RUMP, S.M.

Verified Computation of the Solution of Large Sparse Systems

In this note we summarize some recent results on the computation of verified error bounds for large and sparse systems of equations. The detailed theory and proofs will be published elsewhere.

We focus on the symmetric case. The aim of the methods is to utilize numerical approximate solvers like LDL^T as much as possible. The error estimates are derived in such a way that they hold for any library routine independent on the specific implementation. Numerical examples are given. Comparisons with the LAPACK condition estimator DPBCON are shown and results for Emden's equation.

1. Symmetric linear systems

We use notations and basic facts of interval analysis (cf., for example, [1], [6]). Let a linear system $Ax = b$ be given with

$A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, $A = A^T$ large and sparse or banded,
and let p denote the maximum number of nonzero elements per row or column divided by 2.

We assume n to be large compared to p. For an approximate solution $\tilde{x} \in \mathbb{R}^n$ and an exact solution $\hat{x} \in \mathbb{R}^n$ the following error estimate holds:

$$\|\hat{x} - \tilde{x}\|_\infty \leq \|\hat{x} - \tilde{x}\|_2 \leq \sigma_n(A)^{-1} \cdot \|b - A\tilde{x}\|_2. \tag{1.1}$$

Here, $\sigma_1(A) \geq \ldots \geq \sigma_n(A)$ denote the singular values of A. For A being symmetric positive definite (s.p.d.), in [10] a method for large and sparse or banded matrices has been described with computing time $n \cdot p^2$, which is the complexity of a Cholesky decomposition. The scope of applicability is, as for every linear solver,$\mathrm{cond}(A) \lesssim eps^{-1}$, where eps denotes the relative rounding error unit.

In the following we will describe a method for the large class of linear systems with general symmetric matrix. For $A = LDL^T$ the inertia of A and D are equal by Sylvester's law of inertia. If the inertias of $A - \tilde{\lambda}I$ and $A + \tilde{\lambda}I$ are equal, then $|\lambda(A)| > \tilde{\lambda}$ for every eigenvalue $\lambda(A)$ of A. That means $\sigma_n(A) > \tilde{\lambda}$. For approximate decompositions $A - \tilde{\lambda}I \approx \tilde{L}_1\tilde{D}_1\tilde{L}_1^T$ and $A + \tilde{\lambda}I \approx \tilde{L}_2\tilde{D}_2\tilde{L}_2^T$, the inertia of $A - \tilde{\lambda}I$ and $\tilde{D}_1$ or $A + \tilde{\lambda}I$ and $\tilde{D}_2$ need not be equal because of rounding errors. This problem can be solved by the following theorem.

Theorem 1.1 Let $A \in \mathbb{R}^{n\times n}$, $0 < \tilde{\lambda} \in \mathbb{R}$ and $\tilde{L}_1$, $\tilde{D}_1$, $\tilde{L}_2$, $\tilde{D}_2 \in \mathbb{R}^{n\times n}$ be given. If the inertia of $\tilde{D}_1$ and $\tilde{D}_2$ are equal, then for any matrix norm

$$\sigma_n(A) > \tilde{\lambda} - \max\{\|A - \tilde{\lambda}I - \tilde{L}_1\tilde{D}_1\tilde{L}_1^T\|,\ \|A + \tilde{\lambda}I - \tilde{L}_2\tilde{D}_2\tilde{L}_2^T\|\} \tag{1.2}$$

If all eigenvalues of $\tilde{D}_1$ are positive, then

$$\sigma_n(A) > \tilde{\lambda} - \|A - \tilde{\lambda}I - \tilde{L}_1\tilde{D}_1\tilde{L}_1^T\|. \tag{1.3}$$

Having an approximate decomposition $\tilde{L}\tilde{D}\tilde{L}^T$ of A, a suitable value for $\tilde{\lambda}$ is easily computed by inverse power iteration. After that, we enter a hybrid algorithm. If the approximate decomposition $\tilde{L}_1\tilde{D}_1\tilde{L}_1^T$ of $A - \tilde{\lambda}I$ yields a diagonal matrix $\tilde{D}_1$ with only positive diagonal elements, then we use (1.3) and stop. Otherwise, $A + \tilde{\lambda}I$ is decomposed, too, and (1.2) can be used. The lower bound for $\sigma_n(A)$ is used in (1.1) for the final error estimate. The scope of applicability of the method is again $\mathrm{cond}(A) \lesssim eps^{-1}$.

