

CX1104: Linear Algebra for Computing

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}}_{b}$$

Chap. No : **7.1.1**

Lecture : **Least Squares**

Topic : **Introduction**

Concept : **Consistency in a System of Equations**

Instructor: **A/P Chng Eng Siong**

TAs: **Zhang Su, Vishal Choudhari**

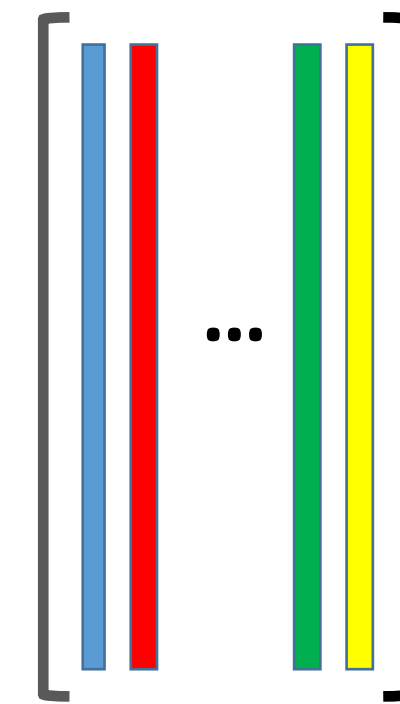
Consistency in a System of Equations

Consider solving the system of equations: $Ax = b$

Note:

- Matrix $A \in R^{M \times N}$, where
 - M denotes no. of rows/equations
 - N denotes no. of columns/unknowns
- $x \in R^N$
- $b \in R^M$
- The above system of equations can either be
 1. consistent (or)
 2. inconsistent.

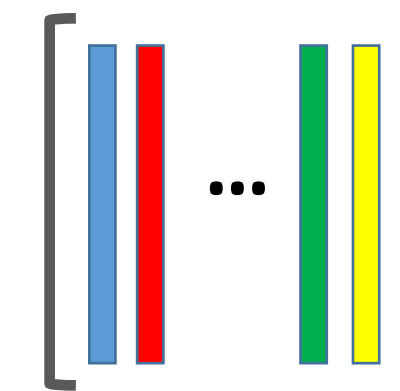
Based on M & N , there exist three cases:



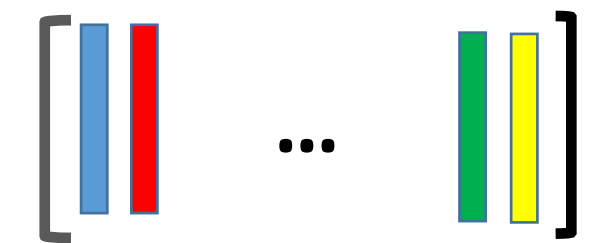
$$M \gg N$$

More equations,
less unknowns.

Hence, **over-determined!**
Typically this will result
in inconsistent system of
equations



$$M \approx N$$

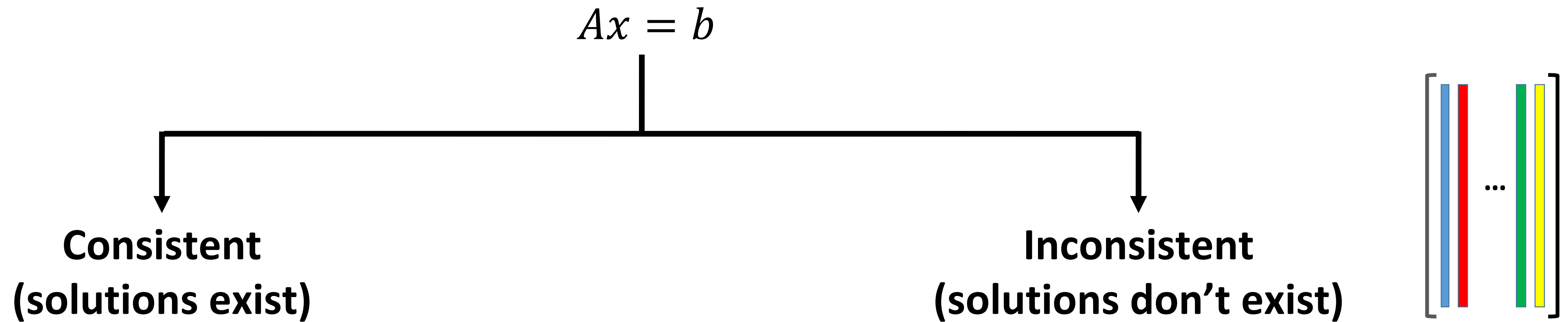


$$M \ll N$$

Less equations,
more unknowns.

Hence, **under-determined!**
Typically, this will
result
in infinite solutions

Consistency in a System of Equations



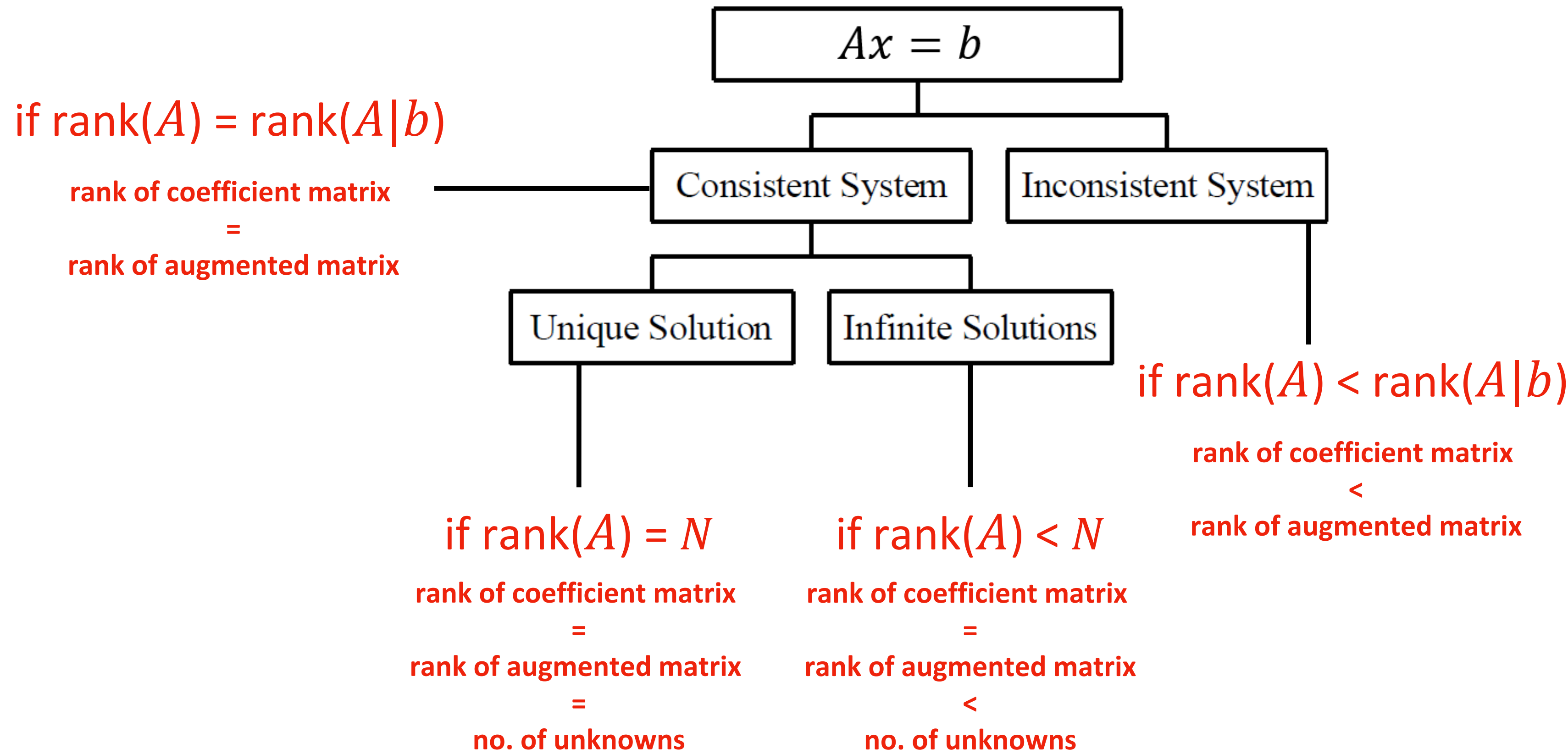
- b is in column space of A , i.e,
 b is formed by linear combinations of A 's columns.
- $\text{Rank}(A) = \text{Rank}(A|b)$, i.e,
rank of A is same as that of the augmented matrix.

- b is NOT in column space of A , i.e,
 b is NOT formed by linear combinations of A 's columns.
- **Typically** occurs when $M \gg N$ (**over-determined**), i.e,
there exist more equations than unknowns.
- The rows of A are dependent but,
their corresponding b values are not consistent.
- $\text{Rank}(A) < \text{Rank}(A|b)$, i.e,
rank of A is less than that of the augmented matrix.

Consistency in a System of Equations

A system of equations can be consistent or inconsistent. What does that mean?

A system of equations $Ax = b$ is consistent if there is a solution, and it is inconsistent if there is no solution. However, consistent system of equations does not mean a unique solution, that is, a consistent system of equation may have a unique solution or infinite solutions.



NOTE: Rank (A) is the maximum number of independent rows or columns of A.

You can find number of independent row or columns by:

1. row reduction process
2. rank(A) in MATLAB

Examples

$\text{rank}(A) = \text{rank}(A|b) = N$

Consistent and Unique Solution

a) The system of equations

$$\begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

is a consistent system of equations as it has a unique solution, that is,

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

```
>> A_b = [ 2 4 6; 1 3 4]

A_b =

     2     4     6
     1     3     4

>> rank(A_b)

ans =

     2
```

Consistent and Having Infinite Solutions

b) The system of equations

$$\begin{bmatrix} 2 & 4 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 6 \\ 3 \end{bmatrix}$$

is also a consistent system of equations but it has infinite solutions as given as follows.

Expanding the above set of equations,

$$\begin{aligned} 2x + 4y &= 6 \\ x + 2y &= 3 \end{aligned}$$

you can see that they are the same equation. Hence any combination of (x,y) that satisfies

$$2x + 4y = 6$$

is a solution. For example $(x,y)=(1,1)$ is a solution and other solutions include $(x,y)=(0.5,1.25)$, $(x,y)=(0, 1.5)$ and so on.

```
>> A_b = [ 2 4 6; 1 2 3]

A_b =

     2     4     6
     1     2     3

>> rank(A_b)

ans =

     1
```

Inconsistent and No solutions Exist

c) The system of equations

$$\begin{bmatrix} 2 & 4 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$$

$\text{rank}(A) < \text{rank}(A|b)$

```
>> A = [2 4; 1 2]

A =

     2     4
     1     2

>> rank(A)

ans =

     1

>> A_b = [ 2 4 6; 1 2 4]

A_b =

     2     4     6
     1     2     4

>> rank(A_b)

ans =

     2
```

$\text{rank}(A) = \text{rank}(A|b) < N$

CX1104: Linear Algebra for Computing

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}}_{b}$$

Chap. No : **7.1.2**

Lecture : **Least Squares**

Topic : **Introduction**

Concept : **The Least Squares Problem**

Instructor: **A/P Chng Eng Siong**

TAs: **Zhang Su, Vishal Choudhari**

Consistency in a System of Equations

$$Ax = b$$

Consistent
(solutions exist)

- b is in column space of A , i.e., b is formed by linear combinations of A 's columns.
- $\text{Rank}(A) = \text{Rank}(A|b)$, i.e., rank of A is same as that of the augmented matrix.

Consider

Inconsistent
(solutions don't exist)

- b is NOT in column space of A , i.e., b is NOT formed by linear combinations of A 's columns.
- **Typically** occurs when $M \gg N$ (**over-determined**), i.e., there exist more equations than unknowns.
- The rows of A are dependent but, their corresponding b values are not consistent.
- $\text{Rank}(A) < \text{Rank}(A|b)$, i.e., rank of A is less than that of the augmented matrix.

$$\begin{bmatrix} | & | & | & \dots & | & | \\ \hline | & | & | & \dots & | & | \\ \hline \end{bmatrix}$$

$$M \gg N$$

Least Squares Solution for Inconsistent Equations

Consider solving the system of equations: $Ax = b$

Note:

- Matrix $A \in \mathbb{R}^{M \times N}$, where
 - M denotes no. of rows/equations
 - N denotes no. of columns/unknowns
- $x \in \mathbb{R}^N$
- $b \in \mathbb{R}^M$
- When $M \gg N$,
 - the system is over-determined
 - the equations may be inconsistent
 - there may be no solution

Best we can do?

Find x such that Ax is as close to b as possible!

If A is $m \times n$ and \mathbf{b} is in \mathbb{R}^m , a **least-squares solution** of $A\mathbf{x} = \mathbf{b}$ is an $\hat{\mathbf{x}}$ in \mathbb{R}^n such that

$$\|\mathbf{b} - A\hat{\mathbf{x}}\| \leq \|\mathbf{b} - A\mathbf{x}\|$$

for all \mathbf{x} in \mathbb{R}^n .

Think of $A\mathbf{x}$ as an *approximation* to \mathbf{b} . The smaller the distance between \mathbf{b} and $A\mathbf{x}$, given by $\|\mathbf{b} - A\mathbf{x}\|$, the better the approximation. The **general least-squares problem** is to find an \mathbf{x} that makes $\|\mathbf{b} - A\mathbf{x}\|$ as small as possible. The adjective “least-squares” arises from the fact that $\|\mathbf{b} - A\mathbf{x}\|$ is the square root of a sum of squares.

