



UNIVERSITÀ
DEGLI STUDI DI TRIESTE



Dipartimento di scienze economiche,
aziendali, matematiche e statistiche
“Bruno de Finetti”

Bayesian Statistics

Introduction to Monte Carlo methods

Leonardo Egidi

A.A. 2018/19

Indice

- 1 Motivations
- 2 Numerical integration
- 3 Accept-reject method
- 4 Monte-Carlo integration

Motivations

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

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

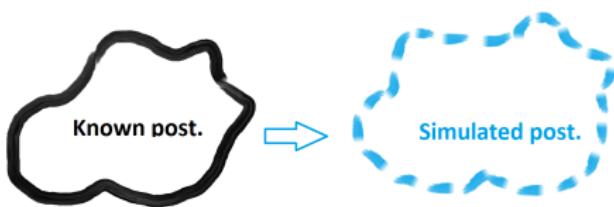
This is **easy for conjugate models: normal likelihood + normal prior, ...**
beta+binomial, Poisson+gamma, multinomial+Dirichlet

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).**

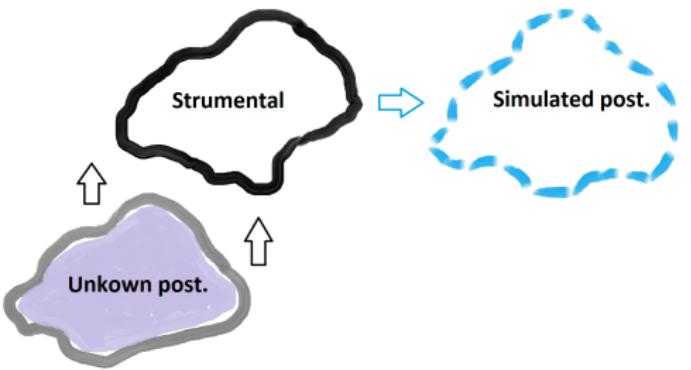
Motivations

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.



Motivations

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

$$E_f[h(X)] = \int_{\mathcal{X}} h(x) \underbrace{f(x)}_{\text{General func. of } x} dx, \quad (2)$$

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

- Numerical integrations
- Asymptotic approximations
- 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).

Indice

- 1 Motivations
- 2 Numerical integration
- 3 Accept-reject method
- 4 Monte-Carlo 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:

Much Concentrated around this value

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

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.

