DELFT UNIVERSITY OF TECHNOLOGY

REPORT 05-06

On the Theory of Deflation and Singular Symmetric Positive Semi-Definite Matrices

J.M. Tang, C. Vuik

ISSN 1389-6520

Reports of the Department of Applied Mathematical Analysis

Delft 2005

On the Theory of Deflation and Singular Symmetric Positive Semi-Definite Matrices

J.M. Tang [1]   C. Vuik [2]

September, 2005

[1] e-mail: j.m.tang@ewi.tudelft.nl
[2] e-mail: c.vuik@ewi.tudelft.nl

Abstract

In this report we give new insights into the properties of invertible and singular deflated and preconditioned linear systems where the coefficient matrices are also symmetric and positive (semi-) definite.

First we prove that the invertible deflated matrix has always a more favorable effective condition number compared to the original matrix. So, in theory, the solution of the deflated linear system converges faster in iterative methods than the original one.

Thereafter, some results are presented considering the singular systems originally from the Poisson equation with Neumann boundary conditions. In practice these linear systems are forced to be invertible leading to a worse (effective) condition number. We show that applying the deflation technique remedies this problem of a worse condition number. Moreover, we derive some useful equalities between the deflated variants of the singular and invertible matrices. Then we prove that the deflated singular matrix has always a more favorable effective condition number compared by the original matrix.

Keywords: singularity, deflation, conjugate gradient method, preconditioning, Poisson equation, spectral analysis, symmetric positive semi-definite matrices.

Contents

Chapter 1

Introduction

In this report we consider the symmetric and positive semi-definite linear system

$$Ax = b, \quad A = [a_{i,j}] \in \mathbb{R}^{n \times n}. \tag{1.1}$$

This linear system (1.1) can be derived for instance after a second-order finite-difference discretization of the 1-D, 2-D or 3-D Poisson equation with Neumann boundary conditions which is

$$\left\{ \begin{array}{rcll} \nabla \cdot \left(\dfrac{1}{\rho(\mathbf{x})} \nabla p(\mathbf{x}) \right) & = & f(\mathbf{x}), & \mathbf{x} \in \Omega, \\ \dfrac{\partial}{\partial \mathbf{n}} p(\mathbf{x}) & = & g(\mathbf{x}), & \mathbf{x} \in \partial\Omega, \end{array} \right. \tag{1.2}$$

where $p, \rho, \mathbf{x}$ and $\mathbf{n}$ denote the pressure, density, spatial coordinates and the unit normal vector to the boundary $\partial\Omega$, respectively.

In this case, A is *singular* and *symmetric positive semi-definite* (SPSP). If $b \in \text{Col } A$ then the linear system (1.1) is consistent and infinite number of solutions exists. Due to the Neumann boundary conditions, the solution x is fixed 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. This means that the absolute value of pressure is not relevant at all, only differences in pressure are meaningful and these are not changed by an arbitrary constant added to the pressure field.

In many computational fluid dynamics packages, see e.g. Patankar [13] and Kaasschieter [4], one would impose an invertible A, denoted by $\widetilde{A}$. This makes solution x unique which can be advantageous in computations, for instance,

- direct solvers like Gauss 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 perturbation of domain

errors whereas the invertible system is always consistent.

- the deflation technique requires an invertible matrix $E := Z^T AZ$ which will be explained later on this report. The choice of Z is straightforward if A is non-singular;

One common way to force invertibility of A in the literature of Computation Fluid Dynamics is to modify the last element of matrix A in the following way:

$$\widetilde{a}_{n,n} = (1 + \sigma) \cdot a_{n,n}, \quad \sigma > 0. \tag{1.3}$$

In fact, a Dirichlet boundary condition is imposed in one point of the domain. Observe that if $\sigma = 0$ would be chosen in the latter expression, then we obtain exactly the original singular problem. This modification results in a symmetric and positive definite linear system

$$\widetilde{A}x = b, \quad \widetilde{A} = [\tilde{a}_{i,j}] \in \mathbb{R}^{n \times n}. \tag{1.4}$$

Presently, direct methods (such as methods based on Cholesky decompositions) are available to solve such a linear system. However, fill-in causes a loss of efficiency for a large and sparse matrix A. For such a case, iterative methods are a better alternative to reduce both memory requirements and computing time.

The most popular iterative method is the Conjugate Gradient (CG) method (see e.g. Golub & Van Loan [2]). It is well-known that the convergence rate of the CG method is bounded as a function of the condition number of matrix $\widetilde{A}$. After k iterations of the CG method, the error is bounded by (cf. Thm. 10.2.6 of [2])

$$||x - x_k||_{\widetilde{A}} \leq 2||x - x_0||_{\widetilde{A}} \left(\frac{\sqrt{\kappa - 1}}{\sqrt{\kappa + 1}} \right)^k, \tag{1.5}$$

where x_0 denotes the starting vector, $\kappa = \kappa(\widetilde{A}) = \lambda_n/\lambda_1$ denotes the spectral condition number of $\widetilde{A}$ and, moreover, the $\widetilde{A}$-norm of x is given by $||x||_{\widetilde{A}} = \sqrt{x^T \widetilde{A} x}$. Therefore, a smaller κ leads to a faster convergence of the CG method.

In practice, it appears that the condition number κ is relatively large, especially if σ is close to 0. Hence, solving (1.4) applying the CG method shows slow convergence to the solution, see also Section 6.7 of [13] and Section 4 of [4]. Instead, a preconditioned system $\widetilde{M}^{-1}\widetilde{A}x = \widetilde{M}^{-1}b$ could be solved, where the SPD preconditioner $\widetilde{M}$ is chosen, such that $\widetilde{M}^{-1}\widetilde{A}$ has a more clustered spectrum or a smaller condition number than that of $\widetilde{A}$. Furthermore, $\widetilde{M}$ must be chosen in such a way that the system $\widetilde{M}y = z$ for every vector z can be solved with less computational work than the original system $\widetilde{A}x = b$. The most easy preconditioner is the so-called diagonal preconditioner defined by $\widetilde{M} = \text{diag}(\widetilde{A})$. A more effective SPD preconditioning strategy in common use is $\widetilde{M} = \widetilde{L}\widetilde{L}^T$ which is an Incomplete Cholesky (IC) factorization of $\widetilde{A}$, defined by Meijerink & Van der Vorst [9]. Since $\widetilde{A}$ is an SPD matrix with $\tilde{a}_{i,j} \leq 0$ for all $i \neq j$, an IC decomposition always exists, see also Kaasschieter [4]. We denote the preconditioned Conjugate Gradient method by PCG and the PCG with the IC preconditioner by ICCG.

In simple practical applications, ICCG shows good convergence relative to other iterative methods (e.g., CG, Gauss-Seidel, SOR). 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 coefficients of the discretized PDE. To remedy the bad convergence of ICCG in more complex models, (eigenvalue) deflation techniques are proposed, originally by Nicolaides [14]. The idea of deflation is to project the extremely large or small eigenvalues of $\widetilde{M}^{-1}\widetilde{A}$ to zero. This leads to a faster convergence of the iterative process, due to Expression (1.5) and due to the fact that the CG method can handle matrices with zero-eigenvalues, see also [4].

The deflation technique has been exploited by several other authors, e.g., Mansfield [7,8], Morgan [10], Vuik *et al.* [1,20,24–26]. A detailed treatment of deflation can also be found in a previous report of the author (Tang [17]). The deflation matrix is defined by

$$\widetilde{P} = I - \widetilde{A}\widetilde{Z}\widetilde{E}^{-1}\widetilde{Z}^T, \quad \widetilde{E} = \widetilde{Z}^T\widetilde{A}\widetilde{Z}, \quad \widetilde{Z} \in \mathbb{R}^{n \times r}, \quad r < n, \tag{1.6}$$

which will be treated more specifically in the next chapter. The resulting linear system which has to be solved is

$$\widetilde{P}\widetilde{M}^{-1}\widetilde{A}x = \widetilde{P}\widetilde{M}^{-1}b. \tag{1.7}$$

In this report, we will concentrate on this latter equation. In particular, we will focus on the deflated-preconditioned system $\widetilde{P}\widetilde{M}^{-1}\widetilde{A}$.

1.1 Objectives of this Report

First we start with comparing the condition number of $\widetilde{M}^{-1}\widetilde{A}$ and the effective condition number of its deflated variant $\widetilde{P}\widetilde{M}^{-1}\widetilde{A}$. It is of importance to show that extending an original preconditioned system with the deflation technique never deteriorates the iterative process.

Moreover, it is known and it will also be shown in this report (Chapter 3) that forcing invertibility of A leads to a worse condition number, i.e.,

$$\kappa(\widetilde{A}) \geq \kappa_{\text{eff}}(A), \tag{1.8}$$

where κ and κ_{eff} denote the standard and effective condition numbers, respectively. As a consequence, the convergence of the CG method applied to the system with A is theoretically faster than with $\widetilde{A}$. In practice, this is indeed the case and it holds also for the preconditioned CG method. In this report, we investigate this issue for the deflated variants of the invertible matrix $\widetilde{M}^{-1}\widetilde{A}$ and singular matrix $M^{-1}A$. Therefore, the effective condition numbers

$\kappa_{\text{eff}}(\widetilde{P}\widetilde{M}^{-1}\widetilde{A})$ and $\kappa_{\text{eff}}(PM^{-1}A)$ will be treated and compared. In addition, 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 in the application of the deflation technique for singular systems. Most articles about deflation, e.g. [1, 7, 8, 10, 20, 24–26], deal only with invertible systems. Applications of deflation to singular systems are described in less articles, see for instance Lynn & Timlake [6] and Verkaik *et al.* [18, 19]. In these articles, some suggestions have been done how to handle singular systems in the deflation technique, but the underlying theory has not yet been developed.

1.2 Outline of this Report

In Chapter 2 and 3 we introduce some notations, assumptions, definitions and preliminary results which will be required through this report.

Chapter 4 deals with the comparison of $\kappa(\widetilde{M}^{-1}\widetilde{A})$ and $\kappa_{\text{eff}}(\widetilde{P}\widetilde{M}^{-1}\widetilde{A})$ for a general invertible SPD matrix $\widetilde{A}$. Moreover, we have seen that forcing invertibility leads to a worse condition number. It will be shown that applying the deflation technique remedies this problem. In the subsequent chapters, we assume A and $\widetilde{A}$ to be matrices from the Poisson equation. In Chapter 5 the proof is given of the equality $\widetilde{P}\widetilde{A} = PA$, which is an unexpected result. Thereafter, in Chapter 6 this is generalized for $\widetilde{P}\widetilde{M}^{-1}\widetilde{A}$ and $PM^{-1}A$, where the diagonal and the Incomplete Cholesky preconditioners are considered. Moreover, a comparison of $\kappa(M^{-1}A)$ and $\kappa_{\text{eff}}(PM^{-1}A)$ will be made in that chapter.

Results of numerical experiments will be presented in Chapter 7 to illustrate the theory. We will end the report with some conclusions in Chapter 8.

Chapter 2

Notations, Assumptions and Definitions

In this chapter some notations, definitions and assumptions will be presented which will be used through this paper.

2.1 Notations of Standard Matrices and Vectors

We first define the following notations for standard matrices and vectors:

$$\begin{array}{lcl}
\mathbf{1}_{p,q} & := & p \times q \text{ unit matrix;} \\
\mathbf{1}_{p} & := & \text{column of } \mathbf{1}_{p,q}; \\
\mathbf{0}_{p,q} & := & p \times q \text{ zero matrix;} \\
\mathbf{0}_{p} & := & \text{column of } \mathbf{0}_{p,q}; \\
\mathbf{e}_{p}^{(r)} & := & r\text{-th column of the } p \times p \text{ identity matrix } I; \\
\mathbf{e}_{p,q}^{(r)} & := & p \times q \text{ matrix with } q \text{ identical columns } \mathbf{e}_{p}^{(r)},
\end{array}$$

with $p, q, r \in \mathbb{N}$.

2.2 Assumptions for Matrices A and $\widetilde{A}$

Through this paper, the $n \times n$ matrices A and $\widetilde{A}$ can be arbitrary chosen provided that they satisfy some assumptions which are given below. First we start with matrix A.

Assumption 2.1. *Matrix A is singular, symmetric and positive semi-definite (SPSD). Moreover, the algebraic multiplicity of the zero-eigenvalue of A is equal to one.*

Assumption 2.2. *Matrix A satisfies $A \cdot \boldsymbol{1}_n = \boldsymbol{0}_n$.*

Next, we give the definition of matrix $\widetilde{A}$.

Definition 2.1. *Let $A = [a_{i,j}]$ be given, which satisfies Assumption 2.1 and 2.2. Then $\widetilde{A} = [\tilde{a}_{i,j}]$ is defined by*

$$\widetilde{a}_{n,n} = (1+\sigma) \cdot a_{n,n}, \quad \sigma > 0, \tag{2.1}$$

and for the other indices i and j

$$\tilde{a}_{i,j} = a_{i,j}. \tag{2.2}$$

Some consequences of this definition can be found in the following two corollaries.

Corollary 2.1. *Matrix $\widetilde{A}$ is invertible, symmetric and positive definite (SPD).*

Corollary 2.2. *Matrix A satisfies $\widetilde{A} \cdot \mathbf{1}_n = \sigma a_{n,n} \cdot \mathbf{e}_n^{(n)}$.*

2.3 Definitions of the Deflation Matrices

In this section the deflation matrices will be defined, but we start with the deflation subspace matrices.

2.3.1 Deflation Subspace Matrices $Z, \widetilde{Z}$ and $\widetilde{Z}_0$

Let the computational domain Ω be divided into open subdomains $\Omega_j,\ j = 1, 2, \ldots, r$, such that $\Omega = \cup_{j=1}^{r} \overline{\Omega}_j$ and $\cap_{j=1}^{r} \Omega_j = \emptyset$ 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:

$$(z_j)_i := \begin{cases} 0, & x_i \in \Omega_h \setminus \overline{\Omega}_{h_j}; \\ 1, & x_i \in \Omega_{h_j}, \end{cases} \tag{2.3}$$

where x_i is a grid point in the discretized domain Ω_h. Define also

$$z_0 = \mathbf{1}_n, \tag{2.4}$$

then it automatically satisfies

$$z_0 \in \operatorname{span}\left\{z_1,\ z_2,\ \ldots,\ z_r\right\}. \tag{2.5}$$

Next, we define the so-called deflation subspace matrices $Z, \widetilde{Z}$ and $\widetilde{Z}_0$ below.

Definition 2.2. *Matrices Z, $\widetilde{Z}$ and $\widetilde{Z}_0$ are defined as follows:*

- $Z := [z_1\ z_2\ \cdots\ z_{r-1}] \in \mathbb{R}^{n \times (r-1)}$*;*
- $\widetilde{Z} := [z_1\ z_2\ \cdots\ z_{r-1}\ z_r] \in \mathbb{R}^{n \times r}$*;*
- $\widetilde{Z}_0 := [z_1\ z_2\ \cdots\ z_{r-1}\ z_0] \in \mathbb{R}^{n \times r}$*.*

Therefore, matrix Z is equal to $\widetilde{Z}$ and $\widetilde{Z}_0$ without their last column, i.e.,

$$\widetilde{Z} = [Z \quad z_r], \quad \widetilde{Z}_0 = [Z \quad z_0]. \tag{2.6}$$

In addition, we also obtain

$$\widetilde{Z} \cdot \mathbf{1}_r = \mathbf{1}_n. \tag{2.7}$$

2.3.2 Deflation Matrices P_r, $\widetilde{P}_r$ and $\widetilde{Q}_r$

The deflation matrices are given below.

Definition 2.3. *Matrices P_r, $\widetilde{P}_r$ and $\widetilde{Q}_r$ are defined as follows:*

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

In the latter expressions, I is the $n \times n$ identity matrix. Moreover, the index 'r' is added as subscript in $P_r, \widetilde{P}_r$ and $\widetilde{Q}_r$ to emphasize the value of r. Note further that $\widetilde{Z}^TA\widetilde{Z}$ is singular, while $E := Z^TAZ$ is invertible so that E^{-1} exists. In addition, also $\widetilde{E}^{-1} := (\widetilde{Z}^T\widetilde{A}\widetilde{Z})^{-1}$ and $\widetilde{E}_0^{-1} := (\widetilde{Z}_0^T\widetilde{A}\widetilde{Z}_0)^{-1}$ always exist, since both $\widetilde{Z}$, $\widetilde{Z}_0$ and $\widetilde{A}$ are full-ranked so also $\widetilde{E}$ and $\widetilde{E}_0$ are full-ranked, see e.g. Horn & Johnson [3].

As special case of $\widetilde{P}_r$ and $\widetilde{Q}_r$, we can take $r = 1$ which leads to

$$\widetilde{Q}_1 = \widetilde{P}_1 = I - \widetilde{A}z_0\widetilde{E}^{-1}z_0^T, \tag{2.8}$$

since $\widetilde{Z} = \widetilde{Z}_0 = z_0$ for $r = 1$. Note that, in contrast to $\widetilde{P}_1$ and $\widetilde{Q}_1$, matrix P_1 does not exist since Z is not defined in this case.

2.4 Eigenvalues and Effective Condition Numbers

Through this report, the eigenvalues λ_i of each arbitrary symmetric $n \times n$ matrix are always ordered increasingly, i.e.,

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n. \tag{2.9}$$

Next, let B be an arbitrary $n \times n$ SPSD matrix with rank $n - r$, so that $\lambda_1 = \ldots = \lambda_r = 0$. Note that all eigenvalues of B are real-valued due to the symmetry of B. Then its effective condition number $\kappa_{\text{eff}}(B)$ are defined as follows:

$$\kappa_{\text{eff}}(B) = \frac{\lambda_n(B)}{\lambda_{r+1}(B)}. \tag{2.10}$$

Since B is singular, $\kappa(B) = \lambda_n(B)/\lambda_1(B)$ is undefined, so the standard condition number makes no sense for singular matrices. Observe further that for an invertible and symmetric matrix C this yields $\kappa(C) = \kappa_{\text{eff}}(C)$.

As special cases we can write the effective condition numbers for P_rA and $\widetilde{P}_r\widetilde{A}$:

$$\kappa_{\text{eff}}(P_rA) = \frac{\lambda_n(P_rA)}{\lambda_r(P_rA)}, \quad \kappa_{\text{eff}}(\widetilde{P}_r\widetilde{A}) = \frac{\lambda_n(\widetilde{P}_r\widetilde{A})}{\lambda_{r+1}(\widetilde{P}_r\widetilde{A})}. \tag{2.11}$$

Chapter 3

Preliminary Results

In this chapter we give some preliminary results from the theory of functional analysis, linear algebra and deflation.

3.1 Results from Functional Analysis

We first start with giving the definition of orthogonal complement and direct sum in terms of Hilbert spaces and subspaces.

