The Two Lenses of Linear Regression

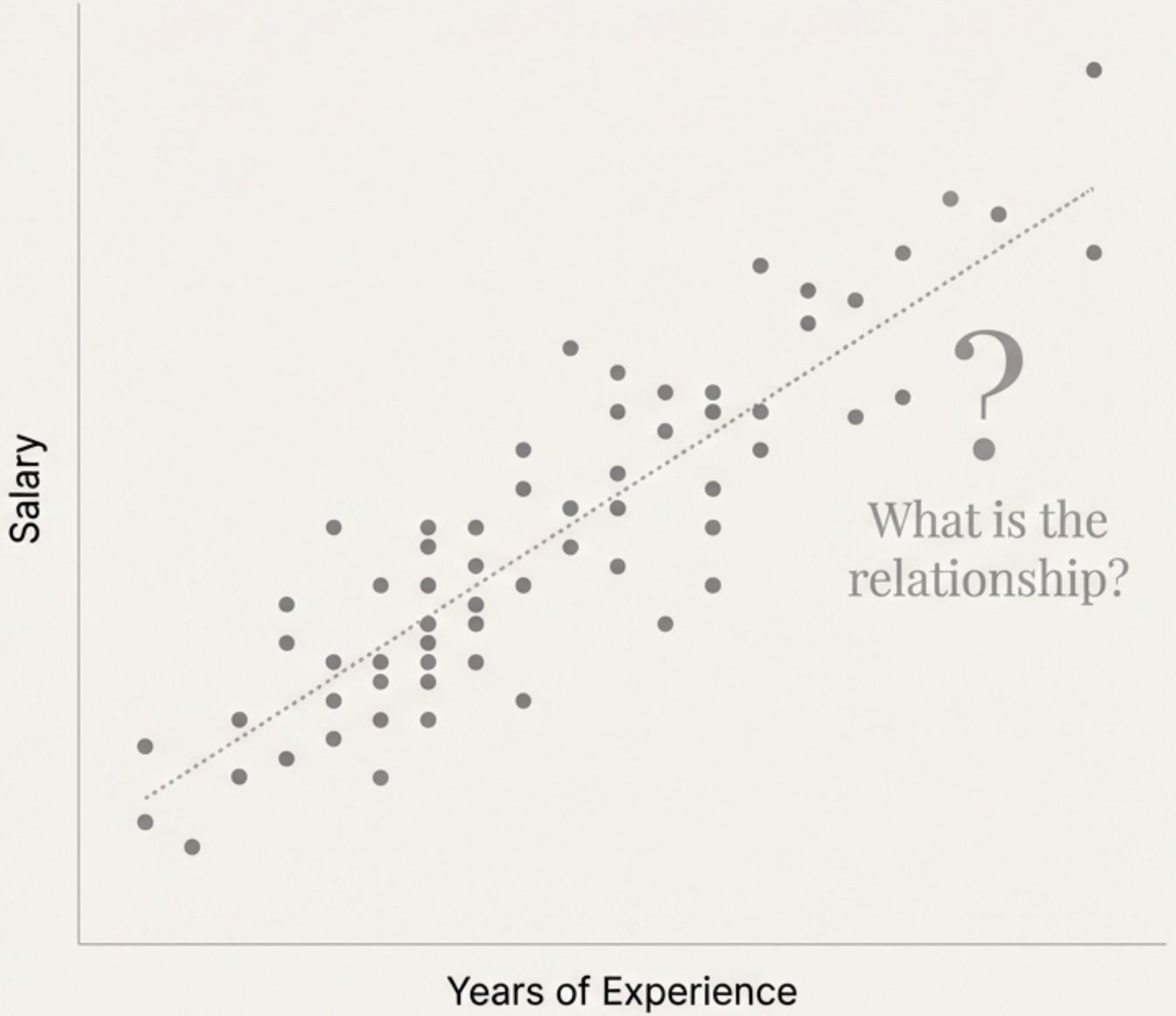
Unifying Geometry and Probability to Reveal a Deeper Truth

The Quest for Prediction

We begin with a fundamental challenge in machine learning: given a set of inputs, how do we predict a continuous output?

Our goal is to uncover the underlying relationship between these variables.

The simplest and most foundational tool for this task is a linear model. We are searching for the “best” straight line that captures the trend in the data.

