stichting
mathematisch
centrum

MC

AFDELING NUMERIEKE WISKUNDE
(DEPARTMENT OF NUMERICAL MATHEMATICS)

NN 24/81 SEPTEMBER

P.W. HEMKER

LECTURE NOTES OF A SEMINAR ON MULTIPLE GRID METHODS

kruislaan 413 1098 SJ amsterdam

Printed at the Mathematical Centre, 413 Kruislaan, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

1980 Mathematics subject classification: 65F10, 65N20

Lecture notes of a seminar on multiple grid methods

by

P.W. Hemker

ABSTRACT

These are lecture notes of a seminar given by the author in the spring of 1981 as an extension of a Capita Selecta course of Prof. P.J. van der Houwen at the University of Amsterdam. In these notes some material has been collected that is basic to the theory of multiple grid and related iteration methods.

In this report the notes are in a preliminary form; neither do they contain all the material that should be included in a multiple grid course, nor are they in their final shape. Therefore, the report is intended for limited distribution only and the author will appreciate comments by readers.

In the first two sections a short introduction to boundary value problems and their discretization is given. Here we find the definitions of relative consistency and convergence in a sequence of related discretizations.

The third and the fourth section are devoted to the Defect Correction Principle. First the basic principle is explained and examples are given. Further different generalizations are treated. In the fifth section the multigrid algorithm is explained in terms of the defect correction principle. A sketch of a convergence theorem is given.

KEY WORDS & PHRASES: *Defect correction, multiple grid methods*

Lecture Notes of a Seminar on Multiple Grid Methods

Contents

1. BOUNDARY VALUE PROBLEMS

In this first section we describe boundary value problems. We give examples and we mention a number of properties that we shall need in the sequel.

Before we make some remarks about boundary-value problems for differential equations, we consider first integral equations of the 2^{nd} kind because of their resemblance with the differential problems in several respects.

Besides the description of the problems, we give a short description of the discretization methods that are used to find their numerical solutions. The Multi-Grid Methods that will be treated in the following sections are efficient means to solve the large systems of equations that arise from these discretizations.

1.1. Integral equations

Let $\Omega \subset \mathbb{R}^n$ be a bounded domain, then the equation

$$(1.1.1) \qquad u(x) = \int_{\Omega} k(x,y,u(y))dy$$

where $k: \Omega \times \Omega \times \mathbb{C}^k \to \mathbb{C}^k$ is a given function and

$$u : \Omega \to \mathbb{C}^k$$

is the unknown function, is called the *Urysohn integral equation*; k is called the *kernel function* of the equation.

If the kernel-function is linear (or, more precisely, affine), then the equation is a *Fredholm integral equation of the* 2^{nd} *kind* and can be written as

$$(1.1.2) \qquad u(x) = \int_{\Omega} k(x,y)u(y)dy + f(x)$$

where

$$u : \Omega \to \mathbb{R}^n$$

is the unknown function and

$$k : \Omega \times \Omega \to \mathbb{C}^{k\times k}$$

$$f : \Omega \to \mathbb{C}^k$$

are given.

Although many properties of the integral equations (1.1.1) and (1.1.2) can be treated for the vector-equation (k>1) and for a more-dimensional domain (n>1), we shall restrict ourselves mainly to scalar equations (k=1) on a one-dimensional domain $\Omega = [0,1]$. Solution methods for $k > 1$ and $n > 1$ are generally analogous to those for $k = 1$ and $n = 1$ but the general treatment would complicate the notation.

Often we consider only integral operators over the field $\mathbb{R}$ instead of over the field $\mathbb{C}$.

Equation (1.1.2) can be symbolically written as

$$(1.1.3) \qquad u = Ku + f,$$

where K denotes the linear integral operator defined by the function k.

If the function k(x,y) is bounded in y and is differentiable in x, then the operator K transforms any integrable function u on Ω into a differentiable function Ku on Ω. We say that the operator K has a smoothing property. This smoothing property is an essential feature for many integral operators and we shall exploit it in the use of our multigrid methods.

In order to formulate this smoothing property in better mathematical terms, we recall a number of theorems and definitions from functional analysis. (cf. Triebel, Smithies)

DEFINITION. A precompact set is a set from which each enumerable infinite sequence contains a convergent subsequence.

DEFINITION. A compact operator from a Banach space into a Banach space is an operator which maps any bounded set into a precompact set.

THEOREM [Triebel thm. 7.4].
Let $\Omega \subset \mathbb{R}^n$ *be a bounded domain and let* $k(x,y) \in \bar{C}(\Omega\times\Omega)$ (*i.e.* k *is a continuous function on the closure of* $\Omega\times\Omega$), *then the operator* $K: \bar{C}(\Omega) \to \bar{C}(\Omega)$ *is a compact operator.*

DEFINITION. A linear integral equation has an L^2-*kernel* if $k(s,t): \Omega \times \Omega \to \mathbb{C}$ satisfies

$$\iint |k(s,t)|^2 \, dsdt = \|K\|^2 < \infty ,$$

$$\int |k(s,t)|ds < \infty \quad \forall t,$$

$$\int |k(s,t)|dt < \infty \quad \forall s.$$

THEOREM [Triebel thm. 7.5].
Let $\Omega \subset \mathbb{R}^n$ *be a bounded domain and let* $k(x,y) \in L^2(\Omega\times\Omega)$ (*i.e.* k *is a square integrable function over* $\Omega \times \Omega$) *then* $K : L^2(\Omega) \to L^2(\Omega)$ *is a compact operator.*

DEFINITION. The *adjoint* K^* of a linear integral operator K is the operator with the kernelfunction $k^*(s,t) = \overline{k(t,s)}$,

K is called *Hermitian* if $K = K^*$

K is called *normal* if $KK^* = K^*K$

$\phi \in L^2(\Omega)$ is called an *eigenfunction* and $\lambda \in \mathbb{C}$ is an *eigenvalue* of K *if*

$$\lambda\phi = K\phi \qquad \phi \neq 0.$$

THEOREM [Triebel thm. 11.2].
A compact operator in a Hilbert space has at most an enumerable infinite set of eigenvalues which may be dense only at $\lambda = 0$. *Each non-zero eigenvalue has a finite multiplicity.*

THEOREM [Smithies, thms. 7.3.1-7.3.3]
For Hermitian operators with L^2-*kernels*
- *eigenvalues are real,*
- *eigenfunctions belonging to distinct eigenvalues are orthogonal to each other,*

- *for each non-zero eigenvalue there is a finite orthogonal base of eigenfunctions: the dimension of the base is the* <u>*multiplicity*</u> *of the eigenvalue*
- *eigenvalues form an enumerable sequence* λ_i *and, counting multiplicity, we have*

(1.1.4) $$\sum_{i=1}^{\infty} \lambda_i^2 = \|K\|^2 < \infty .$$

Usually eigenvalues λ_i and corresponding eigenfunctions ϕ_i are ordered such that

$$|\lambda_1| \geq |\lambda_2| \geq \dots .$$

The set $\{(\lambda_i,\phi_i) \mid \lambda_i \neq 0, \|\phi_i\| = 1\}$ is called a full orthonormal system of K.

<u>REMARK</u>. A full orthonormal system is not necessarily complete. It may even be finite. e.g. with $k(x,y) = p(x).p(y)$ any function Ku must be a scalar multiple of p. Thus there is only one non-zero eigenvalue and the full orthonormal system is $(\|p\|^2, p/\|p\|)$.

<u>REMARK</u>. If the eigenfunctions $\{\phi_i \mid \lambda_i \neq 0\}$ do not span the entire (separable Hilbert-) space, then we can find - orthogonal to the span$\{\phi_i\}$ - a system of orthonormal functions $\{\psi_j\}$ such that $\{\phi_i\} \cup \{\psi_j\}$ span the entire space. We notice that $K\psi_j = 0$ for all ψ_j.

<u>THEOREM</u> [Smithies, thm. 7.4.3].
Any Hermitian L^2*-kernel can be decomposed as*

(1.1.5) $$(Kx,y) = \sum_{n=1}^{\infty} \lambda_n (x,\phi_n)(\phi_n,y)$$

where $\{\lambda_n,\phi_n\}$ *is its (full orthonormal) eigensystem and* $(\cdot,\cdot)$ *denotes the* L^2*-inner product.*

<u>REMARK</u>. For an arbitrary L^2-kernel K both the operators KK^* and K^*K are Hermitian L^2-kernels. These Hermitian operators have the same set of non-negative eigenvalues. Thus we may denote by $\{\sigma_i^2,\phi_i\}$ the eigensystem of KK^* and by $\{\sigma_i^2,\psi_i\}$ the eigensystem of K^*K. The system $\{\sigma_i,\phi_i,\psi_i\}$, with $\sigma_i > 0$, is called the *singular system* of K.

THEOREM [Smithies, thm. 8.3.2].
Let K *be a* L^2*-kernel and* x *and* y L^2*-functions then* (Kx,y) *can be decomposed as*

$$(1.1.6) \qquad (Kx,y) = \sum_{n=1}^{\infty} \sigma_n (x,\phi_n)(\psi_n,y) \ .$$

Further

$$\sum_{i=1}^{\infty} \sigma_i^2 \leq \|K\|^2$$

and

$$\sum_{i=1}^{\infty} \sigma_i^2 = \|K\|^2$$

iff K is normal.

REMARK. The set $\{\chi_n\}_{n=0,1,2,\ldots} = \{1, \sin(\pi x), \cos(\pi x), \sin(2\pi x), \cos(2\pi x), \ldots\}$ is a complete orthonormal system in the Hilbert space $L^2[0,1]$. I.e. any function $x \in L^2[0,1]$ can be expressed as

$$x = \sum_{n=0}^{\infty} (x,\chi_n)\chi_n \ ,$$

with

$$\|x\|^2 = \sum_{n=0}^{\infty} |(x,\chi_n)|^2 \ , \qquad \text{(Parseval equality).}$$

Hence any orthonormal set of (eigen-) functions $\{\phi_j\}$ or $\{\psi_j\}$ can be expressed as

$$\phi_j = \sum_{n=0}^{\infty} a_{jn}\chi_n \qquad \text{with} \quad \sum_n |a_{jn}|^2 = 1.$$

$$\psi_j = \sum_{n=0}^{\infty} b_{jn}\chi_n \qquad \sum_n |b_{jn}|^2 = 1$$

and from (1.1.6) we derive

$$(Kx,y) = \sum_j (x,\phi_j)\sigma_j(\psi_j,y) = \sum_{jnm} (x,\chi_n)\, \bar{a}_{jn}\sigma_j b_{jm}\, (\chi_m,y).$$

The bound $\sum \sigma_j^2 \leq \|K\|^2$ implies $\lim_{j\to\infty} \sigma_j = 0$. From this it follows 1) that high enough frequency components in u (i.e. $\sum_{m=k}^{\infty} (\chi_m,u)\chi_m$ with m large enough) will have an arbitrarily small effect in Ku and 2) that high enough frequency components in Ku will be arbitrarily small.

1.2. Discretization of integral equations

In order to discretize the problem (1.1.2) with $n = k = 1$, $\Omega = [0,1]$, we consider the finite set of points

$$\Omega_h = \{x_0, x_1, \ldots, x_N \mid x_i = i/N\}.$$

To discretize $u : \Omega \to \mathbb{R}$, we consider

$$u_h : \Omega_h \to \mathbb{R},$$

$$u_h = \{u_0, u_1, \ldots, u_N\},$$

and we replace

$$Ku(x) = \int_0^1 k(x,y)u(y)dy$$

by

$$K_h u_h(x) = \sum_{x_\ell \in \Omega_h} w_\ell h(x,x_\ell)u_\ell, \qquad w_\ell \neq 0.$$

The discretized equation (1.1.2) now reads

$$u_m = \sum_{\ell=0}^{N} w_\ell \, k(x_m,x_\ell)u_\ell + f(x_m), \qquad m = 0,1,\ldots,N;$$

which is a simple $(N+1)\times(N+1)$ matrix problem:

$$\sum_{\ell=0}^{N} [\delta_{m,\ell} - w_\ell k(x_m,x_\ell)] \, u_\ell = f(x_m).$$

This discretized equation we denote symbolically by

$$A_h u_h = f_h.$$

How well the values $\{u_i\}$ approximate the values $\{u(x_i)\}$ depends on

1. the number of intervals N,
2. the choice of $\{w_\ell\}$,
3. the functions $k(x,y)$ and $f(x)$.

Typical error estimates are of the form

$$\max_i |u_i - u(x_i)| \le CN^{-p}.$$

Thus, the difficulty we encounter when we want to approximate (accurately) the solution of the integral equation is the large (non-sparse) system of equations to solve.

1.3. Differential equations

Let $\Omega \subset \mathbb{R}^n$, then a differential boundary value problem consists of

(1) a partial differential equation for an unknown $u : \Omega \to \mathbb{R}^k$,
(2) boundary conditions.

Examples of the differential equation are:

1. The Helmholtz equation

$$-\Delta u(x) + cu(x) = g(x) \qquad x \in \Omega \tag{1.3.1}$$

where $\Delta = \sum_{i=1}^{n} \left(\frac{\partial}{\partial x_i}\right)^2$ is the Laplacian operator.
E.g. with n = 2 the equation reads

$$-u_{xx} - u_{yy} + cu = g. \tag{1.3.2}$$

2. The general linear elliptic differential equation of 2^{nd} order.

$$Au \equiv \sum_{i,j=1} a_{ij}(x) \frac{\partial^2 u}{\partial x_i \partial x_i} + \sum_{i=1}^{n} a_i(x) \frac{\partial u}{\partial x_i} + a(x)u = g, \tag{1.3.3}$$

where the functions $a_{ij}, a_i, a : \Omega \to \mathbb{R}$ are the coefficients of the equation. The ellipticity condition is

$$\sum_{i,j=1} a_{ij}(x)\xi_i\xi_j > 0, \qquad \forall x \in \Omega, \quad \forall (\xi_1, \xi_2, \ldots, \xi_n) \in \mathbb{R}^n.$$

3. The biharmonic equation

$$\Delta^2 u(x) = g(x) \qquad x \in \Omega\ . \tag{1.3.4}$$

REMARK. The order (i.e. the highest derivative available in the equation) of an elliptic equation is always even. Usually it is denoted by 2m.

For an elliptic equation to have a unique solution, m (boundary) conditions must be given on the boundary $\partial\Omega$ of Ω.

Examples of boundary conditions are

1) Dirichlet boundary conditions

$$(1.3.5) \qquad (\frac{\partial}{\partial n})^j u(x) = g_j(x) \qquad j = 1,2,\ldots,m-1, \qquad x \in \partial\Omega,$$

where $\frac{\partial}{\partial n} = \vec{n}.\ \nabla u = \vec{n}.\ \mathrm{grad}(u)$ is the outward normal derivative; $\vec{n}$ is the outer normal direction.

2) Neumann boundary conditions, m = 1.

$$(1.3.6) \qquad \frac{\partial u}{\partial n} = h(x), \qquad x \in \partial\Omega.$$

3) Boundary conditions "of the third kind", m = 1.

$$(1.3.7) \qquad \frac{\partial u(x)}{\partial n} + \alpha(x)u(x) = \gamma(x), \qquad x \in \partial\Omega.$$

4) Mixed boundary conditions, m = 1

$$(1.3.8) \qquad \begin{aligned} u(x) &= g(x) \qquad , \quad x \in \partial\Omega_1, \\ \frac{\partial u(x)}{\partial n} &+ \alpha(x)u(x) = \gamma(x), \quad x \in \partial\Omega_2, \end{aligned}$$

with $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2, \qquad \partial\Omega_1 \cap \partial\Omega_2 = \emptyset.$

REMARK. In contrast with the integral operators, differential operators transform smooth (differentiable) functions into less smooth functions. Usually we shall have for $s \geq 0$

$$A : C^{2m+s}(\Omega) \to C^s(\Omega).$$

1.4. The weak formulation of a differential equation

Boundary value problems for differential equations often can be given in a variational formulation.

EXAMPLE. Helmholtz equation with Neumann boundary conditions can be formulated as

$$(1.4.1) \qquad B(u,v) = f(v) \qquad \text{for all } v \in H^1(\Omega),$$

where

$$f(v) = \int_\Omega g(x)v(x)dx + \oint_{\partial\Omega} h(x)v(x)dx$$

is a linear functional in v; $f : H^1(\Omega) \to \mathbb{R}$.

$$B(u,v) = \int_\Omega \nabla u(x).\nabla v(x) + cu(x)v(x)\ dx$$

is a bilinear form on u and v; $B : H^1(\Omega)xH^1(\Omega) \to \mathbb{R}$.

$H^1(\Omega)$ is the linear space of all functions u of which u and ∇u are square integrable.

DEFINITION. The Sobolev space $H^k(\Omega)$, k = 0,1,2,..., is the normed linear space of all (generalized) functions with finite norm $\|u\|_{H^k(\Omega)}$,

$$\|u\|^2_{H^k(\Omega)} = \sum_{|\alpha|\le k} \| (\frac{\partial}{\partial x_1})^{\alpha_1}(\frac{\partial}{\partial x_2})^{\alpha_2}\ldots(\frac{\partial}{\partial x_n})^{\alpha_n} u\|^2_{L^2(\Omega)},$$

$|\alpha| = \sum_{i=1}^n \alpha_i$; α_i are non-negative integers.

REMARK. $H^k(\Omega)$ is a Hilbert-space with inner product

$$(1.4.2) \qquad (u,v)_{H^k(\Omega)} = \sum_{|\alpha|\le k} ((\frac{\partial}{\partial x})^\alpha u, (\frac{\partial}{\partial x})^\alpha v)_{L^2(\Omega)}.$$

$H^0(\Omega) = L^2(\Omega)$.

REMARK. If B is *symmetric*:

$$B(u,v) = B(v,u) \qquad \forall u,v \in H^1(\Omega),$$

and B is *positive definite*:

$$B(u,v) > 0 \quad \forall u \in H^1(\Omega) \quad u \neq 0,$$

then the solution of

$$B(u,v) = f(v) \quad \forall v \in H^1(\Omega)$$

minimizes the functional

$$J(u) = B(u,u) - 2f(u). \tag{1.4.3}$$

EXAMPLE. The Helmholtz equation with $c \geq 0$ and homogeneous Dirichlet boundary conditions is symmetric and positive definite.

REMARK. A function $u \in H^1(\Omega)$
(1) can satisfy the condition

$$B(u,v) = f(v) \quad \forall v \in H^1(\Omega), \tag{1.4.4}$$

(2) and is not necessarily a $C^2(\Omega)$ function.

By partial integration we easily see that any solution of Helmholtz equation (1.3.1) with the Neumann boundary conditions (1.3.6) satisfies the equation (1.4.1). However it is possible that a solution $u \in H^1(\Omega)$ of (1.4.1) exists, which is not a solution $u \in C^2(\Omega)$ of (1.3.1) and (1.3.6). The variational or *weak formulation* (1.4.1) of the boundary value problem *is a generalization* of the classical formulation of the same problem.

REMARK. The equation: find $u \in H^1(\Omega)$ such that

$$B(u,v) = f(v) \quad \forall v \in H^1(\Omega)$$

can be seen as an infinite-dimensional linear system. We can denote this equation as

$$Au = f,$$

where A is a linear operator $A : H^1(\Omega) \to [H^1(\Omega)]^{DUAL}$. The Banach space $[H^1(\Omega)]^{DUAL}$ is also denoted by $H^{-1}(\Omega)$ of all bounded linear functionals on $H^1(\Omega)$. We easily see that

$$H^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega)$$

if we identify functions $f \in L^2(\Omega)$ with the linear functionals $f(v) = (f,v)_{L^2(\Omega)}$. Clearly $L^2(\Omega)$ and $H^{-1}(\Omega)$ contain functions that are not in $H^1(\Omega)$; these functions we can call less smooth function than those in $H^1(\Omega)$. Generalized functions such as the Dirac-delta function, defined by

$$\int_\Omega \delta_x(y)\phi(y)dy = \phi(x),$$

are bounded linear functionals on $H^1(\Omega)$. They can be considered as functions that are contained in $H^{-1}(\Omega)$ but not in $L^2(\Omega)$.

Under sufficient conditions for B (e.g. B is symmetric and positive definite), for any bounded linear functional $f \in H^{-1}(\Omega)$ we can find a solution $u \in H^1(\Omega)$ which satisfies (1.4.4). Then A is *invertible*:

$$A^{-1}: H^{-1}(\Omega) \to H^1(\Omega)$$

exists.
The problem (1.4.4) is called *regular* if A^{-1} is a bounded operator

$$A^{-1}: H^{-1+s}(\Omega) \to H^{1+s}(\Omega), \qquad s > 0.$$

Typically A^{-1} maps less smooth functions into more smooth functions. With a sufficient degree of regularity (i.e. for sufficient large s) we find

$$A^{-1}: L^2(\Omega) \to H^2(\Omega)$$

or

$$A^{-1}: H^{1+s}(\Omega) \to H^{3+s}(\Omega).$$

1.5. Discretization of differential equations

In order to discretize a differential boundary value problem, we can start either from the classical formulation or from the variational

formulation of the differential equation. The former leads to the Finite Difference Method (FDM) for the discretization, the latter to the Finite Element Method (FEM). In a number of cases both discretization methods end up with the same discretization of a given problem. In order not to obscure the notation we restrict ourselves to 2-dimensional scalar problems (n=2, k=1).

The finite difference method

Instead of the original domain of definition Ω, here we consider a finite, discrete, set of points in Ω

$$\Omega_h = \{x_{ij} \mid x_{ij} \in \Omega,\ x_{ij} = (i/N_1, j/N_2)\}.$$

To discretize the function $u : \Omega \to \mathbb{R}$, we consider $u_h : \Omega_h \to \mathbb{R}$

$$u_h = \{u_{ij}\}.$$

The differentials in the original differential equation are replaced by difference approximations.

