

Module 5: Introduction to Monte Carlo

Rebecca C. Steorts

Agenda

- ▶ Motivation
- ▶ Numerical Integration
- ▶ Monte Carlo (The Naive Method)
- ▶ Importance Sampling
- ▶ Rejection Sampling

Intractable Integrals

Goal: approximate

$$\int_{\mathbf{x}} h(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

that is intractable, where $f(\mathbf{x})$ is a probability density.

- ▶ What's the problem? Typically $h(\mathbf{x})$ is messy!
- ▶ Why not use numerical integration techniques?
- ▶ In dimension $d = 3$ or higher, Monte carlo really improves upon numerical integration.

Numerical Integration

- ▶ The most serious problem is the so-called “curse of dimensionality.”
- ▶ Suppose we have a d -dimensional integral.
- ▶ Numerical integration typically entails evaluating the integrand over some grid of points.
- ▶ However, if d is even moderately large, then any reasonably fine grid will contain an impractically large number of points.

Numerical integration

- ▶ Let $d = 6$. Then a grid with just ten points in each dimension will consist of 10^6 points.
- ▶ If $d = 50$, then even an absurdly coarse grid with just *two* points in each dimension will consist of 2^{50} points (note that $2^{50} > 10^{15}$).

What's happening here?

Numerical integration error rates (big Ohh concepts)

- ▶ If $d = 1$ and we assume crude numerical integration based on a grid size n , then we typically get an error of order n^{-1} .
- ▶ For most dimensions d , estimates based on numerical integrations required m^d evaluations to achieve an error of m^{-1} .
- ▶ Said differently, with n evaluations, you get an error of order $n^{-1/d}$.
- ▶ But, the Monte Carlo estimate retains an error rate of $n^{-1/2}$. (The constant in this error rate may be quite large).

Classical Monte Carlo Integration

The generic problem here is to evaluate

$$E_f[h(x)] = \int_{\mathcal{X}} h(x)f(x) dx.$$

The classical way to solve this is generate a sample (X_1, \dots, X_n) from f .

Now propose as an approximation the empirical average:

$$\bar{h}_n = \frac{1}{n} \sum_{j=1}^n h(x_j).$$

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 assumptions¹, the asymptotic variance can be approximated and then can be estimated from the sample (X_1, \dots, X_n) by

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

Finally, by the CLT (for large n),

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

(Technically, it converges in distribution).

¹see Casella and Robert, page 65, for details

Root Mean Squared Error (RMSE)

Assume that X_i are iid observations.

Due to unbiasedness, the root-mean-squared-error (RMSE) equals the standard deviation (square root of the variance) of $\frac{1}{N} \sum X_i$,

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

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

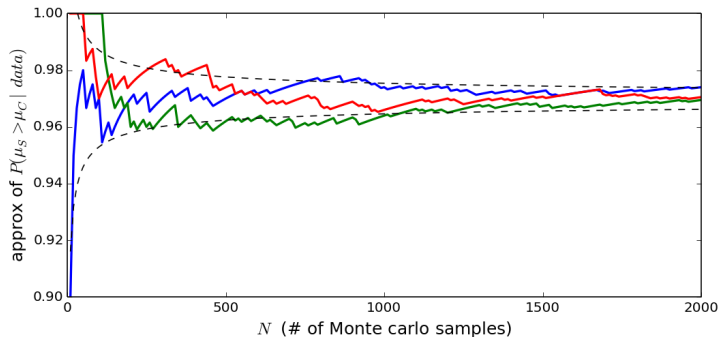


Figure 1: Monte Carlo approximations for an increasing number of samples, N . The red, blue, and green lines indicate three repetitions of the procedure, using different sequences of samples. The dotted lines indicate the true value \pm the RMSE of the Monte Carlo estimator.

