Lecture notes on defect correction

P.W. Hemker

Version: June 28, 1999

RATIO RES TEMPERAT OMNES
©1995

Contents

1 Defect Correct Processes **1**
- 1.1 Introduction . . . 1
- 1.2 Elementary Defect Correction Processes . . . 2
- 1.3 Convergence of the basic defect correction processes . . . 5
  - 1.3.1 Convergence of DCPA . . . 6
  - 1.3.2 Convergence of DCPB . . . 7
  - 1.3.3 Convergence of DCPL . . . 8
- 1.4 More elaborate versions of the principle . . . 8
  - 1.4.1 Non-stationary DCPs . . . 8
  - 1.4.2 Combined DCPs . . . 9
  - 1.4.3 Iterative application of DCPs . . . 10
  - 1.4.4 Recursive application of DCPs . . . 10
  - 1.4.5 Generalisation for nonlinear problems: DCPN . . . 10
- 1.5 Examples of defect correction processes . . . 11

2 Discretisation of continuous equations **17**
- 2.1 Introduction . . . 17
- 2.2 Approximation of spaces . . . 17
- 2.3 Discretisation errors . . . 22
- 2.4 Nested discretisations . . . 26
- 2.5 Galerkin discretisation . . . 29
- 2.6 A more elaborate example . . . 30

3 Defect correction and discretisation **35**
- 3.1 Introduction . . . 35
- 3.2 A fundamental theorem . . . 35
- 3.3 DCP with an approximate inverse of deficient rank . . . 38

4 Example of a two-grid algorithm **43**
- 4.1 Introduction . . . 43
- 4.2 Discretisation . . . 43
- 4.3 The smoothing process . . . 44
- 4.4 The coarse grid correction . . . 45
- 4.5 The two level algorithm . . . 47

5 Multigrid algorithms **55**
- 5.1 Introduction . . . 55
- 5.2 Two-level algorithms . . . 56
- 5.3 Multi-level algorithms . . . 57

5.4 Full multigrid method (FMG) 60

6 Convergence proofs for multigrid methods 65
6.1 The structure of a TGM convergence proof (Hackbusch) 65
6.2 The structure of a MGM convergence proof (Hackbusch) 67
6.3 Tools for a MGM convergence proof 69
6.4 The effect of the smoothing operator 72
6.5 The effect of the coarse grid correction 74
6.6 The convergence proof for the TGM 78
6.7 The convergence proof for the W-cycle and the V-cycle 79

7 Local mode analysis 83
7.1 Fourier transforms of continuous functions 83
7.2 Gridfunctions 84
7.3 The Fourier transform of a gridfunction 85
7.4 The relation between FTs of a function restricted to different grids 86
7.5 Toeplitz operators and their FTs 87
7.6 Consistency of a discrete operator 89
7.7 The smoothing factor for relaxation methods 90
7.8 Restrictions and prolongations 91
7.9 The order of a restriction or a prolongation 93
7.10 Fourier analysis in the case of mutual influencing frequencies 96
7.11 Requirements for transfer operators 97

A Discretisation 99
A.1 Integral equations 99
A.2 Discretisation of integral equations 101
A.3 Differential equations 102
A.4 The weak formulation of a differential equation 104
A.5 Discretisation of differential equations 105

B Function Spaces 109
B.1 Continuous functions and Hölder spaces 109
B.2 Generalised functions and derivatives 110
B.2.1 Distributions 110
B.2.2 Distributions as generalised functions 111
B.2.3 Derivatives of distributions 111
B.3 Sobolev spaces 112
B.3.1 Definition 112
B.3.2 Pivot space and dual spaces 113
B.3.3 Dual spaces of Sobolev spaces 113
B.4 Scales of Banach and Hilbert spaces 114
B.4.1 Scales of Hilbert spaces 114
B.4.2 Scales of Banach spaces 116
B.4.3 Interpolation of operators 116
B.4.4 Simultaneous approximation in Banach scales 117
B.5 Some applications 117
B.5.1 Interpolation with piecewise polynomials 117
B.5.2 Regularity of solutions of the Poisson equation 117

