Deflation in Preconditioned Conjugate Gradient Methods for Finite Element Problems

Fred Vermolen,[1] Kees Vuik,[2] Guus Segal[3]

[1]Department of Applied Mathematical Analysis, Delft University of Technology, Mekelweg 4, 2628 CD Delft, Netherlands F.J.Vermolen@its.tudelft.nl

[2]Department of Applied Mathematical Analysis, Delft University of Technology, Mekelweg 4, 2628 CD Delft, Netherlands C.Vuik@its.tudelft.nl

[3]Department of Applied Mathematical Analysis, Delft University of Technology, Mekelweg 4, 2628 CD Delft, Netherlands g.segal@math.tudelft.nl

1 Introduction

Large linear systems are solved for modelling 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 problems. 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.

In this paper we only consider symmetric positive definite (SPD) matrices. Presently, direct methods (such as an LU-decomposition) and iterative methods are available to solve such a linear system. However, for large sparse coefficient matrices fill-in causes a loss of efficiency (in computer memory and number of floating point operations). For such a case iterative methods are a better alternative. Furthermore, if 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 these problems, the conjugate gradient method is used. We deal with an application from transport in porous media where we encounter extreme contrasts in the coefficients of the partial differential equation. The large contrasts are caused by the layered domain where neighbouring layers in Fig. 1 extreme contrasts in mobility. Here a preconditioning is necessary and we use a standard incomplete Cholesky factorization as a preconditioner

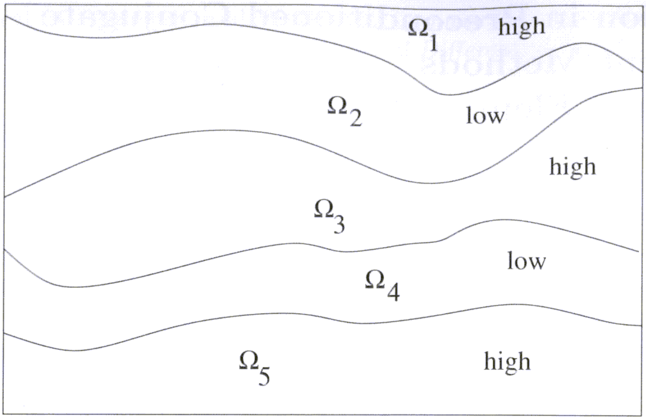

Figure 1. An example of the geometry of the domain with high and low permeability regions.

for the conjugate gradient method (ICCG) to improve convergence behavior. Furthermore, deflation is applied to get rid of remaining extremely small eigenvalues that delay convergence. Vuik et al. [17] proposed a scheme based on physical deflation in which the deflation vectors are continuous and satisfy the original partial differential equation on a subdomain. A different variant involves the so-called algebraic deflation with discontinuous deflation vectors. Here convergence is speeded up. This was also subject of [16]. In that paper a comparison is given between physical deflation vectors and algebraic deflation vectors. Two choices 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 [17] and [16]. The DICCG method has already been successfully used for complicated magnetic field simulations [2]. A related method is recently presented in [11]. In [4] 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 [6], [11]. Therefore insight in a good choice of the deflation vectors can probably be used to devise comparable strategies for coarse grid correction approaches.

We assume that the domain Ω consists of a number of disjoint sets Ω_j, $j = 1, \ldots, m$, such that $\bar{\Omega} = \bigcup_{j=1}^{m} \bar{\Omega}_j$. 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.

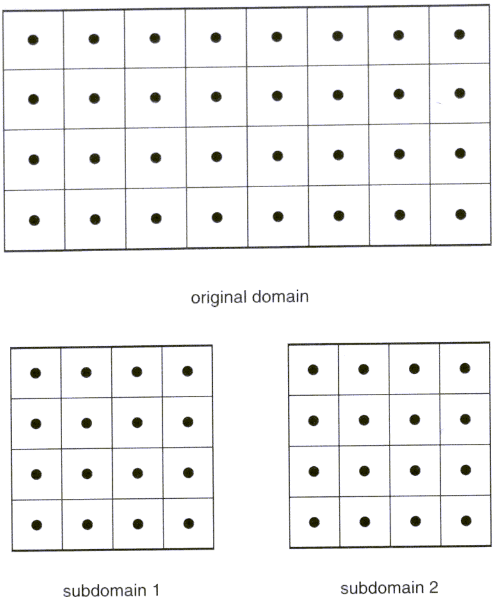

Figure 2. Domain decomposition for a cell centered discretization.

Cell centered

For this discretization the unknowns are located in the interior of the finite volume. The domain decomposition is straightforward as can be seen in Fig. 2. The algebraic deflation vectors z_j are uniquely defined as:

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

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 [12]:

- 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.

For Finite Elements only the last option is commonly used. Note that the vertex oriented decomposition is not well suited to combine with a finite element method. Therefore, we restrict ourselves to the element oriented decomposition, see Fig. 3. As a consequence of this the deflation vectors can overlap at interfaces.

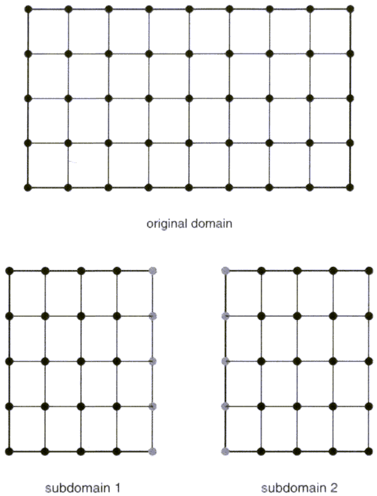

Figure 3. Domain decomposition for a vertex centered discretization. The grey nides are the interface nodes.

In our previous work we always use non-overlapping deflation vectors. In [17], [18] the interface vertices are only elements of the high permeability subdomains, whereas in [4] 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.

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:

$$\left.\begin{aligned} &\frac{\partial S}{\partial t} + \nabla \cdot (\boldsymbol{q}S) = \nabla \cdot (D(S)\nabla S), \\ &\nabla \cdot \boldsymbol{q} = 0, \\ &\frac{1}{\sigma}\boldsymbol{q} + \nabla p - gS\,\boldsymbol{e}_y = \boldsymbol{0}. \end{aligned}\right\} \qquad (\mathrm{P}_0)$$

The above equations are supplemented with appropriate initial and boundary conditions. In above equations S (-), $\boldsymbol{q}$ (m/s) and p (Pa) are the unknown saturation, discharge and pressure respectively. The unit vector in the y-direction is represented by $\boldsymbol{e}_y$ and the contstant of gravity is denoted by g. 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 [9].

For the 2-dimensional case it is favourable [3], [13] to introduce the stream function $\boldsymbol{\psi}$ such that $\boldsymbol{q} = \nabla \times \boldsymbol{\psi}$. Since here ∇ and $\boldsymbol{q}$ respectively work and are given for the (x, y) plane only, it follows that $\boldsymbol{\psi}$ only has a non-constant z-component, i.e. $\boldsymbol{\psi} = \langle 0, 0, \psi \rangle$. For more mathematical details on the existence of such a stream function, we refer to the book of Temam [14]. Further, $\partial_z(\cdot) = 0$ and hence after taking the curl over the third equation of (P_0) we are faced with

$$-\nabla \cdot \left(\frac{1}{\sigma} \nabla \psi\right) = g \frac{\partial S}{\partial y}.$$

