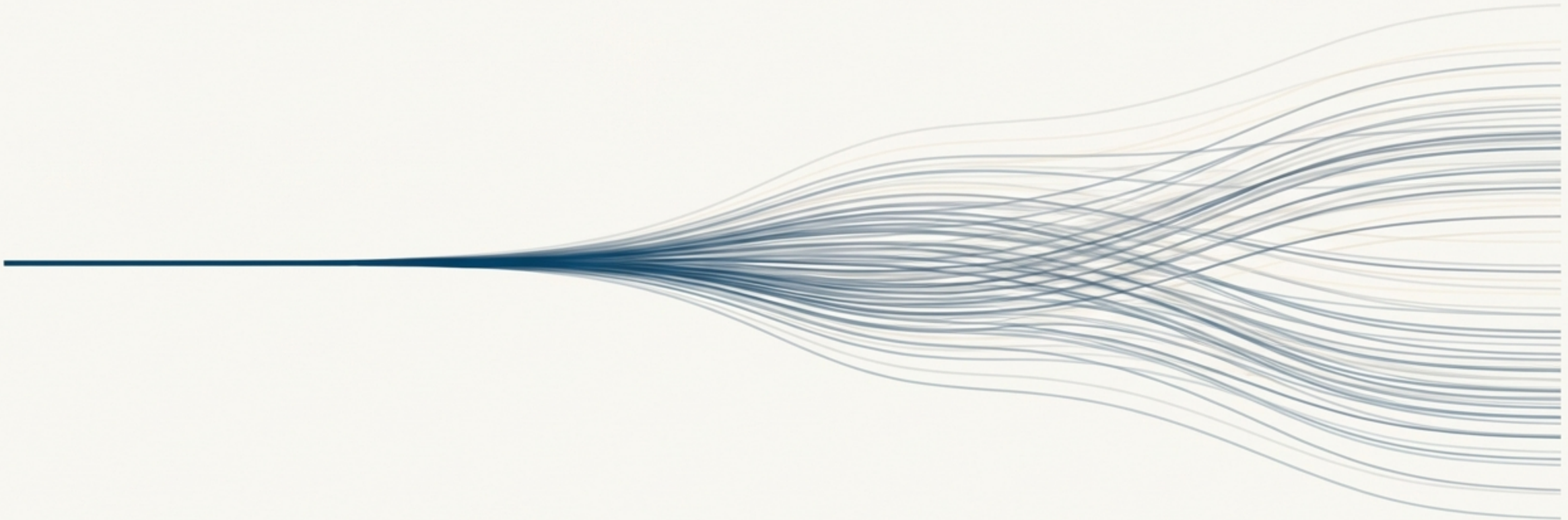
From a Single Truth to a Spectrum of Belief

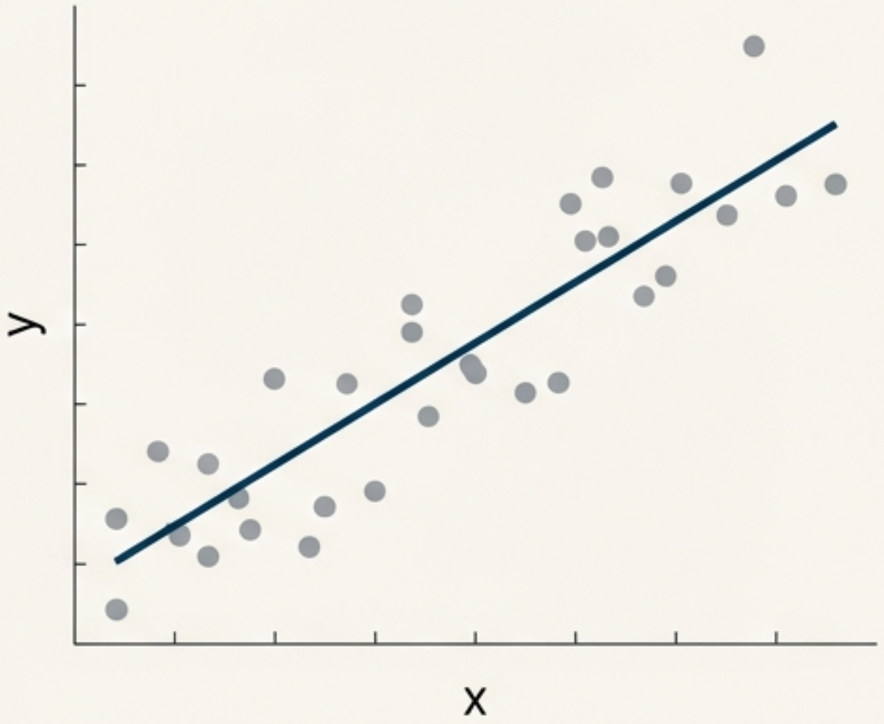
The Bayesian Approach to Linear Regression



