STATISTICS FOR NEUROSCIENCE RESEARCH

9.073 / HST 460

Class 5: Bayesian Methods

Emery N. Brown
Sourish Chakravarty

February 27, 2017

Outline

1. The Bayesian Paradigm

2. Summaries of Posterior Densities

3. Computing the Posterior Density
    a. Analytically (Conjugate Priors)
    b. Gaussian Approximation to the Posterior
    c. Monte Carlo Methods
        1. Importance Sampling
        2. Rejection Method

The Bayesian Paradigm

Assume we have independent observations

$$x = (x_1, \ldots, x_n) \qquad x_i \sim f(x_i \mid \theta)$$

Sampling Probability Density

$$f(x \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Prior Probability Density

$$f(\theta)$$

Posterior Probability Density (Bayes' Rule)

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

Normalizing Constant)

$$f(x) = \int f(\theta) f(x \mid \theta) d\theta$$

The Bayesian Paradigm

Location Summaries

Posterior Mean

$$E(\theta \mid x) = \int \theta f(\theta \mid x) d\theta$$

Posterior Mode

$$\text{mode} = \arg\max_{\theta} f(\theta \mid x)$$

Scale Summary

Posterior Variance

$$Var(\theta \mid x) = \int (\theta - E(\theta \mid x))^2 f(\theta \mid x) d\theta$$

The Bayesian paradigm has the uncertainty defined by the posterior density. The challenges are how to evaluate the posterior density and what summaries to report, e.g. the posterior mean and the posterior mode.

The 95% credibility interval (confidence interval) from a Bayesian posterior says that the probability that the parameter is in the interval is 0.95. Remember In the Bayesian paradigm the parameter is a random variable.

The Bayesian Paradigm

Computing the Posterior Density

1. Analytic (Conjugacy)

2. Numerical

3. Gaussian Approximation

4. Simulations
   a. Importance Sampling Method

   b. Rejection Method

Conjugacy: Given $f(x \mid \theta) = L(\theta)$ then $f(\theta)$ is a conjugate prior for $f(x \mid \theta)$ if $f(\theta)$ is in the same function class as $L(\theta)$.

Joint Density of the Data/Likelihood Function $f(x \mid \theta) = L(\theta)$	Prior Density $f(\theta)$	Posterior Density $f(\theta \mid x)$
Gaussian estimating μ with σ^2 known. $L(\mu) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$	Gaussian μ $\mu \sim N(\mu_0, \sigma_0^2)$	Gaussian in μ
Gaussian estimating σ^2 with μ known. $L(\tau) \propto \tau^{\frac{n}{2}} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$	Gamma in $\tau = (\sigma^2)^{-1}$ $\tau \sim \Gamma(\alpha, \beta)$	Gamma in τ
Binomial for estimating p $L(p) \propto p^k (1-p)^{n-k}$	Beta in p $p \sim beta(\alpha, \beta)$	Beta in p
Poisson for estimating λ $L(\lambda) \propto \lambda^{S_n} \exp(-n\lambda)$	Gamma in λ $\lambda \sim \Gamma(\alpha, \beta)$	Gamma in λ
Gaussian estimating μ and σ^2. $L(\mu, \tau) \propto \tau^{\frac{n}{2}} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$	Gaussian μ $\mu \sim N(\mu_0, \sigma_0^2)$ Gamma in $\tau = (\sigma^2)^{-1}$ $\tau \sim \Gamma(\alpha, \beta)$	Gaussian μ Gamma in $\tau = (\sigma^2)^{-1}$ Approximate Joint Bivariate Gaussian in μ and τ

