Scientific Computing

A practical Companion

5th Notebook

This notebook can be downloaded from the location:
http : //
staff.science.uva.nl / ~walter / SC / Notebooks / SC08 – 5. nb

Author:

Walter Hoffmann (Korteweg-de Vries Institute for Mathematics, UvA)

February - March, 2008

Systems of linear equations I

On the accuracy of the solution

Consider a system of linear equations:

$$\mathbf{A}\mathbf{x} = \boldsymbol{b}\,.$$

If A is invertible, then the system has a unique solution $x = A^{-1}\,b$.

Suppose that $\tilde{x}$ is an approximate solution.

The question "how good is $\tilde{x}$?" can be answered in two different ways.

1. How well does $\tilde{x}$ satisfy the original equations?
2. How large is the difference between $A^{-1}\,b$ and $\tilde{x}$?

Both questions can be answered in a quantitive way if we can measure the size of a vector in some sense.

For this purpose we have the notion of a norm

▽ Norms of vectors and matrices

Defintions:

Vector norms

The 1-norm of a vector x is defined by

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{\boldsymbol{n}} |\boldsymbol{x}_i|\,.$$

The 2-norm of a vector x is defined by

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{\boldsymbol{n}} |\boldsymbol{x}_i|^2}\,.$$

The 2-norm is also known as

Euclidean norm (Euclides [1]), or

spectral norm, or

Frobenius norm (Frobenius [2]).

The ∞-norm (or max norm, or supremum norm) of a vector x is defined by

$$\|\boldsymbol{x}\|_\infty = \max_{i=1,\ldots,n} |\boldsymbol{x}_i|\,.$$

The three vector-norms we have introduced, are special cases of a more general definition of norms, the so called Hölder-norms (Hölder [3]), or p-norms:

$$\| x \|_p = \sqrt[p]{\sum_{i=1}^{n} | x_i |^p} .$$

The supremum norm follows as a special case from this general definition by taking the limit for p to infinity.

Properties

All definitions of norms obey the following rules

(in fact these rules were formulated as axioms, after which the various norms have been constructed)

A1. $\| x + y \| \leq \| x \| + \| y \|$

A2. $\| \alpha x \| = | \alpha | \cdot \| x \|$

A3. $\| x \| = 0 \iff x = \vec{0}$

and as a consequnce from A1 it can be proven that

4. $\| x - y \| \geq \left| \| x \| - \| y \| \right|$

Matrix norms

The following matrix norms also obey the rules that are given in the above properties:

The 1-norm of a matrix A is defined by

$$\| A \|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} \left| A_{ij} \right| .$$

The ∞-norm of a matrix A is defined by

$$\| A \|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} \left| A_{ij} \right| .$$

The 2-norm of a matrix A is defined by

$$\| A \|_2 = \sqrt{\lambda_{max} (A^T A)} .$$

For matrices, the latter norm is called either the 2-norm or the spectral norm

The Euclidean norm or Frobenius norm of a matrix is defined by

$$\| A \|_E = \sqrt{\sum_{i,j=1}^{n} | A_{ij} |^2} \quad .$$

More Properties

Next to the rules A1, A2 and A3, matrices also obey the product rule

$$\| A.B \| \leq \| A \| . \| B \|$$

and probably the most important formula:

$$\| Ax \| \leq \| A \| . \| x \|$$

[1]) Euclides, 325 - 265 BC

[2]) F.G. Frobenius, 1849 - 1917

[3]) O.L. Hölder, 1859 - 1937

▽ Examples:

Vector x and matrix A are given by

$$x = \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix}; \; A = \begin{pmatrix} 3 & -2 \\ -1 & 1 \end{pmatrix};$$

then

$$\| x \|_1 = 6; \; \| x \|_\infty = 3; \; \| x \|_2 = \sqrt{14};$$

$$\| A \|_1 = 4; \; \| A \|_\infty = 5; \; \| A \|_E = \sqrt{15};$$

All these norms were calculated by hart; calculation of the spectral norm for a matrix is much more difficult and can best be done by use of a computer.