Definitions

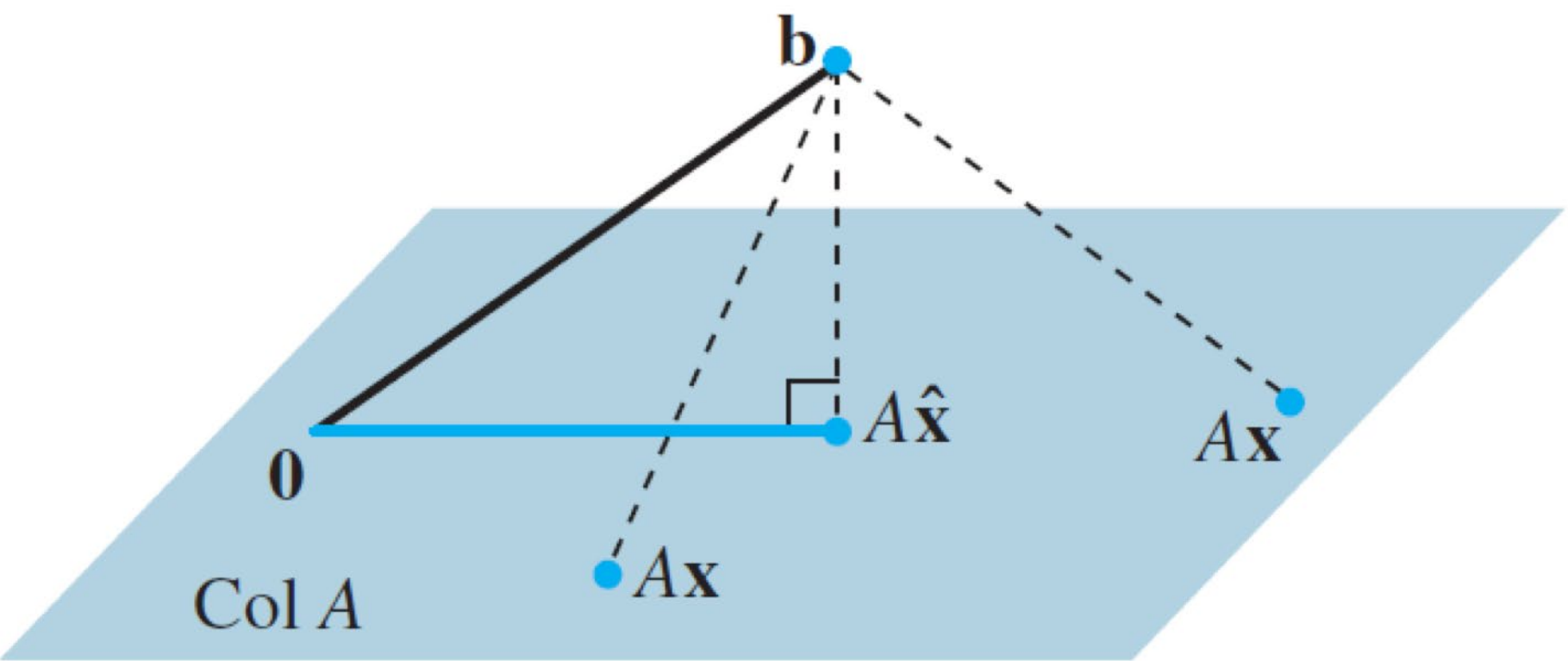


FIGURE 1 The vector \mathbf{b} is closer to $A\hat{\mathbf{x}}$ than to $A\mathbf{x}$ for other \mathbf{x} .

The most important aspect of the least-squares problem is that no matter what \mathbf{x} we select, the vector $A\mathbf{x}$ will necessarily be in the column space, $\text{Col } A$. So we seek an \mathbf{x} that makes $A\mathbf{x}$ the closest point in $\text{Col } A$ to \mathbf{b} . See Fig. 1. (Of course, if \mathbf{b} happens to be in $\text{Col } A$, then \mathbf{b} is $A\mathbf{x}$ for some \mathbf{x} , and such an \mathbf{x} is a “least-squares solution.”)

If a linear system is consistent, then its exact solutions are the same as its least squares solutions, in which case the least squares error is zero.

NOTE:
When the linear system $Ax = b$ is inconsistent, b does not lie in the column space of A .

To explain the terminology in this problem, suppose that the column form of $\mathbf{b} - A\mathbf{x}$ is

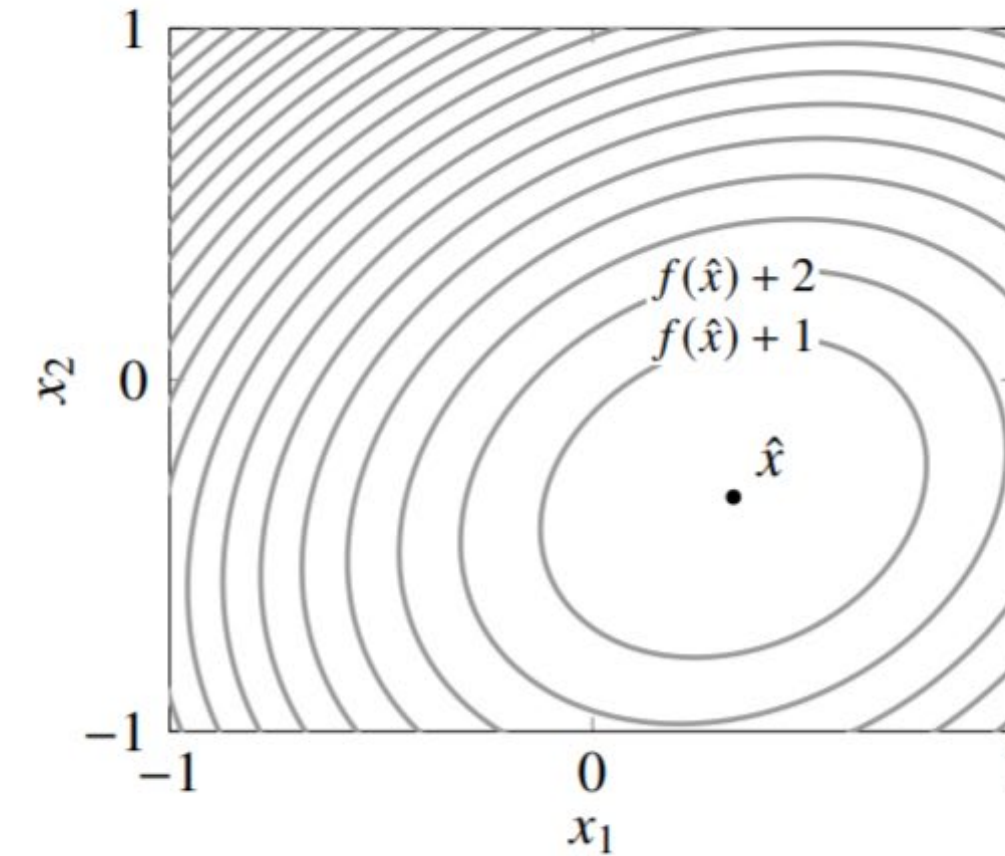
$$\mathbf{b} - A\mathbf{x} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix}$$

The term “least squares solution” results from the fact that minimizing $\|\mathbf{b} - A\mathbf{x}\|$ also has the effect of minimizing $\|\mathbf{b} - A\mathbf{x}\|^2 = e_1^2 + e_2^2 + \cdots + e_m^2$.

Example

Example

$$A = \begin{bmatrix} 2 & 0 \\ -1 & 1 \\ 0 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$



- the least squares solution \hat{x} minimizes

$$f(x) = \|Ax - b\|^2 = (2x_1 - 1)^2 + (-x_1 + x_2)^2 + (2x_2 + 1)^2$$

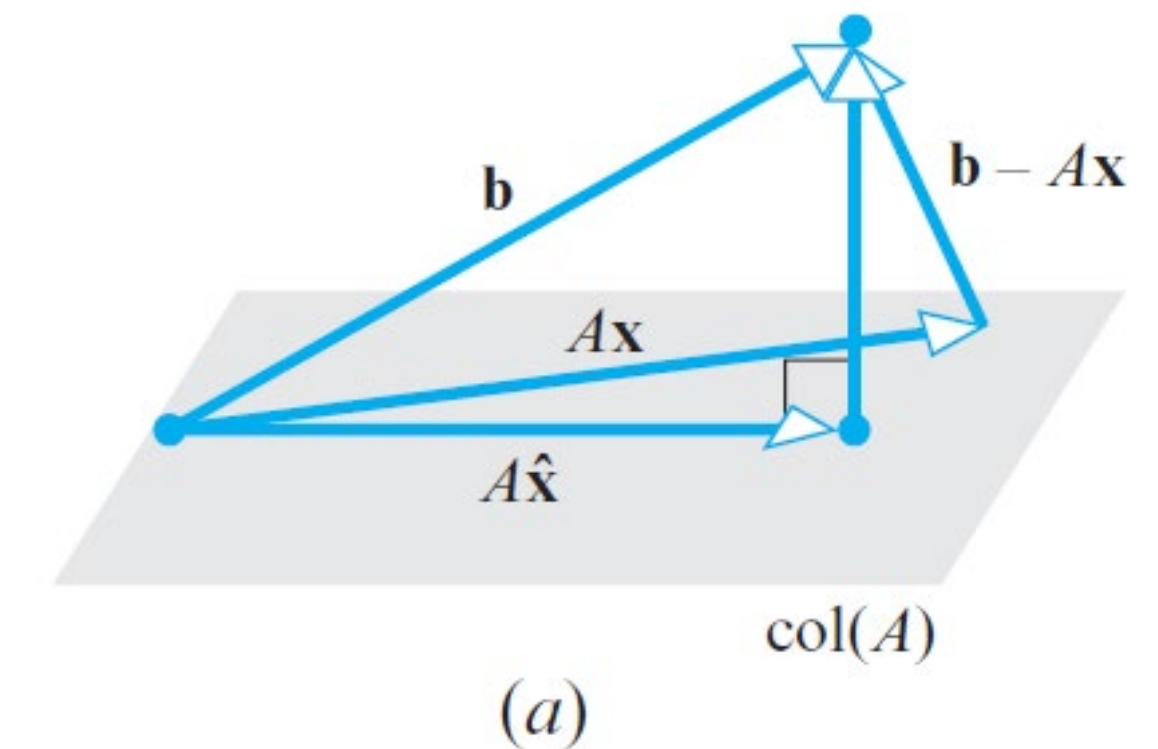
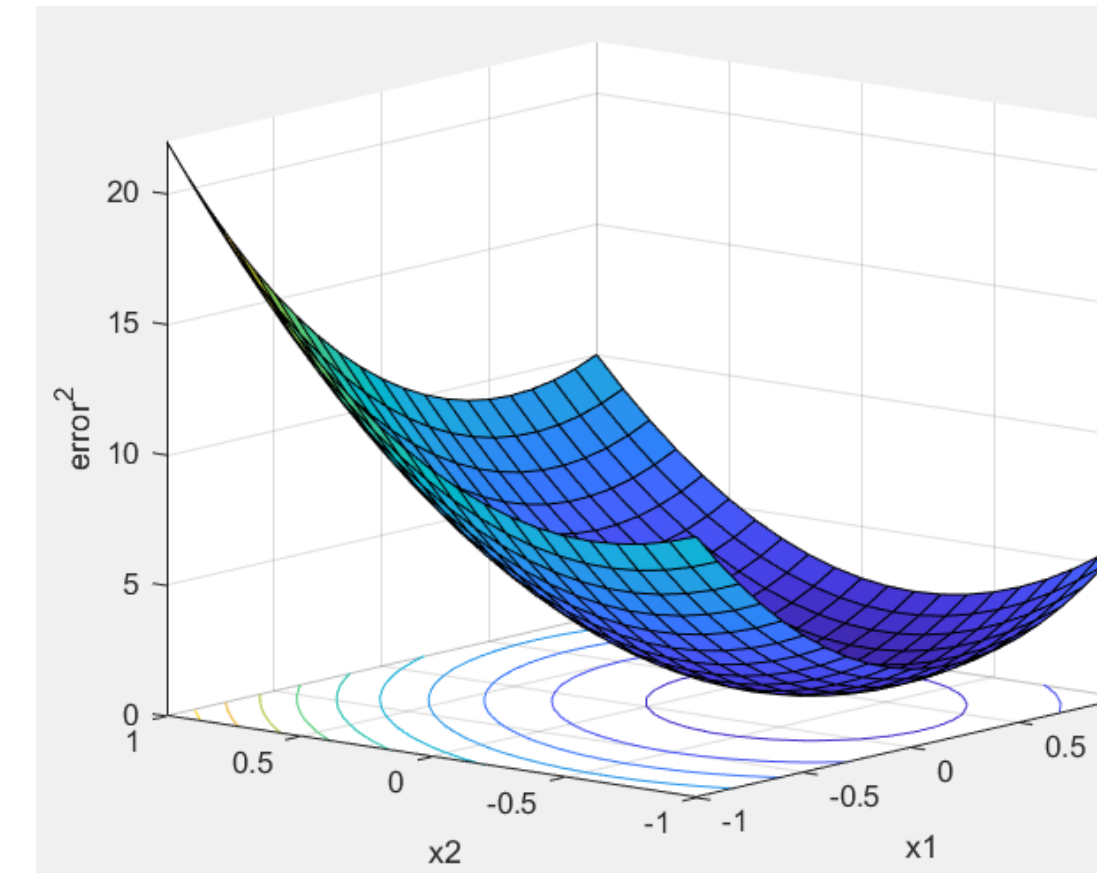
- to find \hat{x} , set derivatives with respect to x_1 and x_2 equal to zero:

$$10x_1 - 2x_2 - 4 = 0, \quad -2x_1 + 10x_2 + 4 = 0$$

solution is $(\hat{x}_1, \hat{x}_2) = (1/3, -1/3)$

Least squares

8.3



opStr =

'x1=0.30, x2=-0.30, err^2=0.680 '

```
%ch6_4_Ex1.m
%Chng Eng Siong, plotting the error wrt x
close all; clear all;
A = [2 0; -1 1; 0 2];
b = [1 0 -1]';
[x1,x2] = meshgrid(-1:0.1:1, -1:0.1:1);
[m,n] = size(x1);
z = zeros(m,n);
for i=1:m
    for j=1:n
        z(i,j) = norm(b - (x1(i,j)*A(:,1)+x2(i,j)*A(:,2))).^2;
    end
end
surfc(x1,x2,z)
xlabel('x1'); ylabel('x2'); zlabel('error^2');

% Lets print the min value and the x vector
minIdx = find(z == min(z(:)));
x1(minIdx), x2(minIdx), z(minIdx)
opStr = sprintf('x1=%0.2f, x2=%0.2f, err^2=%0.3f ',x1(minIdx),x2(minIdx),z(minIdx))
```

