STAT 7650 - Computational Statistics Lecture Slides

Numerical Integration

Elvan Ceyhan

AU

Updated: February, 2025

- Based on parts of: Chapter 5 in Givens & Hoeting (Computational Statistics), and Chapters 4 & 18 of Lange (Numerical Analysis for Statisticians).

Outline

Motivation

While many statistical problems rely on optimization, there are also some that require numerical integration.

- **Bayesian statistics** is almost exclusively integration.
- Data admits a likelihood function $L(\boldsymbol{\theta})$;
- $\boldsymbol{\theta}$ unknown, so assign it a weight function $\pi(\boldsymbol{\theta})$, called the *prior*;
- Combine prior and data using Bayes's formula, to obtain the *posterior*:

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')d\boldsymbol{\theta}'}$$

Need to compute probabilities and expectations (usually with respect to the posterior) — integrals! Some "non-Bayesian" problems may involve integration, e.g., random- or mixed-effects models. Other approaches besides Bayesian and frequentist also exist...

Intuition

There are a number of classical numerical integration techniques, simple and powerful.

Recall from Calculus

- Integral was defined as the limit of sum of areas of rectangles approximating the area under the curve (i.e. function) on small intervals.
- Numerical integration, or quadrature, is based on this definition and refinements thereof.

Basic Principle[1] Approximate the function on a small interval by a "nice" one that you know how to integrate.

- Works well for one- or two-dimensional integrals; for higher-dimensional integrals, other tools are needed.

[1] This principle also motivated the various methods discussed for optimization.

Notation and Partitioning the Interval - I

Consider a function $f(x)$ that we wish to integrate over the interval $[a, b]$.

- Choose a relatively large integer n to divide the interval into small subintervals of width: $h = \frac{b-a}{n}$.
- Define partition points:
  $x_i = a + ih, \quad \text{for } i = 0, 1, \ldots, n-1, \quad \text{and } x_n = b.$

Thus, the explicit list of points is: $x_0 = a$, $x_1 = a + h$, $x_2 = a + 2h$, $x_3 = a + 3h$, $\ldots$, $x_{n-1} = a + (n-1)h$, $x_n = a + nh = b$.

Therefore,

$$\int_a^b f(x)\,dx = \int_{x_0=a}^{x_1} f(x)\,dx + \int_{x_1}^{x_2} f(x)\,dx + \cdots + \int_{x_{n-1}}^{x_n=b} f(x)\,dx$$

$$= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)\,dx.$$

Notation and Partitioning the Interval - II

Key Idea: If $f(x)$ is sufficiently smooth, it can be well-approximated by a simpler function $f_i(x)$ over each subinterval $[x_i, x_{i+1}]$. That is, $f(x) \approx f_i(x)$ over $[x_i, x_{i+1}]$ so that

$$\int_{x_i}^{x_{i+1}} f(x)\,dx \approx \int_{x_i}^{x_{i+1}} f_i(x)\,dx$$

Thus, we can approximate the definite integral using a summation:

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)\,dx \approx \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f_i(x)\,dx.$$

In each subinterval $[x_i, x_{i+1}]$ we insert $m+1$ evaluation points (called nodes) as $x_{i0} = x_i, \ldots, x_{im} = x_{i+1}$.

Numerical Approximation of the Integral

To approximate the integral numerically, we replace each subinterval integral with a weighted sum of function evaluations:

$$\int_{x_i}^{x_{i+1}} f(x)\,dx \approx \sum_{j=0}^{m} A_{ij} f(x_{ij}).$$

where

x_{ij} are selected evaluation points (nodes) within the subinterval $[x_i, x_{i+1}]$,

A_{ij} are corresponding weights that depend on the chosen numerical method,

m is the number of evaluation points per subinterval.

