

On the sensitivity of some spectral preconditioners

L. Giraud* S. Gratton*

CERFACS Technical report TR/PA/04/108
October 2004

Abstract

It is well known that the convergence of the conjugate gradient method for solving symmetric positive definite linear systems depends to a large extent on the eigenvalue distribution. In many cases, it is observed that “removing” the extreme eigenvalues can greatly improve the convergence. Several preconditioning techniques based on approximate eigenelements have been proposed in the past few years that attempt to tackle this problem. The proposed approaches can be split into two main families depending on whether the extreme eigenvalues are moved exactly to one or are shift to close to one. The first technique is often referred to the deflating approach, while the latter is referred to as coarse grid preconditioner by analogy to techniques first used in domain decomposition methods. Many variants exist in the two families that reduce to the same preconditioners if the exact eigenelements are used. In this paper we investigate the behaviour of some of these techniques when the eigenelements are only known approximately. We use the perturbation theory for eigenvalues and eigenvectors to investigate the behaviour of the spectrum of the preconditioned systems using first order approximation. We illustrate the sharpness of the first order approximation and show the effect of the inexactness of the eigenelements on the behaviour of the resulting preconditioner when applied to accelerate the conjugate gradient method.

1 Introduction

In many problems the convergence of Krylov solvers can be significantly slowed down by the presence of small eigenvalues in the spectrum of the matrices involved in the solution of the linear systems. This occurs for instance when the Conjugate Gradient (CG) method is implemented to solve linear systems arising from the discretization of second-order elliptic problems. For symmetric positive definite (SPD) linear systems it is well-admitted that the convergence of CG to solve $Ax = b$ depends to a large extend on the eigenvalue distribution

*CERFACS, 42 av. Gaspard Coriolis, 31057 Toulouse cedex 1, France.

of the coefficient matrix A . This can be illustrated by the bound on the rate of convergence of the CG method given by [10] viz.

$$\|x_k - x^*\|_A \leq 2\|x_0 - x^*\|_A \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k, \quad (1)$$

where $\kappa = \lambda_{\max}(A)/\lambda_{\min}(A)$ is the spectral condition number of A , the A -norm of x is $\|x\|_A = \sqrt{x^T A x}$ and the exact solution is $x^* = A^{-1}b$. This analysis leads to the idea of improving the convergence of CG by using a preconditioner M such that the ratio $\lambda_{\max}(MA)/\lambda_{\min}(MA)$ is less than κ . In this paper, we are interested in a class of two-level preconditioners that exploit some information on the eigenpairs of A . The underlying driving idea of these approaches is to capture in a low dimensional space the modes that do not quickly converge with a first level preconditioner. In order to be efficient and keep the dimension of the low dimensional space reasonably small, these techniques are generally used in combination with a first level preconditioner that does a good job of clustering most eigenvalues near to one with relatively few outliers near the origin [3, 4, 6, 13, 18]. These two-level preconditioners can be split into two main families depending on their effect on the spectrum. They are referred to as deflating preconditioners [4, 8, 9, 14] if they attempt to move to a positive quantity σ the subset of eigenvalues or referred to as coarse grid preconditioners [3, 9] if they only attempt to shift the subset close to σ . The name of those latter techniques comes from domain decomposition and was first introduced in [2]. For this reason $\sigma = 1$ is often considered in practice. An impressive example of the efficiency of a spectral preconditioner is provided by the atmosphere data assimilation area [8]. In this application, nonlinear least-squares problems with more than 10^7 unknowns are daily solved using a Gauss-Newton approach. The sequence of linear least-squares involved in the nonlinear solution scheme are solved by CG. The correspondence between CG and Lanczos is exploited on each linear problem to extract approximate spectral information. This spectral information is used to design a deflating spectral preconditioner for the subsequent linear least-squares problem.

When the exact eigenvectors are used, many of these spectral preconditioners reduce to the same expressions. The aim of this paper is to use the perturbation theory for eigenvalues and eigenvectors to investigate the behaviour of some of these preconditioners when they are constructed using approximate eigenelements. The paper is organized as follows. After the introduction of some notations, Section 3 is devoted to the sensitivity analysis. We first establish in Section 3.2 the theoretical results and illustrate in Section 3.4 the sharpness of the expressions on a set of test matrices. Section 3.3 is devoted to the derivation of the condition number associated with the eigenvalues of the preconditioned matrix. It is also indicated how this information can help screen the eigen-element information to obtain a nice clustering around σ for the spectrum of the preconditioned matrix. In Section 4, we illustrate the behaviour of the preconditioned CG (PCG) for the solution of linear systems associated with matrices from the Harwel-Boeing collection [7] when the spectral information is

approximated. Finally, we conclude with some remarks in Section 5.

2 Spectral preconditioner variants

We first consider one representative of the deflating preconditioners and one of the coarse grid preconditioners. Let $V = [v_1, \dots, v_n] \in \mathbb{R}^{n \times n}$ be an eigen-basis of A and $\{\lambda_i\}_{i=1, \dots, n}$ be the set of corresponding eigenvalues sorted by increasing magnitude. In order to move $\{\lambda_i\}_{i=1, \dots, k}$ to σ , we define the following deflating preconditioner:

$$M^{def} = I + \sum_{i=1}^k \left(\frac{\sigma}{\lambda_i} - 1 \right) \frac{v_i^T v_i}{\|v_i\|^2}. \quad (2)$$

The columns of V also form an eigen-basis of $M^{def}A$. For $V_k = [v_1, \dots, v_k]$, this preconditioner is such that $M^{def}AV_k = \sigma V_k$ and $M^{def}Aw = Aw$ if $V_k^T w = 0$, which shows that M^{def} moves the eigenvalues $\{\lambda_i\}_{i=1, \dots, k}$ to σ and leaves the rest of the spectrum unchanged. This technique is expected to be especially efficient in the case where $\{\lambda_i\}_{i=k+1, \dots, n}$ are already in the neighbourhood of σ , in which case $\lambda_{\max}(MA)/\lambda_{\min}(MA)$ becomes close to one. Because (1) only provides an upper-bound for the convergence rate it might be argued that, if $\{\lambda_i\}_{i=k+1, \dots, n}$ are already close to σ , shifting $\{\lambda_i\}_{i=1, \dots, k}$ to any quantity close to σ (and not necessarily to σ exactly) does still improve the convergence of CG. To this end, if $\{\lambda_i\}_{i=1, \dots, k}$ are small we can use the coarse grid preconditioner

$$M^{coarse} = I + \sigma V_k (\text{diag}(\lambda_i))^{-1} V_k^T, \quad (3)$$

where $\text{diag}(\lambda_i)$ denotes the diagonal matrix with entries λ_i . The columns of V also form an eigen-basis of $M^{coarse}A$. This preconditioner is such that $M^{coarse}Av_i = (\sigma + \lambda_i)v_i$ and $M^{coarse}Aw = Aw$ if $V_k^T w = 0$. That is, the eigenvalues $\{\lambda_i\}_{i=1, \dots, k}$ are shifted to $\sigma + \lambda_i$, while the rest of the spectrum is unchanged. This latter technique is particularly suited when $\{\lambda_i\}_{i=1, \dots, k}$ are small.

