Upperbounds for the GMRES residuals for some CFD - related problems

Jan Valdman
under the supervision of Prof. Henk A. van der Vorst

10.June 1997, Nijmegen, the Netherlands

Abstract

Here we show that the discretization and linearization of Navier - Stokes equations leads to an indefinite linear system of equations. Such a system can be solved and preconditioned by iterative methods working on Krylov subspaces. The rate of convergence of one of these methods, GMRES, is estimated by a max norm of a minimal polynomial corresponding to a domain containing the spectrum of the preconditioned operator. For the case of the real spectrum with both positive and negative eigenvalues we introduce an approach that gives an upper bound close to that for the minimal polynomial.

1 The formulation of the Navier-Stokes equations

Our first task is to derive the Navier-Stokes equations. This can be done by using some basic continuum mechanics principles, and making concrete for the case of a fluid (see [7] for more details).

The first principle is the conservation of mass, which can be expressed by:

$$\frac{D\rho}{Dt} + \rho\nabla \cdot \mathbf{u} = \mathbf{0}, \tag{1}$$

where $\rho = \rho(\mathbf{x}, t)$ is the density of the fluid, t the time, $\mathbf{u} = \mathbf{u}(\mathbf{x}, \mathbf{t})$ the velocity vector and

$$\frac{D\rho}{Dt} = \frac{\partial \rho}{\partial t} + \nabla \rho \cdot \mathbf{u}, \tag{2}$$

is the material derivate of ρ.

The second principle is the balance of momentum (also equation of the motion)

$$\rho \frac{D\mathbf{u}}{Dt} = \rho \mathbf{f} + \nabla \cdot \sigma, \tag{3}$$

where σ is the stress tensor (a matrix) and $\mathbf{f} = \mathbf{f}(\mathbf{x}, t)$ means external forces (the body forces per unit mass, for instance the gravity force).

In order to take the friction in fluid into account we assume a Newtonian fluid (otherwise we would obtain the perfect fluid equation):

$$\sigma = -p\mathbf{I} + 2\mu \left(\mathbf{D} - \frac{1}{3}(\nabla \cdot \mathbf{u})\mathbf{I} \right), \tag{4}$$

with

$$D_{i,j} = \frac{1}{2}(\frac{\partial u_i}{\partial y_j} + \frac{\partial u_j}{\partial y_i}), \tag{5}$$

$\mathbf{D}$ is the symmetrical part of the velocity gradient tensor, and μ is the fluid viscosity.

Substitution of (4) into (3) leads to the Navier-Stokes equations for a compressible fluid.

$$\begin{array}{c} \rho \frac{D\mathbf{u}}{Dt} = \rho \mathbf{f} - \nabla p + \mu \left[\triangle \mathbf{u} + \frac{1}{3} \nabla (\nabla \cdot \mathbf{u}) \right] \\ \frac{D\rho}{Dt} + \rho \nabla \cdot \mathbf{u} = \mathbf{0}. \end{array} \tag{6}$$

These can be again simplified for the case of an incompressible fluid ($\frac{D\rho}{Dt} = 0$)and considering the steady phase of the fluid:

$$\begin{array}{rl} \rho \mathbf{u} \cdot \nabla \mathbf{u} & = \rho \mathbf{f} - \nabla p + \mu \triangle \mathbf{u}, \\ \nabla \cdot \mathbf{u} & = \mathbf{0}. \end{array} \tag{7}$$

The Navier - Stokes equation can be written in dimensionless form by scaling with characteristic quantities U (characteristic velocity) and L (characteristic length) in the following way:

$$\acute{x_i} = \frac{x_i}{L}, \;\; \acute{u_i} = \frac{u_i}{U}, \;\; \acute{p} = \frac{p}{\rho U^2}, \;\; \acute{f_i} = \frac{f_i}{(U^2/L)} \tag{8}$$

Substituting of these quantities in the equation (7) and introducing the dimensionless constant:

$$\nu = \frac{\mu}{\rho U L}, \tag{9}$$

results in the dimensionless form of the Navier-Stokes equations (accents are dropped):

$$\begin{array}{c} \mathbf{u} \cdot \nabla \mathbf{u} - \nu \triangle \mathbf{u} + \nabla p = \mathbf{f} \\ \nabla \cdot \mathbf{u} = 0 \end{array} \tag{10}$$

2 The Oseen problem

Since we have a system of partial differential equations, we have to add boundary conditions. We will restrict ourselves to Dirichlet boundary conditions, in particular homogeneous. Its effect will be that, when using a variational formulation (Galerkin method), boundary integrals vanish (See chapter 3).

$$\begin{array}{c} -\nu \triangle \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \nabla p = \mathbf{f} \\ \nabla \cdot \mathbf{u} = \mathbf{0} \end{array} \quad in \;\; \Omega, \tag{11}$$

subject to the boundary conditions on $\partial\Omega$; $\Omega \subset \mathbf{R^3}$ or ; $\Omega \subset \mathbf{R^3}$.

Since system (11) is nonlinear, we apply a fixed point (or Picard iteration), which reduces it to solving a sequence of linear Oseen problems of the form: given some (divergence free) velocity field, find the velocity $\mathbf{u}$ and pressure satisfying:

$$\begin{array}{c} -\nu \triangle \mathbf{u} + \mathbf{w} \cdot \nabla \mathbf{u} + \nabla p = \mathbf{f} \\ \nabla \cdot \mathbf{u} = 0 \end{array} \quad in \;\; \Omega, \tag{12}$$

with the same boundary conditions as for (11).

3 Obtaining the discrete version - FEM discretization

Rewriting (12) component-wise, leads to 4 equations for 4 unknown functions u_1, u_2, u_3, p:

$$\begin{aligned} -\nu\triangle u_1 + \mathbf{w}\cdot\nabla u_1 + \tfrac{\partial p}{\partial x_1} &= f_1, \\ -\nu\triangle u_2 + \mathbf{w}\cdot\nabla u_2 + \tfrac{\partial p}{\partial x_2} &= f_2, \\ -\nu\triangle u_3 + \mathbf{w}\cdot\nabla u_3 + \tfrac{\partial p}{\partial x_3} &= f_3, \\ \tfrac{\partial u_1}{\partial x_1} + \tfrac{\partial u_2}{\partial x_2} + \tfrac{\partial u_3}{\partial x_3} &= 0. \end{aligned} \tag{13}$$

