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

CX1104: **Linear Algebra for Computing**

Chap. No : **7.1.3**

Lecture : **Least Squares**

Topic : **Solving the Least Squares Problem**

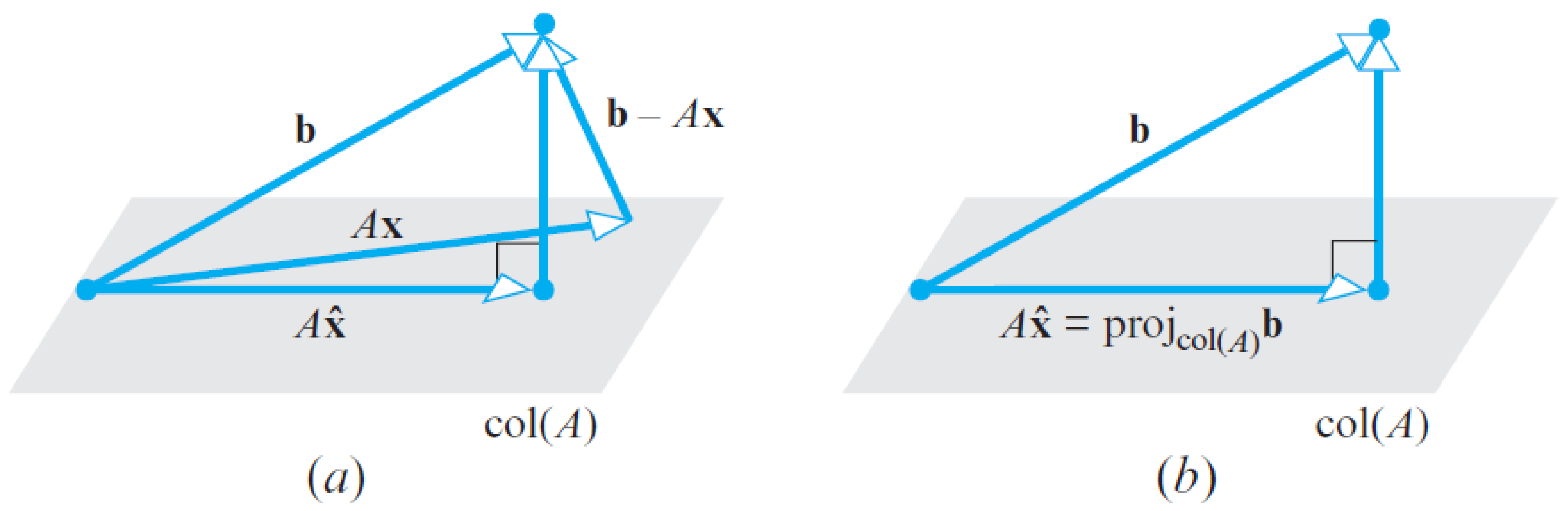
Concept : **Best Approx. Theorem and Normal Equation**

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

TAs: **Zhang Su, Vishal Choudhari**

Rev: 2nd July 2020

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

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

Anton:

6.4 Best Approximation; Least Squares 379

The Normal Equation

$$Ax = b$$

Multiplying both sides by A^T

$$\boxed{A^TA\mathbf{x} = A^T\mathbf{b}}$$

Normal Equation!

The set of least-squares solutions of $A\mathbf{x} = \mathbf{b}$ coincides with the nonempty set of solutions of the normal equations $A^TA\mathbf{x} = A^T\mathbf{b}$.

Proved in the next slide!