CX1104: Linear Algebra for Computing

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}}_{A \quad m \times n} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}}_{x \quad n \times 1} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}}_{b \quad m \times 1}$$

Chap. No : **7.1.3**

Lecture : **Least Squares**

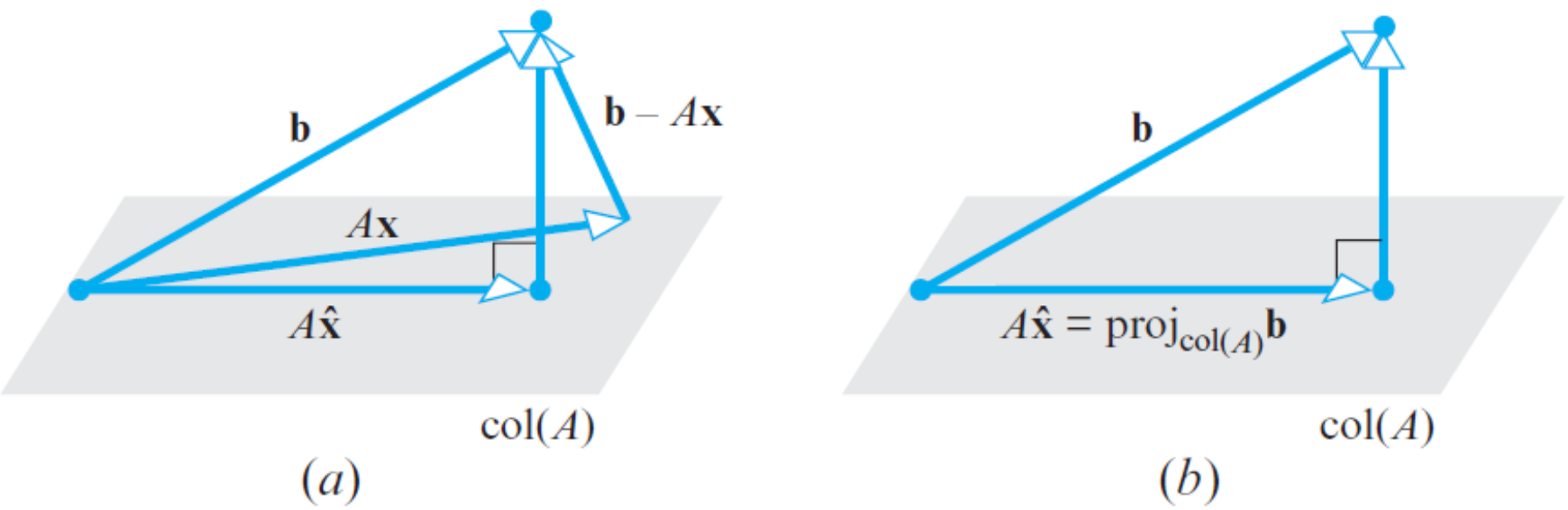
Topic : **Solving the Least Squares Problem**

Concept : **Best Approx. Theorem and Normal Equation**

Instructor: **A/P Chng Eng Siong**

TAs: **Zhang Su, Vishal Choudhari**

Best Approximation Theorem



THEOREM 6.4.1 Best Approximation Theorem

If W is a finite-dimensional subspace of an inner product space V , and if \mathbf{b} is a vector in V , then $\text{proj}_W \mathbf{b}$ is the **best approximation** to \mathbf{b} from W in the sense that

$$\|\mathbf{b} - \text{proj}_W \mathbf{b}\| < \|\mathbf{b} - \mathbf{w}\|$$

for every vector \mathbf{w} in W that is different from $\text{proj}_W \mathbf{b}$.

Proof For every vector \mathbf{w} in W , we can write

$$\mathbf{b} - \mathbf{w} = (\mathbf{b} - \text{proj}_W \mathbf{b}) + (\text{proj}_W \mathbf{b} - \mathbf{w})$$

But $\text{proj}_W \mathbf{b} - \mathbf{w}$, being a difference of vectors in W , is itself in W ; and since $\mathbf{b} - \text{proj}_W \mathbf{b}$ is orthogonal to W , the two terms on the right side of (1) are orthogonal. Thus, it follows from the Theorem of Pythagoras (Theorem 6.2.3) that

$$\|\mathbf{b} - \mathbf{w}\|^2 = \|\mathbf{b} - \text{proj}_W \mathbf{b}\|^2 + \|\text{proj}_W \mathbf{b} - \mathbf{w}\|^2$$

If $\mathbf{w} \neq \text{proj}_W \mathbf{b}$, it follows that the second term in this sum is positive, and hence that

$$\|\mathbf{b} - \text{proj}_W \mathbf{b}\|^2 < \|\mathbf{b} - \mathbf{w}\|^2$$

Since norms are nonnegative, it follows (from a property of inequalities) that

$$\|\mathbf{b} - \text{proj}_W \mathbf{b}\| < \|\mathbf{b} - \mathbf{w}\| \quad \blacktriangleleft$$

The Normal Equation

$$Ax = b$$

Multiplying both sides by A^T

$$A^T A x = A^T b$$

Normal Equation!

The set of least-squares solutions of $Ax = b$ coincides with the nonempty set of solutions of the normal equations $A^T A x = A^T b$.

Proved in the next slide!

Solution of the General Least-Squares Problem

Given A and b as above, apply the Best Approximation Theorem in Section 6.3 to the subspace $\text{Col } A$. Let

$$\hat{b} = \text{proj}_{\text{Col } A} b$$

Because \hat{b} is in the column space of A , the equation $Ax = \hat{b}$ is consistent, and there is an \hat{x} in \mathbb{R}^n such that

$$A\hat{x} = \hat{b} \quad (1)$$

Since \hat{b} is the closest point in $\text{Col } A$ to b , a vector \hat{x} is a least-squares solution of $Ax = b$ if and only if \hat{x} satisfies (1). Such an \hat{x} in \mathbb{R}^n is a list of weights that will build \hat{b} out of the columns of A . See Fig. 2. [There are many solutions of (1) if the equation has free variables.]

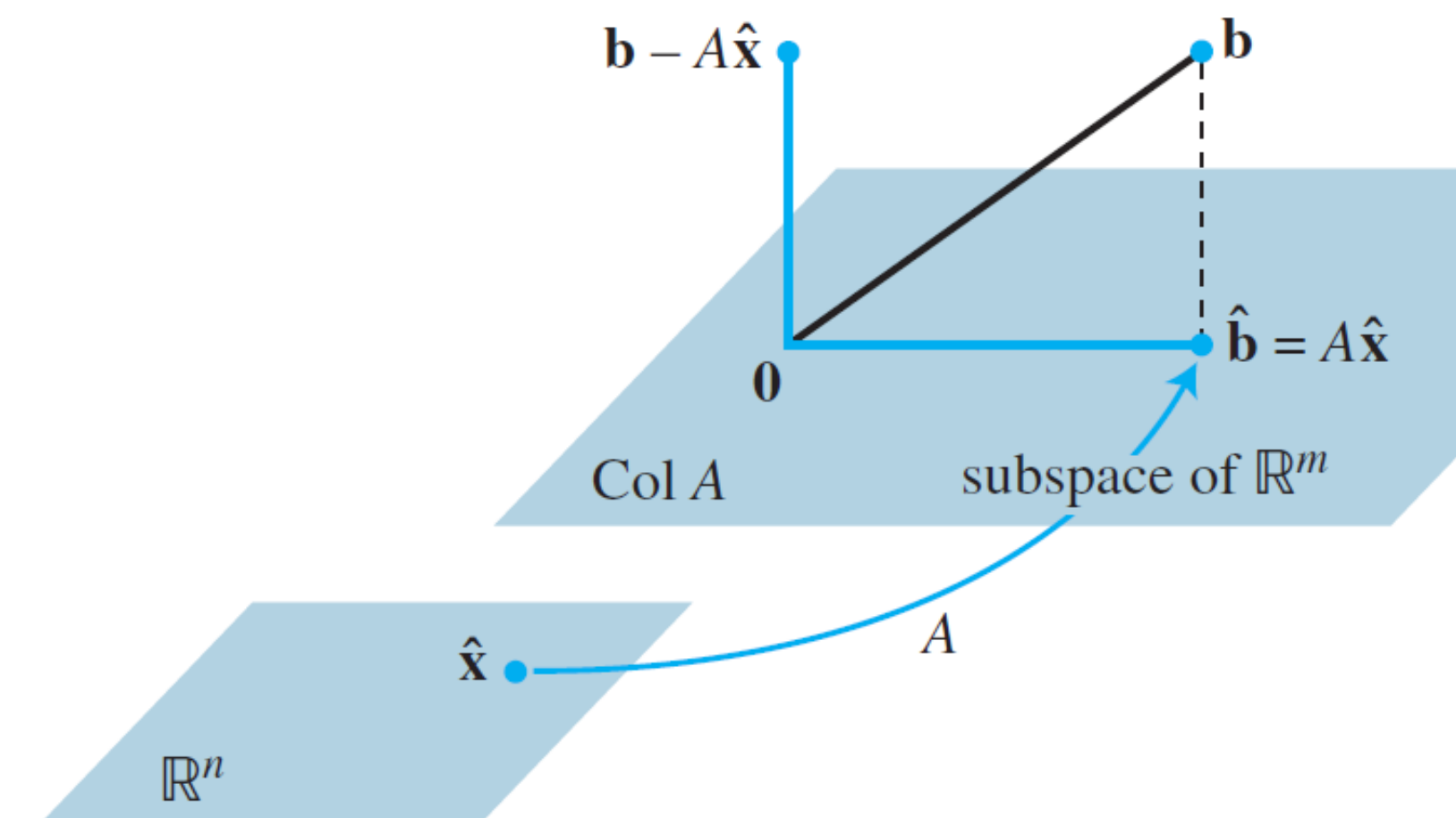


FIGURE 2 The least-squares solution \hat{x} is in \mathbb{R}^n .

Why called “Normal”?

Ref: <https://mathworld.wolfram.com/NormalEquation.html>

Lay, Linear Algebra and its Applications (4th Edition)

6.5 Least-Squares Problems 361

The Normal Equation Proof

Suppose $\hat{\mathbf{x}}$ satisfies $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$. By the Orthogonal Decomposition Theorem in Section 6.3, the projection $\hat{\mathbf{b}}$ has the property that $\mathbf{b} - \hat{\mathbf{b}}$ is orthogonal to $\text{Col } A$, so $\mathbf{b} - A\hat{\mathbf{x}}$ is orthogonal to each column of A . If \mathbf{a}_j is any column of A , then $\mathbf{a}_j \cdot (\mathbf{b} - A\hat{\mathbf{x}}) = 0$, and $\mathbf{a}_j^T (\mathbf{b} - A\hat{\mathbf{x}}) = 0$. Since each \mathbf{a}_j^T is a row of A^T ,

$$A^T (\mathbf{b} - A\hat{\mathbf{x}}) = \mathbf{0} \quad (2)$$

(This equation also follows from Theorem 3 in Section 6.1.) Thus

$$\begin{aligned} A^T \mathbf{b} - A^T A \hat{\mathbf{x}} &= \mathbf{0} \\ A^T A \hat{\mathbf{x}} &= A^T \mathbf{b} \end{aligned}$$

These calculations show that each least-squares solution of $A\mathbf{x} = \mathbf{b}$ satisfies the equation

$$A^T A \mathbf{x} = A^T \mathbf{b} \quad (3)$$

The matrix equation (3) represents a system of equations called the **normal equations** for $A\mathbf{x} = \mathbf{b}$. A solution of (3) is often denoted by $\hat{\mathbf{x}}$.