If one imposes no-flow conditions over the boundary of Ω, then it follows that

$$\psi = 0 \text{ on } \Gamma,$$

where Γ represents the boundary of Ω. This implies that the equations in (P_0) change into

$$\left.\begin{aligned} \frac{\partial S}{\partial t} + \left\langle -\frac{\partial \psi}{\partial y}, \frac{\partial \psi}{\partial x} \right\rangle \cdot \nabla S &= \nabla \cdot (D(S) \nabla S), \\ -\nabla \cdot \left(\frac{1}{\sigma} \nabla \psi\right) &= g \frac{\partial S}{\partial y}, \end{aligned}\right\} \qquad (\mathrm{P}_1)$$

for $(x, y) \in \Omega$. We focus on the solution of the second equation of (P_1) by a Finite Element Method, and hence consider a variational formulation:

$$\left.\begin{aligned} &\text{Find } \psi \in H_0^1(\Omega) \ \ (\psi|_\Gamma = 0) \text{ such that} \\ &\int_\Omega \nabla v \ \cdot \frac{1}{\sigma} \nabla \psi \, \mathrm{d}A \ = \int_\Omega g \frac{\partial S}{\partial y} v \ \mathrm{d}A \ \text{ for all } v \ \in H_0^1(\Omega). \end{aligned}\right\} \qquad (\mathrm{P}_2)$$

In problem (P_2) σ is allowed to be piecewise continuous and hence, the solution ψ is only piecewise smooth. This problem is solved by the use of a standard Galerkin Finite Element Method, with $\psi = \sum_{i=1}^n \psi_i v_i$ $(v_i|_\Gamma = 0)$, with piecewise linear element functions v_i. In our examples we take a layered domain with

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

where we suppose that the closed domain $\overline{\Omega}$ consists of the union of m closed subdomains $\overline{\Omega}_1, \dots, \overline{\Omega}_m$ (see Fig. 1). In some applications, the high and low mobility (σ) respectively correspond to sand and shale layers. We will also use this terminology to refer to the high and low permeability layers. From the Galerkin discretization it follows inmediately that accross an interface the coefficients in the discrete equation varies several orders of magnitude.

Discretization by the use of Galerkin's method results into a matrix-vector equation of type

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

where $A \in \mathbb{R}^{n\times n}$, $\underline{x} \in \mathbb{R}^n$, and $\underline{b} \in \mathbb{R}^n$, respectively, represent the discretization (or stiffness) matrix, solution vector, and right-hand side vector. Using a FEM approach the discretization matrix is sparse, symmetric and positive definite (SPD). Furthermore, the discretization is chosen such that the interfaces between consecutive layers coincide with gridpoints. For the case of large jumps in the coefficient σ the condition of the discretization-matrix is very large. The remainder of the paper is devoted to the efficient solution of the above matrix-vector equation when n is large.

3 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. After k iterations the $\|\cdot\|_A$-norm ($\|\cdot\|_A := \sqrt{(\cdot, A.)}$) of the error is bounded from above by the well-known result of Luenberger, which can also be found in [5]

$$\|\underline{x} - \underline{x}_k\|_A \le 2\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^k \|\underline{x} - \underline{x}_0\|_A, \tag{1}$$

where κ denotes the condition number of the matrix A and $\underline{x}$ is the exact solution of the system. Further, in the above expression, $\underline{x}_0$ and $\underline{x}_k$, respectively, represent the initial estimate of the solution $\underline{x}$ and the result after k conjugate gradient iterations. For the $\|\cdot\|_2$-norm ($\|\cdot\|_2 := \sqrt{(\cdot,\cdot)}$), it can be proven that

$$\sqrt{\lambda_{\min}}\|\cdot\|_2 \le \|\cdot\|_A \le \sqrt{\lambda_{\max}}\|\cdot\|_2. \tag{2}$$

The above inequalities are combined with expression (1) to obtain

$$\|\underline{x} - \underline{x}_k\|_2 \le 2\sqrt{\kappa}\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^k \|\underline{x} - \underline{x}_0\|_2. \tag{3}$$

This estimate is standard and can be found in the book of Golub and van Loan [5]. It gives an upper bound for the error for the $\|\cdot\|_2$-norm.

Vuik et al. [17] 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 Cholesky decomposition 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. This observation is formulated in terms of the following theorem, which is proven by Vuik et al. [17]:

Theorem 1. *Let* $\varepsilon := \dfrac{\sigma_{\min}}{\sigma_{\max}}$ *be small enough,* $D = \text{diag}(A)$ *and let* r *be the number of layers with* σ *of order one between low* σ *layers. Then the diagonally scaled matrix* $D^{-1/2}A\ D^{-1/2}$ *has exactly* r *eigenvalues of order* ε.

We remark here that Theorem 1 is extended to an incomplete Cholesky preconditioning, see Vuik et al [18]. The preconditioning aims at improving the condition of the matrix. However, for this case r small eigenvalues persist. In experiments we use an incomplete Cholesky decomposition as a preconditioner for the symmetric positive definite discretization matrix.

In the next section we consider the DICCG-method as proposed by Vuik et al. [17] for the Laplace problem with extreme contrasts of the coefficients. The aim is to get rid of the remaining very small eigenvalues of the preconditioned matrix $\tilde{A} = L^{-T}L^{-1}A$, where $L^{-T}L^{-1} \approx A^{-1}$ is the IC-preconditioner.

3.1 Deflation

In this subsection we analyze the elimination of the small eigenvalues of A by deflation. Therefore, we first prove that, if the deflation matrix is constructed by eigenvectors of A, then its corresponding eigenvalues are transformed into zero for the product of the deflation matrix and the discretization matrix A. Here we need properties like symmetry and that the deflation matrix is a projection. These properties are proven first. Suppose that λ_i and $\vec{z}_i$, $i \in \{1, \ldots, n\}$, respectively represent eigenvalues and orthonormal eigenvectors of the symmetric discretization matrix $A \in \mathbb{R}^{n \times n}$ (such that $\vec{z}_i^T \vec{z}_j = \delta_{ij}$), and define the matrix $\overline{P} \in \mathbb{R}^{n \times n}$ by

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

Then

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

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

Hence, we state the following result:

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

1. *$\overline{P}A$, for $j > m$, and A have the same eigenvalues λ_j, and the corresponding eigenvalues of A given by λ_j for $j \le m$ 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 Theorem 2. To prove the second statement, we compute

$$\begin{aligned}\overline{P}^2 &= \left(I - \sum_{i=1}^{m} \vec{z}_i\vec{z}_i^{\,T}\right)\left(I - \sum_{i=1}^{m} \vec{z}_i\vec{z}_i^{\,T}\right)\\ &= I - 2\sum_{i=1}^{m} \vec{z}_i\vec{z}_i^{\,T} + \left(\sum_{i=1}^{m} \vec{z}_i\vec{z}_i^{\,T}\right)\left(\sum_{i=1}^{m} \vec{z}_i\vec{z}_i^{\,T}\right)\\ &= I - \sum_{i=1}^{m} \vec{z}_i\vec{z}_i^{\,T} = \overline{P}.\end{aligned}$$

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

From Theorem 2 it follows that $\overline{P}$ and $\overline{P}A$ are singular. Now we introduce the matrix P for a general choice of deflation vectors:

Definition 1. The deflation matrix P is defined by

$$P := I - AZ(Z^TAZ)^{-1}Z^T.$$

We show that P can be written as $\overline{P}$ when the columns of Z are eigenvectors of A.

