Monte Carlo Methods

Interested in

$$E\left[f\left(X\right)\right]=\int f\left(x\right)d\mu\left(x\right)$$

Examples:

- Type I error rate of a hypothesis test
- Mean width of a confidence interval procedure
- Evaluating a likelihood
- Finding posterior mean and variance

Often calculating these will be difficult.

Approximate with

$$\frac{1}{n}\sum_{i=1}^{n}f\left(x_i\right)$$

where $x_1,\ldots,x_n$ is sampled from $\mu\left(X\right)$.

Under certain regularity conditions,

$$\frac{1}{n}\sum_{i=1}^{n}f\left(x_i\right)\rightarrow E\left[f\left(X\right)\right]$$

Issues:

- probability measure being integrated over
- function being integrated
- sampling scheme
- form of convergence

Focus: Sampling Schemes

- Independently and identically distributed (IID)
- Importance sampling
- Sequential importance samplers (SIS)
- Markov Chain Monte Carlo (MCMC)
    - Gibbs sampling
    - Metropolis – Hastings (M-H)
    - Reversible jump
    - Bridge sampling
    - etc
- and so on

Choice is often driven by $\mu(X)$, e.g.,

IID infeasible leads to use of SIS or MCMC

SIS may work but Gibbs sampler has reducible chain

M-W works when SIS has poorly behaved importance sampling weights.

etc

There is no one Monte Carlo approach that will solve every problem.

Want to develop a set of tools that can be used, possibly in combination, for a wide range of problems.

Need to recognize when each of them should work and when modifications are needed.

Basic Simulation Methodology

Pseudo-random numbers:

"Random" numbers generated on computers are not random but generated by deterministic algorithms.

Basic problem: generate u_i, $0 < u_i < 1$, that appear to be an iid sample from the $U(0,1)$ distribution.

Once you have these, you can simulate from "any" distribution.

Uniform deviates

Instead of generating $U(0,1)$, most generators actually generate integers ($U(0,m-1)$ or $U(1,m-1)$) and then convert these to the interval $(0,1)$.

Numerically more stable and faster.

Multiplicative Congruential Generators:

Generate integer sequence $\{k_i\}$ by

$$k_{i+1} = ak_i \bmod m$$

for suitably chosen positive integers a and m, where $b \bmod m$ is the remainder from dividing b by m.

If $a^{m-1} = 1 \bmod m$ and $a^l \neq 1 \bmod m$ for $0 < l < m - 1$ and if k_0 is a positive integer that isn't a multiple of m, then it can be shown that $k_1, \ldots, k_{m-1}$ will be a permutation of $\{1, 2, \ldots, m - 1\}$ (a is said to be a primitive root of unity mod m).

The period of this generator is $m - 1$.

In general the period is the number of values until the generator starts to repeat.

Linear Congruential Generators

$$k_{i+1} = (ak_i + b) \bmod m$$

for suitable integers a, b, and m.

Good generators should have

- long periods
- low (near 0) correlations
- give samples that look uniform

This holds for any generator, not just congruential generators.

Choices for m, a, and b

1) $m = 2^{31} - 1$: largest prime integer that can be stored on most computers
   - $a = 7^5$ (IMSL, early versions of Matlab)
   - a = 950,706,376 (IMSL option, shown to be good by Fishman & Moore)
2) $m = 2^{32}$: number of integers that can be represented on most computers.
   - Can't get full period if b = 0 since m is even.
   - Maximum period of 2^{30} can be achieved if $a = 5 + 8l$ for some l. Common choice is a = 69069

- Can get full period of 2^{30} when $b \neq 0$, such as with $a = 69069$ and $b = 23606797$.

Problems with congruential generators:

- Must have some autocorrelation
- n dimensional uniformity (how close do n tuples of consecutive draws achieve uniformity in the n-dimensional unit cube).
- congruential generators tend to give n-vectors that concentrate near hyperplanes in n-dimensional space for some n.