Moving beyond the ‘best-fit’ line to embrace a distribution of every plausible line.

The World of Point Estimates: Linear Regression as We Know It

The goal of standard linear regression is to find the single set of parameters **θ** that maximizes the likelihood of observing the data, given a model.



Mathematical Objective

Likelihood Function (Gaussian Noise)

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

The Goal: Maximum Likelihood Estimation (MLE)

$$\theta_{\mathrm{ML}} = \mathrm{argmax}_\theta \, p(\mathbf{Y}|\mathbf{X},\theta)$$

Equivalence

Maximizing this likelihood is equivalent to minimizing the sum-of-squared errors:

$$\mathrm{argmin}_\theta \sum (y_n - \theta^{\mathrm{T}}\phi(\mathbf{x}_n))^2$$

The Output

A single vector of parameters, θ_{ML}—a point estimate.

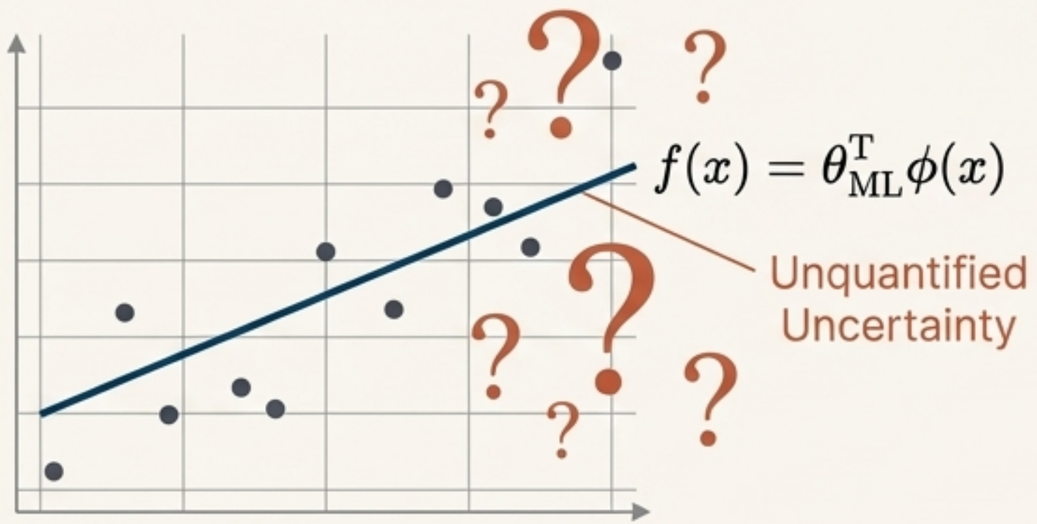
The Limits of a Single Answer

1. How certain are we?

The MLE approach provides a single set of parameters θ_{ML}, but no inherent measure of uncertainty.

If we have sparse data, shouldn't our confidence in the parameters be lower?