Final Approximation:

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n-1} \sum_{j=0}^{m} A_{ij} f(x_{ij}).$$

Outline

Polynomial Approximation

Consider the following sequence of polynomials for $j = 0, \ldots, m$:

$$p_{ij}(x) = \begin{cases} 1, & \text{if } m = 0 \\ \prod_{k=0, k\neq j}^{m} \frac{x - x_{ik}}{x_{ij} - x_{ik}}, & m = 1, 2, \ldots. \end{cases}$$

Then, the mth degree polynomial that interpolates $f(x)$ at the nodes $x_{i0}, \ldots, x_{im}$ is given by:

$$p_i(x) = \sum_{j=0}^{m} p_{ij}(x) f(x_{ij})$$

This polynomial can approximate the integral of $f(x)$ over $[x_i, x_{i+1}]$ as:

$$\int_{x_i}^{x_{i+1}} f(x)dx \approx \int_{x_i}^{x_{i+1}} p_i(x)dx = \sum_{j=0}^{m} \underbrace{\left(\int_{x_i}^{x_{i+1}} p_{ij}(x)dx \right)}_{A_{ij}} f(x_{ij})$$

where A_{ij} are the weights for the approximation.

Riemann Rule: $m = 0$

Approximation by a Constant
Approximate $f(x)$ on $[x_i, x_{i+1}]$ by a constant.

- Here $x_{i0} = x_i$ and $p_{i0}(x) \equiv 1$, so $A_{i0} = h$ and

$$\int_a^b f(x)dx \approx \sum_{i=0}^{n-1} f(x_i)(x_{i+1} - x_i) = h \sum_{i=0}^{n-1} f(x_i).$$

Features of Riemann Rule

- Very easy to program — only need $f(x_0), \ldots, f(x_n)$.
- Can be slow to converge, i.e., many x_i's may be needed to get a good approximation.

Trapezoid Rule: $m = 1$

Linear Function Approximation

Approximate $f(x)$ on $[x_i, x_{i+1}]$ by a linear function. In this case:

$$x_{i0} = x_i,\ x_{i1} = x_{i+1} \text{ and } A_{i0} = A_{i1} = \frac{x_{i+1} - x_i}{2} = \frac{h}{2}.$$

Therefore,

$$\begin{aligned}\int_a^b f(x)dx &\approx \frac{h}{2}\sum_{i=0}^{n-1}(f(x_i) + f(x_{i+1})) = \frac{h}{2}\sum_{i=0}^{n-1}(f(x_i) + f(x_i + h)) \\ &= \frac{h}{2}(f(x_0) + f(x_0 + h) + f(x_0 + 2\,h) + \ldots + f(x_0 + n\,h)) \\ &= \frac{h}{2}(f(a) + f(a + h) + f(a + 2\,h) + \ldots + f(b)) =: T(h)\end{aligned}$$

Features of the Trapezoid Rule

- Still only requires function evaluations at the x_i's.
- More accurate than Riemann because the linear approximation is more flexible than constant.
- Can derive bounds on the approximation error.

Trapezoid Rule (Cont'd)

Euler-Maclaurin Formula:
A general tool to study the precision of the trapezoid rule is the Euler-Maclaurin formula. For a function $g(x)$ twice differentiable, we have:

$$\sum_{t=0}^{n} g(t) = \int_0^n g(t)dt + \frac{1}{2}[g(0) + g(n)] + C_1 g'(t)\Big|_0^n + R_n,$$

where $|R_n| \leq C_2 \int_0^n |g''(t)|dt$.

Implication for Trapezoid Rule:
Letting $g(t) = f(a + ht)$, the trapezoidal approx $T(h)$ becomes:

$$T(h) := h\left[\frac{1}{2}g(0) + g(1) + \ldots + g(n-1) + \frac{1}{2}g(n)\right].$$

This formula helps us understand the error in the trapezoid rule in terms of the function's second derivative, allowing for a precise error estimation.

Trapezoid Rule (Cont'd)

Using the Euler-Maclaurin formula in $T(h)$:

$$T(h) = h\sum_{t=0}^{n} g(t) - \frac{h}{2}[g(0)+g(n)] \approx h\int_0^n g(t)dt + hC_1\left[g'(t)\right]\Big|_0^n$$

$$= h\int_a^b \frac{1}{h} f(x)dx + hC_1\left[hf'(b) - hf'(a)\right]$$

Therefore, the error in the trapezoid rule approximation:

$$\left|T(h) - \int_a^b f(x)dx\right| = O(h^2), \text{ as } h \to 0.$$

Trapezoid Rule (Cont'd)

Can Trapezoid Error $O(h^2)$ Be Improved?
Although our initial derivation suggested an error of $O(h^2)$, the next term in the expansion actually implies an improvement to $O(h^4)$ is possible.