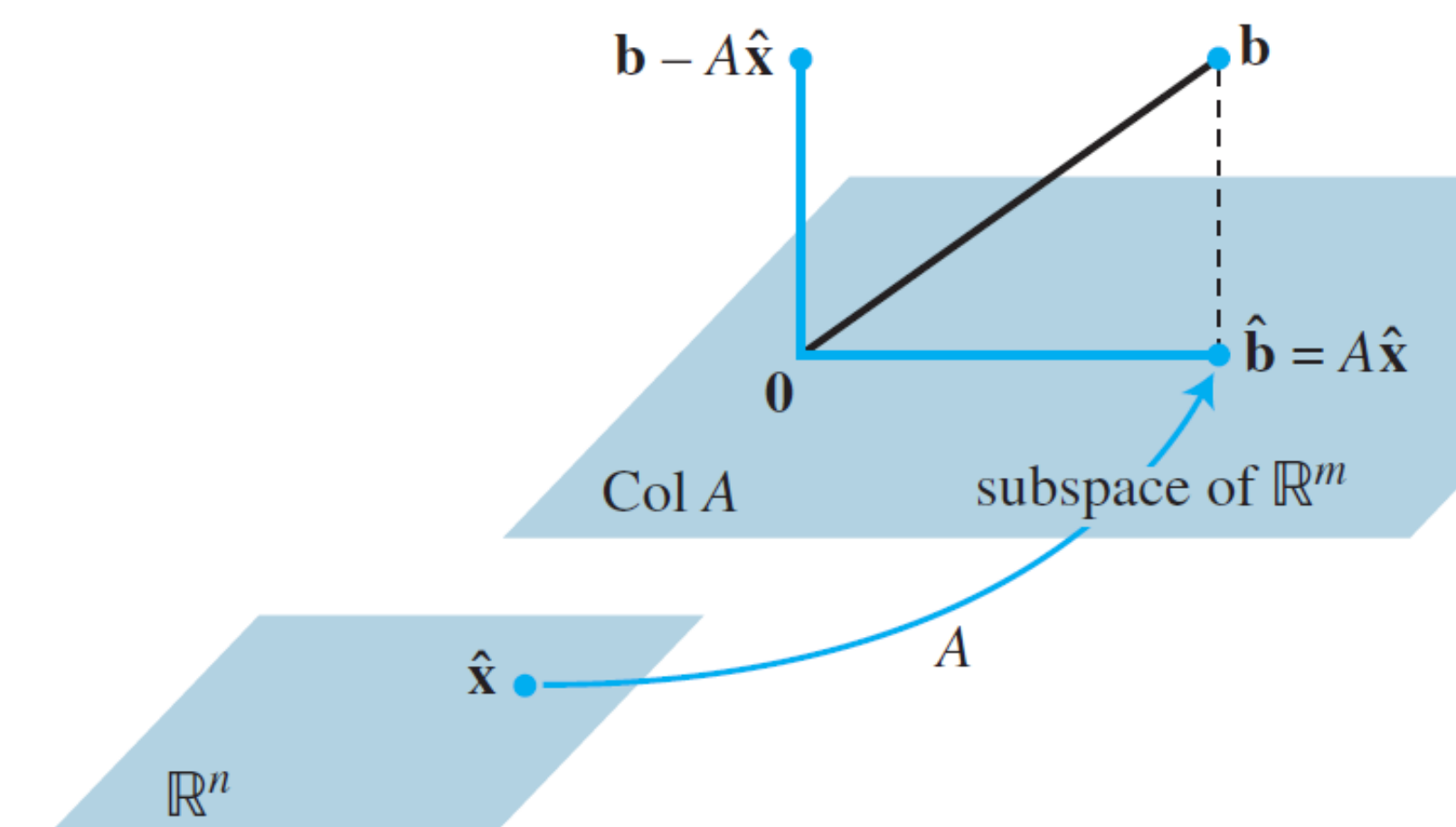


FIGURE 2 The least-squares solution $\hat{\mathbf{x}}$ is in \mathbb{R}^n .

THEOREM 14

Let A be an $m \times n$ matrix. The following statements are logically equivalent:

- The equation $A\mathbf{x} = \mathbf{b}$ has a unique least-squares solution for each \mathbf{b} in \mathbb{R}^m .
- The columns of A are linearly independent.
- The matrix $A^T A$ is invertible.

When these statements are true, the least-squares solution $\hat{\mathbf{x}}$ is given by

$$\hat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b} \quad (4)$$

Examples

EXAMPLE 1 Find a least-squares solution of the inconsistent system $A\mathbf{x} = \mathbf{b}$ for

$$A = \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix}$$

SOLUTION To use normal equations (3), compute:

$$A^T A = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 17 & 1 \\ 1 & 5 \end{bmatrix}$$

$$A^T \mathbf{b} = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix} = \begin{bmatrix} 19 \\ 11 \end{bmatrix}$$

Then the equation $A^T A \mathbf{x} = A^T \mathbf{b}$ becomes

$$\begin{bmatrix} 17 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 19 \\ 11 \end{bmatrix}$$

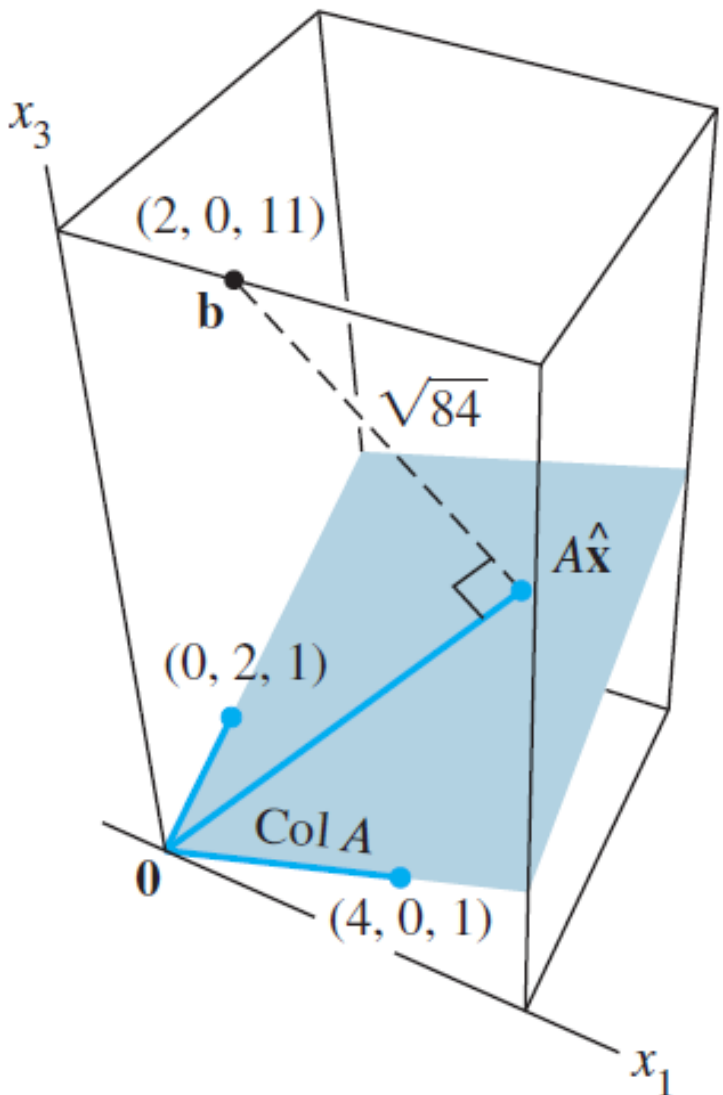


FIGURE 3

Row operations can be used to solve this system, but since $A^T A$ is invertible and 2×2 , it is probably faster to compute

$$(A^T A)^{-1} = \frac{1}{84} \begin{bmatrix} 5 & -1 \\ -1 & 17 \end{bmatrix}$$

and then to solve $A^T A \mathbf{x} = A^T \mathbf{b}$ as

$$\begin{aligned} \hat{\mathbf{x}} &= (A^T A)^{-1} A^T \mathbf{b} \\ &= \frac{1}{84} \begin{bmatrix} 5 & -1 \\ -1 & 17 \end{bmatrix} \begin{bmatrix} 19 \\ 11 \end{bmatrix} = \frac{1}{84} \begin{bmatrix} 84 \\ 168 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \end{aligned}$$

In many calculations, $A^T A$ is invertible, but this is not always the case. The next

Examples

EXAMPLE 3 Given A and \mathbf{b} as in Example 1, determine the least-squares error in the least-squares solution of $A\mathbf{x} = \mathbf{b}$.

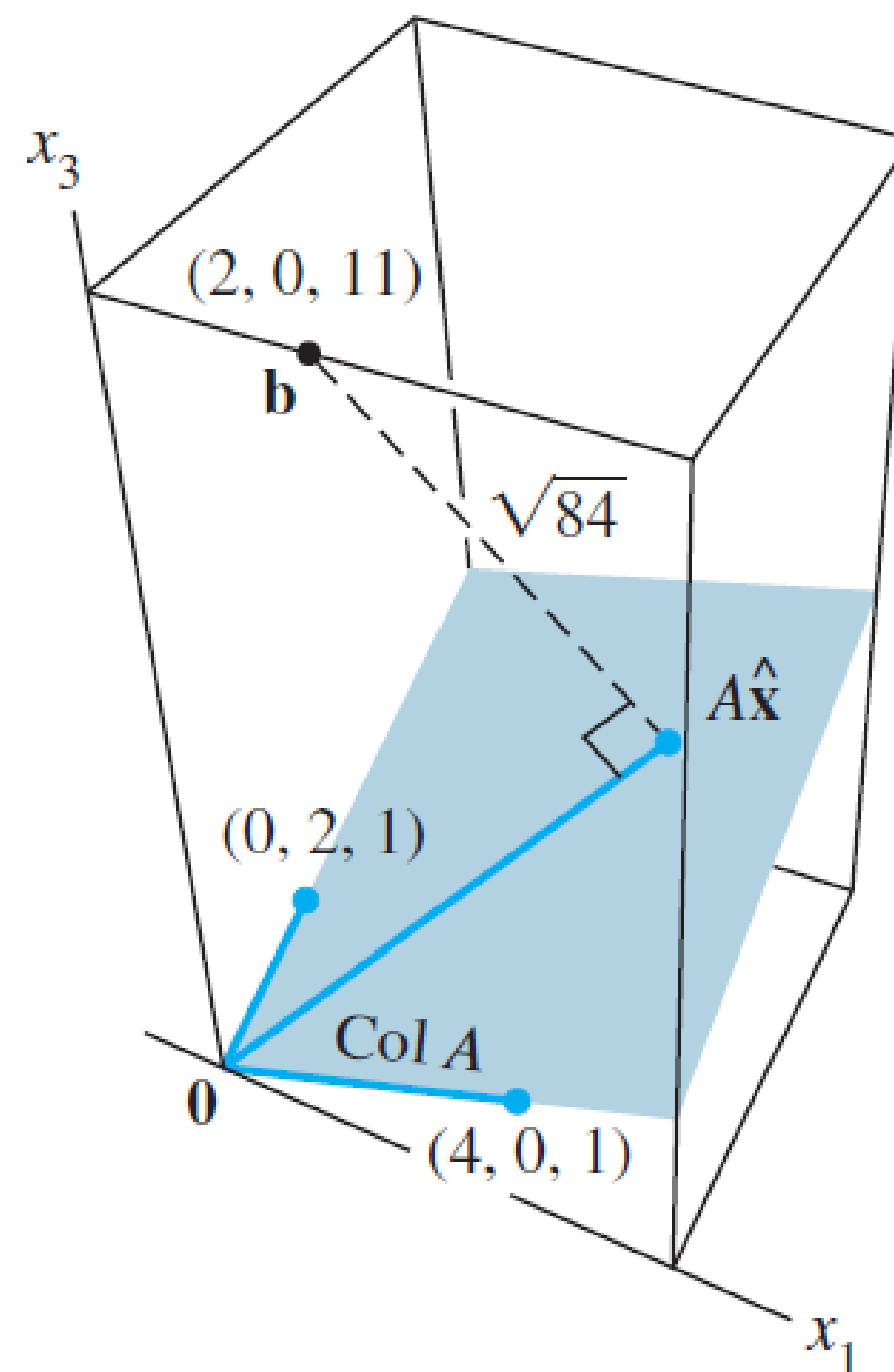


FIGURE 3

SOLUTION From Example 1,

$$\mathbf{b} = \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix} \quad \text{and} \quad A\hat{\mathbf{x}} = \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 3 \end{bmatrix}$$

Hence

$$\mathbf{b} - A\hat{\mathbf{x}} = \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix} - \begin{bmatrix} 4 \\ 4 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ -4 \\ 8 \end{bmatrix}$$

and

$$\|\mathbf{b} - A\hat{\mathbf{x}}\| = \sqrt{(-2)^2 + (-4)^2 + 8^2} = \sqrt{84}$$

The least-squares error is $\sqrt{84}$. For any \mathbf{x} in \mathbb{R}^2 , the distance between \mathbf{b} and the vector $A\mathbf{x}$ is at least $\sqrt{84}$. See Fig. 3. Note that the least-squares solution $\hat{\mathbf{x}}$ itself does not appear in the figure. ■

Examples

EXAMPLE 2 Find a least-squares solution of $A\mathbf{x} = \mathbf{b}$ for

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} -3 \\ -1 \\ 0 \\ 2 \\ 5 \\ 1 \end{bmatrix}$$

- Note the linear dependency in the rows and columns of A :**
- Column 1 = Column 2 + Column 3 + Column 4
 - Rows 1 & 2 are same, but their corresponding b values are different (inconsistent)
 - Rows 3 & 4 are same, but their corresponding b values are different (inconsistent)
 - Rows 5 & 6 are same, but their corresponding b values are different (inconsistent)

SOLUTION Compute

$$A^T A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 2 & 2 & 2 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{bmatrix}$$

$$A^T \mathbf{b} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} -3 \\ -1 \\ 0 \\ 2 \\ 5 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ -4 \\ 2 \\ 6 \end{bmatrix}$$

Note that $A^T A$ is always a square matrix.

The augmented matrix for $A^T A \mathbf{x} = A^T \mathbf{b}$ is

Reduced to

$$\begin{bmatrix} 6 & 2 & 2 & 2 & 4 \\ 2 & 2 & 0 & 0 & -4 \\ 2 & 0 & 2 & 0 & 2 \\ 2 & 0 & 0 & 2 & 6 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & 1 & 3 \\ 0 & 1 & 0 & -1 & -5 \\ 0 & 0 & 1 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$A^T A$ $A^T b$

The general solution is $x_1 = 3 - x_4$, $x_2 = -5 + x_4$, $x_3 = -2 + x_4$, and x_4 is free. So the general least-squares solution of $A\mathbf{x} = \mathbf{b}$ has the form

$$\hat{\mathbf{x}} = \begin{bmatrix} 3 \\ -5 \\ -2 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} -1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Note: Here, there are infinitely many solutions with the same least square error. **Note:** Here, $A^T A$ is not invertible (its determinant is 0).