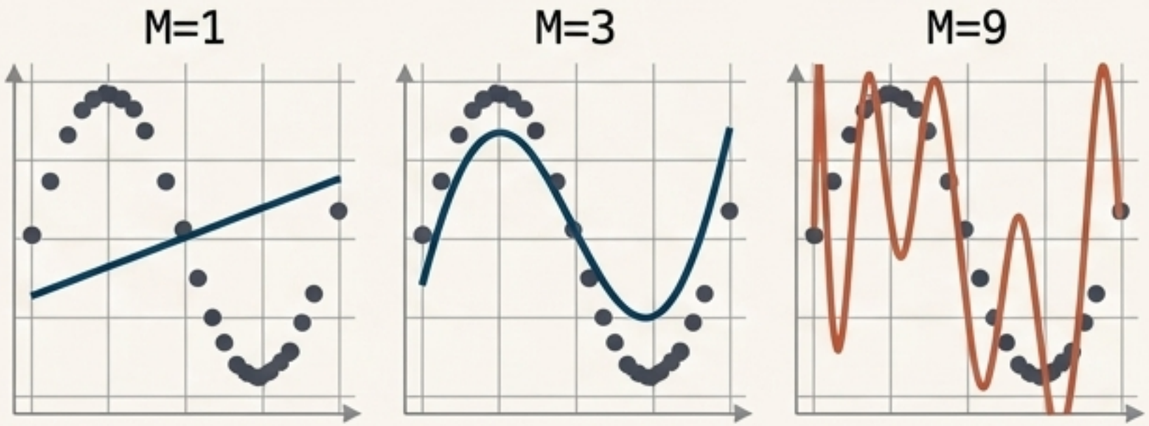
Standard regression doesn't quantify this.



2. What happens when the model is too flexible?

Complex models can perfectly fit the training data but generalize poorly. This is overfitting.

The common solution, regularization (e.g., L2/Ridge), is often introduced as a penalty term without a deep probabilistic justification.



A Fundamental Shift in Philosophy

The Frequentist View (MLE)

Parameters are Fixed Constants

The model parameters θ are unknown but fixed constants of nature. Our goal is to find the single best estimate, θ_{ML}, that explains the data we happened to observe.

Objective: Maximize Likelihood

$$p(\mathbf{Y}|\mathbf{X}, \theta)$$

The Bayesian View

Parameters are Random Variables

The parameters θ are random variables about which we have beliefs. Data is used to update these beliefs, transforming our prior understanding into a posterior one.

Objective: Compute Posterior

$$p(\theta \mid \mathbf{X}, Y)$$

The Engine of Belief Updating: Bayes' Theorem

Likelihood

What the *data says*. The probability of observing D given a specific set of parameters.

Posterior

Our *updated belief* about the parameters after seeing the data D.

$$p(\theta|D) = \frac{p(D|\theta)\ p(\theta)}{p(D)}$$

Prior

Our *initial belief* about the parameters before seeing any data.

Evidence (or Marginal Likelihood)

The probability of the data, averaged over all possible parameters. Acts as a normalization constant.

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

[Updated Belief] is proportional to **[What the Data Says]** times **[Our Initial Belief]**

Building a Bayesian Linear Model

Step 1: Define a Prior on the Parameters.

We start by encoding our initial beliefs. A common choice is a zero-mean **Gaussian prior**, which assumes that smaller parameter values are more likely, encouraging simplicity.

$$p(\boldsymbol{\theta}) = N(\boldsymbol{\theta} \mid \mathbf{0}, \alpha^{-1}\mathbf{I})$$

Step 2: Use the Same Likelihood Function.

The likelihood function is identical to the one used in MLE. It models the data-generating process, assuming Gaussian noise.

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = \prod_n N(\mathbf{y}_n|\boldsymbol{\theta}^T\boldsymbol{\varphi}(\mathbf{x}_n), \sigma^2)$$

Step 3: Compute the Posterior.

Using Bayes' theorem, we combine the prior and the likelihood. For a Gaussian prior and likelihood, the posterior is also conveniently a Gaussian, a property known as conjugacy.

$$p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) = N(\boldsymbol{\theta} \mid \mathbf{m}_N, \mathbf{S}_N)$$

$$\mathbf{S}_N = (\mathbf{S}_0^{-1} + \sigma^{-2}\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$$
$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \sigma^{-2}\boldsymbol{\Phi}^T\mathbf{y})$$

A Principled Origin for Regularization

What happens if we don't compute the full posterior, but just find the *most probable* parameter values? This is Maximum A Posteriori (**MAP**) estimation.

