SYMMETRIC PART PRECONDITIONING
FOR THE CONJUGATE GRADIENT METHOD
IN HILBERT SPACE

O. Axelsson, J. Karátson

Report No. 0204 (February 2002)
Symmetric part preconditioning for the conjugate gradient method in Hilbert space

by O. Axelsson\textsuperscript{1} and J. Karátson\textsuperscript{2}

Dedicated to the memory of Jean-Jacques Lions:
a source of inspiration for rigour in Applied Mathematics

Abstract

The conjugate gradient method for non-symmetric linear operators in Hilbert space is investigated. Conditions on the coincidence of the full and truncated versions, known from the finite-dimensional case, are extended to the Hilbert space setting. The focus is on preconditioning by the symmetric part of the operator, in which case estimates are given for the resulting condition number. An important motivation for this study is given by differential operators, for which the obtained estimates yield mesh independent conditioning properties of the full CGM, and are in fact achieved by the simpler truncated version.

1 Introduction

Conjugate gradient methods have become one of the most widespread ways of solving not only symmetric but also nonsymmetric linear algebraic systems [2, 4]. An early paper dealing with generalized conjugate gradient methods for nonsymmetric systems is [3], and a survey of available methods can be found in [14], which includes also the popular GMRES method [13]. The important issue of automatic truncation of the algorithm for a general initial residual was settled independently in [7] and [16]. A discussion about the role played by the initial residual on truncation can be found in [2] and [4].

In [4] conditions have been given when the full and truncated versions of the CGM coincide. Of special interest is the case when the symmetric part of the operator is used for preconditioning, since in this case the truncated version requires only a single, namely the current search direction. A first study of such methods can be found in [5] and in [17], where a particular algorithm was derived. Important related results for nonsymmetric elliptic problems have been established e.g. in [6, 11]. The aim of this paper is to extend the above results to linear operators in a real Hilbert space, and to give estimates of the resulting condition number using properties of the symmetric and antisymmetric parts.

\textsuperscript{1}\text{Department of Mathematics, University of Nijmegen, 6525ED Nijmegen, The Netherlands; axelsson@sci.kun.nl}
\textsuperscript{2}\text{Department of Applied Analysis, ELTE University, H-1053 Budapest, Hungary; karatson@cs.elte.hu}
This research was done during the second author’s Hungarian post-doc scholarship Magyary Zoltán.
An important motivation for this study is given by differential operators. In this case the matrices obtained from discretization are approximations of the original operator that describes the studied model exactly. Hence the study of the CGM for these operators helps the understanding of the CGM for the discretized problems. In particular, the Sobolev space results give mesh independent estimates of the condition number. Hereby the coincidence of the full and truncated versions of the CGM implies that the corresponding mesh independent convergence properties of the full CGM are achieved by the simpler truncated version.

2 The full and truncated versions of the CGM

The generalized conjugate gradient, least square (GCG-LS) method is defined in [4]. Two versions are discussed: the full version which uses all previous search directions, whereas the truncated version uses only \( s + 1 \) previous search directions (denoted by GCG-LS\( (s) \)), where \( s \) is a nonnegative integer.

In [9] the conjugate gradient method was formulated in Hilbert space. Similarly to the standard CGM for symmetric positive definite systems, we can formulate also the generalized CG methods in a real Hilbert space \( H \). For an equation

\[
Bx = b
\]  

with given \( b \in H \), the methods are constructed as follows.

The full version of the GCG-LS method constructs a sequence of search directions \( d_k \) and simultaneously a sequence of approximate solutions \( u_k \) such that the vectors \( Bd_k \) are linearly independent and \( u_k \) minimizes the residual norm corresponding to (1) in the subspace of the first \( k \) search directions. To construct the search directions, the definition also involves an integer \( s \in \mathbb{N} \), further, we let \( s_k = \min\{k, s\} \) \( (k \geq 0) \). (We note that [4] uses different inner products for the construction of \( (d_k) \) and \( (u_k) \), but they are chosen the same when the coincidence of the full and truncated versions is studied. Hence the construction below is given using one and the same inner product.)
(a) Let $u_0 \in H$ be arbitrary, let $d_0 = -r_0 = -(Bu_0 - b)$; for any $k \in \mathbb{N}$: when $u_k, d_k, r_k$ are obtained, let

(b1) the numbers $\alpha_{k-j}^{(k)}$ ($j = 0, \ldots, k$) be the solution of
\[
\sum_{j=0}^{k} \alpha_{k-j}^{(k)} (Bd_{k-j}, Bd_{k-l}) = -(r_k, Bd_{k-l}) \quad (0 \leq l \leq k);
\]
(b2) $u_{k+1} = u_k + \sum_{j=0}^{k} \alpha_{k-j}^{(k)} d_{k-j};$
(b3) $r_{k+1} = r_k + \sum_{j=0}^{k} \alpha_{k-j}^{(k)} Bd_{k-j};$
(b4) $\beta_{k-j}^{(k)} = (Br_{k+1}, Bd_{k-j})/\|Bd_{k-j}\|^2$ ($j = 0, \ldots, s_k$);
(b5) $d_{k+1} = -r_{k+1} + \sum_{j=0}^{k} \beta_{k-j}^{(k)} d_{k-j}.$

The general truncated version uses only the previous $t_k$ search directions, where

a natural number $t \in \mathbb{N}^+$ is given and $1 \leq t_k \leq \min\{t_{k-1} + 1, t\}$ for all $k$. In [4] it is verified that in this case $\alpha_{k-j}^{(k)} = 0$ ($1 \leq j \leq t_{k-1}$), further, the term $\alpha_{k-t_k}^{(k)} d_{k-t_k}$ is called control term since its relative size indicates whether the truncation of the terms for $j > t_k$ has a minor influence or not.

The truncated version with zero control term is as follows.

