

**Improved Polynomial Bounds and Acceleration of GMRES  
by Solving a min-max Problem on Rectangles,  
and by Deflating**

Nicole Spillane and Daniel B. Szyld

Report 25-04-07  
April 2025. Revised October 2025.

Department of Mathematics  
Temple University  
Philadelphia, PA 19122

This report is available in the World Wide Web at  
<https://https://faculty.cst.temple.edu/~szyld/>  
and at  
<https://hal.science/hal-05022856>



# Improved Polynomial Bounds and Acceleration of GMRES by Solving a min-max Problem on Rectangles, and by Deflating.\*

Nicole Spillane<sup>†</sup> and Daniel B. Szyld<sup>‡</sup>

**Keywords:** GMRES bounds, Min-Max problem on rectangle, Weighted GMRES, Deflated GMRES, left preconditioning, right preconditioning, split preconditioning

**AMS Subject Classification:** 65F10, 65Y05, 68W40

## Abstract

Polynomial convergence bounds are considered for left, right, and split preconditioned GMRES. They include the cases of Weighted and Deflated GMRES for a linear system  $\mathbf{Ax} = \mathbf{b}$ . In particular, the case of positive definite  $\mathbf{A}$  is considered. The well-known polynomial bounds are generalized to the cases considered, and then reduced to solving a min-max problem on rectangles on the complex plane. Several approaches are considered and compared. The new bounds can be improved by using specific deflation spaces and preconditioners. This in turn accelerates the convergence of GMRES. Numerical examples illustrate the results obtained.

## 1 Introduction

Let  $\mathbf{A} \in \mathbb{C}^{n \times n}$  be a non-singular matrix. Our focus is on solving a large sparse linear system of the form

$$\mathbf{Ax} = \mathbf{b}. \quad (1)$$

We consider weighted GMRES with split preconditioning by  $\mathbf{H}_L$  on the left and  $\mathbf{H}_R$  on the right. We refer to this as GMRES preconditioned by  $(\mathbf{H}_L, \mathbf{H}_R)$ . What is meant is that GMRES is applied to

$$\mathbf{H}_L \mathbf{A} \mathbf{H}_R \mathbf{u} = \mathbf{H}_L \mathbf{b} \quad \text{and} \quad \mathbf{x} = \mathbf{H}_R \mathbf{u}. \quad (2)$$

It is assumed that  $\mathbf{A}$ ,  $\mathbf{H}_L$  and  $\mathbf{H}_R$  are all non-singular. Throughout, we use the notation  $\mathbf{H} = \mathbf{H}_R \mathbf{H}_L$  for the combined preconditioner. The term weighted GMRES [16] corresponds to the use of a non-standard inner product within GMRES, here the inner product induced by some Hermitian positive definite (hpd) matrix  $\mathbf{W}$ . Weighted GMRES is studied in [15, 24]; see also [13] where the different inner products are used to study the convergence of GMRES in the 2-norm. As we show in detail in Section 2.1 one bound for the (ideal) preconditioned and weighted GMRES residual is

$$\|\mathbf{H}_L \mathbf{r}_k\|_{\mathbf{W}} \leq \min_{q \in \mathbb{P}_k; q(0)=1} \{\|q(\mathbf{H}_L \mathbf{A} \mathbf{H}_R)\|_{\mathbf{W}}\} \|\mathbf{H}_L \mathbf{r}_0\|_{\mathbf{W}}, \quad (3)$$

where  $\mathbf{r}_0 = \mathbf{b} - \mathbf{Ax}_0$  is the initial residual and  $\mathbb{P}_k$  is the set of polynomials of degree at most  $k$ . Our goal in this paper is to obtain good bounds for the minimization problem in (3) and, at the

---

\*This version dated October 1, 2025

<sup>†</sup>CNRS, CMAP, Ecole Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau Cedex, France (*nicole.spillane@cmap.polytechnique.fr*). This work was supported in part by the ANR JCJC project DARK (research grant ANR-24-CE46-1633).

<sup>‡</sup>Department of Mathematics, Temple University, Philadelphia, PA 19122, USA (*szyld@temple.edu*)

same time, look for appropriate preconditioners and weighting matrices (as well as deflation spaces) to obtain better bounds. To this end, also in Section 2.1, we use the Crouzeix-Palencia result [5], and obtain

$$\min_{q \in \mathbb{P}_k; q(0)=1} \{ \|q(\mathbf{H}_L \mathbf{A} \mathbf{H}_R) \|_{\mathbf{W}} \} \leq (1 + \sqrt{2}) \min_{q \in \mathbb{P}_k; q(0)=1} \max_{z \in FOV^{\mathbf{W}}(\mathbf{H}_L \mathbf{A} \mathbf{H}_R)} |q(z)|,$$

where  $FOV^{\mathbf{W}}$  stands for the field of values in the  $\mathbf{W}$ -inner product.

We also consider Deflated GMRES beginning in Section 3; see, e.g., [19, 23, 34, 42] and references therein for a full description. Recent contributions to deflation for GMRES include [14, 27], as well as our own work [39]. In Section 3.1, we study bounds for Deflated GMRES, and we obtain a similar min-max problem as above, but for a different FOV-type set in  $\mathbb{C}$ .

We are able to bound the sets in  $\mathbb{C}$  for each min-max problem by rectangles, by appropriately choosing the preconditioner, the weighting matrix  $\mathbf{W}$ , and the deflation space. In fact, we consider two different deflation spaces (in Subsections 3.2.1 and 3.2.2), and for each of them, we develop convergence bounds based on the min-max problem in a rectangle. Furthermore, these rectangles are parametrized by a threshold parameter  $\tau$  defining our deflation spaces. Section 4 makes the analysis complete by discussing the solution to the min-max problem on rectangles. The new bounds are an improvement over the linear bounds in [36] and [39] which inspired the present work. This study allows us to choose an appropriate preconditioner, deflation space, and weight matrix for GMRES for specific problems. This is illustrated numerically in Section 5.

To summarize, our contributions in this article consist in a full convergence analysis for split preconditioned GMRES with a non-standard inner product and deflation. In particular the deflation space does not come from approximating eigenvectors of  $\mathbf{A}$ . When  $\mathbf{A}$  is positive definite, *i.e.*, when the Hermitian part of  $\mathbf{A}$  is hpd, the bound depends on a min-max problem on a rectangle in the complex plane that is itself approximated by several methods. Finally, we propose to improve the bound and accelerate convergence by deflation.

## 2 GMRES convergence bounds

### 2.1 The general Case

We begin by stating the minimization property of GMRES in the general case with a combined preconditioner  $\mathbf{H}$ , and with a Hermitian positive definite (hpd) weight matrix  $\mathbf{W}$ .

**Theorem 2.1** (Minimization property). *Consider  $\mathbf{W}$ -weighted GMRES applied to  $\mathbf{Ax} = \mathbf{b}$  and preconditioned by  $(\mathbf{H}_L, \mathbf{H}_R)$ . For any  $k \in \llbracket 0, n \rrbracket$ , the  $k$ -th iterate is characterized by*

$$\mathbf{x}_k = \underset{\mathbf{x} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{H}\mathbf{A}, \mathbf{H}\mathbf{r}_0)}{\operatorname{argmin}} \{ \|\mathbf{H}_L(\mathbf{b} - \mathbf{Ax})\|_{\mathbf{W}} \},$$

where

$$\mathbf{H} := \mathbf{H}_R \mathbf{H}_L \text{ and } \mathcal{K}_k(\mathbf{H}\mathbf{A}, \mathbf{H}\mathbf{r}_0) := \operatorname{span} \left\{ \mathbf{H}\mathbf{r}_0, \mathbf{H}\mathbf{A}\mathbf{H}\mathbf{r}_0, \dots, (\mathbf{H}\mathbf{A})^{k-1} \mathbf{H}\mathbf{r}_0 \right\}.$$

Equivalently, the  $k$ -th residual,  $\mathbf{r}_k = \mathbf{b} - \mathbf{Ax}_k$ , satisfies

$$\|\mathbf{H}_L \mathbf{r}_k\|_{\mathbf{W}} = \min_{q \in \mathbb{P}_k; q(0)=1} \{ \|\mathbf{H}_L q(\mathbf{A}\mathbf{H}) \mathbf{r}_0\|_{\mathbf{W}} \} \leq \min_{q \in \mathbb{P}_k; q(0)=1} \{ \|q(\mathbf{H}_L \mathbf{A} \mathbf{H}_R)\|_{\mathbf{W}} \} \|\mathbf{H}_L \mathbf{r}_0\|_{\mathbf{W}}, \quad (4)$$

where  $\mathbf{r}_0 = \mathbf{b} - \mathbf{Ax}_0$  and  $\mathbb{P}_k$  is the set of polynomials of degree at most  $k$ .

*Proof.* At step  $k$ , GMRES applied to (2) in the  $\mathbf{W}$ -inner product with initial residual  $\mathbf{r}_0$  finds  $\mathbf{H}_R^{-1} \mathbf{x}_k$ , which is the minimizer of  $\|\mathbf{H}_L(\mathbf{b} - \mathbf{Ax}_k)\|_{\mathbf{W}} = \|\mathbf{H}_L \mathbf{r}_k\|_{\mathbf{W}}$  for all possible elements  $\mathbf{x}_k$  of the Krylov subspace  $\mathbf{x}_0 + \mathcal{K}_k(\mathbf{H}\mathbf{A}, \mathbf{H}\mathbf{r}_0) = \mathbf{x}_0 + \mathcal{K}_k(\mathbf{H}_L \mathbf{A} \mathbf{H}_R, \mathbf{H}_L \mathbf{r}_0)$ , where  $\mathbf{r}_k = \mathbf{b} - \mathbf{Ax}_k$ . Then, the minimization over the Krylov subspace and the set of polynomials follow. To obtain the inequality in (4), first note that by definition of the matrix norm,

$$\|\mathbf{H}_L q(\mathbf{A}\mathbf{H}) \mathbf{r}_0\|_{\mathbf{W}} \leq \|\mathbf{H}_L q(\mathbf{A}\mathbf{H}) \mathbf{H}_L^{-1}\|_{\mathbf{W}} \|\mathbf{H}_L \mathbf{r}_0\|_{\mathbf{W}},$$

and that

$$\|\mathbf{H}_L q(\mathbf{A}\mathbf{H})\mathbf{H}_L^{-1}\|_{\mathbf{W}} = \|q(\mathbf{H}_L\mathbf{A}\mathbf{H}_R)\|_{\mathbf{W}}.$$

Indeed, it holds that  $\mathbf{H}_L(\mathbf{A}\mathbf{H})^0\mathbf{H}_L^{-1} = \mathbf{I} = (\mathbf{H}_L\mathbf{A}\mathbf{H}_R)^0$ ,  $\mathbf{H}_L(\mathbf{A}\mathbf{H})\mathbf{H}_L^{-1} = \mathbf{H}_L\mathbf{A}\mathbf{H}_R$ ,  $\mathbf{H}_L(\mathbf{A}\mathbf{H})^2\mathbf{H}_L^{-1} = \mathbf{H}_L\mathbf{A}\mathbf{H}_R\mathbf{H}_L\mathbf{A}\mathbf{H}_R = (\mathbf{H}_L\mathbf{A}\mathbf{H}_R)^2$ , etc.  $\square$

**Remark 2.1.** The Krylov subspace does not depend on the choice of left, right or split preconditioning so long as the combined preconditioner  $\mathbf{H} = \mathbf{H}_R\mathbf{H}_L$  remains the same. What this choice does influence is the norm that gets minimized. So in that sense, moving (multiplicatively) some preconditioning from right to left has the same effect as weighting.

We point out that throughout this article exact arithmetic is assumed. Backward stability analysis as in [9, 29], the influence of a perturbed operator as in [33], or a perturbed deflation operator as in [32] have not yet been considered.

A practical approach for bounding the convergence of GMRES is to consider ideal GMRES, i.e., to find a bound for  $\min_{q \in \mathbb{P}_i; q(0)=1} \{\|\mathbf{H}_L q(\mathbf{A}\mathbf{H})\mathbf{H}_L^{-1}\|_{\mathbf{W}}\}$ . This approximation is discussed in [7] as one of six steps that go into the analysis of GMRES; see [43] for the difference between ideal and worst-case GMRES.

If  $\mathbf{H}_L\mathbf{A}\mathbf{H}_R$  is diagonalizable, the term  $\min_{q \in \mathbb{P}_i; q(0)=1} \{\|q(\mathbf{H}_L\mathbf{A}\mathbf{H}_R)\|_{\mathbf{W}}\}$  can be bounded by a min-max problem where the max is taken over a set that contains the spectrum of  $\mathbf{H}\mathbf{A}$ . For non-normal matrices, a leading constant in the bound appears with the conditioning of eigenvectors of  $\mathbf{H}\mathbf{A}$  in it. Instead, we state here a convergence result where the GMRES min-max problem is posed over the field of values of the coefficient matrix  $\mathbf{A}$  of the linear system. This result, with the leading constant  $(1 + \sqrt{2})$ , is given in [5] where it is shown that the field of values is a  $(1 + \sqrt{2})$ -spectral set. See, also [12, Equation (FOV), page 5] with context and related work. There are also extensions of this bound in [4].

In [5, (1), (3) and Theorem 3.1], it is proved (in particular) that for any bounded linear operator  $A$  in a complex Hilbert space  $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}}, \|\cdot\|_{\mathcal{H}})$  and for any rational function  $f$  with no poles in  $FOV^{\mathcal{H}}(A)$  defined below,

$$\|f(A)\|_{\mathcal{H}} \leq (1 + \sqrt{2}) \sup_{z \in FOV^{\mathcal{H}}(A)} |f(z)|; \quad FOV^{\mathcal{H}}(A) = \{\langle A\mathbf{v}, \mathbf{v} \rangle_{\mathcal{H}}; \mathbf{v} \in \mathcal{H}, \|\mathbf{v}\|_{\mathcal{H}} = 1\}. \quad (5)$$

In the Hilbert space,  $(\mathbb{C}^n, \langle \cdot, \cdot \rangle_{\mathbf{W}}, \|\cdot\|_{\mathbf{W}})$ , we obtain, for any matrix  $\mathbf{B} \in \mathbb{C}^{n \times n}$  and polynomial  $q$ , that

$$\|q(\mathbf{B})\|_{\mathbf{W}} \leq (1 + \sqrt{2}) \sup_{q \in FOV^{\mathbf{W}}(\mathbf{B})} |q(z)|, \quad (6)$$

where the field of values of  $\mathbf{B} \in \mathbb{C}^{n \times n}$  in the  $\langle \cdot, \cdot \rangle_{\mathbf{W}}$  inner product is the set

$$FOV^{\mathbf{W}}(\mathbf{B}) := \left\{ \frac{\langle \mathbf{B}\mathbf{z}, \mathbf{z} \rangle_{\mathbf{W}}}{\langle \mathbf{z}, \mathbf{z} \rangle_{\mathbf{W}}}; \mathbf{z} \in \mathbb{C}^n \setminus \{0\} \right\}.$$

**Remark 2.2.** The field of values of  $\mathbf{B}$  is also, and equally often, called the numerical range of  $\mathbf{B}$ . By the well-known Hausdorff-Toeplitz theorem, see, e.g., [6], the field of values is convex and compact.

