DELFT UNIVERSITY OF TECHNOLOGY

REPORT 01-13

The influence of deflation vectors at interfaces on the deflated conjugate gradient method

F.J. Vermolen and C. Vuik

ISSN 1389-6520

Reports of the Department of Applied Mathematical Analysis

Delft 2001

[The influence of deflation vectors at interfaces on the deflated conjugate gradient method]

[F.J. Vermolen and C. Vuik]

August 21, 2001

Abstract

We investigate the influence of the value of deflation vectors at interfaces on the rate of convergence of preconditioned conjugate gradient methods. Our set-up is a Laplace problem in two dimensions with continuous or discontinuous coefficients that vary in several orders of magnitude. In the continuous case we are interested in the convergence acceleration of deflation on block preconditioners. The finite volume discretization gives a matrix with very distinct eigenvalues and hence many iterations are needed to obtain a solution using conjugate gradients. We use an incomplete Choleski preconditioning for the symmetric discretization matrix. Subsequently, deflation is applied to eliminate the disadvantageous effects to convergence caused by the remaining small eigenvalues. Here we show the effects of several variants of algebraic deflation and we propose an optimal choice for the deflation vectors.

Mathematics Classification: 65F10 (Iterative methods for linear systems), 76S05 (Flow in Porous Media)

Keywords: porous media, Laplace equation, finite differences, conjugated gradients, preconditioning, deflation

1 Introduction

1.1 Scope

Large linear systems occur in many scientific and engineering applications. Often these systems result from a discretization of model equations using Finite Elements, Finite Volumes or Finite Differences. The systems tend to become very large for three dimensional cases. Some models involve both time and space as independent parameters and therefore it is necessary to solve such a linear system efficiently at all time-steps.

Presently, direct methods (such as a LU-decomposition) and iterative methods are available to solve such a linear system. When systems get very large and when a discretization has a band-structure, direct methods do not preserve the matrix structure and fill-in causes a loss of efficiency (in computer memory and number of floating point operations). For this case iterative methods are a good alternative. Furthermore, when a time integration is necessary, then the solution of the preceding time-step can be used as a starting vector for the algorithm to get the result on the next time-step. This too supports the use of iterative methods.

Iterative methods such as Gauss-Seidel, Jacobi, SOR and Chebyshev-methods can be used, however, convergence is in general slow and it is often very expensive to determine good estimates of parameters on which they depend. To avoid this problem, the conjugate gradient method, based on steepest descent, is used on the symmetric positive definite (SPD) discretization matrix. We deal with an application from transport in porous media where we encounter extreme contrasts in the coefficients of the partial differential equation. Here a preconditioning is necessary and we use a standard incomplete Choleski factorization as a preconditioner for the conjugate gradient method (ICCG) to improve convergence behavior. Furthermore, deflation is applied to get rid of remaining eigenvalues that delay convergence. Vuik et al. [11] proposed a scheme based on physical deflation in which the deflation vectors are continuous and satisfy the Laplace equation.

A different variant involves the so-called algebraic deflation with discontinuous deflation vectors. Here convergence is speeded up. This was also subject of [10]. They investigate a comparison between physical deflation vectors and algebraic deflation vectors. In their paper two ways of algebraic deflation vectors are applied:

- algebraic deflation vectors restricted to high permeability layers,
- algebraic deflation vectors for each layer.

From numerical experiments it follows that the second choice gives faster convergence. Furthermore, this option turned out to be more efficient for many applications than the use of physical deflation vectors. Therefore, we limit ourselves to the use of algebraic projection vectors for *each* layer.

For references related to the Deflated ICCG method we refer to the overview given in [11]. The DICCG method has already been successfully used for complicated magnetic field simulations [2]. A related method is recently presented in [8]. In [3] deflation is used to accelerate block-IC preconditioners combined with Krylov subspace methods in a parallel computing environment.

The DICCG method is related to coarse grid correction, which is used in domain decomposition methods [4, 8]. Therefore insight in a good choice of the deflation vectors can probably be used to devise comparable strategies for coarse grid correction approaches.

Since the deflation technique has been successfully used to solve Poisson type problems with strong contrasts in the coefficients and in combination with a parallel block IC preconditioner we aim to generalize the method to a parallel solver for the discretized (in)compressible Navier-Stokes equations. We assume that the domain Ω consists of a number of disjoint sets $\Omega_j, j = 1, ..., m$ such that $\cup_{j=1}^{m} \bar{\Omega}_j = \bar{\Omega}$. The division in subdomains is motivated by jumps in the coefficients and/or the data distribution used to parallelize the solver. For the construction of the deflation vectors it is important which type of discretization is used: cell centered or vertex centered.

Cell centered

For this discretization the unknowns are located in the interior of the finite element or finite volume. The domain decomposition is straightforward as can be seen in Figure 1. The algebraic deflation vectors z_j are uniquely defined as:

$$z_j(\underline{x}_i) = \begin{cases} 1, & \text{for } \underline{x}_i \in \Omega_j \\ 0, & \text{for } \underline{x}_i \in \Omega \setminus \Omega_j. \end{cases}$$

Vertex centered

If a vertex centered discretization is used the unknowns are located at the boundary of the finite volume or finite element. Two different ways for the data distribution are known [9]:

- Element oriented decomposition: each finite element (volume) of the mesh is contained in a unique subdomain. In this case interface nodes occur.
- Vertex oriented decomposition: each node of the mesh is an element of a unique subdomain. Now some finite elements are part of two or more subdomains.

Note that the vertex oriented decomposition is not well suited to combine with a finite element method. Furthermore for interface points it is not uniquely defined to which subdomain they belong. Therefore we restrict ourselves to the element oriented decomposition, see Figure 2. As a consequence of this the deflation vectors overlap at interfaces.

In our previous work we always use non-overlapping deflation vectors. In [11, 12] the interface vertices are elements of the high permeability subdomains, whereas in [3] no interface vertices occur due to a cell centered discretization. The topic of this paper is: how to choose the value of the deflation vectors at interface points in order to obtain an efficient, robust and parallelizable black-box deflation method.

First we briefly present the mathematical model that we use to compare the various deflation vectors. Subsequently we give the algorithm and describe different versions of deflation. This is followed by a description of the numerical experiments.

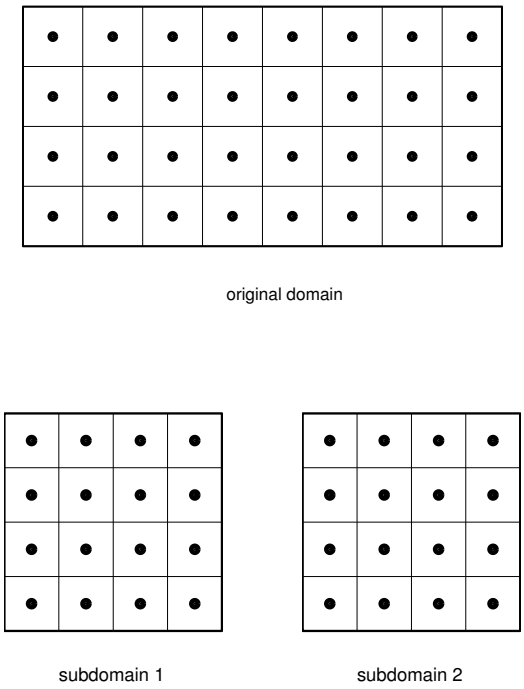

Figure 1: Domain decomposition for a cell centered discretization.

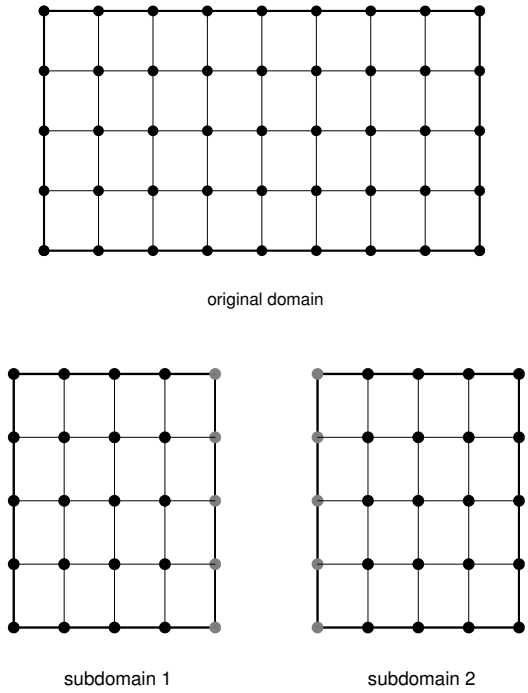

Figure 2: Domain decomposition for a vertex centered discretization. The grey nodes are the interface nodes.

1.2 The mathematical model

We denote the horizontal and vertically downward pointing coordinates by x and y. Flow in porous media is often modelled by the following coupled scaled problem:

$$(P_0)\begin{cases} \dfrac{\partial S}{\partial t} + \nabla \cdot (\underline{q} S) = \nabla \cdot (D(S) \nabla S), \\ \\ \nabla \cdot \underline{q} = 0, \\ \\ \underline{q} + \sigma \nabla p - S \underline{g} = \underline{0}. \end{cases}$$

Above equations are supplemented with appropriate initial and boundary conditions. In above equations S (-), $\underline{q}$ (m/s) and p (Pa) are the unknown saturation, discharge and pressure. The time is denoted by t (s) and σ is the (known) mobility. Porous media mostly consist of several layers where the mobility varies between several orders of magnitude. In this work we take σ as a (piecewise) constant function. For an overview of the equations that occur in modeling flow in porous media we refer to the books of among others Bear [1] and Lake [6].