(a) Let $u_0 \in H$ be arbitrary, let $d_0 = -r_0 = -(Bu_0 - b)$; for any $k \in \mathbb{N}$: when $u_k, d_k, r_k$ are obtained, let

(b1) $\gamma_k = \|Bd_k\|^2,$ $\alpha_k = -\frac{1}{\gamma_k} (Bd_k, r_k);$ (3)
(b2) $u_{k+1} = u_k + \alpha_k d_k;$
(b3) $r_{k+1} = r_k + \alpha_k Bd_k;$
(b4) $\beta_{k-j}^{(k)} = (Br_{k+1}, Bd_{k-j})/\|Bd_{k-j}\|^2$ ($j = 0, \ldots, s_k$);
(b5) $d_{k+1} = -r_{k+1} + \sum_{j=0}^{s_k} \beta_{k-j}^{(k)} d_{k-j}.$

In particular, if $s = 0$ then in (3) the lines (b4) and (b5) reduce to $\beta_k = \frac{1}{\gamma_k} (Br_{k+1}, Bd_k)$ and $d_{k+1} = -r_{k+1} + \beta_k d_k$, respectively.

**Theorem 1** Let $H$ be a real Hilbert space and $B : H \to H$ be a bounded linear operator for which $B + B^* > 0$. Assume that there exists a real polynomial $p_n$ of degree $n$ such that $B^n = p_n(B)$.

If $s \geq n - 1$, then the truncated GCG-LS($s$) method with zero control term for equation (1) coincides with the full version.

**Proof.** For finite-dimensional $H$ the theorem is found in [4], Theorem 4.1. (In fact, there it suffices to assume that $B$ is normal w.r.t. the inner product to ensure
\(B^* = p_n(B)\). Its proof also applies to a general Hilbert space, since the statement concerns the finite-dimensional subspaces spanned by \(d_1, \ldots, d_k\). We note that the main ingredient of the proof uses the properties

\[d_{k-l} = p_{k-l}(B)r_0\]

for some polynomial \(p_{k-l}\) of degree \(k - l\), and that

\[\langle r_{k+1}, Bp_k(B)r_0 \rangle = 0\]

for all polynomials \(p_k\) of degree at most \(k\). (These follow from the construction).

Hence

\[\langle Br_{k+1}, Bd_{k-l} \rangle = \langle r_{k+1}, B^*Bd_{k-l} \rangle = \langle r_{k+1}, Bp_n(B)p_{k-l}(B)r_0 \rangle = 0\]

for \(l \geq n\), i.e. in step (b4) of (2) we have \(\beta^{(k)}_{k-l} = 0\) for \(l \geq n\). Then a brief consideration yields similarly that \(\alpha^{(k)}_{k-j} = 0\) for \(j \geq 1\).

**Remark 1** The proof shows that it suffices to assume that \(B^*r \equiv B\) when \(B\) is not onto.

**Remark 2** In the finite-dimensional case the existence of the polynomial \(p\) in Theorem 1 follows if \(B\) is normal, and the degree \(n\) of \(p\) is at most the order of the matrix \(B\), see [4]. (When \(n\) is minimal, \(p\) and \(n\) are called normal polynomial and normal degree of \(B\), respectively.) Clearly, in the infinite-dimensional case the existence of a normal polynomial cannot be always expected for all normal operators \(B\). If \(B\) is compact, then it can be approximated by its finite-dimensional restrictions, hence for large \(s\) the truncated algorithm will be close to the full one, and the same can be expected for compact perturbations of the identity. The exact formulation of this is not the aim of this paper, since even in the finite-dimensional case the normal degree \(n\) of \(B\) is generally large unless \(n = 1\) (see [4, 7]). Instead, Theorem 1 will be used in the sequel for the case \(n = 1\) under symmetric part preconditioning in order to verify that hereby the truncated GCG-LS(0) method coincides with the full version.

### 3 Preconditioning by the symmetric part in Hilbert space

Let \(H\) be a real Hilbert space with inner product \(\langle \, , \, \rangle\) and corresponding norm \(\| \|\).

In this section we consider a (generally non-symmetric) linear operator \(A : H \to H\) in \(H\), and a corresponding equation

\[Ax = y\]  \quad (4)

with given \(y \in H\). We study separately the cases when \(A\) is a bounded or a densely defined unbounded operator, respectively. In both cases we assume that there exists \(p > 0\) such that

\[\langle Ax, x \rangle \geq p\|x\|^2 \quad (x \in D(A)).\]

(5)

Note that this does not imply the symmetry of \(A\) since \(H\) is real.

We will use the symmetric part \(M = \frac{1}{2}(A + A^*)\) for preconditioning \(A\).
3.1 The case of bounded operators

We consider a bounded linear operator $A : H \to H$ satisfying (5). We define

$$M := \frac{1}{2}(A + A^*), \quad N := \frac{1}{2}(A - A^*),$$

i.e. the symmetric and antisymmetric parts of $A$, respectively.

The following properties are trivial consequences of (5):

**Proposition 1**

(i) $M$ is self-adjoint and $p\|x\|^2 \leq (Mx, x) \leq \|A\|\|x\|^2$ $(x \in H)$;

(ii) $R(M) = H$ and hence $M$ is an isomorphism of $H$;

(iii) the operator $M^{-1}A : H \to H$ exists.

We introduce the energy inner product

$$(u, v)_M = (Mu, v).$$

The notation $\| \cdot \|_M$ is used for the corresponding norm of vectors and also for the related operator norm. We note that part (i) of Proposition 1 gives

$$p\|x\|^2 \leq \|x\|_M^2 \leq \|A\|\|x\|^2 \quad (x \in H)$$

and $H$ remains complete with the equivalent norm $\| \cdot \|_M$ instead of the original one.

We introduce the notation

$$B := M^{-1}A$$

for the operator preconditioned by the symmetric part, and we consider equation (1) with $b = M^{-1}y$ instead of (4).

