Monte Carlo

Rob McCulloch

1. Monte Carlo
2. Transformations
3. Rejection Sampling
4. Importance Sampling
5. SIR: Sampling Importance Resampling

1. Monte Carlo

Often our goal is to understand the distribution of a random variable X (scalar or vector) and/or compute some expectation $E(h(X))$ of some function of X.

If X is known to come from a “nice” family (e.g multivariate normal) and h is linear we can do a lot analytically.

As soon as X is complicated and h is non-linear, we cannot.

To understand X we get draws $X_i,\ i = 1, 2, \ldots, m$ from it's distribution and look at the draws.

We can use monte carlo to estimate an expectation:

$$\mu \equiv E(h(X)) \approx \frac{1}{m} \sum_{i=1}^{m} h(X_i) \equiv \hat{\mu}.$$

In basic Monte Carlo, the draws X_i are iid.

When the draws are iid then $h(X_i)$ are iid so basic law of large number and central limit theorem ideas apply.

$$E(\hat{\mu}) = \mu, \quad Var(\hat{\mu}) = \frac{1}{n} Var(h(X)).$$

$$\hat{\mu} \Rightarrow \mu.$$

$$Var(h(X)) \approx \frac{1}{n} \sum (h(X_i) - \hat{\mu})^2 \equiv s_h^2.$$

So, by the central limit theorem,

$$\hat{\mu} \approx N(\mu, s_h^2/n).$$

Giving monte carlo error:

$$\pm z_{\alpha/2} \frac{s_h}{\sqrt{n}}.$$

E.g. $z_{\alpha/2} = 2.0$.

Monte Carlo, can mean a lot of things.

Two examples of fundamental tools are:

- **rejection sampling**: you are able to compute the density of the distribution $f(x)$ and you want draws from the distribution.
- **importance sample**: you can get draws from a distribution *similar* to that of X and you can reweight these draws to estimate an expectation.

In general, we need tools for drawing from a distribution, given limited knowledge (typically the density) of the distribution.

We will assume that we can get iid draws from the uniform and then learn ways to get iid draws from other distributions given iid uniform draws.

Of course, getting iid uniforms from a computer is a major issue but we will assume we “have a good random number generator”.

Note that when we say *Monte Carlo* we usually mean we are getting iid draws.

Later we will look at *Markov Chain Monte Carlo* where the draws are dependent.

Example

In Bayesian statistics, we can often evaluate:

$$p(\theta \mid y) \propto f(y \mid \theta)\, p(\theta) = L(\theta)\, p(\theta)$$

where p is the prior and L is the likelihood function.

But, we don't know what "kind" of random variable $\theta \mid y$ is and we cannot compute

$$\int L(\theta)\, p(\theta)\, d\theta$$

the normalizing constant.

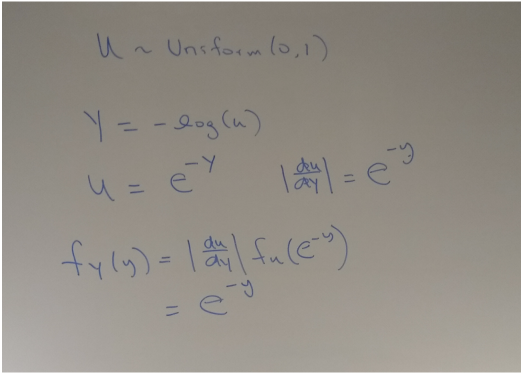
2. Transformations

The *transformation approach* looks for a transformation (function) g such that

$$X = g(U_1, U_2, \ldots, U_p),\ U_p \sim \mathit{uniform}$$

has the desired distributon.

Drawing an Exponential



$U \sim \text{Uniform}(0,1)$

$Y = -\log(u)$

$u = e^{-Y} \qquad \left|\frac{du}{dY}\right| = e^{-y}$

$$f_Y(y) = \left|\frac{du}{dy}\right| f_u(e^{-y})$$
$$= e^{-y}$$

So, to draw and exponential we can draw a uniform and then just compute -log.
We can then rescale it to get a general exponential.

Drawing Gammas and chi-squared

Let $X_j \sim Exp(1)$.

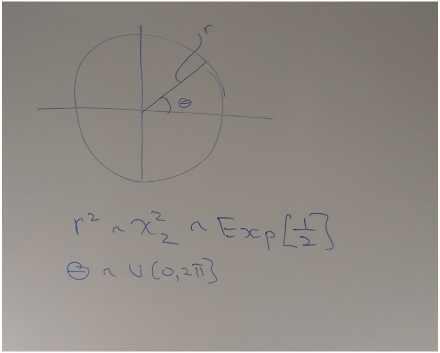
Then

$$Y = 2\sum_{i=1}^{\nu} X_j, \sim \chi^2_{2\nu}$$

$$Y = \beta \sum_{i=1}^{\alpha} X_j, \sim Gamma(\alpha, \beta)$$

For integer α we have a simple way to draw a Gamma.

Drawing a Normal (Box Muller Algorithm)



$r^2 \sim \chi^2_2 \sim Exp[\frac{1}{2}]$

$\Theta \sim U(0, 2\pi]$

Draw U_1 and U_2 uniform.

$$X_1 = \sqrt{-2log(U_1)}\, cos(2\pi U_2), \;\; X_2 = \sqrt{-2log(U_1)}\, sin(2\pi U_2)$$

Then X_1, X_2 are iid N(0,1).

Inverse CDF

A general transformation approach is to use the inverse-CDF.