Why called "Normal"?
Ref: https://mathworld.wolfram.com/NormalEquation.html

Solution of the General Least-Squares Problem

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

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

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

$$A\hat{\mathbf{x}} = \hat{\mathbf{b}} \tag{1}$$

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

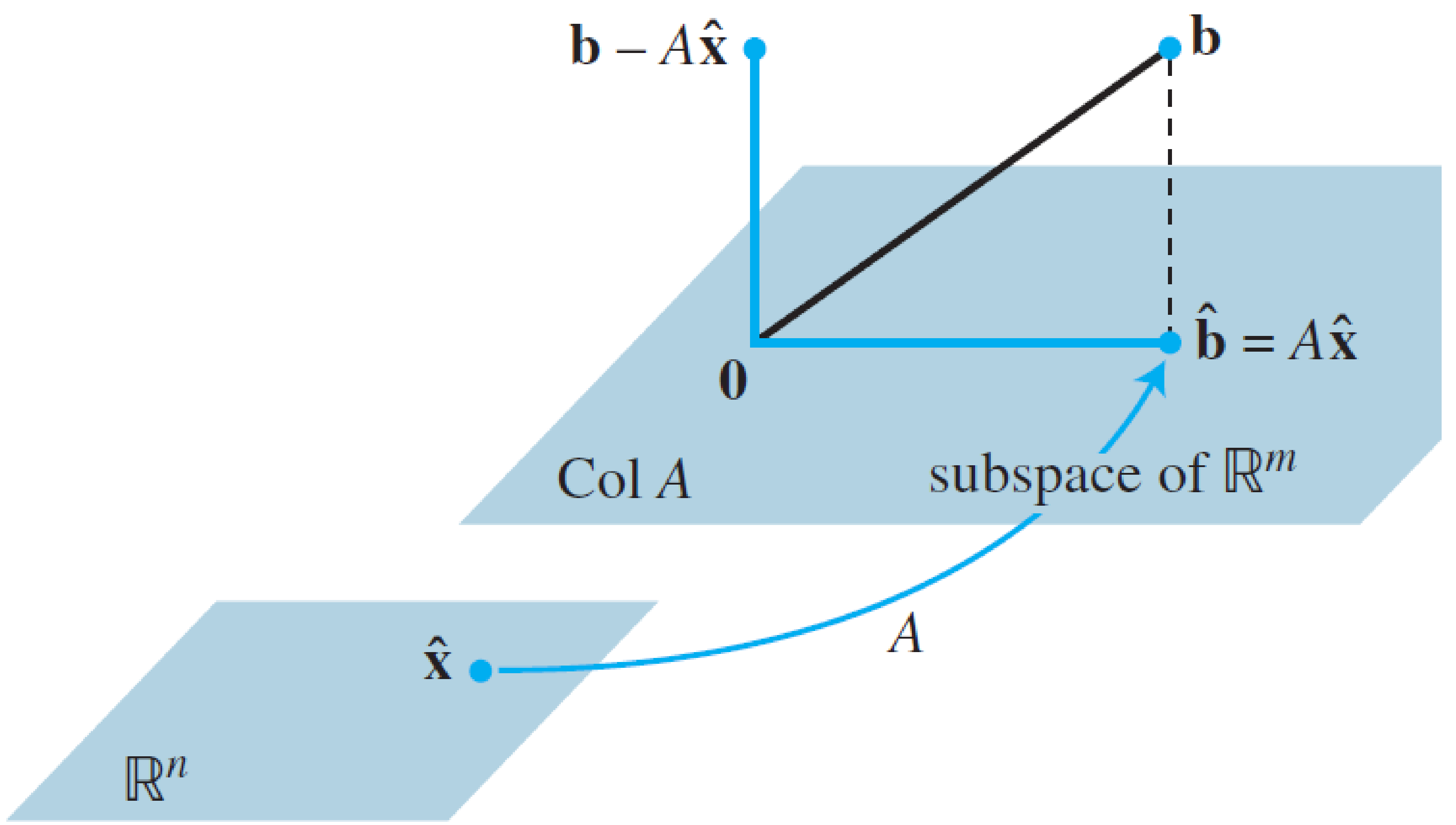


FIGURE 2 The least-squares solution $\hat{\mathbf{x}}$ is in $\mathbb{R}^n$.

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 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,

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

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

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

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

$$A^TA\mathbf{x} = A^T\mathbf{b} \tag{3}$$

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 $\hat{\mathbf{x}}$.

