Lecture 5: Rejection sampling

Let $f(x)$ denote the target density.

1. Generate $x \sim g(x)$
2. Compute $\alpha = \frac{1}{c} \cdot \frac{f(x)}{g(x)}$.
3. Generate $u \sim \mathcal{U}(0,1)$.
4. If $u \leq \alpha$ return x (acceptance step).
5. Otherwise go back to (1) (rejection step).

Note $\alpha \in [0,1]$ and α is called acceptance probability.

Claim: The returned x is distributed according to $f(x)$.

Example: Setting

Suppose we want to sample standard normal random numbers.

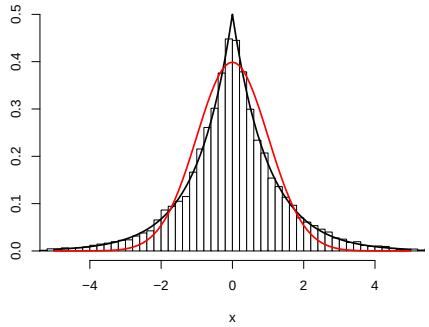
Then

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

As proposal distribution we use a double exponential distribution:

$$g(x) = \frac{\lambda}{2} \exp(-\lambda|x|), \lambda > 0$$

```
> g <- function(x, lambda=1){
+   return(lambda/2 *
+     exp(-lambda * abs(x)))
+ }
> rg <- function(n, lambda){
+   z = rexp(n, lambda)
+   y = sample(c(0,1), n,
+     prob=c(0.5,0.5), replace=TRUE)
+   x = c(z[y==0], -z[y==1])
+   return(x)
+ }
```



Example: Find an efficient bound c

$$\begin{aligned}
\frac{f(x)}{g(x)} &= \frac{\frac{1}{\sqrt{2\pi}} \exp(-1/2x^2)}{\frac{\lambda}{2} \exp(-\lambda|x|)} \\
&= \sqrt{\frac{2}{\pi}} \lambda^{-1} \exp\left(-\frac{1}{2}x^2 + \lambda|x|\right) \\
&\leq \sqrt{\frac{2}{\pi}} \lambda^{-1} \exp\left(\max_{x \in \mathbb{R}} \{-\frac{1}{2}x^2 + \alpha|x|\}\right) \\
&\overset{|x|=\lambda}{=} \sqrt{\frac{2}{\pi}} \lambda^{-1} \exp\left(\frac{1}{2}\lambda^2\right) \\
&\equiv c
\end{aligned}$$

Example: Acceptance probability

Thus the acceptance probability becomes

$$\begin{aligned}
\alpha = \frac{1}{c}\frac{f(x)}{g(x)} &= \frac{\sqrt{\frac{2}{\pi}} \lambda^{-1} \exp\left(-\frac{1}{2}x^2 + \lambda|x|\right)}{\sqrt{\frac{2}{\pi}} \lambda^{-1} \exp\left(\frac{1}{2}\lambda^2\right)} \\
&= \exp\left\{-\frac{1}{2}x^2 - \frac{1}{2}\lambda^2 + \lambda|x|\right\}
\end{aligned}$$

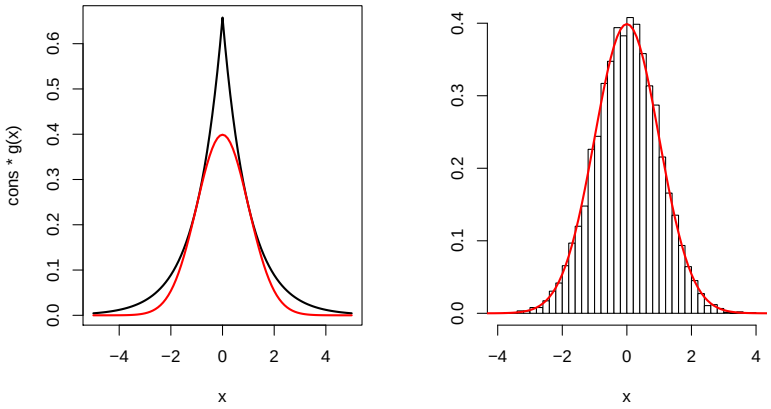
Note, the algorithm is correct for all values of $\lambda > 0$. However, we should choose $\lambda > 0$ so that c becomes as small as possible and consequently α.

