Bayesian Modeling: Foundations and Inference

ELG 5218 - Uncertainty Evaluation in Engineering Measurements and Machine Learning

Miodrag Bolic

University of Ottawa

January 14, 2026

Learning goals

By the end of this lecture, you should be able to:

- Define **uncertainty quantification (UQ)**.
- Specify a Bayesian model using **prior**, **likelihood**, and **posterior**.
- Explain and compute **MLE** and **MAP**.
- Derive and interpret a simple conjugate update (Beta–Binomial).
- Compute and interpret **credible intervals** (central and HPD/HDI).
- Use the posterior to form the **posterior predictive distribution**.

Roadmap

- Introduction to UQ (what / why / how to represent uncertainty)
- Confidence intervals vs credible intervals (CI vs CrI)
- Bayesian inference framework (prior, likelihood, posterior, evidence)
- Conjugate priors and Beta–Binomial example
- Point estimation: MLE and MAP (and connection to regularization)
- Posterior summaries (mean/median/mode, intervals) and prediction (PPD)
- Evidence intuition and what comes next (approximate inference)

What is Uncertainty Quantification?

Uncertainty Quantification (UQ) develops rigorous methods to characterize the impact of "limited knowledge" on quantities of interest.

Two fundamental sources of uncertainty

1. **Aleatoric Uncertainty**: inherent randomness in physical processes (irreducible)
2. **Epistemic Uncertainty**: lack of knowledge that can be reduced with more data/modeling

Key questions in UQ

- What is the expected value of our quantity of interest?
- How much does it vary (variance / standard deviation)?
- What range of values is plausible (intervals)?
- How do we update beliefs when new data arrive?

Example: Heart rate measurement reported as (60 ± 4) beats/minute.

Representing uncertainty

Common ways to express uncertainty:

1. **Standard error**: variability of an estimate
2. **Confidence intervals**: frequentist
3. **Credible intervals**: Bayesian
4. **Probability distributions**: full description of uncertainty
5. **Quantiles / moments**: extracted summaries

A point estimate without uncertainty quantification is incomplete.

A visual language for uncertainty

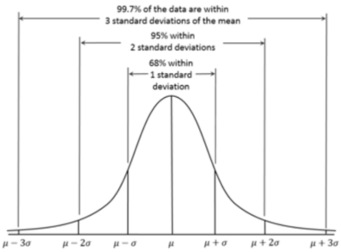


Figure: Normal distribution with interval bands

- Distributions communicate **shape** (skew, multimodality), not only spread.
- Intervals are summaries; distributions are the full story.

Confidence intervals: frequentist perspective

Definition: A 95% confidence interval (CI) is constructed so that:

95% of similarly constructed CIs contain the true parameter value.

Key properties

- The parameter θ is **fixed but unknown**
- The interval is **random** (depends on the sample)
- Interpretation is about **long-run frequency** across repeated sampling

What it does not mean:

$$\mathbb{P}\big(\theta \in [a, b]\big) = 0.95 \quad \textbf{WRONG}$$

Credible intervals: Bayesian perspective

Definition: A 95% credible interval (CrI) satisfies:

$$\mathbb{P}\big(\theta \in [a, b] \mid \text{data}\big) = 0.95$$

Key properties

- The parameter θ is treated as a **random variable**
- The interval is **fixed** (given the posterior)
- Interpretation is **posterior probability** (direct probability statement)

Two common credible intervals

1. **Central (equal-tailed)**: $\alpha/2$ mass in each tail
2. **HPD / HDI**: region(s) with highest posterior density

Comparison: CI vs CrI

Aspect	**Confidence interval**	**Credible interval**
Parameter	Fixed	Random variable
Interval	Random	Fixed (given posterior)
Uses prior?	No	Yes
Meaning	Long-run coverage	Posterior probability
Computation	Often analytic	Often via sampling / numeric

Modeling Data Probabilistically: A Simplistic View

- Assume a dataset $X = \{x_1, \ldots, x_N\}$ is generated from a probabilistic model with unknown parameters θ.
- For i.i.d. observations: $\quad x_1, \ldots, x_N \sim p(x \mid \theta)$.
- Plate notation: shaded nodes = observed; unshaded nodes = unknown / unobserved.
- Goal: estimate the unknowns (here, θ) given the observed data X.
- Use the learned model for prediction:

$$p(x^* \mid \theta) \quad \text{or} \quad p(x^* \mid X).$$

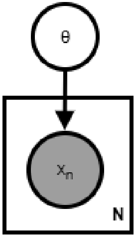
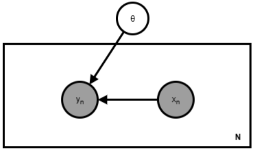