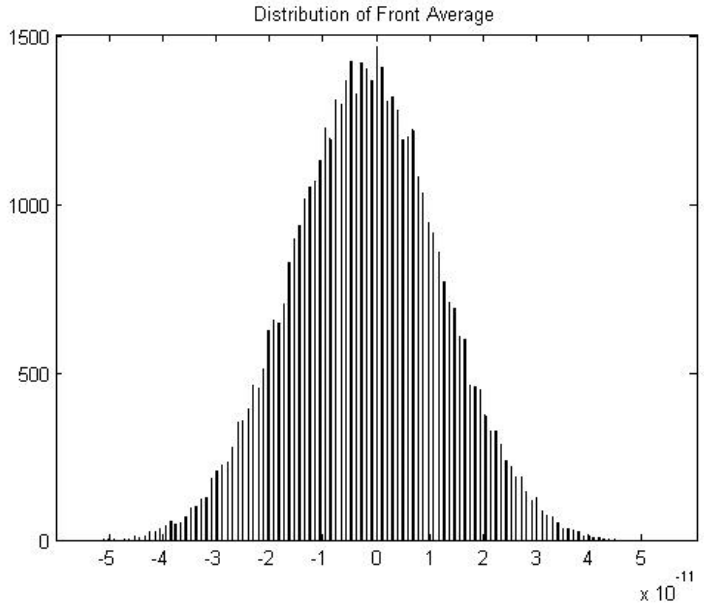
Comments on Conjugate Priors

1. **Conjugacy simplifies the mathematics for computing the posterior.**

2. **The challenge is choosing the prior distribution so that it accurately represents knowledge about the parameter before the experiment.**

Question 10.1 Conduct a Bayesian Analysis of the Variance of the MEG Noise Data.

Assume that $\mu = 0$ **is known.**



Question 10.1 Conduct a Bayesian Analysis of the Variance of the MEG Noise Data.

$$L(\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

$$L(\tau) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

$$= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left\{-\frac{\tau}{2}S_n\right\} \propto \Gamma(\tfrac{n}{2}+1, \tfrac{S_n}{2})$$

where $\tau = \frac{1}{\sigma^2}; S_n = \sum_{i=1}^{n}(x_i-\mu)^2,$ τ **is called the precision.**

Take $f(\tau) = \Gamma(\alpha, \beta)$

$$f(\tau \mid x) = \frac{f(\tau) f(x \mid \tau)}{f(x)} \propto f(\tau)L(\tau) = \Gamma(\alpha + \tfrac{n}{2} + 1, \beta + \tfrac{S_n}{2})$$

is a gamma distribution. Note that the mode of the posterior density or MAP estimate is (See Gaussian Approximation below)

$$\hat{\tau}_{MAP} = \frac{\alpha + \frac{n}{2} + 1 - 1}{\beta + \frac{S_n}{2}} = \frac{\alpha + \frac{n}{2}}{\beta + \frac{S_n}{2}} \approx \frac{n}{S_n} = \frac{1}{\hat{\sigma}^2_{ML}} = \hat{\tau}_{ML}$$

The approximation is true for $n \gg \alpha$ **and** $S_n \gg \beta.$

Gaussian Approximation

Maximum A-Posteriori Estimation

Let $f(x|\theta)$ be the sampling probability density or likelihood and $f(\theta)$ be the prior probability density. The posterior probability density is

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{f(x)} \propto f(\theta)f(x|\theta), \tag{10.31}$$

where $f(x) = \int f(\theta)f(x|\theta)d\theta$.

Recall that the maximum likelihood (ML) estimate of θ is defined as

$$\hat{\theta}_{\text{ML}} = \arg\max_{\theta} f(x|\theta). \tag{10.32}$$

Given $f(\theta)$ we define the **maximum-a-posteriori (MAP)** estimate as of θ as

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} f(\theta|x). \tag{10.33}$$

Simply stated, the MAP estimate is the mode of the posterior density.

Example 10.3 (continued). Gaussian Prior and Gaussian Sampling Density. Before deriving a general Gaussian approximation to a posterior density, let's gain some intuition by considering the Gaussian sampling probability density and prior probability density defined as

$$f(x \mid \theta) = \left(\frac{1}{2\pi\sigma_x^2} \right)^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} \frac{(x-\theta)^2}{\sigma_x^2} \right\}^{\frac{1}{2}} \tag{10.34}$$

$$f(\theta \mid \theta_0) = \left(\frac{1}{2\pi\sigma_\theta^2} \right)^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} \frac{(\theta-\theta_0)^2}{\sigma_\theta^2} \right\}. \tag{10.35}$$

To find the ML estimate of θ we consider the log likelihood

$$\log f(x \mid \theta) = -\frac{1}{2}\log(2\pi\sigma_x^2) - \frac{1}{2}\frac{(x-\theta)^2}{\sigma_x^2}. \tag{10.36}$$