In the previous paragraph the spectral transformations associated with M^{def} and M^{coarse} rely on the facts that (λ_i, v_i) are exact eigenpairs of a SPD matrix. This implies two intensively used properties that are $Av_i = \lambda_i v_i$ and $v_i^T v_j = 0$ if $i \neq j$. We assume now that we only have access to approximate spectral information, and denote by $(\tilde{\lambda}_i, \tilde{v}_i)$ the inexact eigenpairs such that the two latter properties do not necessarily hold. We only suppose that $\tilde{V}_k = [\tilde{v}_1, \dots, \tilde{v}_k]$ is such that $\tilde{V}_k^T \tilde{V}_k$ and $\tilde{V}_k^T A \tilde{V}_k$ are nonsingular. We can write the two above preconditioners in a form that does a weaker usage of the properties of the exact eigenpairs. The inexact “deflating” preconditioner then writes

$$M_1 = I + \tilde{V}_k \left(\sigma \left(\tilde{V}_k^T A \tilde{V}_k \right)^{-1} - \left(\tilde{V}_k^T \tilde{V}_k \right)^{-1} \right) \tilde{V}_k^T. \quad (4)$$

Noticing that $V_k^T AV_k$ is diagonal, we can be tempted to approximate it by the

diagonal of the Rayleigh quotients. This gives

$$M_1^{ral} = I + \tilde{V}_k \left(\sigma \left(\text{diag}(\tilde{v}_i^T A \tilde{v}_i) \right)^{-1} - \left(\tilde{V}_k^T \tilde{V}_k \right)^{-1} \right) \tilde{V}_k^T. \quad (5)$$

Furthermore, because the diagonal elements of M_1 are simply the eigenvalues, we can also use the following preconditioner formulation

$$M_1^{eig} = I + \tilde{V}_k \left(\sigma \left(\text{diag}(\tilde{\lambda}_i) \tilde{v}_i^T \tilde{v}_i \right)^{-1} - \left(\tilde{V}_k^T \tilde{V}_k \right)^{-1} \right) \tilde{V}_k^T. \quad (6)$$

Using the orthogonality property of the eigenvectors M_1 reduces to

$$M_2^{ral} = I + \sum_{i=1}^k \left(\sigma \left(\frac{\tilde{v}_i^T A \tilde{v}_i}{\tilde{v}_i^T \tilde{v}_i} \right)^{-1} - 1 \right) \frac{\tilde{v}_i \tilde{v}_i^T}{\tilde{v}_i^T \tilde{v}_i}, \quad (7)$$

that could also lead to consider

$$M_2^{eig} = I + \sum_{i=1}^k \left(\frac{\sigma}{\tilde{\lambda}_i} - 1 \right) \frac{\tilde{v}_i \tilde{v}_i^T}{\tilde{v}_i^T \tilde{v}_i}. \quad (8)$$

Similarly, for the coarse grid preconditioners we consider the following variants

$$M_3 = I + \sigma \tilde{V}_k \left(\tilde{V}_k^T A \tilde{V}_k \right)^{-1} \tilde{V}_k^T, \quad (9)$$

$$M_3^{ral} = I + \sigma \tilde{V}_k \left(\text{diag}(\tilde{v}_i^T A \tilde{v}_i) \right)^{-1} \tilde{V}_k^T, \quad (10)$$

$$M_3^{eig} = I + \sigma \tilde{V}_k \left(\text{diag}(\tilde{\lambda}_i \tilde{v}_i^T \tilde{v}_i) \right)^{-1} \tilde{V}_k^T. \quad (11)$$

To study the performance of the above preconditioners in the presence of inexact spectral information, we assume that the spectral information is not related to A but to a nearby matrix $A + tE$, where t is a real parameter and $\|E\| = 1$. Let denote $\lambda_i(t)$ and $v_i(t)$ the eigenvalues and eigenvectors of $A + tE$. If A has only simple eigenvalues, it can be shown [16] that the eigenvalues of $A + tE$ are differentiable functions of t in a neighbourhood \mathcal{V} of $t = 0$. If the eigenvectors are normalized using $v_i(t)^T v_i = 1$ the eigenvectors are also differentiable functions of t in a neighbourhood of $t = 0$. Note that none of the preconditioners assume that the eigenvectors have unit length. Indeed, the preconditioners are invariant by any nonzero scaling of the eigenvectors. Therefore the normalization $v_i(t)^T v_i = 1$ can be assumed for the analysis without loss of generality.

Definition 1 We denote by $M_1(t)$, $M_1^{ral}(t)$, $M_1^{eig}(t)$, $M_2^{ral}(t)$, $M_2^{eig}(t)$, $M_3(t)$, $M_3^{ral}(t)$ and $M_3^{eig}(t)$ the preconditioners obtained by setting $\tilde{v}_i = v_i(t)$ and $\tilde{\lambda}_i = \lambda(t)$ in the definitions (4) to (11).

In this paper, we carry out a first order analysis which shows the asymptotic sensitivity of the eigenvalues of the preconditioned matrix for small enough values of the parameter t . The approach is strongly related to the backward error framework since the inaccuracies in v_i and λ_i are modeled by a perturbation of the coefficient matrix A .

3 Sensitivity analysis

3.1 Notation

We introduce some notations and basic results used throughout this paper. For any square matrix $X \in \mathbb{R}^{n \times n}$, let \underline{X}_i denote the $n \times (n-1)$ matrix whose columns are those of X excepted for the i^{th} , that is $\underline{X}_i = [X(:,j)]_{j=1,\dots,n;j \neq i}$. For a $n \times n$ matrix X , $\{\lambda_1(X), \dots, \lambda_n(X)\}$ are the eigenvalues of X where multiple eigenvalues are repeated. We also assume that $|\lambda_1(X)| \leq \dots \leq |\lambda_n(X)|$. The i th eigenvalue of A is denoted by λ_i when there is no possible confusion. Let $A \in \mathbb{R}^{n \times n}$ be a SPD matrix where

$$AV = VD \text{ with } V^T V = I \text{ and } D = \text{diag}(\lambda_i)_{i=1,\dots,n}$$

denotes its spectral decomposition. We assume that all the eigenvalues of A are simple.

For a vector x , $\|x\|^2 = \sum_{i=1}^n x_i^2$ is the Euclidean vector norm, and $\|A\| = \max_{\|x\|=1} \|Ax\|$ is the spectral norm of the matrix A . The operator \circ denotes the Hadamard product: $A \circ B = [a_{ij}b_{ij}] \in \mathbb{C}^{m \times n}$, for A and $B \in \mathbb{C}^{m \times n}$. The spectral norm is submultiplicative with respect to the Hadamard product (see [1, p. 332]): $\|A \circ B\| \leq \|A\| \|B\|$.

Let \tilde{V} and \tilde{D} be defined by $\tilde{V} = [v_1(t), \dots, v_n(t)]$ and $\tilde{D} = \text{diag}(\tilde{\lambda}_i) = \text{diag}(\lambda_i(t))$. For sufficiently small, $t \in \mathcal{V}$, we have

$$(A + tE)\tilde{V} = \tilde{V}\tilde{D}.$$

Note that $\tilde{V}(0) = V$. Using the first order expansion of the eigenvalues and eigenvectors in the direction E [16], we can also write $\tilde{V} = V + \delta V(t) + o(t)$ where the i^{th} column of $\delta V(t)$ is defined by

$$\delta v_i(t) = t \underline{V}_i (\lambda_i I - B_i)^{-1} \underline{V}_i^T E v_i,$$

and $\lambda_i(t) = \lambda_i + \delta \lambda_i(t) + o(t)$, where $\delta \lambda_i(t) = t v_i^T E v_i$ and the $(n-1) \times (n-1)$ diagonal matrix $B_i = \text{diag}(\lambda_j)_{j=1,\dots,n;j \neq i}$. The first order expansion of the eigenvalues of the preconditioned matrices will be expressed in terms of the following $k \times k$ matrices W and Y defined by their (ℓ, s) -entry :

$$W_{\ell,\ell} = 0; W_{\ell,s} = \frac{\sigma}{\lambda_s - \lambda_\ell} \sqrt{\frac{\lambda_\ell}{\lambda_s}} \text{ for } \ell \neq s, \quad (12)$$

$$Y_{\ell,\ell} = 0; Y_{\ell,s} = \frac{\sigma - \lambda_s}{\lambda_s - \lambda_\ell} \sqrt{\frac{\lambda_\ell}{\lambda_s}} \text{ for } \ell \neq s, \quad (13)$$

Similarly we also introduce the $k \times k$ matrix $\Delta = V_k^T E V_k$ and the diagonal matrix $J = \text{diag} \left(-\frac{\sigma}{\lambda_\ell} \right)_{\ell=1, \dots, k}$.