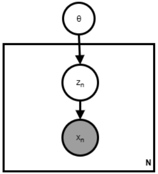
Figure: Simplified plate model for i.i.d. data

Modeling Data Probabilistically

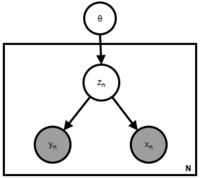
- This basic setup generalizes in many ways.
- Any node (even if observed) that we are uncertain about is modeled by a probability distribution.
- These nodes become the **random variables** of the model.
- The full model is specified via a **joint probability distribution** over all random variables.
- The goal is to **infer unknowns** of the model given the observed data.



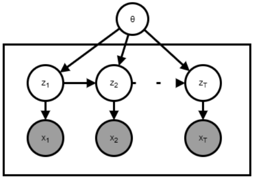
A Simple Supervised Learning Model



A Latent Variable Model for Unsupervised Learning



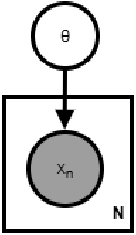
A Latent Variable Model for Supervised Learning



A Latent Variable Model for Sequential Data

Model Specification: Likelihood and Prior

- Probabilistic models require two key ingredients: **likelihood** and **prior**.
- **Likelihood** $p(x \mid \theta)$ ("observation model"): specifies how data is generated and measures data fit (loss) for a given θ.
- **Prior** $p(\theta)$: specifies how plausible parameter values are *a priori*; it often acts like a regularizer.
- Domain knowledge can guide both likelihood and prior choices.



$$p(\theta \mid X) \propto p(X \mid \theta)\, p(\theta)$$

Parameter Estimation vs. Bayesian Inference

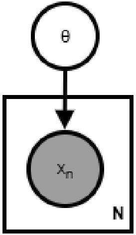
- A simplest approach is **point estimation**: find θ that makes the observed data most likely.

$$\hat{\theta} = \arg\max_{\theta} \log p(X \mid \theta).$$

- But a single point estimate does **not** quantify uncertainty in θ.
- **Bayesian inference** estimates the **full posterior**:

$$p(\theta \mid X) = \frac{p(X \mid \theta)\, p(\theta)}{p(X)} \propto \underbrace{p(X \mid \theta)}_{\text{Likelihood}} \times \underbrace{p(\theta)}_{\text{Prior}}.$$

- The posterior captures uncertainty in θ; we will study point estimation, Bayesian inference, and hybrids.



Posterior $=$ *Likelihood* $\times$ *Prior*
(normalized)

The Bayesian approach: overview

Core philosophy: Probability represents a *degree of belief*, updated with evidence.

Three key components

1. **Prior** $p(\theta)$: initial beliefs about parameters
2. **Likelihood** $p(D \mid \theta)$: probability of data given parameters
3. **Posterior** $p(\theta \mid D)$: updated beliefs after observing data

Goal

- compute posterior, summarize it, and make predictions for new observations $x^\star$

Bayes' theorem

$$\boxed{p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}}$$

Components

- $p(\theta \mid D)$ posterior (what we want)
- $p(D \mid \theta)$ likelihood
- $p(\theta)$ prior
- $p(D)$ evidence / marginal likelihood (normalization)

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta)$$

$$p(D) = \int p(D \mid \theta)\, p(\theta)\, d\theta$$

Bayesian inference pipeline (big picture)

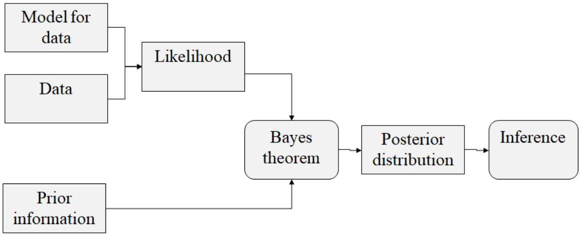


Figure: Bayesian inference

- Modeling: choose $p(D \mid \theta)$ and $p(\theta)$
- Inference: compute/approximate $p(\theta \mid D)$
- Decision/prediction: use posterior and posterior predictive

Prior distribution: encoding beliefs

Purpose: incorporate prior knowledge, stabilize inference, and regularize.

