DEPARTMENT OF MATHEMATICS
UNIVERSITY OF NIJMEGEN The Netherlands

ON THE RATE OF CONVERGENCE OF THE CONJUGATE GRADIENT METHOD FOR LINEAR OPERATORS IN HILBERT SPACE

O. Axelsson and J. Karátson

Report No. 0102 (February 2001)

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF NIJMEGEN
Toernooiveld
6525 ED Nijmegen
The Netherlands

On the rate of convergence of the conjugate gradient method for linear operators in Hilbert space*

O. Axelsson† and J. Karátson‡

Abstract

The rate of convergence of the conjugate gradient method is investigated in Hilbert space. Previous results in $\mathbf{R}^n$ for the sublinear and superlinear rate of convergence, involving various generalized condition numbers, are extended to linear operators. Applications are given to elliptic differential operators, yielding relevant mesh independent estimates including large matrix sizes in the case of discretized boundary value problems.

1 Introduction

The conjugate gradient method has become the most widespread way of solving symmetric positive definite linear algebraic systems since it was first presented in [6]. One of its the most important features is superlinear convergence, first proved in [5] (see also [10]).

A characterization of the convergence of the CGM is given in the book [1] (Chapter 13) and the paper [2]. There are three typical phases: sublinear, linear and superlinear. The description of the different phases relies on different condition numbers. Whereas the standard condition number yields the classical linear convergence result, the other two phases use generalized condition numbers. Namely, the sublinear estimate uses a condition number relative to the initial vector, and the superlinear one uses the K-condition number involving all the eigenvalues of B.

Our aim is to extend the estimates of [1] and [2] to linear operators in Hilbert space. The motivation for this is given by differential operators. In their case the spd matrices obtained from discretization are approximations of the original operator that

*This research was done during the second author's Hungarian post-doc scholarship Magyary Zoltán.

†Department of Mathematics, University of Nijmegen, 6525ED Nijmegen, The Netherlands; axelsson@sci.kun.nl

‡Department of Applied Analysis, ELTE University, H-1053 Budapest, Hungary; karatson@cs.elte.hu

describes the studied model exactly. Hence the study of the CGM for these operators helps the understanding of the CGM for the discretized problems and, especially, gives relevant estimates for large matrix sizes.

The extension of the linear convergence result to Hilbert space has been long known in the same form [4, 5] and can be considered classical. Hence our investigation concerns the other two (sublinear and superlinear) phases.

2 The rate of convergence in $\mathbf{R}^n$

In this section we recall the major convergence results for the CGM for linear systems

$$Ax = a$$

in $\mathbf{R}^n$, where A is a real symmetric spd (symmetric and positive definite) matrix. In general we consider a preconditioned version

$$Bx = b \tag{1}$$

with $B = C^{-1}A$ and $b = C^{-1}a$, where C is also an spd matrix.

The theorems to follow are quoted from [2]. They describe the rate of convergence in the three phases. The estimates use the minimizing property of the CGM:

$$\|e^k\|_A = \min_{P_k \in \pi_k^1} \|P_k(B)e^0\|_A \qquad (k \in \mathbf{N}), \tag{2}$$

where π_k^1 denotes the set of polynomials of degree k such that $P_k(0) = 1$; further, the inner product $\langle x, y\rangle_A = \langle Ax, y\rangle$ and the corresponding norm are used.

(a) The sublinear phase

The estimates rely on the following notion [2]. Let A and B be symmetric and positive definite matrices. The generalized condition number of A with respect to a vector $x \in \mathbf{R}^n$ and the number $\nu \in \mathbf{R} \setminus \{0\}$ is defined by

$$\kappa_\nu(B, x)_A = \frac{\|B^\nu x\|_A}{\|x\|_A} \|B^{-\nu}\|_A \,. \tag{3}$$

Theorem 1 *For any $k = 1, 2, \ldots$ and $s = 1/2, 1, 3/2, \ldots$ there holds the estimate*

$$\frac{\|e^k\|_A}{\|e^0\|_A} \le \left(\frac{s}{k+s}\right)^{2s} \kappa_{-s}(B, e_0)_A \,.$$

The proof of Theorem 1 relies on (2) and the following theorem:

Theorem 2 *For any $k = 1, 2, \ldots$ and $s = 1/2, 1, 3/2, \ldots$ there holds*

$$\min_{P_k \in \pi_k^1} \max_{0 \le x \le 1} |x^s P_k(x)| \le \left(\frac{s}{k+s} \right)^{2s}.$$

In practice often the most relevant values of s in Theorem 1 are $s = 1/2$ and $s = 1$, where the upper bounds are sharp. Hence we introduce the notations

$$F = \kappa_{-1/2}(B, e_0)_A\,, \quad G = \kappa_{-1}(B, e_0)_A$$

for the corresponding generalized condition numbers. Using these, Theorem 1 yields in particular the estimates

$$\frac{\|e^k\|_A}{\|e^0\|_A} \le \frac{F}{2k+1} \qquad \text{and} \qquad \frac{\|e^k\|_A}{\|e^0\|_A} \le \frac{G}{(k+1)^2}\,. \tag{4}$$

In these cases F and G can be related to the (standard) condition number

$$\kappa(B) = \|B\|_A \|B^{-1}\|_A$$

such that more information is gained on the convergence. Namely, we have

Theorem 3 *There holds*

$$1 \le (G^2 + \kappa(B))/(1 + \kappa(B)) \le F^2 \le G \le \kappa(B)^{1/2}(F^2(1 + \kappa(B)^{-1}) - 1)^{1/2}.$$

Remark 1 The case when F is bounded independently of the matrix size n (by a suitable choice of e_0) is analyzed in [2]. Then in (4) we benefit by the uniform $O(1/k)$ estimate in n, further, we obtain a weak dependence of G on n since Theorem 3 yields $G \le const. \cdot \kappa(B)^{1/2}$. This situation turns out to hold when the components for the higher eigenvalue modes are large with respect to the lower modes. A sufficient estimate for this is given in terms of the trace, size and norm of B.