Differentiating with respect to θ gives

$$\frac{\partial \log f(x \mid \theta)}{\partial \theta} = -\frac{(x-\theta)}{\sigma_x^2}, \tag{10.37}$$

and setting the derivative of the log likelihood equal to zero and solving for θ gives

$$\hat{\theta}_{ML} = x. \tag{10.38}$$

To compute the MAP estimate of θ we consider the log posterior probability density

$$\log f(\theta \mid x) \propto \log f(\theta) + \log f(x \mid \theta) \tag{10.39}$$

$$\propto -\frac{1}{2}\left[\frac{(\theta-\theta_0)^2}{\sigma_\theta^2} + \frac{(x-\theta)^2}{\sigma_x^2}\right]. \tag{10.40}$$

Differentiating with respect to θ gives

$$\frac{\partial \log f(\theta \mid x)}{\partial \theta} = -\frac{(\theta-\theta_0)}{\sigma_\theta^2} - \frac{(x-\theta)}{\sigma_x^2}, \tag{10.41}$$

and setting the derivative of log posterior to zero yields

$$\hat{\theta}_{MAP} = \left[\frac{1}{\sigma_\theta^2} + \frac{1}{\sigma_x^2}\right]^{-1}\left[\frac{\theta_0}{\sigma_\theta^2} + \frac{x}{\sigma_x^2}\right] = \frac{\sigma_x^2}{\sigma_\theta^2+\sigma_x^2}\theta_0 + \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_x^2}x \text{ (statistician's view)} \tag{10.42}$$

$$= \theta_0 + \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_x^2}(x-\theta_0) \text{ (control engineer's view)} \tag{10.43}$$

If $\sigma_x^2 \to \infty$, then $\hat{\theta}_{MAP} \to \theta_0$ and there is no information from the likelihood. If $\sigma_\theta^2 \to \infty$, then $\hat{\theta}_{MAP} \to x = \hat{\theta}_{ML}$ and there is no information from the prior. The last expression above is the essential result for deriving the well-known Kalman filter. When the sampling probability density is a point process observation model, then the corresponding form of Eq. 10.43 may be used to derive neural spike train decoding algorithms (Brown et al. 1998).

Gaussian Approximation to a Posterior Probability Density. Let us now analyze the posterior probability density the same way we analyzed the likelihood in **Lecture 9**. Given a posterior probability density $p(\theta \mid x)$ and $\hat{\theta}_{MAP}$, a MAP estimate of θ, expand the log posterior probability density in a Taylor series about θ_{MAP} to obtain

$$\log f(\theta \mid x) = \log f(\hat{\theta}_{MAP} \mid x) + \frac{\partial \log f(\hat{\theta}_{MAP} \mid x)}{\partial \theta}(\theta - \hat{\theta}_{MAP}) + \frac{1}{2}\frac{\partial \log f(\hat{\theta}_{MAP} \mid x)}{\partial \theta^2}(\theta - \hat{\theta}_{MAP})^2 + \ldots \quad (10.44)$$

$$\approx \log f(\hat{\theta}_{MAP} \mid x) + \frac{1}{2}\frac{\partial \log f(\hat{\theta}_{MAP} \mid x)}{\partial \theta^2}(\theta - \hat{\theta}_{MAP})^2, \quad (10.45)$$

because $\frac{\partial \log f(\hat{\theta}_{MAP} \mid x)}{\partial \theta} = 0$ by definition of the MAP estimate. Hence,

$$f(\theta \mid x) \approx f(\hat{\theta}_{MAP} \mid x) \exp\{\frac{1}{2}\frac{\partial^2 \log f(\hat{\theta}_{MAP} \mid x)}{\partial \theta^2}(\theta - \hat{\theta}_{MAP})^2\}. \quad (10.46)$$

gives a general Gaussian approximation to a posterior of probability density.