Types of priors

1. **Uninformative/flat**: minimal information (but still encodes assumptions)
2. **Weakly informative**: gentle regularization (prevents extreme values)
3. **Informative**: genuine prior knowledge (historical data, expert belief)
4. **Conjugate**: chosen for analytic convenience

Key principle: Posterior is a compromise between prior and likelihood; strong data can overwhelm weak priors.

How to choose a prior in practice

- **Domain knowledge:** expert judgment, historical datasets, physics constraints
- **Sensitivity analysis:** change priors and check how conclusions change
- **Prior predictive check:** sample from prior and see if it generates reasonable synthetic data
- **Weak vs strong:** with few observations, the prior matters more

Likelihood function: probability of data given parameters

$$p(D \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$$

Likelihood is a model of the data-generating process

- The likelihood encodes sensor physics + imperfections.
- It is **not** "how likely x is"; it is "how likely y is, if x were true".
- Choosing the likelihood is often the most important modeling decision.

Typical likelihood choices

- **Binomial**: successes in n trials
- **Gaussian**: continuous measurements with additive noise
- **Poisson**: counts / event arrivals

Posterior distribution: updated beliefs

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}$$

Posterior interpretation

- **Mode** (MAP), **mean**, **median**
- **Spread**: uncertainty about θ
- **Quantiles**: credible intervals

Posterior predictive distribution (PPD)

Goal: predict a new observation $x^\star$.

$$p(x^\star \mid D) = \int p(x^\star \mid \theta)\, p(\theta \mid D)\, d\theta$$

Interpretation

- averages predictions over all plausible parameter values
- accounts for parameter uncertainty and observation noise

Plug-in approximation (often used, but weaker)

$$p(x^\star \mid D) \approx p(x^\star \mid \hat{\theta})$$

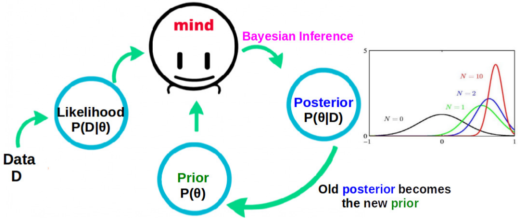


Figure: Bayesian Update

Conjugate priors: computational convenience

Definition: A prior is **conjugate** to a likelihood if the posterior has the same functional form as the prior.

Why it matters

- closed-form posterior (no sampling needed)
- easy to interpret updates
- sequential updating is straightforward

Likelihood	**Prior**	**Posterior**
Binomial / Bernoulli	Beta	Beta
Poisson	Gamma	Gamma
Gaussian (known σ)	Gaussian	Gaussian
Multinomial	Dirichlet	Dirichlet

When conjugacy breaks: optimization and approximation

- Many useful models do **not** have closed-form posteriors.
- Example: logistic regression likelihood (classification).
- Then we rely on:
  - **Optimization**: MAP via Newton / quasi-Newton.
  - **Approximation**: Laplace approximation, variational inference.
  - **Sampling**: MCMC (e.g., NUTS / HMC).

Lecture 2

Gaussian and linear models, Bayesian linear regression, and MAP logistic regression with Newton's method.

Beta–Binomial conjugacy: the canonical example

Model

- Prior: $\theta \sim \mathrm{Beta}(\alpha, \beta)$
- Likelihood: $k \sim \mathrm{Binom}(n, \theta)$
- Posterior: $\theta \mid k \sim \mathrm{Beta}(\alpha + k,\ \beta + n - k)$

Update rule

$$\alpha^\star = \alpha + k, \qquad \beta^\star = \beta + (n - k)$$

Pseudo-count interpretation

- prior contributes $\alpha - 1$ "successes" and $\beta - 1$ "failures"
- data contributes k successes and $n - k$ failures