Now we multiply these equations by arbitrary test functions, integrate over the domain Ω, apply partial integration (Gauss theorem, Green theorem) if necessary, and finally substitute the boundary conditions. If we choose the basis functions of the approximations as test functions, then we end up with the Galerkin equations. As test functions in first instance v_1, v_2, v_3 and q are chosen. Because of the boundary conditions for u_1, u_2, u_3, we assume that:

$$v_1 = v_2 = v_3 = 0 \ \ on \ \ \Gamma. \tag{14}$$

The resulting equations are:

$$\begin{aligned} \textstyle\int_\Omega (-\nu\triangle u_1 + \mathbf{w}\cdot\nabla u_1 + \tfrac{\partial p}{\partial x_1}) v_1 d\Omega &= \textstyle\int_\Omega f_1 v_1 d\Omega, \\ \textstyle\int_\Omega (-\nu\triangle u_2 + \mathbf{w}\cdot\nabla u_2 + \tfrac{\partial p}{\partial x_2}) v_2 d\Omega &= \textstyle\int_\Omega f_2 v_2 d\Omega, \\ \textstyle\int_\Omega (-\nu\triangle u_3 + \mathbf{w}\cdot\nabla u_2 + \tfrac{\partial p}{\partial x_3}) v_3 d\Omega &= \textstyle\int_\Omega f_3 v_3 d\Omega, \\ \textstyle\int_\Omega (\tfrac{\partial u_1}{\partial x_1} + \tfrac{\partial u_2}{\partial x_2} + \tfrac{\partial u_3}{\partial x_3}) q d\Omega &= 0. \end{aligned} \tag{15}$$

We now apply partial integration to the (15), and use the relations:

$$v_i \triangle u_i = div(v_i \nabla u_i) - \nabla u_i \cdot \nabla v_i,$$
$$v_i \frac{\partial p}{\partial x_i} = \frac{\partial p v_i}{\partial x_i} - p \frac{\partial v_i}{\partial x_i}, \tag{16}$$

and together with Gauss theorem, we get:

$$\begin{aligned}
\int_\Omega [\nu \nabla u_1 \cdot \nabla v_1 + (\mathbf{w} \cdot \nabla u_1) v_1 - p \tfrac{\partial v_1}{\partial x_1}] d\Omega &= \int_\Omega f_1 v_1 d\Omega, \\
\int_\Omega [\nu \nabla u_2 \cdot \nabla v_2 + (\mathbf{w} \cdot \nabla u_2) v_2 - p \tfrac{\partial v_2}{\partial x_2}] d\Omega &= \int_\Omega f_2 v_2 d\Omega, \\
\int_\Omega [\nu \nabla u_3 \cdot \nabla v_3 + (\mathbf{w} \cdot \nabla u_3) v_3 - p \tfrac{\partial v_3}{\partial x_3}] d\Omega &= \int_\Omega f_3 v_3 d\Omega, \\
\int_\Omega (\tfrac{\partial u_1}{\partial x_1} + \tfrac{\partial u_2}{\partial x_2} + \tfrac{\partial u_3}{\partial x_3}) q d\Omega &= 0.
\end{aligned} \tag{17}$$

For the construction of the approximation of the solution, u_1, u_2, u_3 and p are written as the linear combination of basis functions (already only a finite number of them):

$$u_i(x) = \sum_{k=1}^{N} u_{ik} \Phi_k(x), \quad i = 1, 2, 3,$$
$$p(x) = \sum_{k=1}^{M} p_k \Psi_k(x).$$

Now we will substitute (according to the Galerkin method):

$$v_1 = v_2 = v_3 = \Phi_k(x), \quad k = 1, 2, \ldots N, \quad q = \Psi_k(x), \quad k = 1, 2, \ldots M.$$

It yields:

$$\int_\Omega \nu (\nabla \sum_{k=1}^{N} u_{1k} \Phi_k) \cdot \nabla \Phi_j d\Omega + \int_\Omega \mathbf{w} \cdot \nabla (\sum_{k=1}^{N} u_{1k} \Phi_k) \Phi_j - \int_\Omega \sum_{k=1}^{M} p_k \Psi_k \frac{\partial \Phi_j}{\partial x_1} d\Omega = \int_\Omega f_1 \Phi_j d\Omega$$
$$\int_\Omega \nu (\nabla \sum_{k=1}^{N} u_{2k} \Phi_k) \cdot \nabla \Phi_j d\Omega + \int_\Omega \mathbf{w} \cdot \nabla (\sum_{k=1}^{N} u_{2k} \Phi_k) \Phi_j - \int_\Omega \sum_{k=1}^{M} p_k \Psi_k \frac{\partial \Phi_j}{\partial x_2} d\Omega = \int_\Omega f_2 \Phi_j d\Omega$$

$$\int_\Omega \nu(\nabla \sum_{k=1}^{N} u_{3k}\Phi_k)\cdot\nabla\Phi_j d\Omega + \int_\Omega \mathbf{w}\cdot\nabla(\sum_{k=1}^{N} u_{3k}\Phi_k)\Phi_j - \int_\Omega \sum_{k=1}^{M} p_k\Psi_k \frac{\partial\Phi_j}{\partial x_3} d\Omega = \int_\Omega f_3\Phi_j d\Omega$$

$$j = 1, 2, \ldots, N$$

$$\int_\Omega (\frac{\partial \sum_{j=1}^{N} u_{1j}\Phi_j}{\partial x_1} + \frac{\partial \sum_{j=1}^{N} u_{2j}\Phi_j}{\partial x_2} + \frac{\partial \sum_{j=1}^{N} u_{3j}\Phi_j}{\partial x_3})\Psi_i d\Omega = 0$$

$$i = 1, 2, \ldots, M.$$

This already represents a system of linear equations of the form:

$$\begin{pmatrix} \nu A + N & 0 & 0 & B_1^T \\ 0 & \nu A + N & 0 & B_2^T \\ 0 & 0 & \nu A + N & B_3^T \\ B_1 & B_2 & B_3 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ p \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ 0 \end{pmatrix}.$$

where

$$\begin{aligned} A(i,j) &= \textstyle\int_\Omega \nabla\Phi_i \cdot \nabla\Phi_j d\Omega \\ N(i,j) &= \textstyle\int_\Omega (\mathbf{w}\cdot\nabla\Phi_i)\Phi_j d\Omega \\ B_k(i,j) &= \textstyle\int_\Omega \frac{\partial\Phi_i}{\partial x_k}\Psi_j d\Omega \;\; k = 1,2,3 \\ f_k(i) &= \textstyle\int_\Omega f_k\Phi_i d\Omega \;\; k = 1,2,3 \\ u_k^T &= [u_{k1}, \ldots, u_{kN}] \\ p^T &= [p_1, \ldots, p_M]. \end{aligned} \tag{18}$$

Remarks: By using the Gauss theorem, it follows that

$$\int_\Omega (\mathbf{w}\cdot\nabla\Phi_i)\Phi_j d\Omega = -\int_\Omega (\mathbf{w}\cdot\nabla\Phi_j)\Phi_i d\Omega,$$

which implies $N^T = -N$, namely that N matrix is skew symetric. Also obviously $A = A^T$ is positive definite, more precisely according to (18) it

is positive semidefinite matrix, it becomes positive definite if we impose Dirichlet boundary conditions.