$A^T A$ may not be invertible if:

- some columns are linearly dependent (i.e. we have redundant features) (as in this example)
 - solution: remove the linear dependency
- too many features ($m < n$)
 - solution: delete some features, there are too many features for the amount of data we have

Ref: http://mlwiki.org/index.php/Normal_Equation

Ref: Andrew Ng discussing this phenomenon-
<https://www.coursera.org/lecture/machine-learning/normal-equation-noninvertibility-zSiE6>

Examples

```
% pg 362 Lay's book, Example 2 - Least Squares, when A'A is singular
close all; clear all;
A = [1 1 0 0; 1 1 0 0; 1 0 1 0; 1 0 1 0; 1 0 0 1; 1 0 0 1];
b = [-3 -1 0 2 5 1]';

AtA = A'*A
rank_Ata = rank(AtA) % A'*A is singular, we check its rank

% Ax = b;
x1 = pinv(A)*b
x2 = inv(A'*A)*A'*b % This is what we think we should do
% compare inv(A'*A) bs pinv(A'*A)
disp("using normal inverseder (A'*A):");
inv(A'*A)
disp("using pinverseder (A'*A):");
pinv(A'*A)

x3 = pinv(A'*A)*A'*b % This is what Andy Ng suggest to d
```

AtA =

6	2	2	2
2	2	0	0
2	0	2	0
2	0	0	2

rank_Ata =

3

x1 =

0.5000
-2.5000
0.5000
2.5000

x2 =

1
-3
0
2

Warning: Matrix is close to singular or badly scaled.
> In [Lay_example2_pg362](#) (line 8)

$A^T A$

ans =

1.0e+15 *			
1.5012	-1.5012	-1.5012	-1.5012
-1.5012	1.5012	1.5012	1.5012
-1.5012	1.5012	1.5012	1.5012
-1.5012	1.5012	1.5012	1.5012

$A^T A$ is non-invertible. Hence MATLAB computes its inverse as a very large value $\Rightarrow \infty$

ans =

0.0938	0.0312	0.0313	0.0313
0.0313	0.3437	-0.1562	-0.1563
0.0312	-0.1562	0.3438	-0.1562
0.0313	-0.1562	-0.1563	0.3438

x3 =

0.5000
-2.5000
0.5000
2.5000

NOTE: Pseudo-inverse (pinv) will be introduced later.

Reference

Some Useful readings:

- 1) Fogel: Learning Goals: find the best solution (by one measure, anyway) of inconsistent equation. Learn to apply the algebra, geometry, and calculus of projections to this problem.
<http://staff.imsa.edu/~fogel/LinAlg/PDF/33%20Least%20Squares.pdf>
- 2) Why normal equation always have a solution: unique or infinite even if A has dependent column
 - a) <https://math.stackexchange.com/questions/2920398/how-do-the-normal-equations-always-have-a-solution>
 - b) <https://math.stackexchange.com/questions/72222/existence-of-least-squares-solution-to-ax-b>
 - c) <https://stats.stackexchange.com/questions/63143/question-about-a-normal-equation-proof>
- 3) Prof Walker, Worcester Polytechnic Institute: https://users.wpi.edu/~walker/MA3257/HANDOUTS/least-squares_handout.pdf

CX1104: Linear Algebra for Computing

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}}_{b}$$

Chap. No : **7.1.4**

Lecture : **Least Squares**

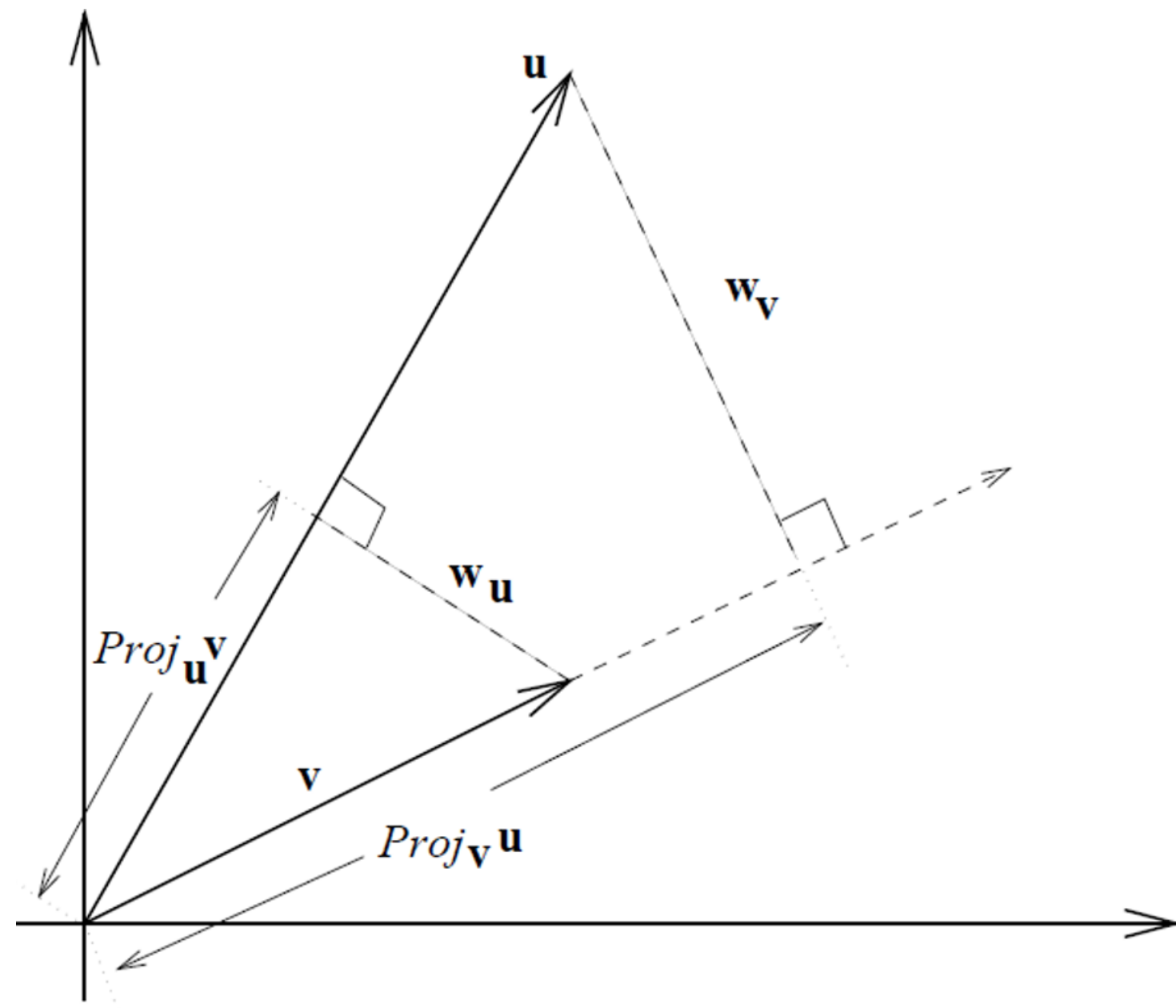
Topic : **Least Squares**

Concept : **Projection Matrix and its Properties**

Instructor: **A/P Chng Eng Siong**

TAs: **Zhang Su, Vishal Choudhari**

Projection matrix for a space spanned by a single vector \mathbf{v}



$$Proj_{\mathbf{v}} \mathbf{u} = \mathbf{v} \left(\frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{v}\|^2} \right),$$

$$\begin{aligned} proj_{\mathbf{v}} \mathbf{u} &= \frac{\mathbf{v} \cdot \mathbf{u}}{\mathbf{v} \cdot \mathbf{v}} \mathbf{v} \\ &= \mathbf{v} \frac{\mathbf{v} \cdot \mathbf{u}}{\mathbf{v} \cdot \mathbf{v}} \end{aligned}$$

Next, use the transpose definition of the inner product followed by the associative property of multiplication. Remember, when performing the dot product, a scalar multiplier may be placed anywhere you wish.

$$\begin{aligned} proj_{\mathbf{v}} \mathbf{u} &= \frac{1}{\mathbf{v}^T \mathbf{v}} \mathbf{v} (\mathbf{v}^T \mathbf{u}) \\ &= \frac{1}{\mathbf{v}^T \mathbf{v}} (\mathbf{v} \mathbf{v}^T) \mathbf{u} \\ &= \frac{\mathbf{v} \mathbf{v}^T}{\mathbf{v}^T \mathbf{v}} \mathbf{u} \end{aligned}$$

The expression $\mathbf{v} \mathbf{v}^T$ is called an **outer product** (the transpose operator is outside the product versus its inside position in the inner product). If we define $P = \frac{\mathbf{v} \mathbf{v}^T}{\mathbf{v}^T \mathbf{v}}$, then the projection formula becomes

$$proj_{\mathbf{v}} \mathbf{u} = P \mathbf{u}, \text{ where } P = \frac{\mathbf{v} \mathbf{v}^T}{\mathbf{v}^T \mathbf{v}}.$$

The matrix P is called the projection matrix. You can project any vector onto the vector \mathbf{v} by multiplying by the matrix P .

Ref: 6.2.2_Orthogonal_Projections

Projection Matrix for col(A) and the Least Squares Solution

Consider solving the system of equations: $Ax = b$

- If b is not in the $\text{col}(A)$, we find the least squares solution \hat{x} .

$$\hat{x} \text{ satisfies: } A^T A \hat{x} = A^T b$$

$$\text{Thus, } \hat{x} = (A^T A)^{-1} A^T b$$

Qn: **What is projection of b in the column space of A ?**

$$\text{Ans: } \hat{b} = \text{Proj}_{\text{Col}(A)} b = A \hat{x}$$

NOTE:

$$\hat{b} = \text{Proj}_{\text{Col}(A)} b = A \hat{x}$$

$$\Rightarrow \hat{b} = \text{Proj}_{\text{Col}(A)} b = A(A^T A)^{-1} A^T b$$

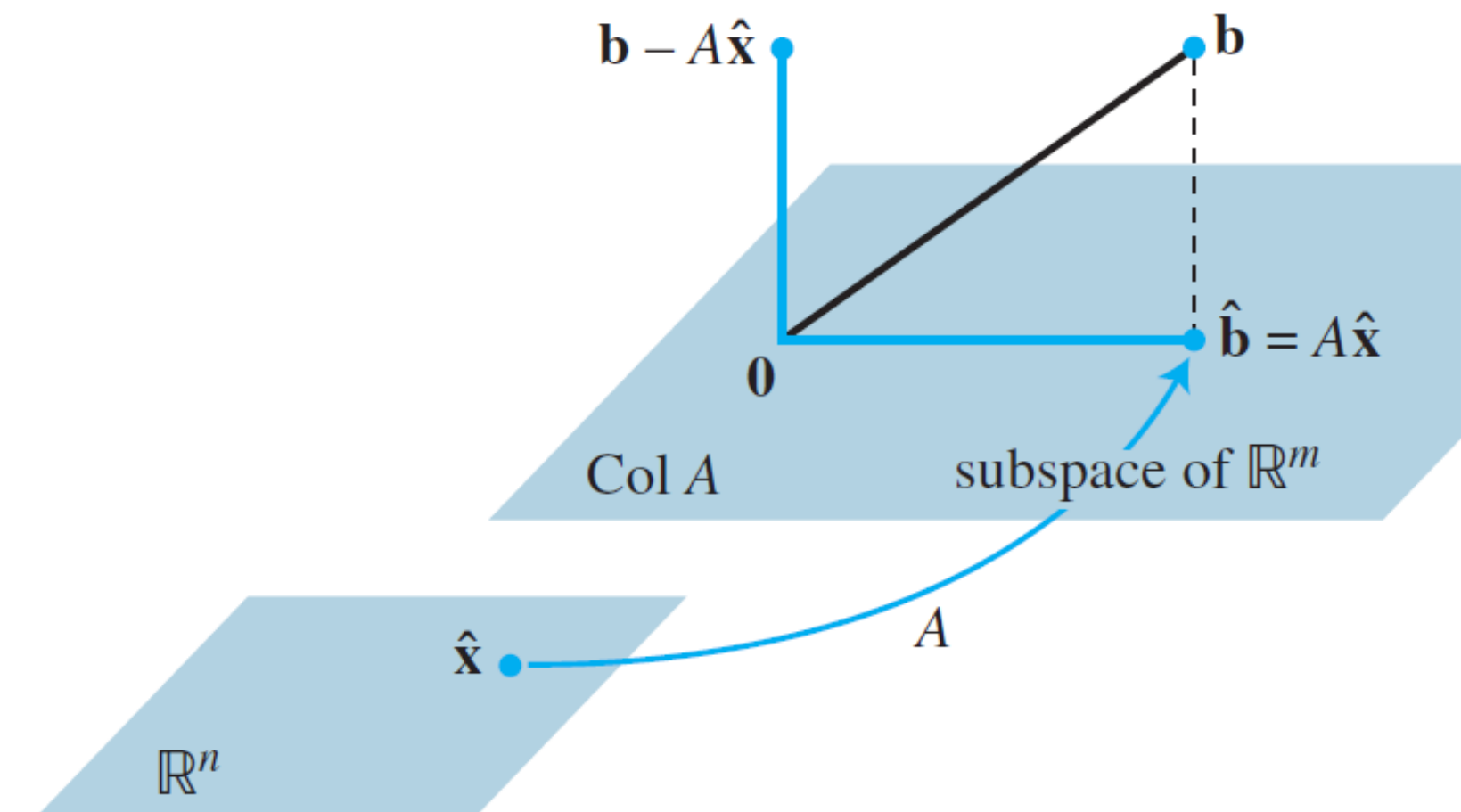


FIGURE 2 The least-squares solution \hat{x} is in \mathbb{R}^n .