Beta distribution: prior intuition

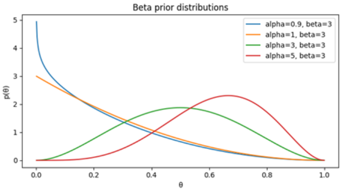


Figure: Beta prior shapes for different α with $\beta = 3$

- larger $\alpha + \beta \Rightarrow$ stronger prior (more concentrated)
- $\alpha > \beta$ biases belief toward larger θ; $\alpha < \beta$ toward smaller θ

Posterior derivation (up to proportionality)

Likelihood

$$p(k \mid \theta) \propto \theta^k (1-\theta)^{n-k}$$

Prior

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Posterior (unnormalized)

$$p(\theta \mid k) \propto \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

Posterior in closed form

$$\theta \mid k, n \sim \mathrm{Beta}(\alpha + k,\ \beta + n - k)$$

Posterior moments

$$\mathbb{E}[\theta \mid k] = \frac{\alpha + k}{\alpha + \beta + n}$$

$$\mathrm{Var}[\theta \mid k] = \frac{(\alpha + k)(\beta + n - k)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$

Prior vs posterior (visual)

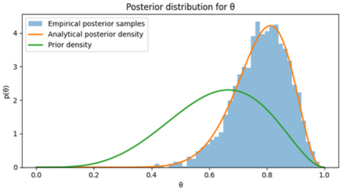


Figure: Prior and posterior density of θ

- Posterior shifts toward parameter values supported by data.
- Posterior becomes more concentrated as n increases.

How dataset size changes uncertainty

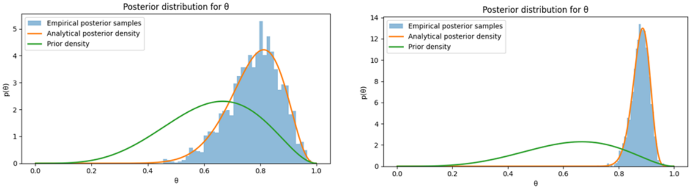


Figure: Posterior densities for $(k, n) = (9, 10)$ on the left and $(k, n) = (90, 100)$ on the right under the same prior.

- Same ratio k/n can imply very different uncertainty depending on n.

Why point estimates still matter

- Many applications require a single parameter value (for deployment simplicity).
- Point estimates are useful **summaries** of the posterior.
- MLE and MAP are the main point-estimation baselines.

Maximum Likelihood Estimation (MLE)

Definition: parameter value maximizing the likelihood.

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \log p(D \mid \theta)$$

Binomial model

$$\log p(k \mid \theta, n) = k \log \theta + (n - k) \log(1 - \theta) + C$$

$$\frac{\partial}{\partial \theta} = \frac{k}{\theta} - \frac{n - k}{1 - \theta} = 0 \quad \Rightarrow \quad \hat{\theta}_{\text{MLE}} = \frac{k}{n}$$

When to use: large sample sizes, purely data-driven estimation.

Maximum A Posteriori (MAP) estimation

Definition: parameter value maximizing the posterior.

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \log \left[p(D \mid \theta) p(\theta) \right]$$

Beta–Binomial MAP

$$\hat{\theta}_{\text{MAP}} = \frac{k + \alpha - 1}{n + \alpha + \beta - 2}$$

- incorporates prior information
- acts like regularization (shrinkage toward the prior)
- still a point estimate (does not capture posterior uncertainty)

MLE vs MAP: practical example

Scenario: coin flip with $k = 7$ successes in $n = 10$ trials.
Prior: $\mathrm{Beta}(\alpha = 5, \beta = 3)$.

$$\hat{\theta}_{\text{MLE}} = \frac{7}{10} = 0.700 \qquad \hat{\theta}_{\text{MAP}} = \frac{7+5-1}{10+5+3-2} = \frac{11}{16} = 0.6875$$

Interpretation

- MLE uses only data.
- MAP balances data and prior (slight pull toward prior mean).
- With much more data, MLE and MAP converge.

MAP as regularization (ML viewpoint)

$$\hat{\theta}_{\text{MAP}} = \arg\min_{\theta} \Big(\underbrace{-\log p(D \mid \theta)}_{\text{data fit}} + \underbrace{-\log p(\theta)}_{\text{regularizer}} \Big)$$

- Gaussian prior on weights $\Rightarrow$ L2 regularization.
- Laplace prior $\Rightarrow$ L1 regularization.
- Prior encodes what parameter values are plausible *before* seeing data.