3.2 First order expansion of the preconditioned matrix eigenvalues

Theorem 1 *The deflating preconditioner $M_1(t)$ is such that the eigenvalues of the preconditioned matrix $M_1(t)A$ are*

$$\begin{cases} \mu_i(t) = \sigma + o(t) & \text{if } i \leq k, \\ \mu_i(t) = \lambda_i(A) + o(t) & \text{if } i > k. \end{cases}$$

Proof of Theorem 1: Let us first mention that the eigenvalues of $M_1(t)A$ are also those of $AM_1(t)$ and then write

$$AM_1(t) = A + \sigma A \tilde{V}_k \left(\tilde{V}_k^T A \tilde{V}_k \right)^{-1} \tilde{V}_k - A \tilde{V}_k \left(\tilde{V}_k^T \tilde{V}_k \right)^{-1} \tilde{V}_k^T.$$

The entry (ℓ, s) of $(\tilde{V}^T \tilde{V})$ is defined by

$$\begin{aligned} \left(\tilde{V}_k^T \tilde{V}_k \right)_{\ell, s} &= (v_\ell^T + \delta v_\ell^T)(v_s + \delta v_s) \\ &= v_\ell^T v_s + v_\ell^T \delta v_s + \delta v_\ell^T v_s + o(t). \end{aligned}$$

Using (29), (30) and the facts that $v_\ell^T v_\ell = 1$, $\underline{V}_\ell^T v_\ell = 0$ and $v_\ell^T \underline{V}_\ell = 0$ we obtain

$$(\tilde{V}_k^T \tilde{V}_k) = I + C + o(t) \quad (14)$$

with $c_{\ell, \ell} = 0$ and

$$c_{\ell, s} = (\lambda_\ell - \lambda_s)^{-1} (v_\ell^T E^T v_s - v_\ell^T E v_s) t \text{ for } \ell \neq s. \quad (15)$$

Consequently

$$\begin{aligned} \left(\tilde{V}_k^T \tilde{V}_k \right)^{-1} &= I - (V_k^T V_k)^{-1} C (V_k^T V_k)^{-1} + o(t) \\ &= I - C + o(t) \end{aligned}$$

and

$$\begin{aligned} A \tilde{V}_k \left(\tilde{V}_k^T \tilde{V}_k \right)^{-1} \tilde{V}_k^T &= A V_k V_k^T + A \delta V_k V_k^T - A V_k C V_k^T \\ &\quad + A V_k \delta V_k^T + o(t) \\ &= A V_k V_k^T + A (\delta V_k V_k^T \\ &\quad + (\delta V_k V_k^T)^T - V_k C V_k^T) + o(t) \end{aligned} \quad (16)$$

Furthermore

$$\begin{aligned}
\tilde{V}_k^T A \tilde{V}_k &= V_k^T A V_k + (\delta V_k^T A V_k) + (\delta V_k^T A V_k)^T + o(t) \\
&= V_k^T A V_k + (\delta V_k^T V_k D_k) + (\delta V_k^T V_k D_k)^T + o(t) \\
&= D_k + G + o(t)
\end{aligned} \tag{17}$$

where

$$g_{\ell,s} = \delta v_\ell^T v_s \lambda_s + \lambda_\ell v_\ell^T \delta v_s$$

and $D_k = \text{diag}(\lambda_i)_{i=1,\dots,k}$. Using (29), (30) and (31) we obtain $g_{\ell,\ell} = 0$ and

$$g_{\ell,s} = (\lambda_\ell - \lambda_s)^{-1} (\lambda_s v_\ell^T E^T v_s - \lambda_\ell v_\ell^T E v_s) t \text{ for } \ell \neq s. \tag{18}$$

Consequently we have

$$\begin{aligned}
(\tilde{V}_k^T A \tilde{V}_k)^{-1} &= (V_k^T A V_k)^{-1} - (V_k^T A V_k)^{-1} G (V_k^T A V_k)^{-1} + o(t) \\
&= D_k^{-1} - D_k^{-1} G D_k^{-1} + o(t).
\end{aligned} \tag{19}$$

and

$$\begin{aligned}
\tilde{V} \left(\tilde{V}_k^T A \tilde{V}_k \right)^{-1} \tilde{V}^T &= V_k D_k^{-1} V_k^T + (\delta V_k D_k^{-1} V_k^T) + (\delta V_k D_k^{-1} V_k^T)^T \\
&\quad - V_k D_k^{-1} G D_k^{-1} V_k^T + o(t).
\end{aligned} \tag{20}$$

Finally combining (16) and (20) we obtain

$$\begin{aligned}
AM_1(t) &= A + A \tilde{V}_k \left((\tilde{V}_k A \tilde{V}_k)^{-1} - (\tilde{V}_k^T \tilde{V}_k)^{-1} \right) \tilde{V}_k \\
&= AM_1(0) + \delta M_1 + o(t)
\end{aligned}$$

with

$$\begin{aligned}
\delta M_1 &= A (\delta V_k (I + \sigma D_k^{-1}) V_k^T + (\delta V_k (I + \sigma D_k^{-1}) V_k^T)^T \\
&\quad - V_k (C + \sigma D_k^{-1} G D_k^{-1}) V_k^T).
\end{aligned}$$

We remind that

$$AM_1(0) = V \text{diag}(\underbrace{\sigma, \dots, \sigma}_k, \lambda_{k+1}, \dots, \lambda_n) V^T.$$

- For $j > k$, the first order approximation of the simple eigenvalues writes

$$\mu_j = \lambda_j + v_j^T \delta M_1 v_j + o(t) = \lambda_j + o(t),$$

since $\delta M_1 v_j = 0$ because (31) and the orthogonality of V .

- For $j \leq k$, the first order approximation of the semi-simple multiple eigenvalue σ writes (see [12, p. 402] or [17])

$$\mu_j = \sigma + \lambda_j (V_k^T \delta M_1 V_k) + o(t),$$

where $(V_k^T \delta M_1 V_k)$ is a $k \times k$ matrix with diagonal entries equal to zero and off-diagonal entries

$$\begin{aligned}
(V_k^T \delta M_1 V_k)_{\ell,s} &= v_\ell^T (A (\delta V_k (I + \sigma D_k^{-1}) V_k^T + (\delta V_k (I + \sigma D_k^{-1}) V_k^T)^T \\
&\quad - V_k (C + \sigma D_k^{-1} G D_k^{-1}) V_k^T) v_s \\
&= \lambda_\ell (v_\ell^T \delta V_k (I + \sigma D_k^{-1}) V_k^T v_s + v_\ell^T V_k ((I + \sigma D_k^{-1}) \delta V_k^T v_s \\
&\quad - v_\ell^T V_k (C + \sigma D_k^{-1} G D_k^{-1}) V_k^T v_s) \\
&= \lambda_\ell (v_\ell^T \delta v_s (\sigma \lambda_s^{-1} + 1) + (\sigma \lambda_\ell^{-1} + 1) \delta v_\ell^T v_s - (c_{\ell,s} + \sigma (\lambda_s \lambda_\ell)^{-1} g_{\ell,s})) \\
&= 0.
\end{aligned}$$

□

Theorem 2 *The deflating preconditioner $M_1^{ral}(t)$ is such that the eigenvalues of the preconditioned matrix $M_1^{ral}(t)A$ are*

$$\begin{cases} \mu_i(t) = \sigma + \lambda_i(W \circ \Delta + W^T \circ \Delta^T)t + o(t) & \text{if } i \leq k, \\ \mu_i(t) = \lambda_i(A) + o(t) & \text{if } i > k. \end{cases}$$

Proof of Theorem 2: We have

$$AM_1^{ral}(t) = A + \sigma A \tilde{V}_k (\text{diag}(\tilde{v}_i^T A \tilde{v}_i))^{-1} \tilde{V}_k - A \tilde{V}_k (\tilde{V}_k^T \tilde{V}_k)^{-1} \tilde{V}_k^T.$$

