MATH 3795
Lecture 6. Sensitivity of the Solution of a Linear System

Dmitriy Leykekhman

Fall 2008

Goals

- Understand how does the solution of $Ax = b$ changes when A or b change.
- Condition number of a matrix (with respect to inversion).
- Vector and matrix norms.

Linear Systems

- Given $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$ we are interested in the solution $x \in \mathbb{R}^n$ of
$$Ax = b.$$
- Suppose that instead of A, and b we are given $A + \Delta A$ and $b + \Delta b$, where $\Delta A \in \mathbb{R}^{n \times n}$ and $\Delta b \in \mathbb{R}^n$. How do these perturbations in the data change the solution of the linear system?
- First we need to understand how to measure the size of vectors and of matrices. This leads to vector norms and matrix norms.

Vector Norms

Definition

A (vector) norm on $\mathbb{R}^n$ is a function

$$\begin{aligned} \|\cdot\|:\ \mathbb{R}^n &\rightarrow \mathbb{R} \\ x &\rightarrow \|x\| \end{aligned}$$

which for all $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ satisfies

1. $\|x\| \geq 0$, $\|x\| = 0 \ \Leftrightarrow x = 0$,
2. $\|\alpha x\| = |\alpha|\|x\|$,
3. $\|x + y\| \leq \|x\| + \|y\|$, (triangle inequality).

Vector Norms

The most frequently used norms on $\mathbb{R}^n$ are given by

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}, \quad \text{2-norm}$$

The MATLAB's build in function $norm(x)$ or $norm(x, 2)$.
More generally for any $p \in [1, \infty)$

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad \text{p-norm.}$$

The MATLAB's build in function $norm(x, p)$ and

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|, \quad \infty\text{-norm.}$$

The MATLAB's build in function $norm(x, inf)$

Vector Norms

Example

Let $x = (1, -2, 3, -4)^T$. Then

$$\begin{aligned} \|x\|_1 &= 1 + 2 + 3 + 4 = 10, \\ \|x\|_2 &= \sqrt{1 + 4 + 9 + 16} = \sqrt{30} \approx 5.48, \\ \|x\|_\infty &= \max\{1, 2, 3, 4\} = 4. \end{aligned}$$

Vector Norms

The boundaries of the unit balls defined by

$$\{x \in \mathbb{R}^n : \|x\|_p \leq 1\}.$$

One can show the following useful inequalities:

Vector Norms

The boundaries of the unit balls defined by

$$\{x \in \mathbb{R}^n : \|x\|_p \leq 1\}.$$

One can show the following useful inequalities:

- $$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1.$$

Vector Norms

The boundaries of the unit balls defined by

$$\{x \in \mathbb{R}^n : \|x\|_p \leq 1\}.$$

One can show the following useful inequalities:

- $$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1.$$
- Let $\|\cdot\|$ is any vector norm on $\mathbb{R}^n$, then

$$\|x+y\| \geq \big|\|x\| - \|y\|\big| \quad \text{for all} \quad x, y \in \mathbb{R}^n.$$

Vector Norms

The boundaries of the unit balls defined by

$$\{x \in \mathbb{R}^n : \|x\|_p \leq 1\}.$$

One can show the following useful inequalities:

- $$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1.$$
- Let $\|\cdot\|$ is any vector norm on $\mathbb{R}^n$, then
  $$\|x+y\| \geq \big|\|x\| - \|y\|\big| \quad \text{for all} \quad x, y \in \mathbb{R}^n.$$
- Cauchy-Schwarz inequality,
  $$x^T y \leq \|x\|_2 \|y\|_2 \quad \text{for all} \quad x, y \in \mathbb{R}^n.$$

Vector Norms

Theorem

Vector norms on $\mathbb{R}^n$ are equivalent, i.e. for every two vector norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on $\mathbb{R}^n$ there exist constants c_{ab}, C_{ab} (depending on the vector norms $\|\cdot\|_a$ and $\|\cdot\|_b$, but not on x) such that

$$c_{ab}\|x\|_b \leq \|x\|_a \leq C_{ab}\|x\|_b \quad \forall x \in \mathbb{R}^n.$$

In particular, for any $x \in \mathbb{R}^n$ we have the inequalities

$$\begin{aligned}
\frac{1}{\sqrt{n}}\|x\|_1 &\leq \|x\|_2 \leq \|x\|_1 \\
\|x\|_\infty &\leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty \\
\|x\|_\infty &\leq \|x\|_1 \leq n\|x\|_\infty.
\end{aligned}$$

Matrix Norms

Definition

A matrix norm on $\mathbb{R}^{m\times n}$ is a function

$$\begin{aligned} \|\cdot\|: \ \mathbb{R}^{m\times n} &\to \mathbb{R} \\ A &\to \|A\|, \end{aligned}$$

which for all $A, B \in \mathbb{R}^{m\times n}$ and $\alpha \in \mathbb{R}$ satisfies

1. $\|A\| \geq 0$, $\|A\| = 0 \ \Leftrightarrow A = 0$ (zero matrix),
2. $\|\alpha A\| = |\alpha| \|A\|$,
3. $\|A + B\| \leq \|A\| + \|B\|$, (triangle inequality).