C Multigrid Convergence **119**
C.1 A multigrid convergence proof without regularity assumption 119

D Conjugate gradients **127**
D.1 The construction with less pain (due to v.d.Sluis) 127
D.2 Convergence of CG 129
D.3 Simple lemmas 129

Index **133**

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 $z_1, z_2, ...$ of approximations for the solution is then generated by

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

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:

$$z_{i+1} = z_i - e_i, \qquad i = 0, 1, 2, ...,$$

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 Δz satisfies the relation:

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

If we could compute Δz 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:

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

and

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

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

Remarks:

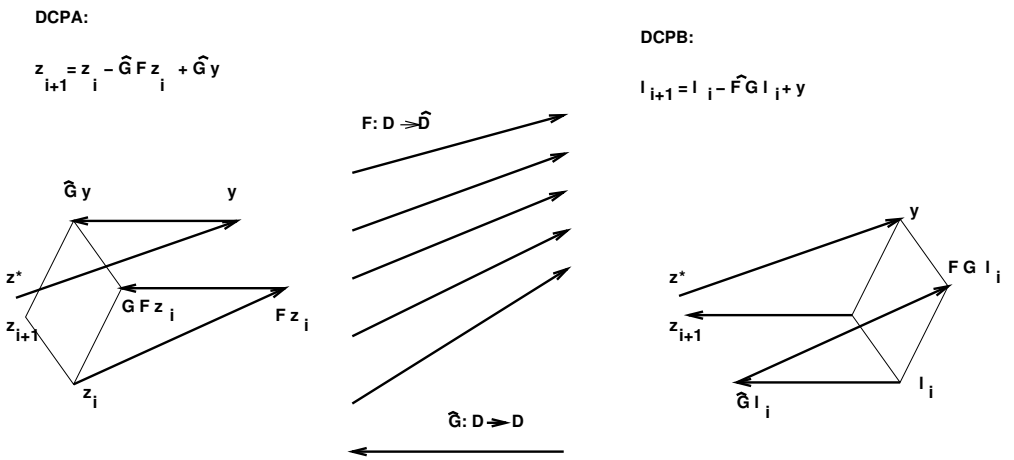

Figure 1.1: DCPA and DCPB

- 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 $\tilde{G}$ needs not 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

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

or, equivalently,

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

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:

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

[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 $\tilde{G}\tilde{F} = I_D$.

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.

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 $\{l_i\}_{i=0,1,2,...}$, and $\{z_i\}_{i=0,1,2,...}$, 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

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

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 $|||F||| \leq 1$.

Apparently

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

and for a linear operator F

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

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

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

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

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

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

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

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

and hence

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

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

$$\begin{array}{l} \|x_p - x_q\| \leq k^q \|x_{p-q} - x_0\| \\ \leq k^q \{\|x_{p-q} - x_{p-q-1}\| + ... + \|x_1 - x_0\|\} \\ \leq k^q \{k^{p-q-1} + ... + k + 1\} \|x_1 - x_0\| \\ \leq \frac{k^q}{1-k} \|x_1 - x_0\|. \end{array} \tag{1.4}$$

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 $\|x_{n+1} - Mu\| = \|Mx_n - Mu\| \leq k\|x_n - u\|$ 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, $v \in D$, $v \neq u$, and $v = Mv$, then $\|u - v\| = \|Mu - Mv\| \leq k\|u - v\|$. This gives $(1 - k)\|u - v\| \leq 0$. Since $1 - k > 0$ we have $\|u - v\| = 0$, i.e. $u = v$.