Romberg's Rule:
Romberg found that manipulating $T(h)$ can cancel the $O(h^2)$ term, enhancing precision to $O(h^4)$:

$$\frac{4T(\frac{h}{2}) - T(h)}{3} = \int_a^b f(x)dx + O(h^4), \quad \text{as } h \to 0.$$

Iterative Improvement:
This technique can be iterated for further improvement (see Section 5.2 in G&H), highlighting a significant advancement in numerical integration strategies.

Simpson Rule: $m = 2$

Quadratic Function Approximation
Approximate $f(x)$ on $[x_i, x_{i+1}]$ by a quadratic function. Similar arguments as above gives the x_i's and A_{ij}'s for the approximation.

Simpson's Rule Approximation
The approximation formula is given by:

$$\int_a^b f(x)dx \approx \frac{h}{6}\sum_{i=0}^{n-1}\left(f(x_i) + 4f\left(\frac{x_i + x_{i+1}}{2}\right) + f(x_{i+1})\right).$$

Accuracy and Simplification

- More accurate than the trapezoid rule with an error of $O(n^{-4})$.
- For even n, the formula simplifies further, enhancing computational efficiency. See Equation (5.20) in G&H and the accompanying `R` code in Canvas for detailed implementation.

Remarks on Numerical Integration

Scalability with m
The approximation accuracy improves as m increases.

Extension to Multi-variable Functions
While the method can be extended to functions of more than one variable, the complexity of the details increases significantly.

Software and Practical Considerations

- In `R`, the function `integrate` is used for one-dimensional integration.
- Numerical methods and corresponding software generally perform very well, yet careful consideration is needed to ensure accuracy and reliability. See Section 5.4 in G&H for a detailed discussion of numerical integration.

Example: Bayesian Analysis of Binomial

Consider $X \sim \text{Bin}(n, \theta)$ where n is known and θ is unknown.

Prior Distribution:
The prior for θ is the semicircle distribution with density:

$$\pi(\theta) = 8\pi^{-1}\left[\frac{1}{4} - \left(\theta - \frac{1}{2}\right)^2\right]^{1/2}, \quad \theta \in [0, 1].$$

Posterior Distribution:
The posterior density is given by:

$$\pi(\theta|x) = \frac{\theta^x(1-\theta)^{n-x}\left[\frac{1}{4} - \left(\theta - \frac{1}{2}\right)^2\right]^{1/2}}{\int_0^1 \theta^x(1-\theta)^{n-x}\left[\frac{1}{4} - \left(\theta - \frac{1}{2}\right)^2\right]^{1/2} d\theta}.$$

Bayes Estimate:
Calculating the Bayes estimate of θ, the posterior mean, requires numerical integration.

Example: Mixture Densities

Mixture distributions are prevalent models, known for their flexibility. They are particularly useful for density estimation and modeling heavy-tailed distributions.

General Mixture Model: Can be represented as:

$$p(y) = \int k(y|x)f(x)dx,$$

where

- $k(y|x)$ is a probability density function (pdf) or probability mass function (pmf) in y for each x.
- $f(x)$ is a pdf or pmf.

Straightforward to verify that $p(y)$ is a pdf or pmf depending on k.

Evaluation of $p(y)$ Evaluating $p(y)$ for each specified y requires integration, e.g. with numerical methods in this chapter.

Example: Mixture Density Model

Suppose we have two normal distributions with pdfs:

$$k_1(y|x) = N(y; \mu_1, \sigma_1^2) \quad \text{and} \quad k_2(y|x) = N(y; \mu_2, \sigma_2^2)$$

and we mix these with proportions $f(x) = \pi$ and $1 - \pi$ respectively, where $0 \leq \pi \leq 1$.

The mixture density $p(y)$ can then be expressed as:

$$p(y) = \pi N(y; \mu_1, \sigma_1^2) + (1 - \pi) N(y; \mu_2, \sigma_2^2)$$

Mixture Density Formulation

For this example, let's choose specific values for our parameters:

- $\mu_1 = 0$, $\sigma_1^2 = 1$ (Standard normal distribution)
- $\mu_2 = 3$, $\sigma_2^2 = 4$ (Normal distribution with mean 3 and variance 4)
- $\pi = 0.5$ (Equally weighted mixture)