Suppose you want to draw X such that

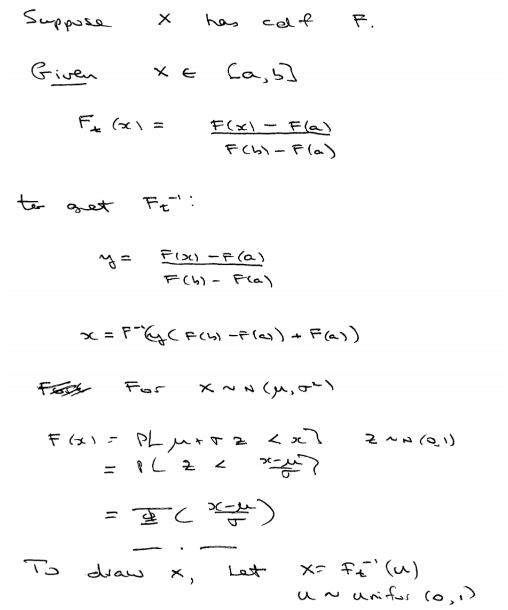
$$F(x) = P(X \leq x)$$

Let $\tilde{X} = F^{-1}(U)$, $U \sim U(0,1)$.

Then,

$$P(\tilde{X} \leq c) = P(F^{-1}(U) \leq c) = P(U \leq F(c)) = F(c).$$

Example: Drawing from a truncated distribution



Suppose X has cdf F.

Given $x \in [a,b]$

$$F_t(x) = \frac{F(x) - F(a)}{F(b) - F(a)}$$

to get F_t^{-1}:

$$y = \frac{F(x) - F(a)}{F(b) - F(a)}$$

$$x = F^{-1}(y(F(b) - F(a)) + F(a))$$

~~For~~ For $X \sim N(\mu, \sigma^2)$

$$F(x) = P[\mu + \sigma z < x] \qquad z \sim N(0,1)$$

$$= P[z < \frac{x-\mu}{\sigma}]$$

$$= \Phi(\frac{x-\mu}{\sigma})$$

— . —

To draw x, Let $X = F_t^{-1}(u)$
$u \sim$ unifor $(0,1)$

3. Rejection Sampling

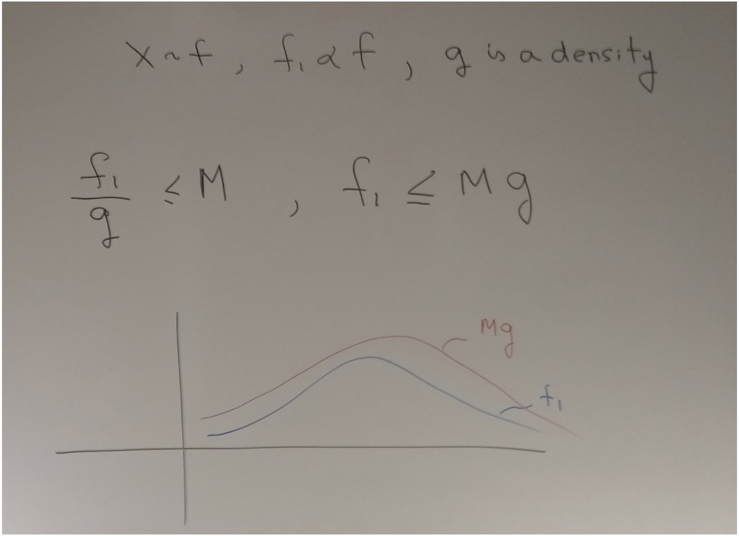
Very often what we do is get draws from a distribution that approximates the distribution we want to draw from and then modify these draws to get what we want.

Rejection sampling is a fundamental method.

Let f be density we want to draw from and suppose we can compute $f_1 \propto f$.

The density g will be our approximation and we assume we can draw from g.

We need to be able to multiply g by some M so that $f_1 \leq M\,g$:



$X \sim f$, $f_1 \propto f$, g is a density

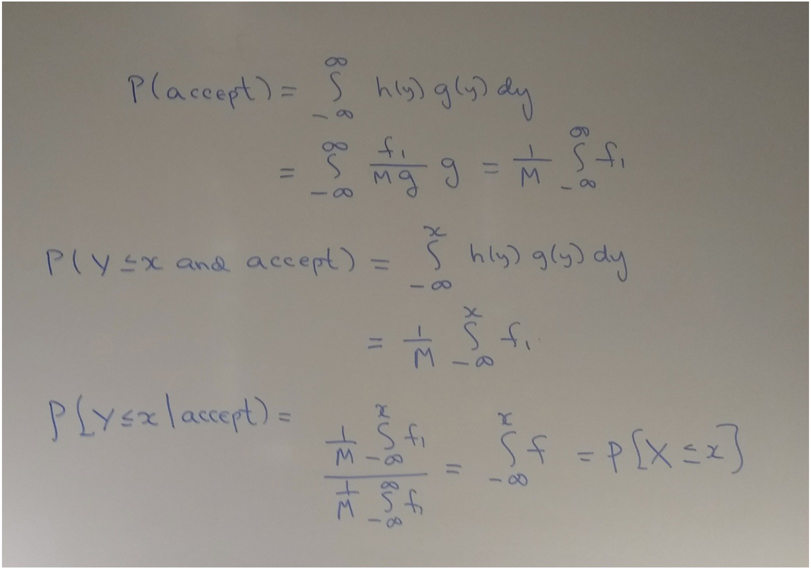
$\frac{f_1}{g} \leq M$, $f_1 \leq Mg$

Rejection Sampling Algorithm

- (i) Draw $Y \sim g$.
- (ii) with probability $h(y) = \frac{f_1(y)}{M\,g(y)}$ let $X = Y$.
- (iii) with probability $1 - h(y)$ return to (ii).

That is, you keep sampling Y from g until it is accepted.

