Error Bounds and Condition # Estimation.

We have seen that whether LU decomposition yields a backward stable algorithm for solving $Ax=b$ depends on whether the pivot growth factor

$$g_{pp} = \frac{\max |U_{ij}|}{\max |A_{ij}|}$$

is small or grows slowly as a function of n.

Proposition. For GEPP, $g_{gepp} \leq 2^{n-1}$.

Proof. When we update $\tilde{A}_{22}$, we use

$$a_{jk} = a_{jk} - l_{ji} u_{ik}$$

where $|l_{ji}| \leq 1$ but there is no bound on $|u_{ik}|$ except $|u_{ik}| \leq \max |a_{ij}|$. Thus a_{jk} could double on this update. $\square$

It turns out that for <u>complete</u> pivoting, one can show

$$g_{gecp} \lesssim \sqrt{n \cdot 2 \cdot 3^{1/2} \cdot 4^{1/3} \cdots n^{1/n-1}} \approx n^{1/2 + \ln n/4}$$

(see Demmel, p 50).

These give us the error bounds

$$\|\delta A\|_\infty \lesssim 3n\epsilon \|L\|_\infty \|U\|_\infty$$

$$\lesssim 3n\epsilon \; n \; n \, g_{gepp} \|A_\infty\|$$

$$\lesssim 3n^3 \epsilon 2^{n-1} \|A_\infty\|$$

since the L^∞ norm of a matrix A is the largest sum of (abs values of) entries in a row of A, and

$$\|\delta A\|_\infty \lesssim 3n^{3\frac{1}{2} + \ln n/4} \epsilon \|A\|_\infty$$

for GECP by the same argument.

Of course, these error bounds are much too large. A better bound comes from the residual estimate. Recall that if $\hat{x}$ is an approximate solution to $Ax = b$, then if

$$r = A\hat{x} - b$$

we can estimate

$$\delta x = \hat{x} - x = A^{-1}(r+b) - A^{-1}b$$
$$= A^{-1}r$$

by

$$\|\delta x\| \leq \|A^{-1}\| \, \|r\|.$$

Now it is easy and cheap to compute r and $\|r\|$, but how can we estimate $\|A^{-1}\|$? It is too slow to compute A^{-1} directly and then calculate its norm.

Instead, we try to use the LU decomposition of A to estimate $\|A^{-1}\|$.

Idea. We want to estimate $\|B\|_1$ for a matrix B. By definition,

$$\|v\|_1 = \sum |v_i|$$

and

$$\|B\|_1 = \max_{x \neq 0} \frac{\|Bx\|_1}{\|x\|_1} = \underbrace{\max_j \sum_{i=1}^{n} |b_{ij}|}_{\text{largest column sum of B.}}$$

So one strategy would be to compute columns of $B = A^{-1}$ and measure their one-norms.

To compute a column, we must compute

$$B e_j = A^{-1} e_j = \vec{x}$$

or equivalently, to solve

$A\vec{x} = e_j$, which is an $O(n^2)$ operation given the LU decomposition of A

Thus computing all n columns would be again $O(n^3)$.

Observations

$$\|B\|_1 = \max_{\|x\|_1 \le 1} \|Bx\|_1 .$$

$\{x \mid \|x\|_1 \le 1\}$ is convex.

$f(x) = \|Bx\|_1$ is a convex function.

Check: $f(\alpha x + (1-\alpha) y) = \|\alpha Bx + (1-\alpha) By\|_1$

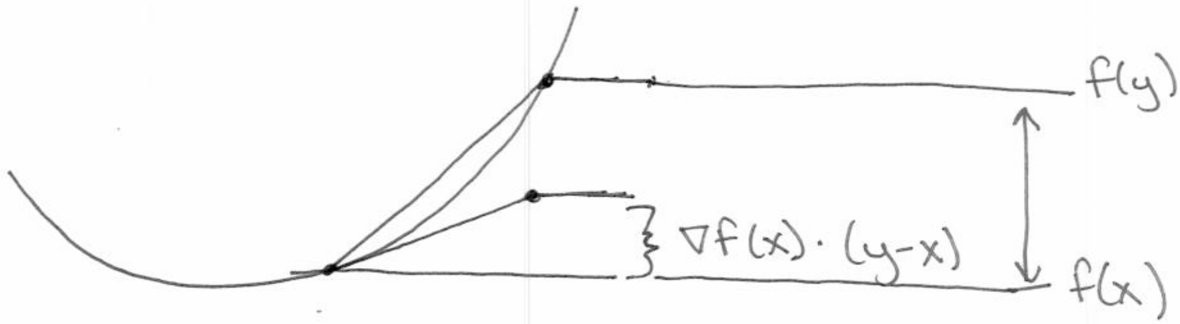
$$\le \alpha \|Bx\|_1 + (1-\alpha) \|By\|_1$$

$$\le \alpha f(x) + (1-\alpha) f(y).$$

(This is just the triangle inequality.)