$\Rightarrow$ Choose the λ that minimises c which is $\lambda = 1$

$$\frac{f(x)}{g(x)} \leq \sqrt{\frac{2}{\pi}} \lambda^{-1} \exp\left(\frac{1}{2}\lambda^2\right) \overset{\lambda=1}{=} \sqrt{\frac{2}{\pi}} \exp\left(\frac{1}{2}\right) \approx 1.32$$

$(1/1.32 \approx 0.7602)$.

Example: Illustration



- Left: Comparison of $f(x)$ versus $c \cdot g(x)$ when $\lambda = 1$.
- Right: Distribution of accepted samples compared to $f(x)$. 10000 samples were generated and 7582 accepted.

Continuation: Standard Cauchy

How can we sample from the semi-unit circle?

Rejection sampling also works when x is a vector:

$$C_f = \{(x_1, x_2) \mid x_1^2 + x_2^2 \leq 1, x_1 > 0\}$$

with

$$f(x_1, x_2) = \frac{1}{\text{area}(C_f)}, \quad (x_1, x_2) \in C_f$$

Let the proposal density be

$$g(x_1, x_2) = \begin{cases} \frac{1}{2} & x_1 \in [0, 1], x_2 \in [-1, 1] \\ 0 & \text{otherwise} \end{cases}$$

Thus the density g is that $x_1 \in \mathcal{U}(0, 1)$, $x_2 \in \mathcal{U}(-1, 1)$ independently.

Standard Cauchy: Rejection sampling algorithm

finished $= 0$

while finished $= 0$ **do**

generate $(x_1, x_2) \sim g(x_1, x_2)$

compute

$$\alpha = \frac{1}{c}\frac{f(x_1,x_2)}{g(x_1,x_2)} = \begin{cases} \frac{1}{c} \cdot \frac{2}{\text{area}(C_f)} \overset{c=\frac{2}{\text{area}(C_f)}}{=} 1, & (x_1, x_2) \in C_f \\ 0, & \text{otherwise} \end{cases}$$

generate $u \sim \mathcal{U}(0, 1)$

if $u \leq \alpha$ **then** finished $= 1$

end if ▷ i.e. If $(x_1, x_2) \in C_f$ finished $= 1$

end while

return x_1, x_2

Standard Cauchy: Summary

Note: To do this algorithm we do not need to know the value of the normalising constant area(C_f).

This is always true in rejection sampling.

Rejection sampling - Acceptance probability

Note: For c to be small, $g(x)$ must be similar to $f(x)$.
The art of rejection sampling is to find a g(x) that is similar to f(x) and which we know how to sample from.

Issues: c is generally large in high-dimensional spaces, and since the overall acceptance rate is $1/c$, many samples will get rejected.

Weighted resampling

A problem when using rejection sampling is to find a legal value for c. An approximation to rejection sampling is the following:

Let, as before:

- $f(x)$: target distribution
- $g(x)$: proposal distribution

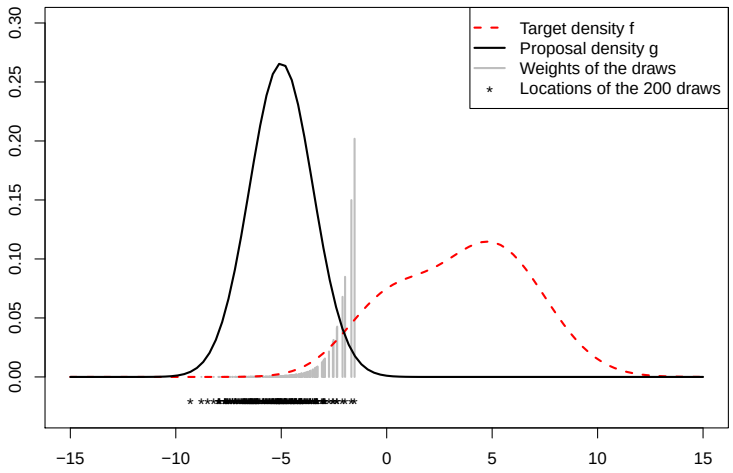
Algorithm

- Generate $x_1, \ldots, x_n \sim g(x)$ iid
- Compute weights
$$w_i = \frac{\frac{f(x_i)}{g(x_i)}}{\sum_{j=1}^{n} \frac{f(x_j)}{g(x_j)}}$$
- Generate a second sample of size m from the discrete distribution on $\{x_1, \ldots, x_n\}$ with probabilities $w_1, \ldots, w_n$.

Comments

- The resulting sample has approximate distribution f
- The resample can be drawn with or without replacement provided that $n >> m$, a suggestion is $n/m = 20$.
- The normalising constant is not needed.
- This approximate algorithm is sometimes called sampling importance resampling (SIR) algorithm.

