Condition and Scaling

1. Condition Numbers
   - measuring the sensitivity of a problem
2. Chemical Equilibria
   - equations determined by chemical reactions
3. Equation and Variable Scaling
   - formulating an optimization problem
   - preconditioning and postconditioning

MCS 563 Lecture 9
Analytic Symbolic Computation
Jan Verschelde, 31 January 2011

Condition and Scaling

1. Condition Numbers
   - measuring the sensitivity of a problem
2. Chemical Equilibria
   - equations determined by chemical reactions
3. Equation and Variable Scaling
   - formulating an optimization problem
   - preconditioning and postconditioning

type of errors

Consider a linear system $A\mathbf{x} = \mathbf{b}$ for $\det(A) \neq 0$.

We compute $\overline{\mathbf{x}}$, an approximation for $\mathbf{x}$.

We distinguish two types of errors:

1. *backward* error: $r = \mathbf{b} - A\overline{\mathbf{x}}$ residual
   $\Rightarrow A\overline{\mathbf{x}} = \mathbf{b} - r$
   The residual is what we should subtract from $\mathbf{b}$ for $\overline{x}$ to be an exact solution.

2. *forward* error: $\mathbf{x} = \overline{\mathbf{x}} + \Delta\mathbf{x}$, correction term
   $\Rightarrow A(\overline{\mathbf{x}} + \Delta\mathbf{x}) = \mathbf{b} \Rightarrow \overline{\mathbf{x}} + \Delta\mathbf{x} = A^{-1}\mathbf{b} \Rightarrow \Delta\mathbf{x} = A^{-1}\mathbf{b} - \overline{\mathbf{x}}$
   The correction term is what should be added to $\overline{\mathbf{x}}$ to obtain an exact solution to $A\mathbf{x} = \mathbf{b}$.

Terminology refers to the solution map A^{-1} in this case from input (A, b) to output $\mathbf{x}$.

derivation of a condition number

Given $A\mathbf{x} = \mathbf{b}$, we solve $\overline{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ with

$$\overline{A} = A + \Delta A, \quad \overline{\mathbf{x}} = \mathbf{x} + \Delta \mathbf{x}, \quad \overline{\mathbf{b}} = \mathbf{b} + \Delta \mathbf{b}.$$

To find a bound for the error $\Delta \mathbf{x}$ on $\mathbf{x}$, we do

$$\begin{array}{rcl} (A + \Delta A)(\mathbf{x} + \Delta \mathbf{x}) & = & \mathbf{b} + \Delta \mathbf{b} \\ -[\qquad A\mathbf{x} & = & \mathbf{b} \qquad] \\ \hline A\Delta \mathbf{x} + \Delta A(\mathbf{x} + \Delta \mathbf{x}) & = & \Delta \mathbf{b} \end{array}$$

Assuming $\det(A) \neq 0$: $\Delta \mathbf{x} = A^{-1}(-\Delta A \overline{\mathbf{x}} + \Delta \mathbf{b})$.

$$||\Delta \mathbf{x}|| \leq ||A^{-1}|| \, (||\Delta A|| \cdot ||\overline{\mathbf{x}}|| + ||\Delta \mathbf{b}||)$$

$$\frac{||\Delta \mathbf{x}||}{||\overline{\mathbf{x}}||} \leq ||A^{-1}|| \cdot ||A|| \left(\frac{||\Delta A||}{||A||} + \frac{||\Delta \mathbf{b}||}{||A|| \cdot ||\overline{\mathbf{x}}||} \right)$$

sensitivity analysis

Consider a linear system $A\mathbf{x} = \mathbf{b}$ for $\det(A) \neq 0$.

The relative error $||\Delta\mathbf{x}||/||\mathbf{x}||$ for $\mathbf{x} = A^{-1}\mathbf{b}$ is bounded by

$$\frac{||\Delta\mathbf{x}||}{||\mathbf{x}||} \leq ||A|| \cdot ||A^{-1}|| \frac{||\Delta A||}{||A||}.$$

This inequality bounds the relative error on the solution by the relative error on the coefficient matrix A and the number $||A|| \cdot ||A^{-1}|| = \mathrm{cond}(A)$, the condition number of A.

If $\mathrm{cond}(A)$ is large, then the linear system is ill conditioned and no matter what algorithm we use to solve it, small errors in the calculations will amplify and cause large errors in $\mathbf{x}$.

$\det(A) = 0 \Leftrightarrow \mathrm{cond}(A) = \infty$ and $A\mathbf{x} = \mathbf{b}$ is ill posed.

condition of an isolated root

Let $f(\mathbf{x}) = \mathbf{0}$ be a system of n equations in n unknowns.