(b) The linear phase

As is well-known, if we only use the fact that the spectrum of B lies in the interval $[a, b]$, where $b = \|B\|_A$, $a = 1/\|B^{-1}\|_A$, then (2) yields linear convergence of the CGM with quotient σ involving the standard condition number:

$$\sigma = (\sqrt{b} - \sqrt{a})/(\sqrt{b} + \sqrt{a}) = (\sqrt{\kappa(B)} - 1)(\sqrt{\kappa(B)} + 1).$$

Remark 2 This linear convergence result has been long known to hold in the same form in Hilbert space [4], hence its study will be not repeated here.

(c) The superlinear phase

A simple superlinear estimate is obtained in [2] using the K-condition number

$$K = K(B) = \left(\frac{1}{n}\text{trace}(B)\right)^n / \det(B) = \left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i\right)^n \left(\prod_{i=1}^{n}\lambda_i\right)^{-1}.$$

Theorem 4 *Let $k < n$ be even and $k \geq 3\ln K$. Then*

$$\frac{\|e^k\|_A}{\|e^0\|_A} \leq \left(\frac{3\ln K}{k}\right)^{k/2}.$$

3 The sublinear phase in Hilbert space

We study the operator equation

$$Ax = a \tag{5}$$

in some real Hilbert space H, where $a \in H$. Here A may be an arbitrary (bounded or unbounded) spd operator, by which we mean the following:

Definition 1 A is an *spd operator* in H if A is self-adjoint and $\langle Ax, x\rangle > 0$ ($x \in D(A), x \neq 0$). If the latter inequality is replaced by $\langle Ax, x\rangle \geq c\|x\|^2$ ($x \in D(A)$) with some constant $c > 0$, then A will be called *uniformly spd.*

The assumption that A is self-adjoint is no loss of generality, since for any densely defined symmetric and strictly positive operator we can construct a self-adjoint extension (see e.g. [8]). On the other hand, the self-adjointness yields the following favourable well-posedness property:

Proposition 1 (see e.g. [8]). *If A is a uniformly spd operator in H, then $R(A) = H$.*

We define, as usual, the energy space H_A of A as the completion of $D(A)$ under the energy inner product

$$\langle u, v\rangle_A = \langle Au, v\rangle.$$

The corresponding norm has the obvious notation $\|\,.\,\|_A$.

Similarly to (1), we generally consider the preconditioned version

$$Bx = b \tag{6}$$

of (5), where $B = C^{-1}A$ with some uniformly spd operator C (and $b = C^{-1}a$). Then B is symmetric and strictly positive in the energy spaces H_A or H_C. In order to make B self-adjoint, the extension of $C^{-1}A$ may be required from $D(A)$ to H_A (resp. H_C) when these do not coincide. Then the solution x is usually looked for in H_A (resp. H_C) instead of $D(B) = D(A)$.

Of special interest in our investigations is the case when the (standard) condition number $\kappa(B) = \|B\|_A\|B^{-1}\|_A$ equals ∞, and thus the classical linear estimate fails. This either holds when B is bounded but B^{-1} is not, or when B is unbounded. We study these cases separately, also allowing both B and B^{-1} to be bounded. (Typical examples are weak elliptic operators with 0 lower bound, and strong elliptic differential operators, respectively.) The generalized condition number with respect to the initial vector in Theorem 1 has its real strength for the first case, when the difficulty with the unboundedness of B^{-1} is avoided by the definition of $\kappa_{-s}(B, e_0)_A$.

3.1 The case of a bounded operator

Our aim here is to extend Theorems 1 and 3 to the operator equation (6) when B is bounded in H_A. Analogously to (3), we define the generalized condition number

$$\kappa_\nu(B, x)_A = \frac{\|B^\nu x\|_A}{\|x\|_A}\|B^{-\nu}\|_A\,. \tag{7}$$

(For any $\nu < 0$, $\kappa_\nu(B, x)_A$ is finite by the boundedness of B, whereas for $\nu > 0$ it is only finite if B^{-1} is bounded.)

In addition, we first verify that the natural spectral estimate is sufficient for B to be bounded in $H = H_A$ even if A and C are unbounded. Hence one can proceed analogously to the finite dimensional case.

(a) Boundedness of the preconditioned operator

We consider the case when $D(A) = D(C) =: D$ and there exists a number $M > 0$ such that

$$\langle Ax, x\rangle \le M\langle Cx, x\rangle \qquad (x \in D). \tag{8}$$

(This is automatically satisfied in $\mathbf{R}^n$.)

This means that B is bounded with respect to $\|\,.\,\|_C$, since (8) is equivalent to

$$\langle Bx, x\rangle_C \le M\|x\|^2 \qquad (x \in D).$$

We prove that the same is true with respect to $\|\,.\,\|_A$:

Proposition 2 *If (8) holds then* $\|B\|_A \le M < \infty$.