The third equation of (P_0) is substituted into the second equation to give

$$\nabla \cdot \left\{ S \underline{g} - \sigma \nabla p \right\} = 0. \tag{1}$$

In the present work we focus on the solution of equation (1) on a square domain. We disregard gravity since it does not complicate the numerical scheme. Then we have to solve the following problem in the open square domain $\Omega := \{(x, y) \in \mathbb{R}^2 : (x, y) \in (0, 1) \times (0, 1)\}$:

$$(P_1)\begin{cases} -\nabla \cdot (\sigma \nabla p) = 0, & \text{for } (x, y) \in \Omega \\ \\ p = 1, & \text{for } y = 0, \\ \\ \dfrac{\partial p}{\partial n} = 0, & \text{for } y = 1 \text{ or } x \in \{0, 1\}. \end{cases}$$

Furthermore, suppose that the closed domain $\overline{\Omega}$ consists of m horizontal closed subdomains, $\overline{\Omega}_1, ..., \overline{\Omega}_m$. Clearly, the exact solution of (P_1) is given by $p(x, y) = 1$ for $(x, y) \in [0, 1] \times [0, 1]$. In our set-up we let σ be discontinuous and we suppose $\sigma \in \{10^{-7}, 1\}$, i.e. for $j \in \mathbb{N} : 2j, 2j + 1 \leq m$ and $(x, y) \in \Omega$:

$$\sigma(x, y) = \begin{cases} \sigma_{\max} = 1, & (x, y) \in \overline{\Omega}_{2j+1} \\ \\ \sigma_{\min} = 10^{-7}, & (x, y) \in \Omega_{2j}. \end{cases}$$

To deal with the discontinuities we look for weak solutions of problem (P_1). We define the set of functions V where for each element v we require that ∇v is square integrable over Ω and where $v = 0$ over $y = 0$, i.e.

$$V := \{v \in L^2(\Omega) : v|_{y=0} = 0\}.$$

Multiplication of $(\mathrm{P}_1)_1$ with a test-function $v \in V$ and integration over Ω, gives

$$-\int_\Omega \nabla \cdot (\sigma \nabla p) v dA = 0.$$

Partial integration and application of the Two-dimensional Divergence Theorem and use of both $v|_{y=0} = 0$ and $\dfrac{\partial u}{\partial n} = 0$ for $y = 1$ or $x \in \{0, 1\}$ gives the following variational (weak) formulation:

$$(WF)\begin{bmatrix} \text{Find } u \in L^2(\Omega) \text{ with } u|_{y=0} = 1, \text{ such that} \\ \\ \int_\Omega \sigma \nabla u \cdot \nabla v dA = 0, \quad \forall v \in V, \\ \\ \text{where } V := \{v \in L^2(\Omega) : v|_{y=0} = 0\}. \end{bmatrix}$$

It can be shown that smooth solutions of (WF) also satisfy problem (P_1). We refine the class of functions V to continuous functions in $\overline{\Omega}$ with compact support near the interface $\ell_j := \overline{\Omega}_j \cap \overline{\Omega}_{j+1}$ in a region $D_j \subset \overline{\Omega}$, i.e.

$$V' := \{v \in C(\Omega) : v|_{\Omega \setminus D_j} = 0\},$$

where we define D_j as the region within $\overline{\Omega}$ where the distance to the interface ℓ_j is less than δ, i.e.

$$D_j := \{(x, y) \in \Omega : dist((x, y), \ell_j) < \delta\}.$$

Combination of $(P_1)_1$ and partial integration of (WF) gives

$$\int_{D_j} \nabla \cdot (\sigma \nabla u v) dA = 0.$$

We split the region D_j into two complimentary subregions D_T and D_B separated by ℓ_j, where $\sigma = \begin{cases} \sigma_T, (x, y) \in D_T \\ \sigma_B, (x, y) \in D_B \end{cases}$ for $(x, y) \in D_j$. With $v \in V'$, hence $v|_{\partial D_j} = 0$, follows, using the two-dimensional Divergence Theorem

$$\int_{\ell_j} \underline{n}_T \cdot (\sigma_T \nabla p_T) v ds = -\int_{\ell_j} \underline{n}_B \cdot (\sigma_B \nabla p_B) v ds. \tag{2}$$

Here $\underline{n}_T$ and $\underline{n}_B$ are unit normal vectors on ℓ outward from D_T and D_B. Using Dubois-lemma, it follows from 2

$$\sigma_T \frac{\partial p_T}{\partial n} = \sigma_B \frac{\partial p_B}{\partial n}.$$

This is the condition that we use to discretize (P_1) and for convenience we consider a medium with straight horizontal layers and hence horizontal discontinuities. We solve system (P_1) with finite differences. The resulting linear system has a symmetric coefficient matrix.

To solve above problem with large contrasts of σ, Vuik et al. [11] proposed and analyzed an efficient algorithm based on a preconditioned conjugated gradient (CG) method. Unfortunately, the presence of very small eigenvalues delays the convergence rate significantly. The main contribution of their algorithm is the use of deflation to annihilate this delay due to the very small eigenvalues. Furthermore, a classical termination criterion becomes robust again. An incomplete Choleski (IC) matrix factorization is used as a preconditioner. New in this report is the definition and analysis of various values of the algebraic deflation vectors at the interfaces.

1.3 Discretization of (P_1) and the discretization matrix

As mentioned in Section 1.1 we solve problem (P_1) using a Finite Difference method. We divide the domain $[0, 1] \times [0, 1]$ into $n \times n$ equidistant grid cells. Let i and j be respectively the i^{th} and j^{th} horizontal and vertical gridnode, so that

- the Dirichlet boundary contains grid-nodes ($j = 0$) and
- the Neumann boundaries lie in the middle of subsequent gridnodes (i.e. between $i = 0$ and $i = 1$, $i = n$ and $i = n + 1$, $j = n$ and $j = n + 1$).

The remainder of the gridnodes are within the domain $[0, 1] \times [0, 1]$. The geometry is shown in Figure 3.

As a result from discretization we obtain the following system of equations

$$A\underline{x} = \underline{b},$$

where $A \in \mathbb{R}^{n^2 \times n^2}$, $\underline{x} \in \mathbb{R}^{n^2}$, $\underline{b} \in \mathbb{R}^{n^2}$ respectively represent the discretization matrix, vector with unknown pressures and vector with the right-hand side. The matrix A is symmetric and positive definite. Furthermore, the discretization is chosen such that the interfaces between subsequent layers coincide with gridpoints.

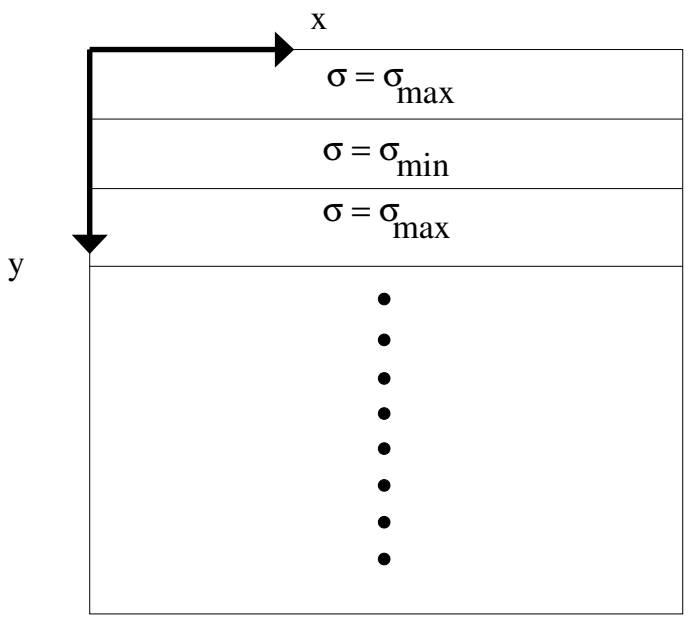

Figure 3: The geometry of the layered domain.

2 Solution of the matrix equation

Since A is symmetric and positive definite, the conjugate gradient method is a natural candidate to solve the matrix equation.

Vuik et al. [11] observe that the number of small eigenvalues (order 10^{-7}) of A is equal to the number of gridnodes in the low mobility layer ($\sigma = 10^{-7}$) plus the number of high permeability layers that have a low permeability layer on top. The conjugate gradient method converges to the exact solution only once all small eigenvalues have been 'discovered'. The number of eigenvalues is reduced by the use of a preconditioner (Incomplete LU or Choleski or even a diagonal scaling). However, still a number of small eigenvalues remain for the preconditioned matrix. These small eigenvalues persist due to the fact that at each interface between low-and high permeability layers a homogeneous Neumann condition is effectively adopted by the sand-layer. This makes the blocks of the discretization matrix that correspond to the sandwiched sandlayers almost singular. The following theorem is proven by Vuik et al. [11]:

Theorem 1: *Let ε be small enough and $D = diag(A)$ and let q be the number of layers with σ of order one between low σ layers, then the diagonally scaled matrix $D^{-1/2}AD^{-1/2}$ has exactly q eigenvalues of order ε.* □

The preconditioning aims at improving the condition of the matrix. However, for this case q small eigenvalues persist. In experiments we use a diagonal preconditioner for the one-dimensional cases, whereas for our two-dimensional experiments an incomplete Choleski decomposition is used as a preconditioner for the symmetric positive definite discretization matrix.

The DICCG-method was proposed by Vuik et al. [11] for the Laplace problem with extreme contrasts of the coefficients. The aim is to get rid of the remaining eigenvalues of the preconditioned matrix $\tilde{A} = L^{-T}L^{-1}A$, where $L^{-T}L^{-1} \approx A^{-1}$ is the IC-preconditioner.

2.1 Deflation

In this subsection we analyze the elimination of the small eigenvalues of $\tilde{A}$ by deflation. Suppose that λ_i and $\underline{z}_i$, $i \in \{1, \ldots, m\}$ respectively represent eigenvalues and orthonormal eigenvectors of

the symmetric discretization matrix $A \in \mathbb{R}^{n^2 \times n^2}$ (such that $\underline{z}_j^T \underline{z}_i = \delta_{ij}$), and define the operator $P \in \mathbb{R}^{n^2 \times n^2}$

$$\overline{P} := I - \sum_{j=1}^{m} \underline{z}_j \underline{z}_j^T, \qquad m \leq n^2$$

then

$$\overline{P} A \underline{z}_i = A\underline{z}_i - \sum_{j=1}^{m} \underline{z}_j \underline{z}_j^T A \underline{z}_i = A\underline{z}_i - \lambda_i \sum_{j=1}^{m} \underline{z}_j \underline{z}_j^T \underline{z}_i = \underline{0}, \qquad \forall\ i \in \{1, \ldots, m\}$$

$$\overline{P} A \underline{z}_k = A\underline{z}_k - \sum_{j=1}^{m} \underline{z}_j \underline{z}_j^T A \underline{z}_k = A\underline{z}_k - \lambda_k \sum_{j=1}^{m} \underline{z}_j \underline{z}_j^T \underline{z}_k = \lambda_k \underline{z}_k, \quad \forall\ k \in \{m+1, \ldots, n^2\}.$$

Hence, we state the following result:

Proposition 2: *Let $\overline{P} := I - \sum_{i=1}^{m} \underline{z}_i \underline{z}_i^T$, where $\underline{z}_i$ and λ_i are respectively orthogonal eigenvectors and eigenvalues of the matrix A, then*

1. *$\overline{P}A$ and A have the same eigenvalues λ_j, $j > m$ and the corresponding eigenvalues of A given by λ_i for $i \leq m$ are all zero for the matrix $\overline{P}A$;*
2. *the matrix $\overline{P}$ is a projection.*

Proof: The first statement is proven by the argument above proposition 1. To prove the second statement, we compute

$$\overline{P}^2 = (I - \sum_{i=1}^{m} \underline{z}_i \underline{z}_i^T)(I - \sum_{i=1}^{m} \underline{z}_i \underline{z}_i^T) = I - 2\sum_{i=1}^{m} \underline{z}_i \underline{z}_i^T + (\sum_{i=1}^{m} \underline{z}_i \underline{z}_i^T)(\sum_{i=1}^{m} \underline{z}_i \underline{z}_i^T) = I - \sum_{i=1}^{m} \underline{z}_i \underline{z}_i^T = \overline{P}.$$