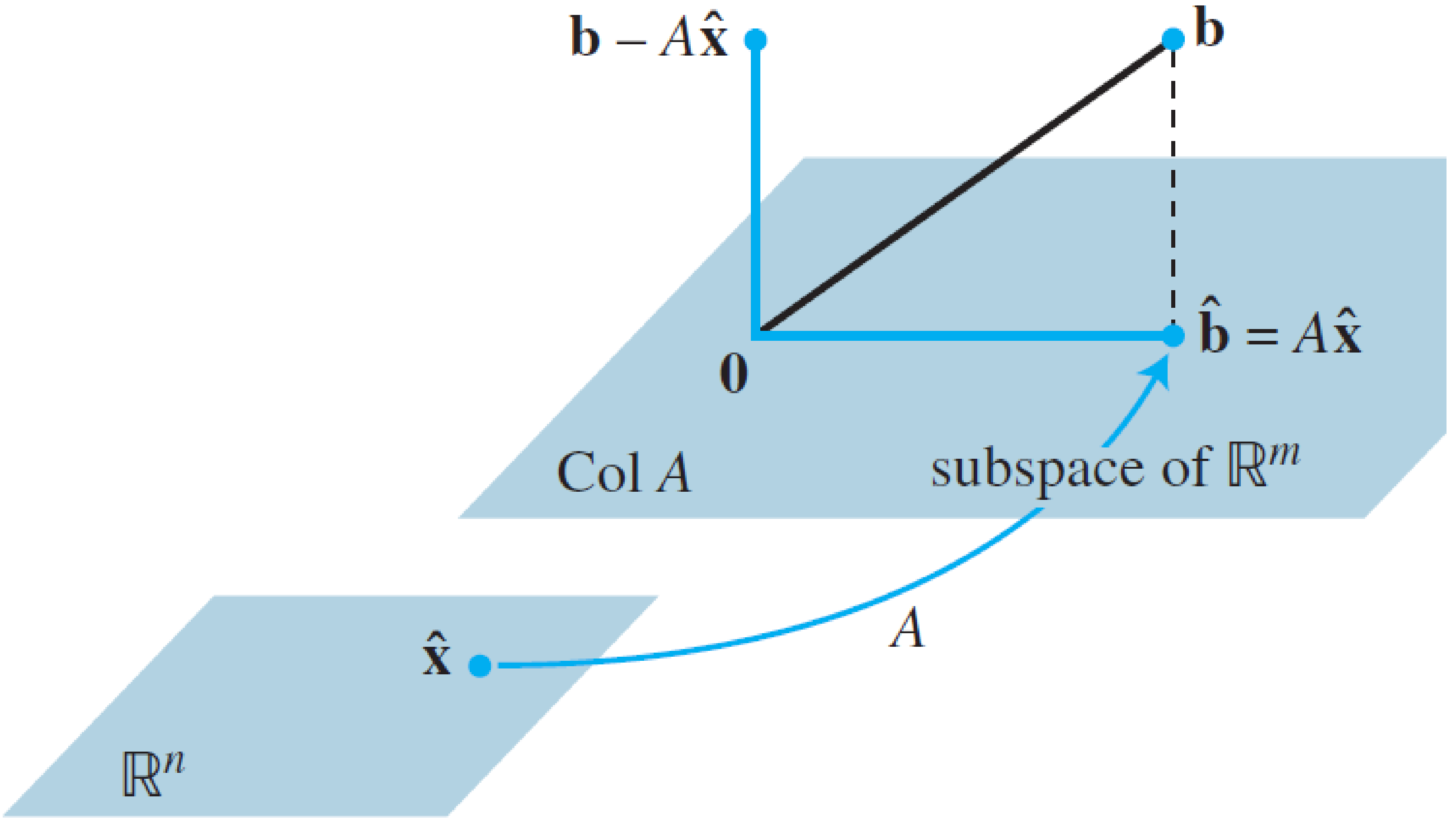


FIGURE 2 The least-squares solution $\hat{\mathbf{x}}$ is in $\mathbb{R}^n$.

