

Accuracy and Stability in View of the Eigenproblem

Jason Slemons and Claus Kirchner

University of Washington

Department of Applied Mathematics

13 April 2006

Overview

- Basics on Floating Point Arithmetic
- Basic Definitions and Results
 - Condition of a problem, condition of a matrix
 - Absolute and relative error
 - Stability of an algorithm
- Perturbation Theory and relation to eigenvalue computation
 - Examples
 - Classical results for square matrices
 - Symmetric eigenproblem

Basics on Floating Point Arithmetic

The number

$$\pm .b_1b_2 \dots b_p \times \beta^e \quad 0 \leq b_i \leq \beta - 1, \quad b_1 \neq 0 \quad (1)$$

is called normalized p -digit floating point number with base β .
 e is called the exponent.

Given a number α that has more than p digits

$$\alpha = \pm .b_1b_2 \dots b_pb_{p+1} \dots b_m \times \beta^e,$$

it has to be *rounded* to fit it into format (1).

Basics on Floating Point Arithmetic

This rounding can be performed as follows:

$$\tilde{\alpha} = \pm .b_1 b_2 \dots \hat{b}_p \times \beta^e$$

where

$$\hat{b}_p = \begin{cases} b_p & \text{if } b_{p+1} < \frac{1}{2}\beta \\ b_p + 1 & \text{if } b_{p+1} \geq \frac{1}{2}\beta \end{cases}$$

It can be shown that

$$\tilde{\alpha} = \alpha(1 + \epsilon)$$

where $|\epsilon| = \frac{1}{2}\beta^{1-p}$ is called the *unit roundoff* or *machine-precision* or *machine- ϵ* .

Basics on Floating Point Arithmetic

The exponent e in the floating point representation determines the largest and smallest representable numbers in magnitude.

Let $-e_{min} \leq e \leq e_{max}$, $e_{min}, e_{max} \in \mathbb{N}$.

For a floating point number α as in (1) we have

$$|\alpha| \leq (1 - \beta^{-1-p}) \cdot \beta^{e_{max}} =: \alpha_{max} \quad (2)$$

$$|\alpha| \geq \beta^{-1-e_{min}} =: \alpha_{min} \quad (3)$$

Overflow occurs if $|\alpha| > \alpha_{max}$.

Underflow occurs if $|\alpha| < \alpha_{min}$.

Example 1: Quadratic Equation

We want to compute the eigenvalues of the matrix

$$A = \begin{pmatrix} .0154 & .2989 \\ .2989 & 6.148 \end{pmatrix}$$

In four-digit arithmetic, this leads to the following quadratic equation

$$\lambda^2 - 6.433\lambda + 0.0095 = 0 \quad (4)$$

The smaller root is given by

$$\lambda = \frac{6.433 - \sqrt{6.433^2 - 4 \cdot 0.0095}}{2} = .1477099614 \cdot 10^{-2} \quad (5)$$

Using formula (5) in four-digit arithmetic ($p = 4$ in (1)), one obtains

$$\tilde{\lambda} = 0.15 \times 10^{-2} \quad (6)$$

Example 1: Quadratic Equation

Using the formula

$$\lambda = \frac{2 \cdot 0.0095}{6.433 + \sqrt{6.433^2 - 4 \cdot 0.0095}} \quad (7)$$

we obtain in four-digit arithmetic

$$\tilde{\lambda} = 0.1477 \times 10^{-2} \quad (8)$$

As one can observe, the error in the computation can be greatly reduced by using formula (7) instead of (5).

This suggests that this particular problem can be *well-posed*:

The quality of the solution depends the method/formula chosen.

Example 2: Solution of a linear system

Consider

$$\begin{pmatrix} 3.000 & 4.127 \\ 1.000 & 1.374 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 15.41 \\ 5.147 \end{pmatrix}$$

Solving for x_2 , we obtain

$$x_2 = \frac{5.147 - \frac{15.41}{3}}{1.374 - \frac{4.127}{3}} = -6.2$$

In four-digit arithmetic, we are led to

$$\tilde{x}_2 = -3.5 \quad \text{or more correctly} \quad \tilde{x}_2 = -.35 \times 10^1$$

Example 2: Solution of a linear system

The matrix in this example is nearly singular (change one coefficient in the fourth digit (4.127 is replaced by 4.122)).

