Bayesian statistics

Laplace's approximation and importance sampling

Leonardo Egidi

2024/2025

Università di Trieste

Table of contents i

Motivations

Gaussian approximation

Laplace's approximation

Numerical integration

Accept-reject method

Monte-Carlo integration

- Classical MC
- Importance sampling

Motivations

Motivations i

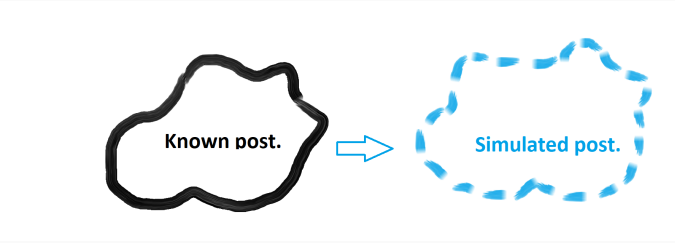
- The entire goal of Bayesian analysis is to compute and extract summaries from the posterior distribution for the parameter θ:

$$\pi(\theta|y) = \frac{\pi(\theta)p(y|\theta)}{\int_{\Theta} \pi(\theta)p(y|\theta)}. \tag{1}$$

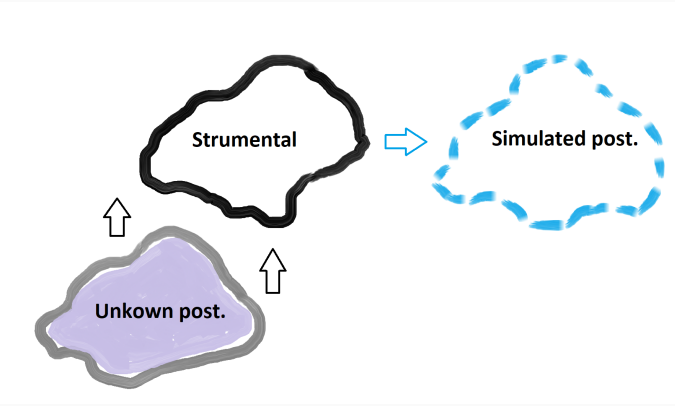
- This is easy for conjugate models: normal likelihood + normal prior, beta + binomial, Poisson + gamma, multinomial + Dirichlet.
- However, in real applications and complex models there is not usually a closed and analytical form for the posterior. The problem is represented by the denominator of (1), the marginal likelihood, usually denoted by $m(y)$.

Motivations ii

- The Bayesian idea is to use *simulation* to generate values from the posterior distribution:
  - *directly* when the posterior is entirely/partially known



  - via some suitable *instrumental* distributions when the posterior is unknown/not analytically available.



- In what follows, we will refer to the evaluation of the general integral:

$$\mathrm{E}_f[h(X)] = \int_{\mathcal{X}} h(x) f(x) dx, \tag{2}$$

Motivations iii

where $f(\cdot)$ is referred as the target distribution, generally untractable/partially tractable. Possible solutions:

- Asymptotic approximations (Laplace, and others).
- Numerical integrations.
- Accept-reject methods.
- Monte Carlo methods: i.i.d. draws from the posterior (or similar) distributions.
- Markov Chain Monte Carlo (MCMC) methods: dependent draws from a Markov chain whose limiting distribution is the posterior distribution (Metropolis-Hastings, Gibbs sampling, Hamiltonian Monte Carlo).
- Variational inference methods: find the approximating posterior within a family, in such a way to minimize the Kullback-Leibler divergence from the true posterior.

Gaussian approximation

Gaussian approximation for the posterior i

- One can show that for many Bayesian models, the posterior distribution approaches a Gaussian form as the number of data points increases.
- We might therefore approximate the posterior by a Gaussian distribution with the same mode, and the same curvature of the density at the mode.
- Suppose that we wish to approximate $\pi(\theta|y) \propto \exp\{h(\theta)\}$, where $h(\theta) = \log \pi(\theta) + \log p(y|\theta)$.
- We find the location of the mode, θ^*, and the Hessian matrix of second derivatives at the mode, $h''(\theta^*)$.
- The approximating Gaussian has mean θ^* and covariance matrix $[-h''(\theta^*)]^{-1}$.

Gaussian approximation for the posterior ii

- To see this, a Taylor series expansion of $\log \pi(\theta|y)$ centered at the posterior mode, θ^*, gives:

$$\log \pi(\theta|y) = \log \pi(\theta^*|y) + \frac{1}{2}(\theta - \theta^*)^T \left[\frac{d^2}{d\theta^2} \log \pi(\theta|y) \right]_{\theta=\theta^*} (\theta - \theta^*) + \ldots, \tag{3}$$

  where the linear term in the expansion is zero because the log-posterior density has zero derivative at its mode.

- Considering (3) as a function of θ, the first term is a constant, whereas the second term is proportional to the logarithm of a normal density, yielding the approximation (*for more technical details see Chapter 4 of BDA book.*)

$$\pi(\theta|y) \approx \mathcal{N}(\theta^*, [-h''(\theta^*)]^{-1}). \tag{4}$$

Gaussian approximation for the posterior iii

- We don't need to know the normalizing constant for the posterior, and (as we'll see) we can get an approximation to this normalizing constant (the marginal likelihood) from the Gaussian approximation.
- We need to be able to find the mode. This is usually easier than any Monte Carlo scheme, assuming that the distribution has a single mode (or at least a dominant mode).
- We need to compute second derivatives of the log posterior density at the mode. This is hard for some models, and becomes unattractive for high dimensional problems.
- Once we have the Gaussian approximation, we can easily find simple moments and quantiles, and we can find the expectation of more complex functions by simple Monte Carlo.
- The adequacy of the approximation may depend on the choice of parameterization, e.g., whether to use $\theta \in (0, \infty)$ or instead reparameterize in terms of $\phi = \log \theta$.

Laplace's approximation

Laplace's method i

- Suppose to deal with a *twice-differentiable* function $g(x)$, such that $h(x) = \log g(x)$, then equivalently $g(x) = e^{h(x)}$. Through the Taylor's theorem, we can expand h around a global maximum for h, $\tilde{x}$,

$$h(x) = h(\tilde{x}) + (x - \tilde{x})h'(\tilde{x}) + \frac{1}{2}(x - \tilde{x})^2 h''(\tilde{x}) + O((x - \tilde{x})^3),$$

such that

$$h(x) \approx h(\tilde{x}) + \frac{1}{2}(x - \tilde{x})^2 h''(\tilde{x}), \tag{5}$$

where $h'(\tilde{x}) = 0$.

