Least squares

Linear Algebra

Department of Computer Engineering

Sharif University of Technology

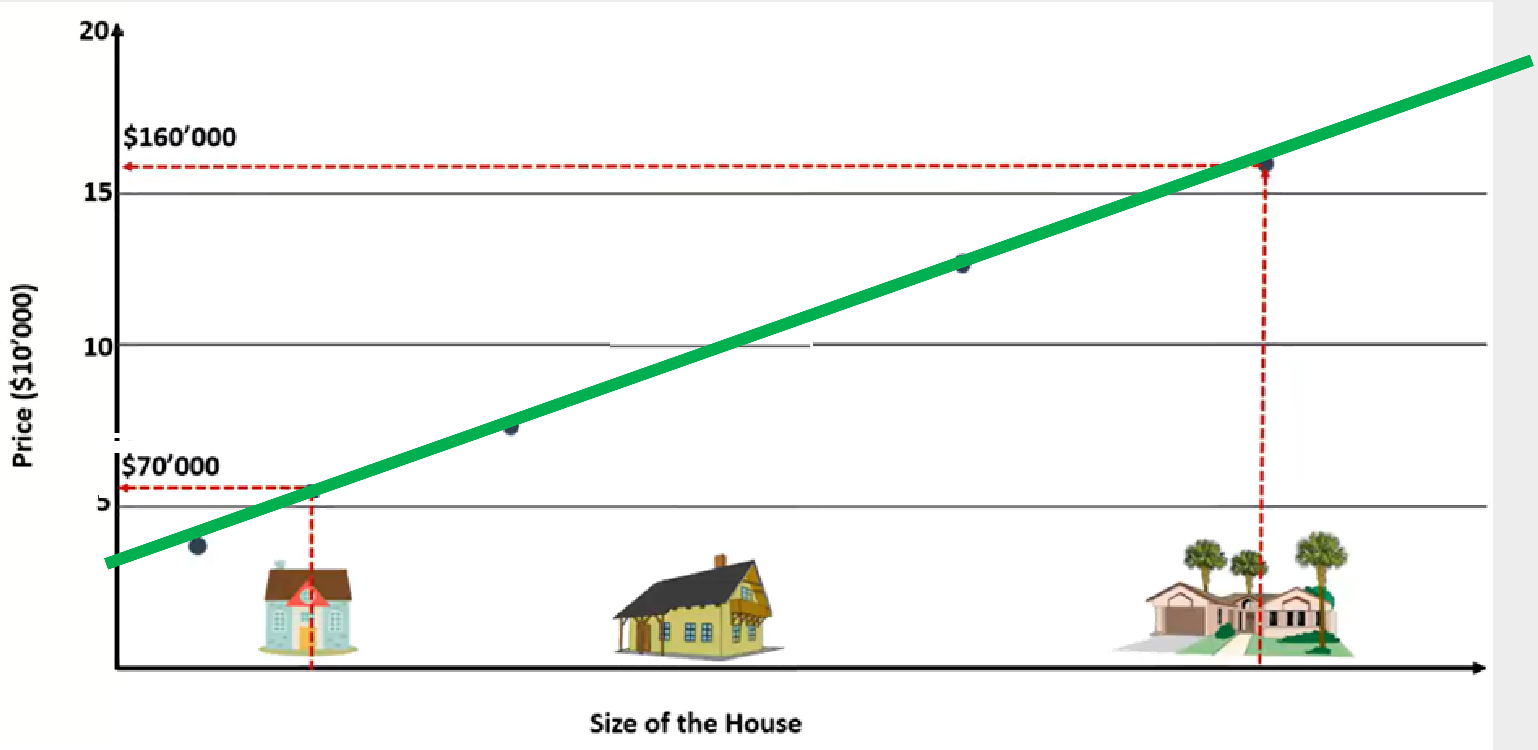