The second last equality results from orthonormality of the eigenvectors. Hence, the matrix $\overline{P}$ is a projection. □

From above proposition follows that $\overline{P}$ and $\overline{P}A$ are singular. The matrix $\overline{P}$ can be re-written:

Proposition 3: *Let $Z \in \mathbb{R}^{n^2 \times m} : Z = (\underline{z}_1 \ldots \underline{z}_m)$ where $\underline{z}_i$ are the orthonormal eigenvectors with eigenvectors λ_i of the matrix A, then the projection $\overline{P}$ can be re-written as $\overline{P} = I - AZ(Z^T AZ)^{-1} Z^T$.*

Proof: Let $\overline{P} = I - AZ(Z^T AZ)^{-1} Z^T$, then

$$\overline{P} = I - A(\underline{z}_1 \ldots \underline{z}_m) \left(\begin{pmatrix} \underline{z}_1^T \\ \ldots \\ \underline{z}_m^T \end{pmatrix} A (\underline{z}_1 \ldots \underline{z}_m) \right)^{-1} \begin{pmatrix} \underline{z}_1^T \\ \ldots \\ \underline{z}_m^T \end{pmatrix} =$$

$$= I - (A\underline{z}_1 \ldots A\underline{z}_m) \begin{pmatrix} \underline{z}_1^T A \underline{z}_1 & \ldots & \underline{z}_1^T A \underline{z}_m \\ \ldots & \ldots & \ldots \\ \underline{z}_m^T A \underline{z}_1 & \ldots & \underline{z}_m^T A \underline{z}_m \end{pmatrix}^{-1} \begin{pmatrix} \underline{z}_1^T \\ \ldots \\ \underline{z}_m^T \end{pmatrix} =$$

$$= I - (\lambda_1 \underline{z}_1 \ldots \lambda_m \underline{z}_m) \mathrm{diag}\left(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_m} \right) \begin{pmatrix} \underline{z}_1^T \\ \ldots \\ \underline{z}_m^T \end{pmatrix} = I - \sum_{i=1}^{m} \underline{z}_i \underline{z}_i^T$$

This proves the statement. □

Corollary 4: The matrix $\overline{P}A$ is symmetric positive semi-definite.

Proof: Symmetry of $\overline{P}A$ is established by, using Proposition 3,

$$(\overline{P}A)^T = (A - AZ(Z^TAZ)^{-1}Z^TA)^T = A - AZ(Z^TAZ)^{-1}Z^TA \qquad (A^T = A).$$

Furthermore, from Proposition 2 and since A is symmetric positive definite, it follows that all eigenvalues of $\overline{P}A$ are nonnegative. Combining symmetry of $\overline{P}A$ and absence of negative eigenvalues, it follows that $\overline{P}A$ is symmetric positive semi-definite. □

Furthermore, the matrix PA, where $P := I - AZ(Z^TAZ)^{-1}Z^T$, is singular for general Z:

$$PAZ = AZ - AZ(Z^TAZ)^{-1}Z^TAZ = AZ - AZ = 0.$$

The columns of Z are eigenvectors of PA with zero value. When we take eigenvectors of A, associated with small eigenvalues of A, for the columns of Z, then the small eigenvalues vanish for PA.

Proposition 5: *Let $P \in \mathbb{R}^{n^2 \times n^2}$ be defined by $P := I - AZ(Z^TAZ)^{-1}Z^T$, with $Z = (\underline{z}_1, \ldots, \underline{z}_m)$, then*

1. *the operator P is a projection for all $Z \in \mathbb{R}^{n^2 \times m}$;*
2. *the null-space of P is spanned by the independent set $\{A\underline{z}_1, \ldots, A\underline{z}_m\}$, i.e. $nul(P) = Span\{A\underline{z}_1, \ldots, A\underline{z}_m\}$.*

Proof:

1. $P^2 = I - 2AZ(Z^TAZ)^{-1}Z^T + AZ(Z^TAZ)^{-1}Z^TAZ(Z^TAZ)^{-1}Z^T = P$.
2. Since $PAZ = 0$, it follows that $A\underline{z}_i \in nul\ P$ and hence $dim\ nul\ P \geq m$. Let $V := Span\{\underline{z}_1, \ldots, \underline{z}_m\}$, then from the Direct Sum Theorem (see for instance [5]) follows $\mathbb{R}^{n^2} = V \oplus V^\perp$, where $V^\perp = \{\underline{y} \in \mathbb{R}^{n^2} : \underline{y} \perp V\}$, hence $dim\ V^\perp = n^2 - m$. Suppose $\underline{y} \in V^\perp$, then

$$P\underline{y} = \underline{y} - AZ(Z^TAZ)^{-1}Z^T\underline{y} = \underline{y}, \qquad \forall \underline{y} \in V^\perp.$$

   Hence $dim\ col\ P \geq n^2 - m$. Since $dim\ nul\ P + dim\ col\ P = n^2$, this implies $dim\ nul\ P \leq m$. Hence, using $dim\ nul\ P \geq m$, we have $dim\ nul\ P = m$. Since $\{A\underline{z}_1, \ldots, A\underline{z}_m\}$ represents a linearly independent set of n vectors in $nul\ P$ and $dim\ nul\ P = n$, it follows from the Basis-Theorem (see for instance [7]) that

$$nul\ P = Span\{A\underline{z}_1, \ldots, A\underline{z}_m\}.$$

This proves the Proposition 5. □

For completeness we present the eigenvalues of the preconditioned matrix $\tilde{A} = L^{-T}L^{-1}A$ in Figure 4 for a one-dimensional case, where $L = diag(A)$ and $n = 99$, $m = 10$ (99 internal gridnodes and 10 subdomains). The condition number of $\tilde{A}$ is improved by elimination of the smaller eigenvalues. The matrix P is referred to as the deflation matrix / operator and the matrix P is chosen such that the small eigenvalues are eliminated. The vectors $\underline{z}_1, \ldots, \underline{z}_m$ are referred to as the projection vectors and they are chosen such that their span approaches the span of the small eigenvectors of $\tilde{A}$. The advantage of working with the matrix $P\tilde{A}$ rather than with $\tilde{A}$ is that the smaller eigenvalues of $\tilde{A}$ are transferred to zero eigenvalues of $\overline{P}A$ which do not influence the convergence of the CG-method.

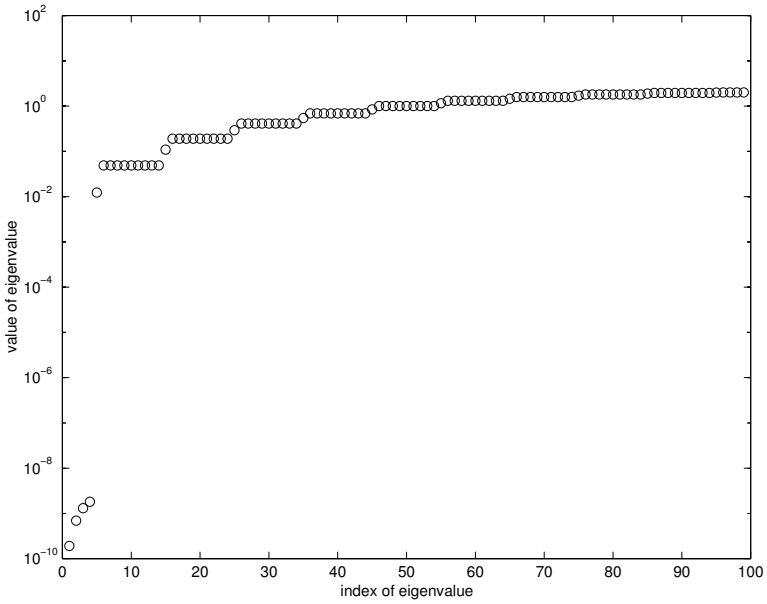

Figure 4: The eigenvalues of the matrix $\tilde{A}$ for the case of high contrasts of the permeability. Above data are given for a one-dimensional example. For more dimensional cases the results are similar.

2.2 Deflated Incomplete Choleski Preconditioned Conjugate Gradients

The elimination of the small eigenvalues of A takes place by using the projection matrix P. We then solve

$$PA\underline{x} = P\underline{b}, \tag{3}$$

with the ICCG-method, where PA is singular. The solution of above equation is not unique. We obtain the solution of the matrix-equation $A\underline{x} = \underline{b}$ by

$$\underline{x} = (I - P^T)\underline{\tilde{x}} + P^T\underline{\tilde{x}},$$

where $\underline{\tilde{x}}$ is the solution of equation (3). In this paragraph we first establish uniqueness of the above $\underline{x}$, given any $\underline{\hat{x}}$ that satisfies equation (3). Equation (3) is written as $P(A\underline{\tilde{x}} - \underline{b}) = \underline{0}$, where $A\underline{\tilde{x}} - \underline{b} \in nul\ P$. Since $PAZ = 0$, the vectors $A\underline{z}_i$ are in the null-space of P, i.e. $A\underline{z}_i \in nul\ P$. Since, we know from Proposition 5 that $nul\ P = Span\{A\underline{z}_1, \ldots, A\underline{z}_m\}$, the vectors in the null-space of P can be written as linear combinations $\underline{w} = \sum_{i=1}^{m} \alpha_i A\underline{z}_i \in nul\ P$. Therefore, it can be seen that one can write for the vector $\underline{\tilde{x}}$:

$$\underline{\tilde{x}} = A^{-1}\underline{b} + A^{-1}\underline{w} = A^{-1}\underline{b} + \sum_{i=1}^{m} \alpha_i \underline{z}_i. \tag{4}$$

Since A is not singular, the first term in the right-hand side is uniquely determined. We investigate the result obtained from multiplication of the second term in the right-hand side with the matrix P^T. For convenience we look at the product $P^T Z$:

$$P^T Z = (I - Z(Z^T AZ)^T Z^T A)Z = Z - Z(Z^T AZ)^{-1} Z^T AZ = 0.$$

Hence, this product is zero and the second term of the right-hand side of equation (4) vanishes, since $Z = (\underline{z}_1 \ldots \underline{z}_m)$. Hence, the vector $P^T\underline{\tilde{x}}$ is unique. This implies that the solution of the matrix equation $A\underline{x} = \underline{b}$ is uniquely determined from

$$\underline{x} = (I - P^T)\underline{\tilde{x}} + P^T\underline{\tilde{x}} = Z(Z^T AZ)^{-1} Z^T\underline{b} + P^T\underline{\tilde{x}} = Z(Z^T AZ)^{-1} Z\underline{b} + P^T\underline{\tilde{x}}.$$

After determination of $\underline{\tilde{x}}$ by solving equation (3), we obtain a unique solution of $A\underline{x} = \underline{b}$ for $\underline{x}$. For completeness, we present the algorithm of the deflated ICCG.