EXAMPLE. With $x_{ij} = (ih, jh)$ e.g. we set

$$\frac{u_{i+1,j} - u_{i-1,j}}{2h} \quad \text{for} \quad u_x$$

and

$$\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h^2} \quad \text{for} \quad u_{yy} \quad \text{etc..}$$

EXAMPLE. [The Helmholtz-equation with Dirichlet boundary conditions]. For each $x_{ij} \in \Omega_h$ we find an equation

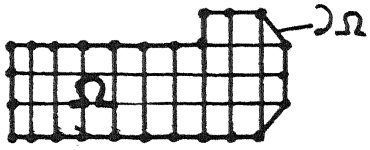
$$(1.5.1) \qquad 4u_{ij} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1} + ch^2 u_{ij} = h^2 f(x_{ij})$$

and for each $x_{ij} \in \partial\Omega_h$

$$(1.5.2) \qquad u_{ij} = g(x_{ij}).$$

In the case that $\partial\Omega$ contains all the "neighbours" of points x_{ij} in Ω_h,

the system (1.5.1) - (1.5.2) determines



as many equations as unknowns u_{ij}. They form a linear system of equations which we denote by

$$A_h u_h = f_h .$$

From this equation $u_h = \{u_{ij}\}$ can be computed. Under suitable conditions the values u_{ij} approximate the values $u(x_{ij})$.

The finite element method

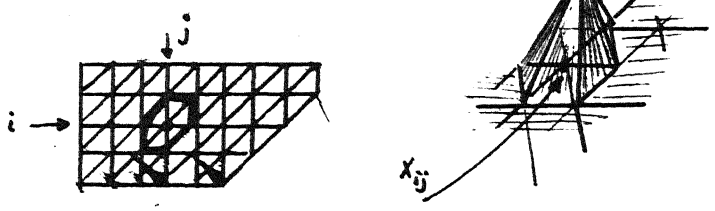
In this case our starting point is the variational formulation: find $u \in H^1(\Omega)$ such that

(1.5.3) $\quad B(u,v) = f(v) \quad$ for all $v \in H^1(\Omega)$.

Now we select from the space $H^1(\Omega)$ a finite dimensional subspace $S_h \subset H^1(\Omega)$ and we replace (1.5.3) by:
find $u_h \in S_h$ such that

$$B(u_h, v_h) = f(v_h) \quad \text{for all } v_h \in S_h .$$

EXAMPLE. We assume that Ω allows a triangulation.



As S_h we consider the space of all functions that are continuous on Ω and linear over all small triangles into which Ω is partitioned. A basis of S_h is formed by the set of "hat-functions" ϕ_{ij} ; ϕ_{ij} is a function which takes the value 1 at x_{ij} and takes the value 0 at all other vertices of triangles. Further ϕ_{ij} is piecewise linear. We see that ϕ_{ij} is nonzero only on those triangles that share the vertex x_{ij} and the (finite) basis in S_h is the set of functions

$$\{\phi_{ij} \mid x_{ij} \in \Omega \cup \partial\Omega\}.$$

We can write any function $u_h \in S_h$ as

$$u_h(x) = \sum_{i,j} u_{ij}\phi_{ij}(x)$$

and the discretized problem reads

$$\sum_{i,j} B(\phi_{ij},\phi_{k\ell})u_{ij} = f(\phi_{k\ell}) \quad \text{for all } \phi_{k,\ell} \in B.$$

This linear system we denote also by

$$A_h u_h = f_h.$$

Under suitable conditions the function u_h approximates the solution u of (1.5.3) and we can find error estimates which are typically of the form

$$\|u_h - u\|_{L^2(\Omega)} \le Ch^2 \|u\|_{H^2(\Omega)} \le Ch^2 \|f\|_{L^2(\Omega)}.$$

2. DISCRETIZATION AND APPROXIMATION

2.1. Discretization of operators and spaces

Let us be given a problem

$$(P) \qquad Fx = y,$$

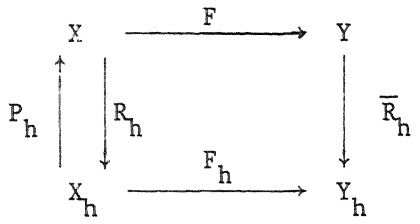
where $F : X \to Y$ and $y \in Y$ are given and X and Y are vector spaces. At first we may think of X and Y as being infinite dimensional function spaces, but - as we shall see later in this section - this is not necessary.

DEFINITION. (*The discretization of a problem*).
The *discretization of the problem* (P) is an associated problem

$$(P_h) \qquad F_h x_h = y_h,$$

where $F_h : X_h \to Y_h$ and $y_h \in Y_h$ are given and X_h and Y_h are (finite dimensional) vector spaces with $\dim(X_h) = \dim(Y_h)$. □

- By selecting $h \in H$, H an *index set*, different discretizations of the same problem are possible.
- The relation between the problem and its discretization is obtained by introducing *surjections* $R_h : X \to X_h$ and $\overline{R}_h : Y \to Y_h$.
- In order to interpret the solution of the discretized problem as an approximation to the solution of the original problem we have to define an *injection* $P_h : X_h \to X$.
- The relation between the different spaces and mappings in a discretization is summarized in the following diagram:



, $h \in H$.

REMARK. Without reference to a particular operator F, we may consider the *discretization of the spaces* X and Y by considering a set of quintuples $(X_h, P_h, R_h, Y_h, \bar{R}_h)$, $h \in H$, where

$\dim(X_h) = \dim(Y_h)$,
$P_h : X_h \to X$ an injection,
$R_h : X \to X_h$ a surjection, and
$\bar{R}_h : Y \to Y_h$ a surjection.

DEFINITION. (*The discretization of a space/an operator*)
The space X_h is also called the discretization of X; Y_h is the discretization of Y and F_h is the discretization of F.

REMARK. (*The index set H; meshwidth*)
Usually, when X is a function space over a domain in $\mathbb{R}^n$, the index h is related to a mesh spacing. Generally, all kinds of mesh spacings are possible and any particular discretization of X can be denoted by an X_h, for some $h \in H$. In particular, if only regular rectangular mesh spacings are considered, H can be identified with (a subset of) a neighbourhood of 0 in $\mathbb{R}^n_+$. In that case, with h_i being the distance between the gridpoints in the i-th direction, $1 \le i \le n$, $h = (h_1, h_2, \ldots, h_n)$ characterizes such a discretization.

To each $h \in H$ we relate a *meshwidth* $|h|$, such that $|h| > 0$. E.g., in the above example we define $|h| = \max_{j=1,\ldots,n} (h_j)$. Usually, we consider families of discretizations where x is such that

$$\forall \varepsilon > 0 \quad \exists h \in H \ni |h| < \varepsilon.$$

We shall often consider properties of tripels $(X_h, P_h, R_h)_{h \in H}$, operators $(F_h)_{h \in H}$ etc., that hold for $|h_j| \xrightarrow[j \to \infty]{} 0$, independently of the choice of the sequence $\{h_j\} \subset H$. These properties we shall denote by $\lim_{h \to 0}$ and, generally, if no confusion is possible, we denote $|h|$ simply by h.

NOTE: The index set H is not only the collection of admissible meshes (mesh-spacings), for - in general - mesh spacings do not determine the discretizations uniquely. It is possible to define different discretizations on the same mesh. It is also possible to define different discretizations with the same spaces X_h and Y_h. E.g. we can construct discretizations with different orders of accuracy on the space of gridfunctions defined on the nodal points of the same regular mesh. A prescription which, given a mesh-spacing (which should satisfy certain conditions), determines the discretizations of a problem (resp. operator or space) is called a *discretization method.*

Discretization errors

Discretization methods are used to approximate the solution of problem (P) by computation of problems (P_h). The difference between the solution $\hat{x}$ of (P) and the solution $\hat{x}_h$ of (P_h) can be called the error caused by the discretization. However, also some measure of the fact that the solution $\hat{x}_h$ only satisfies approximately the problem (P) can be called discretization error. error. Thus, different kinds of "discretization errors" can be introduced.

DEFINITION. (*Local descretization error*)
Let $x \in X$, then the *local discretization error* of x, with respect to a discretization (P_h) is defined by

$$LDE_h(x) = \tau_h(x) = F_h R_h x - \overline{R}_h F x.$$

DEFINITION. (*Global discretization error*)
If $\hat{x}$ denotes the solution of a problem (P) and $\hat{x}_h$ denotes the solution of its discretization (P_h), then the *global discretization error* of $\hat{x} \in X$ is defined by

$$GDE_h(\hat{x}) = \hat{x}_h - R_h \hat{x}.$$

DEFINITION. (*True discretization error*)
If $\hat{x}$ is the solution of (P) and $\hat{x}_h$ the solution of (P_h), then we define the *true discretization error* by

$$TDE_h(\hat{x}_h) = P_h \hat{x}_h - \hat{x}.$$

REMARK. Clearly the global and the true discretization errors can be split into two parts

$$GDE_h(\hat{x}) = (F_h^{-1} y_h - F_h^{-1} \bar{R}_h y) + (F_h^{-1} \bar{R}_h F - R_h)\, \hat{x};$$

$$TDE_h(\hat{x}_h) = (P_h F_h^{-1} y_h - P_h F_h^{-1} \bar{R}_h y) + (P_h F_h^{-1} \bar{R}_h y - F^{-1} y).$$

Hence, if we consider discretizations of the type

$$(P_h^*) \qquad F_h x_h = \bar{R}_h y, \qquad \text{i.e. } y_h := \bar{R}_h y,$$

then we have

$$GDE_h(\hat{x}) = (F_h^{-1} \bar{R}_h F - R_h)\, \hat{x},$$

and

$$TDE_h(\hat{x}_h) = (P_h\, F_h^{-1}\, \bar{R}_h - F^{-1})\, y\,.$$

REMARK. We see that the different kinds of discretization errors are mappings

$$LDE_h : X \to Y_h,$$
$$GDE_h : X \to X_h,$$
$$TDE_h : Y \to X.$$

Sequences of discretizations

DEFINITION. (*A sequence of discretizations*)
A problem (P) has a sequence of discretizations

$$((P_h)) \qquad F_h\, x_h = y_h, \qquad h \in H,$$

if $H = \{h_p\}_{p\in\mathbb{Z}}$ or $H = \{h_p\}_{p\in\mathbb{N}}$ such that $|h_p| \le |h_{p-1}|$ and $\lim_{p\to\infty} |h_p| = 0$.

REMARK. In a sequence of discretizations we denote

$$N_p = \dim(X_{h_p}) = \dim(Y_{h_p}).$$

Of course we have $\lim_{p\to\infty} N_p = \infty$.

DEFINITION. A sequence of discretizations satisfies the *regular relative mesh property* if

$$1 < h_p/h_{p+1} < C \qquad \text{for all } p,$$

with C independent of p.

REMARK. Our definition of the discretization of a problem leaves the possibility that the problem to be discretized is a finite dimensional problem itself (i.e. X and Y are finite dimensional). Hence we can discretize the problem

$$(P_h) \qquad F_h x_h = y_h, \qquad h \in H,$$

to get a discretization

$$(P_H) \qquad F_H x_H = y_H, \qquad H \in H,$$

with $\dim(X_H) = \dim(Y_H) \leq \dim(X_h) = \dim(Y_h)$.

It is clear that, with (P_h) a discretization of (P) and (P_H) a discretization of (P_h), also (P_H) is a discretization of (P). With R_{Hh}, the surjection related with (P_H) as a discretization of (P_h), $R_{Hh} : X_h \to X_H$, we construct $R_H = R_{Hh}R_h$, which is the surjection related with (P_H) as a discretization of (P). Analogously we construct $\bar{R}_H = \bar{R}_{Hh}\bar{R}_h$ and $P_H = P_h P_{hH}$.

DEFINITION. Given two discretizations of the spaces X and Y by $(X_h, Y_h, P_h, R_h, \bar{R}_h)$ and $(X_H, Y_H, P_H, R_H, \bar{R}_H)$, $h,H \in H$, these are called *related discretizations* if surjections R_{Hh} and $\bar{R}_{Hh}$ and an injection P_{hH} exist such that

$$\begin{aligned} R_{Hh} &: X_h \to X_H, & R_{Hh}R_h &= R_H, \\ \bar{R}_{Hh} &: Y_h \to Y_H, & \bar{R}_{Hh}\bar{R}_h &= \bar{R}_H, \\ P_{hH} &: X_H \to X_h, & P_h P_{hH} &= P_H. \end{aligned}$$

REMARK. We see that, if two discretizations (with $h,H \in H$) of the spaces X and Y are related, then the *coarse discretization* (with $H \in H$) can be considered as a discretization of the *fine discretization* (with $h \in H$).

DEFINITION. (*A nested sequence of discretizations*)
A sequence of discretizations is called nested iff each problem (P_{h_p}) is a discretization of a problem $(P_{h_{p+1}})$.

REMARK.
- All discretizations in a nested sequence can be discretizations of an original problem (P).
- In a nested sequence each problem (P_{h_q}) is a discretization of (P_{h_p}) iff $q \leq p$.
- Also without reference to a problem (P) we may consider a sequence or *nested sequence of discretizations of the spaces* X *and* Y.
- Obviously, all discretizations in a nested sequence are related by

$$R_{h_q} = R_{h_q h_{q+1}} R_{h_{q+1} h_{q+2}} \cdots R_{h_{p-2} h_{p-1}} R_{h_{p-1} h_p} R_{h_p},$$

$$\bar{R}_{h_q} = \bar{R}_{h_q h_{q+1}} \cdots \cdots \cdots \bar{R}_{h_{p-1} h_p} \bar{R}_{h_p},$$

$$P_{h_p} = P_{h_{p-1}} P_{h_{p-1} h_{p-2}} \cdots P_{h_{q+2} h_{q+1}} P_{h_{q+1} h_q} P_{h_q}.$$

2.2. Approximation of spaces

DEFINITION. The *approximation of a (normed) linear space* X is a set of triples $\{X_h, P_h, R_h\}_{h \in H}$, where

X_h is a finite dimensional (normed) linear space,

$P_h : X_h \to X$ is a linear injection, and

$R_h : X \to X_h$ is a linear surjection.

P_h and R_h are called *prolongations* and *restrictions* respectively.

REMARK. (*Supplying* X *and* X_h *with norms*)
Usually, the spaces X and X_h, $h \in H$, are *normed* linear spaces. However, we emphasize that the definition of the approximation is independent of these norms. In fact, the same linear spaces will often be supplied with various different norms, and hence properties of the approximations that are expressed in terms of these norms depend on this particular choice. As soon as a particular choice of norms has been made for X and X_h we denote these by $\|\cdot\|_X$ and $\|\cdot\|_{X_h}$.

DEFINITION. For a given $h \in H$, $u \in X$, $u_h \in X_h$ we define:

i) $\|u - P_h u_h\|_X$ the *difference* between u and u_h,

ii) $\|u_h - R_h u\|_{X_h}$ the *discrete difference* between u and u_h,

iii) $\|u - P_h R_h u\|_X$ the *approximation error* of u,

iv) $\|I - P_h R_h\|_{Z \to X}$ the *approximation error* of (X_h, P_h, R_h).

Here Z and X denote the same linear space, which is supplied with different norms $\|\cdot\|_Z$ and $\|\cdot\|_X$.

DEFINITION. A *discrete approximation* of X is the set $\{X_h, R_h\}_{h \in H}$, where X_h are normed linear spaces such that for all $u \in X$

$$\lim_{h \to 0} \|R_h u\|_{X_h} = \|u\|_X.$$

DEFINITION. A sequence $\{u_h \mid u_h \in X_h,\ h \in H\}$ *converges discretely* to $u \in X$ iff

$$\lim_{h \to 0} \|u_h - R_h u\|_{X_h} = 0.$$

DEFINITION. (*A convergent approximation*)
An approximation $\{X_h, P_h, R_h\}_{h \in H}$ of X is called *convergent* iff

$$\lim_{h \to 0} \|u - P_h R_h u\|_X = 0 \qquad \forall u \in X.$$

The largest positive number p for which

$$\|u - P_h R_h u\|_X = O(h^p) \qquad \forall u \in X$$

is called the *order of* approximation (or the order of convergence of the approximation).

REMARK. Clearly, the constant C in the inequality

$$\|u - P_h R_h u\|_X \le C\, h^p$$

depends on u. Generally, estimates are derived in which C depends on some (semi-)norm of u. Then we obtain estimates

$$\|u - P_h R_h u\|_X \le C \|u\|_Z\, h^p,$$

and the convergence property can be expressed as

$$\|I - P_h R_h\|_{Z \to X} = C\, h^p, \text{ for } h \to 0,$$

with C independent of u.
Notice here the essential difference between $\|\cdot\|_X$ and $\|\cdot\|_Z$. Namely, let $Z = X$ and let $N = \text{Kernel}(R_h) \subset X$, $N \neq \{0\}$; then, with $0 \neq u \in N$ we have $\|u - P_h R_h u\|_X = \|u\|_X$ and hence $\|I - P_h R_h\|_{X \to X} \geq 1$.

EXAMPLE 1. (*Finite element approximation in Sobolev spaces*)
A "finite element" is denoted by (K,P,Σ), where K is a closed subset of $\mathbb{R}^n$

K is a closed subset of $\mathbb{R}^n$ with non-empty interior and a Lipschitz continuous boundary ($K \subset \mathbb{R}^n$ is also called 'finite element'.');

P is a set of linearly independent functions defined on K, ($P = \{p_i\}$ is the set of 'basis functions' of the finite element);

Σ is a finite set of linearly independent linear forms defined over P, ($\Sigma = \{\phi_i\}$ is the set of degrees of freedom of the finite element); by definition we assume that Σ is P-unisolvent, i.e. $\dim(\Sigma) = \dim(P) = N$ and for any set of real scalars $\{a_i\}_{i=1}^N$ there exists a unique $p \in \text{Span}(P)$ which satisfies $\phi_i(p) = a_i$, $i = 1,\ldots,N$. [CIARLET, Sect. 2.3].

Let $(\hat{K},\hat{P},\hat{\Sigma})$ be a (master) finite element, for which s denotes the greatest order of partial derivatives occurring in the definition of $\hat{\Sigma}$. If, for some integers $m \geq 0$ and $k \geq 0$ and for some real numbers $p,q \in [1,\infty]$, the following inclusions hold:

(i) $\quad W^{k+1,p}(\hat{K}) \hookrightarrow C^s(\hat{K})$,

(ii) $\quad W^{k+1,p}(\hat{K}) \hookrightarrow W^{m,q}(\hat{K})$,

(iii) $\quad P_k(K) \subset P \subset W^{m,q}(\hat{K})$.

Then, there exists a constant $C(\hat{K},\hat{P},\hat{\Sigma})$ such that for all affine-equivalent finite elements (K,P,Σ) and all functions $v \in W^{k+1,p}(K)$

$$\|v - \Pi_K v\|_{m,q,K} = C(\hat{K},\hat{P},\hat{\Sigma})\,(\text{meas}(K))^{1/q-1/p}\,\frac{h^{k+1}}{\rho^m}\,|v|_{k+1,p,K},$$

where Π_K denotes the P_K-interpolant of the function and

meas(K) : the dx-measure of K,

h_K : the diameter of K,

ρ_K : the diameter of the largest ball contained in K.

[CIARLET, Thm. 3.1.5].

REMARK. (definition of a *quasi-uniform partition*)
A finite element partition is called quasi-uniform if a $C > 0$ exists such that, for all K_e from the finite element partition, we have

$$Ch \le \rho_e \le h_e \le h.$$

EXAMPLE 2. (*The Lagrange finite element approximation for* C^k*-functions*)
Let $\Sigma_e = \{x_e^N\}_{N=1}^{N_e}$ be a k-unisolvent set of nodal points of a finite element $K_e \subset \mathbb{R}^n$; K_e being star-shaped with respect to each nodal point from the set Σ_e. Let $u(x)$ be any function with the properties

(i) $u \in C^k(K_e)$,
(ii) $\mathcal{D}^{k+1}u(x)$ exists for all $x \in K_e$; i.e. the k+1-th order Fréchet-derivative exists for all x in the finite element.

Let $U_e(x)$ be the unique interpolating polynomial of degree $\le k$ of $u(x)$ at Σ_e. Then there exists a positive constant $C = C(n,k,m,\hat{\Sigma}_e)$, independent of u,h,ρ such that for each integer m, $0 \le m \le k$, we have

$$\sup_{x\in K_e} \|\mathcal{D}^m u(x) - \mathcal{D}^m U_e(x)\| = C \sup_{x\in K_e} \|\mathcal{D}^{k+1}u(x)\| \frac{h^{k+1}}{\rho^m}.$$

NOTE. $\Sigma_e = \{x_e^N\}_{N=1}^{N_e}$ is k-unisolvent iff specification of the values of $p \in P_k(K_e)$ at the point x_e^N from Σ_e determines the k-th degree polynomial $p(x)$ uniquely. [ODEN & REDDY, Thm. 6.6].

EXAMPLE 3. Let $\Omega \subset \mathbb{R}^2$ be star-shaped with respect to the points (x_i,y_j); $\{x_i\}$ and $\{y_j\}$ are the nodal points of Π_x and Π_y; Π_x and Π_y being Lagrange projectors of degree k and p respectively. Let $u(x,y) \in C^{(k+1,p+1)}(\Omega)$. Then, for $0 \le m \le k$, $0 \le \ell \le p$,

$$\|D^{(m,\ell)}(u-U_e)\|_\infty \le c_1(m,k) \|D^{(k+1,\ell)}u\|_\infty h^{k+1-m}$$
$$+ c_2(\ell,p) \|D^{(m,p+1)}u\|_\infty h^{p+1-\ell}$$
$$+ c_1(m,k)c_2(\ell,p) \|D^{(k+1,p+1)}u\|_\infty h^{k+p+2-m-\ell},$$

where $U_e = \Pi_x\Pi_y u(x,y) \in P_{k,p}(\Omega)$.
Clearly, the order of convergence of this approximation is $\min(k-m+1,p-\ell+1)$. [ODEN & REDDY, Th, 6.7].

Bounded and stable prolongations and restrictions

DEFINITION. (*A bounded restriction or prolongation*)
Restrictions and prolongations are called *bounded* (*) if

$$\|R_h\|_{X \to X_h} \leq C \quad \text{uniformly in h,}$$

or

$$\|P_h\|_{X_h \to X} \leq C \quad \text{uniformly in h,}$$

respectively.
(*) NOTE: In French literature this property is often called "stability".

DEFINITION. (*A bounded approximation*)
An approximation is called bounded iff its restrictions and prolongations are bounded.

DEFINITION. (*The optimal restriction related to a prolongation*)
Since $P_h : X_h \to X$ is an injection, a *left inverse* operator $\hat{R}_h : X \to X_h$ exists such that $\hat{R}_h P_h = I_h$; I_h being the identity operator on X_h. This $\hat{R}_h$ is called the *optimal restriction* related to P_h.