Definition 3.1. *Let $\mathcal{H}$ be a Hilbert space with an arbitrary inner product $\langle \cdot, \cdot \rangle$ and let $\mathcal{Z}$ be a closed subspace of $\mathcal{H}$. Then the orthogonal complement $\mathcal{Y}$ of $\mathcal{Z}$ is defined by*

$$\mathcal{Y} = \{ y \in \mathcal{H} \mid \langle z, y \rangle = 0 \quad \forall z \in \mathcal{Z} \} . \tag{3.1}$$

In other words, $\mathcal{Z}$ is the subspace orthogonal to $\mathcal{Y}$. Therefore, the orthogonal complement $\mathcal{Y}$ is also often denoted by $\mathcal{Z}^{\perp}$.

Definition 3.2. *Let $\mathcal{X}$ be a vector space and let Y and Z be subspaces of $\mathcal{X}$. Then, $\mathcal{X}$ is said to be the direct sum of $\mathcal{Y}$ and $\mathcal{Z}$, written*

$$\mathcal{X} = \mathcal{Y} \oplus \mathcal{Z}, \tag{3.2}$$

if each $x \in \mathcal{X}$ has a unique representation

$$x = y + z, \tag{3.3}$$

where $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$.

In other words, the direct sum of two subspaces $\mathcal{Y}$ and $\mathcal{Z}$ is the sum of subspaces in which $\mathcal{Y}$ and $\mathcal{Z}$ have only the zero element in common.

Next, using Definitions 3.1 and 3.2 we can derive Theorem 3.1 which says that the union of the subspaces $\mathcal{Y}$ and $\mathcal{Z}$ is exactly $\mathcal{H}$.

Theorem 3.1. *Let $\mathcal{H}, \mathcal{Y}$ and $\mathcal{Z}$ be defined as in Definition 3.1. Then*

$$\mathcal{H} = \mathcal{Y} \oplus \mathcal{Z}. \tag{3.4}$$

Proof. The proof can be found in any elementary functional analysis book, see e.g. pp. 146–147 of Kreyszig [5]. □

Note that $\mathcal{H} = \mathbb{R}^n$ with the standard vector inner product is an Hilbert space and that in this case dim $\mathcal{Y}+$ dim $\mathcal{Z} = n$. This means that if $\mathcal{Z} = \mathbb{R}^r$ with $r < n$ then $\mathcal{Y} = \mathbb{R}^{n-r}$. Moreover, it is easy to see that $\mathcal{Y}$ and $\mathcal{Z}$ are both closed subspaces of $\mathbb{R}^n$, see also [5].

3.2 Results from Linear Algebra

In the following we denote by $\lambda_i(B)$ the eigenvalues of a symmetric $n \times n$ matrix $B = [b_{i,j}]$. Recall that these eigenvalues are ordered increasingly.

Moreover, the p-norm and Frobenius norm for matrices are defined by

$$||B||_F := \sqrt{\sum_{i,j=1}^{n} b_{i,j}^2}, \quad ||B||_p := \sup_{x \neq 0} \frac{||Bx||_p}{||x||_p}. \tag{3.5}$$

In particular, for symmetric matrices the 2-norm satisfies

$$||B||_2 := \sup_{x \neq 0} \frac{||Bx||_2}{||x||_2} = \max \{ \ |\lambda_1(B)| \ , \ |\lambda_n(B)| \ \} . \tag{3.6}$$

It is known that: $||B||_2 \leq ||B||_F \leq \sqrt{n} \cdot ||B||_2$.

Next, we mention well-known properties of the eigenvalues of symmetric matrices which can be found in Section 8.1.2 of Golub & Van Loan [2].

Theorem 3.2. *Let B and $B + E$ be $n \times n$ symmetric matrices. Then*

(i) $\sum_{i=1}^{n} [\ \lambda_i(B+E) - \lambda_i(B) \ ]^2 \leq ||E||_F^2$;

(ii) $\lambda_k(B) + \lambda_1(E) \leq \lambda_k(B+E) \leq \lambda_k(B) + \lambda_n(E), \quad k = 1, 2, \ldots, n$;

(iii) $|\lambda_k(B+E) - \lambda_k(B)| \leq ||E||_2, \quad k = 1, 2, \ldots, n.$

Property (ii) is known as the Wielandt-Hoffman theorem. With the help of this theorem we can immediately derive the following corollary.

Corollary 3.1. *Let A and $\widetilde{A}$ be defined as in Chapter 2. Then the following statements hold:*

(i) $\lim_{\sigma \to 0} \kappa(\widetilde{A}) = \infty$*;*

(ii) if $\sigma a_{n,n} \leq \lambda_2(A)$ *then* $\kappa_{\mathit{eff}}(A) \leq \kappa(\widetilde{A})$*;*

(iii) there exists a $\sigma_0 > 0$ *such that for all* $\sigma < \sigma_0$

$$\kappa_{\mathit{eff}}(A) \leq \kappa(\widetilde{A}) \tag{3.7}$$

hold.

Proof. (i) Taking $B = A$ and $B + E = \widetilde{A}$ in Theorem 3.2(ii) leads to

$$E = \begin{pmatrix} \varnothing & \\ & \\ & \sigma a_{n,n} \end{pmatrix},$$

resulting in

$$\lambda_1(E) = \ldots = \lambda_{n-1}(E) = 0, \quad \lambda_n(E) = \sigma a_{n,n}.$$

As a result of Theorem 3.2(ii) we obtain

$$\lambda_k(A) \leq \lambda_k(\widetilde{A}) \leq \lambda_k(A) + \sigma a_{n,n}, \quad k = 1, 2, \ldots, n.$$

In particular, we have

$$0 \leq \lambda_1(\widetilde{A}) \leq \sigma a_{n,n}, \quad \lambda_n(A) \leq \lambda_n(\widetilde{A}) \leq \lambda_n(A) + \sigma a_{n,n}.$$

This implies

$$\lim_{\sigma \to 0} \kappa(\widetilde{A}) = \lim_{\sigma \to 0} \frac{\lambda_n(\widetilde{A})}{\lambda_1(\widetilde{A})} \geq \lim_{\sigma \to 0} \frac{\lambda_n(A)}{\sigma a_{n,n}} = \infty.$$

(ii) Since $\sigma a_{n,n} \leq \lambda_2(A)$ holds, we have

$$0 \leq \lambda_1(\widetilde{A}) \leq \sigma a_{n,n} \leq \lambda_2(A).$$

Then,

$$\kappa(\widetilde{A}) = \frac{\lambda_n(\widetilde{A})}{\lambda_1(\widetilde{A})} \geq \frac{\lambda_n(A)}{\lambda_2(A)} = \kappa_{\text{eff}}(A).$$

(iii) This statement follows immediately from Property (ii). □

Next, the well-known theorem of Gershgorin (see again Section 8.1.2 of [2]) is given.

Theorem 3.3. *Let* B *be an* $n \times n$ *symmetric matrix and* C *be an* $n \times n$ *orthogonal matrix.*

If $C^T AC = D + F$ where $D = diag(d_1, \ldots, d_n)$ and F has zero diagonal entries, then

$$\lambda(A) \subseteq \bigcup_{i=1}^{n} [d_i - r_i \ , \ d_i + r_i], \tag{3.8}$$

where $r_i := \sum_{j=1}^{n} \ |f_{i,j}|$ for $i = 1, 2, \ldots, n$.

Next, given an SPSD matrix $A \in \mathbb{R}^{n\times n}$ and an SPD matrix $B \in \mathbb{R}^{n\times n}$, we consider the eigenproblem

$$B^{-1}Ax = \lambda x, \tag{3.9}$$

which can be rewritten into

$$(A - \lambda B)x = 0, \tag{3.10}$$

where λ and x are an eigenvalue and corresponding eigenvector of $B^{-1}A$, respectively. The latter problem is known as the symmetric-definite generalized eigenproblem and $A - \lambda B$ is called a pencil, see e.g. Section 8.7 of [2]. In this case, λ and x are known as a generalized eigenvalue and generalized eigenvector of the pencil $A - \lambda B$.

Moreover, the Crawford number $c(A, B)$ of the pencil $A - \lambda B$ is defined by

$$c(A, B) = \min_{||x||_2=1} \ (x^T Ax)^2 + (x^T Bx)^2 > 0. \tag{3.11}$$

The following theorem gives information about the eigenvalues after perturbing matrix B. This theorem is a simplified variant of the origin theorem of Stewart [16], see also Section 8.7 of [2].

Theorem 3.4. *Let the symmetric-definite $n \times n$ pencil $A - \lambda_i B$ have generalized eigenvalues satisfying*

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n. \tag{3.12}$$

Suppose E_B is a symmetric $n \times n$ matrix that satisfy

$$||E_B||_2^2 < c(A, B). \tag{3.13}$$

Then $A - \mu_i(B + E_B)$ is symmetric-definite with generalized eigenvalues

$$\mu_1 \leq \mu_2 \leq \ldots \leq \mu_n, \tag{3.14}$$

satisfying

$$|\arctan{(\lambda_i)} - \arctan{(\mu_i)}| \leq \arctan\left(\frac{||E_B||_2}{c(A, B)}\right), \tag{3.15}$$

for $i = 1, 2, \ldots, n$.

Next, it is a well-known property, see for instance p. 13 of Horn & Johnson [3], that the rank of a matrix is unchanged upon left or right multiplication by a non-singular matrix, see the next theorem.

Theorem 3.5. *Suppose B_1 and B_2 are $n \times n$ invertible matrices and C is an $n \times n$ matrix with rank $n-k,\ k<n$. Then*

$$\textit{rank } C = \textit{rank } B_1C = \textit{rank } CB_2 = \textit{rank } B_1CB_2 = n-k. \tag{3.16}$$

As a consequence,

$$\lambda_i(C) = \lambda_i(B_1C) = \lambda_i(CB_2) = \lambda_i(B_1CB_2) = 0, \quad i = 1, 2, \ldots, k. \tag{3.17}$$

Now, we can derive the following corollary.

Corollary 3.2. *Let M and $\widetilde{M}^{-1}$ be SPD matrices and let A be an SPSD matrix with rank $n-k$. Then,*

$$\lambda_i(M^{-1}A) = \lambda_i(\widetilde{M}^{-1}A) = 0, \quad i = 1, 2, \ldots, k, \tag{3.18}$$

where the eigenvalues are sorted increasingly.

Proof. Note that both M and $\widetilde{M}$ are invertible. Then we obtain immediately

$$\text{rank } M^{-1}A = \text{rank } \widetilde{M}^{-1}A = \text{rank } A = k,$$

resulting in

$$\lambda_i(M^{-1}A) = \lambda_i(\widetilde{M}^{-1}A) = 0, \quad i = 1, 2, \ldots, k.$$

□

Next, for two symmetric $n \times n$ matrices A and B, we can write $A \prec B$ if $A - B$ is positive definite. Now we can give the next theorems, which are Theorem 4.3.1 and Theorem 4.3.6 of Horn & Johnson [3].

Theorem 3.6. *Let A, B be SPD with $A \prec B$, then*

$$\lambda_i(A) > \lambda_i(B), \tag{3.19}$$

for all $i = 1, 2, \ldots, n$.

Theorem 3.7. *Let A, B be symmetric and suppose B has rank t with $t \leq r$. Then*

$$\lambda_i(A) \leq \lambda_{i+r}(A+B), \tag{3.20}$$

for all $i = 1, 2, \ldots, n-r$.

Note that from this lemma we obtain also the inequality

$$\lambda_i(A) \leq \lambda_{i+r}(A - B),$$

since $-B$ is still symmetric and rank $-B = t$ still holds.

Next, Theorem 3.8 (Wilkinson [27], pp. 94–97) is given, which is from the perturbation theory for the symmetric eigenvalue problem (see also Th. 8.1.8 of [2]).

Theorem 3.8. *Suppose $B = A + \tau cc^T$ where $A \in \mathbb{R}^{n \times n}$ is symmetric, $c \in \mathbb{R}^n$ has unit 2-norm and $\tau > 0$. Then*

$$\lambda_i(A) \leq \lambda_i(B) \leq \lambda_{i+1}(A), \quad i = 1, 2, \ldots, n-1. \tag{3.21}$$

Moreover, there exist $m_1, m_2, \ldots, m_n \geq 0$ such that

$$\lambda_i(B) = \lambda_i(A) + m_i\tau, \quad i = 1, 2, \ldots, n, \tag{3.22}$$

with $m_1 + m_2 + \ldots + m_n = 1$.

Using this latter theorem, we can derive Corollary 3.3 which generalizes Corollary 3.1.

Corollary 3.3. *Let A and $\widetilde{A}$ be as defined in Chapter 2. Then,*

$$\kappa(\widetilde{A}) \geq \kappa_{\mathit{eff}}(A), \tag{3.23}$$

for all $\sigma \geq 0$.

Proof. Note that

$$\widetilde{A} = A + \tau cc^T,$$

with

$$c = \mathbf{e}_n^{(n)}, \quad \tau = \sigma \cdot a_{n,n}.$$

So, Theorem 3.8 can be applied. We will show that (i) $\lambda_2(A) \geq \lambda_1(\widetilde{A})$ and (ii) $\lambda_n(A) \leq \lambda_n(\widetilde{A})$, then Inequality (3.23) follows immediately.

(i) Proof of $\lambda_2(A) \geq \lambda_1(\widetilde{A})$. From Eq. (3.21) we have

$$\lambda_i(A) \leq \lambda_i(\widetilde{A}) \leq \lambda_{i+1}(A), \quad i = 1, 2, \ldots, n-1,$$

so in particular

$$\lambda_1(A) \leq \lambda_1(\widetilde{A}) \leq \lambda_2(A).$$

(ii) Proof of $\lambda_n(A) \leq \lambda_n(\widetilde{A})$. From Eq. (3.22) we derive

$$\lambda_i(\widetilde{A}) \geq \lambda_i(A), \quad i = 1, 2, \ldots, n,$$

since $m_i\tau \geq 0$ for all i. In particular,

$$\lambda_n(\widetilde{A}) \geq \lambda_n(A).$$

□

3.3 Results from Deflation

We start with Theorem 2.6 and Lemma 2.9 of Nabben & Vuik [11].

Theorem 3.9. *Let $\widetilde{A}$ and $\widetilde{Z}$ be matrices as defined in Chapter 2. Then*

$$\lambda_1(\widetilde{P}\widetilde{A}) = \lambda_2(\widetilde{P}\widetilde{A}) = \ldots = \lambda_r(\widetilde{P}\widetilde{A}) = 0.$$

This means that the algebraic multiplicity of the zero-eigenvalue of $\widetilde{P}\widetilde{A}$ is equal to r.

Theorem 3.10. *Let $\widetilde{A}$ and $\widetilde{Z}$ be as defined in Chapter 2. Let $\widetilde{Z}_1$ and $\widetilde{Z}_2$ have the same properties as $\widetilde{Z}$ and assume Col($\widetilde{Z}_1$)=Col($\widetilde{Z}_2$). Define $\widetilde{E}_1 := \widetilde{Z}_1^T\widetilde{A}\widetilde{Z}_1$ and $\widetilde{E}_2 := \widetilde{Z}_2^T\widetilde{A}\widetilde{Z}_2$. Define also $\widetilde{P}_1 := I - \widetilde{A}\widetilde{Z}_1\widetilde{E}_1^{-1}\widetilde{Z}_1^T$ and $\widetilde{P}_2 := I - \widetilde{A}\widetilde{Z}_2\widetilde{E}_2^{-1}\widetilde{Z}_2^T$. Then*

$$\widetilde{P}_1\widetilde{A} = \widetilde{P}_2\widetilde{A}$$

and hence,

$$\widetilde{P}_1 = \widetilde{P}_2.$$

As a consequence, $\widetilde{P}\widetilde{A}$ is invariant for permutations, scaling and linear combinations of the columns of $\widetilde{Z}$, as long as the column space of $\widetilde{Z}$ does not change.

Theorem 3.10 can be applied on the deflation matrices $\widetilde{P}_r$ and $\widetilde{Q}_r$ which are defined in the previous chapter, see the next corollary.

Corollary 3.4. *Let $\widetilde{A}$, $\widetilde{P}_r$, $\widetilde{Q}_r$, $\widetilde{Z}$ and $\widetilde{Z}_0$ be matrices defined in Chapter 2. Then,*

$$\widetilde{Q}_r = \widetilde{P}_r.$$

Proof. By substituting $P := \widetilde{P}_r$ and $Q := \widetilde{Q}_r$ in Theorem 3.10, we obtain

$$\widetilde{Q}_r = \widetilde{P}_r,$$

since the conditions rank $\widetilde{Z}_0 =$ rank $\widetilde{Z} = r$ and Col $\widetilde{Z}_0 =$ Col $\widetilde{Z}$ are satisfied. □

Subsequently, Theorem 2.10 of [11] is given below.

Theorem 3.11. *Let $\widetilde{A}$ be as defined in Chapter 2. Let $\widetilde{Z}_1 \in \mathbb{R}^{n \times r}$ and $\widetilde{Z}_2 \in \mathbb{R}^{n \times s}$ with rank $\widetilde{Z}_1 = r$ and rank $\widetilde{Z}_2 = s$. Define again $\widetilde{E}_1 := \widetilde{Z}_1^T \widetilde{A} \widetilde{Z}_1$ and $\widetilde{E}_2 := \widetilde{Z}_2^T \widetilde{A} \widetilde{Z}_2$. Define also $\widetilde{P}_1 := I - \widetilde{A}\widetilde{Z}_1 \widetilde{E}_1^{-1} \widetilde{Z}_1^T$ and $\widetilde{P}_2 := I - \widetilde{A} Z_2 \widetilde{E}_2^{-1} \widetilde{Z}_2^T$. If $Col(\widetilde{Z}_1) \subseteq Col(\widetilde{Z}_2)$ then*

$$\lambda_n(\widetilde{P}_1 \widetilde{A}) \geq \lambda_n(\widetilde{P}_2 \widetilde{A}), \quad \lambda_{r+1}(\widetilde{P}_1 \widetilde{A}) \leq \lambda_{s+1}(\widetilde{P}_2 \widetilde{A}).$$

The consequence of the latter theorem is that the effective condition number of $\widetilde{P}\widetilde{A}$ decreases if we increase the number of deflation vectors, see also Corollaries 3.5 and 3.6.

