty\,.
\end{aligned}$$

■

**Theorem 7.7**

$$(\widehat{R_q u_h})(\omega) = \sum_{p\in[0,q)} \widehat{u_h}(\omega + 2\pi p/H), \text{ for all } \omega\in T_H^n, H = qh, q\in\mathbb{Z}^n, q > 0\,. \qquad (7.14)$$

**Proof:** The proof is left as an exercise. ∎

Theorem 7.7 shows us that, using the restriction $R_q^0$ with $q\in\mathbb{Z}^n, q = (q_1, ..., q_n)$, we get aliasing of $Q = q_1....q_n$ frequencies onto one.

## 7.5 Toeplitz operators and their FTs

Let $A : l_h(\mathbb{Z}_h^n) \to l_h(\mathbb{Z}_h^n)$ be a linear operator. We are interested to know if for any given $\omega\in T_h^n$ the operator $A$ has an eigenvalue $\lambda_\omega$ corresponding with an eigenfunction $e_{h,\omega}$ as defined by (7.9). Suppose it to be true. Then, denoting $A$ by its matrix representation $(a_{mj})$, $m, j\in\mathbb{Z}^n$, the following holds

$$\sum_{j\in\mathbb{Z}^n} a_{mj} e_{h,\omega}(jh) = \lambda_\omega e_{h,\omega}(mh)$$

$$\sum_{j\in\mathbb{Z}^n} a_{mj} e^{ijh\omega} = \lambda_\omega e^{imh\omega},$$

hence

$$\lambda_\omega = \sum_{j\in\mathbb{Z}^n} a_{mj} e^{i(j-m)h\omega} = \sum_{k\in\mathbb{Z}^n} a_{m,m+k} e^{ikh\omega}\,.$$

Because $\lambda_\omega$ is independent of $m$, an $a_{-k}$ should exist such that $a_{m,m+k} = a_{-k}$ for all $m\in\mathbb{Z}^n$. These considerations give rise to the following

**Definition 7.8** A linear operator $A : l_h(\mathbb{Z}_h^n) \to l_h(\mathbb{Z}_h^n)$ whose matrix elements $a_{mj}$ satisfy the relation $a_{m,m+k} = a_{-k}$ for all $m\in\mathbb{Z}^n$ and certain $a_{-k}$, is called a *Toeplitz operator*.

**Lemma 7.9** Let $A : l_h(\mathbb{Z}_h^n) \to l_h(\mathbb{Z}_h^n)$ be a Toeplitz operator with matrix representation $(a_{mj})$. Then for any $\omega\in T_h^n$

$$\lambda_\omega = \sum_{k\in\mathbb{Z}^n} a_{-k} e^{ikh\omega} \qquad (7.15)$$

is an eigenvalue of $A$, corresponding with the eigenfunction $e_{h,\omega}$.

**Definition 7.10** Let $a_h, u_h\in l_h(\mathbb{Z}_h^n)$ be two gridfunctions. The *$a_h$-convolution* of $u_h$, denoted by $a_h \star u_h\in l_h(\mathbb{Z}_h^n)$, is defined by

$$(a_h \star u_h)(mh) = \sum_{j\in\mathbb{Z}^n} a_h((m-j)h)\; u_h(jh) \text{ for all } m\in\mathbb{Z}^n. \qquad (7.16)$$

We can identify the $a_h$-convolution with a Toeplitz operator $A_h$ on $l_h(\mathbb{Z}_h^n)$ in the following sense:

$$A_h : l_h(\mathbb{Z}_h^n) \to l_h(\mathbb{Z}_h^n) \text{ such that } : A_h u_h = a_h \star u_h. \qquad (7.17)$$

I.e. the Toeplitz operator $A_h$ is uniquely identified by the gridfunction $a_h\in l_n(\mathbb{Z}_h^n)$ as an $a_h$-convolution. The gridfunction $a_h$ is related to the matrix elements of $A$ by the relation

$$a_h(-kh) = a_{-k} \text{ for all } k\in\mathbb{Z}^n\,. \qquad (7.18)$$

By (7.18) follows

$$\begin{array}{rl} (Au_h)(mh) & = \sum_{j\in\mathbb{Z}^n} a_{mj}\, u_h(jh) \\ & = \sum_{j\in\mathbb{Z}^n} a_{-(j-m)}\, u_h(jh) \\ & = \sum_{j\in\mathbb{Z}^n} a_h((m-j)h)\, u_h(jh) \\ & = (a_h * u_h)(mh), \quad \text{for all} \quad m\in\mathbb{Z}^n,\ u_h\in l_h. \end{array} \tag{7.19}$$

We say that the gridfunction $a_h$ *generates* the Toeplitz operator $A = a_h \star$ .

**Definition 7.11** The *Fourier transform* $\widehat{A_h} : T_h^n \to \mathbb{C}$ of a *Toeplitz operator* $A_h$ is defined by

$$A_h e_{h,\omega} = \widehat{A_h}(\omega) e_{h,\omega} \,. \tag{7.20}$$

Using (7.15) we see that this means that

$$\widehat{A_h}(\omega) = \lambda_\omega = \sum_{k\in\mathbb{Z}^n} a_h(kh) e^{-ikh\omega}. \tag{7.21}$$

An immediate consequence of this is the following

**Lemma 7.12**

$$\widehat{A_h u_h}(\omega) = \widehat{A_h}(\omega)\widehat{u_h}(\omega). \tag{7.22}$$

Combining (7.21) and definition 7.3 we obtain the following

**Lemma 7.13**

$$\widehat{A_h}(\omega) = \left(\frac{\sqrt{2\pi}}{h}\right)^n \widehat{a_h}(\omega)\,.$$

Examples of Toeplitz operators are linear difference operators with constant coefficients, defined on the space of gridfunctions, such that the same difference equation is applied at each gridpoint (i.e. the matrix elements of this operator satisfy $a_{m,m+k} = a_{-k}$ for all $m\in\mathbb{Z}^n$). Another example is the translation operator $T_{kh}$

$$(T_{kh}u_h)(jh) = u_h((j-k)h)\,. \tag{7.23}$$

A typical property of, for example, the above mentioned linear difference operators, is that the difference equation applied to some gridpoint only relates a few neighbouring gridpoints. Examining the definition of a convolution product

$$(a_h \star u_h)(mh) = \sum_{k\in\mathbb{Z}^n} a_h(kh)\ u_h((m-k)h)$$

we see that this means that the generating gridfunction $a_h$ has *finite support:* $a_h(kh) \neq 0$ only for some $k\in\mathbb{Z}^n$ in a neighbourhood of $0\in\mathbb{Z}^n$.

Suppose now that the Toeplitz operator $A_h$ is generated by a gridfunction $a_h$ with finite support, i.e.

$$\left\{ \begin{array}{ll} a_h(-kh) = s_k & k\in V, \\ a_h(-kh) = 0 & k\notin V, \end{array} \right.$$

for some *finite* set $V \subset \mathbb{Z}^n$, containing $0\in\mathbb{Z}^n$. Then we often associate $A_h$ with a *stencil.* For $n = 2$ such a stencil is given by

$$A_h = \begin{bmatrix} & \vdots & \vdots & \vdots & \\ \cdots & s_{-1,1} & s_{0,1} & s_{1,1} & \cdots \\ \cdots & s_{-1,0} & s_{0,0} & s_{1,0} & \cdots \\ \cdots & s_{-1,-1} & s_{0,-1} & s_{1,-1} & \cdots \\ & \vdots & \vdots & \vdots & \end{bmatrix}_h ,$$

where only non-zero $s_k$ are given. This notation (the *stencil notation*) is very convenient if $V$ is small.

We will now give a few detailed examples of Toeplitz operators and their FT.

**Example 7.14**
Consider a central difference discretisation of the Laplace operator $\Delta$ in one dimension on a regular infinite grid $\mathbb{Z}_h^1$. The matrix representation of the resulting discrete operator $L_h$ is given in Chapter 4. In many cases the stencil representation is more convenient:

$$L_h = \frac{1}{h^2}\,[1,-2,1]_h\,.$$

We can now, very easily, determine $\widehat{L_h}$, using (7.21):

$$\widehat{L_h}(\omega) = \frac{1}{h^2}(e^{ih\omega} + e^{-ih\omega} - 2) = \frac{1}{h^2}(2\cos(h\omega) - 2) = -\frac{4}{h^2}\sin^2(h\omega/2)\,. \tag{7.24}$$

**Example 7.15**
The translation operator $T_{kh}$ (see (7.23)) can be written as a convolution:

$$(T_{kh}u_h)(mh) = u_h((m-k)h) = \sum_{j\in\mathbb{Z}^n} a_h(jh)\,u_h((m-j)h),$$

where $a_h(kh) = 1$, and $a_h(jh) = 0$ if $k \neq j$. Hence we can identify $T_{kh}$ with the following stencil ($n = 1$):

$$T_{kh} = [1, 0, ..., 0, \underline{0}, 0, ..., 0].$$

with $k$ terms on either side of $\underline{0}$.
Its FT is given by

$$\widehat{T_{kh}}(\omega) = e^{-ikh\omega} = \cos(kh\omega) - i\,\sin(kh\omega)\,.$$

Because $|\widehat{T_{kh}}(\omega)| = 1$ for all $\omega\in T_h^1$ we see that application of a translation operators to $u_h$ does not affect the absolute value of $\widehat{u_h}(\omega)$.

## 7.6 Consistency of a discrete operator

For linear differential operators with constant coefficients we can use Fourier analysis to determine the order of consistency of a discretisation of this operator. First we recall the following, well known property . A linear partial differential operator $L$ of order $N$ with constant coefficients, given by

$$L = \sum_{|\alpha|\le N} c_\alpha \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1}...\partial x_n^{\alpha_n}}\,, \quad (|\alpha| = \alpha_1 + ... + \alpha_n)\,,$$

has the property that for any $\xi\in\mathbb{C}^n$ the function $\mathbb{R}^n \to \mathbb{C}: \; x \to e^{ix\cdot\xi}$ is mapped by $L$ onto a multiple of itself:

$$Le^{ix\cdot\xi} = p(\xi)e^{ix\cdot\xi},$$

with $p(\xi) = \sum_{|\alpha|\le N} c_\alpha (i\xi)^\alpha$, the characteristic polynomial. Here $x\cdot\xi = \sum_{k=1}^n x_k\xi_k$ and $\eta^\alpha = \eta_1^{\alpha_1}...\eta_n^{\alpha_n}$. The characteristic polynomial $p$ is called the *symbol* of $L$. In analogy to definition 7.11 we can write

$$\widehat{L}(\omega) = p(\omega)\,.$$

Consider a linear partial differential operator $L$ with constant coefficients and its discretisation $L_h$ on some grid $\mathbb{Z}_h^n$. We are interested in the truncation error

$$\tau_h = L_h R_h - \bar{R}_h L, \quad \tau_h : E \to \hat{E}_h$$

(see Figure 2.1 for a summary of the operators and notation used). $R_h$ and $\bar{R}_h$ are the canonical restriction. For this purpose we consider some arbitrary $\omega \in T_h^n$ and the truncation error of $L_h$ for the function

$$e_\omega(x) = e^{i\omega \cdot x}, \quad e_\omega \in E\,.$$

We find

$$\begin{aligned}\tau_h(e_\omega)(x) &= (L_h R_h e_\omega - \bar{R}_h L e_\omega)(x)\\ &= L_h e^{i\omega jh} - R_h \widehat{L}(\omega) e^{i\omega \cdot x}\\ &= (\widehat{L_h}(\omega) - \widehat{L}(\omega)) e^{ijh\omega},\end{aligned}$$

hence

$$\|\tau_h(e_\omega)\| = |\widehat{L_h}(\omega) - \widehat{L}(\omega)|. \tag{7.25}$$

For given $L$ and $L_h$ this expression can be developed in powers of $h$ and the order of consistency of $L_h$ can be determined.

**Example 7.16** *Consistency order*
We take $L = \Delta$ (in two dimensions),

$$L_h = \frac{1}{h^2} \begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix}_h .$$

We find

$$\begin{aligned}\widehat{L}(\omega) &= -\omega_1^2 - \omega_2^2,\\ \widehat{L_h}(\omega) &= \tfrac{1}{h^2}[2\cos(h\omega_1) + 2\cos(h\omega_2) - 4]\\ &= -\tfrac{1}{h^2}[4\sin^2(\omega_1 h/2) + 4\sin^2(\omega_2 h/2)].\end{aligned}$$

Hence

$$\begin{aligned}|\widehat{L_h}(\omega) - \widehat{L}(\omega)| &= |\frac{4}{h^2}\sin^2(\omega_1 h/2) + \frac{4}{h^2}\sin^2(\omega_2 h/2) - \omega_1^2 - \omega_2^2|\\ &= \frac{1}{12} h^2 |\omega_1^4 + \omega_2^4| + \mathcal{O}(h^4) \quad \text{for} \quad h \to 0\,.\end{aligned}$$

We conclude that the order of consistency of $L_h$ is 2.

## 7.7 The smoothing factor for relaxation methods

Consider a multigrid algorithm where for each coarser mesh the meshwidth is related to finer mesh by $H = 2h$. In Section 7.3 we saw that the range of frequencies that can be represented on the coarser grid is smaller than the range available on the finer grid. For $H = 2h$ the frequencies that cannot be correctly transferred (i.e. without aliasing) to the next coarser grid are all $\omega \in T_h^n \setminus T_{2h}^n$.

If we want the coarse-grid-correction in the multigrid algorithm to be successful, the relaxation steps should damp the amplitudes of the error modes with frequencies $\omega \in T_h^n \setminus T_{2h}^n$. This smoothing property of a given relaxation method with amplification operator $M_h^{REL}$ for the error, is expressed in terms of the so called *smoothing factor* $\mu$:

$$\mu = \sup_{\omega \in T_h^n \setminus T_{2h}^n} |\widehat{M_h^{REL}}(\omega)|\,. \tag{7.26}$$

This smoothing factor $\mu$ describes how well the high-frequency components in the error are reduced by the relaxation method.

If $L_h$ is a Toeplitz operator and $B_h$, an approximate inverse of $L_h$, is a Toeplitz operator as well, then

$$M_h^{REL} = I_h - B_h L_h$$

is a Toeplitz operator, and

$$\widehat{M_h^{REL}}(\omega) = I - \widehat{B_h}(\omega)\widehat{L_h}(\omega). \tag{7.27}$$

**Example 7.17** *computation smoothing factor*
We return to Example 7.14 where we considered

$$L_h = \frac{1}{h^2}[1,-2,1]_h, \quad \widehat{L_h}(\omega) = \frac{1}{h^2}(2\cos(h\omega) - 2).$$

As a relaxation method we choose damped Jacobi relaxation with parameter $\alpha$, i.e. we use

$$B_h = \alpha(\mathrm{diag}(L_h))^{-1} = h^2\,[0, \frac{-\alpha}{2}, 0].$$

We find $\widehat{B_h}(\omega) = \frac{-h^2\alpha}{2}$, hence

$$\widehat{M_h^{REL}}(\omega) = 1 - \widehat{B_h}(\omega)\widehat{L_h}(\omega) = 1 - \alpha + \alpha\cos(h\omega).$$

One can show that $\alpha = 2/3$ yields optimal smoothing of this relaxation method for the given problem. For $\alpha = 2/3$ the smoothing factor is $\mu = 1/3$.

## 7.8 Restrictions and prolongations

In section 7.4 we introduced the canonical restriction $R_h^0$ and the canonical $q$-restrictions $R_q^0 : l_h(\mathbb{Z}_h^n) \to l_{qh}(\mathbb{Z}_{qh}^n)$. Now we show that an arbitrary restriction, where for each gridpoint the same stencil is used) can be described in terms of a Toeplitz operator.