Objective:

$$\theta_{MAP} = \text{argmax}_\theta \, p(\theta|Y,X) = \text{argmax}_\theta \, [\log p(Y|X,\theta) + \log p(\theta)]$$

↓

Substitute Components:

$$\log p(Y|X,\theta) \text{ (from Gaussian Likelihood)} \rightarrow -\frac{1}{2\sigma^2}\sum(y_n - \theta^T\varphi(x_n))^2 + \text{const}$$

$$\log p(\theta) \text{ (from Gaussian Prior)} \rightarrow -\frac{\alpha}{2}\theta^T\theta + \text{const}$$

↓

The Result (equivalent minimization problem):

$$\text{argmin}_\theta \left[\left(\frac{1}{2\sigma^2}\sum(y_n - \theta^T\varphi(x_n))^2 + \frac{\alpha}{2}\theta^T\theta\right]\right.$$

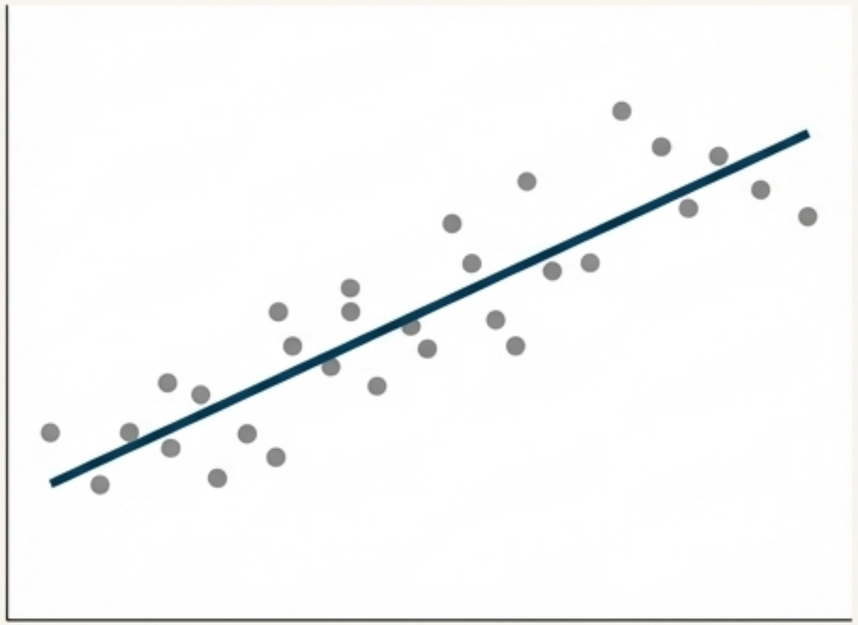
This is mathematically equivalent to minimizing the sum-of-squared errors with an L2 regularization term (Ridge Regression)!

$$\min \|y - \Phi\theta\|^2 + \lambda\|\theta\|^2$$

The Bayesian prior provides a probabilistic justification for L2 regularization. The prior's precision (α) is directly related to the regularization strength (λ).

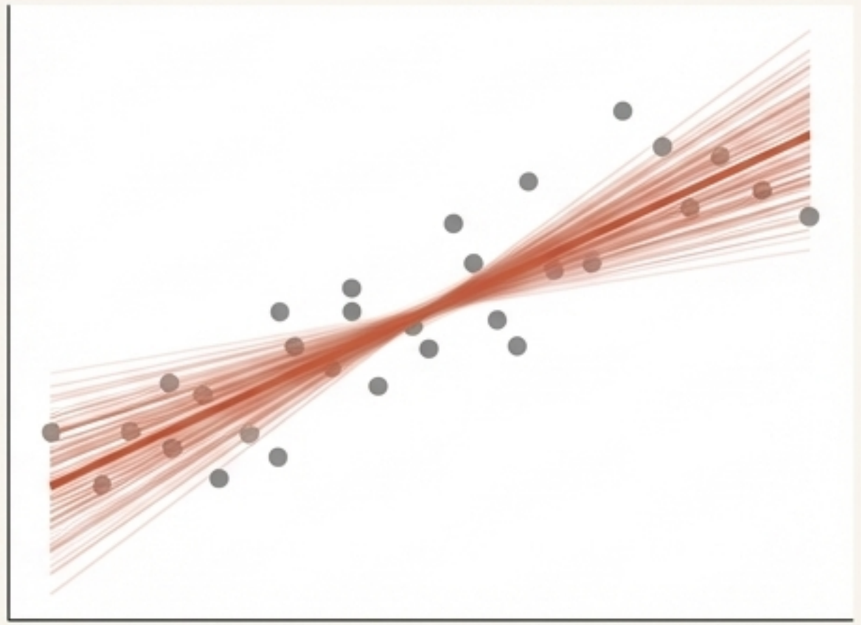
The Output: From a Single Line to a Distribution Over Functions