To obtain eq.(1.3) we have

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

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

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

■

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 ω_i. This is the case e.g. in non-stationary relaxation processes for the solution of linear systems. The value ω_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

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

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

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 σ 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 σ 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

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

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 $\tilde{G}'(\overline{y})$:

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

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 $u \;=\; Fz_i - y$) the form $\tilde{G}'(\overline{y})(Fz_i - y)$ by

$$\mu\tilde{G}(\tilde{\overline{y}} \;+\; (Fz_i - y)/\mu) \;-\; \mu\tilde{G}(\overline{y}).$$

We then obtain the following DCP:

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

Remarks:

- In this new defect correction process one still has the freedom to choose the parameters μ and $\overline{y}$. For the choice $\overline{y} = y$ and $\mu = 1$, DCPN coincides with DCPA.
- For a large enough μ, 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 μ 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

$$F : I\!R^n \to I\!R^n$$

is a square matrix, and $z, y \in I\!R^n$ 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:

$$\begin{cases} LUz_0 &= y \\ z_{i+1} &= z_i - e\ , \quad i = 0, 1, 2, \ldots . \end{cases}$$

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 $\tilde{G} = P^{-1}$. 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:

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

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

$$\left\{ \begin{array}{lcl} x_0 & = & \tilde{G}b, \\ x_{i+1} & = & (I - \tilde{G}A)x_i + \tilde{G}b, \end{array} \right.$$

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

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

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

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

and also

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

Hence we see that

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

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 $\tilde{G}$:

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

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

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

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

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:

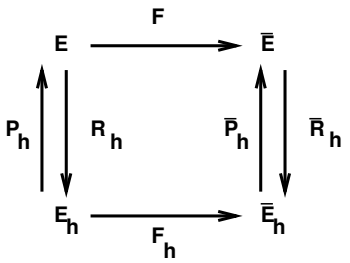

Figure 2.1: Spaces and operators in discretisation

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 Ω 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 ϕ_i is only limited to a (very) small part of Ω.

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

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

as well as

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

or

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

(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

$$F : \ E^\alpha \rightarrow \hat{E}^\alpha, \quad \alpha_0 \leq \alpha \leq \alpha_1.$$

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^α, $\alpha_0 \leq \alpha \leq \alpha_1$ 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 α and the norms on E_h^α are chosen such that

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

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^α, $\alpha_0 \leq \alpha \leq \alpha_1$, is *convergent of order* p if

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

for $\alpha_0 \leq \alpha \leq \alpha_1$.

Remarks:

- If E^α 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 $\tilde{p}$, $0 \leq \tilde{p} \leq p$. 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^α is convergent of order p if p is a number for which

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

- 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 $0 \leq p \leq p^*$. We will call this "the highest order".

- Of course, the value of p in Definition 2.3 depends on the actual α-indexing of the scale of Banach spaces E^α. Usually we consider only the possibilities where E^α are continuous function spaces, with $E^\alpha = C^\alpha(\Omega)$, or where E^α 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

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

is a continuous *linear* mapping

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

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

is bounded.

Remarks:

- The value of σ by this definition depends, again, on the α-indexing of the spaces E^α 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 σ-stable for all $\sigma \geq 0$ since from $C\|L^{-1}y\|_{E^{\alpha-\sigma}} \leq \|L^{-1}y\|_{E^\alpha}$ 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 $E^\alpha = H^\alpha(I\!R^d)$ and $\hat{E}^\alpha = H^{\alpha-2}(I\!R^d)$, $\alpha \geq 2$, then $I : E^\alpha \rightarrow \hat{E}^\alpha$ is bounded and has an unbounded inverse. For, if $u \in H^\alpha(I\!R^d)$, then $\|u\|_{H^{\alpha-2}(\Omega)} \leq \|u\|_{H^\alpha(\Omega)}$ and there doesn't exist a constant $C > 0$ such that

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

$I : E^{\alpha-2} \rightarrow \hat{E}^\alpha$ 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^α and $\hat{E}^\alpha$ the mapping I is 2-stable.