In[112]:=

```
A = {{3, −2}, {−1, 1}};
```

In[113]:=

```
Norm[A]
```

Out[113]=

$$\sqrt{\frac{1}{2}\left(15+\sqrt{221}\right)}$$

In[114]:=

```
N[%]
```

Out[114]=

3.86433

In[115]:=

```
Norm[A, 1]
Norm[A, Infinity]
```

Out[115]=

4

Out[116]=

5

In[117]:=

```
x = {2, −3, 1};
```

In[118]:=

```
Norm[x, Infinity]
Norm[x, 1]
Norm[x]
```

Out[118]=

3

Out[119]=

6

Out[120]=

$\sqrt{14}$

Now, with the knowledge of norms, question 1 can be quantified by calculating a norm of the so called *residual vector* r defined by:

$$r = b - A\tilde{x}\,.$$

Question 2 is much more difficult to answer; we try to estimate a norm of the *error vecor* e defined by

$$e = x - \tilde{x}\,.$$

Residual vector r and error vector e are related via the so called residual equation (which is easy to proof):

$$A\,e = r\,.$$

Independent of performing any calculations at all, one may ask the question

How sensitive is the solution of a linear system for perturbations in its right-hand side and/or in the coefficientmatrix?

Consider once more the equality

$$A\,x = b\,.$$

We see a relation between a matrix A and two vectors, x and b, and what is given or what is to be solved is not of importance: it is a static relation. A similar equation describes the relation between a perturbation of x and a perturbation of b with the same coefficient matrix A:

$$A\,(x + \delta\mathbf{x}) = (b + \delta\mathbf{b})$$

and because of the linearity of matrix multiplication, the perturbations $\delta\mathbf{x}$ and $\delta\mathbf{b}$ obey the relation

$$A\,\delta\mathbf{x} = \delta\mathbf{b}\,.$$

For a further analysis we assume that matrix A is invertible (which is true for many systems of linear equations) so that the same relation between $\delta\mathbf{x}$ and $\delta\mathbf{b}$ can also be expressed by:

$$\delta\mathbf{x} = A^{-1}\,\delta\mathbf{b}$$

and the norms of $\delta\mathbf{x}$ and $\delta\mathbf{b}$ are related by the inequality:

$$\|\,\delta\mathbf{x}\,\| \leq \|\,A^{-1}\,\|\;\|\,\delta\mathbf{b}\,\|\,.$$

Combination with the inequality that can be derived from the original equation:

$$\frac{1}{\|x\|} \leq \|A\| \frac{1}{\|b\|}$$

yields the famous *relative perturbation inequality*:

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\| \, \|A^{-1}\| \, \frac{\|\delta b\|}{\|b\|}$$

Let us try to understand the implications of this inequality for the problem of solving a system of linear equations.

The interpretation is:

A relative perturbation in the right-hand side may be magnified by the quantity $\|A\| \, \|A^{-1}\|$ for the relative perturbation in the solution.

The quantity

$$\kappa(A) = \|A\| \, \|A^{-1}\|$$

is called

the condition number of a matrix with respect to the problem of solving systems of linear equations

or mostly, simply

the condition number of A

For all, but the Frobenius norm, we have $\kappa(A) \geq 1$, which follows from

$$1 = \|I\| = \|A A^{-1}\| \leq \|A\| \, \|A^{-1}\|$$

$\kappa(A)$ can be calculated (or estimated, like is done in many software libraries) and gives a prediction for the accuracy of the solution in cases that the accuracy of the right-hand side vector is given.

If the elements of b are floating-point numbers that have rounding errors, then at least this rounding error may be magnified by $\kappa(A)$ to produce the relative error in the solution.

Consequently

for $\kappa(A)$ larger then $\frac{1}{\text{macheps}}$ ($\approx 10^{16}$),

the relative error in the solution may be more then 100% or in other words,

may be absolutely worthless.

▽ Examples on comparing the solutions of 'nearby' sytems of linear equations.