**Proposition 2** $B$ is a bounded operator w.r.t. $\| \cdot \|_M$.

**Proof.** For any $x \in H$, 

$$\|M^{-1}Ax\|_M = \sup_{\|y\|_M = 1} (M^{-1}Ax, y)_M = \sup_{\|y\|_M = 1} (Ax, y) \leq \sup_{\|y\|_M = 1} \|A\|\|x\|\|y\| \leq \sup_{\|y\|_M = 1} (\|A\|/p)\|x\|_M \|y\|_M = (\|A\|/p)\|x\|_M.$$

**Theorem 2** Let the bounded linear operator $A : H \to H$ satisfy (5) and let $B = M^{-1}A$.

Then the truncated GCG-LS(0) method with zero control term for the equation (1) coincides with the full version.
Proof. Following the finite-dimensional case [4], let $B_M^*$ denote the adjoint of $B$ w.r.t. $\|\cdot\|_M$. Then

$$B_M^* = M^{-1} A^*$$  \hspace{1cm} (10)

since

$$\langle Bx, y \rangle_M = \langle Ax, y \rangle = \langle x, A^*y \rangle_M = \langle x, M^{-1}A^*y \rangle_M \quad (x, y \in H).$$  \hspace{1cm} (11)

Here

$$B = M^{-1}A = M^{-1}(M + N) = I + M^{-1}N,$$

hence

$$B_M^* = M^{-1}A^* = M^{-1}(M - N) = I - M^{-1}N = 2I - B = p(B),$$

and therefore the degree of the polynomial $p$ in Theorem 1 is $n = 1$. Hence Theorem 1 with $s = 0$ yields the required result.

Remark 3 For the preconditioned operator $B = M^{-1}A$, using the relation $\langle Bx, y \rangle_M = \langle Ax, y \rangle$, the iterative sequence in the truncated version (3) with inner product (7) takes the following form,

\[
\begin{cases}
(a) & \text{Let } u_0 \in H \text{ be arbitrary, and let } r_0 \text{ be the solution of } Mr_0 = Au_0 - y; \quad d_0 = -r_0; \\
& \text{for any } k \in \mathbb{N} : \text{ when } u_k, d_k, r_k \text{ are obtained, let } \\
(b1) & e_k \text{ be the solution of } Me_k = Ad_k; \\
& \gamma_k = \langle Ad_k, e_k \rangle, \quad \alpha_k = -\frac{1}{\gamma_k} \langle Ad_k, r_k \rangle; \\
(b2) & u_{k+1} = u_k + \alpha_k d_k; \\
(b3) & r_{k+1} = r_k + \alpha_k e_k; \\
(b4) & \beta_k = \frac{1}{\gamma_k} \langle Ar_{k+1}, e_k \rangle; \\
(b5) & d_{k+1} = -r_{k+1} + \beta_k d_k.
\end{cases}
\]

Now we estimate the conditioning of $B$ using the properties of $M$ and $N$.

Proposition 3 Let the bounded linear operator $A : H \rightarrow H$ satisfy (5). Then

1. $N$ is $M$-bounded, i.e. there exists a number $L > 0$ such that

$$|\langle N x, y \rangle| \leq L \langle Mx, x \rangle^{1/2} \langle My, y \rangle^{1/2} \quad (x, y \in H);$$  \hspace{1cm} (14)

2. there holds

$$\|M^{-1}N\|_M \leq L$$  \hspace{1cm} (15)

with $L$ from (14).
Proof. (1) Similarly to Proposition 2, a (rather pessimistic) bound is
\[ |\langle Nx, y \rangle| \leq \|N\| \|x\| \|y\| \leq (\|N\|/p) \langle Mx, x \rangle^{1/2} \langle My, y \rangle^{1/2}. \]

(2) For any \( x \in H \),
\[ \|M^{-1}Nx\|_M = \sup_{\|y\|_M=1} \langle M^{-1}Nx, y \rangle_M = \sup_{\|y\|_M=1} |\langle Nx, y \rangle| \leq \sup_{\|y\|_M=1} L \|x\|_M \|y\|_M = L \|x\|_M. \]  

Remark 4  
(i) The \( M \)-boundedness of \( N \) means that the bilinear form \( x, y \mapsto \langle Nx, y \rangle \) is bounded w.r.t the norm \( \|\cdot\|_M \). (By part (2) of Proposition 3, this is equivalent to the boundedness of the operator \( M^{-1}N \) w.r.t \( \|\cdot\|_M \).)

(ii) It follows similarly that \( A \) is also \( M \)-bounded, which means that the bilinear form \( x, y \mapsto \langle Ax, y \rangle \) is bounded w.r.t the norm \( \|\cdot\|_M \). (In fact, this is contained in the proof of Proposition 2, which formulates the corresponding statement for the operator \( M^{-1}A \).)

Proposition 4  
The operator \( B \) satisfies
\[ \text{cond}_M(B) \leq 1 + L, \]
where \( \text{cond}_M(B) = \|B\|_M \|B^{-1}\|_M \) and \( L \) is from (14).

Proof. We have
\[ \|x\|_M^2 = \langle Mx, x \rangle = \langle Ax, x \rangle = \langle Bx, x \rangle_M \leq \|Bx\|_M \|x\|_M \quad (x \in H), \]
hence
\[ \|B^{-1}\|_M \leq 1. \]

Further, by (12) \( B = I + M^{-1}N \), hence
\[ \|B\|_M \leq 1 + \|M^{-1}N\|_M \leq 1 + L. \]

3.2 The case of unbounded operators

The boundedness of \( A \) in the previous subsection was used in two major aspects: it implied that \( M \) is onto (hence \( B \) exists) in Proposition 1, and that \( B \) is bounded w.r.t. \( \|\cdot\|_M \) in Proposition 2. For unbounded \( A \), the second property will be ensured by the \( M \)-boundedness of \( A \), whereas the first property will be generalized in two ways: first \( M \) will be assumed to be self-adjoint (\( M^* = M \)) to ensure \( R(M) = H \), second, if \( M \) is only symmetric (which means that \( M^* \) is an extension of \( M \), i.e. \( D(M^*) \supset D(M) \)), then \( B \) will be defined in a weak sense.
In this subsection let $A$ be an unbounded operator in $H$ whose domain is a dense subspace $D(A)$ of $H$. Then the adjoint operator $A^*$ is well-defined on its domain such that
\[ \langle Ax, y \rangle = \langle x, A^*y \rangle \quad (x \in D(A), y \in D(A^*)). \]
The symmetric and antisymmetric parts of $A$ are
\[ M := \frac{1}{2}(A + A^*), \quad N := \frac{1}{2}(A - A^*) \quad \text{with} \quad D(M) = D(N) = D(A) \cap D(A^*). \]

We will assume the $M$-boundedness of $N$ (equivalent to that of $A$, since $A = M + N$). We note that now this is indeed not automatically satisfied, e.g. consider $A = I + N$ with some unbounded antisymmetric operator $N$. Then $M = I$, hence $N$ is not $M$-bounded. (For instance, the densely defined operator $u \mapsto u'$ in $L^2(I)$ for some interval $I \subset \mathbb{R}$ can play the role of $N$.)