From (17) follows that $\text{diag}(\tilde{v}_i^T A \tilde{v}_i) = D_k + o(t)$, which yields

$$\tilde{V}_k (\text{diag}(\tilde{v}_i^T A \tilde{v}_i))^{-1} \tilde{V}_k = V_k D_k^{-1} V_k^T + \delta V_k D_k^{-1} V_k + V_k D_k^{-1} \delta V_k + o(t) \quad (21)$$

Using (16) shows that $AM_1^{ral}(t) = AM_1^{ral}(0) + \delta M_2 + o(t)$, where

$$\begin{aligned}
\delta M_2 &= A (\sigma \delta V_k D_k^{-1} V_k^T + \sigma V_k D_k^{-1} \delta V_k^T - \delta V_k V_k^T - (\delta V_k V_k^T)^T + V_k C V_k^T) \\
&= A (\delta V_k (\sigma D_k^{-1} - I) V_k^T + V_k (\sigma D_k^{-1} - I) \delta V_k^T + V_k C V_k^T).
\end{aligned}$$

- For $j > k$, the first order approximation of the simple eigenvalues writes

$$\mu_j = \lambda_j + v_j^T \delta M_2 v_j + o(t) = \lambda_j + o(t),$$

since $v_j \delta M_2 v_j = 0$ because (31) and the orthogonality of V .

- For $j \leq k$, the first order approximation of the semi-simple multiple eigenvalue σ writes as in the proof of Theorem 1

$$\mu_j = \sigma + \lambda_j (V_k^T \delta M_2 V_k) + o(t),$$

where $(V_k^T \delta M_2 V_k)$ is a $k \times k$ matrix. The diagonal element of $(V_k^T \delta M_2 V_k)$ is

$$\begin{aligned}
(V_k^T \delta M_2 V_k)_{\ell,\ell} &= v_\ell^T A (\delta V_k (\sigma D_k^{-1} - I) V_k^T + V_k (\sigma D_k^{-1} - I) \delta V_k^T + V_k C V_k^T) v_\ell \\
&= \lambda_\ell ((\sigma \lambda_\ell^{-1} - 1) v_\ell^T \delta v_\ell + (\sigma \lambda_\ell^{-1} - 1) \delta v_\ell^T v_\ell + c_{\ell\ell}) \\
&= 0
\end{aligned} \tag{22}$$

where the final equality uses (31) and $c_{\ell\ell} = 0$. For the (ℓ, s) off-diagonal element,

$$\begin{aligned} (V_k^T \delta M_2 V_k)_{\ell,s} &= v_\ell^T A (\delta V_k (\sigma D_k^{-1} - I) V_k^T + V_k (\sigma D_k^{-1} - I) \delta V_k^T + V_k C V_k^T) v_s \\ &= \lambda_\ell ((\sigma \lambda_s^{-1} - 1) v_\ell^T \delta v_s + (\sigma \lambda_\ell^{-1} - 1) \delta v_\ell^T v_s + c_{\ell,s}) \\ &= t v_\ell^T E v_s \frac{\lambda_\ell}{\lambda_s} \frac{\sigma}{\lambda_s - \lambda_\ell} - t v_\ell^T E^T v_s \frac{\sigma}{\lambda_s - \lambda_\ell}. \end{aligned}$$

We consider the diagonal scaling

$$F = D_k^{-1/2} V_k^T \delta M_2 V_k D_k^{1/2}.$$

The matrix F has the same eigenvalues as $V_k^T \delta M_2 V_k$. The end of the proof follows by noticing that the diagonal entries of $F_{\ell,\ell}$ are zeros. The (ℓ, s) off-diagonal entry verifies

$$\begin{aligned} F_{\ell,s} &= \frac{\sigma}{\lambda_s - \lambda_\ell} \left(\sqrt{\frac{\lambda_\ell}{\lambda_s}} v_\ell^T E v_s - \sqrt{\frac{\lambda_s}{\lambda_\ell}} v_\ell^T E^T v_s \right) \\ &= \frac{\sigma}{\lambda_s - \lambda_\ell} \sqrt{\frac{\lambda_\ell}{\lambda_s}} v_\ell^T E v_s + \frac{\sigma}{\lambda_\ell - \lambda_s} \sqrt{\frac{\lambda_s}{\lambda_\ell}} v_\ell^T E v_\ell \\ &= W_{\ell,s} \Delta_{\ell,s} + W_{s,\ell} \Delta_{s,\ell} = (W \circ \Delta + W^T \circ \Delta^T)_{\ell,s} \quad (23) \end{aligned}$$

□

Theorem 3 *The deflating preconditioner $M_1^{eig}(t)$ is such that the eigenvalues of the preconditioned matrix $M_1^{eig}(t)A$ are*

$$\begin{cases} \mu_i(t) = \sigma + \lambda_i((W + J) \circ \Delta + W^T \circ \Delta^T) t + o(t) & \text{if } i \leq k, \\ \mu_i(t) = \lambda_i(A) + o(t) & \text{if } i > k. \end{cases}$$

Proof of Theorem 3: We have

$$AM_1^{eig}(t) = A + \sigma A \tilde{V}_k \left(\text{diag}(\tilde{\lambda}_i) \right)^{-1} \tilde{V}_k^T - A \tilde{V}_k \left(\tilde{V}_k^T \tilde{V}_k \right)^{-1} \tilde{V}_k^T.$$

Since $\text{diag}(\tilde{\lambda}_i \tilde{v}_i^T \tilde{v}_i) = D_k + t \text{diag}(v_i^T E v_i) + o(t)$, we obtain

$$\begin{aligned} \tilde{V}_k \left(\text{diag}(\tilde{\lambda}_i \tilde{v}_i^T \tilde{v}_i) \right)^{-1} \tilde{V}_k^T &= -t V_k^T D_k^{-2} \text{diag}(v_i^T E v_i) V_k^T \\ &\quad + \delta V_k D_k^{-1} V_k^T + V_k D_k^{-1} \delta V_k^T + o(t). \quad (24) \end{aligned}$$

Therefore, $AM_1^{eig}(t) = AM_1^{eig}(0) + \delta M_3 + o(t)$, where

$$\begin{aligned} \delta M_3 &= -t \sigma A V_k D_k^{-2} \text{diag}(v_i^T E v_i) V_k^T + A (\sigma \delta V_k D_k^{-1} V_k^T + \sigma V_k D_k^{-1} \delta V_k^T \\ &\quad - \delta V_k V_k^T - (\delta V_k V_k^T)^T + V_k C V_k^T) \\ &= -t \sigma V_k D_k^{-1} \text{diag}(v_i^T E v_i) V_k^T + \delta M_2. \end{aligned}$$

- For $j > k$, the first order approximation of the simple eigenvalues writes

$$\mu_j = \lambda_j + v_j^T \delta M_3 v_j + o(t) = \lambda_j + o(t),$$

since $v_j \delta M_3 v_j = 0$ because (31) and the orthogonality of V .

- For $j \leq k$, the first order approximation of the semi-simple multiple eigenvalue σ writes as in the proof of Theorem 1

$$\mu_j = \sigma + \lambda_j (V_k^T \delta M_3 V_k) + o(t),$$

where $(V_k^T \delta M_3 V_k)$ is a $k \times k$ matrix. The diagonal element of $(V_k^T \delta M_3 V_k)$ is

$$(V_k^T \delta M_3 V_k)_{\ell, \ell} = (V_k^T \delta M_2 V_k)_{\ell, \ell} - \sigma \frac{v_\ell^T E v_\ell}{\lambda_\ell} t = 0 - \sigma \frac{v_\ell^T E v_\ell}{\lambda_\ell} t,$$

where the final equality uses (22). For the (ℓ, s) off-diagonal element,

$$\begin{aligned} (V_k^T \delta M_3 V_k)_{\ell, s} &= (V_k^T \delta M_2 V_k)_{\ell, s} + (t \sigma D_k^{-1} \text{diag}(v_i^T E v_i))_{\ell, s} \\ &= (V_k^T \delta M_2 V_k)_{\ell, s} \end{aligned}$$

and the result follows from a diagonal scaling by considering the matrix

$$D_k^{-1/2} V_k^T \delta M_3 V_k^T D_k^{1/2}$$

which has the same eigenvalues as $V_k^T \delta M_3 V_k$.

