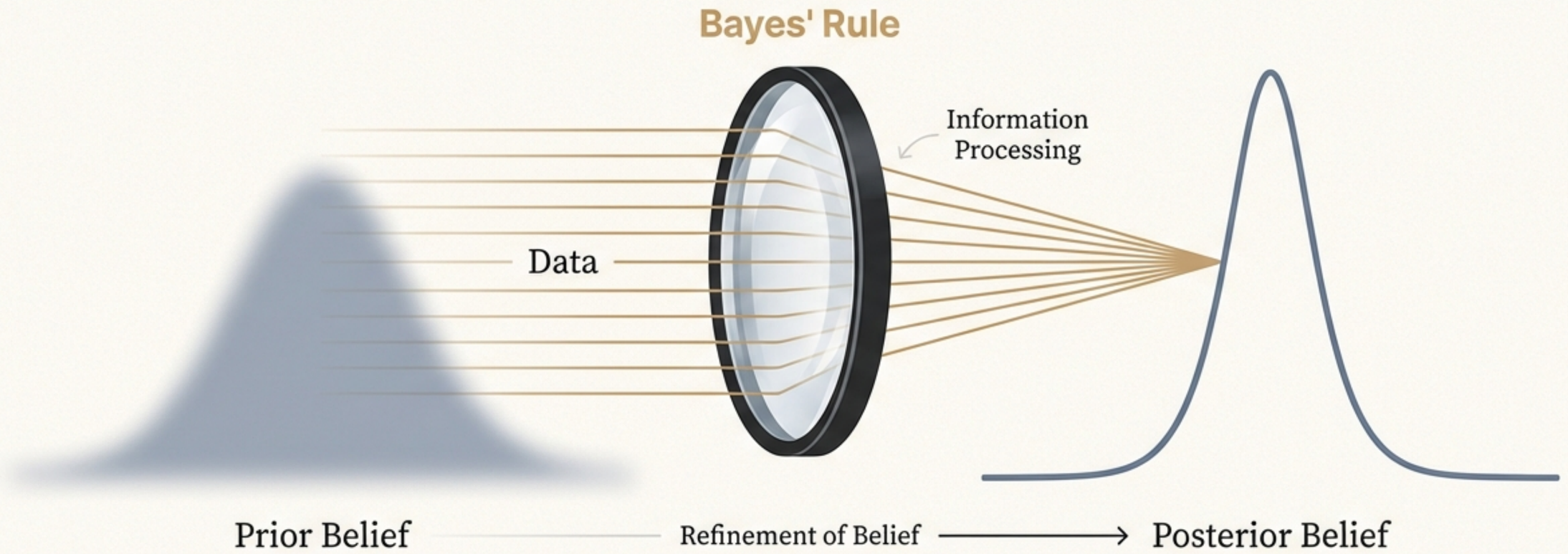
The Engine of Inference

Understanding Bayes' Rule in Machine Learning



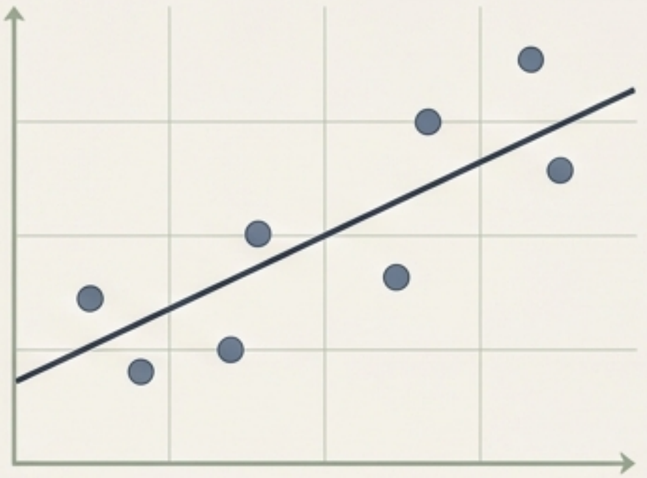
Based on “Mathematics for Machine Learning” by Deisenroth, Faisal, and Ong.

Machine learning is more than finding a single 'best' answer.