(a) **The case of self-adjoint $M$**

Let us assume that $M$ is self-adjoint, i.e. $M^* = M$. Further, since $A$ satisfies (5), there holds
\[ \langle Mx, x \rangle \geq p\|x\|^2 \quad (x \in D(A)). \]
With this, the self-adjointness of $M$ yields the following property:

**Proposition 5** (see e.g. [12]). If $M$ is self-adjoint and satisfies (18), then $R(M) = H$.

Consequently, the operator
\[ B = M^{-1}A \]
exists just as before, now with domain
\[ D(B) = D(A). \]

**Proposition 6** Assume that $N$ is $M$-bounded, i.e. (14) holds for $x, y \in D(M) = D(N)$. Then $B$ is a bounded operator w.r.t. $\| \cdot \|_M$ on $D(M)$.

**Proof.** For any $x \in D(M)$ we can repeat the estimate (16), hence $M^{-1}N$ is bounded w.r.t. $\| \cdot \|_M$. There holds
\[ B = M^{-1}A = M^{-1}(M + N) = I + M^{-1}N \]
on $D(M)$, hence $B|_{D(M)}$ is bounded w.r.t. $\| \cdot \|_M$.

**Remark 5** The operator $B$ can be uniquely extended by preserving boundedness to the energy space $H_M$ of $M$, which is the completion of $D(M)$ w.r.t. $\| \cdot \|_M$ and can be regarded as a subspace of $H$. 

8
Theorem 3 Let the densely defined linear operator $A$ in $H$ satisfy (5) and let us assume that $M$ is self-adjoint and $N$ is $M$-bounded. Let $B = M^{-1}A$.

Then the truncated GCG-LS(0) method with zero control term for the equation (1) coincides with the full version.

Proof. Now it can be verified that

$$B''|_{D(M)} = 2I - B.$$ 

Namely, for any $x \in D(A)$ and $y \in D(M)$

$$\langle Bx, y \rangle_M = \langle Ax, y \rangle = \langle x, A^*y \rangle = \langle x, 2My - Ay \rangle = \langle x, (2I - B)y \rangle_M,$$

i.e. $B''y = (2I - B)y$. Since $R(B) \subset D(M)$, we can use Remark 1 to conclude (similarly to Theorem 2) that Theorem 1 holds with $s = 0$ for $B$ w.r.t $\| \cdot \|_M$.

We note that the iterative sequence takes again the form as in Remark 3.

(b) The case of non self-adjoint $M$

Now we turn to the case when $M$ is only symmetric, i.e. $M^*$ is a proper extension of $M$. In this case Proposition 5 cannot be used, hence $B$ has to be defined in a weak sense instead of (19). We recall that, still, (18) holds and the energy space $H_M$ can be defined as before.

The definition of $B$ in weak sense preserves only the inner product relation connecting $A$ and $B$.

Proposition 7 Assume that $N$ is $M$-bounded, i.e. (14) holds for $x, y \in D(M) = D(N)$. Then there exists a unique bounded linear operator $B : H_M \to H_M$ satisfying

$$\langle Bx, y \rangle_M = \langle Ax, y \rangle \quad (x, y \in D(M)).$$

Proof. The bilinear form $x, y \mapsto \langle Ax, y \rangle$, defined for $x, y \in D(M)$, is bounded w.r.t the norm $\| \cdot \|_M$ by Remark 4. Hence it has a unique bounded extension $\Phi : H_M \times H_M \to \mathbb{R}$. The Riesz theorem implies that there exists a unique bounded operator $B : H_M \to H_M$ satisfying

$$\langle Bx, y \rangle_M = \Phi(x, y) \quad (x, y \in H_M).$$

For $x, y \in D(M)$ this gives the desired result.

Theorem 4 Let the densely defined linear operator $A$ in $H$ satisfy (5) and let $B : H_M \to H_M$ be defined as in Proposition 7. Assume that $N$ is $M$-bounded.

Then the truncated GCG-LS(0) method with zero control term for the equation (1) coincides with the full version.
Proof. Let \( x, y \in D(M) \). Then the calculation in (20) can be repeated, hence

\[
(Bx, y)_M = (x, (2I - B)y)_M \quad (x, y \in D(M)).
\]

This also holds for all \( x \in H_M \) since \( D(M) \) is dense in \( H_M \), hence

\[
B^* y = (2I - B)y \quad (y \in D(M))
\]
or

\[
B^*|_{D(M)} = 2I - B|_{D(M)}.
\]

Since these two operators are bounded, using the density of \( D(M) \) in \( H_M \) again we obtain that

\[
B^* = 2I - B.
\]

Hence, similarly as before, Theorem 1 holds with \( s = 0 \) for \( B \) in \( H_M \).

Remark 6 We note that now, without the decomposition (19), the iterative sequence cannot be written in the form as in Remark 3. Instead, we have to use the original version (3) for \( B \) in \( H_M \) instead of \( H \).

4 Preconditioning nonsymmetric linear elliptic operators in Sobolev space

As an application of the general theory, the conjugate gradient method is used in this section for certain linear elliptic operators on the continuous level in Sobolev spaces. First, using the Hilbert space results of the previous section, it is verified that the truncated GCG-LS(0) method with zero control term coincides with the full version. Then the conditioning of the operator is estimated using the properties of the symmetric and antisymmetric parts \( M \) and \( N \). The main benefit of the latter appears when the CGM is applied to the discretized problem, thereby the Sobolev space results give mesh independent estimates of the condition number. We note that related preconditioning methods have been studied in many papers, where an operator with symmetric principal part (plus sometimes a zeroth-order term) is used as preconditioner for a nonsymmetric equation, see e.g. [6, 11]. The main point in our results concerning the rate of convergence is the following: the coincidence of the full and truncated versions of the CGM implies that the corresponding mesh independent convergence properties of the full CGM are achieved by the simpler truncated version.