Therefore, we can not expect our solution to be very precise, no matter how we compute it.

Thus, this situation is different from the one encountered in example 1.

Basic Definitions and Results

Error definition: Motivation

- Natural choice for assessing the error for two real numbers $\alpha, \beta \in \mathbb{R}$?
- $|\alpha - \beta|$ seems to be reasonable
- $\alpha_1 = 1.234$ and $\beta_1 = 1.233$
approximation is ok
- $\alpha_2 = 0.002$ and $\beta_2 = 0.001$
 β_2 can not be accepted as a good approximation for α_2 .
- Use of $\frac{|\alpha - \beta|}{|\alpha|}$ corresponds more to our intuition.
- 0.00081 vs. 0.5

Basic Definitions and Results

Absolute and Relative Error

Let $A, B \in \mathbb{R}^{n \times n}$ with B regarded as an approximation to A .

The *residual* of B is the matrix $A - B$.

The *absolute* error $e_{abs}(B)$ in B with respect to a matrix-norm $\|\cdot\|$ is given by

$$e_{abs}(B) := \|A - B\|$$

The *relative* error $e_{rel}(B)$ in B with respect to a matrix-norm $\|\cdot\|$ is given by

$$e_{rel}(B) := \frac{\|A - B\|}{\|A\|}$$

Basic Definitions and Results

Condition and stability

Let $f : S \subset \mathbb{K}^n \rightarrow \mathbb{K}^m$ describe a mathematical problem that transfers the input $\xi \in S \subset \mathbb{K}^n$ to an output $f(\xi) \in \mathbb{K}^m$.

Assume we want to solve the problem

$$A\nu = \lambda\nu \tag{9}$$

Then, the input for f is the matrix $A \in \mathbb{K}^{r \times r}$, $n = r^2$.

The function f returns a pair $(\lambda, \nu) \in \mathbb{K} \times \mathbb{K}^r$, i.e., $m = r + 1$.

In this setup, S could correspond to the set of symmetric matrices in $\mathbb{K}^{r \times r}$.

Basic Definitions and Results

Condition and stability

Assume that \tilde{f} describes a numerical algorithm that corresponds to f and let $\tilde{\xi}$ be a perturbation of ξ .

In the example, the numerical method \tilde{f} would be any of the eigenproblem solvers, e.g., Jacobi.

Due to rounding, $\tilde{\xi} = A + E$, where $E \in \mathbb{K}^{r \times r}$ is "small", i.e., $e_{rel}(A + E)$ is small .

We then want to control the following expression

$$\frac{\|f(\xi) - \tilde{f}(\tilde{\xi})\|}{\|f(\xi)\|} \leq \frac{\|f(\xi) - f(\tilde{\xi})\|}{\|f(\xi)\|} + \frac{\|f(\tilde{\xi}) - \tilde{f}(\tilde{\xi})\|}{\|f(\xi)\|}$$

Basic Definitions and Results

Condition of a Problem

For an input $\xi \in S \subset \mathbb{K}^n$, the *relative condition* $\kappa_{rel}(f, \xi)$ of a problem is now defined as the smallest number $\kappa(f, \xi)$ that satisfies

$$\frac{\|f(\xi) - f(\xi + \varepsilon)\|}{\|f(\xi)\|} \leq \kappa(f, \xi) \frac{\|\varepsilon\|}{\|\xi\|}$$

for all $\varepsilon > 0$.

Basic Definitions and Results

Condition: an example

The solution of a linear system of equations $Ax = b$ can be formulated in the above terms.

For simplicity, consider only perturbations in the right hand side:

Let $A \in \mathbb{K}^{r \times r}$ regular. The input $\xi \in S$ is then $b \in \mathbb{K}^r$.

Obviously, $x = f(b) = A^{-1}b$ and $\tilde{x} = f(\tilde{b}) = A^{-1}\tilde{b}$.