Corollary 3.5. *Let $\widetilde{A}, \widetilde{P}_1, \widetilde{P}_2$ be as in Theorem 3.11. Then*

$$\kappa_{\mathit{eff}}(\widetilde{P}_1 \widetilde{A}) \leq \kappa_{\mathit{eff}}(\widetilde{P}_2 \widetilde{A}).$$

Corollary 3.6. *Let $\widetilde{A}$ be as above. Define $\widetilde{Z}_{(i)} = [z_1 \; z_2 \; \cdots \; z_i]$ for $i = 1, 2, \ldots, r$ with $Col(\widetilde{Z}_{(i)}) \subseteq Col(\widetilde{Z}_{(i+1)})$. Moreover, define*

$$\widetilde{P}_{(i)} = I - \widetilde{A}\widetilde{Z}_{(i)} \widetilde{E}_{(i)}^{-1} \widetilde{Z}_{(i)}^T, \quad \widetilde{E}_{(i)} = \widetilde{Z}_{(i)}^T \widetilde{A} \widetilde{Z}_{(i)}.$$

Then:

$$\lambda_2(\widetilde{P}_{(1)} \widetilde{A}) \leq \lambda_3(\widetilde{P}_{(2)} \widetilde{A}) \leq \ldots \leq \lambda_{r+1}(\widetilde{P}_{(r)} \widetilde{A}),$$

and

$$\lambda_n(\widetilde{P}_{(1)} \widetilde{A}) \geq \lambda_n(\widetilde{P}_{(2)} \widetilde{A}) \geq \ldots \geq \lambda_n(\widetilde{P}_{(r)} \widetilde{A}).$$

This yields

$$\kappa_{\mathit{eff}}(\widetilde{P}_{(1)} \widetilde{A}) \geq \kappa_{\mathit{eff}}(\widetilde{P}_{(2)} \widetilde{A}) \geq \ldots \geq \kappa_{\mathit{eff}}(\widetilde{P}_{(r)} \widetilde{A}) = \kappa_{\mathit{eff}}(\widetilde{P}\widetilde{A}).$$

Finally, we end with Theorem 3.12 which gives a useful property of PA and $\widetilde{P}\widetilde{A}$.

Theorem 3.12. *Let $A, \widetilde{A}, P$ and $\widetilde{P}$ be as defined in Chapter 2. Then, both PA and $\widetilde{P}\widetilde{A}$ are SPSD matrices.*

Proof. We prove $\widetilde{P}\widetilde{A}$ to be SPSD. The proof for PA is analogous.

Note first that

$$\widetilde{A}\widetilde{P}^T = \widetilde{A} - \widetilde{A}\widetilde{Z}\widetilde{E}^{-1}\widetilde{Z}^T\widetilde{A} = \widetilde{P}\widetilde{A},$$

and

$$\widetilde{P}^2 = (I - \widetilde{A}\widetilde{Z}\widetilde{E}^{-1}\widetilde{Z}^T)^2 = I - \widetilde{A}\widetilde{Z}\widetilde{E}^{-1}\widetilde{Z}^T = \widetilde{P}.$$

This yields

$$\widetilde{P}\widetilde{A}\widetilde{P}^T = \widetilde{P}^2\widetilde{A} = \widetilde{P}\widetilde{A}.$$

Then, $\widetilde{P}\widetilde{A}$ is symmetric due to

$$(\widetilde{P}\widetilde{A})^T = \widetilde{A}^T\widetilde{P}^T = \widetilde{A}\widetilde{P}^T = \widetilde{P}\widetilde{A}.$$

Moreover, $\widetilde{P}\widetilde{A}$ is positive semi-definite, since by hypothesis $0 < u^T\widetilde{A}u$ for all $u \neq \mathbf{0}_n$, so in particular,

$$0 < (\widetilde{P}^T u)^T\widetilde{A}(\widetilde{P}^T u) = u^T\widetilde{P}\widetilde{A}\widetilde{P}^T u = u^T\widetilde{P}\widetilde{A}u.$$

for $P^T u \neq \mathbf{0}_n$. Hence,

$$0 \leq u^T\widetilde{P}\widetilde{A}u,$$

for all vectors u, see also Frank & Vuik [1]. □

Chapter 4

Comparison of (Effective) Condition Numbers of Deflated Invertible Matrices

In this chapter, first we will prove that the effective condition number of $\widetilde{P}_r\widetilde{A}$ is always lower than the condition number of $\widetilde{A}$ for all choices of $\widetilde{Z}$, see Theorem 4.1.

Theorem 4.1. *Let $\widetilde{A}$ and $\widetilde{P}_r$ be as defined in Chapter 2. Let Z with rank r be arbitrary. Then the following inequality holds:*

$$\kappa_{\mathit{eff}}(\widetilde{P}_r\widetilde{A}) < \kappa(\widetilde{A}). \tag{4.1}$$

Thereafter we proof that it can be generalized in the case of using an SPD preconditioner $\widetilde{M}$, see Theorem 4.2.

Theorem 4.2. *Let $\widetilde{A}$ and $\widetilde{P}_r$ be as defined in Chapter 2. Let $\widetilde{M}$ be an $n \times n$ SPD matrix. Then the following inequality holds:*

$$\kappa_{\mathit{eff}}(\widetilde{M}^{-1}\widetilde{P}_rA) < \kappa(\widetilde{M}^{-1}\widetilde{A}). \tag{4.2}$$

This chapter is organized as follows. We start with some auxiliary results in Section 4.1, which are needed in the proofs of Theorems 4.1 and Theorem 4.2. Thereafter, in Section 4.2 the proof of Theorem 4.1 is given after showing that the inequalities $\lambda_{r+1}(\widetilde{P}_r\widetilde{A}) \geq \lambda_1(\widetilde{A})$ and $\lambda_n(\widetilde{P}_r\widetilde{A}) < \lambda_n(\widetilde{A})$ hold. Finally, we end up with the proof of Theorem 4.2 in the last section.

Important Remarks

- The results given in these chapters, including Theorems 4.1 and 4.2, are applicable for a larger class of matrices than only for $\widetilde{A}$ as defined in Chapter 2. Matrices $\widetilde{A}$ and $\widetilde{M}$ can be replaced by *arbitrary* SPD matrices.

- In the remainder of this chapter, we omit the index r of $\widetilde{P}_r$ since r is always the same. More important, in this whole chapter we do *not* consider the singular matrices A and M but only the invertible matrices $\widetilde{A}$ and $\widetilde{M}$. For the sake of readibility, we will omit the tildes on $\widetilde{P}$, $\widetilde{A}$ and $\widetilde{M}$ in the following. In other words, through this chapter A is an SPD matrix and furthermore M and P are based on this matrix A.

4.1 Auxiliary Results

A set of lemma's, which are needed to prove Theorems 4.1 and 4.2, are given below.

Lemma 4.1. *Let Q be a projection matrix (i.e., $Q^2 = Q$) and let R be an SPD matrix with dimensions $n \times n$ such that QR is symmetric. Then QR is also SPD.*

Proof. By definition, $u^T R u > 0$ for all vectors u. In particular,

$$(Q^T u)^T R (Q^T u) > 0$$

leading to

$$(Q^T u)^T R (Q^T u) = u^T Q R Q^T u > 0.$$

In other words, $QRQ^T = Q(RQ^T)^T = Q^2 R = QR$ is SPD. □

Lemma 4.2. *Matrix $I - P$ is a projector.*

Proof. By definition, $I - P = AZE^{-1}Z^T$ so that

$$(I-P)^2 = AZE^{-1}Z^T AZE^{-1}Z^T = AZE^{-1}EE^{-1}Z^T = AZE^{-1}Z^T = I - P.$$

□

Next, two simple lemma's are given about the rank of a matrix. Recall that a rank of a matrix A is the dimension of the column space of A.

Lemma 4.3. *Let $u = [u_i]$ and $v = [v_i]$ be vectors with length n. Then rank $uv^T = 1$.*

Proof. We have

$$uv^T = [u_1 \ \cdots \ u_n]^T [v_1 \ \cdots \ v_n] = [v_1 u \ \ v_2 u \ \ \cdots \ \ v_n u].$$

Hence, each column is a multiple of the first column. Indeed rank $uv^T = 1$. □

Lemma 4.4. *Define $T := (I - P)A$ with P and A as defined above. Then T is symmetric.*

Proof. Note first that $T = (I - P)A = -AZE^{-1}Z^T A$. Since

$$T^T = (-AZE^{-1}Z^T A)^T = -A^T ZE^{-T}Z^T A = -AZE^{-1}Z^T A = T,$$

matrix T is symmetric. □

We end with the following lemma which says that the preconditioned A, denoted by $\widehat{A}$, is always symmetric and positive definite.

Lemma 4.5. *Let $\widehat{A} := M^{-1/2}AM^{-1/2}$ with M to be SPD. Then $\widehat{A}$ is SPD.*

Proof. Note first that $M^{-1/2}$ exists since M is symmetric positive definite. Obviously, $M^{-1/2}$ is SPD. Now, matrix $\widehat{A}$ is symmetric since

$$\widehat{A}^T = (M^{-1/2}AM^{-1/2})^T = (M^{-1/2})^T A^T (M^{-1/2})^T = M^{-1/2}AM^{-1/2} = \widehat{A}.$$

Moreover, matrix $\widehat{A}$ is positive definite since by definition, $u^T Au > 0$ for all vectors u and in particular,

$$(M^{1/2}u)^T A(M^{1/2}u) > 0$$

leading to

$$u^T M^{1/2}AM^{1/2}u = v^T \widehat{A}v > 0$$

with $v := M^{1/2}u$. □

4.2 Comparison of the (Effective) Condition Numbers of the Matrices A and PA

In this section, the proof of Theorem 4.1 is given. It consists of three steps.

Step 1: Proof of Inequality $\lambda_n(PA) < \lambda_n(A)$.

Note first that

$$A - PA = VA, \quad V := AZE^{-1}Z^T = I - P.$$

$V = I - P$ is a projector due to Lemma 4.2. Obviously, applying the identity $PA = AP^T$, we have that VA is symmetric. Next, since A is SPD, we obtain that VA is also SPD, by using Lemma 4.1. Therefore, by definition, $A \prec PA$ so that

$$\lambda_i(A) > \lambda_i(PA),$$

by Theorem 3.6. Thus in particular:

$$\lambda_n(A) > \lambda_n(PA).$$

Step 2: Proof of Inequality $\lambda_1(A) \leq \lambda_{r+1}(PA)$.

It suffices to prove $\lambda_1(A) \leq \lambda_{r+1}(P_{(1)}A)$ due to Corollary 3.6. We write $Z_{(1)} = z$ since it consists of exactly one vector.

Note first that

$$P_{(1)}A = A - AzE_{(1)}^{-1}z^T A = A + T.$$

with $T = (I - P_{(1)})A = -AzE_{(1)}^{-1}z^T A$. Moreover, since $E_{(1)}^{-1}$ is a scalar, we write $\alpha := -E_{(1)}^{-1} \in \mathbb{R}$. Hence,

$$T = -AzE_{(1)}^{-1}z^T A = \alpha Azz^T A.$$

Obviously, rank $\alpha Azz^T A =$ rank $Azz^T A$. Furthermore, since A is invertible,

$$\text{rank } Azz^T A = \text{rank } zz^T,$$

from Theorem 3.5. Finally,

$$\text{rank } zz^T = 1,$$

due to Lemma 4.3. In order words,

$$\text{rank } T = 1.$$

Moreover, T is symmetric by applying Lemma 4.4, . Hence, the conditions of Lemma 3.7 have been satisfied. By taking $B = T$ in that lemma, we obtain immediately

$$\lambda_1(A) \leq \lambda_2(P_{(1)}A).$$

Step 3: Proof of Theorem 4.1.

In the previous two steps it has been proved that

$$\lambda_1(A) \leq \lambda_{r+1}(PA), \quad \lambda_n(A) > \lambda_n(PA),$$

for all Z with rank $Z = r$. Hence, this leads to

$$\tilde{\kappa}(PA) < \kappa(A).$$

4.3 Comparison of the (Effective) Condition Numbers of the Matrices $M^{-1}A$ and $M^{-1}PA$

As mentioned in the beginning of this chapter, Theorem 4.1 can be generalized for deflated preconditioned systems $M^{-1}PA$ where M is an SPD matrix. This leads to Theorem 4.2 whose the proof can be found below.

Proof of Theorem 4.2. Let $\widehat{A} := M^{-1/2}AM^{-1/2}$. Then $\widehat{A}$ is SPD from Lemma 4.5.

Note that

$$\kappa_{\text{eff}}(M^{-1}PA) = \kappa_{\text{eff}}(M^{-1/2}PAM^{-1/2}) = \kappa_{\text{eff}}(M^{-1/2}PM^{1/2}\widehat{A}) \tag{4.3}$$

and

$$\kappa(M^{-1}A) = \kappa(M^{-1/2}AM^{-1/2}) = \kappa(\widehat{A}) \tag{4.4}$$

using the fact that $\kappa(B_1B_2) = \kappa(B_2B_1)$ (with the standard 2-norm) for two arbitrary invertible symmetric matrices B_1 and B_2.

Next, define $\widehat{P}$ as

$$\widehat{P} := I - \widehat{A}Y\widehat{E}^{-1}Y^T, \quad \widehat{E} := Y^T\widehat{A}Y$$

with $Y := M^{1/2}Z$. Since $M^{1/2}$ is invertible, Y is of rank r. Note further that

$$E = Z^TAZ = (M^{-1/2}Y)^TAM^{-1/2}Y = Y^T\widehat{A}Y = \widehat{E}.$$

Now we obtain

$$\begin{array}{rcl} M^{-1/2}PM^{1/2} & = & M^{-1/2}(I - AZE^{-1}Z^T)M^{1/2} \\ & = & I - M^{-1/2}AZE^{-1}Z^TM^{1/2} \\ & = & I - \widehat{A}M^{1/2}ZE^{-1}Z^TM^{1/2} \\ & = & I - \widehat{A}Y\widehat{E}^{-1}Y^T \\ & = & \widehat{P}. \end{array}$$

Hence, Equation (4.3) can now be rewritten as

$$\kappa_{\text{eff}}(M^{-1}PA) = \kappa_{\text{eff}}(M^{-1/2}PM^{1/2}\widehat{A}) = \kappa_{\text{eff}}(\widehat{P}\widehat{A}). \tag{4.5}$$

From Theorem 4.1 we know that $\kappa_{\text{eff}}(PA) < \kappa(A)$ for arbitrary Z with rank r and for arbitrary SPD matrix A. In particular we can take $P = \widehat{P}$ and $A = \widehat{A}$, since Y is also of rank r and $\widehat{A}$ is SPD. Therefore we obtain

$$\kappa_{\text{eff}}(\widehat{P}\widehat{A}) < \kappa(\widehat{A}),$$

which is equivalent with

$$\kappa_{\text{eff}}(M^{-1}PA) < \kappa(M^{-1}A).$$

□

Chapter 5

Comparison of Deflated Singular and Invertible Matrices

In this chapter, we first show that the problem with a worse condition number is solved, by applying a very simple and cheap deflation technique with only one deflation vector. More precisely, if $\widetilde{P}_1$ is the deflation matrix with one constant deflation vector based on $\widetilde{A}$, then the deflated matrix $\widetilde{P}_1\widetilde{A}$ will be showed to be identical to the original singular 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. Finally, we will compare the (effective) condition numbers of P_rA and A.

The outline of this chapter is as follows. The equality $\widetilde{P}_1\widetilde{A} = A$ will be proved in Section 5.1. In Section 5.2, a set of lemma's is given which are required in Section 5.3 where we will prove $\widetilde{P}_r\widetilde{A} = P_rA$. In the final section, we show that the effective condition number of P_rA is always smaller than the condition number of A.

5.1 Comparison of $\widetilde{P}_1\widetilde{A}$ and A

Before giving the proof of the equality $\widetilde{P}_1\widetilde{A} = A$, we start this section with Lemma 5.1, where it will be shown that $\widetilde{P}_1$ is the identity matrix except for the last row. In addition, $\widetilde{P}_1$ has the properties that the last column is the zero-column and that the matrix consists of only the values 0, 1 and -1.

Lemma 5.1. *Let A, $\widetilde{A}$ and $\widetilde{P}_1$ be defined as in Chapter 2. Then $\widetilde{P}_1$ has the following* *structure:*

$$\widetilde{P}_1 = \begin{bmatrix} 1 & & & & \varnothing \\ & 1 & & & \\ \varnothing & & \ddots & & \\ & & & 1 & \\ -1 & -1 & \cdots & -1 & 0 \end{bmatrix}. \tag{5.1}$$

Proof. For the case of $r = 1$, obviously $Z = z_0$ is a vector and hence, E is a scalar. Therefore, we can rewrite Eq. (2.8) in the following way:

$$\widetilde{P}_1 = I - \alpha \widetilde{A} \mathbf{1}_{n,n}, \tag{5.2}$$

where $\alpha := E^{-1} = 1/E \in \mathbb{R}$ is equal to

$$\alpha = \frac{1}{z_0^T \widetilde{A} z_0} = \frac{1}{\sigma \cdot a_{n,n}},$$

where we have used Corollary (2.2). From this corollary, we obtain also immediately

$$\widetilde{A} \mathbf{1}_{n,n} = \sigma \cdot a_{n,n} \cdot \mathbf{e}_{n,n}^{(n)}, \tag{5.3}$$

resulting in

$$\alpha \widetilde{A} \mathbf{1}_{n,n} = \mathbf{e}_{n,n}^{(n)}.$$

Hence, deflation matrix $\widetilde{P}_1$ as stated in (5.2) is exactly

$$\widetilde{P}_1 = I - \mathbf{e}_{n,n}^{(n)} = \begin{bmatrix} 1 & & & & \varnothing \\ & 1 & & & \\ \varnothing & & \ddots & & \\ & & & 1 & \\ -1 & -1 & \cdots & -1 & 0 \end{bmatrix}.$$

□

Note that Lemma 5.1 still holds if the last row of $\widetilde{A}$ is chosen arbitrary. However, due to the symmetry condition of $\widetilde{A}$, only the last element of $\widetilde{A}$ can be arbitrarily chosen.

Next, applying Lemma 5.1, we obtain the following important theorem which connects the matrices $\widetilde{A}$ and A with the help of the deflation matrix $\widetilde{P}_1$.

Theorem 5.1. *Let $\widetilde{P}_1, A$ and $\widetilde{A}$ be defined as in Chapter 2. Then the following equality holds:*

$$\widetilde{P}_1 \widetilde{A} = A. \tag{5.4}$$

Proof. The exact form of $\widetilde{P}_1$ is given in Lemma 5.1. Obviously, $\widetilde{P}_1\widetilde{A} = A$ for all rows except the last one, since the rows 1 to $n-1$ of $\widetilde{P}_1$ are equal to the corresponding rows of the identity matrix.

