Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational and Applied Mathematics 206 (2007) 603–614
JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS
www.elsevier.com/locate/cam

On deflation and singular symmetric positive semi-definite matrices

J.M. Tang*, C. Vuik

Faculty of Electrical Engineering, Mathematics and Computer Science, Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

Received 13 December 2005; received in revised form 7 June 2006

Abstract

For various applications, it is well-known that the deflated ICCG is an efficient method for solving linear systems with invertible coefficient matrix. We propose two equivalent variants of this deflated ICCG which can also solve linear systems with singular coefficient matrix, arising from discretization of the discontinuous Poisson equation with Neumann boundary conditions. It is demonstrated both theoretically and numerically that the resulting methods accelerate the convergence of the iterative process.

Moreover, in practice the singular coefficient matrix has often been made invertible by modifying the last element, since this can be advantageous for the solver. However, the drawback is that the condition number becomes worse-conditioned. We show that this problem can completely be remedied by applying the deflation technique with just one deflation vector.

MSC: 65F10; 65F50; 65N22

Keywords: Deflation; Conjugate gradient method; Preconditioning; Poisson equation; Spectral analysis; Singular symmetric positive semi-definite matrices

1. Introduction

Recently, moving boundary problems have received much attention in literature due to their applicative relevance in many physical processes. One of the most popular moving boundary problems is modelling bubbly flows, see e.g. [15]. These bubbly flows can be simulated by solving the Navier–Stokes equations using for instance the pressure correction method [5]. The most time-consuming part of this method is solving the symmetric and positive semi-definite (SPSD) linear system on each time step, which is coming from a second-order finite-difference discretization of the Poisson equation with possibly discontinuous coefficients and Neumann boundary conditions:

$$\begin{cases} \nabla \cdot \left(\dfrac{1}{\rho(\mathbf{x})} \nabla p(\mathbf{x}) \right) = f(\mathbf{x}), & \mathbf{x} \in \Omega, \\ \dfrac{\partial}{\partial \mathbf{n}} p(\mathbf{x}) = g(\mathbf{x}), & \mathbf{x} \in \partial\Omega, \end{cases} \tag{1}$$

* Corresponding author.
E-mail addresses: J.M.Tang@tudelft.nl (J.M. Tang), C.Vuik@tudelft.nl (C. Vuik).

0377-0427/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cam.2006.08.015

where p, ρ, $\mathbf{x}$ and $\mathbf{n}$ denote the pressure, density, spatial coordinates and the unit normal vector to the boundary $\partial\Omega$, respectively. The resulting singular linear system is

$$Ax = b, \quad A = [a_{i,j}] \in \mathbb{R}^{n\times n}, \tag{2}$$

where the coefficient matrix A is SPSD. If $b \in \mathrm{Col}(A)$ then the linear system (2) is consistent and infinite number of solutions exists. Due to the Neumann boundary conditions, the solution x is determined up to a constant, i.e., if x_1 is a solution then $x_1 + c$ is also a solution where $c \in \mathbb{R}^n$ is an arbitrary constant vector. This situation presents no real difficulty, since pressure is a relative variable, not an absolute one. In this paper we concentrate on the linear system (2), which can also be derived from other problems besides the bubbly flow problems. The precise requirements can be found in the next section of this paper.

In many computational fluid dynamics packages, see also [1,4,14], one would impose an invertible A, denoted by $\widetilde{A}$. This makes the solution x unique which can be advantageous in computations, for instance,

- direct solvers like Gaussian elimination can only be used to solve the linear systems when A is invertible;
- the original singular system may be inconsistent as a result of rounding errors, whereas the invertible system is always consistent;
- the deflation technique requires an invertible matrix $E := Z^{\mathrm{T}}AZ$ which will be explained later on in this paper. The choice of Z is only straightforward if A is non-singular.

One common way to force invertibility of matrix A in literature is to replace the last element $a_{n,n}$ by $\tilde{a}_{n,n} = (1+\sigma)a_{n,n}$ with $\sigma > 0$. In fact, a Dirichlet boundary condition is imposed at one point of the domain Ω. This modification results in an invertible linear system

$$\widetilde{A}x = b, \quad \widetilde{A} = [\tilde{a}_{i,j}] \in \mathbb{R}^{n\times n}, \tag{3}$$

where $\widetilde{A}$ is symmetric and positive definite (SPD).

The most popular iterative method to solve linear systems like (3) is the preconditioned conjugate gradient (CG) method (see e.g. [3]). After k iterations of the CG method, the error is bounded by (cf. [3, Theorem 10.2.6])

$$\|x - x_k\|_{\widetilde{A}} \leqslant 2\|x - x_0\|_{\widetilde{A}} \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k, \tag{4}$$

where x_0 denotes the starting vector and $\kappa = \kappa(\widetilde{A}) = \lambda_n/\lambda_1$ denotes the spectral condition number of $\widetilde{A}$. Therefore, a smaller κ leads asymptotically to a faster convergence of the CG method. In practice, it appears that the condition number κ is relatively large, especially when σ is close to 0. Hence, solving (3) with the CG method shows slow convergence, see also [4,14]. The same holds if the ICCG method [9] is used. Since $\widetilde{A}$ is an SPD matrix with $\tilde{a}_{i,j} \leqslant 0$ for all $i \neq j$, an incomplete Cholesky (IC) decomposition always exists [4].

ICCG shows good performance for relatively small and easy problems. However, it appears that ICCG still does not give satisfactory results in more complex models, for instance when the number of grid points becomes very large or when there are large jumps in the density of (1). To remedy the bad convergence of ICCG, (eigenvalue) deflation techniques are proposed, originally from Nicolaides [13]. The idea of deflation is to project the extremely large or small eigenvalues of $\widetilde{M}^{-1}\widetilde{A}$ to zero, where $\widetilde{M}^{-1}$ is the IC preconditioner based on $\widetilde{A}$. This leads to a faster convergence of the iterative process, due to Expression (4) and due to the fact that the CG method can handle matrices with zero-eigenvalues, see also [4]. The resulting method is called DICCG.

The deflation technique has been exploited by several other authors, e.g. [2,7,8,10–12]. The resulting linear system which has to be solved is

$$\widetilde{M}^{-1}\widetilde{P}\widetilde{A}x = \widetilde{M}^{-1}\widetilde{P}b, \tag{5}$$

where $\widetilde{P}$ denotes the deflation matrix based on $\widetilde{A}$. We have already mentioned that the ICCG method shows slow convergence after forcing invertibility of A. In this paper, we will investigate this phenomenon for the DICCG method.

