Reaching the superlinear convergence phase of the CG method

by O. Axelsson[1], J. Karátson[2]

Abstract

The rate of convergence of the conjugate gradient method takes place in essentially three phases, with respectively a sublinear, a linear and a superlinear rate. The paper examines when the superlinear phase is reached. To do this, two methods are used. One is based on the K-condition number, thereby separating the eigenvalues in three sets: small and large outliers and intermediate eigenvalues. The other is based on annihilating polynomials for the eigenvalues and, assuming various analytical distributions of them, thereby using certain refined estimates. The results are illustrated for some typical distributions of eigenvalues and with some numerical tests.

1 Introduction

Let A be a symmetric positive definite (spd) matrix. The classical conjugate gradient (CG) method is the most widespread way for the iterative solution of linear systems $Ax = b$, see e.g. [2, 10, 16]. It either uses the standard Euclidean inner product, $\langle x, y\rangle = x^T y$, or in its preconditioned form, the inner product $\langle x, y\rangle = x^T C y$, defined by an spd preconditioning matrix C. Let $\|e\|_A := \langle e, Ae\rangle^{1/2}$. In both its unpreconditioned and preconditioned form, the CG method is an optimal method in the sense that the relative errors satisfy

$$\frac{\|e_k\|_A}{\|e_0\|_A} = \min_{P_k \in \pi_k^1} \max_{\lambda \in S(C^{-1}A)} | P_k(\lambda) | , \tag{1}$$

where $e_k = x_k - x$ is the error vector at the kth iteration step and x_0 is an arbitrarily chosen initial approximation of x. Further, π_k^1 denotes the set of polynomials of degree k normalized at the origin. Here the spectrum $S(C^{-1}A)$ is considered as a disjoint set, i.e. multiplicities of the eigenvalues play no role, see e.g. [1, 2, 9]. The rate of convergence is measured by the ratio in (1) or by its geometric average, $(\|e_k\|_A/\|e_0\|_A)^{1/k}$. For particular initial errors e_0, which are deficient in some parts of the eigenvector space, one can get a faster rate of convergence. In this paper we assume a general initial vector e_0 with no such deficiencies.

Assuming exact arithmetic, it is well-known that the rate of convergence of the method takes place in the three phases: a sublinear, frequently of short duration, an intermediate, where the rate of convergence is nearly linear, and a superlinear phase, where the iteration error decays more rapidly for each new iteration, see e.g. [2, 4, 5]. For an operator theoretical framework to explain the different phases, see [15]. Clearly, there is no sharp

[1]Institute of Geonics AS CR, IT4 Inovations, Ostrava, The Czech Republic, and King Abdulaziz University, Jeddah, Saudi Arabia , owea@it.uu.se.

[2]Department of Applied Analysis & MTA-ELTE NumNet Research Group, ELTE U ment of Analysis, Technical University; Budapest, Hungary; karatson@cs.elte.hu.

boundary for the different phases. We note that if the eigenvalues accumulate at a given number (in our case on the real line) then one immediately enters the superlinear phase, and this also holds for infinite dimensional operators which are compact perturbations of a positive multiple of the identity operator [4, 6, 8, 19].

The above-desribed details serve as a motivation for our paper.

In the present paper we are concerned with estimating the point where the superlinear phase is reached. Thus the main difference of our study compared to the mentioned papers is that we are interested in the point of transition from the linear into the superlinear phase, instead of studying one of these phases. We give bounds for the iteration number where the linear phase more markedly goes over in the superlinear phase. This will be based on various assumptions on the distribution of the eigenvalues. Our new results are based on two types of methods. First we use a proper splitting of the spectrum and estimate via the K-condition number. Then refined estimates are given, based on the annihilating polynomial for the eigenvalues, using the superlinear bound instead of the K-condition number. Here the dependence of the reaching point on the size of the matrix is also studied. The results are illustrated with some numerical tests.

Note finally that the results hold for eigenvalue distributions for both unpreconditioned and preconditioned matrices. In some examples we assume an unpreconditioned matrix, while the case of clustering of eigenvalues occurs typically for certain preconditioners, see [7].

2 Basic estimates and the problem

Upper bounds on the rate of convergence of the method can be obtained by choosing suitable polynomials P_k in (1). As shown, e.g., in [2], the well-known linear and superlinear estimates are obtained in this way.

The classical *linear estimate* for the rate of convergence is based on the best polynomial approximation on an interval,

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq \min_{P_k \in \pi_h^1} \max_{m \leq \lambda \leq M} |P_k(\lambda)|$$

where $m = \min \lambda_i$, $M = \max \lambda_i$. As is well-known, the optimal polynomial equals the scaled Chebyshev polynomial of the first kind,

$$P_k(\lambda) = Q_k(x; m, M) := T_k(\frac{M+m-2\lambda}{M-m}) / T_k(\frac{M+m}{M-m}),$$

where $T_k(x) = \frac{1}{2}[(x+\sqrt{x^2-1})^k + (x-\sqrt{x^2-1})^k]$. This leads to the bound

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq 2 \left(\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}\right)^k \qquad (k = 1, 2, ..., n) \tag{2}$$

for the convergence factor over k iterations, so it leads to a linear rate base root of M/m (i.e. of the spectral condition number of $C^{-1}A$).

One way to obtain a *superlinear estimate* is to write A using the decomposition

$$A = \varrho I + E, \tag{3}$$

where $\varrho > 0$ is a given number, and let

$$\mu_j \quad (j = 1, 2, ..., n)$$

denote the eigenvalues of E in decreasing order, i.e. $|\mu_1| \geq |\mu_2| \geq \dots$. Then the polynomials $\prod_{j=1}^{k}(1 - \frac{\lambda}{\lambda_j})$ yield the bound

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq \max_{i \geq k+1} \prod_{j=1}^{k} \frac{|\mu_j - \mu_i|}{\lambda_j} = \max_{i \geq k+1} \prod_{j=1}^{k} \frac{|\mu_j - \mu_i|}{\varrho + \mu_j} \leq 2^k \prod_{j=1}^{k} \frac{|\mu_j|}{\varrho + \mu_j}. \tag{4}$$

From this, one can get further estimates related to the geometric or arithmetic means:

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq 2\|A^{-1}\| \left(\prod_{j=1}^{k} |\mu_j|\right)^{1/k} \leq \frac{2\|A^{-1}\|}{k} \sum_{j=1}^{k} |\mu_j| \qquad (k = 1, 2, ..., n). \tag{5}$$

Since one has no exact, readily available information on the errors of the CG method, our analytical comparison of the two phases will use the above convergence bounds. (This is, of course, not the case for the numerical tests.) We remark, however, that one can actually compute accurate approximations of the errors, see [18], but only with some additional computations.

Based on (2) and (4), our problem is the following. From which index σ_n on will we have

$$2^k \prod_{j=1}^{k} \frac{|\mu_j|}{\varrho + \mu_j} \leq 2 \left(\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}\right)^k \qquad (k \geq \sigma_n); \tag{6}$$

or more generally, how does σ_n, or rather its upper bound, depend on n asymptotically as $n \to \infty$?

3 Estimates via the K-condition number using a splitting of the spectrum

The aim of this section is to find an estimate of the iteration number k where the average convergence factor starts to decrease. To this end, to improve the bound (2), we shall choose a proper polynomial based on the splitting of the spectrum in three sets. Namely, this choice is based on the assumption that the spectrum of the preconditioned operator contains more or less isolated small eigenvalues, some outlier large eigenvalues and a cluster of intermediate eigenvalues. Clearly, the two boundary points, sep
are not fixed, but can be chosen for convenience.