DEFINITION. (*Stable prolongations*)
Prolongations $\{P_h\}_{h \in H}$ are called *stable* if $\{P_h\}_{h \in H}$ are bounded and

$$\|\hat{R}_h\|_{X \to X_h} \leq C,$$

uniformly in $h \in H$.

REMARK. Since $R_h : X \to X_h$ is a surjection, right-inverse operators $\hat{P}_h : X_h \to X$ exist such that $R_h \hat{P}_h = I_h$. Such $\hat{P}_h$ are called prolongations related to R_h.

DEFINITION. (*The optimal prolongation related to* R_h)
The (a) right-inverse $\hat{P}_h$, related to the restriction R_h, is called the (an) optimal prolongation related to R_h if the norm

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

is minimal.
(Notice that the optimal $\hat{P}_h$ depends on the choice of this norm!)

Analogously an optimal prolongation $\hat{\bar{P}}_h$ related to $\bar{R}_h$ is defined.

DEFINITION. (*Stable restrictions*)
Restrictions $\{R_h\}_{h\in X}$ are called *stable* if $\{R_h\}_{h\in X}$ are bounded and $\{\hat{P}_h\}_{h\in H}$ exist such that

$$\|\hat{P}_h\|_{X_h\to X} \le C,$$

uniformly in $h \in H$.

DEFINITION. (*A stable approximation*)
An approximation $\{X_h,P_h,R_h\}_{h\in H}$ is called *stable* if all P_h and R_h from the approximation are stable.

DEFINITION. (*A prolongation bounded from below*)
A prolongation $P_h : X_h \to X$ is called *bounded from below* if

$$\exists C > 0 \quad \forall v_h \in X_h \quad \|P_h v_h\|_X \ge C \|v_h\|_{X_h}.$$

REMARK. The above definition is equivalent with any one of the following statements:

i) $\exists C > 0 \quad \forall v \neq 0 \quad \dfrac{\|Pv\|}{\|v\|} \ge C;$

ii) $\exists C > 0 \quad \inf\limits_{v\neq 0} \dfrac{\|Pv\|}{\|v\|} \ge C;$

iii) $\inf\limits_{v\neq 0} \dfrac{\|Pv\|}{\|v\|} > 0.$

LEMMA. *If a prolongation* $P : X_h \to X$ *is bounded from below,* $A : Z_h \to X_h$ *and* $B = PA : Z_h \to X$ *then* $\exists C > 0$ *such that* $C\|A\| \le \|B\|$.

PROOF.

$$\|B\| = \|PA\| = \sup_{w\neq 0} \frac{\|PAw\|}{\|w\|} = \sup_{w} \frac{\|PAw\|}{\|Aw\|}\frac{\|Aw\|}{\|w\|} \ge \sup_{w} C\,\frac{\|Aw\|}{\|w\|}$$

$$= C \sup_{w} \frac{\|Aw\|}{\|w\|} = C\,\|A\| \qquad \square$$

THEOREM. *Let* $\hat{R}$ *be the left inverse of a prolongation* P, *then*

i) *if* $\hat{R}$ *is bounded,* $\|\hat{R}\| \neq 0$, *then* P *is bounded from below;*

ii) *if* P *is bounded from below, then* $\hat{R} : PX_h \subset X \to X_h$ *is bounded* (*possibly* $\hat{R} : X \to X_h$ *is not bounded*).

PROOF.

i) By assumption $\hat{R}P = I$, $\|\hat{R}\| \leq C$, $C > 0$,

$$\|v\| = \|\hat{R}Pv\| \leq \|\hat{R}\| \, \|Pv\| \leq C \, \|Pv\| .$$

Therefore P is bounded from below.

ii) Since P is bounded from below

$$\exists C > 0 \quad \forall v \in X_h \quad \|Pv\| \geq C \, \|v\| ;$$

$$\forall v \in X_h \quad \|Pv\| \geq C \, \|v\| = C \, \|\hat{R}Pv\|$$

$$\forall v \in X_h \quad C^{-1} \geq \frac{\|\hat{R}Pv\|}{\|Pv\|}$$

$$C^{-1} \geq \sup_{w \in PX_h} \frac{\|\hat{R}w\|}{\|w\|} = \|\hat{R}\|_{PX_h \to X_h} . \qquad \square$$

DEFINITION. (*A restriction bounded from below*)
A *restriction* $R : V \to W$ is called *bounded from below* if

$$\exists C > 0 \quad \forall 0 \neq w \in RV \quad \exists v \in V \quad Rv = w, \ \frac{\|Rv\|}{\|v\|} \geq C.$$

REMARK. The above definition is equivalent with

$$\inf_{\substack{w \in RV \\ w \neq 0}} \sup_{v \ni Rv = w} \frac{\|Rv\|}{\|v\|} = C > 0.$$

LEMMA. *If a restriction* $R : X \to X_h$ *is bounded from below,* $A : X_h \to Z_h$ *and* $AR = B$ *with* Range(R) = Domain(A), *then*

$$\exists C > 0 \text{ such that } \|B\| \geq C\|A\| .$$

PROOF. Let X = Domain(R), Z_h = Range(R) = Domain(A), then

$$\|B\| = \|AR\| = \sup_{v \in X} \frac{\|ARv\|}{\|v\|} = \sup_{v \in X} \frac{\|ARv\|}{\|Rv\|} \frac{\|Rv\|}{\|v\|} \geq$$

$$\sup_{\substack{v^* \in X \\ Rv^* = Rv = w \in RX \\ \|Rv^*\| \geq C\|v^*\|}} \frac{\|ARv^*\|}{\|Rv^*\|} \frac{\|Rv^*\|}{\|v^*\|} \geq \sup_{w \in RX} \frac{\|Aw\|}{\|w\|} \cdot C = C\|A\| . \qquad \square$$

THEOREM. *Let* $\hat{P}$ *be the right inverse of a restriction* R, *then*

i) *if* $\hat{P}$ *is bounded then* R *is bounded from below;*

ii) *if* R *is bounded from below and* Domain$(\hat{P})$ = Range(R) *then* $\hat{P}$ *is bounded.*

PROOF.

i) By assumption

$$C^{-1} \geq \|\hat{P}\| = \sup_{w} \frac{\|\hat{P}w\|}{\|w\|} \Rightarrow \inf_{w} \frac{\|w\|}{\|\hat{P}w\|} = \inf_{w} \frac{\|R\hat{P}w\|}{\|\hat{P}w\|} \geq C.$$

Hence

$$\exists C > 0 \quad \forall w \in W \quad \exists v^* \in V \quad (v^* = Pw) \quad \|Rv^*\| \geq C\|v^*\|$$

ii) Since Domain$(\hat{P})$ = Range(R) = RV

$$\|\hat{P}\| = \sup_{w} \frac{\|\hat{P}w\|}{\|w\|} \doteq \sup_{w \in RV} \frac{\|\hat{P}w\|}{\|w\|} = \sup_{w \in RV} \frac{\|\hat{P}w\|}{\|R\hat{P}w\|} = \left[\inf_{w \in RV} \frac{\|R\hat{P}w\|}{\|\hat{P}w\|} \right]^{-1}.$$

$$\frac{1}{\|\hat{P}\|} = \inf_{w \in RV} \frac{\|R\hat{P}w\|}{\|\hat{P}w\|} = \inf_{w \in RV} \sup_{v \to RV = w} \frac{\|Rv\|}{\|v\|} = C. \qquad \square$$

2.3. Consistency, convergence and stability of a discretization

DEFINITION. (*Consistency of an operator*)
A sequence of discretizations of an *operator* F *is consistent* on a set $A \subset X$ if

$$(2.3.1) \qquad \lim_{h \to 0} \sup_{u \in A} \|F_h R_h u - \overline{R}_h F u\|_{Y_h} = 0;$$

its *order of consistency* is p if $p \in \mathbb{R}$ is the largest real number p for which

$$(2.3.2) \qquad \|\!| F_h R_h - \overline{R}_h F |\!\|_{A \subset X \to Y_h} = O(h^p) \text{ for } h \to 0.$$

REMARK. These definitions can also be written as

$$\lim_{h \to 0} \sup_{u \in A} \|\tau_h(u)\|_{Y_h} = 0$$

and

$$|||\, LDE_h \,|||_{A \subset X \to Y_h} = O(h^p)$$

respectively.

REMARK. If F and F_h are linear operators and $\{F_h\}_{h \in H}$ is consistent in A, then it is consistent in the whole of X. Hence, for a linear operator the addition "on the set A" is unnecessary. In fact, for non-linear operators we should always be aware that (some) properties only hold "in some neighbourhood".

REMARK. (*consistency of a problem*)
Let $\tilde{u} = F^{-1}y$ be the solution of a problem (P), then a sequence of discretizations of (P) is called *consistent* (of order p) *with the problem* (P) if

$$\|\tau_h(\tilde{u})\|_{Y_h} = O(h^p).$$

REMARK. Notice that, a priori, the consistency depends on the choice of the norms $\|\cdot\|_{X_h}$ and $\|\cdot\|_{Y_h}$. This is similar to the definition of the order of approximation in section 2.2 and it holds as well for the definitions of stability, discrete convergence and convergence which follow below in this section.

DEFINITION. A sequence of discretizations of an operator F is *stable* in a set $B_h = \bar{R}_h B$ with $B \subset Y$, if for all $h \in H$ there exists an F_h^{-1} with

$$|||\, F_h^{-1} \,|||_{B_h \subset Y_h \to X_h}, \tag{2.3.3}$$

uniformly in h.
A sequence of discretizations of a problem (P) is stable if the discrete operators are stable in a neighbourhood of the right-hand side function y.

REMARK. Notice that the definition of stability depends on the neighbourhood B, and the norms $\|.\|_{X_h}$ and $\|\cdot\|_{Y_h}$.

DEFINITION. A sequence of discretizations of an *operator* F *is convergent* (*of order* p) on a set $B \subset Y$ if p is the largest real number $p \in \mathbb{R}$ for which

$$|||\, P_h F_h^{-1} \bar{R}_h - F^{-1} \,|||_{B \subset Y \to X} = O(h^p) \text{ for } h \to 0.$$

REMARK. We assume that the problem (P) has a solution $F^{-1}y$. The sequence of discretizations of *the problem* (P) *is convergent* if

$$\lim_{h\to 0} \| F^{-1}y - P_h F_h^{-1} \overline{R}_h y \|_X = 0;$$

its order of convergence is p if p is the largest real number for which

$$\| F^{-1}y - P_h F_h^{-1} \overline{R}_h y \|_X = O(h^p) \text{ for } h \to 0.$$

DEFINITION. A sequence of discretizations of an operator F is *discrete convergent* in a set $B \subset Y$ if

$$\lim_{h\to 0} \sup_{z\in B} \| R_h F^{-1} z - F_h^{-1} \overline{R}_h z \|_{X_h} = 0.$$

Its *discrete order of convergence* is the largest number $p \in \mathbb{R}$ for which

$$\| \| R_h F^{-1} - F_h^{-1} \overline{R}_h \| \|_{B\subset Y \to X_h} = O(h^p) \text{ for } h \to 0.$$

REMARK. Let $u = F^{-1}y$ be a solution of problem (P), then a sequence of discretizations of (P) is called discrete convergent (of order p) iff $\{F_h\}_{h\in H}$ is discrete convergent (of order p) in a neighbourhood B of y. I.e. if the operators are discrete convergent in a neighbourhood of the solution.

REMARK. Like a vanishing LDE_h for $h \to 0$ is related to consistency of a problem or operator, the discrete convergence and the convergence are related to the global and the true discretization error respectively. To see this, for linear operators F and F_h we consider discretizations of (P) of the type

$$(P_h^*) \qquad F_h u_h = \overline{R}_h y.$$

Then:

1) A sequence of discretizations of the operator F is discrete convergent of order p iff

$$\sup_{u\in X} \| GDE_h(u) \|_{X_h} = O(h^p),$$

because

$$\sup_{u\in X} \|GDE_h(u)\|_{X_h} = \sup_{u\in X} \|u_h - R_h u\|_{X_h} =$$

$$= \sup_{u\in X} \|F_h^{-1} \bar{R}_h F u - R_h u\|_{X_h} =$$

$$= \|F_h^{-1} \bar{R}_h F - R_h\|_{X\to X_h}$$

and hence

$$\sup_{u\in X} \|GDE_h(u)\|_{X_h} \le \|\!| F_h^{-1} \bar{R}_h - R_h F^{-1} |\!\|_{Y\to X_h} \|\!| F |\!\|_{X\to Y},$$

and

$$\|F_h^{-1} \bar{R}_h - R_h F^{-1}\|_{Y\to X_h} \le \|F^{-1}\|_{Y\to X} \sup_{u\in X} \|GDE_h(u)\|_{X_h}.$$

2) A sequence of discretizations (P_h^*) of F is convergent of order p iff

$$\sup_{u_h\in X_h} \|TDE_h(u_h)\|_X = O(h^p),$$

because

$$\sup_{u_h\in X_h} \|TDE_h(u_h)\|_X = \sup_{u_h\in X_h} \|P_h u_h - u\|_X =$$

$$\sup_{y\in Y} \|P_h F_h^{-1} \bar{R}_h y - F^{-1} y\|_X = \|P_h F_h^{-1} \bar{R}_h - F^{-1}\|_{Y\to X}.$$

THEOREM. (*Consistency and stability imply discrete convergence*)
If a sequence of discretizations of a problem (P) *is stable and consistent* (*of order* p) *then it is discrete convergent* (*of order* p).

PROOF.

$$\|R_h F^{-1} y - F_h^{-1} \bar{R}_h y\|_{X_h} = \|R_h u - F_h^{-1} \bar{R}_h F u\|_{X_h} \le$$

$$\le \|\!| F_h^{-1} |\!\|_{B\subset Y_h\to X_h} \|F_h R_h u - \bar{R}_h F u\|_{Y_h} \le C h^p. \qquad \square$$

THEOREM. (*With a bounded and convergent approximation, consistency and stability imply convergence*)
Let $\{X_h, P_h, R_h\}_{h\in H}$ *be a bounded and convergent approximation (of order* q*) of the space* X. *If a sequence* $\{F_h\}_{h\in H}$ *of discretizations of the operator* F *is stable and consistent (of order* p*), then the sequence is convergent (of order* $\min(p,q)$).

PROOF.

$$\| P_h F_h^{-1} \bar{R}_h - F^{-1} \|_{B\subset Y\to X} \le$$

$$\| P_h F_h^{-1} \bar{R}_h - P_h R_h F^{-1} \|_{B\subset Y\to X} + \| P_h R_h F^{-1} - F^{-1} \|_{B\subset Y\to X} \le$$

$$\| P_h \|_{X_h\to X} \| F_h^{-1} \|_{Y_h\to X_h} \| F_h R_h - \bar{R}_h F \|_{Z\to Y_h} \| F^{-1} \|_{B\subset Y\to Z} +$$

$$\| I - P_h R_h \|_{Z\to X} \| F^{-1} \|_{B\subset Y\to Z} \le$$

$$\le C.C.C.\ h^p.C + C\,h^q \le C\,h^{\min(p,q)}. \qquad \square$$

Asymptotic expansions of the local and global discretization error

DEFINITION. The local discretization error *admits an asymptotic expansion in* h if

$$F_h R_h z - \bar{R}_h F z = h^p \bar{R}_h D(z), \qquad D: X \to Y,$$

with

$$D(z) = D_0(z) + hD_1(z) + \ldots + h^{j-1} D_{j-1}(z) + h^j D_j(z;h),$$

$$\| D_k \| = O(1), \quad k = 0,1,\ldots,j.$$

The global discretization error *admits an asymptotic expansion in* h *if*

$$F_h^{-1} \bar{R}_h F z - R_h z = h^p R_h E(z), \qquad E: X \to X,$$

with

$$E(z) = E_0(z) + hE_1(z) + \ldots + h^{j-1} E_{j-1}(z) + h^j E_j(z;h),$$

$$\| E_k \| = O(1), \quad k = 0,1,\ldots,j.$$

(Clearly both discretization errors are of order p: the first discretization is consistent of order p, the second is discrete convergent of order p.)

2.4. Galerkin discretization, relative consistency and convergence

DEFINITION. Given $\{X_h, P_h, R_h, Y_h, \bar{R}_h\}_{h\in H}$, a discretization of the spaces X and Y, we associate with the problem

$$(P) \qquad F\,x = y,$$

the *canonical* or *Galerkin discretization*

$$F_h\, x_h = y_h$$

by taking

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

and

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

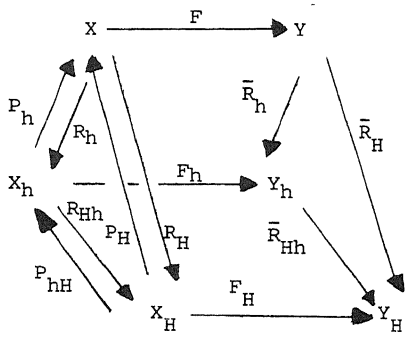
DEFINITION. *(A nested sequence of Galerkin discretizations)* If $\{X_h, Y_h, P_h, R_h, \bar{R}_h\}_{h\in H}$ is a nested sequence of discretizations of X and Y; if (P_h) is a Galerkin discretization of (P) and (P_H) is a Galerkin discretization of (P_h) then (P_H) is a Galerkin discretization of (P):

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

$$Y_H = \bar{R}_{Hh} y_h = \bar{R}_{Hh} \bar{R}_h y = \bar{R}_H\, y.$$

Thus, a nested sequence of discretizations of X and Y uniquely determines a nested sequence of Galerkin discretizations of a problem (P).

REMARK. If $\{X_h, Y_h, P_h, R_h, \bar{R}_h\}_{h\in H}$ is a nested sequence of discretizations of X and Y and if $(P_h)_{h\in H}$ are the corresponding Galerkin discretizations of (P), then the coarse-grid problem (P_H) is a Galerkin discretization of the fine-grid (P_h). They are also called *relative Galerkin* (or *canonical*) *discretizations*. The relation between the different spaces is summarized in the following diagram.



REMARK. The canonical discretization is a discretization of the kind (P_h^*)

$$(P_h^*) \qquad F_h\, x_h = \bar{R}_h\, y,$$

as considered in a remark in section 2.3.

Relative order of approximation, consistency and convergence

Analogous to the definitions of the orders of approximation, consistency and convergence of a sequence of discretizations of an operator (in the previous section), for related discretizations in a nested sequence of discretizations we can define the corresponding relative properties (possibly without reference to the original problem).

Let a coarse and a fine related discretization (in a nested sequence) be characterized by $H > h > 0$, then we define

1) the spaces Z_h and X_h have a *relative order of approximation* p:

$$\| I_h - P_{hH}\, R_{Hh} \|_{Z_h \to X_h} = O(H^p), \text{ for } H \to 0;$$

2) the operators F_h and F_H have a *relative order of consistency* p:

$$\| F_H\, R_{Hh} - \bar{R}_{Hh}\, F_h \|_{A \subset X_h \to Y_H} = O(H^p), \text{ for } H \to 0,$$

3) the operators F_h and F_H have a *relative order of convergence* p:

$$\| F_h^{-1} - P_{hH}\, F_H^{-1}\, \bar{R}_{Hh} \|_{B \subset Y_h \to X_h} = O(H^p), \text{ for } H \to 0,$$

4) the operators F_h and F_H have a *relative order of discrete convergence* p:

$$\| R_{Hh} F_h^{-1} - F_H^{-1} \bar{R}_{Hh} \|_{B_h \subset Y_h \to X_H} = O(H^p), \text{ for } H \to 0.$$

THEOREM. *Let two related discretizations of the same problem 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 discretizations are relatively consistent of order* $\min(p_1,p_2)$.

PROOF.

$$C \| F_H R_{Hh} - \bar{R}_{Hh} F_h \| \le \| F_H R_{Hh} R_h - \bar{R}_{Hh} F_h R_h \| =$$

$$= \| F_H R_H - \bar{R}_H F + \bar{R}_H F - \bar{R}_{Hh} F_h R_h \| =$$

$$\le \| F_H R_H - \bar{R}_H F \| + \| \bar{R}_{Hh} \| \, \| \bar{R}_h F - F_h R_h \|$$

$$= C H^{p_1} + C^{p_2} = C H^{\min(p_1,p_2)}.$$

The first inequality holds by lemma (on p.2.2.23). □

THEOREM. *Let* $(X_h,Y_h,P_h,R_h,\bar{R}_h)$ *and* $(X_H,Y_H,P_H,R_H,\bar{R}_H)$, $h,H \in H$, *be two related discretizations of the spaces* X *and* Y, *with a relative order of approximation (with respect to the norms* $\|\cdot\|_{Z_h}$ *and* $\|\cdot\|_{X_h}$ *) or order* p:

$$\| I_h - P_{hH} R_{Hh} \|_{Z_h \to X_h} \le C H^p.$$

Let F_H *and* F_h *be two relative Galerkin discretizations and let* $\bar{R}_{Hh} : Y_h \to Y_H$ *and* $F_h : A \subset X_h \to Y_h$ *be bounded. Then* F_H *and* F_h *are relatively consistent of order* p.