Another DICCG approach to solve (2) without forcing invertibility is to solve

$$M^{-1}PAx = M^{-1}Pb, \tag{6}$$

Table 1
Notations for standard matrices and vectors where $p, q, s \in \mathbb{N}$

Notation	Meaning
$\mathbf{e}_p^{(s)}$	sth column of the $p \times p$ identity matrix I
$\mathbf{e}_{p,q}^{(s)}$	$p \times q$ matrix with q identical columns $\mathbf{e}_p^{(s)}$
$\mathbf{1}_{p,q}$	$p \times q$ matrix consisting of all ones
$\mathbf{1}_p$	Column of $\mathbf{1}_{p,q}$
$\mathbf{0}_{p,q}$	$p \times q$ matrix consisting of all zeros
$\mathbf{0}_p$	Column of $\mathbf{0}_{p,q}$

where M^{-1} is the IC preconditioner based on A and P is a specific deflation matrix, which should be constructed in such a way that it can deal with the singular matrix A. Most papers on deflation, e.g. [2,7,8,10–12], deal only with invertible systems. Applications of deflation to singular systems are described in [6,18,19]. In these papers, some suggestions have been given how to combine singular systems with a deflation technique, but the underlying theory has not yet been developed. In this paper, relations between the singular matrix A and the invertible matrix $\widetilde{A}$ will be worked out using the deflation matrices P and $\widetilde{P}$, to gain more insight into the application of the deflation technique for singular systems. Moreover, we investigate the two DICCG approaches theoretically, especially by comparing the deflated systems $\widetilde{P}\widetilde{A}$ and PA and thereafter by comparing the preconditioned variants $\widetilde{M}^{-1}\widetilde{P}\widetilde{A}$ and $M^{-1}PA$. Thereafter, numerical experiments will be done to investigate the convergence behavior of these approaches, which will be compared to ICCG.

The outline of this paper is as follows. In Section 2 we introduce some notations, assumptions and definitions. We compare the matrices $\widetilde{P}\widetilde{A}$ and PA in Section 3. Moreover, we investigate the possibilities with deflation to solve the problem of a worse condition number of the coefficient matrix after forcing invertibility. Thereafter, in Section 4 the comparison of $\widetilde{P}\widetilde{A}$ and PA will be generalized to $\widetilde{M}^{-1}\widetilde{P}\widetilde{A}$ and $M^{-1}PA$. Results of numerical experiments will be presented in Section 5 to illustrate the theory. Section 6 is devoted to the conclusions. For more details we refer to [17].

2. Notations, assumptions and definitions

We define some standard matrices and vectors which will be used through this paper, see Table 1.
Subsequently, the $n \times n$ matrix A satisfies two assumptions which are given below.

Assumption 1. Matrix $A \in \mathbb{R}^{n\times n}$ is SPSD and singular. Moreover, the algebraic multiplicity of the zero-eigenvalue of A is equal to one.

Assumption 2. Matrix A satisfies $A\mathbf{1}_n = \mathbf{0}_n$.

Now, matrix $\widetilde{A}$ is defined in the following way.

Definition 1. Let A be given which satisfies Assumptions 1 and 2. Then $\widetilde{A}$ is defined by

$$\widetilde{a}_{n,n} = (1+\sigma)a_{n,n}, \quad \sigma > 0, \tag{7}$$

and for the other indices i and j

$$\widetilde{a}_{i,j} = a_{i,j}. \tag{8}$$

Some consequences of Definition 1 can be found in the following two corollaries.

Corollary 1. *Matrix $\widetilde{A}$ is invertible, SPD.*

Corollary 2. *Matrix* $\widetilde{A}$ *satisfies* $\widetilde{A}\mathbf{1}_{n,n} = \sigma a_{n,n}\mathbf{e}_{n,n}^{(n)}$. *In particular,* $\widetilde{A}\mathbf{1}_n = \sigma a_{n,n}\mathbf{e}_n^{(n)}$.

Next, let the computational domain Ω be divided into open subdomains Ω_j, $j = 1, 2, \ldots, r$, such that $\Omega = \bigcup_{j=1}^{r} \overline{\Omega}_j$ and $\bigcap_{j=1}^{r} \Omega_j = \emptyset$ where $\overline{\Omega}_j$ is Ω_j including its adjacent boundaries. The discretized domain and subdomains are denoted by Ω_h and Ω_{h_j}, respectively. Then, for each Ω_{h_j} with $j = 1, 2, \ldots, r$, we introduce a deflation vector z_j as follows:

$$(z_j)_i := \begin{cases} 0, & x_i \in \Omega_h \backslash \overline{\Omega}_{h_j}, \\ 1, & x_i \in \Omega_{h_j}, \end{cases} \tag{9}$$

where x_i is a grid point of the discretized domain Ω_h and $z_0 = \mathbf{1}_n$. Subsequently, we define the so-called deflation subspace matrices Z, $\widetilde{Z}$ and $\widetilde{Z}_0$.

Definition 2. For $r > 1$, we define $Z := [z_1\ z_2 \cdots z_{r-1}] \in \mathbb{R}^{n\times(r-1)}$, $\widetilde{Z} := [Z\ z_r]$ and $\widetilde{Z}_0 := [Z\ z_0]$. For $r = 1$, only $\widetilde{Z} = [z_r]$ and $\widetilde{Z}_0 = [z_0]$ are defined.

It can be observed that

$$\widetilde{Z}\mathbf{1}_r = \mathbf{1}_n. \tag{10}$$

Finally, the deflation matrices can be defined.

Definition 3. Matrices P_r, $\widetilde{P}_r$, $\widetilde{Q}_r \in \mathbb{R}^{n\times n}$ are defined as follows:

- $P_r := I - AZE^{-1}Z^{\mathrm{T}}, \quad E := Z^{\mathrm{T}}AZ;$
- $\widetilde{P}_r := I - \widetilde{A}\widetilde{Z}\widetilde{E}^{-1}\widetilde{Z}^{\mathrm{T}}, \ \widetilde{E} := \widetilde{Z}^{\mathrm{T}}\widetilde{A}\widetilde{Z};$
- $\widetilde{Q}_r := I - \widetilde{A}\widetilde{Z}_0\widetilde{E}^{-1}\widetilde{Z}_0^{\mathrm{T}}, \quad \widetilde{E}_0 := \widetilde{Z}_0^{\mathrm{T}}\widetilde{A}\widetilde{Z}_0.$

We observe that P_r is based on the singular matrix A, whereas both $\widetilde{P}_r$ and $\widetilde{Q}_r$ are based on the invertible matrix $\widetilde{A}$ which only differ in the deflation subspace matrices. Moreover, note that E, $\widetilde{E}$ and $\widetilde{E}_0$ are relatively small matrices, since $r \ll n$ in general.

