Least Squares

We begin with one of the most fundamental optimization problems: **least squares**. In many scientific and engineering applications, we are given a set of measurements and wish to fit a model to this data. Mathematically, this can often be formulated as follows: we are given a data matrix $A \in \mathbb{R}^{m \times n}$ and a vector of outcomes $\vec{y} \in \mathbb{R}^m$. We seek to find a parameter vector $\vec{x} \in \mathbb{R}^n$ that best explains the data in the sense that it minimizes the discrepancy between our model's predictions, $A\vec{x}$, and the observed outcomes, $\vec{y}$.

Specifically, we aim to minimize the squared Euclidean distance, which is the sum of the squared differences between the components of the vectors. This leads to the following optimization problem:

$$\min_{\vec{x} \in \mathbb{R}^n} \|\vec{y} - A\vec{x}\|_2^2$$

where $\|\vec{v}\|_2 = \sqrt{\sum_{i=1}^n v_i^2}$ is the standard Euclidean norm. Squaring the norm is convenient as it removes the square root and yields a differentiable objective function without changing the location of the minimum.

Theorem (Least Squares Solution)

Let $A \in \mathbb{R}^{m \times n}$ have full column rank, and let $\vec{y} \in \mathbb{R}^m$. The solution to the least squares problem

$$\min_{\vec{x} \in \mathbb{R}^n} \|\vec{y} - A\vec{x}\|_2^2$$

is given by

$$\vec{x}^* = (A^\top A)^{-1} A^\top \vec{y}.$$

Proof. The core idea behind solving the least squares problem is geometric. The set of all possible model predictions, $\{A\vec{x} \mid \vec{x} \in \mathbb{R}^n\}$, forms a subspace of $\mathbb{R}^m$. This subspace is the column space, or **range**, of the matrix A, denoted $\mathcal{R}(A)$. The problem then becomes finding the vector $\vec{z} \in \mathcal{R}(A)$ that is closest to the vector $\vec{y}$.

In general, there is no guarantee that $\vec{y}$ itself lies in $\mathcal{R}(A)$. If it did, we could find an exact solution $\vec{x}$ to the system $A\vec{x} = \vec{y}$. Since this is not always possible, we seek the best approximation. As illustrated below, this best approximation $\vec{z}$ is the **orthogonal projection** of $\vec{y}$ onto the subspace $\mathcal{R}(A)$. Let us define the error vector as $\vec{e} = \vec{y} - \vec{z}$. The condition that $\vec{z}$ is the orthogonal projection of $\vec{y}$ means that $\vec{e}$ is orthogonal to the subspace $\mathcal{R}(A)$.

We must first prove that this orthogonally projected point $\vec{z}$ is indeed the closest point in $\mathcal{R}(A)$ to $\vec{y}$. Consider any other arbitrary point $\vec{u} \in \mathcal{R}(A)$. Our goal is to show that the distance from $\vec{y}$ to $\vec{u}$ is greater than the distance from $\vec{y}$ to $\vec{z}$.

Let's define the vector $\vec{w} = \vec{z} - \vec{u}$. Since both $\vec{z}$ and $\vec{u}$ belong to the subspace $\mathcal{R}(A)$, their difference $\vec{w}$ must also lie in $\mathcal{R}(A)$. We can express the vector from $\vec{u}$ to $\vec{y}$ by decomposing it using $\vec{z}$:

$$\vec{y} - \vec{u} = (\vec{y} - \vec{z}) + (\vec{z} - \vec{u}) = \vec{e} + \vec{w}$$

By construction, the error vector $\vec{e}$ is orthogonal to every vector in the subspace $\mathcal{R}(A)$. Since $\vec{w} \in \mathcal{R}(A)$, it follows that $\vec{e}$ and $\vec{w}$ are orthogonal vectors. This orthogonality allows us to apply the Pythagorean theorem to the squared norms:

$$\|\vec{y} - \vec{u}\|_2^2 = \|\vec{e} + \vec{w}\|_2^2 = \|\vec{e}\|_2^2 + \|\vec{w}\|_2^2$$

Substituting $\vec{e} = \vec{y} - \vec{z}$ and $\vec{w} = \vec{z} - \vec{u}$, we get:

$$\|\vec{y} - \vec{u}\|_2^2 = \|\vec{y} - \vec{z}\|_2^2 + \|\vec{z} - \vec{u}\|_2^2$$

Since we chose $\vec{u}$ to be distinct from $\vec{z}$, the vector $\vec{w} = \vec{z} - \vec{u}$ is non-zero, and thus its squared norm $\|\vec{z} - \vec{u}\|_2^2$ is strictly positive. Therefore,

$$\|\vec{y} - \vec{u}\|_2^2 > \|\vec{y} - \vec{z}\|_2^2$$

This confirms that $\vec{z} = A\vec{x}^*$ is the unique point in $\mathcal{R}(A)$ closest to $\vec{y}$.

Now, we derive the formula for $\vec{x}^*$. The defining property of our solution is that the residual vector, $\vec{y} - A\vec{x}^*$, is orthogonal to the subspace $\mathcal{R}(A)$. For this to be true, the residual vector must be orthogonal to every vector in a spanning set for $\mathcal{R}(A)$. The columns of A form such a spanning set. The condition that a vector is orthogonal to every column of A can be written compactly using the transpose of A:

$$A^\top(\vec{y} - A\vec{x}^*) = \vec{0}$$

Distributing $A^\top$ yields what are known as the **normal equations**:

$$A^\top A\vec{x}^* = A^\top\vec{y}$$

The theorem assumes that A has full column rank. This is a critical condition, as it guarantees that the Gram matrix $A^\top A$ is invertible. Multiplying by the inverse of $A^\top A$ on the left, we isolate $\vec{x}^*$ and arrive at the final solution:

$$\vec{x}^* = (A^\top A)^{-1}A^\top\vec{y}$$

The least squares solution is unique if and only if A has full column rank.

The least squares solution is unique if and only if A has full column rank. □