**Theorem 2.2** (Crouzeix-Palencia bound). *The  $k$ -th residual of  $\mathbf{W}$ -weighted GMRES applied to  $\mathbf{A}\mathbf{x} = \mathbf{b}$  and preconditioned by  $(\mathbf{H}_L, \mathbf{H}_R)$  is bounded by*

$$\frac{\|\mathbf{H}_L \mathbf{r}_k\|_{\mathbf{W}}}{\|\mathbf{H}_L \mathbf{r}_0\|_{\mathbf{W}}} \leq (1 + \sqrt{2}) \min_{q \in \mathbb{P}_k; q(0)=1} \max_{z \in FOV^{\mathbf{W}}(\mathbf{H}_L\mathbf{A}\mathbf{H}_R)} |q(z)|.$$

Moreover,  $FOV^{\mathbf{W}}(\mathbf{H}_L\mathbf{A}\mathbf{H}_R) = FOV^{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L}(\mathbf{A}\mathbf{H})$ , where again  $\mathbf{H} = \mathbf{H}_R\mathbf{H}_L$ .

*Proof.* In the residual bound (4), we apply the Crouzeix-Palencia result (6) to get

$$\|q(\mathbf{H}_L\mathbf{A}\mathbf{H}_R)\|_{\mathbf{W}} \leq (1 + \sqrt{2}) \sup_{z \in FOV^{\mathbf{W}}(\mathbf{H}_L\mathbf{A}\mathbf{H}_R)} |q(z)| = (1 + \sqrt{2}) \max_{z \in FOV^{\mathbf{W}}(\mathbf{H}_L\mathbf{A}\mathbf{H}_R)} |q(z)|.$$

We have replaced the sup by max as we are maximizing the uniform norm of a polynomial over a compact set in a finite dimensional space. The equality between the two weighted field of values comes from their definition.  $\square$

In order to solve the min-max problem in Theorem 2.2, we must characterize the set

$$FOV^{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L}(\mathbf{A} \mathbf{H}) = \left\{ \frac{\langle \mathbf{A} \mathbf{H} \mathbf{z}, \mathbf{z} \rangle_{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L}}{\langle \mathbf{z}, \mathbf{z} \rangle_{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L}}; \mathbf{z} \in \mathbb{C}^n \setminus \{0\} \right\}.$$

## 2.2 The case when $\mathbf{A}$ is positive definite (pd) and $\mathbf{H} = \mathbf{H}_L^* \mathbf{W} \mathbf{H}_L$

Assume that  $\mathbf{A}$  is pd, *i.e.*, that its Hermitian part is pd,  $\mathbf{H}$  is hpd, and the residual is minimized in the  $\mathbf{H}$ -norm, *i.e.*,  $\mathbf{H} = \mathbf{H}_L^* \mathbf{W} \mathbf{H}_L$ , so that  $\|\mathbf{H}_L \mathbf{r}_i\|_{\mathbf{W}} = \|\mathbf{r}_i\|_{\mathbf{H}}$  in (4). Particular cases when this occurs include the following:

- Left preconditioning by hpd  $\mathbf{H}$  and  $\mathbf{W} = \mathbf{H}^{-1}$ ,
- Right preconditioning by hpd  $\mathbf{H}$  and  $\mathbf{W} = \mathbf{H}$ ,
- Split preconditioning by  $\mathbf{H}_L^*$  on the right,  $\mathbf{H}_L$  on the left, and  $\mathbf{W} = \mathbf{I}$ .

The idea to apply GMRES in the inner product induced by the preconditioner has also been explored by [3, 36, 40]. Moreover the authors of [30] propose other combined choices of preconditioner and weighted inner product.

Let  $\mathbf{M}$  and  $\mathbf{N}$  denote respectively the Hermitian and skew-Hermitian parts of  $\mathbf{A}$ :

$$\mathbf{A} = \mathbf{M} + \mathbf{N}, \mathbf{M} := \frac{\mathbf{A} + \mathbf{A}^*}{2} \text{ and } \mathbf{N} := \frac{\mathbf{A} - \mathbf{A}^*}{2}. \quad (7)$$

The eigenvalues of  $\mathbf{H} \mathbf{M}$  are real and positive as a result of  $\mathbf{H}$  and  $\mathbf{M}$  being hpd. The eigenvalues of  $\mathbf{N} \mathbf{H}$  and of  $\mathbf{M}^{-1} \mathbf{N}$  are purely imaginary as a result of  $\mathbf{H}$  and  $\mathbf{M}$  being hpd, and of  $\mathbf{N}$  being skew-Hermitian. Moreover, the non-zero eigenvalues of  $\mathbf{N} \mathbf{H}$  and of  $\mathbf{M}^{-1} \mathbf{N}$  come in complex-conjugate pairs. In what follows,  $\lambda_{\min}(\mathbf{H} \mathbf{M}), \lambda_{\max}(\mathbf{H} \mathbf{M}) \in \mathbb{R}^+ \setminus \{0\}$  denote the extreme eigenvalues of  $\mathbf{H} \mathbf{M}$ , and  $\rho(\cdot)$  denotes the spectral radius of a matrix (the modulus of the eigenvalue of largest modulus). With these assumptions and notation,

$$\begin{aligned} FOV^{\mathbf{H}}(\mathbf{A} \mathbf{H}) &= \left\{ \frac{\langle \mathbf{H} \mathbf{A} \mathbf{H} \mathbf{z}, \mathbf{z} \rangle}{\langle \mathbf{H} \mathbf{z}, \mathbf{z} \rangle}; \mathbf{z} \in \mathbb{C}^n \setminus \{0\} \right\} \\ &\subset \underbrace{\left\{ \frac{\langle \mathbf{H} \mathbf{M} \mathbf{H} \mathbf{z}, \mathbf{z} \rangle}{\langle \mathbf{H} \mathbf{z}, \mathbf{z} \rangle}; \mathbf{z} \in \mathbb{C}^n \setminus \{0\} \right\}}_{\subset \mathbb{R}} + \underbrace{\left\{ \frac{\langle \mathbf{H} \mathbf{N} \mathbf{H} \mathbf{z}, \mathbf{z} \rangle}{\langle \mathbf{H} \mathbf{z}, \mathbf{z} \rangle}; \mathbf{z} \in \mathbb{C}^n \setminus \{0\} \right\}}_{\subset i\mathbb{R}} \\ &= \left\{ \frac{\langle \mathbf{M} \mathbf{z}, \mathbf{z} \rangle}{\langle \mathbf{H}^{-1} \mathbf{z}, \mathbf{z} \rangle}; \mathbf{z} \in \mathbb{C}^n \setminus \{0\} \right\} + \left\{ \frac{\langle \mathbf{N} \mathbf{z}, \mathbf{z} \rangle}{\langle \mathbf{H}^{-1} \mathbf{z}, \mathbf{z} \rangle}; \mathbf{z} \in \mathbb{C}^n \setminus \{0\} \right\}. \end{aligned}$$

Taking  $\mathbf{H}^{1/2}$  to be the square root of  $\mathbf{H}$  and  $\mathbf{y} = \mathbf{H}^{-1/2} \mathbf{z}$ , the fractions can be rewritten as

$$\frac{\langle \mathbf{H}^{1/2} \mathbf{M} \mathbf{H}^{1/2} \mathbf{y}, \mathbf{y} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle} \text{ and } \frac{\langle \mathbf{H}^{1/2} \mathbf{N} \mathbf{H}^{1/2} \mathbf{y}, \mathbf{y} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle},$$

where we recognize Rayleigh quotients for the hpd matrix  $\mathbf{H}^{1/2} \mathbf{M} \mathbf{H}^{1/2}$  and the skew-Hermitian matrix  $\mathbf{H}^{1/2} \mathbf{N} \mathbf{H}^{1/2}$ . Since these matrices are similar, respectively, to  $\mathbf{H} \mathbf{M}$  and  $\mathbf{H} \mathbf{N}$ , it can be concluded that

$$FOV^{\mathbf{H}}(\mathbf{A} \mathbf{H}) \subset [\lambda_{\min}(\mathbf{H} \mathbf{M}), \lambda_{\max}(\mathbf{H} \mathbf{M})] + i[-\rho(\mathbf{H} \mathbf{N}), \rho(\mathbf{H} \mathbf{N})]. \quad (8)$$

We have thus presented a rectangle in  $\mathbb{C}$  that contains the field of values over which the min-max problem is defined. Another rectangle can be obtained by a technique used in our two previous works [36, 39], namely, multiply the fraction in the imaginary term by  $\frac{\langle \mathbf{M} \mathbf{z}, \mathbf{z} \rangle}{\langle \mathbf{M} \mathbf{z}, \mathbf{z} \rangle}$ . Then, for any  $\mathbf{z} \in \mathbb{C}^n \setminus \{0\}$ , by a similar argument

$$\frac{\langle \mathbf{N} \mathbf{z}, \mathbf{z} \rangle}{\langle \mathbf{H}^{-1} \mathbf{z}, \mathbf{z} \rangle} = \frac{\langle \mathbf{N} \mathbf{z}, \mathbf{z} \rangle}{\langle \mathbf{M} \mathbf{z}, \mathbf{z} \rangle} \times \frac{\langle \mathbf{M} \mathbf{z}, \mathbf{z} \rangle}{\langle \mathbf{H}^{-1} \mathbf{z}, \mathbf{z} \rangle} \in i[-\rho(\mathbf{M}^{-1} \mathbf{N}) \lambda_{\max}(\mathbf{H} \mathbf{M}), \rho(\mathbf{M}^{-1} \mathbf{N}) \lambda_{\max}(\mathbf{H} \mathbf{M})]. \quad (9)$$

This gives us the inclusion

$$FOV^{\mathbf{H}}(\mathbf{AH}) \subset [\lambda_{\min}(\mathbf{HM}), \lambda_{\max}(\mathbf{HM})] + i[-\rho(\mathbf{M}^{-1}\mathbf{N})\lambda_{\max}(\mathbf{HM}), \rho(\mathbf{M}^{-1}\mathbf{N})\lambda_{\max}(\mathbf{HM})].$$

It is necessarily less strict than the previous inclusion (8). Indeed, the inclusion  $\frac{\langle \mathbf{N}\mathbf{z}, \mathbf{z} \rangle}{\langle \mathbf{H}^{-1}\mathbf{z}, \mathbf{z} \rangle} \in i[-\rho(\mathbf{HN}), \rho(\mathbf{HN})]$  is tight, whereas (9) is in general not (each term in the product having been bounded independently). We summarize these findings in a theorem.

**Theorem 2.3.** *Consider  $\mathbf{W}$ -weighted GMRES applied to  $\mathbf{Ax} = \mathbf{b}$  and preconditioned by  $(\mathbf{H}_L, \mathbf{H}_R)$ . Under the three conditions that  $\mathbf{A}$  is pd,  $\mathbf{H} := \mathbf{H}_R\mathbf{H}_L$  is hpd and  $\mathbf{H} = \mathbf{H}_L^* \mathbf{W} \mathbf{H}_L$ , the  $k$ -th residual is bounded by*

$$\frac{\|\mathbf{r}_k\|_{\mathbf{H}}}{\|\mathbf{r}_0\|_{\mathbf{H}}} \leq (1 + \sqrt{2}) \min_{q \in \mathbb{P}_k; q(0)=1} \max_{z \in FOV^{\mathbf{H}}(\mathbf{AH})} |q(z)|; \quad FOV^{\mathbf{H}}(\mathbf{AH}) \subset \Omega_1 \subset \Omega_2,$$

where

$$\Omega_1 := [\lambda_{\min}(\mathbf{HM}), \lambda_{\max}(\mathbf{HM})] + i[-\rho(\mathbf{NH}), \rho(\mathbf{NH})]. \quad (10)$$

and

$$\Omega_2 := [\lambda_{\min}(\mathbf{HM}), \lambda_{\max}(\mathbf{HM})] + i[-\rho(\mathbf{M}^{-1}\mathbf{N})\lambda_{\max}(\mathbf{HM}), \rho(\mathbf{M}^{-1}\mathbf{N})\lambda_{\max}(\mathbf{HM})]. \quad (11)$$

(Recall that  $[\lambda_{\min}(\mathbf{HM}), \lambda_{\max}(\mathbf{HM})]$  is a real positive interval that contains all eigenvalues of  $\mathbf{HM}$ , and  $\rho(\cdot)$  denotes the spectral radius of a matrix.)

These bounds depend on

- how well the Hermitian part  $\mathbf{M}$  of  $\mathbf{A}$  is preconditioned by  $\mathbf{H}$  via  $\lambda_{\min}(\mathbf{HM})$  and  $\lambda_{\max}(\mathbf{HM})$ ,
- and
  - either, on how well the skew-Hermitian part  $\mathbf{N}$  of  $\mathbf{A}$  is preconditioned by  $\mathbf{H}$  via  $\rho(\mathbf{NH})$ ,
  - or, on how non-Hermitian the problem is, via,  $\rho(\mathbf{M}^{-1}\mathbf{N})$ .

We defer the solution of the min-max problem on  $\Omega_1$  and  $\Omega_2$  to Section 4. First we consider deflated GMRES.

## 3 Deflated GMRES

### 3.1 The general case

We temporarily relax the assumptions from the previous subsection. Let  $\mathbf{A}$ ,  $\mathbf{H}_R$ ,  $\mathbf{H}_L$  be non-singular  $n \times n$  matrices,  $\mathbf{W}$  be hpd, and  $\mathbf{H} := \mathbf{H}_R\mathbf{H}_L$ .

**Definition 3.1.** *Let  $\mathbf{Y}, \mathbf{Z} \in \mathbb{C}^{n \times m}$  be two full rank matrices. Under the assumption that  $\ker(\mathbf{Y}^*) \cap \text{range}(\mathbf{AZ}) = \{\mathbf{0}\}$ , let*

$$\mathbf{P}_D := \mathbf{I} - \mathbf{AZ}(\mathbf{Y}^*\mathbf{AZ})^{-1}\mathbf{Y}^* \text{ and } \mathbf{Q}_D := \mathbf{I} - \mathbf{Z}(\mathbf{Y}^*\mathbf{AZ})^{-1}\mathbf{Y}^*\mathbf{A}. \quad (12)$$

*These are projection operators called the deflation operators.*

In [34], the projectors  $\mathbf{P}_D$  and  $\mathbf{Q}_D$  are called *sibling projectors*. What is meant is that defining one also defines the other unambiguously. Deflated GMRES is the application of GMRES to the singular system  $\mathbf{P}_D\mathbf{Ax} = \mathbf{P}_D\mathbf{b}$ . Deflation can be applied simultaneously with preconditioning by  $(\mathbf{H}_L, \mathbf{H}_R)$  in which case GMRES is applied to

$$\mathbf{H}_L\mathbf{P}_D\mathbf{AH}_R\mathbf{u} = \mathbf{H}_L\mathbf{P}_D\mathbf{b}; \quad \mathbf{x} = \mathbf{H}_R\mathbf{u}.$$

Assume that  $\mathbf{Y}^*\mathbf{H}^{-1}\mathbf{Z}$  is non-singular, then GMRES does not break down; see [21, Theorem 3.5] for left preconditioning and [39, Theorem 3.2] for right preconditioning. This property

remains true for GMRES in any weighted inner product. Moreover, the residuals of  $\mathbf{W}$ -weighted and  $\mathbf{P}_D$ -deflated GMRES applied to  $\mathbf{Ax} = \mathbf{b}$  with preconditioning by  $(\mathbf{H}_L, \mathbf{H}_R)$  follow the characterization from Theorem 2.1, *i.e.*,  $\mathbf{r}_k = \mathbf{P}_D(\mathbf{b} - \mathbf{Ax}_k)$ , satisfies

$$\begin{aligned} \|\mathbf{H}_L \mathbf{r}_k\|_{\mathbf{W}} &= \min_{q \in \mathbb{P}_k; q(0)=1} \{\|\mathbf{H}_L q(\mathbf{P}_D \mathbf{A} \mathbf{H}) \mathbf{r}_0\|_{\mathbf{W}}\} \\ &= \min_{q \in \mathbb{P}_k; q(0)=1} \{\|q(\mathbf{P}_D \mathbf{A} \mathbf{H}) \mathbf{P}_D \mathbf{r}_0\|_{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L}\} \\ &\leq \min_{q \in \mathbb{P}_k; q(0)=1} \{\|q(\mathbf{P}_D \mathbf{A} \mathbf{H}) \mathbf{P}_D\|_{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L}\} \|\mathbf{r}_0\|_{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L} \\ &= \min_{q \in \mathbb{P}_k; q(0)=1} \{\|q(\mathbf{P}_D \mathbf{A} \mathbf{H} \mathbf{P}_D)\|_{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L}\} \|\mathbf{H}_L \mathbf{r}_0\|_{\mathbf{W}}. \end{aligned}$$