Theorem 3. *Let $Z \in \mathbb{R}^{n\times m}$: $Z = (\vec{z}_1,\dots,\vec{z}_m)$ where $\vec{z}_i$ are the orthonormal eigenvectors with eigenvectors λ_i of the matrix A and let P be defined as in Definition 1. Then $P = \overline{P}$. Moreover, P is a projection for all choices of $Z \in \mathbb{R}^{n\times m}$.*

Proof. If $P = I - A\,Z\,(Z^T A\,Z)^{-1} Z^T$, then

$$\begin{aligned} P &= I - A\,(\vec{z}_1, \dots, \vec{z}_m) \left(\begin{pmatrix} \vec{z}_1^T \\ \vdots \\ \vec{z}_m^T \end{pmatrix} A\,(\vec{z}_1, \dots, \vec{z}_m) \right)^{-1} \begin{pmatrix} \vec{z}_1^T \\ \vdots \\ \vec{z}_m^T \end{pmatrix} \\ &= I - (A\,\vec{z}_1, \dots, A\,\vec{z}_m) \begin{pmatrix} \vec{z}_1^T A\,\vec{z}_1 & \dots & \vec{z}_1^T A\,\vec{z}_m \\ \vdots & \vdots & \vdots \\ \vec{z}_m^T A\,\vec{z}_1 & \dots & \vec{z}_m^T A\,\vec{z}_m \end{pmatrix}^{-1} \begin{pmatrix} \vec{z}_1^T \\ \vdots \\ \vec{z}_m^T \end{pmatrix} \\ &= I - (\lambda_1 \vec{z}_1, \dots, \lambda_m \vec{z}_m)\,\mathrm{diag}\Big(\frac{1}{\lambda_1}, \dots, \frac{1}{\lambda_m}\Big) \begin{pmatrix} \vec{z}_1^T \\ \vdots \\ \vec{z}_m^T \end{pmatrix} \\ &= I - \sum_{i=1}^{m} \vec{z}_i \vec{z}_i^T = \overline{P}. \end{aligned}$$

This proves the first statement. The second statement is proven by direct multiplication:

$$P^2 = I - 2AZ(Z^T A\,Z)^{-1} Z^T + AZ(Z^T AZ)^{-1} Z^T AZ(Z^T AZ)^{-1} Z^T = P.$$

Hence P is a projection. □

Corollary 1. *The matrix PA is symmetric positive semi-definite.*

Proof. Since $A^T = A$, the symmetry of PA is established by

$$(PA)^T = (A - AZ(Z^T A\,Z)^{-1} Z^T A)^T = A - AZ(Z^T AZ)^{-1} Z^T A = PA.$$

Further, P is a projection and A is symmetric positive definite; hence from Lemma 2.1 by Frank and Vuik [4] it follows that PA is positive semi-definite. □

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

$$PA\,Z = A\,Z - A\,Z\,(Z^T A\,Z)^{-1} Z^T A\,Z = A\,Z - A\,Z = 0.$$

This is also shown in Theorem 4.1 in Vuik et al [17]. Consider a matrix A with eigenvalues $\{\lambda_1, \dots, \lambda_m, \lambda_{m+1}, \dots, \lambda_n\}$ and let $Z = (\vec{z}_1, \dots, \vec{z}_m)$ where $A\vec{z}_i = \lambda_i \vec{z}_i$, $i \in \{1, \dots, m\}$ are eigenvectors of A that correspond to eigenvalues $\{\lambda_1, \dots, \lambda_m\}$. Then with P, according to Definition 1, it follows that PA has eigenvalues $\{0, \dots, 0, \lambda_{m+1}, \dots, \lambda_n\}$.

Theorem 4. *Let $P \in \mathbb{R}^{n \times n}$ be defined as in Definition* 1 *with $Z = (\vec{z}_1, \ldots, \vec{z}_m)$. Then the null-space of P is spanned by the independent set $\{A\ \vec{z}_1, \ldots, A\ \vec{z}_m\}$, i.e.* $\text{null}(P) = \text{Span}\{A\vec{z}_1, \ldots, A\ \vec{z}_m\}$.

Proof. Since $PAZ = 0$, it follows that $A\vec{z}_i \in \text{null}\, P$ and hence $\dim \text{null}\, P \geq m$. If $V := \text{Span}\{\vec{z}_1, \ldots, \vec{z}_m\}$, then from the Direct Sum Theorem (see for instance [8]) it follows $\mathbb{R}^n = V\ \oplus V^{\perp}$, where $V^{\perp} = \{\vec{y} \in \mathbb{R}^n : \vec{y} \perp V\}$. Hence, $\dim V^{\perp} = n - m$. Suppose $\vec{y} \in V^{\perp}$, then

$$P\vec{y} = \vec{y} - AZ(Z^T A\ Z)^{-1} Z^T \vec{y} = \vec{y}.$$

Hence, $\dim \text{col}\, P \geq n - m$. Since $\dim \text{null}\, P + \dim \text{col}\, P = n$, this implies

$$\dim \text{null}\, P \leq m.$$

Consequently, using $\dim \text{null}\, P \geq m$, we have $\dim \text{null}\, P = m$. Since $\{A\ \vec{z}_1, \ldots, A\vec{z}_m\}$ represents a linearly independent set of m vectors in $\text{null}\, P$ and $\dim \text{null}\, P = n$, it follows from the Basis-Theorem (see for instance [10]) that

$$\text{null}\, P = \text{Span}\{A\vec{z}_1, \ldots, A\vec{z}_m\}.$$

This proves the theorem. □

The matrix P is referred to as the deflation matrix. The vectors $\vec{z}_1, \ldots, \vec{z}_m$ 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 PA which do not influence the convergence of the CG-method.

3.2 Deflated incomplete Cholesky preconditioned conjugate gradients

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

$$PA\ \vec{\tilde{x}} = P\vec{b}, \tag{4}$$

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

$$\vec{x} = (I\ - P^T)\vec{x} + P^T\vec{x}. \tag{5}$$

Note that $\vec{\tilde{x}}$ is the solution of equation (4). In this paragraph we first establish uniqueness of the above $P^T\vec{\tilde{x}}$, given any $\vec{\tilde{x}}$ that satisfies equation (4). Therefore, we first need to establish that for the projected solution we have $P^T\vec{x} = P^T\vec{\tilde{x}}$ and that hence $P^T\vec{\tilde{x}}$ is unique. Equation (4) is written as

$P(A\,\vec{\tilde{x}} - \vec{b}) = \vec{0}$, where $A\vec{\tilde{x}} - \vec{b} \in \operatorname{null} P$. Since $PA\,Z\ = 0$, the vectors $A\vec{z}_i$ are in the null-space of P, i.e. $A\vec{z}_i \in \operatorname{null} P$. Since, we know from Theorem 4 that $\operatorname{null} P = \operatorname{Span}\{A\vec{z}_1, \dots, A\ \vec{z}_m\}$, the vectors in the null-space of P can be written as linear combinations $\vec{w} = \sum_{i=1}^{m} \alpha_i A\ \vec{z}_i$. Therefore, it can be seen that one can write for the vector $\vec{\tilde{x}}$:

$$\vec{\tilde{x}} = A^{-1}\vec{b} + A^{-1}\vec{w} = A^{-1}\vec{b} + \sum_{i=1}^{m} \alpha_i \vec{z}_i. \tag{6}$$

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 $P^T Z$:

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

Hence, this product is zero and the second term of the right-hand side of equation (6) vanishes after multiplication with P^T, since $Z = (\vec{z}_1, \dots, \vec{z}_m)$. Hence, the vector $P^T\vec{\tilde{x}}$ is unique and we have

$$P^T\vec{\tilde{x}} = P^T A^{-1}\vec{b} = P^T\vec{x}.$$

For the first term of the right-hand side of equation (5) we note that

$$(I\ - P^T)\vec{x} = Z(Z^T AZ)^{-1} Z A\vec{x} = Z(Z^T AZ)^{-1} Z\vec{b}.$$

Hence, this term is also uniquely determined and since the dimension of the matrix $Z^T AZ$ is small, the computation of this term is relatively cheap. This is summarized in the next theorem:

Theorem 5. *Let P be defined as in Definition* 1 *and $\vec{\tilde{x}}$ as in equation* (4)*. Then*

1. *$P^T\vec{\tilde{x}}$ is unique and $P^T\vec{\tilde{x}} = P^T\vec{x}$,*
2. *the unique solution of $A\vec{x} = \vec{b}$ can be written by*

$$\vec{x} = Z(Z^T AZ)^{-1} Z^T\vec{b} + P^T\vec{\tilde{x}},$$

where $\vec{\tilde{x}}$ is a solution of equation (4).

