Analysis of the Deflated Conjugate Gradient Method Based on Symmetric Multigrid Theory

K. Kahl · H. Rittich

May 5, 2014

Abstract Deflation techniques for Krylov subspace methods have seen a lot of attention in recent years. They provide means to improve the convergence speed of the methods by enriching the Krylov subspace with a deflation subspace. The most common approach for the construction of deflation subspaces is to use (approximate) eigenvectors. However, there are many situations where a more general deflation subspace is advisable.

We derive an estimate for the speed of convergence of the deflated conjugate gradient method using theory originally developed for algebraic multigrid methods. Our result holds for general deflation subspaces and is based on the weak approximation property—known from multigrid methods—and a measure of the $A$-invariance of the subspace by the strengthened Cauchy-Schwarz inequality. We give convergence estimates for the case where the deflation space is spanned by eigenvectors of the matrix that are not exactly known. Furthermore, our theory suggests that the techniques developed to construct efficient interpolation operators in algebraic multigrid methods can also be applied to improve deflation subspaces.

Keywords conjugate gradients · deflation · algebraic multigrid · convergence analysis

Mathematics Subject Classification (2010) 65F08 · 65F10

1 Preliminaries

Consider solving the linear system of equations

$$Ax = b,$$

(1.1)

where $A \in \mathbb{K}^{n \times n}$ ($\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$) is self-adjoint and positive definite and $x, b \in \mathbb{K}^n$. In this paper we are interested in the case where the matrix $A$ is large
and sparse. The conjugate gradient (CG) method \cite{11,17} is an iterative method which is often well suited to solve these systems. The speed of convergence of the CG method depends on the condition number $\kappa$ of the matrix $A$, as well as on the distribution of its eigenvalues \cite{17,26}. When the condition number $\kappa$ is large it can become mandatory to precondition the linear system such that a satisfactory speed of convergence is obtained.

One possibility to precondition the CG method is via deflation as introduced by Nicolaides \cite{15} and Dostal \cite{6}; see also \cite{9,10,14,18}. The basic idea of deflation is to “hide” certain parts of the spectrum of the matrix $A$ from the CG method itself, such that the CG iteration “sees” a system that has a much smaller condition number than $A$. The part of the spectrum that is hidden from CG is determined by the deflation subspace $S \subseteq \mathbb{R}^n$.

In \cite{15} the space $S$ is constructed as follows. The variables are combined into aggregates $A_i \subseteq \{1, 2, \ldots, n\}$, $i = 1, \ldots, m$, such that

$$
\bigcup_{i=1}^{m} A_i = \{1, 2, \ldots, n\} \quad \text{and} \quad A_i \cap A_j = \emptyset \quad \text{for } i \neq j.
$$

Then $S$ is spanned by the vectors $v^{(i)}$, $i = 1, \ldots, m$ with

$$
v^{(i)}_j = \begin{cases} 
1 & \text{if } j \in A_i \\
0 & \text{otherwise}
\end{cases}.
$$

This procedure can be motivated in the following way: Assume that the matrix $A$ arises from the discretization of an elliptic partial differential equation by a finite element/difference/volume method. Components of vectors then correspond to grid points of the underlying discretization scheme. When the aggregates $A_i$ are chosen appropriately, $S$ is close to the space consisting of those vectors whose values vary only slowly between neighboring grid points. Those vectors represent in many applications the eigenvectors corresponding to small eigenvalues. By deflating such vectors the corresponding eigenvalues are removed from the spectrum and the deflated CG method behaves as if the matrix had a much smaller condition number. Interestingly, this procedure from \cite{15} is completely analogous to the construction of prolongation operators in (non-smoothed) aggregation based multigrid methods \cite{4}.

Another viable and widely used approach for deflation consists of spanning $S$ directly by the eigenvectors corresponding to the smallest eigenvalues \cite{18}. This immediately leads to the removal of the smallest eigenvalues from the spectrum of $A$. The major drawback of this approach is that it often does not scale when the size of the system increases, because in many cases the number of eigenvalues below a given threshold grows with the size of the system. Thus, as the system size increases, more and more eigenvectors need to be computed to keep the convergence rate at a desired level. However, in the case where only a few very small eigenvalues exist, independent of the system size, this approach works reasonably well.

Recently a combination of the two approaches has been introduced in the context of simulations in Quantum Chromodynamics \cite{14}. Similarly to \cite{15}, aggregates $A_i$ are introduced, but since eigenvectors belonging to small eigenvalues do not necessarily have slowly varying components in this application, a few eigenvectors
$w_1, \ldots, w_\ell$ corresponding to the smallest eigenvalues of the system are computed. Then for every vector $w_j$ and every aggregate $\mathcal{A}_i$ the orthogonal projection $\tilde{w}_j^{(i)}$ of $w_j$ onto

$$V_i := \text{span}\{e^{(j)}: j \in \mathcal{A}_i\}$$

where $e^{(i)}$ with $e^{(j)}_j = \delta_{j,i}$ is the $j$th canonical vector, is computed and the deflation subspace $S$ is spanned by $\tilde{w}_j^{(i)}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, \ell$. This approach has the advantage that it often scales when the size of the system is increased while the number of eigenvectors $\ell$ and the size of the aggregates are chosen to be constant. This particular strategy to define the deflation subspace resembles the setup of adaptive aggregation based algebraic multigrid methods [4], where the prolongation operator is constructed in a similar way.

The case where the deflation subspace is constructed via aggregates [15] is analyzed in [9] for some special classes of matrices. Furthermore, due to the similarities between deflation and multigrid methods there have been a few investigations concerning the connection between deflation and multigrid methods. In [24] it has been shown that the CG method preconditioned with a multigrid method produces the same iterates as a deflated CG method under the following conditions. The multigrid method consists of a $V(0,1)$-cycle, i.e., no pre-smoothing steps and one post-smoothing step is applied. The deflated CG is furthermore preconditioned by the smoother of the multigrid method. The authors of [23] compare the CG method preconditioned with a $V(1,1)$-cycle and a preconditioned deflated CG. The authors describe in which situation which method is more effective. However, the analysis in [23] is only carried out for the case that the deflation subspace is spanned by eigenvectors.

In this paper we contribute to the convergence theory of deflated CG. Our theory is based on techniques used for the analysis of the algebraic multigrid method [3,16,21]. In particular we use the weak approximation property, an important tool in the analysis of these methods. In contrast to previous works [9,23,24] our theory yields estimates for the speed of convergence for general choices of deflation subspaces and without requiring an additional preconditioner. In case prolongation operators from multigrid methods are used to define the deflation subspace, our theory shows that the resulting method behaves like a multigrid method, i.e., the effective condition number is bounded independently of the problem size. Further, we show that our theory simplifies to the known result in case the deflation subspace is spanned by eigenvectors. In addition we give a qualitative bound on the convergence rate of the deflated CG method in the case where the eigenvectors are known only up to a certain accuracy. This result is particularly interesting for the analysis of methods like eigCG (cf. [20]), where the eigenvectors are numerically approximated.

We finish this introduction explaining some basic notation. Assume that $\hat{x} \in \mathbb{K}^n$ is an approximation to $x$, the solution of (1.1). Then the residual $r \in \mathbb{K}^n$ is given by $r = b - A\hat{x}$, the error by $e = x - \hat{x}$. Note that $Ae = r$.

Let $\langle v, w \rangle = \sum_{i=1}^n w_i v_i$ be the euclidean inner product of $v$ and $w$. Since $A$ is self-adjoint and positive definite the $A$-inner product and the $A$-norm are given by

$$\langle v, w \rangle_A := \langle Av, w \rangle$$

and

$$\|v\|_A := \sqrt{\langle v, v \rangle_A}.$$
The rest of the paper is structured as follows. Section 2 gives a short introduction into deflation methods. In Section 3 we analyze the convergence of deflation methods by analyzing the condition number of the deflated matrix, which we estimate by using results of algebraic multigrid theory. In Section 4 we further show that our general convergence result yields the known bounds when eigenvectors are used to span the deflation space. Then we consider the case where the eigenvectors are only approximately known. Moreover, we demonstrate how prolongation operators from the classical algebraic multigrid theory for $M$-matrices can be used to obtain deflation subspaces. We present some numerical experiments confirming the theory in Section 5. Finally in Section 6 we discuss how accurately the projection needed for the deflation has to be computed to maintain overall convergence of the deflated CG method.