If we applied the Crouzeix-Palencia bound of Theorem 2.3 directly to  $\|q(\mathbf{P}_D \mathbf{A} \mathbf{H} \mathbf{P}_D)\|_{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L}$ , the min-max problem would be posed over the field of values of the singular operator  $\mathbf{P}_D \mathbf{A} \mathbf{H} \mathbf{P}_D$ , which includes 0. This is not useful, because of the constraint that  $q(0) = 1$ . Instead we apply (5) in the Hilbert space  $(\text{range}(\mathbf{P}_D), \langle \cdot, \cdot \rangle_{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L}, \|\cdot\|_{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L})$  to the (bounded linear) operator in  $\text{range}(\mathbf{P}_D)$  represented by the matrix  $\mathbf{P}_D \mathbf{A} \mathbf{H} \mathbf{P}_D$ . This gives, for any polynomial  $q$ , that

$$\|q(\mathbf{P}_D \mathbf{A} \mathbf{H} \mathbf{P}_D)\|_{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L} \leq (1 + \sqrt{2}) \sup_{z \in \text{FOV}^{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L}(\mathbf{P}_D \mathbf{A} \mathbf{H}|_{\text{range}(\mathbf{P}_D)})} |q(z)|,$$

where the  $\mathbf{W}$ -field of values of any matrix  $\mathbf{B} \in \mathbb{C}^{n \times n}$  restricted to  $V \subset \mathbb{C}^n$  is defined by

$$\text{FOV}^{\mathbf{W}}(\mathbf{B}|_V) := \left\{ \frac{\langle \mathbf{B} \mathbf{z}, \mathbf{z} \rangle_{\mathbf{W}}}{\langle \mathbf{z}, \mathbf{z} \rangle_{\mathbf{W}}} ; \mathbf{z} \in V \right\}, \quad \text{for } V \subset \mathbb{C}^n.$$

in agreement with the definition in [22, page 268].

**Theorem 3.1.** *Let  $\mathbf{P}_D$  be defined by (12), and assume that the two following conditions hold*

$$\ker(\mathbf{Y}^*) \cap \text{range}(\mathbf{AZ}) = \{\mathbf{0}\} \text{ and } \ker(\mathbf{Y}^*) \cap \text{range}(\mathbf{H}^{-1} \mathbf{Z}) = \{\mathbf{0}\}. \quad (13)$$

*Then, the  $k$ -th residual of  $\mathbf{W}$ -weighted GMRES applied to  $\mathbf{Ax} = \mathbf{b}$ , preconditioned by  $(\mathbf{H}_L, \mathbf{H}_R)$  and deflated by  $\mathbf{P}_D$  is bounded by*

$$\frac{\|\mathbf{r}_k\|_{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L}}{\|\mathbf{r}_0\|_{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L}} \leq (1 + \sqrt{2}) \min_{q \in \mathbb{P}_k; q(0)=1} \max_{z \in \text{FOV}^{\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L}(\mathbf{P}_D \mathbf{A} \mathbf{H}|_{\text{range}(\mathbf{P}_D)})} \{|q(z)|\},$$

where again  $\mathbf{r}_k = \mathbf{P}_D(\mathbf{b} - \mathbf{Ax}_k)$  and  $\mathbf{H} = \mathbf{H}_R \mathbf{H}_L$ .

It remains to characterize the weighted, preconditioned and deflated field of values.

### 3.2 The case when $\mathbf{A}$ is pd and $\mathbf{H} = \mathbf{H}_L^* \mathbf{W} \mathbf{H}_L$ and $\mathbf{Y} = \mathbf{HAZ}$

This is the counterpart of Section 2.2 for the deflated case. Assume that  $\mathbf{A}$  is pd,  $\mathbf{H} = \mathbf{H}_R \mathbf{H}_L$  is hpd, and that the residual norm which is minimized (also the norm for the FOV of interest) is equal to the combined preconditioner  $\mathbf{H}$ , *i.e.*,  $\mathbf{H}_L^* \mathbf{W} \mathbf{H}_L = \mathbf{H}$ . Under these condition, the min-max problem in Theorem 3.1 is posed over

$$\text{FOV}^{\mathbf{H}}(\mathbf{P}_D \mathbf{A} \mathbf{H}|_{\text{range}(\mathbf{P}_D)}) = \left\{ \frac{\langle \mathbf{H} \mathbf{P}_D \mathbf{A} \mathbf{H} \mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{H} \mathbf{x}, \mathbf{x} \rangle} ; \mathbf{x} \in \text{range}(\mathbf{P}_D) \right\}. \quad (14)$$

As in our previous work [39], we assume that  $\mathbf{Y} = \mathbf{HAZ}$  in order to have a projection  $\mathbf{P}_D$  that is  $\mathbf{H}$ -orthogonal, or equivalently self-adjoint with respect to the  $\mathbf{H}$ -inner product, meaning that  $\mathbf{H} \mathbf{P}_D = \mathbf{P}_D^* \mathbf{H}$ , and then

$$\begin{aligned} \text{FOV}^{\mathbf{H}}(\mathbf{P}_D \mathbf{A} \mathbf{H}|_{\text{range}(\mathbf{P}_D)}) &= \left\{ \frac{\langle \mathbf{H} \mathbf{A} \mathbf{H} \mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{H} \mathbf{x}, \mathbf{x} \rangle} ; \mathbf{x} \in \text{range}(\mathbf{P}_D) \right\} \\ &\subset \underbrace{\left\{ \frac{\langle \mathbf{H} \mathbf{M} \mathbf{H} \mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{H} \mathbf{x}, \mathbf{x} \rangle} ; \mathbf{x} \in \text{range}(\mathbf{P}_D) \right\}}_{\subset \mathbb{R}} + \underbrace{\left\{ \frac{\langle \mathbf{H} \mathbf{N} \mathbf{H} \mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{H} \mathbf{x}, \mathbf{x} \rangle} ; \mathbf{x} \in \text{range}(\mathbf{P}_D) \right\}}_{\subset i\mathbb{R}}. \end{aligned} \quad (15)$$

The plan is to assume that  $\mathbf{H}$  is already a good preconditioner for  $\mathbf{M}$  and use deflation to bound  $\frac{\langle \mathbf{H}\mathbf{N}\mathbf{H}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{H}\mathbf{x}, \mathbf{x} \rangle}$  on  $\text{range}(\mathbf{P}_D)$ .

**Remark 3.1.** Equations (14) and (15) hold without the assumption that  $\mathbf{A}$  is positive definite but the field of values may contain 0, e.g., if both  $\mathbf{A}$  and  $-\mathbf{A}$  are not pd.

**Remark 3.2.** The combination of  $\mathbf{A}$  being pd and  $\mathbf{Y} = \mathbf{H}\mathbf{A}\mathbf{Z}$  ensures that the two conditions (13) in Theorem 3.1 are satisfied as long as  $\mathbf{Y}$  is full rank.

We will now consider two different deflation spaces.

### 3.2.1 Spectral deflation space based on HN

**Theorem 3.2.** Consider  $\mathbf{W}$ -weighted GMRES applied to  $\mathbf{A}\mathbf{x} = \mathbf{b}$ , preconditioned by  $(\mathbf{H}_L, \mathbf{H}_R)$  and deflated by  $\mathbf{P}_D$  (defined by (12)). We make the four assumptions that  $\mathbf{A}$  is pd,  $\mathbf{H} := \mathbf{H}_R\mathbf{H}_L$  is hpd,  $\mathbf{H} = \mathbf{H}_L^* \mathbf{W} \mathbf{H}_L$  and  $\mathbf{Y} = \mathbf{H}\mathbf{A}\mathbf{Z}$ .

Moreover, let  $(\lambda^{(k)}, \mathbf{x}^{(k)})_{1 \leq k \leq n}$  be the eigenpairs of the generalized eigenvalue problem

$$\mathbf{N}\mathbf{x}^{(k)} = \lambda^{(k)} \mathbf{H}^{-1} \mathbf{x}^{(k)}. \quad (16)$$

If, for a given  $\tau > 0$ , the columns of  $\mathbf{Y}$  are set to be the vectors  $\{\mathbf{x}^{(k)}; |\lambda^{(k)}| > \tau\}$ , then the  $k$ -th residual is bounded by

$$\frac{\|\mathbf{r}_k\|_{\mathbf{H}}}{\|\mathbf{r}_0\|_{\mathbf{H}}} \leq (1 + \sqrt{2}) \min_{q \in \mathbb{P}_k; q(0)=1} \max_{z \in \Omega_1^\tau} |q(z)|,$$

where

$$\Omega_1^\tau := [\lambda_{\min}(\mathbf{H}\mathbf{M}), \lambda_{\max}(\mathbf{H}\mathbf{M})] + i[-\tau, \tau]. \quad (17)$$

(Recall that  $[\lambda_{\min}(\mathbf{H}\mathbf{M}), \lambda_{\max}(\mathbf{H}\mathbf{M})]$  is the real positive interval that contains all eigenvalues of  $\mathbf{H}\mathbf{M}$ ).

*Proof.* Since  $\mathbf{N}$  is skew-Hermitian and  $\mathbf{H}^{-1}$  is Hermitian positive definite, the eigenvectors  $\mathbf{x}^{(k)}$  can be chosen to form a  $\mathbf{H}^{-1}$ -orthonormal basis of  $\mathbb{C}^n$  and  $\{\mathbf{x}^{(k)}; |\lambda^{(k)}| > \tau\} \perp^{\mathbf{H}^{-1}} \{\mathbf{x}^{(k)}; |\lambda^{(k)}| \leq \tau\}$  (see [39, Lemma 2.2]). Consequently,  $\text{range}(\mathbf{H}\mathbf{P}_D) = \text{span}\{\mathbf{x}^{(k)}; |\lambda^{(k)}| \leq \tau\}$  since

$$\text{range}(\mathbf{P}_D) = \text{range}(\mathbf{Y})^\perp = \left( \text{span}\{\mathbf{x}^{(k)}; |\lambda^{(k)}| > \tau\} \right)^\perp = \text{span}\{\mathbf{H}^{-1}\mathbf{x}^{(k)}; |\lambda^{(k)}| \leq \tau\}.$$

This way, in the application of Theorem 3.1 with (15), the purely imaginary term is

$$\left\{ \frac{\langle \mathbf{H}\mathbf{N}\mathbf{H}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{H}\mathbf{x}, \mathbf{x} \rangle}; \mathbf{x} \in \text{range}(\mathbf{P}_D) \right\} = \left\{ \frac{\langle \mathbf{N}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{H}^{-1}\mathbf{x}, \mathbf{x} \rangle}; \mathbf{x} \in \text{range}(\mathbf{H}\mathbf{P}_D) \right\} \subset i[-\tau, \tau],$$

and the min-max problem can indeed be solved over  $\Omega_1^\tau$ .  $\square$

**Practical limitation.** We have assumed that  $\mathbf{Y} = \mathbf{H}\mathbf{A}\mathbf{Z}$  in order to make sure (13) holds. Thus, in order to set up  $\mathbf{P}_D$ , we need to compute  $\mathbf{A}^{-1}\mathbf{H}^{-1}\mathbf{Y}$ , or a different basis that spans the same space. By definition of  $\mathbf{Y}$ ,  $\text{range}(\mathbf{H}^{-1}\mathbf{Y}) = \text{range}(\mathbf{N}\mathbf{Y})$ . If  $\mathbf{N}$  is non-singular:  $\mathbf{A}^{-1} = (\mathbf{N}(\mathbf{I} + \mathbf{N}^{-1}\mathbf{M}))^{-1} = (\mathbf{I} + \mathbf{N}^{-1}\mathbf{M})^{-1}\mathbf{N}^{-1}$  so that

$$\mathbf{A}^{-1}\mathbf{H}^{-1} \text{range}(\mathbf{Y}) = (\mathbf{I} + \mathbf{N}^{-1}\mathbf{M})^{-1} \text{range}(\mathbf{Y}).$$

It is not clear that we can efficiently compute the space spanned by the columns of  $\mathbf{Z} = (\mathbf{I} + \mathbf{N}^{-1}\mathbf{M})^{-1}\mathbf{Y}$ . We believe this to be a technical assumption that is not essential for the efficiency of the method. In the numerical results in Section 5 we propose setting  $\mathbf{Z} = \mathbf{Y}$  or  $\mathbf{Z} = \mathbf{N}\mathbf{Y}$ . These choices do not satisfy  $\mathbf{Y} = \mathbf{H}\mathbf{A}\mathbf{Z}$  but they do satisfy the two conditions (13) in Theorem 2.1.

### 3.2.2 Spectral deflation space based on $\mathbf{M}^{-1}\mathbf{N}$

Another option is to compute the spectral deflation space from our previous work [39].