In[121]:=

```
A = {{4.1, 2.8}, {9.7, 6.6}};
A // MatrixForm
b = {4.1, 9.7};
b // MatrixForm
```

Out[122]//MatrixForm=

$$\begin{pmatrix} 4.1 & 2.8 \\ 9.7 & 6.6 \end{pmatrix}$$

Out[124]//MatrixForm=

$$\begin{pmatrix} 4.1 \\ 9.7 \end{pmatrix}$$

In[125]:=

```
x = LinearSolve[A, b];
x // MatrixForm
```

Out[126]//MatrixForm=

$$\begin{pmatrix} 1. \\ -2.63678 \times 10^{-15} \end{pmatrix}$$

In[127]:=

```
bp = {4.11, 9.7};
bp // MatrixForm
```

Out[128]//MatrixForm=

$$\begin{pmatrix} 4.11 \\ 9.7 \end{pmatrix}$$

In[129]:=

```
xp = LinearSolve[A, bp];
xp // MatrixForm
```

Out[130]//MatrixForm=

$$\begin{pmatrix} 0.34 \\ 0.97 \end{pmatrix}$$

The large difference between the two solutions of this related linear systems having their right-hand sides

so close together, is explained by the size of the conditionnumber. For quantities that are only given up to two decimals, this is quite large.: $K_\infty(A) = 2249.4$

The residual vector r can be used to estimate a so called *a posteriori* error bound.

Suppose that for given A and b, $\tilde{x}$ is an approximate solution of the system of linear equations

$$A x = b.$$

The following relations come from earlier definitions:

$$\begin{aligned} A\tilde{x} &= b - r \\ \tilde{x} &= x - e \\ \Longrightarrow A\ (x - e) &= b - r \end{aligned}$$

so application of the *relative perturbation* inequality yields:

$$\frac{\|e\|}{\|x\|} \leq \kappa(A) \frac{\|r\|}{\|b\|}$$

A lowerbound for the relative error $\frac{\|e\|}{\|x\|}$ can also be derived:

From **Ae = r** we find

$$\|e\| \geq \frac{1}{\|A\|} \|r\|$$

and from **x = A^{-1} b** we find

$$\frac{1}{\|x\|} \geq \frac{1}{\|A^{-1}\|} \frac{1}{\|b\|}$$

and the combination yields:

$$\frac{\|e\|}{\|x\|} \geq \frac{1}{\kappa(A)} \frac{\|r\|}{\|b\|}$$

▽ The residual vector estimates the error

In[131]:=

```
A = {{1, 0.5}, {2.001, 0.999}};
A // MatrixForm
b = {1.5, 3};
b // MatrixForm
```

Out[132]//MatrixForm=

$$\begin{pmatrix} 1 & 0.5 \\ 2.001 & 0.999 \end{pmatrix}$$

Out[134]//MatrixForm=

$$\begin{pmatrix} 1.5 \\ 3 \end{pmatrix}$$

By some routine, an approximate solution is calculated:

In[135]:=

```
xe = {0, 3};
xe // MatrixForm
```

Out[136]//MatrixForm=

$$\begin{pmatrix} 0 \\ 3 \end{pmatrix}$$

The residual vector is calculated:

In[137]:=

```
r = b − A.xe;
r // MatrixForm
```

Out[138]//MatrixForm=

$$\begin{pmatrix} 0. \\ 0.003 \end{pmatrix}$$

Although the residual vector seems fairly small; the actual error may be much larger. In this example we directly see the correct solution of the system and we know that the error is relatively large. For the relative error we find:

In[139]:=
```
x = {1., 1.};
e = x - xe;
relerr = Norm[e] / Norm[x]
```

Out[141]=
```
1.58114
```

The upperbound for the error should take into account the conditionnumber of the matrix. In many cases computing the inverse of A takes far too much time. We do it here for illustration purposes.