PROOF. Since (8) means $\|A^{1/2}x\|^2 \le M\|C^{1/2}x\|^2$ $(x \in D)$ and there holds $R(C^{1/2}) = D(C^{-1/2}) = H$, hence letting $y = C^{1/2}x$ implies

$$\|A^{1/2}C^{-1/2}y\|^2 \le M\|y\|^2 \qquad (y \in H).$$

The symmetry of $A^{1/2}$ and $C^{-1/2}$ yields

$$\langle C^{-1/2}A^{1/2}y, y\rangle = \langle y, A^{1/2}C^{-1/2}y\rangle \le M^{1/2}\|y\|^2 \qquad (y \in H),$$

whence $\|C^{-1/2}A^{1/2}\| \le M^{1/2}$. Therefore

$$\langle C^{-1}A^{1/2}y, A^{1/2}y\rangle = \|C^{-1/2}A^{1/2}y\|^2 \le M\|y\|^2 \qquad (y \in H),$$

which, if setting $v = A^{-1/2}y$, means $\langle C^{-1}Av, Av\rangle \le M\langle Av, v\rangle$ $(v \in D)$. Hence

$$\langle Bv, v\rangle_A = \langle ABv, v\rangle = \langle AC^{-1}Av, v\rangle = \langle C^{-1}Av, Av\rangle \le M\|v\|_A^2 \qquad (v \in D).$$

This yields $\|B\|_A \le M$ since the assumption $D(A) = D(C) =: D$ implies $D(B) = D(C^{-1}A) = D$. (We note that B has a unique bounded extension to H_A, preserving the norm estimate M). □

Consequently, if (8) holds then for $\nu < 0$ the generalized condition number $\kappa_\nu(B, x)_A$ is finite.

Remark 3 (i) Let A and C be spectrally equivalent, i.e.

$$\inf_{x\in D} \frac{\langle Bx, x\rangle_C}{\|x\|_C^2} = \inf_{x\in D} \frac{\langle Ax, x\rangle}{\langle Cx, x\rangle} =: m > 0.$$

(In other words, there is also a lower estimate in (8) for $\langle Ax, x\rangle$). Then B is uniformly positive and hence B^{-1} is also bounded in H_C, and these hold in H_A as well since $\|\,.\,\|_A$ and $\|\,.\,\|_C$ are now equivalent. Therefore the generalized condition number $\kappa_\nu(B, x)_A$ is also finite for $\nu > 0$, as well as the standard condition number

$$\kappa(B) = \|B\|_A\|B^{-1}\|_A\,.$$

(ii) On the other hand, if $m = 0$, then B^{-1} is unbounded in H_A, $\kappa(B)$ is infinite and $\kappa_\nu(B, x)_A$ is only finite for $\nu < 0$. Hence a particular value of generalizing Theorem 1 is to find a convergence estimate when the linear estimate with $\kappa(B)$ cannot be used.

(b) The sublinear convergence estimate

Theorem 5 *If (8) holds, then for any* $k = 1, 2, \ldots$ *and* $s = 1/2, 1, 3/2, \ldots$ *there holds the estimate*

$$\frac{\|e^k\|_A}{\|e^0\|_A} \le \left(\frac{s}{k+s}\right)^{2s} \kappa_{-s}(B, e_0)_A \,. \tag{9}$$

Proof. We proceed similarly as in the case of $\mathbf{R}^n$ instead of H. We define $\tilde{B} = B/\|B\|_A$ and use $\|B^s\|_A = \|B\|_A^s$. Then the equality

$$\|e^k\|_A = \min_{P_k \in \pi_k^1} \|P_k(B)e^0\|_A \tag{10}$$

(the analogue of (2), which holds in H as well) implies

$$\|e^k\|_A/\|e^0\|_A = \min_{P_k \in \pi_k^1} \|P_k(B)B^s B^{-s} e^0\|_A/\|e^0\|_A$$

$$\le \min_{P_k \in \pi_k^1} \|P_k(\tilde{B})\tilde{B}^s\|_A \|\tilde{B}^{-s}e^0\|_A/\|e^0\|_A \le \min_{P_k \in \pi_k^1} \max_{0 \le x \le 1} |x^s P_k(x)| \kappa_{-s}(B, e_0)_A \,.$$

The required estimate then follows from Theorem 2. □

Remark 4 By Proposition 2 the generalized condition numbers $\kappa_{-s}(B, e_0)_A$ are finite. Especially, for $s = 1/2$ or $s = 1$,

$$F = \frac{\|B^{-1/2}e^0\|_A}{\|e^0\|_A}\|B^{1/2}\|_A \quad \text{and} \quad G = \frac{\|B^{-1}e^0\|_A}{\|e^0\|_A}\|B\|_A \tag{11}$$

yield the estimates $\frac{F}{2k+1}$ and $\frac{G}{(k+1)^2}$ on the right side of (9), respectively.

Theorem 6 *There holds*

$$1 \le (\frac{G^2}{\kappa(B)} + 1)/(\frac{1}{\kappa(B)} + 1) \le F^2 \le G \le \kappa(B)^{1/2}(F^2(1 + \kappa(B)^{-1}) - 1)^{1/2}.$$

(If $\kappa(B) = \infty$ *then* $\frac{1}{\kappa(B)} = 0$ *is understood.)*

Proof. (i) The first inequality is a consequence of $G \ge 1$, which is obvious:

$$\|e^0\|_A = \|BB^{-1}e^0\|_A \le \|B\|_A \|B^{-1}e^0\|_A \,. \tag{12}$$

(ii) If $\kappa(B) = \infty$ then the second inequality reduces to $F \ge 1$, which follows in the same way as (12) with $B^{1/2}$ instead of B. Let $\kappa(B) < \infty$. Then the required inequality can be written as

$$G^2 + \kappa(B) \le F^2(1 + \kappa(B)). \tag{13}$$

Denote as previously by m and M the spectral bounds of B with respect to $\|\,.\,\|_A$, i.e. $\kappa(B) = M/m$. Then, using $\|B^{1/2}\|_A = \|B\|_A^{1/2} = M^{1/2}$, (13) takes the form

$$\frac{\|B^{-1}e^0\|_A^2\, M^2}{\|e^0\|_A^2} + \frac{M}{m} \le \frac{\|B^{-1/2}e^0\|_A^2\, M}{\|e^0\|_A^2}\left(1+\frac{M}{m}\right).$$