4.1 Convection-diffusion equations

We consider convection-diffusion equations

\[
\begin{cases}
-\Delta u + b \cdot \nabla u + cu = f \\
u|_{\partial \Omega} = 0
\end{cases}
\]

(21)
on a bounded domain $\Omega \subset \mathbb{R}^n$, where $b \in W^{1,\infty}(\Omega)^n$ and $c \in L^\infty(\Omega)$, further, there holds the usual coercivity condition
\[ c - \frac{1}{2} \text{div} b \geq 0. \tag{22} \]

In this study we deal only with a regularly perturbed problem.

The Sobolev space $H^1_0(\Omega)$ is endowed with the inner product
\[ \langle u, v \rangle_{H^1_0(\Omega)} = \int_{\Omega} \nabla u \cdot \nabla v, \tag{23} \]
which is the energy inner product corresponding to the minus Laplacian since for $u \in H^2(\Omega) \cap H^1_0(\Omega)$ there holds
\[ \langle u, v \rangle_{H^1_0(\Omega)} = -\int_{\Omega} (\Delta u) v. \tag{24} \]

The weak differential operator $A : H^1_0(\Omega) \to H^1_0(\Omega)$ corresponding to (21) is defined by
\[ \langle Au, v \rangle_{H^1_0(\Omega)} = \int_{\Omega} (\nabla u \cdot \nabla v + (b \cdot \nabla u)v + cuv) \quad (u, v \in H^1_0(\Omega)), \tag{25} \]
and then the weak formulation of problem (21) is
\[ \langle Au, v \rangle_{H^1_0(\Omega)} = \int_{\Omega} f v \quad (v \in H^1_0(\Omega)). \tag{26} \]

We note that
\[ A = I + C \]
where $I : H^1_0(\Omega) \to H^1_0(\Omega)$ is the identity operator and
\[ \langle Cu, v \rangle_{H^1_0(\Omega)} = \int_{\Omega} ((b \cdot \nabla u)v + cuv) \quad (u, v \in H^1_0(\Omega)), \tag{27} \]
or by (24)
\[ Cu = -\Delta^{-1}((b \cdot \nabla u) + cu). \]

**Proposition 8** The adjoint of $A$ is given by
\[ \langle A^* u, v \rangle_{H^1_0(\Omega)} = \int_{\Omega} (\nabla u \cdot \nabla v + u(b \cdot \nabla v) + cuv) \quad (u, v \in H^1_0(\Omega)). \tag{28} \]

The symmetric and antisymmetric parts $M := \frac{1}{2}(A + A^*)$ and $N := \frac{1}{2}(A - A^*)$ are given by
\[ \langle Mu, v \rangle_{H^1_0(\Omega)} = \int_{\Omega} (\nabla u \cdot \nabla v + (c - \frac{1}{2} \text{div} b) uv) \quad (u, v \in H^1_0(\Omega)) \tag{29} \]
and
\[ \langle Nu, v \rangle_{H^1_0(\Omega)} = \frac{1}{2} \int_{\Omega} ((b \cdot \nabla u)v - u(b \cdot \nabla v)) \quad (u, v \in H^1_0(\Omega)), \tag{28} \]
respectively.
Proof. The relations (27) and (29) follow directly from (25). Further, (28) follows from
\[ \int_{\Omega} ((\mathbf{b} \cdot \nabla u)v + u(\mathbf{b} \cdot \nabla v)) = -\int_{\Omega} (\text{div} \mathbf{b}) uv, \]
which is a consequence of the formula
\[ 0 = \int_{\partial \Omega} \mathbf{b} uv \, d\sigma = \int_{\Omega} \text{div} (\mathbf{b} uv) = \int_{\Omega} [(\text{div} \mathbf{b}) uv + (\mathbf{b} \cdot \nabla u)v + u(\mathbf{b} \cdot \nabla v)]. \]

Remark 7 A similar calculation as above implies that the first order terms \((\mathbf{b} \cdot \nabla u)v\) and \(u(\mathbf{b} \cdot \nabla v)\) in (25) and (27), respectively, can be replaced by \(-u \text{div} (\mathbf{b} v)\) and \(-\text{div} (\mathbf{b} u)v\).

Theorem 5 Let \(B = M^{-1}A\). Then the truncated GCG-LS(0) method with zero control term for the problem (26) coincides with the full version.

Proof. It is well-known that the operator \(A\) in (25) is bounded, further, the coercivity condition (22) implies that
\[ (Au,u)_{H^1_0} = (Mu,u)_{H^1_0} = \int_{\Omega} (|\nabla u|^2 + (c - \frac{1}{2} \text{div} \mathbf{b}) u^2) \geq \int_{\Omega} |\nabla u|^2 = ||u||_{H^1_0}^2, \]
that is, \(A\) satisfies (5) with \(p = 1\). Hence Theorem 2 can be applied.

We note that – by (13) – the iterative sequence in the truncated version takes the
(a) Let \( u_0 \in H^1_0(\Omega) \) be arbitrary, let \( r_0 = -d_0 \) be the solution of the problem
\[
\int_\Omega \left( \nabla r_0 \cdot \nabla v + (c - \frac{1}{2} \text{div} \, b)r_0 v \right) = \int_\Omega \left( \nabla u_0 \cdot \nabla v + (b \cdot \nabla u_0) v + cu_0 v - fv \right) \quad (v \in H^1_0(\Omega));
\]
for any \( k \in \mathbb{N} \): when \( u_k, d_k, r_k \) are obtained, let
\[
\begin{align*}
\text{(b1)} & \quad e_k \text{ be the solution of the problem} \\
& \quad \int_\Omega \left( \nabla e_k \cdot \nabla v + (c - \frac{1}{2} \text{div} \, b)e_k v \right) = \int_\Omega \left( \nabla d_k \cdot \nabla v + (b \cdot \nabla d_k) v + cd_k v \right) \quad (v \in H^1_0(\Omega)); \\
& \quad \gamma_k = \int_\Omega \left( \nabla d_k \cdot \nabla e_k + (b \cdot \nabla d_k) e_k + cd_k e_k \right), \\
& \quad \alpha_k = -\frac{1}{\gamma_k} \int_\Omega \left( \nabla d_k \cdot \nabla r_k + (b \cdot \nabla d_k) r_k + cd_k r_k \right); \\
\text{(b2)} & \quad u_{k+1} = u_k + \alpha_k d_k; \\
\text{(b3)} & \quad r_{k+1} = r_k + \alpha_k e_k; \\
\text{(b4)} & \quad \beta_k = \frac{1}{\gamma_k} \int_\Omega \left( \nabla r_{k+1} \cdot \nabla e_k + (b \cdot \nabla r_{k+1}) e_k + c r_{k+1} e_k \right); \\
\text{(b5)} & \quad d_{k+1} = -r_{k+1} + \beta_k d_k.
\end{align*}
\]

Now the conditioning of the operator \( B \) will be estimated using the properties of \( M \) and \( N \).