A Point Estimate



MLE provides one function, $f(x) = \theta_{\mathrm{ML}}^{T}\varphi(x)$. It represents the single “best” model.

A Posterior Distribution



Bayesian inference provides a distribution over all plausible functions. Each line represents a different parameter set θ_i sampled from the posterior, weighted by its probability.

The Prediction: From a Single Value to a Spectrum of Possibilities

Core Idea: To make a prediction for a new input $\mathbf{x}_*$, we don't use a single $\boldsymbol{\theta}$. We average the predictions of *all possible* parameter sets, weighted by their posterior probability.

The Math: This is done by integrating over the posterior parameter distribution:

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{D}) = \int \underbrace{p(\mathbf{y}_*|\mathbf{x}_*, \boldsymbol{\theta})}_{\text{Likelihood}} \underbrace{p(\boldsymbol{\theta}|\mathbf{D})}_{\text{Parameter Posterior}} d\boldsymbol{\theta}$$

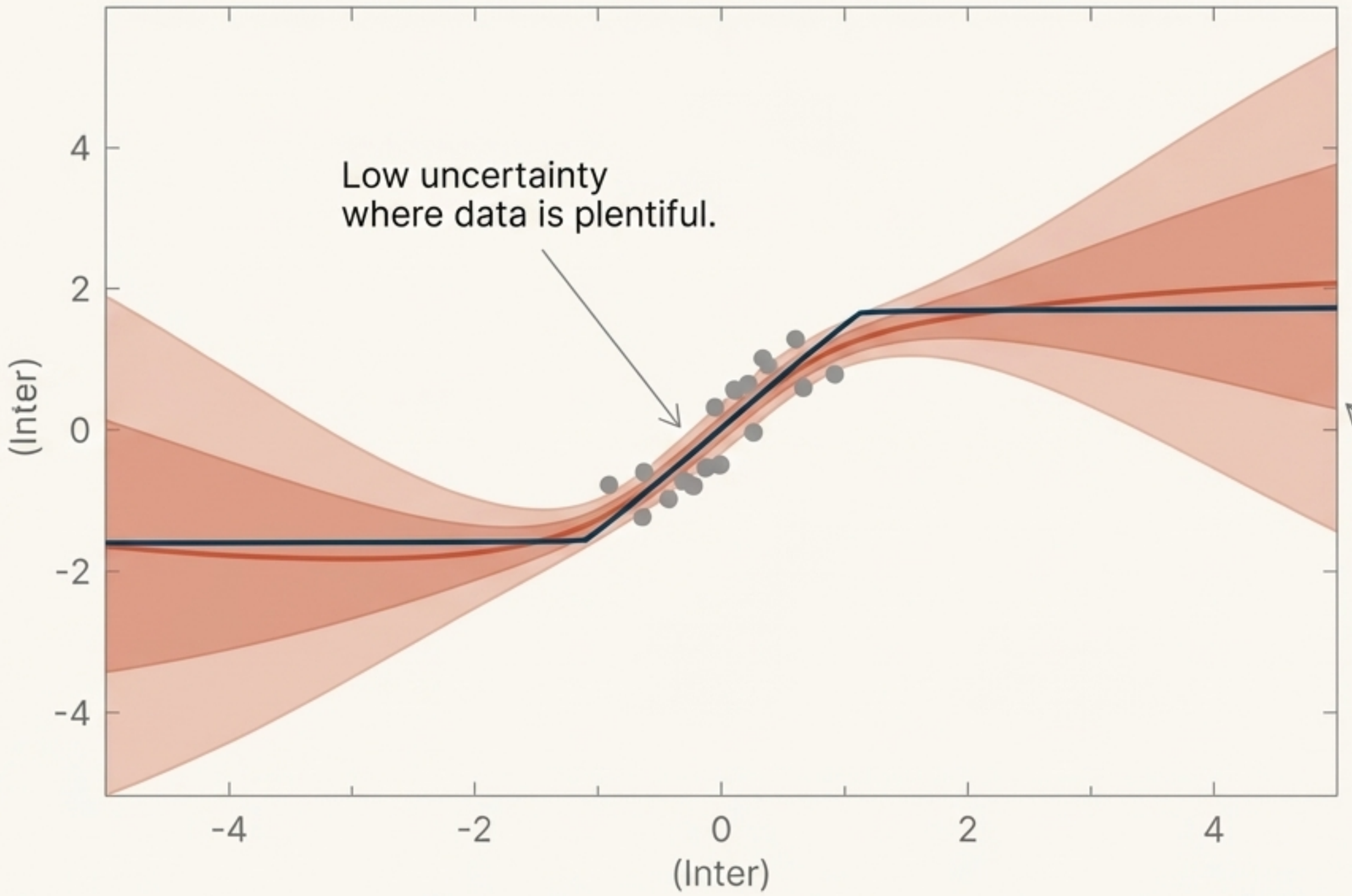
The Result: The result is the **Predictive Distribution**. For Bayesian Linear Regression, this is also a Gaussian:

$$p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{D}) = \mathcal{N}(\mathbf{y}_* \mid \mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}_*), \sigma_N^2(\mathbf{x}_*))$$

$$\sigma_N^2(\mathbf{x}_*) = \sigma^2 + \boldsymbol{\phi}(\mathbf{x}_*)^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_*)$$

The predictive variance $\sigma_N^2(\mathbf{x}_*)$ has two components: the intrinsic noise in the data (σ^2) and the uncertainty in our parameter estimates ($\boldsymbol{\phi}(\mathbf{x}_*)^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_*)$).

Visualizing What the Model Knows (and Doesn't Know)

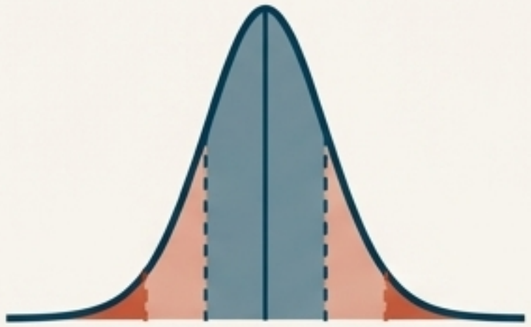


High uncertainty where data is sparse. The model correctly expresses its lack of knowledge.

The Two Paradigms: A Summary

Feature	Frequentist (Maximum Likelihood)	Bayesian
Philosophy	Parameters are fixed, unknown constants. Data is random.	Parameters are random variables representing beliefs.
Objective	Find the single parameter set θ_{ML} that maximizes the likelihood $p(D\|\theta)$.	Compute the full posterior distribution $p(\theta\|D)$ by combining a prior $p(\theta)$ and the likelihood $p(D\|\theta)$.
Output	A point estimate for parameters: θ_{ML}.	A full probability distribution for parameters: $p(\theta\|D)$.
Prediction	A single value prediction: $\hat{y} = \theta_{ML}^T \phi(x)$.	A full predictive distribution: $p(\hat{y}\|x,D)$, which quantifies uncertainty.

What We Gain with the Bayesian Perspective



Principled Uncertainty Quantification

We get a full distribution for parameters and predictions, allowing us to quantify and act on model uncertainty. We know what our model doesn't know.