Given these values, our mixture density model is:

$$p(y) = 0.5N(y; 0, 1) + 0.5N(y; 3, 4)$$

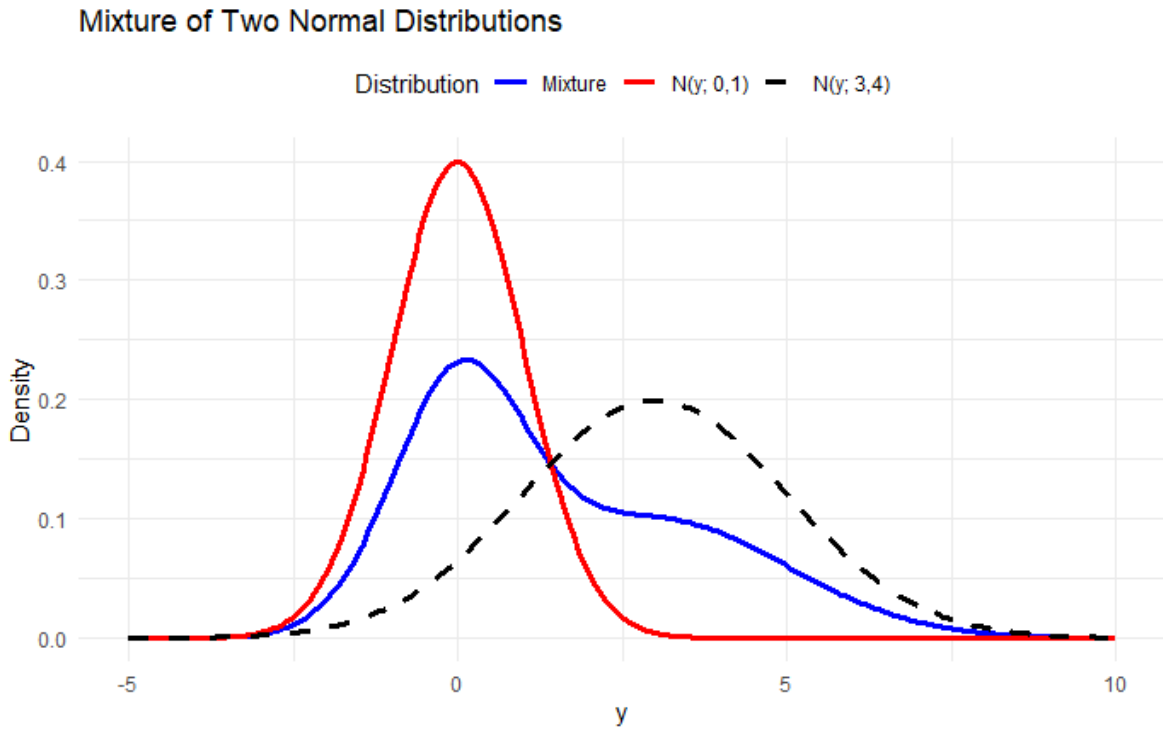

Figure 1: Illustration of the normal mixtures.

Interpretation: This model represents a distribution where half of the data is expected to follow a standard normal distribution, and the other half follows a normal distribution centered at 3 with a larger variance. This mixture can model a scenario where a dataset might be composed of two underlying populations or processes, each with its own normal distribution.

Example 5.1 in G&H: GLMM

Model Description
Consider a Generalized Linear Mixed Model (GLMM) where:

$$Y_{ij} \sim \text{Poi}(\lambda_{ij}), \quad \lambda_{ij} = \exp(\beta_0 + \beta_1 j + \gamma_i),$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. Here, $\gamma_1, \ldots, \gamma_n$ are iid $N(0, \sigma_\gamma^2)$.

Model Parameters: The model parameters are $(\beta_0, \beta_1, \sigma_\gamma^2)$.

Marginal Likelihood
The marginal likelihood for $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma_\gamma^2)$ is given by:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int \left(\prod_{j=1}^{J} \text{Pois}(Y_{ij} | \exp(\gamma_i + \beta_0 + \beta_1 j)) N(\gamma_i | 0, \sigma_\gamma^2) \right) d\gamma_i.$$

Objective
The goal is to maximize $L(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$, necessitating advanced computational methods for integral evaluation and optimization.