**Theorem 3.3.** *Consider  $\mathbf{W}$ -weighted GMRES applied to  $\mathbf{Ax} = \mathbf{b}$ , preconditioned by  $(\mathbf{H}_L, \mathbf{H}_R)$  and deflated by  $\mathbf{P}_D$  (defined by (12)). We make the four assumptions that  $\mathbf{A}$  is pd,  $\mathbf{H} := \mathbf{H}_R \mathbf{H}_L$  is hpd,  $\mathbf{H} = \mathbf{H}_L^* \mathbf{W} \mathbf{H}_L$  and  $\mathbf{Y} = \mathbf{HAZ}$ . Moreover, let  $(\lambda^{(k)}, \mathbf{x}^{(k)})_{1 \leq k \leq n}$  be the eigenpairs of the generalized eigenvalue problem*

$$\mathbf{N}\mathbf{x}^{(k)} = \lambda^{(k)} \mathbf{M}\mathbf{x}^{(k)}. \quad (18)$$

*If for a given  $\tau > 0$ , the columns of  $\mathbf{Z}$  are set to be the vectors  $\{\mathbf{x}^{(k)}; |\lambda^{(k)}| > \tau\}$ , then the  $k$ -th residual is bounded by*

$$\frac{\|\mathbf{r}_k\|_{\mathbf{H}}}{\|\mathbf{r}_0\|_{\mathbf{H}}} \leq (1 + \sqrt{2}) \min_{q \in \mathbb{P}_k; q(0)=1} \max_{z \in \Omega_2^{\tau}} |q(z)|,$$

where

$$\Omega_2^{\tau} = [\lambda_{\min}(\mathbf{HM}); \lambda_{\max}(\mathbf{HM})] + i\lambda_{\max}(\mathbf{HM})[-\tau; \tau]. \quad (19)$$

*Proof.* Since  $\mathbf{N}$  is skew-Hermitian and  $\mathbf{M}$  is Hermitian positive definite, the eigenvectors  $\mathbf{x}^{(k)}$  can be chosen to form an  $\mathbf{M}$ -orthonormal basis of  $\mathbb{C}^n$  and  $\{\mathbf{x}^{(k)}; |\lambda^{(k)}| > \tau\} \perp^{\mathbf{M}} \{\mathbf{x}^{(k)}; |\lambda^{(k)}| \leq \tau\}$  (see [39, Lemma 2.2]). Moreover, it can be noticed that  $\text{range}(\mathbf{AZ}) \subset \text{range}(\mathbf{MZ}) + \text{range}(\mathbf{NZ}) = \text{range}(\mathbf{MZ})$  because the columns in  $\mathbf{Z}$  are eigenvectors of (18). Consequently,

$$\text{range}(\mathbf{P}_D) = \text{range}(\mathbf{Y})^{\perp} = \text{range}(\mathbf{HAZ})^{\perp},$$

and

$$\text{range}(\mathbf{HP}_D) = \text{range}(\mathbf{AZ})^{\perp} = \text{range}(\mathbf{MZ})^{\perp} = \text{span}\{\mathbf{x}^{(k)}; |\lambda^{(k)}| \leq \tau\}$$

This way, in the application of Theorem 3.1 with (15), the purely imaginary term is

$$\begin{aligned} \left\{ \frac{\langle \mathbf{H}\mathbf{N}\mathbf{H}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{H}\mathbf{x}, \mathbf{x} \rangle}; \mathbf{x} \in \text{range}(\mathbf{P}_D) \right\} &= \left\{ \frac{\langle \mathbf{H}\mathbf{N}\mathbf{H}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{H}\mathbf{M}\mathbf{H}\mathbf{x}, \mathbf{x} \rangle} \frac{\langle \mathbf{H}\mathbf{M}\mathbf{H}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{H}\mathbf{x}, \mathbf{x} \rangle}; \mathbf{x} \in \text{range}(\mathbf{P}_D) \right\} \\ &= \left\{ \frac{\langle \mathbf{N}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{M}\mathbf{x}, \mathbf{x} \rangle} \frac{\langle \mathbf{M}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{H}^{-1}\mathbf{x}, \mathbf{x} \rangle}; \mathbf{x} \in \text{range}(\mathbf{HP}_D) \right\} \\ &\subset i\lambda_{\max}(\mathbf{HM})[-\tau; \tau], \end{aligned}$$

and the min-max problem can indeed be solved over  $\Omega_2^{\tau}$ .  $\square$

## 4 Solution of the min-max problem on a rectangle

In our quest for improved bounds for GMRES, we have arrived four times at the solution of the min-max problem over a rectangle of the complex plane. Namely,  $\Omega_1$  and  $\Omega_2$  in Theorem 2.3,  $\Omega_1^{\tau}$  in Theorem 3.2, and  $\Omega_2^{\tau}$  in Theorem 3.3. In all these cases, the rectangle is both on the right of the origin and symmetric around the real axis. In order to complete our GMRES convergence bounds, we address the solution of the min-max problem on such rectangular domains, *i.e.*, the approximation of

$$K_k(\Omega) := \min_{q \in \mathbb{P}_k; q(0)=1} \max_{z \in \Omega} |q(z)|. \quad (20)$$

A well-known and easy to check property is that  $K_k(\Omega)$  remains constant under dilation of  $\Omega$ :  $K_k(\Omega) = K_k(a\Omega)$  for any  $a \in \mathbb{C} \setminus \{0\}$ . Thus, without loss of generality, we can restrict ourselves to the case where

$$\Omega := [1, \mu] + i[-\rho, \rho], \text{ for } \mu, \rho \in \mathbb{R}^+.$$

We proceed by considering different approaches to find a bound for the min-max problem (20).

**Elman bound.** This is a linear bound first proved in [10, 11]. For a pd matrix  $\mathbf{B}$  in a generic norm, the bound reads

$$K_k(\Omega) \leq \left[ 1 - \left( \frac{d(0, FOV(\mathbf{B}))}{\|\mathbf{B}\|} \right)^2 \right]^{k/2},$$

where  $d(0, FOV(\mathbf{B}))$  is the distance between 0 and the field of values of  $\mathbf{B}$ . In our setting, we do not have an equivalent for  $\|\mathbf{B}\|$ , however according to [26, eq. (5.7.21)], it is bounded from below with respect to  $r(\mathbf{B}) := \max\{|z|; z \in FOV(\mathbf{B})\}$  (called the numerical radius of  $\mathbf{B}$ ) since

$$r(\mathbf{B}) \leq \|\mathbf{B}\| \leq 2r(\mathbf{B}).$$

We normalize  $\Omega$  with  $d(0, FOV(\mathbf{B})) = d(0, \Omega) = 1$  and  $r(\mathbf{B}) = \max\{|z|; z \in \Omega\} = \sqrt{\mu^2 + \rho^2}$  so that

$$K_k(\Omega) \leq \left[ 1 - \frac{1}{4(\mu^2 + \rho^2)} \right]^{k/2} \quad (\text{Elman bound}). \quad (21)$$

**Disk bound.** From [28, Theorem 5.2] another linear bound can be obtained. If  $S$  is a non-empty compact set of  $\mathbb{C}$  such that there exists a disk that contains  $S$  but not 0, then there is a disk  $D(a, r)$  of center  $a$  and radius  $r$  that contains  $S$  while minimizing  $r/|a|$ . In this case  $K_1(S) = r/|a|$ . Furthermore, in [28, (5.5) and (5.6)], the case where  $S$  is a segment is considered. In particular, if  $S = [\alpha, \bar{\alpha}]$ , the optimal disk is centered at  $a = |\alpha|^2/\Re(\alpha)$  and of radius  $r = |a - \alpha| = |a - \bar{\alpha}|$  and  $K_1([\alpha, \bar{\alpha}]) = |\Im(\alpha)|/|\alpha|$ . To come back to our rectangle  $\Omega$ , we first consider the segment  $S = [1 - i\rho, 1 + i\rho]$  (i.e.,  $\alpha = 1 - i\rho$ ) which is the left hand side edge of  $\Omega$ , and obtain

$$a = 1 + \rho^2, \quad r = |a - \alpha| = \rho\sqrt{\rho^2 + 1}, \quad \text{and} \quad K_1([1 + i\rho, 1 - i\rho]) = \frac{\rho}{\sqrt{1 + \rho^2}}.$$

The same polynomial is optimal for the min-max problem posed over any compact set that both contains  $[1 - i\rho, 1 + i\rho]$  and is contained in  $D(a, r)$ . For this reason it also holds that

$$K_k(\Omega) \leq \left[ \frac{\rho}{\sqrt{1 + \rho^2}} \right]^k \quad \text{if } \mu \leq 2a - 1 = 2\rho^2 + 1, \quad (22)$$

where the condition comes from enclosing  $\Omega$  in the disk; see Figure 1–left. We could derive similar results for different values of  $\mu$  and  $\rho$ . We have checked numerically that the value given by (22) matches the solution of the min-max problem over a vertical segment proved in [20, Corollary (2.8)], which is cited in [41]. When it applies ( $\mu \leq 2\rho^2 + 1$ ), the linear bound (22) is sharper than (21). Indeed,

$$1 - \frac{1}{4(\mu^2 + \rho^2)} > 1 - \frac{1}{\mu^2 + \rho^2} = \frac{\mu^2 + \rho^2 - 1}{\mu^2 + \rho^2} > \frac{\rho^2}{1 + \rho^2},$$

since  $z \mapsto (z - 1)/z$  is increasing, and since  $\mu > 1$ ,  $\mu^2 + \rho^2 > 1 + \rho^2$ . Table 1 gives some numerical values. The difference between both bounds is particularly significant when  $\mu$  reaches its maximal admissible value  $2\rho^2 + 1$ .

| $(\mu, \rho)$ | (2, 4) | (33, 4) | (2, 10) | (201, 10)                |
|---------------|--------|---------|---------|--------------------------|
| (21) (Elman)  | 0.9937 | 0.9999  | 0.9988  | $1 - 3.09 \cdot 10^{-6}$ |
| (22) (Disk)   | 0.9701 | 0.9701  | 0.9950  | 0.9950                   |

Table 1: Comparison between the bounds for  $K_1(\Omega)$  given by (21) (Elman) and (22) (Disk). A smaller number corresponds to a better approximation.

**Disk-segment bound.** The rectangle can be enclosed in a disk-segment (as shown in Figure 1–center)<sup>1</sup>

$$\Omega \subset D = \{z \in \mathbb{C}; \Re(z) \geq 1 \text{ and } |z| \leq \sqrt{\mu^2 + \rho^2}\}.$$

An application of [2, Lemma 2.2] (see also the discussion in [13]) gives

$$K_k(\Omega) \leq K_k(D) \leq \min \left\{ 2 + \gamma_\beta, \frac{2}{1 - \gamma_\beta^{k+1}} \right\} \gamma_\beta^k; \quad (23)$$

where  $\beta \in [0, \pi/2]$  and  $\gamma_\beta$  are defined by

$$\cos(\beta) = \frac{1}{\sqrt{\mu^2 + \rho^2}} \text{ and } \gamma_\beta = 2 \sin \left( \frac{\beta}{4 - 2\beta/\pi} \right) < \sin(\beta) < 1.$$

(In [2, Lemma 2.2], it is also proved that  $\gamma_\beta^k \leq K_k(D)$  but this is not a lower bound for  $K_k(\Omega)$ .)

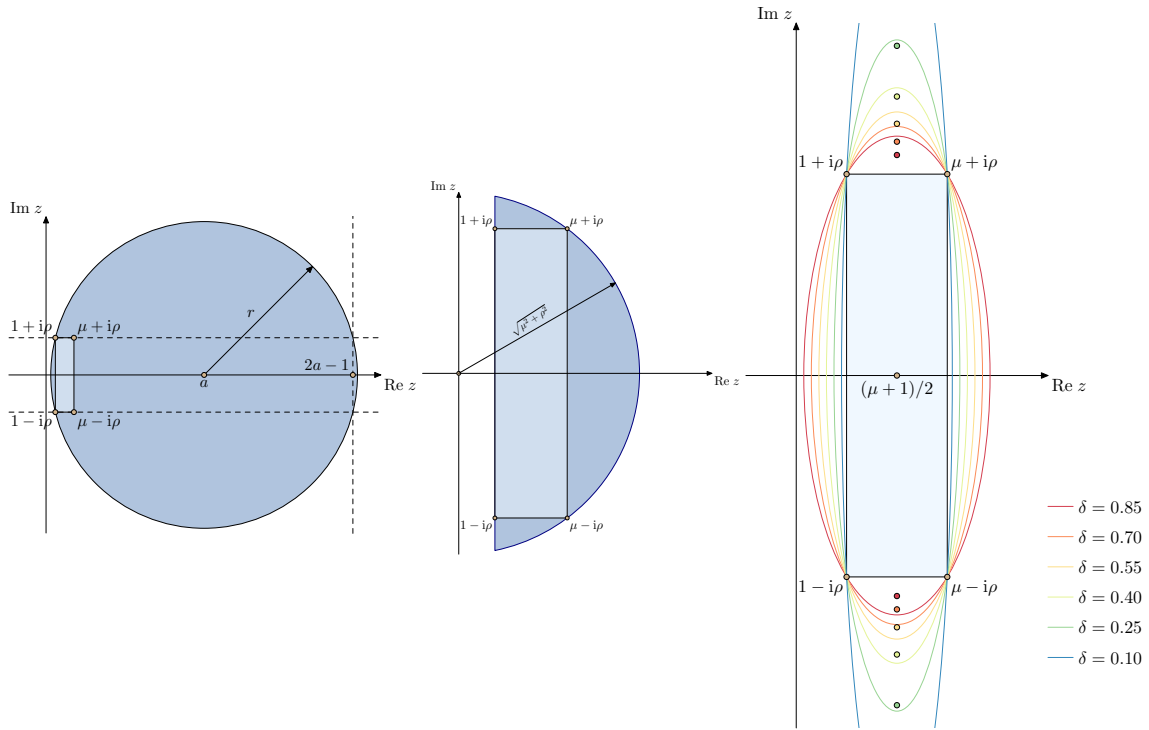


Figure 1:  $\Omega$  is defined by  $\mu = 3$  and  $\rho = 4$ . Left: Disk enclosing  $\Omega$  ( $a = 17$  and  $r = 4\sqrt{17}$ ). Center: Disk-segment circumscribing  $\Omega$ . Right: Family of ellipses  $E(c, d, b)$  circumscribing  $\Omega$ . The dot on the real axis is the center of all ellipses. The other dots are the foci of the ellipses.

**Ellipse bound.** Let  $E$  be an ellipse that circumscribes<sup>2</sup>  $\Omega$ . It is parametrized by  $(c, d, a) \in \mathbb{C}^3$ , with  $c$  its center,  $d$  its focal distance and  $a$  its semi-major axis. For an ellipse  $E$  that

<sup>1</sup>In [2], the field of values is enclosed in a disk-segment obviating the intermediate step of enclosing it in a rectangle. Here, for uniformity in the analysis and comparison we keep the rectangle. We do not expect much difference in the obtained bounds.

<sup>2</sup>We mention that while we consider ellipses bounding rectangles containing the field of values of  $A$ , ellipses bounding rectangles bounding the eigenvalues of  $A$  where studied in [18] in the context of GMRES convergence.

circumscribes our rectangle  $\Omega$ , we first choose  $c = (\mu + 1)/2$  and then determine  $a$  and  $d$ . The Cartesian equation for  $E$  in  $\mathbb{R}^2$  is

$$\frac{(x - c)^2}{\alpha^2} + \frac{y^2}{\beta^2} \leq 1, \text{ for some } \alpha, \beta \in \mathbb{R}^+.$$

A single equation ensures that the vertices  $(c \pm \frac{\mu-1}{2}, \pm\rho)$  of  $\Omega$  (once converted into Cartesian coordinates) are on  $\partial E$ :

$$\frac{(\mu - 1)^2}{4\alpha^2} + \frac{\rho^2}{\beta^2} = 1 \quad \Leftrightarrow \quad \beta = \rho \left[ 1 - \frac{(\mu - 1)^2}{4\alpha^2} \right]^{-1/2} \quad \text{and} \quad \alpha > \frac{\mu - 1}{2}.$$

The circumscribing ellipses are parametrized by  $\alpha > \frac{\mu-1}{2}$ . The origin is outside the ellipse if  $\alpha < \frac{\mu+1}{2}$ . To go back to complex notation, there are two cases. If  $\alpha > \beta$ , the major semi-axis is horizontal so  $a = \alpha$  and  $d = \sqrt{\alpha^2 - \beta^2}$ . If  $\alpha < \beta$ , the major semi-axis is vertical so  $a = i\beta$  and  $d = i\sqrt{\beta^2 - \alpha^2}$ . An illustration of this second case can be seen in Figure 1-right.

