Conditioning of Square Linear Systems

Contents

Vector and Matrix Norms

Vector Norms

The “length” of a vector $\mathbf{v}$ can be measured by its **norm**.

Definition 1 (p-Norm of a Vector)

Let $p \in [1, \infty)$. The p-norm of $\mathbf{v} \in \mathbb{R}^m$ is denoted by $\|\mathbf{v}\|_p$ and is defined by

$$\|\mathbf{v}\|_p = \left(\sum_{i=1}^{m} |v_i|^p \right)^{1/p} .$$

When $p = \infty$,

$$\|\mathbf{v}\|_\infty = \max_{1 \leqslant i \leqslant m} |v_i| .$$

The most commonly used p values are 1, 2, and ∞:

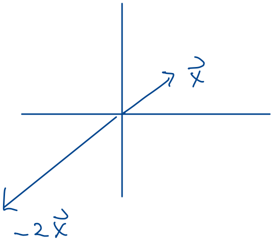
$$\|\mathbf{v}\|_1 = \sum_{i=1}^{m} |v_i| , \quad \|\mathbf{v}\|_2 = \sqrt{\sum_{i=1}^{m} |v_i|^2}.$$

Pythagorean.

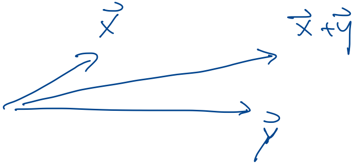
Vector Norms

In general, any function $\|\cdot\| : \mathbb{R}^m \to \mathbb{R}^+ \cup \{0\}$ is called a **vector norm** if it satisfies the following three properties:

1. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$.
2. $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$ for any constant α and any $\mathbf{x} \in \mathbb{R}^m$.
3. $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$. This is called the *triangle inequality*.



$2\|\vec{x}\| = \|-2\vec{x}\|$



Examples Let $\vec{v} = \begin{bmatrix} 6 \\ -3 \\ 93 \\ 5 \end{bmatrix}$.

$$(90+3)^2 = 8100 + 540 + 9$$
$$= 8649$$

① $\|\vec{v}\|_1 = |6| + |-3| + |93| + |5|$

$= 6 + 3 + 93 + 5 = 107$

② $\|\vec{v}\|_2 = \left(|6|^2 + |-3|^2 + |93|^2 + |5|^2\right)^{1/2}$

$= \sqrt{36 + 9 + 93^2 + 25}$

③ $\|\vec{v}\|_\infty = \max\{|6|, |-3|, |93|, |5|\} = 93$

Unit Vectors

$1 \leq p \leq \infty$

- A vector $\mathbf{u}$ is called a **unit vector** if $\|\mathbf{u}\| = 1$.
- Depending on the norm used, unit vectors will be different.
- For instance: $\mathbb{R}^2$

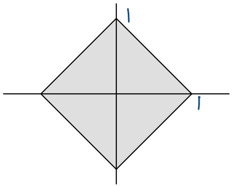


Figure 1: 1-norm

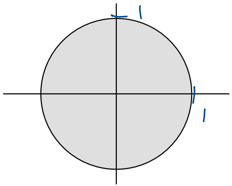


Figure 2: 2-norm

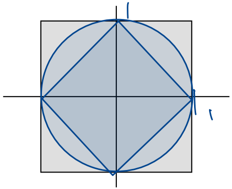


Figure 3: ∞-norm

Why are they shaped like that? (Work in 2-D)

$p=2$ $\quad \vec{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ is a unit vector in 2-norm, that is,

$1 = \|\vec{u}\|_2 = \sqrt{u_1^2 + u_2^2} \implies u_1^2 + u_2^2 = 1$ (unit circle)

$p=1$

$1 = \|\vec{u}\|_1 = |u_1| + |u_2| \longrightarrow \pm u_1 \pm u_2 = 1$

$$\begin{cases} u_1 + u_2 = 1 \\ u_1 - u_2 = 1 \\ -u_1 + u_2 = 1 \\ -u_1 - u_2 = 1 \end{cases}$$

Matrix Norms

The "size" of a matrix $A \in \mathbb{R}^{m\times n}$ can be measured by its **norm** as well. As above, we say that a function $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}^+ \cup \{0\}$ is a **matrix norm** if it satisfies the following three properties:

1. $\|A\| = 0$ if and only if $A = 0$.
2. $\|\alpha A\| = |\alpha|\,\|A\|$ for any constant α and any $A \in \mathbb{R}^{m\times n}$.
3. $\|A + B\| \leqslant \|A\| + \|B\|$ for any $A, B \in \mathbb{R}^{m\times n}$. This is called the *triangle inequality*.