**Definition 7.18** A homogeneous (*weighted*) *q-restriction* of a gridfunction $u_h$ defined on $\mathbb{Z}_h^n$ to a gridfunction $u_H, H = qh$, defined on $\mathbb{Z}_H^n$, denoted by $R_q(a_h)u_h$, is defined by

$$(R_q(a_h)u_h)(jH) = u_H(jH) = \sum_{k\in\mathbb{Z}^n} a_h(kh)\; u_h((qj-k)h). \tag{7.28}$$

The gridfunction $a_h$ generates a Toeplitz operator. If we denote this operator by $A_h$ (i.e. $A_h = a_h\star$) the expression (7.28) can be written in the following form:

$$(R_q(a_h)u_h)(jH) = (R_q^0 A_h u_h)(jqh). \tag{7.29}$$

From this we may conclude, using theorem 7.7 and lemma 7.12, that

$$\mathbf{FT}(R_q(a_h)u_h)(\omega) = \sum_{p\in[0,q)} \widehat{A_h}(\omega + 2\pi p/qh)\; \widehat{u_h}(\omega + 2\pi p/qh) \tag{7.30}$$

The frequencies $\omega + 2\pi p/qh$, $p \neq 0$, we call the *higher harmonics* of the frequency $\omega$. The total number of harmonics ($\omega$ itself included) is

$$Q = q_1.....q_n.$$

Let us first consider the case $n = 1, q = 2$. Then we have

$$\mathbf{FT}(R_q(a_h)u_h)(\omega) = \sum_{p=0}^{1} \widehat{A_h}(\omega + \pi p/h)\,\widehat{u_h}(\omega + \pi p/h).$$

Another notation of this is

$$\widehat{u_H}(\omega) = \mathbf{FT}(R_q(a_h)u_h)(\omega) = (\widehat{A_h}(\omega),\ \widehat{A_h}(\omega+\pi/h)) \begin{pmatrix} \widehat{u_h}(\omega) \\ \widehat{u_h}(\omega+\pi/h) \end{pmatrix}$$

$$=: \ \vec{\mathbf{FT}}(R_q(a_h))(\omega) \cdot \ \vec{\mathbf{FT}}(u_h)(\omega), \quad \omega \in T^1_{qh}.$$

For general $n = 1, q > 1, \omega \in T^n_{qh}$, we write also

$$\mathbf{FT}(R_q(a_h)u_h)(\omega) = \ \vec{\mathbf{FT}}(R_q(a_h))(\omega) \ \ \vec{\mathbf{FT}}(u_h)(\omega), \tag{7.31}$$

where $\vec{\mathbf{FT}}(R_q(a_h))(\omega)$ is a $1 \times Q$-matrix with components $\widehat{A_h}(\omega + 2\pi p/qh)$, $p \in [0,q)$, and $\vec{\mathbf{FT}}(u_h)(\omega)$ a vector with $Q$ components $\widehat{u_h}(\omega + 2\pi p/qh)$, $p \in [0,q)$.

**Example 7.19** $(n = 1, q = 2)$

$$A_h = a_h\star = [1/4, 1/2, 1/4]_h,$$

$$\widehat{A_h}(\omega) = \frac{1}{2} + \frac{1}{2}\cos(\omega h) = \cos^2(\omega h/2),$$

$$\widehat{A_h}(\omega + \pi/h) = \sin^2(\omega h/2),$$

$$\mathbf{FT}(R_q(a_h))(\omega) = (\cos^2(\omega h/2)\,,\, \sin^2(\omega h/2))\cdot$$

**Example 7.20** $(n = 2, q = 2)$

$$A_h = a_h\star = \begin{bmatrix} & 1/8 & \\ 1/8 & 1/2 & 1/8 \\ & 1/8 & \end{bmatrix}_h ,$$

Here $\widehat{A_h}(\omega) = \frac{1}{2}\cos^2(\omega_1 h/2) + \frac{1}{2}\cos^2(\omega_2 h/2)$, and analogously we find $\widehat{A_h}(\omega_1,\ \omega_2 + \pi/h)$, $\widehat{A_h}(\omega_1 + \pi/h,\ \omega_2)$ and $\widehat{A_h}(\omega_1 + \pi/h,\ \omega_2 + \pi/h)$. Thus, we find

$$\mathbf{FT}(R_q(a_h))(\omega) = \frac{1}{2} \begin{pmatrix} \cos^2(\omega_1 h/2) + \cos^2(\omega_2 h/2) \\ \cos^2(\omega_1 h/2) + \sin^2(\omega_2 h/2) \\ \sin^2(\omega_1 h/2) + \cos^2(\omega_2 h/2) \\ \sin^2(\omega_1 h/2) + \sin^2(\omega_2 h/2) \end{pmatrix}.$$

**Definition 7.21** Let $u_H \in l_H(\mathbb{Z}^n_H)$ be a gridfunction defined on $\mathbb{Z}^n_H$, then its *flat q-prolongation* $P^0_q u_H$, $q \in \mathbb{Z}^n$, is the gridfunction $u_h$, $h = H/q$, $u_h \in l_h(\mathbb{Z}^n_h)$, defined by

$$u_h(jh) = (P^0_q u_H)(jh) = \begin{cases} u_H(kqh), & \text{if } j = kq, \\ 0 & \text{otherwise.} \end{cases} \tag{7.32}$$

One can easily verify the following

**Theorem 7.22**

$$\mathbf{FT}(P^0_q u_H)(\omega) = \widehat{u_h}(\omega) = q^{-n}\,\widehat{u_H}(\omega),\ \omega \in T^n_h\cdot \tag{7.33}$$

**Proof:** The proof is left as an exercise. ∎

The interpretation of this theorem is that $\widehat{u_h}(\omega)$ is the periodic continuation of $\widehat{u_H}(\omega)$, except for the factor

$$q^{-n} = Q^{-1}\cdot$$

Analogous to definition 7.18 any homogeneous $q$-prolongation can be described in terms of a Toeplitz operator $A_h = a_h\star$:

$$P_q(a_h) = A_h P_q^0\cdot \tag{7.34}$$

**Example 7.23** *Linear interpolation in one dimension ($n = 1, q = 2$)*

$$A_h = a_h\star\ = [\frac{1}{2}, 1, \frac{1}{2}]_h,$$

$$\widehat{u_h}(\omega) = \mathbf{FT}(A_h P_q^0 u_H)(\omega) = \widehat{A_h}(\omega)\mathbf{FT}(P_q^0 u_H)(\omega)$$

$$= q^{-1}\widehat{A_h}(\omega)\widehat{u_H}(\omega) = \frac{1}{2}(1+\cos(\omega h))\widehat{u_H}(\omega) = cos^2(\omega h/2)\widehat{u_H}(\omega), \quad \omega\in T_h^1\,.$$

(Notice that for $\widehat{u_H}$, originally defined on $T_H^1$, now its periodic continuation on $T_h^1$ is used.) We can write this also as

$$\vec{\mathbf{FT}}(u_h) = \begin{pmatrix} \cos^2(\omega h/2) \\ \sin^2(\omega h/2) \end{pmatrix} \widehat{u_H}\,, \quad \omega\in T_H^1.$$

Analogous to (7.31) we introduce the notation

$$\vec{\mathbf{FT}}(u_h)(\omega) = \ \mathbf{FT}(P_q(a_h))(\omega)\ \ \mathbf{FT}(u_H)(\omega)\,. \tag{7.35}$$

Here $\vec{\mathbf{FT}}(P_q(a_h))$ is a $Q\times 1$- matrix, given by

$$\vec{\mathbf{FT}}(u_h)(\omega) = \ \mathbf{FT}(P_q(a_h)(\omega) = \frac{1}{Q}\begin{pmatrix} \widehat{A_h}(\omega) \\ \vdots \\ \widehat{A_h}(\omega+2\pi/qh) \\ \vdots \end{pmatrix}, \quad \omega\in T_H^n,$$

and $\vec{\mathbf{FT}}(u_h)(\omega)$ a $Q$-vector.

## 7.9 The order of a restriction or a prolongation

In the previous section we saw that a restriction or a prolongation could be associated with a gridfunction $a_h$, an operator $A_h$ or its FT $\widehat{A_h}(\omega)$. Such a FT is a trigonometric polynomial in $\theta = \omega h$. If there is no possibility of confusion in this section we write $A(\theta)$ instead of $\widehat{A_h}(\omega)$ (for restrictions) or $\frac{1}{Q}\widehat{A_h}(\omega)$ for prolongations.

**Definition 7.24** The *(primary) order* or *low-frequency order* of a prolongation $P_q(a_h)$ or of a restriction $R_q(a_h)$, is the largest number $m \geq 0$ for which

$$A(\theta) = 1 + \mathcal{O}(|\theta|^m), \ \text{for}\ |\theta| \to 0\,.$$

**Definition 7.25** For a given grid-coarsening factor $q\in\mathbb{Z}^n$, $H = qh$, the *secondary order* or *high-frequency order* of a prolongation $P_q(a_h)$ or of a restriction $R_q(a_h)$, is the largest number $m \geq 0$ for which

$$A(\theta + 2\pi p/q) = \mathcal{O}(|\theta|^m)\,(|\theta|\to 0),$$

for all $p\in[0,q)^n \subset \mathbb{Z}^n, p \neq 0^n\cdot$

The use of these two definitions will become clear in section 7.11 where primary and secondary orders play a role in the convergence of multigrid methods. We will now give a few examples of the computation of primary and secondary orders of some prolongations and restrictions. In all examples we assume $q = 2$.

**Example 7.26** *Canonical q-restriction* $R_q^0, n = 1$

$$A_h = a_h\star = [0,1,0]_h,$$

$A(\theta) = 1$, hence the primary order is $\infty$ and the secondary order is 0.

**Example 7.27** *Linear interpolation,* $n = 1$

$$A_h = a_h\star = [1/2, 1, 1/2]_h,$$

$$\frac{1}{Q}\widehat{A_h}(\omega) = \cos^2(\omega h/2) \quad \text{(see example 7.23)},$$

$$A(\theta) = \cos^2(\theta/2) = 1 - \frac{1}{4}\theta^2 + \mathcal{O}(\theta^4) \quad \text{for} \quad \theta \to 0\,,$$

hence the primary order is also 2.

$$A(\theta+\pi) = \cos^2(\frac{\theta+\pi}{2}) = \sin^2(\theta/2) = \frac{1}{4}\theta^2 + \mathcal{O}(\theta^4) \quad \text{for} \quad \theta \to 0,$$

hence the secondary order is also 2.

**Example 7.28** *Linear interpolation,* $n = 2$

$$A_h = a_h\star = \begin{bmatrix} 1/2 & 1/2 & \\ 1/2 & 1 & 1/2 \\ & 1/2 & 1/2 \end{bmatrix}_h .$$

We write $h\omega = (\phi, \theta)$ and find

$$\begin{aligned} 4A(\phi,\theta) = \;& 1 + \tfrac{1}{2}e^{i\phi} + \tfrac{1}{2}e^{-i\phi} + \tfrac{1}{2}e^{i\theta} + \tfrac{1}{2}e^{-i\theta} + +\tfrac{1}{2}e^{i(\phi-\theta)} + \tfrac{1}{2}e^{-i(\phi-\theta)} \\ & = 1 + \cos\phi + \cos\theta + \cos(\phi-\theta) \\ & = (1+\cos\theta)(1+\cos\phi) + \sin\phi\sin\theta \\ & = 2\cos^2(\theta/2) 2\cos^2(\phi/2) + 2\sin(\phi/2)\cos(\phi/2)\, 2\sin(\theta/2)\cos(\theta/2) \\ & = 4\cos(\theta/2)\cos(\phi/2)\cdot[\cos(\phi/2)\cos(\theta/2) + \sin(\phi/2)\sin(\theta/2)] \\ & = 4\cos(\theta/2)\cos(\phi/2)\cos((\phi-\theta)/2). \end{aligned}$$

Expanding $A(\phi,\theta)$ in a Taylor series yields

$$\begin{aligned} A(\phi,\theta) & = (1 - \tfrac{1}{2}(\tfrac{\theta}{2})^2 + \mathcal{O}(|\theta|^4)) \cdot (1 - \tfrac{1}{2}(\tfrac{\phi}{2})^2 + \mathcal{O}(|\phi|^4)) \cdot (1 - \tfrac{1}{2}(\tfrac{\phi-\theta}{2})^2 + \mathcal{O}(|\phi-\theta|^4)) \\ & = 1 - \tfrac{1}{4}(\theta^2 + \phi^2 - \phi\theta) + \mathcal{O}((\phi,\theta)^4), \quad \text{for} \quad (\phi,\theta) \to 0\,. \end{aligned}$$

We conclude that the primary order of linear interpolation in two dimensions is 2. For the higher harmonics we find

$$A(\pi-\phi,\theta) = \frac{1}{4}(\theta\phi - \phi^2) + \mathcal{O}((\phi,\theta)^4), \quad \text{for} \;\; (\phi,\theta) \to 0,$$

and similar expressions for $A(\phi, \pi-\theta)$ and $A(\pi-\phi, \pi-\theta)$. We conclude that the secondary order is 2.

**Remark:** The same holds for the linear weighted restriction (7-point restriction).

**Example 7.29** *Bilinear interpolation, $n = 2$*

$$A_h = a_h \star = \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}_h .$$

$$\begin{aligned} A(\phi,\theta) &= 1 + \tfrac{1}{2}e^{i\phi} + \tfrac{1}{2}e^{-i\phi} + \tfrac{1}{2}e^{i\theta} + \tfrac{1}{2}e^{-i\theta} + \\ + & \quad \tfrac{1}{4}e^{i(\phi-\theta)} + \tfrac{1}{4}e^{-i(\phi-\theta)} + \tfrac{1}{4}e^{i(\phi+\theta)} + \tfrac{1}{4}e^{-i(\phi+\theta)} \\ &= 1 + \cos\phi + \cos\theta + \tfrac{1}{2}\cos(\phi-\theta) + \tfrac{1}{2}\cos(\phi+\theta) \\ &= 1 + \cos\phi + \cos\theta + cos\phi\cos\theta \\ &= (1+\cos\phi)\,(1+\cos\theta) \\ &= 4\cos^2(\phi/2)\cos^2(\theta/2)\,. \end{aligned}$$

Expanding $A(\phi,\theta)$ in a Taylor series we find

$$A(\phi,\theta) = (1 - (\frac{\phi}{2})^2 + \mathcal{O}(|\phi|^4))(1 - (\frac{\theta}{2})^2 + \mathcal{O}(|\theta|^4))$$

$$= 1 - \frac{1}{4}(\phi^2 + \theta^2) + \mathcal{O}((\phi,\theta)^4) \quad \text{for} \quad (\phi,\theta) \to 0\,.$$

The primary order of bilinear interpolation is 2. Similarly we find that the secondary order is also 2.

**Remark:** The same holds for the full-weighting (FW) restriction (9-point restriction).

**Example 7.30** *Cubic interpolation, $n = 1$*

$$A_h = a_h \star = \left[-\frac{1}{16}, 0, \frac{9}{16}, 1, \frac{9}{16}, 0, -\frac{1}{16}\right]_h .$$

$$\begin{aligned} A(\phi) &= \tfrac{1}{Q}\widehat{A_h}(\phi/h) = \tfrac{1}{2}\widehat{A_h}(\phi/h) \\ &= \tfrac{1}{16}[8 + 9\cos\phi - \cos 3\phi] \\ &= \tfrac{1}{16}[8 + 9\cos\phi - (4\cos^3\phi - 3\cos\phi)] \\ &= \tfrac{1}{16}[8(1+\cos\phi) + 4\cos\phi(1-\cos^2\phi)] \\ &= \tfrac{1}{16}[16\cos^2(\phi/2) + 4\cos\phi\sin^2\phi] \\ &= \tfrac{1}{16}[16\cos^2(\phi/2) + 16\cos\phi\sin^2(\phi/2)\cos^2(\phi/2)] \\ &= \cos^2(\phi/2)[1 + \cos\phi\sin^2(\phi/2)] \\ &= (1 - \tfrac{1}{2}(\phi/2)^2 + ...)^2[1 + (1 + ...)((\phi/2)^2 + ...)] \\ &= 1 + \mathcal{O}(|\phi|^4) \quad \text{for} \quad \phi \to 0\,. \end{aligned}$$

The primary order of cubic interpolation in 1-D is 4. Similar considerations for $A(\pi - \phi)$ yield secondary order 4.

