Mathematical Foundations for Machine Learning and Data Science

Least-Squares (LS)

Dr. Zubair Khalid
Department of Electrical Engineering
School of Science and Engineering
Lahore University of Management Sciences

https://www.zubairkhalid.org/ee212_2021.html
Outline

- Least-Squares (LS) Formulation
- Formulation of Regression problem as Least-squares problem
- LS Geometric Interpretation
- LS Solution
- Regularized LS
Least-Squares

Formulation:

- We want to find \( x \in \mathbb{R}^n \) given a matrix \( A \in \mathbb{R}^{m \times n} \) and \( y \in \mathbb{R}^m \) related by

\[
y = Ax
\]

- We consider \( m > n \) (over-determined system).

- If solution exists, then the solution is given by \( x = Xy \)
  - where \( X \) is the left inverse of \( A \)

- If solution does not exist. It means there does not exist any \( x \in \mathbb{R}^n \) for which \( Ax = y \).
  - What do we mean by this?
    - Mathematically, \( y \) does not belong to the column space of \( A \).
    - For example, it corresponds to the case when \( y \) corresponds to observations that have not been measured accurately.
    - Recall, each equation of \( Ax = y \) corresponds to hyper-plane. The solution exists if all \( m \) hyper-planes intersect at one point.
Least-Squares

Formulation:

- We want to find \( x \in \mathbb{R}^n \) given a matrix \( A \in \mathbb{R}^{m \times n} \) and \( y \in \mathbb{R}^m \) related by
  \[ y = Ax \]
- No \( x \in \mathbb{R}^n \) for which \( Ax = y \).

How do we handle this case?

- Find \( \hat{x} \in \mathbb{R}^n \) such that \( A\hat{x} \) is closest to \( y \)
- Closest in what sense?
  - Euclidean distance between \( A\hat{x} \) and \( y \) is minimized.
- Define \( r \) as the distance between \( A\hat{x} \) and \( y \) (also known as residual error)
- Minimizing Euclidean distance means minimizing
  \[ \|r\|_2 = \sqrt{\sum_{i=1}^{m} r_i^2} \quad \text{or} \quad \|r\|_2^2 = \sum_{i=1}^{m} r_i^2 \]

Since the solution \( \hat{x} \) minimizes sum of squares of residual error (along each component), it is referred to as least-squares solution.
Least-Squares

Formulation – Optimization Problem:

- Finding Least-squares solution can be represented in the form of following optimization problem.

\[ \hat{x} = \text{minimize} \quad \|Ax - y\|^2_2 \]

Least-squares (LS) objective function
Least-Squares

Application: Linear Regression/Data Fitting in ML and Data Science

- Example: we want to find a relationship between temperature $t$ and electricity consumption $y$ in a town or we want to develop a model which relates electricity consumption and temperature.

In ML, this is Regression: Build a model for Quantitative Prediction on a continuous scale

Here, PROCESS or SYSTEM refers to any underlying physical or logical phenomenon which maps our input data to our observed and noisy output data.
Least-Squares

Application: Linear Regression/Data Fitting in ML and Data Science

First Step – Model Assumption

We assume there is an inherent but unknown relationship between input and output.

\[ y = f(t) = x_1 t + x_2 + n \]

Goal: Given noisy observations, we need to estimate the unknown functional relationship as accurately as possible.

- Learn parameters \( x_1 \) and \( x_2 \) describing our model.
Least-Squares

Application: Linear Regression/Data Fitting in ML and Data Science

**Second Step – Collect Data**

\[ y = f(t) = x_1 t + x_2 + n \]

![Diagram showing the process of collecting data for linear regression](image)

**Training Data**

- First Data Sample: \( \{t(1), y(1)\} \)
- Second Data Sample: \( \{t(2), y(2)\} \)
- \( \ldots \)
- \( m\)-th Data Sample: \( \{t(m), y(m)\} \)
Least-Squares

Application: Linear Regression/Data Fitting in ML and Data Science

Third Step – LS Problem Formulation

- These measurements can be represented in matrix form as

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_m
\end{bmatrix} = \begin{bmatrix}
  t_1 \\
  t_2 \\
  \vdots \\
  t_m
\end{bmatrix} x_1 + \begin{bmatrix}
  1 \\
  1 \\
  \vdots \\
  1
\end{bmatrix} x_2 + \begin{bmatrix}
  n_1 \\
  n_2 \\
  \vdots \\
  n_m
\end{bmatrix}
\]

or

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_m
\end{bmatrix} = \begin{bmatrix}
  t_1 & 1 \\
  t_2 & 1 \\
  \vdots & \vdots \\
  t_m & 1
\end{bmatrix} \begin{bmatrix}
  x_1 \\
  x_2
\end{bmatrix} + \begin{bmatrix}
  n_1 \\
  n_2 \\
  \vdots \\
  n_m
\end{bmatrix}
\]

- Due to noise \( n \), it is evident that we cannot determine \( x \) such that \( Ax = y \)

- We can use LS approach to find \( \hat{x} \in \mathbb{R}^2 \) such that Euclidean distance between \( A\hat{x} \) and \( y \) is minimized.