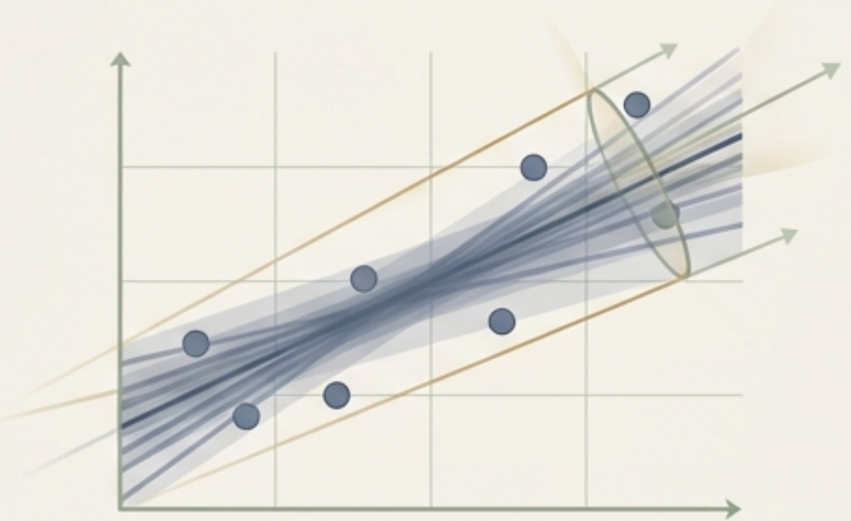
The goal of modern ML is often not to find a single set of parameters, but to reason about uncertainty. As outlined in “Mathematics for Machine Learning” (Chapter 8.4), the challenge is “Probabilistic Modeling and Inference”:

- We start with a model representing our initial beliefs about the world.
- We collect data, which serves as evidence.
- The core task is to update our beliefs in a principled way after observing this evidence.

How do we formalize the process of learning from data to refine our beliefs?



Point Estimate: A single answer



Probabilistic Inference: A distribution of possibilities

A Concrete Challenge: Parameter Estimation in Bayesian Linear Regression

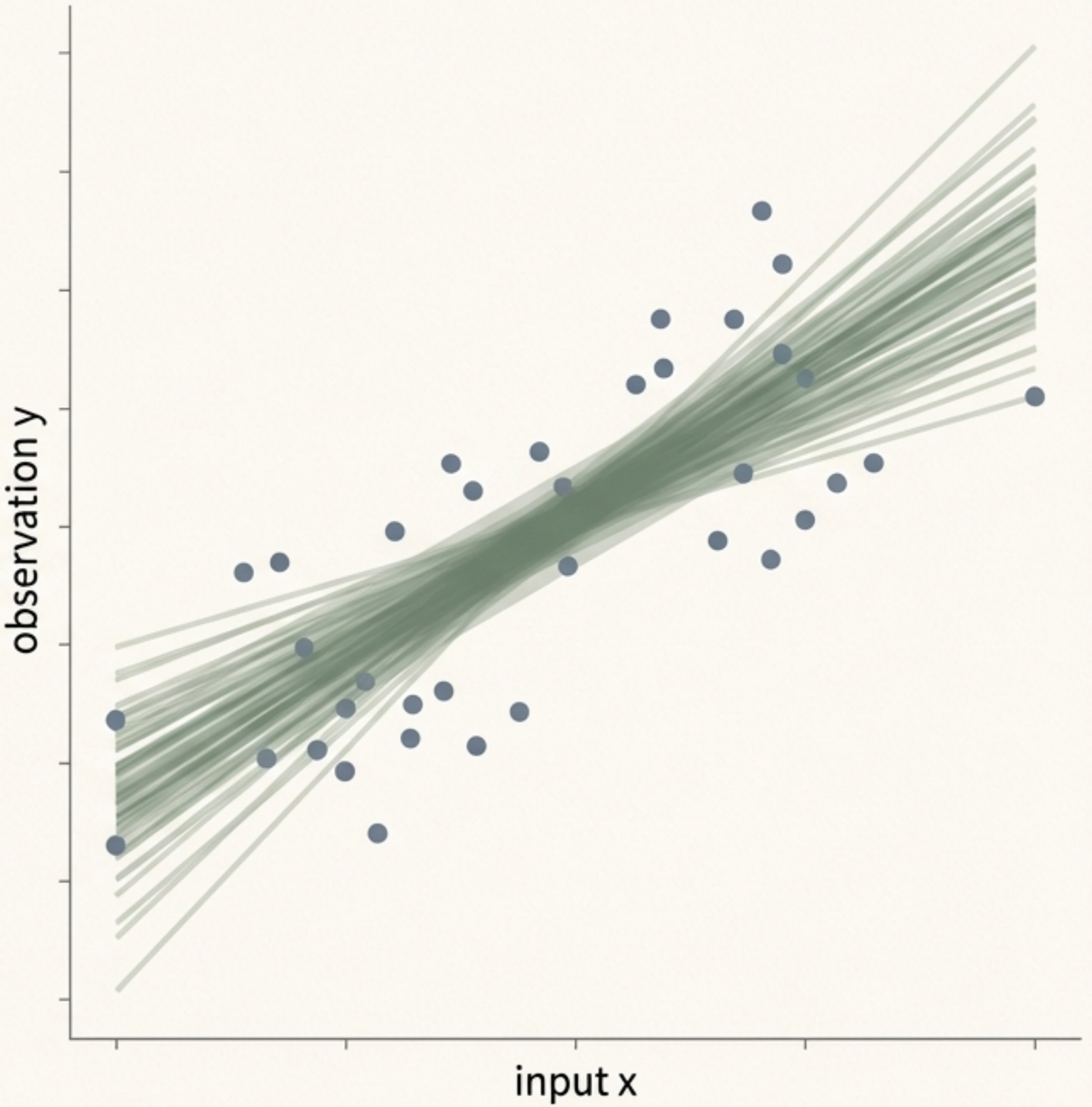
In Chapter 9.3 of MML, we move beyond finding a single best-fit line. Instead, we want to find a *distribution* over possible lines that explains the data.