The analysis of the last row of $\widetilde{P}_1\widetilde{A}$ is as follows. The sum of each column of A is zero due to symmetry and Assumption 2.2, so we obtain immediately

$$a_{n,j} = -\sum_{i=1}^{n-1} a_{i,j}, \quad \forall j. \tag{5.5}$$

By Definition 2.1 we have

$$\sum_{i=1}^{n-1} \widetilde{a}_{i,j} = \sum_{i=1}^{n-1} a_{i,j} \quad \forall j. \tag{5.6}$$

Combining Eqs. (5.5) and (5.6) yields

$$\begin{aligned} (-1,\ -1,\ \ldots,\ -1,\ 0)\ \cdot\ \widetilde{A} \quad &= \quad \left(-\textstyle\sum_{i=1}^{n-1}\widetilde{a}_{i,1},\ -\sum_{i=1}^{n-1}\widetilde{a}_{i,2},\ \ldots,\ -\sum_{i=1}^{n-1}\widetilde{a}_{i,n-1},\ -\sum_{i=1}^{n-1}\widetilde{a}_{i,n}\right) \\ &= \quad (a_{n,1},\ a_{n,2},\ \ldots,\ a_{n,n-1},\ a_{n,n}). \end{aligned}$$

Hence, the last rows of $\widetilde{P}_1\widetilde{A}$ and A are also equal which proves the theorem. □

The consequence of Theorem 5.1 is that, after applying deflation with $r = 1$, the invertible matrix $\widetilde{A}$ becomes the original singular matrix A. Hence, we see that the perturbation parameter σ disappears completely after deflation. This statement can even be made stronger: the results using this deflation technique are independent of the elements of the last row of matrix $\widetilde{A}$. This is a nice result, since matrix $\widetilde{A}$ has been made invertible with the consequence that the perturbation causes a worse condition number. The deflation technique remedies this problem.

Now, intuitively it is clear that subdomain deflation with $r \geq 1$ acting on A and $\widetilde{A}$ leads to the same convergence results, since the constant deflation vector is in the span of the subdomain deflation vectors. In the remaining of this chapter, we will prove this idea. When it is definitely true, it is a favorable result since we can apply both the singular A and invertible $\widetilde{A}$ in our deflation method leading to the same convergence results.

Example 5.1

To illustrate matrices A, $\widetilde{A}$ and $\widetilde{P}_1$, we now consider a simple example with

$$A = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}, \quad \widetilde{A} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1(1+\sigma) \end{bmatrix}.$$

These matrices satisfy the conditions of A and $\widetilde{A}$ as mentioned above. Therefore:

$$\widetilde{P}_1\widetilde{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1(1+\sigma) \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} = A.$$

The spectra and effective condition numbers of the above matrices can be found in Table 5.1.

Eigenvalues	A	$\widetilde{A}_{\sigma=0.001}$	$\widetilde{A}_{\sigma=10}$	$\widetilde{P}_1\widetilde{A}_{\sigma=0.001} = \widetilde{P}_1\widetilde{A}_{\sigma=10}$
λ_1	0	$5.0 \cdot 10^{-4}$	0.5	0
λ_2	2	2.0	3.4	2
λ_3	4	4.0	22.1	4
$\kappa,\ \kappa_{\text{eff}}$	2	8006.0	41.0	2

Table 5.1: Eigenvalue Analysis of Example 1.

From this table, we see that the condition number of $\widetilde{A}$ can be much larger than the effective condition number of A, which can be remedied by applying deflation with $\widetilde{P}_1$.

5.2 Auxiliary Results

In order to prove Theorem 5.1, we need a set of lemma's which are stated below. The most important lemma's are Lemma 5.4 and Lemma 5.9 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 5.2. *Let a symmetric and invertible $n \times n$ matrix $C = [c_{i,j}]$ have the property that*

$$C \cdot \mathbf{1}_n = \alpha \cdot \mathbf{e}_n^{(n)}. \tag{5.7}$$

Then the elements of the last row and last column of C^{-1} have the same values $1/\alpha$, i.e.,

$$c_{n,j}^{-1} = c_{i,n}^{-1} = \frac{1}{\alpha} \quad \forall\, i,\ j. \tag{5.8}$$

Proof. From Eq. (5.7) we obtain

$$\alpha \cdot C^{-1} \cdot \mathbf{e}_n^{(n)} = \mathbf{1}_n,$$

since C is invertible. This leads to

$$c_{n,1}^{-1} \cdot \alpha = 1 \quad \rightarrow \quad c_{n,1}^{-1} = \frac{1}{\alpha},$$

for all $i = 1, 2, \ldots, n$. Due to the symmetry of C, C^{-1} is also symmetric and Eq. (5.8) holds. □

Lemma 5.3. *Let $\widetilde{P}_r$ be the deflation matrix as defined in Chapter 2. Then the last column of $\widetilde{P}_r$ is always zero, i.e.,*

$$\widetilde{P}_r = \begin{bmatrix} \times & \cdots & \times & 0 \\ \times & \cdots & \times & 0 \\ \vdots & & \vdots & \vdots \\ \times & \cdots & \times & 0 \end{bmatrix}. \tag{5.9}$$

Proof. Due to Eq (2.7) and Assumption 2.2, it is easy to see that the rowsums of $\widetilde{A}\widetilde{Z}$ are all equal to zero except for the last one which is $\sigma \cdot a_{n,n}$, i.e.,

$$\widetilde{A}\widetilde{Z} \cdot \mathbf{1}_r = \widetilde{A} \cdot \mathbf{1}_n = \sigma a_{n,n} \cdot \mathbf{e}_n^{(n)} \tag{5.10}$$

and therefore $\widetilde{E} = \widetilde{Z}^T \widetilde{A}\widetilde{Z}$ has the following property:

$$\begin{aligned} \widetilde{Z}^T \widetilde{A}\widetilde{Z} \cdot \mathbf{1}_r &= \widetilde{Z}^T \cdot \sigma a_{n,n} \cdot \mathbf{e}_n^{(n)} \\ &= \sigma a_{n,n} \cdot \mathbf{e}_r^{(r)}, \end{aligned} \tag{5.11}$$

since it is easy to see that $\widetilde{Z}^T \cdot \mathbf{e}_n^{(n)} = \mathbf{e}_r^{(r)}$.

Next, we show that the last column of $\widetilde{E}^{-1}\widetilde{Z}^T$ contains the same elements, namely $1/(\sigma a_{n,n})$. The last column of $\widetilde{Z}^T$ is $\mathbf{e}_r^{(r)}$, so we only have to focus on the last column of $\widetilde{E}^{-1}$. Since Eq. (5.11) holds and $\widetilde{E}$ is both symmetric and invertible, we can take $C := \widetilde{E}$ in Lemma 5.2. Applying this lemma we obtain that this last column of $\widetilde{E}^{-1}$ is a constant vector with element $1/(\sigma a_{n,n})$.

Hence, the last column of $\widetilde{A}\widetilde{Z}\widetilde{E}^{-1}\widetilde{Z}^T$ is exactly $\mathbf{e}_n^{(n)}$ for all values of σ, since for all $i = 1, 2, \ldots, n$ it yields

$$\begin{aligned} (\widetilde{A}\widetilde{Z}\widetilde{E}^{-1}\widetilde{Z}^T)_{i,n} &= \textstyle\sum_{p=1}^r (\widetilde{A}\widetilde{Z})_{i,p} (\widetilde{E}^{-1}\widetilde{Z}^T)_{p,n} = \dfrac{1}{\sigma a_{n,n}} \textstyle\sum_{p=1}^r (\widetilde{A}\widetilde{Z})_{i,p} \\ &= \dfrac{1}{\sigma a_{n,n}} \widetilde{A}\widetilde{Z}\mathbf{1}_r = \dfrac{1}{\sigma a_{n,n}} \sigma a_{n,n} \mathbf{e}_n^{(n)} = \mathbf{e}_n^{(n)}, \end{aligned}$$

where we have again applied Eq. (5.10). Therefore, the last column of $\widetilde{P}_r = I - \widetilde{A}\widetilde{Z}\widetilde{E}^{-1}\widetilde{Z}^T$ is the zero-vector $\mathbf{0}_n$. □

Lemma 5.4. *Let $P_r, \widetilde{P}_r$ and $\widetilde{P}_1$ be matrices as defined in Chapter 2. Then the following equation holds:*

$$\widetilde{P}_r \widetilde{P}_1 = \widetilde{P}_r. \tag{5.12}$$

Proof. In the proof of Lemma 5.1, see Eq. (5.3), we have already derived

$$\widetilde{A} \cdot \mathbf{1}_n = \gamma \cdot \mathbf{e}_{n,n}^{(n)}, \gamma \in \mathbb{R}.$$

From Lemma 5.3 we have the result that the last column of $\widetilde{P}_r$ is $\mathbf{0}_n$. This implies that

$$\widetilde{P}_r \widetilde{A} \cdot \mathbf{1}_n = \mathbf{0}_n,$$

for all values of parameter σ and γ. Using this fact, we obtain immediately

$$\widetilde{P}_r \widetilde{P}_1 = \widetilde{P}_r (I - \alpha \widetilde{A} \mathbf{1}_n) = \widetilde{P}_r - \alpha \widetilde{P}_r \widetilde{A} \mathbf{1}_n = \widetilde{P}_r.$$

□

Next, we know that $\widetilde{Z}_0 = [z_1 \; z_2 \; \cdots \; z_{r-1} \; z_0] \in \mathbb{R}^{n \times r}$. Define now Y as follows:

$$\widetilde{Y} = [z_{r+1} \; z_{r+2} \; \cdots \; z_{n-1} \; z_n] \in \mathbb{R}^{n \times (n-r)}, \tag{5.13}$$

where $z_{r+1}, \ldots, z_n$ are still undefined.

We can employ the theory in terms of Hilbert spaces and subspaces as given in Definition 3.1 of Chapter 3, by considering the column space of these matrices, so we take

$$\mathcal{Z} = \text{Col } \widetilde{Z}_0, \quad \mathcal{Y} = \text{Col } \widetilde{Y}. \tag{5.14}$$

Subsequently, assume that matrix $\widetilde{A}$ is SPD, so $\widetilde{A} = \widetilde{A}^T$ and $x^T \widetilde{A} x > 0$ for all vectors $x \neq 0$ hold. Consider now the $\widetilde{A}-$inner product

$$\langle z, y \rangle_{\widetilde{A}} = z^T \widetilde{A} y. \tag{5.15}$$

In this case, it can be easily seen that this $\widetilde{A}-$inner product is indeed an inner product, since it satisfies the four conditions:

(i) $\langle z, y \rangle_{\widetilde{A}} = z^T \widetilde{A} y = (z^T \widetilde{A} y)^T = y^T \widetilde{A}^T z = y^T \widetilde{A} z = \langle y, z \rangle_{\widetilde{A}}$;

(ii) $\langle z, x + y \rangle_{\widetilde{A}} = z^T \widetilde{A} (x + y) = z^T \widetilde{A} x + z^T \widetilde{A} y = \langle z, x \rangle_{\widetilde{A}} + \langle z, y \rangle_{\widetilde{A}}$;

(iii) $\langle cz, y \rangle_{\widetilde{A}} = c z^T \widetilde{A} y = c \langle z, y \rangle_{\widetilde{A}}$;

(iv) $\langle z, z \rangle_{\widetilde{A}} = z^T \widetilde{A} z > 0$ and $\langle z, z \rangle_{\widetilde{A}} = z^T \widetilde{A} z = 0$ if $z = 0$,

where c is a scalar and x, y, z are vectors of $\mathcal{H} = \mathbb{R}^n$. Hence, Eq. (3.1) holds in particular for the $\widetilde{A}-$inner product:

$$\text{Col } Y = \left\{ y \in \mathbb{R}^n \mid \langle z, y \rangle_{\widetilde{A}} = 0 \quad \forall z \in \text{Col } \widetilde{Z}_0 \right\}. \tag{5.16}$$

Now, we can choose the vectors $z_{r+1}, \dots, z_n$ such that $\mathcal{Y}$ is the orthogonal complement of $\mathcal{Z}$ in the $\widetilde{A}-$inner product. Then the next lemma follows immediately.

Lemma 5.5. *Let $\widetilde{A}$ and $\widetilde{Z}_0$ be as defined in Chapter 2. Then there exists a matrix $\widetilde{Y} := [z_{r+1} \; z_{r+2} \; \cdots \; z_n]$ such that*

- *the columns of $\widetilde{Y}$ and $\widetilde{Z}_0$ are mutually linear independent, i.e., matrix $X := [\widetilde{Y} \;\; \widetilde{Z}_0]$ is invertible;*
- *the following identity holds:*
$$\widetilde{Z}_0^T \widetilde{A} \widetilde{Y} = \boldsymbol{0}_{r,n-r}. \tag{5.17}$$

Proof. Due to Theorem 3.1, we know that a matrix $\widetilde{Y}$ can be found such that

$$\text{Col } H = \text{Col } \widetilde{Y} \oplus \text{Col } \widetilde{Z}_0, \tag{5.18}$$

where Col $\widetilde{Y}$ is an orthogonal complement of Col $\widetilde{Z}_0$. Then, by definition of the direct sum,

$$X := \left[\widetilde{Y} \;\; \widetilde{Z}_0 \right] \tag{5.19}$$

is a square matrix consisting of linear independent columns. Therefore, X is invertible.

Due to Eq. (5.16), we also know that

$$\text{Col } \widetilde{Y} = \left\{ y \in \mathbb{R}^n \mid \langle w, y \rangle_{\widetilde{A}} = 0 \quad \forall w \in \text{Col } \widetilde{Z}_0 \right\}. \tag{5.20}$$

In particular, for each $w \in \widetilde{Z}_0$ and for each $y \in \widetilde{Y}$ we have

$$\langle w, y \rangle_{\widetilde{A}} = w^T \widetilde{A} y = 0. \tag{5.21}$$

Hence,

$$\widetilde{Z}_0^T \widetilde{A} \widetilde{Y} = \mathbf{0}_{r,n-r}. \tag{5.22}$$

□

The latter lemma can also be proven without applying the theory of the functional analysis, see therefore Appendix A.

Lemma 5.6. *Let $\widetilde{A}$, P_r, $\widetilde{Q}_r$ and z_0 be as defined in Chapter 2. Then,*

$$\left[P_r - \widetilde{Q}_r - \boldsymbol{e}_n^{(n)} \cdot \; z_0^T \right] \cdot \widetilde{A} \cdot z_0 = \boldsymbol{0}_n. \tag{5.23}$$

Proof. Expression

$$Z^T \widetilde{A} z_0 = \mathbf{0}_n \tag{5.24}$$

holds, since $\widetilde{A}z_0 = \widetilde{A}\mathbf{1}_n = \sigma a_{n,n} \cdot \mathbf{e}_n^{(n)}$ and the last row of Z consists of zeros. Note further that $\widetilde{E}_0^{-1}\widetilde{Z}_0^T\widetilde{A}\widetilde{Z}_0 = I$, so in particular

$$\widetilde{E}_0^{-1}\widetilde{Z}_0^T\widetilde{A}z_0 = \mathbf{e}_n^{(n)},$$

resulting in

$$\widetilde{Z}_0\widetilde{E}_0^{-1}\widetilde{Z}_0^T\widetilde{A}z_0 = \widetilde{Z}_0 \cdot \mathbf{e}_n^{(n)} = z_0. \tag{5.25}$$

Applying Eqs. (5.24) and (5.25) and Corollary 2.2, we obtain

$$\begin{aligned} \left[\widetilde{A}\widetilde{Z}_0\widetilde{E}_0^{-1}\widetilde{Z}_0^T - AZE^{-1}Z^T\right] \cdot \widetilde{A}z_0 &= \widetilde{A}\widetilde{Z}_0\widetilde{E}_0^{-1}\widetilde{Z}_0^T \cdot \widetilde{A}z_0 \\ &= \widetilde{A}z_0 = \sigma a_{n,n} \cdot \mathbf{e}_n^{(n)}. \end{aligned} \tag{5.26}$$

Note further that $\mathbf{e}_n^{(n)} \cdot \ \mathbf{1}_n^T = \mathbf{e}_{n,n}^{(n)}$ and $\mathbf{e}_{n,n}^{(n)}\mathbf{e}_n^{(n)} = \mathbf{e}_n^{(n)}$. With the help of these equalities, we can derive

$$\mathbf{e}_n^{(n)} \cdot \ \mathbf{1}_n^T \cdot \widetilde{A}z_0 = \sigma a_{n,n} \cdot \mathbf{e}_{n,n}^{(n)} \cdot \mathbf{e}_n^{(n)} = \sigma a_{n,n} \cdot \mathbf{e}_n^{(n)}. \tag{5.27}$$

Finally, equalizing Eqs. (5.26) and (5.27) results in

$$\left[P_r - \widetilde{Q}_r - \mathbf{e}_n^{(n)} \cdot \ z_0^T\right] \cdot \widetilde{A} \cdot z_0 = \mathbf{0}_n,$$

which completes the proof. □

Lemma 5.7. *Let $\widetilde{A}, Z, P_r$ and $\widetilde{Q}_r$ be as defined in Chapter 2. Then,*

$$\left(P_r - \widetilde{Q}_r\right) \cdot \widetilde{A} \cdot Z = \boldsymbol{0}_{n,r-1}. \tag{5.28}$$

Proof. Note first that

$$E^{-1}Z^TAZ = \widetilde{E}^{-1}\widetilde{Z}_0^T\widetilde{A}\widetilde{Z}_0 = I.$$

Since z_j is a column of both Z and $\widetilde{Z}_0$ for all $j = 1, 2, \ldots, r-1$, this yields

$$E^{-1}Z^TAz_i = \widetilde{E}^{-1}\widetilde{Z}_0^T\widetilde{A}z_i = \mathbf{e}_r^{(i)},$$

so the only non-zero element of this vector is located at the i-th position. Hence,

$$Z \cdot E^{-1}Z^TAz_i = Z \cdot \mathbf{e}_r^{(i)} = z_i$$

and

$$\widetilde{Z}_0 \cdot \widetilde{E}^{-1}\widetilde{Z}_0^T\widetilde{A}z_i = \widetilde{Z}_0 \cdot \mathbf{e}_r^{(i)} = z_i, \quad i \neq n.$$

Next, we consider each column z_i of Z separately. Note that

$$\widetilde{A}z_i = Az_i, \quad \forall\, i = 1, 2, \ldots, r-1,$$

since each last element of z_i is zero. Then,

$$\begin{aligned}
\left(P_r - \widetilde{Q}_r\right) \cdot \widetilde{A} \cdot z_i &= \left(\widetilde{A}\widetilde{Z}_0\widetilde{E}_0^{-1}\widetilde{Z}_0^T - AZE^{-1}Z^T\right) \cdot \widetilde{A}z_i, \\
&= \widetilde{A}\widetilde{Z}_0\widetilde{E}_0^{-1}\widetilde{Z}_0^T\widetilde{A}z_i - AZE^{-1}Z^T\widetilde{A}z_i \\
&= \widetilde{A}\widetilde{Z}_0\widetilde{E}_0^{-1}\widetilde{Z}_0^T\widetilde{A}z_i - AZE^{-1}Z^TAz_i \\
&= \widetilde{A}z_i - Az_i \\
&= \mathbf{0}_n,
\end{aligned}$$

for all $i = 1, 2, \ldots, r-1$. Thus, each column of Eq. (5.28) is the zero-vector $\mathbf{0}_n$ and the lemma has been proved. □

Lemma 5.8. *Let P_r and $\widetilde{P}_r$ be matrices as defined above. Then each row of $P_r - \widetilde{P}_r$ contains the same elements, i.e., there exist some parameters $\beta_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$, such that*

$$P_r - \widetilde{P}_r = (\beta_1,\ \beta_2,\ \cdots,\ \beta_n)^T \cdot\ \boldsymbol{1}_n^T \tag{5.29}$$

is satisfied.