Denote the Jacobian matrix of f by J_f
and let $\mathbf{z} \in \mathbb{C}^n$ be an isolated solution of $f(\mathbf{x}) = \mathbf{0}$.

Then, *the relative condition number of the zero* $\mathbf{z}$ *as a solution of* $f(\mathbf{x}) = \mathbf{0}$ is

$$\kappa(f, \mathbf{z}) = ||J_f(\mathbf{z})||_2 ||J_f^{-1}(\mathbf{z})||_2,$$

i.e.: $\kappa(f, \mathbf{z})$ is the condition number of the Jacobian matrix of the polynomials in the system evaluated at $\mathbf{z}$.

Three numbers measure the quality of an approximate solution $\mathbf{z}$: the residual $||f(\mathbf{z})||$, the correction term $||\Delta \mathbf{z}||$ (computed via Newton's method) and an estimate for $\kappa(f, \mathbf{z})$.

Singular Value Decomposition

The singular value decomposition (SVD) of $A \in \mathbb{C}^{n \times n}$ is

$$A = U \Sigma V^H, \quad U^H U = I, V^H V = I,$$

and with singular values σ_i, $i = 1, 2, \ldots, n$:

$$\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n), \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n.$$

Two applications:

1. $||A||_2 = \sigma_1$, for $\det(A) \neq 0$: $||A^{-1}||_2 = \sigma_n^{-1}$

$$\Rightarrow \mathrm{cond}(A) = ||A||_2 \cdot ||A^{-1}||_2 = \frac{\sigma_1}{\sigma_n}.$$

2. if $\sigma_{R+1} < \epsilon$, then $R = \mathrm{Rank}(A, \epsilon)$ and

$$\widehat{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_R, 0, \ldots, 0), \quad \widehat{A} = U \widehat{\Sigma} V^H$$

$\widehat{A}$ is the projection of A onto the space of rank R matrices.

SVD in Octave

```
octave-3.2.3:1> A = rand(2,2) + rand(2,2)*i
A =
   0.10661 + 0.59460i   0.38331 + 0.99751i
   0.06697 + 0.16692i   0.68975 + 0.87405i
octave-3.2.3:2> [U,S,V] = svd(A)
U =
  -0.11645 - 0.72903i   0.14737 + 0.65821i
  -0.29430 - 0.60692i  -0.36257 - 0.64311i
S =
Diagonal Matrix
   1.64097         0
         0   0.29353
V =
  -0.34548 - 0.00000i   0.93843 - 0.00000i
  -0.91733 - 0.19786i  -0.33771 - 0.07284i
octave-3.2.3:3> U*S*V'
```

Condition and Scaling

chemical reactions

Variables in chemical reaction and conservation equations are molar concentrations of the species in the reaction.

For the equilibrium balance between water H_2O, hydrogen H and oxygen O, we derive:

$$H_2O \leftrightharpoons 2H + O$$

total amount T_H of H is conserved
total amount T_O of O is conserved

which leads to the system

$$f(x_H, x_O, x_{H_2O}) = \left\{ \begin{array}{r} k\, x_{H_2O} = x_H^2 x_O \\ 2x_{H_2O} + x_H = T_H \\ x_{H_2O} + x_O = T_O \end{array} \right.$$

where k is a dimensionless stoichiometric constant.

a larger reaction

Involving 11 species, 4 components: O atomic oxygen, H atomic hydrogen, CO carbon monoxide, N atomic nitrogen; and 7 compounds: O_2 molecular oxygen, H_2 molecular hydrogen, N_2 molecular nitrogen, CO_2 carbon dioxide, OH hydroxyl radical, H_2O water vapor, NO nitric oxide.

The reaction equations are

$$
\begin{array}{ll}
O_2 \leftrightharpoons 2O & k_1 x_{O_2} = x_O^2 \\
H_2 \leftrightharpoons 2H & k_2 x_{H_2} = x_H^2 \\
N_2 \leftrightharpoons 2N & k_3 x_{N_2} = x_N^2 \\
CO_2 \leftrightharpoons O + CO & k_4 x_{CO_2} = x_O x_{CO} \\
OH \leftrightharpoons O + H & k_5 x_{OH} = x_O x_H \\
H_2O \leftrightharpoons O + 2H & k_6 x_{H_2O} = x_O x_H^2 \\
NO \leftrightharpoons O + N & k_7 x_{NO} = x_N x_O
\end{array}
$$

conservation equations

and the four equations conserving the total amounts of the four components:

$$
\begin{aligned}
T_H &= x_H + 2x_{H_2} + x_{OH} + 2x_{H_2O} \\
T_C &= x_{CO} + x_{CO_2} \\
T_O &= x_O + x_{CO} + 2x_{O_2} + 2x_{CO_2} + x_{OH} + x_{H_2O} + x_{NO} \\
T_N &= x_N + 2x_{N_2} + x_{NO}.
\end{aligned}
$$