PROOF.

$$\| F_H R_{Hh} - \bar{R}_{Hh} F_h \|_{A \subset Z_h \to Y_h} =$$

$$\| \bar{R}_{Hh} F_h P_{hH} R_{Hh} - \bar{R}_{Hh} F_h \|_{A \subset Z_h \to Y_h} \le$$

$$\| R_{Hh} \|_{Y_h \to Y_H} \cdot \| F_h \|_{A \subset X_h \to Y_h} \cdot \| P_{hH} R_{Hh} - I \|_{Z_h \to X_h} \le$$

$$C \cdot C \cdot C H^p = C H^p. \qquad \square$$

REMARK. In a nested sequence we may consider the coarse discretizations of the problem (P) also as discretizations of the fine discretizations. Thus, e.g. relative convergence of two related discretizations is derived from relative consistency, stability and relative approximation order in the same way as convergence was derived from consistency, stability and approximation order:

$$|||F_h^{-1} - P_{hH}F_H^{-1}\bar{R}_{Hh}||| \leq \|I_h - P_{Hh}\| \; |||F_h^{-1}||| +$$

$$+ \|P_{hH}\| \; |||F_h^{-1}||| \; |||F_H R_{Hh} - \bar{R}_{Hh}F_h|||.$$

REMARK. When we consider related discretizations, the following is a useful identity

$$I_h - P_{hH}F_H^{-1}\bar{R}_{Hh}F_h = (I_h - P_{hH}R_{Hh}) + P_{hH}F_H^{-1}(F_H R_{Hh} - \bar{R}_{Hh}F_h).$$

REMARK. Let $F_h x_h = y_h$ and $F_H x_H = y_H$ be two related canonical discretizations of the same problem, then, for any $\tilde{R}_{Hh} : X_h \to X_H$ we have

$$I_h - P_{hH}F_H^{-1}\bar{R}_{Hh}F_h = (I_h - P_{hH}F_H^{-1}\bar{R}_{Hh}F_h)(I_h - P_{hH}\tilde{R}_{Hh}),$$

and, for any $\tilde{P}_{hH} : X_H \to X_h$ we have

$$I_h - F_h P_{hH}F_H^{-1}R_{Hh} = (I - \tilde{P}_{hH}\bar{R}_{Hh})(I - F_h P_{hH}F_H^{-1}\bar{R}_{Hh}).$$

3. THE DEFECT CORRECTION PRINCIPLE

3.0. Heuristic introduction to the defect correction principle

Often the numerical analyst is faced with the problem of solving an equation

$$Fx = y,$$

where $y \in Y$ and a mapping $F : X \to Y$ are given; X and Y are linear spaces. An element $x \in X$ has to be found such that the equation $Fx = y$ is satisfied. Often we cannot or we will not solve the equation directly because this would exceed our computational capacities. On the other hand we can solve simpler equations that are all similar to the previous equation:

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

for some arbitrary $\tilde{y} \in \tilde{Y} \subset Y$. Sometimes this yields the possibility to solve the original equation by means of an iterative process.

EXAMPLE. Solve the equation $x^2 = 3$. In other words: compute $\sqrt{3}$. We assume that we cannot find the answer immediately, but we can (1.) square the value of a real number (i.e. we can apply the operator F in the equation), and (2.) we can add and (scalar) multiply the real numbers (i.e. we use the fact that $X = Y = \mathbb{R}$ is a linear space). In this example the linear spaces are $X = Y = \mathbb{R}$. The operator $F : \mathbb{R} \to \mathbb{R}$ is defined by $Fx = x^2$ and y is defined by $y = 3.0$. We notice that F is neither surjective nor injective; F is defined on the whole of X, which (in the general case) is not necessary.
If we look for the positive solution of $x^2 = 3$, then we can apply the following iterative process

$$x_0 \in [1,2], \quad \beta \neq 0,$$
$$x_{i+1} = x_i + \beta(3 - (x_i)^2).$$

If the iterands x_i would converge to a value $x^* \in \mathbb{R}$, then we know that it would satisfy

$$x^* = x^* + \beta(3 - (x^*)^2);$$

i.e. we would have found a solution to the original equation. When does the iterative process converge?

$$(x_{i+1} - x^*) = x_i - x^* + \beta[(3 - (x_i)^2) - (3 - (x^*)^2)]$$
$$= (x_i - x^*) + \beta[(x^*)^2 - (x_i)^2]$$
$$= (x_i - x^*)(1 - \beta(x_i + x^*)).$$

This implies that

$$\frac{|x_{i+1} - x^*|}{|x_i - x^*|} = |1 - \beta(x_i + x^*)|;$$

therefore, the condition for convergence is

$$0 < \beta(x_i + x^*) < 2.$$

We know: $1 < x^* < 2$, hence we take x_0 such that $1 \leq x_0 \leq 2$. Now $2 < x_i + x^* < 4$ holds and consequently the process will converge with $0 < \beta < 1/2$.
As a numerical example we take $\beta = 1/4$, $x_0 = 1.5$. Now we find

i	x_i	x_i^2	$3 - x_i^2$
0	1.5	2.25	0.75
1	1.6875	2.84766	0.15234
2	1.72559	2.97765	0.02235
3	1.73117	2.99696	0.00304
4	1.73193	2.99959	0.00041
5	1.73204	2.99995	0.00005
6	1.73205	2.99999	0.00001
7	1.73205	3.00000	0.00000

The convergence factor is $1 - \beta(x_i + x^*) \approx 1 - 1/4.2.\sqrt{3} \approx 1 - 0.866 = 0.134 \approx$ $\approx 1/7$. In many problems we are really pleased by such a convergence factor. Analysing the above process, we write it in the abstract form

$$x_{i+1} = x_i + \beta(y - Fx_i) = (I - \beta F)x_i + \beta y,$$

where $x_i \in X$, $y \in Y$, $F : X \to Y$, $\beta : Y \to X$, $X = Y = \mathbb{R}$. The convergence is derived from

$$|x_{i+1} - x^*| \leq \|| I - \beta F \||\, |x_i - x^*|,$$

from which it is clear that we have a convergent process if $\|| I - \beta F \|| < 1$, i.e. if the operator β is close enough to F^{-1}. In other words β should be a sufficiently close approximation to the solution operator F^{-1}.

3.1. The basic principle

In principle, a defect correction process is an iterative process to solve an equation that we cannot or we do not want to solve directly:

(P) $\qquad Fx = y,$

where $F : A \subset X \to Y$. This short notation means that $F : A \to Y$ is a mapping, A is a subset of X and X and Y are normed linear spaces. In general F is not linear, F is not defined on the whole of X and F is neither injective nor surjective. We assume that there exist subsets $A \subset X$ and $B \subset Y$ such that F is defined on the whole of A, and $\forall y \in B\ \exists x \in A$ such that $Fx = y$ (i.e. the mapping $F : A \to B$ is surjective). In addition we often require that there exists a unique $x \in A$ such that $Fx = y$ (i.e. in addition the mapping $F : A \to B$ is injective and hence it is bijective).

As an introduction to a more formal approach in the following paragraph, we first proceed informally to introduce the notion of "approximate inverse". We assume that we *can* solve some approximations $(\tilde{P})$ of the problem (P), i.e. for all $\tilde{y} \in \tilde{Y} \subset B$ we can solve the equation

$(\tilde{P})$ $\qquad \tilde{F}\tilde{x} = \tilde{y}, \qquad \tilde{x} \in X,$

where $\tilde{F} : X \to \tilde{Y}$ is some "approximation" of the operator F.

Formally we describe this as follows: we assume that for some subset $\tilde{Y} \subset B$, with $y \in \tilde{Y}$, there exists a mapping

$$\tilde{G} : \tilde{Y} \to A,$$

which we shall call the *approximate inverse* of F. The meaning of $\tilde{G}$ is, that for any $\tilde{y} \in \tilde{Y}$ an approximation to the solution of the equation $Fx = \tilde{y}$ is given by $\tilde{G}\tilde{y} \in A$. The mapping $\tilde{G}$ needs not to be linear and is neither necessarily injective nor surjective.

REMARK. If $\tilde{G}$ is *not* surjective, then possibly $x \notin \tilde{G}\tilde{Y}$, with x the solution of $Fx = y$.

REMARK. If $\tilde{G}$ is injective, then an $\tilde{F}: \tilde{G}\tilde{Y} \to \tilde{Y}$ exists such that $\tilde{F}\tilde{G} = I_{\tilde{Y}}$ where $I_{\tilde{Y}}$ is the identity operator on $\tilde{Y}$. Then $\tilde{F}$ is "*an approximation to* F". Here we notice that $\tilde{F}$ is only defined on $\tilde{G}\tilde{Y}$ and not on A!

In a Defect Correction Process the solution of the original problem (P) is found (or approximated) by the iterative application of one (or more) approximate inverse(s) $\tilde{G}$.

In its most elementary form we have two versions of the defect correction process for the solution of (P):

The *first defect correction process* (DCPA)

$$\text{(DCPA)} \qquad \begin{cases} x_0 \in A, \\ x_{i+1} = (I - \tilde{G}F)x_i + \tilde{G}y, \end{cases}$$

with the standard starting value

$$x_0 = \tilde{G}y;$$

and the *second (or dual) defect correction process* (DCPB)

$$\text{(DCPB)} \qquad \begin{cases} \ell_0 \in \tilde{Y}, \quad x_i = \tilde{G}\ell_i, \\ \ell_{i+1} = (I - F\tilde{G})\ell_i + y, \end{cases}$$

with the standard starting value

$$\ell_0 = y.$$

REMARK. DCPA is completely described by $F,\tilde{G},y$ and x_0; (DCPB) is completely described by $F,\tilde{G},y$ and ℓ_0.

REMARK. In order that the above defect correction processes make sense (are well defined) a number of conditions should be satisfied, such as:

for DCPA : $\{x_i\} \subset A$ and $\{Fx_i\} \subset \tilde{Y}$;

for DCPB : $\{\ell_i\} \subset \tilde{Y}$.

Note that $y \in \tilde{Y}$ follows from the definition of $\tilde{G}$, which was defined on $\tilde{Y}$ with $y \in \tilde{Y}$.

REMARK. With DCPA we use the fact that X is a linear space and not the fact that Y is. With DCPB we use the fact that Y is a linear space and not the fact that X is. (Note that both F and $\tilde{G}$ may be non-linear!)

DEFINITION. A value $x^* \in X$ is called a *stationary point* (or a *fixed point*) of an iterative process

$$x_{i+1} = P(x_i, x_{i-1}, \ldots)$$

if x^* satisfies

$$x^* = P(x^*, x^*, \ldots).$$

DEFINITION. The *convergence factor* of an iterative process to a stationary point x^* is defined by

$$\sup_{x_0 \in A} \sup_{i \geq 0} \frac{\|x_{i+1} - x^*\|}{\|x_i - x^*\|}.$$

3.2. The first Defect Correction Process

The first thing we notice when we consider DCPA is that the solution x of (P) is a fixed point of DCPA; moreover, for any stationary point x^* of DCPA, we have

$$(3.2.1) \qquad \tilde{G}F\, x^* = \tilde{G}y = \tilde{G}F\, x\,.$$

(Notice that $x^* \in A$ and $Fx^* \in Y$ are natural assumptions that go with the assumptions of x^* to be a stationary point of DCPA.)

As a direct consequence of (3.2.1) we find the following

THEOREM. *If* DCPA *has a stationary point* $x^* \in X$ *with* $Fx^* \in \tilde{Y}$ *and if* $\tilde{G}$ *is injective, then* $Fx^* = y$ *(i.e. then* x^* *is a solution of* (P)*).*

REMARK. Even, if $\tilde{G}$ is not injective, the solution x of (P) and the fixed point x^* of DCPA are mapped by $\tilde{G}F$ onto the same element of $\tilde{G}\tilde{Y}$ (although we have not necessarily $Fx^* = y = Fx$). In other words: $\tilde{G}$ defines subsets of $\tilde{Y}$ (viz. the sets of points that are mapped to the same point of X) and Fx^* and Fx now are elements of the same subset.

DEFINITION. The *amplification operator of* DCPA is defined as

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

THEOREM. *The convergence factor of* DCPA *to a fixed point* $x^* \in A$, $Fx^* \in \tilde{Y}$, *is bounded by* $|||\, I - \tilde{G}F \,|||_{A\subset X\to X}$.

PROOF. Let x_i be an arbitrary iterand of DCPA, then

$$x_{i+1} - x^* = (I - \tilde{G}F)x_i - (I - \tilde{G}F)x^*.$$

Hence,

$$\|x_{i+1} - x^*\| = \|(I - \tilde{G}F)x_i - (I - \tilde{G}F)x^*\|$$

$$\leq |||\, I - \tilde{G}F \,|||\ \|x_i - x^*\|$$

and

$$\frac{\|x_{i+1} - x^*\|}{\|x_i - x^*\|} \leq |||\, I - GF \,|||_{A\subset X\to X}. \qquad \square$$

If $|||\, I - \tilde{G}F \,||| < 1$, the sequence of iterands of DCPA converges and it might make some sense to call $\tilde{G}$ an approximate inverse of F indeed. If F is injective, we can give the following definition.

DEFINITION. The *approximation error* of $\tilde{G}$ for the solution of (P) is

$$\text{Approx. error}(\tilde{G};F,x) \stackrel{D}{=} \sup_{\xi\in A} \{\|x-\xi\| \,|\, \tilde{G}Fx = \tilde{G}F\xi\}.$$

As a direct consequence of this definition we have for any injective $\tilde{G}$

$$\text{Approx. error}(\tilde{G};F,x) = 0.$$

REMARK. In the special case that $\tilde{G}$ is an *affine mapping*, i.e. if we can write $\tilde{G}y$ as

$$\tilde{G}y = \tilde{G}'y + \tilde{G}0, \qquad \forall y \in Y,$$

where G' is a linear operator, then we may write DCPA as

$$\begin{cases} x_0 \in X, \\ x_{i+1} = x_i - \tilde{G}'(Fx_i - y). \end{cases}$$

3.3. The second Defect Correction Process

If $\ell^* \in \tilde{Y}$ is a stationary point of DCPB, then we clearly have

$$F\tilde{G}\ell^* = y.$$

Hence, we immediately have the following

THEOREM. *If* DCPB *has a stationary point* $\ell^* \in \tilde{Y}$, *then* $\tilde{G}\ell^* = x$ *is a solution of* (P) *in* $\tilde{G}\tilde{Y} \subset X$.

REMARK. If $F : A \to B$ is injective, then $\tilde{G}\ell^*$ is the unique solution of (P).

REMARK. If $\tilde{G} : \tilde{Y} \to A$ is not surjective, then possibly $x \notin \tilde{G}\tilde{Y}$ and hence no $\ell^* \in \tilde{Y}$ exists such that $\tilde{G}\ell^* = x$. In that case no fixed point $\ell^* \in \tilde{Y}$ can exist.

DEFINITION. The *amplification operator of* DCPB is defined as

$$\bar{M} = I - F\tilde{G}.$$

THEOREM. *The convergence factor of* DCPB *to a fixed point* $\ell^* \in \tilde{Y}$ is bounded by $\| I - F\tilde{G} \|_{\tilde{Y} \subset Y \to Y}$.

PROOF.

$$\|\ell_{i+1} - \ell^*\| \leq \| I - F\tilde{G} \| \, \|\ell_i - \ell^*\|. \qquad \square$$

THEOREM. *If* $\tilde{G}$ *is injective, we can define its left-inverse* $\tilde{F}$ *and* DCPB *can be written as*

$$\begin{cases} x_0 \in \tilde{G}\tilde{Y} \\ \tilde{F}\, x_{i+1} = (\tilde{F} - F)x_i + y. \end{cases}$$

PROOF.

$$\tilde{F}\, x_i = \tilde{F}\tilde{G}\, \ell_i = \ell_i,$$

and

$$\tilde{F}\, x_{i+1} = \tilde{F}\, x_i - F\tilde{G}\, \ell_i + y = F\, x_i - \tilde{F}\, x_i + y$$

$$= (\tilde{F} - F)x_i + y. \qquad \square$$

REMARK. In many problems the operator $(\tilde{F} - F)$ can be much simpler to compute than either $\tilde{F}$ or F.

THEOREM. *If $\tilde{G}$ is injective, then the convergence factor of* DCPB *is bounded by*

$$|||\, \tilde{F} - F\, |||_{\tilde{G}\tilde{Y} \subset X \to Y}\ |||\, \tilde{G}\, |||_{\tilde{Y} \subset Y \to X},$$

where $\tilde{F}$ is the left-inverse of $\tilde{G}$.

PROOF.

$$|||\, I - F\tilde{G}\, ||| = |||\, \tilde{F}\tilde{G} - FG\, ||| =$$

$$= \sup \|\, (\tilde{F}\tilde{G} - F\tilde{G})x - (\tilde{F}\tilde{G} - F\tilde{G})y \| \,/\, \|\, x - y \|$$

$$= \sup \|\, \tilde{F}\tilde{G}\, x - F\tilde{G}x - \tilde{F}\tilde{G}y + F\tilde{G}y \| \,/\, \|\, x - y \|$$

$$= \sup \|\, (\tilde{F} - F)\tilde{G}x - (\tilde{F} - F)\tilde{G}y \| \,/\, \|\, x - y \|$$

$$= \sup \frac{\|\, (\tilde{F} - F)\tilde{G}x - (\tilde{F} - F)\tilde{G}y \|}{\|\, \tilde{G}x - \tilde{G}y \|} \cdot \frac{\|\, \tilde{G}x - \tilde{G}y \|}{\|\, x - y \|}$$

$$\leq |||\, \tilde{F} - F\, |||\ |||\, \tilde{G}\, |||\, . \qquad \square$$

REMARK. Clearly, the above bound of the convergence factor can also be expressed, in terms of relative error of $\tilde{F}$ and the condition of $\tilde{F}$, by

$$\frac{\|\, \ell_{i+1} - \ell^* \|}{\|\, \ell_i - \ell^* \|} \leq \frac{|||\, F - \tilde{F}\, |||}{|||\, \tilde{F}\, |||}\ \mathrm{cond}(\tilde{F}).$$

THEOREM. *If* $\tilde{G}$ *is an affine mapping, then the sequences* $\{x_i\}$ *in* (DCPA), *and* $\{x_i\}$ *in* (DCPB), *defined with their standard starting values* $x_0 = \tilde{G}y$ and $\ell_0 = y$, *are identical.*

PROOF. Let $\{\ell_i\}_{i=0,1,2,\ldots}$ and $\{x_i\}_{i=0,1,2,\ldots}$ be defined as in DCPB, then

i) $x_0 = \tilde{G}\,\ell_0 = \tilde{G}y$, and

ii) $$\begin{aligned} x_{i+1} = \tilde{G}\,\ell_{i+1} &= \tilde{G}(\ell_i - F\tilde{G}\,\ell_i + y) \\ &= \tilde{G}\,0 + \tilde{G}'\ell_i - \tilde{G}\,0 - \tilde{G}'F\tilde{G}\,\ell_i + \tilde{G}\,0 + \tilde{G}'y \\ &= \tilde{G}\,\ell_i - \tilde{G}\,F\tilde{G}\,\ell_i + \tilde{G}y \\ &= x_i - \tilde{G}F\,x_i + \tilde{G}y = (I - \tilde{G}F)x_i + \tilde{G}y. \end{aligned}$$

I.e. the values from the sequence $\{x_i\}$ satisfy exactly the generation rules for the sequence $\{x_i\}$ from DCPA. Hence, both sequences are identical. □

REMARK. It is clear from the proof of the last theorem that for general $\tilde{G}$ both processes DCPA and DCPB yield different sequences $\{x_i\}$.

3.4. Further remarks on DCPB

If $\tilde{G}$ in DCPB is not surjective (i.e. possibly $x \notin \tilde{G}\tilde{Y}$, with x the solution of $Fx = y$, and hence possibly there exists no fixed point for DCPB), then sometimes we still can write

(3.4.1) $$\tilde{G} = \tilde{\Gamma}\Delta ,$$

where $\Delta : Y \to \Delta Y$ is a linear projection ($\Delta \tilde{Y} \subset B$), and $\tilde{\Gamma} : \Delta Y \to X$ *is* surjective.

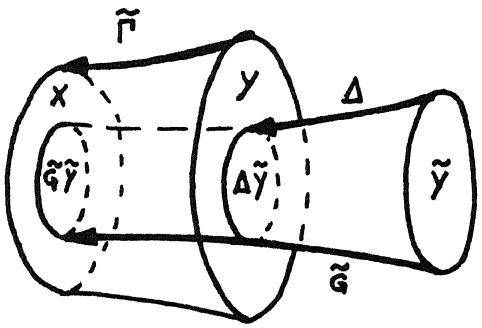


Fig. 3.4.1. The mappings $\tilde{G}$, Δ and $\tilde{\Gamma}$.

The iterands $\{\ell_i\}$ in the iterative process DCPB are all in $\tilde{Y}$. If, instead of $\ell_i \in \tilde{Y}$, we consider their projections $\Delta \ell_i \in \Delta \tilde{Y}$, we get the following iterative process of which all iterands are in $\Delta \tilde{Y}$:

$$\Delta \ell_{i+1} = \Delta \ell_i - \Delta F\tilde{G} \ell_i + \Delta y$$

$$= \Delta \ell_i - \Delta F\tilde{\Gamma} \Delta \ell_i + \Delta y.$$

With the definitions $\lambda_i = \Delta \ell_i$ and $\xi_i = \tilde{\Gamma} \lambda_i$ we get

$$(3.4.2) \qquad \begin{cases} \lambda_{i+1} = \lambda_i - \Delta F \tilde{\Gamma} \lambda_i + \Delta y, \\ \lambda_0 = \Delta \ell_0 = \Delta y. \end{cases}$$

This is exactly the DCPB for the problem:

$$(\Delta P) \qquad \Delta F \xi = \Delta y ,$$

where $\tilde{\Gamma}$ takes the part of the approximating inverse of ΔF. Since, by hypothesis, $\tilde{\Gamma}$ is surjective, this new DCP *has* a fixed point λ^* and the solution (ΔP) is found as $\xi^* = \tilde{\Gamma} \lambda^*$.

REMARK. Notice that $\xi^* \in \tilde{\Gamma} \Delta \tilde{Y} = \tilde{G}\tilde{Y}$. The problem (ΔP) can now be considered as: find $\xi \in \tilde{G}\tilde{Y}$ such that

$$\Delta(F\xi - y) = 0.$$

By application of a projection Δ to the residual of the problem (P), more solutions in X are generated which satisfy the equation. The projection Δ has to become so strong that even a solution becomes in $\tilde{G}\tilde{Y}$. If we find a Δ such that the problem has a solution for all $y \in \tilde{Y}$, we have found a decomposition $\tilde{G} = \tilde{\Gamma}\Delta$ that satisfies the hypotheses.

In the case that the operator $\tilde{\Gamma}$ in the decomposition $\tilde{G} = \tilde{\Gamma} \Delta$ is not only surjective but also injective, we can formulate the following

THEOREM. *If the approximate inverse* $\tilde{G}$ *in* DCPB *can be decomposed as* $\tilde{G} = \tilde{\Gamma} \Delta$, *where* Δ *is a linear projection and* $\tilde{\Gamma} : \Delta \tilde{Y} \to \tilde{G}\tilde{Y}$ *a bijective mapping, then a* $\tilde{\Phi} = (\tilde{\Gamma})^{-1} : \tilde{G}\tilde{Y} \to \Delta \tilde{Y}$ *exists, and a* DCPB *in* $\Delta \tilde{Y}$ *can be formulated:*

$$\begin{cases} \xi_0 \in \tilde{\Gamma}\Delta\tilde{Y} = \tilde{G}\tilde{Y}, \\ \tilde{\Phi}\xi_{i+1} = (\tilde{\Phi} - \Delta F)\xi_i + \Delta y, \end{cases}$$

which has a fixed point $\xi^* \in \tilde{G}\tilde{Y}$ *such that* $\Delta(F\xi^* - y) = 0$.