Gaussian Approximation to the Posterior Density

$$f(\theta \mid x) \doteq N(\hat{\theta}_{MAP}, -\left[\frac{\partial^2 \log f(\hat{\theta}_{MAP} \mid x)}{\partial \theta^2}\right]^{-1})$$

Exact Posterior Density for the Poisson-Gamma Model

We have independent observations

$$x = (x_1, \ldots, x_n) \qquad x_i \sim P(x_i \mid \lambda)$$

Sampling Probability Density

$$L(\lambda) = f(x \mid \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} \exp(-\lambda)}{x_i!} \propto \lambda^{S_n} \exp(-n\lambda)$$

Prior Probability Density

$$f(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda)$$

Posterior Probability Density (Bayes' Rule)

$$f(\lambda \mid x) \propto f(\lambda) f(x \mid \lambda) \propto \lambda^{\alpha-1} \exp(-\beta\lambda) \times \lambda^{S_n} \exp(-n\lambda)$$

$$\propto \lambda^{S_n+\alpha-1} \exp(-(\beta+n)\lambda)$$

$$= \Gamma(S_n + \alpha, \beta + n)$$

$$E(\lambda \mid x) = \frac{S_n + \alpha}{\beta + n} \quad Var(\lambda \mid x) = \frac{S_n + \alpha}{(\beta + n)^2}$$

Gaussian Approximation to the Posterior Density for the Poisson-Gamma Model

$$f(\lambda \mid x) \doteq f(\hat{\lambda}_{MAP} \mid x) \exp\{-\frac{1}{2} \frac{\partial^2 \log f(\hat{\lambda}_{MAP} \mid x)}{\partial \lambda^2} (\lambda - \hat{\lambda}_{MAP})^2\}$$

$$\hat{\lambda}_{MAP} = \frac{S_n + \alpha - 1}{\beta + n} \; ; \; \frac{\partial^2 \log f(\hat{\lambda}_{MAP} \mid x)}{\partial \lambda^2} = -\frac{(n + \beta + 1)^2}{(S_n + \alpha - 1)}$$

$$f(\hat{\lambda}_{MAP} \mid x) \approx \frac{n + \beta + 1}{2\pi (S_n + \alpha - 1)^{\frac{1}{2}}}$$

The approximate mean and variance are

$$E(\lambda \mid x) \approx \frac{S_n + \alpha - 1}{\beta + n + 1} \quad Var(\lambda \mid x) \approx \frac{S_n + \alpha - 1}{(\beta + n + 1)^2}$$

The exact mean and variance are

$$E(\lambda \mid x) = \frac{S_n + \alpha}{\beta + n} \quad Var(\lambda \mid x) = \frac{S_n + \alpha}{(\beta + n)^2}$$

Monte Carlo Methods: Importance Sampling

Importance Sampling

Posterior distributions are all of the form

$$f(\theta \mid x) \propto f(\theta) f(x \mid \theta), \tag{10.58}$$

It may be the case that the prior and sampling distribution are easy to specify but the resulting form of the posterior does not suggest any easy analytic approach to summarizing it. We can generate a summary by Monte Carlo. Suppose we wish to compute for some function $g(\theta)$ the quantity

$$E(g(\theta)) = \int g(\theta) f(\theta \mid x) d\theta, \tag{10.59}$$

We use importance sampling when

1. **the prior and sampling density are easy to evaluate but the posterior may not be.**
2. **there is an importance density that has the same support as the posterior density but is easy to draw samples from and ideally approximates the posterior well.**
3. **we are interested in computing a summary from the posterior such as the its mean and/or variance.**

$$g(\theta) = \theta \ the\, mean$$

$$g(\theta) = (\theta - E(\theta \mid x))^2 \ the\, \mathrm{var}\, iance$$

Monte Carlo Methods: Importance Sampling

Assume we wish to compute

$$E(g(\theta)) = \int g(\theta) f(\theta \mid x) d\theta. \tag{10.A1}$$

Given an importance density $h(\theta)$ that has the same support as $f(\theta \mid x)$, rewrite Eq. 10.A1 as

$$\begin{aligned} E(g(\theta)) &= \int g(\theta) \frac{f(\theta \mid x)}{h(\theta)} h(\theta) d\theta \\ &= f(x)^{-1} \int g(\theta) \frac{f(\theta) f(x \mid \theta)}{h(\theta)} h(\theta) d\theta \\ &= f(x)^{-1} \int g(\theta) w(\theta \mid x) h(\theta) d\theta \end{aligned} \tag{10.A2}$$

where $w(\theta \mid x) = h(\theta)^{-1} f(\theta) f(x \mid \theta)$. Similarly, we can write

$$\begin{aligned} 1 = \int f(\theta \mid x) d\theta &= f(x)^{-1} \int \frac{f(\theta) f(x \mid \theta)}{h(\theta)} h(\theta) d\theta \\ &= f(x)^{-1} \int w(\theta \mid x) h(\theta) d\theta. \end{aligned} \tag{10.A3}$$