**Remark:** One can show that also for cubic interpolation in two dimensions, both primary and secondary order are 4.

**Example 7.31** *Half-weighting, $n = 2$*

$$A_h = a_h \star = \begin{bmatrix} & 1/8 & \\ 1/8 & 1/2 & 1/8 \\ & 1/8 & \end{bmatrix}_h .$$

$$
\begin{array}{ll}
\widehat{A_h}(\phi/h,\theta/h) & = \frac{1}{2} + \frac{1}{8}e^{i\phi} + \frac{1}{8}e^{-i\phi} + \frac{1}{8}e^{i\theta} + \frac{1}{8}e^{-i\theta} \\
 & = \frac{1}{2} + \frac{1}{4}\cos\phi + \frac{1}{4}\cos\theta \\
A(\phi,\theta) & = \frac{1}{Q}\widehat{A_h}(\phi/h,\theta/h) \\
 & = \frac{1}{4}\widehat{A_h}(\phi/h,\theta/h) \\
 & = \frac{1}{16}[2 + cos\phi + \cos\theta] \\
 & = \frac{1}{8}(\cos^2(\phi/2) + \cos^2(\theta/2)) \\
 & = \frac{1}{8}[1 - \frac{1}{2}(\phi/2)^2 + 1 - \frac{1}{2}(\theta/2)^2 + \mathcal{O}((\phi,\theta)^4)] \\
 & = \frac{1}{4} - \frac{1}{64}(\phi^2+\theta^2) + \mathcal{O}((\phi,\theta)^4) \quad \text{for} \quad (\phi,\theta)\to 0,
\end{array}
$$

hence the primary order is 2.

$$
\begin{array}{ll}
A(\pi-\phi,\theta) & = \frac{1}{8}(\sin^2(\phi/2) + \cos^2(\theta/2)) \\
 & = \frac{1}{8}(\frac{1}{2}(\phi/2)^2 + 1 - \frac{1}{2})^2 + \mathcal{O}((\phi,\theta)^4) \\
 & = \frac{1}{8} + \frac{1}{64}(\phi^2-\theta^2) + \mathcal{O}((\phi,\theta)^4) \quad \text{for} \quad (\phi,\theta)\to 0,
\end{array}
$$

hence the secondary order is 0.

## 7.10 Fourier analysis in the case of mutual influencing frequencies

In Section 7.8 we saw that $\widehat{u_h}(\omega)$, $\omega\in T_h^n$ could also be denoted by a $Q$-vector $\vec{\mathbf{FT}}(u_h)(\omega)$, $\omega\in T_H^n$ by taking the harmonic frequencies together

$$
\vec{\mathbf{FT}}(u_h)(\omega) = \begin{pmatrix} \widehat{u_h}(\omega) \\ \vdots \\ \widehat{u_h}(\omega + 2\pi p/qh) \\ \vdots \end{pmatrix}, \quad \omega\in T_H^n = T_{qh}^n,\ p\in[0,q). \tag{7.36}
$$

Consistent with this notation we also introduce the notation

$$
\vec{\mathbf{FT}}(A_h)(\omega) = \text{ diag } (\widehat{A_h}(\omega + 2\pi p/qh)), \quad \omega\in T_{qh}^n,\ p\in[0,q) \tag{7.37}
$$

to replace $\widehat{A_h}(\omega)$, $\omega\in T_h^n$.
**Remark:** $\widehat{A_h}(\omega)\widehat{u_h}(\omega)$, $\omega\in T_h^n$, is now denoted as $\vec{\mathbf{FT}}(A_h)(\omega)\ \vec{\mathbf{FT}}(u_h)(\omega)$, $\omega\in T_H^n$.

The above notation enables us to apply the usual Fourier analysis in the case of mutual influencing of harmonic frequencies. For example, let $A_H$ be a Galerkin-approximation of $A_h$, i.e.

$$
A_H = R_{Hh}A_hP_{hH}, \tag{7.38}
$$

where $R_{Hh}$ and $P_{hH}$ are homogeneous restrictions and prolongations, and $A_h$ a Toeplitz operator. We find

$$
\mathbf{FT}(A_H)(\omega) = \vec{\mathbf{FT}}(R_{Hh})(\omega)\ \vec{\mathbf{FT}}(A_h)(\omega)\ \vec{\mathbf{FT}}(P_{hH})(\omega), \tag{7.39}
$$

where $\vec{\mathbf{FT}}(R_{Hh})(\omega)$ is the $(1\times Q)$-matrix and $\vec{\mathbf{FT}}(P_{hH})(\omega)$ the $(Q\times 1)$- matrix (cf. (7.31) and (7.35)).

Similarly we find for the coarse-grid-correction (CGC) operator

$$
M_h^{CGC} = I_h - P_{hH}L_H^{-1}R_{Hh}L_h \tag{7.40}
$$

the following FT:

$$\vec{\mathbf{FT}}(M_h^{CGC})(\omega) = I - \vec{\mathbf{FT}}(P_{hH})(\omega)\ \mathbf{FT}(L_H^{-1})(\omega)\ \vec{\mathbf{FT}}(R_{Hh})(\omega)\ \vec{\mathbf{FT}}(L_h)(\omega),\ \omega \in T_H^n\,. \quad (7.41)$$

For $n = 2, q = 2$ (hence $Q = 4$) (7.41) is a $(4 \times 4)$- matrix:

Because $\|\widehat{u_h}\|_{L^2(T_h^n)} = \|u_h\|_{l_h^2(\mathbb{Z}_h^n)}$ , we have the following relation for the norms of the amplification operators:

$$\begin{aligned} \|M_h\|_{2\leftarrow 2} &= \sup_{u_h \in l_h^2(\mathbb{Z}_h^n)} \frac{\|M_h u_h\|_2}{\|u_h\|_2} = \\ &= \sup_{u_h} \frac{\| \vec{\mathbf{FT}}(M_h u_h)\|_{L^2}}{\| \vec{\mathbf{FT}}(u_h)\|_{L^2}} = \sup_{\widehat{u_h}} \frac{\| \vec{\mathbf{FT}}(M_h)(\omega)\ \widehat{u_h}(\omega)\|_{L^2(T_h^n)}}{\|\widehat{u_h}(\omega)\|_{L^2(T_h^n)}} \\ &\leq \sup_{\omega \in T_H^n} \| \vec{\mathbf{FT}}(M_h)(\omega)\|_2\,. \end{aligned}$$

Since there is no coupling between non-harmonic-related frequencies we also have

$$\rho(M_h) = \sup_{\omega \in T_H^n} |\lambda_{\max}(\vec{\mathbf{FT}}(M_h)(\omega))|\,. \quad (7.42)$$

## 7.11 Requirements for transfer operators

Let $L$ be an ordinary differential operator of order $M$:

$$L = \sum_{m=0}^{M} c_m \frac{d^m}{dx^m}. \quad (7.43)$$

Its FT or symbol is then given by

$$\widehat{L}(\omega) = \sum_{m=0}^{M} c_m (i\omega)^m, \quad (7.44)$$

(Section 7.6). Let $L_h$ and $L_H$ be two discretisations of $L$ of order $\tilde{p}$ on $\mathbb{Z}_h^n$ and $\mathbb{Z}_H^n$ respectively. Then $\widehat{L_h}(\omega)$ is a trigonometric polynomial of degree $M$ in $\omega$:

$$\widehat{L_h}(\omega) = \sum_{m=0}^{M} \frac{c_m}{h^m} S_m(h\omega), \quad (7.45)$$

where $S_m$ is the FT of the $m^{\text{th}}$ order part of the discrete operator $L_h$. Furthermore, we have

$$|\widehat{L_h}(\omega) - \widehat{L}(\omega)| = \mathcal{O}(h^{\tilde{p}}) \text{ for } h \to 0\,, \quad (7.46)$$

(see (7.25)) and

$$S_m(\omega h) \to (\omega h)^m, \quad \text{for} \quad \omega h \to 0,$$

$$S_m(\omega h) = \mathcal{O}(1), \quad (-\pi \leq \omega h \leq \pi).$$

We consider a fixed $\omega$ and $h \to 0$ and we define $\theta = \omega h$.

Using the above we find:

$$\left\| \frac{\widehat{L_h}(\omega)}{\widehat{L_H}(\omega)} \right\| = \left\| \frac{\widehat{L}(\omega) + \mathcal{O}(h^{\tilde{p}})}{\widehat{L}(\omega) + \mathcal{O}(H^{\tilde{p}})} \right\| = 1 + \mathcal{O}(h^{\tilde{p}})(h \to 0),$$

and
$$\left\| \frac{\widehat{L_h}((\omega+2\pi p/qh)}{\sum_{m=0}^{M} \frac{c_m}{(qh)^m}(\omega qh)^m} \right\| = \frac{\mathcal{O}(h^{-M})}{\mathcal{O}(h^{-M}(\omega h)^M)} = \mathcal{O}(h^{-M}) \text{ for } h \to 0)\,.$$

Now we are interested in the behaviour of the CGC-amplification factors for harmonic frequencies in a neighbourhood of the origin. For simplicity we restrict ourselves to the case $n=1, q=2$. The amplification operator for the error in one CGC-step is given by (7.40), its FT, $\widehat{M_h^{CGC}}(\omega)$, is given in (7.41) and can now be denoted as

$$\widehat{M}(\theta) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} \widehat{P}(\omega) \\ \widehat{P}(\omega+\pi/h) \end{pmatrix} \widehat{L_H^{-1}}(\omega)(\widehat{R}(\omega), \widehat{R}(\omega+\pi/h)) \begin{pmatrix} \widehat{L_h}(\omega) & 0 \\ 0 & \widehat{L_h}(\omega+\pi/h) \end{pmatrix}.$$

If we assume that the primary and secondary orders of the prolongation (restriction) are given by $m_1$ and $m_2$ ($n_1$ and $n_2$), then we find for $\theta \to 0$:

$$\begin{aligned} \widehat{M}(\theta) &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1+\mathcal{O}(\theta^{m_1}) \\ \mathcal{O}(\theta^{m_2}) \end{pmatrix} (1+\mathcal{O}(\theta^{n_1}), \mathcal{O}(\theta^{n_2})) \begin{pmatrix} 1+\mathcal{O}(\theta^{\hat{p}}) & 0 \\ 0 & \mathcal{O}(\theta^{-M}) \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{O}(\theta^{m_1}) + \mathcal{O}(\theta^{n_1}) + \mathcal{O}(\theta^{\hat{p}}) & \mathcal{O}(\theta^{n_2-M}) \\ \mathcal{O}(\theta^{m_2}) & 1+\mathcal{O}(\theta^{n_2+m_2-M}) \end{pmatrix}. \end{aligned}$$

It is clear that convergence requires $m_1 > 0, n_1 > 0$ and $\tilde{p} > 0$. Further, for $\theta \to 0$ the eigenvalues of $M(\theta)$ are

$$\lambda_1 = 1 + \ldots\,, \text{ and } \lambda_2 = \mathcal{O}(\theta^{n_2+m_2-M}).$$

For convergence of the CGC-process $n_2 + m_2 \geq M$ is required. Further, $n_2 \geq M$, $m_2 \geq 0$ is required for $\|M(\theta)\|$ to be bounded. Similarly we derive the requirements $m_2 \geq M$, $n_2 \geq 0$ for $\|\overline{M}(\theta)\|$ to be bounded.

# Appendix A

# Discretisation

## A.1 Integral equations

Let$\Omega \subset \mathbb{R}^n$ be a bounded domain, then the equation

$$u(x) = \int k(x,\ y\ u(y))\ dy \tag{A.1}$$

where $k : \Omega \times \Omega \times \mathbb{C}^m \to \mathbb{C}^m$ is a given function and $u : \Omega \to \mathbb{C}^m$ is the unknown function, is called the *Urysohn integral equation*; $k(x, y, u)$ is called the *kernel function* of the equation.

If the kernel-function is linear (or, more precisely, affine) in the variable $u$, then the equation is a *Fredholm integral equation of the $2^{nd}$ kind* and can be written as

$$u(x) = \int k(x,y)\ u(y)\ dy + f(x) \tag{A.2}$$

where $u : \Omega \to \mathbb{R}^m$ is the unknown function, and $k : \Omega \times \Omega \to \mathbb{C}^{m\times m}$ and $f : \Omega \to \mathbb{C}^m$ are given.

Although many properties of the integral equations (A.1) and (A.2) can be treated for the vector-equation ($m > 1$) and for more-dimensional domain ($n > 1$), we shall restrict ourselves mainly to scalar equations ($m = 1$) on the one-dimensional domain $\Omega = [0, 1]$. Solution methods for $m > 1$ and $n > 1$ are generally analogous to those for $m = 1$ and $n = 1$ but the general treatment would complicate the notation. For simplicity, we also we consider only integral operators over the field $\mathbb{R}$ instead of over the field $\mathbb{C}$.

Equation (A.2) can be symbolically written as

$$u = Ku + f, \tag{A.3}$$

where $K$ denotes the linear integral operator defined by the function $k$.

If the function $k(x,y)$ is bounded in y and is differentiable in $x$, then the operator $K$ transform any integrable function $u$ on into a differentiable function $Ku$ on $\Omega$. We say that the operator $K$ has a smoothing property. This smoothing property is an essential feature for many integral operators and we shall exploit it in the use of our multigrid methods.

In order to formulate this smoothing property in better mathematical terms, we recall a number of theorems and definitions from functional analysis (cf. [24, 21]).

**Definition A.1** A *precompact set* is a set from which each enumerable infinite sequence contains a convergent subsequence.

**Definition A.2** A *compact operator* from a Banach space into a Banach space is an operator which maps any bounded set into a precompact set.

**Theorem A.3** ([24],thm.7.4). Let $\Omega \subset I\!R^n$ be a bounded domain and let $k(x,y) \in \bar{C}(\Omega \times \Omega)$ (i.e. $k$ is a continuous function on the closure of $\Omega \times \Omega$), then the operator $K : \bar{C}(\Omega) \to \mathbb{C}(\Omega)$ is a compact operator.

**Definition A.4** A linear integral equation *has an $L^2$-kernel* if $k(s,t) : \Omega \times \Omega \to \mathbb{C}$ satisfies

$$\int\int \quad |k(s,t)|^2 ds dt \;=\; \|K\|^2 < \infty,$$

$$\int \quad |k(s,t)| ds < \infty \;\; \forall t,$$

$$\int \quad |k(s,t)| dt < \infty \;\; \forall s,$$

**Theorem A.5** ([24],thm.7.5). Let $\Omega \subset I\!R^n$ be a bounded domain and let $k(x,y) \in L^2(\Omega \times \Omega)$ (i.e. $k$ is a square integrable function over $\Omega \times \Omega$), then $K : L^2(\Omega) \to L^2(\Omega)$ is a compact operator.

**Definition A.6** The *adjoint* $K^*$ of a linear integral operator $K$ is the operator with the kernel function $k^*(s,t) = \overline{k(t,s)}$, $K$ is called *Hermitian* if $K = K^*$, $K$ is called *normal* if $KK^* = K^*K$, $\phi \in L^2(\Omega)$ is called an *eigenfunction* and $\lambda \in \mathbb{C}$ is an *eigenvalue* of $K$ *if*

$$\lambda\phi \;=\; K\phi, \;\; \phi \quad \neq 0.$$

**Theorem A.7** ([24],thm.11.2). A compact operator in a Hilbert space has at most an enumerable infinite set of eigenvalues which may be dense only at $\lambda = 0$. Each non-zero eigenvalue has a finite multiplicity.

**Theorem A.8** ([21],thm.7.3.1-7.3.3) For Hermitian operators with $L^2$-kernels
— eigenvalues are real;
— eigenvalues belonging to distinct eigenvalues are orthogonal to each other;
for each non-zero eigenvalue there is a finite orthogonal base of — eigenfunctions: the dimension of the base is the *multiplicity* of the eigenvalue;
eigenvalues form an enumerable sequence $\lambda_i$ — and, counting multiplicity, we have

$$\sum_{i=1} \lambda_i^2 \;=\; \|K\|^2 < \infty. \tag{A.4}$$

Usually eigenvalues $\lambda_i$ and corresponding eigenfunctions $\phi_i$ are ordered such that

$$|\lambda_1| \geq |\lambda_2| \geq \;\cdots\; .$$

The set $\{(\lambda_i, \phi_i) \mid \lambda_i \neq 0, \|\phi_i\| = 1\}$ is called a full orthonormal system of $K$.
**Remark:** A full orthonormal system is not necessarily complete. It may even be finite. e.g. with $k(x,y) = p(x).p(y)$ any function $Ku$ must be a scalar multiple of $p$. Thus there exists only one non-zero eigenvalue and the full orthonormal system is $(\|p\|^2, p/\|p\|)$.
**Remark:** If the eigenfunctions $\{\phi_i |\; \lambda_i \neq 0\}$ do not span the entire (separable Hilbert-) space, then we can find – orthogonal to the Span$\{\phi_i\}$ – a system of orthonormal functions $\{\psi_j\}$ such that $\{\phi_i\} \cup \{\psi_j\}$ span the entire space. We notice that $K\psi_j = 0$ for all $\psi_j$.