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

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

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 $\{P_h \mid P_h : E_h^\alpha \rightarrow E^\alpha,\ h \in \mathcal{H}\}$ 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 $\{R_h \mid R_h : E^\alpha \rightarrow E_h^\alpha,\ h \in H\}$ 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 $x_h \in E_h^\alpha \quad \|x_h\| = \|\hat{R}_h P_h x_h\| \leq \|\hat{R}_h\| \, \|P_h x_h\| \leq C\|P_h x_h\|$ 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,

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

■

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

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

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^α, E^α 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 $\|z^*\| \leq C$,. Then a sequence of discretisations (2.2) has order of convergence p if it has order of discrete convergence p.

Proof:
We have

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

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

[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^α 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

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

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

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

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

$$\begin{aligned} &x^*(t_{i+1}) - x^*(t_i) - h\left(p(t_i)x^*(t_i) + q(t_i)\right) = \\ &x^*(t_i) + hx^{*\prime}(t_i) + \tfrac{1}{2}h^2 x^{\star\prime\prime}(\xi_i) - x^*(t_i) - h(p(t_i)x^*(t_i) + q(t_i)). \end{aligned}$$

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

$$\begin{aligned} &x^*(t_{i+1}) - x^*(t_i) - h\left(p(t_i)x^*(t_i) + q(t_i)\right) = \\ &\tfrac{1}{2}h^2 x^{\star\prime\prime}(t_i) + ... + \tfrac{h^{k-1}}{(k-1)!} x^{\star(k-1)}(t_i) + \mathcal{O}(h^{k-1}). \end{aligned}$$

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 σ 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 τ_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

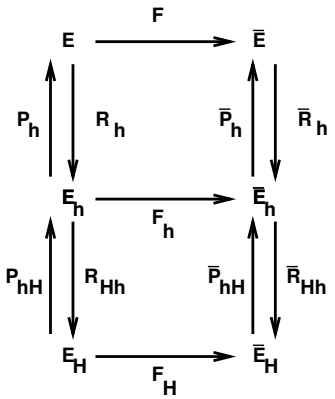

Figure 2.2: Nested discretisations

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

Definition 2.26 If $P_h P_{hH} = P_H$ 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^α and E_H^α, $\alpha_0 \le \alpha \le \alpha$, 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:

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

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

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

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

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

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

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

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

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

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

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

Using this we find

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

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^α 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 Ω (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 $\{\phi_i \mid i = 1, 2, ..., N(h)\}$ 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^α has a sufficiently high order. Furthermore, the set of prolongations $\{P_h \mid P_h : {E_h}^\alpha \to E^\alpha, h \in \mathcal{H}\}$ 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^α 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:

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

■

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^α *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 Ω. Let the problem $Fu = y$ be given by

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

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

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

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

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

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

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

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

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

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

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 Ω 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 ϕ_{ij}^h defined by

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

See Figure 2.4. The function u will be approximated by

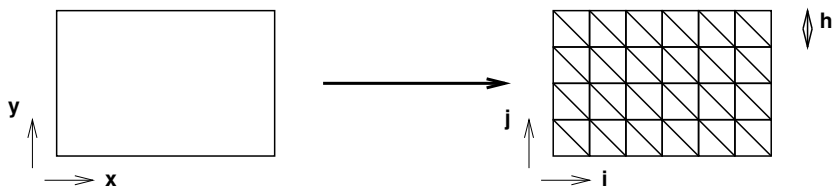

Figure 2.3: A triangularisation of a domain

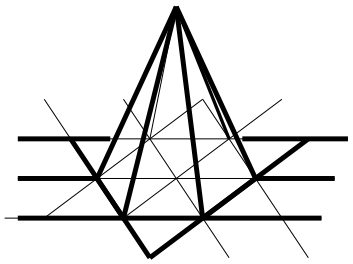

Figure 2.4: A piecewise linear basis function

$$u_h = \sum_{i,j} a^h_{ij}\, \phi^h_{ij}.$$

Notice that $u_h(N^h_{ij}) = a^h_{ij}$, where the constants a^h_{ij} are determined by