Denoting

$$A = \begin{pmatrix} A & & \\ & A & \\ & & A \end{pmatrix}$$

$$N = \begin{pmatrix} N & & \\ & N & \\ & & N \end{pmatrix}$$

$$B = \begin{pmatrix} B_1 & B_2 & B_3 \end{pmatrix}$$

$$u^T = \begin{pmatrix} u_1 & u_2 & u_3 \end{pmatrix}$$

$$f^T = \begin{pmatrix} f_1 & f_2 & f_3 \end{pmatrix},$$

we obtain the desired system of linear equations:

$$\begin{pmatrix} \nu A + N & B^t \\ B & O \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}, \tag{19}$$

where $A = A^T$ also a positive definite matrix, N is the matrix expressing the convection operator and is skew symmetric, i.e. $N = -N^T$, We further assume that the the underlying velocity and pressure approximations are (div-)stable, which means that there exist constants γ, Γ, independent of the mesh size h, such that

$$\gamma^2 \leq \frac{(p, BA^{-1}B^t p)}{(p, Qp)} \leq \Gamma^2, \tag{20}$$

where is the pressure mass matrix (or Gramian matrix of basis functions defining P_h).

Later we will see that the validity of this condition will lead to finding a preconditioner, for which the eigenvalues can be estimated independently on the mesh size.

4 Preconditioning

Elman [1] proposed a block diagonal preconditioning matrix of the form:

$$\begin{pmatrix} F & \\ & \frac{1}{\nu}Q \end{pmatrix}, \tag{21}$$

where $F = \nu A + N$ and Q is the pressure mass matrix coming from FEM discretization. Then the eigenvalues of the preconditioned system are the solutions of the generalised eigenvalue problem:

$$\begin{pmatrix} F & B^T \\ B & O \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \lambda \begin{pmatrix} F & 0 \\ 0 & \frac{1}{\nu}Q \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix}$$

These are given by $\lambda = 1$, with eigenvectors of the form $(u, 0)$ so called discretely divergence free, or

$$\lambda = \frac{1 \pm \sqrt{1 + 4\mu}}{2},$$

where μ comes from the generalised eigenvalue problem for the Schur complement system:

$$BF^{-1}B^T p = \mu \left(\frac{1}{\nu}Q\right) p.$$

The following result provides bounds for the preconditioned operator.

Theorem 1 *The eigenvalues of the discrete Oseen operator preconditioned by (21), are $\lambda = 1$ and eigenvalues enclosed in two rectangular boxes of the form:*

$$\left[\tfrac{1+s_{min}}{2}, \tfrac{1+s_{max}}{2}\right] \times [-t, t] \quad \textit{and} \quad \left[\tfrac{1-s_{max}}{2}, \tfrac{1-s_{min}}{2}\right] \times [-t, t]$$

in the complex plane, where

$$s_{min} = \left(1 + \tfrac{4\gamma^2\nu^2}{\delta^2+\nu^2}\right)^{1/2}, \quad s_{max} = \left[\tfrac{1}{2}(1 + 4\Gamma^2 + \sqrt{1 + 8\gamma^2 + 20})\right]^{1/2},$$

$$t = \frac{\Gamma^2}{(1+\frac{4\gamma^2\nu^2}{\delta^2+\nu^2})^{1/2}}. \textit{ where } \delta > 0 \textit{ such that } \rho(A^{\frac{1}{2}} N A^{\frac{1}{2}}) \leq \delta.$$

Proof : See [1]; recall that constants γ, Γ are specified by (19).

5 The minimal polynomial problem

5.1 GMRES

GMRES is one of the iterative methods for nonsymetric matrices, which lead to the construction of solutions in the so-called Krylov subspace.

The Krylov subspace $K_k(A, r)$ of order k generated by A and r is the subspace spanned by $r, Ar, \cdots, A^{k-1}r$.

Given a linear system, $Ax = b$ with a nonsingular matrix, then the standard Richardson iteration

$$x_k = (I - A)x_{k-1} + b$$

generates an approximate solution in the shifted Krylov subspace

$$x_0 + \{r_0, Ar_0, \cdots, A^{k-1}r_0\},$$

with $r_0 = Ax_0$.

Instead of the standard basis one usually prefers an orthonormal basis $v_0 \cdots v_{k-1}$, which can be computed as follows:

Start with $v_1 = r_0 / \parallel r_0 \parallel_2$, assume that we have already an orthonormal basis $v_1 \ldots v_j$ for $K^j(A; r_0)$, then this basis is expanded by computing $t = Av_j$, and orthogonalizing this vector with respect to $v_1, \ldots v_j$.

If we denote by V_j the matrix with colunms v_1 up to v_j, then it can be showed that such a process of orthogonalization results in:

$$AV_{m-1} = V_m H_{m,m-1},$$

where m by $m-1$ matrix $H_{m,m-1}$ is an upper Hessenberg matrix.

Here we will not give further details, there exists a lot of literature on GMRES, see for instance [3], from which our information was extracted.

It is important for us that the residual corresponding to the solution $x_k = x_0 + \alpha_1 r_0 + \ldots \alpha_k A^{k-1} r_0$, can be reformulated as:

$$\begin{aligned} r_k &= b - Ax_k = b - A(x_0 + \alpha_1 r_0 + \ldots \alpha_k A^{k-1} r_0) \\ &= r_0 - \alpha_1 Ar_0 - \ldots - \alpha_k A^k r_0 \\ &= (I - \alpha_1 A - \ldots - \alpha_k A^k) r_0 \\ &= P(A) r_0, \end{aligned}$$

where $P(\xi) = 1 - \alpha_1\xi - \ldots - \alpha_k\xi^k$, and note that $P(0) = 1$.

So the problem of solving a linear system on a shifted Krylov subspace $x_0 + K^k(A; r_0)$ by the minimum residual approach (the case of GMRES) is equal to finding a polynomial of the k-th degree $p_k^{opt}(A)$, such that

$$\| p_k^{opt}(A)r_0 \| = min_{p \in P_k^1} \| p(A)r_0 \|, \tag{22}$$

where P_k^1 denotes the space of polynomials of the k-th degree, satisfying $p(0) = 1$.

5.2 The minimal polynomial problem on a general domain

Due to the matrix norm multiplicativity (we restrict ourselves to the euclidian norm), we can bound:

$$\| p(A)r_0 \| \leq \| p(A) \| \| r_0 \| .$$

Further assuming that A is diagonalizable, i.e. $J = T^{-1}AT$ is diagonal (containing the eigenvalues of A: $\lambda_1 \ldots \lambda_N$) for some nonsingular matrix T, then we can calculate $p(A)$ as:

$$p(A) = T \begin{pmatrix} p(\lambda_1) & & & & \\ & \cdot & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & p(\lambda_N) \end{pmatrix} T^{-1} \tag{23}$$

Thus:

$$\| p(A) \| \leq \| T \| \| T_{-1} \| max_{\lambda \in \sigma(A)} |(p(\lambda)|.$$

Although the exact spectrum is often unknown a priori, very often by some analysis (as in Theorem 1 or Theorem 3) we can locate a domain I in the complex plane, in which our spectrum is contained: $\sigma(A) \subset I$.
Then:

$$max_{\lambda \in \sigma(A)} |(p(\lambda)| \leq max_{\lambda \in I} |(p(\lambda)| \tag{24}$$