Example 5.1 in G&H: Mixture Components

The marginal likelihood here follows the general mixture model structure: $p(y) = \int k(y|x)f(x)dx$.

where the mixture components are **Component 1:** The conditional distribution $k(y|x)$ is the Poisson likelihood:

$$k(Y_{ij}|\gamma_i) = \text{Pois}(Y_{ij}|\exp(\beta_0 + \beta_1 j + \gamma_i)).$$

Component 2: The mixing distribution $f(x)$ is the normal distribution of the random effects:

$$f(\gamma_i) = N(\gamma_i|0, \sigma_\gamma^2).$$

Mixture Representation: The marginal likelihood integrates out γ_i:

$$L(\theta) = \prod_{i=1}^{n} \int \left(\prod_{j=1}^{J} \text{Pois}(Y_{ij}|\exp(\gamma_i + \beta_0 + \beta_1 j)) N(\gamma_i|0, \sigma_\gamma^2) \right) d\gamma_i.$$

Example 5.1 in G&H (Cont'd)

Log-Likelihood: Taking the logarithm, we obtain the log-likelihood function:

$$\ell(\theta) = \sum_{i=1}^{n} \log \underbrace{\left(\int \left(\prod_{j=1}^{J} \text{Pois}(Y_{ij} | \exp(\gamma_i + \beta_0 + \beta_1 j)) \cdot N(\gamma_i | 0, \sigma_\gamma^2) \right) d\gamma_i \right)}_{L_i(\theta)}.$$

Gradient Evaluation: G&H discuss evaluating the gradient with respect to β_1, which involves computing:

$$\frac{\partial}{\partial \beta_1} L_i(\theta) = \int \left(\sum_{j=1}^{J} j(Y_{1j} - \exp(\gamma_1 + \beta_0 + \beta_1 j)) \right) \times \left(\prod_{j=1}^{J} \text{Pois}(Y_{1j} | \exp(\gamma_1 + \beta_0 + \beta_1 j)) \cdot N(\gamma_1 | 0, \sigma_\gamma^2) \right) d\gamma_1.$$

See the Appendix at the end of the slides for the derivation.

Reproduction of Results: Reproduce Tables 5.2-5.4 using `R` code.

Outline

Very Brief Summary: Gaussian Quadrature

Gaussian Quadrature vs. Newton-Cotes: Gaussian quadrature, as an alternative to Newton-Cotes for numerical integration, is particularly advantageous when integrating with respect to a non-uniform measure, such as in calculating expectations.

Key Concept: The core idea behind Gaussian quadrature is the utilization of a sequence of "orthogonal polynomials" identified by the integration measure, yielding more precise function approximations than Newton-Cotes.

Advantages

- Increased accuracy in approximating integrals, especially for complicated measures.
- Ensures computational efficiency via orthogonal polynomials.

Note: The book provides minimal details on Gaussian quadrature, and it will not be covered in depth here.

Outline

Setup for Laplace Approximation

The Laplace approximation is a powerful method for approximating certain types of integrals, using optimization techniques.

Integral Form: Consider integrals of the form:

$$J_n := \int_a^b f(x)e^{ng(x)}dx, \quad n \to \infty,$$

where $a < b$ can be finite or infinite, and f and g are sufficiently nice functions. Additionally, g has a unique maximizer $\hat{x} = \arg\max g(x)$ within the interval (a, b).

Claim: When n is large, the major contribution to the integral comes from a neighborhood around $\hat{x}$, the maximizer of g.[2] This principle underlies the Laplace approximation's efficiency.

[2] For a proof and further discussion on this claim, see Section 4.7 in Lange.

Formula for Laplace Approximation

Local Approximation: With the premise stated in the "claim", it suffices to restrict the range of integration to a small neighborhood around $\hat{x}$, denoted $\text{nbhd}(\hat{x})$, where the function $g(x)$ can be approximated as:

$$g(x) \approx g(\hat{x}) + g'(\hat{x})(x - \hat{x}) + \frac{1}{2}g''(\hat{x})(x - \hat{x})^2.$$

The linear term vanishes because $g'(\hat{x}) = 0$ at the maximizer.