Assume that, at the kth iteration step, we have chosen p_k smallest eigenvalues and q_k largest. The three subsets are defined by

$$0 < \lambda_1 < \lambda_2 < \ldots < \lambda_{p_k} < a, \qquad \lambda_n > \lambda_{n-1} > \ldots > \lambda_{n-q_k+1} > b\,, \tag{7}$$

and $a = \lambda_{p_k+1} \le \lambda_j \le \lambda_{n-q_k} = b$, where $[a,b]$ is the interval of intermediate eigenvalues. We note that in this section it is more convenient to write the eigenvalues in increasing order.

Then a proper polynomial to annihilate the outlier eigenvalues is

$$P_k(\lambda) = \prod_{j=1}^{p_k}(1-\frac{\lambda}{\lambda_j})\, Q_{k-s_k}(\lambda) \prod_{j=n-q_k+1}^{n}(1-\frac{\lambda}{\lambda_j})\,,$$

where $s_k = p_k + q_k$. Here we use the bound

$$\max_{\{\lambda_i\}} \mid P_k(\lambda) \mid \le \max_{a\le\lambda_i\le b}\{\prod_{j=1}^{p_k} \mid 1-\frac{\lambda_i}{\lambda'_j} \mid \prod_{j=p_k+1}^{s_k}\left(1-\frac{\lambda_i}{\lambda'_j}\right) \max_{a\le\lambda\le b} \mid Q_{k-s_k}(\lambda;a,b) \mid\}\,, \tag{8}$$

where we have denoted the outlier eigenvalues

$$\lambda'_j = \begin{cases} \lambda_j, j = 1,\ldots,p_k \\ \lambda_{n-j+p_k+1}, j = p_k+1, p_k+2, \ldots, s_k\,. \end{cases}$$

Here Q_k is the previously mentioned Chebyshev polynomial and

$$\mid Q_{k-s_k}(\lambda;a,b) \mid \le 2^{1/(k-s_k)}\left(\frac{\sqrt{b/a}-1}{\sqrt{b/a}+1}\right)^{k-s_k}\,.$$

Let $\mu_i = \lambda_i - \frac{a+b}{2},\ \mu'_j = \lambda'_j - \frac{a+b}{2}$.

Since $\lambda'_j > \lambda_i$ and hence $\mu'_j > \mu_j > 0$, for the large outlier eigenvalues the first factor (α) in (8) is bounded by

$$\begin{aligned} \alpha\ :\ &= \prod_{j=1}^{p_k}\frac{\mu_i-\mu'_j}{\lambda'_j}\prod_{j=p_k+1}^{s_k}\frac{\mu'_j-\mu_i}{\lambda'_j} \le 2^{p_k}\prod_{j=1}^{p_k} \mid \mu'_j \mid \prod_{j=p_k+1}^{s_k}\mu'_j/\prod_{j=1}^{s_k}\lambda'_j \\ &\le 2^{p_k}\prod_{j=1}^{s_k} \mid \mu'_j \mid /\prod_{j=1}^{s_k}\lambda'_j \le 2^{p_k}\left(\frac{1}{s_k}\sum_{j=1}^{s_k} \mid \mu'_j \mid\right)^{s_k}/\prod_{j=1}^{s_k}\lambda'_j \\ &= 2^{p_k}K(S_k)\left(\sum_{j=1}^{s_k} \mid \mu'_j \mid /\sum_{j=1}^{s_k}\lambda'_j\right)^{s_k} \end{aligned} \tag{9}$$

where we define the *K-condition number* corresponding to the set $S_k = \{\lambda'_1, \lambda'_2 \ldots, \lambda'_{s_k}\}$ as the ratio

$$K(S_k) := (\frac{1}{s_k}\sum_{j=1}^{s_k}\lambda'_j)^{s_k}/\prod_{j=1}^{s_k}\lambda'_j = (\frac{1}{s_k}\sum_{j=1}^{s_k}\lambda'_j/\{\prod_{j=1}^{s_k}\lambda'_j\}^{1/s_k})^{s_k}\,.$$

For further information about the K-condition number, see e.g. [2, 4, 14, 13].

For the second factor in (9) it holds

$$\begin{aligned}\beta : &= (\sum_{j=1}^{s_k} \mid \mu'_j \mid / \sum_{j=1}^{s_k} \lambda'_j)^{s_k} = \left\{ \left(\sum_{j=1}^{s_k} \mid \lambda'_j - \frac{a+b}{2} \mid / \sum_{j=1}^{s_k} \lambda'_j \right)^{s_k} \right. \\ &\leq \left([p_k \frac{a+b}{2} + \sum_{j=p_k+1}^{s_k} (\lambda'_j - \frac{a+b}{2})] / \sum_{j=1}^{s_k} \lambda'_j \right)^{s_k} \\ &\leq (1-\xi_k)^{s_k} ,\end{aligned}$$

where

$$\xi_k = \frac{q_k - p_k}{q_k + p_k}(a+b) / \frac{2}{s_k} \sum_{j=1}^{s_k} \lambda'_j .$$

If we keep the number p_k of the small outlier eigenvalues fixed and increase the number $q_k = s_k - p_k$ of big outlier eigenvalues, then $\frac{1}{s_k} \sum_{j=1}^{s_k} \lambda'_j$ decreases as k increases. Hence $\beta \to 0$ geometrically with an increasingly smaller factor as k increases. Since the second factor in (8) does not increase if $k - s_k$ and a and b are fixed, or rather also decreases since b decreases when q_k increases, there will occur a superlinear rate of convergence, at least from the point where

$$\frac{1}{1-\xi_k} > 2^{p_k/s_k} K(S_k)^{1/s_k} = 2^{p_k/s_k} \frac{1}{s_k} \sum_{j=1}^{s_k} \lambda'_j / (\prod_{j=1}^{s_k} \lambda'_j)^{1/s_k} .$$

Therefore, the smallest number k where this occurs can be used as an estimate of the number of iterations needed to enter the superlinear rate.

The above separation of eigenvalues illustrates the convergence behaviour of the conjugate gradient method, which typically occurs in there phases. The first, the sublinear phase can be modelled by the first factor $\prod_{j=1}^{p_k}(1 - \frac{\lambda}{\lambda_j})$, where the Fourier series components corresponding to the smoother eigenfunctions are damped out. In the second phase, where the convergence is essentially linear, mainly the intermediate components are damped. After that, one enters the final, superlinear phase. This is first slow, when the decrease of the average convergence factor is small because all components are damped, but it then gradually becomes faster, when the most rapidly oscillating eigenvector components are damped out.

As we can see, the size of the K-condition number plays a role in the above estimate in determining how many iterations are required before one reaches the superlinear convergence rate. Essentially the same analysis as above has appeared in [2, Chapter 13].

Finally, we remark here that in some important practical problems (such as in constrained optimization problems solved by some penalization method), where there are two or more well separated eigenvalue intervals, the rate of convergence of the conjugate gradient method depends essentially only on one of the intervals, see e.g. [
a superlinear rate of convergence for the case where the second interval
eigenvalues, hence separated from the first interval. In the next section, t

number is estimated for some typical distribution of eigenvalues, and corresponding estimates of the number of iterations required to reach the superlinear rate of convergence are shown.

4 Balancing the K-condition number

In this section we use an estimate based on the K-condition number to approximately find the point where one enters the superlinear convergence phase. As has been shown in [2, 14], the following alternate estimate of the rate of convergence of the conjugate gradient method holds,

$$(r^k)^T C^{-1} r^k \leq (K(B)^{1/k} - 1)^k \, (r^k)^T C^{-1} r^0 \qquad (k = 1, 2, \ldots). \tag{10}$$

Note that this estimate is based on a norm defined by the inverse of the preconditioning matrix.