Summarizing the posterior from samples

If we have posterior samples $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$, we can compute:

Central tendency

- mean: $\bar{\theta} = \frac{1}{S} \sum_{s=1}^{S} \theta^{(s)}$
- median: 50th percentile
- mode: most probable value (e.g., MAP)

Dispersion

- variance / standard deviation
- quantiles (IQR, 95% range)

Credible intervals: central (equal-tailed)

Central interval: contains $\alpha/2$ probability in each tail.

$$\mathrm{CrI}_{1-\alpha} = \left[q_{\alpha/2},\ q_{1-\alpha/2}\right]$$

Example (95%):

$$\mathrm{CrI}_{0.95} = \left[q_{0.025},\ q_{0.975}\right]$$

Interpretation:

$$\mathbb{P}\big(\theta \in [q_{0.025}, q_{0.975}] \mid D\big) = 0.95$$

- simple and widely used
- may exclude the mode for skewed posteriors

Credible intervals: highest posterior density (HPD/HDI)

HPD/HDI: region with the highest posterior density containing $(1 - \alpha)$ mass.

Intuition

- all points inside are more credible than points outside
- typically shorter than equal-tailed intervals for skewed posteriors
- may be multi-interval for multimodal posteriors

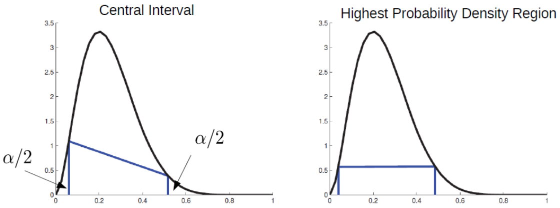


Figure: Central and HPD intervals

Posterior predictive vs plug-in prediction

Posterior predictive

$$p(x^\star \mid D) = \int p(x^\star \mid \theta)\, p(\theta \mid D)\, d\theta$$

- accounts for parameter uncertainty
- often wider (more honest)

Plug-in

$$p(x^\star \mid D) \approx p(x^\star \mid \hat{\theta})$$

- simpler
- can be overconfident

Analytic vs Monte Carlo inference

- Conjugate models: compute posterior in closed form.
- General models: posterior may be intractable $\Rightarrow$ approximate inference.

Two views

- **Analytic:** derive posterior formula directly.
- **Sampling:** approximate posterior with samples (MC, MCMC).

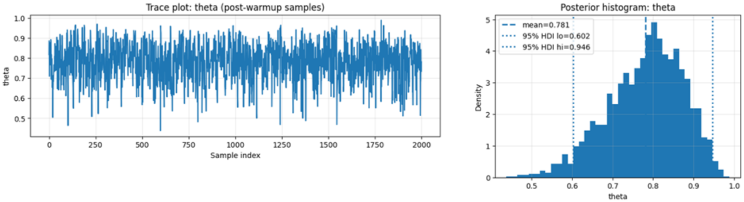


Figure: Posterior samples and their histogram

Evidence and Bayes factors (intuition)

Marginal likelihood (evidence)

$$p(\boldsymbol{X}) = \int p(\boldsymbol{X} \mid \theta)\, p(\theta)\, d\theta$$

- Penalizes overly flexible models automatically ("Occam factor").
- Enables model comparison: $\mathrm{BF}_{10} = \dfrac{p(\boldsymbol{X} \mid M_1)}{p(\boldsymbol{X} \mid M_0)}$.

In conjugate models

You can compute evidence in closed form (e.g., Beta-Binomial). For complex models we approximate.

Key takeaways

- UQ asks: what is plausible, how variable, and how beliefs update with data.
- Bayesian inference: posterior $\propto$ likelihood $\times$ prior.
- MLE and MAP are point estimates; full Bayes keeps a distribution.
- Credible intervals (central/HPD) summarize posterior uncertainty.
- PPD averages over parameter uncertainty to avoid overconfidence.

Acknowledgements/ Document preparation

This document was converted into LaTeX slides and formatted with assistance from ChatGPT (OpenAI). The instructor provided the source text and verified the final structure and wording. The instructor acknowledged that slides 10-13 are used and modified from Piyush Rai, Topics in Probabilistic Modeling and Inference (CS698X), 2019: Course Logistics, Intro to Probabilistic Modeling and Inference. The other material is based on Dr. Rai's course.