STA 360/602L: Module 3.2

Rejection sampling; Importance sampling

Dr. Olanrewaju Michael Akande

Rejection sampling

- Rejection sampling and Importance sampling are one of the first steps into Monte Carlo analysis, in which simulated values from one distribution are used to explore another.
- Simulating from the "wrong distribution" can be incredibly useful as we will see in this module and also later in the course.
- Both are not used very often, but are still of practical interest in
  - fairly small problems, in terms of dimension,
  - in which the density of the distribution of interest can be easily evaluated, but when it is difficult to sample from directly, and
  - when it is relatively easy to identify and simulate from distributions that approximate the distribution of interest.
- Importance sampling and Rejection sampling use the same ideas, but the latter leads to exact corrections and so exact samples from the distribution of interest.

Rejection sampling

- Setup:
  - $p(\theta)$ is some density we are interested in sampling from;
  - $p(\theta)$ is tough to sample from but we are able to evaluate $p(\theta)$ as a function at any point; and
  - $g(\theta)$ is some proposal distribution or importance sampling distribution that is easier to sample from.
- Two key requirements:
  - $g(\theta)$ is easy to sample from; and
  - $g(\theta)$ is easy to evaluate at any point as is the case for $p(\theta)$.
- Usually, the context is one in which $g(\theta)$ has been derived as an analytic approximation to $p(\theta)$; and the closer the approximation, the more accurate the resulting Monte Carlo analysis will be.

Rejection sampling

- Procedure:

  1. Define $w(\theta) = p(\theta)/g(\theta)$.
  2. Assume that $w(\theta) = p(\theta)/g(\theta) < M$ for some constant M. If $g(\theta)$ represents a good approximation to $p(\theta)$, then M should not be too far from 1.
  3. Generate a *candidate* value $\theta \sim g(\theta)$ and **accept** with probability $w(\theta)/M$: if accepted, θ is a draw from $p(\theta)$; otherwise **reject** and try again.
     Equivalently, generate $u \sim U(0,1)$ independently of θ. Then **accept** θ as a draw from $p(\theta)$ if, and only if, $u < w(\theta)/M$.

- For those interested, the proof that all accepted θ values are indeed from $p(\theta)$ is on the next slide. We will not spend time on it.

- Clearly, we need M for this to work. However, in the case of truncated densities, we actually have M.

Proof for simple accept/reject

- We need to show that all accepted θ values are indeed from $p(\theta)$. Equivalently, show that $f(\theta|u < w(\theta)/M) = p(\theta)$.
- By Bayes' theorem,

$$f(\theta|u < w(\theta)/M) = \frac{\Pr(\theta \text{ and } u < w(\theta)/M)}{\Pr(u < w(\theta)/M)} = \frac{\Pr(u < w(\theta)/M \mid \theta)g(\theta)}{\Pr(u < w(\theta)/M)}.$$

- But,
  - $\Pr(u < w(\theta)/M \mid \theta) = w(\theta)/M$ since $u \sim U(0,1)$, and
  - $$\begin{aligned}\Pr(u < w(\theta)/M) &= \int \Pr(u < w(\theta)/M \mid \theta)g(\theta)d\theta \\ &= \int w(\theta)/Mg(\theta)d\theta = 1/M \int w(\theta)g(\theta)d\theta = 1/M \int p(\theta)d\theta = 1/M.\end{aligned}$$
- Therefore,

$$f(\theta|u < w(\theta)/M) = \frac{\Pr(u < w(\theta)/M \mid \theta)g(\theta)}{\Pr(u < w(\theta)/M)} = \frac{w(\theta)/Mg(\theta)}{1/M} = w(\theta)g(\theta) = p(\theta).$$

Rejection sampling for truncated densities

- The inverse CDF method works well for truncated densities but what happens when we can not or prefer not to write down the truncated CDF?
- Suppose we want to sample from $f_{[a,b]}(\theta)$, that is, a known pdf $f(\theta)$ truncated to $[a, b]$.
  - Recall that $f_{[a,b]}(\theta) \propto f(\theta)1[\theta \in [a, b]]$. Using the notation for rejection sampling, $p(\theta) = f_{[a,b]}(\theta)$ and $g(\theta) = f(\theta)$.
  - Set $1/M = \int_a^b f(\theta^\star)\mathrm{d}\theta^\star$, so that M is the normalizing constant of the truncated density.
  - Then, $w(\theta) = p(\theta)/g(\theta) = M1[\theta \in [a, b]] \leq M$ as required.

