Numer. Math. 61, 1–6 (1992)

Numerische Mathematik

A posteriori error estimates for linear equations

Giles Auchmuty *

Department of Mathematics, University of Houston, Houston, TX 77204-3476, USA

Received September 1, 1990 / Revised version received August 5, 1991

Summary. This paper describes upper and lower p-norm error bounds for approximate solutions of the linear system of equations $Ax=b$. These bounds imply that the error is proportional to the quantity $\|r\|_2^2 \|A^T r\|_q^{-1}$ where r is the residual and q is the conjugate index to p. The constant of proportionality is larger than 1 and lies in a specified range. Similar results are obtained for approximations to A^{-1} and solutions of nonsingular linear equations on general spaces.

Mathematics Subject Classification (1991): 15A60, 65F35, 65G99

1 Introduction

In this paper, upper and lower bounds on the p-norm error of approximate solutions of a nonsingular system of linear equations

$$Ax=b \tag{1.1}$$

will be described. In particular, it will be shown that when $\hat{x}$ is the solution of (1.1), $r=A\hat{x}-b$ is the residual, then

$$\|x-\hat{x}\|_p = c\,\frac{\|r\|_2^2}{\|A^T r\|_q} \tag{1.2}$$

where A^T is the transpose of A and c is a constant lying in an interval $[1, C_p(A)]$. The number $C_p(A)$ will be characterized and some simple bounds for it will be described. Throughout this paper $1 \leqq p \leqq \infty$ and $q=p(p-1)^{-1}$ is its conjugate.

Essentially, the p-norm error of an approximate solution x of (1.1) is proportional to, and bounded below by, the quantity

$$a_p(r)=\frac{\|r\|_2^2}{\|A^T r\|_q}. \tag{1.3}$$

* Research was partially supported by NSF Grant DMS8901477

This expression is easily computed and provides an indication of the order of magnitude of the error of a particular approximation. As a consequence, it provides a natural stopping criterion in iterative methods for solving (1.1).

The results obtained using these error bounds provide different information to that obtained using the well-known estimates based on the condition number of A and the residual. Usually this condition number estimate is written

$$\|x-\hat{x}\|_p \leqq K_p(A)\frac{\|\hat{x}\|_p\|r\|_p}{\|b\|_p} \tag{1.4}$$

where $K_p(A)=\|A\|_p\|A^{-1}\|_p$ is the p-norm condition number of A.

The error estimate in Theorem 1 below provides both upper and lower bounds on the error, not just the upper bound of (1.4). Moreover they depend on $\|A^{\mathrm{T}}r\|_q$ as well as $\|r\|_2$ so, unlike (1.4), different error estimates are found for approximate solutions whose residual norms are the same.

2 p-norm error estimates

First consider the case where A is a nonsingular real $n\times n$ matrix and b is a real n-vector. The residual $r(x)$ associated with a vector x is

$$r=r(x):=Ax-b. \tag{2.1}$$

Let $\|\ \|_p$ and $\langle\ ,\ \rangle$ denote the usual p-norm and Euclidean inner product on $\mathbb{R}^n$. When the subscript p is omitted, $\|x\|$ will denote the 2-norm. Terms not defined here should be taken as in [1].

When A is a nonsingular $n\times n$ matrix, define

$$C_p(A):=\sup_{\|y\|_2=1}\|A^{\mathrm{T}}y\|_q\|A^{-1}y\|_p. \tag{2.2}$$

This number is well-defined as it is the maximum value of a continuous function on a compact set.

Theorem 1. *Let A be a nonsingular, $n\times n$ real matrix, $x\neq\hat{x}$ and $1\leqq p\leqq\infty$. Then*

$$a_p(r)\leqq\|x-\hat{x}\|_p\leqq C_p(A)\,a_p(r) \tag{2.3}$$

where a_p, r and $C_p(A)$ are defined by (1.3), (2.1) *and* (2.2) *respectively.*

Proof. From the definition of r, and Hölders inequality,

$$\|r\|_2^2=\langle A^{\mathrm{T}}r,\,x-\hat{x}\rangle\leqq\|A^{\mathrm{T}}r\|_q\|x-\hat{x}\|_p.$$

Upon dividing by $\|A^{\mathrm{T}}r\|_q$, the lower bound in (2.3) follows.

The definition (2.2) implies that

$$C_p(A)=\sup_{y\neq 0}\frac{\|A^{\mathrm{T}}y\|_q\|A^{-1}y\|_p}{\|y\|_2^2}$$

because of the homogeniety of this expression. Substituting r for y, leads to

$$\|A^T r\|_q \|A^{-1} r\|_p \leqq C_p(A) \|r\|_2^2 \tag{2.4}$$

and this is the second inequality in (2.3). □

Note that the upper bound in (2.3) is optimal, in that there is an x in $\mathbb{R}^n$ for which it is attained, because there is an r which yields the maximum value in the definition of $C_p(A)$.