Algorithm 1 (DICCG [11]):
$k = 0, \underline{\tilde{r}}_0 = P\underline{r}_0, \underline{p}_1 = \underline{z}_0 = L^{-T}L^{-1}\underline{\tilde{r}}_0$
while $||\underline{\tilde{r}}_k||_2 > \varepsilon$
$k = k + 1$
$\alpha_k = \dfrac{\underline{\tilde{r}}_{k-1}^T \ \underline{z}_{k-1}}{\underline{p}_k^T PA\underline{p}_k}$
$\underline{x}_k = \underline{x}_{k-1} + \alpha \underline{p}_k$
$\underline{\tilde{r}}_k = \underline{\tilde{r}}_{k-1} - \alpha_k PA\underline{p}_k$
$\underline{z}_k = L^{-T}L^{-1}\underline{\tilde{r}}_k$
$\beta_k = \dfrac{\underline{\tilde{r}}_k^T \ \underline{z}_k}{\underline{\tilde{r}}_{k-1}^T \ \underline{z}_{k-1}}$
$\underline{p}_{k-1} = \underline{z}_k + \beta_k \underline{p}_k$
end while

2.3 Choice of deflation vectors

We define $\partial\Omega_j := \overline{\Omega}_j \backslash \Omega_j$. Let $\underline{z}_j$ be the deflation vector corresponding to the j^{th} layer (subdomain), then we choose the deflation vectors for $j \in \{1, ..., m\}$ as follows

$$z_j(\underline{x}) = \begin{cases} 0, & \text{for } \underline{x} \in \Omega \setminus \overline{\Omega}_j \\ \in [0,1], & \text{for } \underline{x} \in (\partial\Omega_j \cap \partial\Omega_{j-1}) \cup (\partial\Omega_j \cap \partial\Omega_{j+1}) \\ 1, & \text{for } \underline{x} \in \Omega_j. \end{cases}$$

Here $\partial\Omega_0$ and $\partial\Omega_{n+1}$ are the horizontal top $y = 0$ and bottom $y = 1$ boundaries of the square domain Ω. In above expression the value for $\underline{z}_j$ at positions on the boundaries of Ω_j can be chosen in the range $[0, 1]$. We investigate the following choices at the boundaries between subsequent subdomains:

1. **no overlapping** projection vectors of the subdomains:

$$\begin{cases} z_{2j}(\underline{x}) = 0, & \text{for } \underline{x} \in (\partial\Omega_{2j} \cap \partial\Omega_{2j-1}) \cup (\partial\Omega_{2j} \cap \partial\Omega_{2j+1}) \text{ (even, low permeable layer)}, \\ z_{2j+1}(\underline{x}) = 1, & \text{for } \underline{x} \in (\partial\Omega_{2j+1} \cap \partial\Omega_{2j}) \cup (\partial\Omega_{2j+1} \cap \partial\Omega_{2j+2}) \text{ (odd, high permeable layer)} \end{cases}$$

2. **complete overlapping** projection vectors of the subdomains:

$$z_j(\underline{x}) = 1, \qquad \text{for } \underline{x} \in (\partial\Omega_j \cap \partial\Omega_{j-1}) \cup (\partial\Omega_j \cap \partial\Omega_{j+1}) \text{ (for all layers)}.$$

3. **average overlapping** projection vectors of the subdomains:

$$z_j(\underline{x}) = \frac{1}{2}, \qquad \text{for } \underline{x} \in (\partial\Omega_j \cap \partial\Omega_{j-1}) \cup (\partial\Omega_j \cap \partial\Omega_{j+1}) \text{ (for all layers)}.$$

4. **weighted overlapping** projection vectors of the subdomains:

$$\begin{cases} z_{2j}(\underline{x}) = \dfrac{\sigma_{min}}{\sigma_{max} + \sigma_{min}}, & \text{for } \underline{x} \in (\partial\Omega_{2j} \cap \partial\Omega_{2j-1}) \cup (\partial\Omega_{2j} \cap \partial\Omega_{2j+1}) \text{ (even)}, \\ z_{2j+1}(\underline{x}) = \dfrac{\sigma_{max}}{\sigma_{max} + \sigma_{min}}, & \text{for } \underline{x} \in (\partial\Omega_{2j+1} \cap \partial\Omega_{2j}) \cup (\partial\Omega_{2j+1} \cap \partial\Omega_{2j+2}) \text{ (odd)}. \end{cases}$$

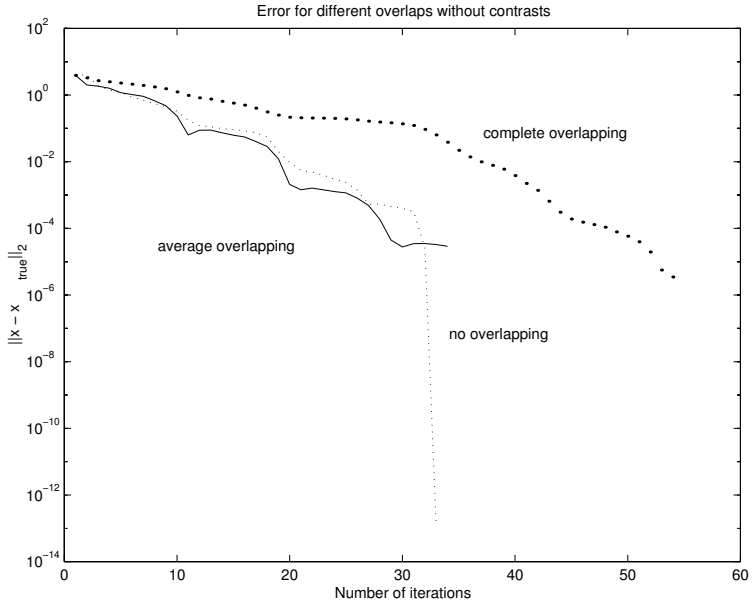

Figure 5: Convergence behavior of the deflated CG-method for $\sigma = 1$ and 10 subdomains.

3 Numerical experiments

We have done experiments with the different overlapping between subsequent subdomains for a one- and two dimensional case. The parameter σ and the number of subdomains have also been varied. We start to report the results from the one-dimensional experiments.

3.1 One dimensional experiments

We start with the case that $\sigma = 1$ for $\underline{x} \in \Omega$. The solution is calculated with non-overlapping, average overlapping, and completely overlapping projection vectors. Note that for this choice of σ the case of weighted projection vectors coincides with average overlapping. The convergence results have been plotted in Figure 5. The experiments were done for 99 internal gridpoints and 10 subdomains. From Figure 5 it can be seen that the average overlap gives fastest convergence. Whereas complete overlapping produces the slowest convergence.

The eigenvalues of the matrix $L^{-T}L^{-1}PA$ (note that L is here a diagonal matrix) for complete overlapping projection vectors are shown in Figure 6. We see that the smallest eigenvalues are in the range of 10^{-15}, which is within the roundoff errors and therefore these eigenvalues can be considered as zero. Hence these eigenvalues are referred as being not *significant*. It can be seen that the *significant* eigenvalues range between less then 10^{-2} and more than 1. The eigenvalues for the same case but with average overlapping are plotted in Figure 7. Here the eigenvalues range between more than 10^{-2} and more than 1. Hence for the case of average (weighted) overlapping the effective condition number of the matrix $L^{-T}L^{-1}PA$ is smaller than for the case of the complete overlapping. We also computed the eigenvalues for the case of no overlapping: the difference found was negligible. This supports the observations from Figure 5.

Figure 8 displays the results that have been obtained for the case that σ varies abruptly: $\sigma \in \{1, 10^{-7}\}$. Here the results are given for the cases of no overlap, complete overlap, weighted overlap and average overlap. It can be seen that the results of weighted overlap and no overlap almost coincide and give the highest convergence rate. Furthermore, the case of average overlapping produces the worst convergence behavior. The convergence rate using completely overlapping projection vectors do not produce any good results either. The observations in Figure 8 differ completely from the observations in Figure 5. However, it can be seen that the use of weighed

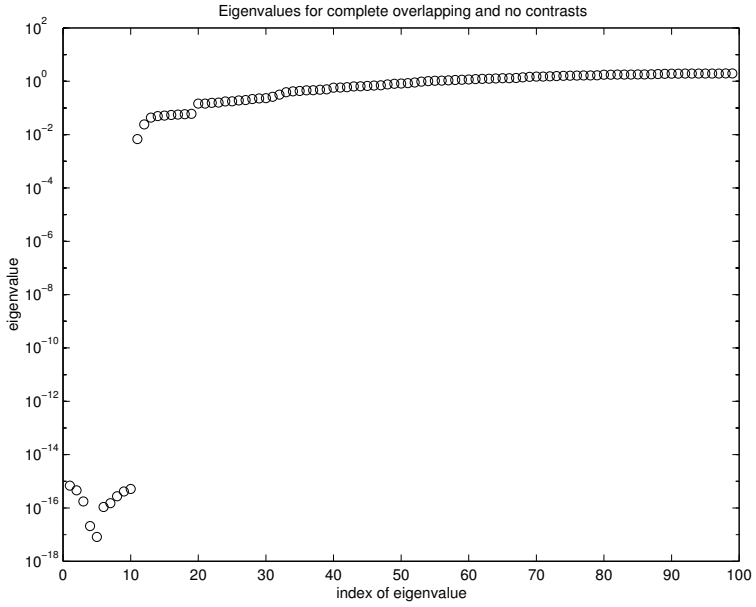

Figure 6: The eigenvalues of the matrix $L^{-T}L^{-1}PA$ for the case of complete overlapping projection vectors and no contrasts for the permeability.

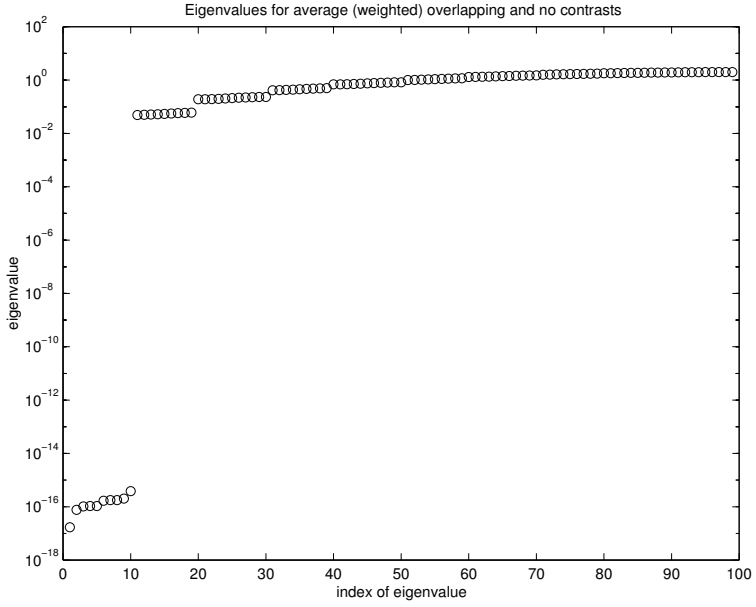

Figure 7: The eigenvalues of the matrix $L^{-T}L^{-1}PA$ for the case of weighted overlapping projection vectors and no contrasts for the permeability.

