Estimates of error norms in the conjugate gradient algorithm

Gérard MEURANT

October, 2008
1. Error norms in solving linear systems
2. Formulas for the $A$–norm of the error in CG
3. Estimates of the $A$–norm of the error
4. Estimates of the $l_2$ norm of the error
5. Relation with finite element problems
6. Numerical experiments
Error norms in solving linear systems

Let $A$ be an SPD matrix of order $n$ and $\tilde{x}$ an approximate solution of

$$Ax = c$$

The residual $r$ is defined as

$$r = c - A\tilde{x}$$

The error $\epsilon$ being defined as $\epsilon = x - \tilde{x}$

$$\epsilon = A^{-1}r$$

The $A$–norm of the error is

$$\|\epsilon\|_A^2 = \epsilon^T A \epsilon = r^T A^{-1} A A^{-1} r = r^T A^{-1} r$$

and the $l_2$ norm is $\|\epsilon\|^2 = r^T A^{-2} r$
bounds can be obtained by running \( N \) iterations of the Lanczos algorithm

\[
\| \mathbf{r} \|^2 (e^1)^T (J_N)^{-i} e^1
\]

however, it does not make to much sense to run Lanczos to bound the error norm of CG!

What can we do for CG?
Formulas for the $A$–norm of the error in CG

Theorem

The square of the $A$–norm of the error at CG iteration $k$ is given by

$$
\| \epsilon^k \|^2_A = \| r^0 \|^2 \left[ (J_n^{-1} e^1, e^1) - (J_k^{-1} e^1, e^1) \right]
$$

where $n$ is the order of the matrix $A$ and $J_k$ is the Jacobi matrix of the Lanczos algorithm whose coefficients can be computed from those of CG. Moreover

$$
\| \epsilon^k \|^2_A = \| r^0 \|^2 \left[ \sum_{j=1}^n \frac{[(z_{(n)}^j)_1]^2}{\lambda_j} - \sum_{j=1}^k \frac{[(z_{(k)}^j)_1]^2}{\theta_{j}^{(k)}} \right]
$$

where $z_{(k)}^j$ is the $j$th normalized eigenvector of $J_k$ corresponding to the eigenvalue $\theta_{j}^{(k)}$. 
Proof.
We have $A\epsilon^k = r^k = r^0 - AV_k y^k$ where $V_k$ is the matrix of the
Lanczos vectors and $y^k$ is the solution of $J_k y^k = \|r^0\| e^1$

$$\|\epsilon^k\|_A^2 = (A\epsilon^k, \epsilon^k) = (A^{-1} r^0, r^0) - 2(r^0, V_k y^k) + (AV_k y^k, V_k y^k)$$

But $A^{-1} V_n = V_n J_n^{-1}$

$$r^0 = \|r^0\| v^1 = \|r^0\| V_n e^1$$

Therefore

$$A^{-1} r^0 = \|r^0\| A^{-1} V_n e^1 = \|r^0\| V_n J_n^{-1} e^1$$

and

$$(A^{-1} r^0, r^0) = \|r^0\|^2 (V_n J_n^{-1} e^1, V_n e^1) = \|r^0\|^2 (J_n^{-1} e^1, e^1)$$
Since $r^0 = \|r^0\| v^1 = \|r^0\| V_k e^1$

\[(r^0, V_k y^k) = \|r^0\| e^1, J_k^{-1} e^1)\]

Finally

\[(A V_k y^k, V_k y^k) = (V_k^T A V_k y^k, y^k) = (J_k y^k, y^k) = \|r^0\| e^1, J_k^{-1} e^1)\]

The second relation is obtained by using the spectral decomposition of $J_n$ and $J_k$.