Here we have assumed that C and A are symmetric and positive definite, and $B := C^{-1}A$. Since the CG method terminates, i.e. in exact arithmetic it needs at most n iterations for a system of order n, this estimate is useful only if

$$K(B)^{1/n} < 2\,, \quad \text{i.e. } \log_2 K(B) < n\,.$$

To illustrate this result, following [2], we consider now some distributions of eigenvalues which can illustrate some typical cases occurring in practice. The first shows that the K-condition number can be very informative for the estimates, while the two others indicate that it can be of less value in other cases. In practice, one normally has a mixture of such eigenvalue distributions. Although these estimates have appeared in nearly the same form in [2, Chapter 13], for completeness we include them also here.

Example 1 (A special case of a compact perturbation).
Let the eigenvalues be $\lambda_j = 1 + 1/j, \quad j = 1, 2, \ldots, n$. Then

$$n^{-1} \sum_{l}^{n} \lambda_j \sim 1 + n^{-1} \ln n + cn^{-1} + O(n^{-2})\,, \ n \to \infty$$

for some positive constant c, and

$$(\prod_{j=1}^{n} \lambda_j)^{1/n} = (n+1)^{1/n}\,.$$

Hence the K-condition number for the whole set of n eigenvalues equals

$$K(B) = \left(1 + n^{-1} \ln n + cn^{-1} + O(n^{-2})\right)^n / (n+1),$$

so

$$\ln K(B) \sim \ln\, n + c + O(n^{-1}) - \ln\,(n+1)$$

and

$$\log_2 K(B) = (\log_2 e) \ln K(B) \sim c \log_2 e + O(n^{-1}) ,$$
$$\text{so} \quad K(B) = e 2^{c+1} + O(n^{-1}) ,$$

i.e., $K(B)^{1/k}$ approaches rapidly a value which is less than 2 already for small values of k. Therefore the superlinear rate of convergence is entered from the very beginning of the iterations.

Example 2 (Arithmetic distribution).
Let $\lambda_j = a + j \frac{b-a}{n}, \quad j = 1, 2, \ldots, n$.
For simplicity, assume that $n/2$ is even. Then

$$K(B) = \left(\frac{b+a}{2}\right)^n / \prod_{j=1}^{n} \lambda_j ,$$

where

$$\begin{aligned} \prod_{j=1}^{n} \lambda_j &= \prod_{j=1}^{n/2} \left(\frac{b+a}{2} - \frac{b-a}{2}\frac{2j}{n}\right)\left(\frac{b+a}{2} + \frac{b-a}{2}\frac{2j}{n}\right) \\ &= \left(\frac{b+a}{2}\right)^n \prod_{j=1}^{n/2} \left(1 - \left(\frac{b-a}{b+a}\frac{2j}{n}\right)^2\right) \end{aligned}$$

Here

$$K(B) > 1 / \prod_{j=1}^{n/2} \left(1 - \frac{1}{4}\left(\frac{b-a}{b+a}\right)^2\right)$$

and

$$\log_2 K(B) > \frac{n}{4} \log_2 \left(1 - \frac{1}{4}\left(\frac{b-a}{b+a}\right)^2\right)^{-1} > O(n),\ n \to \infty .$$

Therefore, in this case, the estimate based on $K(B)$ is inferior to the classical estimate, based on the spectral condition number, $K(B) = b/a$.

In an approximate sense, arithmetic eigenvalue distributions occur for an operator $M + \tau K$, which arises in time-stepping methods for an evolution equation. Here τ is a small time-step, M and K are mass and stiffness matrices, respectively. We then assume that the space domain is separated in vertical stripes and that the horizontal boundary points are located equidistantly.

Using an improved superlinear estimate, the case of arithmetic distribution will be studied further at the end of section 6.

Example 3 (Distribution proportional to a power sequence of the large outlier eigenvalues).

Following [2, p. 588], we estimate first the K-condition number for th
values in S_n distributed as $\lambda_j \sim j^\nu$ $(j = 1, 2, \ldots, n)$ for some $\nu > 1$.
such a distribution can model the eigenvalues of an elliptic boundary valu

well. For instance, the eigenvalues of the operator $-\varepsilon u_{xx} - u_{yy}$ on the unit square with homogeneous Dirichlet b.c., equal

$$\lambda_{k,l} = (\varepsilon k^2 + l^2)\pi^2 \qquad (k, l = 1, 2, \dots).$$

Hence, for small values ε, the smooth eigenvalues (for small k) arising for $l = 1, 2, \dots$ are approximately distributed as constant times l^2. For the corresponding difference matrix for small k, h, l, one has

$$\lambda_{k,l}^{(h)} = \frac{4}{h^2}\Big(\varepsilon \sin^2 \frac{k\pi h}{2} + \sin^2 \frac{l\pi h}{2}\Big) \approx \lambda_{k,l} = (\varepsilon k^2 + l^2)\pi^2.$$

It then holds

$$\frac{1}{n}\sum_{j=l}^{n} j^\nu \sim \frac{1}{\nu+1} n^\nu\,,\ n \to \infty\,.$$

Further, using Stirling's formula,

$$\prod_{j=1}^{n} \lambda_j = (\prod_{j=1}^{n} j)^\nu \sim (2\pi n)^{\nu/2} (\frac{n}{e})^{n\nu}\,, \quad n \to \infty\,.$$

so

$$K(S_n)^{1/n} \sim \frac{e^\nu}{\nu+1}\,, \quad n \to \infty\,. \tag{11}$$

This number is less than 2 only if $e^\nu < 2(\nu+1)$, which holds for a number slightly less than 2. Hence, unless ν is sufficiently small, the rate of convergence based on the K-condition number is less useful also for such an eigenvalue distribution.

Example 4 Assume now that there are p small eigenvalues and further that the large, outlier eigenvalues are distributed as $r\{j^\nu\}_{j=1}^m$, where $r \gg b$, i.e. are located in a separate interval. Then the estimate in (11) shows that there are p iterations plus the number of iterations $O\left((\frac{b}{a})^{1/2}\right)$, corresponding to the interval $[a, b]$. Let $k_0 = p + O\left((\frac{b}{a})^{1/2}\right)$ be the total of these.

The estimate in (11) shows that one enters a superlinear convergence phase with s_k additional iterations, when

$$\frac{1}{1-\xi_k} > 2^{p/s_k}\, K(S_k)^{1/s_k} \tag{12}$$

where

$$\begin{aligned}
\xi_k = \frac{\lambda_{m-s_k} + a}{2} \Big/ \frac{1}{s_k} \sum_{j=m-s_k}^{m} \lambda_j \ \sim \\
\frac{1}{2}(m-s_k)^\nu \Big/ \frac{1}{s_k(\nu+1)} \left(m^{\nu+1} - (m-s_k)^{\nu+1}\right) \\
\sim\ \frac{1}{2}(m-s_k)^\nu \Big/ \frac{m^{\nu+1}}{s_k(\nu+1)} \left(1 - (1 - \frac{s_k}{m})^{\nu+1}\right) \\
\sim\ \frac{1}{2}(m-s_k)^\nu / m^\nu \sim \frac{1}{2}\left(1 - \frac{s_k}{m}\right)^\nu \sim \frac{1}{2}\,,\ m \to \infty,\ s_k \ll r
\end{aligned}$$

It follows readily that since $s_k \ll m$, then $K(S_k) \sim 1$.

Hence, it follows from (12) that one enters the superlinear convergence phase at least when $s_k > p$, that is, for a total number of iterations of $p + k_0 = 2p + O((b/a)^{1/2})$.

This estimate can be useful when $a >> \lambda_1$ and $p << \sqrt{b/\lambda_1}$. We note that in practice one normally uses a preconditioner for the above problem. If one uses an incomplete preconditioning method without modifications (see [2]), then the above results are still applicable.