2 Review of Deflated CG

The CG method is an iterative method for solving linear systems, i.e., given an initial guess $x_0 \in \mathbb{K}^n$ it produces iterates $x_1, x_2, \ldots$ such that $x_i$ converges to the solution $x$ of (1.1). Let $r_0 = b - Ax_0$ and $K_i(A, r_0) := \text{span}\{r_0, Ar_0, A^2r_0, \ldots, A^{i-1}r_0\}$ the $i$th Krylov subspace. The iterate $x_i$ is determined such that $x_i \in x_0 + K_i(A, r_0)$ and the error $e_i = x - x_i$ in the $A$-norm is minimized (cf. [17]), i.e.,

$$
\|e_i\|_A = \min_{\tilde{x} \in x_0 + K_i(A, r_0)} \|x - \tilde{x}\|_A. \tag{2.1}
$$

Note that (2.1) is just the distance in the $A$-norm between $x$ and the affine subspace $x_0 + K_i(A, r_0)$. The convergence of the method can be slow in case of unfavorable spectral properties of the matrix $A$ [26]. In that case the distance is only decreasing slowly. The idea of deflation is to modify the CG method such that the iterates $x_i$ are equal to the solution $x$ on a given subspace $S \subseteq \mathbb{K}^n$ in a certain way. More precisely, the $A$-orthogonal projections of $x_i$ and $x$ onto $S$ coincide. For proper choices of $S$ this will improve the speed of convergence since the affine subspace which contains the iterates may now be much closer to the solution $x$ than the original one. We give a rigorous description of the deflated CG method in the remainder of this section.

Given a certain subspace $S \subseteq \mathbb{K}^n$ we can split the solution $x$ into a component in $S$ and a component in $S^\perp$ via the $A$-orthogonal projection $\pi_A(S) \in \mathbb{K}^{n \times n}$ onto $S$, i.e.,

$$
x = (I - \pi_A(S))x + \pi_A(S)x. \tag{2.2}
$$

Let $V \in \mathbb{K}^{n \times m}$ be a matrix such that its columns form a basis of the subspace $S$. Since

$$
\pi_A(S)x = V(V^*AV)^{-1}V^*Ax = V(V^*AV)^{-1}V^*b \tag{2.3}
$$

we can compute $\pi_A(S)x$—the second term in the right hand side of (2.2)—without explicit knowledge of $x$. The first term of (2.2) can be computed from the solution $\tilde{x}$ of the singular linear system

$$
A(I - \pi_A(S))\tilde{x} = (I - \pi_A(S))^*b, \tag{2.4}
$$

which we call the deflated (linear) system. This is shown by the following lemma.
Lemma 2.1 Using the definitions from above we have:

(i) The following equalities hold
\[ A(I - \pi_A(S)) = (I - \pi_A(S))^* A = (I - \pi_A(S))^* A(I - \pi_A(S)). \]  
(2.5)

(ii) The matrix \( A(I - \pi_A(S)) \) is self-adjoint and positive semi-definite.

(iii) The deflated system (2.4) is consistent, i.e., the right hand side \( (I - \pi_A(S))^* b \) is in the range of \( A(I - \pi_A(S)) \). This implies that the system has at least one solution.

(iv) If \( \hat{x} \) is a solution of the deflated system (2.4) then
\[ (I - \pi_A(S))\hat{x} = (I - \pi_A(S))x, \]  
where \( x \) is the solution of the linear system \( Ax = b \).

Proof Since \( \pi_A(S) = V(V^*AV)^{-1}A \) we have
\[ A(I - \pi_A(S)) = A - AV(V^*AV)^{-1}V^* A = (I - \pi_A(S))^* A, \]
and since \( (I - \pi_A(S)) \) is a projection we also get
\[ A(I - \pi_A(S)) = A(I - \pi_A(S))(I - \pi_A(S)) = (I - \pi_A(S))^* A(I - \pi_A(S)). \]
This proves (i) and also shows that \( A(I - \pi_A(S)) \) is self-adjoint. Due to (2.5) we have
\[ \langle A(I - \pi_A(S))x, x \rangle = \langle (I - \pi_A(S))x, (I - \pi_A(S))x \rangle_A = \| (I - \pi_A(S))x \|_A^2 \geq 0 \]
which gives (ii)

Again due to (2.4) and the fact that \( A \) has full rank we have
\[ \text{range} \left[ A(I - \pi_A(S)) \right] = \text{range} \left[ (I - \pi_A(S))^* A \right] = \text{range} \left[ (I - \pi_A(S))^* \right]. \]
Hence the system (2.4) is consistent which proves (iii) To show (iv) we use (2.4) and (2.5) yielding
\[ A(I - \pi_A(S))\hat{x} = (I - \pi_A(S))^* b = (I - \pi_A(S))^* A x = A(I - \pi_A(S))x. \]

Multiplying with \( A^{-1} \) from the left we obtain (2.6).

Given a solution \( \hat{x} \) for the deflated system (2.4), due to (2.2), (2.3) and (2.6) we obtain the solution \( x \) for the original system as
\[ x = (I - \pi_A(S))x + \pi_A(S)x \]
\[ = (I - \pi_A(S))\hat{x} + V(V^*AV)^{-1}V^* b. \]

We now consider solving the deflated system (2.4). Since \( A(I - \pi_A(S)) \) is positive semi-definite we can apply the CG method. The lack of regularity is no impediment to the standard CG iteration as long as (2.4) is consistent (cf. [12]), which was shown to be the case in Lemma 2.1.

For the purpose of analyzing the method we think of deflated CG as applying the standard CG algorithm to the deflated system (2.4) with the matrix
A(I − π_A(S)). There are various other formulations of deflated CG that are mathematically equivalent (for an overview see [10]) for which our analysis holds as well.

Let \( \mu_1 \geq \cdots \geq \mu_n \geq 0 \) be the eigenvalues of the self-adjoint and positive semi-definite matrix \( A(I − π_A(S)) \). Let \( \ell \in \mathbb{N} \) denote the largest index such that \( \mu_\ell \neq 0 \). The errors of the CG iterates then satisfy

\[
\|e_i\|_A \leq 2 \left( \frac{\sqrt{\kappa_{\text{eff}}} - 1}{\sqrt{\kappa_{\text{eff}}} + 1} \right)^i \|e_0\|_A \quad \text{for } i = 0, 1, 2, \ldots,
\]

where \( \kappa_{\text{eff}} = \frac{\mu_1}{\mu_\ell} \), see [9,18]. We call \( \kappa_{\text{eff}} \) the effective condition number of the deflated matrix \( A(I − π_A(S)) \) to distinguish it from the condition number \( \kappa \) of the original matrix \( A \). Thus in order to give a bound on the convergence rate of deflated CG, it suffices to estimate the largest and smallest non-zero eigenvalue of the matrix \( A(I − π_A(S)) \).

3 Convergence Analysis

In this section we give an estimate for the speed of convergence of the deflated CG method by estimating the effective condition number \( \kappa_{\text{eff}} \) of the matrix \( A(I − π_A(S)) \).

3.1 Eigenvalue Bounds

\( A(I − π_A(S)) \). The largest eigenvalue of the matrix \( A(I − π_A(S)) \) is the maximum of the Rayleigh quotient (cf. [27]), i.e.,

\[
\mu_1 = \max_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\langle A(I − π_A(S))x, x \rangle}{\langle x, x \rangle}.
\]

From the fact that \( Aπ_A(S) = AV(V^*AV)^{-1}V^*A \) is positive semi-definite and thus

\[
\langle A(I − π_A(S))x, x \rangle = \langle Ax, x \rangle − \langle Aπ_A(S)x, x \rangle \leq \langle Ax, x \rangle,
\]

we obtain

\[
\mu_1 \leq \max_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\langle Ax, x \rangle}{\langle x, x \rangle} = \lambda_1 = \|A\|.
\] (3.1)

This gives a simple upper bound for the largest eigenvalue of \( A(I − π_A(S)) \).