Arguments similar to those used to derive (23) conclude the proof. □

Theorem 4 *The deflating preconditioner $M_2^{ral}(t)$ is such that the eigenvalues of the preconditioned matrix $M_2^{ral}(t)A$ are*

$$\begin{cases} \mu_i(t) = \sigma + \lambda_i (Y \circ \Delta + Y^T \circ \Delta^T) t + o(t) & \text{if } i \leq k, \\ \mu_i(t) = \lambda_i(A) + o(t) & \text{if } i > k. \end{cases}$$

Proof of Theorem 4: A first order expansion shows that $AM_2^{ral}(t) = AM_2^{ral}(0) + \delta M_4 + o(t)$, where

$$\delta M_4 = \sum_{i=1}^k \left(\left(\frac{\sigma}{\lambda_i} - 1 \right) (A \delta v_i v_i^T + A v_i \delta v_i^T) - \frac{\sigma}{\lambda_i^2} (\delta v_i^T A v_i + v_i^T A \delta v_i) A v_i v_i^T \right).$$

- For $j > k$, the first order approximation of the simple eigenvalues writes

$$\mu_j = \lambda_j + v_j^T \delta M_4 v_j + o(t) = \lambda_j + o(t),$$

since $v_j^T \delta M_4 v_j = 0$ because (31) and the orthogonality of V .

- For $j \leq k$, the first order approximation of the semi-simple multiple eigenvalue σ writes as in the proof of Theorem 1

$$\mu_j = \sigma + \lambda_j (V_k^T \delta M_4 V_k) + o(t),$$

where $(V_k^T \delta M_4 V_k)$ is a $k \times k$ matrix. The diagonal element of $(V_k^T \delta M_4 V_k)$ is

$$(V_k^T \delta M_4 V_k)_{\ell, \ell} = 0,$$

using the orthogonality of V . For the (ℓ, s) off-diagonal element, using (30) gives

$$\begin{aligned} (V_k^T \delta M_4 V_k)_{\ell, s} &= \left(\frac{\sigma}{\lambda_s} - 1\right) v_\ell^T A \delta v_s + \left(\frac{\sigma}{\lambda_\ell} - 1\right) v_\ell^T A v_\ell \delta v_\ell^T v_s \\ &= (\lambda_s - \lambda_\ell)^{-1} \left(\left(\frac{\sigma \lambda_\ell}{\lambda_s} - \lambda_\ell\right) v_\ell^T E v_s - (\sigma - \lambda_\ell) v_\ell^T E^T v_s \right) t \\ &= (\lambda_s - \lambda_\ell)^{-1} \left(\frac{\lambda_\ell}{\lambda_s} (\sigma - \lambda_s) v_\ell^T E v_s - (\sigma - \lambda_\ell) v_\ell^T E^T v_s \right) t \end{aligned}$$

and the result follows from a diagonal scaling by considering the matrix

$$D_k^{-1/2} V_k^T \delta M_4 V_k D_k^{1/2}$$

which has the same eigenvalues as $V_k^T \delta M_4 V_k$.

Arguments similar to those used to derive (23) conclude the proof. □

Theorem 5 *The deflating preconditioner $M_2^{eig}(t)$ is such that the eigenvalues of the preconditioned matrix $M_2^{eig}(t)A$ are*

$$\begin{cases} \mu_i(t) = \sigma + \lambda_i ((Y + J) \circ \Delta + Y^T \circ \Delta^T) t + o(t) & \text{if } i \leq k, \\ \mu_i(t) = \lambda_i(A) + o(t) & \text{if } i > k. \end{cases}$$

Proof of Theorem 5: A first order expansion shows that $AM_2^{eig}(t) = AM_2^{eig}(0) + \delta M_5 + o(t)$, where

$$\delta M_5 = \sum_{i=1}^k \left(\left(\frac{\sigma}{\lambda_i} - 1\right) (A \delta v_i v_i^T + A v_i \delta v_i^T) - \sigma \frac{v_i^T E v_i}{\lambda_i^2} A v_i v_i^T \right)$$

- For $j > k$, the first order approximation of the simple eigenvalues writes

$$\mu_j = \lambda_j + v_j^T \delta M_5 v_j + o(t) = \lambda_j + o(t),$$

since $v_j^T \delta M_5 v_j = 0$ because (31) and the orthogonality of V .

- For $j \leq k$, the first order approximation of the semi-simple multiple eigenvalue σ writes as in the proof of Theorem 1

$$\mu_j = \sigma + \lambda_j(V_k^T \delta M_5 V_k) + o(t),$$

where $(V_k^T \delta M_5 V_k)$ is a $k \times k$ matrix. The diagonal element of $(V_k^T \delta M_5 V_k)$ is

$$(V_k^T \delta M_5 V_k)_{\ell, \ell} = -\sigma \frac{v_\ell^T E v_\ell}{\lambda_\ell} t,$$

using the orthogonality of V . For the (ℓ, s) off-diagonal element, using (30) gives

$$(V_k^T \delta M_5 V_k)_{\ell, s} = \left(\frac{\sigma}{\lambda_s} - 1\right) v_\ell^T A \delta v_s + \left(\frac{\sigma}{\lambda_\ell} - 1\right) v_\ell^T A v_\ell \delta v_\ell^T v_s = (V_k^T \delta M_4 V_k)_{\ell, s}$$

and the result follows from a diagonal scaling by considering the matrix

$$D_k^{-1/2} V_k^T \delta M_5 V_k D_k^{1/2}$$

which has the same eigenvalues as $V_k^T \delta M_5 V_k$.

Arguments similar to those used to derive (23) conclude the proof. \square

Theorem 6 *The coarse-grid preconditioner $M_3(t)$ (resp. $M_3^{ral}(t)$) is such that the eigenvalues of the preconditioned matrix $M_3(t)A$ (resp. $M_3^{ral}(t)A$) are*

$$\begin{cases} \mu_i(t) = \sigma + \lambda_i(A) + o(t) & \text{if } i \leq k, \\ \mu_i(t) = \lambda_i(A) + o(t) & \text{if } i > k. \end{cases}$$

Proof of Theorem 6: Let

$$AM_3(t) = A + \sigma A \tilde{V}_k \left(\tilde{V}_k^T A \tilde{V}_k \right)^{-1} \tilde{V}_k,$$

from (20) we have

$$\begin{aligned} AM_3(t) &= A + \sigma A V_k (V_k^T A V_k)^{-1} V_k^T + \\ &\quad \sigma A \left(\delta V_k D_k^{-1} V_k^T + (\delta V_k D_k^{-1} V_k^T)^T - V_k D_k^{-1} G D_k^{-1} V_k^T \right) + o(t) \\ &= AM_3(0) + \delta M_6 + o(t) \end{aligned}$$

with

$$\delta M_6 = \sigma A \left(\delta V_k D_k^{-1} V_k^T + (\delta V_k D_k^{-1} V_k^T)^T - V_k D_k^{-1} G D_k^{-1} V_k^T \right). \quad (25)$$

For any j the first order approximation of the simple eigenvalues writes

$$\mu_j = \sigma + \lambda_j + v_j^T \delta M_6 v_j + o(t),$$

with $\delta M_6 v_j = 0$ because (31) and the orthogonality of V . A comparison of (21) and (20) shows that the proof for $M_3^{ral}(t)$ is obtained by setting $G = 0$ in the above proof.