PROOF. Follows immediately from (3.4.2) and Theorem 3.3.

3.5. Another Defect Correction Process for non-linear $\tilde{G}$

In this section we give a generalization of DCPA . In the linear case we can write a defect correction step DCPA

(3.5.1) $\quad x_{i+1} = x_i - \tilde{G}Fx_i + \tilde{G}y$

as

(3.5.2) $\quad x_{i+1} = x_i + \tilde{G}(y - Fx_i).$

For general - nonlinear - $\tilde{G}$, the solution of $Fx = y$ is not a fixed point of the latter iteration. In (3.5.2) the operands of $\tilde{G}$ are in the neightbourhood of zero, whereas in (3.5.1) they are in the neightbourhood of y and Fx_i. An approximation (linearization) of the non-linear DCPA (3.5.1) can be given by

$$x_{i+1} = x_i + \tilde{G}'(\tilde{y})(y - Fx_i),$$

where $\tilde{G}'(\tilde{y})$ denotes the Fréchet derivative of $\tilde{G}$ at $\tilde{y}$, where $\tilde{y}$ is thought to be in the neightbourhood of both y and Fx_i. The Fréchet derivative not being available for computation, we may approximate further

$$\tilde{G}'(\tilde{y})\delta \text{ by } \tilde{G}(\tilde{y}+\delta) - \tilde{G}(\tilde{y}).$$

Also noting that

$$\tilde{G}'(\tilde{y})\delta = \mu\,\tilde{G}'(\tilde{y})(\delta/\mu),$$

we may write down a new Defect Correction Process

(DCPC) $\quad x_{i+1} = x_i + \mu\,\tilde{G}(\tilde{y} + (y - Fx_i)/\mu) - \mu\tilde{G}\tilde{y}.$

In this iteration step the parameters μ and $\tilde{y}$ are still free to choose.

REMARKS. With respect to this new Defect Correction Process we notice:

1. Near a solution of $Fx = y$ the operator $\tilde{G}$ is applied only in the neighbourhood of $\tilde{y}$.
2. In the general case (i.e. for any value of μ and $\tilde{y}$), the solution of $Fx = y$ is a fixed point of DCPC .
3. With $\mu = -1$ and $\tilde{y} = y$, DCPC is identical with DCPA .
4. For arbitrary μ and $\tilde{y}$, with $\tilde{G}$ affine, DCPC is identical with DCPA and hence, by Theorem 3.3. also equivalent with DCPB .
5. The amplification factor of DCPC is given by

$$\frac{\|x_{i+1} - x\|}{\|x_i - x\|} \leq \|I - \tilde{G}'F'\| + \|\tilde{G}'\| \, |||F^*||| + |||\tilde{G}^*||| \, \|F'\| + |||\tilde{G}^*||| \; |||F^*|||,$$

where $\tilde{G}'$ and $\tilde{G}^*$ are defined by

$$\tilde{G}(\tilde{y}+\delta) - \tilde{G}(\tilde{y}) = \tilde{G}'\delta + G^*\delta,$$

with $\tilde{G}'$ linear and $\tilde{G}^*$ such that

$$\frac{\|\tilde{G}^*\delta\|}{\|\delta\|} \to 0 \quad \text{as } \delta \to 0,$$

i.e. $|||G^*|||$ is arbitrarily small in a sufficiently small neighbourhood of $\tilde{y}$. F' and F^* are defined analogously as $F(x+\varepsilon) - F(x) = F'\varepsilon + F^*\varepsilon$. We note that, for Fréchet differentiable F and $\tilde{G}$, by this definition the Lipschitz constants $|||F^*|||$ and $|||G^*|||$ can be taken arbitrarily small if we restrict $\{x_i\}$ to a sufficiently small neighbourhood of x.
Note: by the above definition is $\tilde{G}'$ the Fréchet-derivative of $\tilde{G}$ at $\tilde{y}$ and is F' the Fréchet-derivative of F at x.

3.6. Examples of defect correction processes

Example 1. *The iterative refinement of linear systems.*
In this case the problem (P) is the solution of the finite dimensional linear system

$$Fx = y, \tag{3.6.1}$$

where $F : \mathbb{R}^n \times \mathbb{R}^n$ is a square matrix and $x,y \in \mathbb{R}^n$ are n-vectors.

The approximate inverse $\tilde{G}$ represents the numerical solution by means of (an approximation of) a LU-decomposition, which had been obtained by numerical means and for which we may write

$$LU = F + E; \tag{3.6.1}$$

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

$$\left.\begin{aligned} LU\,x_0 &= y, \\ r_{i+1} &= y - Fx_i, \\ LU\,d_{i+1} &= r_{i+1}, \\ x_{i+1} &= x_i + d_{i+1}, \end{aligned}\right\} \quad i = 0,1,2,\ldots\ . \tag{3.6.2}$$

Clearly, this is DCPA with $\tilde{G} = (F+E)^{-1}$, and because of the linearity of $\tilde{G}$, the process is equivalent to a DCPB. As a result of Theorem 3.3 we know the upperboud of the convergence factor:

$$\frac{\|E\|}{\|F+E\|}\ \mathrm{cond}(F+E).$$

We can also obtain the following convergence result in terms of cond(F).

THEOREM. *The sequence of iterands in* (3.6.2) *converges if*

$$\mathrm{cond}(F)\ \|E\|/\|F\| < 1/2.$$

PROOF.

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

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

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

If $\|F^{-1}E\| < 1$, then

$$\|I - \tilde{G}F\| = \frac{\|F^{-1}E\|}{1 - \|F^{-1}E\|}\ .$$

From $1/2 > \mathrm{cond}(F)\ \|E\|/\|F\| = \|F\|\ \|F^{-1}\|\ \|E\|/\|F\| \geq \|F^{-1}E\|$ the convergence of the DCPA follows immediately. □

EXAMPLE 2. *Iterative methods for the solution of linear systems.*
Many of the well-known iterative methods for the solution of linear systems can easily be recognized as Defect Correction Processes. For all these methods $\tilde{G}$ is linear and, hence, DCPA and DCPB are equivalent. Here we shall identify a number of these methods for the solution of the square linear system $Ax = b$ as Defect Correction Processes.

EXAMPLE 2.1. *The Jacobi-method.*
The Jacobi-method:

$$\mathrm{diag}(A)x_{i+1} = b + (\mathrm{diag}(A) - A)x_i$$

can be written as

$$x_{i+1} = x_i + \tilde{G}(b-Ax_i) = (I-\tilde{G}A)x_i + \tilde{G}b,$$

with $\tilde{G} = (\mathrm{diag}(A))^{-1}$.

EXAMPLE 2.2. *The Gauss-Seidel method.*
Let A be decomposed as $A = L + U$, where U is strict upper-triangular and L is lower triangular; then the Gauss-Seidel process reads

$$Lx_{i+1} = b - Ux_i,$$

i.e. a defect correction process with $\tilde{G} = L^{-1}$.

EXAMPLE 2.3. *The relaxation methods JOR, SOR, RF and GRF.*
All "stationary fully consistent iterative methods of degree one" for the solution of $Ax = b$ can be written as

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

where P is a non-singular matrix (cf. YOUNG [1971]). With $P = pI$, p a scalar and I the identity matrix, it is called a Stationary Richardson method (RF); with P a non-singular diagonal matrix it is a Generalized Stationary Richardson method (GRF); with $P = \omega\tilde{G}$, $\tilde{G}$ as in example 2.1 it is a Jacobi

over-/under-relaxation method (JOR) and with $P = \omega\tilde{G}$, $\tilde{G}$ as in example 2.2 it is is a SOR-method.

EXAMPLE 3. *Modified Newton Iteration.*
In this case the problem (P) is the solution of a non-linear equation

$$Fx = y,$$

with a Fréchet-differentiable operator F. The Fréchet derivative is the linear operator F' such that $\|Fx - F\xi - F'(x-\xi)\| = o(\|x-\xi\|)$. The relation

$$Fx - Fx_i = F'(x-x_i) + o(\|x-x_i\|)$$

or, equivalently

$$x - x_i = (F')^{-1}(y - Fx_i + o(\|x-x_i\|)),$$

suggests the modified Newton iteration:

$$x_{i+1} = x_i + E^{-1}(y - Fx_i),$$

where the non-singular linear operator E is an approximation to F'. Clearly, this is a DCPA with $\tilde{G} = E^{-1}$ and, since E^{-1} is linear, the process can also be written as a DCPB. Here we notice that in a proper Newton-process (not the modified Newton iteration) the approximate Fréchet derivative E is updated during the iteration process. This kind of generalization of the elementary DCP will be treated in section 4.1.

EXAMPLE 4. *An analytic examples*, (cf. STETTER, 1978).
We consider the two-point boundary-value problem

$$(3.6.3) \qquad \begin{cases} x'' - e^x = 0 & \text{on } (-1,+1), \\ x(-1) = x(+1) = 0. \end{cases}$$

This defines the problem

$$Fx = 0$$

where

$$F : C_0^2[-1,+1] \to C(-1,+1).$$

We construct an approximate problem, replacing e^x by $0.99 + 0.81x$ (i.e. a reasonable approximation if $-0.4 \leq x \leq 0.0$). Thus we get the approximate problem $\tilde{F}x = y$, viz.

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

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

$$x(t) = \int_{-1}^{+1} K(t,z)(y(z) + 0.99)dz,$$

for some suitable kernel-function $K(t,z)$. This integral operator defines an approximate inverse $\tilde{G}$ for the problem (3.6.3). With this $\tilde{G}$ we can construct a DCPA or DCPB to find the solution of (3.6.3). Both processes are equivalent since $\tilde{G}$ is an affine operator.

EXAMPLE 5. *A Defect Correction Process for a singular linear system.* We consider the finite-dimensional linear system

$$Ax = b,$$

where A is singular; A is approximated by a nonsingular $\tilde{A}$ and we consider the DCPB

$$\tilde{A}\, x_{i+1} = \tilde{A}x_i - Ax_i + b$$

or, equivalently, the DCPA

$$\begin{cases} x_0 = Bb, \\ x_{i+1} = (I - BA)x_i + b, \end{cases}$$

where $B = \tilde{A}^{-1}$. Generally, x_i can be written as

$$x_i = \sum_{i=0}^{i-1} (I - BA)^j Bb.$$

If we take e.g.

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \qquad \tilde{A} = \begin{pmatrix} \varepsilon & 0 \\ 1 & 1 \end{pmatrix},$$

we have

$$B = \begin{pmatrix} 1/\varepsilon & 0 \\ -1/\varepsilon & 1 \end{pmatrix}, \text{ and } I - BA = \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix};$$

also

$$(I - BA)^j = \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix}$$

and hence

$$x_i = \sum_{j=0}^{i-1} \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1/\varepsilon & 0 \\ -1/\varepsilon & 1 \end{pmatrix} b = \frac{i}{\varepsilon} \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix} b.$$

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

$$I - AB = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Now we take a slightly more general A and a general B:

$$A = \begin{pmatrix} 0 & 0 \\ a & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} p & q \\ r & s \end{pmatrix};$$

The amplification operator I - BA reads

$$I - BA = \begin{pmatrix} 1-aq & q \\ as & 1-s \end{pmatrix}$$

and has the eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 1 - s - aq$. Because of the eigenvalue 1 in the amplification operator, it is clear that no B can be found such that the process will converge. More generally, for arbitrary matrices F or $\tilde{G}$ we know that $\|I - F\tilde{G}\| \geq 1$ and $\|I - \tilde{G}F\| \geq 1$.

EXAMPLE 6. *The non-existence of a fixed point* $\hat{\ell}$, whereas $\hat{x}$ *exists.*
Our original problem $Fx = y$ is to find the solution of the initial value problem

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

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

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

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

$$X = C^1[0,1],$$

$$A = C^1_B[0,1] = \{x \mid x \in X,\ x(0) = 1\},$$

$$Y = C^0[0,1],$$

$$B = \tilde{Y} = Y,$$

$$\tilde{\tilde{GY}} = \{(1+Mt) \mid M \in \mathbb{R}\},$$

$$F\tilde{\tilde{GY}} = F\{(1+Mt)\} = \{M + \lambda + \lambda Mt \mid M \in \mathbb{R}\}.$$

First we apply the DCPB with $\ell_0 = y = 0$, to get

$$\ell_1 = \ell_0 = F\tilde{G}\,\ell_0 + y = \frac{-\lambda^2}{1+\lambda} \cdot (1-t),$$

$$x_1 = \tilde{G}\,\ell_1 = 1 - \frac{\lambda t}{1+\lambda}.$$

By induction we easily show that, for $n = 1,2,\ldots,$

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

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

Now we apply the DCPA to get

$$x_0 = \tilde{G}y = 1 - \frac{\lambda t}{1+\lambda},$$

$$Fx_0 = \frac{\lambda^2}{1+\lambda}\,(1-t),$$

$$\tilde{G}Fx_0 = 1 - \frac{\lambda t}{1+t} = x_0,$$

$$x_1 = x_0 - \tilde{G}Fx_0 + x_0 = x_0,$$

Thus we get $x_n = x_0$ for $n = 0,1,2,\ldots$.

REMARK. Because $\tilde{G}$ is affine, we knew beforehand that the sequences $\{x_n\}$ for DCPA and DCPB are equal. Clearly, $\tilde{G}$ is *not* injective in this example. The fixed point $\hat{x}$ of the DCPA is *not* the solution of the original problem, but we know

$$\tilde{G}F\hat{x} = \tilde{G}Fx = \tilde{G}y.$$

$\tilde{G}$ can be written as $\tilde{G} = \tilde{\Gamma}\Delta$, where Δ is a projection, $\Delta : C^0[0,1] \to \mathbb{R}$ (viz. the restriction to the function value at the point t=1) and the problem solved reads

$$\Delta F \xi = \Delta y ,$$

which has a solution that belongs to $\tilde{G}\tilde{Y}$.

3.7. Defect Correction Processes with an approximate inverse of deficient rank

In this section we consider the linear defect correction process, where both F and $\tilde{G}$ are linear operators $\mathbb{R}^n \to \mathbb{R}^n$; F is bijective (rank(F) = n) and $\tilde{G}$ is of deficient rank (rank(G) = m < n). This is a special case of a DCP with $\tilde{G}$ neither surjective nor injective. We can decompose the n × n matrix $\tilde{G}$ into its singular value decomposition (cf. LAWSON & HANSON [1974])

$$\tilde{G} = U \Sigma V^T, \tag{3.7.1}$$

where U, Σ and V are n × n matrices, U and V are orthonormal and Σ is a non-negative diagonal matrix. Except for the ordering of the elements of Σ (and the corresponding ordering of the columns of U and V), this decomposition is uniquely determined. The diagonal elements of Σ are the singular values and normally they are ordered such that

$$\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \ldots \geq \sigma_n \geq 0.$$

Because rank($\tilde{G}$) = m, we know that $\sigma_1, \sigma_2, \ldots, \sigma_m$ are non-zero and $\sigma_j = 0$, $j = m+1, \ldots, n$.

More generally, for the m-rank matrix $\tilde{G}$ we can write

$$\tilde{G} = P S R, \tag{3.7.2}$$

where $R : \mathbb{R}^n \to \mathbb{R}^m$, $S : \mathbb{R}^m \to \mathbb{R}^m$, $P : \mathbb{R}^m \to \mathbb{R}^n$ and rank(P) = rank(S) = rank(R) = m. Here we can take e.g.:

$P = U_1$: the orthonomral set of the first m columns of U;
$S = \Sigma_1$: a diagonal matrix with elements $\sigma_1, \sigma_2, \ldots, \sigma_m$;
$R = V_1^T$: the orthonormal set of the first m rows of V^T

or we can take arbitrary m-rank matrices P and R, with Range(P) = Range($\tilde{G}$) = Span(U_1) and Kernel(R) = Kernel($\tilde{G}$) = Span(V_2), in which case S is a non-singular full m × m matrix with $S^{-1} = R\, V_1 \Sigma_1^{-1} U_1^T P$.

In order to see the relation with section 3.4 we remark that, in the finite-dimensional linear case considered here, we can construct a decomposition (3.4.1) by taking

$$\tilde{\Gamma} = U\, \tilde{\Sigma}\, V^T, \quad \Delta = V_1 V_1^T,$$

where $\tilde{\Sigma}$ is a diagonal matrix with the first m diagonal elements $\sigma_1, \sigma_2, \ldots, \sigma_m$; for the last n-m elements arbitrary non-zero values can be taken. For these $\tilde{\Gamma}$ and Δ we know that $\tilde{\Gamma}$ is a full rank matrix and Δ is a projector of rank m.

In the decomposition (3.7.2) P is called the prolongation and R is the restriction. Because P and R are full rank matrices: P has a left-inverse $\hat{R} = (U_1^T P)^{-1} U_1^T$ and R has a right-inverse $\hat{P} = V_1 (RV_1)^{-1}$. Moreover, we know that

$$P\hat{R} = \hat{P}R = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix},$$

is a projection operator of rank m.

Now we can consider what happens to the error to the solution or to the residual after one iteration step of the DCP.

I. In order to study this effect on the error of the solution, we consider the defect correction process in the form DCPA . Here the amplification operator is

$$M = I - \tilde{G}F = I - PSRF. \tag{3.7.3}$$

We decompose the error e into two parts: $e_s + e_u$ with $e_s \in$ Range(P) and $e_u \in$ Range(P)$^\perp$ = Kernel($\hat{R}$) = Span(U_2). Analogously we write $Me = (Me)_s + (Me)_u$.

Thus, we have

$$e_s = P\hat{R}\, e_s$$

and

$$e_u = (I - P\hat{R})e_u.$$

Now a simple computation shows

$$(3.7.4) \qquad M\, e_s = M\, P\hat{R}\, e_s = (P\hat{R} - PSRFP\hat{R})\, e_s = P(I - SRFP)\hat{R}\, e_s.$$

We see that the result is again in Range(P). Moreover, we notice that in the special case that $S^{-1} = RFP$ we have $M\, e_s = 0$. More generally, with $S^{-1} = RFP + E$, we have

$$M\, e_s = PSE\hat{R}\, e_s = \tilde{G}\hat{P}E\hat{R}\, e_s.$$

In practice, where $\tilde{G} = PSR$ should be a reasonable approximation to F^{-1}, it is often possible to choose S^{-1} equal to or close to RFP. The contribution from e_u to Me is given by

$$M\, e_u = e_u - \tilde{G}F\, e_u.$$

We see that the second term is again in Range(P), whereas the first term lies in $\text{Range}(P)^{\perp} = \text{Kernel}(\hat{R})$. We conclude that

$$(3.7.5) \qquad \begin{cases} (Me)_s = \tilde{G}\hat{P}E\hat{R}\, e_s - \tilde{G}F\, e_u, \\ (Me)_u = \qquad\qquad\qquad e_u. \end{cases}$$

REMARK. In the context of multi-grid methods (cf. Section 5.), the components in Range(P) are called the *smooth components*, those in $\text{Kernel}(\hat{R})$ the *unsmooth components of the error*.

II. For the residual, the amplification operator is

$$(3.7.6) \qquad \bar{M} = I - F\tilde{G} = I - FPSR.$$

Now we decompose the residual r into two parts $r = r_s + r_u$ with $r_s \in \text{Range}(\hat{P}) = \text{Span}(V_1)$ and $r_u \in \text{Kernel}(R) = \text{Range}(\hat{P})^{\perp} = \text{Span}(V_2)$. Analogously we write

$\bar{M}r = (\bar{M}r)_s + (\bar{M}r)_u$. Again, a simple computation shows

$$(3.7.7) \qquad \begin{cases} (\bar{M}r)_s = \hat{P}E\tilde{R}\tilde{G}\, r_s, \\ (\bar{M}r)_u = -(I - \hat{P}R)\, F\tilde{G}\, r_s + r_u. \end{cases}$$

REMARK. In the context of multi-grid methods, the components in Range($\hat{P}$) are called the *smooth components*, those in Kernel(R) are called the *unsmooth components of the residual*.

REMARK. In the special case that $R = P^T$, we see that

$$\text{Range}(P) = \text{Range}(\hat{P}) = \text{Span}(U_1) = \text{Span}(V_1),$$
$$\text{Kernel}(R) = \text{Kernel}(\hat{R}) = \text{Span}(U_2) = \text{Span}(V_2).$$

In this case the subspace of the smooth (resp. unsmooth) components of the residual is the same as the subspace of the smooth (resp. unsmooth) components of the error.

SUMMARY.

1. The error *in the solution*

$$\begin{array}{lcl} \text{Smooth components} = \text{Range}(P) & \xrightarrow{\tilde{G}\,\hat{P}E\hat{R}} & \text{Range}(P) = \text{Range}(\tilde{G}), \\ \text{Unsmooth components} = \text{Kernel}(\hat{R}) & \xrightarrow[I]{\tilde{G}F} & \text{Kernel}(\hat{R}) = \text{Range}(\tilde{G})^{\perp}. \end{array}$$

2. The error *in the residual*

$$\begin{array}{lcl} \text{Smooth components} = \text{Range}(\hat{P}) & \xrightarrow{\hat{P}E\hat{R}\,\tilde{G}} & \text{Range}(\hat{P}) = \text{Kernel}(\tilde{G})^{\perp}, \\ \text{Unsmooth components} = \text{Kernel}(R) & \xrightarrow[I]{(\hat{P}R-I)F\tilde{G}} & \text{Kernel}(R) = \text{Kernel}(\tilde{G}). \end{array}$$

3. In the case $R = P^T$ we have

$$\text{Range}(P) = \text{Range}(\hat{P}),$$
$$\text{Kernel}(R) = \text{Kernel}(\hat{R}).$$

INTERMEZZO

Before we shall treat extensions of the Defect Correction Principle and introduce multigrid algorithms, we first give a very simple example of a two-grid algorithm. This is a preversion of a multigrid algorithm. This example, which we borrow from HACKBUSCH [1976, 1981], shows a simple two-point boundary value problem and a simple iterative solution method for which the behaviour of the iterative process can be analyzed exactly. In this example the main features of a multigrid algorithm are already visible.