Hamid R. Rabiee rabiee@sharif.edu

Maryam Ramezani maryam.ramezani@sharif.edu

Introduction

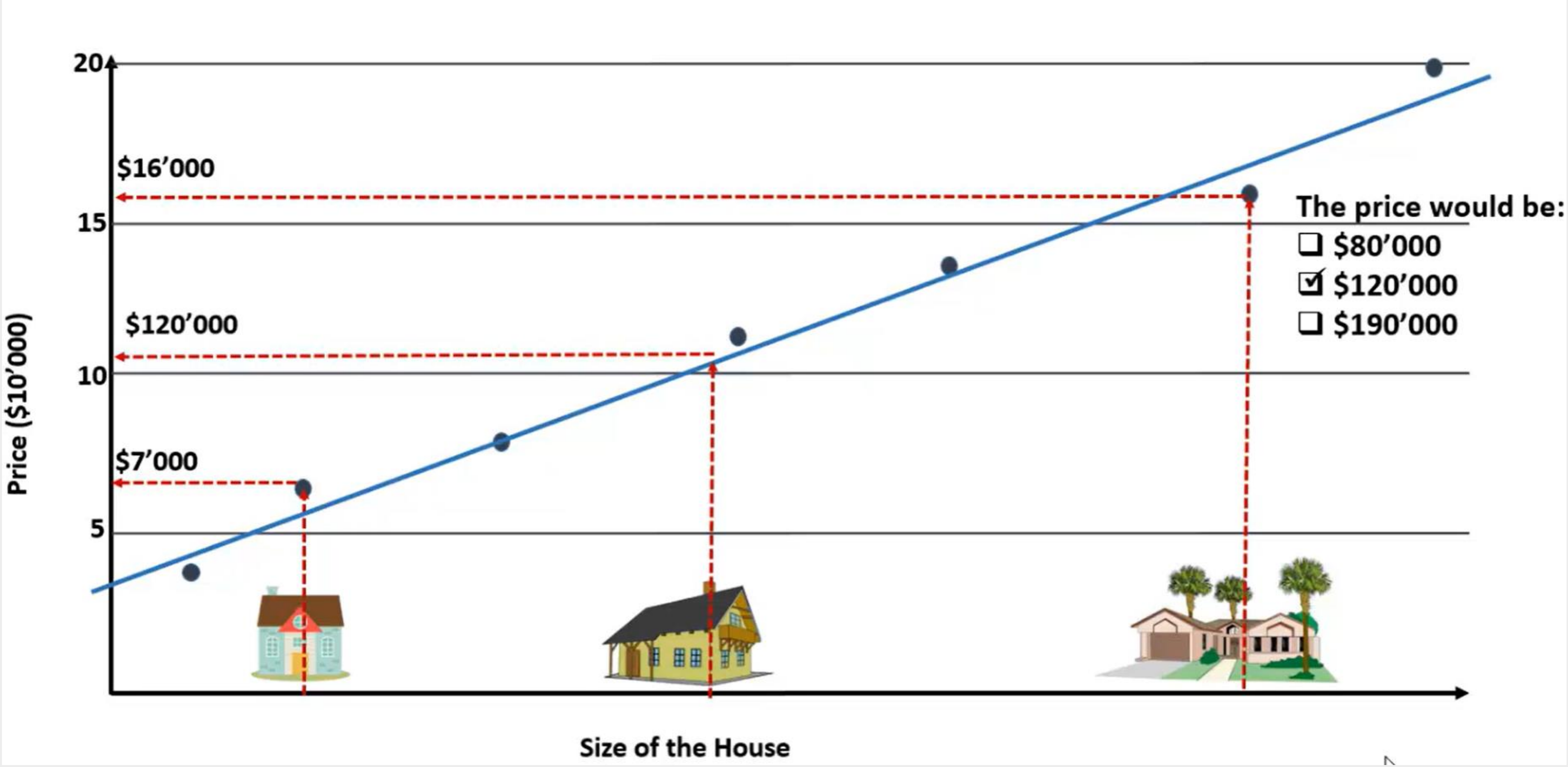
Linear Equation

$Ax = b$ has solution.