This formula is the link between CG and Gauss quadrature. It shows that the square of the A–norm of the error is the remainder of a Gauss quadrature rule for computing $(A^{-1} r^0, r^0)$.
Estimates of the $A$–norm of the error

At CG iteration $k$ we do not know $(J_n^{-1})_{1,1}$

Let $d$ be a given delay integer, an approximation of the $A$–norm of the error at iteration $k - d$ is obtained by

$$\|\epsilon^{k-d}\|_A^2 \approx \|r^0\|^2((J_k^{-1})_{(1,1)} - (J_{k-d}^{-1})_{(1,1)})$$

This can also be understood as writing

$$\|\epsilon^{k-d}\|_A^2 - \|\epsilon^k\|_A^2 = \|r^0\|^2((J_k^{-1})_{(1,1)} - (J_{k-d}^{-1})_{(1,1)})$$

and supposing that $\|\epsilon^k\|_A$ is negligible against $\|\epsilon^{k-d}\|_A$

Another interpretation is to consider that having a Gauss rule with $k - d$ nodes at iteration $k - d$, we use another more precise Gauss quadrature with $k$ nodes to estimate the error of the quadrature rule.
We have to be careful in computing \((J_k^{-1})_{(1,1)} - (J_{k-d}^{-1})_{(1,1)}\)

Let \(j_k = J_k^{-1}e^k\) be the last column of the inverse of \(J_k\); Using the Sherman–Morrison formula

\[
(J_{k+1}^{-1})_{1,1} = (J_k^{-1})_{1,1} + \frac{\eta_{k+1}^2 (j_kj_k^T)_{1,1}}{\alpha_{k+1} - \eta_{k+1}^2 (j_k)_k}
\]

Cholesky factorization of \(J_k\) whose diagonal elements are \(\delta_1 = \alpha_1\) and

\[
\delta_i = \alpha_i - \frac{\eta_i^2}{\delta_{i-1}}, \quad i = 2, \ldots, k
\]

Then

\[
(j_{k})_{1} = (-1)^{k-1} \frac{\eta_2 \cdots \eta_k}{\delta_1 \cdots \delta_k}, \quad (j_{k})_{k} = \frac{1}{\delta_k}
\]
Let $b_k = (J_k^{-1})_{1,1}$

$$b_k = b_{k-1} + f_k, \quad f_k = \frac{\eta_k^2 c_{k-1}^2}{\delta_{k-1}(\alpha_k \delta_{k-1} - \eta_k^2)} = \frac{c_k^2}{\delta_k}$$

where

$$c_k = \frac{\eta_2 \cdots \eta_{k-1}}{\delta_1 \cdots \delta_{k-2}} \frac{\eta_k}{\delta_{k-1}} = c_{k-1} \frac{\eta_k}{\delta_{k-1}}$$

Since $J_k$ is positive definite, $f_k > 0$

Moreover

$$c_k = \frac{\eta_2 \cdots \eta_k}{\delta_1 \cdots \delta_{k-1}} = \frac{\|r^{k-1}\|}{\|r^0\|}$$

and $\gamma_{k-1} = 1/\delta_k$ where $\gamma_{k-1}$ is the CG parameter

($=(r^{k-1}, r^{k-1})/(p^{k-1}, Ap^{k-1})$)
Therefore

$$\| \epsilon^{k-d} \|_A^2 \approx \sum_{j=k-d}^{k-1} \gamma_j \| r^j \|_2^2$$

This gives a lower bound of the error norm

Other bounds can be obtained with the Gauss–Radau and Gauss–Lobatto quadrature rules

**Algorithm CGQL**

Let $x^0$ be given, $r^0 = b - Ax^0$, $p^0 = r^0$, $\beta_0 = 0$, $\alpha_{-1} = 1$, $c_1 = 1$

For $k = 1, \ldots$ until convergence

$$\gamma_{k-1} = \frac{(r^{k-1}, r^{k-1})}{(p^{k-1}, Ap^{k-1})}$$

$$\alpha_k = \frac{1}{\gamma_{k-1}} + \frac{\beta_{k-1}}{\gamma_{k-2}}$$
CGQL (2)

if $k = 1$

$$f_1 = \frac{1}{\alpha_1}$$
$$\delta_1 = \alpha_1$$
$$\bar{\delta}_1 = \alpha_1 - \lambda_m$$
$$\tilde{\delta}_1 = \alpha_1 - \lambda_M$$

else

$$c_k = c_{k-1} \frac{\eta_k}{\delta_{k-1}} = \frac{\|r^{k-1}\|}{\|r^0\|}$$
$$\delta_k = \alpha_k - \frac{\eta_k^2}{\delta_{k-1}} = \frac{1}{\gamma_{k-1}}$$

