Lessons Taught by James Wilkinson

Margaret H. Wright

Computer Science Department
Courant Institute of Mathematical Sciences
New York University

Workshop on "Advances in Numerical Linear Algebra"
School of Mathematics
University of Manchester
Manchester, England
May 29, 2019

Many thanks to Nick Higham, Sven Hammarling, and Françoise Tisseur for organizing the workshop!

The context for this talk. . .

Jim's primary career was as a member of the British Scientific Civil Service at the National Physical Laboratory (NPL) in Teddington, England, which in its heyday was a center for basic research in a wide range of fields of science and engineering. (Alan Turing worked at NPL from 1945–1948.)

Jim was appointed at the NPL in 1946, partly in the General Computing Section and partly in the ACE (Automatic Computing Engine) section to work with Alan Turing. He remained at NPL until he retired in 1980. Starting in 1958, he also taught numerical linear algebra in summer schools at the University of Michigan. (Cleve Moler joined Jim as a lecturer in the Michigan courses in 1966.)

Jim visited Stanford for short periods during the 1960s, hosted by Gene Golub, and wrote several technical reports.

Starting in the early 1970s, Jim regularly visited and lectured at Argonne National Laboratory.

Sometime in the late 1970s, he was officially appointed as a visiting professor in the Computer Science Department at Stanford for one term per year. (The dates given for this vary slightly from source to source.)

Those Stanford courses were the inspiration for this talk, which draws from material contained in Jim's lecture notes. Some of the notes include dates, and from those we know that he taught 'CS 237A', 'CS 237B', and 'CS 238B' in 1977–1982.

The following people (students and/or attendees in Jim's courses) graciously made lecture notes as well as homework problems and solutions available, and also contributed anecdotes and insights—thank you, all!

Petter Bjørstad
William Coughran
Walter Gander
Eric Grosse
Michael Heath
Linda Kaufman
Steven Leon
Randall LeVeque
John Lewis
Stephen Nash
Michael Overton
Michael Saunders
L. N. Trefethen
Raymond Tuminaro
Patrick Worley

The available lecture notes address two broad topics:

1. Linear systems and error analysis

2. Eigenvalues

And there are associated homework problems, some involving computation, and their solutions.

Reading the notes all together makes clear that, although there is overlap in the topics covered, the results are presented sometimes in a different order, sometimes from a different perspective.

As does everyone who teaches, it is evident that Jim continued to rework the material, altering motivation, explanations, and examples.

A question for today: what to include in this 25-minute talk?

Definitely not eigenvalues, a topic that Jim begins by saying

It is our intention to motivate the whole of the background theory of the algebraic eigenvalue problem by consideration of linear differential equations.

For reasons of practicality, this talk will focus on selected parts of "Supplementary notes for CS237b, Winter Quarter 1982".

Following background about linear algebra, Jim turns to computing the solution x of $Ax = b$. Let ϵ denote a modest multiple of machine precision.

An aspect that he mentions early on, and repeatedly, is the "size" of x.

. . . if b is a *random* vector, the probability is very high that $\|x\|$ will be of the order $\|A^{-1}\|\|b\|$ rather than $\|b\|/\|A\|$. Most vectors b give 'large' solutions.

On the other hand, if one takes a random x and computes b from the relation $b = Ax$, the probability will be high that $\|x\|$ is of the order $\|b\|/\|A\|$. Most right-hand sides b produced in this way correspond to 'small' solutions. This result is important in practice because experimentors often construct right-hand sides b from random x in order to have systems of which they know the exact solutions. (Actually they don't know them anyway!)

Jim considers the correctly rounded version of x when x is not representable, namely $x + w$ with $\|w\| \leq \epsilon\|x\|$, and shows that even this paragon of virtue among approximate solutions cannot be expected to give a smaller residual than the exact solution y of $(A + \delta A)y = b$ with $\|\delta A\| \leq \epsilon\|A\|$.

In each set of notes, considerable space is devoted to matrix inversion. Here is a summary of the high points.

The exact inverse X is characterized by four properties, each of which is associated with a measure of excellence of an approximate inverse Y.

Property	Measure
$X - A^{-1} = 0$	1. $\|Y - A^{-1}\|/\|A^{-1}\|$
$X^{-1} - A = 0$	2. $\|Y^{-1} - A\|/\|A\|$
$AX - I = 0$	3. $\|AY - I\|$
$XA - I = 0$	4. $\|YA - I\|$

Let $\kappa = \|A\|\|A^{-1}\|$, the condition number of A.

Jim reviews these four choices of Y with respect to the given measures.