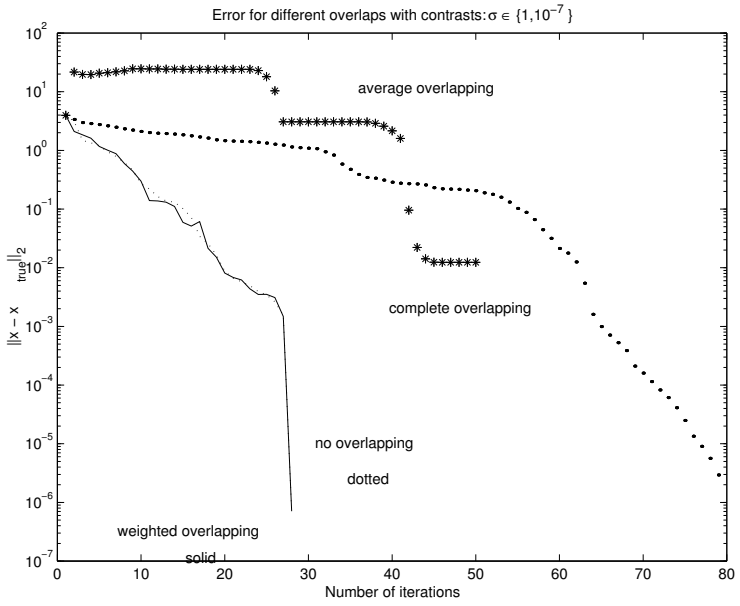

Figure 8: Convergence behavior of the deflated CG-method for the case of permeability contrasts $(10^{-7} - 1)$ and 10 subdomains. Crosses correspond to average overlapping projection vectors. Large dots represent the results for complete overlapping.

overlapping projection vectors produces good results for both cases.

First we show the eigenvalue profile of the case in which we do not apply deflation in Figure 4. It can be seen, consistent to Theorem 1 (Vuik et al. [11]), that four eigenvalues in the order of 10^{-9} persist. This would delay convergence very much. We do not present the results for the residual or error for this case. We computed the eigenvalues of the matrix $L^{-T}L^{-1}PA$ for the cases of complete, average and weighted overlapping projection vectors. The results for complete overlapping are shown in Figure 9. It can be seen that for the case of complete overlapping the *significant* eigenvalues are in the range between about 10^{-3} and 1, which gives an effective condition number $\kappa_{\text{eff}} \approx 9.08 \cdot 10^2$. Whereas for the case of average overlapping still four eigenvalues of the order of 10^{-7} persist, giving an effective condition number $\kappa_{\text{eff}} \approx 2.41 \cdot 10^7$. The slow convergence and *initial* divergence is attributed to these eigenvectors (see Figure 10). Finally we show the profile of the eigenvalues for the case of weighted overlapping in Figure 11. Here it can be seen that the *significant* eigenvalues range between more than 10^{-2} and 1, and the effective condition number $\kappa_{\text{eff}} \approx 4.09 \cdot 10$. Hence the effective condition number is smallest for this case. The large difference between the effective condition numbers from no overlapping / weighted overlap and complete overlap explains the difference in speed of convergence in spite of the absence of significant eigenvalues in the order $10^{-8} - 10^{-6}$. A continuation, not shown in Figure 8, of the calculation for average overlap towards 100 iterations gives four plateaus for the norm of the error. This number corresponds to exactly four small Ritz-values to be found (see Figure 10). We also computed the eigenvalues for the case of no overlapping and the eigenvalues do not differ significantly from those as plotted in Figure 11. This supports the observations made in Figure 8.

Figure 12 shows the results for $\sigma = 1$. Here we vary the number of subdomains (layers) for 99 internal gridpoints. The results are for weighted overlap (hence coincide with average overlap). It can be seen that convergence speeds up as more sub-domains are used. However, the work per iteration increases as well. A similar behavior is observed for the case of complete overlapping, see Figure 13.

In Figure 14 and 15 we show some results for different numbers of subdomains for the case of large contrasts $\sigma \in \{1, 10^{-7}\}$ and respectively weighted ($\approx$ no overlap) and completely overlapping

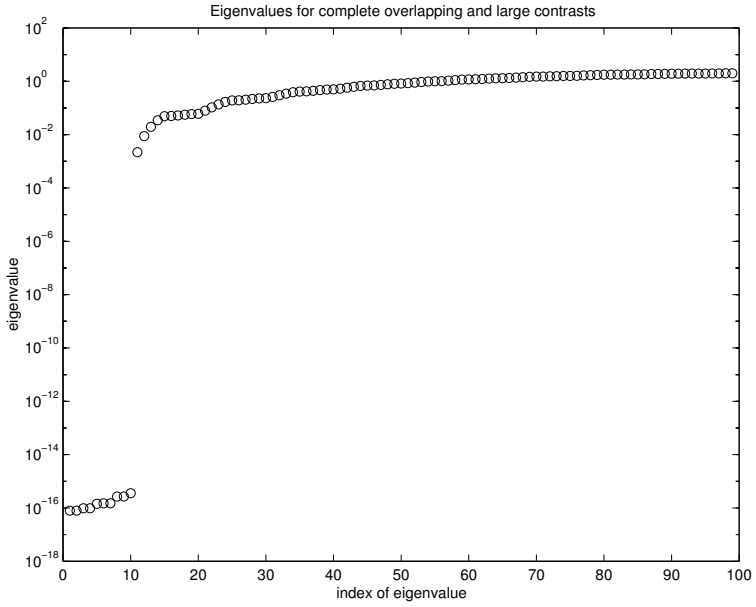

Figure 9: The eigenvalues of the matrix $L^{-T}L^{-1}PA$ for the case of complete overlapping projection vectors and high contrasts for the permeability.

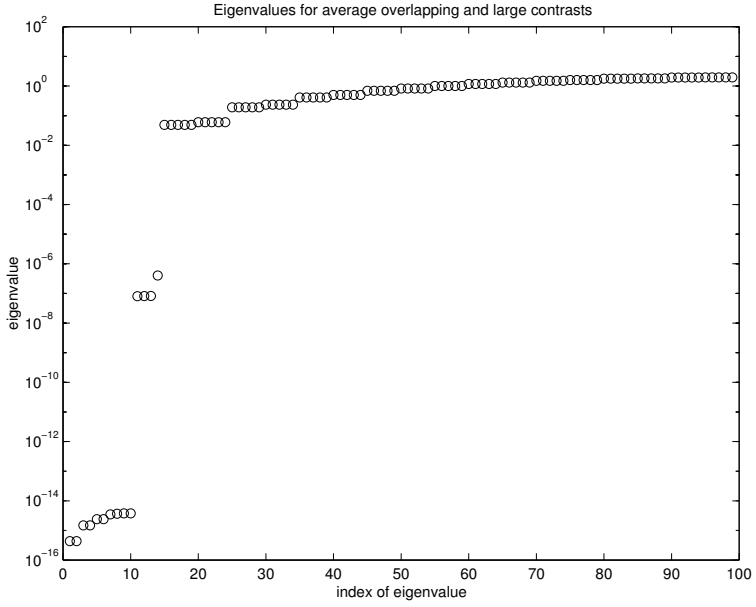

Figure 10: The eigenvalues of the matrix $L^{-T}L^{-1}PA$ for the case of average overlapping projection vectors and high contrasts for the permeability.

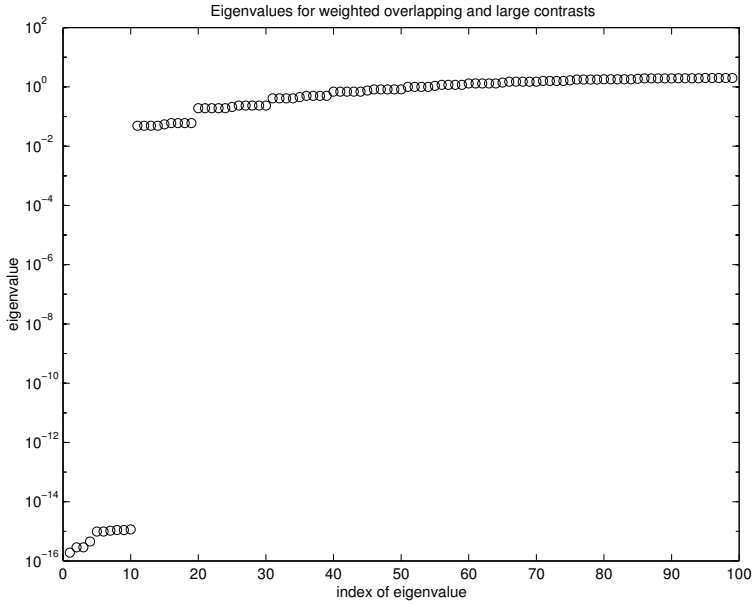

Figure 11: The eigenvalues of the matrix $L^{-T}L^{-1}PA$ for the case of weighted overlapping projection vectors and high contrasts for the permeability.

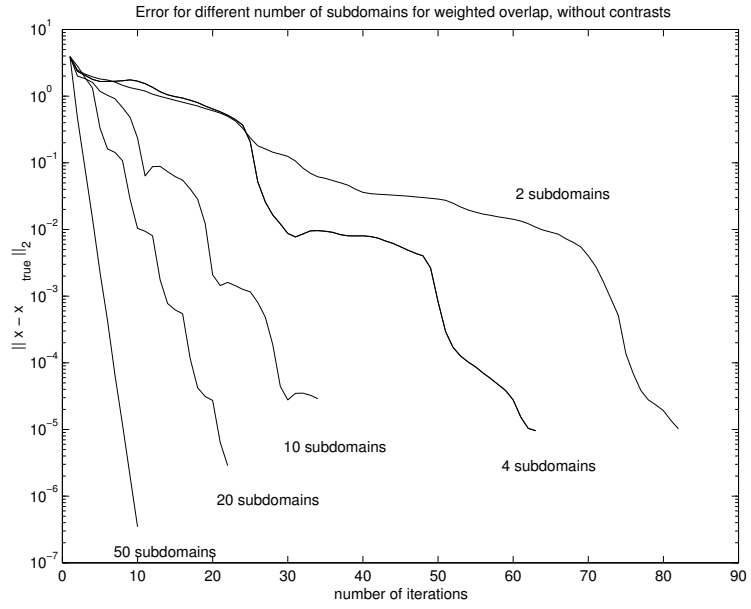

Figure 12: Convergence behavior of the deflated CG-method for $\sigma = 1$. The overlap is weighted (average).