Example: RANDU generator

- IBM SYSTEM/360 generator

$$k_{i+1} = \left(2^{16} + 3\right) k_i \bmod 2^{31}$$

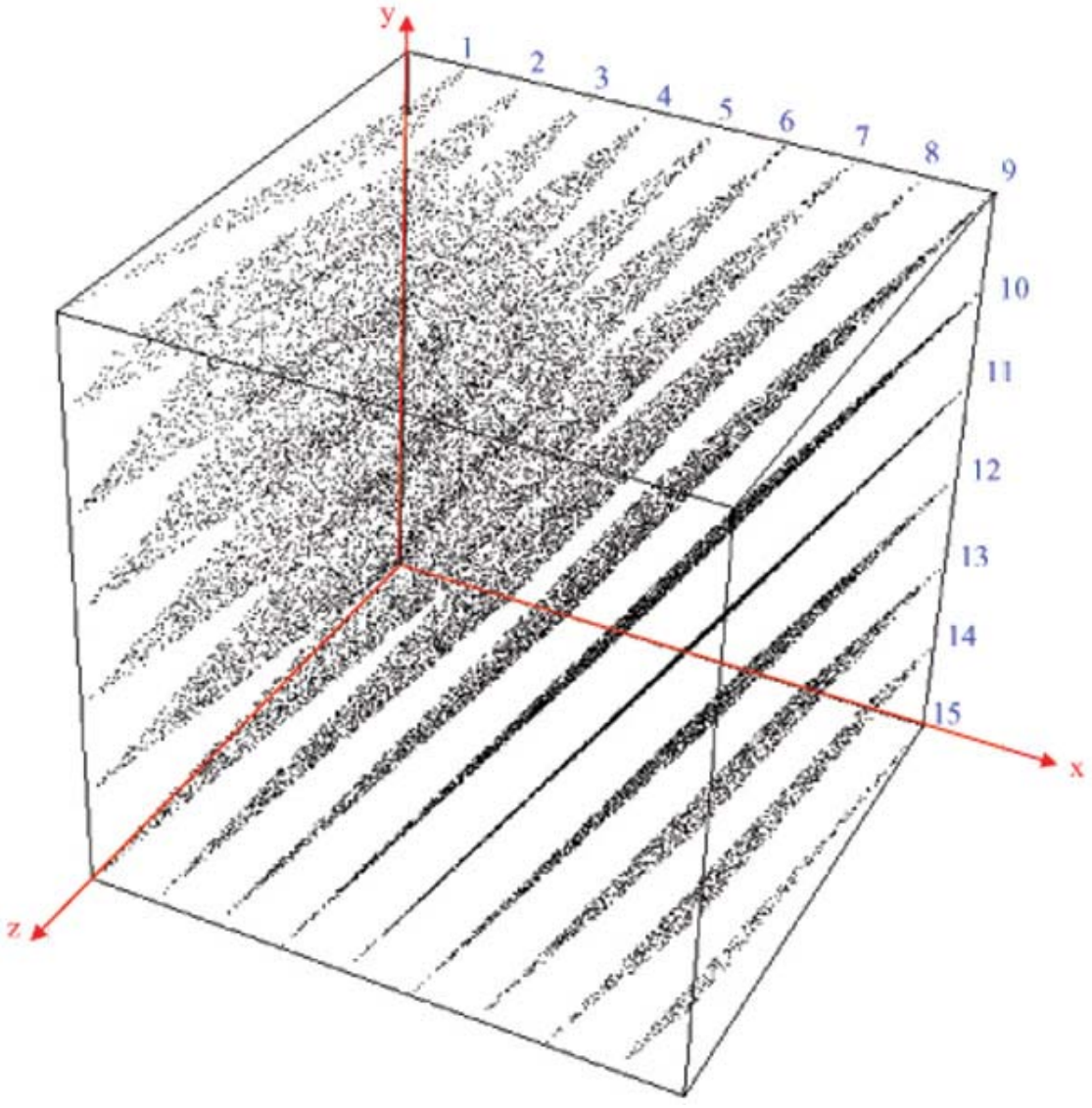
- Has the property that

$$9k_i - 6k_{i+1} + x_{i+2} = 0 \bmod 2^{31}$$

Proof:

$$9x_i - 6ax_i + a^2 x_i = (a-3)^2 x_i$$
$$= 2^{32} x_i = 2^{31} \times 2x_i$$

- Realizations of triples must fall on one of 15 planes, 2^{31} apart



from <http://www.unf.edu/ccec/cis/CIShtml/RANDUinfo.html>

- This generator is still around is system software.

  From HP documentation
  <http://h18009.www1.hp.com/fortran/docs/lrm/lrm0315.htm>

  GSL (Gnu Scientific Library)
  <http://www.gnu.org/software/gsl/manual/gsl-ref_17.html#SEC271>

  In GSL, its there for backward compatibility with old code that people use and historical completeness.

- With congruential generators, the leading bits tend to be more random than the low order bits, so one shouldn't treat subsets of the bits in the representation as separate random numbers
- Standard congruential generators have periods that are too short for some statistical applications.

Shuffling algorithm

- Initialize: $s(i) = u_i; i = 1, \ldots, N$ and set $y = s(N)$.
- Generate a new value u and set $j = \text{int}(yN) + 1$, where $\text{int}(x)$ is the largest integer ≤ x.
- Set $y = s(j)$, $s(j) = u$, and return y as the uniform deviate.

The idea behind this scheme is that a permutation of uniforms is still uniform.

By combining generators with long, but not equal periods, a new generator with a much longer period can be created.

Example: ran2 (Numerical recipies due to L'Ecuyer)

Generator 1 (v_i): a = 40014, m = 2147483563. Uses shuffle algorithm with N = 32.

Generator 2 (w_i): a = 40692, m = 2147483399.

Returns

$$u_i = (v_i - w_i) I(v_i \geq w_i) + (1 - v_i + w_i) I(v_i < w_i)$$

The period of this generator is the product of the periods of the two streams, divided by any common factors.

The period is about $2.3 \times 10^{18} \approx 2^{61}$.

Recursive generators

$$k_{i+1} = a_1 k_i + \ldots + a_l k_{i+1-l} \bmod m$$

Linear combination of the previous l values.

Maximum period: $m^l - 1$

Fibonacci generators

$$u_i = u_{i-17} - u_{i-5}$$
$$u_i = u_{i-97} - u_{i-33}$$

If lagged difference < 0, add 1 to result.

Shift / Tausworthe generators

Based on binary expansion of integers

$$j = \sum_{l=1}^{32} b_l 2^{l-1}$$

The idea is to shift the sequence and then combine it with the original sequence by exclusive or.

As part of the S-Plus generator, they use the following shift generator

```
double ush(j)
  unsigned long *j;
{
  double v = 4294967296;  /* v =
2^32 */
  *j = *j ^ (*j >> 15);
  *j = *j ^ (*j << 17);
  return(*j/v);
}
```

`*j >> 15` shifts the bits right by 15 and replaces bits 1 to 15 with 0. ^ is exclusive or so `*j = *j ^ (*j >> 15)` replaces the vector j with

$$\left(b_{32},\ldots,b_{18},b_{17}+b_{32},\ldots,b_1+b_{16}\right) \bmod 2$$

The S-Plus generator combines this with a congruential generator with $a = 69069$, $m = 2^{32}$ with an exclusive or operation.

R has 6 different uniform generators. The default is the Mersenne Twister, a generalized feedback shift register (GFSR) generator with a period of $2^{19937} - 1 \approx 10^{6000}$. To see the others available in R, see `help(RNGkind)`.

Bottom line: Creating a good generator is an art and a science.

The constants used in congruential generators and the lags used in Fibonacci and Tausworthe generators are not arbitrary. Poor choices can lead to very nonrandom behaviour (such as with RANDU).

Diehard tests

A set of procedures for testing random number generators created by George Marsaglia.

