Module 5: Introduction to Monte Carlo

Rebecca C. Steorts

Agenda

- ▶ Motivation
- ▶ Monte Carlo (The Classical or Naive Method)
- ▶ Inverse CDF Method
- ▶ Rejection Sampling

Sampling Methods

In this module, we will talk about ways of approximating

$$E_X[h(x)] = \int_X h(x)f(x) \, dx$$

which is *intractable*. This means that we cannot compute the integral in closed form.

Why? This means h(x) is a complicated function or we cannot evaluate $f(x)$.

Sampling Methods

$$E_X[h(x)] = \int_X h(x)f(x)\ dx$$

which is *intractable*. This means that we cannot compute the integral in closed form.

Example: Suppose the $h(\cdot)$ is the identity function.

$$E_X[h(x)] = E_X[X] = \int_X xf(x)\ dx$$
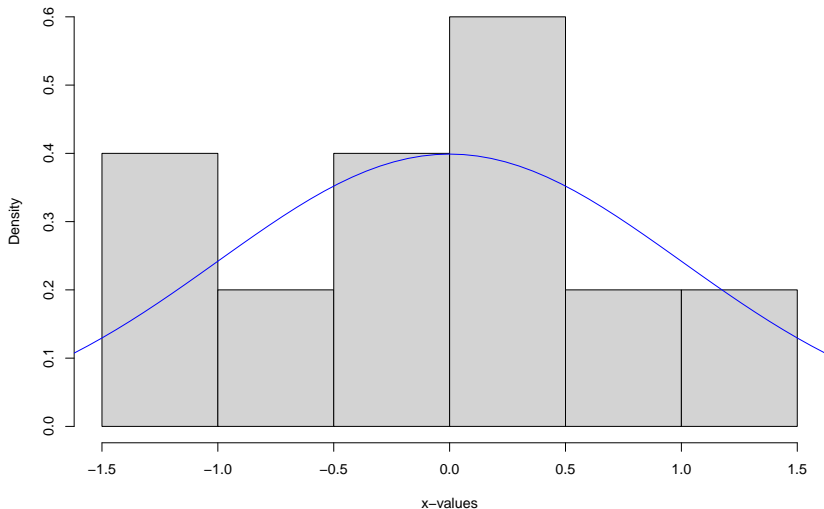
Sampling Methods

Solution: Draw samples $X_1, \ldots, X_n \overset{iid}{\sim} f(x)$.

$$E_X[h(x)] \approx \frac{1}{n} \sum_{j=1}^{n} h(X_j).$$

This is called Monte Carlo sampling (and is the simplest way of producing samples).
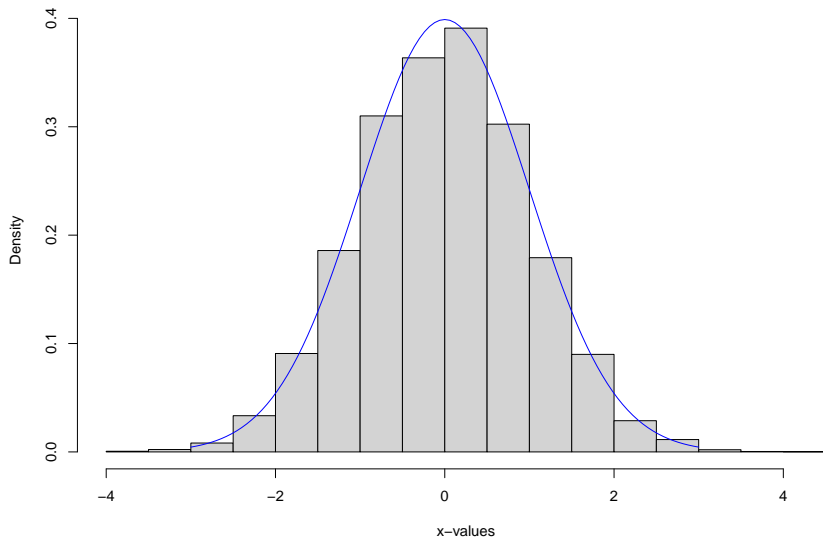
Illustration

Simulate a N(0,1) histogram, and the mean and variance. Compare it to the density in R. Do this for 10 versus 1000 simulations.

Illustration

Simulate a N(0,1) histogram, and the mean and variance. Compare it to the density in R. Do this for 10 versus 1000 simulations.

Illustration

```
mean(rnorm(x))
```

```
## [1] -0.001452919
```

```
var(rnorm(x))
```

```
## [1] 0.9981385
```

What do you observe empirically?

Monte Carlo Sampling

The generic problem here is to evaluate

$$E_f[h(x)] = \int_X h(x)f(x) \, dx.$$

The classical way to solve this is generate samples $(X_1, \ldots, X_n)$ from f.

Now propose as an approximation the empirical average:

$$E_f[h(x)] \approx \frac{1}{n} \sum_{j=1}^{n} h(x_j) := \bar{h}_n.$$

Why? $\bar{h}_n$ converges a.s. (i.e. for almost every generated sequence) to $E_f[h(X)]$ by the Strong Law of Large Numbers.