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:

$$(\mu_S^{(1)}, \lambda_S^{(1)}), \dots, (\mu_S^{(N)}, \lambda_S^{(N)}) \sim \text{NormalGamma}(24.0, 8, 4, 855.0)$$

$$(\mu_C^{(1)}, \lambda_C^{(1)}), \dots, (\mu_C^{(N)}, \lambda_C^{(N)}) \sim \text{NormalGamma}(11.8, 49, 24.5, 6344.0)$$

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 $\frac{1}{N} \sum_{i=1}^N \mathbb{1}(\mu_S^{(i)} > \mu_C^{(i)})$ for $N = 1, 2, \dots$
- ▶ 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 $\alpha = \mathbb{P}(\mu_S > \mu_C \mid \text{data}) = ??$ plus or minus the RMSE of the Monte Carlo estimator, that is, from Equation 1:

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

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

Solution to the ungraded exercise

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

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

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

$$\mathbb{E}(|\mu_S - \mu_C| \mid \text{data}) \approx \frac{1}{N} \sum_{i=1}^N |\mu_S^{(i)} - \mu_C^{(i)}|$$

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

Importance Sampling

Recall that we have a difficult, problem child of a function $h(x)$!

- ▶ Generate samples from a distribution $g(x)$.
- ▶ We then "re-weight" the output.

Note: g is chosen to give greater mass to regions where h is large (the important part of the space).

This is called *importance sampling*.

Importance Sampling

Let g be an arbitrary density function and then we can write

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

This is estimated by

$$\hat{I} = \frac{1}{n} \sum_{j=1}^n \frac{f(X_j)}{g(X_j)} h(X_j) \longrightarrow E_f[h(X)] \quad (3)$$

based on a sample generated from g (not f). Since (2) can be written as an expectation under g , (3) converges to (2) for the same reason the Monte carlo estimator \bar{h}_n converges.

The Variance

$$\text{Var}(\hat{I}) = \frac{1}{n^2} \sum_i \text{Var} \left(\frac{h(X_i)f(X_i)}{g(X_i)} \right) \quad (4)$$

$$= \frac{1}{n} \text{Var} \left(\frac{h(X_i)f(X_i)}{g(X_i)} \right) \implies \quad (5)$$

$$\widehat{\text{Var}}(\hat{I}) = \frac{1}{n} \widehat{\text{Var}} \left(\frac{h(X_i)f(X_i)}{g(X_i)} \right). \quad (6)$$

Simple Example

Suppose we want to estimate $P(X > 5)$, where $X \sim N(0, 1)$.

Naive method (Monte carlo):

- ▶ Generate $X_1 \dots X_n \stackrel{iid}{\sim} N(0, 1)$
- ▶ Take the proportion $\hat{p} = \bar{X} > 5$ as your estimate

Importance sampling method:

- ▶ Sample from a distribution that gives high probability to the "important region" (the set $(5, \infty)$).
- ▶ Do re-weighting.

Importance Sampling Solution

Let $f = \phi_o$ and $g = \phi_\theta$ be the densities of the $N(0, 1)$ and $N(\theta, 1)$ distributions (θ taken around 5 will work). Then

$$p = \int I(u > 5) \phi_o(u) du \quad (7)$$

$$= \int \left[I(u > 5) \frac{\phi_o(u)}{\phi_\theta(u)} \right] \phi_\theta(u) du. \quad (8)$$

In other words, if

$$h(u) = I(u > 5) \frac{\phi_o(u)}{\phi_\theta(u)}$$

then $p = E_{\phi_\theta}[h(X)]$.

If $X_1, \dots, X_n \sim N(\theta, 1)$, then an unbiased estimate is $\hat{p} = \frac{1}{n} \sum_i h(X_i)$.

Naive Method (Monte Carlo) Solution

```
1 - pnorm(5) # gives 2.866516e-07
```

```
## [1] 2.866516e-07
```

```
## Naive method (Monte Carlo)
```

```
set.seed(1)
```

```
mySample <- 100000
```

```
x <- rnorm(n=mySample)
```

```
# pHat is a bernoulli proportion
```

```
pHat <- sum(x>5)/length(x)
```

```
# to calculate the variance,
```

```
# use p(1-p)/number of trials
```

```
sdPHat <- sqrt(pHat*(1-pHat)/length(x)) # gives 0
```

IS Solution

```
set.seed(1)
y <- rnorm(n=mySample, mean=5)
h <- dnorm(y, mean=0)/dnorm(y, mean=5) * I(y>5)
mean(h)                                     # gives 2.865596e-07
```

```
## [1] 2.865596e-07
```

```
sd(h)/sqrt(length(h))                     # gives 2.157211e-09
```

```
## [1] 2.157211e-09
```

What is the difference between the naive (monte carlo) and the IS solution?

