Inference & Causality
Week 2
Session 3

27.10.2025

Lecturer: Narges Chinichian
IU University of Applied Science, Berlin

Course Overview

Check the course hub on Notion for up-to-date information:

https://tinyurl.com/mrcjp79s

Outline of Week 2

- From Discrete to Continuous Bayes
- Conjugate Priors
- Markov Chains

From Discrete to Continuous Bayes

- Last week we learned:
  - Bayes' theorem combines prior × likelihood → posterior
  - "Cookie", "Dice", and "Monty Hall" illustrated our discrete reasoning.

$$P(A|B) = \frac{P(B|A)\,P(A)}{\sum_a P(B|a \in A)P(a \in A)}$$

$$\Downarrow$$

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}$$

When outcomes are continuous, we replace counting with integrating.
The posterior is a density function describing our updated belief about θ.

Why?
Motivation for Continuous Bayes

Problem: Discrete outcomes are easy to count.
But what if θ = “probability of success” or ”level of bias”, which could be any real number between 0 and 1?

“Could we still apply Bayes if there are infinitely many possibilities?”
We replace probability tables with probability densities.

Typical Things We Report from Bayesian Inference

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}$$

Expectation value $E[\theta \mid x] = \int \theta\, f(\theta \mid x)\, d\theta$

Credible interval $[b_l, b_u]$: $\int_{b_l}^{b_u} f\,(\theta \mid x)\, d\theta = 1 - \alpha$

Mode $\mathrm{Mod}(\theta \mid x) = \arg\max_\theta f(\theta \mid x)$

Discrete vs Continuous

Concept	Discrete	Continuous
Example	Dice, Cookies	Biased-coin bias θ, height, noise
Prior	A list of probabilities	A smooth curve (e.g. Beta, Normal)
Update	Multiply & normalize table	Multiply & normalize curves
Evidence	Count outcomes	Integrate likelihood across all θ

Expectation value $\boldsymbol{E[\theta \mid x] = \int \theta\, f(\theta \mid x)\, d\theta}$

- This is the average (expected) value of the parameter θ given the observed data x.
  It minimizes the mean squared error loss and is often used as the “best single-point” estimate of θ if you want the expected value under the posterior.

- **Good for:** Being wrong by the least amount on average.
  (Minimizes squared error — balances all possibilities.)

Credible interval $[b_l, b_\mathrm{u}]$: $\int_{b_l}^{b_u} f\,(\theta \mid x)\, d\theta = 1 - \alpha$

- This defines a range $[b_l, b_u]$ that contains the true parameter with probability $1 - \alpha$ given your model and data.
- For example, a 95% credible interval means that, given the data and the prior, there's a 95% probability that the true θ lies within that interval.
- This is different from a frequentist confidence interval, which says that if we repeated the experiment many times, 95% of such intervals would contain the true value.
- **Good for:** Describing the range where the truth probably lies. (Captures a chosen percentage of your belief, e.g. 95%.)

Mode $\mathrm{Mod}(\theta \mid x) = \arg\max_{\theta} f(\theta \mid x)$

- This is the most probable value of the parameter θ under the posterior — the point where $f(\theta \mid x)$ is highest.
  It minimizes the 0–1 loss function (if you just care about being exactly correct).

- **Good for:** Picking the single most plausible value.
  (Ignores uncertainty — just chooses the peak of belief.)

Game: Guess the Coin

- Imagine a coin but we don’t know if it’s fair.
- Prior belief: “probably fair” (centered at 0.5).
- Toss 5 times → get 4 heads.
- Update belief: shift curve toward right.

Run Notebook 1 of Week 2

How precisely can we estimate θ ?

- The shape of the likelihood tells us how much data “inform” our parameter.

- A steep (narrow) log-likelihood → precise estimate.

- A flat (wide) log-likelihood → uncertain estimate.

Fisher Information

$$\text{likelihood } L(\theta; x) = f(x \mid \theta)$$

$$I(\theta) = -\mathbb{E}\left[\left(\frac{d^2}{d\theta^2} \log L(\theta; X)\right)\right]$$

Fisher Information = expected curvature of the log-likelihood.
Steeper curve → more information → lower variance.