So we end up with a total of 11 equations in 11 unknowns.

constants

The constants for various temperatures:

constants	$T = 1000^o$	$T = 2000^o$	$T = 3000^o$
$\log_{10}(1/k_1)$	24.528	7.289	3.108
$\log_{10}(1/k_2)$	22.206	6.997	3.270
$\log_{10}(1/k_3)$	47.970	15.107	6.942
$\log_{10}(1/k_4)$	24.942	6.825	2.559
$\log_{10}(1/k_5)$	22.120	7.208	3.541
$\log_{10}(1/k_6)$	46.989	14.680	6.791
$\log_{10}(1/k_7)$	32.187	10.285	4.878

with totals T_O = 5.0E–5, T_H = 3.0E–5, T_C = 1.0E–5, and T_N = 1.0E–5.

Approximate coefficients of varying magnitudes...

Condition and Scaling

equation and variable scaling

Equation scaling: multiply equation with constant, e.g.:

$$f(x) = 10^{20}x^2 + 4 \cdot 10^{20}x + 2 \cdot 10^{20} = 0.$$

Variable scaling is needed in

$$f(x) = 10^{-20}x^2 + 4 \cdot x + 2 \cdot 10^{20} = 0.$$

The change of variables $x = 10^{20}z$ turns the second equation into the first equation.

combined scaling

A combined equation and variable scaling method aims

1. to center the coefficients around unity and
2. to reduce the variability of the coefficients.

For $f(x) = 10^{-20}x^2 + 4 \cdot x + 2 \cdot 10^{20} = 0$ let c_1 and c_2:

$$10^{c_2} f(z = 10^{c_1} x) = 10^{e_1} z^2 + 10^{e_2} z + 10^{e_3}.$$

The two objectives are met by minimizing $r(c_1, c_2) = r_1(c_1, c_2) + r_2(c_1, c_2)$ where

$$\begin{cases} r_1(c_1, c_2) = e_1^2 + e_2^2 + e_3^2 \\ r_2(c_1, c_2) = (e_1 - e_2)^2 + (e_1 - e_3)^2 + (e_2 - e_3)^2. \end{cases}$$

Since $r(c_1, c_2)$ is a quadratic and has no maximum, the minimum must occur at the solution of $\frac{\partial r}{\partial c_1} = 0$ and $\frac{\partial r}{\partial c_2} = 0$.

scaling multivariate polynomials

$$f(\mathbf{x}) = \sum_{\mathbf{a} \in A} c_{\mathbf{a}} \mathbf{x}^{\mathbf{a}}, \quad c_{\mathbf{a}} \in \mathbb{C}, \quad \mathbf{x}^{\mathbf{a}} = x_1^{a_1} x^{a_2} \cdots x^{a_n}.$$

We scale $f(\mathbf{x}) = 0$ by multiplication with b^{γ_0} and substitute the variables x_i by $b^{\gamma_i} y_i$, for $i = 1, 2, \ldots, n$.

Then, $b^{\gamma_0} f(x_1 = b^{\gamma_1} y_1, x_2 = b^{\gamma_2} y_2, \ldots x_n = b^{\gamma_n} y_n)$

$$= \sum_{\mathbf{a} \in A} c_{\mathbf{a}} b^{\gamma_0 + \gamma_1 a_1 + \gamma_2 a_2 + \cdots + \gamma_n a_n} y^{\mathbf{a}}$$

$$= \sum_{\mathbf{a} \in A} b^{\log_b(c_{\mathbf{a}}) + \gamma_0 + \langle \boldsymbol{\gamma}, \mathbf{a} \rangle} y^{\mathbf{a}}.$$

where $\langle \boldsymbol{\gamma}, \mathbf{a} \rangle = \gamma_1 a_1 + \gamma a_2 + \cdots + \gamma_n a_n$,
for $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_n)$.

objectives to minimize

We derive conditions on the scaling constants γ_0 and γ.