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
```

Very
wrong
values

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

Indice

1 Motivations

2 Numerical integration

3 Accept-reject method

4 Monte-Carlo integration

Accept-reject method

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

$$E_f[h(\theta)] = \int_{\Theta} h(\theta) f(\theta) d\theta, \quad \text{Spot the change} \quad (3)$$

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



Assume that

Known up to scaling factor

- 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 .
- ② There exists another density $g(\theta)$, an **instrumental** density, such that, for some **big** c , $d(\theta) \leq c \times g(\theta), \forall \theta$.

*g poses an upper bound on d (up to large scaling factor)
or $d/g \leq c$ (ratio is limited) \Rightarrow Support of instrument larger than target*

Accept-reject method

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

Allow for determine it?
(will see in lab)

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

Large this, the more
plausible the value

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

Accept-reject method

Theorem

- (a) The distribution of the accepted value is exactly the target density $f(\theta)$.
- (b) The marginal probability that a single candidate is accepted is K/c .
→ The two const. hyperparameters of the algorithm

Accept-reject method

Proof.

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

$$\begin{aligned}
 F_W(\theta) &= \frac{\Pr(W \leq \theta, Y \leq \frac{d(w)}{c \times g(w)})}{\Pr(Y \leq \frac{d(w)}{c \times g(w)})} = \frac{\int_W \Pr(W \leq \theta, Y \leq \frac{d(w)}{c \times g(w)} | w) g(w) dw}{\int_W \Pr(Y \leq \frac{d(w)}{c \times g(w)} | w) g(w) dw} = \\
 &= \frac{\int_{-\infty}^{\theta} \Pr(Y \leq \frac{d(w)}{c \times g(w)} | w) g(w) dw}{\int_{-\infty}^{+\infty} \Pr(Y \leq \frac{d(w)}{c \times g(w)} | w) g(w) dw} = \frac{\int_{-\infty}^{\theta} \frac{d(w)}{c} dw}{\int_{-\infty}^{+\infty} \frac{d(w)}{c} dw} = \\
 &= \frac{\int_{-\infty}^{\theta} \frac{Kf(w)}{c} dw}{\int_{-\infty}^{+\infty} \frac{Kf(w)}{c} dw} = \int_{-\infty}^{\theta} f(w) dw.
 \end{aligned}$$

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

$$\begin{aligned}
 \Pr(W \text{ accepted}) &= \Pr(Y \leq \frac{d(W)}{c \times g(W)}) = \\
 &= \int_W \Pr(Y \leq \frac{d(W)}{c \times g(W)} | W = w) g(w) dw = \\
 &= \int_W \frac{d(w)}{c} dw = \int_W \frac{K}{c} f(w) dw = \frac{K}{c}
 \end{aligned}$$

A-R algorithm: simulation from a Beta distribution

$d(\theta)$ **TARGET** CONSTANT w.r.t θ (or x)

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

$$f(\theta) = \frac{d(\theta)}{K} = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \xrightarrow{\text{KNOWN (PART OF DENSITY)}} \xrightarrow{\text{NOT UNKNOWN}} \begin{cases} \text{part of beta distro. not dependent upon } g \\ \text{NECESSARILY PART OF DENSITY} \end{cases}$$

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

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

2 uniforms extraction

NB: $\text{Beta}(1,1)$ is $\text{Unif}(0,1)$

The larger c , the smaller the prob. of acceptance

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
```

Maximum of
beta density
(because \int_0^1
already know it)
=

Usually $\text{runif}(\frac{1}{c}, c)$,
otherwise the UNIF arts
the target prob. dist.

Also, tails of
instrumental
leaving (higher)
Then target

A-R algorithm: simulation from a Beta distribution

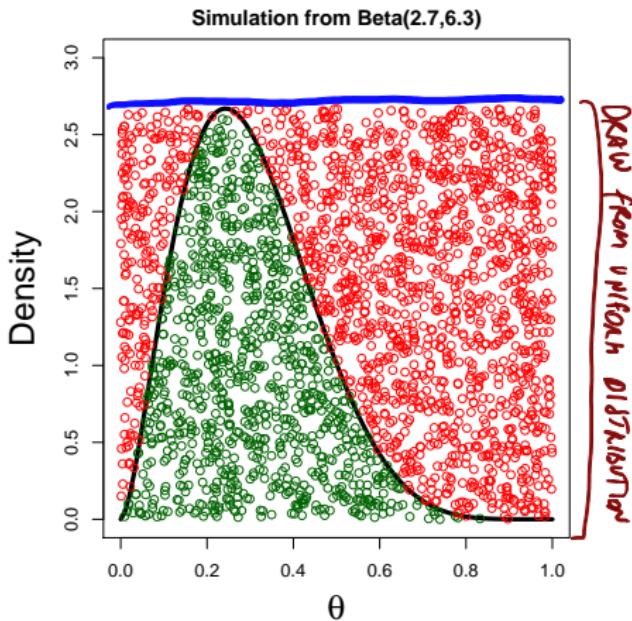
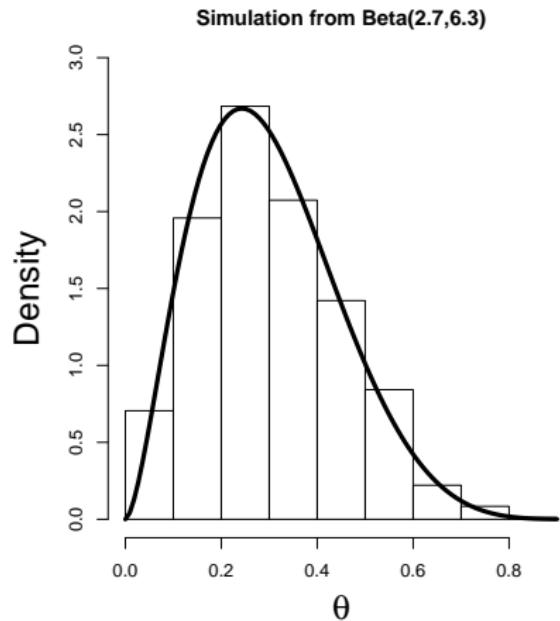