Question

Is there some special choice regarding our importance region?

What happens when we look at the region $u > 100$?

```
set.seed(1)
y <- rnorm(n=mySample, mean=100)
h <- dnorm(y, mean=0)/dnorm(y, mean=100) * I(y>100)
(mean(h))
```

```
## [1] 0
```

```
# versus 2.865596e-07 (u>5)
(sd(h)/sqrt(length(h)))
```

```
## [1] 0
```

```
# versus 2.157211e-09 (u>5)
```

Importance Sampling Reference

For more reading on importance sampling, see the following reference: <https://www.amazon.com/Monte-Statistical-Methods-Springer-Statistics/dp/1441919392>.

This book also covers Monte carlo and Markov chain monte carlo methods extensively, so you may find this as a useful reference.

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 $\text{Uniform}(A)$, where A is complicated.

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

The Mandelbrot

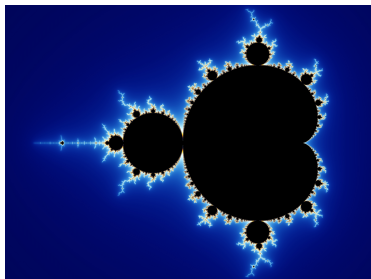


Figure 2: A complicated function A , called the Mandelbrot!

Exercise

- ▶ Suppose $A \subset B$.
- ▶ Let $Y_1, Y_2, \dots \sim \text{Uniform}(B)$ 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

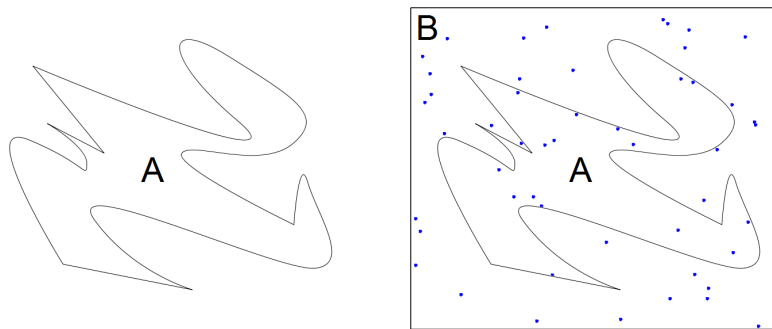


Figure 3: (Left) How to draw uniform samples from region A ? (Right) Draw uniform samples from B and keep only those that are in A .

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!

Suppose the density g is such that for some known constant M ,

$$Mg(x) \geq \ell(x)$$

for all x . Procedure:

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

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

2. Sample $X \sim q$, sample $Y \sim \text{Unif}(0, cq(X))$ (given X)
3. If $Y \leq \tilde{f}(X)$, $Z = X$, **otherwise we reject and return to step (2)**.

Visualizing just f

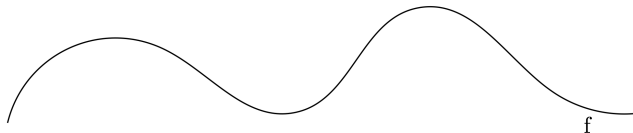


Figure 4: Visualizing just f .

Visualizing just f and \tilde{f}

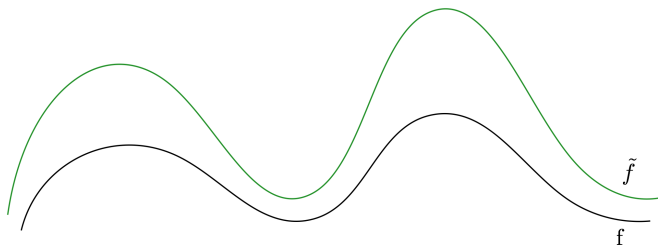


Figure 5: Visualizing just f and \tilde{f} .