$P = A(A^T A)^{-1} A^T$ is the **projection matrix**

The **vector b** can be **projected** into **column space of A** by **multiplying** it with **projection matrix P** .

Projection matrix P maps the actual response values b with predicted values \hat{b} .

Ref: https://en.wikipedia.org/wiki/Projection_matrix

Properties of Projection Matrix

When P is multiplied by vector b , the resulting vector $\hat{b} = Pb = A\hat{x}$ is the least squares (nearest) solution in the column space of A .

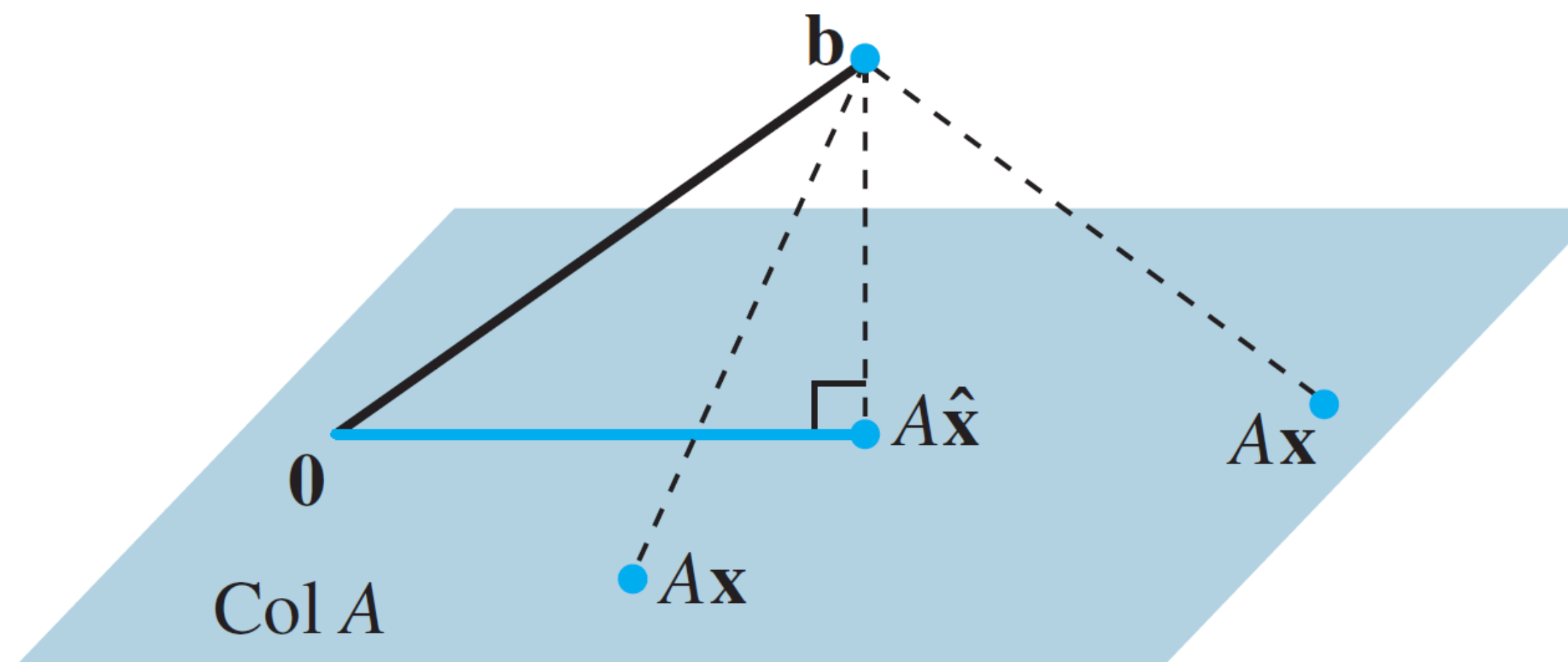


FIGURE 1 The vector \mathbf{b} is closer to $A\hat{\mathbf{x}}$ than to $A\mathbf{x}$ for other \mathbf{x} .

Properties of Projection Matrix:

1. $P^T = P$
2. $P^N = P$ [Idempotent Property]

Properties of Projection Matrix (X)

Lecture 16: Projection matrices and least squares

COURSE HOME

SYLLABUS

CALENDAR

INSTRUCTOR
INSIGHTS

VIDEO LECTURES <

READINGS



Properties of Projection Matrix:

1. $P^T = P$
2. $P^N = P$ [Idempotent Property]

Ref: <https://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/video-lectures/lecture-16-projection-matrices-and-least-squares/>

Example 2: Projection Matrix Properties

4.3 *Projection matrices.* A matrix $P \in \mathbf{R}^{n \times n}$ is called a *projection matrix* if $P = P^T$ and $P^2 = P$.

- (a) Show that if P is a projection matrix then so is $I - P$.
- (b) Suppose that the columns of $U \in \mathbf{R}^{n \times k}$ are orthonormal. Show that UU^T is a projection matrix. (Later we will show that the converse is true: every projection matrix can be expressed as UU^T for some U with orthonormal columns.)
- (c) Suppose $A \in \mathbf{R}^{n \times k}$ is full rank, with $k \leq n$. Show that $A(A^T A)^{-1}A^T$ is a projection matrix.
- (d) If $S \subseteq \mathbf{R}^n$ and $x \in \mathbf{R}^n$, the point y in S closest to x is called the *projection of x on S* . Show that if P is a projection matrix, then $y = Px$ is the projection of x on $\mathcal{R}(P)$. (Which is why such matrices are called projection matrices ...)

Example 1

Solution:

(a) To show that $I - P$ is a projection matrix we need to check two properties:

i. $I - P = (I - P)^T$

ii. $(I - P)^2 = I - P$.

The first one is easy: $(I - P)^T = I - P^T = I - P$ because $P = P^T$ (P is a projection matrix.) To show the second property we have

$$\begin{aligned}(I - P)^2 &= I - 2P + P^2 \\ &= I - 2P + P \quad (\text{since } P = P^2) \\ &= I - P\end{aligned}$$

and we are done.

(b) Since the columns of U are orthonormal we have $U^T U = I$. Using this fact it is easy to prove that $U U^T$ is a projection matrix, *i.e.*, $(U U^T)^T = U U^T$ and $(U U^T)^2 = U U^T$. Clearly, $(U U^T)^T = (U^T)^T U^T = U U^T$ and

$$\begin{aligned}(U U^T)^2 &= (U U^T)(U U^T) \\ &= U(U^T U)U^T \\ &= U U^T \quad (\text{since } U^T U = I).\end{aligned}$$

(c) First note that $(A(A^T A)^{-1} A^T)^T = A(A^T A)^{-1} A^T$ because

$$\begin{aligned}(A(A^T A)^{-1} A^T)^T &= (A^T)^T ((A^T A)^{-1})^T A^T \\ &= A ((A^T A)^T)^{-1} A^T \\ &= A(A^T A)^{-1} A^T.\end{aligned}$$

Also $(A(A^T A)^{-1} A^T)^2 = A(A^T A)^{-1} A^T$ because

$$\begin{aligned}(A(A^T A)^{-1} A^T)^2 &= (A(A^T A)^{-1} A^T) (A(A^T A)^{-1} A^T) \\ &= A ((A^T A)^{-1} A^T A) (A^T A)^{-1} A^T \\ &= A(A^T A)^{-1} A^T \quad (\text{since } (A^T A)^{-1} A^T A = I).\end{aligned}$$

(d) To show that Px is the projection of x on $\mathcal{R}(P)$ we verify that the “error” $x - Px$ is orthogonal to *any* vector in $\mathcal{R}(P)$. Since $\mathcal{R}(P)$ is nothing but the span of the columns of P we only need to show that $x - Px$ is orthogonal to the columns of P , or in other words, $P^T(x - Px) = 0$. But

$$\begin{aligned}P^T(x - Px) &= P^T(x - Px) && (\text{since } P = P^T) \\ &= Px - P^2 x \\ &= 0 && (\text{since } P^2 = P)\end{aligned}$$

and we are done.

Some related lecture notes on Orthogonal projection and Projection Matrix (X)

- 1) Michigan: <http://www.math.lsa.umich.edu/~speyer/417/OrthoProj.pdf>
- 2) Taiwan's lecture on projection matrix:
http://www.ss.ncu.edu.tw/~lyu/lecture_files_en/lyu_LA_Notes/Lyu_LA_2012/Lyu_LA_3_2012.pdf
- 3) Strang: Lecture 17, Spring 2005, Orthogonal Matrix and Gram Schmidt
<https://www.youtube.com/watch?v=0MtwqhlwdrI>
- 4) Berkeley: <https://math.berkeley.edu/~qadeer/teaching/F15Math54/Worksheet%204%20Solutions.pdf>
- 5) Harvard: https://people.math.harvard.edu/~knill/teaching/math19b_2011/handouts/lecture18.pdf
- 6) 2) Georgia Tech: A review of orthogonal projection matrix
<https://textbooks.math.gatech.edu/ila/projections.html>

More related references: Advance (X)

1) MathStackExchange: A projection matrix is unique even with a different set of basis for the same subspace (clever answer using $B = AC$, and solving the projection matrix equation – note this is for full rank in column)

<https://math.stackexchange.com/questions/2189183/show-that-projection-onto-a-subspace-is-unique-even-with-a-different-basis>

3) The confusion of terminology “least squares solution $= x$ ”, to solve for $Ax = y$, the orthogonal projection is \hat{y} , the found Ax for least squares solution

<https://math.stackexchange.com/questions/1298261/difference-between-orthogonal-projection-and-least-squares-solution>

4) Relating SVD to projection matrix

<https://math.stackexchange.com/questions/2033896/least-squares-solutions-and-the-orthogonal-projector-onto-the-column-space/2180194#2180194>

Concrete example showing if $C(A) == C(B)$, then their projection matrixes are the same (X)

Let A and B be 2 matrixes dimension 3x2, spanning the same column space.
We will generate their projection matrix and show they are the same.

```
>> A = [1 1; 2 3; 3 -1]

A =

     1     1
     2     3
     3    -1

>> B = [A(:,1) A(:,1)+A(:,2)]

B =

     1     2
     2     5
     3     2
```

```
>> PA = A*inv(A'*A)*A'

PA =

     0.1232     0.3188     0.0797
     0.3188     0.8841    -0.0290
     0.0797    -0.0290     0.9928

>> PB = B*inv(B'*B)*B'

PB =

     0.1232     0.3188     0.0797
     0.3188     0.8841    -0.0290
     0.0797    -0.0290     0.9928
```

```
>> PA*PA

ans =

     0.1232     0.3188     0.0797
     0.3188     0.8841    -0.0290
     0.0797    -0.0290     0.9928

>> PA*B

ans =

     1.0000     2.0000
     2.0000     5.0000
     3.0000     2.0000

>> PB*A

ans =

     1.0000     1.0000
     2.0000     3.0000
     3.0000    -1.0000
```

```
>> PA_orthogonal = (eye(3)-PA)

PA_orthogonal =

     0.8768    -0.3188    -0.0797
    -0.3188     0.1159     0.0290
    -0.0797     0.0290     0.0072

>> PA_orthogonal*A

ans =

    1.0e-15 *

     0.0833     0.0833
     0.2220     0.2220
     0.3886     0.3886
```

```
>> rank(A)

ans =

     2

>> rank(B)

ans =

     2
```

PA same as PB
as A and B has the same
col space

Shows that A an B are full rank,
and B is constructed using linear combination of A
➔ same space

Orthogonal complement
to col(A)

PA*PA == same
(property of projection matrix)
PA*B == B (shows B in same col space as A)

CX1104: Linear Algebra for Computing

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}}_{b}$$

Chap. No : **7.1.5**

Lecture : **Least Squares**

Topic : **Least Squares**

Concept : **Summary Least Squares Solution**

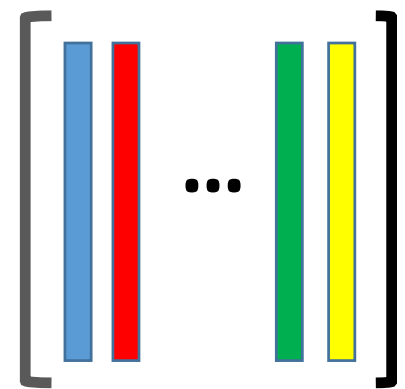
Instructor: **A/P Chng Eng Siong**

TAs: **Zhang Su, Vishal Choudhari**

Solving Least Squares using QR Factorisation and MATLAB

Solving $Ax = b$:

Case 1

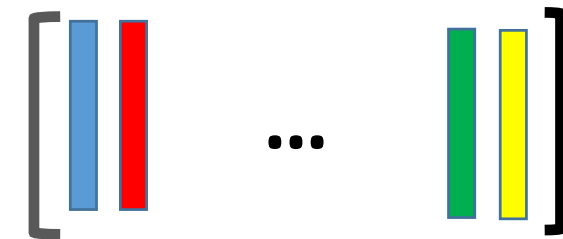


$$M \approx N$$

Say,

- A is square and invertible (full rank)
- then, b is in column space of A

Case 2

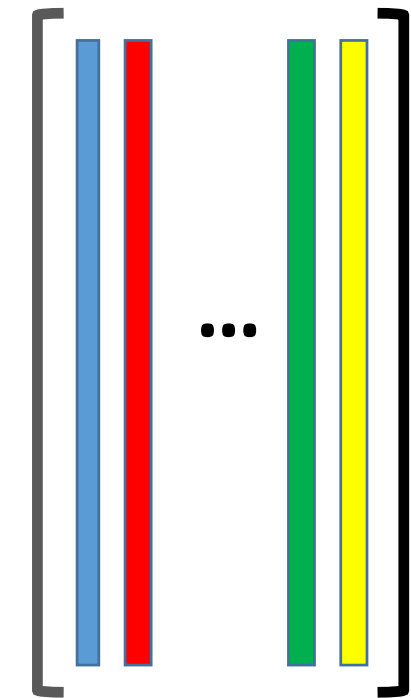


$$M \ll N$$

Under-determined

- As there are more unknowns than equations, infinitely many solutions exist.
- Hence, the goal then becomes to solve for x , such that, $||x||$ is minimised!