We consider the two-point boundary-value problem

$$(1)\qquad \begin{aligned} -u''(x) &= f(x), && x \in \Omega = (0,1),\\ u(0) &= u(1) = 0. \end{aligned}$$

Both with the Finite Difference Method and with the Finite Element Method with piecewise linear test-functions, we find on a uniform mesh $\Omega_h = \{x_i \mid x_i = i/N;\ i = 0,\ldots,N\}$ the discretized problem

$$(2)\qquad L_h u_h = f,$$

with the discrete operator

$$(3)\qquad L_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & \emptyset \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ \emptyset & & -1 & 2 \end{pmatrix}, \quad h = N^{-1},$$

a square $(N-1) \times (N-1)$ matrix, and

$$f_h = (f(x_1), f(x_2), \ldots, f(x_{N-1}))^T$$
$$u_h \quad (u_1, u_2, \ldots, u_{N-1})^T,$$

$(N-1)$-vectors.

First we consider the damped Jacobi-method for the iterative solution of (2). One iteration step in this process reads

$$(4)\qquad u_h^{(i+1)} = u_h^{(i)} - \omega\, D_h^{-1}\,(L_h\, u_h^{(i)} - f_h)$$

where D_h is the main diagonal of A_h and $\omega \in \mathbb{R}$ is the damping parameter. By $\hat{u}_h$ we denote the solution of (2); then the error in the i-th iterand is

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

Further we have

$$e_h^{(i+1)} = M_h^{REL} \quad e_h^{(i)}, \tag{5}$$

where M_h^{REL} is the amplification-operator of the error for this relaxation method:

$$M_h^{REL} = I - \omega D_h^{-1} L_h = \begin{pmatrix} 1 & & \emptyset \\ & \ddots & \\ \emptyset & & 1 \end{pmatrix} - \frac{\omega}{2} \begin{pmatrix} 2 & -1 & & \emptyset \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ \emptyset & & -1 & 2 \end{pmatrix} . \tag{6}$$

The eigensysteem of $\begin{pmatrix} 2 & -1 & & \emptyset \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ \emptyset & & -1 & 2 \end{pmatrix}$ is $\{\lambda_m, \phi_m\}_{m=1,2,\ldots,N-1}$ with eigenvalues

$$\lambda_m = 2 - 2\cos(m\pi h) = 4\sin^2(m\pi h/2) \tag{7}$$

and eigenfunctions

$$\phi_m = \begin{pmatrix} \sin(m\pi h) \\ \sin(2m\pi h) \\ \vdots \\ \sin((N-1)m\pi h) \end{pmatrix} . \tag{8}$$

The eigenvalues of M_h^{REL} are

$$\lambda(M_h^{REL})_m = 1 - \frac{\omega}{2} \cdot 4\sin^2(m\pi h/2) \tag{9}$$

$$= \cos^2(m\pi h/2) + (1-2\omega)\sin^2(m\pi h/2).$$

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

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

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

$$\cos^2(\pi h/2) \geq (M_h^{REL})_m \geq \cos^2((N-1)\pi h/2) = \sin^2(\pi h/2)$$

or

$$1 - \frac{\pi^2 h^2}{4} + O(h^4) \geq \lambda(M_h^{REL})_m \geq \frac{\pi^2 h^2}{4} + O(h^4).$$

We see that slowly varying eigenfunctions (small m) are damped slowly by M_h^{REL}, whereas rapidly varying eigenfunctions (large m) are damped efficiently by M_h^{REL}. After a few iterations with the damped Jacobi relaxation ($\omega = 1/2$), the rapidly varying component in the error will almost vanish, however the slowly varying components will hardly be affected: the error has not become much smaller, but it became much smoother. Application of one step in the (damped Jacobi) relaxation process is therefore also called a *smoothing step*.

Let $\tilde{u}_h$ be an approximation of $\hat{u}_h$ with a smooth error

$$\tilde{e}_h = \tilde{u}_h - \hat{u}_h.$$

Then $\tilde{e}_h$ satisfies the equation

$$L_h\tilde{e}_h = L_h\tilde{u}_h - L_h\hat{u}_h = L_h\tilde{u}_h - f_h = d_h = -r_h, \tag{10}$$

r_h : is the residual of $\hat{u}_h$;
d_h : is the defect of $\hat{u}_h$.
Because $\hat{e}_h$ is a smooth function, we are able to represent it well on a coarser grid.

For this we have to solve e.g.

$$A_H\tilde{e}_H = -r_H \qquad , H = 2h, \tag{11}$$

where (11) is a discretization of (10). For this we need

$$\begin{array}{ccc} X_h & \xrightarrow{L_h} & Y_h \\ {\scriptstyle P_{hH}}\uparrow & & \downarrow{\scriptstyle \bar{R}_{Hh}} \\ X_H & \xrightarrow{L_H} & Y_H \end{array}$$

(i) a restriction $\bar{R}_{Hh}$,
(ii) a coarse-grid operator L_H,
(iii) a prolongation P_{hH}.

For A_H we take an operator similar to (3), only with an mesh H = 2h instead of h. Thus, A_H is a $(\frac{N}{2}-1) \times (\frac{N}{2}-1)$ matrix. (We assume that N is an even integer.)

A simple prolongation, P_{hH}, is found by linear interpolation. This operator P_{hH} is defined by $u_h = P_{hH}u_H$

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

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

$$f_H(x_i) = f_h(x_i),$$

for all $x_i \ \Omega_H$.

Another possible restriction $\bar{R}_{Hh}^*$ could be a weighed restriction, defined by $f_H = \bar{R}_{Hh}^* f_h$, with $f_h \in Y_h$ and $f_H \in Y_H$ such that

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

for all $x_i \in \Omega_H$.

The operators A_H, P_{hH}, $\bar{R}_{Hh}$ and $\bar{R}_{Hh}^*$ can be described explicitly by their matrices

(12a) $$A_H = \frac{1}{H^2}\begin{pmatrix} 2 & -1 & & \emptyset \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ \emptyset & & -1 & 2 \end{pmatrix}$$, an $(\frac{N}{2}-1) \times (\frac{N}{2}-1)$ matrix,

(12b)
$$P_{hH} = \begin{pmatrix} 1/2 & & & & & \\ 1/2 & 1/2 & & & & \\ & 1/2 & 1/2 & & & \\ & & 1/2 & 1/2 & & \\ & & & \ddots & \ddots & \\ & & & & 1/2 & \\ & & & & 1/2 & \end{pmatrix}, \text{ an } (N-1) \times (\tfrac{N}{2}-1) \text{ matrix,}$$

(12c)
$$\overline{R}_{Hh} = \begin{pmatrix} 0 & 1 & 0 & & & \\ & 0 & 1 & 0 & & \\ & & 0 & 1 & 0 & \\ & & & \ddots & \ddots & \ddots \\ & & & & 0 & 1 & 0 \end{pmatrix}, \text{ an } (\tfrac{N}{2}-1) \times (N-1) \text{ matrix}$$

and

(12d)
$$\overline{R}_{Hh}^{*} = \begin{pmatrix} 1/2 & 1 & 1/2 & & & & \\ & 1/2 & 1 & 1/2 & & & \\ & & 1/2 & 1 & 1/2 & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & 1/1 & 1 & 1/2 \end{pmatrix}$$

, an $(\frac{N}{2}-1) \times (N-1)$ matrix.

The vectors u_h and u_H are of the form

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

and

$$u_H = (u_H(H), u_H(2H), \ldots, u_H((\tfrac{N}{2}-1)H))^T.$$

The coarse discretization (11) is used to improve the error in the approximate solution $\tilde{u}_h$ as follows:

1) first we compute $\tilde{e}_h$ from

$$L_H \tilde{e}_H = d_H = R_{Hh} d_h$$

2) then we subtract the error from the approximate solution

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

This correction of the approximate solution is called a *coarse grid correction step* and it can be written as

$$\hat{\tilde{u}} := \tilde{u}_h - P_{hH} L_H^{-1} \overline{R}_{Hh} (L_h \tilde{u}_h - f_h). \tag{13}$$

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

$$M_h^{CGC} = I - P_{hH} L_H^{-1} \overline{R}_{Hh} L_h. \tag{14}$$

The Two Level Algorithm (TLA), which is a preversion of the Multi Level Algorithm, is an iterative process for the solution of (2), in which each step consists of

(i) a number of p smoothing steps,

(ii) a coarse grid correction step,

(iii) a number of q smoothing steps.

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

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

and we consider the effect on a single eigenfunction. Hence for ϕ_m, as given in (8), we shall compute

$$M_h^{TLA} \phi_m = (M_h^{REL})^q (M_h^{CGC}) (M_h^{REL})^P \phi_m.$$

To this end we first compute

$$M_h^{CGC} \phi_m = \phi_m - P_{hH} L_H^{-1} \overline{R}_{Hh} L_h \phi_m \tag{15}$$

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

$$P_{hH} L_{Hh}^{-1} \overline{R}_{Hh} \begin{pmatrix} \sin(\pi mh) \\ \sin(2\pi mh) \\ \sin(3\pi mn \\ \vdots \\ \sin((N-1)\pi mh \end{pmatrix} = P_{hH} L_H^{-1} \begin{pmatrix} \sin(2\pi mh) \\ \sin(4\pi mh) \\ \sin(6\pi mh) \\ \vdots \\ \sin((N-2)\pi mh) \end{pmatrix} =$$

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

$$= \frac{H^2}{4\sin^2(m\pi H/2)} \cdot \frac{1}{2} \begin{pmatrix} \sin(0) + \sin(\pi mH) \\ 2.\sin(\pi mH) \\ \sin(\pi mH) + \sin(2\pi mH) \\ 2.\sin(2\pi mH) \\ \sin(2\pi mH) + \sin(3\pi mH) \\ \vdots \\ 2.\sin((N/2-1)\pi mH) \\ \sin((N/2-1)\pi mH) + \sin(N/2.\pi mh) \end{pmatrix}$$

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

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

$$= \frac{H^2(a\phi_m + b\phi_{N-m})}{4\sin^2(\pi mh)} ,$$

with $a + b = \cos(\pi mh)$ and $a - b = 1$. Hence, $a = \cos^2(\pi mh/2)$ and $b = -\sin^2(\pi mh/2)$. Briefly we write $C_m := \cos(\pi mh/2)$ and $S_m := \sin(\pi mh/2)$. Thus we get

$$P_{hH}\, L_h^{-1}\, \overline{R}_{Hh}\, \phi_m = \frac{h^2(C_m^2\, \phi_m - S_m^2\, \phi_{N-m})}{4 \cdot S_m^2\, C_m^2}$$

and

$$M_h^{CGC}\, \phi_m = \phi_m - \frac{1}{C_m^2}\,(C_m^2\, \phi_m - S_m^2\, \phi_{N-m}) = \frac{S_m^2}{C_m^2}\, \phi_{N-m}. \tag{16}$$

This results in

$$
\begin{aligned}
M_h^{TLA}\,\phi_m &= (M_h^{REL})^q\,(M_h^{CGC})\,(M_h^{REL})^p\,\phi_m \\
&= (M_h^{REL})^q\,(M_h^{CGC})\,\phi_m\,(C_m^2)^p \\
&= (M_h^{REL})^q\,\phi_{N-m}\,S_m^2\,(C_m^2)^{p-1} \\
&= (C_{N-m}^2)^q\,S_m^2\,(C_m^2)^{p-1}\,\phi_{N-m}.
\end{aligned}
\tag{17}
$$

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

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

and

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

We see that low frequencies are converted in high frequencies, v.v. .

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

$$P_{hH}\,L_H^{-1}\,\overline{\overline{R_{Hh}}}^{*}\,\phi_m = P_{hH}\,L_H^{-1}\,\frac{1}{4}\begin{pmatrix} \sin(m\pi h) + 2\sin(2m\pi h) + \sin(3m\pi h) \\ \sin(3m\pi h) + 2\sin(4m\pi h) + \sin)5m\pi h) \\ \cdots\cdots\cdots \\ \sin((k-1)m\pi h) + 2\sin(2km\pi h) + \sin((2k+1)m\pi h) \\ \cdots\cdots\cdots \\ \sin((N-3)m\pi h) + 2\sin((N-2)m\pi h) + \sin((N-1)m\pi h) \end{pmatrix}$$

$$= P_{hH}\,L_H^{-1}\,\frac{1}{4}\begin{pmatrix} \vdots \\ 4\cos^2(m\pi h/2)\,\sin(2km\pi h) \\ \vdots \end{pmatrix}$$

$$= C_m^2\,P_{hH}\,L_H^{-1}\begin{pmatrix} \sin(m\pi H) \\ \sin(2m\pi H) \\ \vdots \\ \sin((\frac{N}{2}-1)m\pi H \end{pmatrix}$$

$$= \frac{C_m^2 H^2 (C_m^2\,\phi_m - S_m^2\,\phi_{N-m})}{4\,\sin^2(\pi m h)} = \frac{h^2 (C_m^2\,\phi_m - S_m^2\,\phi_{N-m})}{4\,S_m^2}.$$

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

and

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

From the last two equalities we conclude

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

$$= \alpha'\phi_m + \beta'\phi_{N-m}, \qquad m = 1,2,\ldots,N/2.$$

We denote this in matrix notation by

$$\begin{pmatrix} \alpha' \\ \beta' \end{pmatrix} = \begin{pmatrix} S_m^2 & S_m^2 \\ S_m^2 & C_m^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = {}_mM_h^{CGC*} \begin{pmatrix} \alpha \\ \beta \end{pmatrix};$$

α denotes the contribution from the low-frequency component ϕ_m and β the contribution from the high-frequency component ϕ_{N-m}.

Similarly we find for the amplification operator of the residual

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

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

$$(19) \qquad {}_m\bar{M}_h^{TLA} = \begin{pmatrix} C_m^2 & 0 \\ 0 & S_m^2 \end{pmatrix}^q \begin{pmatrix} S_m^2 & S_m^2 \\ C_m^2 & C_m^2 \end{pmatrix} \begin{pmatrix} C_m^2 & 0 \\ 0 & S_m^2 \end{pmatrix}^p$$

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

The eigenvalues of this matrix are 0 and

$$S_m^2 C_m^{2p+2q} + C_m^2 S_m^{2p+2q}.$$

Therefore, the spectral radius of $\bar{M}_h^{TLA}$ is

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

with $C_m^2 = \cos^2(\pi mh/2)$ and $S_m^2 = 1-C_m^2$, $m = 1,2,\ldots,N/2$.

We find

$$\begin{aligned} \rho(M_h^{TLA}) &= 1 \quad \text{if } p+q = 0 \\ &= \frac{1}{2} \quad \text{if } p+q = 1 \\ &= \frac{1}{4} \quad \text{if } p+q = 2 \end{aligned} \tag{20a}$$

To compute $\rho(\bar{M}_h^{TLA})$ for large p+q we first see

$$S_m^2 C_m^{2n} \le \max_{0\le t\le 1} (1-t^2)t^{2n} = \left(\frac{1}{1+n}\right)\left(\frac{n}{1+n}\right)^n = \frac{1}{4}\left(1-\frac{1}{1+n}\right)^{n+1} \approx \frac{1}{n}e^{-1}$$

for $n \to \infty$.

This shows that

$$\rho(\bar{M}_h^{TLA}) \approx \frac{1}{(p+q)e} \tag{20b}$$

for $(p+q) \to \infty$,

which describes the convergence rate of the two-level algorithm for large p+q.

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

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

$$= \max_m \sqrt{S_m^4 C_m^{4p+4q} + C_m^{4q} S_m^{4p+q} + S_m^{4q} C_m^{2p+4} + C_m^4 S_m^{4p+4q}}$$

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

Hence,

$$\limsup_{h\to\infty} \|\bar{M}_h^{TLA}\| \sim \max_{0\le t\le \frac{1}{2}} \sqrt{t^2(1-t)^{2p} + (1-t)^2\, t^{2p}}\ \sqrt{(1-t)^{2q} + t^{2q}}\ .$$

If $p = 0$ and $q = 0$ we find

$$\|\bar{M}_h^{TLA}\| = \max_m \sqrt{2(S_m^4 + C_m^4)} \xrightarrow{h\to 0} \sqrt{2}\ .$$

If $q = 0$, $p \neq 0$

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

If $p = 0$, $q \neq 0$

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

$$q = 1 \quad \Rightarrow \quad \frac{1}{2}\ ,$$

$$q \to \infty \quad \Rightarrow \quad \approx \sqrt{2} \max_{0\le t\le 1} \sqrt{t^2(1-t)^{2q}} \approx \frac{\sqrt{2}}{q\,e}\ .$$

If $p,q > 0$

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

$$\lim_{h\to 0} \|\bar{M}_h^{TLA}\| = \max_{0\le t\le \frac{1}{2}} t(1-t)^{p+q} = \frac{1}{p+q} \left(\frac{p+q}{p+q+1}\right)^{p+q+1} \xrightarrow{p+q\to\infty} \frac{1}{e(p+q)}\ .$$

Summarizing we see

$$\begin{aligned} \rho(M_h^{TLA}) = \rho(\bar{M}_h^{TLA}) &\to 1 \quad \text{if} \quad p+q = 0, \qquad \text{for } h \to 0, \\ &\to \frac{1}{2} \quad \text{if} \quad p+q = 1, \\ &\to \frac{1}{4} \quad \text{if} \quad p+q = 2, \\ &\to \frac{1}{e(p+q)} \quad \text{for } (p+q) \to \infty. \end{aligned}$$

$$\|\bar{M}_h^{TLA}\| \quad \begin{cases} \to \sqrt{2} & \text{if } p = 0,\ q = 0 \quad \text{for } h \to 0 \\ \to 1 & \text{if } p \neq 0,\ q = 0 \\ \to \frac{1}{2} & \text{if } p = 0,\ q = 1 \\ \to \frac{\sqrt{2}}{qe} & \text{if } p = 0,\ q \to \infty \\ \to \frac{1}{e(p+q)} & \text{if } p \neq 0 \quad (p+q) \to \infty \end{cases}$$

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

We conclude that relaxations *after* coarse grid corrections are of use for a small norm of the amplification operator of the *residual*, whereas relaxations *before* yield a small norm of the operator of the *error*.

4. EXTENSION OF THE DCP PRINCIPLE

Since a defect correction process is an iterative technique to solve "hard" problems by means of "simpler" ones, we can apply this principle iteratively or recursively again. The "simple" problem $Fx = y$ may be approximated again by an even simpler one, etc. . On the other hand, if we are able to solve a problem, we can try to solve nearby harder problems. In this way we can try e.g. to solve a high-order discretization of a problem by means of a low-order discretization of it. Or we may solve a discretization on a fine grid with the aid of the discretization on a coarser one. Also, starting with a coarse discretization of a continuous problem, we can try to find more and more accurate approximations on finer and finer grids.

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 all over in a defect correction process. Then we describe the iterative and the recursive application of the DCP and in the last subsection we describe how more discretizations of a problem can be applied alternately in order to get a stable and accurate approximation.

4.1. Non-stationary defect correction processes

In order to find a solution to the problem (P) it is not necessary to use one fixed approximate inverse in an iteration process as described in the the preceding section. As we anticipated in the example with Newton's method, it is possible to use different approximate inverses in each iteration step. Then the iteration steps of DCPA and DCPB read respectively

$$(4.1.1) \qquad x_{i+1} = x_i - \tilde{G}_{i+1}Fx_i + \tilde{G}_{i+1}y,$$

and

$$(4.1.2) \qquad \ell_{i+1} = \ell_i - F\tilde{G}_i \ell_i + y.$$

A similar modification of DCPC can be given.

In this way we are able to adapt the approximate inverse during the iteration and we can try to find sequences $\{G_i\}$ in order to accellerate the convergence of the iteration.

REMARK. We see that for general affine operators $\{\tilde{G}_i\}$ we have no longer the equivalence DCPA and DCPB. Instead we see DCPA to be equivalent with the iteration.

$$(4.1.3) \qquad \ell_{i+1} = \tilde{F}_{i+1}\tilde{G}_i\ell_i - F\,\tilde{G}_i\ell_i + y,$$

or DCPB to be equivalent with

$$(4.1.4) \qquad \tilde{F}_{i+1}x_{i+1} = \tilde{F}_i x_i - F\,x_i + y$$

or

$$(4.1.5) \qquad x_{i+1} = \tilde{G}_{i+1}\tilde{F}_i x_i - \tilde{G}_{i+1}F\,x_i + \tilde{G}i+1\ y.$$

Various methods are known to find a proper sequence $\{\tilde{G}_i\}$. Here we mention a few.

EXAMPLE 1. $\tilde{G}_{i+1} = \tilde{G}(x_i)$.
The approximate inverse depends on the last iterand computed. This is the case e.g. in Newton's method for the solution of non-linear equations, where $\tilde{G}(x) = F'(x))^{-1}$, with $F'(x)$ the Fréchet derivative of the operator F in the problem (P).

EXAMPLE 2. $\tilde{G}_i = \tilde{G}(\omega_i)$.
The approximate inverse depends on a single real parameter. 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.

EXAMPLE 3. $\tilde{G}_i \in \{\tilde{G}_1, \tilde{G}_2\}$.

In each iteration step the approximate inverse is chosen from a set of two (or 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 ATKINSON [1976] and BRAKHAGE [1960].) It is also the case in the two level algorithm in the intermezzo.

REMARK. From the pratical point of view (4.1.2) seems to be the more attractive of the two processes (4.1.1) and (4.1.2) because in (4.1.2) $\tilde{G}_i$ appears only once in an iteration step. This implies that only one approximate problem has to be solved, whereas $\tilde{G}_{i+1}$ appears twice in (4.1.1).

4.2. A fixed combination of approximate inverses

In this section we assume that the operator F in (P) and the approximate inverses $\tilde{G}$ and $\approx{G}$ are linear operators. We consider two iteration steps in the non-stationary DCPA in which, in turn, one or the other of two approximate inverses is used. Then the iteration steps

$$x_{i+\frac{1}{2}} = (I - \tilde{G}F)x_i + \tilde{G}y$$

and

$$x_{i+1} = (I - \approx{G}F)x_{i+\frac{1}{2}} + \approx{G}y$$

combine into a single iteration step of the form

$$x_{i+1} = (I - \approx{G}F)(I - \tilde{G}F)x_i + (\approx{G} - \approx{G}F\tilde{G} + \tilde{G})y.$$

This is easily recognized as a new iteration step of the type DCPA, now with the approximate inverse

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

We conclude that a fixed combination of DCPA-steps can be considered as a new DCPA-step with a more complex approximate inverse.

The amplification operator of the new DCPA process is the product of the amplification operators of the elementary processes.