$$f_k = \frac{\eta_k^2 c_{k-1}^2}{\delta_{k-1}(\alpha_k \delta_{k-1} - \eta_k^2)} = \gamma_{k-1} c_k^2$$
CGQL (3)

\[ \bar{\delta}_k = \alpha_k - \lambda_m - \frac{\eta_k^2}{\bar{\delta}_{k-1}} = \alpha_k - \bar{\alpha}_{k-1} \]

\[ \delta_k = \alpha_k - \lambda_M - \frac{\eta_k^2}{\delta_{k-1}} = \alpha_k - \alpha_{k-1} \]

end

\[ x^k = x^{k-1} + \gamma_{k-1}p^{k-1} \]

\[ r^k = r^{k-1} - \gamma_{k-1}Ap^{k-1} \]

\[ \beta_k = \frac{(r^k, r^k)}{(r^{k-1}, r^{k-1})} \]

\[ \eta_{k+1} = \frac{\sqrt{\beta_k}}{\gamma_{k-1}} \]

\[ p^k = r^k + \beta_k p^{k-1} \]
\[ \bar{\alpha}_k = \lambda_m + \frac{\eta_{k+1}^2}{\delta_k} \]

\[ \tilde{\alpha}_k = \lambda_M + \frac{\eta_{k+1}^2}{\delta_k} \]

\[ \check{\alpha}_k = \frac{\bar{\delta}_k \delta_k}{\delta_k - \bar{\delta}_k} \left( \frac{\lambda_M}{\bar{\delta}_k} - \frac{\lambda_m}{\delta_k} \right) \]

\[ \check{\eta}_{k+1}^2 = \frac{\bar{\delta}_k \delta_k}{\delta_k - \bar{\delta}_k} (\lambda_M - \lambda_m) \]

\[ \check{f}_k = \frac{\eta_{k+1}^2 c_k^2}{\delta_k (\bar{\alpha}_k \delta_k - \eta_{k+1}^2)} \]

\[ f_k = \frac{\eta_{k+1}^2 c_k^2}{\delta_k (\alpha_k \delta_k - \eta_{k+1}^2)} \]

\[ \check{f}_k = \frac{\check{\eta}_{k+1}^2 c_k^2}{\delta_k (\check{\alpha}_k \delta_k - \check{\eta}_{k+1}^2)} \]
CGQL (5)

if $k > d$

$$g_k = \sum_{j=k-d+1}^{k} f_j$$

$$s_{k-d} = \| r^0 \|^2 g_k$$

$$\bar{s}_{k-d} = \| r^0 \|^2 (g_k + \bar{f}_k)$$

$$\hat{s}_{k-d} = \| r^0 \|^2 (g_k + \hat{f}_k)$$

end
Proposition

Let $J_k$, $\tilde{J}_k$, $\bar{J}_k$ and $\hat{J}_k$ be the tridiagonal matrices of the Gauss, Gauss–Radau (with $b$ and $a$ as prescribed nodes) and the Gauss–Lobatto rules.

Then, if $0 < a = \lambda_m \leq \lambda_{\text{min}}(A)$ and $b = \lambda_M \geq \lambda_{\text{max}}(A)$,

- $\|r_0\|((J_k^{-1})_{1,1}), \|r_0\|((\tilde{J}_k^{-1})_{1,1})$ are lower bounds of $\|e_0\|^2_A = r_0A^{-1}r_0$,
- $\|r_0\|((\bar{J}_k^{-1})_{1,1})$ and $\|r_0\|((\hat{J}_k^{-1})_{1,1})$ are upper bounds of $r_0A^{-1}r_0$

Theorem

At iteration number $k$ of CGQL, $s_{k-d}$ and $\bar{s}_{k-d}$ are lower bounds of $\|\epsilon^{k-d}\|^2_A$, $\tilde{s}_{k-d}$ and $\hat{s}_{k-d}$ are upper bounds of $\|\epsilon^{k-d}\|^2_A$.
Preconditioned CG

For the preconditioned CG algorithm, the formula to consider is

\[ \| \epsilon^k \|_A^2 = (z^0, r^0)((J_{n-1})_{1,1} - (J_{k-1})_{1,1}) \]

where \( Mz^0 = r^0 \), \( M \) being the preconditioner, a symmetric positive definite matrix that is chosen to speed up the convergence.