Case 3



$$M \gg N$$

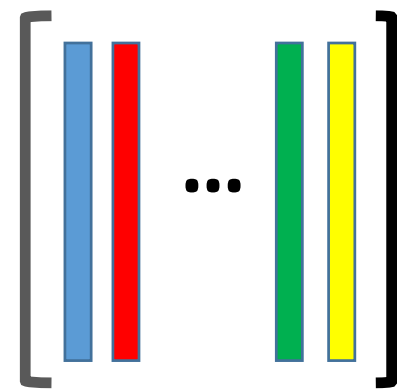
Over-determined

- b may not be in col space of A
- Hence, $b = Ax + \epsilon$
- ϵ models error/noise

Solving Least Squares using QR Factorisation and MATLAB

Solving $Ax = b$:

Case 1

A diagram representing a matrix with columns of different colors. It shows a blue column, a red column, an ellipsis, a green column, and a yellow column, all enclosed in large square brackets. This represents a matrix with multiple columns, where the first two are blue and red, followed by an ellipsis, then green and yellow columns.
$$\begin{bmatrix} \text{blue} & \text{red} & \dots & \text{green} & \text{yellow} \end{bmatrix}$$

$$M \approx N$$

Solution:

As b in column space of A , unique solution exists.
 $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, since A is square and has full rank.

Say,

- A is square and invertible (full rank)
- then, b is in column space of A

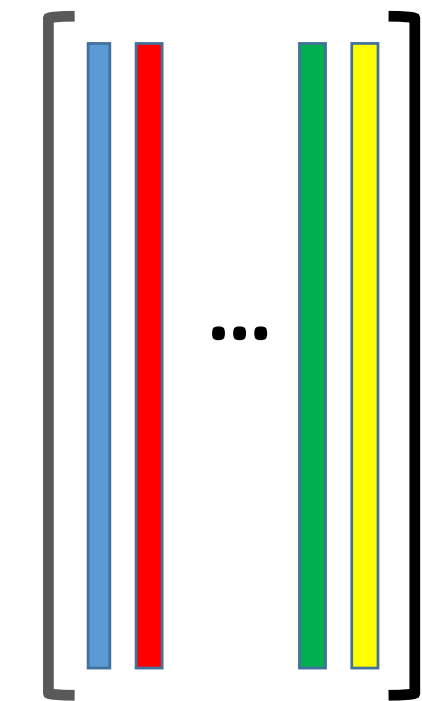
Solving Least Squares using QR Factorisation and MATLAB

For case 3, b is not in col. space of A .

Hence, an estimate of x , denoted by \hat{x} , such that least squares error ($\|b - Ax\|$) is minimised can be found.

Solving $Ax = b$:

Case 3



$M \gg N$

Over-determined

- b may not be in col space of A
- Hence, $b = Ax + \epsilon$
- ϵ models error/noise

Three ways to solve for case 3!

Way 1(X)
Pseudoinverse

$$\hat{x} = \text{pinv}(A) \times b$$

Way 2
Inverting Normal Equation

$$\hat{x} = (A^T A)^{-1} A^T b$$

Way 3
QR

Let $A = QR$, where $Q^T Q = I$.

Hence, $Ax = b$ can be rewritten as:
 $Q^T QRx = Q^T b$ (or)
 $Rx = Q^T b$

Therefore, $\hat{x} = R^{-1} Q^T b$

Revisiting QR Factorisation and Solving $Ax = b$

Consider solving the system of equations: $Ax = b$

Through QR factorisation, A can be written as:

$$A = QR$$

where,

The Q Factor:

- Q is $m \times n$ with orthonormal columns ($Q^T Q = I$)
- If A is square ($m = n$), then Q is orthogonal, i.e., $Q^T Q = Q Q^T = I$

The R Factor:

- R is $n \times n$ upper triangular, with nonzero diagonal elements
- R is nonsingular (diagonal elements are nonzero)

So, $Ax = b$ can be rewritten as: $QRx = b$.

Multiplying both sides by Q^T yields:

$$Q^T QRx = Q^T b \text{ (or)}$$
$$Rx = Q^T b$$

Since R is an upper triangular matrix, x can be solved by:

1. Back-substitution
2. On MATLAB: $x = R^{-1} Q^T b$

Algorithm Complexity

1. compute QR factorization $A = QR$ ($2mn^2$ flops if A is $m \times n$)
2. matrix-vector product $d = Q^T b$ ($2mn$ flops)
3. solve $Rx = d$ by back substitution (n^2 flops)

complexity: $2mn^2$ flops

Ref: "Why use QR to solve $Ax=b$?" by Dr. Peyam

<https://www.youtube.com/watch?v=J41Ypt6Mftc>



CX1104: Linear Algebra for Computing

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}}_{b}$$

Chap. No : **7.2.1**

Lecture : **Least Squares**

Topic : **Applications to Linear Models**

Concept : **Linear in the parameter Models**

Instructor: **A/P Chng Eng Siong**

TAs: **Zhang Su, Vishal Choudhari**

Fitting a line to data

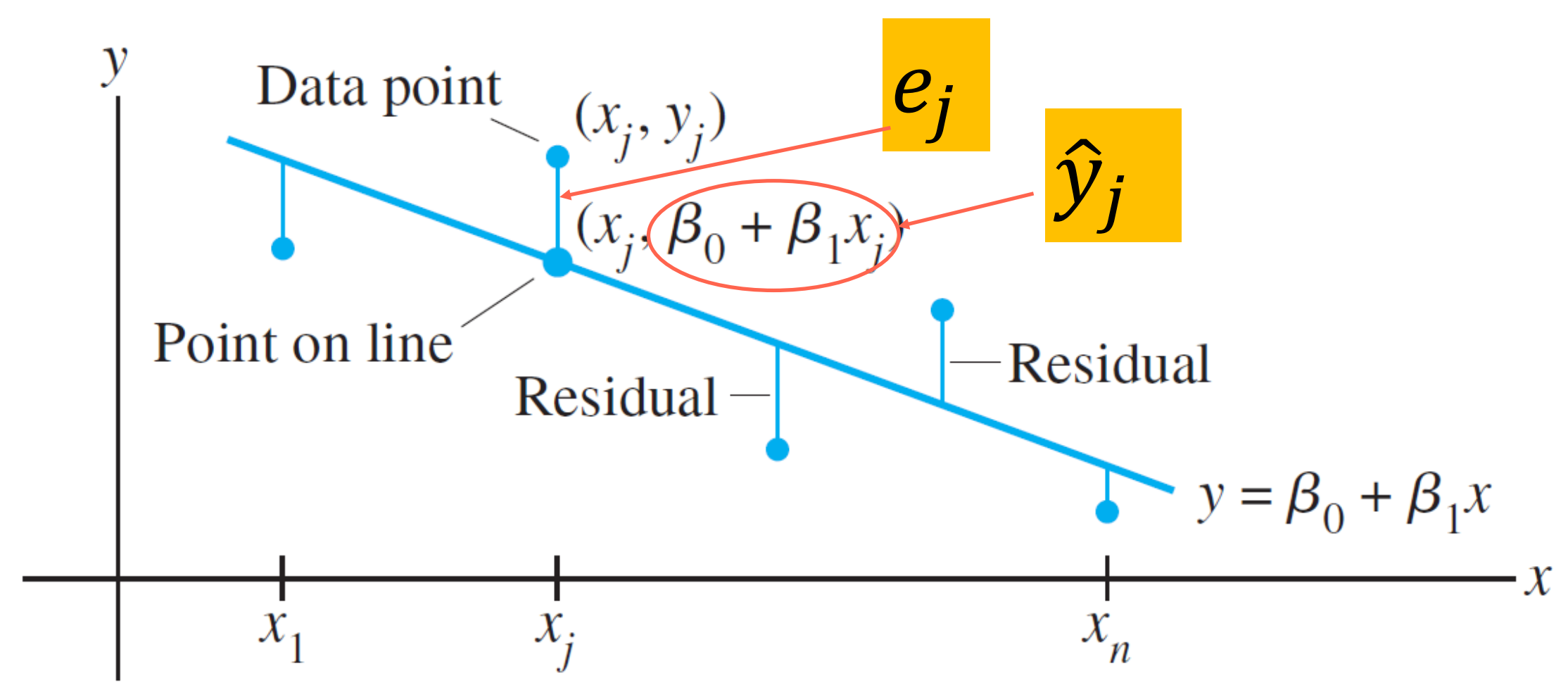


FIGURE 1 Fitting a line to experimental data.

We are given: $\{x_j, y_j\}$ where $j = 1..n$

Example

j	x_i	y_i
1	0.5	9.2
2	1.0	10.4
3	1.5	8.1

Find an equation $y = \beta_0 + \beta_1 x$ that describes the table.

Lay 5e, pg 371

Ref: Chasnov: <https://www.youtube.com/watch?v=RIQBEhLhM8Y>

Least-Squares Lines

The simplest relation between two variables x and y is the linear equation $y = \beta_0 + \beta_1 x$.¹ Experimental data often produce points $(x_1, y_1), \dots, (x_n, y_n)$ that,

¹ This notation is commonly used for least-squares lines instead of $y = mx + b$.

when graphed, seem to lie close to a line. We want to determine the parameters β_0 and β_1 that make the line as “close” to the points as possible.

Suppose β_0 and β_1 are fixed, and consider the line $y = \beta_0 + \beta_1 x$ in Figure 1. Corresponding to each data point (x_j, y_j) there is a point $(x_j, \beta_0 + \beta_1 x_j)$ on the line with the same x -coordinate. We call y_j the *observed* value of y and $\beta_0 + \beta_1 x_j$ the *predicted* y -value (determined by the line). The difference between an observed y -value and a predicted y -value is called a *residual*.

There are several ways to measure how “close” the line is to the data. The usual choice (primarily because the mathematical calculations are simple) is to add the squares of the residuals. The **least-squares line** is the line $y = \beta_0 + \beta_1 x$ that minimizes the sum of the squares of the residuals. This line is also called a **line of regression of y on x** , because any errors in the data are assumed to be only in the y -coordinates. The coefficients β_0, β_1 of the line are called (linear) **regression coefficients**.²

Actual **y** vector

Estimated **y** vector

residual vector ϵ , defined by $\epsilon = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$

$e_j = y_j - \hat{y}_j$, where $\hat{y}_j = \text{estimated } y_j$

Least Squares Solution to linear regression

If the data points were on the line, the parameters β_0 and β_1 would satisfy the equations

Predicted y-value		Observed y-value
$\beta_0 + \beta_1 x_1$	=	y_1
$\beta_0 + \beta_1 x_2$	=	y_2
\vdots		\vdots
$\beta_0 + \beta_1 x_n$	=	y_n

We can write this system as

$$X\boldsymbol{\beta} = \mathbf{y}, \quad \text{where } X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad (1)$$

Of course, if the data points don't lie on a line, then there are no parameters β_0, β_1 for which the predicted y-values in $X\boldsymbol{\beta}$ equal the observed y-values in \mathbf{y} , and $X\boldsymbol{\beta} = \mathbf{y}$ has no solution. This is a least-squares problem, $A\mathbf{x} = \mathbf{b}$, with different notation!

The square of the distance between the vectors $X\boldsymbol{\beta}$ and \mathbf{y} is precisely the sum of the squares of the residuals. The $\boldsymbol{\beta}$ that minimizes this sum also minimizes the distance between $X\boldsymbol{\beta}$ and \mathbf{y} . *Computing the least-squares solution of $X\boldsymbol{\beta} = \mathbf{y}$ is equivalent to finding the $\boldsymbol{\beta}$ that determines the least-squares line in Figure 1.*

Least Squares Solution:

From

$$X\boldsymbol{\beta} = \mathbf{y}$$

Then pre-multiply by X^T to get the normal equation,

$$X^T X \boldsymbol{\beta} = X^T \mathbf{y}$$

Then premultiply by $(X^T X)^{-1}$,

$$\begin{aligned} (X^T X)^{-1} (X^T X) \boldsymbol{\beta} &= (X^T X)^{-1} X^T \mathbf{y} \\ \boldsymbol{\beta} &= (X^T X)^{-1} X^T \mathbf{y} \end{aligned}$$

Example 1

EXAMPLE 1 Find the equation $y = \beta_0 + \beta_1 x$ of the least-squares line that best fits the data points $(2, 1)$, $(5, 2)$, $(7, 3)$, and $(8, 3)$.