- If, in addition to satisfying the three conditions, it satisfies

$$\|AB\| \leqslant \|A\| \, \|B\| \quad \text{for all } A \in \mathbb{R}^{m\times n} \text{ and all } B \in \mathbb{R}^{n\times p},$$

  it is said to be **consistent**.

- If, in addition to satisfying the three conditions, it satisfies

$$\underbrace{\|A\mathbf{x}\|}_{\text{vec.}} \leqslant \underbrace{\|A\|}_{\text{mat.}} \, \underbrace{\|\mathbf{x}\|}_{\text{vec.}} \quad \text{for all } A \in \mathbb{R}^{m\times n} \text{ and all } \mathbf{x} \in \mathbb{R}^{n},$$

  then we say that it is **compatible** with a vector norm.

Induced Matrix Norms

$\|A\vec{x}\|_p \leq \|A\|_p \|\vec{x}\|_p.$

Definition 2 (p-Norm of a Matrix)

Let $p \in [1, \infty]$. The p-norm of $A \in \mathbb{R}^{m \times n}$ is given by

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

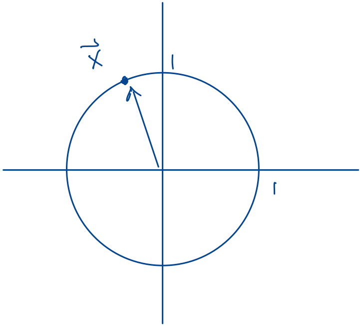
max. over all nonzero vectors

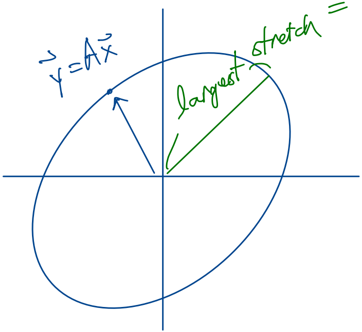
max. over all unit vectors

- The definition of this particular matrix norm is **induced** from the vector p-norm.
- By construction, matrix p-norm is a compatible norm. It is also consistent.
- Induced norms describe how the matrix stretches unit vectors with respect to the vector norm.

<u>Example</u> $p=2$, $n=2$ (i.e., 2-D)

Input $\qquad\qquad A \in \mathbb{R}^{2\times 2} \qquad\qquad$ Output





$\|\vec{x}\|_2 = 1$

$$\|A\|_2 = \max_{\|\vec{x}\|_2 = 1} \|\underline{A\vec{x}}\|_2$$

Induced Matrix Norms

The commonly used p-norms (for $p = 1, 2, \infty$) can also be calculated by

$$\|A\|_1 = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{m} |a_{ij}|, \qquad \texttt{>> max(sum(abs(A), 1))}$$

eigenvalue

singular value

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^{\mathrm{T}}A)} = \sigma_{\max}(A), \qquad \texttt{>> sqrt(max(eig(A' * A)))}$$

$$\|A\|_\infty = \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{n} |a_{ij}|. \qquad \texttt{>> max(sum(abs(A), 2))}$$

In words,

- The 1-norm of A is the maximum of the 1-norms of all column vectors.
- The 2-norm of A is the square root of the largest eigenvalue of $A^{\mathrm{T}}A$.
- The ∞-norm of A is the maximum of the 1-norms of all row vectors.

Non-Induced Matrix Norm – Frobenius Norm

Definition 3 (Frobenius Norm of a Matrix)

The Frobenius norm of $A \in \mathbb{R}^{m \times n}$ is given by

>> sqrt (A(:)' * A(:))

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

- This is not induced from a vector p-norm.
- However, both p-norm and the Frobenius norm are consistent and compatible.
- For compatibility of the Frobenius norm, the vector norm must be the vector 2-norm, that is, $\|A\mathbf{x}\|_2 \leqslant \|A\|_F \|\mathbf{x}\|_2$.

Norms in MATLAB

norm(A) : 2-norm by default.