Generating from non-uniform distributions

For a cumulative distribution function (CDF) $P\left[X \le x\right] = F\left(x\right)$, the inverse CDF is defined by

$$F^{-1}\left(u\right) = \inf\left\{x : F\left(x\right) \le u\right\}$$

For continuous RVs,

$$\begin{aligned} P\left[F\left(X\right) \le u\right] &= P\left[X \le F^{-1}\left(u\right)\right] \\ &= F\left(F^{-1}\left(u\right)\right) = u \end{aligned}$$

so $F\left(X\right) \sim U\left(0,1\right)$. Conversely, if $U \sim U\left(0,1\right)$

$$P\left[F^{-1}\left(u\right) \le x\right] = F\left(x\right)$$

Thus, given an iid $U\left(0,1\right)$ sample $\left\{u_1,\ldots,u_n\right\}$, an iid sample $\left\{x_1,\ldots,x_n\right\}$ from F can be obtained by $x_i = F^{-1}\left(u_i\right)$.

Example: Cauchy

$$F\left(x;\mu,\sigma\right) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\sigma}\right)$$

$$F^{-1}\left(u;\mu,\sigma\right) = \mu + \sigma\tan\left(\pi\left(u - 1/2\right)\right)$$

Example: Exponential

$$F(x;\mu) = 1 - \exp(-x/\mu)$$
$$F^{-1}(u;\mu) = -\mu \log(1-u)$$

Sometimes its easier to work with the survivor function $S(x) = 1 - F(x)$. Since U and $1 - U$ both have uniform distributions, $S^{-1}(u)$ will also be a draw from F.

So

$$S^{-1}(u;\mu) = -\mu \log(u)$$

will also give a draw from an exponential distribution.

Not all distributions (e.g. normal or gamma) have nice closed form expressions for F^{-1}.

The density is usually of a nice form, but often the CDF, and thus its inverse aren't. However there are often good analytical approximations to F^{-1}, so these can be used instead.

For example with the standard normal, a rational function approximation could be used (R and Matlab definitely do, S-Plus probably).

Note that the Inverse CDF method isn't commonly used for most distributions as it tends to be slow.

Functions like log, sin, cos, etc tend to be somewhat expensive to calculate.

Though surprisingly in R, it is the default for normals. However there are 4 other methods available (see `help(RNGkind)`). In S-Plus and Matlab, I don't know what they are doing.

Discrete Distributions:

Suppose that the distribution has support points $s_1, s_2, \ldots, s_k$ (k possibly infinite) and set

$$p_j = \sum_{i=1}^{j} P\left[X = s_i\right] = P\left[X \le s_j\right]$$

Then independent observations x_i can be generated by setting $x_i = s_j$ if $p_{j-1} < u_i \le p_j$ (where $p_0 = 0$).

Essentially this is inverting the CDF.

If k is small, then only a few comparisons need to be made. However if k is big, many comparisons may be needed (if u_i is close to 1).

In this case, other methods are needed.

Relationships with other distributions:

Examples:

- $X \sim N(\mu, \sigma^2)$ then $Y = e^X$ is lognormal
- $X \sim N(0,1)$ then $Y = X^2$ is χ_1^2
- $X_\alpha \sim Gamma(1, \alpha)$, $X_\beta \sim Gamma(1, \beta)$, then $Y = \dfrac{X_\alpha}{X_\alpha + X_\beta} \sim Beta(\alpha, \beta)$

Polar Coordinates and the Normal Distribution

Suppose that X, Y are independent $N(0,1)$ variables and consider the polar coordinates transformation given by

$$X = R\cos\theta, Y = R\sin\theta; \quad R \geq 0, 0 \leq \theta < 2\pi$$

It is easily shown that $\theta \sim U(0, 2\pi)$, $R^2 \sim \chi^2_2$, and they are independent. Also

$$P[R > r] = \left[R^2 > r^2\right] = \exp\left(-r^2/2\right)$$

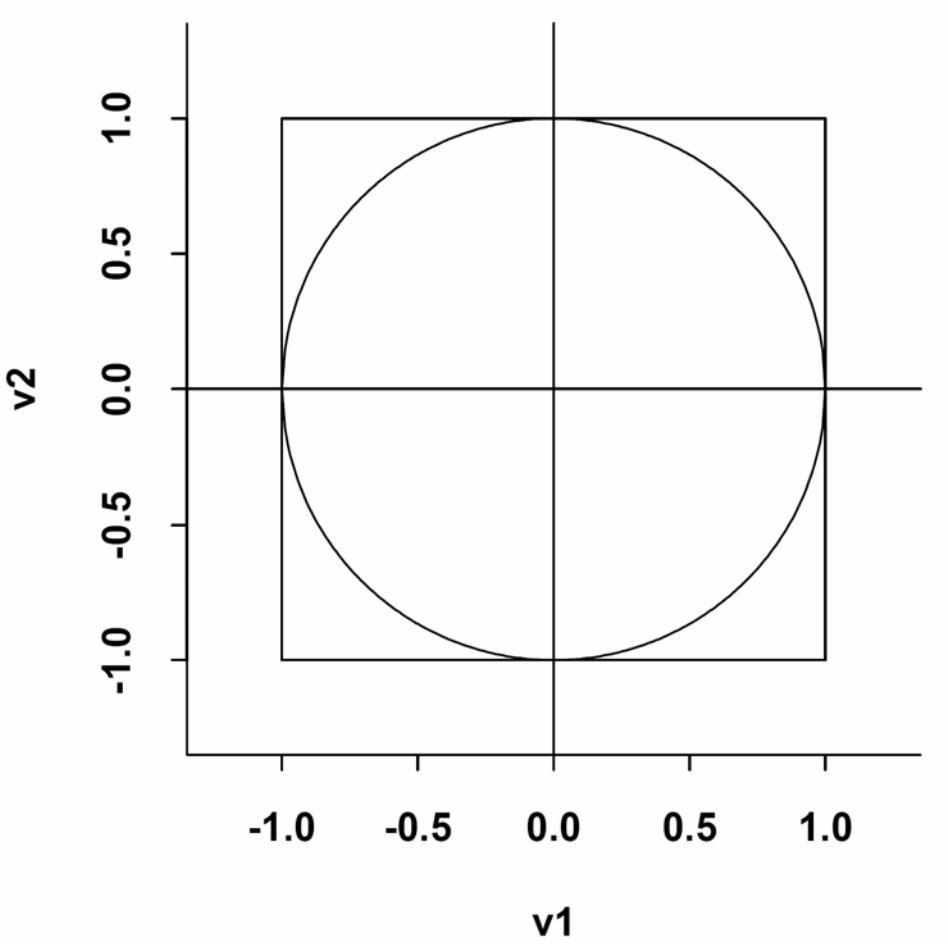
Box-Muller Method

- Generate $u_1, u_2 \sim U(0,1)$
- Set $R = \sqrt{-2\log u_1}$ and $\theta = 2\pi u_2$
- Set $x_1 = R\cos\theta$ and $x_2 = R\sin\theta$

Marsaglia Polar Method

The sin and cos functions (which are slow) can be avoided.