After combination of this with (22), we obtain:

$$\begin{aligned} \| r_k \| = min_{p \in P_k^1} \| p(A) r_0 \| &\leq \left(min_{p \in P_k^1} \| p(A) \| \right) \| r_0 \| \\ &\leq C_E \left(min_{p \in P_k^1} max_{\lambda \in \sigma(A)} \mid p(\lambda) \mid \right) \| r_0 \| \\ &\leq C_E \left(min_{p \in P_k^1} max_{\xi \in I} \mid p(\xi) \mid \right) \| r_0 \|, \end{aligned} \tag{25}$$

or equivalently

$$\frac{\| r_k \|}{\| r_0 \|} \leq C_E \epsilon_k(I), \tag{26}$$

where $C_E = \| T \| \| T^{-1} \|$ and

$$\epsilon_k(I) = \left(min_{p \in P_k^1} max_{\xi \in I} \mid p(\xi) \mid \right). \tag{27}$$

It gives a rise to a definition:

Definition 1 *The polynomial p_{min} of the k-th degreee which satisfies*

$$max_{\xi \in I} |p_{min}(\xi)| = \left(min_{p \in P_k^1} max_{\xi \in I} \mid p(\xi) \mid \right) \tag{28}$$

is called a minimal polynomial of degree k with respect to I.

From (26) one can see the importance of a minimal polynomial, namely having found a 'nice' set I, which contains the spectrum of A and by knowing the minimal polynomial corresponding to I (in practice usually its upper bound), we can estimate the convergence of GMRES.

This basically introduces two problems:

1. Finding as small as possible a 'nice' domaing I containing the spectrum of A.
2. Construction of a convenient upper bound on the minimal polynomial corresponding to I.

In this paper we only pay attention to question 1, our 'nice' will be taken from Theorem 1 and will be even simplified for our further considerations (Theorem 3).

5.3 Some theoretical results concerning the minimal polynomial problem

Definition 2 *The polynomial defined by:*

$$T_k(x) = \begin{cases} cos[kcos^{-1}(x)] & if\ -1 \leq x \leq 1 \\ cosh[kcosh^{-1}(x)] & if\ x \geq 1 \\ (-1)^k cosh[kcosh^{-1}(-x)] & if\ x \leq -1 \end{cases} \tag{29}$$

is called the Chebychev polynomial of degree k.

Remark: Chebychev polynomials posses many important properties, among which we will need these two:

$$|T_k(x)| \begin{cases} \leq 1 & \text{if } -1 \leq x \leq 1 \\ > 1 & \text{elsewhere,} \end{cases} \tag{30}$$

besides

$$T_k[\frac{1}{2}(\nu + \frac{1}{\nu})] = \frac{1}{2}(\nu^k + \frac{1}{\nu^k}). \tag{31}$$

Chebychev polynomials will turn out to be very useful to estimate the minimal polynomial corresponding to our domain I.Here we state the Meinardus-Kaniel theorem concerning operators with positive spectrum. It should be mentioned that the proof of it has a lot in common with our second approach in chapter 6.

Theorem 2 *Suppose* $\sigma(A) \subset [\lambda_-, \lambda_+] \subset R^+$. *Then*

$$\epsilon(\sigma(A)) \leq 2exp(-\frac{2k}{\sqrt{C}}), \tag{32}$$

where $C = \frac{\lambda_+}{\lambda_-}$.

Proof: We consider the shifted and scaled Chebychev polynomial

$$p_k(\xi) = \frac{T_k(\frac{\lambda_+ - \xi}{\lambda_+ - \lambda_-} - \frac{\xi - \lambda_-}{\lambda_+ - \lambda_-})}{T_k(\frac{\lambda_+ + \lambda_-}{\lambda_+ - \lambda_-})}. \tag{33}$$

Since $\frac{\lambda_+ - \xi}{\lambda_+ - \lambda_-} - \frac{\xi - \lambda_-}{\lambda_+ - \lambda_-}$ maps: $M : [\lambda_-, \lambda_+] \rightarrow [-1, 1]$,
then according to (30):

$$|T_k(\frac{\lambda_+ - \xi}{\lambda_+ - \lambda_-} - \frac{\xi + \lambda_-}{\lambda_+ - \lambda_-})| < 1 \quad \forall \xi \in [\lambda_-, \lambda_+]$$

and we have:

$$\epsilon(\sigma(A)) \leq \frac{1}{T_k(\frac{\lambda_+ + \lambda_-}{\lambda_+ - \lambda_-})}. \tag{34}$$

In order to calculate $T_k(\frac{\lambda_+ + \lambda_-}{\lambda_+ - \lambda_-})$, exploiting (31) we have to express:

$$\frac{1}{2}(\nu + \frac{1}{\nu}) = \frac{\lambda_+ + \lambda_-}{\lambda_+ - \lambda_-}, \tag{35}$$

which gives a solution: $\nu = \frac{\sqrt{\lambda_+}+\sqrt{\lambda_-}}{\sqrt{\lambda_+}-\sqrt{\lambda_-}}$

That is how we obtain:

$$T_k(\frac{\lambda_+ + \lambda_-}{\lambda_+ - \lambda_-}) = \frac{1}{2}[(\frac{\sqrt{\lambda_+} + \sqrt{\lambda_-}}{\sqrt{\lambda_+} - \sqrt{\lambda_-}})^k + \frac{\sqrt{\lambda_+} - \sqrt{\lambda_-}}{\sqrt{\lambda_+} + \sqrt{\lambda_-}})^k]).$$

The last formula can be rewritten in terms of $C = \frac{\lambda_+}{\lambda_-}$, as:

$$T_k(\frac{\lambda_+ + \lambda_-}{\lambda_+ - \lambda_-}) = \frac{1}{2}[(\frac{\sqrt{C}+1}{\sqrt{C}-1})^k + (\frac{\sqrt{C}-1}{\sqrt{C}+1})^k] \tag{36}$$

This can be again estimated by (since $C \geq 1$):

$$\begin{aligned} T_k(\frac{\lambda_+ + \lambda_-}{\lambda_+ - \lambda_-}) \geq \frac{1}{2}(\frac{\sqrt{C}+1}{\sqrt{C}-1})^k = \frac{1}{2}[(1 + \frac{1}{\sqrt{C}})(1 + \frac{1}{\sqrt{C}} \ldots)]^k \\ = \frac{1}{2}(1 + \frac{2}{\sqrt{C}} + \ldots)^k \geq \frac{1}{2} exp(\frac{2k}{\sqrt{C}}) \end{aligned} \tag{37}$$

And after substituting (37) into (34) we are ready. □

Remarks:

1. It can be proved [6] that the polynomial given by (33) is exactly the minimal polynomial for the interval $[\lambda_-, \lambda_+]$.

