Bergische Universität Wuppertal

Fachbereich Mathematik und Naturwissenschaften

Institute of Mathematical Modelling, Analysis and Computational
Mathematics (IMACM)

Preprint BUW-IMACM 12/19

K. Kahl and H. Rittich

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

September 2012

http://www.math.uni-wuppertal.de

ANALYSIS OF THE DEFLATED CONJUGATE GRADIENT METHOD BASED ON SYMMETRIC MULTIGRID THEORY

K. KAHL[*] AND H. RITTICH[*]

Abstract. Deflation techniques for Krylov subspace methods and in particular the conjugate gradient method have seen a lot of attention in recent years. They provide means to improve the convergence speed of the methods in a rather straight forward way 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. In addition the result suggests that the techniques developed to construct efficient interpolation operators in algebraic multigrid methods can also be applied to improve deflation subspaces.

Key words. conjugate gradients; deflation; algebraic multigrid; convergence analysis

1. Preliminaries.

Consider solving the linear system of equations

$$Ax = b\,, \tag{1.1}$$

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 [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 κ of the matrix A, more precisely on the distribution of its eigenvalues [17, 22]. When the condition number κ is large it can become necessary 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 [15] and Dostal [6], see also [9, 10, 14, 18]. We mention in particular the paper [18], which derives a variant of Nicolaides' deflated cg that is mathematically equivalent but is algorithmically much closer to the standard cg algorithm. 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 so called *deflation subspace* $\mathcal{S} \subseteq \mathbb{K}^n$.

In [15] the space $\mathcal{S}$ is constructed as follows. The variables are combined into aggregates $\mathfrak{A}_i \subseteq \{1, 2, \ldots, n\}$, $i = 1, \ldots, m$ such that

$$\bigcup_{i=1}^{m} \mathfrak{A}_i = \{1, 2, \ldots, n\} \quad \text{and} \quad \mathfrak{A}_i \cap \mathfrak{A}_j = \emptyset \quad \text{for } i \neq j\,.$$

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

$$v_j^{(i)} = \begin{cases} 1 & \text{if } i \in \mathcal{V}_j \\ 0 & \text{otherwise} \end{cases}.$$

[*] Fachbereich Mathematik und Naturwissenschaften, Bergische Universität Wuppertal, 42097 Wuppertal, Germany {kkahl, rittich}@math.uni-wuppertal.de

This procedure can be motivated in the following way: Assume that the matrix A arises from the discretization of a 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 $\mathfrak{A}_i$ are chosen appropriately, $\mathcal{S}$ 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 [15] is completely analogous to the construction of prolongation operators in (non-smoothed) aggregation based multigrid methods [4].

Another viable and widely used approach for deflation, consists of spanning $\mathcal{S}$ directly by the eigenvectors corresponding to the smallest eigenvalues [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 extremal, i.e., very small eigenvalues exist, independent of the system size, however, this approach works reasonably well.

Recently a combination of the two approaches has been rediscovered in the context of simulations in Quantum Chromodynamics [14]. Similarly to [15], aggregates $\mathfrak{A}_i$ 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 $\mathfrak{A}_i$ the orthogonal projection $\tilde{w}_j^{(i)}$ of w_j onto

$$\mathcal{V}_i = \operatorname{span}\{e^{(j)} : j \in \mathfrak{A}_i\}, \text{ where } e^{(i)} \text{ with } e^{(j)}_\ell = \delta_{j,\ell} \text{ is the } j^{\text{th}} \text{ canonical vector,}$$

is computed and the deflation subspace $\mathcal{S}$ 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 ℓ 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.

Motivated by these similarities to multigrid methods we investigate in this paper more closely the relation between the ranges of good multigrid prolongation operators and good deflation subspaces for the cg method. In doing so we analyze the convergence of deflated cg with techniques known from algebraic multigrid methods, see [3, 16, 20]. The theory of algebraic multigrid measures the quality of a prolongation operator by a *weak approximation property*. We show that the speed of convergence of deflated cg can be estimated using the weak approximation property, showing that prolongation operators that work well in the multigrid setting also yield good results when used to span the deflation subspace. Furthermore, with this choice of the deflation subspace, deflated cg exhibits similar scaling behavior as multigrid methods for many applications. This means that the number of iterations stays constant when we increase the system size due to finer discretizations, and, as opposed to standard (algebraic) multigrid, deflated cg does not require a smoother.

We finish this introduction explaining some basic notation. Assume that $\tilde{x} \in \mathbb{K}^n$ is an approximation to x, the solution of (1.1). Then the residual $r \in \mathbb{K}^n$ is given by

$r = b - A\tilde{x}$, the error by $e = x - \tilde{e}$. Note that $Ae = r$.

Let $\langle x, y\rangle = y^*x$ be the euclidean inner product. Since A is self-adjoint and positive definite we can define the A-inner product and the A-norm by

$$\langle x, y\rangle_A := \langle Ax, y\rangle \quad \text{and} \quad \|x\|_A := \sqrt{\langle x, x\rangle_A}\,.$$

We denote by $\mathbb{K}_k[t]$ the set of polynomials in the variable t with degree less than or equal to k. Let $\mathcal{S} \subseteq \mathbb{K}^n$ be a subspace, then $\mathcal{S}^\perp$ is its orthogonal complement with respect to the 2-inner product and $\mathcal{S}^{\perp_A}$ is its orthogonal complement with respect to the A-inner product. By $\pi(\mathcal{S}) \in \mathbb{K}^{n\times n}$ we denote the orthogonal projection onto $\mathcal{S}$ for the 2-inner product and by $\pi_A(\mathcal{S})$ its counter part for the A-inner product. The distance between a point $x \in \mathbb{K}^n$ and a subspace $\mathcal{S} \subseteq \mathbb{K}^n$ is given by

$$\operatorname{dist}(\mathcal{S}, x) = \min_{y\in\mathcal{S}} \|x - y\|\,.$$

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 matrix $A(I - \pi_A(\mathcal{S})) \in \mathbb{K}^{n\times n}$ which we estimate by using some results from multigrid theory. In Section 4 we further show that our general convergence result yields the known bounds derived for the case, where eigenvectors are directly used to span the deflation space. Moreover we demonstrate how prolongation operators from the classical algebraic multigrid theory for M-matrices can be used to obtain deflation subspaces. Then, we present some numerical experiments confirming the theory in Section 5. Finally in Section 6 we discuss how accurately the the action of the operator $\pi_A(\mathcal{S})$ has to be computed to achive overall convergence of the method.