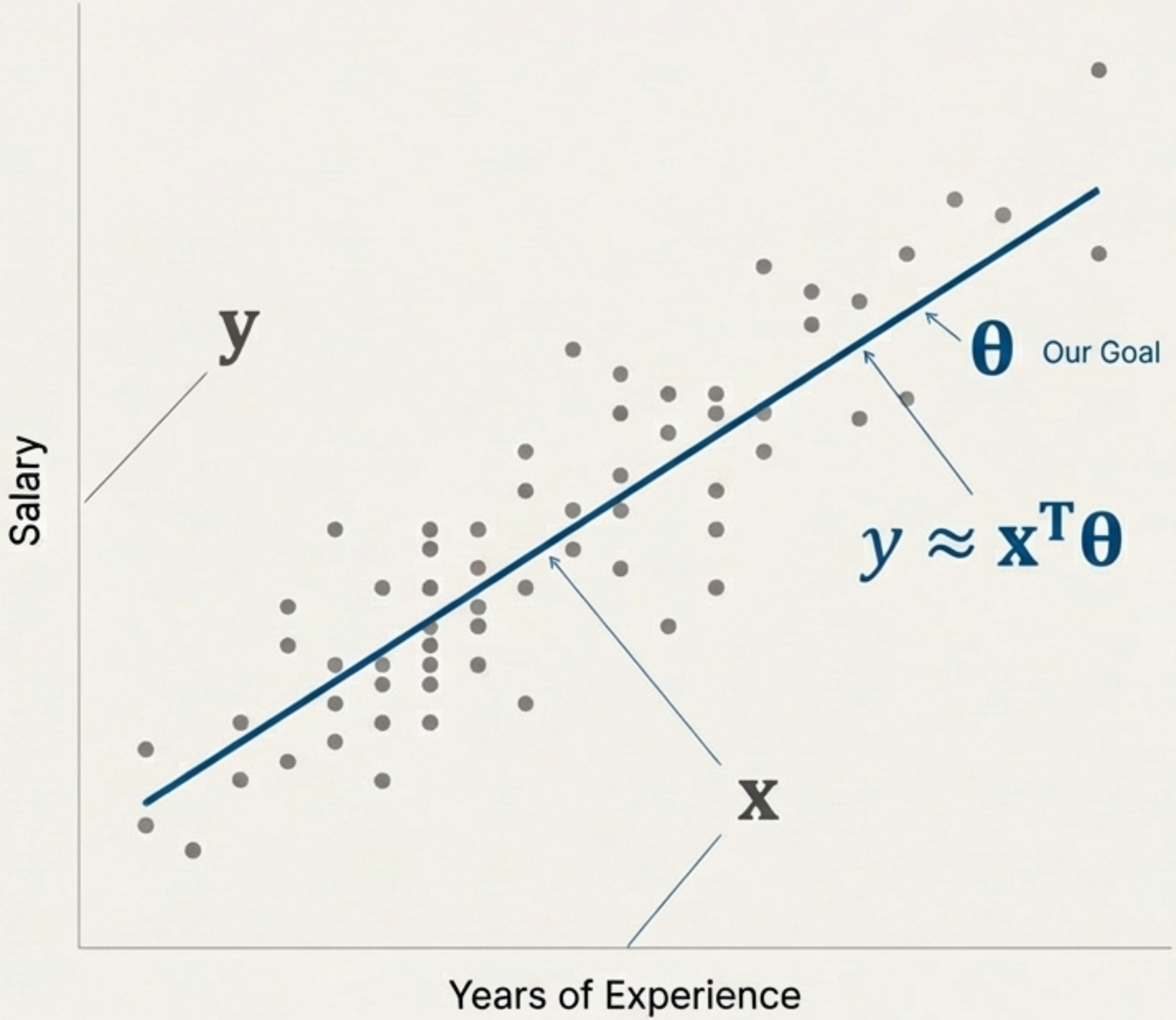


Translating the Visual into a Mathematical Model

To find our line, we must first define it mathematically. We represent our data as input-output pairs (x_n, y_n). Our model, or predictor, is a function f that maps inputs to outputs. For linear regression, this function takes the form:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{x}^{\mathbf{T}}\boldsymbol{\theta}$$