Multiplying by $\|e^0\|_A^2 m/M$ and letting $g^0 = B^{-1/2}e^0$, we obtain

$$\|B^{-1/2}g^0\|_A^2\, Mm + \|B^{1/2}g^0\|_A^2 \le \|g^0\|_A^2\,(M+m). \tag{14}$$

Denote by $E(\lambda)$ the spectral decomposition of B on $\sigma(B) \subset [m, M]$. For any continuous real function f on $\sigma(B)$ and any $x \in H_A$ there holds

$$\|f(B)x\|_A^2 = \int\limits_{\sigma(B)} |f(\lambda)|^2 dE_{x,x}(\lambda)$$

(see e.g. [7]). Here $E_{x,x}$ denotes the measure defined by $E_{x,x}(S) = \langle E(S)x, x\rangle$ for measurable subsets $S \subset \sigma(B)$. Hence, using the notation $dE_0(\lambda) = dE_{g^0,g^0}(\lambda)$, (14) takes the form

$$Mm \int\limits_{\sigma(B)} \frac{1}{\lambda}\, dE_0(\lambda) + \int\limits_{\sigma(B)} \lambda\, dE_0(\lambda) \le (M+m) \int\limits_{\sigma(B)} dE_0(\lambda). \tag{15}$$

Here $\sigma(B) \subset [m, M]$ implies $m \le \lambda \le M$ for all λ, hence

$$\frac{Mm}{\lambda} + \lambda \le M + m,$$

from which (15) follows.

(iii) The inequality $F^2 \le G$ means

$$\frac{\|B^{-1/2}e^0\|_A^2 M}{\|e^0\|_A^2} \le \frac{\|B^{-1}e^0\|_A M}{\|e^0\|_A}.$$

This simply follows from

$$\|B^{-1/2}e^0\|_A^2 = \langle B^{-1}e^0, e^0\rangle_A \le \|B^{-1}e^0\|_A \|e^0\|_A\,.$$

(iv) The final inequality follows from (13) when $\kappa(B) < \infty$, otherwise there is nothing to prove. □

Remark 5 (Analysis of the estimates). The main point in the theorems of this subsection is that the generalized condition numbers $\kappa_{-s}(B, e_0)_A$ are finite. We underline that at the same time there may hold $\kappa(B) = \infty$, in which case the linear estimate does not work.

In practice the obtained Hilbert space theorems help to understand the behaviour of the CGM in the sublinear phase for algebraic systems with large matrix size n. This is because the abstract theorems are asymptotic results for the latter when the size n tends to ∞. In particular, the requirements for the eigenvalue component distribution of the initial vectors in Remark 1 to produce bounded F can be well understood in the context of Theorem 5. Namely, the boundedness of F under increasing the matrix size n corresponds to $F < \infty$ in the Hilbert space theorem. The latter holds automatically whenever e^0 belongs to the space H, which means that if e^0 is expanded into eigenfunction series then the sequence of its coordinates (i.e. the coefficients in the series) tends to 0. That is, for the algebraic systems with large matrix size n the coordinates corresponding to low eigenvalues are small.

We note that for elliptic boundary value problems, such initial functions can be often computed by solving the problem on a "coarse" finite dimensional subspace.

Examples. Let $\Omega \subset \mathbf{R}^N$ be a bounded domain. In the following examples Theorem 5 holds for B.

1. Let $M \geq m > 0$, and $\mathcal{A}(x) = \{a_{ij}(x)\}$ be a spd matrix for all $x \in \Omega$ with eigenvalues between m and M. Let A and C be the elliptic operators

$$Au := -\mathrm{div}(\mathcal{A}(x)\nabla u), \qquad Cu := -\Delta u$$

with $D(A) = D(C) = H^2(\Omega) \cap H_0^1(\Omega)$ in the real Hilbert space $H = L^2(\Omega)$. Then $H_A = H_C = H_0^1(\Omega)$, and $B = C^{-1}A$ is the weak elliptic operator given by

$$\langle Bu, v\rangle_C = \langle Bu, v\rangle_{H_0^1(\Omega)} = \int_\Omega \mathcal{A}(x)\,\nabla u \cdot \nabla v\,.$$

The operator B is uniformly elliptic in $H_0^1(\Omega)$, hence both B and B^{-1} are bounded. (This is also expressed by the spectral equivalence of A and C with bounds m and M.)

Consequently, both the standard and the generalized condition numbers are finite, the latter implying that Theorem (5) holds for B. Here the linear convergence result also holds in virtue of $\kappa(B) < \infty$, see Remark 2.

2. (a) Let $N = 2$,

$$\mathcal{A}_\varepsilon(x) \equiv \mathcal{A}_\varepsilon = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}, \tag{16}$$

where $\varepsilon > 0$ is a constant. We consider the elliptic operators as in Example 1, now with $\mathcal{A}(x)$ from (16). Then

$$\langle A_\varepsilon u, u\rangle = \int_\Omega \left(|\partial_1 u|^2 + \varepsilon\,|\partial_2 u|^2\right), \qquad \langle Cu, u\rangle = \int_\Omega \left(|\partial_1 u|^2 + |\partial_2 u|^2\right),$$

hence it is easy to see that

$$\sup_{u\in D}\frac{\langle B_\varepsilon u,u\rangle_C}{\|u\|_C^2}=\sup_{u\in D}\frac{\langle A_\varepsilon u,u\rangle}{\langle Cu,u\rangle}=1,\qquad \inf_{u\in D}\frac{\langle B_\varepsilon u,u\rangle_C}{\|u\|_C^2}=\inf_{u\in D}\frac{\langle A_\varepsilon u,u\rangle}{\langle Cu,u\rangle}=\varepsilon\,,$$

where $D=H^2(\Omega)\cap H_0^1(\Omega)$.