Figure: On the left plot, the true Beta(2.7, 6.3), and the histogram of the simulated distribution. On the right plot, the pairs (θ^*, U) : the accepted (green) and the discarded (red). $K = 1$.

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

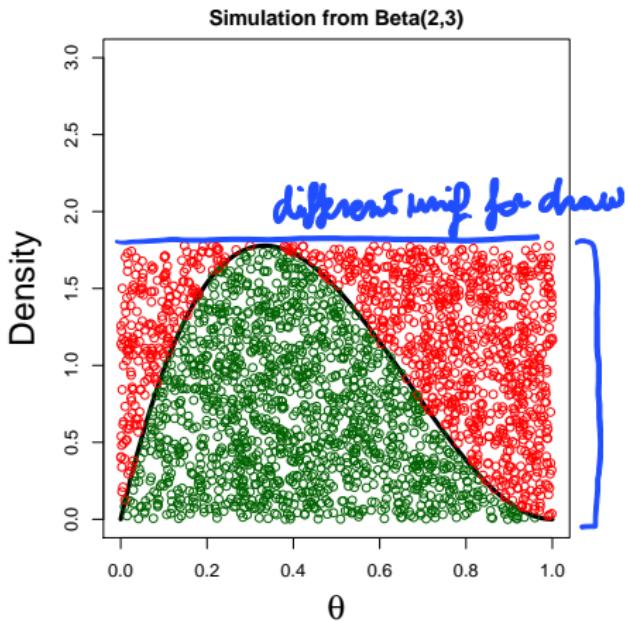
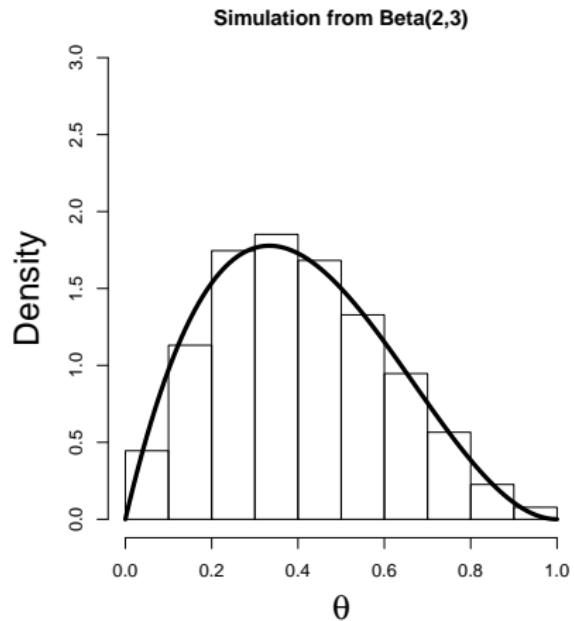


Figure: On the left plot, the true Beta(2,3), and the histogram of the simulated distribution. On the right plot, the pairs (θ^*, U) : the accepted (green) and the discarded (red). $K = 1$.

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 .

Curse of dimensionality

Indice

- 1 Motivations
- 2 Numerical integration
- 3 Accept-reject method
- 4 Monte-Carlo integration

Indice

- 1 Motivations
- 2 Numerical integration
- 3 Accept-reject method
- 4 Monte-Carlo integration
 - Classical MC
 - Importance sampling

Classical Monte Carlo integration

(\forall dist ...) If know generative math,
can know prob. dist.