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

or

$$\int_\Omega \text{grad } u_h \cdot \text{grad } \phi^h_{kl} \; d\Omega + \frac{1}{\epsilon}\int_\Gamma u_h \phi^h_{kl} \; d\Gamma$$

$$= \int_\Omega f_\Omega \phi^h_{kl} d\Omega + \frac{1}{\epsilon}\int_\Gamma f_\Gamma \phi^h_{kl} \; d\Gamma, \qquad \text{for all } k, l,$$

or

$$\sum_{i,j} a^h_{ij} \{ \int_\Omega \text{grad } \phi^h_{ij} \cdot \text{grad } \phi^h_{kl} \; d\Omega + \frac{1}{\epsilon}\int_\Gamma \phi^h_{ij}\, \phi^h_{kl} \; d\Gamma \,\} = \tag{2.18}$$

$$= \int_\Omega f_\Omega \phi^h_{kl} \; d\Omega + \frac{1}{\epsilon}\int_\Gamma f_\Gamma \phi^h_{kl} \; d\Gamma \qquad \text{for all } \; k, l.$$

We define E_h as the set of function values at the nodes N^h_{ij}, 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

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

and $a_h \in I\!R^{N(h)}$ is the vector with coordinates a^h_{ij}. 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 Ω 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 ϕ_{ij}^H are defined in the same way as we previously defined the functions ϕ_{ij}^h, then

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

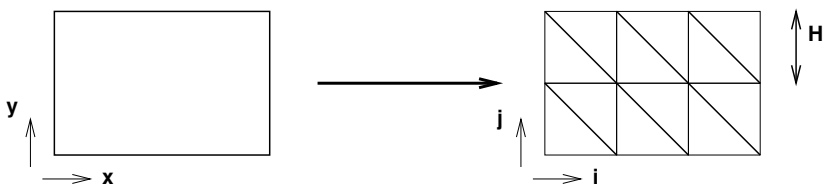

Figure 2.5: A coarser triangulation

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 Ω 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

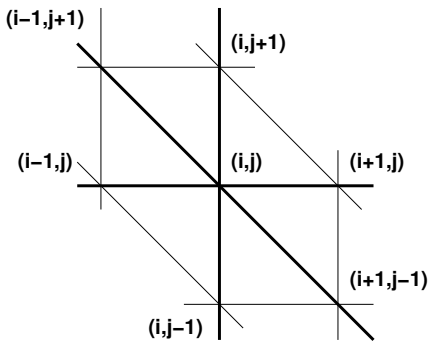

Figure 2.6: Neighbouring nodal points in a triangulation

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

$$\phi^H_{i,j} = \phi^h_{i,j} + 1/2(\phi^h_{i,j-1} + \phi^h_{i+1,j-1} + \phi^h_{i-1,j} + \phi^h_{i+1,j} + \phi^h_{i-1,j+1} + \phi^h_{i,j+1}), \tag{2.22}$$

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:

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

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

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

since $\hat{R}_h$ and $\hat{\tilde{P}}_h$ 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.

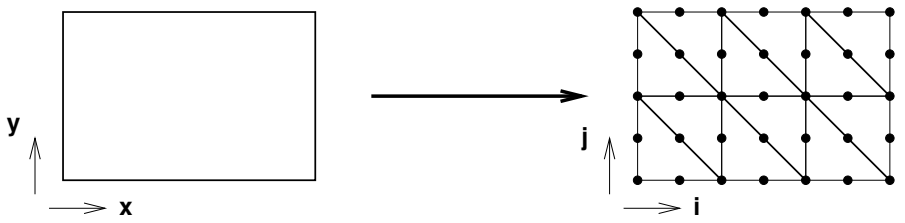

Figure 2.7: Nodal points for a piecewise quadratic approximation

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

Now we can associate with each node N_{ij} a piecewise parabolic function ϕ^h_{ij} 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 $u_h = \Sigma_{ij}\, a^h_{ij}\phi^h_{ij}$, 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 ϕ^h_{ij} 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 $\text{Support}(\phi^h_{ij}) \bigcap \text{Support}(\phi^h_{kl})$ 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.