The point of interest is that the condition number κ obtrudes in the upper bounds on the measures for some choices of Y and some tests.

1. Consider $Y_1 = (A + \delta A)^{-1}$, with $\|\delta A\| \leq \epsilon \|A\|$, so that Y_1 is the exact inverse of a near neighbor of A. Then

$$\text{(test 2)} \quad \frac{\|Y_1^{-1} - A\|}{\|A\|} = \frac{\|\delta A\|}{\|A\|} \leq \epsilon \quad \text{(good), but}$$

$$\text{(test 1)} \quad \frac{\|Y_1 - A^{-1}\|}{\|A^{-1}\|} \leq \frac{\epsilon\kappa}{1 - \epsilon\kappa}. \quad (\kappa \text{ obtrudes}).$$

   The factor κ also obtrudes in tests 3 and 4.

2. What about $Y_2 = A^{-1} + W$, where $\|W\| \leq \epsilon \|A^{-1}\|$, so that Y_2 is a near neighbor of the exact inverse of A? By definition, Y_2 is good for test 1, but κ obtrudes in tests 3 and 4.

3. Y_3 is defined by

$$AY_3 - I = R, \text{ with } \|R\| \leq \epsilon.$$

Then

$$\frac{\|Y_3 - A^{-1}\|}{\|A^{-1}\|} \leq \epsilon \text{ (test 1)} \quad \text{and} \quad \frac{\|A - Y_3^{-1}\|}{\|A\|} \leq \frac{\epsilon}{1-\epsilon} \text{ (test 2).}$$

However,

$$\|Y_3 A - I\| \leq \epsilon\kappa. \quad \text{(test 4)}$$

4. Finally, if Y_4 is defined by $Y_4 A - I = R$ with $\|R\| \leq \epsilon$, Y_4 does well in tests 1 and 2, but κ obtrudes into the bound in test 3.

What about A_R^{-1}, the correctly rounded version of A^{-1}, which satisfies

$$A_R^{-1} = A^{-1} + E, \quad \text{where} \quad \|E\| \leq \epsilon \|A^{-1}\|.$$

How does A_R^{-1} perform on the different tests?

The corresponding relations are

$$AA_R^{-1} = A(A^{-1} + E) = I + AE, \quad \text{with} \quad \|AE\| \leq \|A\|\|E\| \leq \epsilon\kappa$$

so that κ obtrudes and this bound will usually be realistic.

Suppose we want to compute the explicit inverse X.

This is typically done by solving for x_i, the i-th column of X, from

$$Ax_i = e_i,$$

where e_i is the i-th coordinate vector.

Our expectation is that we will obtain y_i (the computed i-th column of X) that satisfies

$$(A + \delta A_i)y_i = e_i \text{ with } \|\delta A_i\| \leq \epsilon \|A\|.$$

Note that in general δA_i will be different for each i.

Jim asks: how 'good' would Y be?

Answer: apply the four tests.

Note that $AY = I - \begin{pmatrix} \delta A_1\, y_1 & \delta A_2\, y_2 & \dots & \delta A_n\, y_n \end{pmatrix} \triangleq I - E$. We have $\|E\| \leq \epsilon \|A\| \|Y\|$, so that

$$\|Y\| \leq \frac{\|A^{-1}\|}{1 - \epsilon\kappa} \qquad \text{and} \qquad \|AY - I\| \leq \frac{\epsilon\kappa}{1 - \epsilon\kappa}.$$

Given these bounds, Y performs just as well on test 3 as the inverse of $A + \delta A$.

For test 4, however, we have $YA - I = A^{-1}(AY - I)A$, so that

$$\|YA - I\| \leq \kappa \|AY - I\| \leq \frac{\epsilon\kappa^2}{1 - \epsilon\kappa},$$

with an 'extra' factor of κ.

The bound involving κ^2 calls for further comments. It is realistic only when the δA_i are random, uncorrelated perturbations satisfying $\|\delta_i A\| \leq \epsilon \|A\|$.

In practice, when computing the columns of an inverse by a *stable direct method*, although the δA_i are all different, they are not a random set. They are related in such a way that we can expect $\|YA - I\|$ to be of the same order of magnitude as $\|AY - I\|$.

The extra factor κ does not materialize.