Rejection sampling for truncated densities

- We can then use the procedure on page 5 to generate the required samples.
- Specifically,
  - For each $i = 1, \ldots, m$, generate $\theta_i \sim f$. If $\theta_i \in [a, b]$, accept θ_i, otherwise **reject** and try again.
  - Easy to show that this is equivalent to accepting each θ_i with probability $w(\theta)/M$.

Example

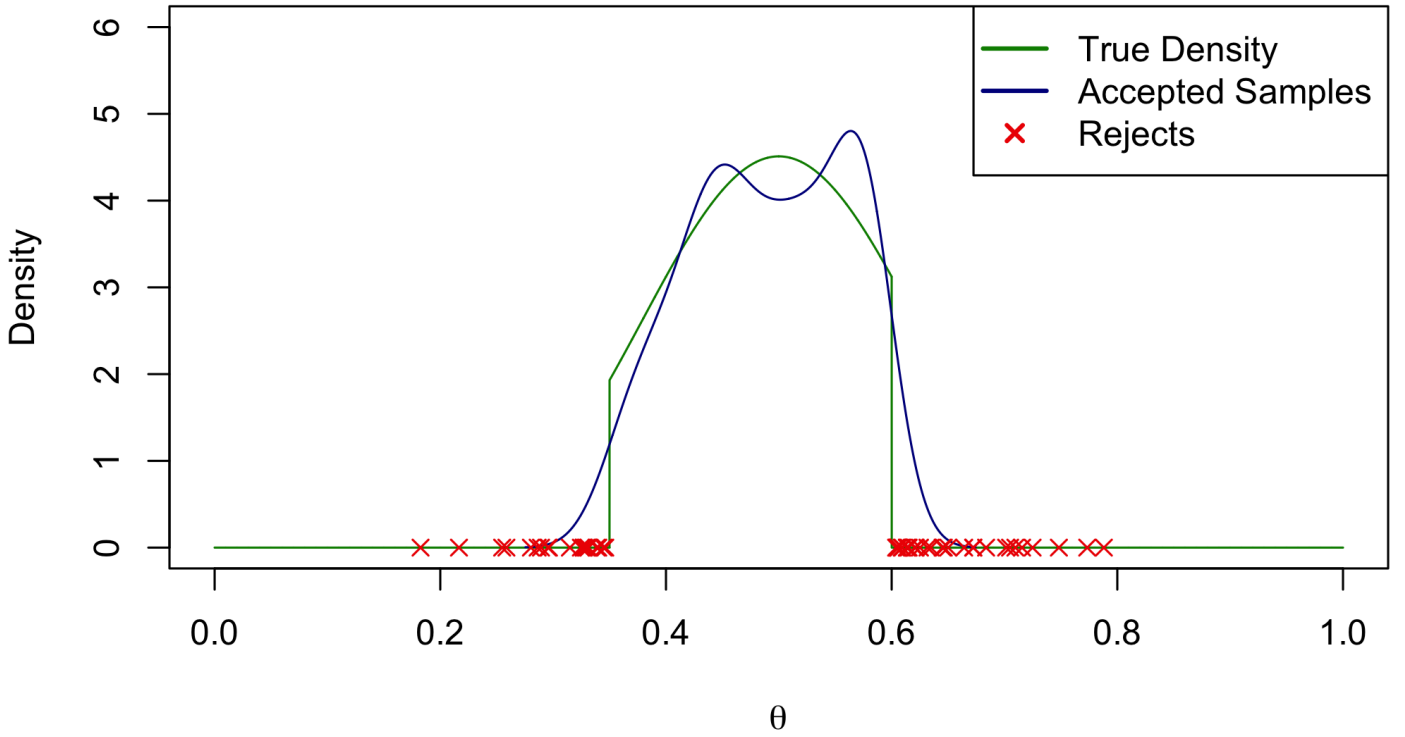
```
#Simple code for using rejection sampling to generate m samples
#from the Beta[10,10] density truncated to (0.35,0.6).
set.seed(12345)
#NOTE: there are more efficient ways to write this code!

#set sample size and reate vector to store sample
m <- 10000; THETA <- rep(0,m)
#keep track of rejects
TotalRejects <- 0; Rejections <- NULL
#now the 'for loop'
for(i in 1:m){
  t <- 0
  while(t < 1){
    theta <- rbeta(1,10,10)
    if(theta > 0.35 & theta < 0.6){
      THETA[i] <- theta
      t <- 1
    } else {
    TotalRejects <- TotalRejects + 1
    Rejections <- rbind(Rejections,theta)
  }
}
}
#Overall acceptance rate:
1 - TotalRejects/(m+TotalRejects)
```

```
## [1] 0.727802
```