2. Review of Deflated CG.

The m^{th} Krylov subspace $\mathcal{K}_m(A, v)$ corresponding to a matrix $A \in \mathbb{K}^{n\times n}$ and a vector $v \in \mathbb{K}^n$ is given by

$$\mathcal{K}_m(A, v) := \operatorname{span}\{v, Av, A^2v, \ldots, A^{m-1}v\} = \{P(A)v : P \in \mathbb{K}_{m-1}[t]\}\,.$$

The cg method generates the iterates $x_1, x_2, x_3, \ldots \in \mathbb{K}^n$ for a given initial guess $x_0 \in \mathbb{K}^n$, where

$$x_i = x_0 + \tilde{e}_i \quad \text{and} \quad \tilde{e}_i \in \mathcal{K}_i(A, r_0)$$

such that the error $e_i = x - x_i$ fulfills

$$\|e_i\|_A = \|x - x_i\|_A = \min_{\tilde{x}\in x_0+\mathcal{K}_i} \|x - \tilde{x}\|_A\,.$$

Note that since $e_i = x - x_i = x - (x_0 + \tilde{e}_i) = e_0 - \tilde{e}_i$, in particular $e_0 = x - x_0$, and putting $\tilde{e} = \tilde{x} - x_0$ this is equivalent to

$$\|e_i\|_A = \|e_0 - \tilde{e}_i\|_A = \min_{\tilde{e}\in\mathcal{K}_i} \|e_0 - \tilde{e}\|_A\,.$$

By definition of the Krylov subspace, and since $\tilde{e}_i \in \mathcal{K}_i$, we can write

$$\tilde{e}_i = \tilde{P}(A)r_0 = \tilde{P}(A)Ae_0$$

with $\tilde{P} \in \mathbb{K}_{i-1}[t]$ and thus

$$e_i = e_0 - \tilde{e}_i = e_0 - \tilde{P}(A)Ae_0 = P(A)e_0 \quad \text{where} \quad P = 1 - t\tilde{P}\,. \tag{2.1}$$

A polynomial P can be written in the form (2.1) if and only if $P \in \mathbb{K}_i[t]$ and $P(0) = 1$. Hence

$$\|e_i\|_A = \min_{\substack{P \in \mathbb{K}_i[t] \\ P(0)=1}} \|P(A)e_0\|_A \,. \tag{2.2}$$

Recall that the matrix A is self-adjoint positive definite so that there exists a unitary matrix $Q \in \mathbb{K}^{n\times n}$ and a corresponding diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n) \in \mathbb{K}^{n\times n}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ yielding the eigen-decomposition $A = Q\Lambda Q^*$. The columns $q_1, \dots, q_n \in \mathbb{K}^n$ of Q form an orthonormal basis of eigenvectors corresponding to the eigenvalues $\lambda_1, \dots, \lambda_n$.

Writing the initial error e_0 as $e_0 = Q\xi = \sum_{j=1}^n \xi_j q_j$ and using $P(A)q_j = P(\lambda_j)q_j$ equation (2.2) yields

$$\|e_i\|_A^2 = \min_{\substack{P \in \mathbb{K}_i[t] \\ P(0)=1}} \|\sum_{j=1}^n \xi_j P(A)e_0\|_A^2 = \min_{\substack{P \in \mathbb{K}_i[t] \\ P(0)=1}} \sum_{j=1}^n |\xi_j|^2 |P(\lambda_j)|^2 \lambda_i \,. \tag{2.3}$$

Thus the cg method implicitly constructs for each iterate x_i a polynomial P_i of degree at most i which interpolates the point $(0, 1)$ and approximates the points $(\lambda_i, 0)$ by minimizing the sum in (2.3) where the values $P_i(\lambda_j)$ are weighted by $|\xi_j|^2\lambda_j$. Clearly, the minimum in (2.3) will be smaller, and the convergence of the cg iteration will thus be faster, if the weights are small, and—given the constraint $P_i(0) = 1$—particularly so if the weights corresponding to small eigenvalues are very small. With this interpretation of cg convergence, deflation can be regarded as a technique to make the weights small by making $|\xi_j|$ small for large j, i.e., small eigenvalues.

Assume that we are given a subspace $\mathcal{S} \subset \mathbb{K}^n$ which contains (approximated) eigenvectors corresponding to the smallest eigenvalues of A. Since deflation requires a first preparatory step before starting the cg iteration, we from now on denote the initial guess by x_{-1} with initial error $e_{-1} = x - x_{-1}$. We want to compute a new vector x_0 with error $\tilde{e}_0$ such that the part of e_0 belonging to $\mathcal{S}$ is "removed". Setting $x_0 = x_{-1} + \tilde{e}_0$ and $e_0 = e_{-1} - \tilde{e}_0$ we thus want $\tilde{e}_0$ to be a projection of e_{-1} onto $\mathcal{S}$. Taking the A-orthogonal projection, i.e., $\tilde{e}_0 = \pi_A(\mathcal{S})e_{-1}$ is particularly adequate: Let the columns of $V \in \mathbb{K}^{n\times m}$ be an arbitrary basis of $\mathcal{S}$, i.e., $\operatorname{range}(V) = \mathcal{S}$. Then

$$\pi_A(\mathcal{S}) = V(V^*AV)^{-1}V^*A \,.$$

Since

$$\tilde{e}_0 = \pi_A(\mathcal{S})e_{-1} = V(V^*AV)^{-1}V^*Ae_{-1} = V(V^*AV)^{-1}V^*r_{-1}$$

we can compute $\tilde{e}_0$ and thus x_0 without explicit knowledge of e_{-1}. Moreover, because of

$$\|e_{-1} - \tilde{e}_0\|_A^2 = \|e_0\|_A^2 = \sum_{i=1}^n |\xi_i|^2 \lambda_i$$

the choice $\tilde{e}_0 = \pi_A(\mathcal{S})e_{-1}$ precisely minimizes the sum of all weights in (2.3).

Computationally, since

$$e_{-1} = \pi_A(\mathcal{S})e_{-1} + (I - \pi_A(\mathcal{S}))e_{-1} = \tilde{e}_0 + e_0$$

and $x = x_{-1} + e_{-1}$, we obtain the solution x of (1.1) once we have computed $(I - \pi_A(\mathcal{S}))e_{-1}$, the A-orthogonal projection of e_{-1} onto $\mathcal{S}^{\perp_A}$. Modifying the cg method to restrict the search directions to the subspace $\mathcal{S}^{\perp_A}$ and thus minimizing the A-norm of the error over the subspace $\mathcal{S}^{\perp_A}$ yields the deflated cg algorithm from [18] which computes the desired projection, described in Algorithm 1.