2. The Meinardus-Kaniel theorem can also be used in the case $[\lambda_-, \lambda_+] \subset R^-$.

3. One might think that the used method in the Meinardus-Kaniel theorem can also be extended for the general case $\lambda_- < 0 < \lambda_+$. However, then we have: $|\frac{\lambda_+ + \lambda_-}{\lambda_+ - \lambda_-}| < 1$, which implies $|T_k(\frac{\lambda_+ + \lambda_-}{\lambda_+ - \lambda_-})| \leq 1$ and we never get a better bound than: $\epsilon(\sigma(A)) > 1$, which is apparently true but useless. In chapter 6, we will discuss better approaches to find convenient bounds, even for this case.

5.4 Our problem

According to Theorem 1, the spectrum of the preconditioned Oseen operator can be contained in a domain I, that consists of the point 1 plus two rectangles in the complex plane, symmetric with respect to $x = \frac{1}{2}$. The reason why the spectrum is in general complex is that the matrix N, coming from the discretization of the convection term, is nonsymmetric (see remarks next to formula (18)).

In this report we only focus on the real spectrum for simplicity, which can be obtained by neglecting the convection term in the Oseen equation. Such an easier problem, the so-called Stokes problem, also often appears in fluid mechanics:

$$\begin{array}{c} -\nu\triangle\mathbf{u} + \nabla p = f \\ \nabla \cdot \mathbf{u} = 0; \end{array} \quad in\ \ \Omega. \tag{38}$$

Then by using a similar technique as in chapter 3, for the Oseen operator, it can be shown that the resulting discrete system is of the form:

$$\begin{pmatrix} \nu A & B^t \\ B & O \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}, \tag{39}$$

and the spectrum of the matrix preconditioned by:

$$\begin{pmatrix} \nu A & \\ & \frac{1}{\nu} Q \end{pmatrix}, \tag{40}$$

can be located (similarly to Theorem 1 for the Oseen operator):

Theorem 3 *The eigenvalues of the discrete Stokes operator preconditioned by (30) are of* $\lambda = 1$ *and those enclosed in two intervals of the form:*

$$\left[\frac{1+s_{min}}{2}, \frac{1+s_{max}}{2}\right] \quad and \left[\frac{1-s_{max}}{2}, \frac{1-s_{min}}{2}\right],$$

where

$$s_{min} = \left(1+4\gamma^2\right)^{1/2}, \quad s_{max} = \left(1+4\Gamma^2\right)^{1/2}.$$

Proof : See [2]; recall that constants γ, Γ are again specified by (19).

Although the constants γ, Γ constants are more of a theoretical value, Theorem 3 shows that our spectrum is located at both sides of a real axis (indefinite problem).

Our problem will be to find some convenient estimates for the minimal polynomial corresponding to such a domain, which is situated at both sides of the real axis. We have already seen (remarks in paragraph (5.3)) that the approach from the proof of the Meinkardus-Kaniel theorem fails in this case (it does not give a suitable estimate).

In the next paragraph we discuss another approach for bounding the minimal polynomial.

6 The minimalization on two disjoint intervals - a novel approach

Our intention now will be to find a polynomial of the k-th degree, with the condition $p(0) = 1$, for which its absolute value on $I = (a,b) \cup (b',a') = I\text{`}$ $I \in R^+, I' \in R^-$ is sufficiently small (more precisely, converges to zero as the degree k of the polynomial goes to infinity). In the following two subsections two different approaches are considered.

6.1 Approach 1 - the product of two Chebyshev polynomials

Let us consider $p_{2\cdot k}(\xi)$ as the product of two Chebyshev polynomials shifted and scaled, to intervals $I = (a,b)$ and $(b',a') = I\text{`}$, respectively.
Thus we have:

$$p_{2\cdot k}(\xi) = \frac{T_k(\frac{b-\xi}{b-a} - \frac{\xi-a}{b-a})T_k(\frac{b'-\xi}{b'-a'} - \frac{\xi-a'}{b'-a\text{`}})}{T_k(\frac{b+a}{b-a})T_k(\frac{b\text{`}+a\text{`}}{b'-a'})}. \tag{41}$$

It is obvious that such a scaled polynomial guarantees the condition $p_{2k}(0) = 1$.

If we consider for instance I, notice that the first $|T_k(\frac{b-\xi}{b-a} - \frac{\xi-a}{b-a})|$ attains one of its maxima (always equal to 1 in absolute value) at the point $x = b$ and the second $|T_k(\frac{b'-\xi}{b'-a'} - \frac{\xi-a'}{b'-a'})|$ is an increasing function on I. It together implies:

$$max_{\xi \in I}|p_{2k}(\xi)| = |\frac{T_k(\frac{b'-b}{b'-a'} - \frac{b-a'}{b'-a'})}{T_k(\frac{b+a}{b-a})T_k(\frac{b'+a'}{b'-a'})}|. \tag{42}$$

For the same reason:

$$max_{\xi \in I'}|p_{2k}(\xi)| = |\frac{T_k(\frac{b-b'}{b-a} - \frac{b'-a}{b-a})}{T_k(\frac{b+a}{b-a})T_k(\frac{b'+a'}{b'-a'})}|. \tag{43}$$

Rewriting: $\frac{b-b'}{b-a} - \frac{b'-a}{b-a} = 1 + \frac{2(a-b')}{b-a}$ and $\frac{b'-b}{b'-a'} - \frac{b-a'}{b'-a'} = 1 + \frac{2(b-a')}{a'-b'}$, we have that

$$max_{\xi \in I \cup I'}|p_{2k}(\xi)| = \frac{max(|T_k(1 + \frac{2(a-b')}{b-a})|, |T_k(1 + \frac{2(b-a')}{a'-b'})|}{|T_k(\frac{b+a}{b-a})T_k(\frac{b'+a'}{b'-a'})|}. \tag{44}$$

To decide which of the polynomials in (44) is greater in absolute value than the other one, we have to compare their arguments $1 + \frac{2(a-b')}{b-a}$ and $1 + \frac{2(b-a')}{a'-b'}$.

It can be shown that depending on the ratio of the lengths of the intervals:

ratio of intervals	relation of polynomials
$b-a = a'-b'$	$\|T_k(1 + \frac{2(a-b')}{b-a})\| = T_k\|(1 + \frac{2(b-a')}{a'-b'})\|$
$b-a > a'-b'$	$\|T_k(1 + \frac{2(a-b')}{b-a})\| < T_k\|(1 + \frac{2(b-a')}{a'-b'})\|$
$b-a < a'-b'$	$\|T_k(1 + \frac{2(a-b')}{b-a})\| > T_k\|(1 + \frac{2(b-a')}{a'-b'})\|$

(45)

For an illustration of the case $b-a = a'-b'$ (the same lengths of intervals I, I'), see Figure 3. (for $k = 4, I = (1.5, 3), I' = (-2, -0.5)$).