Monte Carlo (continued)

Also, under certain assumptions[1], the asymptotic variance v_n can be approximated and then can be estimated from the sample $X = (X_1, \ldots, X_n)$ by

$$v_n = V[\frac{1}{n} \sum_j h(x_j)] = \frac{1}{n^2} \sum_j V[h(x_j)] \approx \frac{1}{n^2} \sum_{j=1}^{n} [h(x_j) - \bar{h}_n]^2.$$

Finally, by the CLT (for large n),

$$\frac{\bar{h}_n - E_f[h(X)]}{\sqrt{v}_n} \overset{\text{approx.}}{\sim} N(0, 1).$$

(Technically, it converges in distribution).

---

[1]see Casella and Robert, page 65, for details

Root Mean Squared Error (RMSE)

Assume that X_i are iid observations. Suppose h is the identity function.

Then

$$
\begin{aligned}
\text{RMSE} &=: \left[\mathbb{E}(|\tfrac{1}{N} \sum X_i - \mathbb{E}X|^2) \right]^{1/2} \\
&= \left[\mathbb{V}(\tfrac{1}{N} \sum X_i) \right]^{1/2} \\
&= \frac{1}{\sqrt{N}} \mathbb{V}(X)^{1/2} = \sigma(X)/\sqrt{N}. \tag{1}
\end{aligned}
$$

The RMSE tells us how far the approximation will be from the true value, on average.

As a practical matter, we need to be able to draw the samples X_i in a computationally-efficient way.
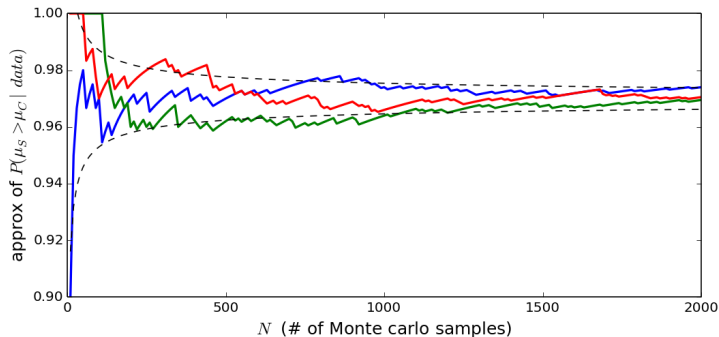
Return of IQ Scores

Figure 1: Monte Carlo approximations for an increasing number of samples, N. The red, blue, and green lines indicate three repetitions of the procedure, using different sequences of samples. The dotted lines indicate the true value $\pm$ the RMSE of the Monte Carlo estimator.

Return of IQ Scores

In Module 4, we saw an example involving the mean change in IQ score μ_S and μ_C of two groups of students (spurters and controls). To compute the posterior probability that the spurters had a larger mean change in IQ score, we drew $N = 10^6$ samples from each posterior:

$$(\mu_S^{(1)}, \lambda_S^{(1)}), \ldots, (\mu_S^{(N)}, \lambda_S^{(N)}) \sim \text{NormalGamma}(24.0, 8, 4, 855.0)$$
$$(\mu_C^{(1)}, \lambda_C^{(1)}), \ldots, (\mu_C^{(N)}, \lambda_C^{(N)}) \sim \text{NormalGamma}(11.8, 49, 24.5, 6344.0)$$

and used the Monte Carlo approximation

$$\mathbb{P}(\mu_S > \mu_C \mid \text{data}) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\mu_S^{(i)} > \mu_C^{(i)}).$$

Return of IQ Scores

- ▶ To visualize this, consider the sequence of approximations $\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\boldsymbol{\mu}_S^{(i)} > \boldsymbol{\mu}_C^{(i)})$ for $N = 1, 2, \ldots$.

- ▶ Figure 1 shows this sequence of approximations for three different sets of random samples from the posterior.

- ▶ We can see that as the number of samples used in the approximation grows, it appears to be converging to around 0.97.

To visualize the theoretical rate of convergence, the figure also shows bands indicating the true value $\alpha = \mathbb{P}(\boldsymbol{\mu}_S > \boldsymbol{\mu}_C \mid \text{data}) =??$ plus or minus the RMSE of the Monte Carlo estimator, that is, from Equation 1:

$$\alpha \pm \sigma(X)/\sqrt{N} =??$$

Simpify this as much as possible for an ungraded exercise (exam II).

Solution to the ungraded exercise

$$\alpha \pm \sigma(X)/\sqrt{N} = \alpha \pm \sqrt{\alpha(1-\alpha)/N}$$
$$= 0.97 \pm \sqrt{0.97(1-0.97)/N}$$

where X has the posterior distribution of $\mathbb{1}(\boldsymbol{\mu_S} > \boldsymbol{\mu_C})$ given the data, in other words, X is a Bernoulli(α) random variable. Recall that the variance of a Bernoulli(α) random variable is $\alpha(1-\alpha)$.