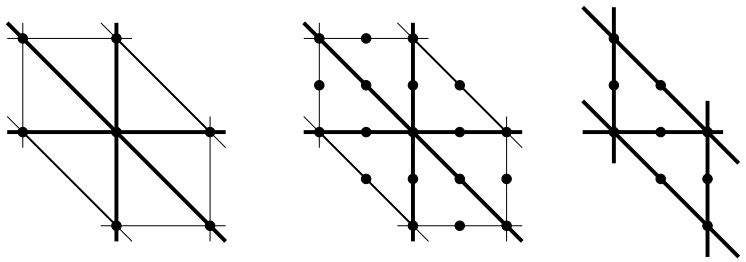

Figure 2.8: Related nodal points for piecewise parabolic approximations

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 $\{E^\alpha | \, \alpha_0 \le \alpha \le \alpha_1\}$, $\{\hat{E}^\alpha | \alpha_0 \le \alpha \le \alpha_1\}$ 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 $\{E_h^\alpha | \alpha_0 \le \alpha \le \alpha_1\}$, $\{\hat{E}_h^\alpha | \alpha_0 \le \alpha \le \alpha_1\}$ are scales of Banach spaces corresponding with $\{E^\alpha | \alpha_0 \le \alpha \le \alpha_1\}$ and $\{\hat{E}^\alpha | \alpha_0 \le \alpha \le \alpha_1\}$ 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 σ-stable for some $\sigma \ge 0$,
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

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

with $\beta_i = \min(\sigma + p^*, i(\sigma + p))$ and C independent of h and further ω, 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 : $\|L_h^{-1}\|_{E_h^{\alpha-\sigma} \leftarrow \hat{E}_h^\alpha} \le C_1$

2. p is a real number for which

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

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

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

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

3. p^* is a real number for which

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

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

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

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

4.

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

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

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

$$\leq \; C_1 C_4 C_5 h^{\min(p^*, (i+1)p)} \|\hat{u}\|_{E^{\omega+\beta_{i+1}}} + C_1 C_3 h^{\min(p^*, (i+1)p)} \|\hat{u}\|_{E^{\omega+\beta_{i+1}}}$$

$$\leq \; C h^{\min(p^*, (i+1)p)} \|\hat{u}\|_{E^{\omega+\beta_{i+1}}} \, .$$

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):

$$\|L_h - L_h^*\|_{\hat{E}_h^\alpha \leftarrow E_h^{\alpha+\tilde{p}}} = \|(L_h - L_h^*) R_h \hat{P}_h\|_{\hat{E}_h^\alpha \leftarrow E_h^{\alpha+\tilde{p}}}$$

$$\leq \; \|L_h R_h - L_h^* R_h\|_{\hat{E}_h^\alpha \leftarrow E^{\alpha+\tilde{p}}} \, \|\hat{P}_h\|_{E^{\alpha+\tilde{p}} \leftarrow E_h^{\alpha+\tilde{p}}}$$

$$\leq \; C_6 \|L_h R_h - \bar{R}_h L + \bar{R}_h L - L_h^* R_h\|_{\hat{E}_h^\alpha \leftarrow E^{\alpha+\tilde{p}}}$$

$$\leq \; C_6 \{\|L_h R_h - \bar{R}_h L\|_{\hat{E}_h^\alpha \leftarrow E^{\alpha+\tilde{p}}} + \|\bar{R}_h L - L_h^* R_h\|_{\hat{E}_h^\alpha \leftarrow E^{\alpha+\tilde{p}}}\}$$

$$\leq \; C_6 \{C_2 h^{\bar{p}} + C_3 h^{\bar{p}}\} \; \leq \; C h^{\bar{p}} \, , \;\text{ for all }\; 0 \leq \tilde{p} \leq p \, .$$