For instance, if $a - b \geq a' - b'$ (which we will assume from now) then:

$$max_{\xi \in I \cup I'} |p_{2k}(\xi)| = \frac{T_k(1 + \frac{2(b-a')}{a'-b'})}{T_k(\frac{b+a}{b-a}) T_k(\frac{b'+a'}{b'-a'})}, \tag{46}$$

which implies

$$\epsilon_{2 \cdot k}(I \cup I') \leq \frac{T_k(1 + \frac{2(b-a')}{a'-b'})}{T_k(\frac{b+a}{b-a}) T_k(\frac{b'+a'}{b'-a'})}. \tag{47}$$

6.2 Why the first approach often fails

Applying (37) to both Chebychev polynomials in the denominator of (47), we can bound (47) by:

$$\epsilon_{2 \cdot k}(I \cup I') \leq 4 T_k(1 + \frac{2(b-a')}{a'-b'}) \cdot exp(-\frac{2k}{\sqrt{C}} - \frac{2k}{\sqrt{C'}}), \tag{48}$$

,

where $C = \frac{b}{a}$, $C' = \frac{b'}{a'}$.

Up to now we have not approximated $T_k(1 + \frac{2(b-a')}{a'-b'})$ which we do as follows.

Let us denote $\alpha = \frac{2(b-a')}{a'-b'}$, and solve the equation

$$1 + \alpha = \frac{1}{2}(\nu + \frac{1}{\nu}). \tag{49}$$

It leads to:

$$\nu^2 - (2 + 2\alpha)\nu + 1 = 0,$$

with solutions:

$$\nu_{1,2} = 1 + \alpha \pm \sqrt{\alpha(\alpha + 2)}.$$

Substituting back for α we get:

$$\nu_{1,2} = \frac{2b - a' - b' \pm 2\sqrt{(b-a')(b-b')}}{a' - b'}. \tag{50}$$

Now we can already express (taking ν_1 from (50)):

$$T_k(1 + \frac{2(b-a')}{a'-b'}) = \frac{1}{2}({\nu_1}^k + \frac{1}{\nu_1^k}).$$

Here, since $\nu_1 > 1$ we can replace $T_k(1 + \frac{2(b-a')}{a'-b'})$ by (for higher values of k, asymptotically):

$$T_k(1 + \frac{2(b-a')}{a'-b'}) \approx \frac{1}{2}(\frac{2b - a' - b' + 2\sqrt{(b-a')(b-b')}}{a'-b'})^k. \qquad (51)$$

And finally substituting (51) in (48) gives (let us remind again this holds for large values of k only):

$$\epsilon_{2\cdot k}(I \cup I') \leq 2 \left(\frac{2b - a' - b' + 2\sqrt{(b-a')(b-b')}}{(a'-b') \cdot exp^2(\sqrt{\frac{a}{b}} + \sqrt{\frac{a'}{b'}})} \right)^k . \qquad (52)$$

Let us see what such estimates give for the case that: $I = [1, b], I' = [1-b, 1-a]$, which is a spectrum typical for the preconditioned Stokes problem, note that instead of working with the isolated eigenvalue $\lambda = 1$ and the interval $[a, b]$, we consider interval $[1, b]$ straight forward.

$I \cup I'$	upper bound on $\epsilon_{2\cdot k}(I \cup I')$
$[-1, -0.1] \cup [1, 2]$	$2 \cdot (1.4524)^k$
$[-3.75, -0.1] \cup [1, 4.75]$	$2 \cdot (2.0677)^k$
$[-49, -4] \cup [1, 50]$	$2 \cdot (2.83)^k$
$[-499, -1] \cup [1, 500]$	$2 \cdot (4.8941)^k$

Results in the Table are calculated with the formula (52). Since our bounds are always in the form of a geometric series with factor greater than 1 the first approach may seem not to work at all.

However, if calculated according to (47) (in terms of Chebychev polynomials), we can get convergence sometimes. More precisely, formula (52) approximates (47) well only in the case $C \gg 1$ (or equivalently $a \ll b$). In order to demonstrate this the comparison between (47) and (52) has been done for the spectra from the Table.

Figure 4 displays the case of a small spectrum (x axis presents the degree of a Chebychev polynomial, y axis gives the values of bounds (47), (52)). Here (47) converges, although (52) diverges. As the spectra are becoming larger (figure 5,6,7), bounds (52) and (47) are getting closer and both give divergence. (by convergence we simply mean $\epsilon_{2\cdot k}(I \cup I') \to 0$ as $k \to \infty$).

The fact that that for a larger spectrum the bound for our product of two scaled and shifted Chebychev polynomial diverges (in the max norm) is certain advantage of this first approach.

6.3 Approach 2 - a polynomial transformation

Suppose again that $I = [a, b] \cup [b', a'] = I', \;\; I \in R^+, I' \in R^-$.

Now let us consider the polynomial $p^{trans}(\xi)$ of degree d, with the two properties:

$$|p^{trans}(\xi)| \leq 1 \quad \text{if } \xi \in I \cup I', \tag{53}$$

$$|p^{trans}(0)| > 1. \tag{54}$$

Then according to (30), $|T_k(p^{trans}(\xi)| \leq 1$ and $|T_k(p^{trans}(0))|$ increases as $k \to \infty$ and consequently

$$p_{k \cdot d}(\xi) = \frac{T_k(p^{trans}(\xi))}{T_k(p^{trans}(0))} \tag{55}$$

is a polynomial of degree $k \cdot d$ satisfying $p_{k \cdot d}(0) = 1$, and

$$max_{\xi \in I \cup I'} |p_{k \cdot d}(\xi)| \leq \frac{1}{T_k(p^{trans}(0))} \to 0 \quad \text{as } k \to \infty. \tag{56}$$

Here arises the basic question, what kind of a polynomial p^{trans} we can identify that satisfies (53) and (54).

It is clear that the choice: $P^{trans}(\xi) = a\xi + b$ fails, since (53) and (54) can not both be satisfied.

Here we will only show that selecting $P^{trans}(\xi)$ as a convenient quadratic function may satisfy (53)and (54).

Suppose again, for instance, (it also holds for the Stokes preconditioned spectrum)

$$b - a \geq a' - b'. \tag{57}$$

We select p^{trans} as the polynomial of second degree satisfying:

$$p^{trans}(a') = 1, \quad p^{trans}(a) = 1, \quad p^{trans}(b) = -1, \tag{58}$$

We can prove (it can be also seen from Figure 8 for the case $a = 1, b = 5, a' = -3, b' = -1$) that (53) and (54) are satisfied.