Algorithm 1 Deflated Conjugate Gradients Method

```
choose $x_{-1} \in \mathbb{K}^n$
$r_{-1} \leftarrow b - Ax_{-1}$
$\tilde{e}_0 \leftarrow V(V^*AV)^{-1}V^*r_{-1}$
$x_0 \leftarrow x_{-1} + \tilde{e}_0$
$r_0 \leftarrow b - Ax_0$
$p_0 \leftarrow r_0 - V(V^*AV)^{-1}V^*Ar_0$
for $i \leftarrow 0, 1, \ldots, n-1$ do
    $\alpha_i \leftarrow \frac{\langle r_i, r_i \rangle}{\langle Ap_i, p_i \rangle}$
    $x_{i+1} \leftarrow x_i + \alpha_i p_i$
    $r_{i+1} \leftarrow r_i - \alpha_i A p_i$
    $\beta_i \leftarrow \frac{\langle r_{i+1}, r_{i+1} \rangle}{\langle r_i, r_i \rangle}$
    $p_{i+1} \leftarrow r_{i+1} - V(V^*AV)^{-1}V^*Ar_{i+1} + \beta_i p_i$
end for
```

There is another, mathematically equivalent, formulation of the method which lends itself more easily for an analysis. To derive the method, we summarize some important properties in the following lemma.

Lemma 2.1. *Consider the singular linear system*

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

which we call the deflated (linear) system. *We have the following properties:*

1. *The following equalities holds*

$$A(I - \pi_A(V)) = (I - \pi_A(V))^* A = (I - \pi_A(V))^* A (I - \pi_A(V))\,. \tag{2.5}$$

2. *The matrix $A(I - \pi_A(\mathcal{S}))$ is self-adjoint and positive semi-definite.*
3. *The system is consistent, i.e., the right hand side $(I - \pi_A(\mathcal{S}))^* b$ is in the range of $A(I - \pi_A(\mathcal{S}))$. This implies that the system has at least one solution.*
4. *If $\hat{x}$ is a solution of (2.4) then*

$$(I - \pi_A(\mathcal{S}))\hat{x} = (I - \pi_A(\mathcal{S}))x\,,$$

where x is the solution of the linear system $Ax = b$.

Proof. Let the columns of the matrix $V \in \mathbb{K}^{n\times m}$ form an arbitrary basis of $\mathcal{S}$. Using the relation

$$A(I - \pi_A(V)) = A - AV(V^*AV)^{-1}V^*A = (I - \pi_A(V))^* A$$

and the fact that $(I - \pi_A(V))$ is a projection yields

$$A(I - \pi_A(V)) = A(I - \pi_A(V))(I - \pi_A(V)) = (I - \pi_A(V))^* A (I - \pi_A(V))\,.$$

This proves the first assertion and also shows that $A(I - \pi_A(V))$ is self-adjoint. Due to (2.5) we have

$$\langle A(I - \pi_A(V))x, x\rangle = \langle (I - \pi_A(V))x, (I - \pi_A(V))x\rangle_A = \|(I - \pi_A(V))x\|_A^2 \geq 0$$

which proves the second assertion.

Again due to (2.5) and the fact that A has full rank we have

$$\operatorname{range}\Big[A(I-\pi_A(\mathcal{S}))\Big] = \operatorname{range}\Big[(I-\pi_A(\mathcal{S}))^* A\Big] = \operatorname{range}\Big[(I-\pi_A(\mathcal{S}))^*\Big] .$$

Hence the system is consistent which proves the third assertion. Using (2.4) and (2.5) yields

$$A(I-\pi_A(\mathcal{S}))\hat{x} = (I-\pi_A(\mathcal{S}))^* b = (I-\pi_A(\mathcal{S}))^* Ax = A(I-\pi_A(\mathcal{S}))x .$$

Multiplying with A^{-1} from the left we get

$$(I-\pi_A(\mathcal{S}))\hat{x} = (I-\pi_A(\mathcal{S}))x . \qquad \square$$

Since

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

we can compute $\pi_A(\mathcal{S})x$ without explicit knowledge of x. Thus due to Lemma 2.1 if we have a solution $\hat{x}$ for the deflated system (2.4) we obtain the solution for the original system by

$$\begin{aligned} x &= (I-\pi_A(\mathcal{S}))x + \pi_A(\mathcal{S})x \\ &= (I-\pi_A(\mathcal{S}))\hat{x} + V(V^*AV)^{-1}V^*b . \end{aligned}$$

We now want to solve the deflated system (2.4). Since the matrix A is positive semidefinite we can apply the cg method. The lack of regularity is no impediment to the standard cg iteration (cf. [12]) as long as (2.4) is consistent, which was shown to be the case in Lemma 2.1.

Assume that $\hat{x}_0, \hat{x}_1, \ldots$ are the iterates and $\hat{r}_0, \hat{r}_1, \ldots$ are the corresponding residuals of the cg method applied to the deflated linear system (2.4). If Algorithm 1 is initialized with $x_0 = \hat{x}_0$ we have the relation (see [10])

$$x_i = (I-\pi_A(\mathcal{S}))\hat{x}_i + V(V^*AV)^{-1}V^*b \qquad \text{and} \qquad r_i = \hat{r}_i . \tag{2.6}$$

Thus Algorithm 1 is mathematically equivalent to solving (2.4) via the cg method and then computing the approximation for the solution by (2.6).

Hence, for the purpose of an analysis we can think of deflated cg as applying the standard cg algorithm with the matrix $A(I-\pi_A(\mathcal{S}))$ to solve (2.4). In a practical implementation Algorithm 1 is to be preferred, since numerically the action of $(I - \pi_A(\mathcal{S}))$ is not computed exactly and we are thus facing an inconsistent system which may introduce instabilities [12]. There are various other formulations of the deflated cg method that are mathematically equivalent (for an overview see [10]) for which our analysis will hold as well.

Let $\mu_1 \geq \cdots \geq \mu_n \geq 0$ be the eigenvalues of the self-adjoint and positive semidefinite matrix $A(I-\pi_A(\mathcal{S}))$. 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 κ_{eff} the effective condition number of the deflated matrix $A(I-\pi_A(\mathcal{S}))$ to distinguish it from the condition number κ of the original matrix A. Thus in order to analyze the convergence of deflated cg it suffices to estimate the largest and smallest *non-zero* eigenvalue of the matrix $A(I-\pi_A(\mathcal{S}))$.

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 κ_{eff} of the matrix $A(I - \pi_A(\mathcal{S}))$.