Return of IQ Scores

Using the same approach, we could easily approximate any number of other posterior quantities as well, for example,

$$\mathbb{P}(\boldsymbol{\lambda}_S > \boldsymbol{\lambda}_C \,|\, \text{data}) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\boldsymbol{\lambda}_S^{(i)} > \boldsymbol{\lambda}_C^{(i)})$$

$$\mathbb{E}(|\boldsymbol{\mu}_S - \boldsymbol{\mu}_C| \,|\, \text{data}) \approx \frac{1}{N} \sum_{i=1}^{N} |\boldsymbol{\mu}_S^{(i)} - \boldsymbol{\mu}_C^{(i)}|$$

$$\mathbb{E}(\boldsymbol{\mu}_S/\boldsymbol{\mu}_C \,|\, \text{data}) \approx \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\mu}_S^{(i)}/\boldsymbol{\mu}_C^{(i)}.$$

Background: inverse CDF method

For a random variable X with c.d.f. F, define the generalized CDF as

$$F^{-1}(u) = min\{x : F(x) \geq u\}.$$

If

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

then

$$F^{-1}(U) \sim X.$$

Illustration

$$X \sim Exponential(1).^2$$

Then

$$f(x) = \exp(-x) \implies F(x) = 1 - \exp(-x).$$

Let $u = 1 - \exp^{-x}$ and solve for u. Then $x = -log(1 - u)$.

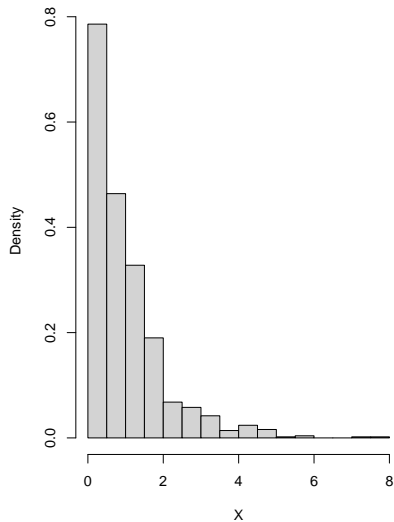Let $U \sim Uniform(0, 1) \implies 1 - U \sim Uniform(0, 1)$.
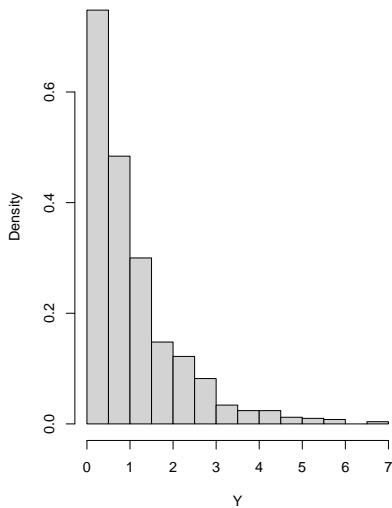
Thus,

$$X \sim -log(U).$$

---

^{2}Applications include the length of time, in minutes, of long distance business teleph

Illustration

Exponenial via Uniform   **Exponenial via R**

Rejection Sampling

Rejection sampling is a method for drawing random samples from a distribution whose p.d.f. can be evaluated up to a constant of proportionality.

Compared with the inverse c.d.f. method, rejection sampling has the advantage of working on complicated multivariate distributions. (see homework)

Difficulties? You must design a good proposal distribution (which can be difficult, especially in high-dimensional settings).

Uniform Sampler

Goal: Generate samples from Uniform(A), where A is complicated.

- $X \sim Uniform(Mandelbrot)$.
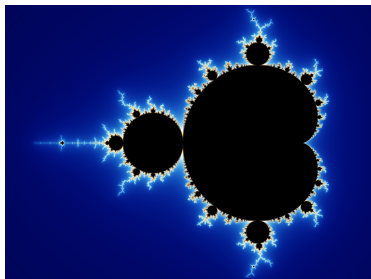- Consider $I_{X(A)}$.

The Mandelbrot

Figure 2: A complicated function A, called the Mandelbrot!

The Mandelbrot

The Mandelbrot set is a set of points in the complex plane.[3]

A point in this plane can be defined using a complex number $c \in C$ such that

$$c = a + bi,$$

where a,b are real numbers and $i = \sqrt{-1}$.

Formally, $c \in C$ belongs to the Mandelbrot set iff

$$lim_{n \to \infty} ||z_{n+1} = z_n^2 + c|| \nrightarrow \infty \quad \text{where} \quad z_o = 0.$$

▶ Note that $|| \cdot ||$ is the Euclidean norm[4]

---

[3]The complex plane is a two-dimensional space with the a vertical imaginary axis, and a horizontal real axis.

[4]This measures how far a point is from it's origin.