We further note that $(I - P^T)Z = Z(Z^T AZ)^{-1} Z^T AZ = Z$. Hence,

$$(I - P^T)(\vec{z}_1, \dots, \vec{z}_m) = (\vec{z}_1, \dots, \vec{z}_m).$$

From this it follows that $\operatorname{Span}\{\vec{z}_1, \dots, \vec{z}_m\} =: V\ \subset \mathbb{R}^n$ is in the eigenspace of eigenvalue $\lambda = 1$ of the matrix $(I - P^T)$.

For completeness, we present the algorithm of the deflated ICCG.

Algorithm 1 (DICCG [17]):

```
$k = 0$, $\vec{\tilde{r}}_0 = P\vec{r}_0$, $\vec{p}_1 = \vec{z}_0 = L^{-T}L^{-1}\vec{\tilde{r}}_0$
**while** $\|\vec{\tilde{r}}_k\|_2 > \varepsilon$
    $k = k + 1$
    $\alpha_k = \dfrac{\vec{\tilde{r}}_{k-1}^T \vec{z}_{k-1}}{\vec{p}_k^T PA\, \vec{p}_k}$
    $\vec{x}_k = \vec{x}_{k-1} + \alpha \vec{p}_k$
    $\vec{\tilde{r}}_k = \vec{\tilde{r}}_{k-1} - \alpha_k PA\, \vec{p}_k$
    $\vec{z}_k = L^{-T}L^{-1}\vec{\tilde{r}}_k$
    $\beta_k = \dfrac{\vec{\tilde{r}}_k^T\ \vec{z}_k}{\vec{\tilde{r}}_{k-1}^T\ \vec{z}_{k-1}}$
    $\vec{p}_{k-1} = \vec{z}_k + \beta_k \vec{p}_k$
**end while**
```

The conjugate gradient method is reported to converge for symmetric positive definite matrices. However, Kaasschieter ([7], Section 2) notes that eigenvalues of a symmetric positive semi-definite matrix that are zero do not contribute to the convergence of the CG-method. Furthermore, he concludes that the singular system can be solved by conjugate gradients as long as system (4) is consistent ($P\vec{b} \in \mathrm{Col}(PA)$). Van der Sluis and van der Vorst [15] note that the convergence may be much faster than bounds (1) and (3) predict when eigenvectors are clustered.

3.3 Choice of deflation vectors

We apply several choices for the deflation vectors to solve the following problem:

$$\left.\begin{aligned} -\nabla\cdot(k\ \nabla\psi) &= 0, \quad (x,y)\in\Omega, \\ \psi &= 1, \quad (x,y)\in\Gamma_D, \\ \frac{\partial\psi}{\partial n} &= 0, \quad (x,y)\in\Gamma_N. \end{aligned}\right\} \qquad (\mathrm{P}_3)$$

The domain Ω is divided into subdomains Ω_j such that $\overline{\Omega} = \bigcup_{i=1}^m \overline{\Omega}_i$ and $\aleph \subseteq \{1,\ldots,m\}$ denotes the set of indices that correspond to subdomains with high permeability, i.e.,

$$k(x,y) = \begin{cases} k_{\max} = 1, & (x,y)\in\Omega_j,\ \ j\in\aleph, \\ k_{\min} = \varepsilon, & (x,y)\in\Omega_j,\ \ j\in\{1,\ldots,m\}\setminus\aleph. \end{cases}$$

Further, we assume that if $\overline{\Omega}_i \cap \overline{\Omega}_j \neq \emptyset$ then the value of k in Ω_i is not equal to the value of k in Ω_j. We take $\varepsilon = 10^{-7}$. For each subdomain Ω_i we introduce a deflation vector $\vec{z}_j$ as follows:

$$z_j(x,y) = \begin{cases} 0 & \text{for } (x,y) \in \Omega \setminus \overline{\Omega}_j, \\ \in [0,1] & \text{for } (x,y) \in \overline{\Omega}_j \setminus (\Omega_j \cup (\Gamma_N \cap \overline{\Omega}_j)), \\ 1 & \text{for } (x,y) \in \Omega_j. \end{cases}$$

Note that in the finite element formulation the Dirichlet boundary points do not participate in the solution of the matrix-vector equation. An example of the geometry is shown in Fig. 1. We investigate the following choices where $\vec{z}_j$ is varied for $(x,y) \in \overline{\Omega}_j \setminus (\Omega_j \cup (\Gamma_N \cap \overline{\Omega}_j))$.

1. **non overlapping** projection vectors:

$$z_j(x,y) = \begin{cases} 1 & \text{for } (x,y) \in \overline{\Omega}_j \setminus (\Omega_j \cup \Gamma), \quad j \in \aleph, \\ 0 & \text{for } (x,y) \in \overline{\Omega}_j \setminus (\Omega_j \cup \Gamma), \quad j \in \{1, \ldots, m\} \setminus \aleph, \end{cases}$$

2. **complete overlapping** projection vectors:

$$z_j(\vec{x}) = 1 \quad \text{for } \overline{\Omega}_j \setminus (\Omega_j \cup \Gamma), \quad j \in \{1, \ldots, m\},$$

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

$$z_j(\vec{x}) = \frac{1}{2} \quad \text{for } \overline{\Omega}_j \setminus (\Omega_j \cup \Gamma), \quad j \in \{1, \ldots, m\},$$

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

$$z_j(x,y) = \begin{cases} \dfrac{k_{\max}}{k_{\max} + k_{\min}} & \text{for } \overline{\Omega}_j \setminus (\Omega_j \cup \Gamma), \quad j \in \aleph, \\ \dfrac{k_{\min}}{k_{\max} + k_{\min}} & \text{for } \overline{\Omega}_j \setminus (\Omega_j \cup \Gamma), \quad j \in \{1, \ldots, m\} \setminus \aleph. \end{cases}$$

Note that weighted overlap is approximated by no overlapping when $k_{\min} \ll k_{\max}$ and by average overlapping whenever $k_{\min} = k_{\max}$. After some theoretical results we investigate the four choices by numerical experiment for both the case that $k_{\min} \ll k_{\max}$ and $k_{\min} = k_{\max}$. Subsequently, we apply the deflation principle to parallel computation.

We consider the case that constrasts are large, i.e. $k_{\min} \ll k_{\max}$, $\varepsilon = 10^{-7}$. We now show that the choice of average overlap is not suitable for this case where $\text{cond}(A) = O(\frac{1}{\varepsilon})$. In the next theorem we refer to areas with $k_{\min}$ and $k_{\max}$ as low mobility and high mobility layers, respectively. Now we will show that the average overlapping projection vectors do not approximate the span of the eigenvectors that belong to eigenvalues that are of the order $O(\varepsilon)$.

Assumption 1. *We assume that the finite element discretization is consistent, which means that discretization error is zero for a constant function.*

We further assume that the off-diagonal entries of the discretization matrix A are non-positive.

Let $\boldsymbol{x}_m$ denote the position of grid point m. A consequence of the above assumption is

$$\sum_{j=1}^{n} a_{mj} = 0 \quad \text{for} \quad \boldsymbol{x}_m \in \Omega \setminus \Gamma_D. \tag{7}$$

Before we state the theorem we introduce the index set $\Pi_i \subset \{1, \ldots, n\}$ as the set of indices of $\vec{v}_i$ (which is the projection vector that corresponds to subdomain Ω_i) that correspond to grid points, which are on the boundary between two consecutive subdomains with $k = k_{\max} = 1$ and $k = k_{\min} = \varepsilon$. Further, we denote the neighbouring grid points of index set Π_i (located in Ω_i) by $\tilde{\Pi}_i$.

Theorem 6. *If the finite element discretization satisfies Assumption* 1 *and if the discretization matrix is irreducible, then for* $D = \text{diag}(a_{11} \ldots a_{nn})$ *we have*

1. $\|D^{-1}A\ \vec{v}_i\|_\infty \sim 1$ *for all* i, *regardless the value of* k *in the subdomain, for average overlapping delation vectors,*
2. $\|D^{-1}A\ \vec{v}_i\|_\infty = O(\varepsilon)$ *for all* i *corresponding to the subdomains* $\overline{\Omega}_i \cap \Gamma_D = \emptyset$ *with* $k = k_{\max} = 1$, *for the cases of non-overlapping, completely overlapping and weighted overlapping deflation vectors,*
3. $\|D^{-1}A\ \vec{v}_i\|_\infty \sim 1$ *for all values of* i *that are not incorporated in part* 2, *i.e.,* $k = k_{\min} = \varepsilon$ *or* $\overline{\Omega}_i \cap \Gamma_D \neq \emptyset$ *and* $k = k_{\max} = 1$.