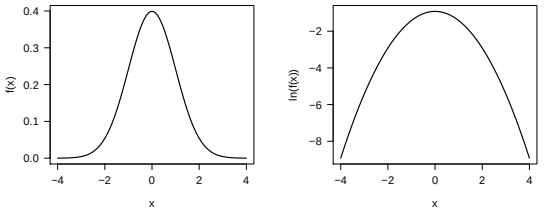
Illustration



Adaptive rejection sampling

This method works only for log concave densities, i.e.

$$\frac{\partial}{\partial x} \ln f(x) \leq 0, \quad \text{for all } x.$$

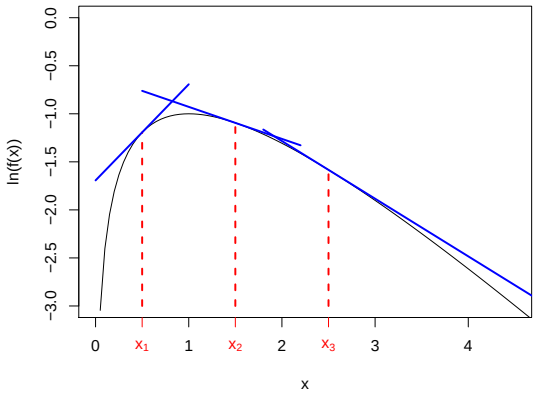


Many densities are log-concave, e.g. the normal, the gamma ($a > 1$), densities arising in GLMs with canonical link.

Basic idea: Form an upper envelope (the upper bound on $f(x)$) adaptively and use this in place of $c \cdot g(x)$ in rejection sampling.

Adaptive rejection sampling (2)

- Start with an initial grid of points $x_1, x_2, \ldots, x_m$ (with at least one x_i on each side of the maximum of $\ln(f(x))$) and construct the envelope using the tangents at $\ln(f(x_i))$, $i = 1, \ldots, m$.
- Draw a sample from the envelop function and if accepted the process is terminated. Otherwise, use it to refine the grid.



Monte Carlo integration

Assumption

It is *easy* to generate independent samples $x^{(1)}, \ldots, x^{(M)}$ from a distribution $f(x)$ of interest.

A Monte Carlo estimate of the mean

$$\mathrm{E}(x) = \int x f(x) dx$$

is then given by

$$\hat{\mathrm{E}}(x) = \frac{1}{M} \sum_{m=1}^{M} x^{(m)}.$$

The strong law of large numbers ensures, that this estimate is consistent. This approach is called Monte Carlo integration

Monte Carlo integration (II)

Monte Carlo integration

Suppose $x^{(1)}, \ldots, x^{(M)}$ is an iid sample drawn from $f(x)$. Then the strong law of large numbers says:

$$\hat{\mathrm{E}}(g(x)) = \frac{1}{M}\sum_{m=1}^{M} g(x^{(m)}) \overset{a.s}{\to} \int g(x)f(x)dx = \mathrm{E}(g(x))$$

Examples

- Using $g(x) = x^2$ we obtain an estimate for $\mathrm{E}(x^2)$.
- An estimate for the variance follows as

$$\widehat{\mathrm{Var}}(x) = \hat{\mathrm{E}}(x^2) - \hat{\mathrm{E}}(x)^2$$

Importance sampling

One of the principal reasons for wishing to sample from complicated probability distributions $f(z)$ is to be able to evaluate expectations with respect to some function $p(z)$:

$$\mathrm{E}(p) = \int p(z)f(z)dz$$

The technique of importance sampling provides a framework for approximating expectations directly but does not itself provide a mechanism for drawing samples from a distribution.

Importance sampling (2)

Importance sampling is based on the use of a proposal distribution $g(x)$ from which it is easy to draw samples.

$$\begin{aligned}\mathrm{E}(p) &= \int p(z)f(z)dz \\ &= \int p(z)\frac{f(z)}{g(z)}g(z)dz\end{aligned}$$

whre $w(z) = f(z)/g(z)$ are known as importance weights.

Importance sampling estimators

The former expression suggests two different importance sampling estimators

$$\hat{\mathrm{E}}(p) = \frac{1}{L}\sum_{l=1}^{L} p(z^{(l)}) \cdot w(z^{(l)}). \tag{1}$$

$$\hat{\mathrm{E}}(p) = \frac{1}{\sum_{l=1}^{L} w(z^{(l)})}\sum_{l=1}^{L} p(z^{(l)}) \cdot w(z^{(l)}). \tag{2}$$

The difference between these two estimates is usually small. The main advantage of the second estimator is that it does not require the normalizing constants of f and g in order to be computed.