It is essential to note that Z cannot be replaced by $\widetilde{Z}$ in the expression of P_r, which is common in the deflation theory for linear systems with invertible coefficient matrices. In our case, deflation matrix P_r with $\widetilde{Z}$ will be undefined, since E becomes singular. It can be observed that all deflation matrices as defined in Definition 3 are well-defined, since E, $\widetilde{E}$ and $\widetilde{E}_0$ are all non-singular.

Subsequently, it is easy to show that $P_r A$ and $\widetilde{P}_r\widetilde{A}$ are SPSD matrices like A. In addition, it is also straightforward to show that these matrices are invariant for permutations, scaling and linear combinations of the columns of the deflation subspace matrices Z and $\widetilde{Z}$, respectively, as long as the column space of Z and $\widetilde{Z}$ does not change. This leads immediately to Lemma 1.

Lemma 1. *Identity* $\widetilde{Q}_r = \widetilde{P}_r$ *holds.*

The resulting linear systems $\widetilde{M}^{-1}\widetilde{P}_r\widetilde{A}x = \widetilde{M}^{-1}\widetilde{P}_r b$ or $M^{-1}P_r Ax = M^{-1}P_r b$ with $\widetilde{M}^{-1}$ and M^{-1} to be IC preconditioners associated to $\widetilde{A}$ and A, respectively, can be solved using the CG method. This method is called DICCG-r.

Definition 4. DICCG-r is defined as the CG method applied to the linear system $\widetilde{M}^{-1}\widetilde{P}_r\widetilde{A}x = \widetilde{M}^{-1}\widetilde{P}_r b$ or $M^{-1}P_r Ax = M^{-1}P_r b$.

Next, the eigenvalues λ_i of a symmetric $n \times n$ matrix are always ordered increasingly, i.e., $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$. Furthermore, let B be an arbitrary $n \times n$ SPSD matrix with rank $n - r$, so that $\lambda_1 = \cdots = \lambda_r = 0$. Note that all eigenvalues of B are real-valued due to the symmetry of B. Then its effective condition number $\kappa_{\mathrm{eff}}(B)$ is defined by $\kappa_{\mathrm{eff}}(B) := \lambda_n(B)/\lambda_{r+1}(B)$.

From [11, Theorem 2.6] we have $\lambda_1(\widetilde{P}_r\widetilde{A}) = \lambda_2(\widetilde{P}_r\widetilde{A}) = \cdots = \lambda_r(\widetilde{P}_r\widetilde{A}) = 0$. Moreover, the effective condition number of $\widetilde{P}_r\widetilde{A}$ decreases if we increase the number of deflation vectors, which follows from [11, Lemma 2.10]. Finally, it is a trivial

result that forcing invertibility of A leads automatically to a worse condition number, i.e., inequality $\kappa(\widetilde{A}) \geqslant \kappa_{\text{eff}}(A)$ holds for all $\sigma \geqslant 0$.

3. Comparison of the deflated singular and invertible matrix

In this section, we first show that the condition number of $\widetilde{A}$ is reduced to the condition number of A by a simple deflation technique. More precisely, if $\widetilde{P}_1$ is the deflation matrix with one constant deflation vector based on $\widetilde{A}$, then the deflated matrix $\widetilde{P}_1\widetilde{A}$ appears to be identical to the original singular matrix A. Thereafter, we show that even the deflated variants of $\widetilde{A}$ and A, denoted by $\widetilde{P}_r\widetilde{A}$ and P_rA respectively, are equal. As a consequence, solving $Ax = b$ and $\widetilde{A}x = b$ with a deflated Krylov iterative method leads in theory to the same convergence results.

3.1. Comparison of $\widetilde{P}_1\widetilde{A}$ and A

Before giving the proof of the equality $\widetilde{P}_1\widetilde{A} = A$, we start with Lemma 2, where it will be shown that $\widetilde{P}_1$ is the identity matrix except for the last row.

Lemma 2. $\widetilde{P}_1 = I - \mathbf{e}_{n,n}^{(n)}$.

Proof. For $r = 1$ we have $\widetilde{P}_1 = I - \widetilde{A}z_0\widetilde{E}^{-1}z_0^{\mathrm{T}}$, where $E^{-1} = 1/(z_0^{\mathrm{T}}\widetilde{A}z_0) = 1/(\sigma a_{n,n})$ using Corollary 2. Hence, $\widetilde{P}_1 = I - (\widetilde{A}\mathbf{1}_{n,n}/\sigma a_{n,n}) = I - \mathbf{e}_{n,n}^{(n)}$. □

Note that $\widetilde{P}_1$ has the properties that the last column is the zero-column and that the matrix consists of only the values 0, 1 and −1. Next, applying Lemma 2, we obtain the following theorem.

Theorem 1. *Equality $\widetilde{P}_1\widetilde{A} = A$ holds.*

Proof. Due to Lemma 2, it is obvious to see that $\widetilde{P}_1\widetilde{A} = A$ holds for all rows except the last one. The analysis of the last row of $\widetilde{P}_1\widetilde{A}$, which is $(\mathbf{e}_n^{(n)} - \mathbf{1}_n)^{\mathrm{T}}\widetilde{A}$, is as follows. Since $\mathbf{1}_n^{\mathrm{T}}A = \mathbf{0}_n^{\mathrm{T}}$ and $(\mathbf{e}_n^{(n)} - \mathbf{1}_n)^{\mathrm{T}}\widetilde{A} = (\mathbf{e}_n^{(n)} - \mathbf{1}_n)^{\mathrm{T}}A$ hold, this yields $(\mathbf{e}_n^{(n)} - \mathbf{1}_n)^{\mathrm{T}}\widetilde{A} = \mathbf{e}_n^{(n)\mathrm{T}}A$. Hence, the last rows of $\widetilde{P}_1\widetilde{A}$and A are also equal which proves the theorem. □

Theorem 1 implies that, after applying deflation with $r = 1$, the invertible matrix $\widetilde{A}$ becomes equal to the original singular matrix A. It even appears that the results using this deflation technique are independent of the elements of the last row of matrix $\widetilde{A}$.