The distance of the coefficients from one is expressed by r_1:

$$r_1(\gamma_0, \gamma) = \sum_{\mathbf{a} \in A} \left(\log_b |c_{\mathbf{a}}| + \gamma_0 + \langle \gamma, \mathbf{a} \rangle \right)^2 .$$

The variability between the coefficients is expressed by r_2:

$$\begin{aligned} r_2(\gamma_0, \gamma) &= \sum_{\mathbf{a}_1 \in A} \sum_{\mathbf{a}_2 \in A \setminus \{\mathbf{a}_1\}} \left(\log_b |c_{\mathbf{a}_1}| + \langle \gamma, \mathbf{a}_1 \rangle \right. \\ &\qquad \left. - \log_b |c_{\mathbf{a}_2}| - \langle \gamma, \mathbf{a}_2 \rangle \right)^2 . \end{aligned}$$

Minimizing $r_1(\gamma_0, \gamma) + r_2(\gamma_0, \gamma)$ is a least squares problem.

a least squares problem

We formulate the problem for a system $f(\mathbf{x}) = \mathbf{0}$ of n equations $f = (f_1, f_2, \ldots, f_n)$ supported on $(A_1, A_2, \ldots, A_n)$.

Instead of one γ_0, we have n constants γ_{i0}, $i = 1, 2, \ldots, n$.

$$\left\{ \begin{array}{rcl} \langle \gamma, \mathbf{a} \rangle + \gamma_{i0} & = & -\log_b |c_{\mathbf{a}}|, \\ & & \mathbf{a} \in A_i, i = 1, 2, \ldots, n \\ \langle \gamma, \mathbf{a}_1 \rangle - \langle \gamma, \mathbf{a}_2 \rangle & = & -\log_b |c_{\mathbf{a}_1}| + \log_b |c_{\mathbf{a}_2}|, \\ & & \mathbf{a}_1 \in A_i, \mathbf{a}_2 \in A_i \setminus \{\mathbf{a}_1\}, i = 1, 2, \ldots, n \end{array} \right.$$

We have $2n$ unknowns, while the number of equations is determined by the number of monomials.

The fewer monomials, the better scaling will work.

Condition and Scaling

preconditioning and postconditioning

Preconditioning = reformulating the given problem into a problem with better numerical condition number.

Postconditioning = interpretation of a solution of the reformulated problem in the original coordinates.

The variable scaling

$$x_i = y_i b^{\gamma_i}, \quad i = 1, 2, \ldots, n$$

could be seen as a floating-point representation where y_i is the fraction and γ_i the exponent.

Summary + Exercises

Floating-point numbers have a fraction and an exponent. By scaling we find a more suitable coordinate system to represent and compute solutions of polynomial systems.

Exercises:

1. Consider the matrix

$$A = \begin{bmatrix} +1 & -1 & -1 & -1 \\ 0 & +1 & -1 & -1 \\ 0 & 0 & +1 & -1 \\ 0 & 0 & 0 & +1 \end{bmatrix}.$$

Compute the condition number with MATLAB or Octave and make a table for larger versions of A, up to $n = 10$ at least. Choose a random right hand side vector $\mathbf{b}$ and solve $A\mathbf{x} = \mathbf{b}$. Do the solutions $\mathbf{x}$ grow in size as the dimension n grows? Choose a random solution vector $\mathbf{x}$ and compute $\mathbf{b} = A\mathbf{x}$. Solve $A\mathbf{x} = \mathbf{b}$. Compare your finding with the previous experiments.

more exercises

2. Consider $f(x) = 10^{-20}x^2 + 4 \cdot x + 2 \cdot 10^{20} = 0$. For a zero z, compute $\gamma(f,z)$ (see Lecture 5) and the corresponding bound on the radius of the disc of guaranteed quadratic convergence of Newton's method. Scale the coefficients of f and do the same computations of γ and the radius on a zero of the scaled polynomial. Compare and interpret the results.
3. Use Gröbner bases to solve the reaction between H_2O, hydrogen H and oxygen O symbolically.
4. Make a Maple worksheet or Sage notebook to model the chemical reaction involving 11 species, using the constants in the table. With this worksheet or notebook you can then generate three different instances of the same problem. Use `phc` to solve these instances. Scaling is available via `phc -s`.

one last exercise

5. Consider the intersection of two circles:

$$f(x_1, x_2, c) = \begin{cases} x^2 + y^2 - 1 = 0 \\ (x - c)^2 + y^2 - 1 = 0. \end{cases}$$

Examine the condition number of one of the solutions of $f(x_1, x_2, c) = \mathbf{0}$ as c goes from 0 to 2.
Is the condition number a continuous function of c?

Homework collection on Monday 14 February at 10AM:
exercises 3, 5, 7 of Lecture 4; exercises 2, 3, 4 of Lecture 5; exercises 3, 4, 5 of Lecture 6; exercises 3, 4, 6, 9 of Lecture 7; and exercises 3, 5, 6 of Lecture 8.