It works!!



$$P(\text{accept}) = \int_{-\infty}^{\infty} h(y)\, g(y)\, dy$$

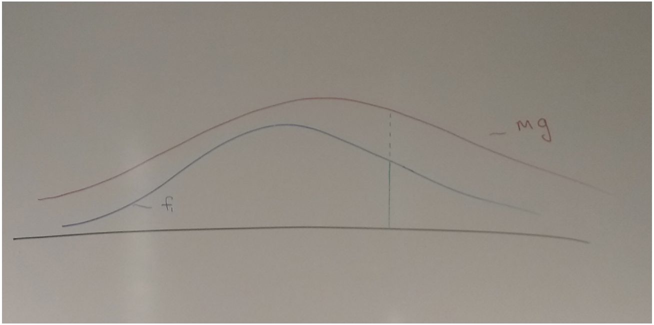
$$= \int_{-\infty}^{\infty} \frac{f_1}{Mg}\, g = \frac{1}{M} \int_{-\infty}^{\infty} f_1$$

$$P(Y \le x \text{ and accept}) = \int_{-\infty}^{x} h(y)\, g(y)\, dy$$

$$= \frac{1}{M} \int_{-\infty}^{x} f_1$$

$$P\{Y \le x \,|\, \text{accept}) = \frac{\frac{1}{M} \int_{-\infty}^{x} f_1}{\frac{1}{M} \int_{-\infty}^{\infty} f_1} = \int_{-\infty}^{x} f = P\{X \le x\}$$

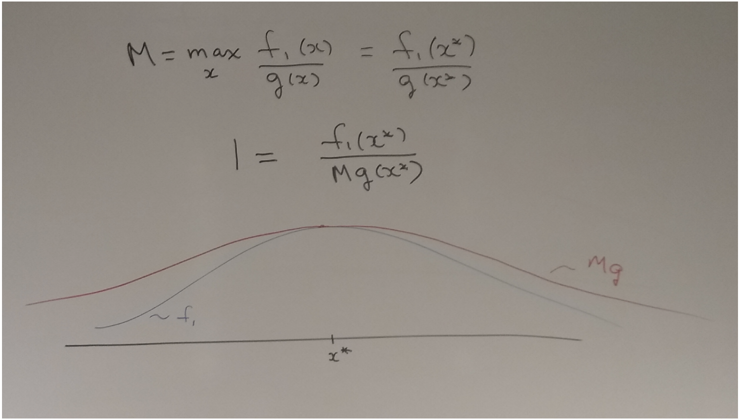
At a drawn y, the probability to accept is $h(y) = \frac{f_1(y)}{M\, g(y)}$.



Equivalently, you accept if

$$U\, M\, g(y) \leq f_1(y), \quad U \sim \mathit{Unif}(0,1).$$

May be efficient to make g fit f_1 snugly by letting $M = \max f_1/g$.



$$M = \max_{x} \frac{f_1(x)}{g(x)} = \frac{f_1(x^*)}{g(x^*)}$$

$$1 = \frac{f_1(x^*)}{Mg(x^*)}$$

Truncated Normal with Rejection Sampling

The inverse CDF approach to drawing a truncated normal is simple and works.

But, computing the inverse CDF is costly (a complex series expansion).

It can be more efficient to use rejection sampling.

We can restrict attention to the case:

$$Z \sim N(0,1), \text{ restricted to } Z \geq a.$$

If we wanted $X \leq a$, then we can flip it using $-X \geq -a$.

If $X \sim N(\mu, \sigma^2)$, then

$$X \geq a \Leftrightarrow \mu + \sigma Z \geq a \Leftrightarrow Z \geq \frac{(a-\mu)}{\sigma}.$$

If a is not far out in the right tail, then you can just draw from a $N(0,1)$ until you get one bigger than a.

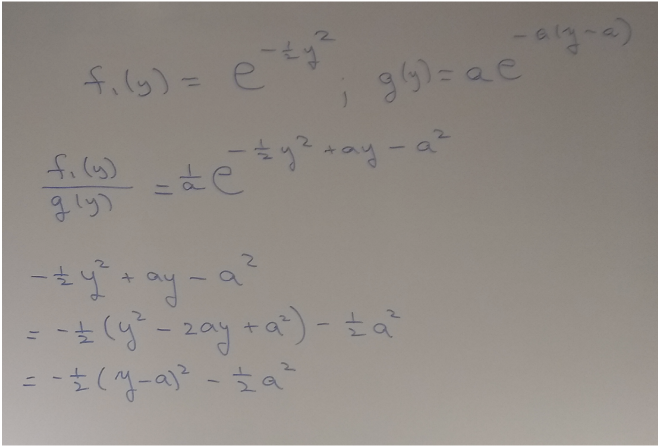
The tricky case is when a is out in the right tail, then you cannot just wait until you get one there.

We use rejection sampling with

$$Y \sim a + Exp(a)$$

so,

$$g(y) = a\, e^{(-a(y-a))}$$



$$f_1(y) = e^{-\frac{1}{2}y^2} \; ; \; g(y) = a e^{-a(y-a)}$$

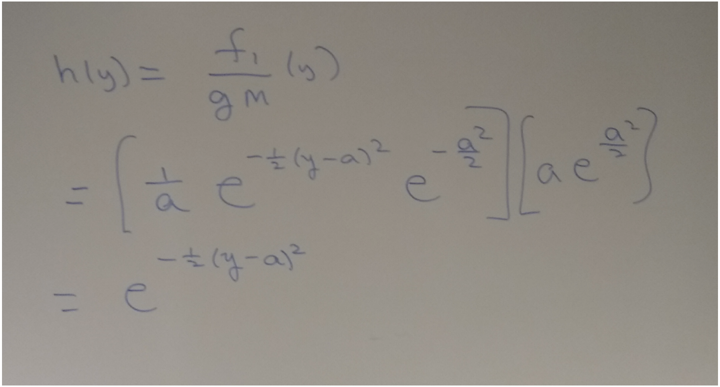
$$\frac{f_1(y)}{g(y)} = \frac{1}{a} e^{-\frac{1}{2}y^2 + ay - a^2}$$

$$-\frac{1}{2}y^2 + ay - a^2$$

$$= -\frac{1}{2}(y^2 - 2ay + a^2) - \frac{1}{2}a^2$$

$$= -\frac{1}{2}(y-a)^2 - \frac{1}{2}a^2$$

So, f_1/g has the optimal value $M = \frac{1}{a}\, e^{-a^2/2}$ at $y = a$.



$$h(y) = \frac{f_1}{gM}(y)$$

$$= \left[\frac{1}{a}\, e^{-\frac{1}{2}(y-a)^2}\, e^{-\frac{a^2}{2}}\right]\left[a\, e^{\frac{a^2}{2}}\right]$$

$$= e^{-\frac{1}{2}(y-a)^2}$$

- Draw $Y \sim Exp(a)$.
- accept with probability $e^{-(1/2)y^2}$.