3.1. Eigenvalue Bounds.

In order to estimate the speed of convergence of the deflated cg method we need estimates for the largest and smallest non-zero eigenvalue of $A(I - \pi_A(\mathcal{S}))$. The largest eigenvalue of the matrix $A(I - \pi_A(\mathcal{S}))$ is the maximum of the Rayleigh quotient (cf. [23]), i.e.,

$$\mu_1 = \max_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\langle A(I - \pi_A(\mathcal{S}))x, x\rangle}{\langle x, x\rangle}\,.$$

In addition, from the fact that $A\pi_A(\mathcal{S}) = AV(V^*AV)^{-1}V^*A$ is positive semi-definite and thus

$$\langle A(I - \pi_A(\mathcal{S}))x, x\rangle = \langle Ax, x\rangle - \langle A\pi_A(\mathcal{S})x, x\rangle \leq \langle Ax, x\rangle\ ,$$

we obtain

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

This gives a simple upper bound for the largest eigenvalue of $A(I - \pi_A(\mathcal{S}))$.

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. [23].

Lemma 3.1. *Let $M \in \mathbb{K}^{n\times n}$ be self-adjoint, $M = UDU^*$ with $U \in \mathbb{K}^{n\times n}$ unitary and $D = \operatorname{diag}(\mu_1, \ldots, \mu_n)$ with $\mu_1 \geq \ldots \geq \mu_n \geq 0$. Then for $k = 1, \ldots, n$*

$$\mu_k = \min_{\substack{x \in \mathbb{K}^n \setminus \{0\} \\ x \perp \operatorname{span}\{u_{k+1}, \ldots, u_n\}}} \frac{\langle Mx, x\rangle}{\langle x, x\rangle}\,.$$

Proof. Let u_j denote the columns of U which form an orthonormal basis of eigenvectors of M. Let $x \perp \operatorname{span}\{u_{k+1}, \ldots, u_n\}$. Then $x = \sum_{j=1}^k \xi_j u_j$ and we have

$$\langle Mx, x\rangle = \langle UDU^*x, x\rangle = \langle DU^*x, U^*x\rangle = \langle D\xi, \xi\rangle$$

with $\xi = (\xi_1, \ldots, \xi_k, 0, \ldots, 0)^T$. Thus

$$\begin{aligned}\langle D\xi, \xi\rangle &= \sum_{j=1}^k \mu_j |\xi_j|^2 \\ &\geq \mu_k \sum_{j=1}^k |\xi_j|^2 \ =\ \mu_k \langle \xi, \xi\rangle\ ,\end{aligned}$$

which yields

$$\langle Mx, x\rangle \geq \mu_k \langle \xi, \xi\rangle = \mu_k \langle U^*U\xi, \xi\rangle = \mu_k \langle U\xi, U\xi\rangle = \mu_k \langle x, x\rangle\ .$$

The assertion now follows since

$$\langle Mu_k, u_k\rangle = \langle \mu_k u_k, u_k\rangle = \mu_k \langle u_k, u_k\rangle$$

and $u_k \perp \operatorname{span}\{u_{k+1}, \ldots, u_n\}$. □

Lemma 3.1 characterizes the smallest non-zero eigenvalue μ_ℓ of $A(I - \pi_A(\mathcal{S}))$ using $\operatorname{span}\{u_{\ell+1}, \ldots, u_n\}$, the kernel of $A(I - \pi_A(\mathcal{S}))$. Since the matrix A has full rank, the kernel of $A(I - \pi_A(\mathcal{S}))$ is the kernel of $(I - \pi_A(\mathcal{S}))$ which is the deflation subspace $\mathcal{S}$. Thus

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

and due to (2.5)

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

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 μ_ℓ of $A(I - \pi_A(\mathcal{S}))$ we introduce some basic ideas of algebraic multigrid convergence analysis.

Algebraic multigrid methods [5, 16, 20] are based on the assumption that the error of a given iterate can be split into highly oscillatory and slowly varying components of the error. The so-called smoother reduces the highly oscillatory components while the coarse grid correction reduces the slowly varying ones. In many applications the highly oscillatory components are spanned by the eigenvectors corresponding to large eigenvalues while the slowly varying 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]) measures how well the coarse grid correction, defined by the *prolongation operator*, is able to reduce the slowly varying error components. In order to measure the quality of interpolation operators we define the *weak approximation property* as follows.

Definition 3.2. *A subspace $\mathcal{S} \subseteq \mathbb{K}^n$ fulfills the* weak approximation property with constant $K \geq 0$ *if*

$$\operatorname{dist}(\mathcal{S}, x)_2^2 \leq \frac{K}{\|A\|} \|x\|_A^2 \quad \textit{for all } x \in \mathbb{K}^n. \tag{3.3}$$

If the diagonal entries a_{ii} of A fulfill $a_{ii} = 1$ then Definition 3.2 coincides with the definition from [2, 3, 16, 20] 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 multi-level one. Unit diagonal entries may be achieved by using $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ instead of A which would be reflected in a change of the constant K.

Assume that $V \in \mathbb{K}^{n \times m}$ is a multigrid prolongation operator. Then often a restriction operator $R \in \mathbb{K}^{m \times n}$ and a constant $C \in \mathbb{R}$ can be determined, such that

$$\|x - VRx\|_2^2 \leq C \, \|x\|_A^2.$$

is derived where C is independent of the size of the matrix (cf. [2, 3, 16, 20]). The following lemma shows that this implies that the weak approximation property is fulfilled for $K = C\,\|A\|$.

LEMMA 3.3. *Let $V \in \mathbb{K}^{n\times m}$, $R \in \mathbb{K}^{m\times n}$ and assume that*

$$\|x - VRx\|_2^2 \le C\,\|x\|_A^2 \quad \textit{for all } x \in \mathbb{K}^n$$

for a constant $C \ge 0$ holds. Then the subspace $\mathcal{S} = \operatorname{range} V$ fulfills the weak approximation property with constant $K = C\,\|A\|$.

Proof. We have for an arbitrary $x \in \mathbb{K}^n$

$$\begin{aligned}\operatorname{dist}(\mathcal{S}, x)_2^2 &= \min_{y\in\mathcal{S}} \|x - y\|_2^2 = \min_{z\in\mathbb{K}^m} \|x - Vz\|_2^2 \\ &\le \|x - VRx\|_2^2 \le C\,\|x\|_A^2 = \frac{K}{\|A\|}\|x\|_A^2\,. \qquad \square\end{aligned}$$

Strictly speaking, any subspace $\mathcal{S}$ fulfills a weak approximation property, just by choosing K large enough, since

$$\operatorname{dist}(\mathcal{S}, x)_2^2 \le \|x\|_2^2 \le \lambda_n \|x\|_A^2 \quad \text{for all } x \in \mathbb{K}^n\,.$$

The interest of Definition 3.2 is in cases where the subspace $\mathcal{S}$ is such that that K is small and in situations where K is constant for a whole family of matrices A and subspaces $\mathcal{S}$, the family of matrices A representing, e.g., different levels of discretization of a continuous operator.