**Theorem A.9** ([21],thm.7.4.3). Any Hermitian $L^2$-kernel can be decomposed as

$$(Kx, y) \;=\; \sum_{n=1}^{\infty} \lambda_n (x, \phi_n)(\phi_n, y) \tag{A.5}$$

where $\{\lambda_n, \phi_n\}$ is its (full orthonormal) eigensystem and $(\cdot,\cdot)$ denotes the $L^2$-inner product.

**Remark:** For an arbitrary $L^2$-kernel $K$, both the operators $KK^*$ and $K^*K$ are Hermitian $L^2$-kernels. These Hermitian operators have the same set of non-negative eigenvalues. Thus we may denote by $\{\sigma_i^2, \phi_i\}$ the eigensystem of $KK^*$ and by $\{\sigma_i^2, \psi_i\}$ the eigensystem of $K^*K$. The system $\{\sigma_i, \phi_i, \psi_i\}$, with $\sigma_i > 0$, is called the *singular system* of $K$.

**Theorem A.10** ([21],thm.8.3.2). Let $K$ be a $L^2$-kernel and let $x$ and $y$ be $L^2$-functions then $(Kx, y)$ can be decomposed as

$$(Kx, y) = \sum_n^\infty \sigma_n (x, \phi_n)(\psi_n, y). \tag{A.6}$$

Further

$$\sum_{i=1}^\infty \sigma_i^2 \leq \|K\|^2$$

and

$$\sum_{i=1}^\infty \sigma_i^2 = \|K\|^2 \quad \text{iff } K \text{ is normal.}$$

**Remark:** The set $\{\chi_n\}_{n=0,1,2,\ldots} = \{1, \sin(\pi x), \cos(\pi x), \sin(2\pi x), \cos(2\pi x), \ldots\}$ is a complete orthonormal system in the Hilbert space $L^2[0,1]$. I.e. any function $x \in L^2[0,1]$ can be expressed as

$$x = \sum_{n=0}^\infty (x, \chi_n)\chi_n,$$

with

$$\|x\|^2 = \sum_{n=0}^\infty |(x, \chi_n)|2, \qquad \text{(Parseval equality).}$$

Hence any orthonormal set of (eigen-) functions $\{\phi_j\}$ or $\{\psi_j\}$ can be expressed as

$$\begin{aligned} \phi_j &= \textstyle\sum_{n=0}^\infty a_{jn}\chi_n \quad \text{with} \quad \sum_n |a_{jn}|^2 = 1. \\ \psi_j &= \textstyle\sum_{n=0} b_{jn}\chi_n \qquad\qquad\; \sum_n |b_{jn}|^2 = 1 \end{aligned}$$

and from (A.6) we derive

$$(Kx, y) = \sum_j (x, \phi_j)\sigma_j(\psi_j, y) = \sum_{jnm} (x, \chi_n)\bar{a}_{jn}\sigma_j b_{jm}(\chi_m, y).$$

The bound $\Sigma\sigma_j^2 \leq \|K\|^2$ implies $\lim_{j\to m} \sigma_j = 0$. From this its follows (i) that high enough frequency components in $u$ (i.e. $\sum_{m=k}^\infty (\chi_m, u)\chi_m$ with $m$ large enough) will have an arbitrarily small effect in $Ku$ and (ii) that high enough frequency components in $Ku$ will be arbitrarily small.

## A.2 Discretisation of integral equations

In order to discretise the problem (A.2) with $n = k = 1, \Omega = [0,1]$, we consider the finite set of points

$$\Omega_h = \{x_0, x_1, \ldots, x_N | x_i = i/N\}.$$

To discretise $u : \Omega \to I\!R$, we consider

$$u_h \ : \ \Omega_h \to I\!R,$$

$$u_h \ = \ \{u_0, u_1, \ldots, u_N\},$$

and we replace

$$Ku(x) \ = \ \int_0^1 k(x,y)u(y)dy$$

by

$$K_h u_h(x) \ = \ \sum_{x_\ell \in \Omega_h} w_\ell k(x, x_\ell) u_\ell, \quad w_\ell \neq 0.$$

The discretised equation (A.2) now reads

$$u_m \ = \ \sum_{\ell=0}^{N} w_\ell k(x_m, x_\ell) u_\ell \ + \ f(x_m), \quad m = 0, 1, \ldots, N,$$

which is a simple $(N+1) \times (N+1)$ matrix problem:

$$\sum_{\ell=0}^{N} [\delta_{m,l} - w_\ell k(x_m, x_\ell)] u_\ell = f(x_m).$$

This discretised equation we denote symbolically by

$$A_h u_h = f_h.$$

How well the values $\{u_i\}$ approximate the values $\{u(x_i)\}$ depends on

1. the number of intervals $N$,
2. the choice of $\{w_\ell\}$,
3. the functions $k(x,y)$ and $f(x)$.

Typically error estimates are of the form

$$\max_i |u_i - u(x_i)| \leq CN^{-p}.$$

Thus, the difficulty we encounter when we want to approximate (accurately) the solution of the integral equation is the large (dense) system of equations to solve.

## A.3 Differential equations

Let $\Omega \subset I\!R^n$, then a differential boundary value problem consists of
(1) a partial differential equation for an unknown $u : \Omega \to I\!R^k$;
(2) boundary conditions.
Examples of the differential equation are:

1. The Helmholtz equation

$$-\Delta u(x) + cu(x) = g(x) \quad x \in \Omega \tag{A.7}$$

where $\Delta = \sum_{i=1}^{n} (\frac{\partial}{\partial x_i})^2$ is the Laplacian operator. E.g. with $n = 2$ the equation reads

$$-u_{xx} - u_{yy} + cu = g. \tag{A.8}$$

2. The general linear elliptic differential equation of second order.

$$Au \equiv \sum_{i,j=1}^{n} a_{ij}(x) \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{n} a_i(x) \frac{\partial u}{\partial x_i} + a(x)u = g, \tag{A.9}$$

where the functions $a_{ij}, a_i, a : \Omega \to I\!R$ are the coefficients of the equation. The ellipticity condition is

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j > 0, \qquad \forall x \in \Omega. \quad \forall(\xi_1, \xi_2, \ldots, \xi_n) \in I\!R^n \; .$$

3. The biharmonic equation

$$\Delta^2 u(x) \;=\; g(x), \qquad x \in \Omega. \tag{A.10}$$

**Remark:** The order (i.e. the highest derivative available in the equation) of an elliptic equation is always even. Usually it is denoted by $2m$. For an elliptic equation to have a unique solution, $m$ (boundary) conditions must be given on the whole boundary $\partial\Omega$ of $\Omega$ . Examples of boundary conditions are

1. Dirichlet boundary conditions

$$(\frac{\partial}{\partial n})^j u(x) \;=\; g_j(x) \;, \quad j = 0, 1, \ldots, m-1, \quad x \in \partial\Omega, \tag{A.11}$$

where $\frac{\partial u}{\partial n} = \mathbf{n}.\nabla(u)$ is the outward normal derivatives; $\mathbf{n}$ is the outer normal direction.

2. Neumann boundary conditions, $m = 1$,

$$\frac{\partial u}{\partial n} = h(x), \qquad x \in \partial\Omega. \tag{A.12}$$

Notice that the Neumann boundary conditions, together with $a(x) \equiv 0$ in (A.9), does *not* give an unique solution. So, additional requirements may be necessary!

3. Boundary conditions "of the third kind", $m = 1$,

$$\frac{\partial u(x)}{\partial n} + \alpha(x)u(x) = \gamma(x), \qquad x \in \partial\Omega. \tag{A.13}$$

4. Mixed boundary conditions, $m = 1$

$$\begin{array}{ll} u(x) = g(x) & , \quad x \in \partial\Omega_1, \\[2ex] \frac{\partial u(x)}{\partial n} + \alpha(x)u(x) = \gamma(x) & , \quad x \in \partial\Omega_2, \end{array} \tag{A.14}$$

with $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2, \quad \partial\Omega_1 \cap \partial\Omega_2 = \emptyset$

**Remark:** In contrast with the integral operators, differential operators transform smooth (differentiable) functions into less smooth functions. Usually, for equations of order $2m$ we shall have for $s \geq 0$

$$A \;:\; C^{2m+s}(\Omega) \to C^s(\Omega).$$

## A.4 The weak formulation of a differential equation

Boundary value problems for differential equations often can be given in a variational formulation.

**Example A.11**
Helmholtz equation with Neumann boundary conditions can be formulated as

$$B(u,v) \;=\; f(v)\;, \qquad \text{for all } v \in H^1(\Omega), \tag{A.15}$$

where

$$f(v) = \int_\Omega v(x)g(x)dx - \oint_{\partial\Omega} v(x)h(x)\,dx$$

is a linear functional in $v; f : H^1(\Omega) \to I\!R$;

$$B(u,v) = \int_\Omega \nabla v(x)\cdot\nabla u(x) + v(x)cu(x)\,dx - \oint_{\partial\Omega} v(x)\;\mathbf{n}\cdot\nabla u(x)\,dx$$

is a bilinear form in $u$ and $v; B : H^1(\Omega)\times H^1(\Omega) \to I\!R$. Here $H^1(\Omega)$ is the linear space of all functions $u$ which $u$ and $\nabla u$ are square integrable (see Section B.3).

For homogeneous Dirichlet boundary conditions we have similar formulas with $H^1$ replaced by $H_0^1$.

**Definition A.12** The Sobolev space $H^k(\Omega), k = 0,1,2,\ldots$, is the normed linear space of all (generalised) functions with finite norm $\|u\|_{H^k(\Omega)}$,

$$\|u\|^2_{H^k(\Omega)} = \sum_{|\alpha|\le k} \|(\frac{\partial}{\partial x_1})^{\alpha_1}(\frac{\partial}{\partial x_n})^{\alpha_2}\cdots(\frac{\partial}{\partial x_n})^{\alpha_n}u\|^2_{L^2(\Omega)},$$

$|\alpha| = \sum_{i=1}^n \alpha_i$ where $\alpha_i$ are non-negative integers.

**Remark:** $H^k(\Omega)$ is a Hilbert-space with inner product

$$(u,v)_{H^k(\Omega)} \;=\; \sum_{|\alpha|\le k}((\frac{\partial}{\partial x})^\alpha u, (\frac{\partial}{\partial x})^\alpha v)_{L^2\Omega}. \tag{A.16}$$

**Remark:** If $B$ is *symmetric*:

$$B(u,v) \;=\; B(v,u) \qquad u,v \in H^1(\Omega),$$

and $B$ is *positive definite*:

$$B(u,v) > 0 \qquad \forall u \in H^1(\Omega) \qquad u \ne 0,$$

then the solution of

$$B(u,v) \;=\; f(v) \qquad \forall v \in H^1(\Omega)$$

minimises the functional

$$J(u) \;=\; B(u,u) - 2f(u). \tag{A.17}$$

**Example A.13**
The Helmholtz equation with $c \ge 0$ and homogeneous Dirichlet boundary conditions is symmetric and positive definite.

**Remark:** Let $B(u,v)$ be the bilinear form associated with (A.9), then a function $u \in H_0^1(\Omega)$ (1) can satisfy the condition

$$B(u,v) \;=\; f(v) \qquad \forall v \in H_0^1(\Omega), \tag{A.18}$$

(2) and is not necessarily a $C^2(\Omega)$ function. By a partial integration we easily see that any solution of Helmholtz equation (A.7) with the Neumann boundary conditions (A.12) satisfies the equation (A.15). However it is possible that a solution $u \in H^1(\Omega)$ of (A.15) exists, which is not a solution $u \in C^2(\Omega)$ of (A.7) and (A.12). The variational or *weak formulation* (A.15) of the boundary value problem *is a generalisation* of the classical formulation of the same problem.

**Remark:** The equation: find $u \in H_0^1(\Omega)$ such that

$$B(u,v) = f(v) \qquad \forall v \in H_0^1(\Omega)$$

can be seen as an infinite-dimensional linear system. We can denote this equation as

$$Au \;=\; f,$$

where $A$ is a linear operator $A : H_0^1(\Omega) \to [H_0^1(\Omega)]^{\mathrm{DUAL}}$. The Banach space $[H_0^1(\Omega)]^{\mathrm{DUAL}}$, i.e. the space of all bounded linear functionals on $H_0^1(\Omega)$, is also denoted by $H^{-1}(\Omega)$. We easily see that

$$H_0^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega)$$

if we identify functions $f \in L^2(\Omega)$ with the linear functionals $f(v) = (f,v)_{L^2(\Omega)}$. Clearly $L^2(\Omega)$ and $H^{-1}(\Omega)$ contain functions that are not in $H_0^1(\Omega)$. We call these functions less smooth than those in $H_0^1(\Omega)$. Generalised functions such as the Dirac-delta function, defined by

$$\int_\Omega \delta_x(y)\phi(y)dy \;=\; \phi(x),$$

are bounded linear functionals on $H_0^1(\Omega)$. They can be considered as functions that are contained in $H^{-1}(\Omega)$ but not in $L^2(\Omega)$.

Under sufficient conditions for $B$ (e.g. $B$ is symmetric and positive definite), for any bounded linear functional $f \in H^{-1}(\Omega)$ we can find a solution $u \in H_0^1(\Omega)$ which satisfies (A.18). Than A is *invertible*:

$$A^{-1} : H^{-1}(\Omega) \to H_0^1(\Omega)$$

exists.
The problem (A.18) is called *regular* if $A^{-1}$ is a bounded operator

$$A^{-1} : H^{-1+s}(\Omega) \to H_0^{1+s}(\Omega), \qquad s > 0.$$

Typically $A^{-1}$ maps less smooth functions into more smooth functions. With a sufficient degree of regularity (i.e. for sufficient large $s$) we find

$$A^{-1} \;:\; L^2(\Omega) \to H^2(\Omega).$$

## A.5 Discretisation of differential equations

In order to discretise a differential boundary value problem, we can start either from the classical formulation or from the variational formulation of the differential equation. The former leads to the Finite Difference Method (FDM) for the discretisation, the latter to the Finite Element Method (FEM). In a number of cases both discretisation methods end up with the same discretisation of a given problem. In order not to obscure the notation we restrict ourselves to 2-dimensional scalar problems ($n = 2, k = 1$).

THE FINITE DIFFERENCE METHOD

Instead of the original domain of definition $\Omega$, here we consider a finite, discrete, set of points in $\Omega$

$$\overline{\Omega_h} \;=\; \{x_{ij} | x_{ij} \in \overline{\Omega},\; x_{ij} \;=\; (i/N_1, j/N_2)\}.$$

To discretise the function $u : \overline{\Omega} \to I\!R$, we consider $u_h : \overline{\Omega_h} \to I\!R$:

$$u_h \;=\; \{u_{ij}\}.$$

The differentials in the original differential equation are replaced by difference approximations.

**Example A.14**
With $x_{ij} = (ih, jh)$ e.g. we set

$$\frac{u_{i+1,j} - u_{i-1,j}}{2h} \qquad \text{for } u_x$$

and

$$\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h^2} \qquad \text{for } u_{yy} \text{ etc..}$$

**Example A.15** *The Helmholtz-equation with Dirichlet boundary conditions.*
For each $x_{ij} \in \overline{\Omega_h} \cap \Omega$ we find an equation

$$4u_{ij} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1} + ch^2 u_{ij} = h^2 f(x_{ij}) \tag{A.19}$$

and for each $x_{ij} \in \partial\Omega_h = \overline{\Omega_h} \backslash \Omega$

$$u_{ij} \;=\; g(x_{ij}). \tag{A.20}$$

As $\partial\Omega$ contains all the "neighbours" of points $x_{ij}$ in $\Omega_h$, the system (A.19)-(A.20) determines as many equations as unknowns $u_{ij}$. They form a linear system of equations which we denote by

$$A_h u_h \;=\; f_h :$$

From this equation $u_h = \{u_{ij}\}$ can be computed. Under suitable conditions the values $u_{ij}$ approximate the value $u(x_{ij})$.