The following auxiliary result will be used to derive a lower bound for the smallest non-zero eigenvalue. It can be regarded as a special case of the min-max (or Courant-Fischer-Weyl) Theorem, cf. [27].

**Lemma 3.1** Let \( M \in \mathbb{K}^{n \times n} \) be self-adjoint with eigendecomposition \( M = UDU^* \), \( U = [u_1| \ldots |u_n] \in \mathbb{K}^{n \times n} \) unitary and \( D = \text{diag}(\mu_1, \ldots, \mu_n) \) with \( \mu_1 \geq \cdots \geq \mu_n \). Then for \( k = 1, \ldots, n \)

\[
\mu_k = \min_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\langle Mx, x \rangle}{\langle x, x \rangle}.
\]

\( x \perp \text{span}\{u_{k+1}, \ldots, u_n\} \).
Proof Apply the discussion from [27] Extremal properties of eigenvalues] to the matrix \((-M)\), e.g. If we choose \(M = A(I - \pi_A(S))\) Lemma 3.1 characterizes the smallest non-zero eigenvalue \(\mu_\ell\) of \(A(I - \pi_A(S))\) using \(\text{span}\{u_{\ell+1}, \ldots, u_n\}\), the kernel of \(A(I - \pi_A(S))\).

Since the matrix \(A\) has full rank, the kernel of \(A(I - \pi_A(S))\) is the kernel of \((I - \pi_A(S))\) which is the deflation subspace \(S\). Thus, denoting the orthogonal complement of \(S\) by \(S^\perp\), we obtain

\[
\mu_\ell = \min_{x \in S^\perp \setminus \{0\}} \frac{\langle A(I - \pi_A(S))x, x \rangle}{\langle x, x \rangle},
\]

and due to (2.5)

\[
\mu_\ell = \min_{x \in S^\perp \setminus \{0\}} \frac{\langle (I - \pi_A(S))^* A(I - \pi_A(S))x, x \rangle}{\langle x, x \rangle}
= \min_{x \in S^\perp \setminus \{0\}} \frac{\langle A(I - \pi_A(S))x, (I - \pi_A(S))x \rangle}{\langle x, x \rangle}
= \min_{x \in S^\perp \setminus \{0\}} \frac{\| (I - \pi_A(S))x \|_A^2}{\|x\|_2^2}. \tag{3.2}
\]

We are now ready to employ techniques developed for the analysis of algebraic multigrid methods to further advance our analysis.

3.2 Weak Approximation Property

In order to estimate the smallest non-zero eigenvalue \(\mu_\ell\) of \(A(I - \pi_A(S))\) we introduce some basic ideas of algebraic multigrid convergence analysis.

Algebraic multigrid methods [5,16,21] are based on the assumption that the error of a given iterate can be split into two parts. One of these parts can be removed easily by a cheap iterative process (e.g. Jacobi or Gauss-Seidel Iteration) which is called a smoother in this context. The other part cannot be removed by the smoother, hence called algebraically smooth, and needs to be resolved on a properly chosen coarse grid. In many applications and for many typical choices of the smoother the part removable by smoothing is spanned by the eigenvectors corresponding to large eigenvalues while the algebraically smooth components are spanned by the eigenvalues corresponding to the small eigenvalues of the matrix. In order to quantify those properties, classical algebraic multigrid theory (cf. [2]) introduces a smoothing property, which measures the quality of the smoother and a weak approximation property which measures how well the coarse grid correction, defined by the prolongation operator, is able to reduce the slowly varying error components. In the context of deflated CG we are only interested in the weak approximation property.

Definition 3.1 A subspace \(S \subseteq \mathbb{K}^n\) fulfills the weak approximation property with constant \(K \geq 0\) if

\[
\dist(S, x)^2 \leq \frac{K}{\|A\|} \|x\|_A^2 \quad \text{for all } x \in \mathbb{K}^n, \tag{3.3}
\]
where the distance between $x$ and the subspace $S$ is given by
\[
\text{dist}(S, x)^2 = \min_{y \in S} \| x - y \|^2.
\]

If the diagonal entries $a_{ii}$ of $A$ fulfill $a_{ii} = 1$ then Definition 3.1 coincides with the definition from [2,16,21] for a weak approximation property. It is called “weak” because it is only sufficient for a two-level convergence theory [16, Section 4.5] instead of a multilevel one. Unit diagonal entries may be achieved by replacing $A$ with $D^{-\frac{1}{2}}A D^{-\frac{1}{2}}$ and a corresponding change of the constant $K$, where $D = \text{diag}(a_{11}, \ldots, a_{nn})$.

In the case of multigrid method we often have the following situation. Assume that $V \in \mathbb{K}^{n \times m}$, $m < n$ is a multigrid prolongation operator and $R \in \mathbb{K}^{m \times n}$ a given (artificial) restriction operator. There exists estimates for a constant $C \in \mathbb{R}$ (see e.g. [21]), such that
\[
\| x - VRx \|^2 \leq C \| x \|^2 A.
\]
The following lemma shows that this implies that the weak approximation property is fulfilled for $K = C \| A \|$.

**Lemma 3.2** Let $V \in \mathbb{K}^{n \times m}$, $R \in \mathbb{K}^{m \times n}$ and assume that
\[
\| x - VRx \|^2 \leq C \| x \|^2 A
\]
holds for a constant $C \geq 0$. Then the subspace $S = \text{range } V$ fulfills the weak approximation property with constant $K = C \| A \|$.

**Proof** For arbitrary $x \in \mathbb{K}^n$ we have
\[
\text{dist}(S, x)^2 = \min_{y \in S} \| x - y \|^2 = \min_{z \in \mathbb{K}^n} \| x - Vz \|^2 \\
\leq \| x - VRx \|^2 \leq C \| x \|^2 A = \frac{K}{\| A \|} \| x \|^2 A.
\]

Strictly speaking, any subspace $S$ fulfills a weak approximation property, just by choosing $K$ large enough, since
\[
\text{dist}(S, x)^2 \leq \| x \|^2 \leq \lambda_n \| x \|^2 A
\]
for all $x \in \mathbb{K}^n$.

The interest of Definition 3.1 is in cases where the subspace $S$ is such that $K$ is small and in situations where $K$ is constant for a whole family of matrices $A$ and subspaces $S$, the family of matrices $A$ representing, e.g., different levels of discretization of a continuous operator. Such bounds on the constant $C$ are known for the case of algebraic multigrid methods (cf. [2,3,16,21]). They can directly be translated into a bound on the constant $K$ due to Lemma 3.2.

The following theorem now gives an estimate for the smallest eigenvalue of the matrix $A(I - \pi_A(S))$ in terms of the constant $K$ of the weak approximation property.

**Theorem 3.1** Let $S \subseteq \mathbb{K}^n$ be a subspace such that the weak approximation property (3.3) is fulfilled with constant $K$. Then the effective condition number $\kappa_{\text{eff}} = \frac{\mu}{\mu^*}$ of the matrix $A(I - \pi_A(S))$ satisfies
\[
\kappa_{\text{eff}} \leq \frac{K}{\xi} \quad \text{where} \quad \xi := \min_{x \in S \setminus \{0\}} \frac{\| x - \pi_A(S)x \|^2_A}{\| x \|^2_A} \in (0, 1].
\]
Proof. Denote by \( \pi(S) \) the orthogonal projection onto \( S \) with respect to the 2-inner product. Then
\[
\|x - \pi(S)x\|_2^2 = \min_{y \in S} \|x - y\|_2^2 = \text{dist}(S, x)_2^2.
\]
(3.5)

For \( x \in S^\perp \) we have \( \pi(S)x = 0 \) and thus due to (3.5)
\[
\|x\|_2^2 = \|x - \pi(S)x\|_2^2 = \text{dist}(S, x)_2^2,
\]
which, using (3.3) gives
\[
\|x\|_2^2 \leq K \|A\| \|x\|_A^2
\]
for \( x \in S^\perp \).