Proof. We start with $\boldsymbol{x}_m \in \Omega_i \cup \overline{(\Omega_i \cap \Gamma_N)}$ and $m \in \{1, \ldots, n\} \setminus (\Pi_i \cup \tilde{\Pi}_i)$, where $\boldsymbol{x}_m$ is not a neighbouring point of a Dirichlet point, then $(v_i)_m$ is equal to its neighbours. From consistency of the discretization it follows

$$(A\ v_i)_m = 0 \quad m \in \{1, \ldots, n\} \setminus (\Pi_i \cup \tilde{\Pi}_i),$$

regardless the value of k in the subdomain i. For $m \in \Pi_i$ we split its neighbours into the sets J_m^H and J_m^H, respectively denoting the neighbouring gridnodes in the high and low permeability layers. Further, we introduce the set J_m^I representing the set of indices corresponding to neighbours of point m with index in Π_i. We proceed with $m \in \Pi_i$, then for the discretization matrix we have

$$\left.\begin{aligned} a_{mj} &\sim 1 && \text{for } j \in J_m^H, \\ a_{mj} &= O(\varepsilon) && \text{for } j \in J_m^L. \end{aligned}\right\}$$

Multiplication of A with vector v_i gives for component m

$$(A\ v_i)_m = \sum_{j=1}^{n} a_{mj}(v_i)_j = a_{mm}(v_i)_m + \sum_{j \in J_m^L \cup J_m^H \cup J_m^I} a_{mj}(v_i)_j.$$

Since $a_{mm} = -\sum_{j=1,\, j\neq m}^{n} a_{mj} \sim 1$ and $(v_i)_m = (v_i)_j$ for $j \in J_m^I$ when the subdomains are layered, the above relation implies

$$(Av_i)_m = \sum_{j \in J_m^L \cup J_m^H} a_{mj}((v_i)_j - (v_i)_m). \tag{8}$$

Now, we estimate $(Av_i)_m$ by the use of equation (8) for the three cases in the theorem.

1. Consider average overlapping deflation vectors. Then

$$\sum_{j \in J_m^L} a_{mj}((v_i)_j - (v_i)_m) = O(\varepsilon).$$

Further, note that from irreducibility of the matrix A it follows that $a_{mj} \sim 1$ for at least one $j \in J_m^H \neq \emptyset$. Hence, we have

$$\sum_{j \in J_m^L} a_{mj}((v_i)_j - (v_i)_m) \sim 1.$$

Therefore, with $D_{mm} = a_{mm} \sim 1$, it follows after using equation (8) that

$$\|D^{-1}A\ \vec{v}_i\|_\infty \sim 1, \quad \text{for all subdomains.}$$

This proves part 1 of the theorem.

2. Consider i corresponding to subdomain $\overline{\Omega_i} \cap \Gamma_D = \emptyset$ with $k = k_{\max} = 1$. Then for $m \in \Pi_i$ we have for all cases of overlap

$$\sum_{j \in J_m^L} a_{mj}((v_i)_j - (v_i)_m) = O(\varepsilon),$$

and

$$\sum_{j \in J_m^H} a_{mj}((v_i)_j - (v_i)_m) = \begin{cases} O(\varepsilon) & \text{for weighted overlap,} \\ 0 & \text{for complete and no overlap.} \end{cases}$$

Substitution of the above relations into equation (8) gives

$$(Av_i)_m = O(\varepsilon).$$

Since $D_{mm} \sim 1$, it follows for non-overlapping, complete overlap and weighted overlap that

$$(D^{-1}A\ v_i)_m = O(\varepsilon) \quad \text{for subdomains where } k\ = k_{\max} = 1.$$

For subdomains Ω_i where $\overline{\Omega}_i \cap \Gamma_D = \emptyset$ and $k = k_{\max} = 1$, it follows, with $(Av_i)_m = 0$ for $m \in \{1, \dots, n\} \setminus \Pi_i$, that

$$\|D^{-1}A\ v_i\|_\infty = O(\varepsilon).$$

This proves part 2 of the theorem.

3. Now we consider Ω_i where $k = k_{\max} = 1$ and with $\overline{\Omega}_i \cap \Gamma_D \neq \emptyset$, then for a certain grid point m that neighbours a Dirichlet grid point, we do not satisfy equation (8), but

$$\sum_{j=1}^{n} a_{mj} \sim 1,$$

since $k = k_{\max} = 1$. This implies that

$$\sum_{j=1}^{n} a_{mj}(v_i)_j \sim 1,$$

and hence

$$\|D^{-1}A\ v_i\|_\infty \sim 1, \quad \text{for } \overline{\Omega}_i \cap \Gamma_D \neq \emptyset \text{ and } k \ = k_{\max} = 1.$$

For $m \in \tilde{\Pi}_i$ we have for a subdomain, where $k = \varepsilon$,

$$\sum_{j=1}^{n} a_{mj}((v_i)_j - (v_i)_m) = O(\varepsilon) \quad \text{for no overlap and weigthed overlap.}$$

Since $a_{mm} = D_{mm} \sim \varepsilon$ for $m \in \tilde{\Pi}_i$ when $k = \varepsilon$, we obtain

$$\|D^{-1}A\ v_i\|_\infty \sim 1, \quad \text{for no overlap or weighted overlap.}$$

For $m \in \Pi_i$ we have for a subdomain, where $k = \varepsilon$,

$$\sum_{j \in J_m^H} a_{mj}((v_i)_j - (v_i)_m) \sim 1 \quad \text{for complete overlap.}$$

For complete overlap, we also have $D_{mm} \sim 1$ for $m \in \Pi_i$ and hence, one obtains

$$\|D_{-1}A\ v_i\|_\infty \sim 1 \quad \text{for complete overlap,}$$

whenever i corresponds to a subdomain where $k = k_{\min} = \varepsilon$. This proves the theorem. □

An important conclusion here is that the average overlapping deflation vectors with $\varepsilon \ll 1$ do not approximate the span of eigenvectors corresponding to the small eigenvalues of $D^{-1}A$. The above theorem is proven for diagonal scaling. We extend the result to an incomplete Cholesky decomposition for algebraic projection vectors with weighted and complete overlap and no overlap. A similar theorem is proven by Vuik et al [18] for physical deflation vectors. This is formulated in the following theorem:

Theorem 7. *If the finite element discretization is consistent under Assumption 1, then*

$$\|L^{-T}L^{-1}A\ \vec{v}_i\|_2 = O(\varepsilon)$$

for non-overlapping, completely and weighted overlapping projection vectors.

Proof. The proof is based on Theorem 6.

$$\|L^{-T}L^{-1}A\ \vec{v}_i\|_2 = \|L^{-T}L^{-1}DD^{-1}A\ \vec{v}_i\|_2 \le \lambda_{\max}(L^{-T}L^{-1}D)\|D^{-1}A\ \vec{v}_i\|_2.$$

Vuik et al [18] (Theorem 2.2) prove that $\lambda(L^{-T}L^{-1}D)$ is bounded also for $\varepsilon \to 0$. Further, since $\|D^{-1}A\vec{v}_i\|_2 \le \sqrt{n}\|D^{-1}A\vec{v}_i\|_\infty$, we obtain

$$\|L^{-T}L^{-1}A\ \vec{v}_i\|_2 \le \lambda_{\max}(L^{-T}L^{-1}D)\sqrt{n}\|D^{-1}A\vec{v}_i\|_\infty.$$

Since we know from Theorem 6

$$\|D^{-1}A\vec{v}_i\|_\infty \le O(\varepsilon)$$

for non-, completely and weighted overlapping projection vectors, we obtain

$$\|L^{-T}L^{-1}A\ \vec{v}_i\|_2 \le \lambda_{\max}(L^{-T}L^{-1}D)\sqrt{n}\,O(\varepsilon).$$

This proves the assertion. □

Note that the above theorem does not say anything for the case of incomplete Cholesky preconditioning and average overlapping projection vectors. The result in Theorem 6 is expected to hold for average overlapping with incomplete Cholesky preconditioning as well. This is also observed by numerical experiments.