THE FINITE ELEMENT METHOD

In this case our starting point is the variational formulation: find $u \in H^1(\Omega)$ such that

$$B(u, v) \;=\; f(v) \qquad \text{for all } v \in H^1(\Omega). \tag{A.21}$$

Now we select from the space $H^1(\Omega)$ a finite dimensional subspace $S_h \subset H^1(\Omega)$ and we replace (A.21) by: find $u_h \in S_h$ such that

$$B(u_h, v_h) \;=\; f(v_h) \qquad \text{for all } v_h \in S_h.$$

**Example A.16**
We assume that $\Omega$ allows a triangulation. As $S_h$ we consider the space of all functions that are continuous on $\Omega$ and linear over all small triangles into which $\Omega$ is partioned.
A basis of $S_h$ is formed by the set of "hat-functions" $\phi_{ij}$, where $\phi_{ij}$ is a function which takes the value 1 at $x_{ij}$ and takes the value 0 at all other vertices of triangles. Further $\phi_{ij}$ is piecewise linear. We see that $\phi_{ij}$ is nonzero only on those triangles that share the vertex $x_{ij}$ and the (finite) basis in $S_h$ is the set of functions

$$\{\phi_{ij} | x_{ij} \in \Omega \cup \partial\Omega\}$$

We can write any function $u_h \in S_h$ as

$$u_h(x) \ = \ \sum_{i,j} u_{ij} \phi_{ij}(x)$$

and the discretised problem reads

$$\sum_{i,j} B(\phi_{ij}, \phi_{k\ell}) u_{ij} = f(\phi_{k\ell}) \text{ for all } \phi_{k,\ell} \in B.$$

This linear system we denote also by

$$A_h u_h \ = \ f_h.$$

Under suitable conditions the function $u_h$ approximates the solution $u$ of (A.21) and we can find error estimates which are typically of the form

$$\|u_h - u\|_{L^2(\Omega)} \le Ch^2 \|u\|_{H^2(\Omega)} \le Ch^2 \|f\|_{L^2(\Omega)}.$$

# Appendix B

# Function Spaces

## B.1 Continuous functions and Hölder spaces

Let $\Omega \subset \mathbb{R}^d$ be a bounded open region in the $d$-dimensional real space $\mathbb{R}^d$. We introduce the notation $\mathbb{K}$ for the space $\mathbb{R}$ or $\mathbb{C}$. By

$$C(\Omega, \mathbb{K}^m) \text{ or } C(\Omega)$$

we denote

$$\{f : \Omega \to \mathbb{K}^m \mid f \text{ continuous}\}.$$

A *multi-index* or *multi-integer* $\alpha = (\alpha_1, ..., \alpha_d)$, $\alpha_i \geq 0$, $i = 1, ..., d$. Its absolute value is defined as

$$|\alpha| = \alpha_1 + \alpha_2 + ... + \alpha_d.$$

We also introduce the following notation for partial derivation with respect to $x_1, ..., x_d$ of order $|\alpha|$:

$$D^\alpha = D_1^{\alpha_1} \cdots D_d^{\alpha_d} = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_d}\right)^{\alpha_d} = (\partial x_1^{\alpha_1})_d^{\alpha_d} = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} ... \partial x_d^{\alpha_d}}.$$

Now we are in a position to define

$$C^k(\Omega) = \{u : \Omega \to \mathbb{K}^m \mid D^\alpha u \in C(\Omega) \text{ for all } |\alpha| \leq k\}$$

$$k = 0, 1, 2, ...\cdot$$

A norm is defined on this space $C^k(\Omega)$ by

$$\|u\|_{C^k(\Omega)} = \sup_{x \in \Omega,\, |\alpha| \leq k} |D^\alpha u(x)|_\infty,$$

with $|\cdot|_\infty$ the supremum norm on $\mathbb{K}^m$.
Next we define for $k = 0, 1, 2, ...$ (cf. [24], p.27):

$$C^k(\bar{\Omega}) = \{u \in C^k(\Omega) \mid D^\alpha u \text{ has a continuous extension on } \partial\Omega \text{ for all } |\alpha| \leq k\},\ k = 0, 1, 2, \ldots.$$

The following space is a Banach space with respect to the norm $\|\cdot\|_{C^k(\Omega)}$ (cf. [24],thm.3.1):

$$\bar{C}^k(\Omega) = \{u \in C^k(\bar{\Omega}) \mid \|u\|_{C^k(\Omega)} < \infty\}.$$

We now extend the concept of the spaces $\bar{C}^k(\Omega)$ to spaces $\bar{C}^s(\Omega)$, $s \geq 0$, of (possibly) fractional order $s$. For this purpose we assume :

$$u \in C(\Omega),\ \lambda \in [0,1),\ k \in \mathbb{N}\,.$$

We define

$$\|u\|_{C^\lambda(\Omega)} = \max\{\|u\|_{C^0(\Omega)}, \sup_{x,y\in\bar{\Omega},x\neq y} \frac{|u(x)-u(y)|}{|x-y|^\lambda}\};$$

and for $u \in C^k(\Omega)$ we define

$$\|u\|_{C^{k+\lambda}(\Omega)} = \max_{|\alpha|\leq k} \{\|D^\alpha u\|_{C^\lambda(\Omega)}\}\,.$$

Using the above we are now able to define

$$\bar{C}^{k+\lambda}(\Omega) = \{u \in \bar{C}^k(\Omega) \mid \ \|u\|_{C^{k+\lambda}} < \infty\}\,.$$

Now the spaces $\bar{C}^s(\Omega)$ are defined for all $s \geq 0$. For $s \in \mathbb{N}$ the definition coincides with the old definition. Finally we define:

$$\bar{C}_0^s(\Omega) = \{u \in \bar{C}^s(\Omega) \,|\, u \text{ has compact support in } \Omega\}\,.$$

**Remark:**

1. The spaces $\bar{C}^s(\Omega)$, $s \geq 0$, are Banach spaces.
2. For $s_1 > s_2 \geq 0$ we have $\bar{C}^{s_1}(\Omega) \subset \bar{C}^{s_2}(\Omega)$. We can now construct so called *scales of embedded Banach spaces* to connect two given spaces $\bar{C}^k(\Omega)$ and $\bar{C}^l(\Omega)$, $k, l \in \mathbb{N}$, $l > k$, so that for any $s \in \mathbb{R}, k < s < l$ we have $\bar{C}^l(\Omega) \subset \bar{C}^s(\Omega) \subset \bar{C}^k(\Omega)$.

## B.2 Generalised functions and derivatives

### B.2.1 Distributions

Let $\Omega \subset \mathbb{R}^d$ be the domain on which we want to define our functions. By $C_0^\infty(\Omega)$ we denote all (classical) functions, defined on $\Omega$ that are continuous and differentiable (in the classical sense) an infinite number of times in all directions. Further, all functions in $C_0^\infty(\Omega)$ have a compact support, i.e. for $\phi \in C_0^\infty(\Omega)$

$$\text{support}(\phi) = \{x \in \Omega \mid \phi(x) \neq 0\}$$

is a compact subset of $\Omega$.

**Definition B.1** The linear space of *distributions*, $\mathcal{D}$, is the dual space of $C_0^\infty(\Omega)$. I.e. the elements of $\mathcal{D}$ are the bounded linear functionals on $C_0^\infty(\Omega)$.

**Remark:** Thus, we have for $T \in \mathcal{D}$ that $T : C_0^\infty(\Omega) \to \mathbb{R}$ is a linear mapping. We may denote the duality pairing by $T(\phi)$, and we have

$$T(\phi + \psi) = T(\phi) + T(\psi), \quad \forall \phi, \psi \in C_0^\infty(\Omega)$$

$$T(\alpha\phi) = \alpha T(\phi), \quad \forall \alpha \in I\!R, \ \forall \phi \in C_0^\infty(\Omega).$$

Further, $\mathcal{D}$ is a linear space by:

$$T, S \in \mathcal{D} \Rightarrow T + S \in \mathcal{D}, \ \alpha T \in \mathcal{D},$$

and

$$(T + S)(\phi) = T(\phi) + S(\phi),$$
$$(\alpha T)(\phi) = \alpha T(\phi).$$

We may introduce a topology in $\mathcal{D}$ by defining

$$\lim_{n\to\infty} T_n = T \quad \text{in } \mathcal{D}$$

as

$$\lim_{n\to\infty} T_n(\phi) = T(\phi) \quad \forall \phi \in C_0^\infty(\Omega).$$

### B.2.2 Distributions as generalised functions

We notice that each $L^p(\Omega)$-function, $1 \le p \le \infty$, can be identified with a distribution by the following process.
Let $f \in L^p(\Omega)$ be arbitrary, then we associate with $f$ the distribution $T_f$ by

$$T_f(\phi) = \int f(x)\ \phi(x)\ d\Omega.$$

We see that $T_{f+g} = T_f + T_g$ and $T_{\alpha f} = \alpha T_f$, and hence the linear space $L^p(\Omega)$ can be identified with a subset of $\mathcal{D}$. Notice that $T_f$ with $T_f(\phi) = 0$ for all $\phi \in C_0^\infty(\Omega)$ corresponds with the zero-function in $L^p(\Omega)$.

We also see that $L^p(\Omega) \subset \mathcal{D}$ is a proper subset, because there are also elements in $\mathcal{D}$ that do not belong to $L^p(\Omega)$, e.g.

$$T_f(\phi) = \phi(x_0), \quad \text{with } x_0 \in \Omega \text{ fixed},$$

is clearly an element from $\mathcal{D}$ (because it is a linear functional on $C_0^\infty(\Omega)$) but it cannot be associated with an $L^p(\Omega)$-function.

### B.2.3 Derivatives of distributions

**Definition B.2** The partial derivative $D_i T$ of a distribution $T \in \mathcal{D}$, $(1 \le i \le d)$ is defined by

$$D_i T(\phi) = -T\left(\frac{\partial\phi}{\partial x_i}\right) \quad \forall \phi \in C_0^\infty(\Omega).$$

We notice that $\phi \in C_0^\infty$ and hence its derivative $\partial\phi/\partial x_i$ exists and is an element of $C_0^\infty(\Omega)$ itself. Further the operator $\partial/\partial x_i$ is linear and hence $D_i T$ is a well defined distribution.

**Remark:** If $T_f$ is identified with a differentiable function $f$, and $T_{f_i} \in \mathcal{D}$ is identified with its derivative $\partial f/\partial x_i$, then

$$D_i T_f = T_{f_i}.$$

This means that, if $f$ is a differentiable function, then the generalised function associated with the derivative is the generalised derivative of the function. In other words: we have generalised the notion of derivative, and –in the generalised sense– any (generalised) function

is differentiable.
**Proof:** We have to show

$$D_i T_f(\phi) = T_{f_i}(\phi) \quad \forall \phi \in C_0^\infty(\Omega)\,,$$

which is equivalent with

$$-T_f(D_i\phi) = T_{f_i}(\phi) \quad \forall \phi \in C_0^\infty(\Omega)\,,$$

which is equivalent with

$$-\int f\ D_i\phi\ d\Omega = \int f_i\ \phi\ d\Omega \quad \forall \phi \in C_0^\infty(\Omega)\,,$$

which is equivalent with

$$-\int f\,\frac{\partial \phi}{\partial x_i}\ d\Omega = \int \frac{\partial f}{\partial x_i}\ \phi\ d\Omega \quad \forall \phi \in C_0^\infty(\Omega)\,.$$

The latter is obviously true because of Gauss' theorem: with $n_i = \cos\alpha$, $\alpha$ the angle between the normal at the boundary and the $x_i$-axis,

$$\int \frac{\partial f}{\partial x_i}\,\phi\, d\Omega = \oint f\,\phi\, n_i ds - \int f\,\frac{\partial \phi}{\partial x_i}\ d\Omega$$

and $\phi(x) = 0$ at the boundary of $\Omega$. ∎

Of course, as a function $f \in \mathcal{D}$ can be in $L^p(\Omega)$ or not, depending on whether or not we can find a real number $\int |f|^p d\Omega$, which is zero if and only if $f$ is the zero-function, similarly a derivative of a function can be an $L^p(\Omega)$-function or not. An important class of functions is the set of functions which, together with all its derivatives of order $k$ belong to $L^p(\Omega)$. This leads to the notion of a Sobolev space, $W^{p,k}(\Omega)$, as we shall see in the next section.

## B.3 Sobolev spaces

### B.3.1 Definition

We denote the spaces of $p$-Lebesgue integrable functions $u : \Omega \to \boldsymbol{K}^m$ by

$$L^p(\Omega, \boldsymbol{K}^m) \text{ or, for short, } L^p(\Omega),\ p \geq 1.$$

These $L^p$ spaces are Banach spaces with respect to the standard $L^p$-norm $\|\cdot\|_{L^p(\Omega)}$ . We now define the *Sobolev spaces* $W^{p,k}(\Omega)$ to be the closure of $\bar{C}_0^\infty(\mathbb{R}^d)$ with respect to the following norm:

$$\|u\|_{W^{p,k}} = \left\{ \sum_{|\alpha|\leq k} \|D^\alpha u\|^p_{L^p(\Omega)} \right\}^{1/p} .$$

The Sobolev space $W_0^{p,k}(\Omega)$ is the closure of $\bar{C}_0^\infty(\Omega)$ with respect to the same norm. Finally we define the *Sobolev spaces:*

$$H^k(\Omega) = W^{2,k}(\Omega)$$

and

$$H_0^k(\Omega) = W_0^{2,k}(\Omega).$$

**Remark:** The Sobolev spaces $H^k(\Omega)$ and $H_0^k(\Omega)$ are Hilbert spaces with respect to the inner product

$$(u,v)_k = \sum_{|\alpha|\le k} \int D^\alpha u \overline{D^\alpha v} \, d\Omega \, .$$

**Remark:** For $k_1, k_2 \in I\!N$, $k_1 > k_2$ we have $W^{p,k_1}(\Omega) \subset W^{p,k_2}(\Omega)$ for all $p \geq 1$.

Again we can define spaces $H^s(\Omega)$, for real $s \geq 0$, such that $H^{s_1}(\Omega) \subset H^{s_2}(\Omega)$ whenever $s_1 > s_2 \geq 0$, and such that the spaces $H^s(\Omega)$ coincide with the previously defined spaces $H^k(\Omega)$ whenever $s = k \in I\!N$. We give more details in Section B.4. For a proper definition of these Hilbert spaces $H^s(\Omega)$ see also [17].

### B.3.2 Pivot space and dual spaces

Let $U$ be a Hilbert space with respect to the inner product $(\cdot,\cdot)_U$. The dual space $U'$ of $U$, i.e. the space of all bounded linear functionals on $U$, can be identified with $U$ according to the following

**Theorem B.3** (Riesz): For every continuous (i.e. bounded) linear functional $f$ on a Hilbert space $U$, with inner product $(\cdot,\cdot)_U$, there is a *unique* $u \in U$ such that $f(x) = (x,u)_U$ for all $x \in U$.

**Definition B.4** $V \subset U$ is called a *continuously embedded* subspace of $U$ if $V$ is a subspace of $U$ such that there exist a norm $\|v\|_V$ on $V$ and a constant $C > 0$ with

$$\|v\|_U \;\leq\; C \, \|v\|_V \quad \text{for all } v \in V,$$

i.e. the embedding operator has finite norm $\|I\|_{U \leftarrow V}$.