Laplace's method ii

- We evaluate then the following integral using (5):

$$
\begin{aligned}
\int_a^b g(t)dt &= \int_a^b e^{h(t)}dt \\
&\approx \int_a^b e^{h(\tilde{x})+\frac{1}{2}(t-\tilde{x})^2 h''(\tilde{x})}dt \\
&= e^{h(\tilde{x})}\int_a^b \underbrace{e^{\frac{1}{2}(t-\tilde{x})^2 h''(\tilde{x})}}_{\propto\ \mathcal{N}(\tilde{x},-h''(\tilde{x})^{-1})} dt \qquad (6) \\
&= e^{h(\tilde{x})} 2\pi^{1/2}[-h''(\tilde{x})]^{-1/2} \\
&= e^{h(\tilde{x})}\sqrt{\frac{2\pi}{-h''(\tilde{x})}},
\end{aligned}
$$

where $\sigma^2 = [-h''(\tilde{x})]^{-1}$ is the variance of the Gaussian density in the integral.

Laplace's method iii

- Suppose now to deal with a definite integral over $\mathbb{R}^p$. In this case:

$$\int_a^b g(t)dt_1 \dots dt_p \approx \; e^{h(\tilde{x})}(2\pi)^{p/2}|-h''(\tilde{x})|^{-1/2}, \tag{7}$$

where $\Sigma = (D^2[-h(\tilde{x})])^{-1}$, and $|-h''(\tilde{x})|$ is the determinant of the (minus) Hessian matrix with (i,j)-th element equal to $-\frac{\partial^2 h}{\partial x_i \partial x_j}\Big|_{x=\tilde{x}}$.

Laplace's approximation for Bayesian inference i

- Consider now the marginal likelihood,

$$\begin{aligned} m(y) &= \int p(y|\theta)\pi(\theta)d\theta \\ &= \int e^{-n\left(-\frac{1}{n}\log p(y|\theta)-\frac{1}{n}\log \pi(\theta)\right)}d\theta. \end{aligned} \tag{8}$$

- Assume $h(\theta) = -\frac{1}{n}\log p(y|\theta) - \frac{1}{n}\log \pi(\theta)$. We could now use a *maximum a posteriori* $\hat{\theta}$ and the result in (7) to approximate the marginal likelihood:

$$\begin{aligned} m(y) &= \int e^{-nh(\theta)}d\theta \\ &\approx e^{-nh(\hat{\theta})}(2\pi)^{p/2}|\Sigma|^{1/2}n^{-p/2} \\ &= n^{-p/2}p(y|\hat{\theta})\pi(\hat{\theta})(2\pi)^{p/2}|\Sigma|^{1/2}, \end{aligned} \tag{9}$$

Laplace's approximation for Bayesian inference ii

where $\Sigma = (D^2(h(\hat{\theta}))^{-1}$ is the inverse of the Hessian of h evaluated at $\hat{\theta}$. This expansion is accurate to order $O(1/n)$, since we only consider the first order terms of the Laplace approximation.

- This approximation is used for example in model selection, where computing the marginal likelihood analytically can be hard unless there is conjugacy.
- Computing the Laplace approximation requires finding the maximum a posteriori probability $\hat{\theta} = \underset{\theta \in \Theta}{\text{argmax}}\{-h(\theta)\}$, which can be done using a standard method such as gradient search. It also requires computing the second derivative matrix and inverting it to obtain Σ. This is usually the harder quantity to calculate.

Computing the posterior mean with Laplace approximation

- We compute the posterior mean using the Laplace's approximation:

$$
\begin{aligned}
\mathrm{E}[\theta|y] = & \int \theta\pi(\theta|y)d\theta = \frac{\int \theta p(y|\theta)\pi(\theta)d\theta}{\int p(y|\theta)\pi(\theta)d\theta} \\
\approx & \frac{\int \theta e^{h(\hat{\theta})+\frac{1}{2}(\theta-\hat{\theta})h''(\hat{\theta})}d\theta}{\int e^{h(\hat{\theta})+\frac{1}{2}(\theta-\hat{\theta})h''(\hat{\theta})}d\theta} \\
= & \frac{\int \theta\sqrt{\frac{2\pi}{-h''(\hat{\theta})}}p(\theta|\hat{\theta},[-h''(\hat{\theta})]^{-1})d\theta}{\int \sqrt{\frac{2\pi}{-h''(\hat{\theta})}}p(\theta|\hat{\theta},[-h''(\hat{\theta})]^{-1})d\theta} \\
= & \int \theta p(\theta|\hat{\theta},[-h''(\hat{\theta})]^{-1}) = \hat{\theta},
\end{aligned}
\tag{10}
$$

where $p(\cdot|\mu,\sigma^2)$ denotes the density of a Gaussian distribution with mean μ and variance σ^2.

Laplace and BIC

- If we take the logarithm of Equation (9) and replace the maximum a posteriori with the maximum likelihood estimate, then we obtain:

$$\log m(y) \approx \log p(y|\hat{\theta}) + \log \pi(\hat{\theta}) - \frac{p}{2} \log n + \frac{p}{2} \log 2\pi + \frac{1}{2} \log |\Sigma|. \quad (11)$$

- The BIC score only retains the terms that vary in n, since asymptotically the terms that are constant in n do not matter. Dropping the constant terms we get

$$\log m(y) \approx \log p(y|\hat{\theta}) - \frac{p}{2} \log n, \quad (12)$$

  and the BIC is obtained by multiplying Equation (7) by -2.
- In the model selection problem, we pick the model with the lowest BIC score. Frequentist analysis shows that the BIC score is an asymptotically consistent model selection procedure under weak conditions.
- The BIC score is part of a family of competing penalized likelihood scores that also include the AIC (which is not consistent asymptotically) and the DIC.

Numerical integration

Numerical integration

- Numerical integration methods often fails to spot the region of importance for the function to be integrated.
- For example, consider a sample of ten Cauchy rv's y_i $(1 \leq y_i \leq 10)$ with location parameter $\theta = 350$. The marginal distribution of the sample under a flat prior is:

$$m(y) = \int_{-\infty}^{+\infty} \prod_{i=1}^{10} \frac{1}{\pi} \frac{1}{1 + (y_i - \theta)^2} d\theta$$

- The R function `integrate` does not work well! In fact, it returns a wrong numerical output (see next slide) and fails to signal the difficulty since the error evaluation is absurdly small. Function `area` may work better.
- Let's evaluate now the

Numerical integration: Cauchy example

```
set.seed(12345)
rc = rcauchy(10) + 350
 lik = function(the) {
 u = dcauchy(rc[1] - the)
 for (i in 2:10) u = u * dcauchy(rc[i] - the)
 return(u)}
 integrate(lik, -Inf, Inf)

[1] 3.728903e-44 with absolute error < 7.4e-44
 integrate(lik, 200, 400)

[1] 1.79671e-11 with absolute error < 3.3e-11
```

We need to know the range where the likelihood is not negligible. Moreover, numerical integration cannot easily face multidimensional integrals.