Two major classes of numerical problems that arise in statistical inference are *optimization* problems and *integration* problems.



Suppose we need to calculate:

$$E_f[h(X)] = \int_{\mathcal{X}} h(x)f(x)dx, \quad \begin{matrix} \xrightarrow{\text{f must be known}} \\ \text{(But that's not usually the case)} \end{matrix} \quad (4)$$

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?

• Approximate by a large sample of draws

If $|I| < \infty$ and X_1, X_2, \dots, X_S are i.i.d $\sim f$, then by the Strong Law of Large Numbers, we have that the empirical mean is consistent for $E_f[h(X)]$

Monte Carlo est.

$$\widehat{E_f[h(X)]} = \frac{1}{S} \sum_{s=1}^S h(X_s) \rightarrow \text{true value } E_f[h(X)] \text{ in probability, as } S \rightarrow \infty \quad (5)$$

Classical Monte Carlo integration

The variance of $\widehat{E_f[h(X)]}$ is

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

and it can be approximated by

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

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

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

Example: Normal mean with Cauchy prior

Consider: *if $y \sim N, \mu | \theta \sim N \Rightarrow \text{Only } \mu$*
But if $y | \theta \sim N, \theta \sim \text{Cauchy}$, Only don't know

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

The posterior mean for a single observation y is:

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

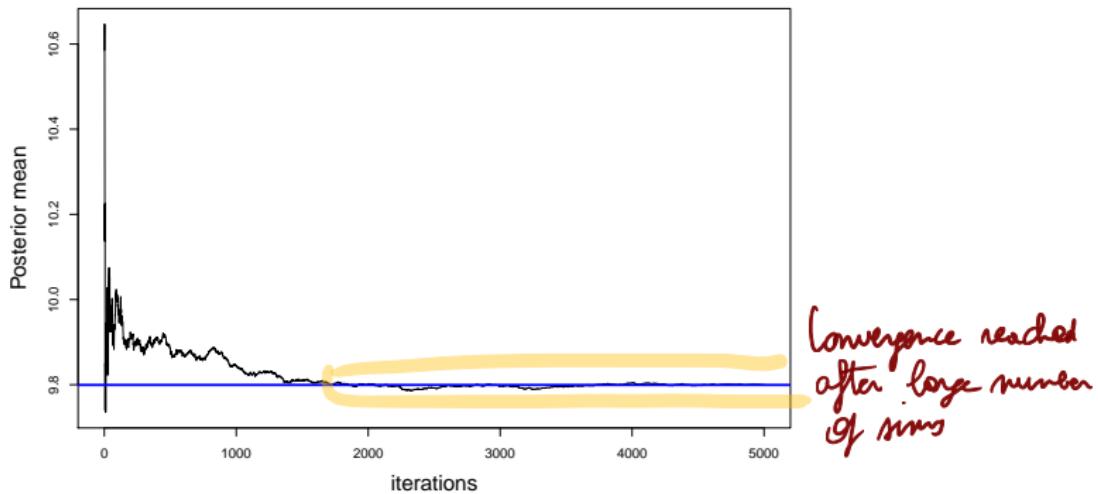
We could draw $\theta_1, \dots, \theta_S$ from $\mathcal{N}(y, 1)$ and compute:

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

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
```



Indice

- 1 Motivations
- 2 Numerical integration
- 3 Accept-reject method
- 4 Monte-Carlo integration
 - Classical MC
 - Importance sampling

Importance sampling

Importance sampling is based on the following representation:

$$\begin{aligned}
 E_f[h(X)] &= \int_{\mathcal{X}} h(x)f(x)dx = \\
 &= \int_{\mathcal{X}} h(x) \frac{f(x)}{g(x)} g(x)dx = E_g\left[h(X) \frac{f(X)}{g(X)}\right]
 \end{aligned}
 \tag{6}$$

Shift attention to g (known)

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

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

$$E_f^{is}[h(X)] = \frac{1}{S} \sum_{s=1}^S h(x_s) \frac{f(x_s)}{g(x_s)} = \frac{1}{S} \sum_{s=1}^S h(x_s) \underbrace{w(x_s)}_{\substack{\text{Weighted mean} \\ \text{Weights}}}, \tag{7}$$

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

Importance sampling

Note that classical Monte Carlo and importance sampling both produce unbiased estimator for the integral (4), but:

Import. Sampling

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

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



We can work on g in order to minimize the variance of (7). The constraint that $\text{supp}(h \times f) \subset \text{supp}(g)$ is absolute in that using a smaller support truncates the integral (4) 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.

Importance sampling

- IS variance is finite only when

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

- 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 (4).
- Note also that

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

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::

$$E_{\pi(\theta|y)}[h(\theta)] = \frac{\int_{\Theta} h(\theta) \underbrace{p(y|\theta)}_{\text{lik. prob.}} \pi(\theta) d\theta}{\int_{\Theta} p(y|\theta) \pi(\theta) d\theta} = \int_{\Theta} h(\theta) \frac{p(y|\theta) \pi(\theta)}{p(y)} d\theta, \quad (8)$$

↑
posterior

Marg. lik.

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 $\theta_1, \dots, \theta_S$ i.i.d. from $g(\theta)$ an IS estimator for (8) is given by:

$$E_{\pi(\theta|y)}^{is}[h(\theta)] = \frac{\frac{1}{S} \sum_{s=1}^S h(\theta_s) \frac{p(y|\theta_s) \pi(\theta_s)}{p(y) g(\theta_s)}}{\frac{1}{S} \sum_{s=1}^S \frac{p(y|\theta_s) \pi(\theta_s)}{p(y) g(\theta_s)}} \quad \text{UNNORMALIZED POSTERIOR DISTRIBUT.} \quad (9)$$

IS for Bayesian inference: location of a t -distribution

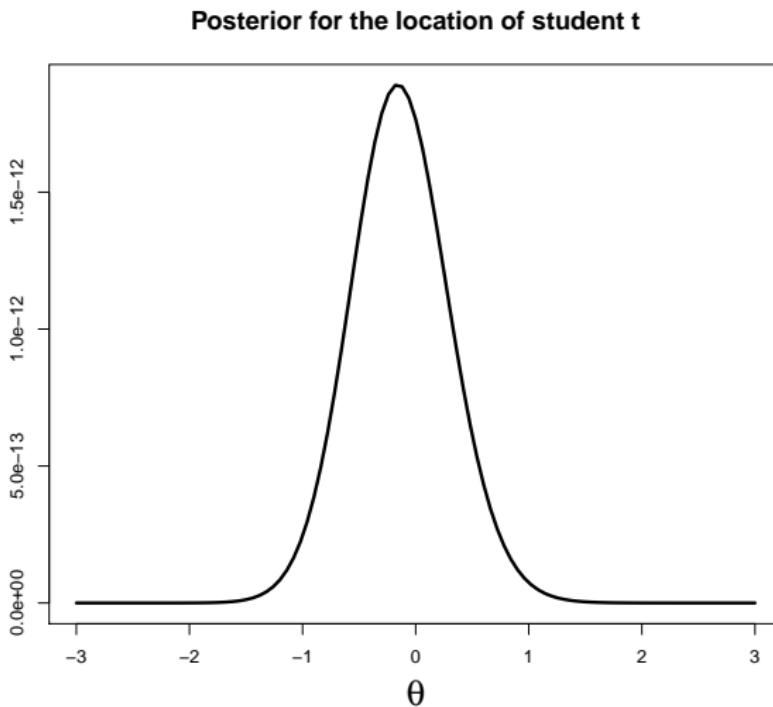
Let y_1, \dots, y_n be an i.i.d. sample from a student- t with fixed degrees of freedom:

```
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:

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

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:

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

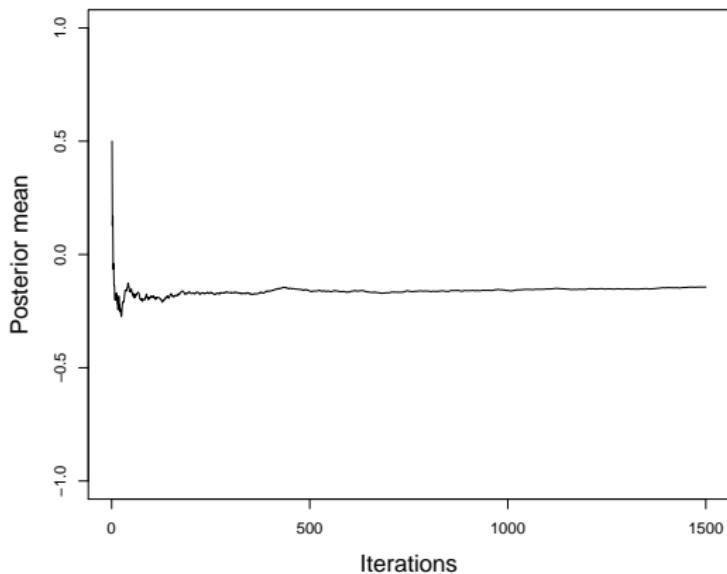
Given S draws from $g(\theta)$, estimate the posterior mean by:

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

IS for Bayesian inference: location of a t -distribution

```
t.medpost = function(nsim, data, 1) {  
  sim <- data[1] + rt(nsim, 3)  
  n <- length(data)  
  s <- c(1:n)[-1]  
  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?

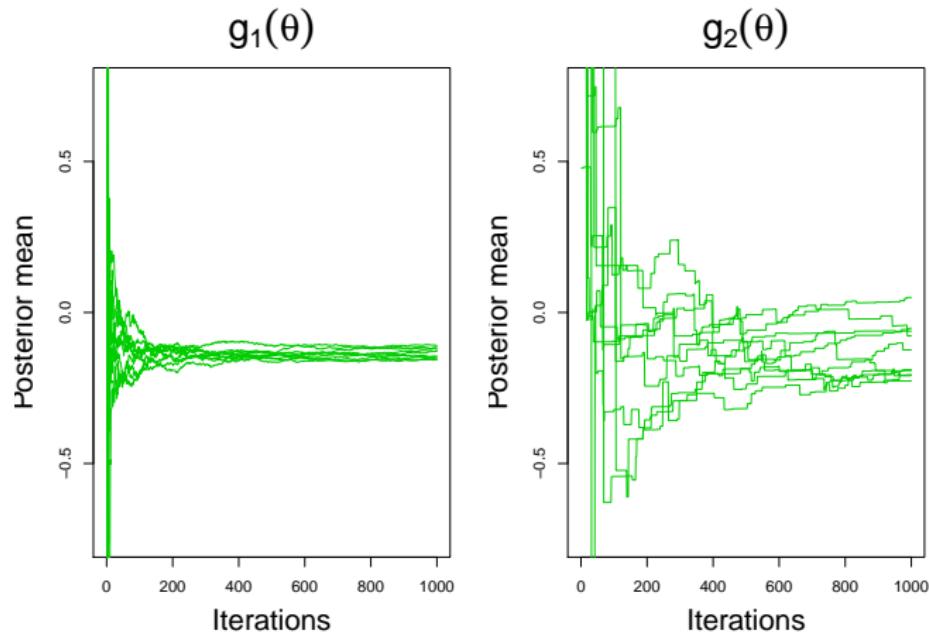
IS for Bayesian inference: location of a t -distribution

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

```
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)}
```

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

```
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 5 from *Bayesian computation with R*, J. Albert
- Chapter 3 and 5 from *Introducing Monte Carlo Methods with R*, C. Robert and G. Casella.