Here B_ε and B_ε^{-1} are bounded. However, let us now vary ε and consider a singular perturbation problem for the family A_ε by letting $\varepsilon\to 0$. Then the lower bound ε of B_ε deteriorates and the standard condition number $\kappa(B_\varepsilon)$ tends to ∞. Therefore the quotient of the classical linear convergence result also deteriorates. On the other hand, since B_ε is uniformly bounded in ε, therefore the generalized condition numbers $\kappa_{-s}(B_\varepsilon,e_0)_{A_\varepsilon}$ (and hence the estimate in Theorem 5) remain bounded as $\varepsilon\to 0$.

As an underlying limiting case, we may let $\varepsilon=0$. Then by Remark 3, B_0 is bounded in H_A but B_0^{-1} is not, hence $\kappa(B_0)=\infty$, whereas $\kappa_{-s}(B_0,e_0)_A$ is finite, its value being the limit of $\kappa_{-s}(B_\varepsilon,e_0)_{A_\varepsilon}$ as $\varepsilon\to 0$. That is, Theorem 5 even holds for B_0.

(b) Let $H=H_0^1(\Omega)$ with $\langle u,v\rangle_{H_0^1(\Omega)}=\int_\Omega\nabla u\cdot\nabla v$, A be defined by

$$\langle Au,v\rangle_{H_0^1(\Omega)}=\int_\Omega uv\qquad (u,v\in H_0^1(\Omega)).\tag{17}$$

We let $C=I$, hence $B=A$. Then by Green's formula $Au=(-\Delta)^{-1}u\quad(u\in H_0^1(\Omega))$, where $-\Delta$ is understood with homogeneous Dirichlet boundary conditions. Further, (17) implies that $H_A=L^2(\Omega)$. As is well-known, $B=(-\Delta)^{-1}$ is compact in $L^2(\Omega)$ and its eigenvalues tend to 0. Hence B but not B^{-1} is bounded in H_A.

Consequently, $\kappa(B)=\infty$ and the classical linear convergence result cannot be used, whereas $\kappa_{-s}(B,e_0)_A$ is finite and hence Theorem 5 applies to B.

3.2 The case of an unbounded operator

We consider the case when $B=C^{-1}A$ is still unbounded in H_A. Then Theorem 5 gives no result since $\kappa_{-s}(B,e_0)_A=\infty$. Therefore we suitably modify the generalized condition number (7) to extend Theorem 5.

Definition 2 Let $s>0$, $k\in\mathbf{N}$. For any $e^0\in H$ we introduce

$$V_k=V_{s,k,e^0}:=span\{B^{j/2}e^0\}_{j=-2s,\dots,2k}$$

and

$$\kappa_{-s,k}(B,e^0)_A=\frac{\|B^{-s}e^0\|_A}{\|e^0\|_A}\|B_{|V_k}\|_A^s\,,$$

where $B_{|V_k}$ denotes the restriction of B to V_k.

Theorem 7 *For any $k = 1, 2, \ldots$ and $s = 1/2, 1, 3/2, \ldots$ there holds the estimate*

$$\frac{\|e^k\|_A}{\|e^0\|_A} \leq \left(\frac{s}{k+s}\right)^{2s} \kappa_{-s,k}(B, e^0)_A \,. \tag{18}$$

PROOF. For simplicity we write V instead of V_k in the proof. Let $Q = Q_{s,k,e^0}$ denote the orthogonal projection to the subspace V, and let $\hat{B} = (B^{1/2}Q)^2$. Then for any $P_k \in \pi_k^1$

$$P_k(B)e^0 = P_k(\hat{B})e^0, \tag{19}$$

since for all $i = 0, 1, ..., k$ there holds $B^i e^0 = (B^{1/2})^{2i} e^0 = (B^{1/2}Q)^{2i} e^0 = \hat{B}^i e^0$ (which follows from Q being equal to the identity on V). Similarly,

$$e^0 = B^s B^{-s} e^0 = (B^{1/2}Q)^{2s} B^{-s} e^0 = \hat{B}^s B^{-s} e^0. \tag{20}$$

From here we can proceed similarly as in Theorem 5. Defining $\tilde{B} = \hat{B}/\|\hat{B}\|_A$, the equalities (10), (19) and (20) imply

$$\|e^k\|_A/\|e^0\|_A = \min_{P_k \in \pi_k^1} \|P_k(B)e^0\|_A/\|e^0\|_A = \min_{P_k \in \pi_k^1} \|P_k(\hat{B})e^0\|_A/\|e^0\|_A =$$

$$\min_{P_k \in \pi_k^1} \|P_k(\hat{B})\hat{B}^s B^{-s} e^0\|_A/\|e^0\|_A \leq \min_{P_k \in \pi_k^1} \|P_k(\tilde{B})\tilde{B}^s\|_A \|\hat{B}^s\|_A \|B^{-s}e^0\|_A/\|e^0\|_A \,. \tag{21}$$

Here

$$\min_{P_k \in \pi_k^1} \|P_k(\tilde{B})\tilde{B}^s\|_A \leq \min_{P_k \in \pi_k^1} \max_{0 \leq x \leq 1} |x^s P_k(x)| \leq \left(\frac{s}{k+s}\right)^{2s}$$

from Theorem 2, just as in Theorem 5. Further,

$$\|\hat{B}^s\|_A = \|(B^{1/2}Q)^{2s}\|_A = \|B^{1/2}Q\|_A^{2s} = \|B^{1/2}{}_{|V}\|_A^{2s} \leq \|B_{|V}\|_A^s,$$

hence

$$\|\hat{B}^s\|_A \|B^{-s}e^0\|_A/\|e^0\|_A \leq \|B_{|V}\|_A^s \|B^{-s}e^0\|_A/\|e^0\|_A = \kappa_{-s,k}(B, e^0)_A \,.$$