□

Theorem 7 *The coarse-grid preconditioner $M_3^{eig}(t)$ is such that the eigenvalues of the preconditioned matrix $M_3^{eig}(t)A$ are*

$$\begin{cases} \mu_i(t) = \sigma + \lambda_i(A) - \sigma \frac{v_i^T E v_i}{\lambda_i(A)} t + o(t) & \text{if } i \leq k, \\ \mu_i(t) = \lambda_i(A) + o(t) & \text{if } i > k. \end{cases}$$

Proof of Theorem 7: Let

$$AM_3^{eig}(t) = A + \sigma A \tilde{V} \text{diag}(\tilde{\lambda}_i \tilde{v}_i^T \tilde{v}_i)^{-1} \tilde{V},$$

from (24) we have

$$\begin{aligned} AM_3^{eig}(t) &= AM_3^{eig}(0) \\ &\quad - \sigma V_k^T D_k^{-1} \text{diag}(v_i^T \delta A v_i) V_k^T + \sigma A \delta V_k D_k^{-1} V_k^T + \sigma A V_k D_k^{-1} \delta V_k^T + o(t) \\ &= AM_3^{eig}(0) + \delta M_7 + o(t), \end{aligned}$$

with

$$\delta M_7 = \sigma V_k^T D_k^{-1} \text{diag}(v_i^T \delta A v_i) V_k^T + \sigma A \delta V_k D_k^{-1} V_k^T + \sigma A V_k D_k^{-1} \delta V_k^T.$$

For any j the first order approximation of the simple eigenvalues writes

$$\mu_j = \sigma + \lambda_j + v_j^T \delta M_7 v_j + o(t),$$

with $v_j^T \delta M_7 v_j = -\sigma \frac{v_j^T E v_j}{\lambda_j} t$ because (31) and the orthogonality of V .

□

Remark 1 *All the theoretical study has been made assuming that all the eigenvalues of A are simple. Actually, the results are still true if some of the λ_i for $i > k$ are multiple (i.e. the ones that are not targeted by the preconditioners).*

3.3 Sensitivity and backward errors

For all the preconditioners considered in this paper, the eigenvalues of the preconditioned matrices write $\mu_i(t) = \mu_i(0) + \lambda_i ((X_1 + X_2) \circ \Delta + X_1^T \circ \Delta^T) t + o(t)$, where the X_i are matrices depending on selected preconditioner and the targeted eigenvalues. We summarize the various values of the matrices X_1 and X_2 for the different preconditioners in Table 1. We can therefore define a condition number κ_i for the eigenvalue μ_i in the direction of E [15] by

$$\kappa_i = \lim_{u \rightarrow 0} \sup_{0 < |t| < u} \frac{|\mu_i(t) - \mu_i(0)|}{|t|} = |\lambda_i ((X_1 + X_2) \circ \Delta + X_1^T \circ \Delta^T)|. \quad (26)$$

Taking norms and using the submultiplicativity of the spectral norm with respect to the Hadamard product yields

$$\begin{aligned}
\kappa_i &\leq \|(X_1 + X_2) \circ \Delta + X_1^T \circ \Delta^T\| \\
&\leq (2\|X_1\| + \|X_2\|)\|\Delta\| = (2\|X_1\| + \|X_2\|)\|V_k^T E V_k\| \\
&\leq (2\|X_1\| + \|X_2\|),
\end{aligned} \tag{27}$$

where we have used that $\|E\| = 1$. Equations (26) and (27) show that if the entries of X_1 and X_2 are small, the condition number of the eigenvalues μ_i is small. By inspecting the equalities (12) and (13) follows that asymptotically for $t \rightarrow 0$,

- the preconditioners M_1 , M_3 and M_3^{ral} are stable (i.e. X_1, X_2 are the zero matrix),
- the preconditioners M_1^{ral} , M_2^{ral} , M_1^{eig} and M_2^{eig} present an instability if for some (s, ℓ) , the ratio $\frac{\sigma}{\lambda_s - \lambda_\ell} \sqrt{\frac{\lambda_\ell}{\lambda_s}}$ is large. In the above statement we have assumed that λ_s is far from σ , which seems to be a reasonable assumption as otherwise we would not have targeted this eigenvalue. This instability happens for instance if some eigenvalues are clustered or small and isolated.

In Table 1 we summarize the situation where a high sensitivity of the eigenvalues is expected. For the solution of linear systems, it is possible to combine backward

Prec	X_1	X_2	some cases of ill-conditioning
M_1	0	0	none
M_1^{ral}	W	0	cluster, small
M_1^{eig}	W	J	cluster, small
M_2^{ral}	Y	0	cluster, small
M_2^{eig}	Y	J	cluster, small
M_3	0	0	none
M_3^{ral}	0	0	none
M_3^{eig}	0	J	small

Table 1: Matrices X_1 and X_2 for the spectral preconditioners and some cases of ill-conditioning. The terms “cluster” and “small” refer respectively to the presence of cluster or of small isolated eigenvalues.

error with sensitivity analysis to obtain an estimate of the forward error on the solution [11], under the assumption that the dependency of the solution on the matrix is not too nonlinear. Similar assumptions yield in our framework, that if the eigenpairs used in the preconditioners are exact eigenpairs of $A + tE$, we can estimate the perturbation on the eigenvalues of the preconditioners induced by the use of inexact spectral information by the quantity $(2\|X_1\| + \|X_2\|)$.

If the eigenelements $(\tilde{\lambda}_i, \tilde{v}_i)$ are given they can readily be considered as exact eigenpairs of the matrix

$$A - (A\tilde{V}_k - \tilde{D}_k\tilde{V}_k)\tilde{V}_k^\dagger, \quad (28)$$

where \dagger denotes the More-Penrose inverse, i.e. $tE = -(A\tilde{V}_k - \tilde{D}_k\tilde{V}_k)\tilde{V}_k^\dagger$. Therefore an estimate for the gap $|\mu_i - \tilde{\mu}_i|$ is $(2\|X_1\| + \|X_2\|)\|(A\tilde{V}_k - \tilde{D}_k\tilde{V}_k)\tilde{V}_k^\dagger\|$. We note that if \tilde{V}_k has orthogonal columns, \tilde{V}_k^\dagger reduces to \tilde{V}_k^T , and an upper-bound for the gap is $(2\|X_1\| + \|X_2\|)\|A\tilde{V}_k - \tilde{D}_k\tilde{V}_k\|$.

This analysis shows that the gap $|\mu_i - \tilde{\mu}_i|$ depends on two quantities:

- the quantities $(2\|X_1\| + \|X_2\|)$ which characterizes the mathematical sensitivity of the eigenvalues of the preconditioned matrix,
- the residual error $\|(A\tilde{V}_k - \tilde{D}_k\tilde{V}_k)\tilde{V}_k^\dagger\|$ which from (28) is an upper-bound for the backward error associated with the approximate eigenpairs.

For a targeted gap ρ_{targ} , the first order analysis shows that the residual error has to be small enough so that its product with the sensitivity term does not exceed ρ_{targ} . A screening of spectral information could be based on this idea.

3.4 Numerical illustrations

To illustrate and assess the correctness of the first order expansions given in the previous section we consider 10×10 matrices with prescribed eigenvalues. We take $\sigma = 1$, $A = QDQ^T$, where Q is the orthogonal matrix whose columns are the eigenvectors of the 2D-Laplacian. The matrix D is chosen to illustrate the high sensitivity of the preconditioners when the eigenvalues of A are small or clustered. We use the preconditioners to shift the 3 smallest eigenvalues of A . We consider the perturbed matrix $A + tE$, where E is a random matrix with unit spectral norm and t is the perturbation parameter. We compute with the routine `eig` of MATLAB the eigenpairs $(\lambda_i(t), v_i(t))$ of $A + tE$. The 3 eigenpairs corresponding to those selected on the original matrix are used to construct the approximate “deflating” and “coarse grid” preconditioners. An additional eigenvalue computation, still with `eig`, is performed on the preconditioned matrix to obtain the eigenvalues $\tilde{\mu}_i$. The estimates using the first order expansions of Theorems 1 to 7 are denoted by $\bar{\mu}_i$.