In[142]:=
```
kappa = Norm[A] * Norm[Inverse[A]]
```

Out[142]=
```
4168.
```

In[143]:=
```
erbnd = kappa * Norm[r] / Norm[b]
```

Out[143]=
```
3.72797
```

Consider also perturbations in the coefficient matrix

Until now, the 'nearby' systems of linear equations were only different in their right-hand sides. The bounds on the 'differences' ('errors', 'perturbations') were all stated without considering changes in the coefficient matrix.

The situation becomes a little bit more complicated (but not much) when we also consider perturbations in the matrix elements.

To finish this chapter, we formulate a theorem that describes the effects of perturbations in both the elements of the right-hand side and the matrix elements; it is a more general version of the *relative perturbation* inequality.

Next to equation

$$\mathbf{A}\mathbf{x} = \boldsymbol{b}$$

we consider the perturbed equation

$$(\boldsymbol{A} + \delta\mathbf{A})(\boldsymbol{x} + \delta\mathbf{x}) = (\boldsymbol{b} + \delta\mathbf{b}).$$

and we look for an expression that gives a bound on $\frac{||\delta\mathbf{x}||}{||\mathbf{x}||}$ in terms of the other quantities.

We require that A + δA is a nonsingular matrix; like A itself.

A sufficient condition is:

$$\| \delta \mathbf{A} \| \, \| A^{-1} \| \, (= \rho) < 1$$

The inequality that bounds the relative error in the solution can now be formulated as

$$\frac{\| \delta \mathbf{x} \|}{\| x \|} \leq \kappa (A) \left(\frac{\| \delta \mathbf{A} \|}{\| A \|} + \frac{\| \delta \mathbf{b} \|}{\| b \|} \right) \frac{1}{1 - \rho}$$

Systems of linear equations II
Calculating a solution

To compute the solution of a given system of linear equations on a computer, we distinguish between two classes of methods and for both classes a multitude of algorithms do exist.

The two classes are

1. Direct methods
2. Iterative methods

All methods from the two classes have their advantages and disadvantages;

there is no favourite method for all linear equations problems.

The main difference between the methods in both classes is:

In direct methods the matrixelements are changed during computation

With iterative methods the matrix elements remain unchanged.

In this course emphasis is on iterative methods.

Of direct methods, we only mention the main characteristics.

The elements of the coefficient matrix are used to find (expressions for) other matrices L, U, Q, ...

such that matrix A can be factored as a product of two or more matrices; for instance

$$A = P L U$$

with matrices that are in some sense simpler than A. The computation of these matrices has in general time complexity $O(n^3)$; the memory usage is such that all elements of the matrix factors can be stored in the locations where originally the elements of A were stored, plus possibly some O(n) elements extra.

The general way of solving the linear system after the factorization has been accomplished, is as follows:

$$A x = b \iff$$
$$P L U x = b.$$

The following steps have time complexity O (n^2) in general; they consist of simple vector operations and evaluate the following relations:

Compute successively **y** and **z** such that

$$P\,y = b;\ L\,z = y;\ U\,x = z.$$

At a first glance it seems as if again systems of equations have to be solved, but the matrixfactors are always such that solving the systems is a trivial matter. The factors are either triangular matrices, orthogonal matrices or diagonal matrices.

Convince yourself that finally vector x that is found in the last step satisfies Ax = b.

The methods that belong to the class of direct methods are among others:

Gaussian eleimination (algorithmically equivalent to LU factorization),

Cholesky factorization,

QR factorization,

SingularValue Decomposition.

(to name just a few)

All methods have in common that good stable implementations do exist; the implementations is almost always

backward stable, which means:

The calculated solution is an
exact solution of *a* nearby problem

Or in mathematical language: the calculated solution $\tilde{x}$ = x+δx is always the exact solution of a system of equations (A + δA) (x+δx) = (b+δb) and the bounds on ||δA|| and ||δb|| are only determined by the method chosen (and its implementation). In general they are always bounded by O(n^2) $\times$ macheps with relatively small constants for the coefficients of n^2 and n.

The bottomline is,

A small residual vector is almost always guaranteed;

and as a consequence:

The error in the solution is mainly determined by the conditionnumber of the given matrix.

Close all sections