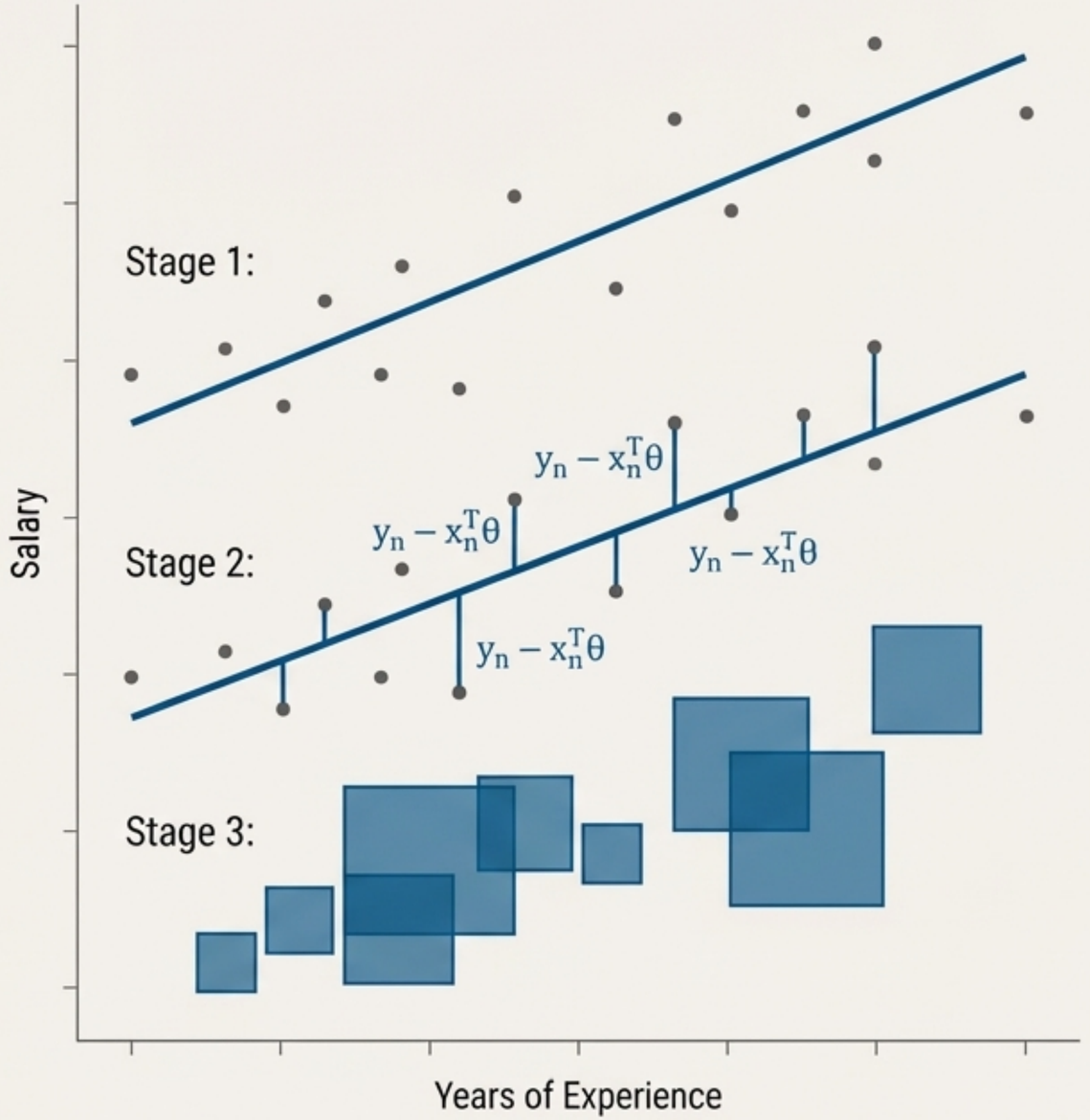
$\mathbf{x} \in \mathbb{R}^D$ is the input feature vector.
$\mathbf{y} \in \mathbb{R}$ is the observed output.
$\boldsymbol{\theta} \in \mathbb{R}^D$ is the parameter vector we need to find. These parameters define the specific line (its slope and intercept).



Lens 1: A World of Geometry and Error

Our first approach is to view this as an optimization problem. The “best” line is the one that is geometrically closest to all of our data points simultaneously.

We need a way to measure the total “error” or “misfit” between our model’s predictions ($x_n^T\theta$) and the actual data (y_n). The most common method is to measure the vertical distance for each point, square it, and sum these squares. This is the principle of **least squares**.