Integral Transformation: The integral J_n can be approximated by:

$$J_n \approx e^{ng(\hat{x})} \int_{\text{nbhd}(\hat{x})} f(x) e^{\frac{1}{2}ng''(\hat{x})(x-\hat{x})^2} dx,$$

which can be written as:
$= e^{ng(\hat{x})} \int_{\text{nbhd}(\hat{x})} f(x) e^{-\frac{1}{2}[-ng''(\hat{x})](x-\hat{x})^2} dx.$

Importance: This approximation effectively reduces the integral to a more manageable form, focusing on the significant contributions

Formula for Laplace Approximation (Cont'd)

Continuing from the previous slide, we have:

$$J_n \approx e^{ng(\hat{x})} \int_{\text{nbhd}(\hat{x})} f(x) e^{-\frac{1}{2}[-ng''(\hat{x})](x-\hat{x})^2} dx.$$

Two Observations

- Since $\hat{x}$ is a maximizer, $g''(\hat{x}) < 0$.
- In a small neighborhood, $f(x) \approx f(\hat{x})$.

Refined Approximation: Thus, the integral J_n can be further approximated as:

$$\begin{aligned} J_n \approx & f(\hat{x}) e^{ng(\hat{x})} \int_{\text{nbhd}(\hat{x})} e^{-\frac{1}{2}[-ng''(\hat{x})](x-\hat{x})^2} dx \\ \approx & \sqrt{2\pi} f(\hat{x}) e^{ng(\hat{x})} \left[-ng''(\hat{x})\right]^{-\frac{1}{2}}. \end{aligned}$$

This formula provides a concise approximation for J_n, emphasizing the crucial role of the maximizer $\hat{x}$ and simplifying the calculation of complex integrals.

Example: Stirling's Formula

Stirling's formula provides an approximation for factorials, crucial for various mathematical and statistical applications.

Gamma Function Representation: The factorial can be expressed as a gamma function:

$$n! = \Gamma(n+1) = \int_0^\infty z^n e^{-z} dz.$$

Change of Variable: By changing the variable $x = z/n$, we obtain:

$$n! = n^{n+1} \int_0^\infty e^{n \log x - nx} dx, \quad g(x) = \log x - x,$$

where $g(x)$ has a maximizer $\hat{x} = 1$ within the interval $(0, \infty)$.

Laplace Approx.: For large n, the Laplace approximation yields:

$$n! \approx n^{n+1} e^{-n\, g(1)} \sqrt{2\pi} (-n g''(1))^{-\frac{1}{2}} = \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}.$$

Laplace Approximation in Bayesian Inference

Bayesian Setup: In Bayesian analysis, we often want to compute the posterior distribution of a parameter θ given observed data y. According to Bayes' theorem:

$$\pi(\theta \mid y) = \frac{L(\theta \mid y)\, p(\theta)}{\int L(\theta \mid y) p(\theta)\, d\theta},$$

where:

- $L(\theta \mid y)$ is the **likelihood** of the data given the parameter.
- $p(\theta)$ is the **prior** on θ.
- $p(y) = \int L(\theta \mid y) p(\theta)\, d\theta$ is the **normalizing constant**.
- $\pi(\theta \mid y)$ is the **posterior** distribution.

For complex models, the posterior $\pi(\theta \mid y)$ may be difficult to compute directly, primarily due to the intractability of the normalization constant $p(y)$. Instead, we apply the **Laplace approximation**, which approximates the posterior with a Gaussian distribution centered at its mode.

Laplace Approximation in Bayesian Inference - Steps

I- Identifying the Function to Approximate

Taking the logarithm of the posterior:

$$\log \pi(\theta \mid y) = \log L(\theta \mid y) + \log p(\theta) - \log p(y).$$

Define $g(\theta) = \log L(\theta \mid y) + \log p(\theta)$, which represents the **log unnormalized posterior** (ignoring the constant $\log p(y)$).

II- Finding the Mode (MAP Estimator)

The Laplace approximation is built around expanding $g(\theta)$ near its maximum. The maximum a posteriori (MAP) estimate is:

$$\hat{\theta} = \arg\max_{\theta} g(\theta),$$

found by solving:

$$\frac{d}{d\theta} g(\theta) = 0.$$

Laplace Approximation in Bayesian Inference - Steps

III- Second-Order Taylor Expansion Around $\hat{\theta}$

Use a **second-order Taylor expansion** around $\hat{\theta}$:

$$g(\theta) \approx g(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T H(\hat{\theta}) (\theta - \hat{\theta}),$$

where

$$H(\hat{\theta}) = \frac{d^2}{d\theta^2} g(\theta)\Big|_{\theta=\hat{\theta}}$$

is the **Hessian matrix** evaluated at $\hat{\theta}$.