A short computation shows, that by our definition for relative condition we have

$$\kappa_{rel}(f, \xi) = \kappa_{rel}(A) := \|A\| \|A^{-1}\| \quad (10)$$

which is independent of the input $\xi = b$.

Basic Definitions and Results

Forward stability for linear systems

An algorithm is called *normwise stable w.r.t. forward analysis*, if

$$\frac{\|f(\xi) - \tilde{f}(\xi)\|}{\|f(\xi)\|} = \mathcal{O}(\kappa_{rel}(f, \xi) \cdot \epsilon) \quad (11)$$

where ϵ is the machine-epsilon.

Note: For linear systems, $f(\xi) = x$ and $\tilde{f}(\xi) = \tilde{x}$. Thus, (11) becomes

$$\frac{\|x - \tilde{x}\|}{\|x\|} = \mathcal{O}(\kappa_{rel}(A) \cdot \epsilon)$$

In forward analysis, the idea is to explicitly track the rounding errors introduced by \tilde{f} .

As one can imagine, this is a rather tedious business.

Basic Definitions and Results

Backward Stability

An algorithm \tilde{f} is called *normwise stable w.r.t. backward analysis* if, for any ξ , it produces an output $\tilde{f}(\xi)$, such that

$$\tilde{f}(\xi) = f(\xi + \Delta\xi) \quad (12)$$

such that $\Delta\xi$ is small; the definition of "small" depends on the context.

Note: If there were no rounding errors, $\Delta\xi = 0$.

Perturbation Theory

Use of backward error estimates in the linear system case

Backward stability implies forward stability.

Loosely speaking, the following rule of thumb holds:

$$\text{forward error} \leq \text{condition number} \times \text{backward error}$$

One can show:

QR- and Jacobi-algorithm are backward stable.

For the solution of linear systems we have:

Cramer's rule for $n = 2$ and Gauss-Jordan elimination is forward stable but *not* backward stable.

Basic Definitions and Results

Backward Error Result for linear systems

Consider the problem of solving $Ax = b$.

Rigal and Gaches (1967) investigated the *normwise backward error* η :

$$\eta_{(|A|,|b|)}(y) := \min\{\delta \mid (A + \Delta A)y = b + \Delta b, \\ \|\Delta A\| \leq \delta \| |A| \|, \|\Delta b\| \leq \delta \| |b| \| \}$$

In our notation, $y = \tilde{f}(x)$, $\Delta x = (\Delta A, \Delta b)$. We want $\eta_{(|A|,|b|)}(y)$ to be small.

Their classical result then reads :

$$\eta_{(|A|,|b|)}(y) = \frac{\|b - Ay\|}{\| |A| \| \|y\| + \| |b| \|} \quad (13)$$

Perturbation Theory

Back to Eigenvalues and Eigenvectors

We want to solve the following equation numerically

$$Ax = \lambda x \quad (14)$$

with one of the methods presented in the last session.

Clearly, the input is $A \in \mathbb{R}^{n \times n}$ and the result is one eigenvector, a set of eigenvectors or eigenvectors and eigenvalues, depending on the chosen method.

Once more, due to rounding, (14) will be transformed into

$$(A + E)x = \lambda x \quad (15)$$

with $|e_{ij}| \leq \delta = \frac{1}{2}\beta^{1-p}$.

Perturbation Theory

Problems for an example: Eigenvectors

Consider the following matrices:

$$A := \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \quad \tilde{A} := A + E = \begin{pmatrix} \lambda_1 & \varepsilon \\ \varepsilon & \lambda_2 \end{pmatrix}$$

Then the eigenvectors r_{λ_i} of A and $r_{\tilde{\lambda}_i}$ of \tilde{A} can be chosen as

$$\begin{aligned} r_{\lambda_1} &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} & r_{\tilde{\lambda}_1} &= \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ r_{\lambda_2} &= \begin{pmatrix} 0 \\ 1 \end{pmatrix} & r_{\tilde{\lambda}_2} &= \begin{pmatrix} 1 \\ -1 \end{pmatrix} \end{aligned}$$

Due to this instability, one is interested in relations between the invariant subspaces corresponding to a pair of eigenvalues $(\lambda_i, \tilde{\lambda}_i)$.