Underlying the Box-Muller method is to pick uniform angle independently of a radius. This can be recast in terms of picking points in the unit circle



Let $v_1, v_2 \sim U(-1,1)$ such that $v_1^2 + v_2^2 \le 1$

Let θ be the counterclockwise angle from the positive v_1 axis to the point (v_1, v_2). By symmetry, $\theta \sim U(0, 2\pi)$ and

$$\cos\theta = \frac{u_1}{\sqrt{u_1^2 + u_2^2}} \text{ and } \sin\theta = \frac{u_2}{\sqrt{u_1^2 + u_2^2}}$$

Also $P\left[v_1^2 + v_2^2 \le u\right] = \pi u/\pi = u$, so $v_1^2 + v_2^2 \sim U(0,1)$

Finally θ and $v_1^2 + v_2^2$ are independent (again by symmetry).

Thus, given (v_1, v_2), two normal deviates are given by

$$u = v_1^2 + v_2^2,\ w = \sqrt{-2\log u / u},$$

$$x_1 = v_1 w,\ x_2 = v_2 w$$

This is an example of the Acceptance-Rejection Method.

For this example, the fraction of pairs (v_1, v_2) that are accepted = $\pi/4$ = 0.785, the ratio of the area of the circle to the square.

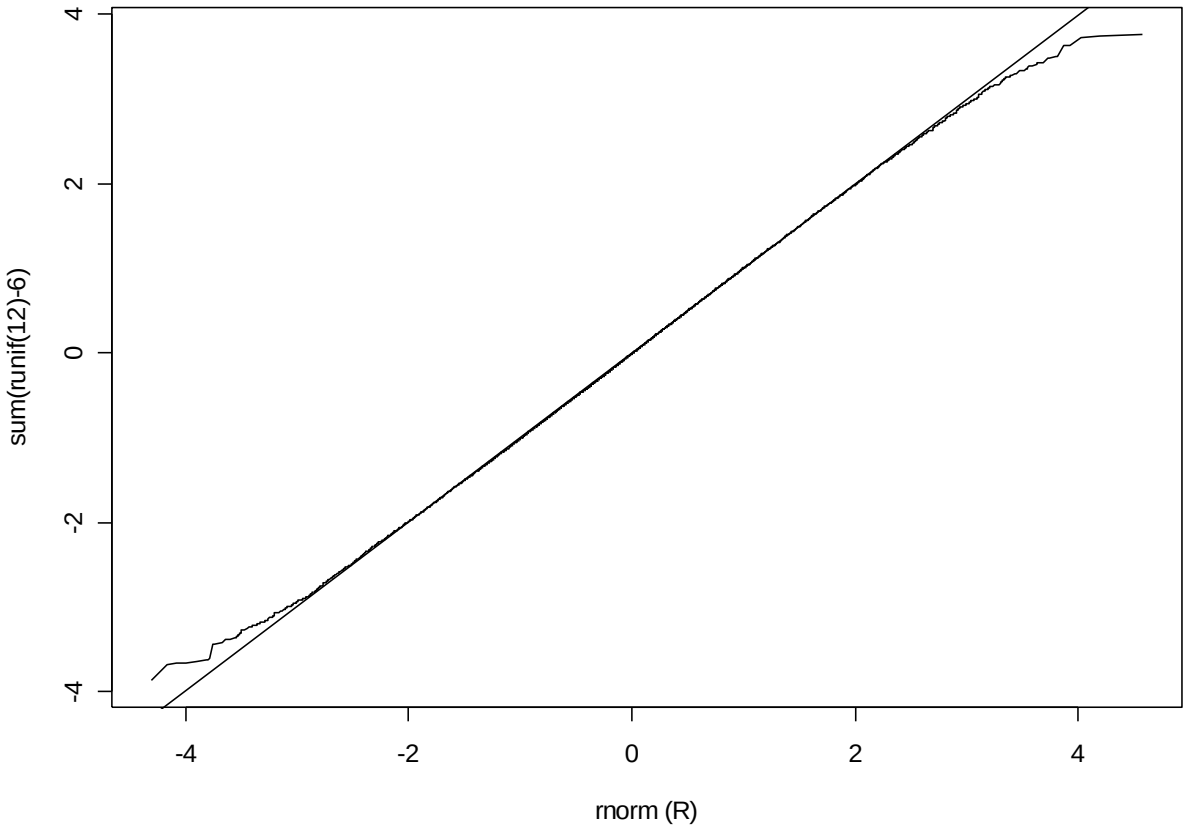
An old generator $N(0,1)$

Let $u_1, \ldots, u_{12} \sim U(0,1)$ and $z = \sum u_i - 6$

$$E[u_i] = \frac{1}{2}; \mathrm{Var}(u_i) = \frac{1}{12}$$

$$E\left[\sum u_i\right] = 6; \mathrm{Var}\left(\sum u_i\right) = 1$$

So z has mean 0 and standard deviation 1 and is approximately normal (very approximately) by the “Central Limit Theorem”



This generator has tails which are too light.

An example of a consequence of this would be to incorrectly estimate the coverage rate of a confidence interval procedure.

Generating Random Deviates

Often there are no direct ways of sampling from a desired distribution (e.g. inverse cdf or relationship with other distributions).

So we need other approaches to generation for other distributions.

Acceptance-Rejection (von Neumann, 1951)

Want to simulate from a distribution with density $f(x)$.