In a first experiment we choose $D = \text{diag}(1, 2, \dots, 10)$, and decide to move 1, 2 and 3. In Table 2, we report on the distance $\tilde{\mu}_i - \bar{\mu}_i$, for $i = 1, \dots, 4$ the four smallest eigenvalues of the preconditioned system. We see that for a perturbation size of $t = 10^{-2}$ the first order estimation accurately approximate the computed eigenvalues both the multiple ones (first three) and the simple one (fourth): $|\tilde{\mu}_i - \bar{\mu}_i|$ has several order of magnitude less than t .

As mentioned in Section 3.3 the first order expansions reveal the sensitivity of the spectrum of the preconditioned matrices to the magnitude and clustering of the targeted eigenvalues.

Precond	$ \tilde{\mu}_1 - \bar{\mu}_1 $	$ \tilde{\mu}_2 - \bar{\mu}_2 $	$ \tilde{\mu}_3 - \bar{\mu}_3 $	$ \tilde{\mu}_4 - \bar{\mu}_4 $
M_1	3.8e-07	7.1e-08	2.8e-08	3.8e-08
M_1^{ral}	1.5e-06	1.5e-07	1.2e-06	3.8e-08
M_1^{eig}	2.3e-05	3.4e-07	3.5e-07	3.9e-08
M_2^{ral}	1.4e-06	8.1e-08	9.6e-07	3.8e-08
M_2^{eig}	2.3e-05	6.7e-07	5.1e-07	3.9e-08
M_3	5.7e-07	2.5e-07	3.9e-05	3.9e-05
M_3^{ral}	2.0e-06	3.0e-07	3.9e-05	4.0e-05
M_3^{eig}	2.2e-05	2.2e-07	1.2e-05	1.3e-05

Table 2: Difference between the computed and first order estimate of the eigenvalues for $t = 10^{-2}$.

We first investigate the case of a small isolated eigenvalue. For the matrix D , we consider $\lambda_1 = 10^{-4}$, $\lambda_2 = 10^{-2}$ and $\lambda_3 = 10^{-1}$; the seven other eigenvalues are in the neighbourhood of 2. In Table 3 we display, for the perturbation size $t = 10^{-5}$, the exact radius $\rho = \max_{i=1,2,3} |1 - \mu_i|$, the computed radius $\tilde{\rho} = \max_{i=1,2,3} |1 - \tilde{\mu}_i|$ and the estimated radius $\bar{\rho} = \max_{i=1,2,3} |1 - \bar{\mu}_i|$. These quantities measure the radius of the cluster obtained around 1. If the exact eigenpairs were available we would have ρ equal to zero for the deflating preconditioners and equal to $10^{-1} = \max\{\lambda_1, \lambda_2, \lambda_3\}$ for the coarse grid preconditioners. The comparison of the first two results shows again the sharpness of the estimations given by the first order approximations that is fairly accurate for all the preconditioners. Inspecting the first and third column we see that M_1 does a fairly good job in moving the eigenvalues close to one. This nice clustering deteriorates when one uses the Rayleigh quotients for approximating the eigenvalues (i.e. M_1^{ral} and M_2^{ral}) and is even worse when we use the perturbed eigenvalues computed by `eig` (i.e. M_1^{eig} and M_2^{eig}). Regarding the coarse grid preconditioners, we see that both M_3 and M_3^{ral} behave similarly as if they were constructed with exact eigenpairs: they shift λ_i to $1 + \lambda_i$. The preconditioner M_3^{eig} is slightly more sensitive to perturbations. In addition, we see that the coarse grid preconditioners give rise to similar cluster radius as M_1^{eig} and M_2^{eig} . We remind that M_1^{eig} and M_2^{eig} are supposed to better cluster around 1, because they are designed to translate to 1 exactly and not to $1 + \lambda_i$.

In order to study the sensitivity of the spectrum of the preconditioners when the targeted eigenvalues are clustered we define the matrix D by taking its 3 smallest diagonal entries in a cluster of radius 10^{-4} around 10^{-1} ; the others seven eigenvalues are in the neighbourhood of 2. We consider a perturbation of size $t = 10^{-5}$ for the numerical experiments reported in Table 4. The first and third columns show that M_1 is efficient in clustering the targeted eigenvalues close to 1. The other variants of the deflating approaches behave rather poorly as $\tilde{\rho}$ is about 10^{-2} . Regarding the coarse grid preconditioners, M_3 behave the best. The two others are still performing well as $\tilde{\rho}$ does not grow much beyond

Precond	$\tilde{\rho}$	$\bar{\rho}$	ρ
M_1	3.7e-07	0.0e+00	0.0e+00
M_1^{ral}	2.0e-03	2.0e-03	0.0e+00
M_1^{eig}	1.8e-01	2.1e-01	0.0e+00
M_2^{ral}	2.0e-03	2.0e-03	0.0e+00
M_2^{eig}	1.8e-01	2.1e-01	0.0e+00
M_3	1.0e-01	1.0e-01	1.0e-01
M_3^{ral}	1.0e-01	1.0e-01	1.0e-01
M_3^{eig}	1.7e-01	2.1e-01	1.0e-01

Table 3: Small isolated eigenvalue : computed and estimated cluster radius for $t = 10^{-5}$.

its exact value ρ .

Precond	$\tilde{\rho}$	$\bar{\rho}$	ρ
M_1	4.2e-10	0.0e+00	0.0e+00
M_1^{ral}	2.5e-02	2.7e-02	0.0e+00
M_1^{eig}	2.5e-02	2.7e-02	0.0e+00
M_2^{ral}	2.3e-02	2.4e-02	0.0e+00
M_2^{eig}	2.3e-02	2.4e-02	0.0e+00
M_3	1.0e-01	1.0e-01	1.0e-01
M_3^{ral}	1.2e-01	1.0e-01	1.0e-01
M_3^{eig}	1.2e-01	1.0e-01	1.0e-01

Table 4: Clustered eigenvalues : computed and estimated cluster radius for $t = 10^{-5}$.

These experimental results are in agreement with what could be expected from Table 1.

4 Use of inexact preconditioners in CG

In this section we illustrate the effect of using the approximate eigenpairs on the convergence behaviour of PCG. In that respect we consider the 685×685 BUS685 matrix, denoted B_{685} , from the Harwell-Boeing collection. We compute an Incomplete Cholesky factorization (IC) CC^T with threshold $4 \cdot 10^{-1}$, which is our first level preconditioner and consequently $\sigma = 1$. We apply the various spectral preconditioners to the matrix $A = C^{-1}B_{685}C^{-T}$. As in the previous series of experiments, we use the eigenvectors of the perturbed matrix $A + tE$ to build the preconditioners. Consequently we use eigenelements that have a backward error of the order of t .

The right-hand side is chosen so that the solution of $Ax = b$ is the vector

of all ones : $x = (1, \dots, 1)^T$, $b = Ax$. For the numerical experiments the initial guess is the zero vector and we decide to stop the PCG iterations when the normalized unpreconditioned residual is reduced by 10^{-9} , so that the stopping criterion is independent of the preconditioner. Even though this quantity might be a by-product of the PCG solver we explicitly compute the unpreconditioned residual to decide when to stop the iterations.

In Table 5 we report on the number of PCG iterations for the various deflating and coarse-grid preconditioners for $t = 10^{-12}$ and $t = 10^{-3}$. We vary from one to ten the number ne of eigenlements used to build the preconditioners. The eigenvalues of A obtained with `eig` are reported in Table 6. For the smallest perturbation, that is $t = 10^{-12}$, all the preconditioners behave exactly the same. Because the IC preconditioner has already clustered many of the eigenvalues close to one, moving the smallest eigenvalues (that vary from $5.67 \cdot 10^{-4}$ to $3.05 \cdot 10^{-2}$ see Table 6) exactly to one or shifting them by one leads to the same behaviour of PCG. However, when a perturbation is applied, that is when the eigenlements are less accurately computed, some differences appear. Both M_1 and M_3 perform similarly and outperform the others. Then the various variants that approximate the eigenvalues using Rayleigh quotients perform similarly. The worse behaviour is observed for the variants that make use of the approximate eigenvalues. Although not reported in that paper on all the experiments we have run with PCG in our study exhibit the same trend.