Since this expansion is analogous to the log of a normal distribution, exponentiating both sides yields:

$$\pi(\theta \mid y) \approx N(\hat{\theta},\ -[H(\hat{\theta})]^{-1}).$$

That is, the Laplace approximation substitutes the posterior with a normal distribution centered at $\hat{\theta}$, with covariance given by the inverse Hessian of $g(\theta)$.

Interpretation of the Laplace Approximation

- Particularly useful for θ that is **high-dimensional**, where direct integration is infeasible.
- The **precision matrix** of the Gaussian approximation is $-H(\hat{\theta})$, hence the covariance is $[-H(\hat{\theta})]^{-1}$.
- Works well when the posterior is **unimodal** and roughly **Gaussian-shaped** near $\hat{\theta}$.

Example: Bayesian Posterior Expectations

In Bayesian analysis, we compute posterior expectations to infer parameter values based on observed data.

Bayesian Framework
Given:

- $L(\theta)$ as the likelihood based on n iid observations.
- $\pi(\theta)$ as a prior density.

The posterior expectation is defined as:

$$\mathbf{E}[h(\theta)|\text{data}] = \frac{\int h(\theta)L(\theta)\pi(\theta)d\theta}{\int L(\theta)\pi(\theta)d\theta}.$$

Example: Bayesian Posterior Expectations (Cont'd)

Laplace Approximation Application
For large n, applying Laplace's approximation to both the numerator and denominator simplifies the posterior expectation to:

$$\mathbf{E}[h(\theta)|\text{data}] \approx h(\hat{\theta}),$$

where $\hat{\theta}$ is the maximum likelihood estimate (MLE).

Importance
This result demonstrates that the proximity of the posterior mean to the MLE in large sample scenarios is not coincidental (e.g., previous binomial example that showed posterior mean close to MLE), highlighting the efficiency of Laplace approximation in Bayesian inference.

Example: Bayesian Logistic Regression

Consider a Bayesian logistic regression model:

$$Y_i \sim \text{Ber}(\sigma(\boldsymbol{\beta}^T \mathbf{x}_i)),$$

where the logistic function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$

We assume a Gaussian prior on the parameter vector $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}) = N(\mathbf{0}, \tau^2 I).$$

Posterior Distribution:

$$\pi(\boldsymbol{\beta}|\mathbf{y}) \propto \prod_{i=1}^{n} \sigma(\boldsymbol{\beta}^T \mathbf{x}_i)^{y_i} (1 - \sigma(\boldsymbol{\beta}^T \mathbf{x}_i))^{1-y_i} \cdot \exp\left(-\frac{1}{2\tau^2} \|\boldsymbol{\beta}\|^2\right).$$

Example: Bayesian Logistic Regression

Posterior Distribution:

$$\pi(\beta|\mathbf{y}) \propto \prod_{i=1}^{n} \sigma(\beta^T \mathbf{x}_i)^{y_i} (1 - \sigma(\beta^T x_i))^{1-y_i} \cdot \exp\left(-\frac{1}{2\tau^2}\|\beta\|^2\right).$$

Computing this posterior is intractable for large datasets, so we approximate it using the **Laplace method**:

1. Find the MAP estimate $\hat{\beta}$ by maximizing $\log p(\mathbf{y}|\beta) + \log p(\beta)$.
2. Compute the Hessian $H(\hat{\beta})$ of the log-posterior.
3. Approximate $\pi(\beta|\mathbf{y})$ with $N(\hat{\beta}, -H(\hat{\beta})^{-1})$.

This provides an efficient Gaussian approximation to the posterior, enabling predictions and uncertainty quantification without requiring expensive MCMC sampling.

Remarks on Laplace Approximation

Error Rate: The error in the Laplace approximation is $O(n^{-1})$, indicating high accuracy for large sample sizes. This error rate can be further improved with additional refinements.

Core Idea: Locally, the target integrals can be approximated by Gaussian integrals, simplifying their evaluation significantly.

Extensions to Multivariate Cases: The Laplace approximation is not limited to univariate integrals; it extends naturally to multivariate integrals, where it becomes even more useful.

