THE DEFECT CORRECTION PRINCIPLE

P.W. Hemker
Mathematical Centre
Amsterdam
The Netherlands

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

Hardback 0-906783-12-7 /82 $2.50 + .10
Paperback 0-906783-09-7 /82 $2.50 + .10

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

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}.

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_o; (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

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

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

$$\tilde{G}F\,x^* = \tilde{G}y = \tilde{G}F\,x. \tag{2.1}$$

(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 (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 $\tilde{G}'$ is a linear operator, then we may write DCPA as

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

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

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

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

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.

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

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

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

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

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

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

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

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

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 $\{x_i\}$.

4. Further remarks on DCPB

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

$$\tilde{G} = \tilde{\Gamma}\Delta, \tag{4.1}$$

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

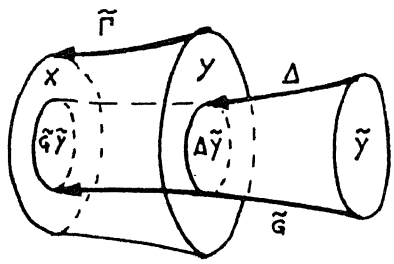

Fig. 4.1. The mappings $\tilde{G}$, Δ and $\tilde{\Gamma}$.

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 $\Lambda \ell_i \in \Lambda \tilde{Y}$, we get the following iterative process of which all iterands are in $\Delta \tilde{Y}$:

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

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

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

$$(4.2) \qquad \begin{cases} \lambda_{i+1} = \lambda_i - \Delta F \tilde{\Gamma} \lambda_i + \Delta y, \\ \lambda_0 = \Delta \ell_0 = \Delta y. \end{cases}$$

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 (4.2) and Theorem 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

$$x_{i+1} = x_i - \tilde{G}Fx_i + \tilde{G}y \tag{5.1}$$

as

$$x_{i+1} = x_i + \tilde{G}(y - Fx_i). \tag{5.2}$$

For general - nonlinear - $\tilde{G}$, the solution of $Fx = y$ is not a fixed point of the latter iteration. In (5.2) the operands of $\tilde{G}$ are in the neightbourhood of zero, whereas in (5.1) they are in the neightbourhood of y and Fx_i. An approximation (linearization) of the non-linear DCPA (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 neighbourhood 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

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

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 bounded 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.

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

$$Fx = y, \tag{6.1}$$

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

$$LU = F + E; \tag{6.1}$$

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

$$\left.\begin{aligned} LU\,x_0 &= y, \\ r_{i+1} &= y - F x_i, \\ LU\,d_{i+1} &= r_{i+1}, \\ x_{i+1} &= x_i + d_{i+1}, \end{aligned}\right\} \quad i = 0,1,2,\ldots . \tag{6.2}$$

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 3.4 we know the upperbound 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* (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

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

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

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 (6.3). With this $\tilde{G}$ we can construct a DCPA or DCPB to find the solution of (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

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

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

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

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

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

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:

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

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

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

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

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

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

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 $n = 1,2,\ldots,$

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

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

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 $\tilde{G}\tilde{Y}$.

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])

$$\tilde{G} = U \Sigma V^T, \tag{7.1}$$

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

$$\tilde{G} = P S R, \tag{7.2}$$

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 4 we remark that, in the finite-dimensional linear case considered here, we can construct a decomposition (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 (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 $\hat{P} = V_1 (R V_1)^{-1}$. 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

$$M = I - \tilde{G}F = I - PSRF. \tag{7.3}$$

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

$$M\, e_s = M\, P\hat{R}\, e_s = (P\hat{R} - PSRFP\hat{R})\, e_s = P(I - SRFP)\hat{R}\, e_s. \tag{7.4}$$

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

$$\begin{cases} (Me)_s = \tilde{G}\hat{P}E\hat{R}\, e_s - \tilde{G}F\, e_u, \\ (Me)_u = \qquad\qquad\quad e_u. \end{cases} \tag{7.5}$$

REMARK. In the context of multi-grid methods (cf.Hemker 1981), the components in Range(P) are called the *smooth components*, those in Kernel($\hat{R}$) the *unsmooth components of the error.*

II. For the residual, the amplification operator is

$$\bar{M} = I - F\tilde{G} = I - FPSR. \tag{3.7.6}$$

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

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

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

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

2. The error *in the residual*

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

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

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

EXTENSIONS OF THE DEFECT CORRECTION PRINCIPLE

P.W. Hemker
Mathematical Centre
Amsterdam
The Netherlands

0. Introduction

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.