Enveloping q over f

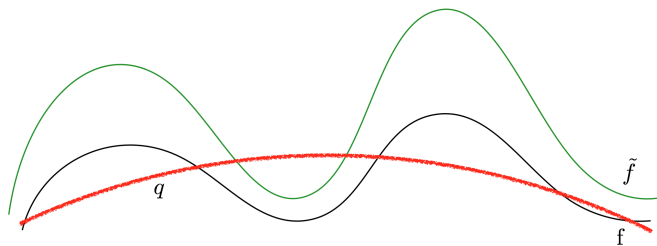


Figure 6: Visualizing f and \tilde{f} . Now we look at enveloping q over f .

Enveloping cq over \tilde{f}

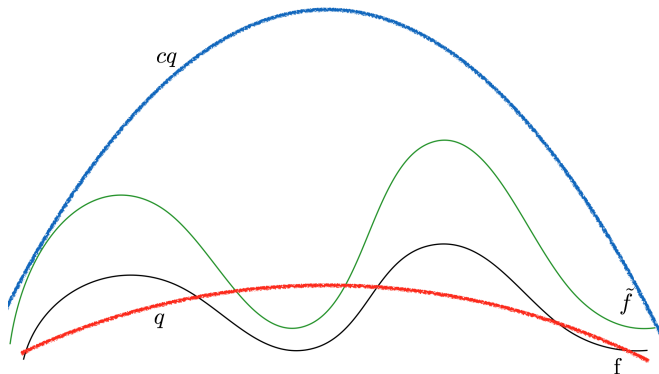


Figure 7: Visualizing f and \tilde{f} . Now we look at enveloping cq over \tilde{f} .

Recalling the sampling method and accept/reject step

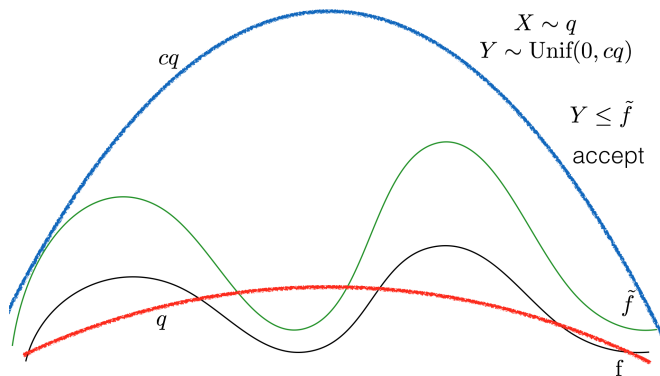


Figure 8: Recalling the sampling method and accept/reject step.

Entire picture and an example point X and Y

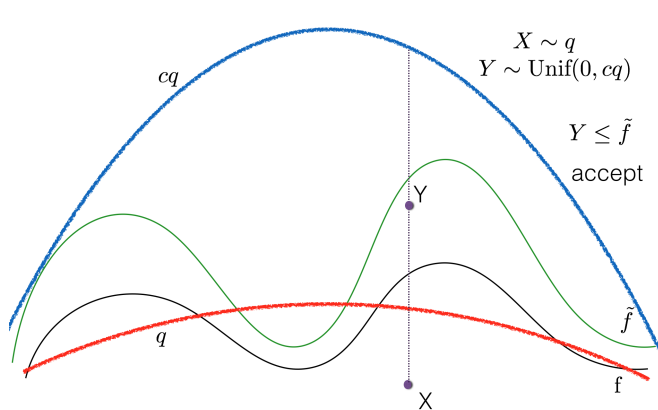


Figure 9: Entire picture and an example point X and Y .