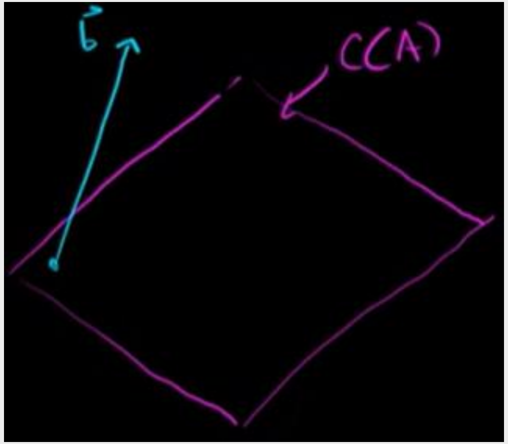


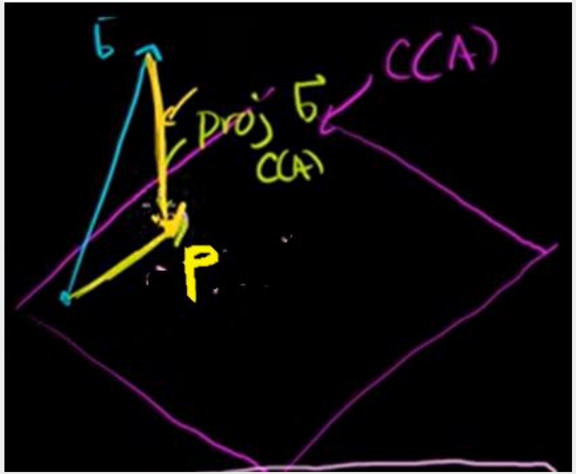
Least Squares Error Correction

$Ax = b$ has no solution.



- A is $m \times n$ matrix
- $Ax = b$ has no solution -> b is not in the $C(A)$ why?





How to solve the problem?

- Bad News: $Ax = b$ has no solution
- Good News: $A\hat{x} = p$ has solution
  - Unique
  - Many

Review

- 4 Subspaces:
  - Column Space $C(A)$
  - Null Space $N(A)$
  - Row Space $C(A^T)$
  - Null Space of A^T= Left Null Space of A = $\mathrm{N}(A^T)$

Theorem

- Orthogonality of the Row Space and the Null Space
- Orthogonality of the Column Space and the Left Null Space

Review

- Projection a vector on a vector
    - Column space of matrix?
    - Rank of matrix?
    - Is the matrix symmetric?
    - Power two of this matrix?
- Projection a vector on a plane

Fill this page with my notes on the board 😁

Lets think!!!

$$P = A(A^T A)^{-1} A^T$$

- Think about Ps when:
  - s is in the column space of A
  - s is in the orthogonal complement space of A

  - Geometry?
  - Math?

Fill this page with my notes on the board 😁

Good News!

- Fill this page with my notes on the board 😁
  - Least square in $\mathbb{R}^2$ and regression!!!
  - Error
  - Outlier

Look another way!!

- $(A^T A)^{-1} A^T$ is the left inverse of A
- $A(A^T A)^{-1} A^T$ is the projection matrix on C(A)

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

What will happen when A is an invertible matrix?

Another problem?

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

Theorem

- If A has linearly independent columns, then $A^T A$ is invertible.

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

$$= A^{\dagger} b$$

pseudo-inverse of a left-invertible matrix

Therefore, when $A^T A$ is invertible, $\hat{x}$ is the unique solution. This often happens when for D number of variables and N number of equations, we have $D \ll N$.

What will happen when $A^T A$ is not an invertible matrix? (when N $< D$)

When A^TA is not an invertible matrix?

X^TX will not be invertible when $N < D$. To illustrate why we have infinite number of solutions, consider in a two-dimensional problem $(D = 2)$ we have only one training sample $\boldsymbol{x}_1 = [1, -1], y_1 = 1$. We can see $\boldsymbol{w} = [a + 1, a]$ for any $a \in \mathbb{R}$ will get 0 training error:

$$\boldsymbol{w}^T\boldsymbol{x}_1 = a + 1 - a = 1 = y_1.$$

This is true for any problem with $N < D$—in this case, you can always find a vector in the null space of X (a vector such that $X\boldsymbol{v} = 0$), and then for a solution $\boldsymbol{w}^*$, any vector with $\boldsymbol{w}^* + a\boldsymbol{v}$ with $a \in \mathbb{R}$ will get the same square error with $\boldsymbol{w}^*$. This case $(N < D)$ is also called the **under-determined** problem, since you have too many degree of freedom in your problem and don't have enough constraints (data).