2. Error estimates

Our methods needs an error estimate on $\|B - \tilde{L}\tilde{D}\tilde{L}^T\|$ for an approximate decomposition $\tilde{L}\tilde{D}\tilde{L}^T$ of $B \in \mathbb{R}^{n\times n}$. For this purpose, a priori error estimates have been derived being independent on the specific implementation (for

example, order of execution). The heart is an error estimate for sums of floating point numbers. Define

$$\widetilde{a} = fl(a) \quad :\Leftrightarrow \quad \exists\, \varepsilon, \delta \in \mathbb{R}, |\varepsilon| \le eps,\ |\eta| \le eta \ : \quad \widetilde{a} = a \cdot (1+\varepsilon) + \eta \tag{2.1}$$

with machine constants *eps* and *eta* representing the relative rounding error and underflow error, respectively. We define recursively

$$\widetilde{a} = fl^{k+1}(a) \ :\Leftrightarrow \quad \widetilde{a} = fl\big(fl^k(a)\big) \quad \text{for } 1 \le k \in \mathbb{N}.$$

For a floating point sum $\widetilde{s} := c_1 \boxplus c_2 \boxplus \ldots \boxplus c_n$ of floating point numbers $c_1, \ldots, c_n$ one can show

$$\widetilde{s} = \sum_{i=1}^{n} fl^l(c_i) \quad \text{and} \quad c_k = fl^l(s) - \sum_{\substack{i=1 \\ i \ne k}}^{n} fl^l(c_i) \quad \text{for every } 1 \le k \le n.$$

Here, l depends on the order of summation. For usual summation *in any order* we have $l = n-1$, for recursive blockwise summation with blocksize b we have $l = 2(b-1)\lceil \log_b n \rceil$, and using a precise scalar product [5] yields $l = 1$. With this the following rigorous error estimate for an approximate floating point decomposition $\widetilde{L}\widetilde{D}\widetilde{L}^T$ can be proved:

$$|B - \widetilde{L}\widetilde{D}\widetilde{L}^T|_{ij} \le \frac{l+2}{1-eps} \cdot \{\mu \cdot eps + (p+1) \cdot eta\} \quad \text{with} \quad \mu := \max_i \sum_k \widetilde{L}_{ik}^2 |\widetilde{D}_{kk}|.$$

This estimates also holds in the presence of underflow and it even holds for a computer arithmetic without guard digit. Moreover, it can be shown to be very sharp for components (i,j) with positive $\widetilde{D}_{ii}$, $\widetilde{D}_{jj}$. But in many practical situations there are few negative eigenvalues of B and therefore few negative diagonal elements of $\widetilde{D}$. For those, the residual $|B - \widetilde{L}\widetilde{D}\widetilde{L}^T|$ can be calculated explicitly, whereas for the others the estimation above can be used. This yields very sharp estimates for $\|B - \widetilde{L}\widetilde{D}\widetilde{L}^T\|$.

We only mention that similar estimates have been derived for decompositions like Cholesky, LU, LDM^T and others and also for the product of triangular matrices. This is, for example, also useful in the approach described in [10].

We summarize that our estimates are valid for library routines using specific implementations like row-, or column-, or blockwise versions, and others, they hold for almost any computer arithmetic, they are valid in the presence of underflow and include higher order terms, and they are independent on the dimension, and only mildly dependent on the number of nonzero elements per row and column.

3. Nonlinear systems

Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be continuous, $\widetilde{x} \in D$, $\emptyset \ne X \subseteq \mathbb{R}^n$ compact and convex with $\widetilde{x} + X \subseteq D$. We are looking for some $\widehat{x} \in D$ with $f(\widehat{x}) = 0$. We locally linearize f at $\widetilde{x}$. Using slope expansions or, if f is differentiable, a Jacobian (cf. [6]), an interval matrix $S(\widetilde{x}, X) \in \mathbb{IR}^{n \times n}$ can be calculated satisfying

$$\forall\, x \in \widetilde{x} + X \quad f(x) \in f(\widetilde{x}) + S(\widetilde{x}, X) \cdot X. \tag{3.1}$$

We want to stress that the process of computing the expansion matrix $S(\widetilde{x}, X)$ can be fully automated for large classes of functions. Using the Krawczyk operator [4]

$$K(\widetilde{x}, X) := -R \cdot f(\widetilde{x}) + \{I - R \cdot S(\widetilde{x}, X)\} \cdot X$$

with some $R \in \mathbb{R}^{n \times n}$ for the defect equation $f(\widetilde{x} + x) = 0$, one can show [9]

$$K(\widetilde{x}, X) \subseteq X \text{ and } R \text{ regular} \quad \Rightarrow \quad \exists\, \widehat{x} \in \widetilde{x} + X : f(\widehat{x}) = 0. \tag{3.2}$$