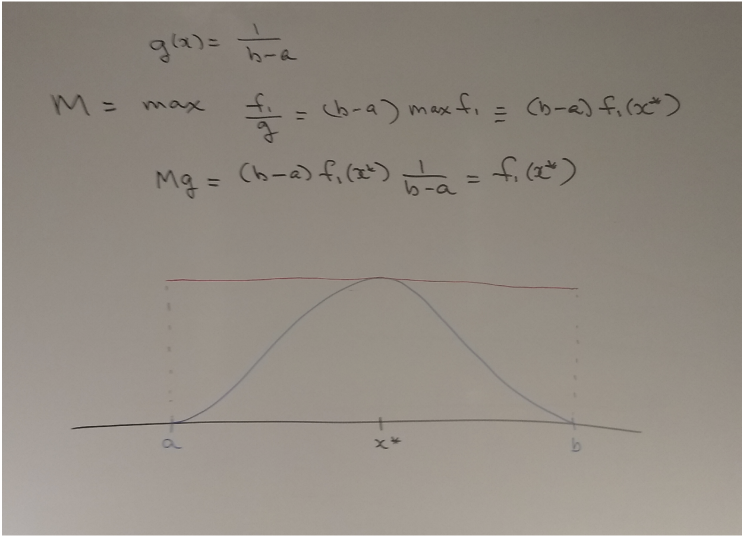
Inutition for Rejection Sampling

The math for rejection sampling was pretty easy.

But, we can make this algorithm more intuitive.

First we consider a very special case.

Simple but informative special case, $X \in (a, b)$, $g \sim U(a, b)$.



$$g(x) = \frac{1}{b-a}$$

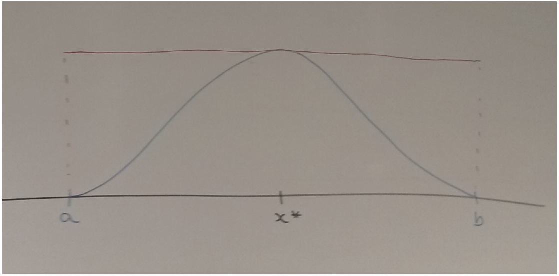
$$M = \max \frac{f_1}{g} = (b-a) \max f_1 \equiv (b-a) f_1(x^*)$$

$$Mg = (b-a) f_1(x^*) \frac{1}{b-a} = f_1(x^*)$$

$h(y) = \frac{f_1(y)}{(Mg(y))} = \frac{f_1(y)}{f_1(x^*)}$.

For $U \sim Unif(0,1)$, accept if $U \leq h(y)$ or $Uf_1(x^*) \leq f_1(y)$.

- Draw $y = U_1 \sim Unif(a,b)$.
- Draw $U_2 \sim Unif(0, f_1(x^*))$.
- Accept U_1 as X if $U_2 \leq f(U_1)$.



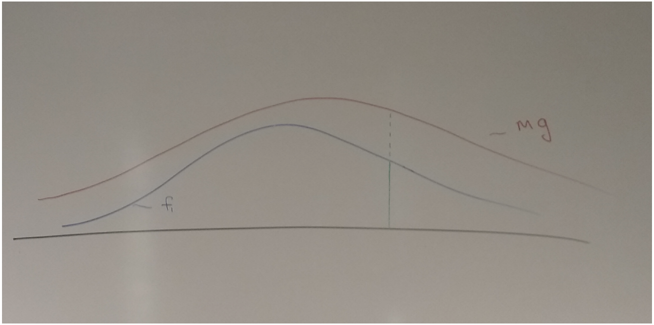
Hit or Miss

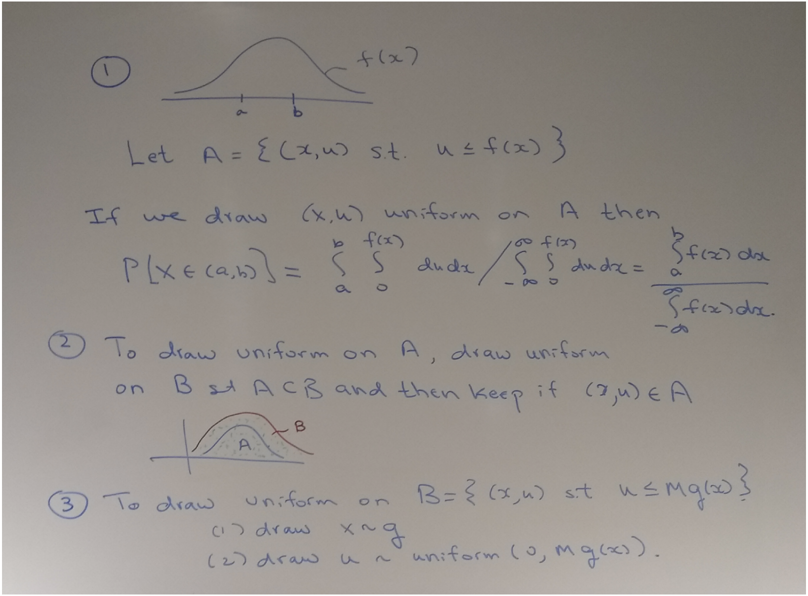
Draw (U_1, U_2) uniformly on $(a,b) \times (0, f_1(x^*))$.
Keep $X = U_1$ if (U_1, U_2) is below f_1.