Such a polynomial is obviously (by Lagrange interpolation):

$$p^{trans}(\xi) = 1 \cdot \frac{(\xi - a)(\xi - b)}{(a' - a)(a' - b)} + 1 \cdot \frac{(\xi - a')(\xi - b)}{(a - a')(a - b)} - 1 \cdot \frac{(\xi - a')(\xi - a)}{(b - a')(b - a')} \quad (59)$$

with

$$p^{trans}(0) = \frac{ab}{(a' - a)(a' - b)} + \frac{(a'b}{(a - a')(a - b)} - \frac{a'a}{(b - a')(b - a')}. \quad (60)$$

Finally, according to (27) we can write:

$$\epsilon_{2 \cdot k}(I \cup I') \leq \frac{1}{T_k(\frac{ab}{(a'-a)(a'-b)} + \frac{a'b}{(a-a')(a-b)} - \frac{a'a}{(b-a')(b-a')})} \quad (61)$$

The denominator of (61) can be bounded again by an exponential function (as in (49),(50),(51)). To demonstrate how the second approach works, see Table below (for the same spectra as for the first approach):

$I \cup I'$	upper bound on $\epsilon_{2 \cdot k}(I \cup I')$
$[-1, -0.1] \cup [1, 2]$	$2 \cdot (0.6485)^k$
$[-3.75, -0.1] \cup [1, 4.75]$	$2 \cdot (0.8623)^k$
$[-49, -4] \cup [1, 50]$	$2 \cdot (0.9252)^k$
$[-499, -1] \cup [1, 500]$	$2 \cdot (0.9920)^k$

In general, no matter what spectrum we have (as long as it is contained in $I \cup I', I \in R^+, I' \in R^-$) with the second approach we are always able to to bound $\epsilon_k(I \cup I')$ (and consequently the rate of convergence of GMRES) by some convergent geometrical series.

Another property of the second approach is that the conditions (53) and (54) do not determine the transformation polynomial $p^{trans}(\xi)$ uniquely.

For this second approach there are still many open questions, such as:

1. In the class of all quadratic polynomials satisfying (53) and (54), find $p^{trans}_{optimal}(\xi)$ for which the value in 0 is maximal. Then our estimate on $\epsilon_k(I \cup I')$ can be sharper.

2. Does the use of a polynomial transformation of degree $d > 2$ lead to any improvement in the $\epsilon_k(I \cup I')$ estimate ?

7 Conclusions

We have seen that the preconditioning of the Oseen problem leads to an indefinite operator with, in generall, a complex spectrum. If neglecting the convection term, we obtain an easier problem, the so-called Stokes problem, for which the preconditioned operator has a real spectrum. We have proposed two methods for estimating the max norm of the minimal polynomial associated with a domain that contains the spectrum. Although the first approach turns out not to be always useful (it may only work in cases of a small spectrum), the second approach, using a quadratic polynomial transformation is applicable for any real domain of the type $I \cup I', I \in R^+, I' \in R^-$, its further development needs further research.

8 References

[1] Howard Elman, David Silvester, *Fast Nonsymetric Iterations and Preconditioning for Navier-Stokes Equation , UMCP-CSD:CS-TR-3283, 1996*

[2] *Howard Elman, David Silvester, Andrew J. Warhen, Iterative Methods for Problems in Computational Fluid Mechanics, UMCP-CSD:CS-TR-3675, 1996*

[3] *G.L.G Sleijpen, H.A. van der Vorst, Krylov subspace methods for large linear systems of equation, Department of mathematics, University Utrecht(1993)*

[4] *Jan Modersitzki, Gerald Opfer, Faber versus minimal polynomials on annular sectors, Hamburger Beitrage zur Angewandten Mathematik, 1994*

[5] *Gerhard Starke, R.S. Varga, A hybrid Arnordi-Faber iterative method for nonsymetric systems of linear equations, Institute of Computational Mathematics, Kent State University, 1991*

[6] *O. Axelsson, V.A.Barker, Finite Element Solution of Boundary Value Problems, Academic Press, 1984*

[7] *C. Cuvelier, A.Segal, A.A. van Steenhoven, Finite Element Methods and Navier Stokes equations, D.Reidel Publishing Company, 1986*