5 Further estimates for clustering eigenvalues

In this section we give estimates for clustering eigenvalues using the superlinear bound (4) instead of the K-condition number. We consider first a specific and then a more general distribution, and the goal is to show that in such cases σ_n does not increase as $n \to \infty$. We note that in this section it is more convenient to write the eigenvalues in decreasing order.

5.1 An example: eigenvalues proportional to a power sequence

In this first example, we consider eigenvalues clustering around ϱ with a difference proportional to a power sequence. This is an extension of Example 1 of section 4. For such distributions one has a uniform superlinear estimate [2, 7], hence the iteration is expected to enter the superlinear phase quickly. Indeed, as seen below, the index of entering is independent of the matrix size, and for powers not smaller than 1 one can consider the superlinear phase as almost immediate.

Proposition 5.1 *Let $A = \varrho I + E$, let μ_j $(j = 1, 2, \dots)$ be the eigenvalues of E and $R := \|E\|$. Let $\alpha > 0$ be given,*

$$\mu_j := \frac{R}{j^\alpha} \quad (j \in 1, \dots, n).$$

Then $\sigma_n \le 2^{1/\alpha} \cdot e + 1$ independently of n. In particular, if $\alpha \ge 1$ then $\sigma_n \le 5$.

Proof. Here the factor 2^k can be omitted in (4) (and hence also on the l.h.s. of (6)), since all μ_j are positive, hence the estimate $|\mu_j - \mu_i| \le 2|\mu_j|$ used there can be replaced by $|\mu_j - \mu_i| \le \mu_j$. Then for all k, using that $k! \ge (k/e)^k$,

$$\prod_{j=1}^{k} \frac{\mu_j}{\varrho + \mu_j} = \prod_{j=1}^{k} \frac{R}{\varrho j^\alpha + R} \le \Big(\frac{R}{\varrho}\Big)^k \prod_{j=1}^{k} \frac{1}{j^\alpha} = \Big(\frac{R}{\varrho}\Big)^k \frac{1}{(k!)^\alpha} \le \Big(\frac{Re^\alpha}{\varrho k^\alpha}\Big)^k.$$

On the other hand, the quotient on the right of (6) appears on the l.h.s. of (16) (to come below), which shows that it is not less than $R/2\varrho$. Hence, replacing the quotient on the right of (6) by $R/2\varrho$, it suffices that

$$\Big(\frac{Re^\alpha}{\varrho k^\alpha}\Big)^k \le 2\Big(\frac{R}{2\varrho}\Big)^k, \quad \text{that is} \quad \Big(\frac{2e^\alpha}{k^\alpha}\Big)^k \le 2.$$

Let σ_n denote the integer part of $2^{1/\alpha} \cdot e + 1$. Then all $k \geq \sigma_n$ satisfy $k \geq 2^{1/\alpha} \cdot e$, hence $\frac{2e^\alpha}{k^\alpha} \leq 1$, which ensures (13). If $\alpha \geq 1$ then for all $k \geq 5$ we have $(\frac{2e}{k})^k \leq 2$, and here $\frac{e}{k} \leq 1$, hence we obtain (13). ∎

5.2 General clustering of eigenvalues

One can consider any sequence of numbers that tends to 0 and the eigenvalues of the given matrix are the first n terms of this sequence. Then the following holds:

Proposition 5.2 *Let $(\mu_j) \subset \mathbf{R}$ be a sequence such that $|\mu_j| \to 0$ monotonically as $j \to \infty$. If the eigenvalues of the $n \times n$ matrix E are $\mu_1, \ldots, \mu_n$, then σ_n is bounded independently of n.*

Proof. Rewriting (6), we need

$$2\Big(\prod_{j=1}^{k} \frac{|\mu_j|}{\varrho + \mu_j}\Big)^{1/k} \leq 2^{1/k}\, \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \qquad (k \geq \sigma_n). \tag{14}$$

The l.h.s. is twice the geometric mean sequence of $\frac{|\mu_j|}{\varrho+\mu_j}$. The latter tends to 0 as $j \to \infty$, hence so does the l.h.s. On the other hand, the r.h.s. has the limit $\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}$, which is positive (unless the trivial case when A is a constant times identity). Hence the l.h.s. becomes smaller after sufficiently large k, say, for $k \geq \sigma$ where $\sigma \in \mathbf{N}^+$ is independent of k and n. That is, for given $n \geq \sigma$ the l.h.s. of (14) is smaller than its r.h.s. if $\sigma \leq k \leq n$, i.e. $\sigma_n \leq \sigma$. Hence σ_n is bounded for $n \geq \sigma$, and is obviously also bounded for $n = 1, \ldots, \sigma$ (since the latter is a finite range). ∎

6 Estimates for uniformly distributed eigenvalues

In this section we study uniformly distributed eigenvalues, i.e. with no clustering around a particular point. We will find that the obtained bound on σ_n now grows unboundedly as $n \to \infty$.

In the case without clustering, there is no special a priori value of ϱ. Hence, in order to decompose a given A as in (3), it is natural to define a symmetric choice of ϱ:

$$\varrho := \frac{M+m}{2}, \qquad \text{then} \quad \|E\| = \max|\mu_i| = \frac{M-m}{2} =: R,$$

i.e. R is the maximal deviation of the eigenvalues of A from ϱ. Here

$$\varrho + R = M, \qquad \varrho - R = m \qquad \text{and hence} \quad R < \varrho. \tag{15}$$

The corresponding reformulation of the linear convergence quotient on the right of (6) is

$$\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} = \frac{\sqrt{\varrho + R} - \sqrt{\varrho - R}}{\sqrt{\varrho + R} + \sqrt{\varrho - R}} = \frac{R}{\varrho + \sqrt{\varrho^2 - R^2}}.$$

Assumption 6.1: the values of $|\mu_j|$ (ordered with nonincreasing moduli) follow a common distribution for all n, i.e. there exists a 'density' function $d : [0,1] \to \mathbf{R}^+$ such that

$$|\mu_j| = d\big(\frac{j}{n}\big) \qquad (j = 1, 2, ..., n). \tag{17}$$

There is no limitation in assuming that d is differentiable,

$$d(0) = R, \quad d' \le 0. \tag{18}$$

(We note that the following results also hold if (17) is valid only asymptotically as $n \to \infty$.)

6.1 Estimates using the arithmetic mean

Here, using (5), inequality (6) is modified to

$$\left(\frac{2\|A^{-1}\|}{k} \sum_{j=1}^{k} |\mu_j|\right)^k \le 2\left(\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}\right)^k \qquad (k \ge \sigma_n). \tag{19}$$

This leads to a rougher estimate of σ_n, but is easier to handle, and it will still provide the characteristic asymptotics of σ_n as $n \to \infty$.