The Mandelbrot

Formally, $c \in C$ belongs to the Mandelbrot set iff

$$lim_{n \to \infty} ||z_{n+1} = z_n^2 + c|| \nrightarrow \infty \quad \text{where} \quad z_o = 0.$$

- ▶ We have a re-cursive function.[5]
- ▶ Conjugate distributions – out the window!
- ▶ We're going to need to do something numerical!

---

[5]To read more about fractals, see https://www.kth.se/social/files/5504b42
ff276543e4aa5f5a1/An_introduction_to_the_Mandelbrot_Set.pdf.

Exercise

▶ Suppose $A \subset B$.

▶ Let $Y_1, Y_2, \ldots \sim \text{Uniform(B)}$ iid and

▶ $X = Y_k$ where $k = \min\{k : Y_k \in A\}$,

Then it follows that

$$X \sim \text{Uniform}(A).$$

Proof: Exercise. Hint: Try the discrete case first and use a geometric series.
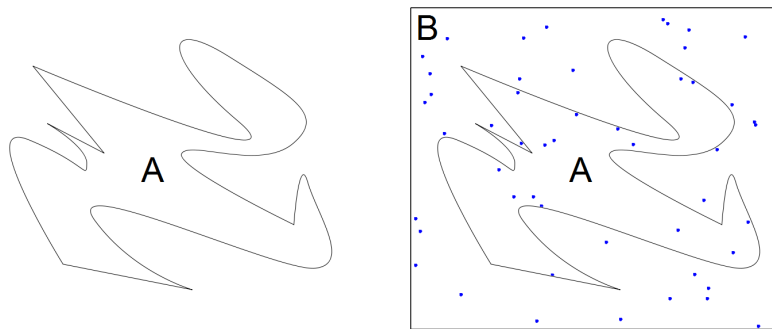
Drawing Uniform Samples

Figure 3: (Left) How to draw uniform samples from region A? (Right) Draw uniform samples from B and keep only those that are in A.

General Rejection Sampling Algorithm

Goal: Sample from a complicated pdf $f(x)$.

Suppose that

$$f(x) = \tilde{f}(x)/\alpha, \alpha > 0$$

.

Assumption: f is difficult to evaluate, $\tilde{f}$ is easy! Why? α may be very difficult to calculate even computationally.

1. Choose a proposal distribution q such that $c > 0$ with

$$cq(x) \geq \tilde{f}(x).$$

2. Sample $X \sim q$, sample $Y \sim \text{Unif}(0, c\ q(X))$ (given X)

3. If $Y \leq \tilde{f}(X), Z = X$, otherwise we reject and return to step (2).
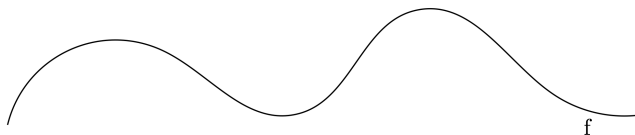
Output: $Z \sim f(x)$

Visualizing just f

Figure 4: Visualizing just f.
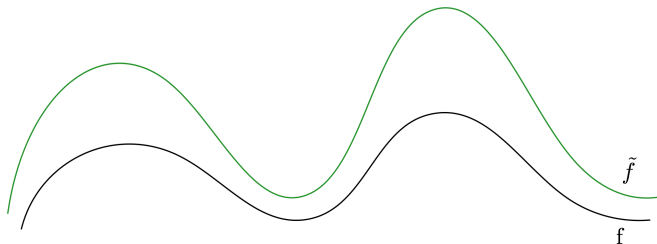
Visualizing just f and $\tilde{f}$

Figure 5: Visualizing just f and $\tilde{f}$.
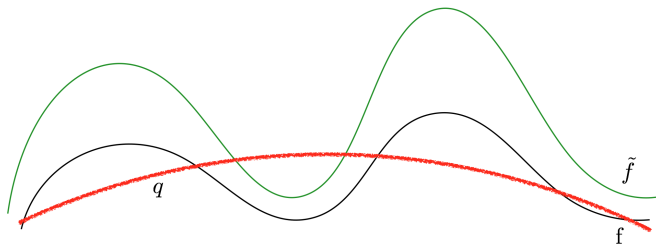
Enveloping q over f

Figure 6: Visualizing f and $\tilde{f}$. Now we look at enveloping q over f.
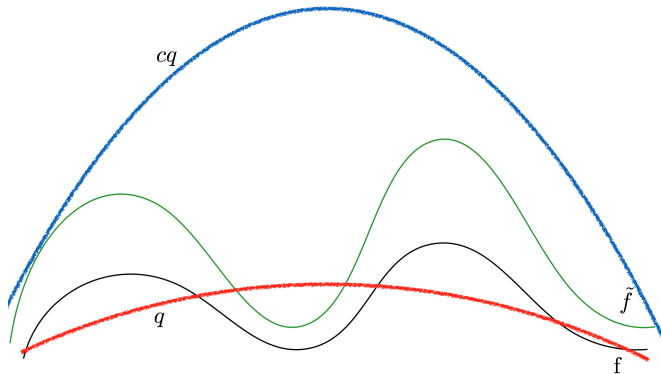
Figure 7: Visualizing f and $\tilde{f}$. Now we look at enveloping *cq* over $\tilde{f}$.