Boundary Maximizers: A variant of the Laplace approximation caters to situations where the maximizer of g is on the boundary, which adapts the integral to resemble exponential or gamma integrals.

Further Reading: For more details, especially on the boundary case, see Section 4.6 of Lange.

Outline

Remarks on Numerical Integration Methods

- **Quadrature methods** are highly effective for numerical integration, particularly in one to two dimensions. However, their efficacy diminishes in higher dimensions due to the "curse of dimensionality," which necessitates an exponentially growing number of grid points for accurate approximation.
- The **Laplace approximation** remains practical in higher dimensions but is specifically suited for certain types of integrals. Statistically relevant integrals often fall into this category; so, Laplace approximation is a valuable tool in statistical analysis.
- **Monte Carlo Methods:** For integrals in higher dimensions, Monte Carlo methods become preferable due to their:
  - General ease of implementation.
  - Approximation accuracy, which is notably independent of the dimensionality of the integral.
- We will delve into Monte Carlo methods in greater detail later.

APPENDIX

Derivation of the Gradient for β_1 in Example 5.1 of G&H

Log-Likelihood:
Taking the logarithm of the marginal likelihood, we obtain the log-likelihood function:

$$\ell(\theta) = \sum_{i=1}^{n} \log \underbrace{\left(\int \left(\prod_{j=1}^{J} \text{Pois}(Y_{ij} \mid e^{\gamma_i + \beta_0 + \beta_1 j}) \cdot N(\gamma_i \mid 0, \sigma_\gamma^2) \right) d\gamma_i \right)}_{L_i(\theta)}.$$

Steps for Gradient Evaluation (w.r.t. β_1):

1. **Differentiate the log-likelihood.**

$$\frac{\partial}{\partial \beta_1} \ell(\theta) = \sum_{i=1}^{n} \frac{1}{L_i(\theta)} \cdot \frac{\partial L_i(\theta)}{\partial \beta_1}.$$

2. **Pull the derivative inside the integral.** By regularity conditions (interchanging differentiation and integration), we write:

$$\frac{\partial L_i(\theta)}{\partial \beta_1} = \int \frac{\partial}{\partial \beta_1} \left[\prod_{j=1}^{J} \text{Pois}(Y_{ij} \mid e^{\gamma_i + \beta_0 + \beta_1 j}) \cdot N(\gamma_i \mid 0, \sigma_\gamma^2) \right] d\gamma_i.$$

3. **Differentiate the Poisson term.** Recall:

$$\text{Pois}(Y_{ij} \mid \lambda_{ij}) = \frac{\lambda_{ij}^{Y_{ij}} e^{-\lambda_{ij}}}{Y_{ij}!}, \quad \lambda_{ij} = e^{\gamma_i + \beta_0 + \beta_1 j}.$$

Taking $\partial / \partial \beta_1$ introduces a factor of

$$j(Y_{ij} - \lambda_{ij})$$

due to the chain rule on $\lambda_{ij} = e^{\gamma_i + \beta_0 + \beta_1 j}$.

4. **Combine factors.** Within the integrand, $N(\gamma_i \mid 0, \sigma_\gamma^2)$ does not depend on β_1. Consequently,

$$\frac{\partial L_i(\theta)}{\partial \beta_1} = \int \Big(\sum_{j=1}^{J} j \big[Y_{ij} - e^{(\gamma_i + \beta_0 + \beta_1 j)} \big] \Big) \Big(\prod_{j=1}^{J} \text{Pois}(Y_{ij} \mid e^{\gamma_i + \beta_0 + \beta_1 j}) \Big) \cdot N(\gamma_i \mid 0, \sigma_\gamma^2) \, d\gamma_i.$$

5. **Insert into the sum.** Plugging this back into the partial derivative of $\ell(\theta)$ yields:

$$\frac{\partial}{\partial \beta_1} \ell(\theta) = \sum_{i=1}^{n} \frac{1}{L_i(\theta)} \int \Big(\sum_{j=1}^{J} j \big[Y_{ij} - e^{(\gamma_i + \beta_0 + \beta_1 j)} \big] \Big) \Big(\prod_{j=1}^{J} \text{Pois}(Y_{ij} \mid e^{\gamma_i + \beta_0 + \beta_1 j}) \Big) \cdot N(\gamma_i \mid 0, \sigma_\gamma^2) \, d\gamma_i.$$