Takeaways

- ▶ What is Monte Carlo (The naive method)
- ▶ Why do we use Monte Carlo over numerical integration?
- ▶ Importance sampling
- ▶ Rejection sampling

Detailed Takeaways

- ▶ Why do we use Monte Carlo?
- ▶ Why do we use importance sampling?
- ▶ Why do we use rejection sampling?
- ▶ When would we use a particular method above over another for a case study? This would be problem specific. So, if you know something about the important region at hand, you would typically use importance sampling. Rejection sampling and importance sampling will typically out perform naive Monte Carlo in terms of providing a lower variance estimate.
- ▶ In the next modules, we will learn about Markov chain Monte Carlo algorithms (MCMC), which are used for working in high dimensional spaces.
- ▶ Monte Carlo, importance sampling, and rejection sampling do not work well in high dimensions and thus we turn to MCMC. The trade-off is that 1) MCMC is a bit trickier to learn, 2) it will be slower to implement in practice, and 3) we can NEVER be assured that our sampler will reach the true target.

Exercise (Lab 5)

Consider the function

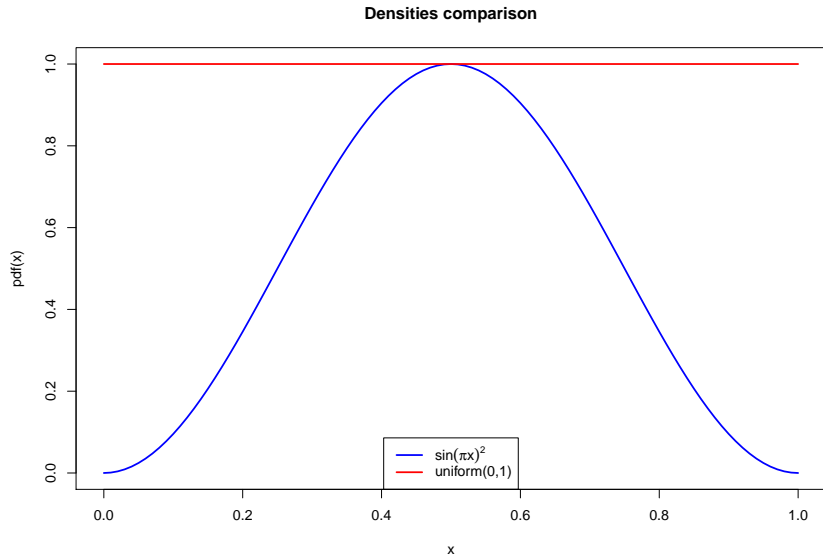
$$f(x) \propto \sin^2(\pi x), x \in [0, 1]$$

1. Plot the densities of $f(x)$ and the $\text{Unif}(0,1)$ on the same plot.
2. According to the rejection sampling approach sample from $f(x)$ using the $\text{Unif}(0,1)$ 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 $\text{Beta}(2,2)$ as an enveloping function. Compare your results with results in Task 3.
5. Do you recommend the Uniform or the $\text{Beta}(2,2)$ 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)
```

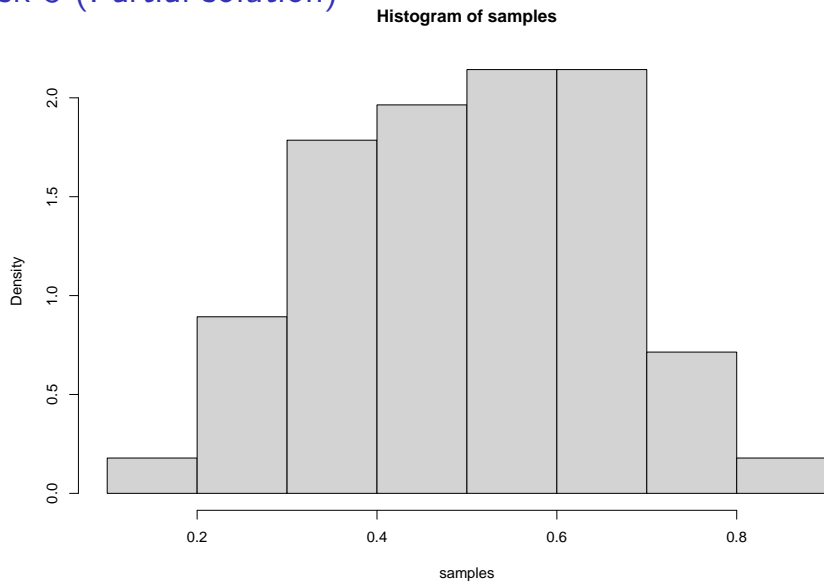
Task 1



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)