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

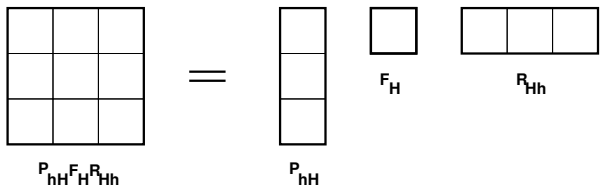

Figure 3.1: f.3.3.1

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

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

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 $\text{Range}(P) \subset E_h$. 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 $\text{Range}(P)^T$.

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

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

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

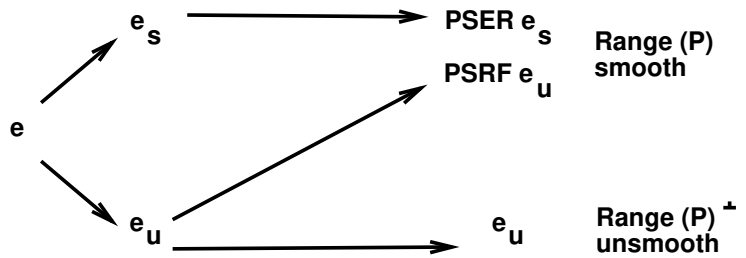

Figure 3.2: f.3.3.2

residual.

Definition 3.6 Let $\hat{\bar{P}}$ be a prolongation related to $\bar{R}$, i.e. a right inverse, e.g. $\hat{\bar{P}} = \bar{R}^T(R\bar{R}^T)^{-1}$. 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

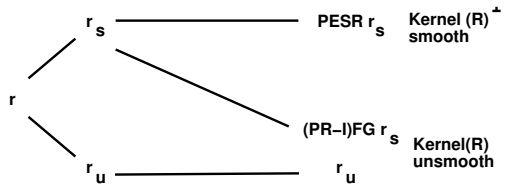

Figure 3.3: f.3.3.3

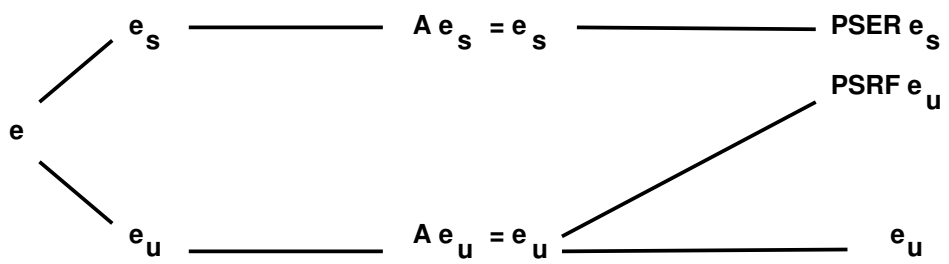

Figure 3.4: f.3.3.4

Galerkin approximation is small. Figure 3.4 illustrates this point. The reader may easily check 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.

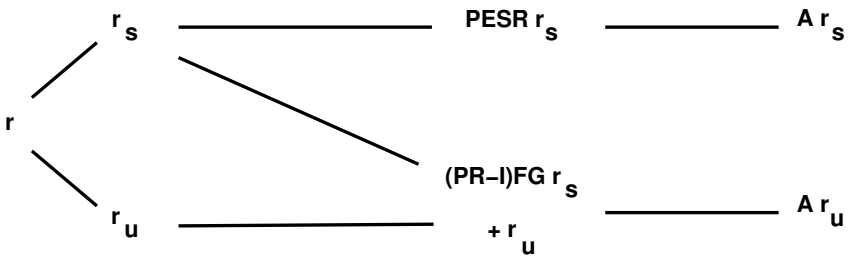

Figure 3.5: f.3.3.5

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.