Accept-reject method

Accept-reject method i

- Suppose we need to evaluate the following integral, but we cannot directly sample from the target density:

$$\mathrm{E}_f[h(\theta)] = \int_\Theta h(\theta) f(\theta) d\theta, \tag{13}$$

where $h(\cdot)$ is a parameter function and $f(\cdot)$ is the target distribution (in Bayesian inference, this is usually the posterior).

- Assume that
  1. $f(\theta)$ is continuous and such that $f(\theta) = d(\theta)/K$, and we know how to evaluate $d(\theta) \Rightarrow$ we know the functional form of f up to a *multiplicative constant*.
  2. There exists another density $g(\theta)$, an instrumental density, such that, for some big c, $d(\theta) \leq c \times g(\theta), \forall \theta$.

Accept-reject method ii

- It is possible to show that the following algorithm will generate values from the target density $f(\theta)$:

A-R algorithm

1. draw a candidate $W = w \sim g(w)$ and a value $Y = y \sim \text{Unif}(0, 1)$.
2. If

$$y \leq \frac{d(w)}{c \times g(w)},$$

set $\theta = w$, otherwise reject the candidate w and go back to step 1.

Theorem

(a) *The distribution of the accepted values is exactly the target density $f(\theta)$.*

(b) *The marginal probability that a single candidate is accepted is K/c.*

Accept-reject method iii

Proof.

(a) The cdf of $W|[Y \leq \frac{d(w)}{c \times g(w)}]$ can be written as:

$$F_W(\theta) = \frac{\Pr(W \leq \theta, Y \leq \frac{d(w)}{c \times g(w)})}{\Pr(Y \leq \frac{d(w)}{c \times g(w)})} = \frac{\int_W \Pr(W \leq \theta, Y \leq \frac{d(w)}{c \times g(w)} | w) g(w) dw}{\int_W \Pr(Y \leq \frac{d(w)}{c \times g(w)} | w) g(w) dw} =$$

$$= \frac{\int_{-\infty}^{\theta} \Pr(Y \leq \frac{d(w)}{c \times g(w)} | w) g(w) dw}{\int_{-\infty}^{+\infty} \Pr(Y \leq \frac{d(w)}{c \times g(w)} | w) g(w) dw} = \frac{\int_{-\infty}^{\theta} \frac{d(w)}{c} dw}{\int_{-\infty}^{+\infty} \frac{d(w)}{c} dw} =$$

$$= \frac{\int_{-\infty}^{\theta} \frac{Kf(w)}{c} dw}{\int_{-\infty}^{+\infty} \frac{Kf(w)}{c} dw} = \int_{-\infty}^{\theta} f(w) dw. \square$$

(b) The probability that a single candidate $W = w$ will be accepted is

$$\Pr(W \text{ accepted}) = \Pr(Y \leq \frac{d(W)}{c \times g(W)}) =$$

$$= \int_W \Pr(Y \leq \frac{d(W)}{c \times g(W)} | W = w) g(w) dw =$$

$$= \int_W \frac{d(w)}{c} dw = \int_W \frac{K}{c} f(w) dw = \frac{K}{c}. \square$$

$\square$

A-R algorithm: simulation from a Beta distribution

- Suppose we need to draw values from a Beta(a, b), our f, but we only have a random number generator for the interval (0,1), a Unif$(0, 1)$, our instrumental distribution g. Both the distributions have support (0,1), then we have:

$$f(\theta) = \frac{d(\theta)}{K} = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\text{B(a,b)}},$$

  where B(a,b) is the Beta function with arguments a and b and $K = 1$.

- The AR steps are:
  - draw $\theta^* \sim g = \text{Unif}(0, 1)$, $U \sim \text{Unif}(0, 1)$.
  - we accept $\theta = \theta^*$ iff $U \leq \frac{d(\theta^*)}{c \times g(\theta^*)}$.
  - otherwise, go back to step 1

A-R algorithm: simulation from a Beta distribution

```
Nsims=2500
#parameters
a=2.7; b=6.3
#find optimal c
c=optimise(f=function(x) {dbeta(x,a,b)},
           interval=c(0,1), maximum=TRUE)$objective
u=runif(Nsims, max=c)
theta_star=runif(Nsims)
theta=theta_star[u<dbeta(theta_star,a,b)]
# accept prob
1/c

[1] 0.3745677
```