General Hit or Miss

Draw (U_1, U_2) uniformly on $\{(U_1, U_2)\ s.t.\ U_2 \leq Mg(U_1)\}$.

Keep $X = U_1$ if $U_2 \leq f_1(U_1)$.





①

Let $A = \{(x,u) \text{ s.t. } u \le f(x)\}$

If we draw (x,u) uniform on A then

$$P[X \in (a,b)] = \int_a^b \int_0^{f(x)} du\,dx \Big/ \int_{-\infty}^{\infty} \int_0^{f(x)} du\,dx = \frac{\int_a^b f(x)\,dx}{\int_{-\infty}^{\infty} f(x)\,dx}.$$

② To draw uniform on A, draw uniform on B st $A \subset B$ and then keep if $(x,u) \in A$

③ To draw uniform on $B = \{(x,u) \text{ s.t } u \le Mg(x)\}$

(1) draw $x \sim g$

(2) draw $u \sim$ uniform $(0, Mg(x))$.

Squeezed Rejection Sampling

Sometimes it is costly to evaluate $f_1(y)$.

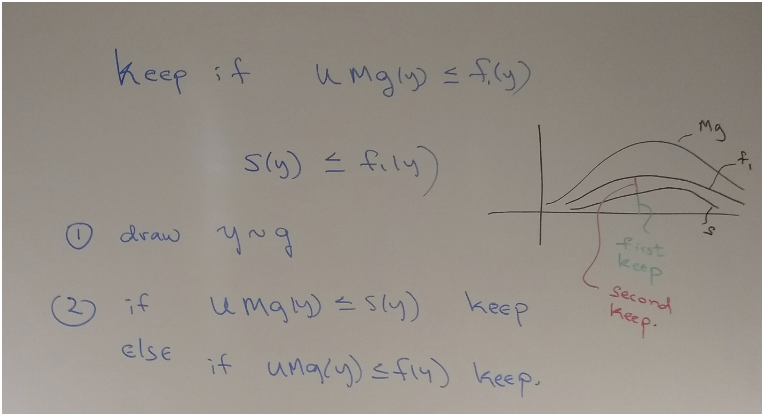
If we can find a function s such that $s(y) \leq f_1(y)$ then we can just check s first.

That is, we accept if

$$UMg(y) \leq f_1(y)$$

But,

$$UMg(y) \leq s(y) \;\Rightarrow\; UMg(y) \leq s(y) \leq f_1(y).$$



Keep if $uMg(y) \le f_1(y)$

$s(y) \le f_1(y)$

① draw $y \sim g$

② if $uMg(y) \le s(y)$ keep

else if $uMg(y) \le f(y)$ keep.

Mg
f_1
s
first keep
second keep.

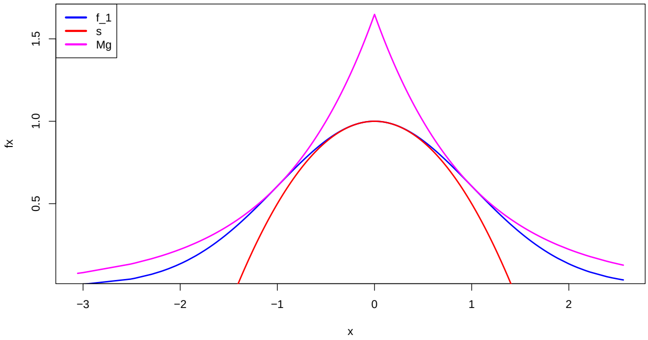
Example

$$f_1(x) = e^{(-x^2/2)}, \;\; g(y) = \frac{1}{2}\, e^{(-|x|)}.$$

g is the *double exponential*.

Then we can use,

$$s(x) = (1 - \frac{x^2}{2}) \leq e^{(-x^2/2)}$$


f_1
s
Mg
fx
0.5
1.0
1.5
x
−3
−2
−1
0
1
2

Adaptive Rejection Sampling

To use rejection sampling you have to come up with a clever envelope g.

Adaptive rejection sampling gives us an *automatic* way of constructing an efficient envelope!!

Everytime you do an accept/reject step you have to do an evaluation of $f(x)$ (or $f_1(x)$).

The idea of *adaptive rejection sampling* is that you can use the information to update an envelope $g(x)$ that you can draw from.

The method assumes that f is log convcave, that is, $log(f(x))$ is concave.

As an example, suppose $X \sim N(0,1)$.

$$f(x) \propto \exp(-\frac{1}{2}\,x^2).$$

and then

$$log(f(x)) = -\frac{1}{2}\,x^2$$

which is concave.

The method constructs a *piecewise linear* $h(x)$ such that

$$log(f(x)) \leq h(x)$$

and then,

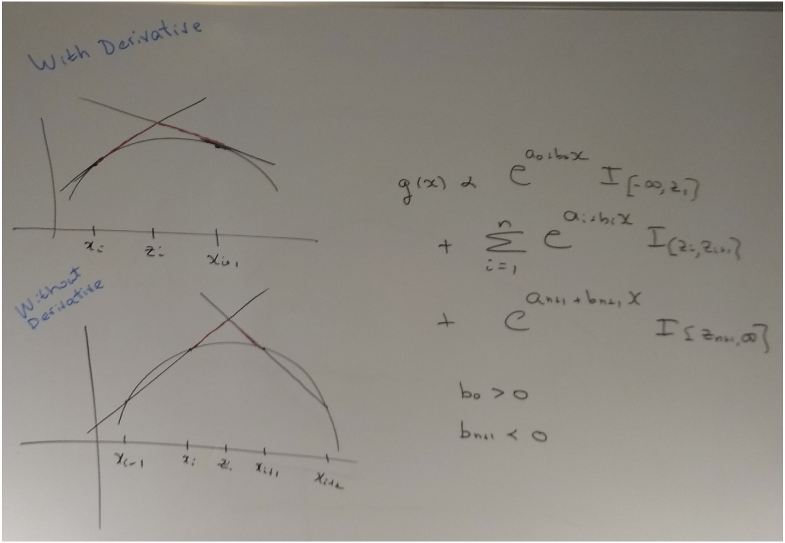
$$f(x) \leq \exp(h(x)) = g(x) \propto \tilde{g}(x).$$

And it turns out we can sample from $\tilde{g}(x)$ fairly easily.