4 Numerical experiments

For the experiments we consider test-problem (P_3) for a rectangular domain Ω. It easily follows that the solution of this test-problem is $\psi = 1$. The problem is solved by the use of a Finite Element method. We have done experiments with the different overlapping between subsequent subdomains. The parameter k is allowed to have large jumps. To illustrate the need for deflation, when k has large jumps, we start with the setting where we have horizontal layers of alternating permeability where $\varepsilon = 10^{-7}$. Here we take seven layers and 3200 (80×40) elements per layer and show a convergence result obtained from a Cholesky preconditioned conjugate gradient method when no deflation is applied in Fig. 4. It can be seen in the graph of the residual $\|\vec{b} - A\vec{x}_k\|_2$ that convergence is fast at the early stages. Subsequently, a non-monotonic behaviour is observed which is due to the presence of three eigenvalues that are of the order of 10^{-7} in accordance to Theorem 1.

These eigenvalues are due to the three sand layers that are sandwiched between low permeability shale layers. The convergence speed of the exact error $\|\vec{x}_k - \vec{x}_{\text{true}}\|_2$, however, is slow at the early stages and hence the solution has a poor quality then. Further for the sake of illustration, we plot the smallest eigenvalue of the preconditioned discretization matrix as a function of the iteration number. Only after 'discovery' of all the (small) eigenvalues convergence sets in. We see that although the problem is small, convergence is poor. To illustrate the increase of speed of convergence by the use of deflation, we present the results from deflation for the same problem in Fig. 5. Here physical deflation has been used, where only for each high permeability layer, i.e. sand layer, that is surrounded by shale-layers a deflation vector is used. From the results it can be seen that the smallest eigenvalue is of the order of 0.01 and that convergence of the residual is monotonous. Further, the exact error converges fast from the start. The computation has been finished in about 99 iterations instead 290 iterations in Fig. 4. So it is clear that deflation increases the convergence speed. Similar results can be found in the paper of Vuik et al [17].

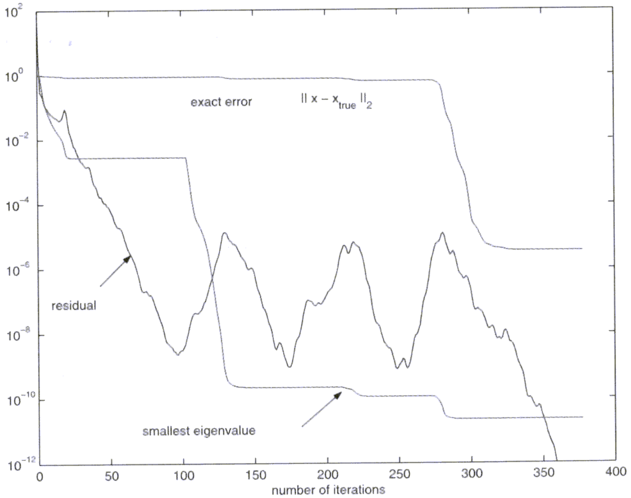

Figure 4. Convergence behavior of the incomplete Cholesky preconditioned CG-method for 7 subdomains with large jumps of the coefficient k in (P$_3$). The $\| \cdot \|_2$-norm of the residual and error have been presented. The number of elements is 3200 per layer.

4.1 Algebraic deflation vectors

We consider a comparison between deflation with algebraic projection vectors without overlap and deflation with physical projection vectors when contrasts in the permeability are very high $\varepsilon = 10^{-7}$. The results have been plotted in Figs. 5 and 6. From both figures it is clear that algebraic deflation without

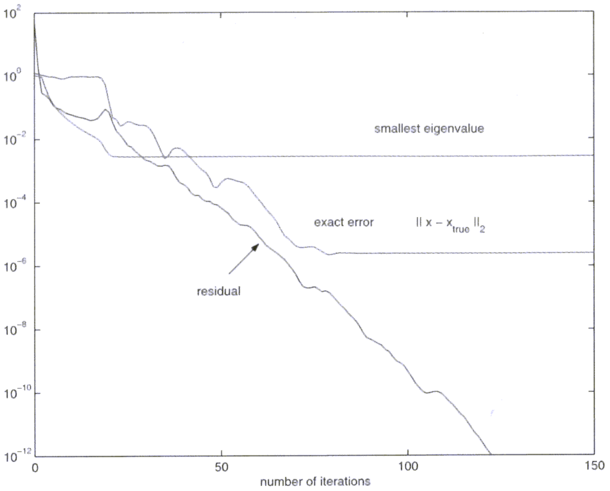

Figure 5. Convergence behavior of the deflated incomplete Cholesky preconditioned CG-method for 7 subdomains with large jumps of the coefficient k in (P$_3$). The $\| \cdot \|_2$-norm of the residual and error have been presented. Here results obtained by the use of physical deflation are shown. The number of elements is 3200 per layer.

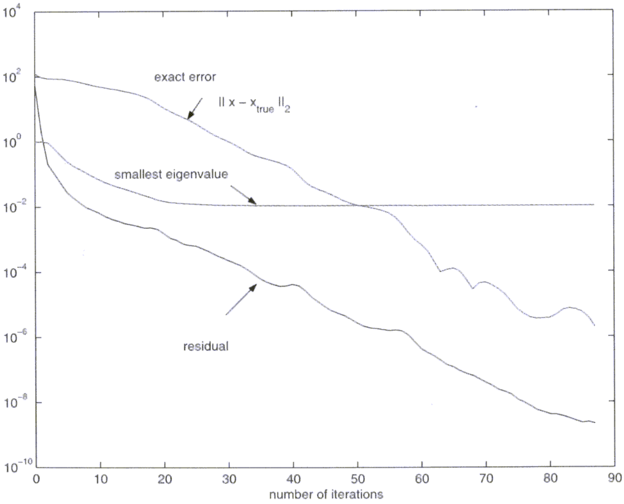

Figure 6. Convergence behavior of the deflated incomplete Cholesky preconditioned CG-method for 7 subdomains with large jumps of the coefficient k in (P$_3$). The $\| \cdot \|_2$-norm of the residual and error have been presented. Here results obtained by the use of algebraic deflation are shown. Algebraic deflation is here by no overlap for the projection vectors. The number of elements is 3200 per layer.

overlap, where for each layer there is a projection vector, requires fewer iterations than physical deflation. In Table 1 we present the number of iterations needed and computing time to obtain convergence for different number of elements per subdomain.

Table 1. Computation time and speed of convergence for physical and algebraic deflation vectors for several mesh-sizes.