Our plan is to apply a numerical method to maximize $f(x)$ on $\{x \mid \|x\|_1 \le 1\}$.

Now for any convex function



we have

$$f(y) - f(x) \geq \nabla f(x) \cdot (y - x)$$

So we can get a lower bound on $f(y)$ by

$$f(x) + \nabla f(x) \cdot (y - x) \leq f(y)$$

We will step according to this method (gradient ascent). But how do we compute ∇f?

$$f(x) = \sum_i \left| \sum_j b_{ij} x_j \right|$$

if $\sum_j b_{ij} x_j \neq 0$, let $s_i = \text{sign} \sum_j b_{ij} x_j = \pm 1$.

So

$$f(x) = \sum_{ij} s_i b_{ij} x_j$$

and

$$\frac{\partial f}{\partial x_k} = \sum_i s_i b_{ik}$$

so if S is the vector of signs

$$\nabla f = S^T B = (B^T S)^T$$

Now if $A = LU$ then $A^T = U^T L^T$ so $U^T L^T$ is the LU decomposition of A^T (the <u>upper</u> triangular matrix L^T is now the unit one, but who cares).

So we can ~~solve~~ find

$$(A^{-1})^T S = x \Rightarrow (A^T)^{-1} S = x$$

$$\Rightarrow S = A^T x$$

by solving $S = A^T x$ using $A^T = U^T L^T$.

We are left with

Algorithm (Hager's condition estimator)

Choose any x with $\|x\|_1 = 1$.

repeat {

  let $\omega = Bx$, $S = \text{sign}(\omega)$, $z = B^T S$.

  if $\|z\|_\infty \leq z^T x$ then

    return $\|\omega\|_1$

  else

    set $x = e_j$ where $|z_j| = \|z\|_\infty$

}

This is a little opaque when you first see it, so let's prove it works.

Theorem. When $\|w\|_1$ is returned, $\|w\|_1 = \|Bx\|_1$ is a local max of $f(x) = \|Bx\|_1$. Otherwise $\|Be_j\| > \|Bx\|$ so the algorithm has made progress.

Proof. Suppose $\|w\|_1$ is returned. We know $\|z\|_\infty \leq z^T x$. Now near x (as long as we don't change any signs),

$$f(x) = \|Bx\|_1 = \sum_{i,j} s_i b_{ij} x_j$$

is linear in x so

$$f(y) = f(x) + \nabla f(x) \cdot (y-x)$$

Now suppose y is near x and $\|y\|_1 = 1$. We want

$$\nabla f(x) \cdot (y-x) = z^T (y-x) \leq 0.$$

~~Bx~~

But

$$z^T(y-x) = z^T y - z^T x$$
$$= \sum z_i y_i - z^T x$$
$$\leq \sum |z_i||y_i| - z^T x$$
$$\leq \|z\|_\infty \|y\|_1 - z^T x$$
$$\leq \|z\|_\infty - z^T x \leq 0 \quad \text{as desired.}$$

Now suppose $\|z\|_\infty > z^T x$. We must show that if $\tilde{x} = e_j \cdot \text{sign}(z_j)$ where $|z_j| = \|z\|_\infty$, then

$$f(\tilde{x}) \geq f(x) + \nabla f(x) \cdot (\tilde{x} - x) \qquad \text{(convexity of } f\text{)}$$
$$\geq f(x) + z^T(\tilde{x} - x)$$
$$\geq f(x) + z^T \tilde{x} - z^T x$$
$$\geq f(x) + |z_j| - z^T x$$
$$\geq f(x) + \|z\|_\infty - z^T x$$
$$> f(x).$$

□

Experiments show that this is generally within a factor of 2 of the true condition number.

How do we use this in practice?

We know

$$\underset{\left(\text{in each entry}\right)}{\text{error}} = \frac{\|\delta x\|_\infty}{\|\hat{x}\|_\infty} \leq \|A^{-1}\|_\infty \frac{\|r\|_\infty}{\|\hat{x}\|_\infty}$$

Now $\|A^{-1}\|_\infty$ = max row sum while $\|A\|_1$ = max column sum so if we apply the condition estimator to find $\|(A^{-1})^T\|_1 = \|A^{-1}\|_\infty$, we can compute the rhs estimate.