On $V'$, the dual space of $V$, we define the following norm

$$\|v'\|_{V'} = \sup_{v \in V,\, v \neq 0} \frac{|v'(v)|}{\|v\|_V} \, .$$

Let $V \subset U$ be a continuously embedded dense subspace of $U$. For any $u \in U$ the mapping $v \to (u,v)_U$ is a linear functional on $V$. It is bounded because for all $v \in V$:

$$|(u,v)_U| \;\leq\; \|u\|_U \|v\|_U \;\leq\; C \|u\|_U \, \|v\|_V \;=\; C' \, \|v\|_V \, .$$

In the above sense we have then have $V \subset U = U' \subset V'$. With respect to the $U$ inner-product $U'$ can be identified with $U$ (Riesz), but $V'$ is *not* identified with $V$. Of course, $V'$ can be identified with $V$ with respect to the $V$ inner-product. If we consider $U$ inner-products, the space $U$ is called the *pivot-space*.

### B.3.3 Dual spaces of Sobolev spaces

We can apply the above for Sobolev spaces. We take $L^2(\Omega) = H^0(\Omega)$ as pivot space. Introducing the notation

$$H^{-k}(\Omega) = (H_0^k(\Omega))'$$

we obtain the following nesting of spaces

$$\bar{C}_0^\infty(\Omega) \subset ... \subset H_0^2(\Omega) \subset H_0^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega) \subset H^{-2}(\Omega) \subset ... \subset \mathcal{D}(\Omega),$$

with $\mathcal{D}(\Omega)$ the space of generalised functions, i.e. the bounded functionals on $\bar{C}_0^\infty(\Omega)$ .

## B.4 Scales of Banach and Hilbert spaces

In Section B.1 we introduced Banach spaces $\bar{C}^s(\Omega)$, $s \geq 0$, based on the Banach spaces $\bar{C}^k(\Omega), k \in I\!N$. In Section B.3 we mentioned the possibility of introducing Hilbert spaces $H^s(\Omega), s \in I\!R, s \geq 0$, which exist –in some sense– between the Hilbert spaces $H^k(\Omega), k \in I\!N$. Here we will give some more details about how such spaces can be inserted.

The spaces $H^s(\Omega)$ and $\bar{C}^s(\Omega)$ have two important properties which we use frequently: For $s_1, s_2 \in I\!R$, $s_1 > s_2 \geq 0$ we have

1. $H^{s_1}(\Omega) \subset H^{s_2}(\Omega)$ dense, and $\bar{C}^{s_1}(\Omega) \subset \bar{C}^{s_2}(\Omega)$ dense.
2. $\|v\|_{H^{s_2}} \leq \|v\|_{H^{s_1}}$ for all $v \in H^{s_1}(\Omega)$ and $\|v\|_{\bar{C}^{s_2}} \leq \|v\|_{\bar{C}^{s_1}}$ for all $v \in \bar{C}^{s_1}(\Omega)$.

Property 1 means that we have scales of embedded Banach (Hilbert) spaces, viz. $\{\bar{C}^s(\Omega)\}_{s\geq 0}$ and $\{H^s(\Omega)\}_{s\geq 0}$. Many results in these lecture notes are valid for both the $H^s(\Omega)$ and $\bar{C}^s(\Omega)$ spaces. To simplify our notation, then we introduce a scale of Banach (Hilbert) spaces

$$\{E^\alpha\}_{\alpha_0 \leq \alpha \leq \alpha_1} ,$$

which satisfies properties 1 and 2. The reader may substitute any scale of embedded Banach (Hilbert) spaces that satisfy properties 1 and 2. One possible choice is to substitute $E^\alpha = H^\alpha(\Omega)$.

**Definition B.5** A Banach space $H_1$ is called *normally embedded* in the Banach space $H_0$ if:

- $H_1$ is a dense subspace of $H_0$
- $\|x\|_{H_0} \leq \|x\|_{H_1}$ for all $x \in H_1$.

### B.4.1 Scales of Hilbert spaces

We now formulate a theorem on the existence of a scale of Hilbert spaces which connects two given Hilbert spaces.

**Theorem B.6** (cf. [15],thm.1.12): If a Hilbert space $H_1$ is *normally embedded* in a separable Hilbert space $H_0$, then there is a *unique* scale of Hilbert spaces which connects $H_1$ and $H_0$, i.e. $\exists \{H_s\}, 0 \leq s \leq 1 : H_1 \subset H_s \subset H_0$ for all $0 < s < 1$ and $H_s$ satisfies properties 1 and 2.
**Proof:** By virtue of the inequality, for an arbitrary $x \in H_1$,

$$|(x,y)_{H_0}| \;\leq\; \|y\|_{H_0}\|x\|_{H_0} \;\leq\; \|y\|_{H_0}\|x\|_{H_1} .$$

For a fixed $y \in H_0$ the functional $x \to (x,y)_{H_0}$ is bounded on $H_1$, and so by the Riesz theorem there is a uniquely determined element $z \in H_1$, such that $(x,y)_{H_0} = (x,z)_{H_1}$, for all $x \in H_1$. We write $z = Vy$. Then $V : H_0 \to H_1$ with

$$(x,y)_{H_0} = (x,Vy)_{H_1} , \text{ for all } x \in H_1).$$

Because $H_1 \subset H_0$, we have defined a $V : H_0 \to H_0$. Moreover, if $\|Vy\| \neq 0$,

$$\|Vy\|_{H_1} \;=\; \|z\|_{H_1} \;=\; \frac{(z,z)_{H_1}}{\|z\|_{H_1}} \;=\; \frac{|(z,y)_{H_0}|}{\|z\|_{H_1}} \;\leq$$

$$\leq\; \frac{|(z,y)_{H_0}|}{\|z\|_{H_0}} \;\leq\; \frac{\|y\|_{H_0}\cdot\|z\|_{H_0}}{\|z\|_{H_0}} \;=\; \|y\|_{H_0} .$$

Consequently $\|Vy\|_{H_0} \leq \|Vy\|_{H_1} \leq \|y\|_{H_0}$, hence $V$ is bounded. We also find that $V$ is self-adjoint and nonnegative: for $x, y \in H_0$ we have $(Vx, y)_{H_0} = (Vx, Vy)_{H_1} = (Vy, \bar{V}x)_{H_1} = (Vy, \bar{x})_{H_0} = (x, Vy)_{H_0}$ and $(Vy, y)_{H_0} = (Vy, Vy)_{H_1} \geq 0$. Also, if $Vy = 0$, then $(x, y)_{H_0} = 0$ for every $x \in H_1$, and by the density of $H_1$ we have $y = 0$. Thus, $V$ is positive.

Summarizing, we find that the operator $V : H_0 \to H_0$ is linear, defined on the whole of $H_0$, bounded, self-adjoint and positive. Hence $V^s$ is defined[1] for $s = 1/2$ and for $s = -1$.

We can now write $V^{-1} = U$. Then D($U$): = Domain ($U$) = Range ($V$) $\subset H_1$ and (B.6) can be written in the form

$$(Uz, x)_{H_0} = (z, x)_{H_1} \text{ ( for all } x \in H_1).$$

The domain of $U^{1/2}$ can be obtained by closing D(U) in the norm

$$\|U^{1/2}z\|_{H_0} = \sqrt{(Uz, z)_{H_0}} = \sqrt{(z, z)_{H_1}} = \|z\|_{H_1}.$$

Thus, $D(U^{1/2})$ is contained in $H_1$ and is closed in it. We show that $D(U^{1/2})$ coincides with $H_1$. If it was not the case, there would exist $x_0 \in H_1$ such that $(z, x_0)_{H_1} = 0$ for all $z \in D(U^{1/2})$ and, in particular, for all $z \in D(U)$. Then it follows from (B.6) that $(Uz, x_0)_{H_0} = 0\,(z \in D(U))$, and consequently $x_0 = 0$, since Range$(U) = H_0$. Finally, it follows from (B.6) that $\|U^{1/2}z\|_{H_0} = \|z\|_{H_1} (z \in H_1)$. It is now possible to consider the scale of Hilbert spaces

$$\{D(U^{\phi/2})\}_{-1 \leq \phi \leq 1},$$

which connects $H_0$ and $H_1$, because

$$D(U^{-1/2}) = D(V^{1/2}) = H_0,$$

and

$$D(U^{1/2}) = H_1.$$

For more details and a proof that this scale is unique, we refer to [15]. ∎

### A relation with the spectrum

It is convenient to recognise relations with spectral theory. We mention the following facts. (See, e.g., [20].)

(1) Every normal operator $N$ in $H$ has a unique spectral decomposition which satisfies

$$(Nx, y) = \int_{\sigma(M)} \lambda \, dE_{x,y}(\lambda) \qquad \forall x \in \mathrm{Dom}(N), y \in H$$

(2) $H_1$ is the domain of an (unbounded) positive self-adjoint operator $\Lambda : H_1 \to H_0$ such that $\|u\|_{H_1} = \|\Lambda u\|_{H_0}$. For simplicity we consider the case[2] where the spectrum of $\Lambda$ is discrete and the eigenvectors form a complete orthonormal basis for $H_0$. Then

$$\begin{aligned} u &= \sum_i (u, \phi_i)\, \phi \quad \forall u \in H_0 \\ \Lambda u &= \sum_i \lambda_i u, \phi_i)\, \phi \quad \forall u \in H_1 \end{aligned}$$

and we know

$$\|u\|^2_{H_0} = \sum (u, \phi_i)^2$$

[1] See, e.g., [16, Thm. 9.10-1, p.541] or [20, Thm. 13.31].

[2] This is the case when $H_1 \subset H_0$ is compactly embedded ?

and

$$\|u\|_{H_1}^2 = \sum \lambda_i^2 (u, \phi_i)^2 \, .$$

Now we can also define $H_s$, $s \geq 0$ to be the subset of the elements from $H_0$ with finite norm $\| \cdot \|_{H_s}$:

$$\|u\|_{H_1}^2 = \sum \lambda_i^{2s} (u, \phi_i)^2 \, .$$

### B.4.2 Scales of Banach spaces

Let $B_1 \subset B_0$ be Banach spaces, densely and continuously embedded, with norms $\| \cdot \|_0$, and $\| \cdot \|_1$. A Banach space B is called an *intermediate space* if $B_1 \subset B \subset B_0$. We can introduce a scale of intermediate spaces by introducing, with $q, p \in [1, \infty)$

$$K_q(t, u) = \inf_{u_0 + u_1 = u} \{ \|u_0\|_0^q + t^q \|u_1\|_1^q \}^{1/q} \, ,$$

$$\|u\|_{B_{p,s}} = \left[ \int_0^\infty \{ t^{-s} K_q(t, u) \}^p \frac{dt}{t} \right]^{1/p} \, ,$$

$$B_{p,s} = \left\{ u \in B_0, \, \|u\|_{B_{p,s}} < \infty \right\} \, .$$

**Theorem B.7** $B_{p,s}$ is a Banach space, intermediate between $B_0$ and$B_1$.

**Theorem B.8**

$$\|u\|_{B_{p,s}} = \leq [ps(1-s)]^{-1/p} \, \|u\|_0^{1-s} \|u\|_1^s$$

**Proof:** A straightforward proof is found in [6, p.142] ∎

**Theorem B.9** For the case that $B_0$ and $B_1$ are Hilbert spaces, we can take $p = q = 2$, and we find

$$\|u\|_s = \|u\|_{B_{p,s}} \, .$$

**Proof:** A straightforward proof is found in [6, p.144] ∎

### B.4.3 Interpolation of operators

**Theorem B.10** Let $B_1 \subset B_0$ and $\tilde{B}_1 \subset \tilde{B}_0$ be densely and continuously embedded, and let $L : B_i \to \tilde{B}_i$ with

$$\|Lu\|_{\tilde{B}_i} \leq C_i \, \|u\|_{B_i} \, , \quad i = 0, 1$$

then

$$\|Lu\|\|_{\tilde{B}_{p,s}} \leq C_0^{1-s} \, C_1^s \, \|u\|_{B_{p,s}} \, .$$

**Proof:** A straightforward proof is found in [6, p.145] ∎

### B.4.4 Simultaneous approximation in Banach scales

**Theorem B.11** Let $B_1 \subset B_0$ be densely and continuously embedded, and let $\forall \epsilon > 0\, \exists S_\epsilon \subset B_1$. Let $p \geq 1$, $s \in (0,1)$, and

$$\inf_{v \in S_\epsilon} \quad \epsilon^s \|u - v\|_{B_{p,s}} \leq \epsilon \, \|u\|_1 \quad \forall u \in B_1 \, ,$$

Then $\exists C(s) > 0$

$$\inf_{v \in S_\epsilon} \quad \|u - v\|_0 + \epsilon^s \|u - v\|_{B_{p,s}} \leq C(s) \, \epsilon \, \|u\|_1 \quad \forall u \in B_1 \, .$$

**Proof:** The theorem was proved in [7]. A straightforward proof is found in [6, p.146-147]. ∎

## B.5 Some applications

### B.5.1 Interpolation with piecewise polynomials

We define

$$\mathbb{Z}_h = \{ jh \, | j \in \mathbb{Z} \}, \, h > 0.$$

$\mathbb{Z}_h$ is a 1-dimensional uniform grid with meshwidth $h$. Let $\Omega \subset \mathbb{R}$, $D = \Omega \bigcap \mathbb{Z}_h$. The space of functions on the grid $\mathbb{Z}_h$ is denoted by

$$l_h(\mathbb{Z}_h) = \{ u_h : \mathbb{Z}_h \to \mathbb{R} \} \, .$$

The restriction $R_h : C^0(\Omega) \to l_h(D)$ is simply a surjection, i.e. :

$$(R_h u)(jh) = u(jh) \text{ for all } j \in \mathbb{Z} \text{ such that } jh \in D.$$

The prolongation $P_h : l_h(D) \to C^0(\Omega)$ uses interpolation with piecewise polynomials of degree $k$. We find

$$\|I - P_h R_h\|_{C^0(\Omega) \leftarrow C^{k+1}(\Omega)} \; = \sup_{u \in C^{k+1}(\Omega) \, , \, u \neq 0} \frac{\|u - P_h R_h u\|_{C^0(\Omega)}}{\|u\|_{C^{k+1}(\Omega)}} \leq C \, h^{k+1},$$

hence we have a convergent approximation of order $k+1$ (c.f. 2.3). A similar statement can be proved for the Sobolev spaces:

$$\|I - P_h R_h\|_{H^0(\Omega) \leftarrow H^{k+1}(\Omega)} \; = \sup_{u \in H^{k+1}(\Omega) \, , \, u \neq 0} \frac{\|u - P_h R_h u\|_{H^0(\Omega}}{\|u\|_{H^{k+1}(\Omega)}} \leq C \, h^{k+1}.$$

### B.5.2 Regularity of solutions of the Poisson equation

We consider the following PDE:

$$\begin{cases} \Delta u = f, & \text{on } \Omega, \\ u = 0, & \text{on } \partial\Omega. \end{cases}$$

Let $u \in H^2(\Omega) \bigcap H_0^1(\Omega)$ be a solution of this problem. We then have

$$\Delta u = f,$$

hence

$$(v, \Delta u)_{L^2(\Omega)} = (v, f)_{L^2(\Omega)} \text{ for all } v,$$

or

$$-(\nabla v, \nabla u)_{L^2(\Omega)} = (v, f)_{L^2(\Omega)} \text{ for all } v.$$

If we now take $v = u$ we obtain

$$(\nabla u, \nabla u)_{L^2(\Omega)} = -(u, f)_{L^2(\Omega)}.$$

The norm on $H^{-1}(\Omega)$ is given by

$$\|w\|_{H^{-1}(\Omega)} = \sup_{v \in H_0^1(\Omega),\, v \neq 0} \frac{(v, w)_{L^2}}{\|v\|_{H^1(\Omega)}} \text{ for all } w \in H^{-1}(\Omega).$$

Combining this definition with the above we obtain:

$$\|\nabla u\|^2_{H^0(\Omega)} \leq \|u\|_{H^1(\Omega)} \|f\|_{H^{-1}(\Omega)}.$$

By using the Poincaré inequality we find:

$$\|u\|^2_{H_0^1(\Omega)} \leq \|\nabla u\|^2_{H^0(\Omega)} + \|u\|^2_{H^0(\Omega)} \leq (C+1)\, \|\nabla u\|^2_{H^0(\Omega)},$$

and hence

$$\|u\|_{H^1(\Omega)} \leq (C+1)\, \|f\|_{H^{-1}(\Omega)} \text{ for some } C > 0.$$

We would like to generalise this result to something like

$$\|u\|_{H^{2+s}(\Omega)} \leq (C+1)\, \|f\|_{H^s(\Omega)}.$$

for some $C > 0$ and $-1 \leq s \leq s_{\max}$. However, this is not always possible. The number $s$ is a measure for the *regularity* of the solution $u$. We have given the proof for $s = -1$. Under conditions, $s > -1$ may hold. We shall not treat proofs for the theorems that make this statement more precise. Smoothness of the boundary (and more generally smoothness of the coefficients) plays a crucial role in such regularity proofs.