3.2. Comparison of $\widetilde{P}_r\widetilde{A}$ and P_rA

Theorem 2 is the main result of this section. In order to prove this theorem, a set of lemmas is required which is stated below. The most important lemmas are Lemmas 3 and 6 which show that deflation matrix $\widetilde{P}_r$ is invariant by right-multiplication with deflation matrix $\widetilde{P}_1$ and that deflated systems $\widetilde{P}_rA$ and P_rA are identical.

Lemma 3. $\widetilde{P}_r\widetilde{P}_1 = \widetilde{P}_r$ *holds.*

Proof. Note first that by using $\widetilde{Z}^{\mathrm{T}}\widetilde{A}\widetilde{Z}\mathbf{1}_r = \widetilde{Z}^{\mathrm{T}}\sigma a_{n,n}\mathbf{e}_n^{(n)} = \sigma a_{n,n}\mathbf{e}_r^{(r)}$ it is obvious that the last column of $\widetilde{E}^{-1}$ is equal to$(1/(\sigma a_{n,n}))\mathbf{1}_n$. Then, it can easily be seen that the last column of $\widetilde{A}\widetilde{Z}\widetilde{E}^{-1}\widetilde{Z}^{\mathrm{T}}$ is exactly $\mathbf{e}_n^{(n)}$ for all values of σ, resulting in the fact that the last column of $\widetilde{P}_r$ is the zero-vector $\mathbf{0}_n$. Then, $\widetilde{P}_r\widetilde{A}\mathbf{1}_n = \mathbf{0}_n$ for all $\sigma > 0$, where we have also applied Corollary 2. Hence, this implies $\widetilde{P}_r\widetilde{P}_1 = \widetilde{P}_r(I - \alpha\widetilde{A}\mathbf{1}_n) = \widetilde{P}_r - \alpha\widetilde{P}_r\widetilde{A}\mathbf{1}_n = \widetilde{P}_r$. □

Lemma 4. *There exists a matrix $\widetilde{Y} := [z_{r+1}\ z_{r+2}\ \cdots\ z_n]$ such that*

- *matrix $X := [\widetilde{Y}\ \widetilde{Z}_0]$ is invertible*;
- *the identity $\widetilde{Z}_0^{\mathrm{T}}\widetilde{A}\widetilde{Y} = \mathbf{0}_{r,n-r}$ holds.*

Proof. It is always possible to find a matrix $\widetilde{Y}$ such that $\mathrm{Col}(H) = \mathrm{Col}(\widetilde{Y}) \oplus \mathrm{Col}(\widetilde{Z}_0)$, where $\mathrm{Col}(\widetilde{Y})$ is an orthogonal complement of $\mathrm{Col}(\widetilde{Z}_0)$. Then, by definition of the direct sum, $X := [\widetilde{Y}\ \widetilde{Z}_0]$ is an invertible matrix. Furthermore, it is known that $\mathrm{Col}(\widetilde{Y}) = \{y \in \mathbb{R}^n \mid \langle w, y\rangle_{\widetilde{A}} = 0\ \forall w \in \mathrm{Col}(\widetilde{Z}_0)\}$. In particular, for each $w \in \widetilde{Z}_0$ and for each $y \in \widetilde{Y}$ we have $\langle w, y\rangle_{\widetilde{A}} = w^{\mathrm{T}}\widetilde{A}y = 0$, which yields $\widetilde{Z}_0^{\mathrm{T}}\widetilde{A}\widetilde{Y} = \mathbf{0}_{r,n-r}$. □

Lemma 5. *The following equalities holds*:

(i) $(P_r - \widetilde{Q}_r - \mathbf{e}_n^{(n)} z_0^{\mathrm{T}})\widetilde{A}z_0 = \mathbf{0}_n$;
(ii) $(P_r - \widetilde{Q}_r)\widetilde{A}Z = \mathbf{0}_{n,r-1}$.

Proof. (i) Note first that $Z^{\mathrm{T}}\widetilde{A}z_0 = \mathbf{0}_n$ and $\widetilde{Z}_0\widetilde{E}_0^{-1}\widetilde{Z}_0^{\mathrm{T}}\widetilde{A}z_0 = \widetilde{Z}_0\mathbf{e}_n^{(n)} = z_0$ hold. Combining these facts with Corollary 2 implies

$$\left(\widetilde{A}\widetilde{Z}_0\widetilde{E}_0^{-1}\widetilde{Z}_0^{\mathrm{T}} - AZE^{-1}Z^{\mathrm{T}}\right)\widetilde{A}z_0 = \widetilde{A}\widetilde{Z}_0\widetilde{E}_0^{-1}\widetilde{Z}_0^{\mathrm{T}}\widetilde{A}z_0 = \widetilde{A}z_0 = \sigma a_{n,n}\mathbf{e}_n^{(n)}. \tag{11}$$

With the help of the equalities $\mathbf{e}_n^{(n)}\mathbf{1}_n^{\mathrm{T}} = \mathbf{e}_{n,n}^{(n)}$ and $\mathbf{e}_{n,n}^{(n)}\mathbf{e}_n^{(n)} = \mathbf{e}_n^{(n)}$, we derive

$$\mathbf{e}_n^{(n)}\mathbf{1}_n^{\mathrm{T}}\widetilde{A}z_0 = \sigma a_{n,n}\mathbf{e}_{n,n}^{(n)}\mathbf{e}_n^{(n)} = \sigma a_{n,n}\mathbf{e}_n^{(n)}. \tag{12}$$

Finally, equalizing Eqs. (11) and (12) results in $(P_r - \widetilde{Q}_r - \mathbf{e}_n^{(n)}\ z_0^{\mathrm{T}})\widetilde{A}z_0 = \mathbf{0}_n$.

(ii) It is clear that $ZE^{-1}Z^{\mathrm{T}}Az_i = Z\mathbf{e}_r^{(i)} = z_i$ and $\widetilde{Z}_0\widetilde{E}^{-1}\widetilde{Z}_0^{\mathrm{T}}\widetilde{A}z_i = \widetilde{Z}_0\mathbf{e}_r^{(i)} = z_i$ for $i \neq n$. Note that $\widetilde{A}z_i = Az_i$ for all $i = 1, 2, \ldots, r-1$. As a consequence,

$$(P_r - \widetilde{Q}_r)\widetilde{A}z_i = \widetilde{A}\widetilde{Z}_0\widetilde{E}_0^{-1}\widetilde{Z}_0^{\mathrm{T}}\widetilde{A}z_i - AZE^{-1}Z^{\mathrm{T}}Az_i = \widetilde{A}z_i - Az_i = \mathbf{0}_n,$$

for all $i = 1, 2, \ldots, r-1$, which gives immediately $(P_r - \widetilde{Q}_r)\widetilde{A}Z = \mathbf{0}_{n,r-1}$. □

Lemma 6. *Equality* $\widetilde{P}_r A = P_r A$ *holds*.