When $p=2$, a weaker form of the upper bound in (2.3) can be derived using Kantorovich' inequality, which says that if B is a positive definite, symmetric, matrix, then

$$\langle By, y\rangle \langle B^{-1} y, y\rangle \leqq K_1 \|y\|^2 \tag{2.5}$$

with $K_1=(\lambda_1+\lambda_n)^2(4\lambda_1\lambda_n)^{-1}$. Here $\lambda_1(\lambda_n)$ is the largest, (smallest) eigenvalue of B. See Luenberger [3], Sect. 7.6, or Householder [2], Sect. 3.4, for proofs of (2.5). Let $B=AA^{\mathrm{T}}$, then $\lambda_1=\sigma_1^2$, $\lambda_n=\sigma_n^2$ where $\sigma_1, (\sigma_n)$ is the largest, (smallest) singular value of A. Then

$$\|A^{\mathrm{T}} y\|_2 \|A^{-1} y\|_2 \leqq K \|y\|_2^2 \tag{2.6}$$

with

$$K=\frac{\sigma_1^2+\sigma_n^2}{2\sigma_1\sigma_n}=\frac{1}{2}\left(K_2(A)+\frac{1}{K_2(A)}\right) \tag{2.7}$$

and where $K_2(A)=\dfrac{\sigma_1}{\sigma_n}$ is the 2-norm condition number of A.

From the definition of $C_2(A)$, one must have

$$C_2(A) \leqq K \leqq K_2(A) \tag{2.8}$$

when K is given by (2.6) as $C_2(A)$ is the smallest number for which (2.6) holds.

The absolute error bounds in (2.3) lead to the relative error bounds

$$\frac{\|r\|_2^2}{\|A^{\mathrm{T}} r\|_q \|x\|_p + C\|r\|^2} \leqq \frac{\|x-\hat{x}\|_p}{\|\hat{x}\|_p} \leqq \frac{C\|r\|^2}{\|A^{\mathrm{T}} r\|_q \|x\|_p - C\|r\|^2} \tag{2.9}$$

with $C=C_p(A)$ and provided the last denominator is positive. These are a direct consequence of the triangle inequality.

The lower bound in (2.3) may often be improved. From the Hahn-Banach theorem

$$\|x-\hat{x}\|_p = \|A^{-1} r\|_p = \sup_{z \neq 0} \frac{\langle A^{-1} r, z\rangle}{\|z\|_q}.$$

Let $z=A^{\mathrm{T}} y$ then, since A is nonsingular,

$$\|x-\hat{x}\|_p = \sup_{y \neq 0} \frac{\langle r, y\rangle}{\|A^{\mathrm{T}} y\|_q}. \tag{2.10}$$

The lower bound in (2.3) corresponds to the choice $y=r$ in this expression. For a particular equation, there may well be a better choice of y.

The lower bound in (2.3) may also be generalized to the case where the matrix is singular or non-square. Let B be an $m \times n$ matrix, f an m-vector and consider the least squares problem of solving

$$B^{\mathrm{T}}(Bx-f)=0. \tag{2.12}$$

Let $\tilde{x}$ be a solution of (2.12) and $r:=Bx-f$ be the residual associated with an approximate solution. If $B\tilde{x}=\tilde{f}$ then

$$r=B(x-\tilde{x})+d$$

where $d=\tilde{f}-f$ lies in the null space of B^{T}. Thus

$$\|r\|_2^2=\langle B(x-\tilde{x}), r\rangle+\|d\|_2^2$$

so rearranging, and using Hölder's inequality as before yields

$$\|x-\tilde{x}\|_p \geqq \frac{\|r\|_2^2-\|d\|_2^2}{\|B^{\mathrm{T}} r\|_q}. \tag{2.13}$$

Here $\|d\|_2$ is the Euclidean distance of f from the range of B.

3 Evaluation and properties of $C_p(A)$

The quantity $C_p(A)$ defined by (2.2) depends only on p and A and obeys $C_p(\alpha A) = C_p(A)$ for any nonzero scalar α.

Despite appearences, it is not necessary to find A^{-1} to compute $C_p(A)$. Substitute Aw for y in (2.2), then using homogeneity,

$$\begin{aligned} C_p(A) &= \sup_{\|Aw\|_2=1} \|A^{\mathrm{T}} Aw\|_q \|w\|_p \\ &= \sup_{w \neq 0} \frac{\|A^{\mathrm{T}} Aw\|_q \|w\|_p}{\|Aw\|_2^2} \\ &= \sup_{\|w\|_p=1} \frac{\|A^{\mathrm{T}} Aw\|_q}{\|Aw\|_2^2}. \end{aligned} \tag{3.1}$$

These formulae provide different characterizations of $C_p(A)$ as the value of a maximization problem of a continuous function subject to an equality constraint. When $1<p<\infty$, the functions involved are continuously differentiable. When $p=1$ or ∞, this is a nonlinear programming problem wih "non-smooth" functions.