The key problem is the choice of the preconditioner R. The optimal choice would be $R = \mathrm{mid}\big(S(\widetilde{x}, X)\big)^{-1}$, as has been shown by Neumaier [6] and Rex and Rohn, see [8]. However, if $S(\widetilde{x}, X)$ is large and sparse, then, in general, the midpoint inverse is full. This seems to forbid this choice of R. Nevertheless, we define

$$R := \mathrm{mid}\big(S(\widetilde{x}, X)\big)^{-1}$$

and show a posteriori the regularity of $\mathrm{mid}(S(\tilde{x}, X))$. For every $x \in X$

$$\begin{aligned}\{I - R \cdot S(\tilde{x}, X)\} \cdot x &= R \cdot \left\{R^{-1} - \left[\mathrm{mid}(S(\tilde{x}, X)) \pm \mathrm{rad}(S(\tilde{x}, X))\right]\right\} \cdot x \\ &= R \cdot \left[\pm \mathrm{rad}(S(\tilde{x}, X))\right] \cdot x.\end{aligned}$$

We use a lower estimate on the smallest singular value of $\mathrm{mid}(S(\tilde{x}, X))$, thereby showing regularity of $\mathrm{mid}(S(\tilde{x}, X))$ and the existence of R. Then, using 2-norms, the following theorem can be proved.

Theorem 3.1. Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be continuous, $\tilde{x} \in D$ and $0 < \rho \in \mathbb{R}$ with

$$X := \{ x \in \mathbb{R}^n \mid \|x\| \leq \rho \} \quad \text{and} \quad \tilde{x} + X \subseteq D.$$

Let $S(\tilde{x}, X) \in \mathbb{IR}^{n \times n}$ be given with

$$\forall\, x \in \tilde{x} + X \quad f(x) \in f(\tilde{x}) + S(\tilde{x}, X) \cdot X.$$

If $0 < \tau \in \mathbb{R}^n$ satisfies $\tau \leq \sigma_n\left(\mathrm{mid}(S(\tilde{x}, X))\right)$, and

$$\|f(\tilde{x})\| + \|\mathrm{rad}(S(\tilde{x}, X))\| \cdot \rho \leq \tau \cdot \rho,$$

then there exists $\hat{x} \in \tilde{x} + X$ with $f(\hat{x}) = 0$.

4. Examples

As a first example consider a random lower triangular matrix L of bandwidth 7 with uniformly distributed entries in [0,1] and define $A := LL^T$. Then A is s.p.d. of bandwidth 14. The right hand side b is computed such that the exact solution of the linear system satisfies $\hat{x}_i := (-1)^{i+1}/i$. In the following table we display the dimension n, the computed error bound and the condition number of A. Moreover, we display the ratio ρ of the computing time for the LAPACK [2] condition estimator DPBCON for banded s.p.d. matrices and the *total* computing time for our algorithm (including approximation and verification).

n	$\frac{\|\hat{x} - \tilde{x}\|_\infty}{\|\hat{x}\|_\infty}$	$\mathrm{cond}(A)$	ρ
1 000	$2.0 \cdot 10^{-9}$	$2.5 \cdot 10^{7}$	1
10 000	$2.4 \cdot 10^{-6}$	$3.5 \cdot 10^{10}$	40
100 000	$3.0 \cdot 10^{-5}$	$4.3 \cdot 10^{11}$	1 100

We see that the error diveded by the condition number is approximately equal to the relative rounding error unit *eps*. If the input data are perturbated in the last bit, these error bounds are best possible. With the lower estimate on $\sigma_n(A)$ we also obtain an upper estimate on the 2-condition number of A. The ratio ρ shows that in the examples this *verified* upper bound for the condition number is computed much faster than an approximation by DPBCON. This has also been observed by Korn [3].

The second example is Emden's equation

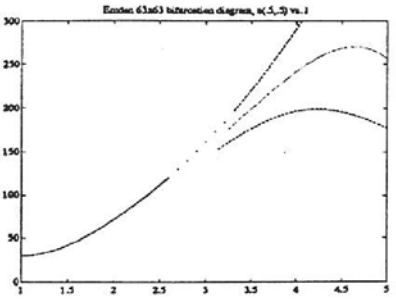
$$-\Delta U = U^2 \quad \text{with} \quad U = 0 \text{ on } \delta\Omega, \quad \Omega = (0, l^{-1}) \times (0, l).$$

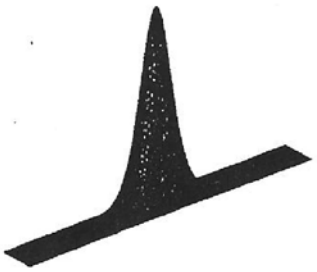
We use central difference quotients as a discretisation with m_1, m_2 inner grid points. We do *not* restrict Ω on $(0, \frac{1}{2} \cdot l^{-1}) \times (0, \frac{1}{2} l)$, because we want to test the method for ill-conditioned problem. We solve the discretized problem, and display the dimension N of the nonlinear system, the number *iter* of inverse power iterations in order to obtain $\tilde{\lambda}$, $\hat{u}_m = \hat{u}(\frac{1}{2} \cdot l^{-1}, \frac{1}{2} \cdot l)$, $\mathrm{cond}\left(\mathrm{mid}(S(\tilde{x}, X))\right)$, and the maximum error e satisfying $\|\hat{u} - \tilde{u}\| \leq e \cdot \|\tilde{u}\|$ for a computed approximate solution $\tilde{u}$.