Problem Statement

Given: A dataset of inputs **x** and observations **y**.

Prior Belief: We have some initial assumption about the distribution of the model parameters θ. This is our $P(\theta)$.

Goal: After observing the data, what is our *updated* belief about the parameters? We want to find the posterior distribution, $P(\theta \mid y, x)$.



The Language of Uncertainty: Sum and Product Rules

To formally update our beliefs, we need two fundamental rules of probability from MML Chapter 6.3:

Sum Rule

The probability of a variable x is found by marginalizing (summing over) all possibilities of another variable y:

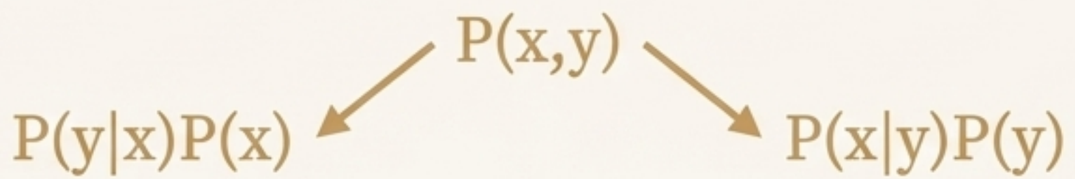
$$P(x) = \sum_{y} P(x, y)$$

Product Rule

The joint probability of x and y can be expressed in two ways using conditional probability:

$$P(x, y) = P(y|x)P(x) = P(x|y)P(y)$$

Key Insight: The symmetry of the Product Rule is the key. It shows we can relate P(y|x) to P(x|y). This simple algebraic relationship is the foundation of inference.



From First Principles: Deriving the Engine of Inference

Step 1: The starting point from the previous slide is shown clearly:

We start with the two forms of the Product Rule:

$$P(\mathbf{y}|\mathbf{x})P(\mathbf{x}) = P(\mathbf{x}|\mathbf{y})P(\mathbf{y})$$

Step 2: The algebraic manipulation is shown.

By dividing both sides by P(x), we can isolate P(y|x):

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$$\boxed{P(y|x) = \frac{P(x|y)P(y)}{P(x)}}$$

Conclusion*: This is Bayes' Rule. It provides the exact mechanism for 'inverting' a conditional probability—allowing us to calculate `P(hypothesis | data)` from `P(data | hypothesis)`. This is precisely the tool we need to update our beliefs.

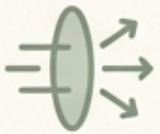
The Anatomy of Belief Updating

Posterior = (Likelihood × Prior) / Evidence

$$P(\theta \mid D) = \frac{P(D \mid \theta) \times P(\theta)}{P(D)}$$

Prior `P(θ)`

Our initial belief about the hypothesis θ before seeing any data D.



Likelihood `P(D | θ)`

The probability of observing the data D given our hypothesis θ. It measures how well our hypothesis explains the evidence.

Posterior `P(θ | D)`

Our updated belief about θ after observing the data. This is what we compute.