Have a look at notebook 2 of this week!

The Curse of Dimensionality

- When a model contains several continuous parameters like $\theta_1, \theta_1, \dots, \theta_k$, the posterior is defined over a k-dimensional space:

$$\boldsymbol{\theta} = (\theta_1, \theta_1, \dots, \theta_k)$$

$$f(\boldsymbol{\theta} \mid x) = \frac{f(x \mid \boldsymbol{\theta})\, f(\boldsymbol{\theta})}{f(x)},$$

- where

$$f(x) = \int_{\Theta_1} \dots \int_{\Theta_k} f\,(x, |, \theta_1 \dots \theta_k)\, f(\theta_1, \dots, \theta_k)\, d\theta_1 \cdots d\theta_k$$

- That denominator, the evidence, involves an integral over all combinations of parameter values.
  As the number of parameters grows, the number of integration points grows exponentially — this is known as the curse of dimensionality.
  Even simple Gaussian integrals quickly become intractable when $k > 4$–5.

What Are Conjugate Priors?

- We already knew the importance of **prior**.
- When our **prior** and **likelihood** come from the same "family," updating beliefs is easy.
- the **posterior** has the same shape as the **prior**, just with new parameters.
- If you choose a prior that is conjugate to the likelihood (e.g. Beta–Binomial, Gamma–Poisson, Normal–Normal),
- the math simplifies, and those integrals have closed-form updates.

Conjugate Priors

make Bayesian updating simple — no need for integrals.

- We can “see” how data shapes beliefs.
- Useful for real-time or small data cases (counting, binary success, rates).
- Historically, the first Bayesian models used them before computers could sample.

Conjugate Game

- Let’s play a quick game:
- Each of you has to take a paper from one of the stacks.
- You need to find the conjugate distribution of you and form a group.
- Read up on what are those distributions and explain them with one good example.
- You could use notebook 3 of this week and make a glossary for your distributions (optional).

Review of some relevant distributions

Likelihood	Model params	Conjugate prior
Bernoulli	p (probability)	**Beta**
Binomial (known trials (m))	p (probability)	**Beta**
Negative binomial (known failures (r))	p (probability)	**Beta**
Poisson	λ (rate)	**Gamma**
Exponential	λ (rate)	**Gamma**
Normal (with known variance σ^2)	μ (mean)	**Normal**
Normal (with known mean μ)	σ2 (variance)	**Inverse Gamma**

Read more on: https://en.wikipedia.org/wiki/Conjugate_prior

But in real models (nonlinear regression, neural nets, hierarchical models), no such analytical simplification exists.

When we can’t integrate analytically

- We must approximate the posterior:
- by sampling (drawing representative points),
- or by approximating it with a simpler shape.
- For low-dimensional problems, we can evaluate the posterior at many θ values and take a weighted average.
- But as dimensionality increases, the required number of samples explodes.

What is the Monte Carlo Method?

Let’s have a look at notebook 4 of today.

Let's Play a Markov Chain Game Together.

Modern Approach: Markov Chain Monte Carlo (MCMC)

- To handle realistic models, we use algorithms that sample efficiently from high-dimensional posteriors without evaluating the full multidimensional integral explicitly.
- Markov Chain Monte Carlo (MCMC) methods (such as Metropolis–Hastings, Gibbs sampling, and later Hamiltonian Monte Carlo) approximate the integral by simulating a “chain” that explores the posterior space.

- These methods let us compute expectations, variances, or credible intervals numerically:

$$E[g(\theta) \mid x] \approx \frac{1}{N}\sum_{i=1}^{N} g\left(\theta^{(i)}\right),$$

where each $\theta^{(i)}$ is a draw from the posterior.

Congratulations!
We finished unit 1 of this course.

Don’t forget to read unit one of your course book for more detailed understanding of this unit.

Online Test Unit 1

- Now you should be ready to take Online Test for Unit 1 on your mycampus platform.

Homework

- Exercise: Fill out the exercises on notebooks 1 and 2 and 4 for this week, commit your answers and submit .