The well-known bound  $K_k(E)$  for the min-max problem over the ellipse  $E$  comes from maximizing a (near-optimal) scaled Chebyshev polynomial given *e.g.*, in [31, equation (6.119)]:

$$K_k(E) \leq \frac{C_k\left(\frac{a}{d}\right)}{\left|C_k\left(\frac{c}{d}\right)\right|}, \text{ reached by } \hat{C}_k : z \mapsto \frac{C_k\left(\frac{c-z}{d}\right)}{C_k\left(\frac{c}{d}\right)} \text{ at } z = c + a, \quad (24)$$

where  $C_k$  is the Chebyshev polynomial of the first kind of degree  $k$  and it has implicitly been assumed that  $a/d$  is real, either because  $a$  and  $d$  are both real (*horizontal ellipse*) or because they are both imaginary (*vertical ellipse*). The asymptotic convergence rate, given in [31, (6.121)], is

$$\frac{C_k\left(\frac{a}{d}\right)}{\left|C_k\left(\frac{c}{d}\right)\right|} \approx \left| \frac{a + \sqrt{a^2 - d^2}}{c + \sqrt{c^2 - d^2}} \right|^k. \quad (25)$$

In Figure 2 we plot the asymptotic convergence rate  $|(a + \sqrt{a^2 - d^2})/(c + \sqrt{c^2 - d^2})|$  with respect to  $c - \alpha$ , the distance from the origin to the ellipse, for various choices of  $\Omega$ . Generally speaking, the asymptotic convergence rate deteriorates (*i.e.*, increases) when  $\Omega$  gets larger. We observe that in each case there is an optimal value of  $\alpha$  which minimizes the asymptotic convergence rate. This reflects the fact that there is a trade-off between the ellipse not becoming too tall and it not becoming too close to the origin. In Figure 3, the bound given by (24) is plotted with respect to the polynomial order  $k$  for various choice of ellipses (parametrized by the distance from the ellipse to 0) and two values of  $\rho$ . When  $\Omega$  gets taller, the value of  $K_k(\Omega)$  is reduced at a slower pace when  $k$  grows: for example, at iteration 150, the best bound for  $K_{150}$  is 0.12 when  $\rho = 40$  whereas it is  $2.2 \cdot 10^{-4}$  when  $\rho = 10$ . These best bounds are reached by the ellipse that has the optimal convergence rate. From now on when we consider the bound that comes from enclosing  $\Omega$  in an ellipse, we choose the near-optimal ellipse found by selecting the  $\alpha$  that minimizes the asymptotic convergence rate out of 100 values evenly spread out over  $[(\mu - 1)/2, (\mu + 1)/2]$ .

**Remark 4.1.** Figure 4 shows a contour plot of  $|\hat{C}_k|$  for  $k = 1, 2, 3$  and for the optimal ellipse. The plot is over the rectangle that encloses the ellipse (which in turn encloses  $\Omega$ ). We observe that the functions have largest magnitude toward the corners but these points are not in the ellipse (or in  $\Omega$ ). To improve the bound (24), we also tried numerically maximizing  $|\hat{C}_k|$  over  $\Omega$  instead of over the ellipse. This does not significantly improve the result so we do not report results here. The bound obtained with this technique is larger for  $k = 3$  than for  $k = 2$ . The fact that  $\hat{C}_k$  is not in general optimal goes back to [17].

**Conformal Mapping bound.** In [2, Proof of Lemma 2.2], as a first step in the proof of (23), a lower and upper bound for  $K_k(C)$  is proved for any convex compact set  $C$  such that  $0 \notin C$  that contains at least two elements. Letting  $\phi$  denote the Riemann conformal mapping

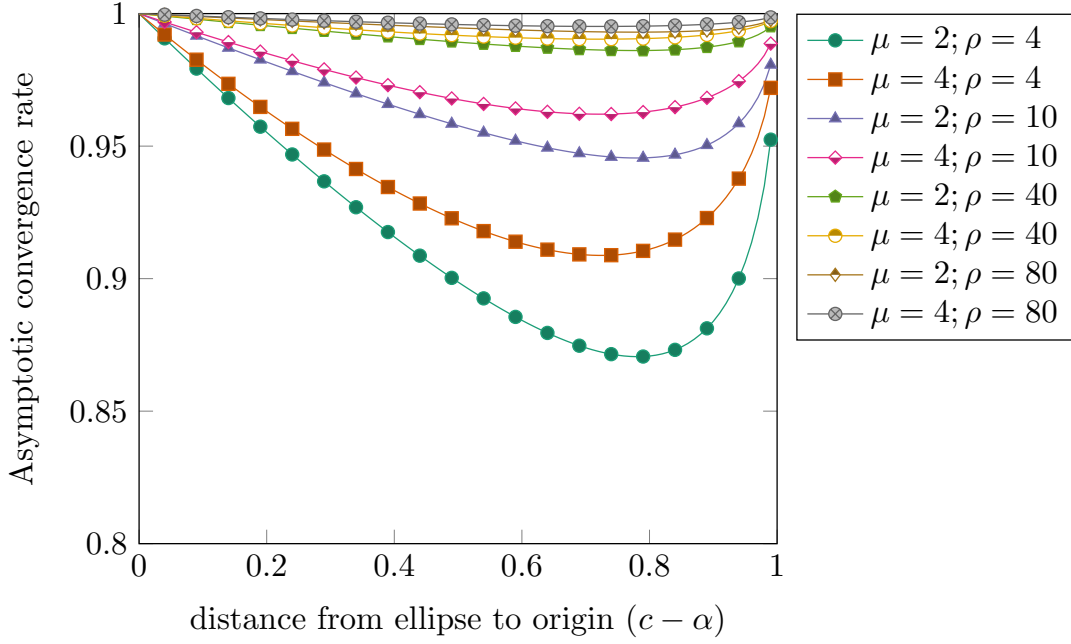
Asymptotic convergence rate for various  $\Omega$  and various ellipses

Figure 2: For various choices of  $\Omega$  parametrized by  $\mu$  and  $\rho$ , asymptotic convergence rate  $|(a + \sqrt{a^2 - d^2})/(c + \sqrt{c^2 - d^2})|$  from (25) with respect to the distance from the ellipse to zero. For each  $\Omega$  there is an optimal ellipse that gives the best (*i.e.*, lowest) convergence rate.

from  $\bar{\mathbb{C}} \setminus C$  onto the exterior of the closed unit disk with  $\phi(\infty) = \infty$ , the result is that, for any  $k \geq 1$ ,

$$\gamma^k \leq K_k(C) \leq \min \left\{ 2 + \gamma, \frac{2}{1 - \gamma^{k+1}} \right\} \gamma^k \leq 3\gamma^k; \text{ where } \gamma := \frac{1}{\phi(0)}. \quad (26)$$

This is a tight bound since a factor less than 3 separates the left- and right- hand sides. For the rectangle  $\Omega$  we have evaluated  $\gamma$  numerically thanks to the (Schwarz-Christoffel) SC-toolbox [8] for Matlab. This takes only four lines of code, as follows.

```
Omega = polygon([1+1i*rho 1-1i*rho mu-1i *rho mu+1i*rho]);
M = extermat(Omega);
invM = inv(M); %Map from the exterior of Omega to the interior of the unit disk
gamma = abs(invM(0))
```

Indeed, `invM` maps to the interior of the unit disk. To map to the exterior of the unit disk and obtain  $\phi$ , `invM` should be composed with  $z \mapsto 1/z$ . This inverse cancels out with the inverse in the definition of  $\gamma$ . Schwarz-Christoffel conformal mapping applies to any polygon  $p$  so this technique is more general than the solution of the min-max problem on a rectangle.

**Faber polynomial bound.** As a final bound we include part of estimate [1, (11)]. For any  $k \geq 1$

$$K_k(C) \leq \frac{2}{|F_k^E(0)|}; \text{ where } F_k^E \text{ is the } k\text{-th Faber polynomial for } C. \quad (27)$$

The Faber polynomial is the polynomial part of the Laurent expansion at infinity of  $\phi^n$ . When  $C$  is polygonal (*e.g.*, rectangular), it can also be computed by the SC-toolbox as follows.

```
F = faber(p,k); %k is the polynomial degree
```

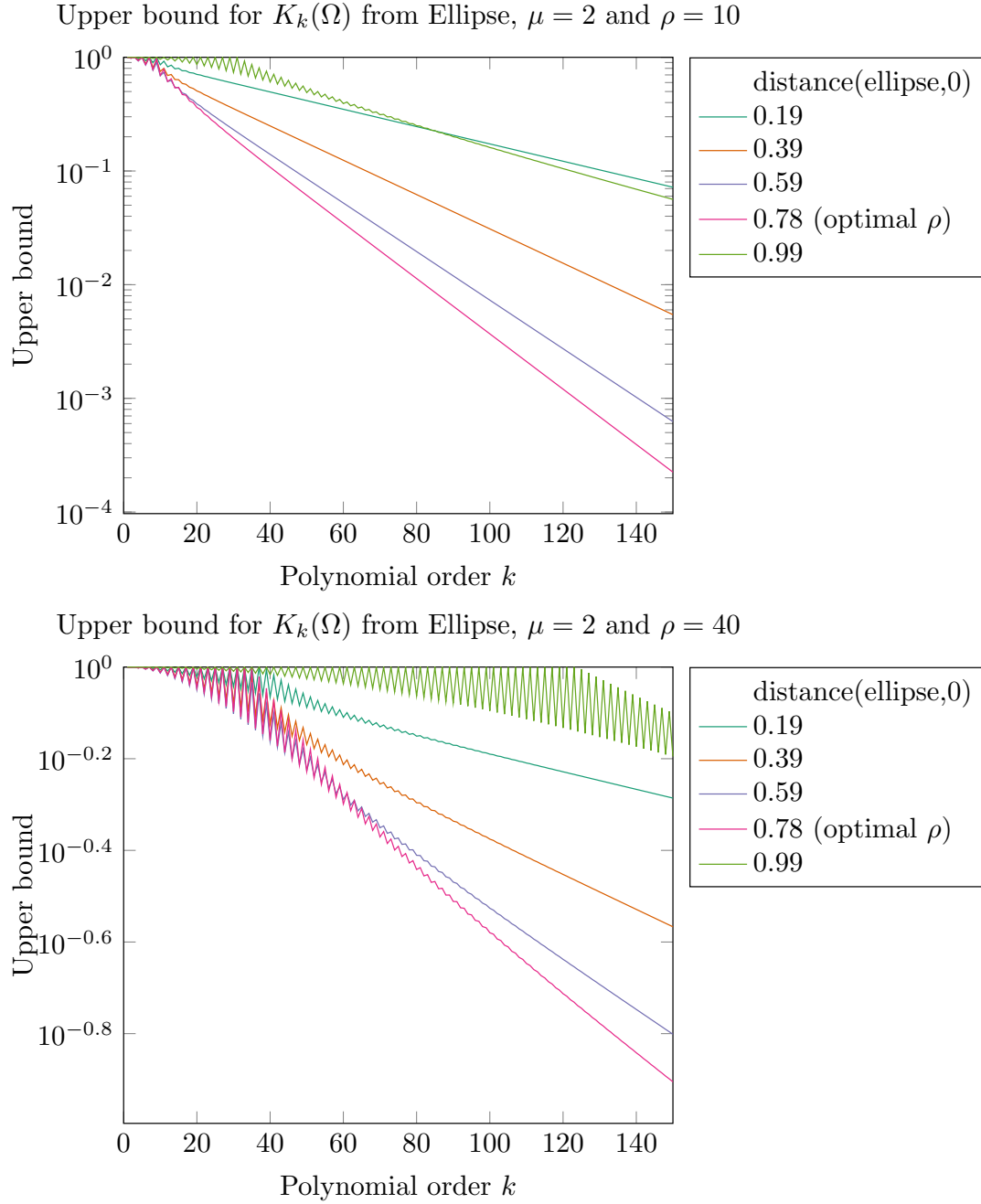


Figure 3: Bound for  $K_k(\Omega)$  computed using (24) for different choices of the enclosing ellipse including the one that gives the optimal convergence rate. Top:  $(\mu, \rho) = (2, 10)$  – Bottom:  $(\mu, \rho) = (2, 40)$ . The vertical axis is not the same for the two plots.

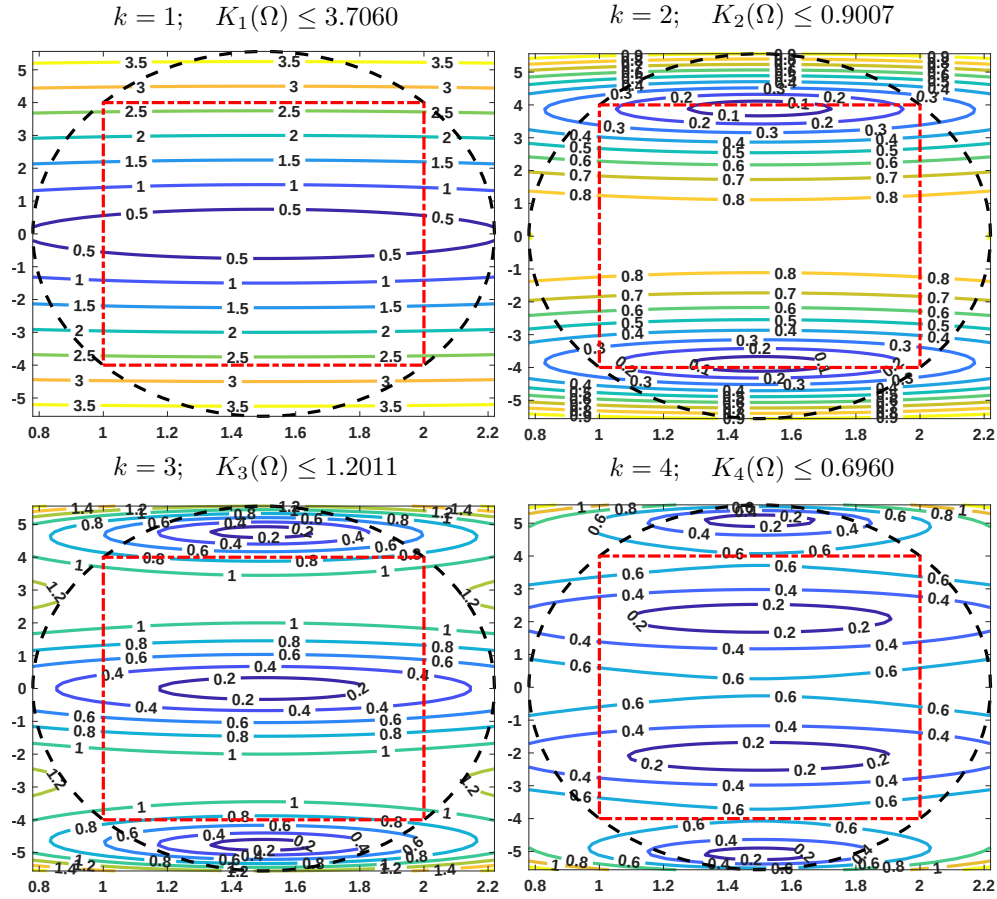


Figure 4: Contour plot (in the imaginary plane) of  $|\hat{C}_k|$ , the modulus of the polynomial from which the bound in (24) is obtained for polynomial orders  $k = 1, 2, 3, 4$ . The red dot-dashed rectangle shows  $\Omega$  (for  $\mu = 2$  and  $\rho = 4$ ). The black dashed ellipse is the optimal ellipse from which  $\hat{C}_k$  is defined. The resulting upper bound for  $K_k(\Omega)$  is also reported. The real and imaginary axes have different scales.

**Comparison between all the bounds.** Figures 5–8 show the values given by all bounds (‘Disk’ (22), ‘Ellipse’ (24), ‘Disk-segment’ (23), ‘Conformal map’ (26) and ‘Faber’ (27)) with respect to the polynomial order  $k$ . If the value returned by a bound is larger than 1, it is set to 1. If the value returned by a bound at polynomial order  $k$  is larger than at  $k - 1$ , we replace it by the value at  $k - 1$ . This way there are no distracting oscillations in the plots.

Each Figure corresponds to a different choice of  $\Omega$ . Figure 5 corresponds to  $\Omega = [1, 2] \times [-4, 4]$ . It is observed that for  $k = 100$  the ‘Conformal map’ and ‘Faber’ bounds give the best estimate, followed by ‘Disk-segment’ and ‘Ellipse’ (which are several orders of magnitude larger). The ‘Disk’ bound is very pessimistic at  $k = 100$ . However, to make this picture more complete, the bottom of Figure 5 shows the same data up to the order  $k = 10$ . It becomes apparent that, the bounds ‘Disk-segment’, ‘Conformal Map’ and ‘Faber’, are lower than 1 only for  $k \geq 5$  or 6. This is a consequence of the multiplicative constant. The bounds ‘Disk’ and ‘Ellipse’ do provide a bound even for these small values of  $k$ . In fact, ‘Ellipse’ is the best bound up to  $k = 9$ .