Substituting these into (21), the theorem is proved. □

Remark 6 In general Theorem 7 does not give a common bound for all k, since for fixed e^0 the condition numbers $\kappa_{-s,k}(B, e^0)_A$ may grow unboundedly as V_k increases with k. On the other hand, if e^0 is suitably chosen then we may still obtain information from this theorem. Namely, let us consider the case when B^{-1} is compact, i.e. B has eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq ... \to \infty$ with corresponding eigenvectors v_j forming a complete orthonormal system in H_A. Trivially, if e^0 has only finitely many

non-zero coordinates then V_k remains finite-dimensional and $\kappa_{-s,k}(B,e^0)_A$ bounded. More generally, we can choose a sequence of vectors $e^0 = e^0_{(k)}$ with infinitely many non-zero coordinates such that the corresponding condition numbers $\kappa_{-s,k}(B,e^0_{(k)})_A$, corresponding to the kth iterates $e^k = e^k_{(k)}$, are bounded as k increases. This is illustrated below for $s = 1/2$, i.e. for the condition number

$$F_{k,e^0} := \kappa_{-1/2,k}(B,e^0_{(k)})_A = \frac{\|B^{-1/2}e^0\|_A}{\|e^0\|_A}\|B_{|V_k}\|_A^{1/2} \tag{22}$$

corresponding to (11).

Proposition 3 *Let B^{-1} be compact, and denote the eigenvalues of B by $0 < \lambda_1 \leq \lambda_2 \leq ... \to \infty$ with corresponding eigenvectors v_j normalized with respect to $\|\,.\,\|_C$. We choose $r \in \mathbf{N}^+$ such that*

$$\lambda_r \geq 1.$$

Let $0 < \alpha_1 \leq \alpha_2 \leq ...$ be a real sequence such that $\sum_j \lambda_j^{-\alpha_j}$ converges, and let

$$S := \sum_{j=r+1}^{\infty} \lambda_j^{-\alpha_j}.$$

Let

$$e^0_{(k)} = \sum_{j=1}^{\infty} c_k^j v_j \qquad (k \in \mathbf{N}^+),$$

where the sequences $(c_k^j)_{j\in\mathbf{N}^+}$ satisfy the following: $c_k^r \neq 0$ is arbitrary, and there exists a constant $\gamma > 0$ such that

$$\left(\frac{c_k^j}{c_k^r}\right)^2 \leq \begin{cases} \gamma & \text{if } j < r \\ \frac{\lambda_r}{S\lambda_j^{\alpha_j+2k+3}} & \text{if } j > r. \end{cases}$$

Then $F_{k,e^0_{(k)}}$, defined as in (22), is bounded as $k \to \infty$.

Proof. Since B is uniformly positive, it follows that $\|B^{-1/2}e^0_{(k)}\|_A/\|e^0_{(k)}\|_A \leq \|B^{-1/2}\|_A < \infty$ is bounded in k. Hence for the boundedness of $F_{k,e^0_{(k)}}$ it suffices that $\|B_{|V_k}\|_A$ is bounded in k, where now $V_k = V_{1/2,k,e^0_{(k)}} := span\{B^{j/2}e^0\}_{j=-1,...,2k}$. There holds

$$\|B_{|V_k}\|_A = \max\{\frac{\|Bv\|_A}{\|v\|_A} : v \in V_k\} = \max\{\frac{\|B^{\frac{l+1}{2}}e^0_{(k)}\|_A}{\|B^{\frac{l-1}{2}}e^0_{(k)}\|_A} : l = 0,1,\cdots 2k+1\}.$$

Since $\langle v_j, v_i\rangle_A = \lambda_j\langle Cv_j, v_i\rangle = \lambda_j\delta_{ij}$, we obtain

$$\|B_{|V_k}\|_A = \max\left\{\frac{\sum_{j=1}^{\infty}(c_k^j)^2\lambda_j^{l+2}}{\sum_{j=1}^{\infty}(c_k^j)^2\lambda_j^l} : l = 0,1,\cdots 2k+1\right\}^{1/2}.$$

We verify that for all $k \in \mathbf{N}^+$ and $l = 0, 1, \cdots 2k+1$, there holds

$$\sum_{j=1}^{\infty} (c_k^j)^2 \lambda_j^{l+2} \leq M \sum_{j=1}^{\infty} (c_k^j)^2 \lambda_j^l, \tag{23}$$

where

$$M := (\gamma/\lambda_r) \sum_{j=1}^{r-1} \lambda_j^{l+2} + \lambda_r^2 + 1.$$

In fact, using $\lambda_j \geq 1$ $(j \geq r)$, we have

$$\sum_{j=1}^{\infty} (c_k^j)^2 \lambda_j^{l+2} = \sum_{j=1}^{r-1} (c_k^j)^2 \lambda_j^{l+2} + (c_k^r)^2 \lambda_r^{l+2} + \sum_{j=r+1}^{\infty} (c_k^j)^2 \lambda_j^{l+2}$$

$$\leq (c_k^r)^2 \gamma \sum_{j=1}^{r-1} \lambda_j^{l+2} + (c_k^r)^2 \lambda_r^{l+2} + (c_k^r)^2 \frac{\lambda_r}{S} \sum_{j=r+1}^{\infty} \frac{\lambda_j^{l+2}}{\lambda_j^{\alpha_j + 2k+3}}$$

$$\leq (c_k^r)^2 \left(\gamma \sum_{j=1}^{r-1} \lambda_j^{l+2} + \lambda_r^{l+2} + \frac{\lambda_r^l}{S} \sum_{j=r+1}^{\infty} \frac{\lambda_j^{2k+3}}{\lambda_j^{\alpha_j + 2k+3}} \right)$$

$$= (c_k^r)^2 \left(\gamma \sum_{j=1}^{r-1} \lambda_j^{l+2} + \lambda_r^{l+2} + \lambda_r^l \right) = M (c_k^r)^2 \lambda_r^l \leq M \sum_{j=1}^{\infty} (c_k^j)^2 \lambda_j^l. \qquad \square$$