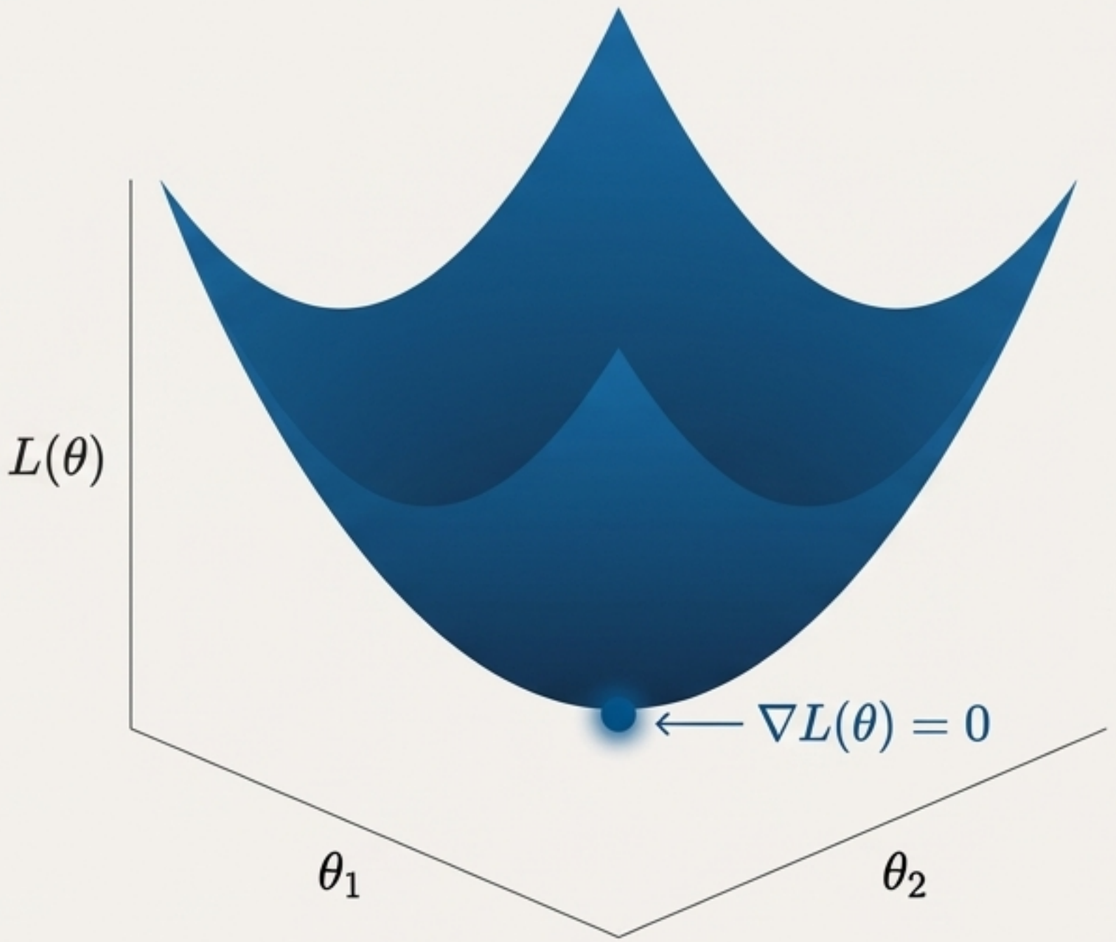
Finding the Bottom of the Error Valley

Our goal is to find the parameters θ that minimize the **sum of squared errors**, also known as the loss function $L(\theta)$:

$$L(\theta) = \sum_n (y_n - \mathbf{x}_n^T \theta)^2$$

To find the minimum of this function, we compute its gradient with respect to θ and set it to zero. This identifies the point where the slope of the loss function is flat—the bottom of the valley.

$$\nabla L(\theta) = 0$$



The Geometric Solution: The Normal Equations

Setting the gradient of the loss function $L(\theta) = \|\mathbf{y} - \mathbf{\Phi}\theta\|^2$ to zero yields the following analytical solution:

$$\nabla L(\theta) = -2\mathbf{\Phi}^T(\mathbf{y} - \mathbf{\Phi}\theta) = 0$$
$$\mathbf{\Phi}^T\mathbf{\Phi}\theta = \mathbf{\Phi}^T\mathbf{y}$$

This gives us the celebrated **Normal Equations**, which provide a closed-form expression for the optimal parameters θ:

$$\boxed{\theta = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y}}$$

Where $\mathbf{\Phi}$ is the **design matrix**, an N × D matrix where each row is an input vector $\mathbf{x}_n^T$. The term $(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$ is known as the Moore-Penrose pseudo-inverse.