In order to emphasize that the 'accumulation of errors' is not the important feature we consider approximate inverses for an ill-conditioned 2×2 example:

$$A = \begin{pmatrix} 0.8623 & 0.7312 \\ 0.8177 & 0.6935 \end{pmatrix}, \quad \text{with} \quad \kappa \approx 2.36 \times 10^4.$$

Here are A^{-1} and an approximate inverse Y:

$$A^{-1} = \frac{10^4}{1.0281} \begin{pmatrix} 0.6935 & -0.7312 \\ -0.8177 & 0.8623 \end{pmatrix}$$

$$Y = 10^4 \begin{pmatrix} 0.6745 & -0.7112 \\ -0.7954 & 0.8387 \end{pmatrix}.$$

We have (exactly)

$$AY = \begin{pmatrix} 0.2487 & -0.1032 \\ -0.7125 & 0.9021 \end{pmatrix} \quad \text{and} \quad YA = \begin{pmatrix} 0.7311 & -0.2280 \\ -0.6843 & 0.4197 \end{pmatrix}$$

so that $\|AY - I\|$ and $\|YA - I\|$ are $O(1)$.

We cannot expect them to be small because A^{-1} is so large that the mere act of rounding to 4 decimals introduces errors of order unity that are almost certain to be uncorrelated.

The exact inverse of Y is given by

$$Y^{-1} = \frac{1}{0.1467}\begin{pmatrix} 0.8387 & 0.7112 \\ 0.7954 & 0.6745 \end{pmatrix} = \begin{pmatrix} 5.7171\ldots & 4.8479\ldots \\ 5.4219\ldots & 4.5978\ldots \end{pmatrix}.$$

Hence Y^{-1} differs from A by quantities of $O(1)$. This is to be expected given that $\kappa = O(10^4)$. However, Y^{-1} is wrong in rather a strange way. To four significant decimals, $Y^{-1} = 6.630A$, i.e. Y^{-1} is almost a multiple of A.

Can you explain why?

Now we consider an approximate inverse Z in which z_1 and z_2 are the first and second columns of the inverses of

$$A+\delta A_1 = \begin{pmatrix} 0.8624 & 0.7323 \\ 0.8177 & 0.6935 \end{pmatrix} \quad \text{and} \quad A+\delta A_2 = \begin{pmatrix} 0.8623 & 0.7312 \\ 0.8177 & 0.6936 \end{pmatrix},$$

where $\|\delta A_i\| = 10^{-4}$ and $\delta_1 A \neq \delta_2 A$.

The exact z_1 and z_2 are

$$z_1 = \frac{10^4}{1.7216} \begin{pmatrix} 0.6935 \\ -0.8177 \end{pmatrix} \approx 10^4 \begin{pmatrix} 0.4028 \\ -0.4750 \end{pmatrix}$$

$$z_2 = \frac{10^4}{1.8904} \begin{pmatrix} 0.7312 \\ 0.8623 \end{pmatrix} \approx 10^4 \begin{pmatrix} -0.3868 \\ 0.4561 \end{pmatrix}.$$

Using rounded versions of z_1 and z_2 as the columns of Z,

$$AZ = \begin{pmatrix} 0.1444 & -0.3732 \\ -0.4294 & 0.1899 \end{pmatrix}$$

and $\|AZ - I\| = O(1)$. But now

$$ZA = \begin{pmatrix} 310.4808 & 262.8156 \\ -366.3953 & -310.1465 \end{pmatrix},$$

and $\|ZA - I\|$ is larger by a factor of order 1000. Although $z_1 \approx 0.5972x_1$ and $z_2 \approx 0.5439x_2$, the difference in the multiples does not affect AZ because z_1 and z_2 are not 'mixed'. But in ZA, each element involves one element from z_1 and one from z_2, and the fact that the multiples are different is disastrous.

One peripheral result is that, when an inverse of a matrix with $\kappa = O(1/\epsilon)$ is calculated *using a stable algorithm*, the multiples in each column are almost exactly the same.

Leaving matrix inverses and moving on to error analysis, Jim repeats that error analysis has concentrated too much on the details of floating point, which is unfortunate because the important features of the results are often really quite independent of these details. They would remain the same even if we were considering computation with integers and the old errors were blunders.

In this spirit, he introduces a formalism in which a quantity to be computed is defined in terms of known quantities by a very simple equation. If the new computed quantity is inserted into the equation, a discrepancy may arise because of computing errors. But the discrepancy is merely the sum of all errors made during the computation; there is no interaction.

We do not concern ourselves with what the various quantities would have been if we had produced them exactly. This apparently trivial remark is of vital importance.

A familiar relation in Gaussian elimination is $\bar{M}\bar{A}^{(n)} = A + \Delta$, which can be looked at in two different ways.

We can say that we have computed $\bar{M}$ and $\bar{A}^{(n)}$ and in order to assess our performance we are looking at $\bar{M}\bar{A}^{(n)} - A$. This is a perfectly natural approach. It pervades the whole of numerical analysis. the whole of numerical analysis.