Remark 7 The main point in the proposition is that the coordinates of $e^0_{(k)}$ are decreasing increasingly rapidly. That is, for large k we have to choose e^0 such that it has small enough coordinates with respect to the extreme eigenvalues that tend to ∞ and cause $\kappa(B) = \infty$. This is the counterpart of the bounded case when $\kappa(B) = \infty$ was due to B^{-1} unbounded, and e^0 had to be chosen to have small coordinates with respect to the extreme eigenvalues tending to 0 (see Remark 5). For elliptic problems, such initial functions may be computed by first using some smoothing iteration method.

4 The superlinear phase in Hilbert space

The superlinear convergence result has been extended to Hilbert space in [4] and [5] for the special case when the strictly positive operator B has the form

$$B = \lambda I + L, \tag{24}$$

where $\lambda > 0$, I is the identity operator and L is a compact self-adjoint linear operator. In the sequel we prove that the estimate in Theorem 4 via the K-condition

number can be suitably generalized to (24) when L is a Hilbert-Schmidt operator (cf. Remark 8). The main consequence of this extension is that the K-condition numbers corresponding to the discretizations of B have a common bound when the matrix size n tends to ∞. See also [1] (Example 13.6) for the eigenvalue distribution $\lambda_i = 1 + 1/i$.

It is reasonable to consider B of the form (24) an already preconditioned operator. Hence, in contrast to Section 2, we can study for simplicity the CGM estimate in a general Hilbert space H (disregarding that it is especially obtained as an energy space $H = H_A$).

Theorem 4 can be generalized for (24) as follows.

Theorem 8 *Let H be a Hilbert space, L be a compact self-adjoint linear operator on H with eigenvalues μ_i $(i \in \mathbf{N})$ such that $\mu_i \to 0$. Let $\lambda > 0$ and B defined as in (24):*

$$B = \lambda I + L.$$

Assume that B is strictly positive, i.e. $\mu_i > -\lambda$ $(i \in \mathbf{N})$. Let

$$K_n := \left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i\right)^n \left(\prod_{i=1}^{n}\lambda_i\right)^{-1} \qquad (k \in \mathbf{N}),$$

where $\lambda_i = \lambda + \mu_i$ $(i \in \mathbf{N})$ are the eigenvalues of B.

If $\sum_i \mu_i^2 < \infty$, then

(1) the sequence K_n is bounded, further, $K := \sup_{n\in\mathbf{N}} K_n \le \exp(\sum_{i=1}^{\infty} \mu_i^2/2a\lambda^2)$, where $a = \inf_i \lambda_i/\lambda$;

(2) if $k \in \mathbf{N}$ is even and $k \ge 3\ln K$, then there holds

$$\frac{\|e^k\|}{\|e^0\|} \le \left(\frac{3\ln K}{k}\right)^{k/2}.$$

Proof. (1) Using notation $\rho_i = \frac{\mu_i}{\lambda}$ (> -1), and that $(1+t)^n \le \exp(nt)$ $(t \ge -1, n \in \mathbf{N}^+)$, we obtain

$$K_n = \left(1 + \frac{1}{n}\sum_{i=1}^{n}\rho_i\right)^n \prod_{i=1}^{n}(1+\rho_i)^{-1} \le \prod_{i=1}^{n}\frac{\exp(\rho_i)}{1+\rho_i}.$$

This implies

$$\ln K_n \le \sum_{i=1}^{n}\left[\rho_i - \ln(1+\rho_i)\right] \le \sum_{i=1}^{n}\rho_i^2/2a = \sum_{i=1}^{n}\mu_i^2/2a\lambda^2,$$

since $\rho_i = \frac{\lambda_i}{\lambda} - 1 \geq a - 1$ and $x - \ln(1+x) \leq \frac{x^2}{2a}$ for $x \in [a-1, \infty)$ (which follows by comparing the derivatives).

(2) Let $k = 2m \in \mathbf{N}$ be even, $k \geq 3 \ln K$. Using the eigenfunction expansion of e^0, it is easy to see that (10) implies

$$\|e^k\| = \min_{P_k \in \pi_k^1} \|P_k(B)e^0\| \leq \min_{P_k \in \pi_k^1} \max_{j \in \mathbf{N}^+} |P_k(\lambda_j)| \, \|e^0\| .$$

Let $n \in \mathbf{N}^+$, $n > k$ be arbitrary. Introducing, as in [2], the polynomial

$$P_k(t) = \prod_{i=1}^{m} \left(1 - \frac{t}{\lambda_i}\right) \left(1 - \frac{t}{\lambda_{n+1-i}}\right) \qquad (t \in \mathbf{R}),$$

which vanishes at $\lambda_1, \lambda_2, \cdots, \lambda_m$ and $\lambda_{n+1-m}, \cdots, \lambda_{n-1}, \lambda_n$, we obtain

$$\|e^k\| / \|e^0\| \leq \max_{j \in \mathbf{N}^+} |P_k(\lambda_j)| = \max \left\{ \max_{m<j<n+1-m} |P_k(\lambda_j)|, \ \max_{j>n} |P_k(\lambda_j)| \right\} . \qquad (25)$$

The two max terms can be estimated as follows. Since $k \geq 3 \ln K_n$, the proof of Theorem 4 (see [2]) and then part (1) of the present theorem yield that

$$\max_{m<j<n+1-m} |P_k(\lambda_j)| \leq \left(\frac{3 \ln K_n}{k}\right)^{k/2} \leq \left(\frac{3 \ln K}{k}\right)^{k/2} .$$