σ applications of the same approximate inverse

We can describe the DCPA in matrix notation by

$$\begin{pmatrix} x_{i+1} \\ y \end{pmatrix} = \begin{pmatrix} I-\tilde{G}F & \tilde{G} \\ \emptyset & I \end{pmatrix} \begin{pmatrix} x_i \\ y \end{pmatrix}.$$

σ times an application of the same iteration step yields

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

Thus, we see that one iteration step which consists of σ applications of DCPA-steps results in a DCPA with the amplification operator

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

and the approximate inverse

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

Since the operators F and $\tilde{G}$ are linear, we may look at the combined process as a DCPB as well; its approximate inverse being the same as for the DCPA, of course, and with the amplification operator

$$\bar{M} = FMF^{-1} = (I - F\tilde{G})^{\sigma}.$$

4.3. Iterative application of DCP

It is possible not only to change the approximate inverse $\tilde{G}$ during the iteration process, often it makes sense also to substitute different operators F_k for F during iteration. In general, the operators $\{F_k\}_{k=1,2,\ldots}$ will be simple to evaluate in the beginning of the iteration and they will converge in some sense to the "target" operator F, the operator of the original problem, as the iteration proceeeds.

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

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

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

iteration of (P_k). This way of looking at the changing F_k yields a criterion for the number of iterations that has to be spent to approximate the solution of (P_k); viz. the iterand $x_{k,i}$ in the DCP for the solution of (P_k) should not approximate x_k^*, the solution of (P_k), better than the solution of (P_k) is itself an approximation to the solution of (P_{k+1}); i.e. we should not iterate the DCP for (P_k) further than until

$$\| x_{k,i} - x_k^* \| \approx \| x_k^* - x_{k+1}^* \| .$$

EXAMPLES 1a and 1b. One example of the iterative application of a DCP is the IUDeC (Iteratively Updated Defect Correction) process described by STETTER [1978]. Here $\{F_k\}$ are discrete approximations of higher and higher order to an analytic operator F. The approximate inverse $\tilde{G} = F_0^{-1}$ is kept constant during the process.

Another example is the Full Multigrid Method (BRANDT [1979]), in which $\{F_k\}$ are discretizations on finer and finer nets of an analytic operator F.

One way to create a sequence of problems (P_k) is Galerkin approximations of a "target" problem (P):

$$(P_k) \qquad \bar{R}_k \, F \, P_k \, x_k = \bar{R}_k \, y.$$

Then the different discretizations are determined by $\{\bar{R}_k, P_k\}$.

EXAMPLE 2. *Global interpolation.*
Here $\bar{R}_k = \bar{R}_h$ is independent of k,

$$\bar{R}_h : C(\Omega) \to \ell_h(\Omega_h)$$

is the restriction of a continuous function to its values on a set of nodal points Ω_h. The prolongation P_k is global piecewise polynomial

$$P_k : \ell_h(\Omega_h) \to C(\Omega)$$

of order k: the set of nodal values is interpolated to a continuous piecewise polynomial function defined on Ω. (Finite element interpolation.)

EXAMPLE 3. *Local interpolation.*
We take $\bar{R}_k = \bar{R}_h$ as in example 2. Now P_k is local interpolation in the neighbourhood of nodal points. I.e. $P_k \, u_h$ is a function which is (only) defined

in (small) neighbourhoods of nodal points from Ω_h. The value of $P_k u_h$ and its derivatives at $\xi_h \in \Omega_h$ are determined from the values of u_h at $\xi_h \in \Omega_h$ and a number of neighbouring nodal points by taking (divided) difference quotients around ξ_h. In this case $P_k u_h$ is not necessarily a function defined on the whole of Ω [and the operator F is only applied on (open) neighbourhoods of points in Ω_h].

The accuracy of successive approximations in a DCP iteration with different discretizations of the same problem

Let us consider (different) discretizations of the problem Fx = y, viz.

$$F_h^i x_h = y_h, \text{ with } F_h^i : X_h \to Y_h \text{ for all } i = 0,1,2,\ldots,$$

and let X, X_h, Y and Y_h be related by

$$R_h : X \to X_h \text{ and } \bar{R}_h : Y \to Y_h.$$

Let the order of consistency of the discretizations be p_i, and let the first discretization be stable. We will study the iterative application of DCPA, with the equations $F_h^i i_h = y_h = \bar{R}_h y$ to solve in the i-th iteration step and with the same approximate inverse $\tilde{G}_h = (F_h^0)^{-1}$ in all iteration steps. Then the DCPA reads

$$\begin{cases} u_0 = \tilde{G}_h y_h = \tilde{G}_h \bar{R}_h y, \\ u_{i+1} = (I_h - \tilde{G}_h F_h^i) u_i + \tilde{G}_h y_h. \end{cases}$$

We are going to estimate the relative error of approximation for a finite number of number of iteration steps:

$$k_i = \| u_i - R_h x \| / \| x \|.$$

THEOREM. *For the relative error of approximation in the i-th iteration step of the iterative* DCPA *process*:

$$k_i = \| u_i - R_h x \| / \| x \|,$$

we have

$$k_0 \le \|\tilde{G}_h\| \, \|\bar{R}_h F - F_h^0 R_h\| = O(h^{p_0})$$

$$k_i \le \|\tilde{G}_h\| \, \|\bar{R}_h F - F_h^{i-1} R_h\| + \|\tilde{G}_h\| \, \|F_h^0 - F_h^{i-1}\| \, k_{i-1}$$

$$= O(h^{\min_{0\le j\le i}(p_j+(i-j)p_0)}), \quad i = 1,2,\ldots .$$

PROOF.

$$u_0 - R_h x = \tilde{G}_h \bar{R}_h y - R_h x = \tilde{G}_h(\bar{R}_h F - F_h^0 R_h) x.$$

The given estimate now follows from the stability of F_h^0 (i.e. $\tilde{G}_h$ is uniformly bounded) and the consistency of F_h^0.

$$u_{i+1} - R_h x = u_i - R_h x - \tilde{G}_h F_h^i u_i + \tilde{G}_h y_h$$

$$= u_i - R_h x + \tilde{G}_h(\bar{R}_h F x - F_h^i R_h x + F_h^i R_h x - F_h^i u_i)$$

$$= (I_h - \tilde{G}_h F_h^i)(u_i - R_h x) + \tilde{G}_h(\bar{R}_h F - F_h^i R_h) x.$$

Hence, for $i = 0,1,2,\ldots,$

$$k_{i+1} \le \|I_h - G_h F_h^i\| \cdot k_i + \|\tilde{G}_h\| \, \|\bar{R}_h F - F_h^i R_h\| .$$

Here again, the estimate follows from the stability of F_h^0 and the consistency of F_h^i. □

COROLLARY. *If*

$$\begin{cases} p_i \ge (i+1)p_0 & (i < n) \\ p_i = p_n & (i \ge n) \end{cases}$$

then

$$k_i = O(h^{\min(p_n,(i+1)p_0)}).$$

4.4. Recursive application of DCP

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 activate new (simpler to solve) DCP. In this way we can construct a recursive construction of DCPs in which only on the lowest level of recursion a very simple equation is to be solved.

Independently, this is probably not a real meaningful constuction, but in combination with non-stationary processes, where also other (non-recursive) approximate inverses are available, it describes the essentials of the multigrid algorithm.

Such a combination of a non-stationary process with some recursive approximate inverses can be described by the following sequence of DCPs.

$$
\begin{array}{lll}
DCP_1: & x := x - \tilde{G}_1 \ (F_1 x - f_1) & \tilde{G}_j \qquad\qquad\qquad j = 1,2,\ldots,n, \\
DCP_2: & x := x - \tilde{G}_{2,i}(F_2 x - f_2) & \\
\vdots & \vdots & \tilde{G}_{j,i} \in \{\tilde{G}_j,\ F_{j-1}^{-1}\}, \\
\vdots & \vdots & \qquad\qquad\qquad j = 2,3,\ldots,n. \\
DCP_n: & x := x - \tilde{G}_{n,i}(F_n x - f_n) &
\end{array}
$$

A full use of the sequence of DCPs is made by combining also the iterative application: first DCP_1 is solved and its solution is used as a starting value for DCP_2 etc.. In a multigrid context

$$DCP_1,\ DCP_2,\ldots,\ DCP_n,$$

are processes to solve operator equations, discretized on finer and finer grids. The complete iterative process is called: Full Multigrid Algorithm (BRANDT [1979]).

4.5. Mixed Defect Correction Processes

Up to now we have considered DCPs where each time one final target problem

(4.5.1)(P) $Fx = y, \qquad F : X \to Y$

was solved. In this section we treat the possibility of two (or more) different target problems:

(4.5.2)(P1) $F_1 x_1 = y_1, \qquad F_1 : X_1 \to Y_2,$

(P2) $F_2 x_2 = y_2, \qquad F_2 : X_1 \to Y_2,$

to be used in *one* iteration process. Behind the screen both procedures (P1) and (P2) probably are two approximations of an original problem (P), but the operator F is not used in the algorithmic procedure.

We introduce first the approximate inverses $\tilde{G}_1$ and $\tilde{G}_2$ of the operators F_1 and F_2 respectively. We assume that F_1, F_2, $\tilde{G}_1$ and $\tilde{G}_2$ are linear. Then we introduce the *Mixed Defect Correction Process*

(MDCP)
$$\begin{cases} u_{i+\frac{1}{2}} = u_i - \tilde{G}_1(F_1 u_i - y_1), \\ u_{i+1} = u_{i+\frac{1}{2}} - \tilde{G}_2(F_2 u_{i+\frac{1}{2}} - y_2). \end{cases}$$

Thus, the complete iteration step reads

(4.5.3) $$u_{i+1} = (I - \tilde{G}_2F_2)(I - \tilde{G}_1F_1)u_i + (I - \tilde{G}_2F_2)\tilde{G}_1y_1 + \tilde{G}_2y_2.$$

We find for MDCP the "amplification operator of the error"

(4.5.4) $$M = (I - \tilde{G}_2F_2)(I - \tilde{G}_1F_1).$$

A stationary point $\hat{u}$ of (MDCP) satisfies

(4.5.5) $$(I - M)\hat{u} = (I - \tilde{G}_2F_2)\tilde{G}_1y_1 + \tilde{G}_2y_2.$$

In the case that y_1 and y_2 can be written as $y_1 = \bar{R}_1 y$ and $y_2 = \bar{R}_2 y$, $\bar{R}_1 : Y \to Y_1$, $\bar{R}_2 : Y \to Y_2$, equation (4.5.5) is equivalent with

(4.5.6) $$(\tilde{G}_2F_1 + \tilde{G}_2F_2 - \tilde{G}_2F_2\tilde{G}_1F_1)u = -(\tilde{G}_1\bar{R}_1 + \tilde{G}_2\bar{R}_2 + \tilde{G}_1F_2\tilde{G}_1\bar{R}_1)y.$$

If equation (4.5.5) has a unique solution $\hat{u}$, this $\hat{u}$ is the stationary point of (MDCP) and with the error defined by

$$e_i = u_i - \hat{u},$$

the operator M has again the property

$$e_{i+1} = Me_i.$$

For an arbitrary w we know

$$(4.5.7) \qquad (I - M)w = (I - \tilde{G}_2F_2)\tilde{G}_1F_1w + \tilde{G}_2F_2w$$

and by (4.5.5) we find

$$(4.5.8) \qquad (I - M)(w - \hat{u}) = (I - \tilde{G}_2F_2)\tilde{G}_1(F_1w - y_1) + \tilde{G}_2(F_2w - y_2).$$

THEOREM

(i) *Let* (P_1) *and* (P_2) *be two discretizations of* (P) *with*

$$R : X \to X_1; \quad \bar{R}_1 : Y \to Y_1; \quad \bar{R}_2 : Y \to Y_2;$$

and such that $y_1 = \bar{R}_1 y$ *and* $y_2 = \bar{R}_2 y$;

(ii) *Let the local discretization error of the discretizations* (P_1) *and* (P_2) *of the problem* (P) *be respectively of order* p_1 *and* p_2;

(iii) *Let the approximate operators* $\tilde{F}_k = \tilde{G}_k^{-1}$, $F_k : X_1 \to Y_k$, $k = 1,2$, *be stable discretizations of* F *and let* $\tilde{F}_k$ *be consistent with* F_k, $k = 1,2$, *of order* $q_k > 0$;

Let $\tilde{u} \in X$ *be the solution of* (P) *and let* $\hat{u}$ *be a stationary point of* (MDCP), *then*

$$\|\hat{u} - R\tilde{u}\| \le C\, h^{\min(q_2+p_1,p_2)}.$$

PROOF. From (iii) it follows that, with k = 1,2,

$$\|\tilde{F}_k - F_k\| \le C\, h^{q_k}, \quad \|\tilde{G}_k\| \le C \text{ unif. in h.}$$

Hence, for k = 1,2 we have

$$\|I - \tilde{G}_k F_k\| \leq \|\tilde{G}_k\| \|\tilde{F}_k - F_k\| \leq C \cdot C\, h^{q_k} \xrightarrow{h \to 0} 0.$$

Thus,

$$\|M\| \leq \|I - \tilde{G}_1 F_1\| \|I - \tilde{G}_2 F_2\| \leq C < 1$$

for h small enough, and

$$\|I - M)^{-1}\| < C$$

for h small enough.
From (ii) it follows that the truncation errors of the discretization with respect to the solution $\tilde{u}$ are of order p_1 and p_2 respectively:

$$\tau_k = y_h - F_k R\tilde{u} = \bar{R}_h F\tilde{u} - F_k R\tilde{u} = (\bar{R}_k F - F_k R)\tilde{u} = \tau_k(\tilde{u})$$

$$\|\tau_k\| = \|\tau_k(\tilde{u})\| \leq C\, h^{p_k}.$$

From (4.5.8) we derive

$$(I - M)(R\tilde{u} - \hat{u}) = (I - \tilde{G}_2 F_2)\tilde{G}_1 (F_1 R\tilde{u} - y_1) - \tilde{G}_2 (\tilde{F}_2 R\tilde{u} - y_2)$$

$$= -(I - \tilde{G}_2 F_2)\tilde{G}_1 \tau_1 + \tilde{G}_2 \tau_2.$$

Hence

$$\|R\tilde{u} - \hat{u}\| \leq \|(I - M)^{-1}\| \|\tilde{G}_2\| \{\|\tilde{F}_2 - F_2\| \|\tilde{G}_1\| \|\tau_1\| + \|\tau_2\|\}$$

$$\leq c \cdot c \{ C\, h^{q_2} \cdot C \cdot h^{p_1} + h^{p_2} \}$$

$$< C\, h^{\min(p_1+q_2, p_2)}. \qquad \square$$

REMARK. The theorem can easily be generalized for more different target problems

$$(P_k) \qquad F_k x_k = y_k \qquad k = 1,2,\ldots,\ell.$$

With $\tilde{G}_k$ an approximate inverse of F_h, $\tilde{F}_k = \tilde{G}_k^{-1}$ and $M_k = (I - \tilde{G}_k F_k)$ we get for the multiple MDCP

$$\text{(MDCP)} \qquad \begin{cases} u_{i+k/\ell} = u_{i+(k-1)/\ell} - \tilde{G}_k(F_k u_{i+(k-1)/\ell} - G_k), \\ \qquad\qquad\qquad\qquad\qquad k = 1,2,\ldots,\ell. \end{cases}$$

The amplification operator of the error is

$$M = M_\ell M_{\ell-1} \cdots M_2 M_1.$$

We find

$$(I - M)(\hat{u} - R\tilde{u}) = \sum_{k=1}^{\ell} M_\ell M_{\ell-1} \cdots M_{k+1} \tilde{G}_k \tau_k$$

and hence

$$\begin{aligned} \|\hat{u} - R\tilde{u}\| &\le \|(I - M)^{-1}\| \sum_{k=1}^{\ell} \|\tilde{G}_\ell\| \|\tilde{F}_\ell - F_\ell\| \cdots \|\tilde{G}_{k+1}\| \|\tilde{F}_{k+1} - F_{k+1}\| \|\tilde{G}_k\| \|\tau_k\| \\ &\le C \sum_{k=1}^{\ell} c\, h^{q_\ell + q_{\ell-1} + \ldots + q_{k+1} + p_k} \\ &= C\, h^{p^*} \end{aligned}$$

with $p^* = \min_{k=1,\ldots,\ell} (p_k + \sum_{k+1}^{\ell} q_j)$.

5. THE PRINCIPLE OF THE MULTIPLE GRID ALGORITHM

5.1. The two-level algorithm TLA

The two-level algorithm is a non-stationary defect correction process in which only two different approximate inverses are used:

(1) some relaxation method (e.g. Jacobi, Gauss-Seidel, the incomplete LU-decomposition iteration, etc.) on the fine grid, and

(2) a coarse grid correction.

The approximate inverse in the coarse grid correction for the solution of $L_h x_h = f_h$ is given by $\tilde{G}_h = P_{hH} L_H^{-1} \bar{R}_{Hh}$. Thus a coarse grid correction step in the two-level algorithm reads

$$x_{i+1} = x_i + P_{hH} L_H^{-1} \bar{R}_{Hh}(f_h - L_h x_i). \tag{5.1.1}$$

One step in the two-level algorithm consists of p relaxation sweeps of the relaxation method chosen, a coarse-grid correction step and again q relaxation sweeps of the relaxation method. Such a step of the linear two-level algorithm for the solution of $L_h u_h = f_h$ is described on the following ALGOL-like program.

```
proc TLA = (ref gridf uh, gridf fh) void:
begin
      to p do relax (uh,fh) od;

      d := restrict (Lh * uh - fh);
      solve (LH,v,d);     # solves L_H * v = d  #
      uh := uh - prolongate  v ;

      to q do relax (uh,fh) od
end;
```

To this procedure the right-hand side f_h and an approximate solution u_h are given; by the procedure the given u_h (i.e. u_i) is updated and changed into the new iterand u_{i+1}. Clearly, the amplification operator of one step of this linear two-level algorithm is given by

$$(5.1.2) \qquad M_h^{TLA} = M_h^{TLA,p,q} = (I_h - B_h L_h)^q (I_h - P_{hH} L_H^{-1} \bar{R}_{Hh} L_h)(I_h - B_h L_h)^p,$$

where B_h is the approximate inverse of the relaxation process. The subscript h denotes that the operator is related to the solution of the discrete problem $L_h x_h = f_h$. The superscripts p and q, denoting the number of pre- and post-relaxations, are omitted in M_h^{TLA} if no confusion is possible.
In equation (5.1.2) we recognize the amplification operators of the relaxation process:

$$(5.1.3) \qquad \begin{aligned} M_h^{REL} &= (I_h - B_h L_h), \text{ or} \\ \bar{M}_h^{REL} &= (I_h - L_h B_h); \end{aligned}$$

and we may write

$$(5.1.4) \qquad M_h^{TLA,p,q} = (M_h^{REL})^q (L_h^{-1} - P_{hH} L_H^{-1} \bar{R}_{Hh}) (\bar{M}_h^{REL})^p L_h,$$

or

$$(5.1.5) \qquad \bar{M}_h^{TLA,p,q} = L_h (M_h^{REL})^q (L_h^{-1} - P_{hH} L_H^{-1} \bar{R}_{Hh}) (\bar{M}_h^{REL})^p.$$

We notice that the operator

$$L_h^{-1} - P_{hH} L_H^{-1} \bar{R}_{Hh}$$

determines the relative convergence between the operators L_h and L_H.

The principle of a convergence proof for the TLA

Following the convergence proof for the multi-grid method as given by HACKBUSCH [1980] and references therein] we unravel here first the convergence of the TLA. Sufficient conditions for this convergence are formulated .
In a later stage it will be shown that these conditons are satisfied when particular multigrid methods are applied to certain discretized (e.g. elliptic) boundary value problems.

In this section we assume that $L_h : X_h \to Y_h$ and $L_H : X_H \to Y_H$ are related discretizations of an operator $L : X \to Y$. L_h is the fine and L_H is the coarse discretization, with meshwidths h and H respectively.

DEFINITION. If B_h is the approximate inverse related to some relaxation process for the solution of the equation $L_h y_h = f_h$, then the relaxation process has a *proper smoothing property of order* α if

$$\|L_h(I - B_h L_h)^\nu\| \leq C_0(\nu)\, h^{-\alpha} \tag{5.1.6}$$

with $C_0(\nu)$ independent of h

$C_0(\nu) \to 0$ as $\nu \to \infty$.

Possibly $\nu \in [0, \nu_{max}(h)]$, and $\nu_{max}(h) \to \infty$ as $h \to 0$.

REMARK. The proper smoothing property can also be written

$$\|L_h(M_h^{REL})^\nu\| \leq C_0(\nu)\, h^{-\alpha} \tag{5.1.7}$$

where M_h^{REL} is the error-amplification operator of the relaxation. We can write the property in terms of the residual-amplification operator as well:

$$\|(\bar{M}_h^{REL})^\nu L_h\| \leq C_0(\nu)\, h^{-\alpha}. \tag{5.1.8}$$

EXAMPLE. In the intermezzo we saw for the operator A_h in equation (2) together with damped Jacobi relaxation

$$\|L_h(I - B_h L_h)^\nu\| = \sup_m \left\| \frac{4}{h^2} \begin{pmatrix} s_m^2 & 0 \\ 0 & c_m^2 \end{pmatrix} \begin{pmatrix} c_m^2 & 0 \\ 0 & s_m^2 \end{pmatrix}^\nu \right\|$$

$$\leq \max_{t \in [0,1]} \frac{4}{h^2} t(1-t)^\nu \approx \frac{4}{h^2} \frac{1}{\nu e} \quad \text{for } \nu \to \infty.$$

Hence, for this operator A_h, damped Jacobi has the proper smoothing property of order 2.

REMARK. The addition $\nu \in [0, \nu_{max}(h)]$, with $\nu_{max}(h) \to \infty$ as $h \to 0$, means that it is not necessary that the inequality (5.1.6) holds uniformly in h and ν.