Alternatively we could say that if we take $A + \Delta$ and do Gaussian elimination (or the equivalent factorization) exactly we shall end up precisely with $\bar{M}$ and $\bar{A}$. Historically the second viewpoint was the one which motivated me and led to my introduction of the term 'backward' error analysis. Its heuristic value to me was enormous and greatly increased the impact of the results at the time but I doubt whether I would stress it quite so much if I were starting from scratch.

Trivial example 1: Consider the equations

$$x + 2y + 3z = 1$$
$$2x + y + 5z = 2$$
$$3x + 2y + 4z = 3.$$

and assume that, in the very first step, I blunder and get $m_{21} = 3$ rather than 2, but do everything else correctly.

If I attempt to trace the effect of this, we find that the whole path of the subsequent computation is altered in rather a miserable way.

Instead it can be said that a_{21} was wrongly taken as 3 rather than 2. If everything else is done correctly, the result will be the exact solution of an original system in which $a_{31} = 3$.

The computing error is reflected straight back to the original system and is uninfluenced by anything else.

A second, less trivial example: let $\epsilon = 10^{-4}$, and suppose that we attempt to solve the following system without pivoting, using three-decimal digit arithmetic:

$$\begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

The exact solution is

$$x_2 = \frac{1-2\epsilon}{1-\epsilon} \quad \text{and} \quad x_1 = \frac{1}{1-\epsilon} \quad \text{(very close to } (1,1)\text{)}.$$

We have $m_{21} = 10^4$ (exactly) and the reduced system is

$$\begin{pmatrix} 10^{-4} & 1 \\ 0 & -10^4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -10^4 \end{pmatrix}, \quad \text{giving} \quad x = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

We are sadly adrift. The original system of equations is *very* well conditioned. A poor result is not acceptable.

How can this failure be described in terms of backward error analysis?

$$\text{exact } a_{22}^{(2)} = 1 - 10^4; \quad \text{computed } \bar{a}_{22}^{(2)} = -10^4$$

$$\text{exact } b_2^{(2)} = 1 - 10^4; \quad \text{computed } \bar{b}_2^{(2)} = -10^4.$$

The computed values $\bar{a}_{22}^{(2)}$ and $\bar{b}_2^{(2)}$ are exactly equal to what they would be if $a_{22}^{(1)} = 0$ and $b_2^{(1)} = 0$. Thus we have computed the exact reduction of

$$\begin{pmatrix} 10^{-4} & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{1}$$

and, since we did not make any other independent errors, we will obtain the exact solution of (1).

It follows from this analysis that exactly the same reduced system (1) would be the result if a_{22} and b_2 take a whole range of values. But, because of the need to subtract 10^4, they are 'lost'. This is the rationale for pivoting, expressed through the backward viewpoint.

Notice that the failure occurs although the equation

$$-10^4 x_2 = -10^4$$

is a *very good* equation. It gives x_2 almost exactly. It is the original exact equation that lets us down. The original equations are a miserable pair.

Who but Jim would describe this innocuous relation fondly as "a very good equation" and two innocent equations as "a miserable pair"?

To the best of my knowledge, the "blunder" motivation for error analysis has not been widely adopted.

But I am considering it for my next class on numerical computing!

And it seems appropriate to repeat the dedication to Jim in *Numerical Linear Algebra and Optimization* (1991):

Jim's fundamental contributions influenced literally all fields of scientific computing, and were matched by his intellectual generosity and personal warmth. ...his masterly papers and books ...offer new insights each time they are read. Jim's formal teachings and lectures were a never-failing source of understanding and inspiration. A definitive measure of Jim's pedagogical skills was his unsurpassed ability to deliver an informative and *entertaining* talk on error analysis!

In conclusion, please enjoy a homework problem from 1982:

3. In my lectures I introduced the concept of 'small' solutions of $Ax = b$, i.e., solutions of the order of magnitude $\|b\|/\|A\|$) and 'large' solutions (i.e., solutions of the order of magnitude $\|A^{-1}\|\|b\|$). Consider the two following physical problems.

(i) We wish to solve Laplace's equation $\nabla^2 u = 0$ on a unit square with given boundary values $u = v(x, y)$. We solve it by discretization using the familiar pattern

```
     -1
-1    4  -1
     -1
```

If the equations are $Ax = b$ you may assume that $\kappa(A) = O(1/h^2)$ where h is the mesh size. Show that, whatever v may be, the corresponding b will give an x which is right at the 'small' end of the range.

(ii) We now wish to solve Poisson's equation $\nabla^2 u = \rho$ with $u = 0$ on the boundary. Assuming that $\rho(x, y)$ is a 'reasonably sensible function', show that b is such that as $h \to 0$ the corresponding x is right at the 'large' end of the range.