A-R algorithm: simulation from a Beta distribution

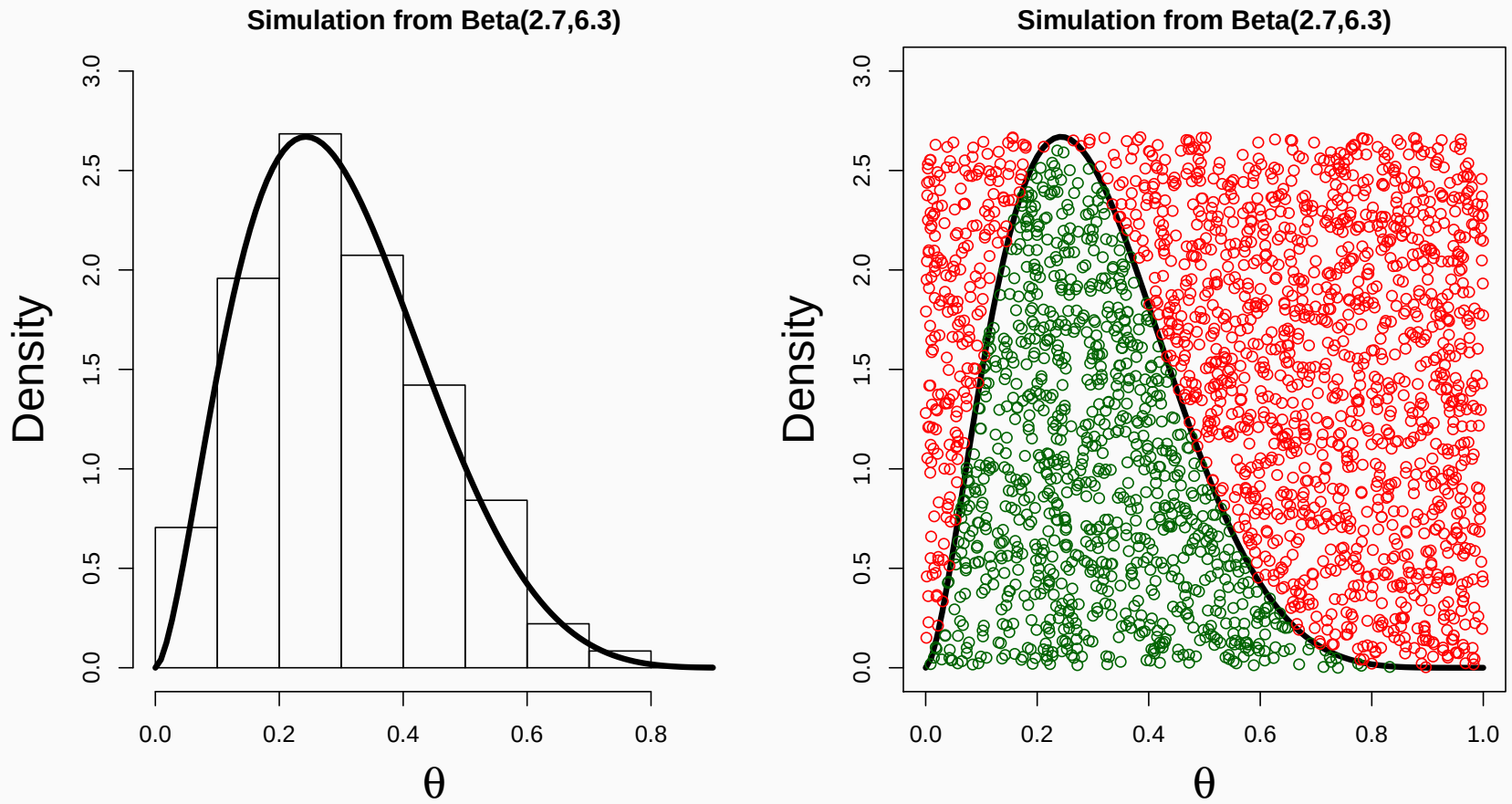


Figure 1: On the left plot, the true Beta$(2.7, 6.3)$, and the histogram of the simulated distribution. On the right plot, the pairs (θ^*, U): the accepted (green) and the discarded (red). $K = 1$.

A-R algorithm: simulation from a Beta distribution

```
Nsims=2500
#beta parameters
a=2; b=3
#find optimal c
c=optimise(f=function(x) {dbeta(x,a,b)},
           interval=c(0,1), maximum=TRUE)$objective
u=runif(Nsims, max=c)
theta_star=runif(Nsims)
theta=theta_star[u<dbeta(theta_star,a,b)]
#accept prob
1/c
[1] 0.5625
```

A-R algorithm: simulation from a Beta distribution

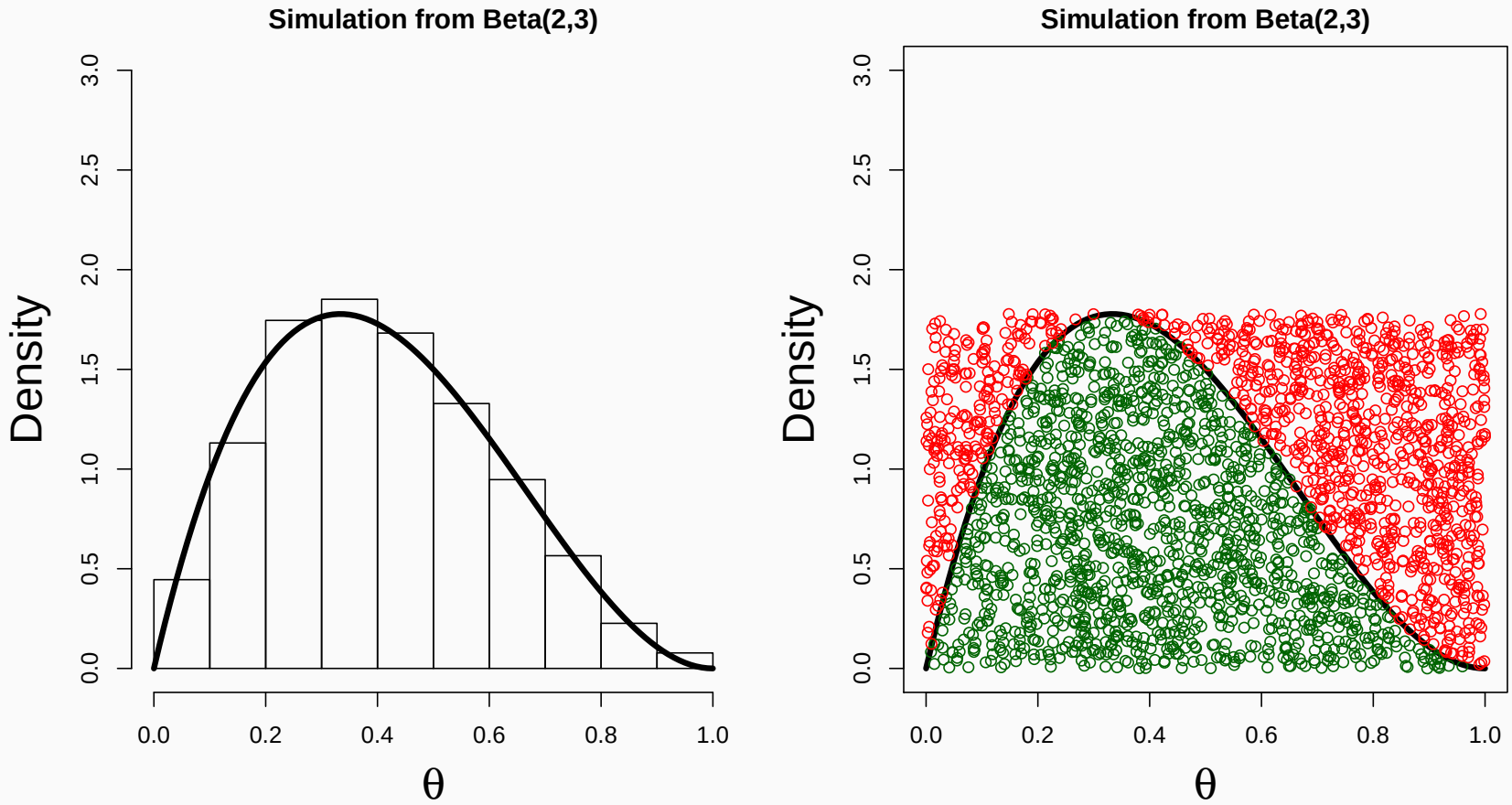


Figure 2: On the left plot, the true Beta$(2, 3)$, and the histogram of the simulated distribution. On the right plot, the pairs (θ^*, U): the accepted (green) and the discarded (red). $K = 1$.