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

THEOREM 3.4. *Let $\mathcal{S} \subseteq \mathbb{K}^n$ be a subspace such that the weak approximation property* (3.3) *is fulfilled with constant K. Then the effective condition number $\kappa_{\text{eff}} = \frac{\mu_1}{\mu_\ell}$ of the matrix $A(I - \pi_A(\mathcal{S}))$ satisfies*

$$\kappa_{\text{eff}} \le \frac{K}{\xi} \qquad \textit{where} \qquad \xi := \min_{x\in\mathcal{S}^\perp\setminus\{0\}} \frac{\|x - \pi_A(\mathcal{S})x\|_A^2}{\|x\|_A^2} \in (0, 1]\,. \tag{3.4}$$

Proof. Denote by $\pi(\mathcal{S})$ the orthogonal projection onto $\mathcal{S}$ with respect to the 2-inner product. Then

$$\|x - \pi(\mathcal{S})x\|_2^2 = \min_{y\in\mathcal{S}} \|x - y\|_2^2 = \operatorname{dist}(\mathcal{S}, x)_2^2\,. \tag{3.5}$$

For $x \in \mathcal{S}^\perp$ we have $\pi(\mathcal{S})x = 0$ and thus due to (3.5)

$$\|x\|_2^2 = \|x - \pi(\mathcal{S}))x\|_2^2 = \operatorname{dist}(\mathcal{S}, x)_2^2$$

which, using (3.3) gives

$$\|x\|_2^2 \le \frac{K}{\|A\|}\|x\|_A^2 \quad \text{for } x \in \mathcal{S}^\perp\,.$$

By applying this estimate to the denominator in (3.2) we obtain

$$\mu_\ell \ge \frac{\|A\|}{K} \min_{x\in\mathcal{S}^\perp\setminus\{0\}} \frac{\|x - \pi_A(\mathcal{S})x\|_A^2}{\|x\|_A^2} = \frac{\|A\|}{K}\xi\,. \tag{3.6}$$

Hence, by (3.1) and (3.6),

$$\kappa_{\text{eff}} = \frac{\mu_1}{\mu_\ell} \le \frac{\|A\|}{\frac{\|A\|}{K}\xi} = \frac{K}{\xi}\,.$$

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

$$\|x - \pi_A(\mathcal{S})x\|_A^2 = \min_{y \in \mathcal{S}} \|x - y\|_A^2\,.$$

Hence $\|x - \pi_A(\mathcal{S})x\|_A^2 \le \|x\|_A^2$, which proves $\xi \le 1$. □

Now that we have derived a connection between the effective condition number and the weak approximation property used in the analysis of algebraic multigrid methods, it remains to interpret the quantity ξ which is what we do in the next section.

3.3. Strengthened Cauchy-Schwarz Inequality. We first introduce the concept of abstract angles between subspaces which are defined by the *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

$$|\,\langle u,v\rangle\,| \le \gamma \sqrt{\langle u,u\rangle}\sqrt{\langle v,v\rangle} \quad \forall u \in H_1,\ \forall v \in H_2 \tag{3.7}$$

[7, Theorem 2.1]. Equation (3.7) is called *strengthened Cauchy-Schwarz inequality* and γ 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

$$|\,\langle u,v\rangle\,| \le \gamma\,\langle u,u\rangle^{\frac{1}{2}}\,\langle v,v\rangle^{\frac{1}{2}} \le \tfrac{\gamma}{2}\left[\langle u,u\rangle + \langle v,v\rangle\right]\,,$$

since for any two numbers a, b we have $|ab| \le \frac{1}{2}(|a|^2 + |b|^2)$. Hence

$$\begin{aligned}(1-\gamma)\left[\langle u,u\rangle + \langle v,v\rangle\right] &\le \left[\langle u,u\rangle + \langle v,v\rangle\right] - \gamma\left[\langle u,u\rangle + \langle v,v\rangle\right]\\ &\le \left[\langle u,u\rangle + \langle v,v\rangle\right] - 2\,|\,\langle u,v\rangle\,|\\ &\le \left[\langle u,u\rangle + \langle v,v\rangle\right] - 2\,\mathrm{Re}\,\langle u,v\rangle\\ &= \langle u+v, u+v\rangle\ .\end{aligned}$$

Taking the infimum over all $v \in H_2$ yields

$$(1-\gamma)\|u\|^2 \le \inf_{v \in H_2} \|u+v\|^2 \quad \forall u \in H_1\,. \tag{3.8}$$

We now apply this general result in the case where $H_1 = \mathcal{S}^\perp$, $H_2 = \mathcal{S}$ and $\langle\cdot,\cdot\rangle$ is the A-inner product, like in [1] and [8]. The A-orthogonal projection $\pi_A(\mathcal{S})u$ yields the vector in $\mathcal{S}$ which is closest to u in the A-norm. Thus the infimum in (3.8) is obtained for $v = -\pi_A(\mathcal{S})u$ and therefore

$$(1-\gamma)\|u\|_A^2 \le \|u - \pi_A(\mathcal{S})u\|_A^2 \quad \forall u \in \mathcal{S}^\perp\,.$$

This yields a bound for ξ as

$$\xi = \min_{x \in \mathcal{S}^\perp \setminus \{0\}} \frac{\|x - \pi_A(\mathcal{S})x\|_A^2}{\|x\|_A^2} \ge (1-\gamma)\,. \tag{3.9}$$

Using (3.9) we can show the following lemma which states that we can interpret ξ as a measure of approximate A-invariance of the subspace $\mathcal{S}$, i.e., a small value of γ, and thus large value for ξ, indicates that $A\mathcal{S}$ is close to $\mathcal{S}$.

LEMMA 3.5. *If $\mathcal{S}$ is A-invariant, i.e., $A\mathcal{S} = \mathcal{S}$, then $\gamma = 0$ and thus $\xi = 1$.*

Proof. Since the subspace $\mathcal{S}$ is A-invariant, we have $Av \in \mathcal{S}$ and thus

$$\langle u, v \rangle_A = \langle u, Av \rangle = 0 \quad \forall u \in \mathcal{S}^{\perp},\ \forall v \in \mathcal{S}.$$

This gives $\gamma = 0$ and thus $\xi = 1$. □

We can now formulate our main result.