Proof. We have $(\widetilde{P}_r - P_r)A = \mathbf{0}_{n,n}$, if each row of $P_r - \widetilde{P}_r$ contains the same elements. In other words, after defining $B := (\beta_1,\ \beta_2, \ldots, \beta_n)^{\mathrm{T}}\mathbf{1}_n^{\mathrm{T}}$, it suffices to prove that there exist some parameters $\beta_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$, such that

$$(P_r - \widetilde{Q}_r - B)C = \mathbf{0}_{n,n} \tag{13}$$

is satisfied with C to be an arbitrary invertible matrix. Then, we will obtain immediately $P_r - \widetilde{P}_r = B$, since $\widetilde{Q}_r = \widetilde{P}_r$ holds due to Lemma 1.

The proof is as follows. Take $C = \widetilde{A}[\widetilde{Z}_0\ \widetilde{Y}]$, where $\widetilde{Y} = [z_{r+1}\ z_{r+2}\cdots z_n]$ has the properties that the set $\{z_i,\ i = 0, 1, \ldots, n,\ i \neq r\}$ is linearly independent and $z_0^{\mathrm{T}}\widetilde{A}\widetilde{Y} = \mathbf{0}_{n-r}$ is satisfied. Using Lemma 4, matrix $\widetilde{Y}$ with these properties can always be constructed.

Next, note that from the construction of $\widetilde{Y}$, we can derive $Z^{\mathrm{T}}\widetilde{A}z_j = \mathbf{0}_{r-1}$ resulting in $(P_r - \widetilde{Q}_r)\widetilde{A}\widetilde{Y} = \mathbf{0}_{n,n-r}$. Note also that due to Lemma 5, we have $(P_r - \widetilde{Q}_r)\widetilde{A}Z = \mathbf{0}_{n,r-1}$. In addition, we have also $z_0^{\mathrm{T}}\widetilde{A}Z = \mathbf{0}_r^{\mathrm{T}}$ and $z_0^{\mathrm{T}}\widetilde{A}\widetilde{Y} = \mathbf{0}_{n-r}^{\mathrm{T}}$, using $z_0^{\mathrm{T}}\widetilde{A}\widetilde{Y} = \mathbf{0}_{n-r}$ and $\widetilde{A}z_i = Az_i,\ \forall i = 1, 2, \ldots, n,\ i \neq r$. This yields $B[\widetilde{A}Z\ \widetilde{A}\widetilde{Y}] = \mathbf{0}_{n,n}$. Combining these results, this implies $(P_r - \widetilde{Q}_r - B)[\widetilde{A}Z\widetilde{A}\widetilde{Y}] = \mathbf{0}_{n,n-1}$ for all β_i. Moreover, $(P_r - \widetilde{Q}_r - B)\widetilde{A}z_0 = \mathbf{0}_n$ holds due to Lemma 5 by taking $\beta_1 = 1$ and $\beta_2 = \beta_3 = \cdots = \beta_n = 0$.

Thus, $(P_r - \widetilde{Q}_r - B)C = \mathbf{0}_{n,n}$ with $\beta_1 = 1,\ \beta_2 = \beta_3 = \cdots = \beta_n = 0$ is satisfied and thereby the proof of the lemma has been completed. □

Finally, Theorem 2 shows that the deflated singular system based on A is equal to the deflated variant of the invertible system $\widetilde{A}$, which is a rather unexpected result. The consequence of the theorem is that two different variants of the deflated linear systems can be solved with theoretically the same convergence rate.

Theorem 2. *$\widetilde{P}_r\widetilde{A} = P_r A$ holds for all $\sigma > 0$ and $r \geqslant 1$.*

Proof. Theorem 1, Lemmas 3 and 6 give us the equalities $\widetilde{P}_1\widetilde{A} = A$, $\widetilde{P}_r\widetilde{P}_1 = \widetilde{P}_r$ and $\widetilde{P}_r A = P_r A$, which hold for all $\sigma > 0$ and $r \geqslant 1$. Hence, $\widetilde{P}_r\widetilde{A} = \widetilde{P}_r\widetilde{P}_1\widetilde{A} = \widetilde{P}_r A = P_r A$. □

4. Comparison of the preconditioned deflated singular and invertible matrix

In the previous section we have shown that $\widetilde{P}_r\widetilde{A} = P_r A$ holds. However, in general, the preconditioned variant of this equality is not valid, i.e., $\widetilde{M}^{-1}\widetilde{P}_r\widetilde{A} \neq M^{-1}P_r A$. Recall that $\lim_{\sigma\to 0}\kappa(\widetilde{A}) = \infty$, whereas obviously $\lim_{\sigma\to 0}\kappa_{\text{eff}}(\widetilde{P}_r\widetilde{A}) = \kappa_{\text{eff}}(P_r A)$.

The topic of this section is to show that $\lim_{\sigma\to 0}\kappa_{\text{eff}}(\widetilde{M}^{-1}\widetilde{P}_r\widetilde{A}) = \kappa_{\text{eff}}(M^{-1}P_r A)$ holds, where we restrict ourselves to the IC preconditioners. First, we deal with the comparison of the effective condition numbers of $M^{-1}A$ and $\widetilde{M}^{-1}A$ and thereafter we generalize these results to $M^{-1}P_r A$ and $\widetilde{M}^{-1}\widetilde{P}_r\widetilde{A}$.