From a purely geometric and optimization standpoint, we have found our answer. This feels like a complete solution.

Lens 2: A Universe of Probability and Belief

Let's re-examine our problem. What if the data we observe is not perfect? What if our model $y = \mathbf{x}^T\theta$ represents the true underlying relationship, but our measurements y_n are corrupted by random noise ϵ?

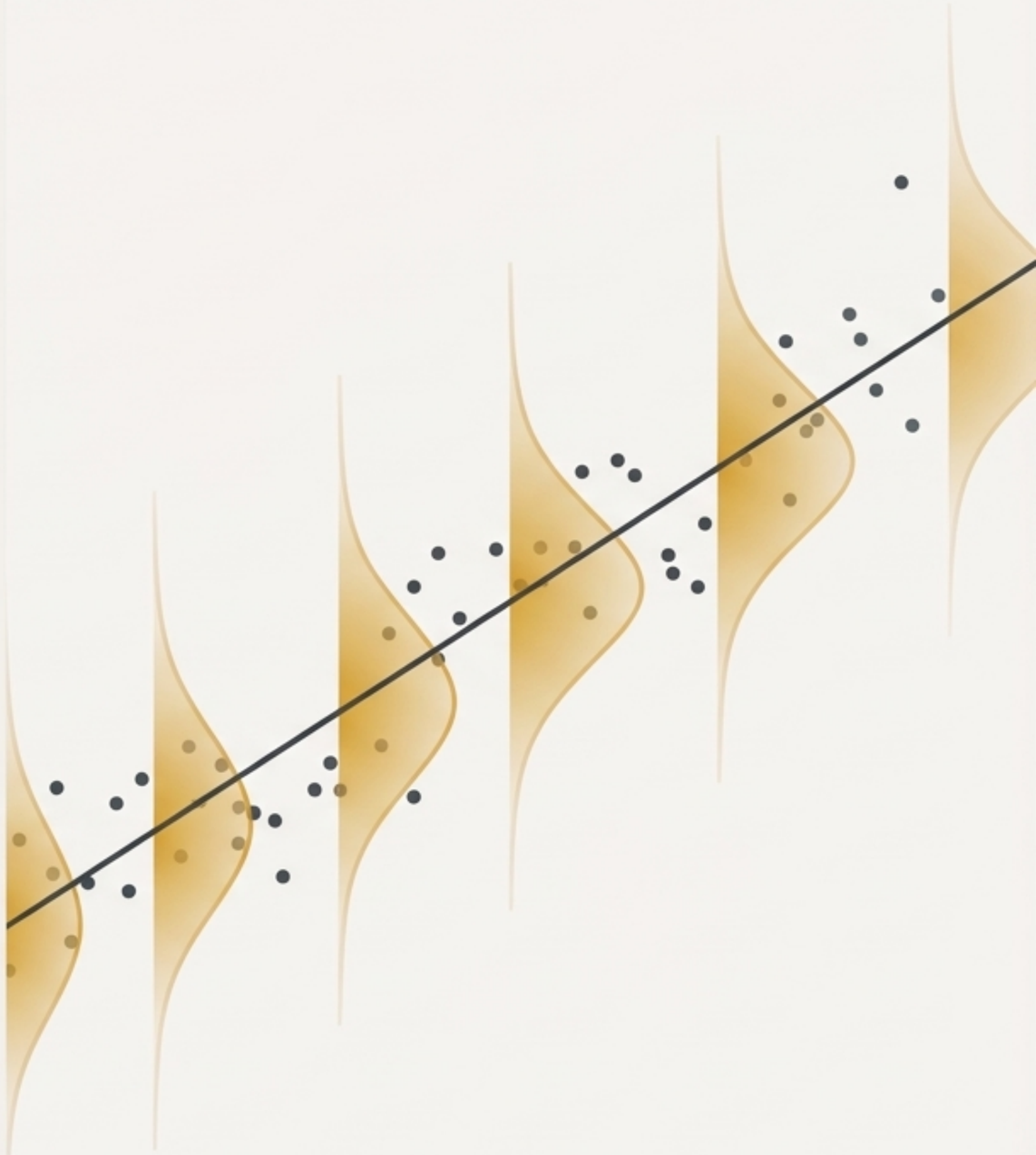
$$y = \mathbf{x}^T\theta + \epsilon$$

A common and powerful assumption is that this noise is drawn from a Gaussian (Normal) distribution with zero mean and variance σ^2.