The arithmetic mean of the $|\mu_j|$ will be denoted by

$$a_k := \frac{1}{k} \sum_{j=1}^{k} |\mu_j| \,.$$

Let us rewrite (19) as follows. Using $\|A^{-1}\| = \frac{1}{m} = \frac{1}{\varrho - R}$ and (16), taking the kth root of (19) and reordering, we must have

$$a_k \le 2^{1/k} \, \frac{\varrho - R}{2} \cdot \frac{R}{\varrho + \sqrt{\varrho^2 - R^2}} \qquad (k \ge \sigma_n). \tag{20}$$

The basic idea is as follows: since $a_k := \frac{1}{k} \sum_{j=1}^{k} |\mu_j| = \frac{1}{k} \sum_{j=1}^{k} d\big(\frac{j}{n}\big) = \frac{1}{k} \sum_{j=1}^{k} d\big(\frac{k}{n} \cdot \frac{j}{k}\big)$, we have

$$a_k \approx I_k := \int_0^1 d\big(\frac{k}{n}\, x\big) \, \mathrm{d}x,$$

and, as is easily seen, denoting $\delta := \max\limits_{[0,1]} |d'|$, one has

$$|a_k - I_k| \le \frac{\delta}{2k}. \tag{21}$$

Now let us introduce the following notations:

$$D(x) := \int_0^x d(t) \, \mathrm{d}t, \qquad \delta(x) := \frac{D(x)}{x} \qquad (0 < x \le 1).$$

Here $\delta(0) := \lim_0 \delta(x) = D'(0) = d(0)$. Then

$$I_k = \int_0^1 d\big(\frac{k}{n}\, x\big)\, \mathrm{d}x = \frac{n}{k} \cdot D\big(\frac{k}{n}\big) = \delta\big(\frac{k}{n}\big),$$

hence by (20) and (21) we need

$$\delta\big(\frac{k}{n}\big) \le -\frac{\delta}{2k} + 2^{1/k}\, \frac{\varrho - R}{2} \cdot \frac{R}{\varrho + \sqrt{\varrho^2 - R^2}} \qquad (k \ge \sigma_n) \tag{22}$$

to ensure that (19) holds.

Proposition 6.1 *Under assumption* (18) *the function* δ *decreases.*

Proof. We have

$$(x\, d(x))' = d(x) + x\, d'(x) \le d(x) = D'(x),$$

hence

$$x\, d(x) \le D(x) \qquad (0 \le x \le 1)$$

since they coincide at 0. Therefore $d(x) \le \delta(x)$. We then have for all x

$$d(x) = D'(x) = (x\, \delta(x))' = \delta(x) + x\, \delta'(x) \ge d(x) + x\, \delta'(x),$$

hence $\delta'(x) \le 0$. ∎

Theorem 6.1 *Let us assume* (17)-(18), *and let*

$$\int_0^1 d(t)\, \mathrm{d}t < \ P_0 := \ \frac{\varrho - R}{2} \cdot \frac{R}{\varrho + \sqrt{\varrho^2 - R^2}}. \tag{23}$$

Then there exist $0 < c_0 < 1$ *and* $N_0 \in \mathbf{N}^+$ *such that if* $n \ge N_0$ *then*

$$\sigma_n \le c_0 \cdot n.$$

Proof. Proposition 6.1 implies that the minimum of δ on $[0,1]$ is $\delta(1) = \int_0^1 d(t)\, \mathrm{d}t$, i.e. we have $\min \delta < P_0$. Let us choose some β such that $\min \delta < \beta < P_0$. First, since δ decreases on $[0,1]$, there exists $0 < c_0 < 1$ such that $\delta(x) \le \beta$ if $x \ge c_0$, therefore

$$\delta\big(\frac{k}{n}\big) \le \beta \quad \text{if} \quad k \ge c_0 \cdot n.$$

Further, the r.h.s. of (22) tends to P_0 as $k \to \infty$, hence there exists $k_0 \in \mathbf{N}^+$ such that it becomes at least β if $k \ge k_0$. Hence (22) is valid if k satisfies both of the above, i.e. if we choose $N_0 \in \mathbf{N}^+$ such that $c_0 N_0 \ge k_0$, and let $n \ge N_0$ and $k \ge c_0 \cdot n$ be ar

Example: *Power order distribution.*

(a) Let

$$|\mu_j| = R\Big(\frac{n-j}{n}\Big)^\alpha \quad (j=1,2,...,n), \quad \text{i.e.} \qquad d(x) = R(1-x)^\alpha \quad (x\in[0,1]). \tag{24}$$

To a certain extent such eigenvalues still cluster at 0 if $\alpha > 1$, but for fixed j the value of $|\mu_j|$ grows with n in contrast to the assumptions section 5.

Then $\int_0^1 d(t)\,\mathrm{d}t = \frac{R}{\alpha+1}$, hence if $\alpha + 1 > \frac{R}{P_0}$ then (23) holds and hence Theorem 6.1 is valid.

(b) In particular, if we consider a linear distribution (which is the most uniform distribution), i.e. $\alpha = 1$ and $d(x) := R(1-x)$, then the following difficulty is met. Let us vary R between 0 and ϱ, and study the multiplier of R on the r.h.s. of (23). First, it equals $\frac{1}{4}$ if $R = 0$. Further, it is easy to see that it decreases as $R \to \varrho$ and becomes 0 for $R = \varrho$, i.e. it is at most $\frac{1}{4}$, whereas the l.h.s. of (23) is $\int_0^1 d = \frac{R}{2}$, which is impossible.

In the sequel we return to sharper estimates than the quite rough one with arithmetic mean, so as to see if there is still a superlinear phase for distributions close to linear.

6.2 Logarithmic estimates

Although the CG estimates in (5) yield a nice and easy bound, they contain steps which can be rough, hence we return to estimate (4). Moreover, one can improve it by a factor up to $\sqrt{2}$. Namely, let us divide the eigenvalues into two groups:

$$\mu_j^+ > 0 \quad (j=1,\dots,n^+), \qquad \mu_1^+ \ge \cdots \ge \mu_{n^+}^+ > 0,$$
$$\mu_j^- < 0 \quad (j=1,\dots,n^-), \qquad |\mu_1^-| \ge \cdots \ge |\mu_{n^-}^-| > 0.$$

For given $k \le n$, we consider the first k eigenvalues with greatest moduli and denote by k^+ and k^- the number of positive and negative ones, respectively, and finally $k^* := \max\{k^+, k^-\}$. Then in (4) the estimate $|\mu_j - \mu_i| \le 2|\mu_j|$ can be replaced by $|\mu_j - \mu_i| \le \mu_j$, respectively $|\mu_j|$, for positive and negative values of μ_j, i.e. without factor 2, whenever μ_j and μ_i have the same sign. Thus the r.h.s. of (4) can be replaced by modifying (4) as

$$\frac{\|e_k\|_A}{\|e_0\|_A} \le \max_{i\ge k+1} \prod_{j=1}^{k} \frac{|\mu_j - \mu_i|}{\varrho + \mu_j} \le 2^{k^*} \prod_{j=1}^{k} \frac{|\mu_j|}{\varrho + \mu_j} \tag{25}$$

where $k/2 \le k^* \le k$.

Now, if we use this estimate instead of the arithmetic mean, then we must replace the term

$$\frac{2\|A^{-1}\|}{k} \sum_{j=1}^{k} |\mu_j| = \frac{2}{\varrho - R}\, a_k \qquad \text{by} \qquad 2^{k^*/k} \left(\prod_{j=1}^{k} \frac{|\mu_j|}{\varrho + \mu_j} \right)^{1/k} =: 2^{k^*/k}\, b_k$$

in (20), which then becomes

$$b_k \le 2^{-\frac{k^*-1}{k}} \cdot \frac{R}{\varrho + \sqrt{\varrho^2 - R^2}} \qquad (k \ge \sigma_n).$$

Here we have

$$b_k = \left(\prod_{j=1}^{k} \frac{|\mu_j|}{\varrho + \mu_j}\right)^{1/k} = \exp\Big(\frac{1}{k}\sum_{j=1}^{k} \ln \frac{|\mu_j|}{\varrho + \mu_j}\Big),$$

and now the exponent will be approximated by logarithmic integrals, in the vein of subsection 6.1.