The algorithm of computing the IC preconditioner can be found in for instance of [3, Section 10.3.2]. The lower triangular part of the resulting matrix A is L and the IC preconditioner is formed by $M = LL^{\mathrm{T}}$. Analogously, $\widetilde{M} = \widetilde{L}\widetilde{L}^{\mathrm{T}}$ can be computed from $\widetilde{A}$. Obviously, the IC preconditioners of A and $\widetilde{A}$ are the same except the last element, since L and $\widetilde{L}$ differ only in the last element, i.e.,

$$\widetilde{M} - M = \beta \mathbf{e}_n^{(n)}\mathbf{e}_n^{(n)\mathrm{T}}, \quad \beta \in \mathbb{R}. \tag{14}$$

By definition of the IC preconditioner, we derive $m_{n,n} = a_{n,n}$ and $\tilde{m}_{n,n} = \tilde{a}_{n,n}$ and consequently $\beta = \tilde{m}_{n,n} - m_{n,n} = \tilde{a}_{n,n} - a_{n,n} = \sigma a_{n,n}$. This implies

$$\lim_{\sigma\to 0}\beta = \lim_{\sigma\to 0}\sigma a_{n,n} = 0. \tag{15}$$

To prove the main theorem of this section, Lemma 7 is required. It gives information about the eigenvalues after perturbation of matrix G. This lemma is a simplified variant of the original theorem of [16], see also [3, Section 8.7].

Lemma 7. *Let $F \in \mathbb{R}^{n\times n}$ be an SPSD matrix and $G \in \mathbb{R}^{n\times n}$ be an SPD matrix. Let $F - \lambda_i G$ be the symmetric-definite $n \times n$ pencil with $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$. Suppose E_G is a symmetric $n \times n$ matrix that satisfy $\|E_G\|_2^2 < c(F, G)$. Then $F - \mu_i(G + E_G)$ is symmetric-definite with $\mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_n$, satisfying*

$$|\arctan(\lambda_i) - \arctan(\mu_i)| \leqslant \arctan\left(\frac{\|E_G\|_2}{c(F, G)}\right), \quad i = 1, 2, \ldots, n, \tag{16}$$

where the Crawford number $c(F, G)$ of the pencil $F - \lambda G$ is defined by $c(F, G) = \min_{\|x\|_2=1}(x^{\mathrm{T}}Fx)^2 + (x^{\mathrm{T}}Gx)^2 > 0$.

Now it can be proven that the effective condition numbers of $M^{-1}A$ and $\widetilde{M}^{-1}A$ are the same, if $\sigma \to 0$, see Theorem 3.

Theorem 3. *Let M^{-1} and $\widetilde{M}^{-1}$ be the corresponding IC preconditioners to A and $\widetilde{A}$. Then $\lim_{\sigma\to 0}\kappa_{\text{eff}}(\widetilde{M}^{-1}A) = \kappa_{\text{eff}}(M^{-1}A)$.*

Proof. Note first that A is SPSD while both M and $\widetilde{M}$ are SPD matrices. This implies $\lambda_i(M^{-1}A) = \lambda_i(M^{-1/2}AM^{-1/2})$ and $\lambda_i(\widetilde{M}^{-1}A) = \lambda_i(\widetilde{M}^{-1/2}A\widetilde{M}^{-1/2})$ for all $i = 1, 2, \ldots, n$. Therefore, the eigenvalues of both systems $M^{-1}A$ and $\widetilde{M}^{-1}A$ are all real-valued.

Now, the proof consists of three steps.

Step 1: Transforming eigenproblems to generalized eigenproblems. We deal with the eigenproblems

$$M^{-1}Av = \lambda v, \quad \widetilde{M}^{-1}Aw = \mu w. \tag{17}$$

These can be rewritten into $(A - \lambda M)v = 0$ and $(A - \mu\widetilde{M})w = 0$, which are generalized eigenproblems. Due to Eq. (14), expression $M + E_M = \widetilde{M}$ can be derived, where $E_M = \beta\mathbf{e}_n^{(n)}\mathbf{e}_n^{(n)\mathrm{T}}$, $\beta \in \mathbb{R}$ is a symmetric matrix. This gives $\|E_M\|_2 = \max\{|\lambda_1(E_M)|,\ |\lambda_n(E_M)|\} = \beta$.

Step 2: Satisfying conditions of Lemma 7. Note that perturbation matrix E_M is symmetric. Moreover, $\|E_M\|_2^2 < c(A, M)$ holds due to the fact that there exists a parameter $\sigma_0 > 0$ such that for all $\sigma < \sigma_0$ yields $\beta^2 < c(A, M)$, since the Crawford number $c(A, M)$ does obviously not depend on σ. Hence, the conditions of Lemma 7 are satisfied.

Step 3: Application of Lemma 7. Note first that $\lim_{\sigma\to 0}\beta = 0$ from Eq. (15). This implies $\lim_{\sigma\to 0}(\beta/c(A, M)) = (1/c(A, M))\lim_{\sigma\to 0}\beta = 0$, so also

$$\lim_{\sigma\to 0} \arctan\left(\frac{\beta}{c(A, M)}\right) = 0. \tag{18}$$

Now, the eigenvalues of (17) are related by Eq. (16) from Lemma 7, which is $|\arctan(\lambda_i) - \arctan(\mu_i)| \leqslant \arctan(\|E_M\|_2/c(A, M))$. Therefore, applying Eq. (18), this implies $\lim_{\sigma\to 0}\arctan(\lambda_i) = \arctan(\mu_i)$, resulting in $\lim_{\sigma\to 0}\lambda_i = \mu_i$, since the arctan-operator is bijective and continuous. Hence, the theorem follows immediately. □

Next, we compare the effective condition numbers of $M^{-1}P_rA$ and $\widetilde{M}^{-1}\widetilde{P}_r\widetilde{A}$. Recall that both A and P_rA are SPSD matrices. So in particular, we can substitute P_rA into A in Theorem 3. Since $\widetilde{P}_r\widetilde{A} = P_rA$ holds due to Theorem 2, this gives us the next theorem.

Theorem 4. *Let M^{-1} and $\widetilde{M}^{-1}$ be the corresponding IC preconditioners to A and $\widetilde{A}$. Then* $\lim_{\sigma\to 0}\kappa_{\text{eff}}(\widetilde{M}^{-1}\widetilde{P}_r\widetilde{A}) = \kappa_{\text{eff}}(M^{-1}P_rA)$.

Theorem 4 states that although $\widetilde{M}^{-1}\widetilde{P}_r\widetilde{A}$and $M^{-1}P_rA$ are not identical, their effective condition number are equal for sufficiently small perturbation σ. As a result, two different variants of the preconditioned deflated linear systems can be solved with theoretically the same convergence rate.

5. Numerical experiments

In this section we give the results of some numerical experiments. These experiments will illustrate the theoretical results obtained in the previous sections.