We can use Eqs. 10.A2 and 10.A3 to write Eq. 10.A1 as

$$\begin{aligned} E(g(\theta)) = \frac{E(g(\theta))}{1} &= \frac{f(x)^{-1} \int g(\theta) w(\theta \mid x) h(\theta) d\theta}{f(x)^{-1} \int w(\theta \mid x) h(\theta) d\theta} \\ &= \frac{\int g(\theta) w(\theta \mid x) h(\theta) d\theta}{\int w(\theta \mid x) h(\theta) d\theta}. \end{aligned} \tag{10.A4}$$

Monte Carlo Methods: Importance Sampling

We can approximate the numerator and denominator in Eq. 10.A4 respectively by n draws from $h(\theta)$ as

$$\int g(\theta)w(\theta \mid x)h(\theta)d\theta \doteq \frac{\sum_{i=1}^{n} g(\theta_i)w(\theta_i \mid x)}{n} \tag{10.A5}$$

$$\int w(\theta \mid x)h(\theta)d\theta \doteq \frac{\sum_{i=1}^{n} w(\theta_i \mid x)}{n}.$$

Therefore, we have the importance sampling approximation of $E(g(\theta))$ as

$$E(g(\theta)) \doteq \frac{\sum_{i=1}^{n} g(\theta_i)w(\theta_i \mid x)}{\sum_{i=1}^{n} w(\theta_i \mid x)}. \tag{10.A6}$$

We can simulate Eq. 10.59 (10.A1) with the following algorithm

Algorithm 10.2 (Importance Sampling)