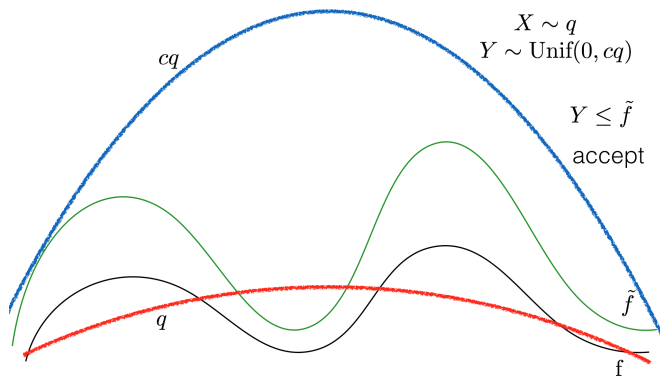# Recalling the sampling method and accept/reject step

Figure 8: Recalling the sampling method and accept/reject step.

Entire picture and an example point X and Y

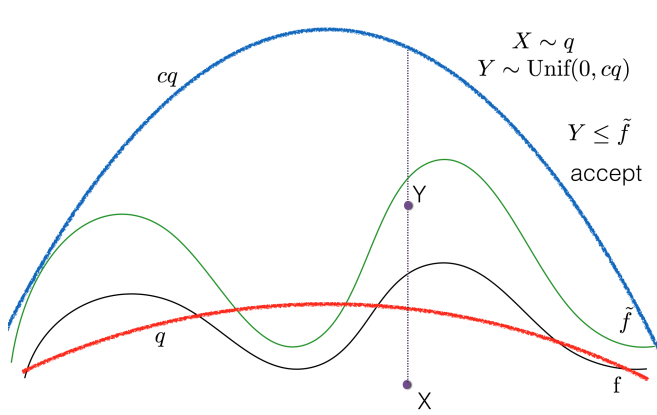Figure 9: Entire picture and an example point X and Y.

Takeaways

- What is Monte Carlo (The naive method)
- Rejection sampling
- Inverse CDF method

Detailed Takeaways

▶ Why do we use Monte Carlo?
▶ Why do we use rejection sampling?
▶ In the next modules, we will learn about Markov chain Monte Carlo algorithms (MCMC), which are used for working in high dimensional parameters spaces.

Exercise (Lab 5)

Consider the function

$$f(x) \propto \sin^2(\pi x), x \in [0, 1]$$

1. Plot the densities of $f(x)$ and the Unif(0,1) on the same plot.

2. According to the rejection sampling approach sample from $f(x)$ using the Unif(0,1) pdf as an enveloping function.

3. Plot a histogram of the points that fall in the acceptance region. Do this for a simulation size of 10^2 and 10^5 and report your acceptance ratio. Compare the ratios and histograms.

4. Repeat Tasks 1 - 3 for Beta(2,2) as an enveloping function. Compare your results with results in Task 3.

5. Do you recommend the Uniform or the Beta(2,2) as a better enveloping function (or are they about the same)? If you were to try and find an enveloping function that had a high acceptance ratio, which one would you try and why?

Task 1

```
# density function for f(x)
densFun <- function(x) {
  return(sin(pi*x)^2)
}
x <- seq(0, 1, 10^-2)
```