THEOREM 3.6. *Let $\mathcal{S} \subseteq \mathbb{K}^n$, $\mathcal{S} = \mathrm{range}(V), V \in \mathbb{K}^{n\times m}$ and let V fulfill the weak approximation property (3.3) with constant K. Let $\mathcal{S} \subseteq \mathbb{K}^n$ be the subspace spanned by the columns of V. Furthermore let $\gamma \in [0, 1)$ be the smallest constant such that*

$$|\langle u, v \rangle_A| \le \gamma \langle u, u \rangle_A^{\frac{1}{2}} \langle v, v \rangle_A^{\frac{1}{2}} \quad \forall u \in \mathcal{S}^{\perp},\ \forall v \in \mathcal{S}\,. \tag{3.10}$$

Then the effective condition number $\kappa_{\mathrm{eff}} = \frac{\mu_1}{\mu_\ell}$ of the matrix $A(I - \pi_A(V))$ satisfies

$$\kappa_{\mathrm{eff}} \le \frac{K}{(1-\gamma)}\,. \tag{3.11}$$

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 $A(I - \pi_A(\mathcal{S}))$ which depends solely on the weak approximation constant K and the measure ξ on the A-invariance of the deflation subspace $\mathcal{S}$.

4. Applications.

In this section we apply our theory developed so far to the classical case where the deflation subspace $\mathcal{S}$ is spanned by the eigenvectors corresponding to the $(n-k)$ smallest eigenvalues. In addition we consider the case where $\mathcal{S}$ is the range of a prolongation operator from the the classical algebraic multigrid method described in [5, 16, 20].

4.1. The Case of Exact Eigenvalue Deflation.

Let $\mathcal{S}$ be spanned by the eigenvectors corresponding to the $(n-k)$ smallest eigenvalues, e.g., $V = [q_{k+1}|\ldots|q_n]$, where $k \in \mathbb{N}$. It has been shown for this case (cf. [9, Section 1]) that $\kappa_{\mathrm{eff}} = \frac{\lambda_1}{\lambda_k}$. To demonstrate the quality of the bound $\frac{K}{(1-\gamma)}$ from Theorem 3.6 we now show that in this case we actually have $\frac{K}{(1-\gamma)} = \kappa_{\mathrm{eff}}$, i.e., the bound is best possible.

We first consider ξ. Since the subspace $\mathcal{S}$ is A-invariant, we have $\xi = 1$ due to Lemma 3.5.

We now compute the smallest value for K, such that the weak approximation property (3.3) is fulfilled. If $\pi(\mathcal{S})$ is the orthogonal projection onto $\mathcal{S}$ (in the 2-inner product) then

$$\|x - \pi(\mathcal{S})x\|_2^2 = \min_{y\in\mathcal{S}} \|x - y\|_2^2 = \mathrm{dist}(\mathcal{S}, x)_2^2\,.$$

We expand x in terms of the orthonormal eigenvectors q_i of A,

$$x = \sum_{i=1}^{n} \xi_i q_i\,.$$

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

$$\operatorname{dist}(\mathcal{S}, x)_2^2 = \|x - \pi(V)x\|_2^2 = \| \sum_{i=1}^{k} \xi_i q_i + \sum_{i=k+1}^{n} (\xi_i - \xi_i) q_i \|_2^2 = \sum_{i=1}^{k} |\xi|^2 \,.$$

This yields

$$\|x\|_A^2 = \sum_{i=1}^{n} |\xi_i|^2 \lambda_i \geq \sum_{i=1}^{k} |\xi_i|^2 \lambda_i \geq \lambda_k \sum_{i=1}^{k} |\xi_i|^2 = \lambda_k \operatorname{dist}(\mathcal{S}, x)_2^2 \,.$$

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 3.6 is $\frac{K}{(1-\gamma)} = \frac{\lambda_1}{\lambda_k}$, i.e., the bound is equal to the effective condition number κ_{eff} and thus best possible.

4.2. Deflation Subspaces for M-Matrices. The classical algebraic multigrid method [5, 16, 20] was specifically designed for the case that $A \in \mathbb{R}^{n \times n}$ is an M-matrix, i.e., A is symmetric positive definite and $a_{ij} \leq 0$ for $i \neq j$. In this section we show that we can use the prolongation operators constructed in the classical algebraic multigrid method as the operator V which spans the deflation subspace $\mathcal{S}$ and derive *a priori* bounds on the effective condition number following the analysis done in [16, 20].

The construction of prolongation operators is done by inspecting the graph $G(A)$ of a matrix $A \in \mathbb{K}^{n \times n}$ which is given by

$$\begin{aligned} G(A) = (W, E) \quad \text{where} \quad & W = \{1, 2, \ldots, n\} \\ & E = \{(i,j) \in W \times W : a_{ij} \neq 0, i \neq j\} \,. \end{aligned}$$

The neighborhood of a node $i \in W$ is given by

$$N_i := \{j \in W : (i,j) \in E\} \,.$$

To construct the prolongation operator, or equivalently the deflation subspaces, we split the variables W into *coarse* and *fine* variables $\mathcal{C}$ and $\mathcal{F}$, such that $W = \mathcal{C} \,\dot{\cup}\, \mathcal{F}$ and $N_i \cap \mathcal{C} \neq \emptyset$ for $i \in \mathcal{F}$. The coarse variables have a direct representation on the coarse grid, or equivalently the deflation subspace, while the fine variables interpolate from the coarse ones. For simplicity of notation assume that the variables in $\mathcal{C}$ have a smaller index than those in $\mathcal{F}$, i.e., $\mathcal{C} = \{1, 2, \ldots, m\}$, $\mathcal{F} = \{m+1, m+2, \ldots, n\}$. For every fine variable $i \in \mathcal{F}$ choose a set of variables $P_i \subseteq N_i \cap \mathcal{C}$. The value for the variable i is then interpolated from the variables in P_i. Defining the interpolation weights

$$w_{ik} = \alpha_i \frac{-a_{ik}}{a_{ii}} \quad \text{with} \quad \alpha_i = \frac{\sum_{k \in N_i} a_{ik}}{\sum_{k \in P_i} a_{ik}} \quad \text{for} \quad i \in \mathcal{F}$$

yields the prolongation operator $V \in \mathbb{R}^{n \times m}$ by

$$\left(V e^c\right)_i = \begin{cases} e_i^c & \text{for } i \in \mathcal{C} \\ \sum_{k \in P_i} w_{ik} e_k^c & \text{for } i \in \mathcal{F} \end{cases} . \tag{4.1}$$

Under the assumption that the $\mathcal{C}/\mathcal{F}$-splitting is reasonably well chosen the classical multigrid theory yields an estimate for the constant K of the weak approximation property (3.3).