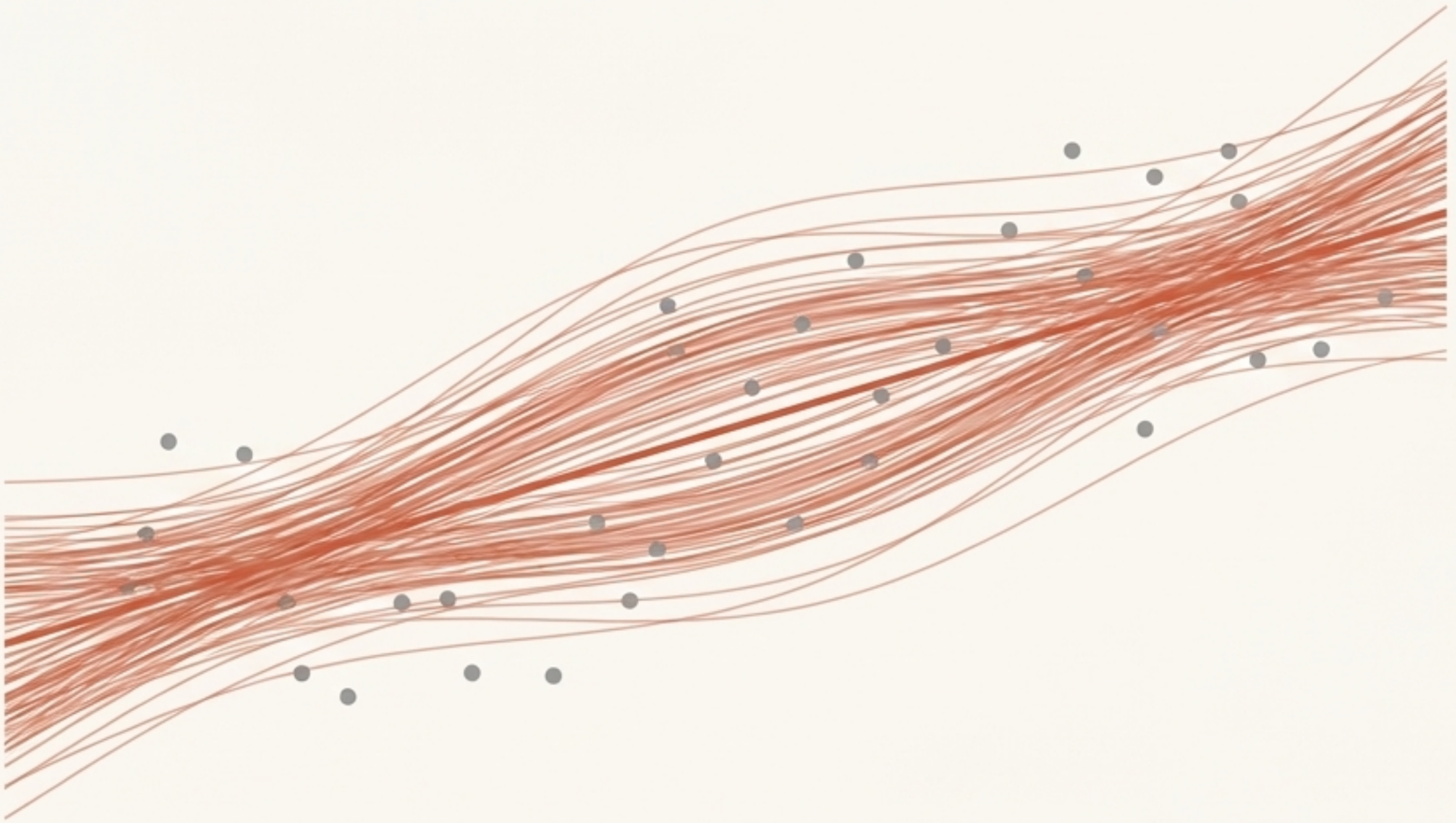
A Probabilistic View of Regularization

Regularization is no longer an ad-hoc penalty. It is the logical consequence of incorporating prior knowledge into the model, with a clear probabilistic interpretation.

A Framework for Model Comparison

The 'evidence' or 'marginal likelihood' term p(D) provides a principled way to compare different models and guard against overfitting, embodying a form of Occam's razor.

Beyond a Single Truth



Bayesian inference transforms linear regression from a method for finding the ‘one correct’ line into a system for reasoning about all plausible lines. It replaces the certainty of a single answer with the wisdom of a quantified belief.