Good News!!! (When A^TA is not an invertible matrix)

$$\hat{x} = (A^TA)^{-1}A^Tb$$

will have infinite number of solutions in this case

In fact, given any real $m \times n$-matrix A, there is always a unique x^+ of minimum norm that minimizes $\|Ax - b\|^2$, even when the columns of A are linearly dependent.

the following approach to find the **minimum-norm solution** $\boldsymbol{w}^+$: Let $\mathcal{W} = \operatorname{argmin}_{\boldsymbol{w}} \|X\boldsymbol{w} - \boldsymbol{y}\|^2$ denote the set of solutions, we aim to find the minimum norm solution that

$$\boldsymbol{w}^+ = \operatorname*{argmin}_{\boldsymbol{w} \in \mathcal{W}} \|\boldsymbol{w}\|_2. \tag{4}$$

SVD Next slide

Least squares problem

Theorem

- given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, find vector $x \in \mathbb{R}^n$ that minimizes

$$||Ax - b||^2 = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} A_{ij}x_j - b_i\right)^2$$

- ''least squares'' because we minimize a sum of squares of affine functions:

$$||Ax - b||^2 = \sum_{i=1}^{m} r_i(x)^2, \qquad r_i(x) = \sum_{j=1}^{n} A_{ij}x_j - b_i$$

- the problem is also called the linear least squares problem

Least squares and linear equations

Important

$$\text{minimize } \left|\left|Ax - b\right|\right|^2$$

solution of the least squares problem: any $\hat{x}$ that satisfies

- $$||A\,\hat{x} - b|| \le ||Ax - b|| \quad \text{for all } x$$

Note

$\hat{r} = A\hat{x} - b$ is the residual vector

if $\hat{r} = 0$, then $\hat{x}$ solves the linear equation $Ax = b$

if $\hat{r} \neq 0$, then $\hat{x}$ is a least squares approximate solution of the equation

in most least squares applications, $m > n$ and $Ax = b$ has no solution

Solving least squares problems (Method 1)

Example

- Normal equations of the least squares problem $A^T A x = A^T b$
  - Coefficient matrix $A^T A$ is the ......
  - Equivalent to $\nabla f(x) = 0$ where $f(x) =$
  - All solutions of the least squares problem satisfy the normal equations

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

Look at board I am writing in vector and matrix form with derivation

Example

- Rewrite least squares solution using QR factorization $A = QR$

- Complexity: $2mn^2$

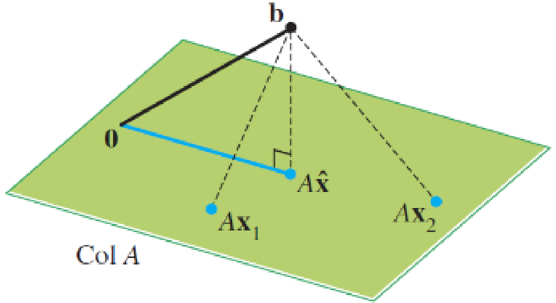
Algorithm: Least squares via QR factorization

Input: $A : m \times n$ left-invertible
Input: $b : m \times 1$
output: $x_{LS} : n \times 1$
Find QR factorization $A = QR$
Compute $Q^T b$
Solve $Rx_{LS} = Q^T b$ using back substitution

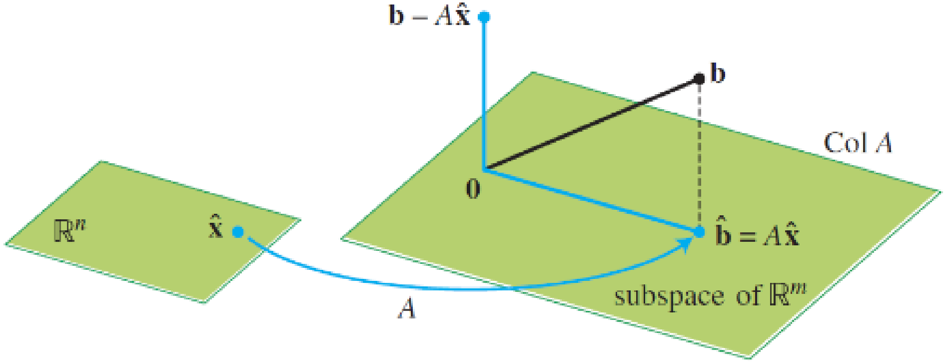
- Identical to algorithm for solving $Ax = b$ for square invertible A, but when A is tall, gives least squares approximate solution

Note

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



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



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

Solution of a least squares problem

Theorem

- A has linearly independent columns, then below vector is the unique solution of the least squares problem

$$\text{minimize } ||Ax - b||^2$$

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

$$= A^{\dagger} b$$

pseudo-inverse of a left-invertible matrix

- Proof?