We assume we have a set of function evaluations:

$$\{(x_i, f(x_i)) : x_i < x_{i+1}\}.$$

There are two schemes for constructing the linear envelope for $\log(f(x))$ depending on whether or not the derviative is available.

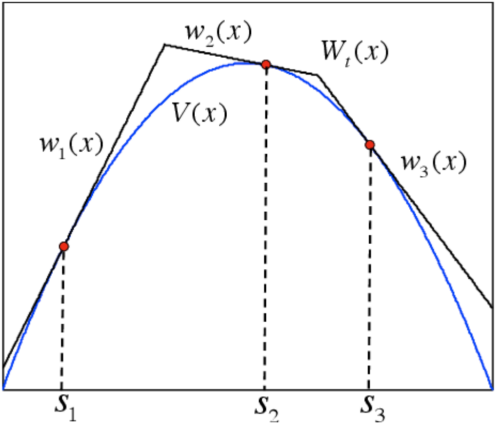


$$g(x) \propto e^{a_0+b_0x} I_{[-\infty, z_1]} + \sum_{i=1}^{n} e^{a_i+b_ix} I_{(z_i, z_{i+1}]} + e^{a_{n+1}+b_{n+1}x} I_{[z_{n+1}, \infty]}$$

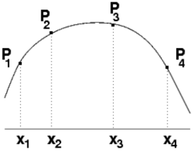
$b_0 > 0$

$b_{n+1} < 0$

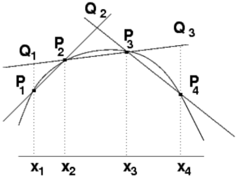
ARS with the derivative.



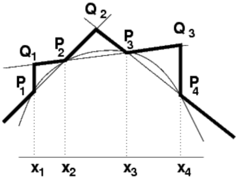
ARS without derivative.



Graph of *log(f)*

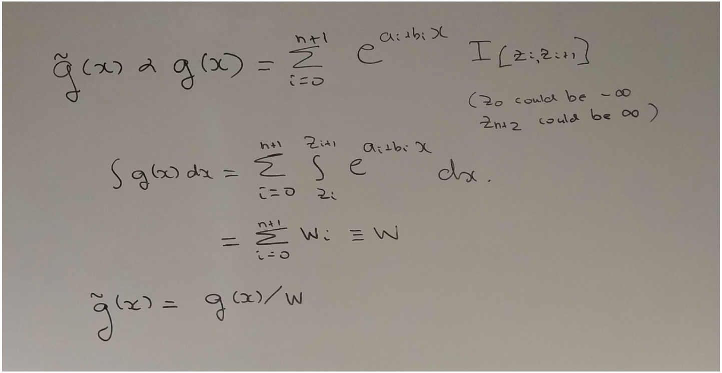


Join with lines



Zigzag crown shown with thick line

With either construction we have:



$$\tilde{g}(x) \propto g(x) = \sum_{i=0}^{n+1} e^{a_i + b_i x} I_{[z_i, z_{i+1}]}$$

(z_0 could be $-\infty$
z_{n+2} could be ∞)

$$\int g(x)\,dx = \sum_{i=0}^{n+1} \int_{z_i}^{z_{i+1}} e^{a_i + b_i x}\,dx.$$

$$= \sum_{i=0}^{n+1} w_i \equiv W$$

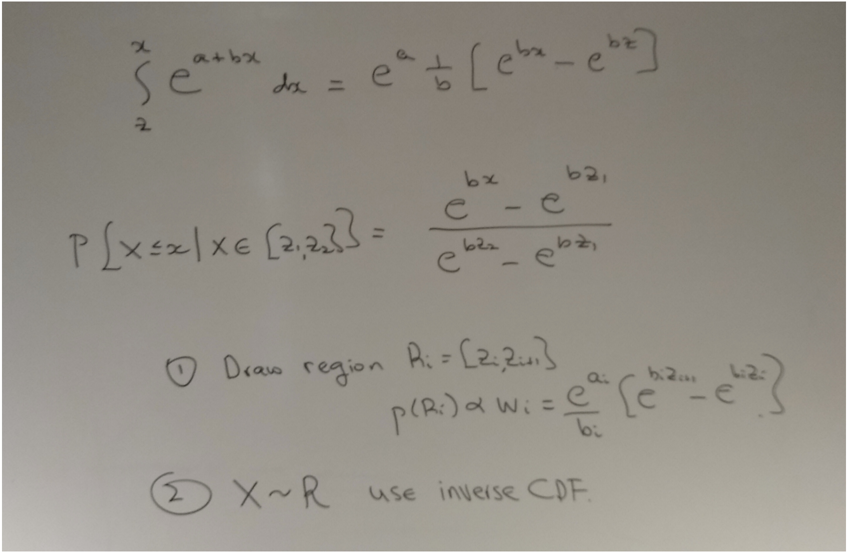
$$\tilde{g}(x) = g(x)/W$$

Let $R_i = [z_i, z_{i+1}]$.

Let $R_i = [z_i, z_{i+1}]$.