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

This implies that for a given x, the output y is also a random variable:

$$p(y \mid \mathbf{x}, \theta, \sigma^2) = \mathcal{N}(y \mid \mathbf{x}^T\theta, \sigma^2)$$



The Probabilistic Goal: Maximum Likelihood

Given our model, we want to find the parameters θ that make our observed dataset $D = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$ the most probable outcome. This is the Principle of Maximum Likelihood Estimation (MLE).

Assuming our data points are independent and identically distributed (i.i.d.), the total likelihood of the dataset is the product of the individual probabilities:

$$p(Y \mid \mathbf{X}, \theta, \sigma^2) = \prod_n p(y_n \mid \mathbf{x}_n, \theta, \sigma^2) = \prod_n \mathcal{N}(y_n \mid \mathbf{x}_n^T \theta, \sigma^2)$$

Find θ_{ML} that maximizes this function:

$$\theta_{\mathrm{ML}} = \mathrm{argmax}_\theta \, p(Y \mid \mathbf{X}, \theta, \sigma^2)$$

The Strategic Power of the Log-Likelihood

The Problem with Products

Maximizing a long product of probabilities is numerically unstable (can lead to underflow) and mathematically cumbersome to differentiate.

The Logarithmic Solution

Since the logarithm is a strictly monotonically increasing function, maximizing $f(x)$ is equivalent to maximizing $\log(f(x))$. We can therefore maximize the **log-likelihood** instead.

$$\log\left[\prod_n \mathcal{N}(y_n \mid x_n^T\theta, \sigma^2)\right] \xrightarrow{\log} \sum_n \log \mathcal{N}(y_n \mid x_n^T\theta, \sigma^2)$$

$$\sum_n \left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_n - x_n^T\theta)^2\right]$$

Instead of maximizing the likelihood, we can minimize the **negative log-likelihood** (L).

The Grand Reveal: Two Paths Converge

As we seek the parameters $\boldsymbol{\theta}$ that minimize the **negative log-likelihood**, something remarkable happens.

The Geometric Objective

Minimize Sum of Squared Errors

$$L(\boldsymbol{\theta}) = \sum_n (y_n - \mathbf{x}_n^T \boldsymbol{\theta})^2$$

The Probabilistic Objective

Minimize Negative Log-Likelihood

$$L(\boldsymbol{\theta}) = \frac{N}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_n (y_n - \mathbf{x}_n^T \boldsymbol{\theta})^2$$

To find the optimal $\boldsymbol{\theta}$, we only need to minimize the term that depends on $\boldsymbol{\theta}$. The first term and the scaling factor do not change the location of the minimum.

The Synthesis: Why This Convergence Matters

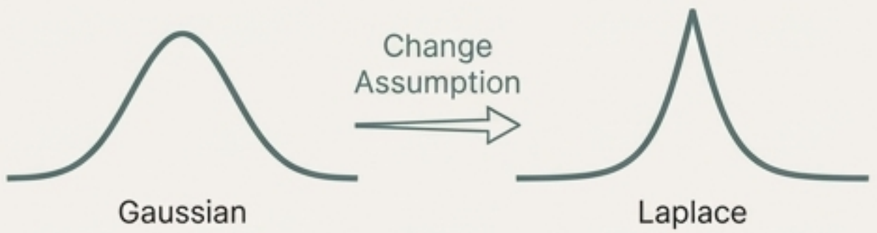
The Deeper Truth: The method of least squares is not just an arbitrary choice for measuring error. It is mathematically equivalent to assuming that the noise in our data is independent and identically distributed according to a Gaussian distribution. This provides a powerful probabilistic justification for a fundamentally geometric method.

Justification

We now have a deeper reason to trust least squares regression in many real-world scenarios where noise is often approximately normal.

A Path Forward

This equivalence provides a recipe for creating new models. If we believe our data's noise follows a different distribution (e.g., a Laplace distribution), we can simply replace the Gaussian likelihood with a Laplace likelihood. This leads to a different loss function (L1 or "least absolute deviations" regression), which is more robust to outliers.



Our simple straight line is built on a profound and elegant foundation, revealing the deep connections that unify the fields of optimization and statistics.