Figure 6 is for  $\Omega = [1, 33] \times [-4, 4]$ , the behavior is very similar to that observed in Figure 5 except the values of the bounds: when  $\mu = 33$ , the best bound after  $k = 100$  is approximately  $10^{-4}$ , versus  $10^{-9}$  when  $\mu = 2$ .

Figure 7 is for  $\Omega = [1, 2] \times [-10, 10]$ . We can again make similar comments on the comparison between the bounds and the best final value is again approximately  $10^{-4}$ . The ‘Ellipse’ bound is the best for  $k \leq 15$ .

Finally, Figure 8 is for  $\Omega = [1, 2] \times [-40, 40]$ . This time the final value is significantly worse: the best estimate with  $k = 100$  is larger than  $10^{-1}$ . Making the rectangle taller significantly worsens the bound.

The takeaway is that bounds that approximate correctly the asymptotic convergence rate (‘Conformal map’, ‘Faber’) give the best result after a large number of iterations but ‘Disk’ and ‘Ellipse’ are the only ones that are guaranteed to be informative at every iteration (even the first). ‘Disk-segment’ gives the correct convergence rate but on the disk-segment (which is sometimes much larger than  $\Omega$ ). It does have the advantage over ‘Conformal Map’ and ‘Faber’ that the formula is more direct.

## 5 Numerical Results

In this section, the problem considered is the convection-diffusion-reaction problem posed in  $\mathcal{D} = [-1, 1]^2$ . The strong formulation of the problem is:

$$\begin{aligned} c_0 u + \operatorname{div}(\mathbf{a}u) - \operatorname{div}(\nu \nabla u) &= f \text{ in } \mathcal{D}, \\ u &= 0 \text{ on } \partial \mathcal{D}. \end{aligned}$$

The variational formulation is: Find  $u \in H_0^1(\mathcal{D})$  such that

$$\underbrace{\int_{\mathcal{D}} \left( \left( c_0 + \frac{1}{2} \operatorname{div} \mathbf{a} \right) uv + \nu \nabla u \cdot \nabla v \right)}_{\text{symmetric part}} + \underbrace{\int_{\mathcal{D}} \left( \frac{1}{2} \mathbf{a} \cdot \nabla uv - \frac{1}{2} \mathbf{a} \cdot \nabla vu \right)}_{\text{skew-symmetric part}} = \int_{\mathcal{D}} f v,$$

for all  $v \in H_0^1(\mathcal{D})$ . The reaction coefficient  $c_0 > 0$  and viscosity  $\nu > 0$  are assumed to be constant over  $\mathcal{D}$ . The right hand side and convection field are chosen as

$$f(x, y) = \exp(-2.5(x^2 + (y + 0.8)^2)), \text{ and } \mathbf{a}(x, y) = \eta \pi \begin{pmatrix} -y - 0.8 \\ x \end{pmatrix} \text{ with } \eta = 100.$$

It can be remarked that  $\operatorname{div} \mathbf{a} = 0$ . For the parameters in the problem, we set

$$c_0 = 1, \nu = 1, \text{ and } \eta = 100.$$

The problem is discretized by Lagrange  $\mathbb{P}_1$  finite elements on a triangular mesh with 8643 vertices and 16948 triangles (see Figure 9 – left). GMRES is implemented in Octave while the

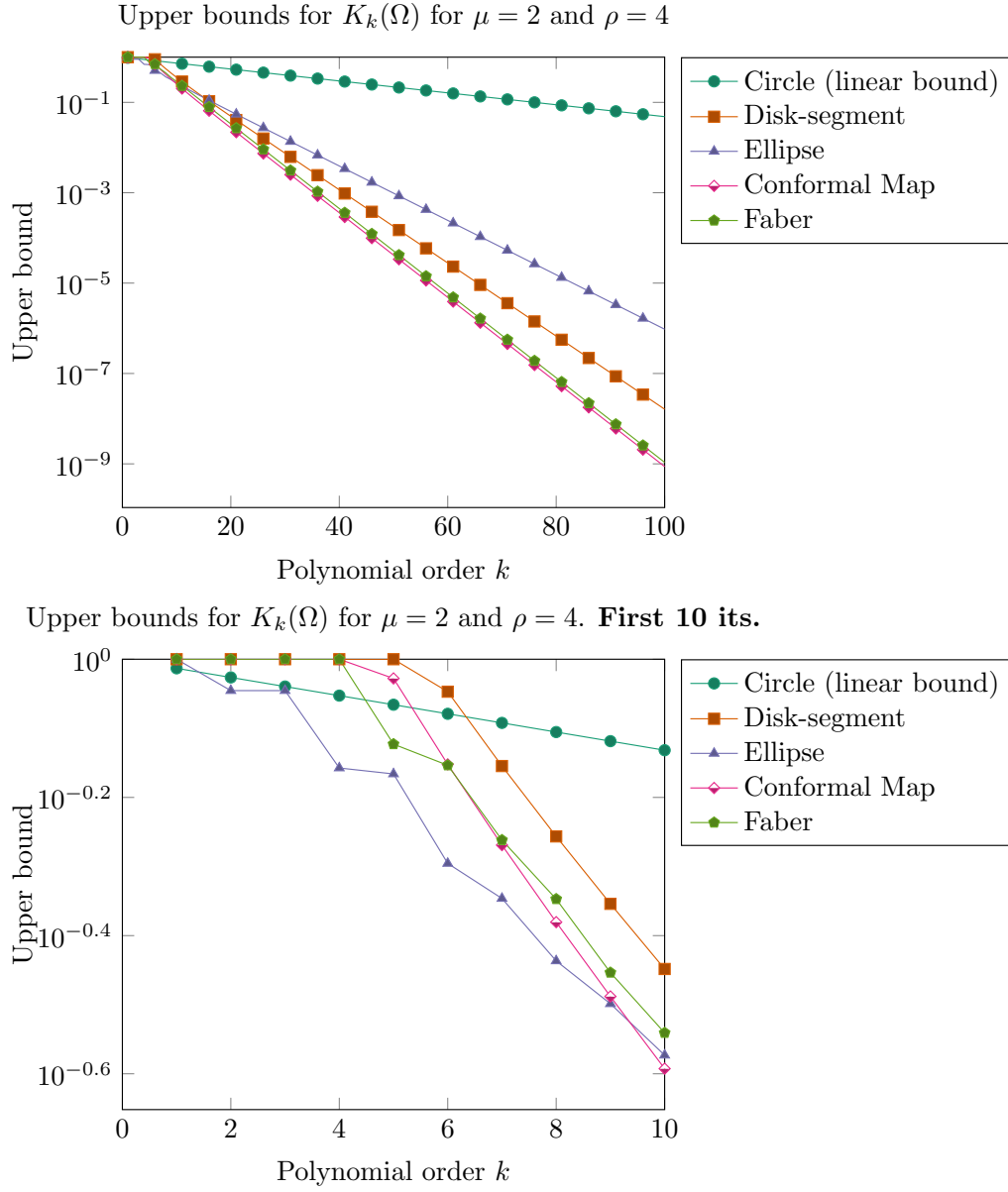
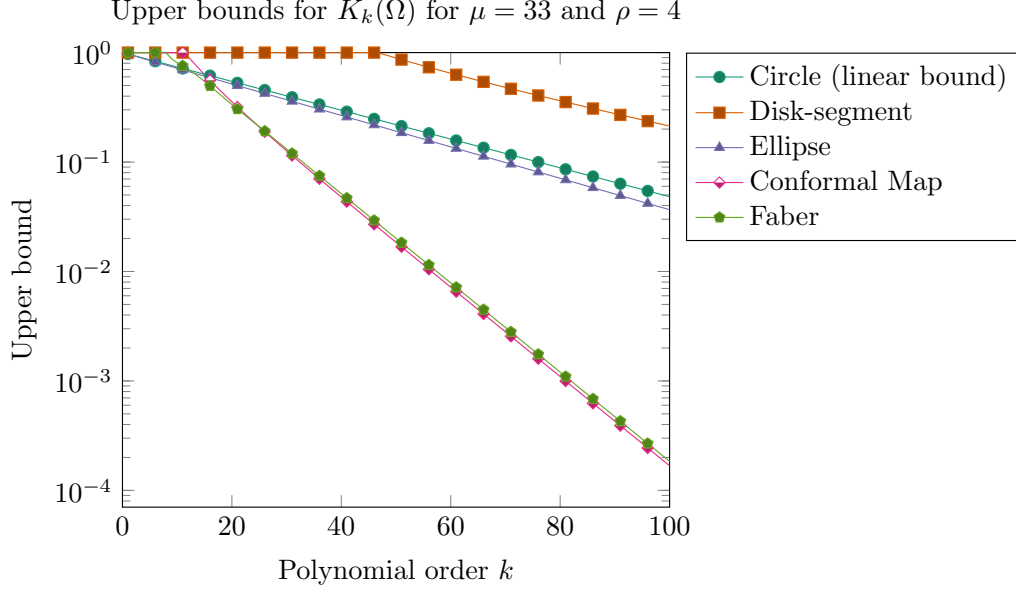


Figure 5: Comparison between all the bounds for  $\mu = 2$  and  $\rho = 4$ . The bottom plot is a zoom on the first 10 iterations.

Figure 6: Comparison between all the bounds for  $\mu = 33$  and  $\rho = 4$ .

finite element matrices and right hand side are assembled by FreeFem++ [25]. All iteration counts correspond to the number of iterations needed to reach  $\|\mathbf{r}_i\|_{\mathbf{H}} < 10^{-10} \|\mathbf{b}\|_{\mathbf{H}}$  starting from a zero initial vector. The Dirichlet boundary condition has been enforced by elimination. Let  $(\phi_i)_{1 \leq i \leq n}$  denote the  $\mathbb{P}_1$  finite element basis corresponding to the mesh. The problem matrix splits into

$$\mathbf{A} = \mathbf{M} + \mathbf{N}, \text{ with } \mathbf{M} \text{ spd and } \mathbf{N} \text{ skew-symmetric,}$$

where the entries of  $\mathbf{M}$  and  $\mathbf{N}$  are

$$\mathbf{M}_{ij} = \int_{\mathcal{D}} (c_0 \phi_i \phi_j + \nu \nabla \phi_i \cdot \nabla \phi_j), \text{ and } \mathbf{N}_{ij} = \eta \int_{\mathcal{D}} \left( \frac{1}{2} \mathbf{a} \cdot \nabla \phi_i \phi_j - \frac{1}{2} \mathbf{a} \cdot \nabla \phi_j \phi_i \right).$$

The positive definiteness of  $\mathbf{M}$  is guaranteed by the assumption that  $c_0$  and  $\nu$  are positive. This is the same setup as in [39], and with a change of variables as in [36].

**Choice of preconditioner and inner product.** As a preconditioner  $\mathbf{H}$ , we select a domain decomposition (DD) preconditioner based on a partition of the mesh into  $N = 16$  subdomains (as shown in Figure 9 (right)). In detail,  $\mathbf{H}_{\text{DD}}$  is the Additive Schwarz domain decomposition method with the GenEO coarse space [37, 38]. The condition number of the resulting preconditioned operator is bounded by

$$\lambda_{\min}(\mathbf{H}\mathbf{M}) \geq \left(1 + \frac{k_0}{v}\right)^{-1}; \quad \lambda_{\max}(\mathbf{H}\mathbf{M}) \leq k_0; \text{ and } \kappa(\mathbf{H}\mathbf{M}) \leq k_0 \left(1 + \frac{k_0}{v}\right),$$

where  $k_0$  ( $= 3$  here) denotes the maximal number of subdomains that each mesh element belongs to [35, Theorem 4.40] and  $v$  is a parameter that has been set to 0.15. The constant in the bound does not depend on the total number  $N$  of subdomains or the mesh parameter  $h$ .

The preconditioner is spd. It is applied on the right ( $\mathbf{H}_R = \mathbf{H}$  and  $\mathbf{H}_L = \mathbf{I}$ ) and the weight is  $\mathbf{W} = \mathbf{H}$ .

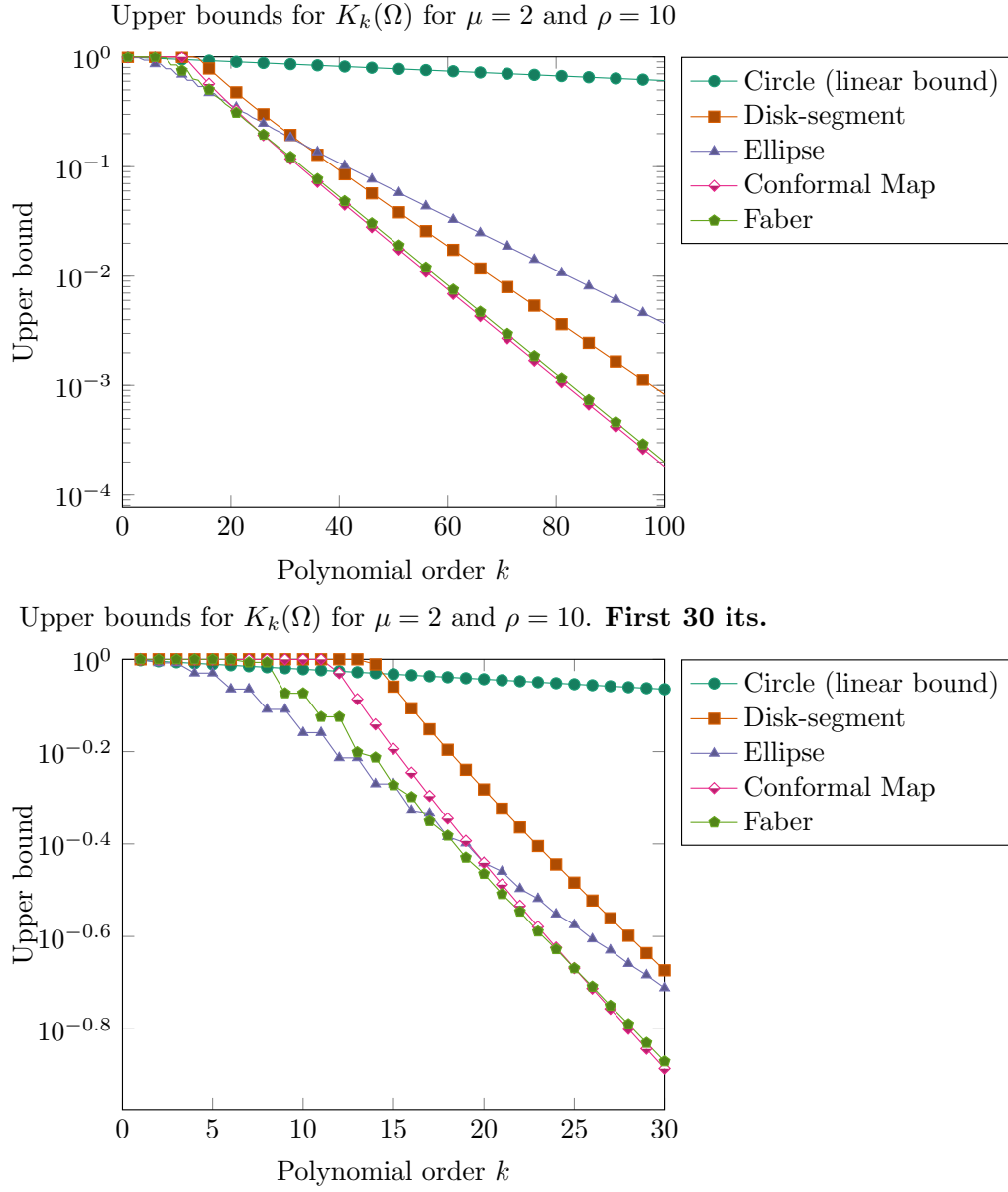


Figure 7: Comparison between all the bounds for  $\mu = 2$  and  $\rho = 10$ . The bottom plot is a zoom on the first 30 iterations.

