Math 221 Final Project: A Tighter Forward Error Estimate

Vasily Volkov and Sage Briscoe

December 7, 2007

The Problem

Our goal is to substitute LAPACK's estimate of forward error $|A^{-1}| \cdot |r| + O(\epsilon)$ with a hopefully tighter estimate $|A^{-1}r| + O(\epsilon)$.

We present a numerically robust expression for the $O(\epsilon)$ term and experimentally validate the new method.

Given A, b and $\widehat{x}$, a proposed solution to $Ax = b$, we estimate the forward error $\Delta x = \widehat{x} - x$:

$$|\Delta x| = |\widehat{x} - x| = |A^{-1}r| \leqslant |A^{-1}| \cdot |r|,$$

where $r = A\widehat{x} - b$. We can compute $\||A^{-1}| \cdot |r|\|_\infty$ by Hager's estimator. If we take into account that the floating point residual $\widehat{r} = \mathrm{fl}(A\widehat{x} - b)$ differs from r by as much as $\Delta r = r - \widehat{r}$, $|\Delta r| \leqslant \epsilon(n+1)(|A| \cdot |\widehat{x}| + |b|)$, where ϵ is the machine precision, then

$$\begin{aligned}
|A^{-1}r| &\leqslant |A^{-1}| \cdot |r| \\
&= |A^{-1}| \cdot |\widehat{r} + \Delta r| \\
&\leqslant |A^{-1}| \cdot (|\widehat{r}| + |\Delta r|) \\
&\leqslant |A^{-1}| \cdot (|\widehat{r}| + \epsilon(n+1)(|A| \cdot |\widehat{x}| + |b|)).
\end{aligned}$$

Our Proposed Error Estimate

Given A, b and $\widehat{x}$ as before, as well as L and U corresponding to an LU decomposition of A, we will attempt to create a tighter bound on the forward error. If $r = \widehat{r} + \Delta r$, then

$$|A^{-1}r| = |A^{-1}(\widehat{r} + \Delta r)| \leqslant |A^{-1}\widehat{r}| + |A^{-1}| \cdot |\Delta r|.$$

The term $|A^{-1}| \cdot |\Delta r|$ does not differ from as in the LAPACK's bound, so we need to bound $|A^{-1}\widehat{r}|$.
Let ΔA be the backward error of LU-factorization, i.e. $LU = A + \Delta A$, and let $\widehat{f} = \mathrm{fl}(A^{-1}\widehat{r}) = (A + \widehat{\Delta A})^{-1}\widehat{r}$ where $A + \widehat{\Delta A} = (L + \Delta L)(U + \Delta U)$. Then $\widehat{f} = (U + \Delta U)^{-1}(L + \Delta L)^{-1}\widehat{r}$ is readily found using triangular solves.

Then, since $\widehat{r} = (L + \Delta L)(U + \Delta U)\widehat{f}$, we see

$$\begin{aligned} A^{-1}\widehat{r} = &= A^{-1}(L + \Delta L)(U + \Delta U)\widehat{f} \\ &= A^{-1}(LU + \Delta LU + L\Delta U + \Delta L\Delta U)\widehat{f} \\ &= A^{-1}(A + \Delta A + \Delta LU + L\Delta U + \Delta L\Delta U)\widehat{f} \\ &= \widehat{f} + A^{-1}(\Delta A + \Delta LU + L\Delta U + \Delta L\Delta U)\widehat{f} \end{aligned}$$

and

$$\begin{aligned} |A^{-1}\widehat{r}| &\leqslant |\widehat{f}| + |A^{-1}| \cdot (|\Delta A| + |\Delta LU| + |L\Delta U| + |\Delta L\Delta U|)|\widehat{f}| \\ &\leqslant |\widehat{f}| + |A^{-1}| \cdot (|LU - A| + 3\epsilon(n+1)|L| \cdot |U|)|\widehat{f}|. \end{aligned}$$

Putting it together we get

$$
\begin{aligned}
|A^{-1}r| &\leqslant |A^{-1}\widehat{r}| + |A^{-1}| \cdot |\Delta r| \\
&\leqslant |\widehat{f}| + |A^{-1}| \cdot (|LU - A| + 3\epsilon(n+1)|L| \cdot |U|)|\widehat{f}| + \\
&\qquad\qquad\qquad\qquad |A^{-1}| \cdot \epsilon(n+1)(|A| \cdot |\widehat{x}| + |b|) \\
&= |\widehat{f}| + |A^{-1}| \cdot \\
&\qquad (|LU - A| \cdot |\widehat{f}| + \epsilon(n+1)(3|L| \cdot |U| \cdot |\widehat{f}| + |A| \cdot |\widehat{x}| + |b|)
\end{aligned}
$$

and then

$$||A^{-1}r||_\infty \leqslant ||\widehat{f}||_\infty + |||A^{-1}| \cdot |\xi|||_\infty$$

for $\xi = |LU - A| \cdot |\widehat{f}| + \epsilon(n+1)(3|L| \cdot |U| \cdot |\widehat{f}| + |A| \cdot |\widehat{x}| + |b|)$.

Where we do Better

Our estimate for the forward error differs from the LAPACK error bound only in the way we bound $|A^{-1}\widehat{r}|$, so to find an example where our bound performs better than the LAPACK bound we need to find an A and r such that $||A^{-1}r|| << |||A^{-1}| \cdot |r|||$.

Thus we seek examples where A is ill-conditioned. If r is in the direction of the largest singular value of A, σ_1, and $|r|$ is in the direction of a smaller singular value then our bound will be tighter.

Experimental Results

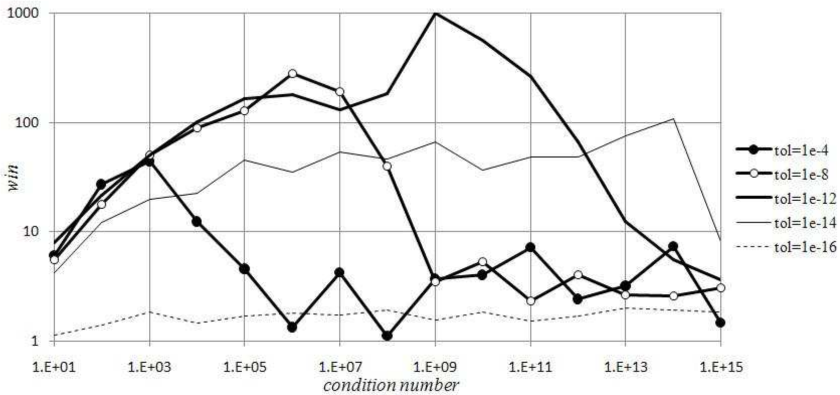
$n = 10$
matrix A has singular values $1, \alpha, \alpha^2, \ldots$ for $\alpha < 1$
$\kappa(A) = 1/\alpha^{n-1}$
measure how much our bound is tighter than LAPACK's bound
take maximum over 100 tests.
compute $|A^{-1}| \cdot |w|$ is using explicit matrix inversion, to avoid underestimation in Hager's estimator.

Experimental Results

Solve $Ax = b$ using a perturbed LU-factorization, b is random

tol: size of perturbation

win: LAPACK's bound divided by our bound



Experimental Results

SVD: $A = U\Sigma V^T$

"largest": $b = v_1$

"top": $b = c^T(v_1, ..., v_{n/2})$, c is a random vector