THEOREM 14

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

a. The equation $A\mathbf{x} = \mathbf{b}$ has a unique least-squares solution for each $\mathbf{b}$ in $\mathbb{R}^m$.

b. The columns of A are linearly indpendent.

c. The matrix A^TA is invertible.

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

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

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

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:

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

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

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

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

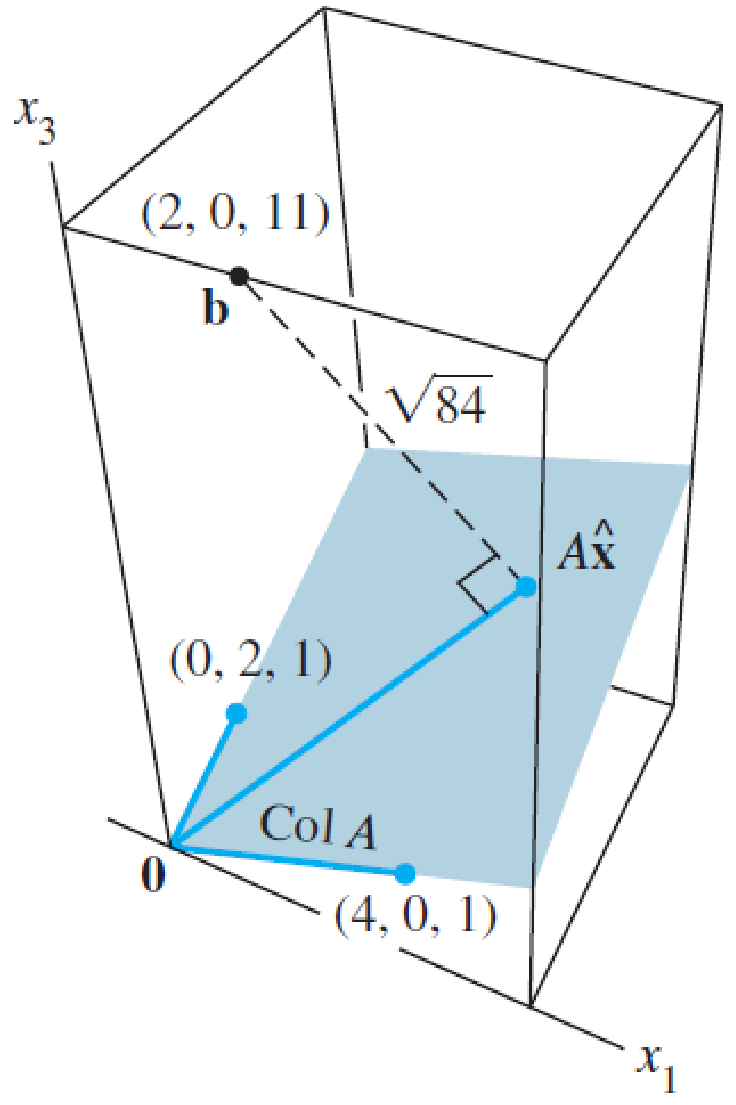


FIGURE 3

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

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

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

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

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

In many calculations, A^TA 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 $A\mathbf{x} = \mathbf{b}$.