SOLUTION Use the x -coordinates of the data to build the design matrix X in (1) and the y -coordinates to build the observation vector \mathbf{y} :

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix}$$

For the least-squares solution of $X\boldsymbol{\beta} = \mathbf{y}$, obtain the normal equations (with the new notation):

$$X^T X \boldsymbol{\beta} = X^T \mathbf{y}$$

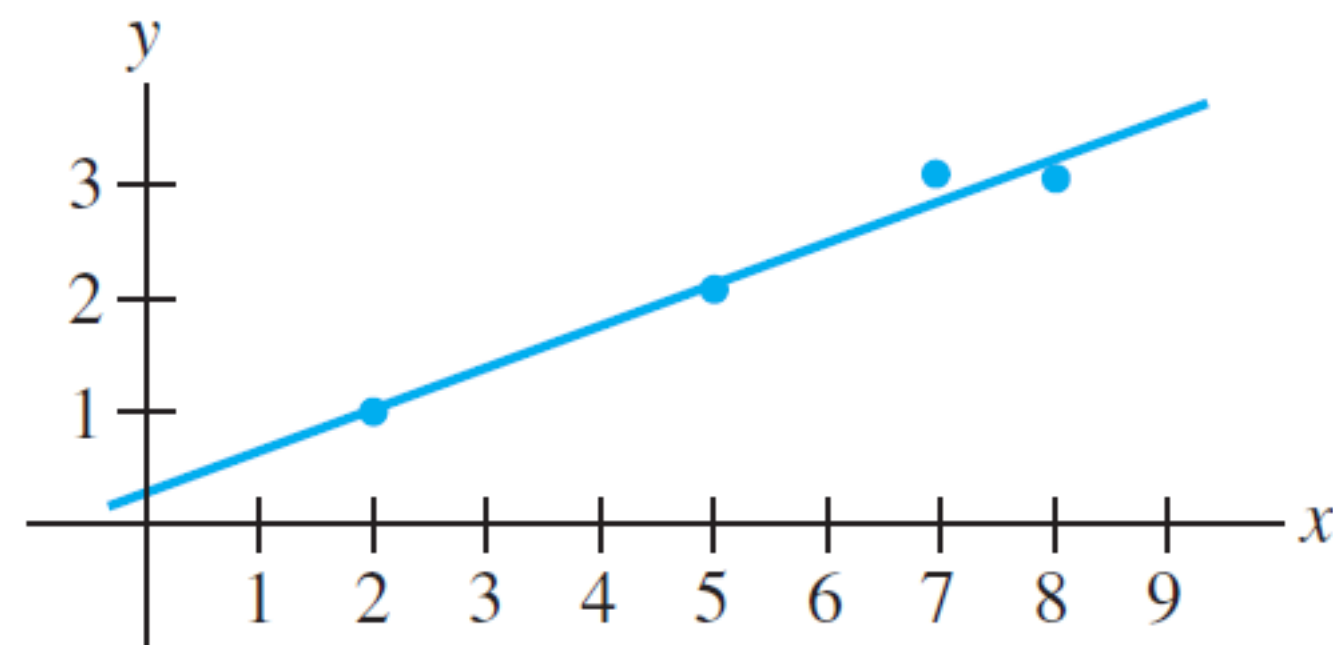


FIGURE 2 The least-squares line
 $y = \frac{2}{7} + \frac{5}{14}x$.

That is, compute

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix}$$
$$X^T \mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 9 \\ 57 \end{bmatrix}$$

The normal equations are

$$\begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 9 \\ 57 \end{bmatrix}$$

Hence

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix}^{-1} \begin{bmatrix} 9 \\ 57 \end{bmatrix} = \frac{1}{84} \begin{bmatrix} 142 & -22 \\ -22 & 4 \end{bmatrix} \begin{bmatrix} 9 \\ 57 \end{bmatrix} = \frac{1}{84} \begin{bmatrix} 24 \\ 30 \end{bmatrix} = \begin{bmatrix} 2/7 \\ 5/14 \end{bmatrix}$$

Thus the least-squares line has the equation

$$y = \frac{2}{7} + \frac{5}{14}x$$

Least Squares Fit to Other Curves (X)

Least-Squares Fitting of Other Curves

When data points $(x_1, y_1), \dots, (x_n, y_n)$ on a scatter plot do not lie close to any line, it may be appropriate to postulate some other functional relationship between x and y .

The next two examples show how to fit data by curves that have the general form

$$y = \beta_0 f_0(x) + \beta_1 f_1(x) + \dots + \beta_k f_k(x) \quad (2)$$

where f_0, \dots, f_k are known functions and β_0, \dots, β_k are parameters that must be determined. As we will see, equation (2) describes a linear model because it is linear in the unknown parameters.

Linear in the parameter model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

residual vector $\boldsymbol{\epsilon}$, defined by $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$

Any equation of this form is referred to as a **linear model**. Once \mathbf{X} and \mathbf{y} are determined, the goal is to minimize the length of $\boldsymbol{\epsilon}$, which amounts to finding a least-squares solution of $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$. In each case, the least-squares solution $\hat{\boldsymbol{\beta}}$ is a solution of the normal equations

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$$

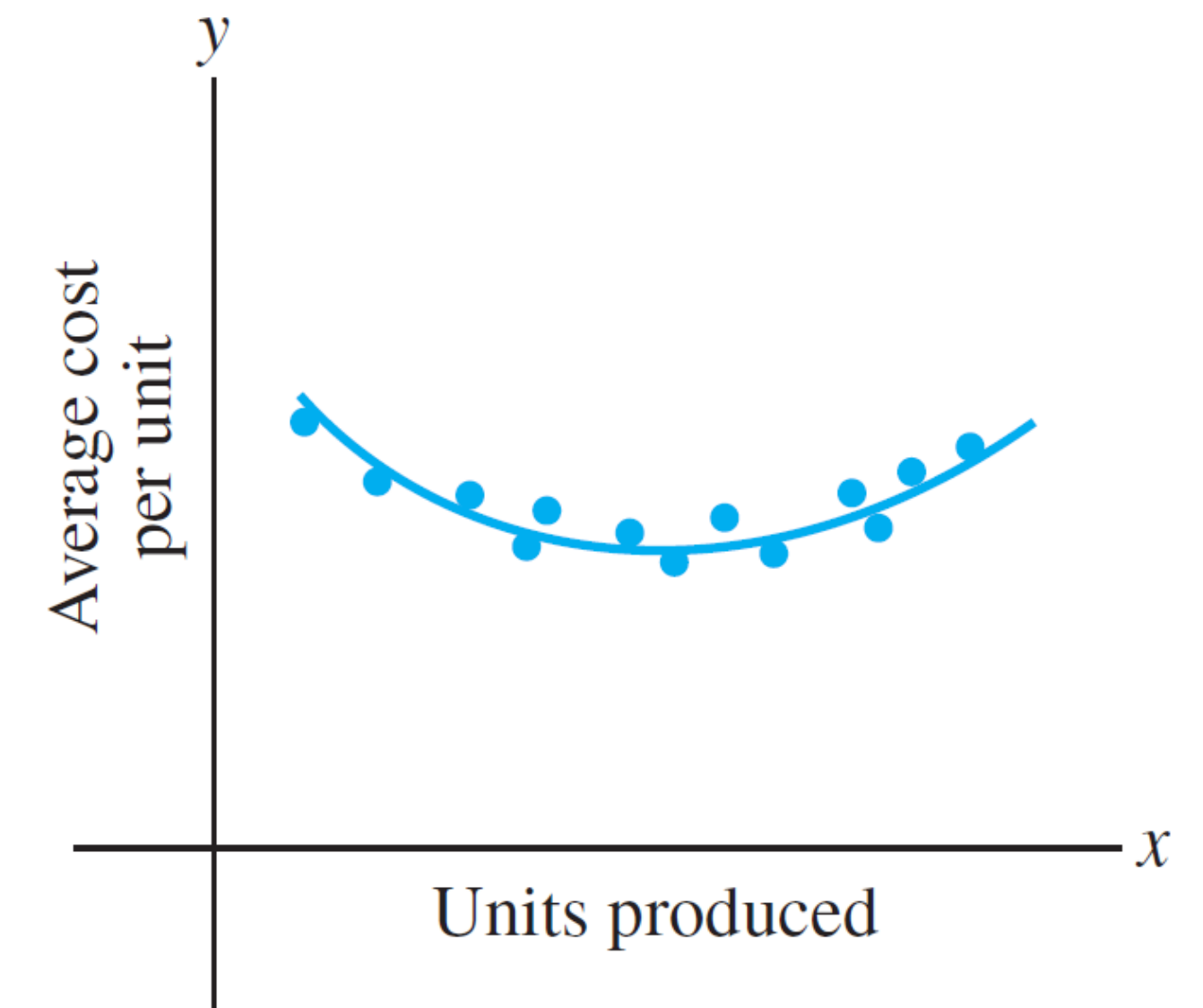


FIGURE 3

Average cost curve.

Example 2: fitting to curves (X)

EXAMPLE 2 Suppose data points $(x_1, y_1), \dots, (x_n, y_n)$ appear to lie along some sort of parabola instead of a straight line. For instance, if the x -coordinate denotes the production level for a company, and y denotes the average cost per unit of operating at a level of x units per day, then a typical average cost curve looks like a parabola that opens upward (Figure 3). In ecology, a parabolic curve that opens downward is used to model the net primary production of nutrients in a plant, as a function of the surface area of the foliage (Figure 4). Suppose we wish to approximate the data by an equation of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{3}$$

Describe the linear model that produces a “least-squares fit” of the data by equation (3).

SOLUTION Equation (3) describes the ideal relationship. Suppose the actual values of the parameters are $\beta_0, \beta_1, \beta_2$. Then the coordinates of the first data point (x_1, y_1) satisfy an equation of the form

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon_1$$

where ϵ_1 is the residual error between the observed value y_1 and the predicted y -value $\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$. Each data point determines a similar equation:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \epsilon_n \end{aligned}$$

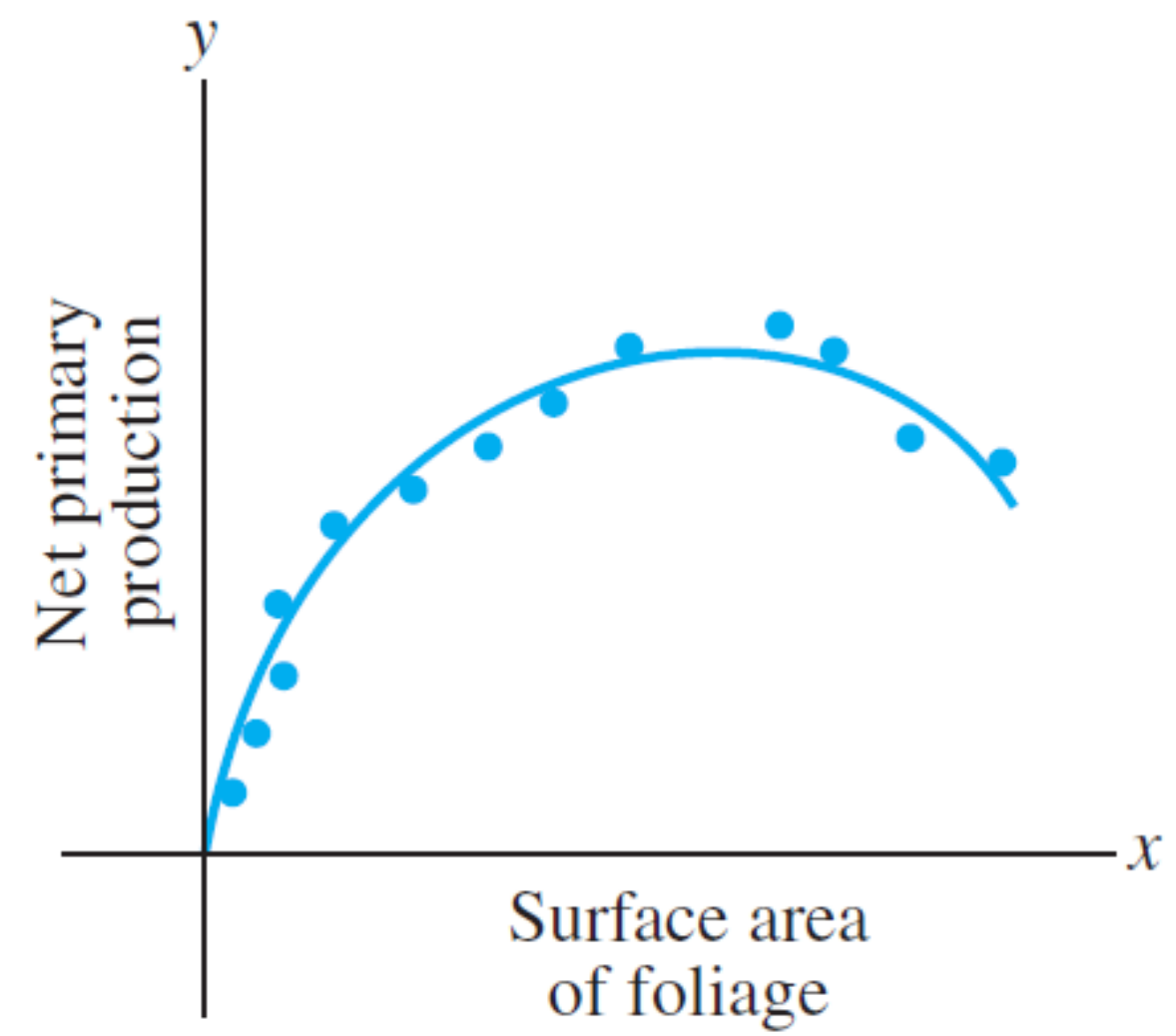


FIGURE 4
Production of nutrients.

It is a simple matter to write this system of equations in the form $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. To find X , inspect the first few rows of the system and look for the pattern.

$$\begin{aligned} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\ \mathbf{y} &= X \boldsymbol{\beta} + \boldsymbol{\epsilon} \end{aligned}$$

Observation
vector

Design
matrix

Parameter
vector

Residual
vector