$\text{Sum} = 0$

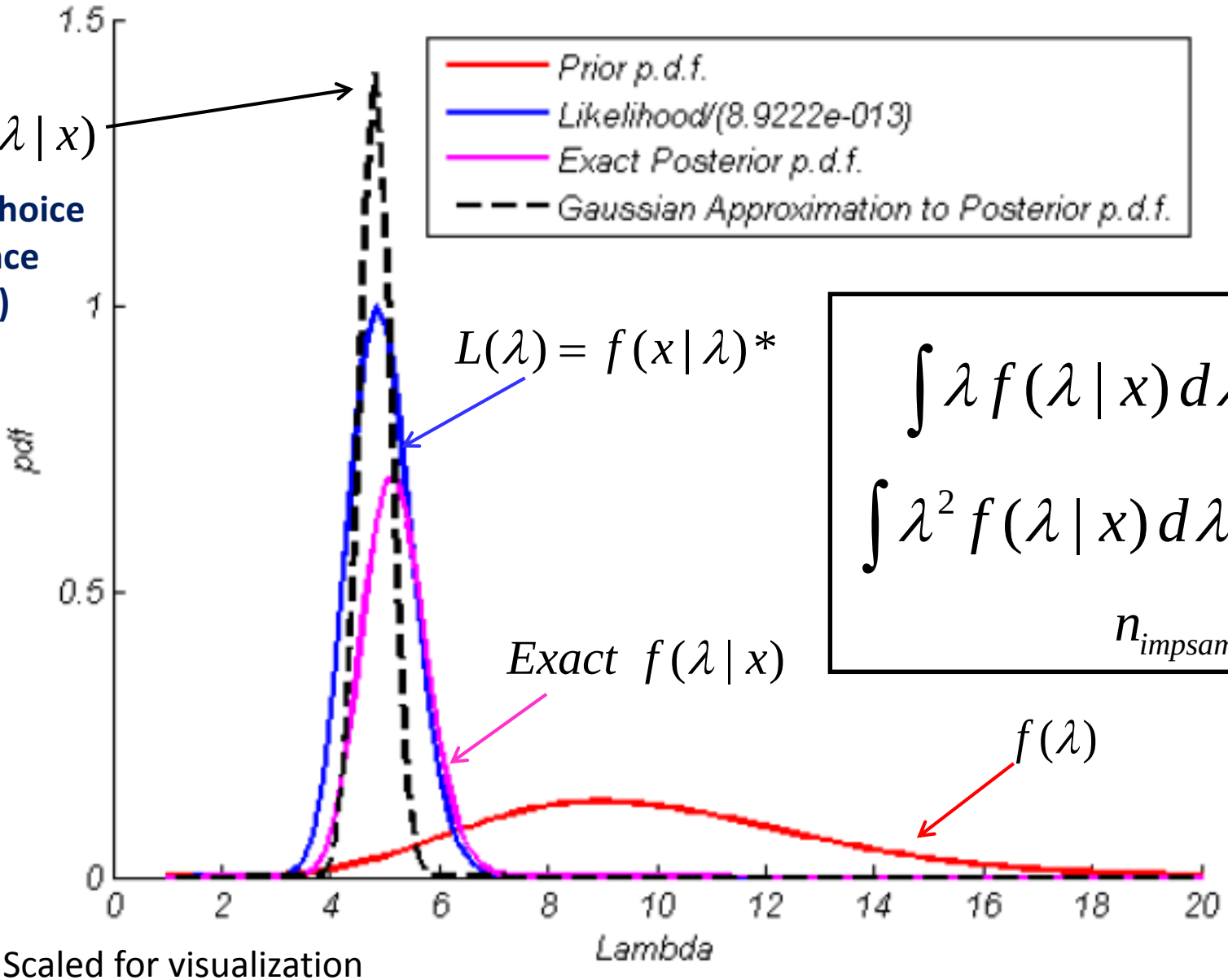
$W = 0$

$\text{For } j = 1,\ldots,10,000$

1. Draw θ_j from $h(\theta)$
2. Compute $w(\theta_j) \mid x) = \frac{f(\theta_j)f(x \mid \theta_j)}{h(\theta_j)}$ and $g(\theta_j)$
3. $\text{Sum} \leftarrow \text{Sum} + w(\theta_j \mid x)g(\theta_j)$
4. $W \leftarrow W + w(\theta_j \mid x)$

Compute $E[g(\theta)] \doteq W^{-1}\text{Sum}$

Example of Importance Sampling Method



Approx. $f(\lambda \mid x)$

(Candidate choice for Importance Distribution)

$L(\lambda) = f(x \mid \lambda)^*$

Exact $f(\lambda \mid x)$

$f(\lambda)$

$$\int \lambda f(\lambda \mid x)\, d\lambda \approx 4.49$$

$$\int \lambda^2 f(\lambda \mid x)\, d\lambda \approx 22.07$$

$$n_{impsamp} = 1000$$

*Scaled for visualization

Monte Carlo Methods: Rejection Method

Rejection Method

Suppose that $f(\theta \mid x)$ can be approximated by $h(\theta)$, a probability density which has the same support or whose support contains the support of $f(\theta \mid x)$. Suppose there is a constant c such that

$$\frac{f(\theta \mid x)}{h(\theta)} \leq c \tag{10.61}$$

for all θ. We then have the following algorithm for simulating from $f(\theta \mid x)$. It is called the rejection method.

Algorithm 10.3 (Rejection Method).