Solving least squares problems

Example

a 3×2 matrix with “almost linearly dependent” columns

$$A = \begin{bmatrix} 1 & -1 \\ 0 & 10^{-5} \\ 0 & 0 \end{bmatrix}, \qquad b = \begin{bmatrix} 0 \\ 10^{-5} \\ 1 \end{bmatrix},$$

round intermediate results to 8 significant decimal digits

- Solve using both methods
  - Which one is more stable? Why?

Review: Linear-in-parameters model

Note

- we choose the model $\hat{f}(x)$ from a family models

$$\hat{f}(x) = \theta_1 f_1(x) + \theta_2 f_2(x) + \cdots + \theta_p f_p(x)$$

model parameters

scalar valued basis functions (chosen by us)

Solution of weighted least squares

Example

weighted least squares is equivalent to a standard least squares problem

-

$$\text{minimize} \quad \left\| \begin{bmatrix} \sqrt{\lambda_1}A_1 \\ \sqrt{\lambda_2}A_2 \\ \vdots \\ \sqrt{\lambda_k}A_k \end{bmatrix} x - \begin{bmatrix} \sqrt{\lambda_1}b_1 \\ \sqrt{\lambda_2}b_2 \\ \vdots \\ \sqrt{\lambda_k}b_k \end{bmatrix} \right\|^2$$

- Solution is unique if the $stacked\ matrix$ has linearly independent columns
- Each matrix A_i may have linearly dependent columns (or be a wide matrix)
- if the stacked matrix has linearly independent columns, the solution is

$$\hat{x} = \left(\lambda_1 A_1^T A_1 + \cdots + \lambda_k A_k^T A_k\right)^{-1} (\lambda_1 A_1^T b_1 + \cdots + \lambda_k A_k^T b_k)$$

Lagrange multiplier

Example

$$f(x) = \min(x_1 x_2)$$

$$g(x) = 1 - \mathrm{x}_1 - \mathrm{x}_2$$

$$g(x) = 0$$

$$L(x, \lambda) = f(x) + \lambda g(x)$$

$$\nabla_x f(x) = 0$$

Constrained Least Square

Example

- $\begin{cases} \min\limits_{x} ||Ax-b||^2 & A: m \times n \\ s.t. \quad Cx = d & C: p \times n \end{cases}$

$$L(x,\lambda) = ||Ax-b||^2 + \lambda^T(Cx-d)$$

$$\begin{cases} \nabla_x L = 2A^T Ax - 2A^T b + C^T \lambda = 0 \\ \nabla_\lambda L = Cx - d = 0 \end{cases} \rightarrow \begin{bmatrix} 2A^T A & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} 2A^T b \\ d \end{bmatrix}$$

Note

- #equations: $n+p$ #Unkowns: $n+p$
- KKT equations
- Least Square problem is a KKT problem with $A = I, b = 0$

Least Squares Regression

Note

- Remember the regression model (affine function) :

$$\hat{f}(x) = x^T\beta + v$$

- The prediction error for example i is:

$$\begin{aligned} r^{(i)} &= y^{(i)} - \hat{f}\left(x^{(i)}\right) \\ &= y^{(i)} - \left(x^{(i)}\right)^T\beta - v \end{aligned}$$

- The MSE is :

$$\frac{1}{N}\sum_{i=1}^{N}(r^{(i)})^2 = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \left(x^{(i)}\right)^T\beta - v)^2$$

- choose the model parameters v, β that minimize the MSE

$$\frac{1}{N}\sum_{i=1}^{N} \left(y^{(i)} - \left(x^{(i)}\right)^T \beta - v\right)^2$$

this is the least square problem: minimize $\left\|A\theta - y^d\right\|^2$ with

$$A = \begin{bmatrix} 1 & \left(x^{(1)}\right)^T \\ 1 & \left(x^{(2)}\right)^T \\ \vdots & \vdots \\ 1 & \left(x^{(N)}\right)^T \end{bmatrix}, \qquad \theta = \begin{bmatrix} v \\ \beta \end{bmatrix}, \qquad y^d = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}$$

we write the solution as $\hat{\theta} = (\hat{v}, \hat{\beta})$

Least Squares Regression

Example

$$\hat{f}(x) = \theta_1 + \theta_2 x + \theta_3 x^2 + \ldots + \theta_p x^{p-1}$$

- a linear-in-parameters model with basis functions…….
- least squares model fitting in matrix notation?