The Gauss rule estimate is

\[ \| \epsilon^{k-d} \|_A^2 \approx \sum_{j=k-d}^{k-1} \gamma_j(z^j, r^j) \]
Estimates of the $l_2$ norm of the error

Theorem

$$
\| \epsilon^k \|^2 = \| r^0 \|^2 [(e^1, J_n^{-2} e^1) - (e^1, J_k^{-2} e^1)] \\
+ (-1)^k 2 \eta_{k+1} \frac{\| r^0 \|}{\| r_k \|} (e^k, J_k^{-2} e^1) \| \epsilon^k \|^2_A
$$

Corollary

$$
\| \epsilon^k \|^2 = \| r^0 \|^2 [(e^1, J_n^{-2} e^1) - (e^1, J_k^{-2} e^1)] - 2 \frac{(e^k, J_k^{-2} e^1)}{(e^k, J_k^{-1} e^1)} \| \epsilon^k \|^2_A
$$

This can be computed introducing a delay and using a QR factorization of $J_k$
Relation with finite element problems

Suppose we want to solve a PDE

\[ L u = f \quad \text{in } \Omega \]

\( \Omega \) being a two or three–dimensional bounded domain, with appropriate boundary conditions on \( \Gamma \) the boundary of \( \Omega \).

As a simple example, consider the PDE

\[ -\Delta u = f, \quad u|_{\Gamma} = 0 \]

This problem is naturally formulated in the Hilbert space \( H^1_0(\Omega) \)

\[ a(u, v) = (f, v), \quad \forall v \in V = H^1_0(\Omega) \]

where \( a(u, v) \) is a self–adjoint bilinear form
\[ a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, dx \]

and

\[ (f, v) = \int_{\Omega} fv \, dx \]

There is a unique solution \( u \in H^1_0(\Omega) \)
The approximate solution is sought in a finite dimensional subspace \( V_h \subset V \) as

\[ a(u_h, v_h) = (f, v_h), \ \forall v_h \in V_h \]

The simplest method triangulates the domain \( \Omega \) (with triangles or tetrahedrons of maximal diameter \( h \)) and uses functions which are linear on each element
Using basis functions $\phi_i$ which are piecewise linear and have a value 1 at vertex $i$ and 0 at all the other vertices

$$v_h(x) = \sum_{j=1}^{n} v_j \phi_j(x)$$

The approximated problem is equivalent to a linear system $Au = c$, where

$$[A]_{i,j} = a(\phi_i, \phi_j), \quad c_i = (f, \phi_i)$$

The matrix $A$ is symmetric and positive definite. The solution of the finite dimensional problem is

$$u_h(x) = \sum_{j=1}^{n} u_j \phi_j(x)$$

We use CG to solve the linear system
We have two sources of errors, the difference between the exact and approximate solution $u - u_h$ and $u_h - u_h^{(k)}$, the difference between the approximate solution and its CG computed value (not speaking of rounding errors)

Of course, we desire the components of $u - u_h^{(k)}$ to be small. This depends on $h$ and on the CG stopping criterion

The problem of finding an appropriate stopping criterion has been studied by Arioli and al

Let $\|v\|_a^2 = a(v, v)$ and $u^*_h \in V_h$ be such that

$$\|u_h - u^*_h\|_a^2 \leq h^2 t \|u_h\|_a^2$$

Then

$$\|u - u^*_h\|_a \leq \|u - u_h\|_a + \|u_h - u^*_h\|$$

$$\leq h^t \|u\|_a + (1 + h^t) \|u - u_h\|_a$$
If $t > 0$ and $h < 1$

$$\|u - u_h^*\|_a \leq h^t \|u\|_a + 2\|u - u_h\|_a$$

Therefore, if $u_h^* = u_h^{(k)}$ and we choose $\|u_h - u_h^*\|_a$ such that $h^t \|u\|_a$ is of the same order as $\|u - u_h\|_a$ we have

$$\|u - u_h^*\|_a \approx \|u - u_h\|_a$$

We have

$$\|v_h^{(k)}\|_a = \|v^k\|_A$$

Let $\zeta_k$ be an estimate of $\|\varepsilon^k\|_A^2$, Arioli’s stopping test is