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

$$x_{i+1} = x_i - \tilde{G}_{i+1}Fx_i + \tilde{G}_{i+1}y, \tag{1.1}$$

Hardback 0-906783-12-7 /82 $2.50 + .10
Paperback 0-906783-09-7 /82 $2.50 + .10

and

$$\ell_{i+1} = \ell_i - F\tilde{G}_i \ell_i + y. \tag{1.2}$$

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.

$$\ell_{i+1} = \tilde{F}_{i+1}\tilde{G}_i\ell_i - F\,\tilde{G}_i\ell_i + y, \tag{1.3}$$

or DCPB to be equivalent with

$$\tilde{F}_{i+1}x_{i+1} = \tilde{F}_i x_i - F\,x_i + y \tag{1.4}$$

or

$$x_{i+1} = \tilde{G}_{i+1}\tilde{F}_i x_i - \tilde{G}_{i+1}F\,x_i + \tilde{G}i+1\ y. \tag{1.5}$$

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].)

REMARK. From the pratical point of view (1.2) seems to be the more attractive of the two processes (1.1) and (1.2) because in (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 (1.1).

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 $\tilde{\tilde{G}}$ 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

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

combine into a single iteration step of the form

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

This is easily recognized as a new iteration step of the type DCPA, now with the approximate inverse

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

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.

g 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

$$\overline{M} = FMF^{-1} = (I-F\tilde{G})^{\sigma}.$$

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 [1977]), 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

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.

$$
\begin{array}{lll}
DCP_1: & x := x - \tilde{G}_1 (F_1 x - f_1) & \tilde{G}_j \qquad\qquad\quad j = 1,2,\ldots,n, \\
DCP_2: & x := x - \tilde{G}_{2,i} (F_2 x - f_2) & \\
\vdots & \vdots & \tilde{G}_{j,i} \in \{\tilde{G}_j,\ F_{j-1}^{-1}\}, \\
\vdots & \vdots & \qquad\qquad\quad j = 2,3,\ldots,n. \\
DCP_n: & x := x - \tilde{G}_{n,i} (F_n x - f_n) &
\end{array}
$$

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 [1977]).

5. Mixed Defect Correction Processes

Up to now we have considered DCPs where each time one final target problem

(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:

(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*

$$\text{(MDCP)} \qquad \begin{cases} u_{i+\frac{1}{2}} = u_i - \tilde{G}_1(F_1 u_i - y_1), \\ u_{i+1} = u_{i+\frac{1}{2}} - \tilde{G}_2(F_2 u_{i+\frac{1}{2}} - y_2). \end{cases}$$

Thus, the complete iteration step reads

$$(5.3) \qquad u_{i+1} = (I - \tilde{G}_2F_2)(I - \tilde{G}_1F_1)u_i + (I - \tilde{G}_2F_2)\tilde{G}_1y_1 + \tilde{G}_2y_2.$$

We find for MDCP the "amplification operator of the error"

$$(5.4) \qquad M = (I - \tilde{G}_2F_2)(I - \tilde{G}_1F_1).$$

A stationary point $\hat{u}$ of (MDCP) satisfies

$$(5.5) \qquad (I - M)\hat{u} = (I - \tilde{G}_2F_2)\tilde{G}_1y_1 + \tilde{G}_2y_2.$$

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 (5.5) is equivalent with

$$(5.6) \qquad (\tilde{G}_2F_1 + \tilde{G}_2F_2 - \tilde{G}_2F_2\tilde{G}_1F_1)u = (\tilde{G}_1\bar{R}_1 + \tilde{G}_2\bar{R}_2 - \tilde{G}_1F_2\tilde{G}_1\bar{R}_1)y.$$

If equation (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

$$(5.7) \qquad (I-M)w = (I-\tilde{G}_2F_2)\tilde{G}_1F_1w + \tilde{G}_2F_2w$$

and by (5.5) we find

$$(5.8) \qquad (I-M)(w-\hat{u}) = (I-\tilde{G}_2F_2)\tilde{G}_1(F_1w-y_1) + \tilde{G}_2(F_2w-y_2).$$