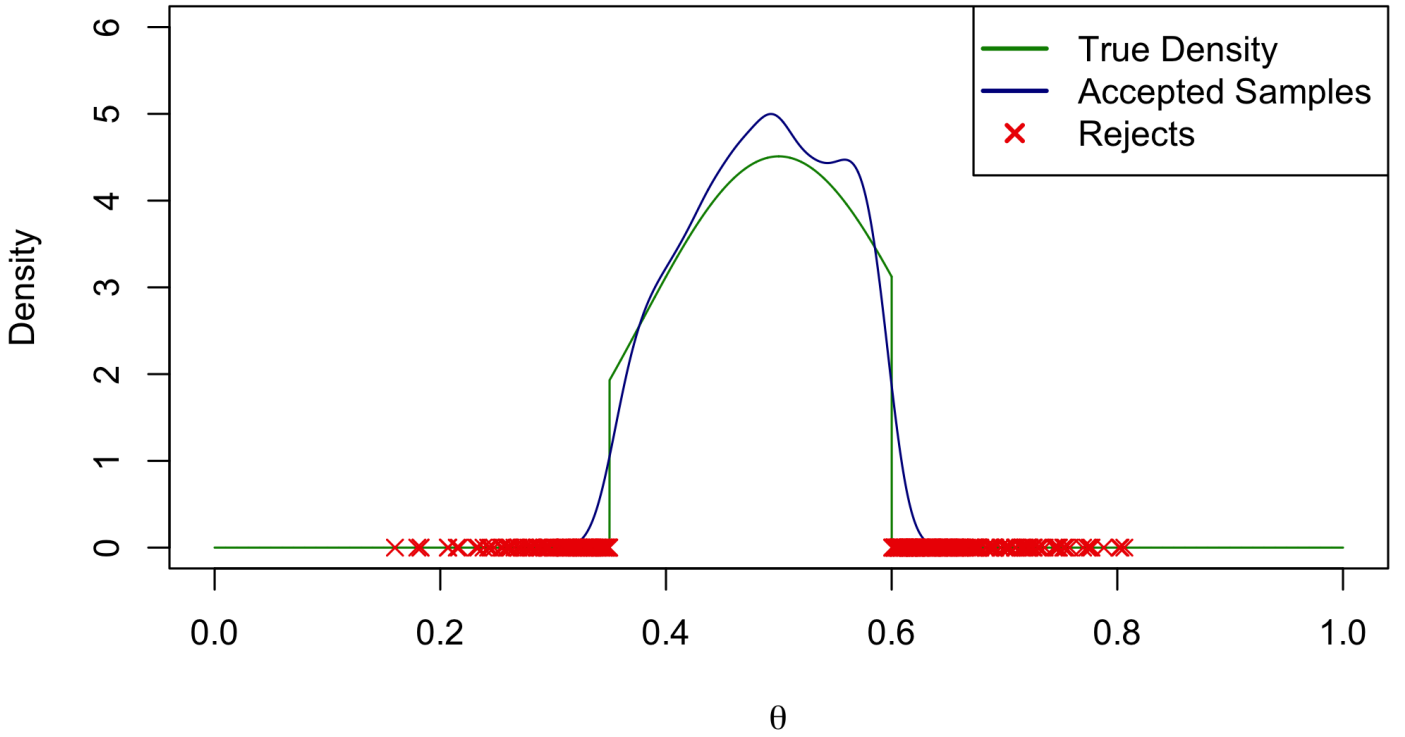
Example

How does our sample compare to the true truncated density? $m = 100$



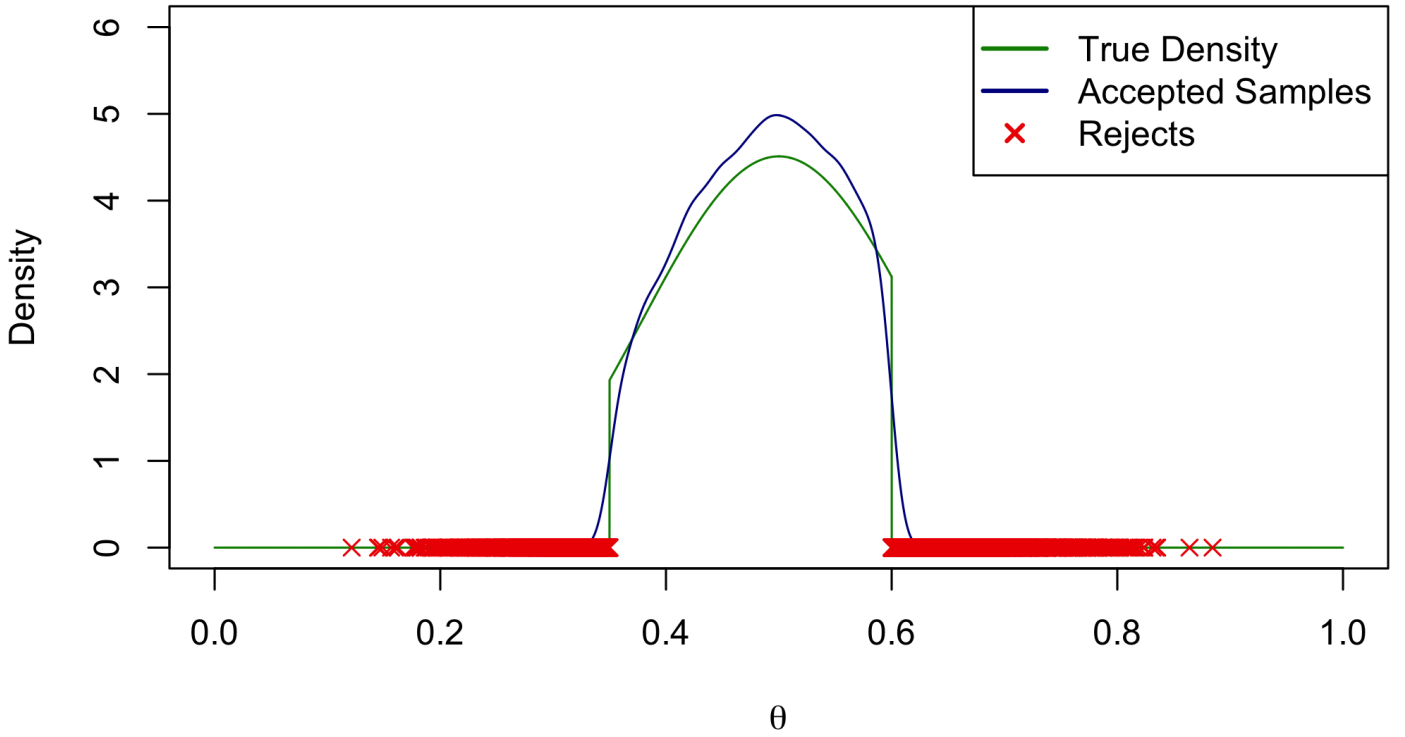
Example

How does our sample compare to the true truncated density? $m = 1000$



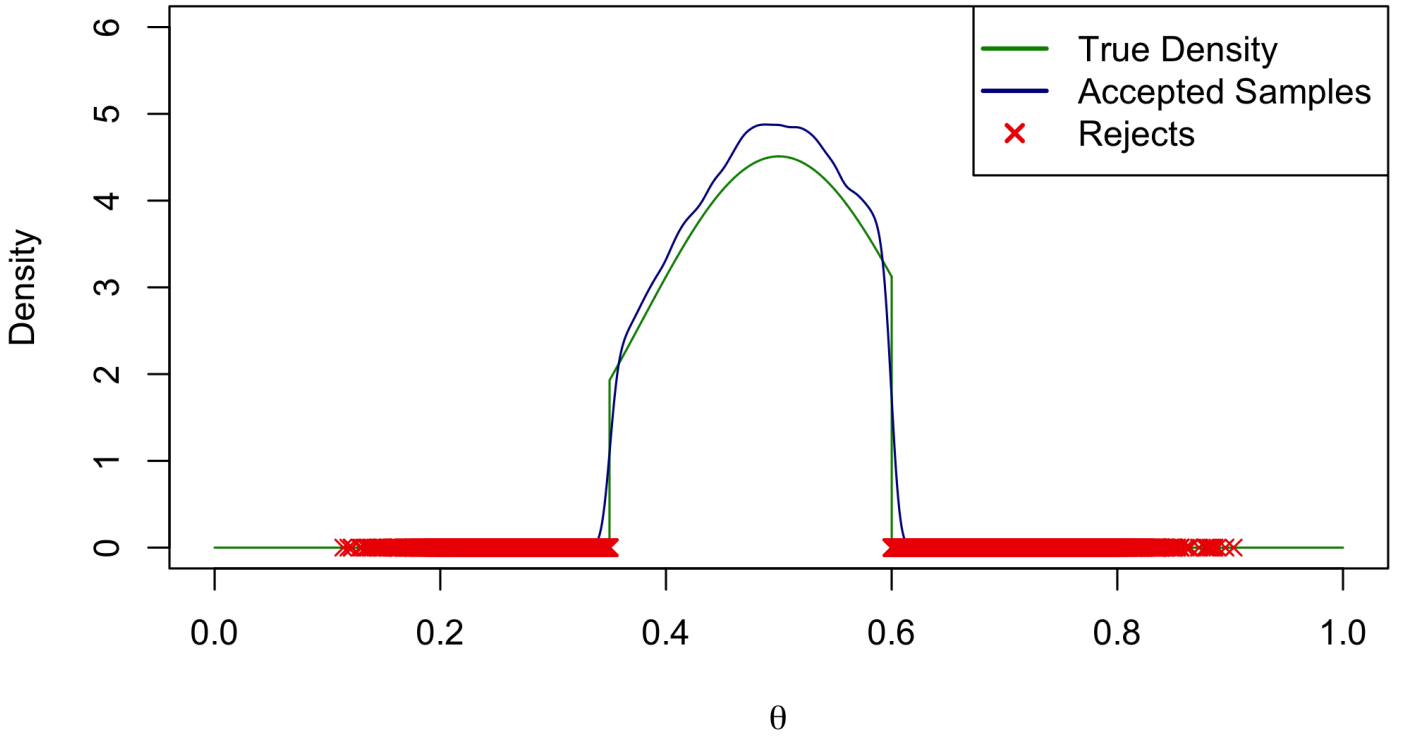
Example

How does our sample compare to the true truncated density? $m = 10000$



Example

How does our sample compare to the true truncated density? $m = 100000$



Comments

- Clearly less efficient than the inverse CDF method, which we already know how to use for this particular problem.
- When you can write down the truncated CDF, use the inverse CDF method instead.
- When you cannot, rejection sampling can be a possible alternative, as are many more sampling methods which we will not cover in this course.
- Anyway, generally, rejection sampling can still be very useful.
- Importance sampling is another related sampling method but we will not spend time on it. If you are interested, take a look at the next few slides. If not, feel free to skip.

Optional content from here on...

Importance sampling

- Interest lies in expectations of the form (instead of the actual samples)

$$H = \int h(\theta)p(\theta)d\theta,$$