First, let us replace (17) by two functions. We assume that there exist two 'density' functions in the following way. Let $\delta^+ := \frac{n^+}{n}$ and $\delta^- := \frac{n^-}{n}$, then $0 < \delta^\pm < 1$ and $\delta^+ + \delta^- = 1$. The functions $d^+ : [0, \delta^+] \to \mathbf{R}^+$ and $d^- : [0, \delta^-] \to \mathbf{R}^+$ must then satisfy

$$\mu_j^+ = d^+\big(\frac{j}{n}\big) \quad (j = 1, 2, ..., n^+), \qquad |\mu_j^-| = d^-\big(\frac{j}{n}\big) \quad (j = 1, 2, ..., n^-). \tag{27}$$

As in (18), we assume that $d^\pm$ are differentiable and

$$(d^\pm)' \le 0. \tag{28}$$

For given $k \le n$, we consider the first k eigenvalues with greatest moduli and denote by k^+ and k^- the number of positive and negative ones as before. Then

$$\frac{1}{k}\sum_{j=1}^{k} \ln \frac{|\mu_j|}{\varrho + \mu_j} = \frac{1}{k}\sum_{j=1}^{k^+} \ln \frac{\mu_{j^+}}{\varrho + \mu_{j^+}} + \frac{1}{k}\sum_{j=1}^{k^-} \ln \frac{|\mu_{j^-}|}{\varrho + \mu_{j^-}}.$$

Here for both terms we can repeat the arguments of subsection 6.1. For this, let us introduce the following notations:

$$g^\pm(x) := \ln \frac{d^\pm(x)}{\varrho \pm d^\pm(x)}, \qquad G^\pm(x) := \int_0^x g^\pm(t)\,\mathrm{d}t, \qquad \gamma^\pm(x) := \frac{G^\pm(x)}{x} \qquad (0 < x \le \delta^\pm).$$

Now, first,

$$\frac{1}{k}\sum_{j=1}^{k^+} \ln \frac{\mu_{j^+}}{\varrho + \mu_{j^+}} = \frac{1}{k}\sum_{j=1}^{k^+} \ln \frac{d^+\big(\frac{j}{n}\big)}{\varrho + d^+\big(\frac{j}{n}\big)} \approx \int_0^{\frac{k^+}{k}} g^+\big(\frac{k}{n}\,x\big)\,dx = \frac{n}{k}\,G^+\big(\frac{k^+}{n}\big) = \frac{k^+}{k}\,\gamma^+\big(\frac{k^+}{n}\big),$$

and similarly for the second term, hence

$$b_k \approx J_k := \exp\Big(\frac{k^+}{k}\cdot\gamma^+\big(\frac{k^+}{n}\big) + \frac{k^-}{k}\cdot\gamma^-\big(\frac{k^-}{n}\big)\Big) \tag{29}$$

where one in fact has again

$$|b_k - J_k| \le \frac{\hat\delta}{2k}. \tag{30}$$

Here the functions $t \mapsto \ln \frac{t}{\varrho \pm t}$ are increasing, and by (28) $d^\pm$ are decreasing, hence their compositions $g^\pm$ are decreasing. Then Proposition 6.1 implies that the similarly derived functions $\gamma^\pm$ are decreasing too, and hence have their minima at $\delta^\pm = \frac{n^\pm}{n}$. Assume now also that the positive and negative eigenvalues are distributed asymptotic[illegible] i.e.

$$\lim \frac{k^\pm}{k} = \delta^\pm$$

uniformly in n as $n \to \infty$. Then also $\frac{k^*}{k} \to \delta^* := \max(\delta^+, \delta^-)$, and following the proof of Theorem 6.1, for $k \geq k_0$ the r.h.s. of (29) satisfies

$$J_k \leq \hat{\beta} < \hat{\beta}_0 := \exp\Big(\delta^+ \cdot \gamma^+\big(\delta^+\big) \ + \delta^- \cdot \gamma^-\big(\delta^-\big)\Big) = \exp\Big(G^+\big(\delta^+\big) \ + G^-\big(\delta^-\big)\Big) \qquad (31)$$

if $k \geq \hat{c}_0 n$. Comparing with (26), and since $\lim 2^{-\frac{k^*-1}{k}} = 2^{-\delta^*}$, we obtain the following result: if

$$\hat{\beta}_0 < 2^{-\delta^*} \ \cdot \frac{R}{\varrho + \sqrt{\varrho^2 - R^2}}, \qquad (32)$$

then (26) holds for $n \geq N_0$ and $k \geq c_0 \cdot n$.

One can rewrite $\hat{\beta}_0$ in a more visible form. Namely, let us define the function $\tilde{d} : [0,1] \to \mathbf{R}^+$ as

$$\tilde{d}(x) := \begin{cases} -d^-(x) & \text{if } \ x \in [0,\, \delta^-]; \\ d^+(x - \delta^+) & \text{if } \ x \in [\delta^-,\, 1]. \end{cases} \qquad (33)$$

(Then by (28), $\tilde{d}$ is increasing on $[0,1]$.) Here

$$\hat{\beta}_0 = \exp\Big(\int_0^{\delta^-} \ln \frac{d^-(t)}{\varrho - d^-(t)}\, \mathrm{d}t + \int_0^{\delta^+} \ln \frac{d^+(t)}{\varrho + d^+(t)}\, \mathrm{d}t\Big)$$

$$= \exp\Big(\int_0^{\delta^-} \ln \frac{|\tilde{d}(t)|}{\varrho + \tilde{d}(t)}\, \mathrm{d}t + \int_{\delta^-}^1 \ln \frac{\tilde{d}(x)}{\varrho + \tilde{d}(x)}\, \mathrm{d}x\Big) = \exp\Big(\int_0^1 \ln \frac{|\tilde{d}(x)|}{\varrho + \tilde{d}(x)}\, \mathrm{d}x\Big)$$

where in the first term the transformation $x = t + \delta^-$ has been used. Hence, altogether, we have proved:

Theorem 6.2 *Let us assume* (27)-(28)*, and let*

$$\exp\Big(\int_0^1 \ln \frac{|\tilde{d}(x)|}{\varrho + \tilde{d}(x)}\, dx\Big) < 2^{-\delta^*} \ \cdot \frac{R}{\varrho + \sqrt{\varrho^2 - R^2}}. \qquad (34)$$

Then there exist $0 < c_0 < 1$ *and* $N_0 \in \mathbf{N}^+$ *such that if* $n \geq N_0$ *then*

$$\sigma_n \leq c_0 \cdot n.$$

Remark 6.1 To apply Theorem 6.2, it suffices to order all eigenvalues monotonically from the extreme negative to extreme positive values, i.e., using the function $\tilde{d} : [0,1] \to \mathbf{R}^+$ in (33), we write $\mu_j = \tilde{d}\big(\frac{j}{n}\big) \ \ (j = 1, 2, ..., n)$.