Warning: Matrix- and vector-norms are denoted by the same symbol $\|\cdot\|$. However, as we will see shortly, vector-norms and matrix-norms are computed very differently. Thus, before computing a norm we need to examine carefully whether it is applied to a vector or to a matrix. It should be clear from the context which norm, a vector-norm or a matrix-norm, is used.

Matrix Norms. First Approach.

- View a matrix $A \in \mathbb{R}^{m \times n}$ as a vector in $\mathbb{R}^{mn}$, by stacking the columns of the matrix into a long vector.

Matrix Norms. First Approach.

- View a matrix $A \in \mathbb{R}^{m \times n}$ as a vector in $\mathbb{R}^{mn}$, by stacking the columns of the matrix into a long vector.
- Apply the vector-norms to this vectors of length mn.

Matrix Norms. First Approach.

- View a matrix $A \in \mathbb{R}^{m \times n}$ as a vector in $\mathbb{R}^{mn}$, by stacking the columns of the matrix into a long vector.
- Apply the vector-norms to this vectors of length mn.
- This will give matrix norms. For example if we apply the 2-vector-norm, then

$$\|A\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2 \right)^{1/2}.$$

This is called the **Frobenius norm**.
(We will use $\|A\|_2$ to denote a different matrix norm.)

Matrix Norms. First Approach.

- View a matrix $A \in \mathbb{R}^{m \times n}$ as a vector in $\mathbb{R}^{mn}$, by stacking the columns of the matrix into a long vector.
- Apply the vector-norms to this vectors of length mn.
- This will give matrix norms. For example if we apply the 2-vector-norm, then

$$\|A\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2 \right)^{1/2}.$$

  This is called the **Frobenius norm**.
  (We will use $\|A\|_2$ to denote a different matrix norm.)
- This approach is not very useful.

Matrix Norms. Second Approach.

- We want to solve linear systems $Ax = b$.
  Find a vector x such that if we multiply A by this vector (we apply A to this vector), then we obtain b.

Matrix Norms. Second Approach.

- We want to solve linear systems $Ax = b$.
  Find a vector x such that if we multiply A by this vector (we apply A to this vector), then we obtain b.
- View a matrix $A \in \mathbb{R}^{m\times n}$ as a linear mapping, which maps a vector $x \in \mathbb{R}^n$ into a vector $Ax \in \mathbb{R}^m$

$$\begin{aligned} A: \ \mathbb{R}^n &\longrightarrow \mathbb{R}^m \\ x &\longrightarrow Ax. \end{aligned}$$

Matrix Norms. Second Approach.

- We want to solve linear systems $Ax = b$.
  Find a vector x such that if we multiply A by this vector (we apply A to this vector), then we obtain b.
- View a matrix $A \in \mathbb{R}^{m \times n}$ as a linear mapping, which maps a vector $x \in \mathbb{R}^n$ into a vector $Ax \in \mathbb{R}^m$

$$\begin{aligned} A: \ \mathbb{R}^n &\rightarrow \mathbb{R}^m \\ x &\rightarrow Ax. \end{aligned}$$

- How do we define the size of a linear mapping?

Matrix Norms. Second Approach.

- We want to solve linear systems $Ax = b$.
  Find a vector x such that if we multiply A by this vector (we apply A to this vector), then we obtain b.
- View a matrix $A \in \mathbb{R}^{m \times n}$ as a linear mapping, which maps a vector $x \in \mathbb{R}^n$ into a vector $Ax \in \mathbb{R}^m$

$$\begin{aligned} A: \ \mathbb{R}^n &\to \mathbb{R}^m \\ x &\to Ax. \end{aligned}$$

- How do we define the size of a linear mapping?
- Compare the size of the image $Ax \in \mathbb{R}^m$ with the size of x. This leads us to look at

$$\sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

  Here $Ax \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$ are vectors and $\| \cdot \|$ are vector norms (in $\mathbb{R}^m$ and $\mathbb{R}^n$).

Matrix Norms

- Let $p \in [1, \infty]$. The following identities re valid

$$\sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \sup_{\|x\|_p = 1} \|Ax\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p = 1} \|Ax\|_p$$

Matrix Norms

- Let $p \in [1, \infty]$. The following identities re valid

$$\sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \sup_{\|x\|_p = 1} \|Ax\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p = 1} \|Ax\|_p$$

- One can show

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}. \tag{1}$$

Note that on the left hand side in (1) the symbol $\|\cdot\|_p$ refers to the p-matrix-norm, while on the right hand side in (1) the symbol $\|\cdot\|_p$ refers to the p-vector-norm applied to the vectors $Ax \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$, respectively.