Evidence `P(D)`

The marginal probability of the data, $\int P(D \mid \theta)P(\theta)\, d\theta$. It acts as a normalizing constant to ensure the posterior is a valid probability distribution.

A Concrete Example: Medical Diagnosis

Problem Setup

Problem

A patient tests positive for a rare disease. What is the **actual probability** they have the disease?

Inputs

- **Prior** P(Disease): The prevalence of the disease in the population is 0.1% (or 0.001).
- **Likelihood** P(Pos | Disease): The test is 99% accurate (true positive rate).
- **Likelihood** P(Pos | No Disease) is 1% (false positive rate).

Calculation

Applying Bayes' Rule

We want to find the Posterior P(Disease | Pos).

$$P(\text{Disease} \mid \text{Pos}) = \frac{P(\text{Pos} \mid \text{Disease})\,P(\text{Disease})}{P(\text{Pos})}$$

Calculation

- Numerator: $0.99 \times 0.001 = 0.00099$
- Evidence $P(\text{Pos}) = P(\text{Pos}|\text{Disease})P(\text{Disease}) + P(\text{Pos}|\text{No Disease})P(\text{No Disease})$ $= (0.99 \times 0.001) + (0.01 \times 0.999) = 0.00099 + 0.00999 = 0.01098$
- Posterior: $\frac{0.00099}{0.01098} \approx 0.09$

Result: Even with a positive test, there is only a 9% chance the patient has the disease. The low prior probability dominates the outcome.

Closing the Loop: Bayesian Linear Regression Revisited

Let's map the terms of Bayes' Rule directly onto the parameter estimation problem from MML 9.3. We want to find our updated belief about parameters θ after seeing data $D = (\mathbf{y}, \mathbf{x})$.

The Bayesian Formulation:

$$P(\theta \mid \mathbf{y}, \mathbf{x}) = \frac{P(\mathbf{y} \mid \mathbf{x}, \theta) P(\theta)}{P(\mathbf{y} \mid \mathbf{x})}$$

Direct Mapping

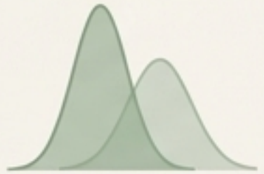
Posterior $P(\theta \mid \mathbf{y}, \mathbf{x})$ → Our goal. The distribution over parameters *after* seeing the data. It represents our refined knowledge.

Likelihood $P(\mathbf{y} \mid \mathbf{x}, \theta)$ → How probable are our observed $\mathbf{y}$ values, given the inputs $\mathbf{x}$ and a specific set of parameters θ? This is typically defined by our model's assumptions (e.g., Gaussian noise).

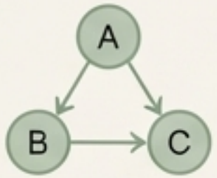
Prior $P(\theta)$ → Our initial belief about the parameters θ *before* we see any data. This allows us to incorporate domain knowledge.

The Engine for a Universe of Probabilistic Models

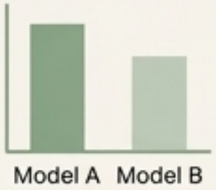
The belief-updating mechanism of Bayes' Rule is not limited to regression. It is the fundamental principle behind many of the central machine learning problems discussed in MML:



- **Density Estimation (Ch 11)**
  In models like Gaussian Mixture Models, Bayesian methods can be used to infer the parameters of the mixture components.



- **Directed Graphical Models (Ch 8.5)**
  Bayes' Rule defines how information (evidence) propagates through the nodes of a probabilistic graph, allowing for complex inference.



- **Model Selection (Ch 8.6)**
  The ‘evidence' term in Bayes' Rule, P(D), can be used to compare different models and select the one that best explains the data.

Conclusion

Bayes' Rule provides a unified mathematical framework for incorporating evidence and managing uncertainty across a vast landscape of machine learning models.

More Than a Formula: A Framwork for Rational Inference

1. Bayes' Rule **is the formal language for updating beliefs.** It provides a principled way to move from a prior state of knowledge to a posterior state informed by evidence.

2. **It connects what we knew with what we saw.** The Prior ($P(\theta)$) and the Likelihood ($P(D|\theta)$) are explicitly combined to produce the Posterior ($P(\theta|D)$).

3. **It is the mathematical core of learning under uncertainty.** In machine learning, 'learning' is often synonymous with using data to reduce uncertainty about our models of the world. Bayes' Rule is the engine that drives this process.