Need to find a "dominating" or majorizing distribution $g(x)$ where g is easy to sample from and

$$f(x) \le cg(x) = h(x)$$

for all x and some constant $c > 1$.

Sampling scheme

1) Sample x from $g(x)$ and compute the ratio

$$r(x) = \frac{f(x)}{cg(x)} = \frac{f(x)}{h(x)} \leq 1$$

2) Sample $u \sim U(0,1)$

If $u \leq r(x)$ accept and return x

If $u > r(x)$ reject and go back to 1)

Note that step 2) is equivalent to flipping a biased coin with success probability r.

The resultant sample is a draw from $f(x)$.

Proof:

Let I be the indicator of whether a sample x is accepted. Then

$$P[I=1] = \int P[I=1 | X=x] g(x) dx$$

$$= \int \frac{f(x)}{cg(x)} g(x) dx = \frac{1}{c}$$

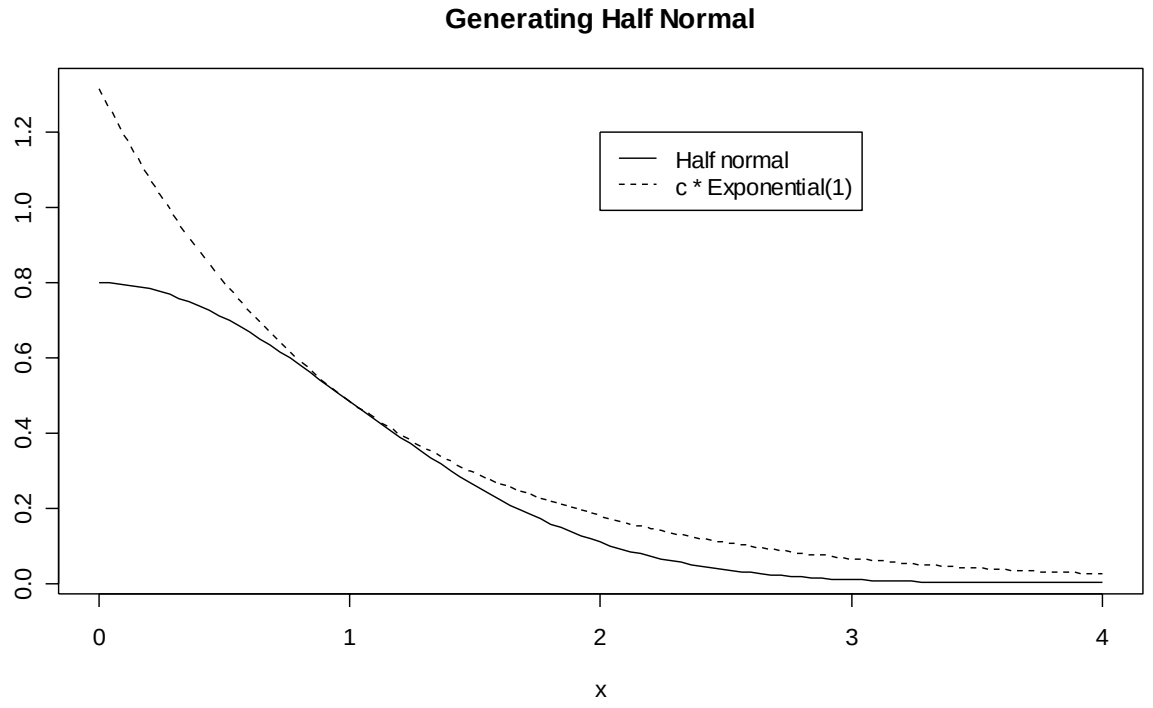
Next,

$$p\left(x\middle|I=1\right)=\frac{f\left(x\right)}{cg\left(x\right)}g\left(x\right)\Big/P\left[I=1\right]$$

$$=\frac{cf\left(x\right)}{c}=f\left(x\right)$$

See Flury (1990) for a more geometrical proof.

Its based on the idea of drawing uniform points $\left(x,y\right)$ under the curve $h\left(x\right)$ and only accepting the points that also lie under the curve $f\left(x\right)$.



The number of draws needed until an acceptance occurs is Geometric$(1/c)$ and thus the expected number of draws until a sample is accepted is c.

The acceptance probability satisfies

$$\frac{1}{c} = \frac{\int f(x)\,dx}{\int cg(x)\,dx} = \frac{\text{Area under } f(x)}{\text{Area under } h(x)}$$

One consequence of this is that c should be made as small as possible to minimize the number rejections.

The optimal c is given by

$$c = \sup \frac{f(x)}{g(x)}$$

Note that the best c need not be determined, just one that satisfies

$$f(x) \le cg(x) = h(x)$$

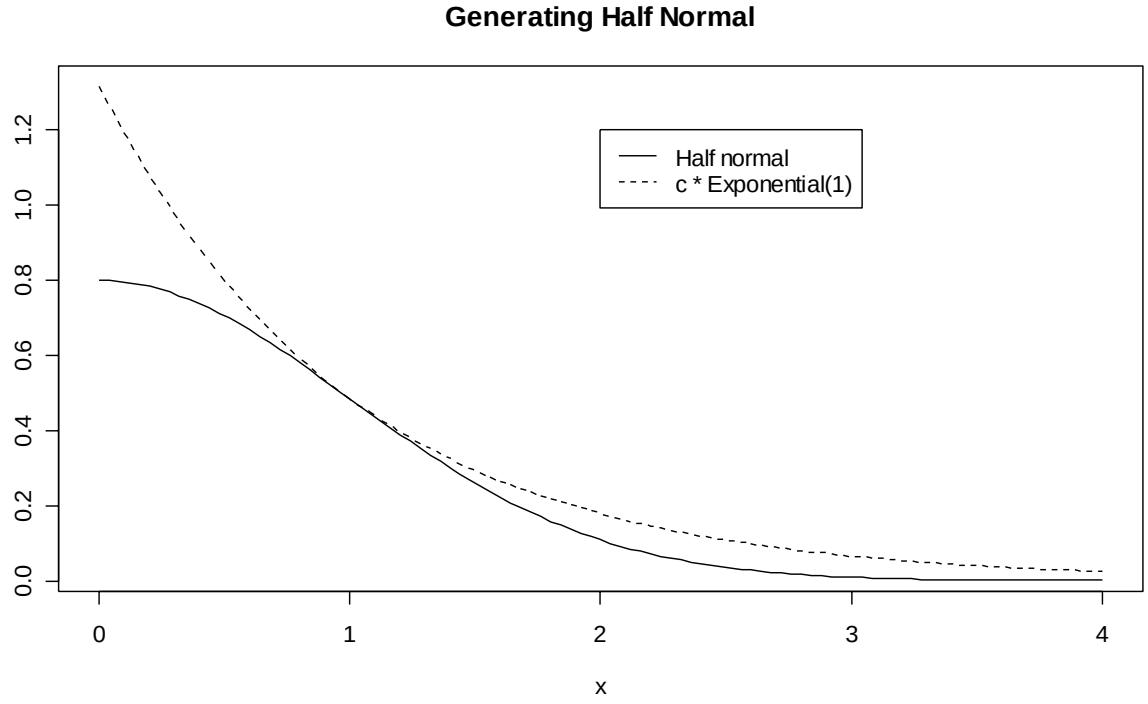
for all x.