Matrix Norms

For the most commonly used matrix-norms (1) with $p = 1$, $p = 2$, or $p = \infty$, there exist rather simple representations.

Let $\|\cdot\|_p$ be the matrix norm defined in (1), then

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| \quad \text{(maximum column norm)};$$

$$\|A\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}| \quad \text{(maximum row norm)};$$

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

where $\lambda_{max}(A^T A)$ is the largest eigenvalue of $A^T A$.

Matrix Norms

Example

Let

$$A = \begin{pmatrix} 1 & 3 & -6 \\ -2 & 4 & 2 \\ 2 & 1 & -1 \end{pmatrix}.$$

Then

$$\begin{aligned} \|A\|_1 &= \max 5, 8, 9 = 9, \\ \|A\|_\infty &= \max 10, 8, 4 = 10, \\ \|A\|_2 &= \sqrt{\max\{3.07, 23.86, 49.06\}} \approx 7.0045, \\ \|x\|_F &= \sqrt{76} \approx 8.718. \end{aligned}$$

Matrix Norms

Two important inequalities.

Theorem

For any $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times k}$ and $x \in \mathbb{R}^n$, the following inequalities hold.

$$\|Ax\|_p \le \|A\|_p \|x\|_p \quad \textit{(compatibility of matrix and vector norm)}$$

and

$$\|AB\|_p \le \|A\|_p \|B\|_p \quad \textit{(submultiplicativity of matrix norms)}$$

Note that for the identity matrix I,

$$\|I\|_p = \max_{x\neq 0} \frac{\|Ix\|_p}{\|x\|_p} = 1.$$

Compare this with the first approach in which we view I as a vector of length n^2. For example the Frobenius norm (2-vector norm) is

$$\|I\|_F = \sqrt{n}.$$

Error Analysis

- Let
$$Ax = b \tag{2}$$
be the original system, where $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$.

Error Analysis

- Let

$$Ax = b \tag{2}$$

be the original system, where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$.

- Let

$$(A + \Delta A)\tilde{x} = b + \Delta b \tag{3}$$

be the perturbed system, where $\Delta A \in \mathbb{R}^{n\times n}$ and $\Delta b \in \mathbb{R}^n$ represent the perturbations in A and b, respectively.

Error Analysis

- Let
$$Ax = b \tag{2}$$
be the original system, where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$.
- Let
$$(A + \Delta A)\tilde{x} = b + \Delta b \tag{3}$$
be the perturbed system, where $\Delta A \in \mathbb{R}^{n\times n}$ and $\Delta b \in \mathbb{R}^n$ represent the perturbations in A and b, respectively.
- What is the error $\Delta x = \tilde{x} - x$ between the solution x of the exact linear system (7) and the solution ex perturbed linear system (8).
- Use a representation
$$\tilde{x} = x + \Delta x.$$

Error Analysis. Perturbation in b only

The original linear system,

$$Ax = b,$$

where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. The perturbed linear system

$$A(x + \Delta x) = b + \Delta b,$$

where $\Delta b \in \mathbb{R}^n$ represents the perturbations in b.
Subtracting we get

$$A\Delta x = \Delta b, \quad \text{or} \quad \Delta x = A^{-1}\Delta b.$$

Error Analysis. Perturbation in b only

The original linear system,

$$Ax = b,$$

where $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$. The perturbed linear system

$$A(x + \Delta x) = b + \Delta b,$$

where $\Delta b \in \mathbb{R}^n$ represents the perturbations in b.
Subtracting we get

$$A\Delta x = \Delta b, \quad \text{or} \quad \Delta x = A^{-1}\Delta b.$$

Take norms:

$$\|\Delta x\| = \|A^{-1}\Delta b\| \leq \|A^{-1}\|\|\Delta b\|. \tag{4}$$

To estimate relative error, note that $Ax = b$ and as a result

$$\|b\| = \|Ax\| \leq \|A\|\|x\| \quad \Rightarrow \quad \frac{1}{\|x\|} \leq \|A\|\frac{1}{\|b\|}. \tag{5}$$

Combining (4) and (5) we get

$$\frac{\|\Delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\Delta b\|}{\|b\|}. \tag{6}$$

Error Analysis. Perturbation in b only

Definition

The (p-) condition number $\kappa_p(A)$ of a matrix A (with respect to inversion) is defined by

$$\kappa_p(A) = \|A\|_p \|A^{-1}\|_p.$$

Set $\kappa_p(A) = \infty$ is A is not invertible. MATLAB's build in function $cond(A)$.