We draw from $\tilde{g}$ by

- First drawing the R_i with probability $\frac{w_i}{w}$.
- Then drawing $X \mid X \in R_i$.



$$\int_z^x e^{a+bx}\,dx = e^a \frac{1}{b}\left[e^{bx} - e^{bz}\right]$$

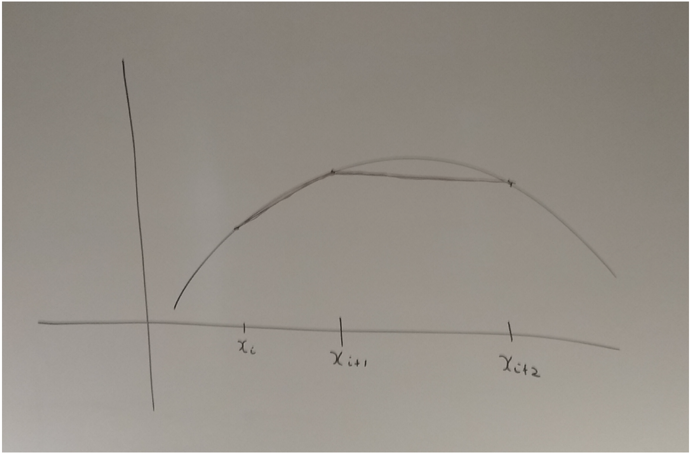
$$P\left\{X \le x \mid X \in [z_1, z_2]\right\} = \frac{e^{bx} - e^{bz_1}}{e^{bz_2} - e^{bz_1}}$$

① Draw region $R_i = [z_i, z_{i+1}\}$

$$p(R_i) \propto w_i = \frac{e^{a_i}}{b_i}\left\{e^{b_i z_{i+1}} - e^{b_i z_i}\right\}$$

② $X \sim R$ use inverse CDF.

Using the concavity, we can also construct a lower envelope squeezing function:



4. Importance Sampling

Often the goal may be expressed as the estimation of the expectation of a function:

$$E_f(h(X)) = \int h(x)\, f(x) dx \equiv \mu.$$

Importance sampling follows from the identity:

$$\int h(x)\, f(x) dx = \int h(x)\, \frac{f(x)}{g(x)}\, g(x) dx = E_g(w(X)h(X)).$$

with

$$w(X) = \frac{f(X)}{g(X)}.$$

So, for X_i, iid $\sim g$,

$$\hat{\mu}_{IS} = \frac{1}{m} \sum_{i=1}^{m} w(X_i)\, h(X_i)$$

estimates μ, with

$$w(X_i) = \frac{f(X_i)}{g(X_i)}.$$

Intuitively, the weights tell us when f would have given X more (or less) weight than g.

Why would you do this?

- You can't draw from f, but you can draw from g.
- Even if you can draw from f, it might be smart to get draws where h is big.

The optimal g is

$$g^*(x) = \frac{|h(x)|\, f(x)}{\int |h(x)|\, f(x) dx} \propto |h(x)|\, f(x)$$

Typically, you can't actually use g^*, but it does give the basic intuition that it makes sense to draw where *both* f and $|h|$ are big.

For example, if $h(x) \geq 0$ then the normalizing constant is exactly what you are trying to estimate.

If $h(x) \geq 0$, then $g^* = f\, h/(\int f\, h)$,

$$w\, h = (f/g^*)h = (\int f\, h)\frac{f}{f\, h}h = (\int f\, h),$$

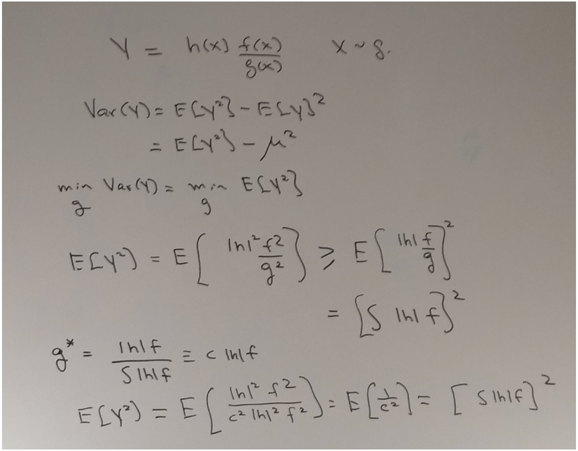
which has 0 variance.

Why is g^* optimal?

Have to min $E(Y^2)$ for $Y = hf/g$, $Y \sim g$.

Then find a lower bound for $E(Y^2)$ using Jensen's inequality.

Then show the bound is obtained at g^*.



$$Y = h(x)\frac{f(x)}{g(x)} \qquad X \sim g.$$

$$\mathrm{Var}(Y) = E\{Y^2\} - E\{Y\}^2$$

$$= E\{Y^2\} - \mu^2$$

$$\min_g \mathrm{Var}(Y) = \min_g E\{Y^2\}$$

$$E\{Y^2\} = E\left[\frac{|h|^2 f^2}{g^2} \right] \geq E\left[|h| \frac{f}{g} \right]^2$$

$$= \left\{ \int |h| f \right\}^2$$

$$g^* = \frac{|h| f}{\int |h| f} \equiv c\,|h| f$$

$$E\{Y^2\} = E\left[\frac{|h|^2 f^2}{c^2 |h|^2 f^2} \right] = E\left[\frac{1}{c^2}\right] = \left[\int |h| f \right]^2$$

Typically, you want to choose g so that the weights, $w(x) = f(x)/g(x)$ are well behaved.

If a few weights dominate the sum, then you know you can't be sure of the result.

If you did it again, you could get something completely different.

Often this means you want g to have heavier tails than f.
You definitely want g to cover the "effective support" of f.

Helps to have $h(x)$ small, when $w(x) = f(x)/g(x)$ are large, this way when f likes x a lot more than g, it is for a negligible h.

Normalized Importance Sampling

A second version of importance sampling divides the weights by their sum so that things look like a weighted average.