If $\zeta_k \leq \eta^2 ( (u^k)^T r^0 + c^T u^0 )$ then stop

The parameter $\eta$ is chosen as $h$ or $\eta^2$ as the maximum area of the triangles in 2D
Numerical experiments

F3, \( d = 1 \), \( \log_{10} \) of the \( A \)-norm of the error (plain), Gauss (dashed), Gauss–Radau(\( \lambda_{\text{min}} \)) (dot–dashed)
F3, $d = 5$, zoom of $\log_{10}$ of the $A$–norm of the error (plain), Gauss (dashed), Gauss–Radau (dot–dashed)
For the Gauss–Radau upper bound we use a value of $a = 0.02$ whence the smallest eigenvalue is $\lambda_{min} = 0.025$.

F4, $n = 900, \ d = 1, \ \log_{10}$ of the $A$–norm of the error (plain), Gauss (dashed), Gauss–Radau (dot–dashed)
Adaptive algorithm for the smallest eigenvalue

F4, $n = 900$, $d = 1$, est. of $\lambda_{min}$, $\log_{10}$ of the $A$–norm of the error (plain), Gauss (dashed), Gauss–Radau (dot–dashed)
Another example

\[-\text{div}(\lambda(x, y)\nabla u) = f, \quad u|_{\Gamma} = 0\]

Finite differences in the unit square

\[\lambda(x, y) = \frac{1}{(2 + p \sin \frac{x}{\eta})(2 + p \sin \frac{y}{\eta})}\]

We use $p = 1.8$ and $\eta = 0.1$

We compute $f$ such that the solution is $u(x, y) = \sin(\pi x) \sin(\pi y)$
CG2, $d = 1$, $n = 10000$, $\log_{10}$ of the $A$–norm of the error (plain), Gauss (dashed), Gauss–Radau (dot–dashed), $a = 10^{-4}$, $\lambda_{\text{min}} = 2.3216 \times 10^{-4}$
CG2, $d = 1$, $n = 10000$, IC(0), $\log_{10}$ of the $A$–norm of the error (plain), Gauss (dashed)
Since we are using finite difference and we have multiplied the right hand side by $h^2$, we modify the Arioli’s criteria to

If $\zeta_k \leq 0.1 \times (1/n)^2((x^k)^T r^0 + c^T x^0)$ then stop

where $\zeta_k$ is an estimate of $\|\epsilon^k\|_A^2$

When using $n = 10000$, the $A$–norm of the difference between the “exact” solution of the linear system (obtained by Gaussian elimination) and the discretization of $u$ is $n_u = 5.6033 \times 10^{-5}$

With the stopping criterion, we do 226 iterations and we have $n_x = 9.5473 \times 10^{-5}$

Using an incomplete Cholesky preconditioner IC(0) we do 47 iterations and obtain $n_x = 5.6033 \times 10^{-5}$
Anti–Gauss estimates

- Anti–Gauss quadrature rules can also be used to obtain estimates of the $A$–norm of the error
- Using the anti–Gauss rule is interesting since it does not need any estimate of the smallest eigenvalue of $A$
- Anti–Gauss estimates may fail since sometimes we have to take square roots of negative values

We use the generalized anti–Gauss rule with a parameter $\gamma$:

- We start from a value $\gamma_0$
- When at some iteration we find a value $\delta_k < 0$ in the Cholesky factorization of the Jacobi matrix we decrease the value of $\gamma$ until we find a positive definite matrix
- At most we will find $\gamma = 0$ and recover the Gauss rule
F4, $d = 1$, $n = 900$, $\log_{10}$ of the $A$–norm of the error (plain), anti–Gauss
$\gamma_0 = 1$ (dashed), $\gamma_0 = 0.7$ (dot–dashed)
Bound of the $l_2$ norm of the error

F4, $d = 1$, $n = 900$, $\log_{10}$ of the $l_2$ norm of the error (plain), Gauss (dashed)


B. Fischer and G.H. Golub, *On the error computation for polynomial based iteration methods*, in Recent advances in


G.H. Golub and G. Meurant, Matrices, moments and quadrature II or how to compute the norm of the error in iterative methods, BIT, v 37 n 3, (1997), pp 687–705