On the other hand, let $\lambda_- = \min\{\lambda_j : j \in \mathbf{N}^+\}$, $\lambda_+ = \max\{\lambda_j : j \in \mathbf{N}^+\}$. Then for any $i = 1, ..., m$ or $i = n+1-m, ..., n-1$ there holds $|\lambda_i - \lambda_j| \leq \lambda_+ - \lambda_-$ for $j > n$ and $\lambda_i \geq \lambda_-$. Hence

$$\max_{j>n} |P_k(\lambda_j)| = \max_{j>n} \prod_{\substack{1 \leq i \leq m \\ n+1-m \leq i \leq n}} \frac{|\lambda_i - \lambda_j|}{\lambda_i} \leq \left(\frac{\lambda_+ - \lambda_-}{\lambda_-}\right)^{m-1} \frac{\max_{j>n} |\lambda_j - \lambda_n|}{\lambda_n}.$$

This estimate holds for all $n \in \mathbf{N}^+$, hence we can let $n \to \infty$. Then the first factor is constant and the second tends to 0 since $\lim_{n\to\infty} \lambda_n = \lambda$. Consequently, (25) implies the required estimate. □

Remark 8 The condition $\sum_i \mu_i^2 < \infty$ in Theorem 8 means that L is a so-called *Hilbert-Schmidt operator* (see e.g. [9]). Then the Hilbert-Schmidt norm of L is defined as

$$|||L||| := \left(\sum_{i=1}^{\infty} \mu_i^2\right)^{1/2}$$

(and there also holds $|||L|||^2 = \sum_i \|Le_i\|^2$ for any complete orthonormal system $\{e_i\}$). Hence the estimate for the K-condition number in the theorem is

$$K := \sup_{n \in \mathbf{N}} K_n \leq \exp(|||L|||^2 / 2a\lambda^2).$$

Example. Let $N \leq 3$, $\Omega \subset \mathbf{R}^N$ be a bounded domain. Let $M \geq m > 0$, and $\mathcal{A}(x) = \{a_{ij}(x)\}$ be a spd matrix for all $x \in \Omega$ with eigenvalues between m and M. Let $q(x) \geq 0$. We assume that $\mathcal{A} \in L^\infty(\Omega, \mathbf{R}^{N\times N})$, $q \in L^\infty(\Omega)$. Let $H = H_0^1(\Omega)$ with the inner product

$$\langle u, v\rangle_{\mathcal{A}} = \int_\Omega \mathcal{A}\nabla u \cdot \nabla v.$$

Let the operator B be defined by

$$\langle Bu, v\rangle_{\mathcal{A}} = \int_\Omega (\mathcal{A}\nabla u \cdot \nabla v + quv) \qquad (u, v \in H_0^1(\Omega)).$$

Then Theorem 8 holds for B.

Namely, we have

$$B = I + L, \quad \text{where} \quad \langle Lu, v\rangle_{\mathcal{A}} = \int_\Omega quv \qquad (u, v \in H_0^1(\Omega)).$$

It is well-known that L is compact and self-adjoint on $H_0^1(\Omega)$. Further, the variational characterization of the eigenvalues yields that replacing q by a pointwise greater function, the eigenvalues of L are also replaced by greater ones. Hence they can be estimated by the eigenvalues τ_j of the operator

$$\langle Tu, v\rangle_{\mathcal{A}} = \beta \int_\Omega uv \qquad (u, v \in H_0^1(\Omega)),$$

where $\beta = \|q\|_\infty$. The well-known estimate (see e.g. [3]) is

$$\tau_j \leq const. \cdot j^{-2/N} \leq const. \cdot j^{-2/3} \qquad (j \in \mathbf{N}^+),$$

since $N \leq 3$.

Consequently, the eigenvalues μ_j of L are estimated by

$$\mu_j \leq const. \cdot j^{-2/3} \qquad (j \in \mathbf{N}^+),$$

and hence

$$\sum_j \mu_j^2 \leq const. \cdot \sum_j j^{-4/3} < \infty.$$

That is, the conditions of Theorem 8 are satisfied.

5 Conclusions

It has been shown that previous results in $\mathbf{R}^n$ for the sublinear and superlinear rate of convergence of the conjugate gradient method, involving various generalized condition numbers, can be extended to Hilbert space. In the sublinear phase the results

even include the case when B or B^{-1} is unbounded. Applications of the generalized theorems to elliptic differential operators have been given. In this case the obtained results are asymptotic when discretization is refined, hence they are relevant for large matrix sizes. Moreover, the convergence results for the operators themselves yield mesh independent conditioning estimates for the discretized problems.

References

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

[2] Axelsson, O., Kaporin, I., On the sublinear and superlinear rate of convergence of conjugate gradient methods, to appear in *Numerical Algorithms.*

[3] Courant, R, Hilbert, D., *Methods of Mathematical Physics II.*, Wiley Classics Library, J. Wiley & Sons, 1989.

[4] Daniel, J.W., The conjugate gradient method for linear and nonlinear operator equations, *SIAM J. Numer. Anal.*, 4 (1967) No.1., 10-26.

[5] Hayes, R.M., Iterative methods of solving linear problems in Hilbert space, *Nat. Bur. Standards Appl. Math. Ser*, 39 (1954), 71-104.

[6] Hestenes, M.R., Stiefel, E., Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Standards*, Sect. B, 49 (1952) No.6., 409-436.

[7] Rudin, W., *Functional Analysis*, McGraw-Hill, 1991.

[8] Riesz F., Sz.-Nagy B., *Vorlesungen über Funktionalanalysis*, Verlag H. Deutsch, 1982.

[9] Weidman, R., *Linear Operators in Hilbert Space,* Springer, 1976.

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