Example: Generating from the half normal distribution.

$$f(x) = 2\phi(x) I(x \geq 0)$$
$$= \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) I(x \geq 0)$$

Lets use an Exp(1) as the dominating density

$$g(x) = e^{-x} I(x \geq 0)$$



The optimal c is

$$c = \sqrt{\frac{2}{\pi}} \exp(1/2) \approx 1.315489$$

Thus the acceptance –rejection scheme is

1) Draw $x \sim \text{Exp}(1)$

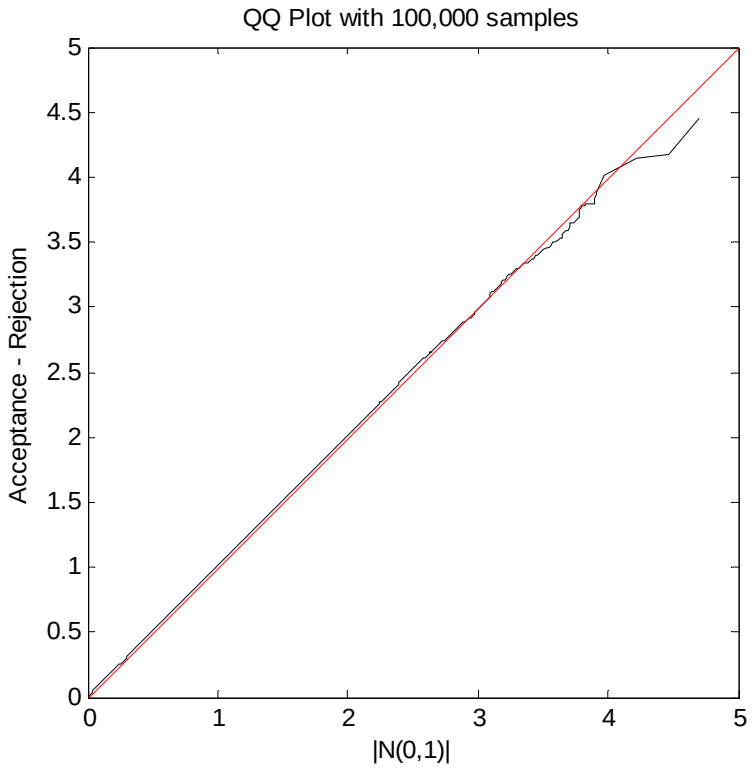
$$r(x) = \exp\left(-0.5(x-1)^2\right)$$

2) Draw $u \sim U(0,1)$

If $u \leq r(x)$ accept and return x

If $u > r(x)$ reject and go back to 1)

Note that this scheme isn't needed for this example as the half normal distribution is the distribution of the absolute value of $N(0,1)$



In the above is was assumed that $f(x)$ was a density function. In fact $f(x)$ only needs to be known up to a multiplicative constant

$$l(x) = bf(x)$$

where b may be unknown.

One place where this is useful is with posterior distributions as

$$p(x|y) \propto \pi(x) f(y|x)$$

The normalizing constant may be difficult to determine exactly.

However it is not necessary to do so. Modify the procedure as follows.

Find c such that

$$l(x) \leq cg(x) = h(x)$$

for all x and some constant $c > 1$.

Sampling scheme

1) Sample x from $g(x)$ and compute the ratio

$$r(x) = \frac{l(x)}{cg(x)} = \frac{l(x)}{h(x)} \leq 1$$

2) Sample $u \sim U(0,1)$

If $u \leq r(x)$ accept and return x

If $u > r(x)$ reject and go back to 1)

Do everything the same except use $l(x)$ instead of $f(x)$

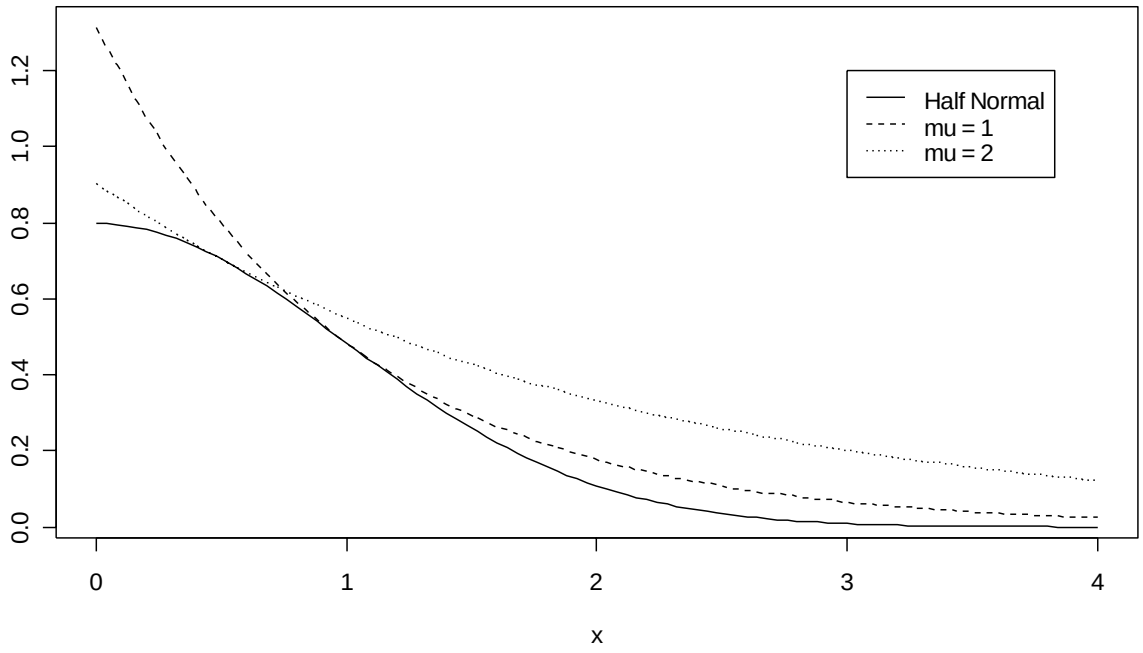
The acceptance probability for this scheme is b/c.