Remark 6.2 Let us now consider the special case when the spectrum of E is symmetric w.r.t. zero, i.e. if $n^+ = n^-$ and

$$\mu_j^- = -\mu_j^+ \qquad (j = 1, \ldots, n^+).$$

Then the factor 2^{k^*} in (25) can be further improved. Namely, if k is even then (25) can be replaced by

$$\max_{i\ge k+1}\prod_{j=1}^{k}\frac{|\mu_j-\mu_i|}{\varrho+\mu_j}=\max_{i\ge k+1}\prod_{j=1}^{k/2}\frac{|(\mu_j^+-\mu_i)(\mu_j^++\mu_i)|}{(\varrho+\mu_j^+)(\varrho+\mu_j^-)}\le\prod_{j=1}^{k/2}\frac{|\mu_j^+\mu_j^-|}{(\varrho+\mu_j^+)(\varrho+\mu_j^-)}=\prod_{j=1}^{k}\frac{|\mu_j|}{\varrho+\mu_j},$$

where we have used the inequality

$$|(\mu_j^++\mu_i)(\mu_j^+-\mu_i)|=(\mu_j^+)^2-\mu_i^2\le(\mu_j^+)^2=|\mu_j^+\mu_j^-|.$$

That is, the factor 2^{k^*} disappears. If k is odd then the above estimate is used for $j\le k-1$ and the original estimate $|\mu_j-\mu_i|\le 2|\mu_j|$ is used for $j=k$, hence the product is multiplied by a factor 2. Altogether, for any k we obtain

$$\frac{\|e_k\|_A}{\|e_0\|_A}\le\max_{i\ge k+1}\prod_{j=1}^{k}\frac{|\mu_j-\mu_i|}{\varrho+\mu_j}\le 2\prod_{j=1}^{k}\frac{|\mu_j|}{\varrho+\mu_j},\tag{35}$$

i.e. k^* is replaced by 1 in the exponent of 2. Then the factor $\lim 2^{-\frac{k^*-1}{k}}=2^{-\delta^*}$ in (32) is replaced by $2^0=1$, i.e. (32) is replaced by

$$\hat\beta_0<\frac{R}{\varrho+\sqrt{\varrho^2-R^2}}.\tag{36}$$

Accordingly, Theorem 6.2 is modified as follows: if

$$\exp\Big(\int_0^1\ln\frac{|\tilde d(x)|}{\varrho+\tilde d(x)}\,dx\Big)<\frac{R}{\varrho+\sqrt{\varrho^2-R^2}},\tag{37}$$

then there exist $0<c_0<1$ and $N_0\in\mathbf{N}^+$ such that if $n\ge N_0$ then $\sigma_n\le c_0\cdot n$.

Example: *Linear distribution of eigenvalues.* This corresponds to the function

$$\tilde d(x):=R(2x-1),$$

which is symmetric w.r.t. $1/2$, and its values on $[0,1]$ grow from $-R$ to R. One can show with elementary but tedious calculations that condition (37) is then satisfied, and hence the statement of Theorem 6.2 holds.

Indeed, using notation $p:=\frac{m}{R}$, one can then verify that

$$\int_0^1\ln\frac{|\tilde d(x)|}{\varrho+\tilde d(x)}\,dx=\frac{1}{2}\Big(p\ln p-(p+2)\ln(p+2)\Big)\qquad\text{and}$$

$$\frac{R}{\varrho+\sqrt{\varrho^2-R^2}}=1/\Big(p+1+\sqrt{p(p+1)}\Big),$$

i.e. for (37) one must prove

$$-\frac{1}{2}\Big(p\ln p-(p+2)\ln(p+2)\Big)>\ln\Big(p+1+\sqrt{p(p+1)}\Big)\qquad(p>0).$$

Using the transformation $t:=\sqrt{\frac{p+2}{p}}$ and with some calculation, the above inequality becomes equivalent to the condition

$$\psi(t):=\frac{2t^2}{t^2-1}\ln t-2\ln(1+t)+\ln 2>0\quad(t>1).$$

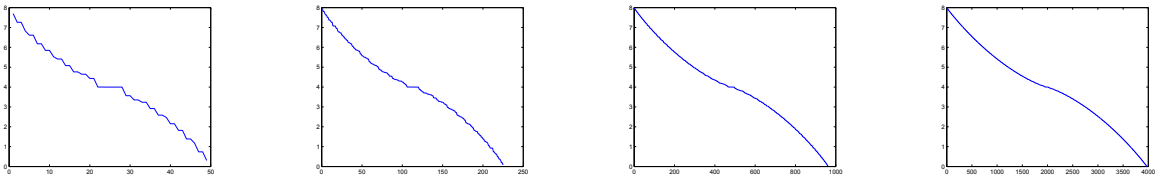
One can verify that $\psi'>0$ and $\lim_{t\to 1+}\psi(t)=1-\ln 2\approx 0.307$, which justifies the desired inequality.

7 Numerical illustration

To compare with the theoretical estimates, some numerical experiments have been performed on the unit square.

First we consider as test problem the Poisson equation with homogeneous Dirichlet boundary condition, discretized on a uniform grid. The eigenvalue distribution of the arising discrete Laplacian satisfies Assumption 6.1 approximately, as shown by Figure 1.

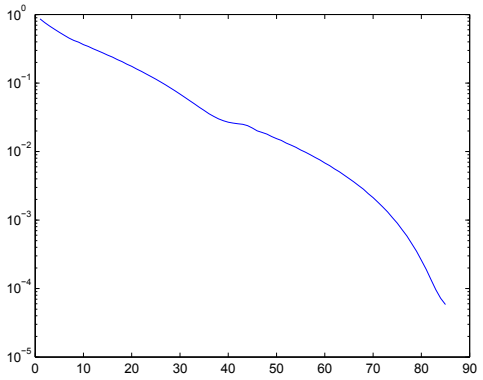
Figure 1: Eigenvalue distribution for the discrete Laplacian



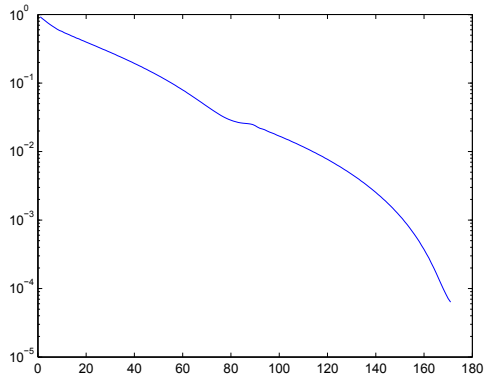
(a) $1/h=8$ (b) $1/h=16$ (c) $1/h=32$ (d) $1/h=64$

The theorems of section 6 state that the entering index σ_n can grow at most proportionally to the matrix size n. The experiments show that the shape of the curves looks very similar, but with growing number of iterations for σ_n to more markedly enter a superlinear convergence phase, see Figure 2. It is seen that σ_n grows slower than the bound, namely it is proportional to $1/h=O(\sqrt{n})$, i.e., as is well-known and follows from [2], it grows at the same rate as the upper bound for the number of iterations of the CG method to reach a fixed relative accuracy.

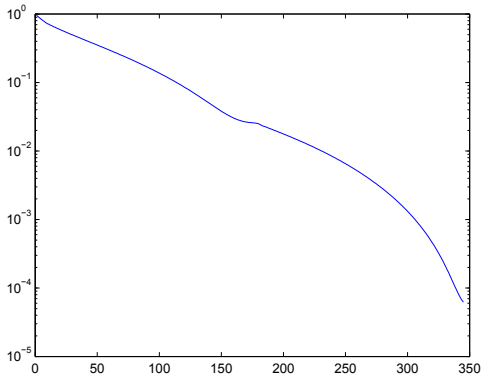
Figure 2: Convergence history, discrete Laplacian, $\varepsilon_n = \|e_n\|_A / \|e_0\|_A$,



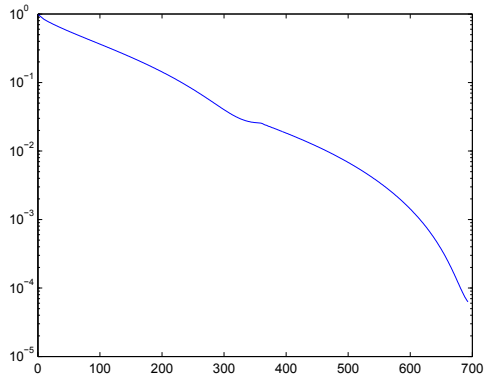
(a) $1/h = 64$



(b) $1/h = 128$



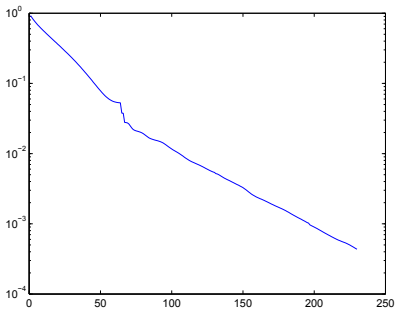
(c) $1/h = 256$



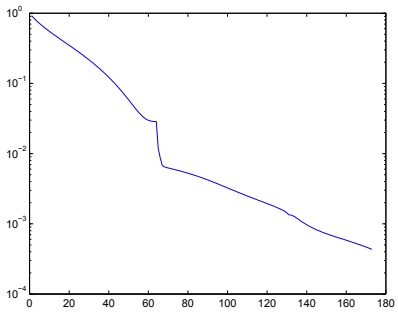
(d) $1/h = 512$

Now we consider an anisotropic equation $-\varepsilon u_{xx} - u_{yy} = f$ instead of the Poisson equation, and vary ε. For $1/h = 128$ and tolerance 10^{-4}, the results are shown in Figure 3.