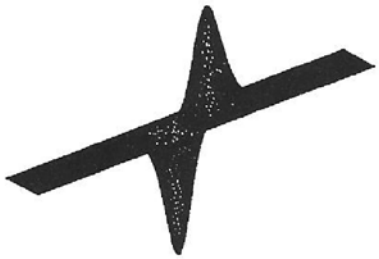
l	N	m_1	m_2	*iter*	$\hat{u}_m$	*cond*	e
1	32385	255	127	3	29.3	$1.3 \cdot 10^{4}$	$7.7 \cdot 10^{-13}$
2	32385	255	127	2	71.5	$8.0 \cdot 10^{6}$	$4.5 \cdot 10^{-10}$
2.5	32385	255	127	2	111.7	$7.9 \cdot 10^{9}$	$4.6 \cdot 10^{-7}$

Again, $e/\mathrm{cond} \approx eps$ demonstrating the quality of the inclusion. For larger values of l the condition increases above the critical value 10^{10}. It can be shown that due to the nonlinearities, an inclusion using a fixed point approach is *not* possible for condition numbers beyond 10^{10}.

Up to now the system matrix had 1 negative eigenvalue in all examples. For larger values of l other branches occur, with one or more negative eigenvalues. The first graph shows $\widehat{u}_m$ plotted vs. l. For $l \geq 3$ the upper two branches have 1, the lower has 2 negative eigenvalues.







The second plot shows a (verified) solution $\widehat{u}_1$ for $l = 3.5$. Optically, there is no difference to a second solution $\widehat{u}_2$, whereas the third plot shows the difference $\widehat{u}_1 - \widehat{u}_2$. Searching for other solution was inspired by Plum [7]. The author also thanks him for many fruitful discussions about Emden's equation.

Finally, we want to mention that for larger values of l Emden's equation frequently produces ill-conditioned numerical examples, where numerical difficulties are hard to detect. Consider $l = 3.1$, $m_1 = 127$, $m_2 = 31$, thus $N = 3937$. Starting with some $\widetilde{u}^0$ we perform Newton steps, and obtain the following results:

k	$\|f(\widetilde{u}^k)\|/\|\widetilde{u}^k\|$
1	$1.3 \cdot 10^{-3}$
2	$9.8 \cdot 10^{-5}$
3	$2.6 \cdot 10^{-7}$
4	$3.1 \cdot 10^{-12}$

This seems to indicate convergence, but the next iterate

k	$\|f(\widetilde{u}^k)\|/\|\widetilde{u}^k\|$
5	$1.2 \cdot 10^{-4}$

decovers the intrinsic difficulty of the problem.

5. References

1 ALEFELD, G.; HERZBERGER, J.: Introduction to Interval Computations. Academic Press, New York 1983.
2 ANDERSEN, E.; BAI, Z.; BISCHOF, C.; DEMMEL, J.; DONGARRA, J.; DU CROZ, J.; GREENBAU, A.; HAMMARLING, S.: LAPACK User's Guide. SIAM Publications, Philadelphia, 1992.
3 KORN, C.F.: Die Erweiterung von Software-Bibliotheken zur effizienten Verifikation der Approximationslösung linearer Gleichungssysteme. Dissertation, Universität Basel, 1993.
4 KRAWCZYK, R.: Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken. In: Computing 4 (1969), 187–201.
5 KULISCH, U.; MIRANKER, W.L.: Computer Arithmetic in Theory and Practice. Academic Press, New York 1981.
6 NEUMAIER, A.: Interval Methods for Systems of Equations. In: Encyclopedia of Mathematics and its Applications, Cambridge University Press, 1990.
7 PLUM, M.: private communication.
8 REX, G.: Zum Regularitätsnachweis von Matrizen. Talk at the GAMM annual conference 1994 at Braunschweig.
9 RUMP, S.M.: Solving Algebraic Problems with High Accuracy. Habilitationsschrift. In: KULISCH, U.W.; MIRANKER, W.L. (eds.): A New Approach to Scientific Computation, Academic Press, New York 1983.
10 RUMP, S.M.: Validated Solution of Large Linear Systems. In: ALBRECHT, R.; ALEFELD, G.; STETTER, H.J. (eds.): Computing Supplementum 9, Validation Numerics, Springer 1993, 191–212.

Anschrift: RUMP, S.M., Technische Universität Hamburg-Harburg, Technische Informatik III, Eißendorfer Straße 38, D-21071 Hamburg