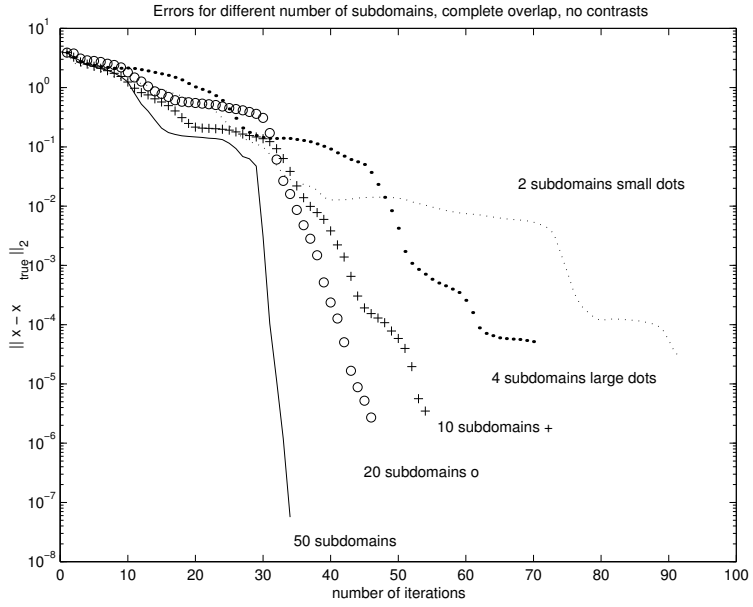

Figure 13: Convergence behavior of the deflated CG-method for $\sigma = 1$. The overlap is complete.

projection vectors. The observations are similar to the preceding curves.

Figure 16 and 17 display results for different contrasts of the permeability. Figures 16 and 17 respectively correspond to complete and weighted overlapping. It can be seen that the convergence rate is very sensitive to the permeability contrasts for the case that complete overlapping projection vectors are used. Higher contrasts delay convergence significantly. Whereas, the dependency of the convergence rate almost vanishes for the case where weighted overlapping is used. Experiments in which the permeability contrast is varied for average overlapping deflation vectors reveal a qualitatively similar and quantitatively stronger behavior as for the case of complete overlap. We show these experiments in Figure 18. Finally we did some experiments for the case of no overlapping projection vectors. The results are shown in Figure 19. We see that the largest convergence speed is reached for high contrasts, although this difference is not large. From an oblivious point of view one would expect that the introduction of larger abrupt contrasts implies that more iterations are necessary for convergence.

3.1.1 Three different permeabilities

The final one-dimensional calculations involve a setting with 10 subdomains. The total number of internal gridpoints is again 99. Instead of two different permeabilities, we have three different permeabilities, such that for $j \in \mathbb{N} : 3j, 3j+1, 3j+2 \leq m$

$$\sigma = \begin{cases} 1, & \text{for } \underline{x} \in \Omega_{3j+1}, \\ 10^{-3}, & \text{for } \underline{x} \in \Omega_{3j+2}, \\ 10^{-7}, & \text{for } \underline{x} \in \Omega_{3j}. \end{cases}$$

Figure 20 shows the convergence of deflated CG-method for the cases of no overlapping, completely overlapping, weighted overlapping and averaged overlapping deflation vectors. It can be seen that the case of weighted overlapping exhibits the fastest convergence. Due to high contrasts this closely resembles the case of no overlapping deflation vectors. Furthermore, the cases of average and complete overlapping deflation vectors give slow convergence.

In Figure 21 the eigenvalues of the matrix $L^{-T}L^{-1}PA$ for the case of no overlapping deflation vectors are shown. The smallest eigenvectors in the range of 10^{-13} or smaller are within the

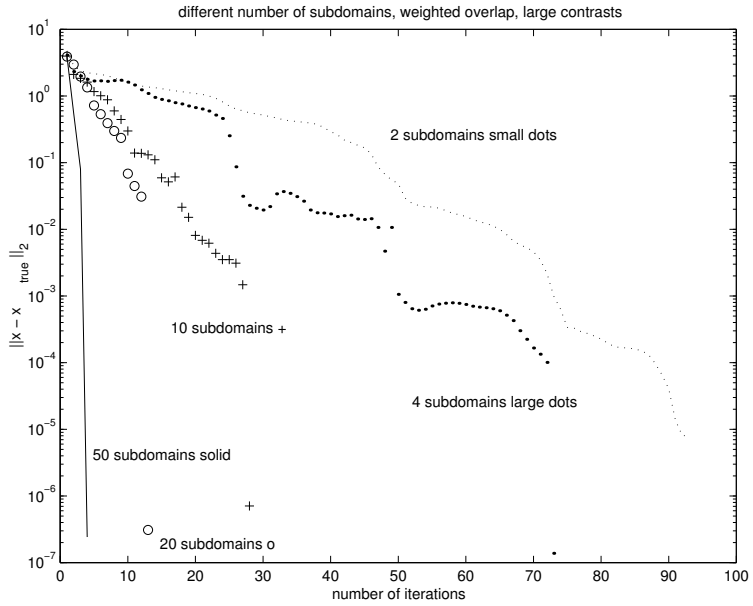

Figure 14: Convergence behavior of the deflated CG-method with permeability contrasts ($10^{-7} - 1$). The overlap is weighted (hence almost equal to no overlapping).

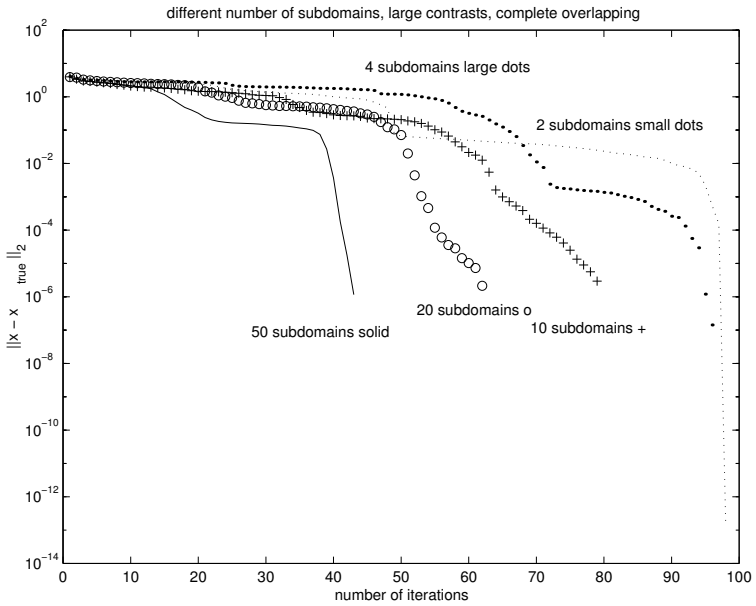

Figure 15: Convergence behavior of the deflated CG-method with permeability contrasts ($10^{-7} - 1$). The overlap is complete.

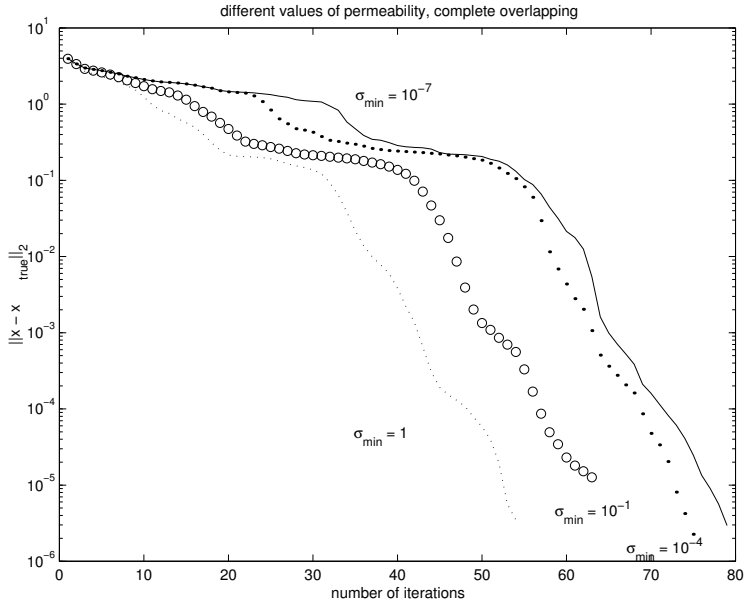

Figure 16: Convergence behavior of the deflated CG-method for different values of $\sigma_{\min}$. The overlap is complete.

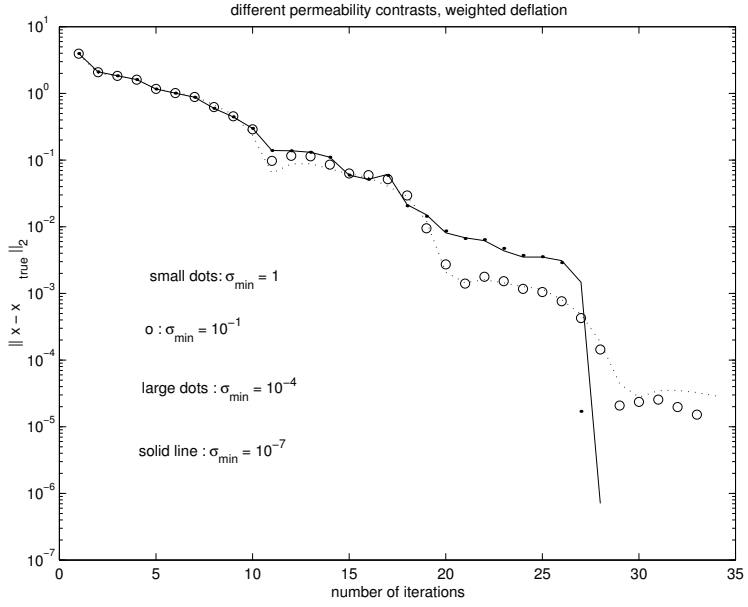

Figure 17: Convergence behavior of the deflated CG-method for different values of $\sigma_{\min}$. The overlap is weighted.

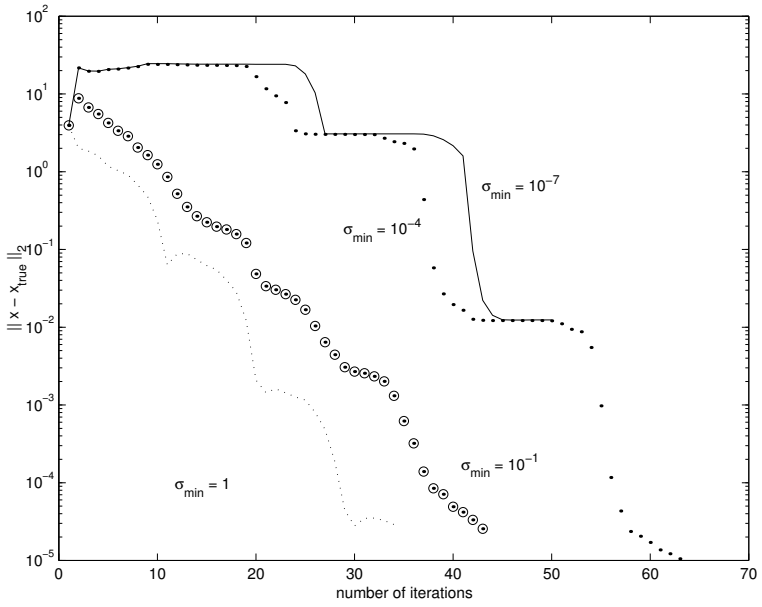

Figure 18: Convergence behavior of the deflated CG-method for different values of $\sigma_{\min}$. The overlap is average.