A-R algorithm: simulation from a Beta distribution

Comments:

- The probability of accepting the candidate θ^* is higher in the second case, since a Beta$(2, 3)$ is more similar to a Unif$(0, 1)$ than a Beta$(2.7, 6.3)$.
- c must be chosen in such a way that the condition $d(\theta) \leq c \times g(\theta)$ is verified for all θ.
- K has been fixed to 1, since all the distribution π to be sampled from is completely known.
- In general, g needs to have thicker tail than d for d/g to remain bounded for all θ. For instance, normal g cannot be used to sample from a Cauchy d. You can do the opposite of course.
- One criticism of the A-R method is that it generates *useless* simulations from the proposal g when rejecting, even those necessary to validate the output as being generated from the target f.

Monte-Carlo integration

Classical Monte Carlo integration i

- Two major classes of numerical problems that arise in statistical inference are *optimization* problems and *integration* problems.
- Suppose we need to calculate:

$$\mathrm{E}_f[h(X)] = \int_{\mathcal{X}} h(x) f(x) dx, \tag{14}$$

where $f(\cdot)$ is a probability density and $h(\cdot)$ is a function of x. When an analytical solution is not possible, how do we approximate this integral?

- If $|I| < \infty$ and $X_1, X_2, \ldots, X_S$ are i.i.d $\sim f$, then the Strong Law of Large Numbers implies that the empirical mean is consistent for $\mathrm{E}_f[h(X)]$

$$\widehat{\mathrm{E}_f[h(X)]} = \frac{1}{S} \sum_{s=1}^{S} h(X_s) \to \mathrm{E}_f[h(X)] \text{ in probability, as } S \to \infty \tag{15}$$

Classical Monte Carlo integration ii

- The variance of $\widehat{\mathrm{E}_f[h(X)]}$ is

$$\mathrm{Var}(\widehat{\mathrm{E}_f[h(X)]}) = \frac{1}{S}\int_{\mathcal{X}}[h(x) - \mathrm{E}_f[h(x)]]^2 f(x)dx$$

  and it can be approximated by

$$\hat{V} = \frac{1}{S}\sum_{s=1}^{S}[h(x_s) - \widehat{\mathrm{E}_f[h(X)]}]^2.$$

- When S is large (approximately) for the Central Limit Theorem we have that:

$$\frac{\widehat{\mathrm{E}_f[h(X)]} - \mathrm{E}_f[h(X)]}{\sqrt{\hat{V}}} \sim \mathcal{N}(0,1).$$

Example: Normal mean with Cauchy prior

- Consider:

$$y|\theta \sim \mathcal{N}(\theta, 1), \quad \theta \sim \text{Cauchy}(0, 1).$$

- The posterior mean for a single observation y is:

$$\mathrm{E}(\theta|y) = \frac{\int_{-\infty}^{+\infty} \frac{\theta}{1+\theta^2} e^{-(y-\theta)^2/2} d\theta}{\int_{-\infty}^{+\infty} \frac{1}{1+\theta^2} e^{-(y-\theta)^2/2} d\theta}.$$

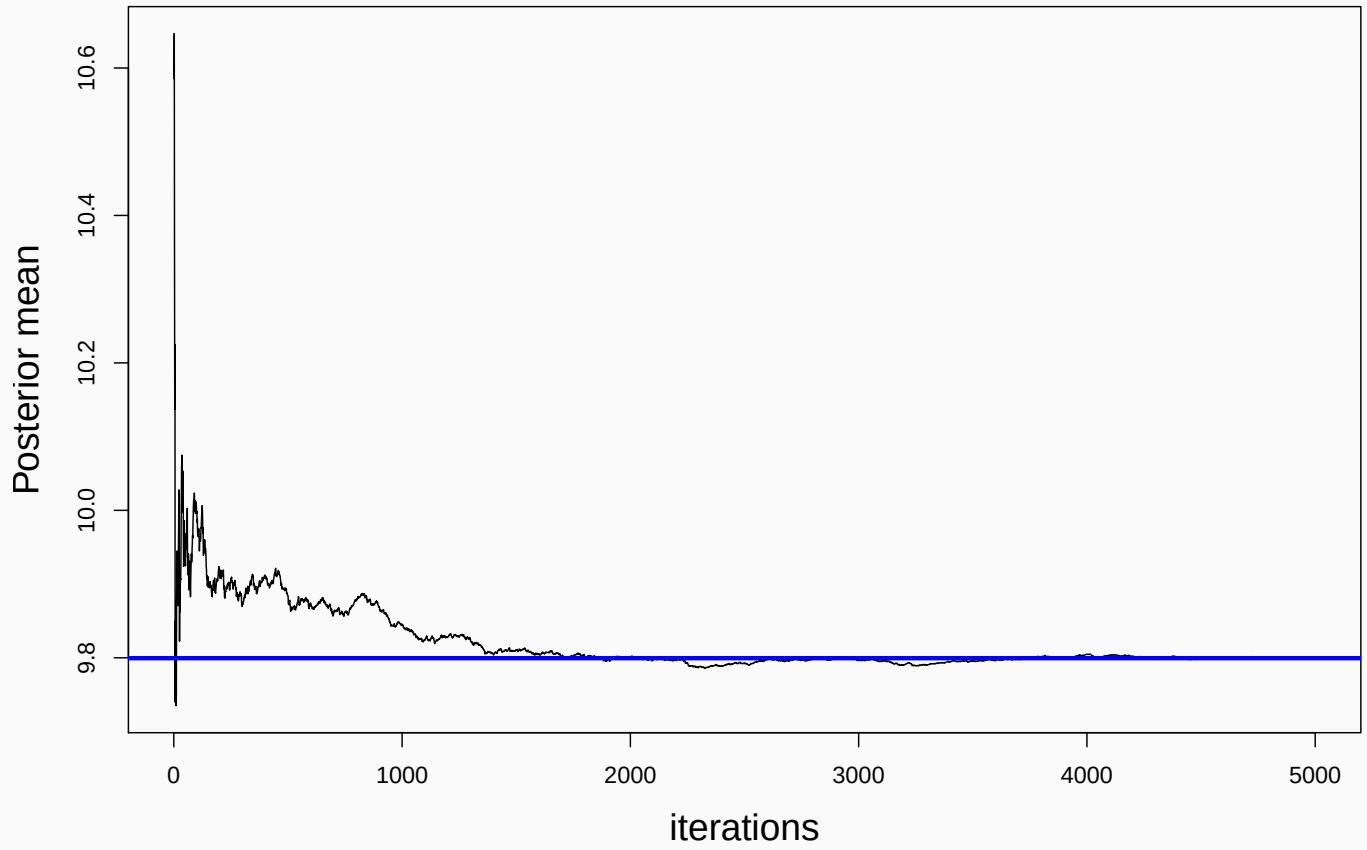
- We could draw $\theta_1, \ldots, \theta_S$ from $\mathcal{N}(y, 1)$ and compute:

$$\hat{\mathrm{E}}(\theta|y) = \frac{\sum_{s=1}^{S} \frac{\theta_s}{1+\theta_s^2}}{\sum_{s=1}^{S} \frac{1}{1+\theta_s^2}}$$

- The effect of the prior is to pull a little bit the estimate of θ toward 0.