In addition to the constant c chosen, the distribution $g(x)$ will also affect the acceptance rate. (c is chosen conditional on $g(x)$)

A good choice $g(x)$ will normally be "close to" $f(x)$. You want to minimize the separation between the two densities.

Often will look for much member of a parametric family will minimize c.

For example, for the Half normal problem, which $\text{Exp}(\mu)$, will minimize $c(\mu)$.



In fact $\mu = 1$ will minimize $c(\mu)$ for this problem.

Note that so far I'm been appeared to have been focusing on continuous random variables.

In fact acceptance-rejection works fine with discrete random variables and with variables over more than one dimension.

The proof goes through in this more general place by replacing integration over a density to integration over a more general measure.

For discrete random variables, you get a sum over the probability mass function.

With higher dimensional problems, the majorization constants tend to be higher, implying the procedure is less efficient.

Log-concave densities

There is a class of densities where it is easy to set up an acceptance-rejection scheme.

It the case when the log of the density is concave on the support of the distributions

If $f(x)$ is log concave, any tangent line to $\log f(x)$ will lie above $\log f(x)$ (call it $l(x) = a + bx$).

Thus $h(x) = e^{l(x)} = e^a e^{bx}$ lies above $f(x)$.

$h(x)$ looks like a scaled exponential density.

This suggests that exponential distributions can be used as the majorizing distributions.

A strictly log concave density is unimodal.

The mode may occur at either endpoint or in the middle.

If the mode occurs at an endpoint, a single exponential can be used (as with the half normal example)

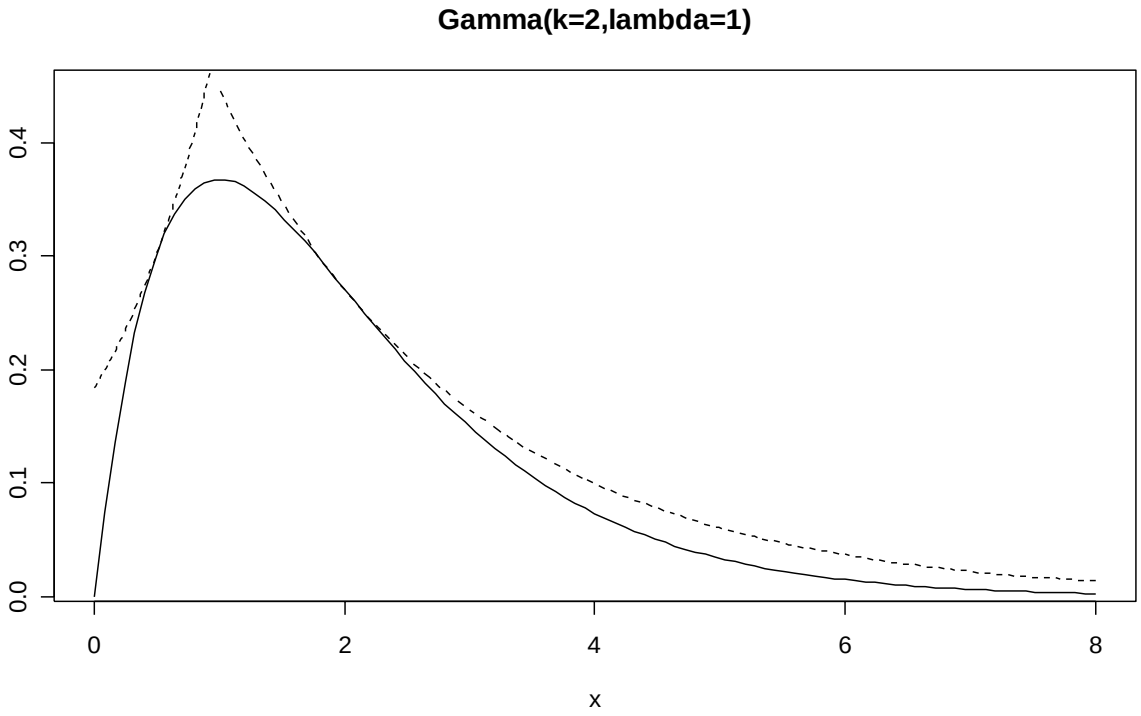
If the mode occurs in the middle of the range, two exponential envelopes are needed (one for left of the mode, the other for the right of the mode)

Example: Gamma $\left(k = 2, \lambda = 1\right)$

The mode for a Gamma is $\left(k-1\right)\lambda$ (so 1 for this example)

Left side: $g_l\left(x\right) = \frac{1}{\mu_l}\exp\left(\left(x-1\right)/\mu_l\right) I\left(x < 1\right)$

Right side: $g_r\left(x\right) = \frac{1}{\mu_r}\exp\left(-\left(x-1\right)/\mu_r\right) I\left(x \geq 1\right)$



The choice of μ_l and μ_r depend on where you want the majorized densities $g_l(x)$ and $g_r(x)$ to be tangent to $f(x)$

In the above the tangent points are $x_l = 0.5$ and $x_r = 2$.

