Math 56 Compu & Expt Math, Spring 2014: Homework 3 Debrief

April 21, 2014

Please study this, and all the debriefings. I consider them part of the required reading for the course. You will learn a lot, especially if you redo the HW you lost points on.

1. 3+3+2 = 8 pts

   (a) The 3 rows of the linear system are the conditions that: eliminates the coefficient of $f(x)$, equates the coefficient of $f'(x)$ to unity, and eliminates the coefficient of $f''(x)$. The "stencil" is solved to be $[-3\ 4\ \ -1]/2h$. See eg Max.

   (b) It's best to test your formula to check it, e.g. like this:
   ```
   f=@(x)sin(x); fp=@(x)cos(x); h=1e-5; x=1; (-3*f(x)+4*f(x+h)-f(x+2*h))/2/h - fp(x)
   ```
   This gives error of about `3e-11`. Note I chose around the best $h = O(\varepsilon_{\text{mach}}^{1/3})$. Pawan did this with a nice log-log convergence plot covering all h.

   (c) Taylor's theorem with bounds on the $f'''(q)$ terms. Most of you had problems staying rigorous here, forgetting the absolute value signs. It was ok if you presented it in $O(h^2)$ form if you showed why the bound involving a sum of two f''' values tends to a constant (basically, you need f''' continuous).

2. 3+4 = 7 pts.

   (a) This was as in lecture (there we did $x_1 + x_2$).

   (b) This is tricky (good practise understanding the concept) and comes out backwards from what you first might expect from conditioning. Intuitively: $\varepsilon_{\text{mach}}$ relative error in $\cos x$ near $x = 0$ requires a large change in input x to account for it, because cos is "flat" there, ie $\cos(\varepsilon) \approx 1 - \varepsilon^2/2$. But near $x = \pi/2$, relative error in cos can easily be accounted for by $\varepsilon_{\text{mach}}$ relative change in x.

   It was also good if you used a model $\tilde{\cos}(x) = \cos(x(1+\varepsilon_1))(1+\varepsilon_2)$, showing cos cannot even be relatively accurate around $\pi/2$ (precisely because $\kappa \to \infty$); but it still is BS, as eg Max and Aron explain.

3. 2+3+3 = 8 pts

   (a) The eigvals of A are complex conjugate pair (typical of mostly rotation matrices) of equal magnitude, 2, which must be less than $\|A\|$ (can you prove this? Pawan did).

   (b) You may notice a shortcut that the sqrt of the *smallest* eigenvalue of A^TA tells you the reciprocal of $\|A^{-1}\|$. This is because $1/\lambda_{\min}(A^TA) = 1/\lambda_{\min}(AA^T) = \lambda_{\max}((AA^T)^{-1}) = \lambda_{\max}((A^{-1})^TA^{-1})$. The middle step comes since inverting a matrix inverts its eigenvalues.

   Thus $\kappa(A)$ is the sqrt of ratio of largest to smallest eigenvalue of A^TA.

   (c) See anyone's picture: the ellipse longest semi-axis is $\|A\|$, and shortest is $1/\|A^{-1}\|$ (why? max growth under A^{-1} must be min growth under A), so $\kappa(A)$ is the ratio of major to minor axes, as James explains. Michael showed that the unit circle *rotates* as it becomes the ellipse.

4. 3+3+3+3 = 12 pts. The point of this question is really that $\kappa(A)$ controls (via the Backwards Stability Theorem) only the *worst-case* errors, not always even the typical error.

   (a) Shows matrices with same norm may have wildly different $\kappa(A)$. In fact I chose `A1` to have *singular values* `10.^(-(0:99)/7)` which makes the ratio of largest to smallest 1.39×10^{14}.

(b) Here `bvec` was chosen to be unit magnitude, in the direction of `A1`'s largest ellipse semi-axis (growth output direction). However, perturbations of **b** are amplified by the reciprocal of the smallest growth factor of A, which was the reciprocal of the tiny 10^{-14}. Thus relative errors are huge. Consistent with theorem.

(c) Isn't it fascinating that now the relative error due to random perturbations of **b** is only around 10^{-14}, and that moreover this is typical of random RHS vectors? See Pawan for verbal explanation, and Matthew for picture. Note this doens't contradict the bounds, but is just surprisingly mall.

The random case is subtle: a random RHS vector **b** has $A^{-1}\mathbf{b}$ very large, $O(\|A^{-1}\|)$, so relative errors in **x** are small even though the absolute errors are large $O(10^{-2})$.

BONUS See discussion above, and draw yourself a sphere mapping under A to a very elongated ellipse. The clue is that **x** is size 1 in (b) (so that `bvec` aligns with output ellipse semi-major axis), but the largest possible size 1.39×10^{14} in (c) (so that `cvec` aligns with semi-minor axis).

Exactly the same issue comes up in Midterm 1, 2013, bonus for #4, explained in the solutions.

To understand better a *random* RHS vector, say **d**, it falls at random angle on a sphere, so that $\mathbf{x} = A^{-1}\mathbf{d}$ typically lies some fraction out along the longest semiaxis of A^{-1}, so is of typical size $\|A^{-1}\|$. This is like the case of **c**. See Matthew's nice pictures!

However, if instead it were the **x** that were chosen randomly, and used to generate $\mathbf{d} = A\mathbf{x}$, then the linear system solve generically would give behavior as in (b), because the growth factor of such an **x** is only $O(1)$.

(d) See eg Eli or James proof. Note despite appearances, Cauchy-Schwartz is *not* relevant here.

5. 5+4 = 9 pts. [Dan's question]

(a) Everyone made the correct plot and attributed the discrepancy to some form of round-off error. Many of you incorrectly attributed the discrepancy to the number of floating point operations that occurred in each method. This argument doesn't make sense. For instance, one could not argue that computing 9 via 3×3 and $1+1+1+1+1+1+1+1+1$ should lead to larger error in the second case because more floating point operations took place. The source of the error is (as usual) catastrophic cancellation. For a nice explanation see Max.

(b) Almost everyone saw the remedy here. One interesting thing observed by Michael and Max is that scaling the coefficients by the larger root and then applying the quadratic formula produced the right answer. Can anyone explain this? (+1 if you are able to...)