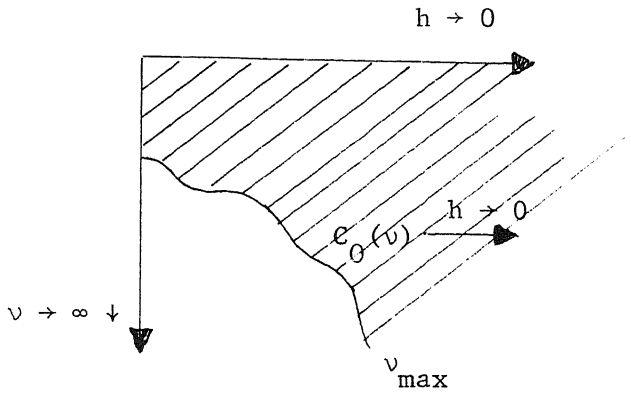


Figure 5.1. The region in the h-ν-plane where the inequality (5.1.6) should be satisfied is the shadowed area.

The convergence of the TLA algorithm is proved by showing that $\|M_h^{TLA}\|_{X_h \to X_h}$ or $\|\bar{M}_h^{TLA}\|_{Y_h \to Y_h}$ is less than one.

THEOREM 5.1.1. *If*

(i) *the operators* L_h *and* L_H *are relative convergent of order* α,

(ii) *the relaxation process for* L_h *satisfies a proper smoothing property of order* α, *and*

(iii) *the discretizations* L_h *and* L_H *satisfy the regular relative mesh property.*

Then the error-amplification operator of the corresponding TLA *satisfies*

$$\|M^{TLA,p,0}\|_{X_h \to X_h} \leq C \cdot C_0(p),$$

where C *is independent of* h *and* $C_0(\nu) \to 0$ *as* $\nu \to \infty$. *Possibly* $\nu \in [0, \nu_{max}(h)]$ *with* $\nu_{max}(h) \to \infty$ *as* $h \to 0$.

PROOF.

$$M^{TLA,p,0} = (I_h - P_{hH} L_H^{-1} \bar{R}_{Hh} L_h)(I_h - B_h L_h)^p$$

$$= (L_h^{-1} - P_{hH} L_H^{-1} \bar{R}_{Hh}) L_h (I_h - B_h L_h)^p.$$

$$\|M^{TLA,p,0}\|_{X_h \to X_h} \leq \|L_h^{-1} - P_{hH} L_H^{-1} \bar{R}_{Hh}\|_{Y_h \to X_h} \|L_h (I_h - B_h L_h)^p\|_{X_h \to Y_h}$$

$$\leq \quad C\, H^{\alpha} \quad . \quad C_0(p) h^{-\alpha}$$

$$= C \,.\, C_0(p) \,.\, (H/h)^{\alpha} \leq C \,.\, C_0(p). \qquad \square$$

Analogously we find a theorem for the residual amplification operator.

THEOREM 5.1.2. *If the conditions* (i), (ii) *and* (iii) *of theorem* 5.1.1 *are satisfied, then the residual amplification operator of the corresponding* TLA *satisfies*

$$\|\bar{M}^{TLA,0,q}\|_{Y_h \to Y_h} \leq C \,.\, C_0(q),$$

where C *is independent of* h *and* $C_0(\nu) \to 0$ *as* $\nu \to \infty$. *Possibly* $\nu \in [0, \nu_{max}(h)]$ *with* $\nu_{max}(h) \to \infty$ *as* $h \to 0$.

PROOF.

$$\bar{M}^{TLA,0,q} = (I_h - L_h B_h)^q (I_h - L_h P_{hH} L_H^{-1} \bar{R}_{Hh})$$

$$= L_h (I_h - B_h L_h)^q (L_h^{-1} - P_{hH} L_H^{-1} \bar{R}_{Hh}).$$

$$\|\bar{M}^{TLA,0,q} \leq \|L_h (I_h - B_h L_h)^q\|_{X_h \to Y_h} \|L_h^{-1} - P_{hH} L_H^{-1} \bar{R}_{Hh}\|_{Y_h \to X_h}$$

$$\leq \quad C_0(q)\, h^{-\alpha} \quad . \quad C\, H^{\alpha}$$

$$= \quad C \,.\, C_0(q)(H/h)^{\alpha} \quad \leq \quad C \,.\, C_0(q). \qquad \square$$

REMARK. If, in addition to the conditions of theorem 5.1.1, we have $\|M_h^{REL}\|_{X_h \to X_h} \leq C$ uniformly in h, we find

$$\|M_h^{TLA,p,q}\|_{X_h \to X_h} \leq C \,.\, C(p),$$

and analogously with $\|\bar{M}_h^{REL}\|_{Y_h \to Y_h} \leq C$ uniformly in h, we find

$$\|\bar{M}_h^{TLA,p,q}\|_{Y_h \to Y_h} \leq C \,.\, C(q).$$

5.2. The linear multi-level algorithm MLA

Whereas in the two-level algorithm we have to solve a coarse-grid problem in each iteration step, in the multi-level algorithm we solve this problem approximately by the application of a few iteration steps of the same multi-level algorithm on the coarse level.

As was explained in section 4.4, by recursion we now have to solve a discretized problem directly only on the very coarsest grid. When σ iteration steps of the multi-level algorithm are used to approximate L_H^{-1}, this multi-level algorithm is described in the following ALGOL-like program

```
proc MLA = (ref grid uh, gridf fh) void:
begin
      to p while ...    do relax(uh,fh) od;

      d := restrict (fh - Lh * uh);
      if level of uh = 1
      then solve (LH,v,d)
      else v := 0;
           to sigma while ...
           do MLA (v,d) od
      fi;
      uh := uh - prolongate v;

      to q while ... do relax (uh,fh) od
end;
```

By while ... we denote in the program that some iterations may be terminated sooner, depending on the speed of convergence or other conditions that can be checked during the computation. Multigrid algorithms that make use of this possibility are said to have an *adaptive strategy*; algorithms where the iterations are controled only by the fixed numbers p, σ and q are said to have a *fixed strategy*. Although the adaptive strategy may be very efficient (cf. BRANDT [1979]), the fixed strategy is better accessible for a theoretical analysis.

For some fixed strategies, we show in figure 5.2 how is switched between the different levels of discretization. We see that - essentially - most relaxation sweeps are performed on the lower levels.

One iteration step in the multi-grid process (i.e. one call of the procedure MLA) is also called one *cycle* of the multi-grid process. Iterative application is also refered to as cycling. One multi-grid cycle is called a V-*cycle* if $\sigma = 1$; if $\sigma = 2$ it is called a W-*cycle*.

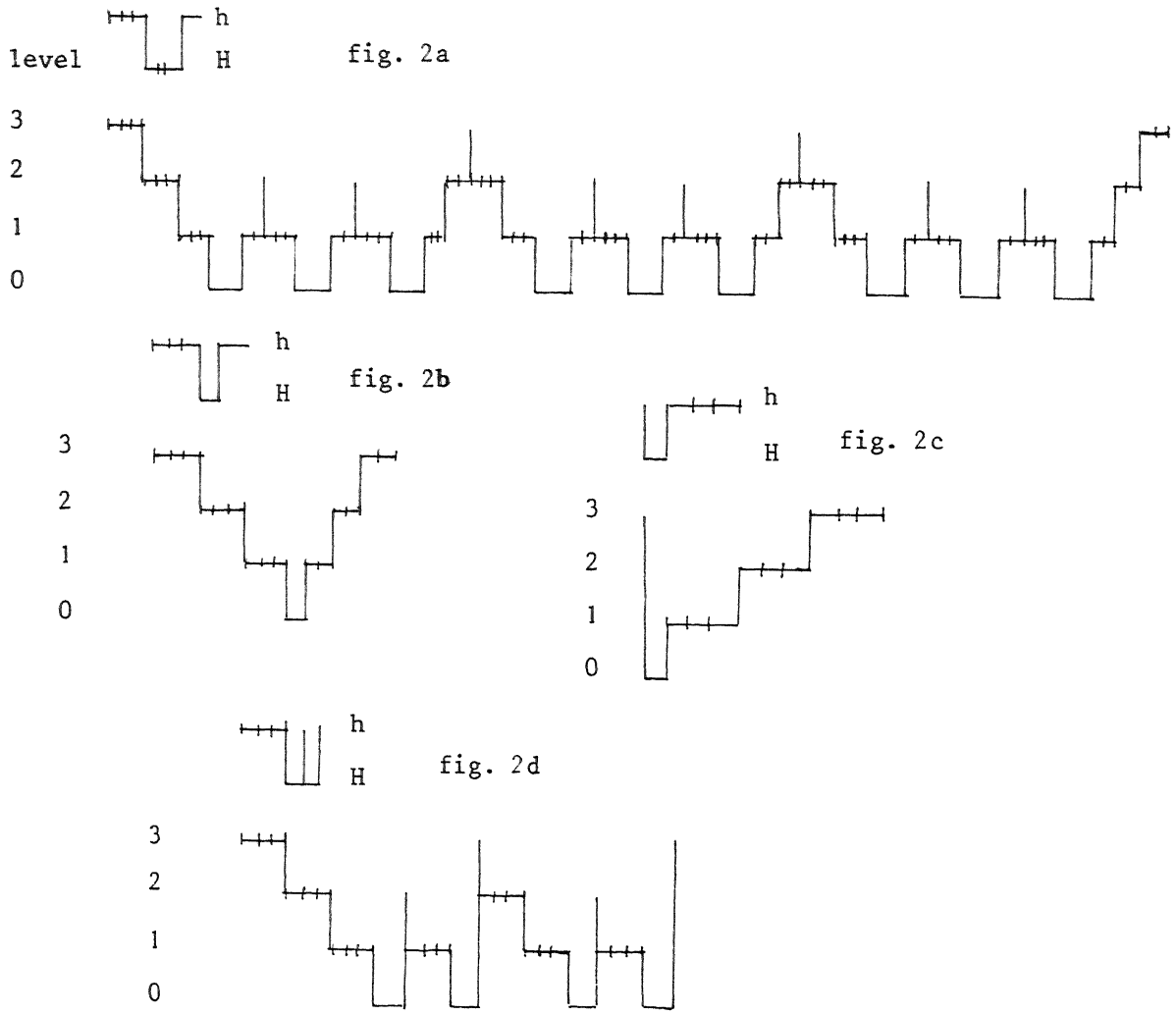


Figure 5.2. The recursive structure of multi-grid algorithms with a fixed strategy.

In all diagrams the number of levels is 4, the very coarsest level is denoted by 0. In each diagram 1a, 1b, 1c or 1d, the basic structure on the two levels h and H is given as well as the recursive structure on one cycle at level 3. Segments between tick-marks on a level > 0 denote the execution of a relaxation step on this level; a segment on level 0 denotes the direct solution on the coarsest level.

The different structures shown are:

1a. A general structure with $p = 3$, $\sigma = 3$ and $q = 2$.

1b. A structure with $\sigma = 1$ (NICOLAIDES [1979] $p = 3$, $q = 2$.

1c. A structure with $\sigma = 1$, $p = 0$ (FREDERICKSON [1975] $q = 3$.

1d. A structure with $q = 0$ (HACKBUSCH [1979] $p = 3$, $\sigma = 2$.

We denote the amplification operator of a multi-level iteration step on the h-level of discretization by M_h^{MLA}, or - if we want to specify the number of relaxation sweeps and the number of coarse grid correction steps - we denote it by

$$M_h^{MLA,p,q} \quad \text{or} \quad M_h^{MLA,p,q,\sigma} .$$

The same amplification operator on the next coarser level we denote by M_H^{MLA}. In the multi-grid cycle the approximate inverse of the coarse grid correction is not given by $P_{hH}\, L_H^{-1}\, \bar{R}_{Hh}$, because in the algorithm L_H^{-1} is approximated by application of σ steps of a defect correction process. The amplification operator of this DCP is given by M_H^{MLA}. Hence, as was shown in section 4.2, the approximate inverse of the σ iteration steps together is given by

$$(I - (M_H^{MLA})^{\sigma})\, L_H^{-1} .$$

Consequently, the amplification operator of the coarse grid correction in MLA is

$$(I - P_{hH}(I - (M_H^{MLA})^{\sigma}) L_H^{-1}\, \bar{R}_{Hh}\, L_h),$$

and we have

$$\begin{aligned} M_h^{MLA} &= M_h^{MLA,p,q,\sigma} \\ &= (M_h^{REL})^q (I - P_{hH}(I - (M_H^{MLA})^{\sigma}) L_H^{-1}\, \bar{R}_{Hh}\, L_h)(M_h^{REL})^p \\ &= M_h^{TLA,p,q} + (M_h^{REL})^q\, P_{hH} (M_H^{MLA})^{\sigma}\, L_H^{-1}\, \bar{R}_{Hh}\, L_h (M_h^{REL})^p \end{aligned} \tag{5.2.1}$$

and

$$(5.2.2)\qquad \bar{M}_h^{MLA} = \bar{M}_h^{TLA,p,q} + L_h(M_h^{REL})^q\, P_{hH}(M_H^{MLA})^\sigma\, L_H^{-1}\, \bar{R}_{Hh}\, (\bar{M}_h^{REL})^p$$

$$= \bar{M}_h^{TLA,p,q} + (\bar{M}_h^{REL})^q\, L_h\, P_{hH}\, L_H^{-1}\, (\bar{M}_H^{MLA})^\sigma\, \bar{R}_{Hh}\, (\bar{M}_h^{REL})^p.$$

The principle of a convergence proof for the MLA

THEOREM 5.2.1. *If the conditions* i), ii) *and* iii) *of theorem* 5.1.1 *are satisfied and in addition*

iv) $P_{hH} : X_H \to X_h$ *is* h*-uniformly bounded and bounded from below,*

v) $(M_h^{REL})^p$ *and* $(M_h^{REL})^q : X_h \to X_h$ *are* h*-uniformly bounded (for all* p *sufficiently large if* $q = 0$*), and*

vi) p *is sufficiently large.*

Then

$$\|M_h^{MLA,p,q,\sigma}\|_{X_h\to X_h} \le \|M_h^{TLA,p,q}\|_{X_h\to X_h} + C\,\|M_H^{MLA}\|^\sigma_{X_H\to X_H}.$$

PROOF. From equation (5.2.1) it follows that

$$\|M_h^{MLA,p,q,\sigma}\| \le \|M_h^{TLA,p,q}\| +$$

$$+ \|(M_h^{REL})^q\|\,\|P_{hH}\|\,\|M_H^{MLA}\|^\sigma\,\|L_H^{-1}\,\bar{R}_{Hh}\,L_h\,(M_h^{REL})^p\|.$$

Further,

$$-P_{hH}\,L_H^{-1}\,\bar{R}_{Hh}\,L_h(M_h^{REL})^p = (L_h^{-1} - P_{hH}L_H^{-1}\,\bar{R}_{Hh} - L_h^{-1})L_h(M_h^{REL})^p$$

$$= (L_h^{-1} - P_{hH}\,L_H^{-1}\,\bar{R}_{Hh})L_h(M_h^{REL})^p - (M_h^{REL})^p$$

$$= M_h^{TLA,p,0} - (M_h^{REL})^p.$$

Since P_{hH} is h-uniformly bounded from below, a $C > 0$ exists, independent of h, such that

$$\|L_H^{-1}\,\bar{R}_{Hh}\,L_h\,(M_h^{REL})^p\| \le C\,\|M_h^{TLA,p,0} - (M_h^{REL})^p\|,$$

and hence

$$\|M_h^{MLA,p,q,\sigma}\| \le \|M_h^{TLA,p,q}\| +$$

$$+ C \|(M_h^{REL})^q\| \|P_{hH}\| \|M_H^{MLA}\|^\sigma \|M_h^{TLA,p,0} - (M_h^{REL})^p\|$$

$$\le \|M_h^{TLA,p,q}\| + C \|(M_h^{REL})^q\| \|M_H^{MLA}\|^\sigma \{\|M_h^{TLA,p,0}\| + \|(M_h^{REL})^p\|\}$$

$$\le \|M_h^{TLA,p,q}\| + C \|M_h^{MLA}\|^\sigma \{C \cdot C_0(p) + C\}$$

$$\le \|M_h^{TLA,p,q}\| + C \|M_h^{MLA}\|^\sigma. \qquad \square$$

THEOREM 5.2.2. *If the conditions* i), ii) *and* iii) *of theorem* 5.1.1 *are satisfied and in addition*

iv) $\bar{R}_{Hh} : Y_h \to Y_H$ *is* h*-uniformly bounded and bounded from below*

v) $(\bar{M}_h^{REL})^p$ *and* $(\bar{M}_h^{REL})^q : Y_h \to Y_H$ *are* h*-uniformly bounded (for all* q *sufficiently large if* p = 0), *and*

vi) q *is sufficiently large.*

Then

$$\|\bar{M}_h^{MLA,p,q,\sigma}\|_{Y_h \to Y_h} \le \|\bar{M}_h^{TLA,p,q}\|_{Y_h \to Y_h} + C \|\bar{M}_H^{MLA}\|^\sigma_{Y_H \to Y_H}.$$

PROOF. The proof is similar to that of theorem 5.2.1. From equation (5.2.2) follows

$$\|\bar{M}_h^{MLA,p,q,\sigma}\| \le \|\bar{M}_h^{TLA,p,q}\| +$$

$$+ \|(\bar{M}_h^{REL})^q L_h P_{hH} L_H^{-1}\| \|\bar{M}_H^{MLA}\|^\sigma \|\bar{R}_{Hh}\| \|(M_h^{REL})^p\|,$$

and

$$(\bar{M}_h^{REL})^q L_h P_{hH} L_H^{-1} \bar{R}_{Hh} = \bar{M}_h^{TLA,0,q} - (\bar{M}_h^{REL})^q.$$

Since $\bar{R}_{Hh}$ is bounded from below

$$\|(\bar{M}_h^{REL})^q L_h P_{hH} L_H^{-1}\| \le C\{\|\bar{M}_h^{TLA,0,q}\| + \|(\bar{M}^{REL})^q\|\}$$

and hence

$$\|\bar{M}_h^{MLA,p,q,0}\| \le \|\bar{M}_h^{TLA,p,q}\| +$$

$$C \cdot (\|\bar{M}_h^{TLA,0,q}\| + \|(\bar{M}^{REL})^q\|)\|\bar{M}_H^{MLA}\|^\sigma \|\bar{R}_{Hh}\| \|(M_h^{REL})^p\|$$

$$\le \|\bar{M}_h^{TLA,p,q}\| + \{C \cdot C_0(q) + C\}\|\bar{M}_H^{MLA}\|^\sigma$$

$$\le \|\bar{M}_h^{TLA,p,q}\| + C\,\|\bar{M}_H^{MLA}\|^\sigma . \qquad \square$$

COROLLARY. *Using the results of section* 5.1 *we immediately conclude from the hypotheses of theorem* 5.2.1 *or theorem* 5.2.2 *that*

$$\|M_h^{MLA,p,q,\sigma}\|_{X_h \to X_h} \le C \cdot C_0(p) + C\|M_H^{MLA}\|^\sigma_{X_H \to X_H} ,$$

or

$$\|\bar{M}_h^{MLA,p,q,\sigma}\|_{Y_h \to Y_h} \le C \cdot C_0(q) + C\|\bar{M}_H^{MLA}\|^\sigma_{Y_H \to Y_H} .$$

Here the constants C, $C_0(p)$ *and* $C_0(q)$ *are independent of the meshwidth* h.

We denote the sequence of discretizations used in the MLA by

$$L_{h_k} X_{h_k} = f_{h_k}, \qquad k = 0,1,2,\ldots ,$$

with $|h_0| \ge |h_1| \ge \ldots$, or briefly by

$$L_k X_k = f_k, \qquad k = 0,1,2,\ldots .$$

Here k is called the *level* of the discretization.
We define

$$K_k = \|M_{h_k}^{TLA}\| \tag{5.2.3}$$

and

$$\tilde{K}_k = \|M_{h_k}^{MLA}\| \qquad k = 1,2,\ldots . \tag{5.2.4}$$

If the conditions i) - vi) hold for the discretizations on all levels $1,2,..,k,...$, then we easily derive from theorem 5.2.1 the recursive relation for $\tilde{K}_k$, viz.

$$(5.2.5) \qquad \begin{cases} \tilde{K}_1 = K_1 \\ \tilde{K}_k \leq K_k + C\, \tilde{K}_{k-1}^{\sigma}. \end{cases}$$

From theorem 5.1.1 we know that $K_k \leq C \cdot C(p)$ independent of the level k.

A similar recursive is obtained for $\|\bar{M}_{h_k}^{MLA}\|$.

The amount of work in one cycle of the multi-level algorithm

If we sum up the work that is done in one cycle of the MLA-algorithm on level level k, we find

1) p+q relaxation sweeps on level k
2) 1 $f_h - L_h U_h$ residual computation on level k
3) 1 application of R_{Hh}
4) 1 application of P_{hH}
5) 1 subtraction $P_{hH} v_H$ from u_h, and
6) σ application of a MLA-cycle at level k-1.

We notice that, for a differential problem, all computations in 1) - 5) require $O(N_k)$ operations if N_k is the number of unknowns in the system $L_k u_k = f_k$.

Hence we write for the amount of work in one MLA-cycle on level k:

$$W_k^{MLA} \leq (p+q)\, C^{REL}\, N_k + C\, N_k + \sigma\, W_{k-1}^{MLA},$$

where C^{REL} denotes the constant depending on the relaxation method.
If we assume that $N_{k-1} = \rho N_k$, (usually $\rho = 2^{-d}$, where d is the dimension of the original problem), we find

$$W_k^{MLA} \leq [(p+q)C^{REL} + C]N_k[1 + \sigma\rho + (\sigma\rho)^2 + \ldots + (\sigma\rho)^{k-1}] + \sigma^k\, W_0^{MLA},$$

where W_0 denotes the (approximate) solution of the problem on the coarsest grid. Hence with $\sigma\rho < 1$ we find

$$W_k^{MLA} \le \frac{N_k}{1-\sigma\rho}\,[(p+q)C^{REL}+C]+\sigma^{k-1}\,W_0$$

$$\le N_k\left[\frac{(p+q)C^{REL}+C}{1-\sigma\rho}+\frac{W_0}{\sigma N_0}\right].$$

This means that W_k^{MLA} is proportional to N_k, i.e. as long as $\sigma < N_k/N_{k-1}$ for $k = 1,2,\ldots$ the amount of work for one multi-grid cycle is proportional to the number of unknowns in the discretization $L_k u_k = f_k$.

The above operation count holds for a discretized differential equation, where the number of operations for a relaxation or a L_h-evaluation is proportional to N. For the solution of systems arising from integral equations, the work for L -evaluation is generally $O(N^2)$. The same reasoning for discretized Fredholm integral equations therefore shows that the amount of work in a multi-grid cycle is $O(N^2)$.