Proof. Define $\widetilde{Q}_r$ as in Chapter 2. Then from Lemma 3.4, we obtain immediately

$$\widetilde{Q}_r = \widetilde{P}_r.$$

Therefore, we are allowed to replace $\widetilde{P}_r$ by $\widetilde{Q}_r$ in this lemma. Now, it suffices to show that

$$\left[P_r - \widetilde{Q}_r - (\beta_1,\ \beta_2,\ \cdots,\ \beta_n)^T \cdot\ \mathbf{1}_n^T\right] \cdot C = \mathbf{0}_{n,n}, \tag{5.30}$$

where C is an arbitrary invertible matrix. Obviously, after multiplication of the latter expression with C^{-1}, we would exactly obtain Eq. (5.29).

The proof is as follows. First take

$$C = \widetilde{A} \cdot \left[\widetilde{Z}_0 \ \ \widetilde{Y}\right],$$

where $\widetilde{Y} = [z_{r+1} \ \ z_{r+2} \ \ \cdots \ \ z_n]$ with the following two properties:

- the set $z_1,\ z_2,\ \ldots,\ z_{r-1},\ z_0,\ z_{r+1},\ \ldots,\ z_n$ is linear independent;
- the equation $\widetilde{Z}_0^T\widetilde{A}\widetilde{Y} = \mathbf{0}_{r,n-r}$ holds.

Using Lemma 5.5, matrix $\widetilde{Y}$ with these properties can always be constructed. As a consequence of $\widetilde{Z}_0^T \widetilde{A}\widetilde{Y} = \mathbf{0}_{r,n-r}$, we obtain in particular

$$z_0^T \widetilde{A}\widetilde{Y} = \mathbf{0}_{n-r}^T. \tag{5.31}$$

Next, observe that

$$\widetilde{A} \cdot [z_1 \;\; z_2 \;\; \cdots \;\; z_{r-1}] = A \cdot [z_1 \;\; z_2 \;\; \cdots \;\; z_{r-1}],$$

since the last element of the vectors $z_i,\; i = 1, 2, \ldots, r-1$ is zero and all columns of A and $\widetilde{A}$ are identical except for the last column. Then, the last element of z_i is zero for all $i = 1, 2, \ldots, n$ except for $i = r$, because

- by construction the last element of the vectors $z_i,\; i = 1, 2, \ldots, r-1$, are zero;
- Equality $\widetilde{A}z_0 = \sigma a_{n,n} \cdot \mathbf{e}_n^{(n)}$ holds due to Corollary (2.2). Combining this with Eq. (5.31) results in zeros for the last element of z_i where $i = r+1, r+2, \ldots, n$. More detailed, $z_0^T \widetilde{A}\widetilde{Y}$ can only be zero if the last row of $\widetilde{Y}$ is zero, since only the last element of $z_0^T \widetilde{A}$ is non-zero.

Therefore, we obtain immediately

$$\widetilde{A}z_i = Az_i, \quad \forall\; i = 1, 2, \ldots, n, \quad i \neq r. \tag{5.32}$$

Next, define $C_0 := \widetilde{A} \cdot z_0,\; C_1 := \widetilde{A} \cdot Z$ and $C_2 := \widetilde{A} \cdot \widetilde{Y}$. Then we have $C = [C_0 \;\; C_1 \;\; C_2]$. To prove Eq. (5.30), we distinguish two cases which will be shown seperately.

- Case 1: $\left[P_r - \widetilde{Q}_r - (\beta_1,\; \beta_2,\; \cdots,\; \beta_n)^T \cdot\; \mathbf{1}_n^T\right] \cdot C_0 = \mathbf{0}_n.$
- Case 2: $\left[P_r - \widetilde{Q}_r - (\beta_1,\; \beta_2,\; \cdots,\; \beta_n)^T \cdot\; \mathbf{1}_n^T\right] \cdot [C_1 \;\; C_2] = \mathbf{0}_{n,n-1}.$

Case 1. The proof is given in Lemma 5.6 by taking

$$\beta_1 = 1, \quad \beta_2 = \beta_3 = \ldots = \beta_n = 0.$$

Case 2. The proof of Case 2 consists of three steps, where all β_i can be *arbitrarily* chosen.

- Using Assumption 2.2, Eqs. (5.31) and (5.32) this gives

  $$z_0^T \widetilde{A}Z = \mathbf{0}_r^T, \quad z_0^T \widetilde{A}\widetilde{Y} = \mathbf{0}_{n-r}^T, \tag{5.33}$$

  or equivalently,

  $$\mathbf{1}_n^T \cdot C_1 = \mathbf{0}_r^T, \quad \mathbf{1}_n^T \cdot C_2 = \mathbf{0}_{n-r}^T. \tag{5.34}$$

  Hence this yields

  $$\left[[\beta_1 \;\; \beta_2 \;\; \cdots \;\; \beta_n]^T \cdot\; \mathbf{1}_n^T\right] \cdot [C_1 \;\; C_2] = \mathbf{0}_{n,n}.$$

- The equality $\left[P_r - \widetilde{Q}_r\right] \cdot C_1 = \mathbf{0}_{n,r-1}$ holds, using Lemma 5.7 and noting that $Z = \widetilde{A}^{-1}C_1$.

- By construction of $\widetilde{Y}$, the identity $\widetilde{Z}_0^T\widetilde{A}z_j = \mathbf{0}_r$ holds for all $j = r+1, r+2, \ldots, n$. Therefore, also $Z^T\widetilde{A}z_j = \mathbf{0}_{r-1}$ holds, since $Z \in \widetilde{Z}$. As a result, we have
$$\left[\widetilde{A}\widetilde{Z}_0\widetilde{E}_0^{-1}\widetilde{Z}_0^T - AZE^{-1}Z^T\right] \cdot \widetilde{A}z_j = \mathbf{0}_n, \quad j = r+1, r+2, \ldots, n,$$
and hence,
$$\left[P_r - \widetilde{Q}_r\right] \cdot C_2 = \mathbf{0}_{n,n-r}.$$

Thus, combining Cases 1 and 2, the following equation is satisfied:

$$\left[P_r - \widetilde{Q}_r - (\beta_1,\ \beta_2,\ \cdots,\ \beta_n)^T \cdot\ \mathbf{1}_n^T\right] \cdot C = \mathbf{0}_{n,n},$$

with $\beta_1 = 1,\ \beta_2 = \beta_3 = \ldots = \beta_n = 0$ and thereby the proof of the lemma has been completed. □

Lemma 5.9. *Let $P_r, \widetilde{P}_r$ and A be as defined as in Chapter 2. Then,*

$$\widetilde{P}_rA = P_rA. \tag{5.35}$$

Proof. In Lemma 5.8, it has been shown that each row i of $B = [b_{i,j}] := (\widetilde{P}_r - P_r)$ has the same elements, i.e.,

$$B = (\beta_1,\ \beta_2,\ \cdots,\ \beta_n)^T \cdot\ \mathbf{1}_n^T, \quad \beta_i \in \mathbb{R},\ i = 1, 2, \ldots, n.$$

Then $BA = (\widetilde{P}_r - P_r)A = \mathbf{0}_{n,n}$ will hold, since each columnsum of A is zero from Assumption 2.2, i.e.,

$$(BA)_{i,j} = \sum_{p=1}^{n} b_{i,p}a_{p,j} = \beta_i \sum_{p=1}^{n} a_{p,j} = \beta_i \cdot 0 = 0.$$

□

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

After giving the lemma's and their proofs in the previous section, the main theorem and its proof will be presented in this section. Theorem 5.2 shows that the deflated singular system based on A is equal to the deflated variant of the invertible system $\widetilde{A}$. This is a rather unexpected result, since Z consists of one vector less compared to $\widetilde{Z}$.

Theorem 5.2. *Let $P_r, \widetilde{P}_r, A$ and $\widetilde{A}$ be matrices as defined in Chapter 2. Then,*

$$\widetilde{P}_r \widetilde{A} = P_r A, \tag{5.36}$$

for all $\sigma > 0$ and $r \geq 1$.

Proof. By applying Theorem 5.1, Lemma 5.4 and Lemma 5.9, we obtain the following three equalities:

$$\widetilde{P}_1 \widetilde{A} = A, \quad \widetilde{P}_r \widetilde{P}_1 = \widetilde{P}_r, \quad \widetilde{P}_r A = P_r A, \tag{5.37}$$

which hold for all $\sigma > 0$ and $r \geq 1$. Hence,

$$\widetilde{P}_r \widetilde{A} = \widetilde{P}_r \widetilde{P}_1 \widetilde{A} = \widetilde{P}_r A = P_r A.$$

□

We illustrate Theorem 5.2 and its corresponding lemma's in Example 5.2.

Example 5.2

Let

$$A = \begin{bmatrix} 1 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 1 \end{bmatrix}, \quad \widetilde{A} = \begin{bmatrix} 1 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 1(1+\sigma) \end{bmatrix}.$$

Obviously, A is SPSD and $A \cdot \mathbf{1}_n = \mathbf{0}_n$ holds, whereas $\widetilde{A}$ is SPD and $A \cdot \mathbf{1}_n = \sigma \cdot \mathbf{e}_n^{(n)}$ holds.

Constructing Z and $\widetilde{Z}$ with $r = 2$ leads to

$$Z = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \widetilde{Z} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

Now we can derive some auxiliary matrices:

$$AZ = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix}, \quad \widetilde{A}\widetilde{Z} = \begin{bmatrix} 0 & 0 \\ 1 & -1 \\ -1 & 1 \\ 0 & \sigma \end{bmatrix},$$

and

$$E = Z^T A Z = 1, \quad \widetilde{E} = \widetilde{Z}^T \widetilde{A} \widetilde{Z} = \begin{bmatrix} 1 & -1 \\ -1 & 1+\sigma \end{bmatrix},$$

which result in

$$E^{-1} = 1, \quad \widetilde{E}^{-1} = \frac{1}{\sigma}\begin{bmatrix} 1+\sigma & 1 \\ 1 & 1 \end{bmatrix}.$$

In this case, we have

$$E^{-1}Z^T = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}, \quad \widetilde{E}^{-1}\widetilde{Z}^T = \frac{1}{\sigma}\begin{bmatrix} 1+\sigma & 1+\sigma & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Moreover,

$$AZE^{-1}Z^T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \widetilde{A}\widetilde{Z}\widetilde{E}^{-1}\widetilde{Z}^T = \frac{1}{\sigma}\begin{bmatrix} 0 & 0 & 0 & 0 \\ -\sigma & -\sigma & 0 & 0 \\ \sigma & \sigma & 0 & 0 \\ \sigma & \sigma & \sigma & \sigma \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ -1 & -1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

and hence,

$$P_2 = I - AZE^{-1}Z^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \widetilde{P}_2 = I - \widetilde{A}\widetilde{Z}\widetilde{E}^{-1}\widetilde{Z}^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ -1 & -1 & -1 & 0 \end{bmatrix}.$$

Note that parameter σ has completely disappeared from the latter expression. Now we can derive the following:

$$P_2A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

and

$$\widetilde{P}_2\widetilde{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ -1 & -1 & -1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 1(1+\sigma) \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix} = P_2A.$$

Thus indeed: $P_2A = \widetilde{P}_2\widetilde{A}$ (Theorem 5.2). Note further that P_2A gives two decoupled Neumann problems, but in general this does not hold for $n > 4$.

Furthermore, we can also show that $\widetilde{P}_2 A = P_2 A$ (Lemma 5.4):

$$\widetilde{P}_2 A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ -1 & -1 & -1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix} = P_2 A.$$

Next, $\widetilde{P}_2 \widetilde{P}_1 = \widetilde{P}_2$ (Lemma 5.9) can be verified:

$$\widetilde{P}_2 \widetilde{P}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ -1 & -1 & -1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ -1 & -1 & -1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ -1 & -1 & -1 & 0 \end{bmatrix} = \widetilde{P}_2.$$

We end up with noting that $P_2 \widetilde{P}_1 \neq P_2$ and $\widetilde{P}_2 \widetilde{A} \neq P_2 \widetilde{A}$.

5.4 Comparison of the Effective Condition Numbers of PA and A

In Chapter 4 we have already proved for the invertible matrix $\widetilde{A}$ that (cf. Eq. (4.1))

$$\kappa_{\text{eff}}(\widetilde{P}_r \widetilde{A}) < \kappa(\widetilde{A}). \tag{5.38}$$

In the next theorem, we will show that such a inequality can be derived for the singular matrix A.

Theorem 5.3. *Let A and P_r be as defined in Chapter 2. Let Z with rank r be arbitrary. Then the following inequality holds:*

$$\kappa_{\mathit{eff}}(P_r A) \leq \kappa_{\mathit{eff}}(A). \tag{5.39}$$

Proof. From Theorems 5.1 and 5.2 we have

$$\widetilde{P}_r \widetilde{A} = P_r A, \quad \widetilde{P}_1 \widetilde{A} = A.$$

This implies

$$\kappa_{\text{eff}}(\widetilde{P}_r \widetilde{A}) = \kappa_{\text{eff}}(P_r A), \quad \kappa_{\text{eff}}(\widetilde{P}_1 \widetilde{A}) = \kappa_{\text{eff}}(A). \tag{5.40}$$

From Corollary 3.6 we know

$$\kappa_{\text{eff}}(\widetilde{P}_r \widetilde{A}) \leq \kappa_{\text{eff}}(\widetilde{P}_1 \widetilde{A}). \tag{5.41}$$

Finally, combining Eqs. (5.40) and (5.41) gives

$$\kappa_{\text{eff}}(P_r A) \leq \kappa_{\text{eff}}(A).$$

□

The generalization of this theorem where a preconditioner is included in (5.39) can be found in the next chapter.

Chapter 6

Comparison of Preconditioned Deflated Singular and Invertible Matrices

In the previous chapter we have shown that $\widetilde{P}_r\widetilde{A} = P_rA$ holds. However, in general, the preconditioned variant of this equality does not hold, i.e., $\widetilde{M}^{-1}\widetilde{P}_r\widetilde{A} \neq M^{-1}P_rA$. Moreover, we have seen in Chapter 3 that $\lim_{\sigma\to 0}\kappa(\widetilde{A}) = \infty$, whereas obviously $\lim_{\sigma\to 0}\kappa_{\text{eff}}(\widetilde{P}_r\widetilde{A}) = \kappa_{\text{eff}}(P_rA)$. The question in this chapter is:

$$\lim_{\sigma\to 0}\kappa_{\text{eff}}(\widetilde{M}^{-1}\widetilde{P}_r\widetilde{A}) = \kappa_{\text{eff}}(M^{-1}P_rA)? \tag{6.1}$$

This is the same as showing that

$$\lim_{\sigma\to 0}\kappa_{\text{eff}}(\widetilde{M}^{-1}P_rA) = \kappa_{\text{eff}}(M^{-1}P_rA) \tag{6.2}$$

holds. We restrict ourselves to the standard diagonal and incomplete Cholesky (IC) preconditioners in our proofs. These are denoted by D and M_{IC} when they are based on A and these are denoted by $\widetilde{D}$ and $\widetilde{M}_{IC}$ when they are based on $\widetilde{A}$.

This chapter is organized as follows. Section 6.1 deals with the comparison of $D^{-1}A$ and $\widetilde{D}^{-1}A$. In Section 6.2 the comparison of $M_{IC}^{-1}A$ and $\widetilde{M}_{IC}^{-1}A$ is given. We generalize these results and comparisons to $D^{-1}P_rA$ and $\widetilde{D}^{-1}P_rA$ and also to $M_{IC}^{-1}P_rA$ and $\widetilde{M}_{IC}^{-1}P_rA$ in Section 6.3. In the last section, we end with a comparison of the (effective) condition numbers of $M^{-1}P_rA$ and $M^{-1}P_rA$ for general preconditioner M.

6.1 Comparison of $D^{-1}A$ and $\widetilde{D}^{-1}A$

The diagonal preconditioners D and $\widetilde{D}$ are defined as follows:

$$D := \text{diag}\,(A)\,, \quad \widetilde{D} := \text{diag}\left(\widetilde{A}\right). \tag{6.3}$$

Note that both D and $\widetilde{D}$ are SPD matrices. Next, we define the diagonal-preconditioned systems

$$Q := D^{-1/2}AD^{-1/2}, \quad \widetilde{Q} := \widetilde{D}^{-1/2}A\widetilde{D}^{-1/2}. \tag{6.4}$$

Note that both systems are singular systems since A is singular, i.e., $\lambda_1(Q) = \lambda_1(\widetilde{Q}) = 0$.

Now, in this section we compare Q and $\widetilde{Q}$ and their spectra. For Krylov iterative methods this is equivalent to comparing $D^{-1}A$ and $\widetilde{D}^{-1}A$, since the spectra in both cases are identical.