By applying this estimate to the denominator in (3.2) we obtain
\[
\mu_{\ell} \geq \frac{\|A\|}{K} \min_{x \in S^\perp \setminus \{0\}} \frac{\|x - \pi_A(S)x\|_A^2}{\|x\|_A^2} = \frac{\|A\|}{K} \xi.
\]
(3.6)

Hence, by (3.1) and (3.6),
\[
\kappa_{\text{eff}} = \frac{\mu_1}{\mu_{\ell}} \leq \frac{\|A\|}{\|A\| \xi} = \frac{K}{\xi}.
\]

It remains to show that \( \xi \in (0, 1] \). Since \( \pi_A(S)x = x \) if and only if \( x \in S \), we have \( \|x - \pi_A(S)x\|_A^2 > 0 \) for \( x \in S^\perp \setminus \{0\} \). Thus \( \xi > 0 \). The \( A \)-orthogonal projection minimizes the distance in the \( A \)-norm, i.e.,
\[
\|x - \pi_A(S)x\|_A^2 = \min_{y \in S} \|x - y\|_A^2.
\]
Hence \( \|x - \pi_A(S)x\|_A^2 \leq \|x\|_A^2 \), which proves \( \xi \leq 1 \).

Now that we have derived a connection between the effective condition number and the weak approximation property it remains to interpret the quantity \( \xi \) which is what we do in the next section in terms of the strengthened Cauchy-Schwarz inequality.

3.3 Strengthened Cauchy-Schwarz Inequality

For two subspaces \( H_1, H_2 \subseteq \mathbb{K}^n \) with \( H_1 \cap H_2 = \{0\} \) there exists a constant \( \gamma \in [0, 1) \) such that
\[
|\langle u, v \rangle| \leq \gamma \sqrt{\langle u, u \rangle \langle v, v \rangle} \quad \forall u \in H_1, \forall v \in H_2
\]
(3.7)

Theorem 2.1. Equation (3.7) is called strengthened Cauchy-Schwarz inequality and \( \gamma \) can be interpreted as the abstract angle between \( H_1 \) and \( H_2 \).

Inequality (3.7) implies that for \( u \in H_1, v \in H_2 \) we have
\[
|\langle u, v \rangle| \leq \gamma \sqrt{\langle u, u \rangle \langle v, v \rangle} \leq \frac{\gamma}{2} [\langle u, u \rangle + \langle v, v \rangle],
\]
since for any two numbers \(a, b\) we have \(|ab| \leq \frac{1}{2}(|a|^2 + |b|^2)\). Hence
\[
(1 - \gamma) \left( \langle u, u \rangle + \langle v, v \rangle \right) = \langle u, u \rangle + \langle v, v \rangle - \gamma \left( \langle u, u \rangle + \langle v, v \rangle \right) \\
\leq \langle u, u \rangle + \langle v, v \rangle - 2 \left| \langle u, v \rangle \right| \\
\leq \langle u, u \rangle + \langle v, v \rangle - 2 \Re \langle u, v \rangle \\
= (u + v, u + v).
\]
Taking the infimum over all \(v \in H_2\) yields
\[
(1 - \gamma) \|u\|^2 \leq \inf_{v \in H_2} \|u + v\|^2 \quad \forall u \in H_1.
\tag{3.8}
\]
We now apply this general result in the case where \(H_1 = S^\perp, H_2 = S\) and \(\langle \cdot, \cdot \rangle\) is the \(A\)-inner product, like in \([1]\) and \([8]\). The \(A\)-orthogonal projection \(\pi_A(S)u\) yields the vector in \(S\) which is closest to \(u\) in the \(A\)-norm. Thus the infimum in \((3.8)\) is obtained for \(v = -\pi_A(S)u\) and therefore
\[
(1 - \gamma)\|u\|^2_A \leq \|u - \pi_A(S)u\|^2_A \quad \forall u \in S^\perp.
\]
This yields a bound for \(\xi\) as
\[
\xi = \min_{x \in S^\perp \setminus \{0\}} \frac{\|x - \pi_A(S)x\|^2_A}{\|x\|^2_A} \geq (1 - \gamma). \tag{3.9}
\]
Using \((3.9)\) we can show the following lemma which states that we can interpret \(\xi\) as a measure of approximate \(A\)-invariance of the subspace \(S\), i.e., a small value of \(\gamma\), and thus value of \(\xi\) close to one, indicates that \(AS\) is close to \(S\).

**Lemma 3.3** If \(S\) is \(A\)-invariant, i.e., \(AS = S\), then \(\gamma = 0\) and thus \(\xi = 1\).

**Proof** Since the subspace \(S\) is \(A\)-invariant, we have \(Av \in S\) and thus
\[
\langle u, v \rangle_A = \langle u, Av \rangle = 0 \quad \forall u \in S^\perp, \forall v \in S.
\]
This gives \(\gamma = 0\) and thus \(\xi = 1\).

We can now formulate our main result.

**Theorem 3.2** Let \(S \subseteq \mathbb{R}^n\) fulfill the weak approximation property \((3.3)\) with constant \(K\). Furthermore, let \(\gamma \in [0, 1)\) be the smallest constant such that
\[
|\langle u, v \rangle_A| \leq \gamma \langle u, u \rangle_A^{\frac{1}{2}} \langle v, v \rangle_A^{\frac{1}{2}} \quad \forall u \in S^\perp, \forall v \in S. \tag{3.10}
\]
Then the effective condition number \(\kappa_{\text{eff}}\) of the matrix \(A(I - \pi_A(S))\) satisfies
\[
\kappa_{\text{eff}} \leq \frac{K}{(1 - \gamma)}. \tag{3.11}
\]

**Proof** Equation \((3.11)\) follows from \((3.4)\) and \((3.9)\).

This theorem gives a bound on the effective condition number of the deflated matrix \(A(I - \pi_A(S))\) which depends solely on the weak approximation constant \(K\) and the measure \(\xi\) on the \(A\)-invariance of the deflation subspace \(S\).
4 Applications

In this section we derive bounds on the effective condition number using Theorem 3.2 for different choices of the deflation subspace. First we show that in the classical situation of a deflation subspace spanned by eigenvectors the bound known from the literature \[9\] is retrieved. Moreover, we discuss the case where \(S\) is spanned by approximations of the eigenvectors. We also show that we can obtain estimates for the effective condition number for deflation subspaces which are given by the range of prolongation operators from algebraic multigrid methods \[5, 10, 21\].

4.1 Exact Eigenvalue Deflation

Let \(q_1, q_2, \ldots, q_n\) be an orthonormal basis consisting of eigenvectors of the matrix \(A\), s.t. \(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n\) are the corresponding eigenvalues. Furthermore, let \(S\) be spanned by the eigenvectors corresponding to the \((n - k)\) smallest eigenvalues, e.g., \(S = \text{span}\{q_k+1, \ldots, q_n\}\), where \(k \leq n\). It has been shown for this case (cf. \[9, Section 1\]) that \(\kappa_{\text{eff}} = \frac{\lambda_1}{\lambda_k}\). To demonstrate the quality of the bound \(K(1 - \gamma)\) from Theorem 3.2 we now show that in this case we actually have \(\frac{K}{(1 - \gamma)} = \kappa_{\text{eff}}\), i.e., the bound \(\frac{K}{(1 - \gamma)}\) for \(\kappa_{\text{eff}}\) is best possible.