THEOREM

(i) *Let* (P_1) *and* (P_2) *be two discretizations of* (P) *with*

$$R : X \to X_1; \quad \overline{R}_1 : Y \to Y_1; \quad \overline{R}_2 : Y \to Y_2;$$

and such that $y_1 = \overline{R}_1 y$ *and* $y_2 = \overline{R}_2 y$;

(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

$$\|I - \tilde{G}_k F_k\| \le \|\tilde{G}_k\| \|\tilde{F}_k - F_k\| \le C \cdot C\, h^{q_k} \xrightarrow{h \to 0} 0.$$

Thus,

$$\|M\| \le \|I - \tilde{G}_1 F_1\| \|I - \tilde{G}_2 F_2\| \le C < 1$$

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:

$$\tau_k = y_h - F_k R\tilde{u} = \bar{R}_h F\tilde{u} - F_k R\tilde{u} = (\bar{R}_k F - F_k R)\tilde{u} = \tau_k(\tilde{u})$$

$$\|\tau_k\| = \|\tau_k(\tilde{u})\| \le C\, h^{p_k}.$$

From (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

$$\|R\tilde{u} - \hat{u}\| \le \|(I - M)^{-1}\| \|\tilde{G}_2\| \{\|\tilde{F}_2 - F_2\| \|\tilde{G}_1\| \|\tau_1\| + \|\tau_2\|\}$$

$$\le c \cdot c \{C\, h^{q_2} \cdot C \cdot h^{p_1} + h^{p_2}\}$$

$$< C\, h^{\min(p_1+q_2, p_2)}. \qquad \square$$

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

$$\text{(MDCP)} \qquad \begin{cases} u_{i+k/\ell} = u_{i+(k-1)/\ell} - \tilde{G}_k(F_k u_{i+(k-1)/\ell} - G_k), \\ \qquad\qquad\qquad\qquad k = 1,2,\ldots,\ell. \end{cases}$$

The amplification operator of the error is

$$M = M_\ell M_{\ell-1} \cdots M_2 M_1 .$$

We find

$$(I - M)(\hat{u} - R\tilde{u}) = \sum_{k=1}^{\ell-1} M_\ell M_{\ell-1} \cdots M_{k+1} \tilde{G}_k \tau_k + \tilde{G}_\ell \tau_\ell$$

and hence

$$\begin{aligned} \|\hat{u} - R\hat{u}\| &\le \|(I - M)^{-1}\| \sum_{k=1}^{\ell} \|\tilde{G}_\ell\| \|\tilde{F}_\ell - F_\ell\| \cdots \|\tilde{G}_{k+1}\| \|\tilde{F}_{k+1} - F_{k+1}\| \|\tilde{G}_k\| \|\tau_k\| \\ &\le C \left(\sum_{k=1}^{\ell} c\, h^{q_\ell + q_{\ell-1} + \ldots + q_{k+1} + p_k} + h^{p_\ell} \right) \\ &\le C\, h^{p^*} \end{aligned}$$

with $p^* = \min_{k=1,\ldots,\ell} (p_k + \sum_{j=k+1}^{\ell} q_j)$.

REFERENCES

K.E. ATKINSON, *A survey of Numerical Methods for the solution of Fredholm Integral Equations of the Second Kind*, SIAM, Philadelphia, Pa., 1976.

H. BRAKHAGE, *Über die numerische Behandlung von Integralgleichungen nach der Quadraturformalmethode*, Numer Math. 2 (1960) 183-196.

A. BRANDT, *Multilevel adaptive solutions to boundary-value problems*, Math. Comp. 31 (1977) 333-390.

W. HACKBUSCH, *Bemerkungen zur iterierten Defektkorrektur und zu ihrer Kombination mit Mehrgitterverfahren*, Rev. Roumaine Math. Pures Appl. 26 (1981) 1319-1329.

P.W. HEMKER, *Introduction to multigrid methods*, Nw. Arch. Wisk. 29 (1981) 71-101.

C.L. LAWSON & R.J. HANSON, *Solving least squares problems*, Prentice Hall Inc., N.J., 1974.

H. STETTER, *The defect correction principle and discretization methods*, Num. Math. 29 (1978) 425-443.

D.M. YOUNG, *Iterative solution of large linear systems*, Academic Press, 1971.