(iii) Can you explain how it is in (i) that x is always 'small' although v is arbitrary?

(iv) What is the flaw in the following argument in connexion with (i):

(a) The computed solution is the solution of $(A + \delta A)\bar{x} = b$ with $\|\delta A\| \leq \epsilon f(n)\|A\|$, where $f(n)$ is rather unimportant.

(b) $r = b - A\bar{x} = \delta A\bar{x}$. Hence $\|r\| \leq \|\delta A\|\,\|\bar{x}\|$.

(c) $\bar{x}$ is at the small end and hence $\|\bar{x}\| = O(\|b\|/\|A\|)$.

(d) Hence $\|r\| = O(\|\delta A\|\,\|b\|/\|A\|)$, i.e., $O(\epsilon f(n)\|b\|)$.

(Don't be too critical about the use of $O(\cdot)$, that is *not* the point at issue.

(e) The elements of $\|b\|$ come *directly* from the boundary values v and hence we have the exact solution of a problem with boundary values $\bar{v}$ *very* close to v (i.e., with $\|v - \bar{v}\|$ independent of $\kappa(A)$).

(v) For problem (i) defining r to be $b - A\bar{x}$, what is the discretized physical problem to which we *do* have the exact solution? (You may describe this physical problem in terms of the components r_i of r.)

(vi) If X is the correctly rounded inverse of A, how good would Xb be for problems (i) and (ii) respectively as regards accuracy? (Since A^{-1} is not sparse, the efficiency is not good anyway.)

Here is Jim's solution to problem 3 in Homework 2, Winter 1982.

3.

```
    v1      v2      v3      v4
v16    1       2       3       4    v5
v15    5       6       7       8    v6
v14    9      10      11      12    v7
v13   13      14      15      16    v8
    v12    v11     v10      v9
```

(i) For Laplace's equation we see that for an internal point typically

$$4x_6 - x_5 - x_2 - x_7 - x_{10} = 0.$$

An internal value could not therefore be greater than all other values. The maximum is therefore achieved at a boundary point. For boundary points we have typically

$$4x_5 - x_1 - x_6 - x_9 = v_{15} \quad \text{(non-corner point)}$$

$$4x_1 - x_2 - x_5 = v_1 + v_{16} \quad \text{(corner point)}.$$

Obviously $x_i < \max(v_i)$ $(\forall i, j)$. This is the discretized version of the maximum principle. Similarly, $x_i > \min(v_j)$ $(\forall i, j)$. This means

$$\|x\|_\infty \leq \|v\|_\infty = \|b\|_\infty.$$

Hence, however irregular the boundary values may be, $\|x\|_\infty$ is tied down by $\|b\|$ although $\|A^{-1}\| = O(1/h^2) \to \infty$ as $h \to 0$. The solution is right at the small end.

(ii) For Poisson's equation the R.H.S. corresponding to point 'i' is $h^2\rho(x, y)$ at that point. However, as $h \to 0$, the solution $x \to$ solution of the continuous problem, i.e., for a given function $\rho(x, y)$ it is tending to a fixed function $u(x, y)$. But $\|b\|_\infty = O(h^2)$ and $\|A^{-1}\| = O(1/h^2)$. The solution is $O(1)$ i.e., $O(\|A^{-1}\|\|b\|)$. The solution is right up at the large end.

(iii) Although v is quite arbitrary the R.H.S. of the system of equations is always very special. If $1/h = n + 1$ we have n^2 equations and for $(n - 2)^2$ of these the R.H.S. element is zero. It is therefore not at all surprising that if b is expanded in terms of the eigenvectors of A, those corresponding to the small eigenvalues (eigenvalues are singular values here because A is positive definite) are absent. You will find it instructive to look at these vectors and to see that b is deficient in them.

(iv) It is (e) that is nonsense. The residuals are all small, indeed at the level of machine precision relative to $\|b\|$, but these residuals occur in all n^2 positions, i.e., they include all the $(n-2)^2$ positions in which the R.H.S. is zero. Hence in addition to a small relative perturbation in $\|v\|$ we have the other $(n-2)^2$ perturbations.

(v) The solution we have is the exact solution of the physical problem $\nabla^2 u = \rho$ with u given by v_i (for appropriate values of i) on the boundary and ρ given by r_i/h^2 at any internal point. Hence the very small residual corresponds to a charge distribution ρ which is $1/h^2$ times as large.

(vi) Since (i) has a small solution Xb would be disappointing for the Laplace problem. For (ii) Xb would be very good.