The total area under $h(x) = c_l g_l(x) + c_r g_r(x)$ is $c = c_l + c_r$

For the example, $c_l = 0.5$ and $c_r = 0.8925206$, so the rejection rate for this sampler is just under 30%.

To determine which exponentials to use, involves solving the systems (for given x_l and x_r)

$$f(x_l) = c_l g_l(x_l) \qquad f(x_r) = c_r g_r(x_r)$$
$$f'(x_l) = c_l g_l'(x_l) \qquad f'(x_r) = c_r g_r'(x_r)$$

Solving gives

$$\lambda_l = \frac{f'(x_l)}{f(x_l)} \qquad \lambda_r = -\frac{f'(x_r)}{f(x_r)}$$

$$c_l = \frac{f(x_l)^2}{f'(x_l)} e^{-\lambda_l (x_l - m)} \qquad c_r = -\frac{f(x_r)^2}{f'(x_r)} e^{\lambda_r (x_r - m)}$$

where $\lambda_i = 1/\mu_i$

The optimal choices for x_l and x_r can be found by minimizing c_l and c_r separately.

Discrete log concave distributions

Random variable defined on non-negative integers

Log concave defined as

$$\log f(x) \geq \frac{1}{2}\left[\log f(x-1) + \log f(x+1)\right]$$

which is equivalent to

$$f(x)^2 \geq f(x-1)\, f(x+1)$$

for all integers x.

A possible majorizing distribution in the discrete case is the geometric distribution

$$P[X = x] = p(1-p)^x; \quad x = 0, 1, 2, \ldots$$

See Lange for choices of p_l, p_r, x_l, x_r

Ratio Method

This is another method that is useful when the distribution you are interested in $f(x)$, is only known up to an unknown constant, $h(x) = cf(x)$

Define

$$S_h = \left\{(u,v) : 0 < u \le \sqrt{h(v/u)}\right\}$$

If this set is bounded, we can draw uniform points from this set to generate *X*.

Proposition 20.5.1

Suppose

$$k_u = \sup_x \sqrt{h(x)}$$

and

$$k_v = \sup_x |x| \sqrt{h(x)}$$

are both finite. Then the rectangle $[0, k_u] \times [-k_v, k_v]$ encloses S_h.

If $h(x) = 0$ for $x < 0$, then the rectangle $[0, k_u] \times [0, k_v]$ encloses S_h.

If the point (U,V) sampled uniformly from the enclosing set falls with in S_h, then the ratio $X = V/U$ is distributed according to $f(x)$.

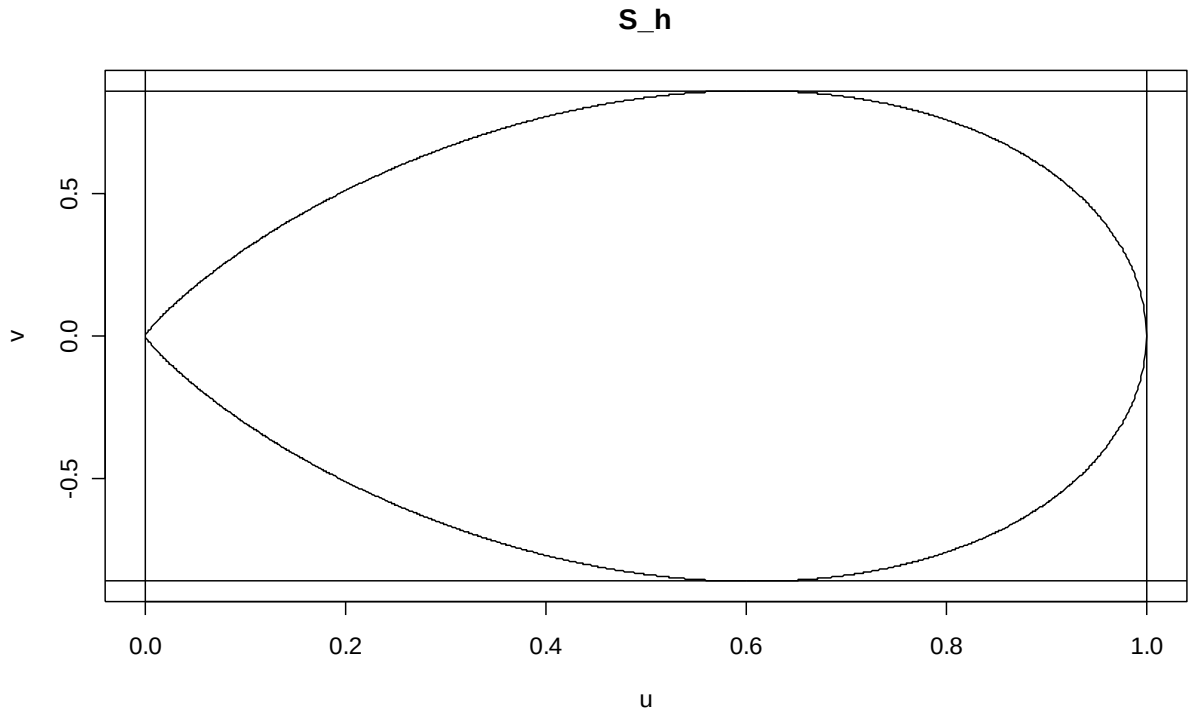
Example: Standard normal

$$h(x) = \exp\left(-\frac{1}{2}x^2\right)$$

$$\sqrt{h(x)} = \exp\left(-\frac{1}{4}x^2\right)$$

$$k_u = \sqrt{h(0)} = 1$$

$$k_v = \sqrt{2}\exp(-0.5)$$



Generate

$$U \sim U(0,1)$$

$$V \sim U\left(-\sqrt{2}\exp(-0.5), \sqrt{2}\exp(-0.5)\right)$$

Accept $X = V/U$ if

$$U \leq \exp\left(-0.25 V^2/U^2\right)$$

The acceptance rate for this procedure is about 62.6%