6.1.1 Perturbation Matrix E

Matrices $D^{-1/2}$ and $\widetilde{D}^{-1/2}$ can be written out:

$$D^{-1/2} = \begin{pmatrix} \frac{1}{\sqrt{a_{1,1}}} & & & \\ & \frac{1}{\sqrt{a_{2,2}}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{a_{n,n}}} \end{pmatrix}, \quad \widetilde{D}^{-1/2} = \begin{pmatrix} \frac{1}{\sqrt{a_{1,1}}} & & & \\ & \frac{1}{\sqrt{a_{2,2}}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{(1+\sigma)\cdot a_{n,n}}} \end{pmatrix}.$$

Then, $\widetilde{D}^{-1/2} = D^{-1/2}R = RD^{-1/2}$ where $R = \mathrm{diag}\left(1,\ 1,\ \ldots,\ 1, \frac{1}{\sqrt{1+\sigma}}\right)$. Next, $RAR - A$ is as follows:

$$\begin{aligned} RAR - A &= \begin{pmatrix} a_{1,1} & \cdots & a_{1,n-1} & \frac{a_{1,n}}{\sqrt{1+\sigma}} \\ \vdots & & \vdots & \vdots \\ a_{n-1,1} & \cdots & a_{n-1,n-1} & \frac{a_{n-1,n}}{\sqrt{1+\sigma}} \\ \frac{a_{n,1}}{\sqrt{1+\sigma}} & \cdots & \frac{a_{n,n-1}}{\sqrt{1+\sigma}} & \frac{a_{n,n}}{1+\sigma} \end{pmatrix} - A \\ &= \begin{pmatrix} & & & a_{1,n}\left(\frac{1}{\sqrt{1+\sigma}} - 1\right) \\ & \varnothing & & \vdots \\ & & & a_{n-1,n}\left(\frac{1}{\sqrt{1+\sigma}} - 1\right) \\ a_{n,1}\left(\frac{1}{\sqrt{1+\sigma}} - 1\right) & \cdots & a_{n,n-1}\left(\frac{1}{\sqrt{1+\sigma}} - 1\right) & a_{n,n}\left(\frac{1}{1+\sigma} - 1\right) \end{pmatrix}. \end{aligned}$$

Furthermore, perturbation matrix E is defined by $E = [e_{i,j}] := \widetilde{Q} - Q$ and can be worked out in the following way:

$$\begin{aligned} E &= \widetilde{Q} - Q \\ &= \widetilde{D}^{-1/2}A\widetilde{D}^{-1/2} - D^{-1/2}AD^{-1/2} \\ &= D^{-1/2}(RAR - A)D^{-1/2}. \end{aligned}$$

And hence,

$$\begin{aligned}
E &= D^{-1/2}(RAR-A)D^{-1/2}\\
&= \begin{pmatrix} & & & \frac{a_{1,n}}{\sqrt{a_{1,1}a_{n,n}}}\left(\frac{1}{\sqrt{1+\sigma}}-1\right)\\ & \varnothing & & \vdots\\ & & & \frac{a_{n-1,n}}{\sqrt{a_{n-1,n-1}a_{n,n}}}\left(\frac{1}{\sqrt{1+\sigma}}-1\right)\\ \frac{a_{n,1}}{\sqrt{a_{n,n}a_{1,1}}}\left(\frac{1}{\sqrt{1+\sigma}}-1\right) & \cdots & \frac{a_{n,n-1}}{\sqrt{a_{n,n}a_{n-1,n-1}}}\left(\frac{1}{\sqrt{1+\sigma}}-1\right) & \left(\frac{1}{1+\sigma}-1\right) \end{pmatrix}\\
&= \begin{pmatrix} & & & \frac{a_{1,n}\left(1-\sqrt{1+\sigma}\right)}{\sqrt{a_{1,1}a_{n,n}(1+\sigma)}}\\ & \varnothing & \vdots & \\ & & & \frac{a_{n-1,n}\left(1-\sqrt{1+\sigma}\right)}{\sqrt{a_{n-1,n-1}a_{n,n}(1+\sigma)}}\\ \frac{a_{n,1}\left(1-\sqrt{1+\sigma}\right)}{\sqrt{a_{1,1}a_{n,n}(1+\sigma)}} & \cdots & \frac{a_{n,n-1}\left(1-\sqrt{1+\sigma}\right)}{\sqrt{a_{n-1,n-1}a_{n,n}(1+\sigma)}} & \frac{-\sigma}{1+\sigma} \end{pmatrix}.
\end{aligned}$$

Observe that E is symmetric, so we obtain

$$E = \begin{pmatrix} & & & e_{n,1}\\ & \varnothing & \vdots & \\ & & & e_{n,n-1}\\ e_{n,1} & \cdots & e_{n,n-1} & e_{n,n} \end{pmatrix}, \tag{6.5}$$

where

$$e_{n,n} = \frac{-\sigma}{1+\sigma}, \quad e_{n,j} = \frac{a_{n,j}\left(1-\sqrt{1+\sigma}\right)}{\sqrt{a_{j,j}a_{n,n}(1+\sigma)}}, \quad j = 1,\ldots,n-1. \tag{6.6}$$

This perturbation matrix E has the following properties:

- only the last row and column contain non-zero elements, more stronger: only m elements (independent of the sizes of E) located in the last row and column are non-zero elements where m is the number of diagonals in A;

- the last element of E is negative, while the other non-zero elements are all positive;

- if $\sigma = 0$ then we have $E = 0$ as expected;

- E is indefinite which can be derived in several ways, for instance with Theorem 3.2(ii) by taking $k = 1$:

$$\lambda_1(Q) + \lambda_1(E) \leq \lambda_1(\widetilde{Q}) \leq \lambda_1(Q) + \lambda_n(E) \quad \rightarrow \quad \lambda_1(E) \leq 0 \leq \lambda_n(E).$$

Moreover, we can derive $||E||_F^2$:

$$\begin{aligned} ||E||_F^2 &= e_{n,n}^2 + 2\sum_{j=1}^{n-1} e_{n,j}^2 \\ &= \left(\frac{-\sigma}{1+\sigma}\right)^2 + 2\left(\frac{1-\sqrt{1+\sigma}}{\sqrt{a_{n,n}(1+\sigma)}}\right)^2 \sum_{p=1}^{n-1}\left(\frac{a_{n,p}}{\sqrt{a_{p,p}}}\right)^2, \end{aligned}$$

where we have used the fact that E is symmetric. We work this latter expression out:

$$||E||_F^2 = \frac{\sigma^2}{(1+\sigma)^2} + \frac{2(1-\sqrt{1+\sigma})^2}{a_{n,n}(1+\sigma)} \sum_{p=1}^{n-1} \frac{a_{n,p}^2}{a_{p,p}}. \tag{6.7}$$

This can be simplified to

$$||E||_F^2 = \frac{\sigma^2 + \theta(1+\sigma)(1-\sqrt{1+\sigma})^2}{(1+\sigma)^2}, \tag{6.8}$$

with

$$\theta = \frac{2}{a_{n,n}} \sum_{p=1}^{n-1} \frac{a_{n,p}^2}{a_{p,p}}. \tag{6.9}$$

Note that, if A consists of m nonzero diagonals, then the sum in Eq. (6.9) consists of m terms.

Moreover, note that

$$\frac{\sigma^2}{(1+\sigma)^2} = \mathcal{O}(\sigma^2), \quad \frac{(1-\sqrt{1+\sigma})^2}{(1+\sigma)} = \mathcal{O}(\sigma^2), \quad \sigma \to 0,$$

which can be easily derived with Taylor expansions. Therefore, Eq. (6.8) can be rewritten into

$$||E||_F^2 = \mathcal{O}(\sigma^2) + \theta \cdot \mathcal{O}(\sigma^2) = (1+\theta) \cdot \mathcal{O}(\sigma^2).$$

Example 6.1

We consider the singular matrix A derived from the 3-D Poisson equation with Neumann boundary conditions as described in Chapter 1. Furthermore, it is assumed that there is only one fluid in the neighbourhood of the last grid point. Then,

- A and $\widetilde{A}$ consist of 7 non-zero diagonals;

- the non-diagonal elements of the last row are all the same. Let $\alpha := a_{n,n}$, then for all nonzero $a_{n,j}$ we have $a_{n,j} = -\alpha/3$.

In this case, simple analysis can be done for θ to estimate the order of this parameter. We get

$$\theta = \frac{2}{a_{n,n}} \sum_{p=1}^{n-1} \frac{a_{1,p}^2}{a_{p,p}} = \frac{2}{\alpha} \sum_{p=1}^{n-1} \frac{(\alpha/3)^2}{\alpha} = \frac{6(\alpha/3)^2}{\alpha^2} = \frac{2}{3}.$$

Hence, parameter θ is of $\mathcal{O}(1)$ and we obtain

$$||E||_F^2 = \mathcal{O}(\sigma^2).$$

6.1.2 Eigenvalue analysis of Q and $\widetilde{Q}$

To deal with the spectra of Q and $\widetilde{Q}$, we apply Theorem 3.2(i) which gives

$$\sum_{i=1}^{n} \left[\ \lambda_i(\widetilde{Q}) - \lambda_i(Q) \ \right]^2 \leq ||E||_F^2 = \frac{\sigma^2 + \theta(1+\sigma)(1-\sqrt{1+\sigma})^2}{(1+\sigma)^2}. \qquad (6.10)$$

Observe that the RHS of Eq. (6.10) does not depend on n. Moreover, due to this expression and the fact that $(\lambda_i(\widetilde{Q}) - \lambda_i(Q))^2 \geq 0$, we have that $\lambda_i(\widetilde{Q}) \to \lambda_i(Q)$ for $\sigma \to 0$ which holds for all i. In other words, for sufficiently small σ, the eigenvalues of Q and $\widetilde{Q}$ resemble each other very well, since the RHS of (6.10) approaches zero, see also Table 6.1.

σ	$\lvert\lvert E\rvert\rvert_F^2$		
1	$2.5 \cdot 10^{-1}$	+	$8.6 \cdot 10^{-2}\ \theta$
10^{-3}	$1.0 \cdot 10^{-6}$	+	$2.5 \cdot 10^{-7}\ \theta$
10^{-6}	$1.0 \cdot 10^{-12}$	+	$2.5 \cdot 10^{-13}\ \theta$

Table 6.1: Value of $||E||_F^2$ for several choices of σ.

In the next section we investigate the condition numbers of Q and $\widetilde{Q}$ to complete the whole spectral analysis.

6.1.3 Condition Numbers of Q and $\widetilde{Q}$

In Theorem 6.1 we prove that the condition numbers of Q and $\widetilde{Q}$ are more or less the same if σ is sufficiently small.

Theorem 6.1. *Let Q and $\widetilde{Q}$ as defined above. Then,*

$$\lim_{\sigma \to 0} \kappa_{\mathit{eff}}(\widetilde{Q}) = \kappa_{\mathit{eff}}(Q). \qquad (6.11)$$

Proof. The proof consists of four parts.

- *Application of Theorem 3.2.* Applying Theorem 3.2(ii) to Q and $\widetilde{Q}$ leads to

$$\lambda_k(Q) + \lambda_1(E) \leq \lambda_k(\widetilde{Q}) \leq \lambda_k(Q) + \lambda_n(E), \quad k = 1, 2, \ldots, n.$$

In particular we have

$$\lambda_2(Q) + \lambda_1(E) \leq \lambda_2(\widetilde{Q}) \leq \lambda_2(Q) + \lambda_n(E)$$

and

$$\lambda_n(Q) + \lambda_1(E) \leq \lambda_n(\widetilde{Q}) \leq \lambda_n(Q) + \lambda_n(E)$$

resulting in

$$\frac{\lambda_n(Q) + \lambda_1(E)}{\lambda_2(Q) + \lambda_n(E)} \leq \kappa_{\text{eff}}(\widetilde{Q}) = \frac{\lambda_n(\widetilde{Q})}{\lambda_2(\widetilde{Q})} \leq \frac{\lambda_n(Q) + \lambda_n(E)}{\lambda_2(Q) + \lambda_1(E)}. \tag{6.12}$$

- *Proof of $\kappa_{\mathit{eff}}(\widetilde{Q}) \leq \kappa_{\mathit{eff}}(Q)$.* First we give bounds for $\lambda_1(E)$ and $\lambda_n(E)$. Note first that

$$e_{n,n} - \sum_{p=1}^{n-1} e_{n,p} < -e_{n,j}, \quad \forall\, j = 1, \ldots, n,$$

using Eqs. (6.5) and (6.6). Now, we apply the theorem of Gershgorin (see Theorem 3.3) which leads to

$$e_{n,n} - \sum_{p=1}^{n-1} e_{n,p} \leq \lambda_1(E)$$

and

$$\lambda_n(E) \leq \max\left\{ e_{n,n} + \sum_{p=1}^{n-1} e_{n,p} \;,\; e_{n,n} \;,\; e_{n,n-1} \;,\; \ldots \;,\; e_{n,1} \right\}.$$

This gives

$$\frac{\lambda_n(Q) + \lambda_n(E)}{\lambda_2(Q) + \lambda_1(E)} \leq \frac{\lambda_n(Q) + \max\left\{ e_{n,n} + \sum_{p=1}^{n-1} e_{n,p} \;,\; e_{n,n} \;,\; e_{n,n-1} \;,\; \ldots \;,\; e_{n,1} \right\}}{\lambda_2(Q) + e_{n,n} - \sum_{p=1}^{n-1} e_{n,p}}.$$

Obviously, if $\sigma \rightarrow 0$, then

$$\frac{\lambda_n(Q) + \max\left\{ e_{n,n} + \sum_{p=1}^{n-1} e_{n,p} \;,\; e_{n,n} \;,\; e_{n,n-1} \;,\; \ldots \;,\; e_{n,1} \right\}}{\lambda_2(Q) + e_{n,n} - \sum_{p=1}^{n-1} e_{n,p}} \rightarrow \frac{\lambda_n(Q)}{\lambda_2(Q)},$$

since each term of E approaches zero for small σ. Hence, for the RHS inequality of Eq. (6.12) this implies

$$\lim_{\sigma \rightarrow 0} \kappa_{\text{eff}}(\widetilde{Q}) = \lim_{\sigma \rightarrow 0} \frac{\lambda_n(\widetilde{Q})}{\lambda_2(\widetilde{Q})} \leq \lim_{\sigma \rightarrow 0} \frac{\lambda_n(Q) + \lambda_n(E)}{\lambda_2(Q) + \lambda_1(E)} = \frac{\lambda_n(Q)}{\lambda_2(Q)} = \kappa_{\text{eff}}(Q). \tag{6.13}$$

- *Proof of $\kappa_{\mathit{eff}}(\widetilde{Q}) \geq \kappa_{\mathit{eff}}(Q)$.* We can repeat the whole procedure as given in the above proof of $\kappa_{\text{eff}}(\widetilde{Q}) \leq \kappa_{\text{eff}}(Q)$. This yields

$$\frac{\lambda_n(Q) + e_{n,n} - \sum_{p=1}^{n-1} e_{n,p}}{\lambda_2(Q) + \max\left\{e_{n,n} + \sum_{p=1}^{n-1} e_{n,p}\ ,\ e_{n,n}\ ,\ e_{n,n-1}\ ,\ \ldots\ ,\ e_{n,1}\right\}} \leq \frac{\lambda_n(Q) + \lambda_n(E)}{\lambda_2(Q) + \lambda_1(E)}.$$

Obviously, if $\sigma \rightarrow 0$, then

$$\frac{\lambda_n(Q) + \max\left\{e_{n,n} + \sum_{p=1}^{n-1} e_{n,p}\ ,\ e_{n,n}\ ,\ e_{n,n-1}\ ,\ \ldots\ ,\ e_{n,1}\right\}}{\lambda_2(Q) + e_{n,n} - \sum_{p=1}^{n-1} e_{n,p}} \rightarrow \frac{\lambda_n(Q)}{\lambda_2(Q)}.$$

Therefore, for the left inequality of Eq. (6.12) we get

$$\lim_{\sigma \rightarrow 0} \frac{\lambda_n(Q) + \lambda_n(E)}{\lambda_2(Q) + \lambda_1(E)} = \frac{\lambda_n(Q)}{\lambda_2(Q)} = \kappa_{\text{eff}}(Q) \leq \lim_{\sigma \rightarrow 0} \kappa_{\text{eff}}(\widetilde{Q}) = \lim_{\sigma \rightarrow 0} \frac{\lambda_n(\widetilde{Q})}{\lambda_2(\widetilde{Q})}. \qquad (6.14)$$

- *Proof of $\kappa_{\mathit{eff}}(\widetilde{Q}) = \kappa_{\mathit{eff}}(Q)$.* By combining Eqs. (6.13) and (6.14), we obtain finally

$$\lim_{\sigma \rightarrow 0} \kappa_{\text{eff}}(\widetilde{Q}) = \kappa_{\text{eff}}(Q).$$

□

6.2 Comparison of $M_{IC}^{-1}A$ and $\widetilde{M}_{IC}^{-1}A$

In the previous section, we have based the analysis on $D^{-1}A$ and $\widetilde{D}^{-1}A$. In this section, we consider $M_{IC}^{-1}A$ and $\widetilde{M}_{IC}^{-1}A$. For the sake of simplicity we omit the underscript 'IC' through this section, so the IC-preconditioners are denoted by $M = [m_{i,j}]$ and $\widetilde{M} = [\tilde{m}_{i,j}]$. Below, we will show that

$$\lim_{\sigma \rightarrow 0} \ \kappa_{\text{eff}}(\widetilde{M}^{-1}A) = \kappa_{\text{eff}}(M^{-1}A) \qquad (6.15)$$

hold.

6.2.1 Connection between M and $\widetilde{M}$

The algorithm of computing the IC-preconditioner can be found in for instance Section 10.3.2 of Golub and Van Loan [2] and for completeness this algorithm is also given below.

The lower triangular part of the resulting matrix A is L and the IC-preconditioner is formed by $M = LL^T$. Analogously, $\widetilde{M} = \widetilde{L}\widetilde{L}^T$ can be formed from $\widetilde{A}$.

Obviously, the IC-preconditioners of A and $\widetilde{A}$ are the same except the last element, since

Algorithm 1 Construction of the IC-preconditioner for Matrix A

```
Given matrix $A = [a_{i,j}]$
for $k = 1, \ldots, n$ do
  $a_{k,k} := \sqrt{a_{k,k}}$
  for $i = k + 1, \ldots, n$ do
    if $a_{i,k} \neq 0$ then
      $a_{i,k} = a_{i,k}/a_{k,k}$
    end if
  end for
  for $j = k + 1, \ldots, n$ do
    for $i = j, \ldots, n$ do
      if $a_{i,j} \neq 0$ then
        $a_{i,j} = a_{i,j} - a_{i,k}/a_{j,k}$
      end if
    end for
  end for
end for
```

L and $\widetilde{L}$ differ only in the last element. In other words:

$$\widetilde{M} - M = \begin{pmatrix} & & \\ & \varnothing & \\ & & \beta \end{pmatrix}, \quad \beta \in \mathbb{R}. \tag{6.16}$$

Note that, since $m_{n,n} = a_{n,n}$ and $\tilde{m}_{n,n} = \tilde{a}_{n,n}$ hold by definition of the IC-preconditioner and

$$\beta = \tilde{m}_{n,n} - m_{n,n} = \tilde{a}_{n,n} - a_{n,n} = \sigma a_{n,n}, \tag{6.17}$$

we obtain immediately

$$\lim_{\sigma \to 0} \beta = \lim_{\sigma \to 0} \sigma a_{n,n} = 0. \tag{6.18}$$

6.2.2 Condition Numbers of $M^{-1}A$ and $\widetilde{M}^{-1}A$

Below it will be proved that the effective condition numbers of $M^{-1}A$ and $\widetilde{M}^{-1}A$ are the same if the perturbation σ is asymptotically zero, i.e., if $\sigma \to 0$, see Theorem 6.2.