	Physical deflation	Algebraic deflation		
Number of elements	Time (s)	Time (s)	N_{phys}	N_{alg}
50	0.02	0.02	16	15
200	0.07	0.07	28	26
800	0.46	0.39	52	44
3200	3.49	2.92	99	84
12800	24.20	18.72	173	140
41200	171.18	129.61	298	255

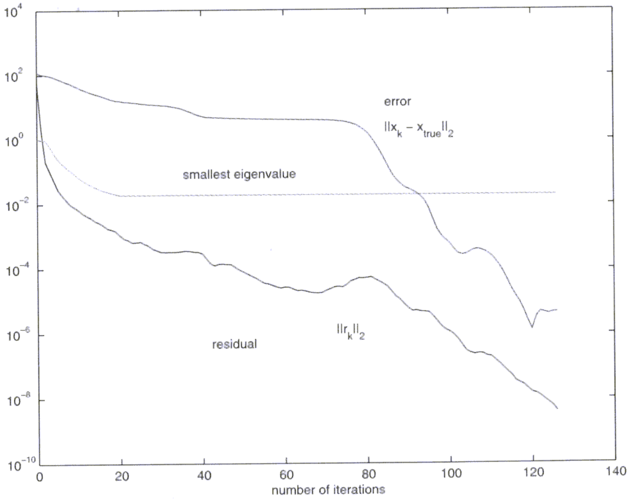

Figure 7. Convergence behavior of the deflated incomplete Cholesky preconditioned CG-method for 7 subdomains with large jumps of the coefficient k in (P$_3$). The $\|\cdot\|_2$-norm of the residual and error have been presented. Here results obtained by the use of algebraic deflation are shown. Algebraic deflation is here by complete overlap for the projection vectors. The number of elements is 3200 per layer.

From Table 1 it follows that algebraic deflation vectors without overlap gives a better convergence than physical deflation vectors, especially when the number of elements becomes large. Furthermore, the CPU-time is larger for physical deflation vectors, which is due to solving the homogeneous partial differential equation when the projection vector is determined. Note, however, that the number of projection vectors is larger for algebraic deflation.

From Fig. 5 and 6 it can also be seen that the smallest absolute value of the non-zero eigenvalues is less small when algebraic deflation is used. This is assumed to be the cause for the increase of speed of convergence when algebraic projections vectors are used.

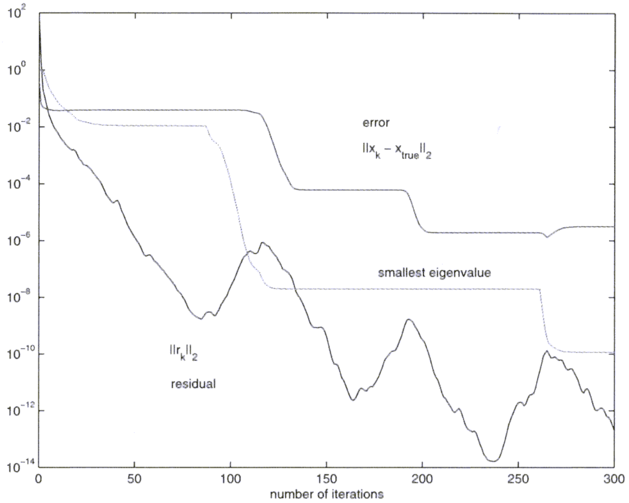

Figure 8. Convergence behavior of the deflated incomplete Cholesky preconditioned CG-method for 7 subdomains with large jumps of the coefficient k in (P_3). The $\|\cdot\|_2$-norm of the residual and error have been presented. Here results obtained by the use of algebraic deflation are shown. Algebraic deflation is here by average overlap for the projection vectors. The number of elements is 3200 per layer.

For the sake of illustration we show the evolution of the residual, smallest eigenvalue and error during the conjugate gradient iterations for the various implementations of deflation. In Fig. 7 we show the convergence for a system of seven subdomains with 3200 triangular elements. Here the projection vectors are chosen with complete overlap at the interface points. From Figs. 6 and 7 we see that the choice of complete overlapping projection vectors gives a slower convergence than non overlapping projection vectors. Subsequently, we show convergence with deflation vectors chosen with average overlap at the interfaces in Fig. 10. It can be seen that the smallest eigenvalue for average overlap is in the order of 10^{-8} by which convergence is deteriorated. From Figs. 6, 7 and 8 it is concluded that no overlapping deflation vectors give the best choice when contrasts are huge. Furthermore, average overlap gives the worst convergence behaviour. This is further illustrated by Fig. 9, where the exact error is plotted for the different options of deflation during the iterations. Computations with weighted overlapping projection vectors give the same results as for no-overlap in the case of sharp contrasts in the permeability, which follows from the definition of the several overlappings.

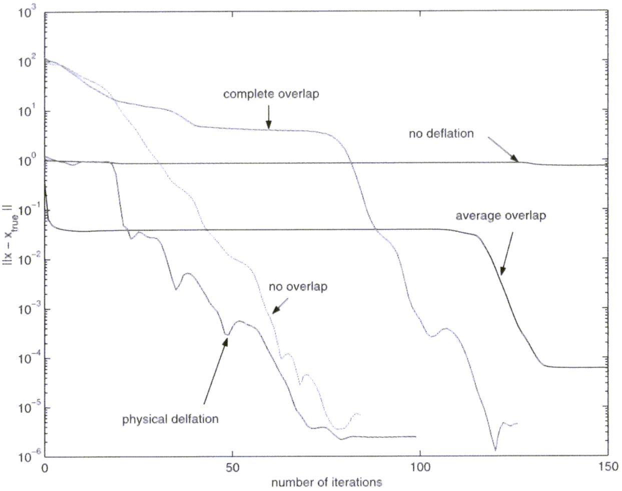

Figure 9. Convergence behavior of the deflated incomplete Cholesky preconditioned CG-method for 7 subdomains *with* large jumps of the coefficient k in (P$_3$). The $\|\cdot\|_2$-norm of error is presented for the various choices of deflation. The number of elements is 3200 per layer.

Similar computations, without sharp contrasts, $\varepsilon = 1$, are presented in Fig. 10. This figure indicates that average overlap gives the best results, although the differences are not as striking as for the case where there are large contrasts. Note that weighted overlap and average overlap are equivalent here. This difference is small due to the absence of extremely small eigenvalues. Further, the computations with complete overlap gives the poorest convergence. From the figures and computations with weighted overlap, whose results are omitted in this paper, it is concluded that weighted overlap always gives a good convergence, since it mimics no overlap when constrasts are high and average overlap when constrasts do not exist. This is an important insight for future parallelization of the deflated preconditioned conjugate gradient method. We further note that computations with weighted overlap give the same results as for average overlapping for the case of no contrasts of the permeability.

We conclude that weighted overlap is most robust and this will always give the best choice for use in a 'blackbox' algorithm.

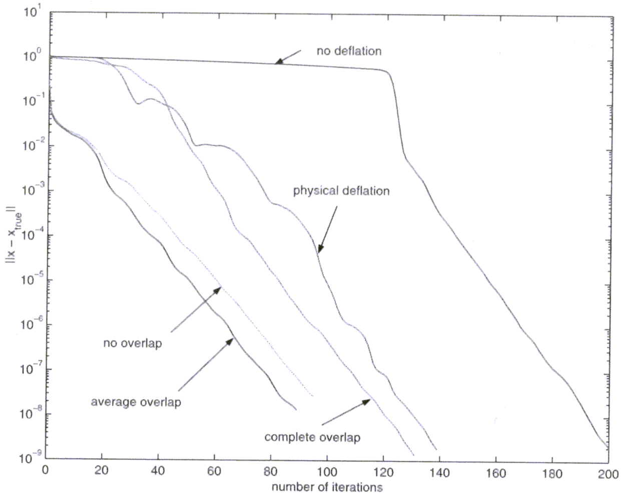

Figure 10. Convergence behavior of the deflated incomplete Cholesky preconditioned CG-method for 7 subdomains *without* large jumps of the coefficient k in (P_3). The $\| \cdot \|_2$-norm of the error is presented for the various choices of deflation. The number of elements is 3200 per layer.

4.2 Parallelization

The deflated preconditioned conjugate gradient method is very suitable for parallelization. Parallelization is still a topic of research. The first results are given here, which have been computed for a layered domain as in Fig. 1. In the first series of numerical experiments, we show the convergence behaviour when the number of subdomains is increased, using a constant number of grid points, so the total number of gridnodes increases. Subsequently, we show the convergence behaviour for the case that the number of blocks increases such that the total number of gridnodes remains constant.