We consider the 3-D Poisson problem as given in Eq. (1) with two fluids Λ_0 and Λ_1, see also [15]. Specifically, we consider two-phase bubbly flows with air and water in a unit domain. In this case, ρ is piecewise constant with a relatively high contrast:

$$\rho = \begin{cases} \rho_0 = 1, & \mathbf{x} \in \Lambda_0, \\ \rho_1 = 10^{-3}, & \mathbf{x} \in \Lambda_1, \end{cases}$$

where Λ_0 is water, the main fluid of the flow around the air bubbles, and Λ_1 is the region inside the bubbles. In the first part of the numerical experiments, we choose $m = 2^3 = 8$ bubbles with the same radii. In Fig. 1 one can find the geometry of this test case.

The resulting singular linear system $Ax = b$ and also the invertible linear system $\widetilde{A}x = b$ are ill-conditioned due to the presence of the bubbles. We apply ICCG and DICCG$-r$ to solve the linear system. A random starting vector x_0 and the relative tolerance $\|M^{-1}P(b - A\tilde{x}_k)\|_2/\|M^{-1}b - Ax_0\|_2$ are chosen to be smaller than $\varepsilon = 10^{-8}$. It is easy to see that this choice of relative tolerance for DICCG is equivalent to the relative tolerance of $\|M^{-1}(b - Ax_k)\|_2/\|M^{-1}b - Ax_0\|_2$ for ICCG. We vary the perturbation parameter σ and the number of deflation vectors r in our experiments.

5.1. Results considering the invertible linear systems

The results of the above described test problem with $\widetilde{A}$ can be found in Tables 2 and 3. In the case of ICCG, the results of the singular matrix A are added for comparison.

From this table, one observes immediately that the results of DICCG$-r$ are completely independent of σ, as expected from the previous sections. Furthermore, if $\sigma = 0$ then the original singular problem has been solved. In this case, we see that the required number of iterations for ICCG is equal to the number for DICCG $- 1$ when the problem with arbitrary $\sigma > 0$ is solved. Moreover, note that increasing the number of deflation vectors r leads to a non-decreasing number of iterations for DICCG $- r$. All these observations are in agreement with the theoretical results.

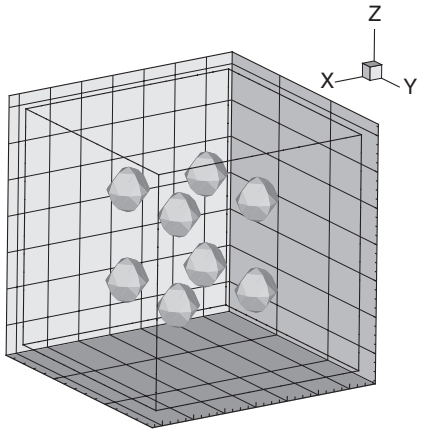

Fig. 1. Geometry of an air–water problem with eight air bubbles in the domain.

Table 2
Number of iterations of ICCG to solve the invertible linear system $\widetilde{A}x = b$ with $m = 2^3$ bubbles

σ	$n = 32^3$	$n = 64^3$
0	118	200
10^{-1}	163	329
10^{-3}	170	350

Table 3
Number of iterations of DICCG–r to solve the invertible linear system $\widetilde{A}x = b$ with $m = 2^3$ bubbles

σ	r	$n = 32^3$	$n = 64^3$
10^{-1}	1	118	200
10^{-1}	2^3	57	106
10^{-1}	4^3	57	106
10^{-3}	1	118	200
10^{-3}	2^3	57	106
10^{-3}	4^3	57	106

Table 4
Number of iterations of ICCG to solve the invertible linear system $\widetilde{A}x = b$ with $m = 3^3$ bubbles and $n = 32^3$

σ	# Iterations
0	160
10^{-1}	234
10^{-3}	254

In Fig. 2(a) one can find a plot of the residuals of ICCG and DICCG $- r$ for our test case. From this figure, it can be observed that ICCG shows an erratic convergence behavior, while DICCG $- r$ converges almost monotonically. Apparently, the approximations of the eigenvectors corresponding to the small eigenvalues are very good. Moreover, we note that the residuals of DICCG $- 2^3$ and DICCG $- 4^3$ coincide. However, from Tables 4 and 5, it appears that if we take $m = 3^3$ bubbles, then the results with $r = 4^3$ is much better than with $r = 2^3$.

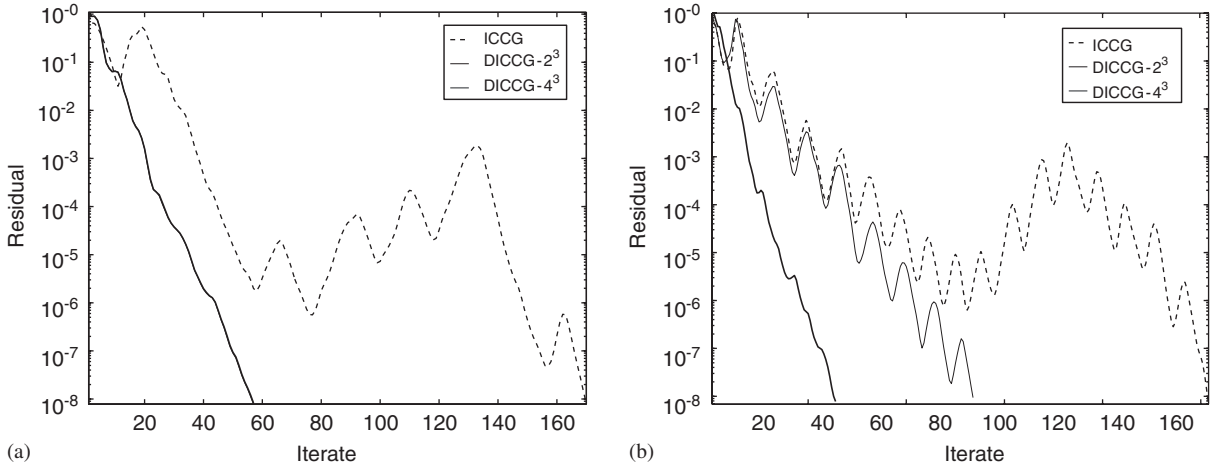

Fig. 2. Plots of the update residuals of ICCG, DICCG $-\ 2^3$ and DICCG $-\ 3^3$ in the test cases with $n = 32^3$ and $\sigma = 10^{-3}$: (a) $m = 2^3$ bubbles; (b) $m = 3^3$ bubbles.