- Write

$$H = \int h(\theta)w(\theta)g(\theta)d\theta \quad \text{with} \quad w(\theta) = p(\theta)/g(\theta)$$

that is, $\mathbb{E}[h(\theta)]$ under $p(\theta)$ is just $\mathbb{E}[h(\theta)w(\theta)]$ under $g(\theta)$.

- Using direct Monte Carlo integration

$$\bar{h} = \frac{1}{m}\sum_{i=1}^{m} w(\theta_i)h(\theta_i).$$

where $\theta_1, \ldots, \theta_m \overset{ind}{\sim} g(\theta)$. We are sampling from the "wrong" distribution.

Importance sampling

- The measure of "how wrong" we are at each simulated θ_m value is the importance weight

$$w(\theta_i) = p(\theta_i)/g(\theta_i).$$

- These ratios weight the sample estimates $h(\theta_i)$ to "correct" for the fact that we sampled the wrong distribution.
- See Lopes & Gamerman (Ch 3.4) and Robert and Casella (Ch. 3.3) for discussions of convergence and optimality.
- Clearly, the closer g is to p, the better the results, just as we had with rejection sampling.

Importance sampling

- Key considerations:
  - MC estimate $\bar{h}$ has the expectation H; and is generally almost surely convergent to H (under certain conditions of course but we will not dive into those).