4.2.1 Increase of the size of the domain of computation

Here we consider a rectangular domain Ω^n that consists of the equisized subdomains $\Omega_1, \ldots, \Omega_n$, $\overline{\Omega}^n = \overline{\Omega}_1 \cup \ldots \cup \overline{\Omega}_n$, $\overline{\Omega}^{n+1} = \overline{\Omega}^n \cup \overline{\Omega}_{n+1}$. In this section we take the permeability constant over the whole domain. Further we use algebraic projection vectors with average overlap. We compare the results from the following computation methods:

1. Sequential ICCG without deflation (SICCG);
2. Sequential ICCG with deflation (SDICCG);
3. Parallel ICCG without deflation (PICCG);
4. Parallel ICCG with deflation (PDICCG).

The number of iterations needed for convergence is plotted as a function of the number of subdomains for all four approaches in Fig. 11. The number of grid points per subdomain is 80 × 80. It is seen that the parallel ICCG

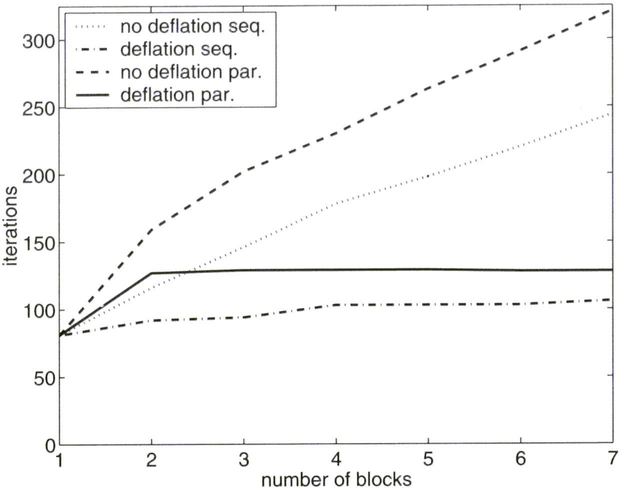

Figure 11. The number of iterations as the number of subdomains increases for the various methods of computation.

(PICCG) requires more iterations than the sequential ICCG (SICCG). This is a common observation. For both cases the number of iterations needed for convergence increases rapidly as the size of the domain of computation increases. However, for both deflated methods the number of iterations is lower than for the non-deflated methods. Further, as the number of subdomains, i.e. the size of the domain of computation, increases the number of needed iterations becomes independent of the number of subdomains.

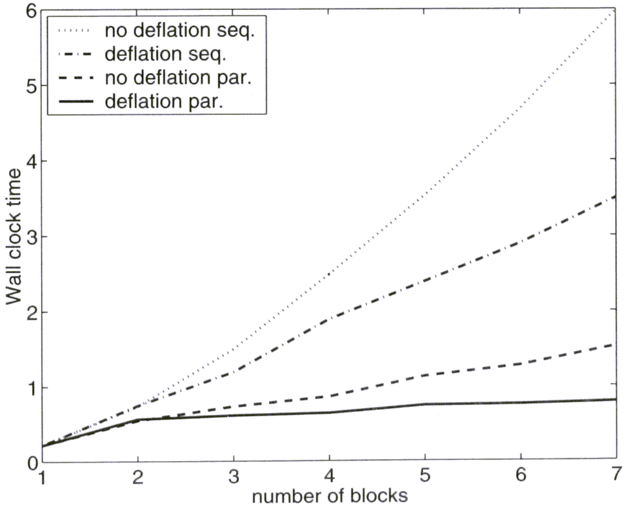

Figure 12. The wall-clock time as the number of subdomains increases for the various methods of computation.

As a further illustration we show the wall-clock time for the four different approaches in Fig. 12. The wall-clock time is measured on a Beowulf cluster. It can be seen from this figure that the wall-clock time is significantly smaller for the deflated ICCG (SDICCG) for the sequential computation. So, for sequential computations deflation is attractive to increase the speed of computation. Further, parallelization gives a significant speed-up, but as the number of subdomains increases, the wall-clock time continues to increase when no deflation is used (PICCG). However, if deflation is used in the parallel computations, the wall-clock time decreases and even becomes almost independent of the number of added subdomains. It is therefore concluded in this section that deflation accellerates computation in both a sequential and parallel computer environment, if the solution of an elliptic problem as in (P_3) is computed.

4.2.2 Increase of the number of deflation vectors in a given domain

Given a square domain, we increase the number of deflation vectors. Further, we assume that the permeability is constant over the whole domain. We use 150 × 150 elements over the domain.

We present the results of the computations in Table 2 where we use the methods SDICCG, DICCG and PDICCG (sequential with deflation, parallel without deflation and parallel with deflation) for the cases of 3 and 6 blocks. The number of iterations for the convergence of the SICCG method (sequential, without deflation) is 142 iterations.

Table 2. Number of iterations for a square domain with 150 × 150 elements.

Number of blocks	SDICCG	PICCG	PDICCG
3	106	198	141
6	101	198	138

We remark that we used average overlap for the projection vectors in the deflated method. From the above table it is seen that deflation gives a reduction of the number of required iterations for the sequential computations (compare column 2 with the 142 iterations for the SICCG method). Further, the number of iterations increases when a parallel method is applied. However, deflation applied in a parallelized method reduces the number of iterations again. Hence, deflation is again recommended to use in both sequential and parallelized computations.

5 Conclusions

We investigated various choices of deflation vectors, which are used in the Deflated ICCG 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. Further, deflation makes the parallel preconditioned conjugate gradient method scalable: the wall-clock time becomes invariant with respect to the number of blocks if the number of blocks is increased and the number of gridnodes per block is constant.
- For the case of no contrasts of the permeability between subsequent layers, it is observed that average overlap between subsequent deflation vectors is superior to no overlapping. Here the use of complete overlapping projection vectors is unsuitable. Whereas, for cases with large contrasts of permeability the use of average overlapping projection vectors is not suitable. This is caused by the fact that the span of the projection vectors approximates the span of the eigenvectors that belong to the small eigenvectors badly.
- 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 for both the presence and absence of sharp contrasts until now.

References

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

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

[3] G. De Josselin de Jong, *Singularity distributions for the analysis of multiple flyuid flow in porous media*, Journal of Geothermal Research **65** (1960), 3739–3758.

[4] J. Frank, C. Vuik, *On the construction of deflation-based preconditiones*, SIAM J. Sci. Comput. **23** (2001), 442–562.

[5] G. H. Golub, C. F. van Loan, *Matrix Computations*, Third edition, The Johns Hopkins University Press, Baltimore, 1996.

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

[7] E. F. Kaasschieter, *Preconditioned conjugate gradients for solving singular systems*, J. Comput. Appl. Math. **24** (1988), 265–275.

[8] E. Kreyszig, *Introductory Functional Analysis with Applications*, Willey, New York, 1989.

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

[10] D. C. Lay, *Linear Algebra and Its Applications*, Addison-Wesley, Longman Scientific, Reading, Massachusetts, 1996.

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

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

[13] G. J. M. Pieters, *Stability analysis for a saline layer formed by uniform using finite elements*, Report RANA 01-07, Eindhoven University of Technology, Eindhoven, 2001.

[14] R. Temam, *Navier-Stokes Equations, Theory and Numerical Analysis*, Elsevier Science Publishers, Amsterdam, 1984.

[15] A. van der Sluis, H. van der Vorst, *The rate of convergence of conjugate gradients*, Numer. Math. **48** (1986), 543–560.

[16] C. Vuik, A. Segal, L. el Yaakoubi and E. Dufour, *A comparison of various deflation vectors applied to elliptic problems with discontinuous coefficients*, Appl. Numer. Math. **41** (2002), 219–233.

[17] C. Vuik, A. Segal, and J. A. Meijerink, *An efficient preconditioned DG method for the solution of a class of layered problems with extreme contasts in the coefficients*, J. Comput. Phys. **152** (1999), 385–403.

[18] 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*, J. Comput. Phys. **172** (2001), 426–450.