**Proposition 9** Consider (28)-(29). The operator \( N \) is \( M \)-bounded with constant
\[
L = \lambda^{-1/2} \|b\|_\infty,
\]
where \( \|b\|_\infty = \sup_{\Omega} |b| = \sup_{\Omega} (\sum_i b^2_i)^{1/2} \) and \( \lambda > 0 \) is the smallest eigenvalue of \(-\Delta\) on \( H^2(\Omega) \cap H^1_0(\Omega) \).

**Proof.** We have
\[
\int_\Omega |(b \cdot \nabla u) v| \leq \|b \cdot \nabla u\|_{L^2} \|v\|_{L^2} \leq \|b\|_\infty \|\nabla u\|_{L^2} \|v\|_{L^2} \leq \lambda^{-1/2} \|b\|_\infty \|\nabla u\|_{L^2} \|\nabla v\|_{L^2} \leq \lambda^{-1/2} \|b\|_\infty \langle Mu, u \rangle^{1/2} \langle Mv, v \rangle^{1/2}.
\]
The same estimate holds if \( u \) and \( v \) are exchanged and hence also for \( \langle Nu, v \rangle \).
The constant $L$ in (31) gives a concrete value for the bound required in the similar general setting in (14). Hence we can apply Proposition 4:

**Corollary 1** The operator $A$ in (25) satisfies the estimate

$$\text{cond}_M(B) \leq 1 + \lambda^{-1/2} \|b\|_\infty,$$

where $\text{cond}_M(B) = \|B\|_M \|B^{-1}\|_M$.

**Remark 8** The main benefit of the estimate (32) appears when the truncated GCG-LS(0) method is applied to a finite element discretization of (26). Namely, the related condition numbers for the discretized problems are estimated by (32) in a mesh independent way, and can be a priori determined before discretization. The number $\|b\|_{L^\infty}$ can be calculated directly from the given coefficients, further, a possible estimate for $\lambda$ (see e.g. [15]) is

$$\lambda \geq n\pi^2/diam(\Omega)^2.$$ (33)

**Remark 9** Under mild restrictions one can use the strong forms of the operators $A$ and $M$ to define the iteration. Namely, assume that $\Omega$ is $C^2$-diffeomorphic to a convex domain, $b \in C^1(\Omega)^n$ and $c \in C(\Omega)$. Then $c - \frac{1}{2} \text{div} b \in C(\Omega)$, hence the operator

$$Mu = -\Delta u + (c - \frac{1}{2} \text{div} b) u \quad (u \in H^2(\Omega) \cap H^1_0(\Omega))$$

satisfies $R(M) = L^2(\Omega)$ due to a regularity result [8]. One can also define the original operator in $H^2(\Omega) \cap H^1_0(\Omega)$: let

$$Au = -\Delta u + b \cdot \nabla u + cu \quad (u \in H^2(\Omega) \cap H^1_0(\Omega)).$$ (35)

In this case $M$ is a self-adjoint unbounded operator and one can apply Theorem 3. The iterative sequence almost coincides with (30), namely, in (30) the auxiliary equations in (a) and (b) take the form

$$\begin{align*}
(a) \quad -\Delta r_0 + (c - \frac{1}{2} \text{div} b) r_0 &= -\Delta w_0 + b \cdot \nabla w_0 + cw_0 - f, \quad r_0|_{\partial \Omega} = 0, \\
(b) \quad -\Delta e_k + (c - \frac{1}{2} \text{div} b) e_k &= -\Delta d_k + b \cdot \nabla d_k + cd_k, \quad e_k|_{\partial \Omega} = 0
\end{align*}$$

with $r_0, e_k \in H^2(\Omega)$ (and by induction $r_k, e_k, u_k \in H^2(\Omega)$). The benefit of the discussed $H^2$-regularity appears in the FEM solution of (36), where one can use the FEM error estimates that contain the $H^2$-norm of $r_0$ and $e_k$.

If the above regularity assumptions do not hold, then Theorem 4 applies to the operators (34)–(35), in which case the obtained iteration coincides with the previously discussed weak form.
4.2 Linearized diffusion problems

(a) The original nonlinear problem

Let us consider the nonlinear diffusion equation

\[
\begin{aligned}
- \text{div} (a(x, w) \nabla w) &= f \\
\varepsilon w |_{\partial \Omega} &= 0
\end{aligned}
\]  

(37)

on a bounded domain \( \Omega \subset \mathbb{R}^n \) \((n = 2 \text{ or } 3)\) with the following conditions: \( a \in C^2(\Omega \times \mathbb{R}) \), the partial derivatives \( a_w(x, w), a_{ww}(x, w) \) and \( a_{xw}(x, w) = \{a_{x_i w}(x, w)\}_{i=1}^n \) are bounded, and there holds

\[
0 < \alpha \leq a(x, w) \leq \tilde{\alpha} \quad (x \in \Omega, w \in \mathbb{R})
\]

(38)

with suitable constants \( \alpha, \tilde{\alpha} \). The source term \( f \) satisfies \( f \in L^2(\Omega) \) and \( f \geq 0 \).

The weak formulation reads

\[
\int_{\Omega} a(x, w) \nabla w \cdot \nabla v = \int_{\Omega} fv \quad (v \in H_0^1(\Omega)).
\]

(b) The linearized equation

In the sequel we fix \( w \in H_0^1(\Omega) \) and study the linearized equations around \( w \). That is, given \( w \) and some \( g \in L^2(\Omega) \), one seeks \( u \in H_0^1(\Omega) \) such that

\[
\int_{\Omega} (a(x, w) \nabla u \cdot \nabla v + u (a_w(x, w) \nabla w \cdot \nabla v)) = \int_{\Omega} gv \quad (v \in H_0^1(\Omega)).
\]

(39)

Typically such equations appear in the course of a Newton-like linearization, in which case

\[
w = w_k
\]

(40)

is the \( k \)th Newton iterate and \( g \) also comes from the previous iteration.