Task 1

Densities comparison
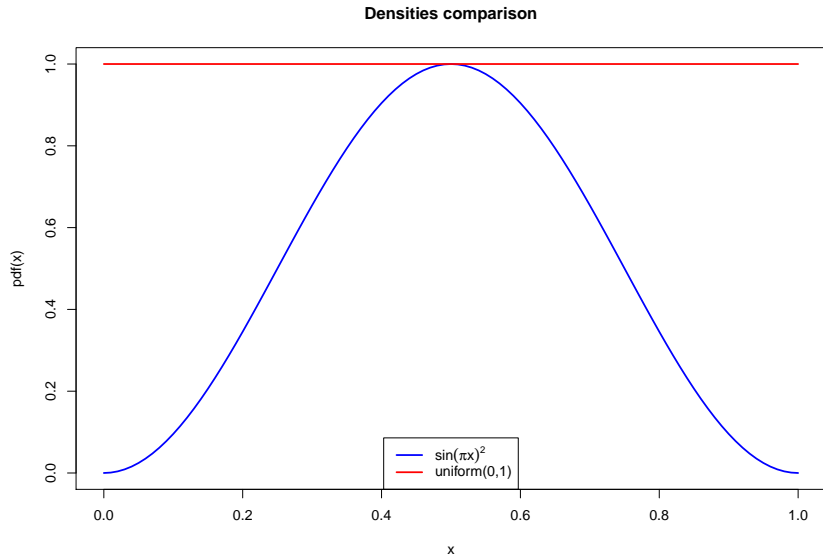
Task 2

```
numSim=10^2
samples = NULL
for (i in 1:numSim) {
  # get a uniform proposal
  proposal <- runif(1)
  # calculate the ratio
  densRat <- densFun(proposal)/dunif(proposal)
  #accept the sample with p=densRat
  if ( runif(1) < densRat ){
    #fill our vector with accepted samples
    samples <- c(samples, proposal)
  }
}
```
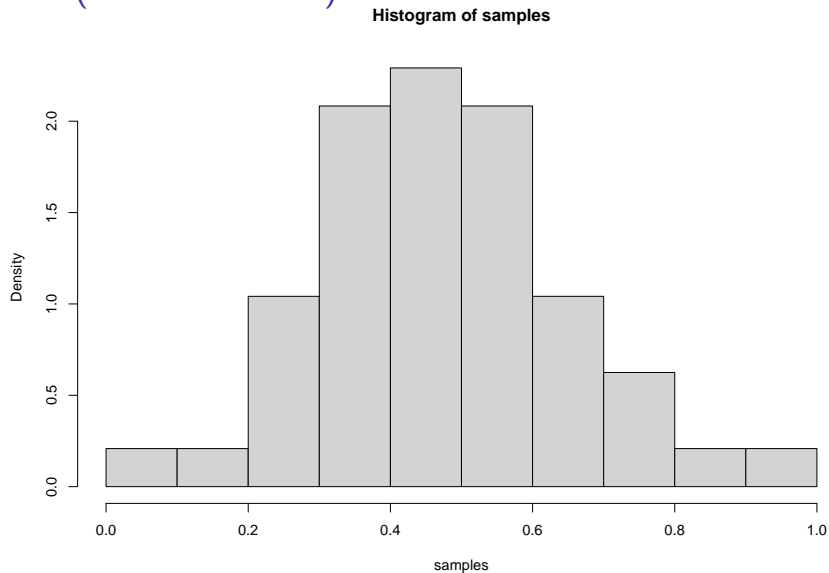
Task 3 (Partial solution)

Histogram of samples

[1] "Acceptance Ratio: 0.48"

Task 2 – 4 (Partial Solution)

```
sim_fun <- function(f, envelope = "unif", par1 = 0,
                     par2 = 1, n = 10^2, plot = TRUE){

  r_envelope <- match.fun(paste0("r", envelope))
  d_envelope <- match.fun(paste0("d", envelope))
  proposal <- r_envelope(n, par1, par2)
  density_ratio <- f(proposal) / d_envelope(proposal, par1, par2)
  samples <- proposal[runif(n) < density_ratio]
  acceptance_ratio <- length(samples) / n
  if (plot) {
    hist(samples, probability = TRUE,
         main = paste0("Histogram of ", n, " samples from ",
                       envelope, "(", par1, ",", par2, ").\n",
                       Acceptance ratio: ",
                       round(acceptance_ratio,2)), cex.main = 0.75)
  }
  list(x = samples, acceptance_ratio = acceptance_ratio)
}
```

Task 2 – 4 (Partial Solution)
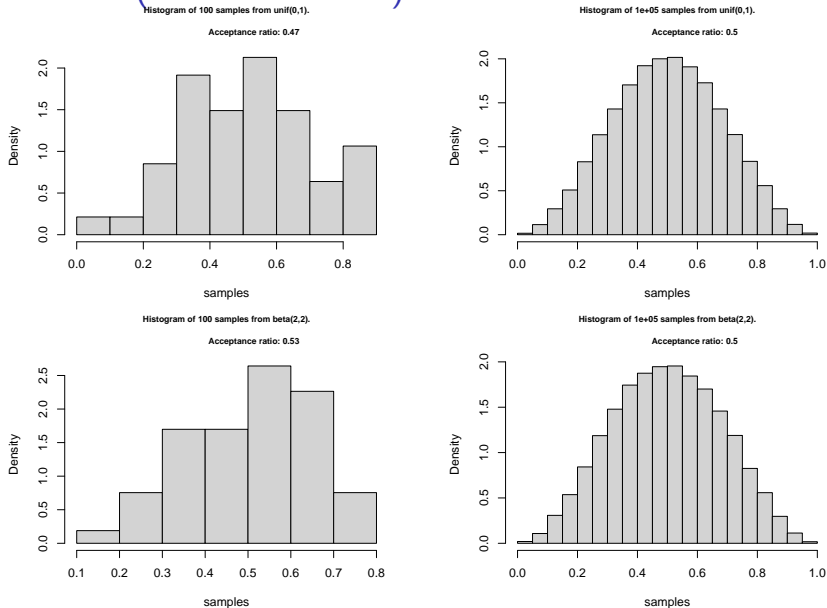
Figure 2: Rejection sampling for 100 versus 100,000 simulations

Takeaways

1. What do you notice about the enveloping functions and the acceptance ratio as the number of samples is large?

2. What does this tell you about the uniform proposal versus the beta proposal in this specific application?

Exercise (Lab 6)

Consider

$$I = \int_{-\infty}^{\infty} \exp(-x^4) \, dx.$$

Tasks (Lab 6)

1. Task 1: Find a closed form solution to *I* and evaluate this.
2. Task 2: Approximate *I* using Monte carlo.
3. Task 3: Approximate *I* using importance sampling.