Theorem 4.1. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric weakly diagonally dominant M-matrix, i.e., $a_{ij} \le 0$ for $i \ne j$ and $\sum_j a_{ij} \ge 0$ for all i. If for fixed $\tau \ge 1$ a $\mathcal{C}/\mathcal{F}$-splitting exists such that for each $i \in \mathcal{F}$ there is a $P_i \in \mathcal{C} \cap N_i$ with*

$$\sum_{k\in P_i} |a_{ik}| \ge \frac{1}{\tau} \sum_{j\in N_i} |a_{ij}| \quad \text{for } i \in \mathcal{F} \tag{4.2}$$

then there exists a matrix $R \in \mathbb{R}^{m\times n}$ such that the operator V from (4.1) fulfills

$$\|e - VRe\|_D^2 \le \tau \|e\|_A^2$$

where $D = \operatorname{diag} A$ is the diagonal matrix containing the diagonal entries a_{ii} of A.

Proof. Define the restriction $R \in \mathbb{R}^{m\times n}$ as

$$R = \begin{bmatrix} I_m \\ 0 \end{bmatrix}$$

where $I_m \in \mathbb{R}^{m\times m}$ is the identity matrix. Thus R maps a coarse variables to itself and a fine variables to zero. Then the result follows from [20, Theorem A.4.3]. □

Corollary 4.2. *Under the same assumptions as in Theorem 4.1 we have that V fulfills the weak approximation property (3.3) with*

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

Proof. Directly follows from Lemma 3.3, the fact that $(\min_i a_{ii}) \|x\|_2^2 \le \|x\|_D^2$ for $x \in \mathbb{R}^n$ and Theorem 4.1. □

Example 4.3. *Let $N \in \mathbb{N}$ be odd. Consider the discrete 9-point Laplacian, i.e., the block tridiagonal matrix*

$$A = \begin{bmatrix} B & C & & \\ C & B & \ddots & \\ & \ddots & \ddots & C \\ & & C & B \end{bmatrix} \in \mathbb{R}^{N^2\times N^2}$$

with $B, C \in \mathbb{R}^{N\times N}$,

$$B = \begin{bmatrix} 8 & -1 & & \\ -1 & 8 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 8 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} -1 & -1 & & \\ -1 & -1 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & -1 \end{bmatrix}$$

The graph $G(A)$ of A is a regular $N \times N$ grid with added diagonal connections, see figure 4.1. We set $\mathcal{C}$ as the variables in odd rows and columns ($\square$), $\mathcal{F}$ as the remaining variables ($\diamond$, $\bullet$ and $\star$).

The requirement (4.2) is equivalent to $\tau \ge \alpha_i$ for $i \in \mathcal{F}$. A straight forward computation yields $\alpha_i = 2$ for $i \in \bullet$, $\alpha_i = 4$ for $i \in \diamond$ and $\alpha_i = \frac{5}{2}$ for $i \in \star$. Hence $\tau = 4$ fulfills (4.2). Due to Gershgorin's theorem [23] the eigenvalues λ_i of A fulfill $\lambda_i \in [0, 16]$ and thus $\|A\| \le 16$ and due to (4.2) the weak approximation property is fulfilled for $K = 8$.

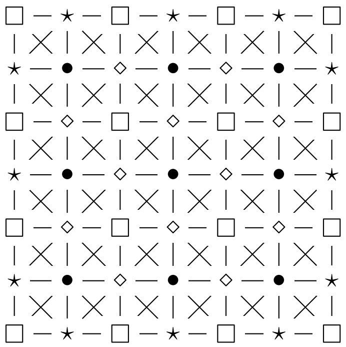

FIGURE 4.1. *The graph of A split into $\mathcal{C}$ and $\mathcal{F}$.*

p	Iterations	Residual	Error
4	8	$5.64429 \cdot 10^{-7}$	$1.53486 \cdot 10^{-7}$
5	8	$8.4304 \quad \cdot 10^{-7}$	$4.64197 \cdot 10^{-7}$
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 $N = 2^p - 1$.

5. Numerical Experiments. To carry out our numerical experiments we consider the matrix V from (4.1) with A, $\mathcal{C}$ and $\mathcal{F}$ from example 4.3. Thus the deflation subspace $\mathcal{S}$ is range V.

5.1. Independence of the Grid Size. In Example 4.3 we have seen that the constant K is independent of the grid size N. Thus we expect the method to converge in a constant number of iterations if the abstract angle γ is independent of N. Hence we measure the number of iterations for different sizes $N = 2^p - 1$ of the linear system.

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 is listed in table 5.1 where we observe that the number of iterations stays constant.

5.2. Numerical Computation of the Constants. In this subsection we want to verify our theory by numerically computing the condition numbers κ, κ_{eff} the constants K and γ and the estimate for the effective condition number $\frac{K}{1-\gamma}$ for a small linear system. This computation is only possible for small linear systems since it involves the computation of eigenvectors of dense matrices which are of the same size as the linear system.

We are interested in the smallest K such that the weak approximation property with constant K holds. For small linear systems such K can be computed numerically as follows: From Definition 3.2 it follows that K is given by

$$K = \|A\| \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\operatorname{dist}(\mathcal{S}, x)_2^2}{\|x\|_A^2} \,.$$

Let the columns of $W \in \mathbb{K}^{n \times m}$ form an orthonormal basis of $\mathcal{S}^\perp$. Then the orthogonal

	A	$A(I-\pi_A(V))$	
λ_1	0.0577	6.0708	μ_1
λ_n	11.9616	11.9586	μ_ℓ
κ	207.3403	1.9698	κ_{eff}
		1.9703	K
		0.3369	γ
		2.9715	$\frac{K}{1-\gamma}$

TABLE 5.2
Condition number of the linear systems for $N = 2^5 - 1$.

projection $(I - \pi(\mathcal{S}))$ onto $\mathcal{S}^\perp$ fulfills $(I - \pi(\mathcal{S})) = WW^*$ and thus due to (3.5)

$$K = \|A\| \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\|(I - \pi(\mathcal{S}))x\|_2^2}{\|x\|_A^2} = \|A\| \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\|WW^*x\|_2^2}{\|x\|_A^2} \,. \tag{5.1}$$

Using the eigendecomposition $A = Q\Lambda Q^*$ yields

$$\|x\|_A^2 = \langle Ax, x\rangle = \langle Q\Lambda Q^* x, x\rangle = \left\langle \Lambda^{\frac{1}{2}} Q^* x, \Lambda^{\frac{1}{2}} Q^* x \right\rangle = \|\Lambda^{\frac{1}{2}} Q^* x\|_2^2 \,. \tag{5.2}$$

Since we are interested in the supremum over all vectors in $\mathbb{K}^n \setminus \{0\}$ we can substitute x by $Q\Lambda^{-\frac{1}{2}} z$ in (5.1). This yields due to (5.2) that

$$K = \|A\| \sup_{z \in \mathbb{K}^n \setminus \{0\}} \frac{\|WW^* Q \Lambda^{-\frac{1}{2}} z\|_2^2}{\|z\|_2^2} = \|A\| \|WW^* Q \Lambda^{-\frac{1}{2}}\|^2 \,.$$

The matrix norms on the right hand side can numerically be computed via the singular value decomposition. The constant γ from the strengthened Cauchy-Schwarz inequality is 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 for the condition number κ_{eff} of the deflated matrix in this example.