Importance sampling: Summary

As with rejection sampling, the success of importance sampling depends crucially on how well the proposal distribution $g(x)$ matches the target distribution $f(x)$.

Bayesian concept

... The essence of the Bayesian approach is to provide a mathematical rule explaining how you change your existing beliefs in the light of new evidence. In other words, it allows scientists to combine new data with their existing knowledge or expertise. ...

The Economist, September 30th 2000

Bayes Theorem I

named after the English theologian and mathematician Thomas Bayes [1701–1761]

The theorem relies on the asymmetry of the definition of conditional probabilities:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \Rightarrow \mathrm{P}(A \cap B) = \mathrm{P}(B)\,\mathrm{P}(A|B) \qquad (3)$$

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} \Rightarrow \mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B|A) \qquad (4)$$

for any two events A and B under regularity conditions, i.e. $\mathrm{P}(B) \neq 0$ in (3) and $\mathrm{P}(A) \neq 0$ in (4).

Bayes Theorem II

Thus, from $\mathrm{P}(A|B)\,\mathrm{P}(B) = \mathrm{P}(B|A)\,\mathrm{P}(A)$ follows

Bayes Theorem

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\mathrm{P}(B)} \overset{\text{Law of tot. prob.}}{=} \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\mathrm{P}(B|A)P(A) + \mathrm{P}(B|\bar{A})\,\mathrm{P}(\bar{A})}$$

More general, let $A_1, \ldots, A_n$ be *exclusive* and *exhaustive* events, then

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(B|A_i)\,\mathrm{P}(A_i)}{\sum_{i=1}^{n} \mathrm{P}(B|A_i)\,\mathrm{P}(A_i)}$$

Interpretation

$\mathrm{P}(A_i)$ prior probabilities

$\mathrm{P}(A_i|B)$ posterior probabilities

After observing B the prob. of A_i changes from $\mathrm{P}(A_i)$ to $\mathrm{P}(A_i|B)$.

Towards inference

A more general formulation of Bayes theorem is given by

$$f(X = x|Y = y) = \frac{f(Y = y|X = x)f(X = x)}{f(Y = y)}$$

where X and Y are random variables.

(Note: Switch of notation from P(.) to $f(.)$ to emphasise that we do not only relate to probabilities of events but to general probability functions of the random variables X and Y.)

Even more compact version

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}.$$

Posterior distribution

The posterior distribution is the most important quantity in Bayesian inference. It contains all information about the unknown parameter θ after having observed the data $X = x$.

Let $X = x$ denote the observed realisation of a random variable or random vector X with density function $f(x|\theta)$. Specification of a prior distribution with density function $f(\theta)$ allows to compute the density function of the posterior distribution using Bayes theorem:

$$\begin{aligned} f(\theta|x) &= \frac{f(x|\theta)f(\theta)}{f(x)} \\ &= \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}. \end{aligned}$$

For discrete parameter space the integral has to be replaced with a sum.

Posterior distribution (II)

Since the denominator in

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

does not depend on θ, the density of the posterior distribution is proportional to

$$\underbrace{f(\theta|x)}_{\text{Posterior}} \propto \underbrace{f(x|\theta)}_{\text{Likelihood}} \times \underbrace{\mathrm{p}(\theta)}_{\text{Prior}}$$

where $1/\int f(x|\theta)f(\theta)d\theta$ is the corresponding normalising constant to ensure $\int f(\theta|x)d\theta = 1$.

Reminder:

A likelihood approach uses only the likelihood and calculated Maximum Likelihood estimate (MLE), defined as the particular value of θ that maximises the likelihood.

Bayesian point estimates

Statistical inference about θ is based solely on the posterior distribution $f(\theta|x)$. Suitable point estimates are location parameters, such as:

- Posterior mean $\mathrm{E}(\theta|x)$:
$$\mathrm{E}(\theta|x) = \int \theta f(\theta|x)d\theta.$$
- Posterior mode $\mathrm{Mod}(\theta|x)$:
$$\mathrm{Mod}(\theta|x) = \arg\max_{\theta} f(\theta|x)$$
- Posterior median $\mathrm{Med}(\theta|x)$ is defined as the value a which satisfies
$$\int_{-\infty}^{a} f(\theta|x)d\theta = 0.5 \quad \text{and} \quad \int_{a}^{\infty} f(\theta|x)d\theta = 0.5$$