Gamma density

Before proceeding with the tasks, let's recall that one variant of the Gamma(a,b) density can be written as follows:

$$\text{Gamma}(x|a, b = \text{rate}) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \mathbb{1}(x > 0)$$

for $a, b > 0$,

Task 1

For the sake of comparison, we can derive the true value using substitution and the gamma function. We will use the substitution $u = x^4$ and $\text{Gamma}(x|a, b = \text{rate}) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \mathbb{1}(x > 0)$ for $a, b > 0$,

$$
\begin{aligned}
\int_{-\infty}^{\infty} \exp(-x^4)\, dx &= 2 \int_0^{\infty} \exp(-x^4)\, dx \\
&= 2 \int_0^{\infty} \frac{\exp(-u)}{4u^{3/4}}\, du \\
&= 2^{-1} \int_0^{\infty} u^{1/4-1} e^{-u}\, du \\
&= \frac{\Gamma(1/4)}{1^{1/4}} \times 2^{-1} \int_0^{\infty} u^{1/4-1} e^{-u} \times \frac{1^{1/4}}{\Gamma(1/4)}\, du \\
&= \frac{\Gamma(1/4)}{2} \\
&= 1.813.
\end{aligned}
$$

Task 2 (Monte Carlo)

In this task, we perform Monte carlo. Let $y = \sqrt{2}x^2$. We will perform u-substitution to evaluate the integral with Monte Carlo (verify this calculation on your own)[6]

$$I = 2^{-5/4} \int_0^\infty \sqrt{\frac{2\pi}{y}} 2\phi(y)\, dy.$$

The function $2\phi(y)$ is the density of the normal distribution truncated (or folded) at zero. We can sample from this distribution by taking samples from the standard normal distribution and then taking their absolute value. Note that if $X \sim N(0, 1)$ we see for any $c > 0$

$$
\begin{aligned}
P(|X| < c) &= P(-c < X < c) \\
&= 2(\Phi(c) - 1/2) \\
&= 2\Phi(c) - 1,
\end{aligned}
$$

Task 3 (Importance Sampling)

We can multiply and divide the integral by a density that has a support equal to the area over which we are integrating. An obvious and easy choice is the standard normal density, ϕ:

$$\int_{-\infty}^{\infty} \exp(-x^4)\, dx = \int_{-\infty}^{\infty} \frac{\exp(-x^4)}{\phi(x)} \phi(x)\, dx.$$

We can therefore evaluate the integral by sampling from a standard normal and averaging the values evaluated in $\exp(-x^4)/\phi(x)$. Thus, we will perform re-weighting, and thus, utilizing importance sampling.

Comparison of Methods

The results of 10,000 simulations using the three methods described
above are summarized in Table 1; the true value is included for
comparison. Of these methods, multiplying and dividing by the
standard normal density and then sampling from this density seems
to yield the best estimate, which is both closer to the true value and
has lower standard error. The other two methods are comparable in
both their estimates and standard errors.

	Mean	SE
True value	1.81	
Truncated Normal (Monte Carlo)	1.79	0.06
Full Normal (IS)	1.81	0.01

Table 1: Comparison of the Monte Carlo estimate for the value of the
integral using various methods with 10,000 draws for each method. As we
can see under importance sampling, the estimate is closer to the true value
and the SE is also lower.

Comparison of Methods

```
## pdf
##   2

## pdf
##   2

## pdf
##   2

## % latex table generated in R 4.0.3 by xtable 1.8-4 packa
## % Tue Jan 31 10:18:40 2023
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & Mean & SE \\
##   \hline
## True value & 1.81 &  \\
##   row.names
##                                 Normal & 1.79 & 0^{51}06^{67} \\
```
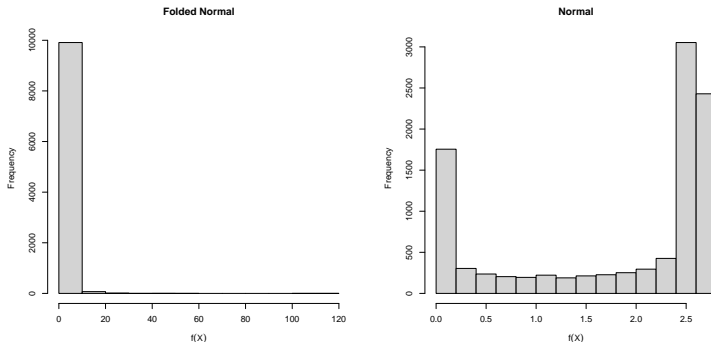
Histograms

Figure 10: Histograms for the various Monte Carlo simulations.

An ideal histogram would be highly centered around the true value of 1.81. For the folded normal, we see that there are some very large observations that skew the distribution. The normal method also results in a strange histogram, with values concentrated near the edges.