6. Influence of the Accuracy of Computations. The deflated cg method involves the solution of the "inner" linear system

$$(V^*AV)z_{i+1} = V^*Ar_{i+1} \tag{6.1}$$

in every step of the 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 (we had $m = n/4$ in the numerical experiments of Section 5), so that solving (6.1) will be done using an "inner" iteration. Its accuracy will be decisive for the convergence process of the overall iteration. This can be motivated explained as follows.

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

$$A\underbrace{(I - V(V^*AV)^{-1}V^*A)}_{=(I-\pi_A(\mathcal{S}))}\hat{x} = (I - AV(V^*AV)^{-1}V^*)b \,.$$

In here, the kernel of $(I - \pi_A(\mathcal{S}))$ is $\mathcal{S}$, the column range of V. Let us now assume that we are given an approximation $M \in \mathbb{K}^{m \times m}$ 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 thus loose the property of having a considerable number of zero eigenvalues, an essential ingredient to estimate the effective condition number κ_{eff} in Theorem 3.6. Even worse, the eigenvalues that would be mapped to zero by the exact A-orthogonal projection now remain as small non-zero eigenvalues making the condition number of the approximately deflated system potentially larger than that of the original system.

Initially we split the error into $e_{-1} = e_0 + \tilde{e}_0$, where $e_0 \in \mathcal{S}^{\perp_A}$ and $\tilde{e}_0 \in \mathcal{S}$. If this is done inexactly, e_0 will not be in $\mathcal{S}^{\perp_A}$, but its A-orthogonal projection on $\mathcal{S}$ will be small, i.e., we have

$$e_0 = e_0^{\text{ex}} + e_0^{\text{err}} \quad \text{with} \quad e_0^{\text{ex}} \in \mathcal{S}^{\perp_A},\, e_0^{\text{err}} \in \mathcal{S} \quad \text{and} \quad \|e_0^{\text{err}}\| \leq \rho\,,$$

for some tolerance $\tau \geq 0$. The eigenvectors corresponding to zero eigenvalues of the exactly deflated matrix $A(I - \pi_A(\mathcal{S}))$ span the subspace $\mathcal{S}$. Thus, in the case where we work with the matrix $A(I - VMV^*A)$, its eigenvectors u_ℓ corresponding to very small eigenvalues μ_ℓ will be close to $\mathcal{S}$, i.e.,

$$u_\ell = u_\ell^{\text{ex}} + u_\ell^{\text{err}} \quad \text{with} \quad u_\ell^{\text{ex}} \in \mathcal{S},\, u_\ell^{\text{err}} \in \mathcal{S}^{\perp_A} \quad \text{and} \quad \|u_\ell^{\text{err}}\| \leq \rho\,.$$

Hence those components in the expansion of e_0 in the basis of eigenvectors of $A(I - VMV^*A)$ corresponding to small eigenvalues are very small. Together with (2.3) and the subsequent discussion this explains that the cg method does not "see" these error components as long as they are substantially smaller than the other error components, resulting in an initial phase of fast convergence. However, when the norm of the current error approaches ρ, the error components belonging to the small eigenvalues will not be negligible anymore, and the cg iteration slows down dramatically.

The question remains how to determine a suitable stopping criterion for the inner iteration based on the stopping criterion of the outer iteration

$$\|r_i\| \leq \tau\,\|b\| =: \varepsilon \tag{6.2}$$

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

$$\|r_i\| \leq \tau\,\|b\| =: \varepsilon\,,$$

to achieve (6.2)? A first strategy may be to set $\tau^{\text{c}} = \varepsilon$ 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 $p \in \mathbb{K}^n$ in the outer cg iteration, we can use the results from [19, 21]. There it is suggested to use

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

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. In this paper we derived a convergence estimate for the deflated cg method based on algebraic multigrid theory. We have shown that our theory recovers the exact result for the convergence estimate for eigenvector deflation. 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 pave the way 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 the 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 suitable 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. *Applied Mathematics and Computation*, 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(5):1570–1592, 2001.

[4] M. Brezina, R. D. Falgout, S. MacLachlan, T. A. Manteuffel, S. F. McCormick, and J. W. Ruge. Adaptive smoothed aggregation (αSA) multigrid. *SIAM Rev.*, 47(2):317–346, 2005.

[5] W. L. Briggs, V. E. Henson, and S. F. McCormick. *A Multigrid Tutorial.* SIAM, 2 edition, 2000.

[6] Z. Dostál. Conjugate gradient method with preconditioning by projector. *Int. J. Comput. Math.*, 23(3-4):315–323, 1988.

[7] V. Eijkhout and P. S. Vassilevski. The role of the strengthened Cauchy-Buniakowskiĭ-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. H. Gutknecht, J. Liesen, and R. Nabben. A framework for deflated and augmented Krylov subspace methods. *submitted*, 2012. arXiv:1206.1506.

[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. Sci. Comput.*, 23(6):2008–2040 (electronic), 2002.

[14] M. Lüscher. Local coherence and deflation of the low quark modes in lattice QCD. *Journal of High Energy Physics*, 2007(07):081, 2007.

[15] R. A. Nicolaides. Deflation of conjugate gradients with applications to boundary value problems. *SIAM J. Numer. Anal.*, 24(2):355–365, 1987.

[16] J. W. Ruge and K. Stüben. Algebraic multigrid. In *Multigrid methods*, volume 3 of *Frontiers Appl. Math.*, pages 73–130. SIAM, Philadelphia, PA, 1987.

[17] Y. Saad. *Iterative Methods for Sparse Linear Systems.* Society for Industrial Mathematics, 2 edition, 2003.

[18] Y. Saad, M. Yeung, J. Erhel, and F. Guyomarc'h. A deflated version of the conjugate gradient algorithm. *SIAM J. Sci. Comput.*, 21(5):1909–1926, 2000.

[19] V. Simoncini and D. B. Szyld. Theory of inexact Krylov subspace methods and applications

to scientific computing. *SIAM J. Sci. Comput.*, 25:454–477, 2003.

[20] K. Stüben. An introduction to algebraic multigrid. In U. Trottenberg, C. W. Oosterlee, and A. Schüller, editors, *Multigrid*. Academic Press, 2001.

[21] 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.

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

[23] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford University Press, 1967.