```
## [1] "Acceptance Ratio: 0.56"
```

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)

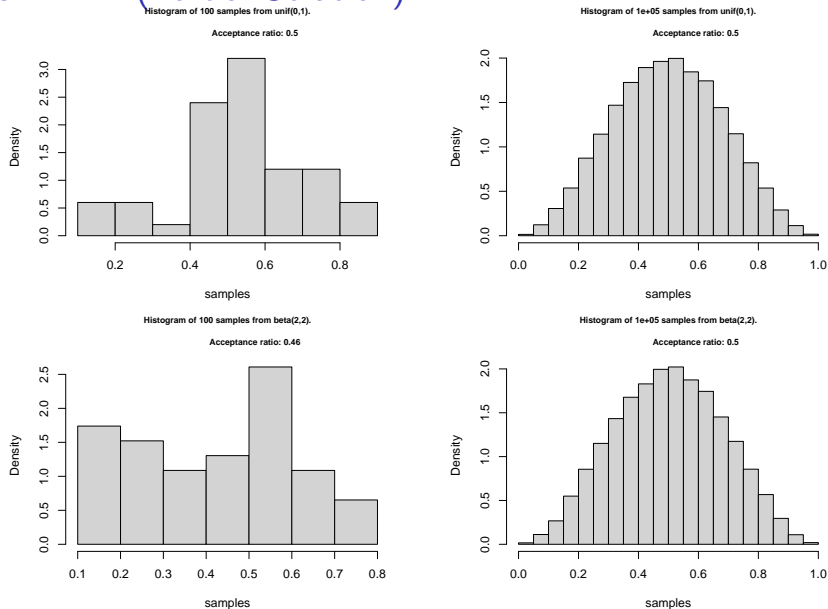


Figure 2: Rejection sampling for 100 versus 100,000 simulations

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)²

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

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 4.0.2 by xtable 1.8-4 package
```

```
## % Wed Sep 30 20:21:58 2020
```

```
## \begin{table}[ht]
```

```
## \centering
```

```
## \begin{tabular}{rrr}
```

```
## \hline
```

```
## & Mean & SE \\\
```

```
## \hline
```

```
## True value & 1.81 & \\\
```

```
## row.names
```

```
##
```

Histograms

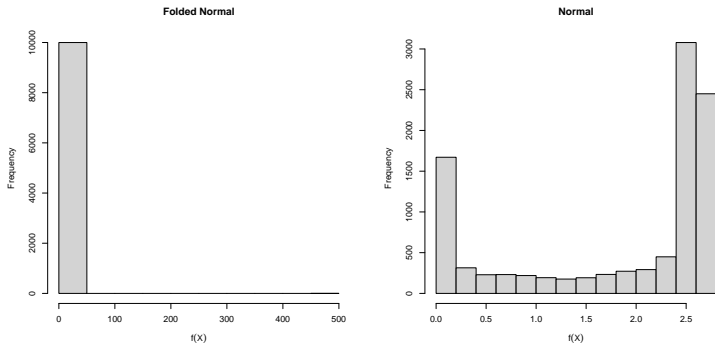


Figure 10: Histograms for the various Monte Carlo simulations.

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

- ▶ Closed notes, closed book.
- ▶ You will be given cover page with distributions.
- ▶ Please make sure that you arrive early and not late.
- ▶ 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.

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.

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 improper 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 hierarchical 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

- ▶ Two quizzes (ungraded).
- ▶ Practice exercises: <https://github.com/resteorts/modern-bayes/tree/master/exercises>
- ▶ Review homework exercises
- ▶ Next Tuesday, there there will be a timed question at the start of the review session that is meant to help prepare you for the exam. We will go over it in class.