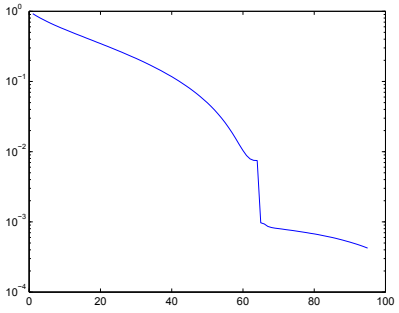
Figure 3: Convergence history, $-\varepsilon u_{xx} + u_{yy} = f$; $1/h = 128$, $TOL = 10^{-4}$



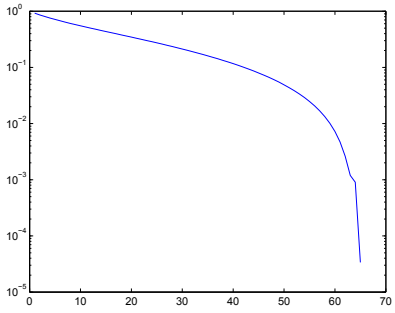
(a) $\varepsilon = 0.01$



(b) $\varepsilon = 0.001$



(c) $\varepsilon = 0.0001$



(d) $\varepsilon = 0.00001$

These tests show that for not very small values of the parameter ε, the iterations stay long in the sublinear convergence phase. As was indicated in Section 3, this can be explained by the presence of many small eigenvalues, since for each small eigenvalue corresponding to the regular term $-u_{yy}$ in the operator, there corresponds several of the larger eigenvalues for the singular perturbation term $-\varepsilon u_{xx}$, typically of the order ε/h^2 or smaller.

However, when ε decreases further, the iteration process first sees only the small eigenvalues corresponding to the regular term; on the other hand, when the corresponding iteration errors are sufficiently damped out, one enters a very sharp, i.e. short-lived superlinear convergence phase. This gets interrupted when the iteration errors get so small that they become influenced also by the eigenvalues corresponding to the singular perturbation term. After some further iterations one then enters a second superlinear convergence phase.

If ε is very small, then the iteration process is not influenced at all by the perturbation of the eigenvalues corresponding to the singular perturbation term, until one has reached a very small relative iteration error . (The latter phase is not seen in the figure since the tolerance was achieved before that.)

As can be seen, our analytical estimates of σ_n are somewhat rough, but in general give at least some indication where the superlinear rate of convergence is reached. Except for the anisotropic singular perturbation problem, there is no sharp point where this occurs, but one enters the superlinear phase gradually. This was also indicated in Section 3.

In our examples the number of iterations are not very large and no reorthogonalization has been used. We remark that for extremely large number of iterations, rounding errors due to finite precision arithmetic may influence the results, making the theoretical estimates less valuable.

8 Conclusions

In order to examine when the superlinear rate of convergence starts in a conjugate gradient iteration, two types of methods have been used. One method, based on the K-condition number, shows results of less interest in general, since this number appears to be greater than the order n of the system for certain distributions of eigenvalues. Using refined estimates, based on the annihilating polynomial for the eigenvalues, and assuming various forms of the distributions of eigenvalues, it has been shown that one can get sharper and more interesting results. The theoretical results are completed by the numerical tests.

Acknowledgements. The authors gratefully acknowledge the help of Tamás Kurics for providing the numerical experiments. The work of the first author was supported by the European Regional Development Fund in the IT4 Innovations Centre of Excellence project (CZ E.1.05/1.1.00/02.0070).

References

[1] O. Axelsson, Solution of linear systems of equations: iterative methods. In ed. V. A. Barker, *Sparse Matrix Techniques*, Berlin, Heidelberg, New York: Springer, Verlag, 1976.

[2] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, 1994.

[3] O. Axelsson, V. A. Barker, *Finite Element Solution of Boundary Value Problems*. Theory and Computation. SIAM Classics in Applied Mathematics 35 SIAM Philadelphia, 2001.

[4] O. Axelsson, I. Kaporin, On the sublinear and superlinear rate of convergence of conjugate gradient methods, *Numer. Algor.* 25 (2000), no. 1-4, 1–22.

[5] O. Axelsson, J. Karátson, On the rate of convergence of the conjugate gradient method for linear operators in Hilbert space, *Numer. Funct. Anal.* 23 (2002), no. 3-4, 285-302.

[6] O. Axelsson, J. Karátson, Superlinearly convergent CG methods via equivalent preconditioning for nonsymmetric elliptic operators, *Numer. Math.* 99 (2004), no. 2, 197–223.

[7] O. Axelsson, J. Karátson, Mesh independent superlinear PCG rates via compact-equivalent operators, *SIAM J. Numer. Anal.*, 45 (2007), no. 4, 1495-1516.

[8] O. Axelsson, J. Karátson, Equivalent operator preconditioning for elliptic problems, *Numer. Algor.* 50 (2009), 297–380.

[9] O. Axelsson, G. Lindskog, On the eigenvalue distribution of a class of preconditioning methods, *Numer. Math.* 48 (1986), no. 5, 479–498.

[10] A. Greenbaum, *Iterative methods for solving linear systems*, Frontiers in Applied Mathematics, 17, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[11] A. Greenbaum, Comparison of splittings used with the conjugate gradient algorithm, *Numer. Math.* 33 (1979), no. 2, 181–193.

[12] M. R. Hestenes, E. Stiefel, Methods of conjugate gradients for solving linear systems, *J. Research Nat. Bur. Standards* 49 (1952), 409–436.

[13] I. E. Kaporin, An alternative approach to the estimation of the number of iterations in the conjugate gradient method (Russian), *Numerical methods and software* (Russian), 55–72, Akad. Nauk SSSR, Otdel Vychisl. Mat., Moscow, 1990.

[14] I. E. Kaporin, New convergence results and preconditioning strateg
jugate gradient method, *Numer. Linear Algebr. Appl.* 1 (1994), no. 2

[15] O. Nevanlinna, *Convergence of iterations for linear equations,* Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 1993.

[16] Y. Saad, *Iterative methods for sparse linear systems* (second edition), Society for Industrial and Applied Mathematics, Philadelphia, PA, 2003.

[17] A. van der Sluis, H. van der Vorst, The rate of convergence of conjugate gradients, *Numer. Math.* 48 (1986), 543-560.

[18] Strakoš, Z., Tichý, P., On error estimation in the conjugate gradient method and why it works in finite precision computations, *Electron. Trans. Numer. Anal.* 13, 56-80, 2002.

[19] R. Winther, Some superlinear convergence results for the conjugate gradient method, *SIAM J. Numer. Anal.*, 17 (1980), 14-17.