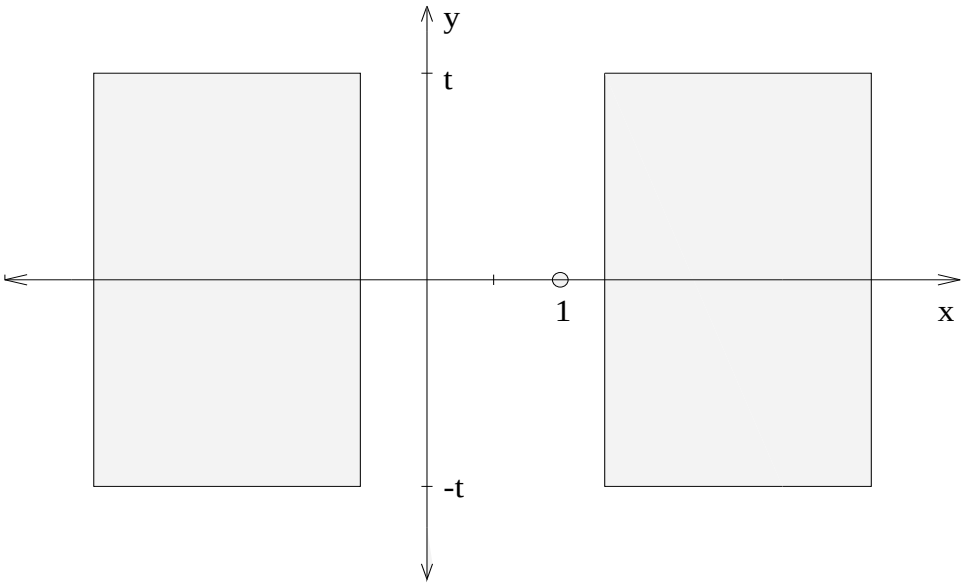

Figure 1: Example of the spectrum in Theorem 1

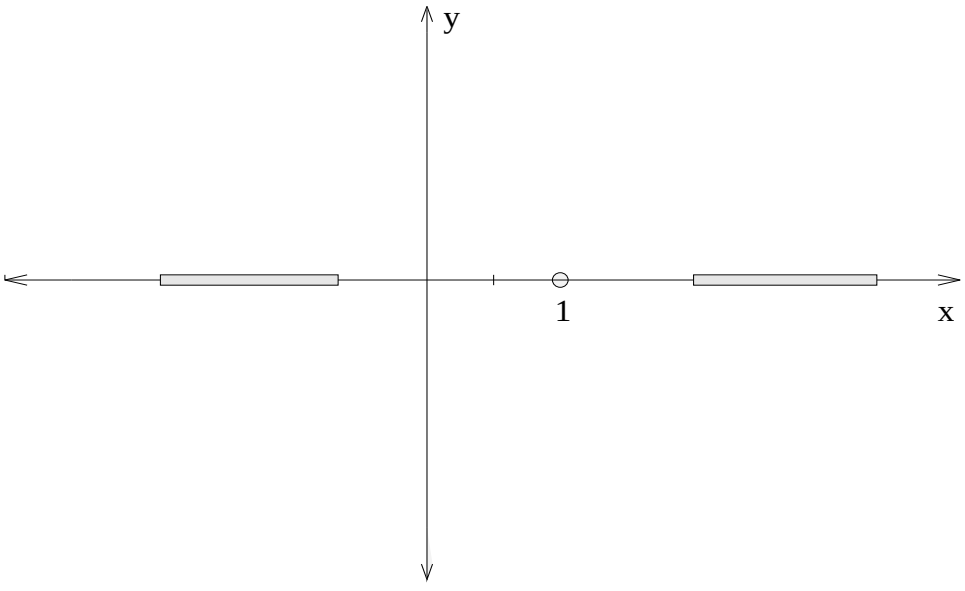

Figure 2: Example of the spectrum in Theorem 3

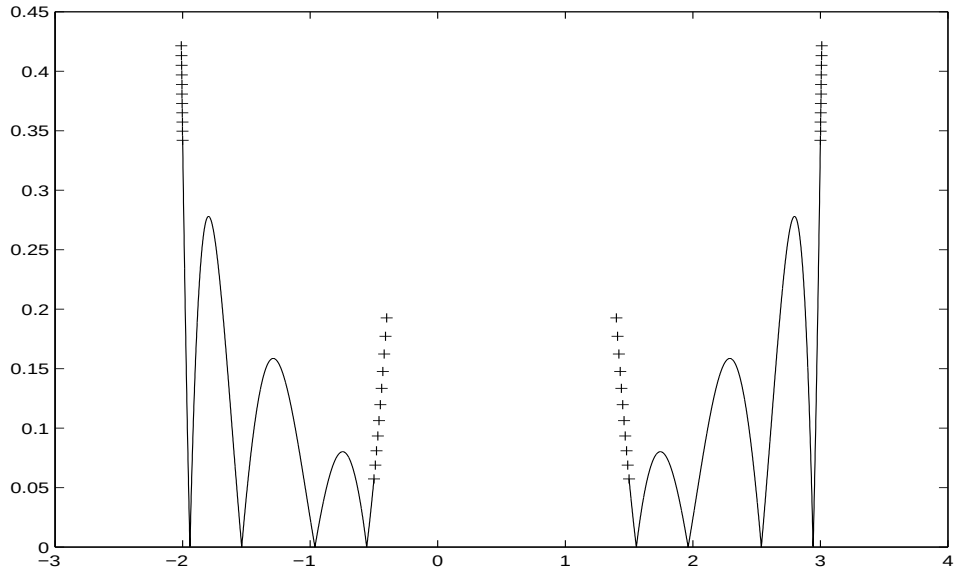

Figure 3: Symmetry of $|p_8(\xi)|$,for the case $I = (1.5, 3), I' = (-2, -0.5)$

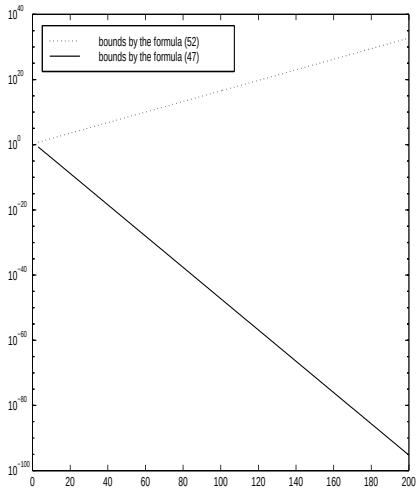

Figure 4: $[-1,-0.1]\cup[1,2]$

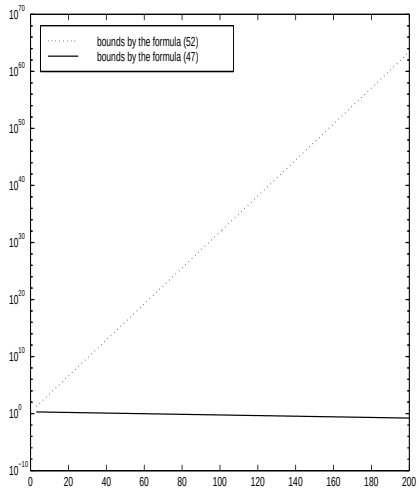

Figure 5: $[-3.75,-0.1]\cup[1,4.75]$

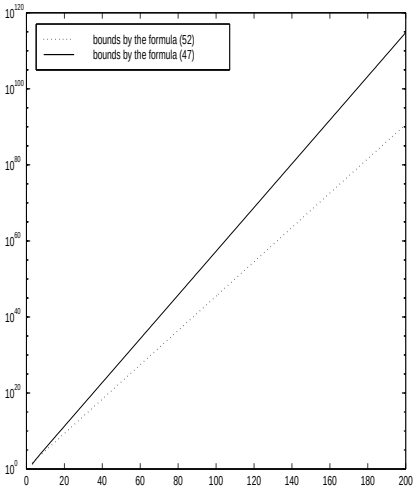

Figure 6: $[-49,-4]\cup[1,50]$

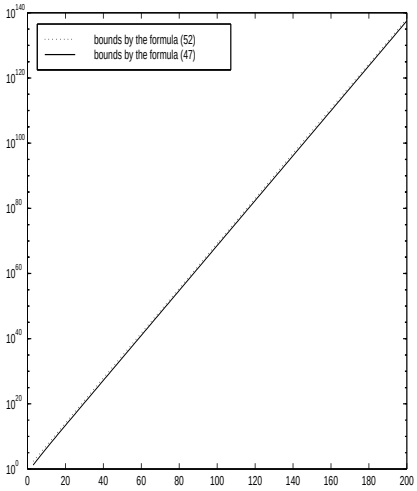

Figure 7: $[-499,-1]\cup[1,500]$

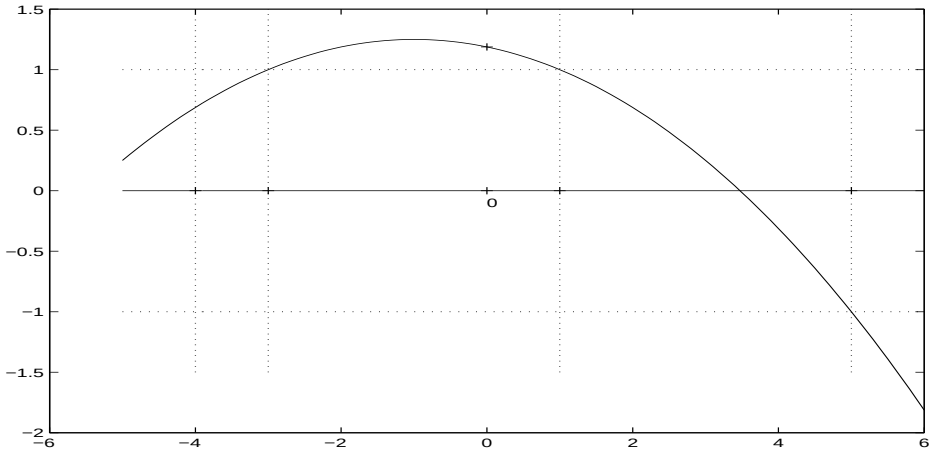

Figure 8: Parabolic transformation, $I \cup I' = [-4, -3] \cup [1, 5]$