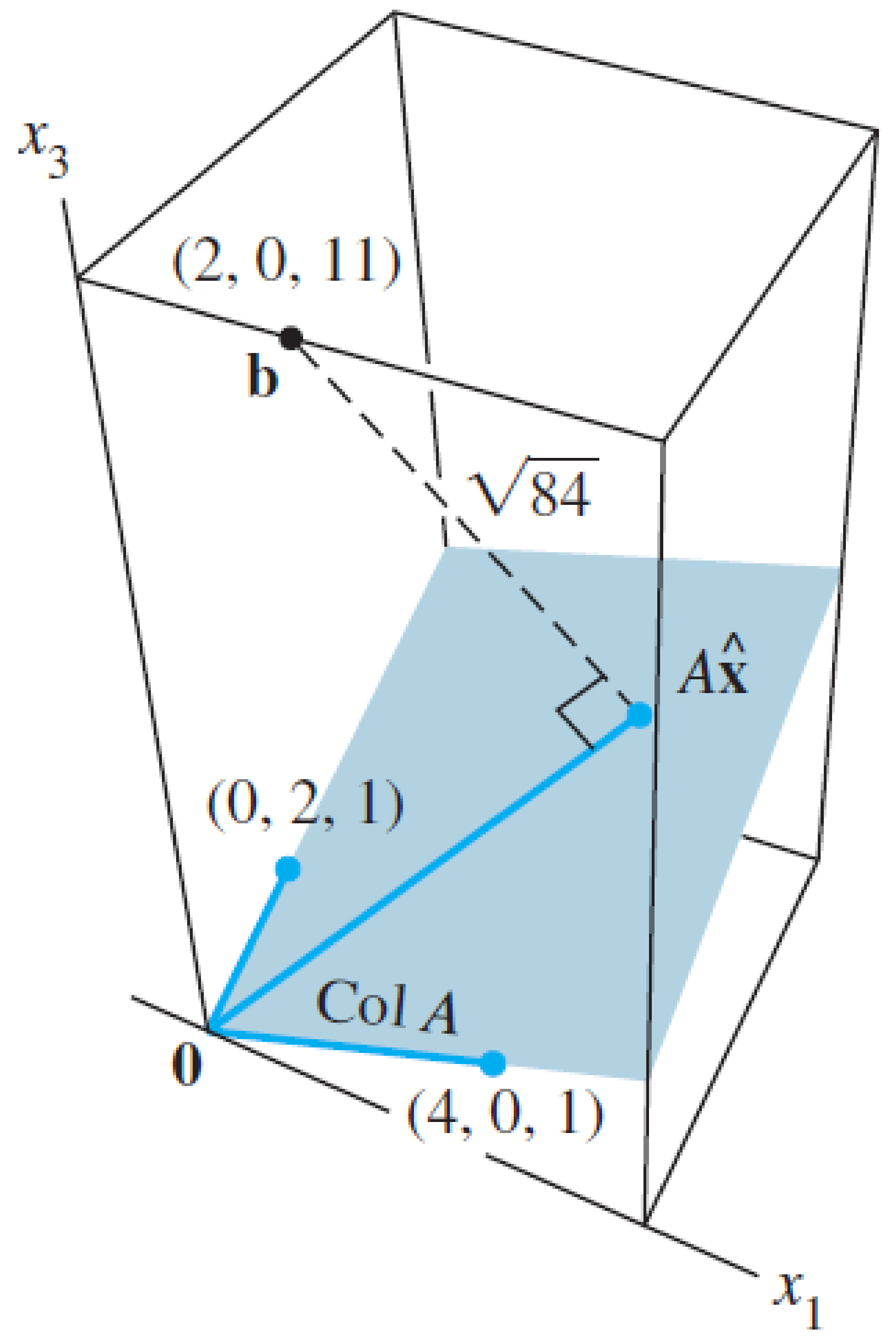


FIGURE 3

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. ■

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

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

$$A^TA = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 6 & 2 & 2 & 2 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{bmatrix}}_{A^TA}$$

Note that A^TA is always a square matrix.

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

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

6.5 Least-Squares Problems **361**

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

Reduced to

$$\left[\begin{array}{cccc|c} \multicolumn{4}{c}{} & \\ 6 & 2 & 2 & 2 & 4 \\ 2 & 2 & 0 & 0 & -4 \\ 2 & 0 & 2 & 0 & 2 \\ 2 & 0 & 0 & 2 & 6 \end{array}\right] \sim \begin{bmatrix} 1 & 0 & 0 & 1 & 3 \\ 0 & 1 & 0 & -1 & -5 \\ 0 & 0 & 1 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

A^TA $\quad$ A^Tb

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^TA is not invertible (its determinant is 0).

A^TA 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 inverser (A'*A):");
inv(A'*A)
disp("using pinverser (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
```

$A^T A$

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

```
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 $\Longrightarrow \infty$

```
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.