Example: Normal mean with Cauchy prior

```
set.seed(12345)
theta = rnorm(5000, 10, 1)
I = sum(theta/(1 + theta^2))/sum(1/(1 + theta^2))
I
[1] 9.793254
```



Importance sampling i

- Importance sampling is based on the following representation:

$$\begin{aligned} \mathrm{E}_f[h(X)] &= \int_{\mathcal{X}} h(x) f(x) dx = \\ &= \int_{\mathcal{X}} h(x) \frac{f(x)}{g(x)} g(x) dx = \mathrm{E}_g \left[h(X) \frac{f(X)}{g(X)}, \right] \end{aligned} \tag{16}$$

where g is an arbitrary density function, called instrumental distribution, whose support is greater than $\mathcal{X}$.

Importance sampling ii

- Given a sequence $X_1, \ldots, X_S$ i.i.d. from g we can estimate the integral above by

$$\mathrm{E}_f^{is}[h(X)] = \frac{1}{S}\sum_{s=1}^{S} h(x_s)\frac{f(x_s)}{g(x_s)} = \frac{1}{S}\sum_{s=1}^{S} h(x_s)w(x_s), \tag{17}$$

  where $w(x) = f(x)/g(x)$ is called importance function.

- Note that classical Monte Carlo and importance sampling both produce unbiased estimator for the integral (14), but:

$$\mathrm{Var}(\widehat{\mathrm{E}_f[h(X)]}) = \frac{1}{S}\int_{\mathcal{X}} [h(x) - \mathrm{E}_f[h(x)]]^2 f(x)dx$$

$$\mathrm{Var}(\mathrm{E}_f^{is}[h(X)]) = \frac{1}{S}\int_{\mathcal{X}} [h(x)\frac{f(x)}{g(x)} - \mathrm{E}_f[h(x)]]^2 g(x)dx$$

Importance sampling iii

- We can work on g in order to minimize the variance of (17). The constraint that $\text{supp}(h \times f) \subset \text{supp}(g)$ is absolute in that using a smaller support truncates the integral (14) and thus produces a biased result.
- It puts very little restriction on the choice of the instrumental distribution g, which can be chosen from distributions that are either easy to simulate or efficient in the approximation of the integral.
- IS variance is finite only when

$$\mathrm{E}\left[h(X)^2 \frac{f(X)^2}{g(X)^2}\right] = \int_{\mathcal{X}} h(x)^2 \frac{f(x)^2}{g(x)^2} dx < \infty$$

- Densities g with lighter tails than f, ($\sup f/g = \infty$) are not good proposals because they can lead to infinite variance.
- When $\sup f/g = \infty$ the weights $f(x_i)/g(x_i)$ may take very high values and few values x_i influence the estimate of (14).

Importance sampling iv

- Note also that

$$\mathrm{E}_g\left[h(X)^2\frac{f(X)^2}{g(X)^2}\right] = \int_{\mathcal{X}} h(x)^2\frac{f(x)^2}{g(x)^2}\,dx$$

the ratio $f(x)/g(x)$ should be bounded when $f(x)$ is not negligible...hence the modes of $f(x)$ and $g(x)$ should be close each other.

Importance sampling for Bayesian inference

- In Bayesian inference we need to compute quantities coming from the posterior distribution, such as::

$$\mathrm{E}_{\pi(\theta|y)}[h(\theta)] = \frac{\int_\Theta h(\theta)p(y|\theta)\pi(\theta)d\theta}{\int_\Theta p(y|\theta)\pi(\theta)}d\theta = \int_\Theta h(\theta)\frac{p(y|\theta)\pi(\theta)}{p(y)}d\theta, \quad (18)$$

  where $\pi(\theta)$ is the prior, $p(y|\theta)$ is the likelihood function and $p(y) = \int_\Theta p(y|\theta)\pi(\theta)d\theta$, the marginal likelihood, is often *unknown*.

- Given $\theta_1, \ldots, \theta_S$ i.i.d. from $g(\theta)$ an IS estimator for (18) is given by:

$$\mathrm{E}^{is}_{\pi(\theta|y)}[h(\theta)] = \frac{S^{-1}\sum_{s=1}^S h(\theta_s)\frac{p(y|\theta_s)\pi(\theta_s)}{p(y)g(\theta_s)}}{S^{-1}\sum_{s=1}^S \frac{p(y|\theta_s)\pi(\theta_s)}{p(y)g(\theta_s)}} \quad (19)$$

IS for Bayesian inference: location of a t-distribution

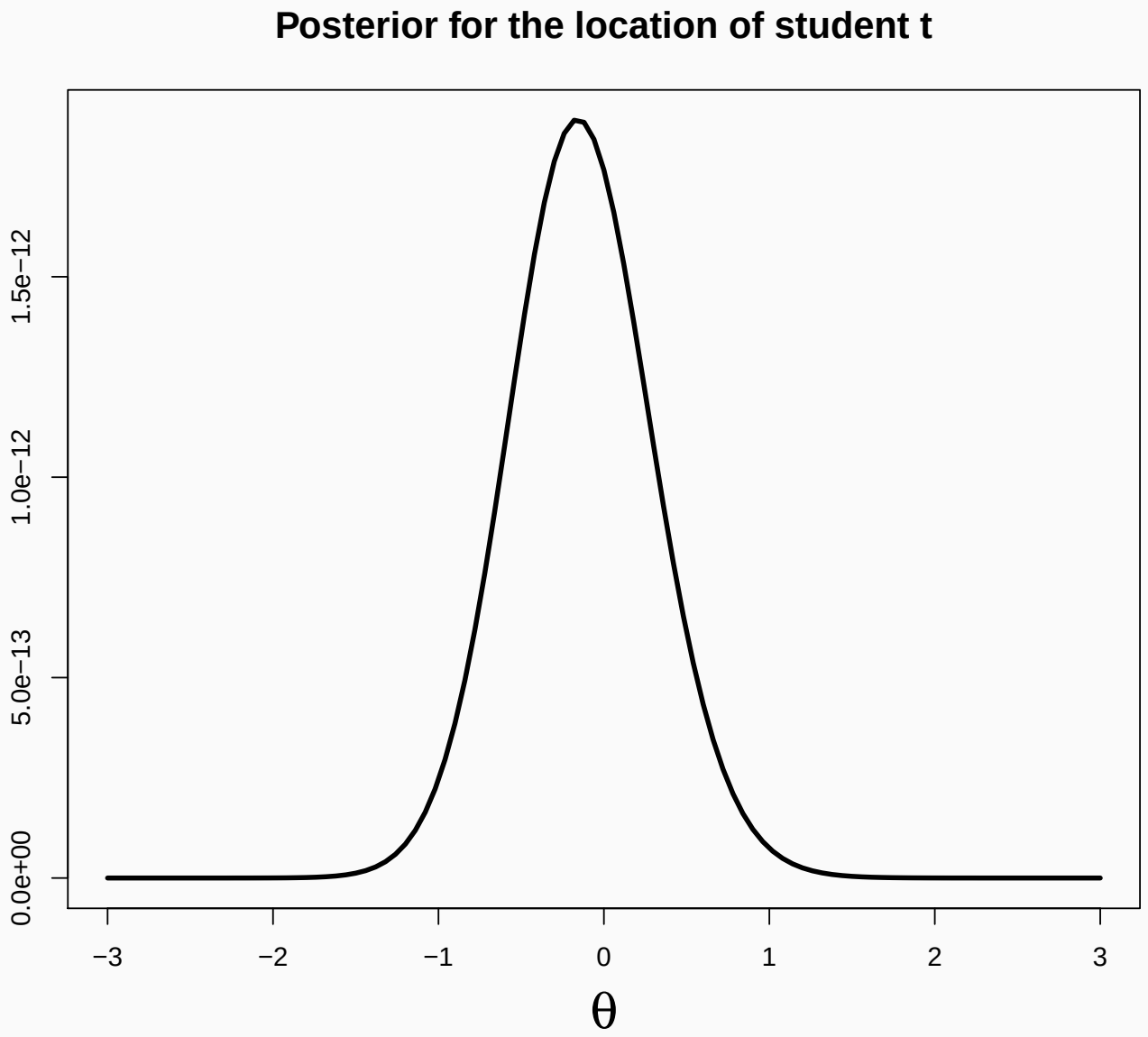
- Let $y_1, \ldots . y_n$ be an i.i.d. sample from a student-t with fixed degrees of freedom:

  ```
  y.t <- rt(n=9, df =3)
  ```