If

$$Ax = b,$$

and

$$A(x + \Delta x) = b + \Delta b,$$

then the relative error between the solutions obeys

$$\frac{\|\Delta x\|}{\|x\|} \leq \kappa_p(A) \frac{\|\Delta b\|}{\|b\|}.$$

Error Analysis. General Case.

- Let
$$Ax = b \tag{7}$$
be the original system, where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$.
- Let
$$(A + \Delta A)(\Delta x + x) = b + \Delta b \tag{8}$$
be the perturbed system, where $\Delta A \in \mathbb{R}^{n\times n}$ and $\Delta b \in \mathbb{R}^n$ represent the perturbations in A and b, respectively.
- If $\|A^{-1}\|_p \|\Delta A\|_p < 1$, then
$$\frac{\|\Delta x\|_p}{\|x\|_p} \leq \frac{\kappa_p(A)}{1 - \kappa_p(A)\frac{\|\Delta A\|_p}{\|A\|_p}} \left(\frac{\|\Delta A\|_p}{\|A\|_p} + \frac{\|\Delta b\|}{\|b\|} \right). \tag{9}$$

If $\kappa_p(A)$ is small, we say that the linear system is **well conditioned.** Otherwise, we say that the linear system is **ill conditioned.**

Error Analysis. Example. Hilbert Matrix

Example

Hilbert Matrix $H \in \mathbb{R}^{n\times n}$ with entries

$$h_{ij} = \int_0^1 x^{i+j-2}\, dx = \frac{1}{i+j-1}.$$

For $n = 4$,

$$H = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{pmatrix}.$$

$$H^{-1} = \begin{pmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{pmatrix}.$$

Error Analysis. Example. Hilbert Matrix.

Example

We compute that the condition number of a Hilbert matrix grows very fast with n. For $n = 4$

$$\|H\|_1 = \frac{25}{12} \quad \|H^{-1}\|_1 = 13620, \quad \kappa_1(H) = 28375,$$

$$\|H\|_\infty = \frac{25}{12} \quad \|H^{-1}\|_\infty = 13620, \quad \kappa_\infty(H) = 28375,$$

$$\|H\|_2 \approx 1.5 \quad \|H^{-1}\|_2 \approx 1.03 * 10^4, \quad \kappa_2(H) \approx 1.55 * 10^4.$$

Error Analysis. Example. Hilbert Matrix.

Example

We consider the linear systems

$$Hx = b.$$

For given n we set $x_{ex} = (1, \dots, 1)^T \in \mathbb{R}^n$, and compute $b = Hx_{ex}$. Then we compute the solution of the linear system $Hx = b$ using the LU-decomposition and compute the relative error between exact solution x_{ex} and computed solution x.

n	$\kappa_\infty(H)$	$\frac{\|x_{ex}-x\|_\infty}{\|x_{ex}\|_\infty}$
4	$2.837500e+004$	$2.958744e-013$
5	$9.436560e+005$	$5.129452e-012$
6	$2.907028e+007$	$5.096734e-011$
7	$9.851949e+008$	$2.214796e-008$
8	$3.387279e+010$	$1.973904e-007$
9	$1.099651e+012$	$4.215144e-005$
10	$3.535372e+013$	$5.382182e-004$

Error Analysis.

- If we use finite precision arithmetic, then rounding causes errors in the input data. Using m-digit floating point arithmetic it holds that

$$\frac{|x - fl(x)|}{|x|} \leq 0.5 * 10^{-m+1}.$$

- Thus, if we solve the linear system in m-digit floating point arithmetic, then, as rule of thumb, we may approximate the the input errors due to rounding by

$$\frac{\|\Delta A\|}{\|A\|} \approx 0.5 * 10^{-m+1}, \quad \frac{\|\Delta b\|}{\|b\|} \approx 0.5 * 10^{-m+1}$$

- If the condition number of A is $\kappa(A) = 10^{\alpha}$, then

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{10^{\alpha}}{1 - 10^{\alpha-m+1}}(0.5 * 10^{-m} + 0.5 * 10^{-m}) \approx 10^{\alpha-m}.$$

  Provided $10^{\alpha-m+1} < 1$.
- **Rule of thumb:** If the linear system is solved in m-digit floating point arithmetic and if the condition number of A is of the order 10^{α}, then only $m - \alpha - 1$ digits in the solution can be trusted.

Summary.

- If the condition number of a matrix A is large, then small errors in the data may lead to large errors in the solution.
- Rule of thumb: If the linear system is solved in m-digit floating point arithmetic and if the condition number of A is of the order 10^{α}, then only $m - \alpha - 1$ digits in the solution can be trusted.