Theorem 6.2. *Let A and $\widetilde{A}$ be matrices as defined in Chapter 2. Let M^{-1} and $\widetilde{M}^{-1}$ be their corresponding IC-preconditioners. Then*

$$\lim_{\sigma \to 0} \kappa_{\mathit{eff}}(\widetilde{M}^{-1}A) = \kappa_{\mathit{eff}}(M^{-1}A). \tag{6.19}$$

Proof. Note first that A is SPSD while both M and $\widetilde{M}$ are SPD matrices. Then,

$$\lambda_i(M^{-1}A) = \lambda_i(M^{-1/2}AM^{-1/2}), \quad \lambda_i(\widetilde{M}^{-1}A) = \lambda_i(\widetilde{M}^{-1/2}A\widetilde{M}^{-1/2}),$$

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

$$M^{-1}Av = \lambda v, \quad \widetilde{M}^{-1}Aw = \mu w, \tag{6.20}$$

which can be rewritten into

$$(A - \lambda M)v = 0, \quad (A - \mu\widetilde{M})w = 0, \tag{6.21}$$

which are generalized eigenproblems, see also Chapter 3. Due to Eq. (6.16), the following expression can be derived:

$$M + E_M = \widetilde{M},$$

where

$$E_M = \begin{pmatrix} & \varnothing & \\ & & \\ & & -\beta \end{pmatrix}, \quad \beta \in \mathbb{R}$$

is a symmetric matrix. This gives

$$||E_M||_2 = \max\{\ |\lambda_1(E_M)|\ ,\ |\lambda_n(E_M)|\ \} = \beta.$$

Step 2: Satisfying Conditions of Theorem 3.4. Before we can apply Theorem 3.4, the corresponding conditions have to be satisfied:

- Perturbation matrix E_M is symmetric;
- The Crawford number $c(A, M)$ does obviously not depend on σ. Obviously, there exists a parameter $\sigma_0 > 0$ such that for all $\sigma < \sigma_0$ yields

  $$\beta^2 < c(A, M). \tag{6.22}$$

  Hence,

  $$||E_M||_2^2 < c(A, M).$$

Step 3: Application of Theorem 3.4. Theorem 3.4 can be applied since all conditions have been satisfied. Note first that $\lim_{\sigma\to 0}\ \beta = 0$ from Eq. (6.18). This implies

$$\lim_{\sigma\to 0}\frac{\beta}{c(A,M)} = \frac{1}{c(A,M)}\lim_{\sigma\to 0}\beta = 0,$$

so also

$$\lim_{\sigma\to 0}\arctan\left(\frac{\beta}{c(A,M)}\right) = 0. \tag{6.23}$$

Now, the eigenvalues of (6.20) are related by Eq. (3.15) from Theorem 3.4, which is

$$|\arctan(\lambda_i) - \arctan(\mu_i)| \leq \arctan\left(\frac{||E_M||_2}{c(A,M)}\right).$$

Therefore, applying Eq. (6.23), 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,

$$\lim_{\sigma\to 0}\kappa_{\text{eff}}(\widetilde{M}^{-1}A) = \lim_{\sigma\to 0}\frac{\lambda_n}{\lambda_2} = \frac{\mu_n}{\mu_2} = \kappa_{\text{eff}}(M^{-1}A). \tag{6.24}$$

□

6.3 Generalization to Deflated Systems

In the previous sections we have shown

$$\lim_{\sigma\to 0}\kappa_{\text{eff}}(\widetilde{D}^{-1}A) = \lim_{\sigma\to 0}\kappa_{\text{eff}}(D^{-1}A), \quad \lim_{\sigma\to 0}\kappa_{\text{eff}}(\widetilde{M}^{-1}A) = \kappa_{\text{eff}}(M^{-1}A). \tag{6.25}$$

Moreover, we have already mentioned that both A and P_rA are SPSD matrices, see also Theorem 3.12. So in particular, we can subsitute P_rA into A in Eq. (6.25), which implies

$$\lim_{\sigma\to 0}\kappa_{\text{eff}}(\widetilde{D}^{-1}P_rA) = \lim_{\sigma\to 0}\kappa_{\text{eff}}(D^{-1}P_rA), \quad \lim_{\sigma\to 0}\kappa_{\text{eff}}(\widetilde{M}^{-1}P_rA) = \kappa_{\text{eff}}(M^{-1}P_rA). \tag{6.26}$$

In other words, the theory given in the previous two sections still holds if we replace A by PA in the whole analysis.

6.4 Comparison of the (Effective) Condition Numbers of $M^{-1}PA$ and $M^{-1}A$

In Chapter 4 we have already proved for non-singular matrices that

$$\kappa_{\text{eff}}(\widetilde{M}^{-1}\widetilde{P}\widetilde{A}) \leq \kappa(\widetilde{M}^{-1}\widetilde{A}).$$

In this section we show that

$$\kappa_{\text{eff}}(M^{-1}PA) \leq \kappa(M^{-1}A), \tag{6.27}$$

also hold for the *singular* matrix A, see the next theorem.

Theorem 6.3. *Let A and P be matrices as defined in Chapter 2. Let M be the corresponding IC-preconditioner of A. Then*

$$\kappa_{\mathit{eff}}(M^{-1}PA) \leq \kappa(M^{-1}A). \tag{6.28}$$

Proof. Note first that M is invertible. With the identities $\widetilde{P}\widetilde{A} = PA$ and $\widetilde{P}_1\widetilde{A} = A$ we can immediately derive

$$\kappa_{\text{eff}}(M^{-1}\widetilde{P}\widetilde{A}) = \kappa_{\text{eff}}(M^{-1}PA), \quad \kappa_{\text{eff}}(M^{-1}\widetilde{P}_1\widetilde{A}) = \kappa_{\text{eff}}(M^{-1}A).$$

So, it suffices to prove the following inequality:

$$\kappa_{\text{eff}}(M^{-1}\widetilde{P}\widetilde{A}) \leq \kappa(M^{-1}\widetilde{P}_1\widetilde{A}).$$

This latter inequality holds due to Theorem 2.12 of Nabben & Vuik [11], which is a generalization of Theorem 3.11. □

Chapter 7

Numerical Experiments

In this chapter we give the results of some numerical experiments which are done by using FORTRAN. These results will illustrate the theoretical results obtained in the previous chapters.

7.1 Problem Setting

We consider the 3-D Poisson problem as given Eq. (1.2) with two fluids Λ_0 and Λ_1. Specifically, we consider two-phase bubbly flows with air and water in an unit domain. In this case, ρ is piecewise constant with a relatively high contrast:

$$\rho = \begin{cases} \rho_0 = 1, & \mathbf{x} \in \Lambda_0, \\ \rho_1 = 10^{-3}, & \mathbf{x} \in \Lambda_1, \end{cases} \tag{7.1}$$

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 Figure 7.1 one can find a plot in the case of such a problem with $m = 8$ bubbles.

7.2 Results of ICCG and DICCG$-k$

Eight bubbles are chosen in the domain (i.e., $m = 8$) in all test cases. The resulting $n \times n$ singular linear system $Ax = b$ and also the resulting $n \times n$ invertible linear system $\widetilde{A}x = b$ are ill-conditioned due to the presence of these bubbles. We apply ICCG and DICCG$-k$ to solve this linear system, where DICCG$-k$ denotes DICCG with k deflation vectors. The relative tolerance [1] of the iterative method is $\epsilon = 10^{-8}$. Moreover, the number of bubbles m and the number of grid points $n = n_x n_y n_z$ are taken constant. We vary the perturbation parameter σ

[1] In Appendix B one can find more details of the relative tolerance and the termination criterions of both ICCG and DICCG.

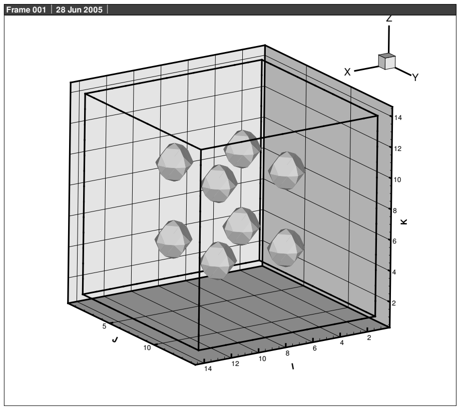

Figure 7.1: Geometry of an air-water problem with eight air bubbles in the domain.

and the number of deflation vectors k in our experiments. The results can be found in Table 7.1.

	# ICCG Iterations	
σ	$n = 32^3$	$n = 64^3$
0	118	200
10^{-1}	163	329
10^{-3}	174	350

		# DICCG$-k$ Iterations	
σ	k	$n = 32^3$	$n = 64^3$
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 7.1: Number of iterations of ICCG and DICCG$-k$ to solve the invertible linear system $\widetilde{A}x = b$ with $m = 2^3$ bubbles.

The results of DICCG$-k$ are completely independent of σ, as expected from the previous chapters. 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. This is in agreement with the theoretical results found in Chapters 5 and 6.

Moreover, note that increasing the number of deflation vectors k leads to a non-decreasing number of iterations for DICCG$-k$, which agrees again the theoretical results found in Chapters 5 and 6.

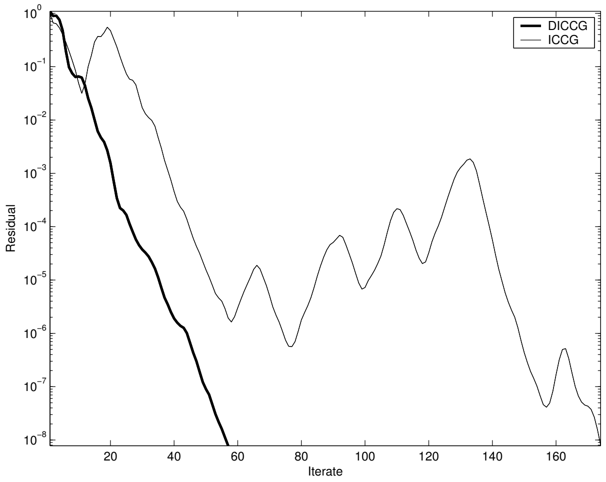

Figure 7.2: Plot of the update residuals of both ICCG and DICCG for the test case with $n = 32^3$, $\sigma = 10^{-3}$ and $k = m = 2^3$.

In Figure 7.2 one can find a plot of the residuals of ICCG and DICCG−k for one test case. From the figure we can see that ICCG shows an erratic convergence behavior, while DICCG converges almost monotonically. Apparently, the approximations of the eigenvectors corresponding to the small eigenvalue are very good. This is due to the fact that for $k = 2^3$ each bubble is in the interior of a subdomain corresponding to one of the 2^3 deflation vectors. In this case, the deflation technique is very efficient, in spite of the relatively low number of deflation vectors. This explains also the same results of DICCG−k for $k = 2^3$ and $k = 4^3$, as can be seen in Table 7.1. It appears that if we take $m = 3^3$ instead of $m = 2^3$ bubbles, then the results with $k = 4^3$ is much better than with $k = 2^3$ (see Table 7.2), since now none of the bubbles is contained in one of the subdomains.

In Figure 7.3 one can find a plot of the residuals of ICCG and DICCG−k for one test case. It can be observed that both plots are very 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. Figure 7.2). 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.

7.3 Results of DICCG−k for Singular Systems

In the above experiments, we have not yet tested DICCG−k in cases for singular linear systems. That will be the subject of this section.

We define the deflation technique for singular systems as in Chapter 2. In this case,

σ	ICCG
0	160
10^{-1}	234
10^{-3}	254

σ	k	DICCG$-k$
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 7.2: Number of iterations of ICCG and DICCG$-k$ to solve the invertible linear system $\widetilde{A}x = b$ with $m = 3^3$ bubbles and $n = 32^3$.

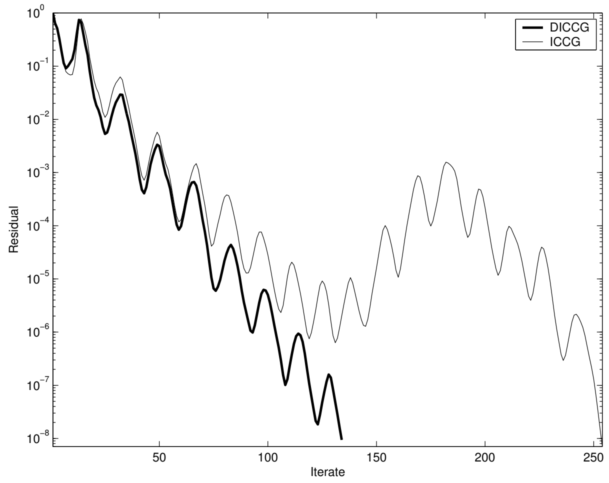

Figure 7.3: Plot of the update residuals of both ICCG and DICCG for the test case with $n = 32^3$, $\sigma = 10^{-3}$, $k = 2^3$ and $m = 3^3$.

for instance DICCG-8 applies 7 instead of 8 deflation vectors. Note that DICCG-1 is not defined in this case. Now, the results can be found in Table 7.3.

From Table 7.3 we observe immediately that these results are the same as the results of the corresponding test cases with invertible matrices. Indeed, the two approaches of the deflation technique for both the singular and invertible matrices are equivalent, as earlier seen in Chapters 5 and 6.

	# Iterations of DICCG$-k$	
k	$n = 32^3$	$n = 64^3$
1	–	–
2^3	57	106
4^3	57	106

Table 7.3: Number of iterations of DICCG$-k$ to solve the singular linear system $Ax = b$.

7.4 Modified Matrix $\widetilde{A}$

From Definition 2.1 we know that $\tilde{a}_{n,n} = (1 + \sigma) \cdot a_{n,n}$. We can also choose other locations on the main diagonal of A to perturb. In fact, we can take

$$\tilde{a}_{i,i} = (1 + \sigma) \cdot a_{i,i}, \quad 1 \le i \le n, \tag{7.2}$$

and leaving the other elements of A untouched. The resulting $\widetilde{A}$ is invertible and will be denoted by $\widetilde{A}^{(i)}$. The question is whether or not the above results will change for various matrices $\widetilde{A}^{(i)}$. Some results can be found in Table 7.4.

i	DICCG-2^3	DICCG-4^3
1	57	57
15	57	57
$n - 15$	57	57
$n - 1$	57	57
n	57	57

Table 7.4: Number of iterations of DICCG$-k$ to solve the invertible linear system $\widetilde{A}^{(i)}x = b$ for $n = 32^2$ and $\sigma = 10^{-3}$.

Obviously, DICCG$-k$ does not depend on the value of i. So, each diagonal element of A can be chosen to perturb in order to obtain an invertible matrix.

Next, we take

$$\tilde{a}_{n,n} = (1 + \sigma) \cdot a_{n,n} + \gamma, \quad \gamma \ge 0, \tag{7.3}$$

and subsequently we do the same analysis as done above. The results can be found in Table 7.5.

For $\gamma < 10^{-4}$ the same results of DICCG$-k$ can be observed, while for $\gamma > 10^{-4}$ the convergence of DICCG$-k$ is slower. This confirms the theory of Chapter 6, since we have seen that only for sufficiently small perturbations in $a_{n,n}$ the eigenvalues of the systems $\widetilde{M}^{-1}\widetilde{P}\widetilde{A}$ and $M^{-1}PA$ are more or less equal.

γ	DICCG-2^3	DICCG-4^3
0	57	57
10^0	57	57
10^2	57	57
10^4	58	58
10^6	73	73

Table 7.5: Number of iterations of DICCG$-k$ to solve the invertible linear system $\widetilde{A}x = b$ for $n = 32^2$ using $\tilde{a}_{n,n} = (1 + \sigma) \cdot a_{n,n} + \gamma$.

7.5 Further Analysis

In this section, we investigate the deflation methods in more detail using numerical experiments. First we consider more severe termination criteria and real residuals, thereafter we investigate matrix Z and we end up with some alternative choices for the deflation vectors.

7.5.1 Termination Criteria and Real Residuals

We have seen that the deflation technique can be applied for both the singular and invertible matrices which results in identical results. In this section, we investigate the real residuals and we test the convergence for more severe termination criterions. The results can be found in Table 7.6 where we use $n = 32^3$.

	Invertible System		Singular System	
ϵ	# Iterations	Real Residual	# Iterations	Real Residual
10^{-8}	57	$3.4 \cdot 10^{-4}$	57	$3.4 \cdot 10^{-4}$
10^{-12}	87	$4.2 \cdot 10^{-8}$	87	$4.1 \cdot 10^{-8}$
10^{-14}	102	$1.4 \cdot 10^{-8}$	102	$1.0 \cdot 10^{-8}$

Table 7.6: Number of iterations of DICCG$-k$ to solve the singular and invertible linear system.

For various ϵ, the results considering the number of iterations and the real residuals are more or less identical in cases of both the singular and invertible matrices. Therefore, both deflation methods have the same performance.

7.5.2 Modified Matrix Z

We have already mentioned that $\widetilde{Z} = [Z \;\; z_r]$, so in fact we have omitted the last column of $\widetilde{Z}$ to construct Z. Now, an idea is to omit other columns of $\widetilde{Z}$ instead of the last column. Therefore, we define $Z^{(m)}$ as follows:

$$Z^{(m)} = [z_1 \; \cdots \; z_{m-1} \; z_{m+1} \; \cdots \; z_r], \quad 1 < m < r - 1, \tag{7.4}$$

and for $m = 1$ and $m = r - 1$:

$$Z^{(1)} = [z_2 \ \cdots \ z_r], \quad Z^{(r-1)} = [z_1 \ \cdots \ z_{r-2} \ z_r], \tag{7.5}$$

Some results can be found below where we apply the same test case as above with $n = 32^2, 64^2$ and $m = 5$.

m	DICCG-4		DICCG-16	
	$n = 32^2$	$n = 64^2$	$n = 32^2$	$n = 64^2$
1	45	116	61	127
$r - 1$	63	130	61	127
r	53	128	60	125

Table 7.7: Number of iterations of DICCG$-k$ to solve the singular linear system $Ax = b$ with $Z^{(m)}$ instead of Z.

From the table we can observe that the convergence results depend on the value of m. The original method $(m = r)$ is the best one when the number of deflation vectors is sufficiently large.

7.5.3 Alternative Choices for the Deflation Vectors

For the singular systems we have used matrix Z instead of $\widetilde{Z}$, since $\widetilde{Z}^T A \widetilde{Z}$ is singular while $Z^T AZ$ is an invertible matrix. Now, the question is whether there are another choices for the deflation vectors which are better than Z. We test the following subspace deflation matrices:

$$Z_m := [Z \ \ v_m], \quad m \in \mathbb{N}, \tag{7.6}$$

where

$$v_m = z_r - \sum_{p=1}^{m} \mathbf{e}_n^{(n-p+1)}. \tag{7.7}$$

In other words, Z_m is identical to $\widetilde{Z}$ except the last column whose last m elements are zero. Some results can be found below where we apply the same test case as above with $n = 32^3$.

m	DICCG-8 (Original: 57 iter.)	DICCG-64 (Original: 57 iter.)
1	72	72
2	75	75
3	76	76

Table 7.8: Number of iterations of DICCG$-k$ to solve the singular linear system $Ax = b$ with Z_m instead of Z.