- Let be θ the location parameter (in the simulation $\theta = 0$) and take $\pi(\theta) \propto 1$. Then the posterior for θ is:

$$\pi(\theta|y) \propto \prod_{i=1}^{n} [3 + (y_i - \theta)^2]^{-2}$$

IS for Bayesian inference: location of a t-distribution



IS for Bayesian inference: location of a t-distribution

- Consider the posterior mean:

$$\mathrm{E}(\theta|y) = \frac{\int_\Theta \theta \prod_{i=1}^n [3 + (y_i - \theta)^2]^{-2} d\theta}{\int_\Theta \prod_{i=1}^n [3 + (y_i - \theta)^2]^{-2} d\theta}$$

  Possible strategies for computation:

  - draws from the prior are not proper (the prior is improper)
  - draws from the posterior are not possible (we are not able to do them)
  - draws from the components $g(\theta) \propto p(y_i|\theta)$? maybe...

IS for Bayesian inference: location of a t-distribution

- For example take:

$$g(\theta) \propto p(y_i|\theta) \propto [3 + (y_i - \theta)^2]^{-2}.$$

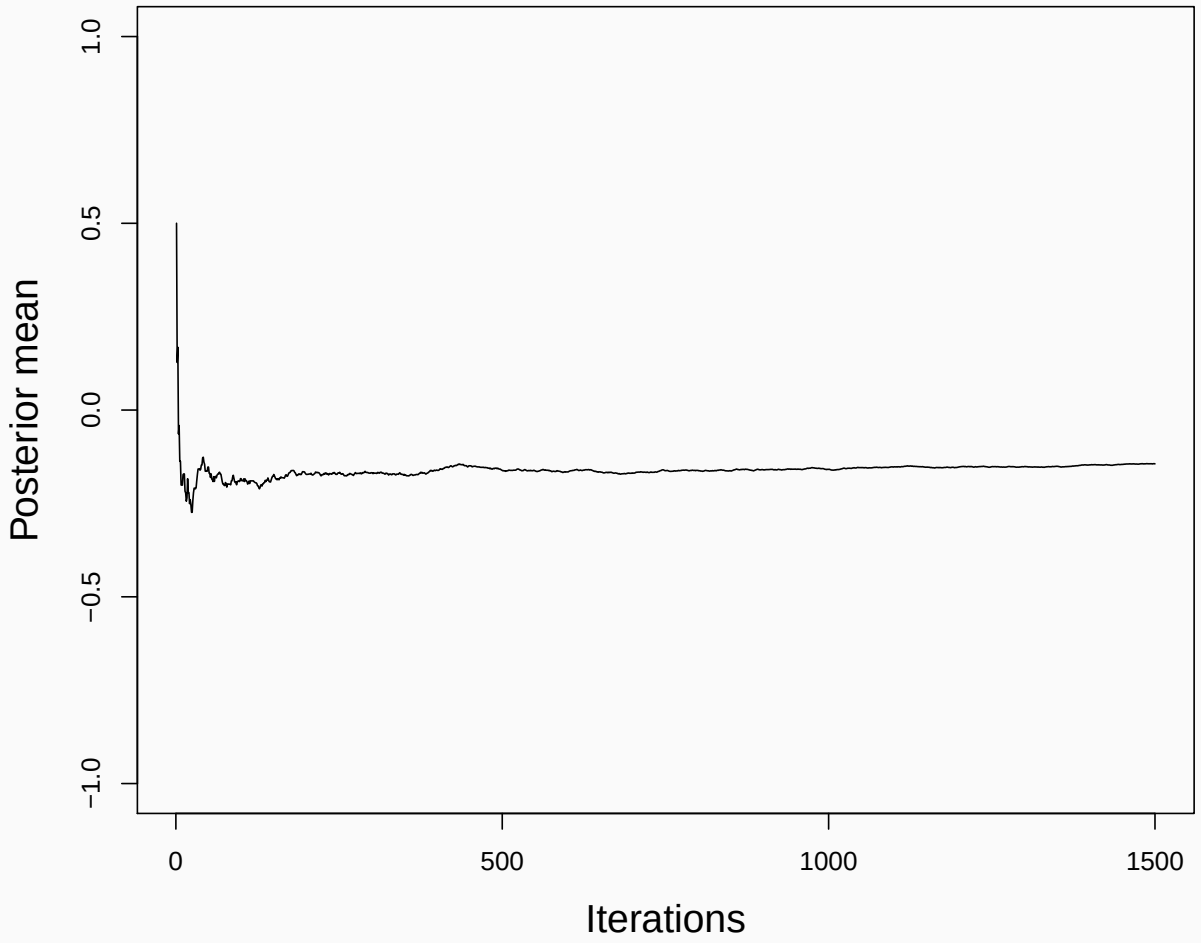
- Given S draws from $g(\theta)$, estimate the posterior mean by:

$$\mathrm{E}^{is}(\theta|y) = \frac{\sum_{s=1}^{S} \theta_s \frac{\prod_{i=1}^{n}[3+(y_i-\theta)^2]^{-2}}{[3+(y_i-\theta)^2]^{-2}}}{\sum_{s=1}^{S} \frac{\prod_{i=1}^{n}[3+(y_i-\theta)^2]^{-2}}{[3+(y_i-\theta)^2]^{-2}}} = \frac{\sum_{s=1}^{S} \theta_s \prod_{i=1}^{n}[3 + (y_i - \theta)^2]^{-2}}{\sum_{s=1}^{S} \prod_{i=1}^{n}[3 + (y_i - \theta)^2]^{-2}}$$