Table 5
Number of iterations of DICCG–r to solve the invertible linear system $\widetilde{A}x = b$ with $m = 3^3$ bubbles and $n = 32^3$

σ	r	# Iterations
10^{-1}	1	160
10^{-1}	2^3	134
10^{-1}	4^3	64
10^{-3}	1	160
10^{-3}	2^3	134
10^{-3}	4^3	64

Table 6
Number of iterations of DICCG $-\ r$ to solve the singular linear system $Ax = b$ and the invertible linear system $\widetilde{A}x = b$ with $m = 2^3$ bubbles

# Iterations		
r	$n = 32^3$	$n = 64^3$
(a) $Ax = b$		
2^3	57	106
4^3	57	106
(b) $\widetilde{A}x = b$ *with* $\sigma = 10^{-1}$		
2^3	57	106
4^3	57	106
(c) $\widetilde{A}x = b$ *with* $\sigma = 10^{-3}$		
2^3	57	106
4^3	57	106

In Fig. 2(b) one can find a plot of the residuals of ICCG and DICCG $-\ r$ for our test case. Now, the residuals of DICCG $-\ 4^3$ decrease more or less monotonically, whereas the residuals of both ICCG and DICCG $-\ 2^3$ are still erratic. Obviously, in this case the small eigenvalues are worse approximated by the deflation technique compared by the case

with $m = 2^3$ bubbles (cf. Fig. 2(a)). The reason is not only the position of the bubbles with respect to the subdomains, but also the increased number of bubbles is more difficult to treat with a constant number of deflation vectors.

5.2. Comparison of the results between singular and invertible linear systems

In the above experiments, we have not yet tested DICCG $- r$ in cases for singular linear systems. In Table 6 we have compared these to the results using the invertible linear systems. Recall that in the singular case, DICCG $- r$ applies $r - 1$ instead of r deflation vectors. Note further that DICCG $- 1$ is not defined in this case.

From Table 6 we observe immediately that the results considering singular matrices (Table 6(a)) are the same as the results of the corresponding test cases with invertible matrices (Tables 6(b) and (c)). Indeed, the two different approaches of the deflation technique considering both the singular and invertible matrices are equivalent, which confirms the theory.

6. Conclusions

In this paper, we have analyzed a singular matrix coming from for instance the Poisson equation. This matrix can be made invertible by modifying the last element, while the solution of the resulting linear system is still the same. Invertibility of the matrix gives several advantages for the iterative solver. The drawback, however, is that the condition number becomes worse compared to the effective condition number of the singular matrix. It appears that this problem with a worse condition number can completely be remedied by applying the deflation technique with just one deflation vector.

Moreover, the deflated singular and invertible matrices have been related to each other. For special choices of the deflation vectors, these matrices are even identical. These results can also be generalized for the preconditioned singular and invertible matrices. This means that the two variants of deflated and preconditioned linear systems results in the same convergence of the iterative process. Results of numerical experiments confirm the theoretical results and show the good performance of the iterative method including the deflation technique. In literature deflation methods have already been proven to be efficient for invertible linear systems. In this paper we have shown that they can also be easily adapted for singular linear systems.

References

[1] P. Bochev, R.B. Lehoucq, On the finite element solution of the pure Neumann problem, SIAM Rev. 47 (1) (2005) 50–66.
[2] J. Frank, C. Vuik, On the construction of deflation-based preconditioners, SIAM J. Sci. Comput. 23 (2001) 442–462.
[3] G.H. Golub, C.F. van Loan, Matrix Computations, third ed., The John Hopkins University Press, Baltimore, MA, 1996.
[4] E.F. Kaasschieter, Preconditioned conjugate gradients for solving singular systems, J. Comput. Appl. Math. 24 (1988) 265–275.
[5] J.J.I.M. van Kan, A second-order accurate pressure correction method for viscous incompressible flow, SIAM J. Sci. Statist. Comput. 7 (1986) 870–891.
[6] M.S. Lynn, W.P. Timlake, The use of multiple deflations in the numerical solution of singular systems of equations with applications to potential theory, SIAM J. Numer. Anal. 5 (2) (1968) 303–322.
[7] L. Mansfield, On the conjugate gradient solution of the Schur complement system obtained from domain decomposition, SIAM J. Numer. Anal. 27 (1990) 1612–1620.
[8] L. Mansfield, Damped Jacobi preconditioning and coarse grid deflation for conjugate gradient iteration on parallel computers, SIAM J. Sci. Statist. Comput. 12 (1991) 1314–1323.
[9] J.A. Meijerink, H.A. van der Vorst, An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix, Math. Comp. 31 (1977) 148–162.
[10] R.B. Morgan, A restarted GMRES method augmented with eigenvectors, SIAM J. Matrix Anal. Appl. 16 (1995) 1154–1171.
[11] R. Nabben, C. Vuik, A comparison of deflation and coarse grid correction applied to porous media flow, SIAM J. Numer. Anal. 42 (2004) 1631–1647.
[12] R. Nabben, C. Vuik, A comparison of deflation and the balancing Neumann–Neumann preconditioner, SIAM J. Sci. Comp. 27 (2006) 1742–1759.
[13] R.A. Nicolaides, Deflation of conjugate gradients with applications to boundary value problems, SIAM J. Matrix Anal. Appl. 24 (1987) 355–365.
[14] S.V. Patankar, Numerical heat transfer and fluid flow, Series in Computer Methods in Mechanics and Thermal Science, McGraw-Hill, New York, 1980.
[15] S.P. van der Pijl, A. Segal, C. Vuik, P. Wesseling, A mass-conserving level-set method for modelling of multi-phase flows, Internat. J. Numer. Meth. Fluids 47 (4) (2005) 339–361.

[16] G.W. Stewart, Perturbation bounds for the definite generalized eigenvalue problem, Linear Algebra Appl. 23 (1979) 69–85.
[17] J.M. Tang, C. Vuik, On the theory of deflation and singular symmetric positive semi-definite matrices, Report 05-06, Delft University of Technology, Department of Applied Mathematical Analysis, ISSN 1389-6520, 2005.
[18] J. Verkaik, Deflated Krylov–Schwarz domain decomposition for the incompressible Navier–Stokes equations on a colocated grid, Master's Thesis, TU Delft, 2003.
[19] J. Verkaik, C. Vuik, B.D. Paarhuis, A. Twerda, The deflation accelerated Schwarz method for CFD, Computational Science-ICCS 2005: Fifth International Conference, Atlanta, GA, USA, May 22–25, 2005, Proceedings Part I, Springer, Berlin, 2005, pp. 868–875.