Since the subspace \(S\) is \(A\)-invariant, we have \(\xi = 1\) due to Lemma 3.3. We now compute the smallest value for \(K\), such that the weak approximation property (3.3) is fulfilled. If \(\pi(S)x\) is the orthogonal projection onto \(S\) (in the 2-inner product) then

\[
\|x - \pi(S)x\|^2 = \min_{y \in S} \|x - y\|^2 = \text{dist}(S, x)^2.
\]

Expanding \(x\) in terms of the orthonormal eigenvectors \(q_i\) of \(A\) gives

\[
x = \sum_{i=1}^{n} \xi_i q_i.
\]

Then the orthogonal projection \(\pi(S)x\) of \(x\) onto \(S\) fulfills \(\pi(S)x = \sum_{i=k+1}^{n} \xi_i q_i\) and thus

\[
\text{dist}(S, x)^2 = \|x - \pi(S)x\|^2 = \|\sum_{i=k+1}^{n} \xi_i q_i\|^2 = \sum_{i=k+1}^{n} \|\xi_i q_i\|^2 = \sum_{i=k+1}^{n} |\xi_i|^2.
\]

This yields

\[
\|x\|^2_A = \sum_{i=1}^{n} |\xi_i|^2 \lambda_i \geq \sum_{i=k+1}^{n} |\xi_i|^2 \lambda_i \geq \lambda_k \sum_{i=k+1}^{n} |\xi_i|^2 = \lambda_k \text{dist}(S, x)^2. \tag{4.1}
\]

Hence the weak approximation property (3.3) holds with \(K = \frac{|A|}{\lambda_k} = \frac{\lambda_1}{\lambda_k}\).

Putting things together we see that the bound from Theorem 3.2 is \(\frac{K(1 - \gamma)}{\lambda_1} = \frac{\lambda_1}{\lambda_k}\), i.e., the bound is equal to the effective condition number \(\kappa_{\text{eff}}\) and thus best possible.
4.2 Inexact Eigenvalue Deflation

In many practical applications the eigenvectors corresponding the smallest eigenvalues are unknown and need to be approximated numerically. Thus the question arises how precisely we have to determine those eigenvectors to achieve fast convergence. In this section we establish a new qualitative convergence estimate for this setting. We show that the effective condition number $\kappa_{\text{eff}}$ converges towards the optimal effective condition number from Section 4.1 at least linearly w.r.t. the size of the eigenvalue approximation error.

Let

$$Q_1 = [q_{k+1}|q_{k+2}| \ldots |q_n] \quad \text{and} \quad Q_2 = [q_1|q_2| \ldots |q_k], \quad \text{i.e.,} \quad (4.2)$$

range $Q_1 = S$ and range $Q_2 = S^\perp$. Hence $S$ is the space spanned by the eigenvectors corresponding to the $k$ smallest eigenvalues. Let us now assume that we are given inexact eigenvectors. Define the perturbed deflation space by $\tilde{S} = \text{range}(Q_1 + E_1)$ where $E_1 \in \mathbb{K}^{n \times (n-k)}$ represents the perturbation of the eigenvectors. To compute a bound on the constant $K$ from the weak approximation property (3.3) and the constant $\gamma$ from the strengthened Cauchy-Schwarz inequality (3.10) we need the following lemma.

**Lemma 4.1** Let $Q_1, E_1 \in \mathbb{K}^{n \times (n-k)}$, $Q_2 \in \mathbb{K}^{n \times k}$, $Q = [Q_1|Q_2]$ be unitary $\|E_1\|_2 < 1$ and $\tilde{S} := \text{range}(Q_1 + E_1)$. Then there exist $W_1 \in \mathbb{K}^{n \times (n-k)}$ and $W_2 \in \mathbb{K}^{n \times k}$, s.t.

(i) the matrix $\tilde{Q} = [Q_1 + W_1|Q_2 + W_2]$ is unitary,

(ii) range($Q_1 + W_1$) $= \tilde{S}$,

(iii) range($Q_2 + W_2$) $= \tilde{S}^\perp$ and

(iv) $\delta := \max\{\|W_1\|_2, \|W_2\|_2\} \leq \frac{(1 + \sqrt{2})\|E_1\|_F}{1 - \|E_1\|_2}$.

Here $\|E_1\|_F$ denotes the Frobenius norm of the matrix $E_1$.

**Proof** Let $E = [E_1|0] \in \mathbb{K}^{n \times n}$ and let

$$Q + E = \tilde{Q} \tilde{R}$$

be the $QR$ decomposition of the matrix $(Q + E)$, where $\tilde{Q} \in \mathbb{K}^{n \times n}$ is unitary, $\tilde{R} \in \mathbb{K}^{n \times n}$ is upper triangular with positive diagonal entries. Define $W \in \mathbb{K}^{n \times n}$, $W_1 \in \mathbb{K}^{n \times (n-k)}$ and $W_2 \in \mathbb{K}^{n \times k}$ via

$$\tilde{Q} = Q + W = [Q_1 + W_1|Q_2 + W_2],$$

and note that

range($Q_1 + W_1$) $= \text{range}(Q_1 + E_1) = \tilde{S}$ as well as range($Q_2 + W_2$) $= \tilde{S}^\perp$.

Since $\|E\|_2 = \|E_1\|_2 < 1$ we can apply [22] Theorem 1.6 to the $QR$ decompositions of the matrices $Q\tilde{I}$ of $Q$ and $\tilde{Q}\tilde{R}$ of $\tilde{Q} + E$ to obtain

$$\|W\|_F \leq \frac{(1 + \sqrt{2})\|E\|_F}{1 - \|E\|_2}.$$

Using $\|E\|_2 = \|E_1\|_2$, $\|E\|_F = \|E_1\|_F$ as well as $\|W_1\|_2 \leq \|W_1\|_F \leq \|W\|_F$ and $\|W_2\|_2 \leq \|W_2\|_F \leq \|W\|_F$ completes the proof.
Lemma 4.2 Let the assumptions of Lemma 4.1 be fulfilled and \( Q_1, Q_2 \) be given by (4.1). Then the subspace \( \tilde{S} \) fulfills the weak approximation property with constant

\[
K = \left[ \sqrt{\kappa_\text{opt}} + \sqrt{2(2\delta + \delta^2)} \right]^2, \quad \text{where} \quad \kappa_\text{opt} = \frac{1}{\lambda_1}.
\]

Proof With \( \pi(\tilde{S}) = \tilde{Q}_1 \tilde{Q}_1^* \), we obtain

\[
\text{dist}(\tilde{S}, x)_2 = \| x - \tilde{Q}_1 \tilde{Q}_1^* x \|_2 = \| x - (Q_1 + W_1)(Q_1^* + W_1^*)^* x \|_2 \\
\leq \| x - Q_1 \tilde{Q}_1^* x \|_2 + \| Q_1 W_1^* x \|_2 + \| W_1 \tilde{Q}_1^* x \|_2 + \| W_1 W_1^* x \|_2 \cdot (4.3)
\]

Inspecting the terms on the right hand side individually yields

\[
\| x - Q_1 \tilde{Q}_1^* x \|_2 = \| x - \pi(S) x \|_2 = \text{dist}(S, x)_2 \leq \frac{1}{\sqrt{\lambda_n}} \| x \|_A \quad (4.4a)
\]
due to (4.1). Furthermore, we have \( \| Q_1 W_1^* x \|_2 \leq \| Q_1 \|_2 \| W_1^* \|_2 \| x \|_2 \) where \( \| Q_1 \|_2 = 1 \), \( \| W_1 \|_2 = \| W_1^* \|_2 \leq 1 \), \( \| x \|_2 \leq \frac{1}{\sqrt{\lambda_n}} \| x \|_A \). Thus

\[
\| Q_1 W_1^* x \|_2 \leq \frac{1}{\sqrt{\lambda_n}} \delta \| x \|_A \cdot (4.4b)
\]

Analogously,

\[
\| W_1 \tilde{Q}_1^* x \|_2 \leq \frac{1}{\sqrt{\lambda_n}} \delta \| x \|_A \quad \text{and} \quad \| W_1 W_1^* x \|_2 \leq \frac{1}{\sqrt{\lambda_n}} \delta^2 \| x \|_A \cdot (4.4c)
\]

Inserting (4.4a) to (4.4c) into (4.3) yields

\[
\text{dist}(\tilde{S}, x)_2 \leq \left[ \frac{1}{\sqrt{\lambda_n}} + \frac{1}{\sqrt{\lambda_n}} (2\delta + \delta^2) \right] \| x \|_A ,
\]

and thus

\[
\text{dist}(\tilde{S}, x)_2^2 \leq \frac{1}{\lambda_1} \left[ \frac{1}{\sqrt{\lambda_n}} + \frac{1}{\sqrt{\lambda_n}} (2\delta + \delta^2) \right]^2 \| x \|_A^2 ,
\]

which proves the assertion.