$t = 10^{-12}$										
ne	1	2	3	4	5	6	7	8	9	10
All Prec	161	147	132	118	106	95	88	81	81	79
$t = 10^{-3}$										
ne	1	2	3	4	5	6	7	8	9	10
M_1	168	154	142	129	122	114	107	104	101	98
M_1^{ral}	168	155	152	142	144	139	133	131	126	121
M_1^{eig}	168	156	154	141	143	140	134	194	183	183
M_2^{ral}	168	155	151	139	138	133	125	124	121	117
M_2^{eig}	168	156	152	140	139	133	125	184	173	171
M_3	168	154	142	129	121	114	107	104	101	98
M_3^{ral}	168	155	152	139	138	133	125	124	121	117
M_3^{eig}	168	156	152	140	139	133	125	184	174	171

Table 5: # iterations for PCG for perturbed eigenpairs on the BUS 685 matrix.

5 Concluding remarks

We use the perturbation theory for eigenvalues and eigenvectors to investigate the behaviour deflating and coarse preconditioners for SPD linear systems. Our analysis shows a better stability of the preconditioners M_1 and M_3 compared

λ_1	λ_2	λ_3	λ_4	λ_5
5.67e-04	2.91e-03	3.58e-03	4.71e-03	6.19e-03
λ_6	λ_7	λ_8	λ_9	λ_{10}
8.35e-03	1.85e-02	2.05e-02	3.04e-02	3.05e-02

Table 6: The ten smallest eigenvalues targeted by the spectral preconditioners on the BUS 685 matrix.

to the other preconditioning variants that exploit some additional properties that are only true for exact eigenpairs. We have established theoretical results that provide us with first order approximations of the eigenvalues of the preconditioned matrix. These estimates give sharp approximations. They reveal the possible instabilities that might lead to poor preconditioners if inexact eigen-information is used. They show that targeting small eigenvalues or small clusters may require a backward stable calculation of the eigenelements. An important result of this work is that the efficiency of a spectral preconditioner should not be assessed only using exact eigenpairs. In practice these preconditioners may be built using approximate information, that is computed either with an eigensolver or obtained by some approximations as in domain decomposition [5]. Unsymmetric spectral preconditioners exist and similar results can be developed; this topic will be the scope of a future work.

References

- [1] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1995.
- [2] J. H. Bramble, J. E. Pasciak, and A. H. Schatz. The construction of preconditioners for elliptic problems by substructuring I. *Math. Comp.*, 47(175):103–134, 1986.
- [3] B. Carpentieri, I. S. Duff, and L. Giraud. A class of spectral two-level preconditioners. *SIAM Journal on Scientific Computing*, 25:749–765, 2003.
- [4] B. Carpentieri, L. Giraud, and S. Gratton. Additive and multiplicative spectral two-level preconditioners for general linear systems. Technical Report TR/PA/04/38, CERFACS, Toulouse, France, 2004.
- [5] T. F. Chan and T. P. Mathew. *Domain Decomposition Algorithms*, volume 3 of *Acta Numerica*, pages 61–143. Cambridge University Press, Cambridge, 1994.
- [6] I. S. Duff, L. Giraud, J. Langou, and E. Martin. Using spectral low rank preconditioners for large electromagnetic calculations. Technical Report TR/PA/03/95, CERFACS, Toulouse, France, 2003. Also Technical report

RAL-TR-2003-023, Preliminary version of the paper to appear in Int J. Numerical Methods in Engineering.

- [7] I. S. Duff, R. G. Grimes, and J. G. Lewis. Sparse matrix test problems. *ACM Transactions on Mathematical Software*, 15:1–14, 1989.
- [8] M. Fisher. Minimization algorithms for variational data assimilation. In *Proc. ECMWF seminar "Recent developments in numerical methods for atmospheric modelling*, pages 364–385, September 1998.
- [9] J. Frank and C. Vuik. On the construction of deflation-based preconditioners. *SIAM Journal on Scientific Computing*, 23:442–462, 2001.
- [10] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996.
- [11] N. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, 2002. Second edition.
- [12] P. Lancaster and M. Tismenetsky. *The Theory of Matrices*. Academic Press, New York, 1985. Second Edition with Applications.
- [13] R. Nabben and C. Vuik. A comparison of deflation and coarse grid correction applied to porous media flow. Report 03-10, Delft University of Technology, Department of Applied Mathematical Analysis, Delft, 2003.
- [14] A. Padiy, O. Axelsson, and B. Polman. Generalized augmented matrix preconditioning approach and its application to iterative solution of ill-conditioned algebraic systems. *SIAM Journal on Matrix Analysis and Applications*, 22(3):793–818, 2000.
- [15] J. R. Rice. A theory of condition. *SIAM Journal on Numerical Analysis*, 3:287–310, 1966.
- [16] G. W. Stewart. *Matrix algorithms. Volume II: Eigensystems*. Society for Industrial and Applied Mathematics, 2001.
- [17] J. Sun. Multiple eigenvalue sensitivity. *Linear Algebra and its Applications*, 138:183–211, 1990.
- [18] H. Waisman, J. Fish, R. S. Tuminaro, and J. Shadid. The generalized global basis method. *Int. Journal of Numerical Methods in Engineering*, 61(8):1243–1269, 2004.

A Some useful equalities

$$v_\ell^T \delta v_s = t(\lambda_s - \lambda_\ell)^{-1} v_\ell^T E v_s \text{ if } \ell \neq s \quad (29)$$

$$\delta v_\ell^T v_s = t(\lambda_\ell - \lambda_s)^{-1} v_\ell^T E^T v_s \text{ if } \ell \neq s \quad (30)$$

$$v_\ell^T \delta v_\ell = \delta v_\ell^T v_\ell = 0 \quad (31)$$

Proof: For $\ell \neq s$

$$\begin{aligned}
v_\ell^T \delta v_s &= v_\ell^T \underline{V}_s (\lambda_s I - B_s)^{-1} \underline{V}_\ell^T E v_s \\
&= (0 \dots \delta_{j,t} \dots 0) (\lambda_s I - B_s)^{-1} \underline{V}_\ell^T E v_s \\
&= (\lambda_s - \lambda_\ell)^{-1} (0 \dots \delta_{j,t} \dots 0) \underline{V}_\ell^T E v_s \\
&= t (\lambda_s - \lambda_\ell)^{-1} v_\ell^T E v_s
\end{aligned}$$

Similarly

$$\begin{aligned}
\delta v_\ell^T v_s &= v_\ell^T E^T \underline{V}_\ell (\lambda_\ell I - B_\ell)^{-T} \underline{V}_\ell v_s \\
&= v_\ell^T E^T \underline{V}_\ell (\lambda_\ell I - B_\ell)^{-T} (0 \dots \delta_j, s \dots 0)^T \\
&= v_\ell^T E^T \underline{V}_\ell (0 \dots \delta_j, s \dots 0)^T (\lambda_\ell - \lambda_s)^{-1} \\
&= t v_\ell^T E^T v_s (\lambda_\ell - \lambda_s)^{-1}
\end{aligned}$$

For $\ell = s$,

$$\begin{aligned}
\delta v_\ell^T v_\ell &= v_\ell^T \underline{V}_\ell (\lambda_\ell I - B_\ell)^{-1} \underline{V}_\ell^T E v_\ell \\
&= 0 \text{ because } v_\ell^T \underline{V}_\ell = (0 \dots 0) \\
&= v_\ell^T \delta v_\ell
\end{aligned}$$

□