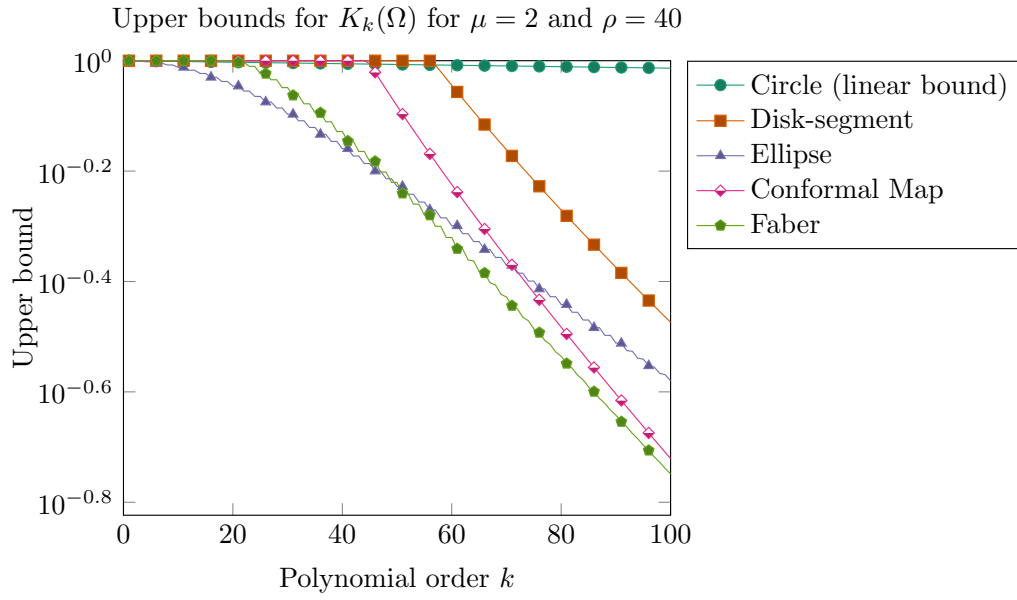


Figure 8: Comparison between all the bounds for  $\mu = 2$  and  $\rho = 40$ . After 100 iterations, the best of the bounds is still above  $10^{-1}$ .

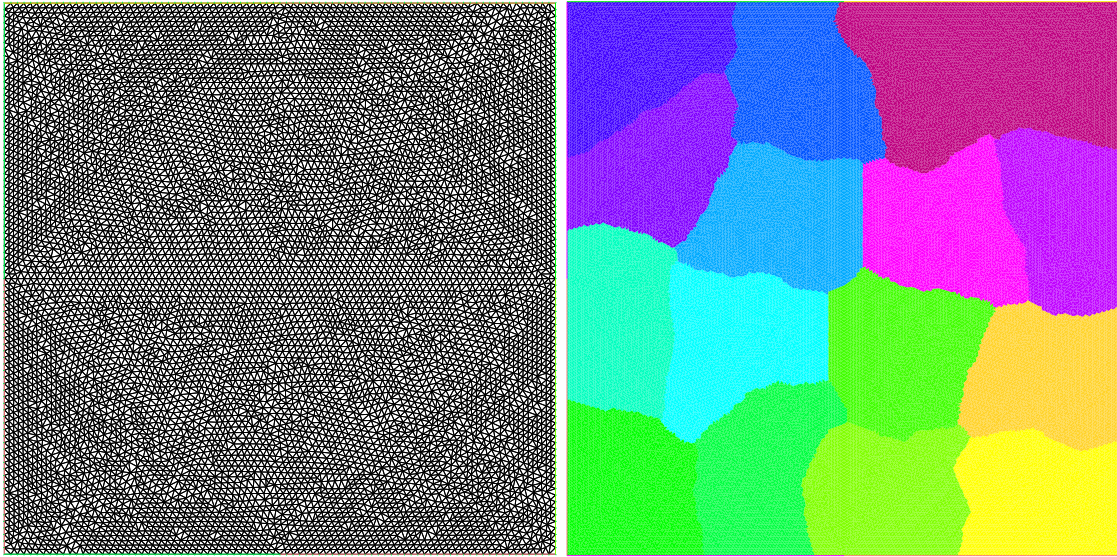


Figure 9: Left: Mesh consisting of 8643 vertices and 16948 triangles. Right: Partition into 16 subdomains.

**Deflation Operator.** We aim to illustrate the convergence results in Theorems 3.2 and 3.3 for the two spectral deflation spaces. Once  $\mathbf{M}$ ,  $\mathbf{N}$  and the domain decomposition operators that go into  $\mathbf{H}$  have been assembled by FreeFem++, they are imported into Octave. The preconditioner  $\mathbf{H}$  is never assembled into a matrix. Then the generalized eigenvalue problem  $\mathbf{N}\mathbf{x} = \lambda\mathbf{M}\mathbf{x}$  or  $\mathbf{N}\mathbf{x} = \lambda\mathbf{H}^{-1}\mathbf{x}$  is partially solved by *eigs*: the eigenpairs corresponding to the eigenvalues of largest magnitude are approximated. Following Definition 3.1, the deflation operator is fully defined by the choice of  $\mathbf{Y}$  and  $\mathbf{Z}$ . We give more detail below.

**First deflation space (Theorem 3.3).** Solve (18) (i.e.,  $\mathbf{N}\mathbf{z}^{(j)} = \lambda_j\mathbf{M}\mathbf{z}^{(j)}$ ) and order the eigenpairs in non-increasing modulus order. Let  $m \in \mathbb{N}$  be the desired dimension of the deflation space. Assuming that  $m$  is even, let  $\mathbf{Z}$  be

$$\mathbf{Z} := \left[ \Re(\mathbf{z}^{(1)}) | \Im(\mathbf{z}^{(1)}) | \Re(\mathbf{z}^{(3)}) | \Im(\mathbf{z}^{(3)}) | \dots | \Re(\mathbf{z}^{(m-1)}) | \Im(\mathbf{z}^{(m-1)}) \right],$$

where it is meant that the vectors are the columns of  $\mathbf{Z}$ . We could equivalently have set  $\mathbf{Z} = \left[ \mathbf{z}^{(1)} | \mathbf{z}^{(2)} | \dots | \mathbf{z}^{(m-1)} | \mathbf{z}^{(m)} \right]$  but then the deflation operator is complex whereas the linear system and preconditioner are real, an unnecessary numerical expense. The reason both choices are equivalent is that  $\mathbf{M}$  is spd and  $\mathbf{N}$  is skew-symmetric so the eigenvalues, which are purely imaginary, come in complex conjugate pairs (with the possible exception being 0), with complex conjugate pairs of eigenvectors. Finally, we choose  $\mathbf{Y} = \mathbf{H}\mathbf{A}\mathbf{Z}$  and this completes the definition of the deflation operator  $\mathbf{P}_D$ .

**Second deflation space (Theorem 3.2).** Solve (16) (i.e.,  $\mathbf{N}\mathbf{x}^{(j)} = \lambda_j\mathbf{H}^{-1}\mathbf{x}^{(j)}$ ) and order the eigenvalues in non-increasing modulus order. Let  $m \in \mathbb{N}$  be the desired dimension of the deflation space. Assuming that  $m$  is even, let the matrix  $\mathbf{Y}$  be

$$\mathbf{Y} = \left[ \Re(\mathbf{x}^{(1)}) | \Im(\mathbf{x}^{(1)}) | \Re(\mathbf{x}^{(3)}) | \Im(\mathbf{x}^{(3)}) | \dots | \Re(\mathbf{x}^{(m-1)}) | \Im(\mathbf{x}^{(m-1)}) \right],$$

where we mean that the vectors are the columns of  $\mathbf{Y}$ . For  $\mathbf{Z}$ , we study three possible choices

- $\mathbf{Z} = \mathbf{A}^{-1}\mathbf{N}\mathbf{Y}$ , so that  $\text{range}(\mathbf{Z}) = \text{range}(\mathbf{A}^{-1}\mathbf{N}\mathbf{Y}) = \text{range}(\mathbf{A}^{-1}\mathbf{H}^{-1}\mathbf{Y})$ . This satisfies the technical assumption  $\mathbf{Y} = \mathbf{H}\mathbf{A}\mathbf{Z}$  but it is unrealistic to apply  $\mathbf{A}^{-1}$ . Note that  $\mathbf{P}_D$  can actually be assembled with  $\mathbf{A}\mathbf{Z} = \mathbf{N}\mathbf{Y}$  but the knowledge of  $\mathbf{Z}$  is needed to compute the solution to the original (non-deflated) problem.
- $\mathbf{Z} = \mathbf{Y}$ .
- $\mathbf{Z} = \mathbf{N}\mathbf{Y}$ .

The last two choices do not satisfy the technical assumption  $\mathbf{Y} = \mathbf{H}\mathbf{A}\mathbf{Z}$  but they are numerically feasible.

**Results.** The spectra of both generalized eigenvalue problems are shown in Figure 10 where the modulus of the first 600 eigenvalues has been plotted. We observe that there is a lot of similarity between both curves. The table in Figure 10 also gives the numerical radii of  $\mathbf{M}^{-1}\mathbf{N}$  and  $\mathbf{H}\mathbf{N}$  as well as some particular eigenvalues that confirm the closeness. These eigenvalues are used in the discussion below.

Figure 11 shows the  $\mathbf{H}$ -weighted spectrum and the  $\mathbf{H}$ -weighted field of values of the preconditioned operator  $\mathbf{A}\mathbf{H}$ . The axes have different scales: the field of values is actually tall and skinny. The rectangle  $\Omega$  that encloses  $\text{FOV}^{\mathbf{H}}(\mathbf{A}\mathbf{H})$  is

$$\Omega_1 = [0.21; 3.00] \times i[-48.9; 48.9] = 0.21 ([1; 14.3] \times i[-233; 233]).$$

This is exactly  $\Omega_1$  from Theorem 2.3 (convergence without deflation).

Following the work in Section 4, the best bound for the min-max problem on  $\Omega_1$  for  $k = 200$  is  $K_{200}(\Omega_1) \leq 0.85$ . The corresponding GMRES residual bound is  $\|\mathbf{r}_{200}\|_{\mathbf{H}}/\|\mathbf{r}_0\|_{\mathbf{H}} \leq (1 + \sqrt{2})K_{200}(\Omega_1) \leq 2.04$ . Unfortunately any bound larger than 1 is not useful. Next we apply Theorem 3.2. If  $m = 100$  vectors of  $\mathbf{H}\mathbf{N}$  are deflated, the bound computed from  $\lambda_{101}$  is

| Theorem 3.2 and deflation of eigenvectors of $\mathbf{HN}$ (with $\mathbf{Y} = \mathbf{HAZ}$ ) |               |                     |               |                      |
|------------------------------------------------------------------------------------------------|---------------|---------------------|---------------|----------------------|
|                                                                                                | iteration 100 |                     | iteration 180 |                      |
|                                                                                                | Bound         | Experiment          | Bound         | Experiment           |
| Without deflation                                                                              | 2.29          | 0.013               | 2.08          | $2.3 \cdot 10^{-4}$  |
| Deflation of $m = 100$ vectors                                                                 | 1.02          | $9.4 \cdot 10^{-4}$ | 0.29          | $7.3 \cdot 10^{-7}$  |
| Deflation of $m = 300$ vectors                                                                 | 0.43          | $7.3 \cdot 10^{-6}$ | 0.061         | $9.7 \cdot 10^{-11}$ |

| Theorem 3.3 and deflation of eigenvectors of $\mathbf{M}^{-1}\mathbf{N}$ |               |                     |               |                      |
|--------------------------------------------------------------------------|---------------|---------------------|---------------|----------------------|
|                                                                          | iteration 100 |                     | iteration 180 |                      |
|                                                                          | Bound         | Experiment          | Bound         | Experiment           |
| Without deflation                                                        | 2.40          | 0.013               | 2.38          | $2.3 \cdot 10^{-4}$  |
| Deflation of $m = 100$ vectors                                           | 2.20          | $1.2 \cdot 10^{-3}$ | 1.98          | $1.0 \cdot 10^{-6}$  |
| Deflation of $m = 300$ vectors                                           | 1.97          | $1.2 \cdot 10^{-5}$ | 0.96          | $2.1 \cdot 10^{-10}$ |

Table 2: Numerical comparison of theoretical bounds and GMRES residuals. ‘Bound’: relative residual bound as given in Theorem 3.2 (top) and Theorem 3.3 (bottom) at iterations 100 and 180. ‘Experiment’: relative residual at iterations 100 and 180 of GMRES without deflation and with deflation of  $m = 100$  eigenvectors and  $m = 300$  eigenvectors of either  $\mathbf{HN}$  (top), or  $\mathbf{M}^{-1}\mathbf{N}$  (bottom).

$K_{200}([0.21; 3.00] \times i[-11.7; 11.7]) \leq 0.07$ . If  $m = 300$  vectors of  $\mathbf{HN}$  are deflated, the bound computed from  $\lambda_{301}$  is  $K_{200}([0.21; 3.00] \times i[-7.1; 7.1]) \leq 0.02$ . It is predicted that convergence improves when more deflation vectors are added. The estimates from Theorem 3.3 for deflation of eigenvectors of  $\mathbf{M}^{-1}\mathbf{N}$  can also be computed in the same way (see below). Without deflation, Theorem 2.3 introduces  $\Omega_2 = [0.21; 3.00] \times 3.00 * i[-64.5; 64.5] = [0.21; 3.00] \times i[-193.5; 193.5]$  for which the bound on the min-max problem resulting from the strategy in Section 4 is  $K_{200}(\Omega_2) \leq 0.98$ . This is worse than previously, as expected, because  $\Omega_2$  from Theorem 3.3 is a superset of  $\Omega_1$  above.

Finally, the convergence of GMRES on the proposed test case is analyzed. A first check is that the residuals satisfy the bounds from Theorems 3.2 and 3.3. Explicit values are given in Table 2. It follows that the achieved residual norms are indeed below the bounds given in Theorems 3.2 and 3.3. It is also observed that the theoretical bounds are not tight compared to the experiment to the extent that they do not give a good approximation of the residual. What the theoretical bounds do indicate is how deflation improves convergence and in this sense the theoretical analysis is successful.

Convergence curves are presented in Figure 12. There, it is confirmed that the spectral deflation that we have proposed significantly accelerates convergence. In fact, without deflation or with  $m = 100$ , there is no convergence in less than 200 iterations, in the sense that the relative residual norm is larger than  $10^{-9}$ , while for  $m = 300$ , convergence is achieved. There is hardly any difference in convergence between deflating eigenvectors of  $\mathbf{M}^{-1}\mathbf{N}$  and deflating eigenvectors of  $\mathbf{HN}$ . There is also very little difference between the three variants in the second case. This is good news: the variants  $\mathbf{Y} = \mathbf{Z}$  and  $\mathbf{NY} = \mathbf{Z}$  which lack the technical assumption  $\mathbf{Y} = \mathbf{HAZ}$  are just as efficient numerically.