The next step is to determine the constant \( \gamma \) from the strengthened Cauchy-Schwarz inequality. This involves the following auxiliary result.

Lemma 4.3 Let \( M \in \mathbb{K}^{m \times n} \) and \( x \in \mathbb{K}^n \). Furthermore let \( U \in \mathbb{K}^{m \times n} \) have orthonormal columns. Then

\[
\| Mx \|_A \leq \sqrt{\alpha} \| M \|_2 \| Ux \|_A
\]

Proof We have

\[
\| Mx \|_A = \| A^{1/2} Mx \|_2 \leq \| A^{1/2} \|_2 \| M \|_2 \| x \|_2 = \sqrt{\lambda_1} \| M \|_2 \| x \|_2 = \sqrt{\lambda_1} \| M \|_2 \| Ux \|_2 , \quad \text{and} \quad \| Ux \|_2 \leq \frac{1}{\sqrt{\lambda_n}} \| Ux \|_A .
\]
The sought-after estimate for $\gamma$ is given by the following Lemma.

**Lemma 4.4** Let the assumptions of Lemma 4.1 be fulfilled, $Q_1, Q_2$ be given by (4.2) and $\sqrt{\kappa \delta} < 1$. Then the strengthened Cauchy-Schwarz inequality (3.10) is fulfilled for

$$\gamma = \frac{2\sqrt{\kappa \delta} + \kappa \delta^2}{(1 - \sqrt{\kappa \delta})^2} = 1 - \frac{1 - 4\sqrt{\kappa \delta}}{(1 - \sqrt{\kappa \delta})^2},$$

and $\gamma \leq 1$ if $\sqrt{\kappa \delta} < \frac{1}{4}$.

**Proof** Let $u \in \tilde{S}$ and $v \in \tilde{S}^\perp$. There exist $x \in \mathbb{K}^{n-k}$ and $y \in \mathbb{K}^k$ s.t. $u = \tilde{Q}_1 x$ and $v = \tilde{Q}_2 x$. We have

$$\langle u, v \rangle_A = \langle \tilde{Q}_1 x, \tilde{Q}_2 y \rangle_A = \langle (Q_1 + W_1) x, (Q_2 + W_2) y \rangle_A = \langle Q_1 x, Q_2 y \rangle_A + \langle Q_1 x, W_2 y \rangle_A + \langle W_1 x, Q_2 y \rangle_A + \langle W_1 x, W_2 y \rangle_A. \tag{4.5}$$

The columns of $Q_1$ are eigenvectors of $A$, thus range $AQ_1 = \text{range } Q_1 \perp \text{range } Q_2$.

Therefore

$$\langle Q_1 x, Q_2 y \rangle_A = 0. \tag{4.6a}$$

By the Cauchy-Schwarz inequality we have

$$\langle Q_1 x, W_2 y \rangle_A \leq \|Q_1 x\|_A \|W_2 y\|_A$$

and using Lemma 4.3 (for $M = W_2, U = Q_2$) yields

$$\langle Q_1 x, W_2 y \rangle_A \leq \|Q_1 x\|_A \sqrt{\kappa} \|W_2\|_2 \|Q_2 y\|_A,$$

and thus due to $\|W_2\|_2 \leq \delta$

$$\langle Q_1 x, W_2 y \rangle_A \leq \sqrt{\kappa \delta} \|Q_1 x\|_A \|Q_2 y\|_A. \tag{4.6b}$$

Analogously we get

$$\langle W_1 x, Q_2 y \rangle_A \leq \sqrt{\kappa \delta} \|Q_1 x\|_A \|Q_2 y\|_A, \tag{4.6c}$$

$$\langle W_1 x, W_2 y \rangle_A \leq \kappa \delta^2 \|Q_1 x\|_A \|Q_2 y\|_A. \tag{4.6d}$$

Inserting (4.6a) to (4.6d) into (4.5) yields

$$\langle u, v \rangle_A \leq (2\sqrt{\kappa \delta} + \kappa \delta^2) \|Q_1 x\|_A \|Q_2 y\|_A. \tag{4.7}$$

By Lemma 4.3 we have

$$\|(Q_1 + W_1) x\|_A \geq \|Q_1 x\|_A - \|W_1 x\|_A \geq (1 - \sqrt{\kappa \delta})\|Q_1 x\|_A, \tag{4.8a}$$

and, analogously,

$$\|(Q_2 + W_2) y\|_A \geq (1 - \sqrt{\kappa \delta})\|Q_2 y\|_A. \tag{4.8b}$$

From (4.7), (4.8a) and (4.8b) therefore finally get

$$\langle u, v \rangle_A \leq \frac{2\sqrt{\kappa \delta} + \kappa \delta^2}{(1 - \sqrt{\kappa \delta})^2} \|Q_1 + W_1\|_A \|Q_2 + W_2\|_A = \gamma \|u\|_A \|v\|_A.$$
Combining the results on the weak approximation property and the strengthened Cauchy-Schwarz inequality we now obtain our bound on the effective condition number $\kappa_{\text{eff}}$.

**Theorem 4.1** Let the assumptions of Lemma 4.2 be fulfilled and $\delta < \frac{1}{4\sqrt{\kappa}}$ then

$$\kappa_{\text{eff}} \leq \left[ \sqrt{\kappa_{\text{opt}}} + \delta \left( \frac{1}{4} + 2\sqrt{\kappa} \right) \right]^2 = \kappa_{\text{opt}} + O(\delta) \quad \text{for} \quad \delta \to 0.$$  

**Proof** Due to Theorem 3.2, Lemma 4.2 and Lemma 4.4

$$\kappa_{\text{eff}} \leq \left[ \sqrt{\kappa_{\text{opt}}} + \sqrt{\kappa} (2\delta + \delta^2) \right]^2 = \left[ \sqrt{\kappa_{\text{opt}}} + \sqrt{\kappa} (2\delta + \delta^2) \right]^2 \frac{(1 - \sqrt{\kappa}\delta)^2}{1 - 4\sqrt{\kappa}\delta}. \quad (4.9)$$

Since $0 \leq \delta < \frac{1}{4\sqrt{\kappa}}$ we have $(2\delta + \delta^2) < (2\delta + \frac{1}{4\sqrt{\kappa}})$ and $(1 - \sqrt{\kappa}\delta)^2 < 1$, thus

$$\kappa_{\text{eff}} \leq \left[ \sqrt{\kappa_{\text{opt}}} + \delta \left( \frac{1}{4} + 2\sqrt{\kappa} \right) \right]^2 \frac{1}{1 - 4\sqrt{\kappa}\delta}.$$  

To obtain the asymptotic expansion for $\delta \to 0$ we note that

$$\frac{1}{1 - 4\sqrt{\kappa}\delta} = 1 + O(\delta) \quad \text{for} \quad \delta \to 0,$$

which gives

$$\kappa_{\text{eff}} \leq \left[ \sqrt{\kappa_{\text{opt}}} + \delta \left( \frac{1}{4} + 2\sqrt{\kappa} \right) \right]^2 (1 + O(\delta)) = \kappa_{\text{opt}} + O(\delta).$$

Lemma 4.1 gives a bound for $\delta$ in terms of $\|E_1\|$, namely $\delta \leq \frac{1 + \sqrt{2}}{1 - \|E_1\|_2} \|E_1\|_F$. Since $\frac{1}{1 - \|E_1\|_2} = 1 + O(\|E_1\|_2)$ for $\|E_1\|_2 \to 0$ we therefore have

$$\delta \leq \left( 1 + \sqrt{2} \right) (1 + O(\|E_1\|_2)) \|E_1\|_F = O(\|E_1\|_F) \quad \text{for} \quad \delta \to 0.$$  

Using $\|E_1\|_2 \leq \|E_1\|_F$ we thus obtain the following Corollary to Theorem 4.1 which expresses the effective condition number $\kappa_{\text{eff}}$ in terms of the initial perturbation of the eigenvectors $q_{k+1}, \ldots, q_n$.