This example has demonstrated the use of LS for solving linear regression problem.
Least-Squares

Recap:
- We want to find $x \in \mathbb{R}^n$ given a matrix $A \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$ related by
  \[ y = Ax \]

Formulation – Geometric Interpretation:
- Solution does not exist if $y$ does not belong to the column space of $A$.
  - To understand this statement, consider $m = 3$ and $n = 2$.

- Given $y$ and columns, $a_1$ and $a_2$, of $A$ are indicated. The column space $\mathcal{C}(A)$ is represented by a plane.

- Clearly, there does not exist $x \in \mathbb{R}^2$ such that $Ax = y$

- Can you find $\hat{x}$ such that
  \[ A\hat{x} \text{ is closest to } y \text{ in least-squares sense (Euclidean distance minimized)}? \]
  \[ A\hat{x} = a_1\hat{x}_1 + a_2\hat{x}_2 \text{ is closest to } y \text{ in least-squares sense?} \]
Least-Squares

Solution: Geometrically

We require:

\[ a_1 \hat{x}_1 + a_2 \hat{x}_2 \] is closest to \( y \) in least-squares sense.

- \( A\hat{x} = a_1 \hat{x}_1 + a_2 \hat{x}_2 \in C(A) \), that is, it represents a point on the plane.

- The solution \( \hat{x} \) for which \( a_1 \hat{x}_1 + a_2 \hat{x}_2 \) is closest to \( y \) is indicated in green.

- Residual error \( r = A\hat{x} - y \) is indicated in blue.

- \( \hat{x} \) is determined for which \( r \) is minimized.

- In other words, we require \( r \) and \( A\hat{x} \) to be orthogonal to every column of \( A \), that is,

\[
\begin{align*}
\quad a_1^T r &= 0 \\
\quad a_2^T r &= 0
\end{align*}
\]

\( \Rightarrow \ A^T r = 0 \quad \Rightarrow \ A^T (A\hat{x} - y) = 0 \quad \Rightarrow \ A^T A\hat{x} = A^T y \)

\[ \Rightarrow \ \hat{x} = (A^T A)^{-1} A^T y \]

Least-squares (LS) solution
Least-Squares

Solution: Verification

\( \hat{x} \) is indeed a LS solution, that is,

\[
\|Ax - y\| \geq \|A\hat{x} - y\|, \quad \forall x \quad \text{(residual error is minimum for } x = \hat{x})
\]

Residual error for any \( x \) \hspace{1cm} \text{Residual error for } \hat{x} \text{ (LS solution)}

Proof:

\[
\|Ax - y\|^2 = \|Ax + \hat{x} - A\hat{x} - y\|^2 \\
= \|Ax - A\hat{x}\|^2 + \|A\hat{x} - y\|^2 + 2(Ax - A\hat{x})^T(A\hat{x} - y) \\
= \|Ax - A\hat{x}\|^2 + \|A\hat{x} - y\|^2 + 2(x - \hat{x})^T A^T (A\hat{x} - y) \\
= \|Ax - A\hat{x}\|^2 + \|A\hat{x} - y\|^2 \\
\geq \|A\hat{x} - y\|^2
\]

Since \( A^T r = A^T (A\hat{x} - y) = 0 \) \hspace{1cm} \text{Since } \|Ax - A\hat{x}\|^2 \geq 0
Least-Squares

**Summary:**

- We want to find $x \in \mathbb{R}^n$ given a matrix $A \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$ related by
  
  $$y = Ax$$

- Least-squares (LS) solution

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

- Requires $A^T A$ to be invertible. In other words, we require columns of $A$ to be linearly independent.
Least-Squares

Regularization:

What if:

- $A^T A$ is not invertible or $A^T A$ is poorly conditioned.

- One solution could be to apply PCA to drop the columns of $A$.

- Other solution that is frequently used is Tikhonov regularization, that is, add a small value along the diagonal of $A^T A$ to make it invertible. With this regularization, LS solution can be modified as

$$\hat{x} = (A^T A + \lambda I)^{-1} A^T y$$

Regularized Least-squares (LS) solution.

- Here $\lambda$ is a scalar known as regularization parameter. Usually, we choose $\lambda = 0.01, 0.05$. 

LUMS
A Not-for-Profit University