$\vec{v}^T\vec{v} = [v_1 \cdots v_n]\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = v_1^2 + \cdots + v_n^2$

- Vector p-norms can be easily computed:

```
norm(v, 1)       % = sum(abs(v))
norm(v, 2)       % = sqrt(v'*v)  if v is a column
norm(v, 'inf')   % = max(abs(v))
```

- The same function `norm` is used to calculate matrix p-norms:

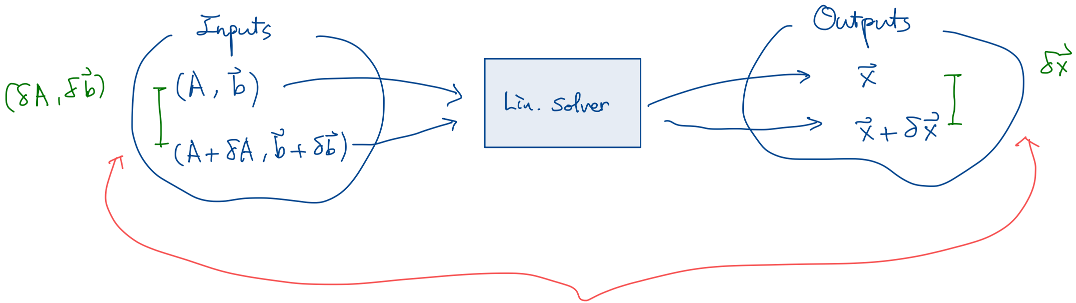
```
norm(A, 1)        % = max(sum(abs(A), 1))
norm(A, 2)        % = max(sqrt(eig(A'*A)))
norm(A, Inf)      % = max(sum(abs(A), 2))
```

- To calculate the Frobenius norm:

```
norm(A, 'fro')    % = sqrt(A(:)'*A(:))
                  % = norm(A(:), 2)
```

Conditioning

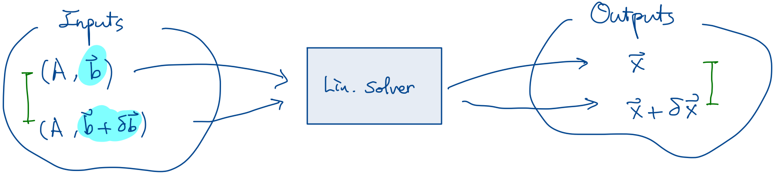
Conditioning of Solving $A\vec{x} = \vec{b}$



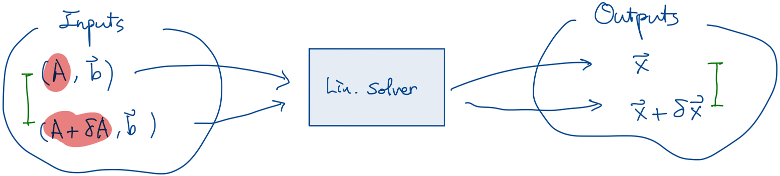
Measure sizes of errors using vector & matrix norms.

Conditioning (sensitivity) is the ratio of these two.

Case 1


Inputs
$(A, \vec{b})$
$(A, \vec{b}+\delta\vec{b})$
Lin. solver
Outputs
$\vec{x}$
$\vec{x}+\delta\vec{x}$

Case 2


Inputs
$(A, \vec{b})$
$(A+\delta A, \vec{b})$
Lin. solver
Outputs
$\vec{x}$
$\vec{x}+\delta\vec{x}$

Conditioning of Solving Linear Systems: Overview

Inputs: A, $\vec{b}$

Outputs: $\vec{x}$

- Analyze how robust (or sensitive) the solutions of $\underline{A\mathbf{x} = \mathbf{b}}$ are to perturbations of A and $\mathbf{b}$.
- For simplicity, consider separately the cases where

1. $\mathbf{b}$ changes to $\mathbf{b} + \delta\mathbf{b}$, while A remains unchanged, that is

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

2. A changes to $A + \delta A$, while $\mathbf{b}$ remains unchanged, that is

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

- Assume the matrix norm used is consistent and compatible.

Sensitivity to Perturbation of RHS

($\vec{b} \to \vec{b} + \delta\vec{b}$)

Case 1. $A\mathbf{x} = \mathbf{b} \ \rightarrow \ A(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$

- Bound $\|\delta\mathbf{x}\|$ in terms of $\|\delta\mathbf{b}\|$:

abs. err. of output

abs. err. of input.

$$\cancel{A\mathbf{x}} + A\delta\mathbf{x} = \cancel{\mathbf{b}} + \delta\mathbf{b}$$

$$A\delta\mathbf{x} = \delta\mathbf{b} \qquad \Longrightarrow \qquad \|\delta\mathbf{x}\| \leqslant \|A^{-1}\|\|\delta\mathbf{b}\|.$$

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

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

compatibility.