Obviously, the method with Z_m is not better than the original deflation method using Z.

It appears that the larger m the worse the convergence of the iterative process.

Chapter 8

Conclusions

In this report, we develop the theory of deflation and both singular and invertible SPSD matrices. The main results are given below.

First, we have shown that the effective condition number of the deflated singular system is always better than the effective condition number of the original singular system. This result can also be generalized for preconditioned singular systems. These results are attractive for Krylov iterative methods, since a more favorable (effective) condition number leads to faster convergence of the solution.

Next, we have supposed the 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 has completely been 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. Also these results can be generalized for the preconditioned singular and invertible matrices. This means that two variants of deflated and preconditioned linear systems can be solved resulting in the same convergence results.

Results of numerical experiments confirm the theoretical results and show the good performance of the deflation technique.

Bibliography

[1] J. Frank, C. Vuik, *On the construction of deflation-based preconditioners*, SIAM Journal on Scientific Computing, **23**, pp. 442–462, 2001.

[2] G.H. Golub, C.F. van Loan, *Matrix Computations*, Third Edition, The John Hopkins University Press, Baltimore, Maryland 21218, 1996.

[3] R. Horn, C. Johnson, *Matrix Analysis*, Cambridge University Press, USA Edition, 1990.

[4] E. F. Kaasschieter, *Preconditioned Conjugate Gradients for solving singular systems*, J. Comp. Appl. Math., **24**, pp. 265–275, 1988.

[5] E. Kreyszig, *Introductory Functional Analysis with Applications*, New York: Wiley, 1989.

[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)**, pp. 303–322, 1968.

[7] L. Mansfield, *Damped Jacobi preconditioning and coarse grid deflation for Conjugate Gradient iteration on parallel computers*, SIAM J. Sci. Stat. Comput., **12**, pp. 1314–1323, 1991.

[8] L. Mansfield, *On the Conjugate Gradient Solution of the Schur Complement System Obtained from Domain Decomposition*, SIAM J. Num. Anal., **27**, pp. 1612-1620, 1990.

[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*, Mathematics of Computation, **31**, pp. 148–162, 1977.

[10] R.B. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Analysis and Applications, **16**, pp. 1154–1171, 1995.

[11] R. Nabben, C. Vuik, *A comparison of Deflation and Coarse Grid Correction applied to porous media flow*, SIAM J. Numer. Anal., **42**, pp. 1631–1647, 2004.

[12] R. Nabben, C. Vuik, *A comparison of Deflation and the balancing Neumann-Neumann preconditioner*, Delft University of Technology, Department of Applied Mathematical Analysis, Report 04-09, ISSN 1389-6520, 2004.

[13] S. V. Patankar, *Numerical Heat Transfer and Fluid Flow*, Series in Comput. Meth. in Mechanics and Thermal Science, McGraw-Hill, New York, 1980.

[14] R.A. Nicolaides, *Deflation of Conjugate Gradients with applications to boundary value problems*, SIAM J. Matrix Anal. Appl., **24**, pp. 355–365, 1987.

[15] A. van der Sluis, H.A. van der Vorst, *The rate of convergence of Conjugate Gradients*, Num. Math., **48**, pp. 543–560, 1986.

[16] G. W. Stewart, *Perturbation bounds for the definite generalized eigenvalue problem*, Linear Algebra Appl., **23**, pp. 69–85, 1979.

[17] J.M. Tang, *Parallel Deflated CG Methods applied on Moving Boundary Problems*, Delft University of Technology, Department of Applied Mathematical Analysis, Report 05-02, ISSN 1389-6520, 2004.

[18] J. Verkaik, *Deflated Krylov-Schwarz Domain Decomposition for the Incompressible Navier-Stokes Equations on a Colocated Grid*, Master's thesis (see `http://ta.twi.tudelft.nl/nw/users/vuik/numanal/verkaik_afst.pdf`), TU Delft, 2003.

[19] J. Verkaik, C. Vuik, B.D. Paarhuis, A. Twerda *The Deflation Accelerated Schwarz Method for CFD*, Computational Science-ICCS 2005: 5th International Conference, Atlanta, GA, USA, May 22-25, 2005, Proceedings Part I, see also `http://ta.twi.tudelft.nl/nw/users/vuik/papers/Ver05VPT.pdf`), Springer Berlin, pp. 868-875 , 2005.

[20] F. Vermolen, C. Vuik, A. Segal, *Deflation in preconditioned Conjugate Gradient methods for finite element problems*, In: Conjugate Gradient and Finite Element Methods (Ed: M. Krizek and P. Neittaanmaki and R. Glowinski and S. Korotov), Springer, Berlin, pp. 103–129, 2004.

[21] C. Vuik, J. Frank, *Coarse grid acceleration of a parallel block preconditioner*, Future Generation Computer Systems, **17**, pp. 933–940, 2001.

[22] C. Vuik, J. Frank, A. Segal, *A parallel block-preconditioned GCR method for incompressible flow problems*, Future Generation Computer Systems, **18**, pp. 31–40, 2001.

[23] C. Vuik, J. Frank, F.J. Vermolen, *Parallel Deflated Krylov methods for incompressible flow*, in: Parallel Computational Fluid Dynamics: Practice and Theory, pp. 381–388, 2002.

[24] C. Vuik, A. Segal, L. El Yaakoubi, E. Dufour, *A comparison of various deflation vectors applied to elliptic problems with discontinuous coefficients*, Applied Numerical Mathematics, **41**, pp. 219–233, 2002.

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

[26] C. Vuik, A. Segal, J.A. Meijerink, G.T. Wijma, *The construction of projection vectors for a Deflated ICCG method applied to problems with extreme contrasts in the coefficients*, Journal of Computational Physics, **172**, pp. 426–450, 2001, (see also Shell Report EP2000-8019).

[27] J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

Appendix A

Proofs of two Lemma's

In this appendix we proof first some elementary observations (Lemma A.1). Thereafter, we show that there exists a matrix Y such that $[Y\ \widetilde{Z}_0]$ is invertible and $\widetilde{Z}_0^T\widetilde{A}Y = \mathbf{0}_{r,n-r}$ holds (Lemma A.2).

Lemma A.1. *Let $\widetilde{A}, \widetilde{Z}_0$ and $\widetilde{Z}$ be matrices as in Chapter 2. Let $Y \in \mathbb{R}^{n\times(n-r)}$ be a matrix such that $(\widetilde{A}\widetilde{Z})^T Y = \mathbf{0}_{r,n-r}$ and define $X := [Y\ \widetilde{Z}]$. Then,*

(i) $\widetilde{A}\widetilde{Z}$ has rank r;

(ii) identical pivot positions of $\widetilde{Z}$ and $\widetilde{A}\widetilde{Z}$ can be chosen;

(iii) there exists an Y such that X is full-ranked, i.e., $\exists\, Y$: rank $X = n$.

(iv) after replacing $\widetilde{Z}$ by $\widetilde{Z}_0$ in all above expressions in this lemma, the properties (i)–(iii) still hold.

Proof. (i) An elementary observation since rank $\widetilde{A} = n$ and rank $\widetilde{Z} = r$, see e.g. p. 13 of Horn & Johnson [3].

(ii) Denote the position corresponding to the last non-zero element of a column of $\widetilde{Z}$ by (p, q) (i.e., row p and column q). Obviously, pivot position (p, q) can always be chosen as pivot position since it is a non-zero element. Below we prove that the position (p, q) of $\widetilde{A}\widetilde{Z}$ can also be chosen as pivot position by showing that this element is always a non-zero element.

The proof is as follows. Let $\tilde{z}_{p,q}$ the last non-zero element of $\widetilde{Z}$ in column $q \neq n$. Then,

$$(\widetilde{A}\widetilde{Z})_{p,q} = \sum_{i=1}^{n} \tilde{a}_{p,i}\cdot \tilde{z}_{i,q} = \sum_{i=1}^{p} \tilde{a}_{p,i}\cdot \tilde{z}_{i,q} = \tilde{a}_{p,1}\cdot \tilde{z}_{1,q} + \ldots + \tilde{a}_{p,p}\cdot \tilde{z}_{p,q}. \tag{A.1}$$

Since the main diagonal elements of $\widetilde{A}$ are non-zero, we have that

$$\tilde{a}_{p,p}\cdot \tilde{z}_{p,q} \neq 0.$$

Due to Corollary 2.2, this means that $(\widetilde{A}\widetilde{Z})_{p,q} = 0$ only holds if all non-zero elements of $\widetilde{A}$ in row $p > 1$ contributes to the sum from Eq. (A.1). However, for all rows $p < n$, we know that there exists a $k \geq 1$ such that $a_{p,p-k} \neq 0$. Since this element does not contribute to the sum from Eq. (A.1) because $\tilde{z}_{p-k,q} = 0$ for all $k \geq 1$, this results in

$$(\widetilde{A}\widetilde{Z})_{p,q} = \sum_{i=1}^{p} \tilde{a}_{p,i} \cdot \tilde{z}_{i,q} = \tilde{a}_{p,1} \cdot \tilde{z}_{1,q} + \ldots + \tilde{a}_{p,p} \cdot \tilde{z}_{p,q} > 0, \quad \forall\, p > 1.$$

Furthermore, we know that the sum of the last row of $\widetilde{A}$ is $\sigma \cdot a_{n,n} \neq 0$. Hence, for position $(p,q) = (n,n)$ it yields

$$(\widetilde{A}\widetilde{Z})_{n,n} \neq 0,$$

where we also assume that the value of the diagonal elements has different sign compared to the other elements.

Thus, pivot position (p,q) of $\widetilde{Z}$ is a non-zero position of $\widetilde{A}\widetilde{Z}$ and therefore it can also be chosen as a pivot position of $\widetilde{A}\widetilde{Z}$, since $\widetilde{Z}$ and $\widetilde{A}\widetilde{Z}$ have the same dimensions and rank $\widetilde{A}\widetilde{Z} =$ rank $\widetilde{Z} = r$.

(iii) Note first that $\widetilde{Z}$ and $\widetilde{A}\widetilde{Z}$ have both rank r. From Property (ii) we already know that the same pivot positions of $\widetilde{Z}$ and $\widetilde{A}\widetilde{Z}$ can be chosen. Since span $\left\{\widetilde{A}\widetilde{Z}\right\} \perp$ span $\{Y\}$, Y is an orthogonal complement of $\widetilde{A}\widetilde{Z}$, which can be chosen such that Y has pivot positions different of the common pivot positions of Z by definition of orthogonal complements, see also pp. 16–17 of Horn & Johnson [3]. Due to Property (ii), these pivot positions of Y differ also from AZ. Then, matrix Y is linear independent of both span$\left\{\widetilde{Z}\right\}$ and span$\left\{\widetilde{A}\widetilde{Z}\right\}$, although Y is no orthogonal complement of span$\left\{\widetilde{Z}\right\}$ in general. As a consequence, both matrices $[Y\ \widetilde{A}\widetilde{Z}]$ and $[Y\ \widetilde{Z}]$ are full-ranked and hence rank $X = n$ is achieved.

(iv) Since $\widetilde{Z}_0$ and $\widetilde{Z}$ only differs in the last column and $z_0 \subset \text{span}\{z_1,\ z_2,\ \ldots,\ z_r\}$ for all $r > 1$, the proof is analogous to the proofs of above. □

Lemma A.2. *Let $\widetilde{A}$ and $\widetilde{Z}_0$ be defined as in the previous lemma. Then there exists a matrix $Y := [z_{r+1}\ z_{r+2}\ \cdots\ z_n]$ such that*

- *the columns of Y and $\widetilde{Z}_0$ are mutually linear independent, i.e., matrix $X := [Y\ \ \widetilde{Z}_0]$ is invertible;*
- *the following identity holds:*

$$\widetilde{Z}_0^T \widetilde{A} Y = \boldsymbol{0}_{r,(n-r)}. \tag{A.2}$$

Proof. Note first that $\widetilde{Z}_0$ has rank r since the columns of $\widetilde{Z}_0$ are linear independent by construction.

We start the proof with a simple case when $\widetilde{Z}_0$ is an $\widetilde{A}-$invariant subspace and thereafter we prove the lemma when $\widetilde{Z}_0$ is derived from subdomain deflation as in Section 2.

- $\widetilde{Z}_0$ *is* $\widetilde{A}-$*invariant* ($\widetilde{A}\widetilde{Z}_0 \subseteq \widetilde{Z}_0$). By definition, $\widetilde{Z}_0$ is an $\widetilde{A}-$invariant subspace if for all $z \in \widetilde{Z}_0$ we have $\widetilde{A}z \in \widetilde{Z}_0$, which will be denoted by $\widetilde{A}\widetilde{Z}_0 \subseteq \widetilde{Z}_0$. For instance $\widetilde{Z}_0$ is $\widetilde{A}-$invariant when it consists of eigenvectors of $\widetilde{A}$.

  Since $\widetilde{A}$ has full rank and $\widetilde{Z}_0$ has rank r, we may choose the remaining space Y in the orthogonal complement of $\text{span}\left\{\widetilde{Z}_0\right\}$, i.e., $Y^T Z = \mathbf{0}_{n-r,r}$. As a consequence,

  $$Y^T \widetilde{A}\widetilde{Z}_0 \subseteq Y^T \widetilde{Z}_0 = \mathbf{0}_{n-r,r}.$$

  By taking the transpose of the latter expression, we have

  $$\widetilde{Z}_0^T \widetilde{A} Y = \mathbf{0}_{r,n-r}.$$

- $\widetilde{Z}_0$ *is from subdomain deflation* ($\widetilde{A}\widetilde{Z}_0 \nsubseteq \widetilde{Z}_0$). Because now matrix $\widetilde{Z}_0$ is from subdomain deflation, $\widetilde{Z}_0$ is definitely not $\widetilde{A}-$invariant, just 'nearly $\widetilde{A}-$invariant'.

  Note first that if $\widetilde{Z}_0^T \widetilde{A} Y = \left(\widetilde{A}\widetilde{Z}_0\right)^T Y = \mathbf{0}_{n-r,r}$, then in fact each column of $\widetilde{A}\widetilde{Z}_0$ is orthogonal to any column of Y, i.e., $w^T \widetilde{A} Y = \mathbf{0}_r^T$. By definition of orthogonal complement such a Y can always be constructed and it is not unique in general. We show that for some choices of Y this leads to an invertible X.

  Since $\widetilde{A}$ has full rank and $\widetilde{Z}_0$ has rank r, $\widetilde{A}\widetilde{Z}_0$ has again rank r from Lemma A.1(i). Denote the columns of $\widetilde{A}\widetilde{Z}_0$ by $[w_1 \; w_2 \; \cdots \; w_r]$. Applying Lemma A.1(iii)–(iv), we may choose the remaining space Y in the orthogonal complement of $\text{span}\{w_1, \; w_2, \; \cdots, \; w_r\}$ such that

  $$\text{rank}\,\begin{bmatrix} Y & \widetilde{Z}_0 \end{bmatrix} = \text{rank}\,\begin{bmatrix} Y & \widetilde{A}\widetilde{Z}_0 \end{bmatrix} = n. \tag{A.3}$$

  Then X is invertible and the lemma has been proved.

□

Remark. Note that the conditions of both $\widetilde{A}$ and $\widetilde{Z}_0$ have to be satisfied in Lemma A.2 to apply this lemma. Take for instance the following matrix:

$$\widetilde{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Obviously, A is SPD and invertible. Next, if $\widetilde{Z}_0 = (1,0)^T$, then $\widetilde{A}\widetilde{Z}_0 = (0,1)^T$. To satisfy $\widetilde{Z}_0 \widetilde{A} Y = \mathbf{0}_{r,n-r}$, we have to choose $Y = (1,0)^T$. However, in this case the identity $\widetilde{Z}_0 = Y$ has been derived resulting in a singular matrix $[Y \;\; \widetilde{Z}_0]$. Apparently, Lemma A.2 does not hold in this example. Clearly, the conditions for $\widetilde{A}$ and $\widetilde{Z}_0$ (as described in Lemma A.2) are required to apply the lemma.

Appendix B

Termination criterions of ICCG and DICCG

In the original ICCG we solve the system

$$M^{-1}Ax_k = M^{-1}b \tag{B.1}$$

in each iterate k, where M is the preconditioner. It is common to use the following termination criterion in the iterative process:

$$\frac{||M^{-1}(b - Ax_k)||}{||M^{-1}b||} < \epsilon, \tag{B.2}$$

where we assume that we start with the zero starting vector, i.e., $x_0 = \mathbf{0}_n$. Note that the LHS of (B.2) is called the relative tolerance of the method.

On the other hand, in DICCG we solve the singular system

$$M^{-1}PA\tilde{x}_k = M^{-1}Pb, \tag{B.3}$$

where $\tilde{x}_k$ is the non-unique solution in each iterate k. The unique solution can be made using

$$x_k = ZE^{-1}Z^Tb + P^T\tilde{x}_k. \tag{B.4}$$

However, extra cost is required in order to compute (B.4) and form x_k in each iterate k, whereas we would like to choose the same termination criterion as in ICCG for a comparison between ICCG and DICCG in our numerical experiments. Fortunately, it is easy to show that (B.2) is equivalent to

$$\frac{||M^{-1}P(b - A\tilde{x}_k)||}{||M^{-1}b||} < \epsilon, \tag{B.5}$$

due to Theorem B.1. Hence the computation (B.4) can be avoided in the DICCG-iterates. Note that the LHS of (B.5) is the relative tolerance of the method and differs from the relative tolerance of ICCG as showed in (B.2).

Theorem B.1. *Assume that A, M, P, b are defined as above. Let x_k and $\tilde{x}_k$ the solutions in iterate k of the ICCG and DICCG method, respectively. Then we have the identity*

$$\frac{||M^{-1}(b - Ax_k)||}{||M^{-1}b||} = \frac{||M^{-1}P(b - A\tilde{x}_k)||}{||M^{-1}b||}. \tag{B.6}$$

Proof. It suffices to show that

$$b - Ax_k = P(b - A\tilde{x}_k). \tag{B.7}$$

Using Exp. (B.4) we have

$$b - Ax_k = b - A(ZE^{-1}Z^T b + P^T \tilde{x}_k) = (I - AZE^{-1}Z^T)b - AP^T \tilde{x}_k$$

and hence,

$$b - Ax_k = Pb - AP^T \tilde{x}_k. \tag{B.8}$$

Note that $AP^T = PA$ since

$$AP^T = A - AZE^{-1}Z^T A = PA. \tag{B.9}$$

Combining Eqn. (B.8) and (B.9) it yields

$$b - Ax_k = P(b - A\tilde{x}_k).$$

which is exactly (B.7). □