Exam I

- ▶ Open notes, open book.
- ▶ This is a closed exam to speaking to others except your instructor and teaching assistants until the exam grades have been released to all students.
- ▶ You will be given cover page with distributions.
- ▶ This exam will be held virtually in case that anyone falls ill or needs to quarantine.
- ▶ Please make sure that you arrive early and not late to the zoom session so that you do not miss any announcements.
- ▶ Please only ask questions privately through the chat and do not broadcast these to everyone.
- ▶ Material covers: Modules 1–4.
- ▶ Material covers lectures (slides and written material in class), labs, and homeworks.
- ▶ You will be given 30 minutes after the exam to upload one PDF document to Gradescope and assign pages.

Exam I

▶ Under the syllabus, you are not allowed to speak to anyone regarding the exam except the instructor until the results (grades) are released back to you.

Exam Dates II and III

These dates are tentative on the calendar and I will announce these more formally in the next few weeks. Please be careful regarding booking travel during class times.

Module I (Recap)

- ▶ Bayes Theorem
- ▶ Cast of characters
- ▶ Conjugacy
- ▶ Marginal and posterior predictive distributions
- ▶ Here, we looked at very simple applied examples regarding polling and sleep to motivate the use of conjugacy.

Module II (Recap)

- ▶ Decision Theory
- ▶ Loss functions
- ▶ Bayes Risk
- ▶ Frequentist Risk
- ▶ Integrated Risk
- ▶ Here, we looked at a resource allocation problem with a non-trivial loss function.

Module III (Recap)

- ▶ Univariate Normal distribution
- ▶ Properties of the normal distribution
- ▶ Normal-Uniform
- ▶ The uniform is an example of an imporoper prior
- ▶ Normal-Normal conjugacy
- ▶ The precision
- ▶ What happens to the Normal-Normal posterior as the sample size gets large?
- ▶ The applied example here was about Dutch heights of women and men and looking at bi-modality.

Module IV (Recap)

- ▶ The Normal-Gamma conjugacy
- ▶ This module was the first time we saw a three-layer hierarhical model
- ▶ This was a very long derivation
- ▶ The applied example that went with this model was IQ scores since we had two different populations with different means and precisions.

Review Materials

- Practice exercises: https://github.com/resteorts/modern-bayes/tree/master/exercises
- Review homework exercises
- I highly recommend that you work all these problems on your own and make sure that you understand the solutions (which are provided).

In class notes

Derivation of bounds can be found here:
https://github.com/resteorts/modern-bayes/blob/master/lecturesModernBayes20/lecture-5/05-class-notes/derivation-of-bounds.pdf

Notes on importance sampling can be found here:
https://github.com/resteorts/modern-bayes/blob/master/lecturesModernBayes20/lecture-5/05-class-notes/importance-sampling.pdf

Appendix

The appendix sketches out a proof of the general rejection sampler, which you do not need to know for the exam.

Lemma 1

Lemma: If

$$X \sim q \quad Y \mid X \sim \text{Unif}(0, cq) \implies (X, Y) \sim Unif(B),$$

where $B = \{(x, y), \quad x \in R^d, \quad 0 < y < cq(x)\}$.

Proof:

a.) If $y \notin (0, cq)$, then $p(x, y) = p(y \mid x)p(x) = 0$.

b.) Else, $p(x, y) = p(y \mid x)p(x) = \frac{1}{cq(x)} \times q(x) = \frac{1}{c}$.

Lemma 2

If $(X, Y) \sim \text{Unif}(A)$, where $A = \{(x, y) : x \in R^d, 0 < y < \tilde{f}(x)\}$, then $X \sim f$.

Proof: It follows that $m(A) = \int \tilde{f}(x) dx = \int \alpha f(x) dx = \alpha$.

Consider $1 = \int_A b \, dx dy = b \int [\int_0^{\tilde{f}(x)} dy] dx = b \int \tilde{f}(x) dx = b\alpha \implies b = 1/\alpha$.

Then $p(x) = \int p(x, y) dy = \int \frac{1}{\alpha} I(0 < y < \tilde{f}(x)) dy = \frac{1}{\alpha} \int_0^{\tilde{f}(x)} dy = \frac{1}{\alpha} \tilde{f}(x) = f(x)$.

Proposition

Suppose f and q are pdfs on $\mathbb{R}^d$ such that

$$f(x) = \tilde{x}/\alpha, \alpha > 0$$

and

$$cq(x) \geq \tilde{f}(x) \forall x \in \mathbb{R}^d.$$

If

$$X_1, X_2, \ldots, \sim q,$$
$$Y_k \mid X_k \sim Unif(0, cq(X_k)),$$

and

$$Z = X_k \text{ where } K = min\{k : Y_k \leq \tilde{f}(X_k)\}$$

then

$$Z \sim f.$$

Proof of Proposition

The proposition follows by Lemma 1 and Lemma 2.

Example

Find the value of I