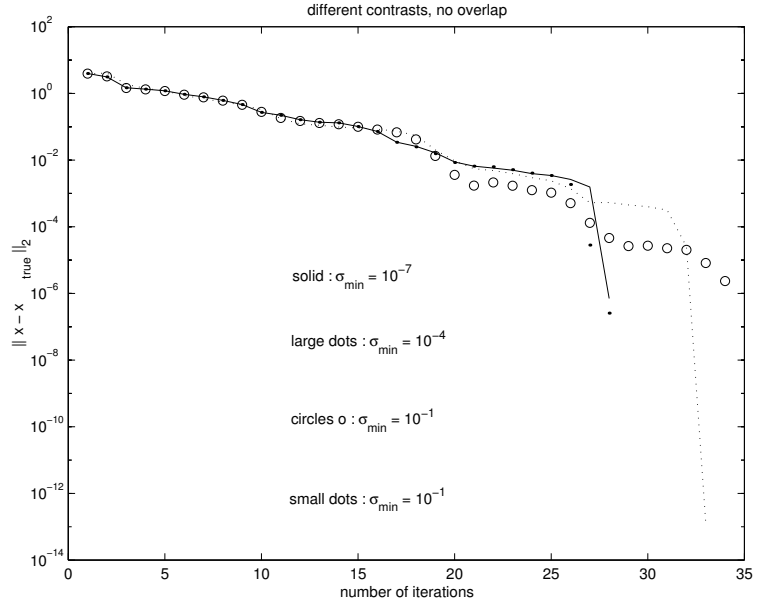

Figure 19: Convergence behavior of the deflated CG-method for different values of $\sigma_{\min}$. There is no overlap.

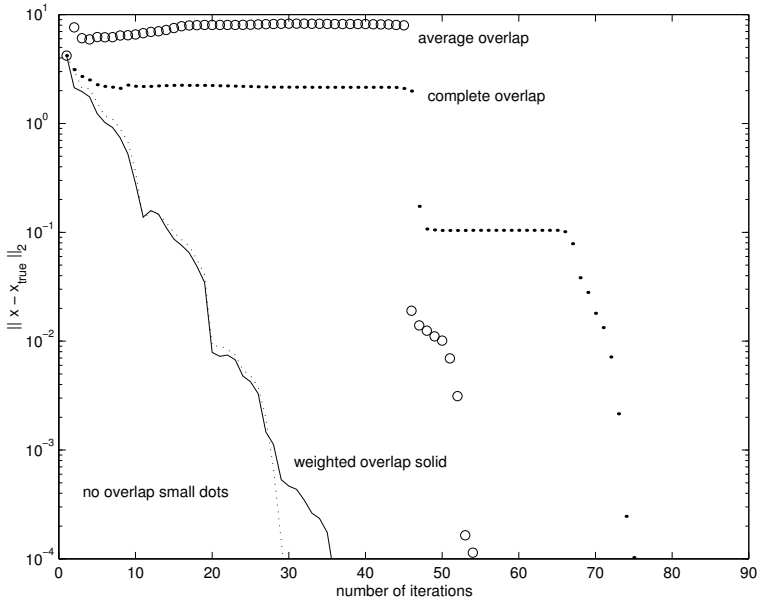

Figure 20: Convergence behavior of the deflated CG-method for different overlapping of deflation vectors.

machine precision and hence we assume that these eigenvectors do not delay convergence, see Figure 20. The eigenvalues for weighted overlapping deflation vectors have been plotted in Figure 22, they do not differ much from the case of no overlapping deflation vectors. The cases of complete and average overlapping deflation vectors are plotted in Figures 23 and 24. Here some eigenvalues are in the order of 10^{-7} and 10^{-5}.

3.2 Two-dimensional experiments

Figure 25 displays the convergence of the deflated ICCG method for 41×41 internal gridnodes and 7 horizontal subdomains. The subdomains have large contrasts for the parameter σ. The calculations have been done with the different overlaps. It can be seen that the data for no overlapping and weighted overlapping almost coincide, as to be expected. Furthermore, average overlapping produces the worst results. This is attributed to the spectrum of the matrix $L^{-T}L^{-1}PA$. Complete overlapping, however, gives convergence. In Figure 9 it can be seen that the eigenvalue pattern does not contains any small eigenvalues for the one-dimensional case. Nevertheless, the speed of convergence is lower than the case of no overlapping and weighted overlapping. The observations are similar to the one-dimensional case.

The results for 41×41 internal gridnodes and 7 horizontal subdomains and absence of large contrasts in the parameter α have been shown in Figure 26. The calculations have been done for different overlaps. Now, the results for weighted overlap and average overlap coincide. These results do not differ much from the results obtained by no overlap. The results for complete overlap are worst but converge and here the difference is not as striking as for the case where we have large contrasts in the parameter σ.

We varied the number of subdomains for a grid of 41×41 internal gridnodes. The results are shown in Figures 27 and 28 for high and no permeability contrasts. It can be seen that the speed of convergence is less sensitive to the number of subdomains than for the one-dimensional case.

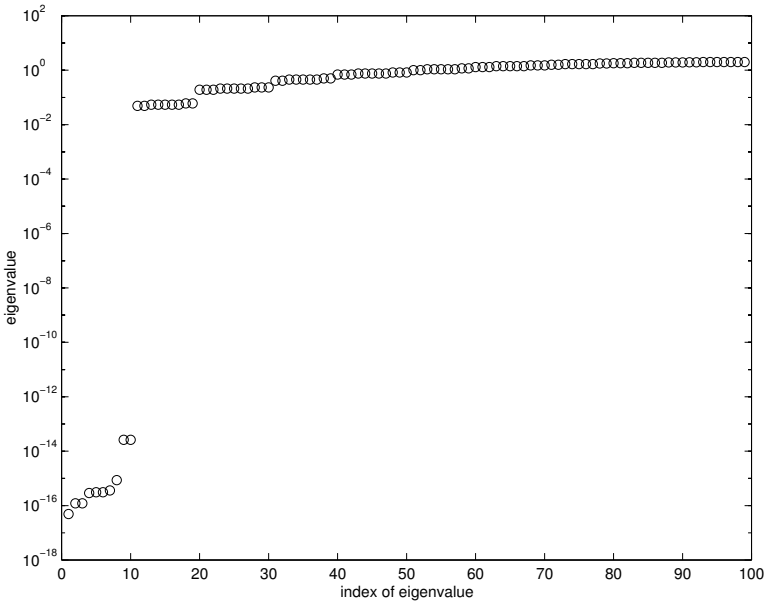

Figure 21: The eigenvalues of the matrix $L^{-T}L^{-1}PA$ for the case of no overlapping deflation vectors.

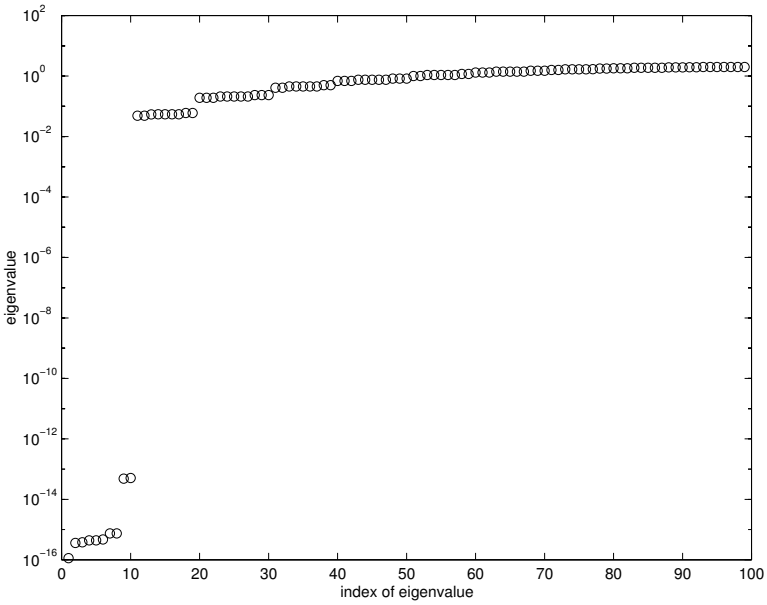

Figure 22: The eigenvalues of the matrix $L^{-T}L^{-1}PA$ for the case of weighted overlapping deflation vectors.

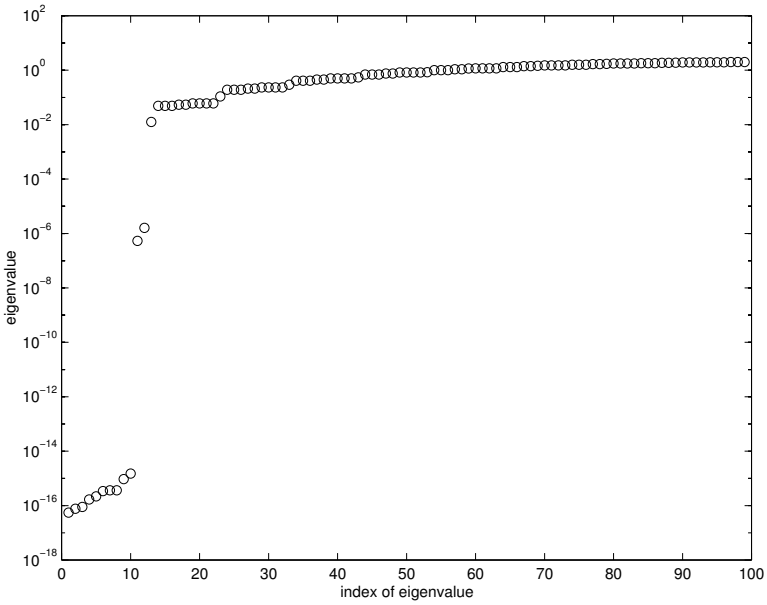

Figure 23: The eigenvalues of the matrix $L^{-T}L^{-1}PA$ for the case of completely overlapping deflation vectors.