IS for Bayesian inference: location of a t-distribution

```
t.medpost = function(nsim, data, l) {
   sim <- data[l] + rt(nsim, 3)
   n <- length(data)
   s <- c(1:n)[-l]
   num <- cumsum(sim * sapply(sim,
    function(theta) t.lik(theta, data[s])))
   den <- cumsum(sapply(sim,
    function(theta) t.lik(theta, data[s])))
   num/den
   }
 media.post <- t.medpost(nsim = 1500, data = y.t,
                         l = which(y.t == median(y.t)))
 media.post[1500]
[1]-0.1440603
```

IS for Bayesian inference: location of a t-distribution



The convergence seems to be reached even after a few observations. What if we sample from other g's?

IS for Bayesian inference: location of a t-distribution

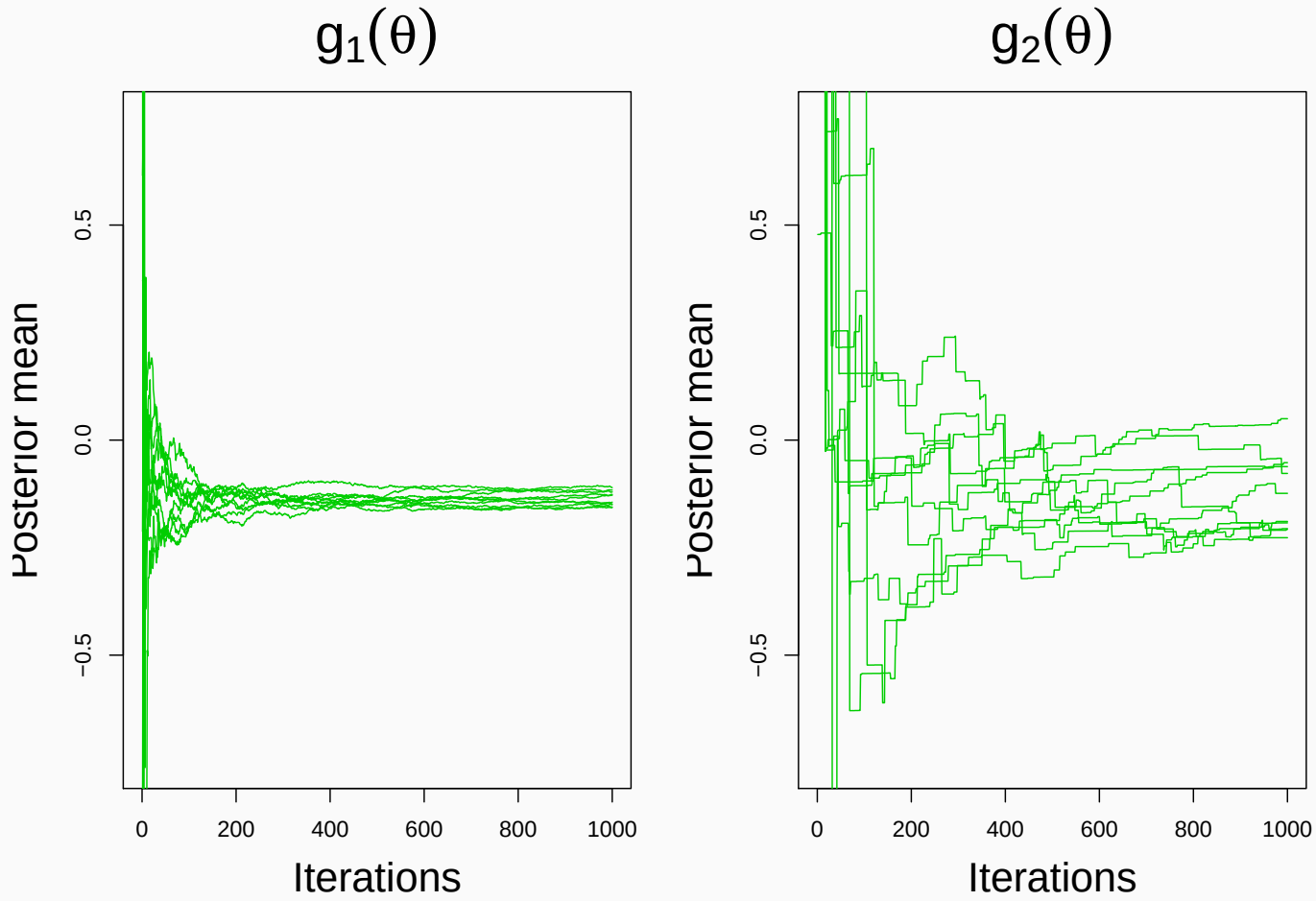
$g_1(\theta) \propto p(y_{(n/2)}|\theta)$

```
par(mfrow = c(1, 2))
  plot(c(0, 0), xlim = c(0, 1000),
       ylim = c(-0.75,0.75), type = "n", ylab = "Posterior mean",
       xlab="Iterations", main =)
  for (i in 1:10) {
    lines(x = c(1:1000), y = t.medpost(nsim = 1000,
          data = y.t, l = which(y.t == median(y.t))), col = 3)}
```

$g_2(\theta) \propto p(y_{(n)}|\theta)$

```
  plot(c(0, 0), xlim = c(0, 1000), ylim = c(-0.75,  0.75),
         type = "n", ylab = "Posterior mean",
         xlab ="Iterations")
  for (i in 1:10) {
    lines(x = c(1:1000), y = t.medpost(nsim = 1000,
                data = y.t, l = which(y.t == max(y.t))), col = 3)}
```

IS for Bayesian inference: location of a t-distribution



There is greater variability and slower convergence if we sample from the distribution of the maximum.

Further reading

Further reading:

- Chapter 4 from *Bayesian Data Analysis*, A. Gelman et al.
- Chapter 5 from *Bayesian computation with R*, J. Albert
- Chapter 3 and 5 from *Introducing Monte Carlo Methods with R*, C. Robert and G. Casella.