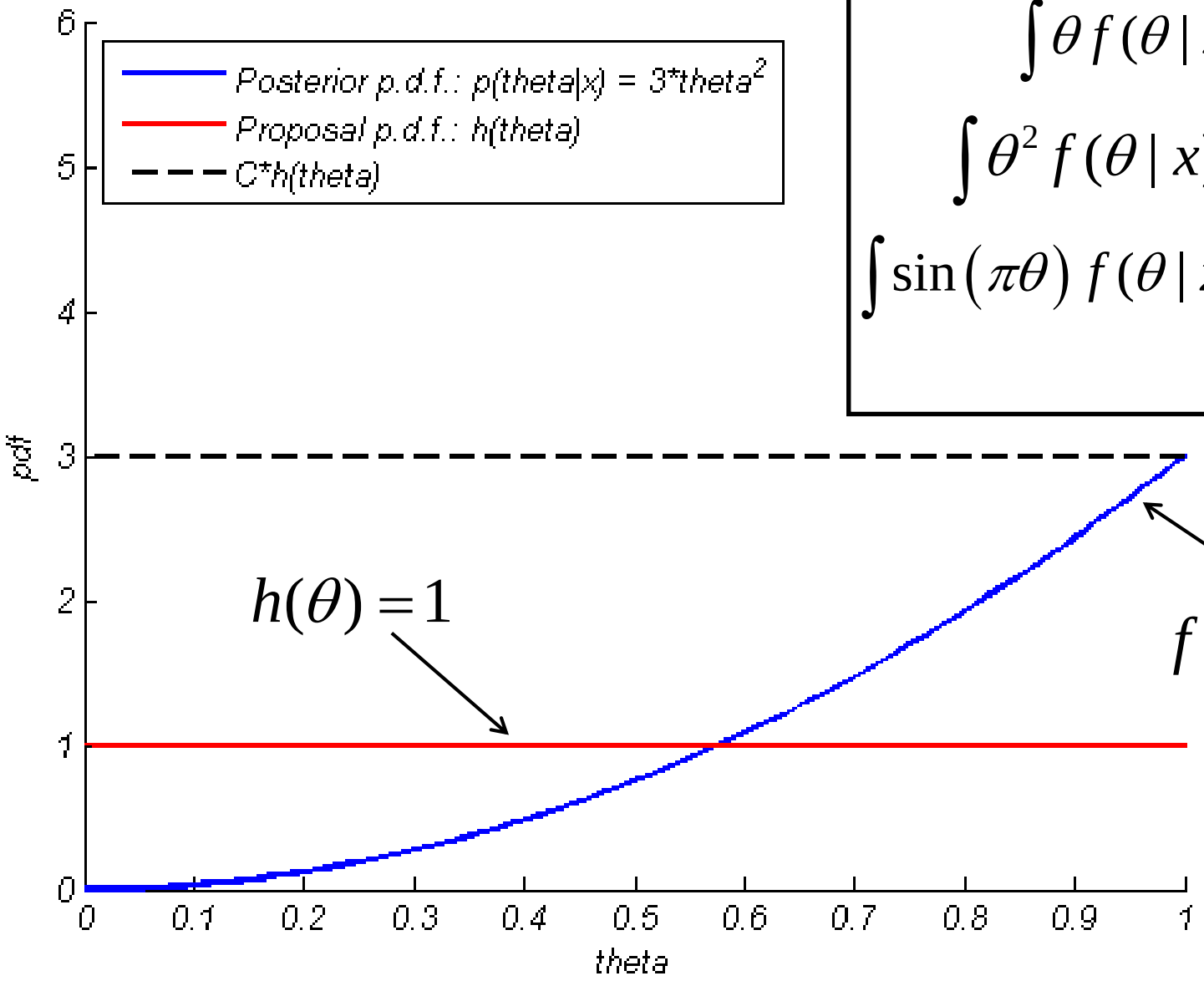
Pick n large $j = 1; i = 1$.

1. Draw θ_i^* from $h(\theta)$
2. Draw U_i from $U(0,1)$
3. If $U_i \leq \frac{f(\theta_i^* \mid x)}{ch(\theta_i^*)}$ take $\theta_j = \theta_i^*$
4. If $j = n$ stop else $j \leftarrow j+1, i \leftarrow i+1$ Return to Step 1

This algorithm was originally due to John Von Neumann. We accept the value θ_i^* with probability $f(\theta_i^* \mid x)/ch(\theta_i^*)$ by generating a uniform $(0,1)$ random variable U and accepting if $U_i \leq f(\theta_i^* \mid x)/ch(\theta_i^*)$.

When $j = n$; the quantity n/i measures the efficiency of the algorithm.

Example of Rejection Methods



$$\int \theta f(\theta \mid x)\, d\lambda \approx 0.7504$$

$$\int \theta^2 f(\theta \mid x)\, d\lambda \approx 0.5998$$

$$\int \sin(\pi\theta) f(\theta \mid x)\, d\lambda \approx 0.5704$$

$$n_{rejsamp} = 1000$$

III. Summary

Bayesian methods provide a powerful framework for statistical analysis that enjoys the optimality properties of the likelihood approach. They also provide a principled approach for addressing the important problem of combining information from different sources.

Complex Bayesian calculations in higher dimensions are computed by simply combining the analytic, approximate and Monte Carlo techniques we have learned here.