Perturbation Theory

Problems for an example: Eigenvalues

Consider the following matrix:

$$A := \begin{pmatrix} \lambda_1 & b \\ 0 & \lambda_2 \end{pmatrix}$$

and the different perturbations

$$E_1 := \begin{pmatrix} 0 & 0 \\ \varepsilon & 0 \end{pmatrix} \quad E_2 := \begin{pmatrix} \varepsilon & 0 \\ 0 & 0 \end{pmatrix} \quad E_3 := \begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}$$

We have $\|E_i\|_\infty = \varepsilon$. The eigenvalues λ_j^i of $A + E_i$ are

$$\begin{aligned} \lambda_{1,2}^1 &= \frac{1}{2} \left(\lambda_1 + \lambda_2 \pm \sqrt{(\lambda_1 - \lambda_2)^2 + 4b\varepsilon} \right) \\ \lambda_1^2 &= \lambda_1 + \varepsilon, & \lambda_2^2 &= \lambda_2 \\ \lambda_1^3 &= \lambda_1 + \varepsilon, & \lambda_2^3 &= \lambda_2 + \varepsilon \end{aligned}$$

Perturbation Theory

Stability of Eigenvalues

Let $A \in \mathbb{R}^{n \times n}$. If A has a simple eigenvalue λ with corresponding left and right eigenvectors x and y , then for sufficiently small $\|\Delta A\|$ there is an eigenvalue μ of $A + \Delta A$ with

$$\mu = \lambda + \frac{y^* \Delta A x}{y^* x} + \mathcal{O}(\|\Delta A\|^2) \quad (16)$$

Consider once more

$$A := \begin{pmatrix} \lambda_1 & b \\ 0 & \lambda_2 \end{pmatrix} \quad A + \Delta A := \begin{pmatrix} \lambda_1 + \epsilon & b \\ 0 & \lambda_2 \end{pmatrix}$$

and let $\lambda_1 \neq \lambda_2$. One eigenvector can be chosen as $x = (1, 0)^T$; then $y = (1, 0)$ and (16) reads

$$\mu = \lambda_1 + \epsilon + \mathcal{O}(\epsilon^2)$$

Perturbation Theory

Stability of Eigenvalues

Another classical result from Bauer and Fike (1960) reads:

Let μ be an eigenvalue of $A + \Delta A$ and $X^{-1}AX = \text{diag}(\lambda_1, \dots, \lambda_n)$.
Then

$$\min_{\lambda \in \lambda(A)} |\lambda - \mu| \leq \kappa_p(X) \|\Delta A\|_p \quad (17)$$

We choose A and $\Delta A = E_2$ from the previous example; then

$$X = \begin{pmatrix} 1 & b \\ 0 & \lambda_2 - \lambda_1 \end{pmatrix} \implies X^{-1}AX = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

With $\lambda_1 := 2\varepsilon$, $\lambda_2 := 100 + 2\varepsilon$, $b := 50$ and $p := \infty$, (17) reads

$$\min_{\lambda \in \{2\varepsilon, 100+2\varepsilon\}} |\lambda - \mu| \leq 150\varepsilon$$

Perturbation Theory

Moral for Eigenvalues

- + Norms can provide good bounds
- + Condense the information of $m \times n$ number in a single number
- + Perturbation results can be easily interpreted and first insights can be gained.
- Norms ignore structure in the form both of scaling and sparsity
- Norms can not account for the size of perturbations distributed among the matrix' elements
- Norm bounds often lack sharpness

Perturbation Theory

Accuracy results on symmetric positive definite matrices

Weyl's inequality:

Let the eigenvalues of a perturbation E be given by

$$\nu_1 \geq \nu_2 \geq \dots \geq \nu_n$$

Then we have for the eigenvalues $\tilde{\lambda}_i$ of $A + E$

$$\tilde{\lambda}_i \in [\lambda_i + \nu_n, \lambda_i + \nu_1]$$

This can be recast into

$$\max |\tilde{\lambda}_i - \lambda_i| \leq \|E\|_2 \quad (18)$$

Perturbation Theory

Accuracy results on symmetric positive definite matrices

Assuming that $\lambda_1 \geq \dots \geq \lambda_n$ and $\|E\| \leq \eta\|A\|$, the best relative perturbation bound one can derive from (18) is:

$$\frac{|\tilde{\lambda}_i - \lambda_i|}{\lambda_i} \leq \eta \kappa_{rel,2}(A) \quad (19)$$

In case $\lambda_i \ll \lambda_1$, λ_i can undergo a large relative change.

Consider

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 10^{-12} \end{pmatrix} \quad E := \begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}$$

Then (19) reads (for $\eta \geq \varepsilon$)

$$\frac{|\tilde{\lambda}_2 - 10^{-12}|}{10^{-12}} \leq \eta \cdot 1 \cdot 10^{12} \quad \text{or} \quad |\tilde{\lambda}_2 - 10^{-12}| \leq \eta$$

Perturbation Theory

Accuracy results on symmetric positive definite matrices

By componentwise analysis, Demmel and Veselić (1992) could derive a potentially much smaller bound:

Let $A \in \mathbb{R}^{n \times n}$ be s.p.d. and write $A = DHD$, where $D = \text{diag}(A)^{\frac{1}{2}}$. Let the symmetric perturbation $\Delta A = D\Delta HD$ satisfy $\|\Delta H\|_2 = \varepsilon < \lambda_n(H)$. Then

$$\frac{|\tilde{\lambda}_i - \lambda_i|}{\lambda_i} \leq \kappa_{rel,2}(H)\varepsilon \quad (20)$$

Due to a famous result of van der Sluis, $\kappa_{rel,2}(H) \ll \kappa_{rel,2}(A)$ is possible.

Perturbation Theory

Accuracy results on symmetric positive definite matrices

Consider again

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 10^{-12} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 10^{-6} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 10^{-6} \end{pmatrix}$$

$$E := \begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 10^{-6} \end{pmatrix} \begin{pmatrix} \varepsilon & 0 \\ 0 & 10^{12}\varepsilon \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 10^{-6} \end{pmatrix}$$

The bound (20) reads then

$$\frac{|\tilde{\lambda}_2 - 10^{-12}|}{10^{-12}} \leq 1 \cdot \varepsilon \quad \text{or} \quad |\tilde{\lambda}_2 - 10^{-12}| \leq 10^{-12}\varepsilon$$

where $10^{12}\varepsilon < 1$ is required. For $\varepsilon = 10^{-16}$ we have

$$|\tilde{\lambda}_2 - 10^{-12}| \leq 10^{-16} \quad |\tilde{\lambda}_2 - 10^{-12}| \leq 10^{-28}$$

References

- [1] *What Every Computer Scientist Should Know About Floating-Point Arithmetic*, David Goldberg, Computing Surveys, March 1991
- [2] *Accuracy and Stability of Numerical Algorithms*, Nicholas J. Higham, SIAM, 1996
- [3] *Introduction to Matrix Computations*, G. W. Stewart, Academic Press, New York, San Francisco, London, 1973
- [4] *On the computability of a given solution with the data of a linear system*, J. L. Rigal and J. Gaches, J. Assoc. Comput. Mach. 14, 1967, no. 3, pp. 543-548

References

- [5] *Matrix perturbation theory*, G. W. Stewart and Ji-guang Sun, Academic Press, London, 1990
- [6] *The rotation of eigenvectors by a perturbation. III*, Chandler Davis and W.M. Kahan, SIAM J. Numer. Anal. 7 (1970), no. 1, 1-46
- [7] *A survey of componentwise perturbation theory in numerical linear algebra*, Nicholas J. Higham, in Proceedings of Symposia in Applied Mathematics, vol. 48, pp. 49-77, 1994
- [8] *Jacobi's method is more accurate than QR*, J. Demmel and K. Veselić, SIAM J. Matrix Anal. Appl. 13, pp. 1204-1245, 1992