We have,

$$\mu = \frac{\int h(x)\, f(x)\, dx}{\int f(x)\, dx} = \frac{\int h(x)(f(x)/g(x))g(x)\, dx}{\int (f(x)/g(x))g(x)\, dx}.$$

Which motivates

$$\hat{\mu} = \frac{\sum w(X_i)\, h(X_i)}{\sum w(X_i)}, \quad X_i \sim g.$$

Or

$$\hat{\mu} = \sum w^*(X_i)h(X_i), \quad w^*(X_i) = \frac{w(X_i)}{\sum w(X_j)}.$$

Why Normalize

Sometimes we can only compute f and/or g up to a proportionality constant.

A basic example is Bayesian statistics where $X = \theta$ and

$$f(\theta \mid y) \propto L(\theta)\, p(\theta) \equiv f_1(\theta).$$

Then

$$w(\theta) = \frac{L(\theta)p(\theta)}{g(\theta)}$$

An example of this is prior sensitivity.

Suppose we have prior $p_1(\theta)$ and prior $p_2(\theta)$.
Let $p_i(\theta \mid y)$ be the posterior obtained from prior $p_i(\theta)$ and likelihood $L(\theta)$.

We have developed an algorithm for drawing from $p_1(\theta \mid y)$, the prior 1 posterior.

If we want an expectation with respect to the prior 2 posterior, we can draw from the prior 1 posterior and use

$$g(\theta) = p_1(\theta \mid y) \propto L(\theta)\, p_1(\theta), \;\; f(\theta) = p_2(\theta \mid y) \propto L(\theta)\, p_2(\theta).$$

Then

$$w(\theta) \propto \frac{L(\theta)\, p_2(\theta)}{L(\theta)\, p_1(\theta)} = \frac{p_2(\theta)}{p_1(\theta)}.$$

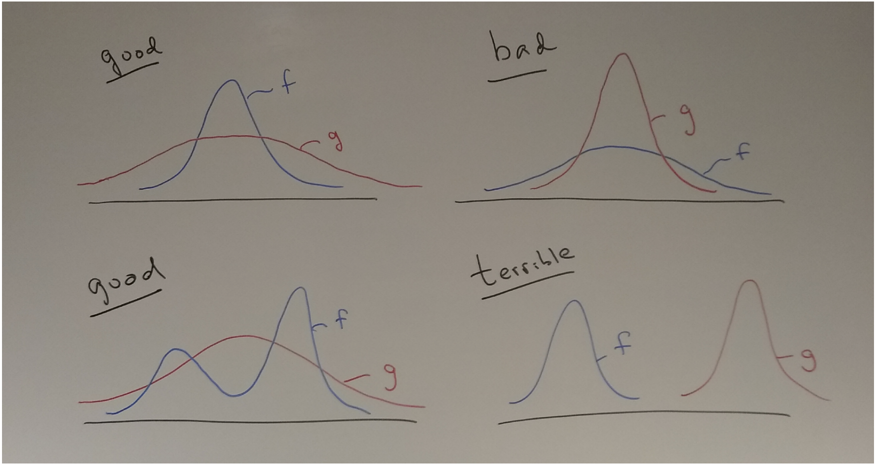
If you know the normalizing constants so that computation of $w(x) = f(x)/g(x)$ is not an issue then which is better can depend on h.

You can show that the standardized version can be better when $w(X)$ and $w(X)\,h(X)$ are strongly positively correlated.

Note that the normalized version may be slightly biased, but the basic version is unbiased.

Note:

you can't reweight to a different place !!!!!



5. SIR: Sampling Importance Resampling

Let δ_x put mass 1 on x.

Given iid draws $X_i \sim P$, the empirical distribution is

$$\hat{P} = \sum \frac{1}{n} \delta_{X_i}.$$

Our basic monte carlo idea is that $\hat{P}$ approximates P so that,

$$\int h(x) dP(x) \approx \int h(x) d\hat{P}(x) = \frac{1}{n} \sum_{i=1}^{n} h(X_i).$$

Note: If P has density f, then $\int h(x) dP(x) = \int h(x) f(x) dx$. In this case, $\hat{P}$ is a discrete approximation to the continuous distribution corresponding to f.

The normalized version has the nice interpretation that if we let

$$\hat{P}_w(x) = \sum w^*(X_i)\,\delta_{X_i}$$

then the weighted version is just the expecion with respect to $\hat{P}_w$ and we can think of $\hat{P}_w$ as an approximation to the distribution corresponding to $f(x)$ in that, for any h

$$E_P(h(X)) = \int h(x)\,dP(x) \approx \sum h(X_i)\,w^*(X_i) = \int h(x)\,d\hat{P}_w.$$

SIR: Sampling Importance Resampling

Sometimes it makes like simpler if the weights are uniform.

We can get an iid sample approximating P by drawing from $\hat{P}_w$.

- Let X_i be iid from g, $i = 1, 2, \ldots, m$.
- Let $w_i^*(X_i) = \frac{f(X_i)/g(X_i)}{\sum_{j=1}^m f(X_j)/g(X_j)},\ i = 1, 2, \ldots, m$.
- $\hat{P}_w = \sum_{i=1}^m w_i^*(X_i)\,\delta_{X_i}$.
- Draw $\tilde{X}_i,\ i = 1, 2, \ldots, n$, iid from $\hat{P}_w$.
- $\tilde{P} = \sum_{i=1}^n \frac{1}{n}\,\delta_{\tilde{X}_i}$.

Note:

In principle need large n and m.

Should have $n/m \to 0$, that is, a big m.

G & H:
"We have sometimes found $n/m < \frac{1}{10}$ tolerable so long as the resulting sample does not contain too many replicates of any initial draw."

Example: Bayesian Statistics without Tears (Gelfand and Smith)

A Bayesian model: $\{f(y \mid \theta), p(\theta)\}$.

Let $g = p$ and $f \propto f(y \mid \theta)\, p(\theta)$.

- Draw θ_i from the prior.
- Reweight draws by $w = f/g = f(y \mid \theta) = L(\theta)$.

Beautiful idea, typically does not work.

Typically you hope the data is informative so that the likelihood is much more concentrated than the prior.