# Appendix C

# Multigrid Convergence

## C.1 A multigrid convergence proof without regularity assumption

In this appendix we give the convergence proof as found in [5]. We use the notaton that is used in the notes on Subspace Correction Techniques' and we supplement this with notation from our MG-notes: $M^{REL}$, $M^{CGC}$ denote the error amplification operators for relaxation and coarse-grid correction, possibly with subscripts denoting the level. Similarly we have $M^{TLA}$, $M^{TLA,p,q}$, $M^{MLA}$, $M^{MLA,p,q,\sigma}$. In general, we have for these operators the relation known from DCP-theory:

$$M = I - BA\,,$$

where $A$ is the target operator and $B$ the approximate inverse: $B^{REL}$, $B^{CGC}$, $B^{TLA}$ or $B^{MLA}$, respectively.

We start with a nested sequence of Hilbert spaces

$$V_0 \subset V_1 \subset \cdots \subset V_j\,,$$

and with $a(\cdot,\cdot)$ and $(\cdot,\cdot)_k$ symmetric positive definite bilinear forms on $V_k$, $(k = 0, \cdots, j)$. We consider the MG algorithm for the solution of the problem: given $f \in V_j$, find $v \in V_j$ satisfying

$$a(v,\phi) = (f,\phi)_j \quad \forall \phi \in V_j\,.$$

With the forms $a(\cdot,\cdot)$ we associate the operators $A_k : V_k \to V_k$ defined by

$$(A_k v, w) = a(v,w) \quad \forall v, w \in V_k\,.$$

Sometimes we also use $(\cdot,\cdot) = (\cdot,\cdot)_j$, $A = A_j$ and $(\cdot,\cdot)_A = a(\cdot,\cdot)$. We further introduce $R_k$ and $\overline{R}_k$ as the orthogonal projections onto $V_k$, respectively with respect to $(\cdot,\cdot)_A$ and $(\cdot,\cdot)_k$. With $\overline{R}_{k-1,k} : V_k \to V_{k-1}$ we denote the operator $\overline{R}_{k-1}$ restricted to $V_k$.

**Lemma C.1** On $V_k$ we have $\overline{R}_{k-1,k} A_k = A_{k-1} R_{k-1}$.
**Proof:** On $V_k$ we have $A_k = A_k R_k$, and hence $\overline{R}_{k-1,k} A_k = \overline{R}_{k-1,k} A_k R_k = \overline{R}_{k-1,k} \overline{R}_k A = \overline{R}_{k-1} A = A_{k-1} R_{k-1}$. ∎

**Remark:** In a slightly different notation, the above result can be written as $\overline{R}_{Hh} A_h = A_H R_H$ or

$$R_H = A_H^{-1} \overline{R}_{Hh} A_h\,.$$

For the convergence proof we are interested in the behaviour of $M^{MLA}$. Here we restrict ourselves to

$$M^{MLA} = M^{MLA,p,q,\sigma} = M^{MLA,1,0,1}\,,$$

because the theory by Bramble c.s. yields no improved results for W-cycles compared to V-cycles, and because the more general case $M^{MLA,p,q,1}$ is easily reduced to the case for $M^{MLA,1,0,1}$. Our starting point is, of course, the two-level algorithm:

$$\begin{aligned} M_h^{TLA} = M_h^{CGC} M_h^{REL} &= (I_h - A_H^{-1}\overline{R}_{Hh}A_h)(I_h - B_h^{REL}A_h) \\ &= (I_h - R_H)(I_h - B_h^{REL}A_h)\,. \end{aligned}$$

**Lemma C.2** If the CGC is realised by $\sigma$ sweeps of the MLA on the coarser grid, we have

$$M_h^{CGC,\sigma} = (I - R_H) + (M_H^{MLA})^\sigma R_H\,.$$

**Proof:** $M_h^{CGC,\sigma} = I_h - B_h^{CGC,\sigma} R_{Hh} A_h$, where $B_h^{CGC,\sigma}$ is the approximate inverse realised by $\sigma$ sweeps. By DCP eq.(1.6) we know

$$B_h^{CGC,\sigma} = \left[I - (I - B_H^{MLA}A_H)^\sigma\right] A_H^{-1}\,,$$

so that

$$\begin{aligned} I_h - M_h^{CGC,\sigma} &= \left[I - (I - B_H^{MLA}A_H)^\sigma\right] A_H^{-1}\overline{R}_{Hh}A_h \\ &= \left[I - (I - B_H^{MLA}A_H)^\sigma\right] R_H \\ &= R_H - (M_H^{MLA})^\sigma R_H\,. \end{aligned}$$

■

**Remark:** It follows that

- $\sigma = 0 \Rightarrow M_h^{CGC,0} = I_h$.
- $\sigma = \infty \Rightarrow M_h^{CGC,\infty} = M_h^{CGC} = I_h - R_h = I_h - A_H^{-1}\overline{R}_{Hh}A_h$.
- $\sigma = 1 \Rightarrow M_h^{CGC,1} = (I - R_H)(I - B_H^{MLA}A_H)R_H = I - B_H^{MLA}\overline{R}_{Hh}A_h$.

**Lemma C.3** With $T_k = (I - M_k^{REL})R_k$, $(k = 1, \cdots, j)$ and $T_0 = R_0$ we can write

$$M_j^{MLA} = M^{MLA,1,0,1} = (I - T_0)(I - T_1)\cdots(I - T_j)\,.$$

**Proof:**

$$\begin{aligned} M_k^{MLA} &= I - B_k^{MLA}A_k \\ &= M_k^{CGC,1}M_k^{REL} \\ &= (I - B_{k-1}^{MLA}\overline{R}_{k-1,k}A_k)M_k^{REL}\,. \end{aligned}$$

Hence

$$\begin{aligned} I - B_k^{MLA}A_kR_k &= I - R_k + (I - B_{k-1}^{MLA}\overline{R}_{k-1,k}A_k)M_k^{REL}R_k \\ &= I - R_k + (I - B_{k-1}^{MLA}A_{k-1}R_{k-1})M_k^{REL}R_k \\ &= (I - B_{k-1}^{MLA}A_{k-1}R_{k-1})(I - R_k + M_k^{REL}R_k) \\ &= (I - B_{k-1}^{MLA}A_{k-1}R_{k-1})(I - T_k)\,, \end{aligned}$$

where we use $R_{k-1}(I - R_k) = 0$, and

$$\begin{aligned} M_j^{MLA} &= I - B_j^{MLA} A_j = I - B_j^{MLA} A_j R_j \\ &= (I - T_0)(I - T_1) \cdots (I - T_j) \, . \end{aligned}$$

∎

**Remark:** In the following discussion we frequently use

$$T_k = (I - M_k^{REL}) R_k, \quad (k = 1, \cdots, j); \quad T_0 = R_0 \, ,$$

which is an A-selfadjoint operator iff $M^{REL}$ is A-selfadjoint (*what we assume from this point!*), as well as

$$E_k = (I - B_k^{MLA} A_k R_k)^T, \quad (k = 0, \cdots, j); \quad E_{-1} = I \, .$$

We notice that $E_j = (M_j^{MLA})^T$ and from the previous lemma we see

$$E_k = (I - T_k) E_{k-1} \, .$$

**Theorem C.4** If we assume:

- $\exists Q_k : V_j \to V_k$, $k = 0, \cdots, j$, with $Q_j = I$ such that
$$k = 1, \cdots, j; \quad \lambda_k = \rho(A_k) \, . \|(Q_k - Q_{k-1}) u\|_k^2 \le \frac{C_1}{\lambda_k} \|u\|_A^2, \tag{C.1}$$
- $$\|Q_k u\|_A^2 \le C_2 \|u\|_A^2, \quad k = 0, \cdots, j \, . \tag{C.2}$$
- $$\tilde{M}_k = \mathrm{Range}(B_k^{REL}) \supset \mathrm{Range}(Q_k - Q_{k-1}) \tag{C.3}$$
- $$\frac{\|u\|_k^2}{\lambda_k} \le C_R \, (B_k^{REL} u, u)_k \qquad \forall u \in \tilde{M}_k \tag{C.4}$$

Then

$$\|M^{MLA,1,0,1} u\|_A^2 \le \delta_j \|u\|_A^2 \, ,$$

with $\delta_j = 1 - \frac{1}{j \; [\sqrt{2(1+C_2)} + \sqrt{C_R C_1}]^2}$.

**Remark:** If we consider the usual second order symmetric elliptic equation with solution in $H_0^1(\Omega)$ and

$$c_0 \|v\|_1 \le a(v,v) \le \|v\|_1 \, .$$

for $Q_k$ we may take the $L^2(\Omega)$-projection on $V_k$. Then, for a quasi-uniform triangulation we know for all $v \in H_0^1(\Omega)$ that $\|I - Q_n\|_m \le C h_k^{p-m} \|v\|_p$, $0 \le m < p \le 2$, and hence[1]

$$\|(I - Q_k) v\| \le C h_k \|v\|_1 \, ,$$

[1] The first assumption in the theoren follows by $\|(Q_k - Q_{k-1}) u\|_k \le \|(Q_{k-1} - I) u\|_k + \|(Q_k - I) u\|_k \le C h_k \|v\|_1$, and $\lambda_k \approx h_k^{-2}$.

$$\|Q_k v\|_1 \leq C\|v\|_1 \, .$$

With $(\cdot,\cdot)_k$ a discrete innerproduct which satisfies for all $v \in V_k$

$$c\|v\| \leq (v,v)_k \leq C\|v\| \, ,$$

with $c$ and $C$ independent of $k$, and for the Richardson method as relaxation, with $B^{REL} = \frac{1}{\overline{\lambda}_k} I$ where $\overline{\lambda}_k = \mathcal{O}(h_k^{-2})$, then the assumptions are satisfied.

**Proof:** The proof is in two parts: first we show that the theorem holds if

$$\|u\|_A^2 \leq Cj \sum_{k=0}^{j} (E_{k-1}u, T_k E_{k-1}u)_A \, , \tag{C.5}$$

and then we verify this condition.

Because $E_k = (I - T_k)E_{k-1}$ we know that $E_{k-1} = E_k + T_k E_{k-1}$, and

$$\begin{aligned}
\|E_{k-1}u\|_A^2 &= \|E_k u\|_A^2 + \|T_k E_{k-1}u\|_A^2 + 2\,(E_k u, T_k E_{k-1}u)_A \\
&= \|E_k u\|_A^2 + \|T_k E_{k-1}u\|_A^2 + 2\,((I-T_k)E_{k-1}u, T_k E_{k-1}u)_A \\
&= \|E_k u\|_A^2 + (E_{k-1}u, T_k E_{k-1}u)_A^2 \\
&\quad +((I-T_k)E_{k-1}u, T_k E_{k-1}u)_A \, .
\end{aligned}$$

Using that the spectrum of $T_k$ lies in the interval $[0,1]$, we know that the last term is positive, and hence

$$\begin{aligned}
(E_{k-1}u, T_k E_{k-1}u)_A &\leq \|E_{k-1}u\|_A^2 - \|E_k u\|_A^2 \, , \\
\sum_{k=0}^{j} (E_{k-1}u, T_k E_{k-1}u)_A &\leq \|u\|_A^2 - \|E_j u\|_A^2 \, .
\end{aligned}$$

Assuming (C.5) we derive

$$\|u\|_A^2 \leq Cj \left( \|u\|_A^2 - \|E_j u\|_A^2 \right) \, ,$$

from which we see

$$Cj \; \|E_j u\|_A^2 \leq Cj\|u\|_A^2 - \|u\|_A^2 \, ,$$

or

$$\|E_j u\|_A^2 \;\leq\; \frac{Cj-1}{Cj} \|u\|_A^2$$

and hence

$$\|M^{MLA}u\|_A^2 \;\leq\; \left(1 - \frac{1}{Cj}\right) \|u\|_A^2 \, .$$

Which was to be proved for $C = 1 + \sqrt{C_2} + \sqrt{C_1 C_R}$.

Now the relation (C.5) is left to be proved. For this we use

$$u = \sum_{k=1}^{j} (Q_k - Q_{k-1})u + Q_0 u \, .$$

$$\|u\|_A^2 \;=\; \sum_{k=1}^{j} (u, (Q_k - Q_{k-1})u)_A + (u, Q_0 u)_A$$

$$
\begin{aligned}
&= \sum_{k=1}^{j} (E_{k-1}u, (Q_k - Q_{k-1})u)_A \\
&\quad + \sum_{k=1}^{j} ((I - E_{k-1})u, (Q_k - Q_{k-1})u)_A + (u, Q_0 u)_A
\end{aligned} \tag{C.6}
$$

The first part of this sum gives

$$
\begin{aligned}
&\sum_{k=1}^{j} (E_{k-1}u, (Q_k - Q_{k-1})u)_A \\
= &\sum a(E_{k-1}u, (Q_k - Q_{k-1})u) \\
= &\sum (AE_{k-1}u, (Q_k - Q_{k-1})u)_k \\
= &\sum (\overline{R}_k AE_{k-1}u, (Q_k - Q_{k-1})u)_k \\
= &\sum (A_k R_k E_{k-1}u, (Q_k - Q_{k-1})u)_k \\
\leq &\sum \|A_k R_k E_{k-1}u\|_k \; \|(Q_k - Q_{k-1})u\|_k \\
\leq &\sum \|A_k R_k E_{k-1}u\|_k \; \sqrt{C_1/\lambda_k}\, \|u\|_A \\
= &\sqrt{C_1}\, \|u\|_A \sum \lambda_k^{-1/2} \|\overline{R}_k A_k R_k E_{k-1}u\|_k
\end{aligned}
$$

We make use of the relation between different vector norms: $\|w\|_1 \leq \sqrt{n}\|w\|_2$.

$$
\begin{aligned}
&\sqrt{C_1}\, \|u\|_A \sum \|A_k R_k E_{k-1}u\|_k \lambda_k^{-1/2} \\
\leq &\sqrt{jC_1}\, \|u\|_A \left( \sum \|A_k R_k E_{k-1}u\|_k^2 \lambda_k^{-1} \right)^{1/2} \\
= &\sqrt{jC_1}\, \|u\|_A \left( \sum C_R \; (B_k^{REL} A_k R_k E_{k-1}u, A_k R_k E_{k-1}u)_k \right)^{1/2} \\
= &\sqrt{jC_1}\, \|u\|_A \left( \sum C_R \; (B_k^{REL} A_k R_k E_{k-1}u, R_k E_{k-1}u)_A \right)^{1/2} \\
= &\sqrt{jC_1}\, \|u\|_A \left( \sum C_R \; (B_k^{REL} A_k R_k E_{k-1}u, E_{k-1}u)_A \right)^{1/2} \\
= &\sqrt{jC_1}\, \|u\|_A \left( \sum C_R \; ((I - M_k^{REL}) R_k E_{k-1}u, E_{k-1}u)_A \right)^{1/2} \\
= &\sqrt{jC_1 C_R}\; \|u\|_A \left( \sum (T_k E_{k-1}u, E_{k-1}u)_A \right)^{1/2}
\end{aligned}
$$

The remaining part of the sum (C.6) yields

$$
\begin{aligned}
&(u, Q_0 u)_A + \sum_{k=1}^{j} ((I - E_{k-1})u, (Q_k - Q_{k-1})u)_A \\
= &(u, Q_0 u)_A + \sum_{k=1}^{j} (u, Q_k u)_A - \sum_{k=1}^{j} (E_{k-1}u, Q_k u)_A \\
&- \sum_{k=1}^{j} (u, Q_{k-1}u)_A + \sum_{k=1}^{j} (E_{k-1}u, Q_{k-1}u)_A
\end{aligned}
$$

$$
\begin{aligned}
&= (u, Q_j u)_A - \sum_{k=1}^{j} (E_{k-1}u, Q_k u)_A + \sum_{k=0}^{j-1} (E_k u, Q_k u)_A \\
&= (u, Q_j u)_A - (E_{j-1}u, Q_j u)_A + (E_0 u, Q_0 u)_A \\
&\quad + \sum_{k=1}^{j-1} ((E_k - E_{k-1})u, Q_k u)_A \\
&= ((I - E_{j-1})u, u)_A - \sum_{k=1}^{j-1} ((E_{k-1} - E_k)u, Q_k u)_A \\
&= ((I - E_0)u, u)_A + \sum_{k=1}^{j-1} ((E_{k-1} - E_k)u, (I - Q_k)u)_A \\
&= (T_0 u, u)_A + \sum_{k=1}^{j-1} ((E_{k-1} - E_k)u, (I - Q_k)u)_A \\
&= (T_0 u, u)_A + \sum_{k=1}^{j-1} (T_k E_{k-1}u, (I - Q_k)u)_A \,.
\end{aligned}
$$

Where we used $E_k - E_{k-1} = -T_k E_{k-1}$ and $I - E_{j-1} = \sum_0^{j-1} T_k E_{k-1}$, and $E_0 = I - T_0 = I - R_0$, so that $(E_0 u, Q_0 u)_A = ((I - R_0)u, Q_0) = 0$.