**Corollary 4.1** Under the assumptions of Theorem 4.1, $\kappa_{\text{eff}} = \kappa_{\text{opt}} + O(\|E_1\|_F)$.

4.3 Deflation Subspaces for $M$-Matrices

The classical algebraic multigrid method [5,16,21] was specifically designed for the case that $A \in \mathbb{R}^{n \times n}$ is a symmetric $M$-matrix, i.e., $A$ is symmetric positive definite and $a_{ij} \leq 0$ for $i \neq j$.

Given a splitting of the variables $\{1, \ldots, n\} = \mathcal{C} \cup \mathcal{F}$ the direct interpolation operator $V : \mathbb{R}^{\mathcal{C}} \to \mathbb{R}^{\mathcal{C}}$ of the algebraic multigrid method [21] fulfills

$$\|e - V Re\|_D^2 \leq \tau \|e\|_A^2,$$
where $D = \text{diag}(A)$. In here $\tau$ depends only on the matrix coefficients and the splitting. Consequently, it has been shown in [21] that there exist classes of matrices for which $\tau$ is bounded and which contain matrices of arbitrary size. Using this result and Lemma 3.2 we find that (3.3) holds with

$$K = \frac{\|A\|}{\min_i a_{ii}} \tau$$

and thus is bounded independently of the matrix size, as long as $\|A\|$ is bounded.

**Example 4.2** Consider the bi-linear quadratic finite element discretization on a uniform $N \times N$ grid of the Laplace equation on $[0,1]^2$ with Dirichlet boundary conditions which is given by the stencil

$$
\begin{pmatrix}
-1 & -1 & -1 \\
-1 & 8 & -1 \\
-1 & -1 & -1
\end{pmatrix}.
$$

Straightforward calculation shows that if $C$ is given by full-coarsening, i.e., choosing all grid points on even rows and columns as coarse variables, $\tau = 4$ and thus $K = 8$ since by Gershgorin’s theorem $\|A\| \leq 16$.

5 Numerical Experiments

This section contains numerical experiments to illustrate the developed theory and the quality of the derived convergence estimates. In our experiments we consider the matrix $A$ and the direct interpolation operator $V$ of Example 4.2 and use the deflation subspace $S = \text{range } V$. Among the various mathematically equivalent ways to implement the deflated CG method we choose the one from [18] for its numerical stability.

5.1 Independence of the Grid Size

The discussion in Example 4.2 showed that 8 is an upper bound for the constant $K$, independent of the grid size $N$. By Theorem 3.2 we thus expect the method to converge in a constant number of iterations provided the abstract angle $\gamma$ is independent of $N$.

We choose a random right hand side $b$ such that the solution $x$ fulfills $\|x\| = 1$. Then we run the deflated CG method [18] until the residual $r_i$ of the $i^{th}$ iterate satisfies $\|r_i\| \leq 10^{-6}$. The number of iterations for different grid sizes $N = 2^p - 1$ is listed in Table 5.1. We observe that the number of iterations stays indeed constant.

5.2 Numerical Computation of the Constants

In order to verify our theory we numerically compute the condition numbers $\kappa$, $\kappa_{\text{eff}}$ and the constants $K$ and $\gamma$, then comparing $\kappa_{\text{eff}}$ with its estimate $\frac{K}{\gamma}$. This approach is only feasible for small systems since it involves the computation of eigenvectors of dense matrices which are of the same size as the linear system.
Analysis of the Deflated CG Method Based on Symmetric Multigrid Theory

| p | Iterations | Residual Error |
|---|------------|----------------|
| 4 | 8          | 5.64429 \cdot 10^{-7} | 1.53486 \cdot 10^{-7} |
| 5 | 8          | 8.4304 \cdot 10^{-7}  | 4.64197 \cdot 10^{-7} |
| 6 | 9          | 5.27683 \cdot 10^{-7} | 1.63928 \cdot 10^{-6} |
| 7 | 9          | 6.11646 \cdot 10^{-7} | 5.74174 \cdot 10^{-6} |
| 8 | 9          | 6.36346 \cdot 10^{-7} | 2.56345 \cdot 10^{-5} |
| 9 | 9          | 6.57814 \cdot 10^{-7} | 6.60374 \cdot 10^{-5} |

Table 5.1 Number of iterations where \( N = 2^p - 1 \).

\[
A(I - \pi_A(S)) = \lambda_1 \begin{bmatrix} 0.0577 \\ 6.0708 \\ 0.3369 \\ 2.9715 \end{bmatrix} + \mu_1 
\]

\[
A = \begin{bmatrix} 10^{-2} \\ 10^{-1} \\ 1 \end{bmatrix} = \begin{bmatrix} 0.0577 \\ 6.0708 \\ 0.3369 \\ 2.9715 \end{bmatrix} \]

\[
\|x\|_A^2 = \langle Ax, x \rangle = \langle QAQ^*x, x \rangle = \langle A^{\frac{1}{2}}Q^*x, A^{\frac{1}{2}}Q^*x \rangle = \|A^{\frac{1}{2}}Q^*x\|^2. \tag{5.1}
\]

\[
K = \|A\|_2 \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\text{dist}(S, x)^2}{\|x\|_A^2} = \|A\|_2 \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|WW^*x\|^2}{\|x\|_A^2}. \tag{5.2}
\]

Using the eigendecomposition \( A = Q\Lambda Q^* \) yields

\[
\|x\|_A^2 = \langle Ax, x \rangle = \langle QAQ^*x, x \rangle = \langle A^{\frac{1}{2}}Q^*x, A^{\frac{1}{2}}Q^*x \rangle = \|A^{\frac{1}{2}}Q^*x\|^2. \tag{5.2}
\]

Since we are interested in the supremum over all vectors in \( \mathbb{R}^n \setminus \{0\} \) we can substitute \( x \) by \( QA^{-\frac{1}{2}}z \) in (5.1). Due to (5.2) this yields

\[
K = \|A\|_2 \sup_{z \in \mathbb{R}^n \setminus \{0\}} \frac{\|WW^*QA^{-\frac{1}{2}}z\|^2}{\|z\|_2^2} = \|A\|_2 \|WW^*QA^{-\frac{1}{2}}\|_2. \]

The matrix norms on the right hand side can numerically be computed via the singular value decomposition. The constant \( \gamma \) from the strengthened Cauchy-Schwarz inequality can be computed by the method from [13] which is also based on the singular value decomposition.

The results for \( N = 2^5 - 1 \) are given in Table 5.2. We see that \( \frac{K}{1-\gamma} \) is a good estimate – it is only a factor of 1.5 off – for the condition number \( \kappa_{\text{eff}} \) of the deflated matrix in this example.

5.3 Perturbation for Eigenvector Deflation

In order to illustrate the discussion of the perturbation theory from Section 4.2 we consider a matrix \( A \in \mathbb{R}^{100 \times 100} \) with a simple eigenvalue \( 10^{-2} \) and a multiple eigenvalue of one. We choose the deflation subspace \( S = \text{span}\{v_1\} \), where \( v_1 \) is
the eigenvector corresponding to the eigenvalue $10^{-2}$. We perturb the deflation subspace by a vector orthogonal to $v_1$. In Figure 5.1 we compare the estimate \( (4.9) \) to the actual effective condition number.

We note that the estimate is relatively pessimistic. This is expected due to the fact that the derivation required many bounds on 2-norms in terms of $A$-norms and vice versa. Nevertheless, the asymptotic behavior is reproduced correctly by the estimate.

6 Influence of the Accuracy of Computations

The deflated CG method involves the solution of the “inner” linear system

\[
(V^*AV)z_{i+1} = V^*Ar_{i+1}
\]  

in every iteration. If the dimension $m$ of the deflated subspace, i.e., the number of columns of $V$ is small, we can solve (6.1) exactly up to numerical errors, e.g., by using a factorization of $V^*AV$. Often, however, $m$ will be large (e.g., $m = n/4$ in the numerical experiments of Section 5), so that solving (6.1) will be done using an “inner” iteration. Its accuracy is decisive for the convergence process of the overall iteration. This can be motivated as follows.