The corresponding linear operator \( A : H_0^1(\Omega) \rightarrow H_0^1(\Omega) \) is defined by

\[
\langle Au, v \rangle_{H_0^1} = \int_{\Omega} (a(x, w) \nabla u \cdot \nabla v + u (a_w(x, w) \nabla w \cdot \nabla v)) \quad (u, v \in H_0^1(\Omega)).
\]

(41)

Then equation (39) is written as

\[
\langle Au, v \rangle_{H_0^1} = \int_{\Omega} gv \quad (v \in H_0^1(\Omega)).
\]

For simplicity, we assume that the regularity \( w \in H^2(\Omega) \) is satisfied. Further, we introduce the notation

\[
b(w) = a_w(x, w) \nabla w
\]

(42)
and assume that the coercivity condition

\[ \text{div} b(w) \leq 0 \]  

holds. This corresponds to (22) since now \( c = 0 \), i.e. there is no term of order zero. (We note that (43) expresses that the property \( \text{div} \ (a(x, w) \nabla w) \leq 0 \), which follows from (37), is preserved when \( a \) is replaced by \( a_w \).) Using (42), formula (41) reduces to

\[
(Au, v)_{H^1_0} = \int_{\Omega} (a(x, w) \nabla u \cdot \nabla v + w (b(w) \cdot \nabla v)) \quad (u, v \in H^1_0(\Omega)).
\]

Before formulating the symmetric and antisymmetric parts of \( A \), we first verify that the relations (41) and (43) are well-defined, i.e. there exist \( A \) as a bounded operator in \( H^1_0(\Omega) \) and \( \text{div} \ b(w) \) as an element of \( L^2(\Omega) \).

**Proposition 10**  
(1) The relation (41) defines a bounded operator \( A : H^1_0(\Omega) \rightarrow H^1_0(\Omega) \).

(2) There holds

\[ \text{div} b(w) \in L^2(\Omega). \]

**Proof.**  
(1) There holds the Sobolev embedding

\[ H^1(\Omega) \subset L^4(\Omega) \]  

for \( n \leq 4 \) (see [1]), hence \( u \) and \( \nabla w \) in (41) are in \( L^4(\Omega) \). Using the corresponding estimate for \( H^1_0(\Omega) \):

\[ \|u\|_{L^4} \leq K_4 \|u\|_{H^1_0} \quad (u \in H^1_0(\Omega)) \]

with a suitable constant \( K_4 > 0 \), the terms in the integral on the r.h.s. of (41) can be estimated as follows:

\[
\int_{\Omega} |a(x, w) \nabla u \cdot \nabla v| \leq \tilde{\alpha} \|u\|_{H^1_0} \|v\|_{H^1_0},
\]

and

\[
\int_{\Omega} |a(a_w(x, w) \nabla w \cdot \nabla v)| \leq \|a_w\|_{L^\infty} \|u\|_{L^1} \|\nabla w\|_{L^4} \|\nabla v\|_{L^2} \leq K_4 \|a_w\|_{L^\infty} \|u\|_{H^1_0} \|\nabla w\|_{L^4} \|v\|_{H^1_0}. \]

Hence the integral can be estimated by

\[
(\tilde{\alpha} + K_4 \|a_w\|_{L^\infty} \|\nabla w\|_{L^4}) \|u\|_{H^1_0} \|v\|_{H^1_0} \quad (u, v \in H^1_0(\Omega)),
\]

i.e. it defines a bounded bilinear form on \( H^1_0(\Omega) \) in \( u \) and \( v \). Then the Riesz theorem ensures the existence of a unique bounded operator defined by (41).

(2) We have

\[
\text{div} b(w) = \text{div} \ (a_w(x, w) \nabla w) = a_w(x, w) \Delta w + a_{xw}(x, w) \cdot \nabla w + a_{ww}(x, w) |\nabla w|^2.
\]
Here $a_w(x,w)$, $a_{vw}(x,w)$ and $a_{ww}(x,w)$ are bounded by assumption. On the other hand, $w \in H^2(\Omega)$ yields that $\Delta w$ and $\nabla w$ are in $L^2(\Omega)$, further, using (45), we have $\nabla w \in H^1(\Omega) \subset L^4(\Omega)$ and hence $|\nabla w|^2 \in L^2(\Omega)$. Therefore each term is in $L^2(\Omega)$, which implies that their sum $\text{div } b(w)$ is also in $L^2(\Omega)$.

**Proposition 11** The adjoint of $A$ is given by

$$\langle A^* u, v \rangle_{H_0^1} = \int_{\Omega} \left( a(x,w) \nabla u \cdot \nabla v + (b(w) \cdot \nabla u) v \right) \quad (u, v \in H_0^1(\Omega)).$$

The symmetric and antisymmetric parts $M := \frac{1}{2}(A + A^*)$ and $N := \frac{1}{2}(A - A^*)$ are given by

$$\langle Mu, v \rangle_{H_0^1} = \int_{\Omega} \left( a(x,w) \nabla u \cdot \nabla v - \frac{1}{2} \left( \text{div } b(w) \right) uv \right) \quad (u, v \in H_0^1(\Omega))$$

and

$$\langle Nu, v \rangle_{H_0^1} = \frac{1}{2} \int_{\Omega} \left( u (b(w) \cdot \nabla v) - (b(w) \cdot \nabla u) v \right) \quad (u, v \in H_0^1(\Omega)),$$

respectively.

**Proof.** The relations (48) and (50) follow directly from (41), further, (49) is obtained similarly as in Proposition 8.

**Remark 10** Similarly to Remark 7, the first order terms $u(b(w) \cdot \nabla v)$ and $(b(w) \cdot \nabla u)v$ in (44) and (48), respectively, can be replaced by $-\text{div } (b(w)u)v$ and $-u \text{div } (b(w)v)$.

**Theorem 6** Let $B = M^{-1}A$. Then the truncated GCG-LS(0) method with zero control term for the problem (39) coincides with the full version.