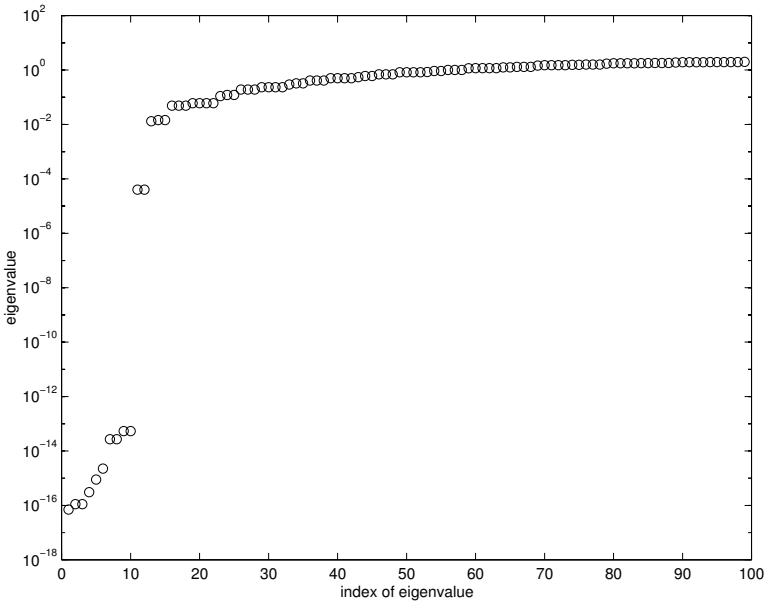

Figure 24: The eigenvalues of the matrix $L^{-T}L^{-1}PA$ for the case of average overlapping deflation vectors.

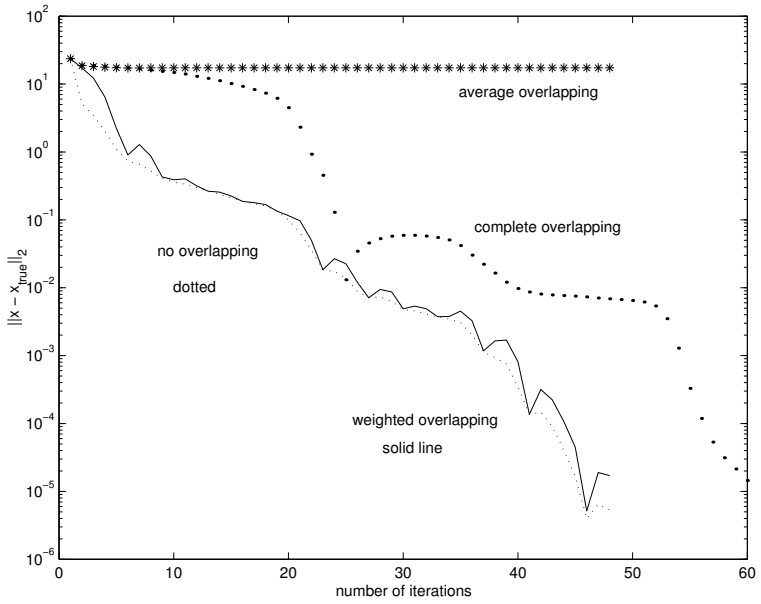

Figure 25: Convergence of the Deflated ICCG method for 41 x 41 internal nodes and high contrasts ($\sigma_{min} = 10^{-7}, \sigma_{max} = 1$). The computations are done with average, weighted, complete overlap and without overlap.

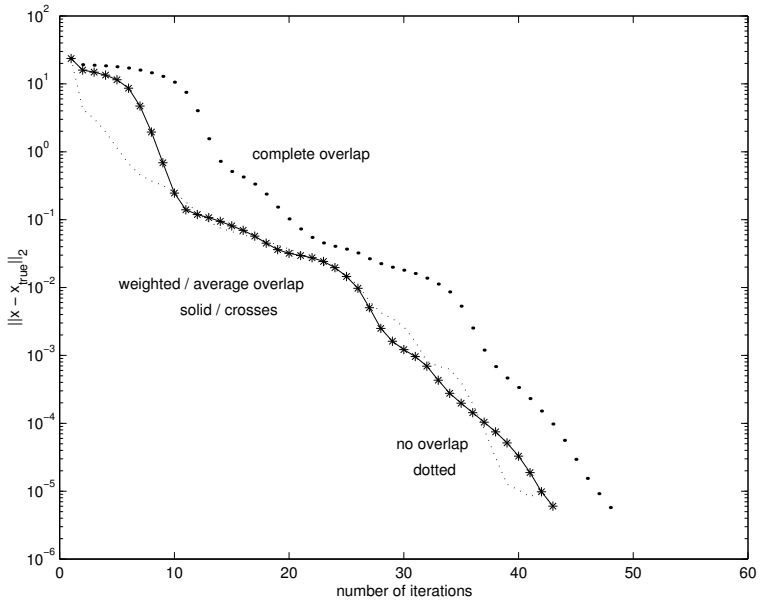

Figure 26: Convergence of the Deflated ICCG method for 41 x 41 internal nodes and no contrasts ($\sigma_{min} = 1, \sigma_{max} = 1$). The computations are done with average, weighted, complete overlap and without overlap.

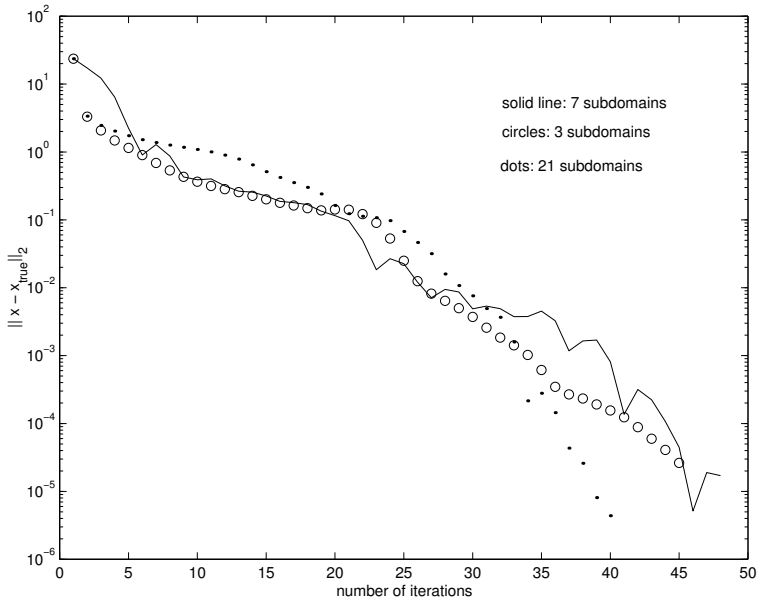

Figure 27: Convergence of the Deflated ICCG method for 41 x 41 internal nodes and high contrasts ($\sigma_{min} = 10^{-7}, \sigma_{max} = 1$). The computations are done with weighted (no) overlap. The number of subdomains is varied: 3, 7 and 21.

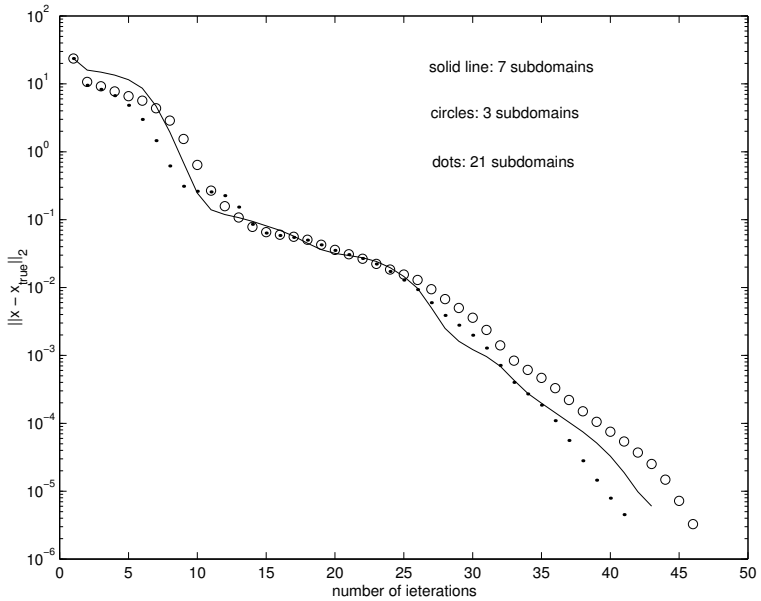

Figure 28: Convergence of the Deflated ICCG method for 41 x 41 internal nodes and no contrasts ($\sigma_{min} = 1, \sigma_{max} = 1$). The computations are done with weighted (average) overlap. The number of subdomains is varied: 3, 7 and 21.

4 Conclusions

We investigated various choices of deflation vectors, which are used in the Deflated CG method. It is found that the choice of the deflation vectors at the interfaces plays a crucial role in the convergence rate. Summarized, the following is concluded so far:

- As the domain is divided into more subdomains, the number of iterations needed for convergence decreases. This effect has been observed for $\sigma = 1$. Furthermore, this observation does not depend on the choice of the values of the deflation vectors at the boundaries. A drawback in sequential calculations is that more subdomains (more deflation vectors) give larger computation times per iteration. However, for parallel computing wall clock time can be gained.
- We introduce the method of 'weighted overlap', which mimics average and no overlap for respectively the cases of no contrasts and very large contrasts of the permeability. It is observed that this choice gives the best convergence behavior until now.

References

[1] J. Bear. *Dynamics of Fluids in Porous Media*. Elsevier, New York, 1972.

[2] H. De Gersem and K. Hameyer. A deflated iterative solver for magnetostatic finite element models with large differences in permeability. *Eur. Phys. J. Appl. Phys.*, 13:45–49, 2000.

[3] J. Frank and C. Vuik. On the construction of deflation-based preconditioners. MAS-R 0009, CWI, Amsterdam, 2000. to appear in SIAM J. Sci. Comput.

[4] C. B. Jenssen and P. Å. Weinerfelt. Coarse grid correction scheme for implicit multiblock Euler calculations. *AIAA Journal*, 33(10):1816–1821, 1995.

[5] E. Kreyszig. *Introductory functional analysis with applications*. Wiley, New-York, 1989.

[6] L.W. Lake. *Enhanced Oil Recovery*. Prentice-Hall, Englewood Cliffs, 1989.

[7] D.C. Lay. *Linear algebra and its applications*. Addison-Wesley, Longman Scientific, Reading, Massachusetts, 1996.

[8] A. Padiy, O. Axelsson, and B. Polman. Generalized augmented matrix preconditioning approach and its application to iterative solution of ill-conditioned algebraic systems. *SIAM J. Matrix Anal. Appl.*, 22:793–818, 2000.

[9] E. Perchat, L. Fourment, and T. Coupez. Parallel incomplete factorisations for generalised Stokes problems: application to hot metal forging simulation. Report, EPFL, Lausanne, 2001.

[10] C. Vuik, A. Segal, L. el Yaakoubi, and E. Dufour. A comparison of various deflation vectors applied to elliptic problems with discontinuous coefficients. Report of the department applied mathematical analysis 01-03, Delft University of Technology, Delft, 2001.

[11] C. Vuik, A. Segal, and J.A. Meijerink. An efficient preconditioned CG method for the solution of a class of layered problems with extreme contrasts in the coefficients. *J. Comp. Phys.*, 152:385–403, 1999.

[12] C. Vuik, A. Segal, J.A. Meijerink, and G.T. Wijma. The construction of projection vectors for a Deflated ICCG method applied to problems with extreme contrasts in the coefficients. *Journal of Computational Physics*, to appear, 2001.