  - $\mathbb{V}[\bar{h}]$ is often going to be finite in cases in which, generally, $w(\theta) = p(\theta)/g(\theta)$ is bounded and decays rapidly in the tails of $p(\theta)$.
  - Thus, superior MC approximations, are achieved for choices of $g(\theta)$ whose tails dominate those of the target $p(\theta)$.
  - That is, importance sampling distributions should be chosen to have tails at least as fat as the target (think normal distribution vs t-distribution).
  - Obviously require the support of $g(\theta)$ to be the same as, or contain, that of $p(\theta)$.
- These also clearly apply to rejection sampling too.

Importance sampling

- Problems in which $w(\theta) = p(\theta)/g(\theta)$ can be computed are actually rare.
- As you will see when we move away from conjugate distributions, we usually only know $p(\theta)$ up to a normalizing constant.
- When this is the case, simply "re-normalize" the importance weights, so that

$$\bar{h} = \frac{1}{m}\sum_{i=1}^{m} w_i h(\theta_i) \quad \text{where} \quad w_i = \frac{w(\theta_i)}{\sum_{i=1}^{m} w(\theta_i)}.$$

- Generally, in importance sampling, weights that are close to uniform are desirable, and very unevenly distributed weights are not.

What's next?

Move on to the readings for the next module!