**Proof.** Theorem 2 can be applied, since by Proposition 10 the operator $A$ in (41) is bounded, further, the positivity and coercivity conditions (38) and (43) imply that

$$\langle Au, u \rangle_{H_0^1} = \langle Mu, u \rangle_{H_0^1} = \int_{\Omega} \left( a(x,w) |\nabla u|^2 - \frac{1}{2} \left( \text{div } b(w) \right) u^2 \right) \geq \alpha \int_{\Omega} |\nabla u|^2 = \alpha \|u\|^2_{H_0^1},$$

i.e. $A$ satisfies (5) with $p = \alpha$.

We note that – by (13) – the iterative sequence in the truncated version takes the
form

(a) Let \( u_0 \in H^1_0(\Omega) \) be arbitrary, let \( r_0 = -d_0 \) be the solution of the problem

\[
\int_{\Omega} \left( a(x, w) \nabla r_0 \cdot \nabla v - \frac{1}{2} (\text{div} b(w)) r_0 v \right) = \int_{\Omega} \left( a(x, w) \nabla u_0 \cdot \nabla v + u_0 (b(w) \cdot \nabla v) + cu_0 v - gv \right) \quad (v \in H^1_0(\Omega));
\]

for any \( k \in \mathbb{N} \) : when \( u_k, d_k, r_k \) are obtained, let

(b1) \( \epsilon_k \) be the solution of the problem

\[
\int_{\Omega} \left( a(x, w) \nabla \epsilon_k \cdot \nabla v - \frac{1}{2} (\text{div} b(w)) \epsilon_k v \right) = \int_{\Omega} \left( a(x, w) \nabla d_k \cdot \nabla v + d_k (b(w) \cdot \nabla v) + cd_k \epsilon_k \right) \quad (v \in H^1_0(\Omega));
\]

\[
\gamma_k = \frac{1}{\gamma_0} \int_{\Omega} \left( a(x, w) \nabla r_k \cdot \nabla \epsilon_k + d_k (b(w) \cdot \nabla \epsilon_k) + cd_k \epsilon_k \right),
\]

\[
\alpha_k = -\frac{1}{\gamma_0} \int_{\Omega} \left( a(x, w) \nabla \epsilon_k \cdot \nabla r_k + d_k (b(w) \cdot \nabla r_k) + cd_k \epsilon_k \right);
\]

(b2) \( u_{k+1} = u_k + \alpha_k d_k \);

(b3) \( r_{k+1} = r_k + \epsilon_k d_k \);

(b4) \( \beta_k = \frac{1}{\gamma_0} \int_{\Omega} \left( a(x, w) \nabla r_{k+1} \cdot \nabla \epsilon_k + r_{k+1} (b(w) \cdot \nabla \epsilon_k) + cr_{k+1} \epsilon_k \right) \);

(b5) \( d_{k+1} = -r_{k+1} + \beta_k d_k \).

Now the conditioning of the operator \( B \) will be estimated using the properties of \( M \) and \( N \).

**Proposition 12** Consider (49)-(50). The operator \( N \) is \( M \)-bounded with constant

\[
L = \frac{K_4}{\alpha} \|a_w\|_{L^\infty} \|\nabla w\|_{L^4}.
\]

**Proof.** By the estimate (47), we have

\[
\int_{\Omega} |u(b(w) \cdot \nabla v)| \leq K_4 \|a_w\|_{L^\infty} \|\nabla w\|_{L^4} \|u\|_{H^1_0} \|v\|_{H^1_0},
\]

and the same is obtained for \( |v(b(w) \cdot \nabla u)| \) by exchanging \( u \) and \( v \). Hence

\[
|\langle Nu, v \rangle_{H^1_0} | \leq K_4 \|a_w\|_{L^\infty} \|\nabla w\|_{L^4} \|u\|_{H^1_0} \|v\|_{H^1_0} \quad (u, v \in H^1_0(\Omega)).
\]

Here (51) implies

\[
\|u\|_{H^1_0} \leq \alpha^{-1/2} (Mu, u)_{H^1_0},
\]

18
hence
\[ |\langle Nu, v \rangle_{H^2_0} | \leq L \langle Mu, u \rangle_{H^2_0}^{1/2} \langle Mv, v \rangle_{H^2_0}^{1/2}, \quad (u, v \in H^1_0(\Omega)) \]
with \( L \) in (52).

**Corollary 2** The operator \( A \) in (41) satisfies the estimate
\[ \text{cond}_M(B) \leq 1 + L, \quad (53) \]
where \( \text{cond}_M(B) = \|B\|_M \|B^{-1}\|_M \) and \( L \) is from (52).

**Remark 11** The value of \( L \) in (52) yields a mesh independent estimate for (53) when applied for the discretized problems, and can be a priori determined before discretization. Namely, here \( w \) is taken as a fixed known function (typically a previous Newton iterate as in (40), and the unknown function is now \( u \)), hence the numbers \( \alpha \), \( \|a_w\|_{L^\infty} \) and \( \|\nabla w\|_{L^2} \) can be calculated directly from the given coefficients and \( w \), further, \( K_4 \) can be estimated in terms of the domain.

The estimate of \( K_4 \) can be carried out in the following way. We use an inequality from [10]:
\[ \|u\|_{L^4}^4 \leq 2\|u\|_{L^2}^2 \|\nabla u\|_{L^2}^2, \quad (u \in H^1_0(\Omega)), \]
and that
\[ \lambda \|u\|_{L^2}^2 \leq \|\nabla u\|_{L^2}^2 = \|u\|_{H^1_0}^2, \quad (u \in H^1_0(\Omega)) \]
where \( \lambda > 0 \) is the smallest eigenvalue of \( -\Delta \) on \( H^2(\Omega) \cap H^1_0(\Omega) \). Hence
\[ \|u\|_{L^4}^4 \leq (2/\lambda) \|u\|_{H^1_0}^3, \quad (u \in H^1_0(\Omega)), \]
that is,
\[ K_4^4 \leq 2/\lambda. \]
A possible estimate for \( \lambda \) is (33), by use of which we obtain
\[ K_4^4 \leq 2 \text{diam}(\Omega)^2/n\pi^2. \]

**References**