- Sensitivity in terms of relative errors:

$$\frac{\text{rel. err. of output}}{\text{rel. err. of input}} = \frac{\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|}}{\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}} = \frac{\|\delta\mathbf{x}\|\,\|\mathbf{b}\|}{\|\delta\mathbf{b}\|\,\|\mathbf{x}\|} \leqslant \frac{\|A^{-1}\|\,\|\delta\mathbf{b}\| \cdot \|A\|\,\|\mathbf{x}\|}{\|\delta\mathbf{b}\|\,\|\mathbf{x}\|} = \boxed{\|A^{-1}\|\|A\|.}$$

Sensitivity to Perturbation of Matrix

($A \longrightarrow A + \delta A$)

Case 2. $A\mathbf{x} = \mathbf{b} \;\rightarrow\; (A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}$

To obtain the bound below from here, We need a technical result called the Neumann's series theorem.

- Bound $\|\delta\mathbf{x}\|$ now in terms of $\|\delta A\|$:

$$\cancel{A\mathbf{x}} + A\delta\mathbf{x} + (\delta A)\mathbf{x} + (\delta A)\delta\mathbf{x} = \cancel{\mathbf{b}}$$
$$A\delta\mathbf{x} = -(\delta A)\mathbf{x} - (\delta A)\delta\mathbf{x}$$
$$\delta\mathbf{x} = -A^{-1}(\delta A)\mathbf{x} - A^{-1}(\delta A)\delta\mathbf{x} \implies \|\delta\mathbf{x}\| \lesssim \|A^{-1}\|\|\delta A\|\,\|\mathbf{x}\|.$$

(first-order truncation) as $\|\delta A\| \to 0$.

- Sensitivity in terms of relative errors:

$$\frac{\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|}}{\frac{\|\delta A\|}{\|A\|}} = \frac{\|\delta\mathbf{x}\|\,\|A\|}{\|\delta A\|\,\|\mathbf{x}\|} \lesssim \frac{\|A^{-1}\|\|\delta A\|\,\|\mathbf{x}\| \cdot \|A\|}{\|\delta A\|\,\|\mathbf{x}\|} = \boxed{\|A^{-1}\|\|A\|.}$$

just as above

Matrix Condition Number

- Motivated by the previous estimations, we define the **matrix condition number** by

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

(as long as it is consistent & compatible.)

where the norms can be any p-norm or the Frobenius norm.

- A subscript on κ such as 1, 2, ∞, or F(robenius) is used if clarification is needed.

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

$$\kappa_F(A) = \|A^{-1}\|_F \|A\|_F$$

Matrix Condition Number (cont')

Interpretation

- We can write

Case 1 Case 2

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \kappa(A)\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}, \qquad \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \kappa(A)\frac{\|\delta A\|}{\|A\|},$$

$\kappa(A)$ is the magnification ratio of errors.

where the second inequality is true only in the limit of infinitesimal perturbations δA. ($\lesssim$)

- The matrix condition number $\kappa(A)$ is equal to the condition number of solving a linear system of equation $A\mathbf{x} = \mathbf{b}$.
- The exponent of $\kappa(A)$ in scientific notation determines the approximate number of digits of accuracy that will be lost in calculation of $\mathbf{x}$.

e.g. If $\kappa(A) = 7.5 \times 10^{6}$, then about 6 sig. digits will be lost in the comp. of $\vec{x}$ (from $A\vec{x} = \vec{b}$).

- Since $1 = \|I\| = \|A^{-1}A\| \leqslant \|A^{-1}\|\|A\| = \kappa(A)$, a condition number of 1 is the best we can hope for.
- If $\kappa(A) > \boxed{\text{eps}}^{-1}$, then for computational purposes the matrix is singular.

$\approx 10^{16}$

Condition Numbers in MATLAB

Likewise, norm(A) = norm(A, 2)

- Use `cond` to calculate various condition numbers:

```
cond(A)              % the 2-norm; or  cond(A, 2)
cond(A, 1)           % the 1-norm
cond(A, Inf)         % the infinity-norm
cond(A, 'fro')       % the Frobenius norm
```

- A condition number estimator (in 1-norm)

```
condest(A)           % faster than cond
```

- The fastest method to estimate the condition number is to use `linsolve` function as below:

```
[x, inv_condest] = linsolve(A, b);
fast_condest = 1/inv_condest;
```