## 6 Conclusion

We have presented an analysis of GMRES based on the Crouzeix-Palencia result that the field of values is a  $(1 + \sqrt{2})$  spectral set. The role of (left, right) preconditioning, weighting and deflation has been made explicit in the bounds. Two spectral deflation spaces were studied. Either the high-frequency eigenvectors of  $\mathbf{M}^{-1}\mathbf{N}$  are deflated (as in [39]) or the high-

frequency eigenvectors of  $\mathbf{H}\mathbf{N}$  are deflated. Unless the inverse of  $\mathbf{M}$  is known, we would always recommend the second option. Indeed, with an iterative eigensolver, only applications of  $\mathbf{H}$  and  $\mathbf{N}$  are necessary (whereas with the first choice, the action of  $\mathbf{M}^{-1}$  must be computed, or approximated, many times). Theoretical results and numerical experiments show that deflation of these vectors indeed accelerates convergence of GMRES in terms of iterations.

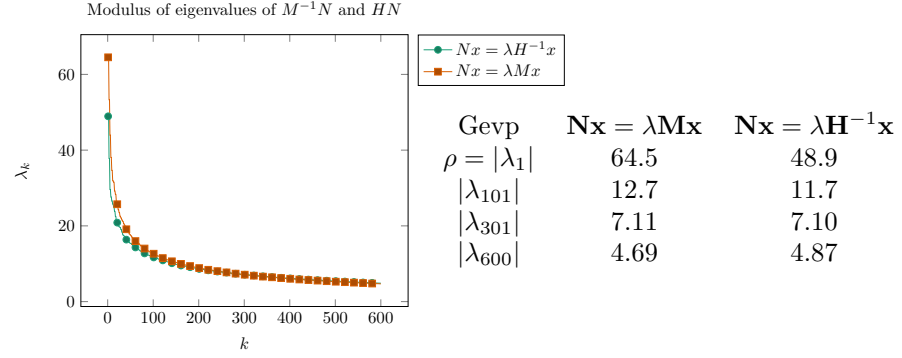


Figure 10: Upper part of the spectrum (first 600 eigenvalues) for the generalized eigenvalue problems (Gevps) that define the two deflation spaces.

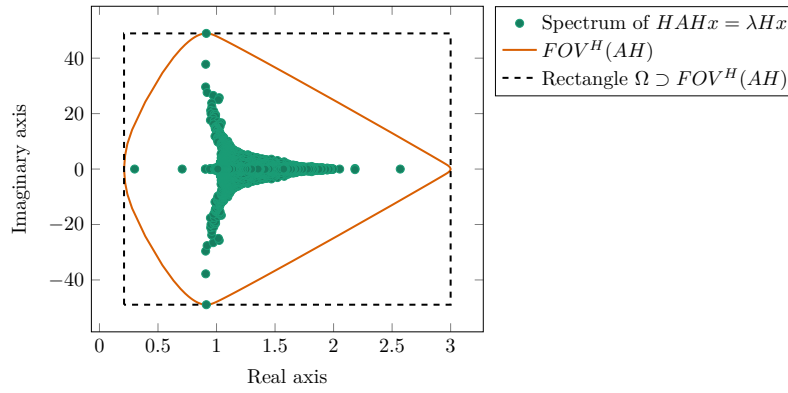
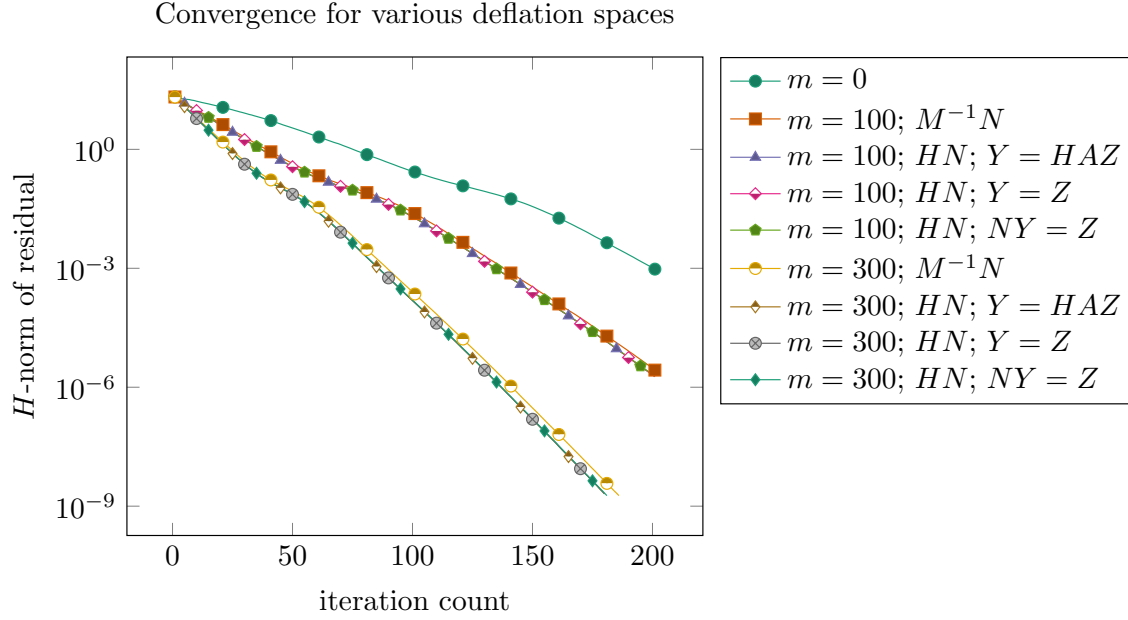


Figure 11:  $\mathbf{H}$ -weighted spectrum and  $\mathbf{H}$ -weighted field of values of the preconditioned operator without any deflation ( $m = 0$ ). (Remark that the axis have different scales.)



Summary of all iteration counts (and final residuals, either at convergence or after 200 iterations)

| $m = 0$                             |                                   | $m = 100$                                          |                                   |                                   |
|-------------------------------------|-----------------------------------|----------------------------------------------------|-----------------------------------|-----------------------------------|
| $\mathbf{Nx} = \lambda \mathbf{Mx}$ |                                   | $\mathbf{Nx} = \lambda \mathbf{H}^{-1} \mathbf{x}$ |                                   |                                   |
|                                     |                                   | $\mathbf{Y} = \mathbf{HAZ}$                        | $\mathbf{Y} = \mathbf{Z}$         | $\mathbf{NY} = \mathbf{Z}$        |
| > 200<br>( $1.02 \cdot 10^{-3}$ )   | > 200<br>( $3.01 \cdot 10^{-6}$ ) | > 200<br>( $2.05 \cdot 10^{-6}$ )                  | > 200<br>( $2.05 \cdot 10^{-6}$ ) | > 200<br>( $2.09 \cdot 10^{-6}$ ) |

| $m = 300$                           |                                                    |                                 |                                 |
|-------------------------------------|----------------------------------------------------|---------------------------------|---------------------------------|
| $\mathbf{Nx} = \lambda \mathbf{Mx}$ | $\mathbf{Nx} = \lambda \mathbf{H}^{-1} \mathbf{x}$ |                                 |                                 |
|                                     | $\mathbf{Y} = \mathbf{HAZ}$                        | $\mathbf{Y} = \mathbf{Z}$       | $\mathbf{NY} = \mathbf{Z}$      |
| 186<br>( $2.14 \cdot 10^{-9}$ )     | 180<br>( $2.04 \cdot 10^{-9}$ )                    | 180<br>( $2.05 \cdot 10^{-9}$ ) | 181<br>( $1.85 \cdot 10^{-9}$ ) |

Figure 12: Convergence without a deflation space ( $m = 0$ ) and with deflation of  $m = 100$  or  $m = 300$ . Both choices of deflation space are considered. In the case where the generalized eigenvalue problem is  $\mathbf{Nx} = \lambda \mathbf{H}^{-1} \mathbf{x}$ , all three variants for defining  $\mathbf{Z}$  with respect to  $\mathbf{Y}$  are considered (and these curves are hardly distinguishable). The table summarizes the number of iterations needed to reduce the residual by a factor  $10^{10}$  and, in parenthesis, either the converged residual or the residual after 200 iterations when convergence is not yet achieved.

## Acknowledgements

The authors thank Nick Trefethen for his help with the SC-toolbox for computing the convergence bounds based on Conformal mapping and Faber polynomials. They also thank the two referees for their careful reading of the manuscript and their many comments and suggestions which helped improved the presentation.

## References

- [1] B. Beckermann. Numerical range, GMRES and Faber polynomials. *C. R., Math., Acad. Sci. Paris*, 340:855–860, 2005.
- [2] B. Beckermann, S. A. Goreinov, and E. E. Tyrtyshnikov. Some remarks on the Elman estimate for GMRES. *SIAM J. Matrix Anal. Appl.*, 27:772–778, 2006.
- [3] T. F. Chan, E. Chow, Y. Saad, and M. C. Yeung. Preserving symmetry in preconditioned Krylov subspace methods. *SIAM J. Sci. Comput.*, 20:568–581, 1999.
- [4] M. Crouzeix and A. Greenbaum. Spectral sets: numerical range and beyond. *SIAM J. Matrix Anal. Appl.*, 40:1087–1101, 2019.
- [5] M. Crouzeix and C. Palencia. The numerical range is a  $(1 + \sqrt{2})$ -spectral set. *SIAM J. Matrix Anal. Appl.*, 38:649–655, 2017.
- [6] C. Davis. The Hausdorff-Toeplitz theorem explained. *Can. Math. Bull.*, 14:245–246, 1971.
- [7] T. A. Driscoll, K.-C. Toh, and L. N. Trefethen. From potential theory to matrix iterations in six steps. *SIAM Rev.*, 40:547–578, 1998.
- [8] T. A. Driscoll and L. N. Trefethen. *Schwarz-Christoffel mapping*, volume 8 of *Camb. Monogr. Appl. Comput. Math.* Cambridge: Cambridge University Press, 2002.
- [9] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš. Numerical stability of GMRES. *BIT*, 35:309–330, 1995.
- [10] S. C. Eisenstat, H. C. Elman, and M. H. Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.*, 20:345–357, 1983.
- [11] H. C. Elman. *Iterative methods for large, sparse, nonsymmetric systems of linear equations*. PhD thesis, Department of Computer Science, Yale University, 1982.
- [12] M. Embree. How Descriptive are GMRES Convergence Bounds? Preprint, arXiv:2209.01231 [math.NA] (2022), 2022.
- [13] M. Embree. Extending Elman’s Bound for GMRES. *Linear Algebra Appl.*, 276:54–70, 2025.
- [14] M. Embree, J. A. Henningsen, J. Jackson, and R. B. Morgan. Polynomial Approximation to the Inverse of a Large Matrix. Preprint, arXiv:2502.18317 [math.NA] (2025), 2025.
- [15] M. Embree, R. B. Morgan, and H. V. Nguyen. Weighted inner products for GMRES and GMRES-DR. *SIAM J. Sci. Comput.*, 39:S610–S632, 2017.
- [16] A. Essai. Weighted FOM and GMRES for solving nonsymmetric linear systems. *Numer. Algorithms*, 18:277–292, 1998.
- [17] B. Fischer and R. Freund. Chebyshev polynomials are not always optimal. *J. Approx. Theory*, 65(3):261–272, 1991.
- [18] B. Fischer, A. Ramage, D. Silvester, and A. Wathen. On parameter choice and iterative convergence for stabilised discretisations of advection–diffusion problems. *Computer Methods Appl. Mech. Eng.*, 179:179–195, 1999.
- [19] J. Frank and C. Vuik. On the construction of deflation-based preconditioners. *SIAM J. Sci. Comput.*, 23:442–462, 2001.

- [20] R. Freund and S. Ruscheweyh. On a class of Chebyshev approximation problems which arise in connection with a conjugate gradient type method. *Numer. Math.*, 48:525–542, 1986.
- [21] L. García Ramos, R. Kehl, and R. Nabben. Projections, deflation, and multigrid for nonsymmetric matrices. *SIAM J. Matrix Anal. Appl.*, 41:83–105, 2020.
- [22] L. García Ramos and R. Nabben. On the spectrum of deflated matrices with applications to the deflated shifted Laplace preconditioner for the Helmholtz equation. *SIAM J. Matrix Anal. Appl.*, 39:262–286, 2018.
- [23] A. Gaul, M. H. Gutknecht, J. Liesen, and R. Nabben. A framework for deflated and augmented Krylov subspace methods. *SIAM J. Matrix Anal. Appl.*, 34:495–518, 2013.
- [24] S. Güttel and J. Pestana. Some observations on weighted GMRES. *Numer. Algorithms*, 67:733–752, 2014.
- [25] F. Hecht. New development in FreeFem++. *J. Numer. Math.*, 20:251–265, 2012.
- [26] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge: Cambridge University Press, 2nd ed. edition, 2013.
- [27] R. B. Morgan, T. Whyte, W. Wilcox, and Z. Yang. Two-Grid Deflated Krylov Methods for Linear Equations. *Electr. Trans. Numer. Anal.*, 63:129–149, 2025.
- [28] G. Opfer and G. Schober. Richardson’s iteration for nonsymmetric matrices. *Linear Algebra Appl.*, 58:343–361, 1984.
- [29] C. C. Paige, M. Rozložník, and Z. Strakoš. Modified Gram-Schmidt (MGS), least squares, and backward stability of MGS-GMRES. *SIAM J. Matrix Anal. Appl.*, 28:264–284, 2006.
- [30] J. Pestana and A. J. Wathen. On the choice of preconditioner for minimum residual methods for non-Hermitian matrices. *J. Comput. Appl. Math.*, 249:57–68, 2013.
- [31] Y. Saad. *Iterative methods for sparse linear systems*. Philadelphia, PA: SIAM Society for Industrial and Applied Mathematics, 2nd ed. edition, 2003.
- [32] J. A. Sifuentes, M. Embree, and R. B. Morgan. GMRES convergence for perturbed coefficient matrices, with application to approximate deflation preconditioning. *SIAM J. Matrix Anal. Appl.*, 34:1066–1088, 2013.
- [33] 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.
- [34] K. M. Soodhalter, E. de Sturler, and M. E. Kilmer. A survey of subspace recycling iterative methods. *GAMM-Mitteilungen*, 43(4):e202000016, 2020.
- [35] N. Spillane. *Robust domain decomposition methods for symmetric positive definite problems*. PhD thesis, Université Pierre et Marie Curie, Paris, France, 2014.
- [36] N. Spillane. Hermitian preconditioning for a class of non-Hermitian linear systems. *SIAM J. Sci. Comput.*, 46(3):A1903–A1922, 2024.
- [37] N. Spillane, V. Dolean, P. Hauret, F. Nataf, C. Pechstein, and R. Scheichl. A robust two-level domain decomposition preconditioner for systems of PDEs. *C. R. Math. Acad. Sci. Paris*, 349:1255–1259, 2011.
- [38] N. Spillane, V. Dolean, P. Hauret, F. Nataf, C. Pechstein, and R. Scheichl. Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps. *Numer. Math.*, 126:741–770, 2014.
- [39] N. Spillane and D. B. Szyld. New convergence analysis of GMRES with weighted norms, preconditioning, and deflation, leading to a new deflation space. *SIAM J. Matrix Anal. Appl.*, 45:1721–1745, 2024.
- [40] G. Starke. Field-of-values analysis of preconditioned iterative methods for nonsymmetric elliptic problems. *Numer. Math.*, 78:103–117, 1997.

- [41] D. B. Szyld and O. B. Widlund. Variational analysis of some conjugate gradient methods. *East-West J. Numer. Math.*, 1:51–74, 1993.
- [42] J. M. Tang, R. Nabben, C. Vuik, and Y. A. Erlangga. Comparison of two-level preconditioners derived from deflation, domain decomposition and multigrid methods. *J. Sci. Comput.*, 39:340–370, 2009.
- [43] K.-C. Toh. GMRES vs. ideal GMRES. *SIAM J. Matrix Anal. Appl.*, 18:30–36, 1997.