Continuing, again using $\|w\|_1 \le \sqrt{n}\|w\|_2$,

$$
\begin{aligned}
& (T_0 u, u)_A + \sum_{k=1}^{j-1} (T_k E_{k-1}u, (I - Q_k)u)_A \\
\le\ & (T_0 u, u)_A + \sum \|T_k E_{k-1}u\|_A \ \|(I - Q_k)u\|_A \\
\le\ & (T_0 u, u)_A + \left((j-1) \sum \|T_k E_{k-1}u\|_A^2 \ \|(I - Q_k)u\|_A^2\right)^{1/2} \\
\le\ & (T_0 u, u)_A + \left(2(j-1) \sum_{k=1}^{j-1} \|T_k E_{k-1}u\|_A^2 \ (\|u\|_A^2 + \|Q_k u\|_A^2\right)^{1/2} \\
\le\ & (T_0 u, u)_A + \left(2(j-1)(1 + C_2) \sum_{k=1}^{j-1} \|T_k E_{k-1}u\|_A^2 \ \|u\|_A^2\right)^{1/2} \\
\le\ & \left(2(j-1)(1 + C_2) \sum_{k=0}^{j-1} \|T_k E_{k-1}u\|_A^2 \ \|u\|_A^2\right)^{1/2} \\
\le\ & \sqrt{2j(1 + C_2)}\, \|u\|_A \left(\sum_{k=0}^{j-1} \|T_k E_{k-1}u\|_A^2\right)^{1/2} \\
=\ & \sqrt{2j(1 + C_2)}\, \|u\|_A \left(\sum_{k=0}^{j-1} (T_k E_{k-1}u, , T_k E_{k-1}u)_A\right)^{1/2} \\
\le\ & \sqrt{2j(1 + C_2)}\, \|u\|_A \left(\sum_{k=0}^{j-1} (T_k E_{k-1}u, , E_{k-1}u)_A\right)^{1/2}
\end{aligned}
$$

where we used again $\sigma(T) \subset [0, 1]$. Combining the results for the first and the second part, we prove equation (C.5):

$$
\begin{aligned}
\|u\|_A^2 &\leq \sqrt{jC_1C_R}\;\|u\|_A \left(\sum_{k=1}^{j}(T_kE_{k-1}u, E_{k-1}u)_A\right)^{1/2} \\
&\quad + \sqrt{2j(1+C_2)}\,\|u\|_A \left(\sum_{k=0}^{j-1}(T_kE_{k-1}u, , E_{k-1}u)_A\right)^{1/2}, \\
&= (\sqrt{jC_1C_R} + \sqrt{2j(1+C_2)})\,\|u\|_A \left(\sum_{k=0}^{j-1}(T_kE_{k-1}u, , E_{k-1}u)_A\right)^{1/2}, \\
\|u\|_A^2 &\leq j\left(\sqrt{C_1C_R} + \sqrt{2(1+C_2)}\right)^2 \sum_{k=0}^{j-1}(T_kE_{k-1}u, , E_{k-1}u)_A\,,
\end{aligned}
$$

■

# Appendix D

# Conjugate gradients

## D.1 The construction with less pain (due to v.d.Sluis)

Consider a linear system with symmetric and positive definite matrix $A$

$$Ax = b\,.$$

Let $K_i$ denote the $i$-dimensional Krylov space[1], $K_0 = \{0\}$ and $K_i = \mathrm{Span}(b, Ab, A^2b, \cdots, A^{i-1}b)$. Without loss of generality starting from $x_0 = 0$, we want to determine approximate solutions $x_1$, $x_2$, $\cdots$ such that $x_i$, with respect to some suitable inner product, is the best approximation of the solution $x$ in $K_i$.

We call the inner product $P$. Hence, we want $x_i$ to be the $P$-projection of $x$ on $K_i$:

$$x - x_i \perp_P K_i\,. \tag{D.1}$$

In order to find such a projection, it is convenient to have a $P$-orthogonal sequence $p_1, p_2, \cdots$ with $p_i \in K_i$ for all $i$. As, $p_1, p_2, \cdots$ *is a $P$-orthogonal basis for $K_i$,* we have

$$x_i = x_{i-1} + \frac{(x, p_i)_P}{(p_i, p_i)_P}\, p_i\,,$$

or

$$x_i = x_{i-1} + \alpha_i\, p_i \tag{D.2}$$

with

$$\alpha_i = \frac{(x, p_i)_P}{(p_i, p_i)_P}\,. \tag{D.3}$$

It seems that we have to know $x$ to compute $\alpha_i$, but we can take e.g. the $A$-inner product for the $P$-inner product. Then

$$\alpha_i = \frac{(x, p_i)_P}{(p_i, p_i)_P} = \frac{(x, p_i)_A}{(p_i, p_i)_A} = \frac{(Ax, p_i)}{(Ap_i, p_i)} = \frac{(b, p_i)}{(Ap_i, p_i)}\,. \tag{D.4}$$

Hence, we take this as a proper choice for the inner product.

As we can find the sequence $\{p_i\}$ e.g. by application of Gram-Schmidt orthogonalisation of the sequence $b, Ab, A^2b, \cdots$, it is clear that the sequence $\{x_i\}$ can be computed.

In the conjugate gradient process the sequence $\{p_i\}$ and the residuals $\{r_i\}$ are generated in a clever way. Let

$$r_{i+1} = b - Ax_i \tag{D.5}$$

[1]The notation is such that in the sequences $\{x_i\}$, $\{r_i\}$, $\{p_i\}$, the elements $x_i$, $r_i$, $p_i$, are found in $K_i$.

and notice that, because $A\,K_i \subset K_{i+1}$, we have

$$r_{i+1} \in K_{i+1}\,. \tag{D.6}$$

Further, because of (D.1) we have $(r_{i+1}, K_i) = (A(x - x_i), K_i) = (x - x_i, K_i)_A = 0$, or in other words

$$r_{i+1} \perp K_i \tag{D.7}$$

Thus, all $r_i$ are orthogonal to each other w.r.t. the usual inner product.

Because $AK_{i-1} \subset K_i$ we also have $(r_{i+1}, K_{i-1})_A = 0$, i.e.

$$r_{i+1} \perp_A K_{i-1} \tag{D.8}$$

Hence, if $r_{i+1} \neq 0$ (otherwise we have found the solution), we have by (D.6) and (D.7) $r_{i+1} \in K_{i+1}$ and $r_{i+1} \notin K_i$. Therefore, we can determine $p_{i+1}$ for the next iteration step by orthogonalising $r_{i+1}$ w.r.t. $K_i$. Because of (D.8) we know

$$r_{i+1} \perp_A p_j\,, \qquad \text{for} \quad j = 1, 2, \cdots, i-1\,,$$

and in the $\{p_j\}$-expansion of $r_{i+1}$ we only have to take into account $p_i$:

$$p_{i+1} = r_{i+1} + \beta_{i+1}\, p_i\,, \tag{D.9}$$

$$\beta_{i+1} = -\frac{(r_{i+1}, p_i)_A}{(p_i, p_i)_A}\,. \tag{D.10}$$

Taking $p_1 = b$, the process (D.2) to (D.10) can be executed for $i = 1, 2, \cdots$, and it stops for $i \leq \dim(A)$. This describes the conjugate gradient process.

Further we have the following interesting relations.

$$\begin{gathered}(x, p_i)_A = (A(x - x_{i-1}), p_i) + (x_{i-1}, p_i)_A = (r_i, p_i) + 0 = \\ (r_i, r_i) + \beta_i (r_i, p_{i-1}) = (r_i, r_i)\,,\end{gathered}$$

and because

$$-r_{i+1} + r_i = A(x_i - x_{i-1}) = \alpha_i\, A\, p_i$$

we have

$$-\frac{(r_{i+1}, p_i)_A}{(p_i, p_i)_A} = -\frac{(r_{i+1}, (r_i - r_{i+1})/\alpha_i)}{(p_i, (r_i - r_{i+1})/\alpha_i)} = \frac{(r_{i+1}, r_{i+1})}{(p_i, r_i)}\,.$$

This leads to the following, efficient process

$$\begin{aligned}
x_0 &= 0, \\
r_1 = p_1 &= b, \\
\alpha_i &= \tfrac{(r_i, r_i)}{(Ap_i, p_i)}, \\
x_i &= x_{i-1} + \alpha_i\, p_i, \\
r_{i+1} &= r_i - \alpha_i\, A p_i, \\
\beta_{i+1} &= \tfrac{(r_{i+1}, r_{i+1})}{(r_i, r_i)}, \\
p_{i+1} &= r_{i+1} + \beta_{i+1}\, p_i\ .
\end{aligned}$$

## D.2 Convergence of CG

Let $u_k$ ve the $k$-th CG-approximation with initial guess $u_0$, then [Golub & VanLoan]

$$\|u-u_k\|_A \le 2\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^k \|u-u_0\|_A\,,$$

with $\kappa(A)=\lambda_{\max}(A)/\lambda_{\min}(A)$.

Assume the matrx $B$ to be symmetric w.r.t the inner product $(\cdot,\cdot)$. The operator of the preconditioned system $BAu=Bf$ is selfadjoint w.r.t. $(\cdot,\cdot)_A$ or $(\cdot,\cdot)_{B^{-1}}$. With $B=SS^T$, we get

$$S^TASv = S^Tf\,, \quad \text{and} \quad u=Sv\,.$$

The operator $S^TAS$ is selfadjoint w.r.t. $(\cdot,\cdot)$.

## D.3 Simple lemmas

**Lemma D.1** If $B$ is selfadjoint with respect to $(\cdot,\cdot)_A$ and

$$\|I-BA\|_A=\rho<1$$

then $B$ is symmetric positive definite (?) and

$$\kappa(BA)\le\frac{1+\rho}{1-\rho}\,.$$

**Proof:** Let $Y=I-BA$, then $1=(1-\rho)(1+\rho+\rho^2+\cdots)=(1-\rho)\|(1+\|Y\|+\|Y\|^2+\cdots)\ge(1-\rho)\|(1+Y+Y^2+\cdots)=(1-\rho)\|\|(I-Y)^{-1}\|=(1-\rho)\|\|(BA)^{-1}\|$. Hence, $\|(BA)^{-1}\|\le(1-\rho)^{-1}$.

Further $\|BA\|\le\|I\|+\|BA-I\|=1+\rho$. Hence

$$\kappa(BA)=\frac{\lambda_{\max}(BA)}{\lambda_{\min}(BA)}=\|BA\|\;\|(BA)^{-1}\|\le\frac{1+\rho}{1-\rho}\,.$$

■

**Lemma D.2** If $B$ is a preconditioner for $A$ then

$$u^{(k+1)}=u^{(k)}-\omega B(Au^{(k)}-f)$$

is convergent for $\omega\in(0,2/\rho(BA))$, and

$$\omega_{\text{opt}}=\frac{2}{\lambda_{\min}(BA)+\lambda_{\max}(BA)}$$

with the convergence factor

$$\frac{\kappa(BA)-1}{\kappa(BA)+1}=\rho\,.$$

**Proof:** See: [10, Teubner 1991, p.89] ■

The convergence rate of the preconditioned CG-method is

$$\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\le\frac{\sqrt{\frac{1+\rho}{1-\rho}}-1}{\sqrt{\frac{1+\rho}{1-\rho}}+1}=\frac{1-\sqrt{1-\rho^2}}{\rho}<\rho\,.$$

**Proof:** See: [10, Teubner 1991, p.257] ■

# Bibliography


[1] K.E. Atkinson. Iterative variants of the nystrom method for the numerical solution of integral equations. *Numer. Math.*, 22:17–33, 1973.

[2] R.E. Bank and C.G. Douglas. Sharp estimates for multigrid rates of convergence with general smoothing and acceleration. *SIAM J. Numer. Anal.*, 22:617–633, 1985.

[3] K. Böhmer, P. Hemker, and H.J. Stetter. The defect correction approach. *Computing Suppl.*, 5:1–32, 1984.

[4] H. Brakhage. Ueber die numerische behandlung von integralgleichungen nach der quadraturformelmethode. *Numer. Math.*, 2:183–196, 1960.

[5] J. H. Bramble, J. E. Pasciak, J. Wang, and J. Xu. Convergence estimates for multigrip algorithms without regularity assumptions. *Math. Comp.*, 57:23–45, 1991.

[6] J.H. Bramble. *Multigrid Methods.* Longman Scientific & Technical, 1993.

[7] J.H. Bramble and R. Scott. Simultaneous approximation in scales of banach spaces. *Math. Comp.*, 31:9447–954, 1985.

[8] A. Brandt and N. Dinar. Multigrid solutions to elliptic flow problems. In S.V. Parter, editor, *Numerical Methods for Partial Differential Equations*, New York, etc., 1979. Academic Press.

[9] W. Hackbusch. *Multigrid Methods and Applications.* Springer Verlag, 1985.

[10] W. Hackbusch. *Iterative Solution of Large Sparse Systems of Equations.* Springer–Verlag, Berlin, 1993.

[11] R.W. Hamming. *Digital Filters.* Prentice Hall, Inc., Englewood Cliffs, 1977.

[12] C. Johnson. *Numerical solutions of partial differential equations by the finite element method.* Cambridge Univ. Press, 1987.

[13] M.C. Joshi and R.K. Bose. *Some Topics in Nonlinear Functional Analysis.* Wiley Eastern Ltd., New Delhi etc., 1985.

[14] Y. Katznelson. *An Introduction to Harmonic Analysis.* Wiley, New York, etc., 1968.

[15] S.G. Krein, Ju.I. Petunin, and E.M. Semenov. *Interpolation of Linear Operators.* American Mathematical Society, Providence, RI, 1982.

[16] E. Kreyszig. *Introductory Functional Analysis with Applications.* John Wiley & Sons, 1978.

[17] J.L Lions and E. Magenes. *Problemes aux Limites Non Homogenes*. Dunod, Paris, 1968.

[18] J.T. Oden and J.N. Reddy. *An Introduction to the Mathematical Theory of Finite Elements*. John Wiley, 1976.

[19] A. Papoulis. *The Fourier Integral and its Applications*. McGraw-Hill, New York, etc., 1962.

[20] W. Rudin. *Functional Analysis*. Tata McGraw-Hill Publ. Comp., New Delhi, 1973.

[21] Smithies. *Integral Equations Analysis*. Opzoeken !!!, 1900.

[22] H.J. Stetter. The defect correction principle and discretization methods. *Numer. Math.*, 29:425–443, 1978.

[23] G. Strang and G.J. Fix. *An Analysis of the Finite Element Method*. Prentice Hall, Inc., 1973.

[24] H. Triebel. *Höhere Analysis*. VEB Deutscher Verlag der Wissenschaften, Berlin, 1972.

[25] K. Yosida. *Functional Analysis*. Springer Verlag, Berlin, etc., 1974.

[26] D. Young. *Iterative Solution of Large Linear Systems*. Academic Press, New York, London, 1971.

# Index