There are some simpler estimates of $C_p(A)$. Given $1 \leqq p, s \leqq \infty$, let

$$\|B\|_{ps}=\sup_{x \neq 0} \frac{\|Bx\|_p}{\|x\|_s}.$$

From the definition (2.2),

$$C_p(A) \leqq \|A^{\mathrm{T}}\|_{q2} \|A^{-1}\|_{p2} = \|A\|_{2p} \|A^{-1}\|_{p2} \tag{3.2}$$

as $\|B^{\mathrm{T}}\|_{q2} = \|B\|_{2p}$ when p, q are conjugate.

Consider the problem of extremizing the quadratic form

$$g(y) = \langle A^{\mathrm{T}} A y, y \rangle = \|Ay\|^2$$

on the unit sphere $S_p = \{y \in \mathbb{R}^n : \|y\|_p = 1\}$.

Let

$$\alpha_p := \inf_{S_p} g(y), \qquad \gamma_p = \inf_{S_p} g(y) \tag{3.4}$$

then

$$\alpha_p \|y\|_p^2 \leqq g(y) \leqq \gamma_p \|y\|_p^2$$

for all y. This implies that

$$\|A\|_{2p} = \sqrt{\gamma_p} \quad \text{and} \quad \|A^{-1}\|_{p2} = \alpha_p^{-\frac{1}{2}}$$

so (3.2) yields

$$C_p(A) \leqq \tilde{C}_p(A) := \left(\frac{\gamma_p}{\alpha_p}\right)^{\frac{1}{2}}. \tag{3.5}$$

In general it is easier to evaluate $\tilde{C}_p(A)$ than $C_p(A)$ since it is easy to extremize this quadratic form g. When $p=2$, then $\alpha_2 = \sigma_n^2$, $\gamma_2 = \sigma_1^2$ so (3.5) leads to (2.8) again.

4 Error estimates for general linear equations

The preceding analysis may be generalized to obtain upper and lower error estimates for approximate solutions of general, nonsingular, linear problems.

Let X be a normed vector space over a field F, Y be an inner product space over F and $A: X \to Y$ be a continuous linear operator with a bounded inverse. None of X, Y, F need be complete.

Let X^* be the dual space of X with the dual norm

$$\|h\|_* = \sup_{x \neq 0} \frac{|h(x)|}{\|x\|}$$

where $\|\ \|$ denotes the norm on X, $\langle\ ,\ \rangle$ will be the inner product on Y and the adjoint operator $A^*: Y \to X^*$ is defined by

$$(A^* y)(x) = \langle Ax, y \rangle \tag{4.1}$$

for all x in X and y in Y. A^* is a bounded, continuous, linear operator whenever A is. Define

$$C(A) := \sup_{\|y\|_2 = 1} \|A^* y\|_* \|A^{-1} y\|. \tag{4.2}$$

This is finite as both A^{-1} and A^* are continuous linear operators.

Our interest is in solving equation (1.1) with b given in Y, and the error is expressed in terms of the residual

$$r = r(x) = Ax - b \tag{4.3}$$

and the functional

$$a(r) = \frac{\|r\|^2}{\|A^* r\|_*}. \tag{4.4}$$

Using the same arguments as in Theorem 1, we prove

Theorem 2. *Assume X, Y, A as above. If $\hat{x}$ is the unique solution of* (1.1), *x is any vector in X, then*

$$a(r) \leqq \|x - \hat{x}\| \leqq C(A)\, a(r) \tag{4.5}$$

where $C(A)$, r and $a(r)$ are defined by (4.2)–(4.4).

This theorem may be used to obtain error estimates for linear equations over general fields including the rational or complex numbers. It may be used to obtain error bounds on matrix inverses by taking $X = Y = M_n(F)$, with the usual inner product and $b = I_n$ to be the identity. The dual norms on X here could be the 1 and ∞ norms as well as the inner product norm.

These error bounds also apply to nonsingular linear operator equations where Y is a Hilbert space.

Acknowledgements. I would like to thank Roland Glowinski and V.G. Hart for useful comments and discussions on these topics. I would also like to thank the Department of Mathematics of the University of Queensland for their support and hospitality during the period when this work was initiated.

References

1. Golub, G.H., van Loan, C.F. (1983): Matrix Computations. Johns Hopkins University Press, Baltimore
2. Householder, A.S. (1975): The Theory of Matrices in Numerical Analysis. Dover Publications, New York
3. Luenberger, D.G. (1984): Linear and Nonlinear Programming, 2nd ed. Addison-Wesley, Reading