Recall that we can think of deflated CG as applying the CG method to the linear system

\[
A(I - V(V^*AV)^{-1}V^*)\tilde{x} = (I - AV(V^*AV)^{-1}V^*)b.
\]  

In here, the kernel of $I - \pi_A(S)$ is $S$, the column range of $V$. To outline what possibly could go wrong if we solve (6.1) inaccurately, let us assume we are solving the system by a fixed number of iterations of a stationary method. That means that we are given an approximation $M \in \mathbb{R}^{m \times m}$ for the matrix $(V^*AV)^{-1}$. If we
then replace \((V^*AV)^{-1}\) by \(M\) in \(I - V(V^*AV)^{-1}V^*A\), the operator is usually not a projection anymore, and \(\mathcal{S}\) is not its kernel. That is, in general the matrix of the approximately deflated system

\[ A(I - VMV^*A) \]

is non-singular. The approximately deflated system will therefore loose the property of having a considerable number of zero eigenvalues, an essential ingredient to estimate the effective condition number \(\kappa_{\text{eff}}\) in Theorem 3.2. Even worse, the eigenvalues that would be mapped to zero by the exact \(A\)-orthogonal projection now remain as small non-zero eigenvalues which leads to a potentially larger condition number of the approximately deflated system than the one of the original system.

Within the deflated CG method the matrix \(A(I - \pi_A(\mathcal{S}))\) is only used to compute matrix vector products. From the previous discussion we might conclude that we should compute this matrix vector product as accurately as possible. However, it is known that the speed of convergence of the CG method does not suffer from inexact matrix vector products, if accuracy is high enough \([19,25]\). Hence the question remains how to determine a suitable stopping criterion for the inner iteration based on the stopping criterion of the outer iteration

\[ \|r_i\| \leq \tau \|b\| =: \varepsilon \quad (6.2) \]

for some \(0 < \tau \ll 1\). More precisely, how do we have to choose \(\tau^c\) for the inner stopping criterion

\[ \|r_i^c\| \leq \tau^c \|b^c\| , \]

to achieve (6.2)? A first strategy may be to set \(\tau^c = \varepsilon\) but it turns out that we can relax this requirement. Since, our problem is equivalent to the question, how accurately we have to compute the matrix vector product \(A(I - \pi_A(\mathcal{S}))p\) for \(p \in \mathbb{R}^n\) in the outer CG iteration, we can use the results from \([19,25]\). There it is proposed to use

\[ \tau^c = \max \left\{ \frac{\varepsilon}{\|r_i\|}, \varepsilon \right\} \cdot c \quad \text{with} \quad 0 < c \leq 1 \]

in the \(i\)th outer iteration. That is the relative tolerance for the inner iteration can be relaxed while the outer iteration advances.

7 Conclusions

We derived a convergence estimate for the deflated CG method using techniques from algebraic multigrid theory and have shown that our theory recovers the results for the convergence estimate for deflation using exactly known eigenvectors. In addition, our theory allowed us to derive a bound on the effective condition number with respect to perturbations of exact eigenvectors.

By combining the deflation ansatz with the ideas of algebraic multigrid we not only gave a proof of convergence for deflation subspaces spanned by multigrid prolongation operators, but also allowed for the theoretical analysis of more general deflation subspaces that are not necessarily \(A\)-invariant and do not need to be spanned by (approximate) eigenvectors. In this manner all tools developed for
the construction of efficient (algebraic) multigrid interpolation operators can be facilitated to construct improved deflation subspaces.

Finally, the developed theory suggests that using the multigrid coarse-grid correction in a deflated conjugate gradient method yields a scalable—in the sense of constant number of iterations when increasing the resolution of the discretization—iterative method without the need of constructing a additional smoothing iteration. This might be attractive in situations where such an iteration is hard to come by or not known all together.

References

1. O. Axelsson and I. Gustafsson. Preconditioning and two-level multigrid methods of arbitrary degree of approximation. *Math. Comp.*, 40(161):219–242, 1983.
2. A. Brandt. Algebraic multigrid theory: The symmetric case. *Appl. Math. Comput.*, 19(1-4):23–56, 1986.
3. M. Brezina, A. J. Cleary, R. D. Falgout, V. E. Henson, J. E. Jones, T. A. Manteuffel, S. F. McCormick, and J. W. Ruge. Algebraic multigrid based on element interpolation (AMGe). *SIAM J. Sci. Comput.*, 22(3):1570–1592, 2001.
4. M. Brezina, R. D. Falgout, S. MacLachlan, T. A. Manteuffel, S. F. McCormick, and J. W. Ruge. Adaptive smoothed aggregation (oSA) multigrid. *SIAM Rev.*, 47(2):317–346, 2005.
5. W. L. Briggs, V. E. Henson, and S. F. McCormick. *A Multigrid Tutorial*. SIAM, 2nd edition, 2000.
6. Z. Dostál. Conjugate gradient method with preconditioning by projector. *Int. J. Comput. Math.*, 29(3-4):315–323, 1988.
7. V. Eijkhout and P. S. Vassilevski. The role of the strengthened Cauchy-Buniakowski-Schwarz inequality in multilevel methods. *SIAM Rev.*, 33(3):405–419, 1991.
8. R. D. Falgout, P. S. Vassilevski, and L. T. Zikatanov. On two-grid convergence estimates. *Numer. Linear Algebra Appl.*, 12(5-6):471–494, 2005.
9. J. Frank and C. Vuik. On the construction of deflation-based preconditioners. *SIAM J. Sci. Comput.*, 23(2):442–462, 2001.
10. A. Gaul, M. Gutknecht, J. Liesen, and R. Nabben. A framework for deflated and augmented krylov subspace methods. *SIAM J. Matrix Anal. Appl.*, 34(2):495–518, 2013.
11. M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49:409–436 (1953), 1952.
12. E. F. Kaasschieter. Preconditioned conjugate gradients for solving singular systems. *J. Comput. Appl. Math.*, 24(1-2):265–275, 1988.
13. A. V. Knyazev and M. E. Argentati. Principal angles between subspaces in an A-based scalar product: algorithms and perturbation estimates. *SIAM J. Comput.*, 23(6):454–477, 2003.
14. Y. Saad, M. Yeung, J. Erhel, and F. Guyomarc’ch. A deflated version of the conjugate gradient algorithm. *SIAM J. Sci. Comput.*, 21(5):1909–1926, 2000.
15. A. Stathopoulos and K. Orginos. Computing and deflating eigenvalues while solving multiple right-hand side linear systems with an application to quantum chromodynamics. *SIAM J. Sci. Comput.*, 32(1):439–462, 2010.
16. K. Stüben. An introduction to algebraic multigrid. In U. Trottenberg, C. W. Oosterlee, and A. Schüller, editors, *Multigrid*. Academic Press, 2001.
17. Ji Guang Sun. Perturbation bounds for the Cholesky and QR factorizations. *BIT*, 31(2):341–352, 1991.
23. J. M. Tang, S. P. MacLachlan, R. Nabben, and C. Vuik. A comparison of two-level preconditioners based on multigrid and deflation. *SIAM J. Matrix Anal. Appl.*, 31(4):1715–1739, 2010.

24. J. M. Tang, R. Nabben, C. Vuik, and Y. A. Erlangga. Comparison of two-level preconditioners derived from deflation, domain decomposition and multigrid methods. *J. Sci. Comput.*, 39(3):340–370, 2009.

25. J. van den Eshof and G. L. G. Sleijpen. Inexact Krylov subspace methods for linear systems. *SIAM J. Matrix Anal. Appl.*, 26:125–153, 2004.

26. A. van der Sluis and H. A. van der Vorst. The rate of convergence of conjugate gradients. *Numer. Math.*, 48(5):543–560, 1986.

27. J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford University Press, 1967.