Importance Sampling and Variance Reduction

Mark M. Fredrickson (`mfredric@umich.edu`)

Computational Methods in Statistics and Data Science (Stats 406)

AR Reject Rates

Recall two different **candidate distributions** in order to sample from a **truncated standard Normal** on [0,1].

- Standard uniform rejected about **15% of the candidates**.
- The density $g(y) = (2/3)(2 - y)$ rejected about **4%**

Could we do even better and reject 0%?

Changing variables

The reason we usually need collections of X is that we want to estimate:

$$\mathrm{E}\left(h(X)\right) = \int_{-\infty}^{\infty} h(x)f(x)\,dx$$

Remember the trick we used for integrating arbitrary functions:

$$\int_{-\infty}^{\infty} h(x)f(x)\,dx = \int_{-\infty}^{\infty} h(x)f(x)\frac{g(x)}{g(x)}\,dx = \mathrm{E}\left(h(Y)\frac{f(Y)}{g(Y)}\right)$$

for random variable Y with density $g(y)$.

Example: Tail probabilities

Using our rejection techniques is difficult to estimate **tail probabilities**

$$P(X \geq x)$$

since we get relatively few random samples from the tail.

Ex.: For $Z \sim N(0,1)$, what is $P(Z \geq 4.5) = E(I(Z \geq 4.5))$?

```
> k <- 100000
> sum(rnorm(k) >= 4.5)
```

```
[1] 1
```

Using a shifted exponential

Consider instead drawing from $Y = 4.5 + \text{Exp}(1)$ so that

$$g(y) = \exp(-(y - 4.5)), \quad y > 4.5$$

$$\mathrm{E}\left(I(Z \geq 4.5)\right) = \mathrm{E}\left(\frac{I(Y \geq 4.5)\phi(Y)}{g(Y)}\right) = \mathrm{E}\left(\frac{\phi(Y)}{g(Y)}\right)$$

```
> ys <- rexp(k) + 4.5
> ratios <- dnorm(ys) / (dexp(ys - 4.5))
> mean(ratios)

[1] 3.416e-06

> (truep <- pnorm(4.5, lower.tail = FALSE))

[1] 3.398e-06
```

Variances

Recall, we prefer estimators that are **efficient** (have lower variance).

Variance of the MC estimator (true, not estimated):

```
> truep * (1 - truep) / k
```

```
[1] 3.398e-11
```

Estimated variance of the importance sampling version:

```
> var(ratios) / k
```

```
[1] 1.951e-16
```

Terminology

We call using Monte Carlo to estimate $\mathrm{E}\,(h(Y)f(Y)/g(Y))$ as **importance sampling**.

- We call the distribution of Y the "envelope."
- The density of Y, $g(y)$, is the "importance function."
- We call the ratio $f(Y)/g(Y)$ the "importance weights."

Note: The importance weights $f(Y)/g(y)$ are very similar to the ratios we computed for the accept-rejection algorithm.

Picking Envelope Distribution

Recall that we had the strict requirement for accept-reject sampling:

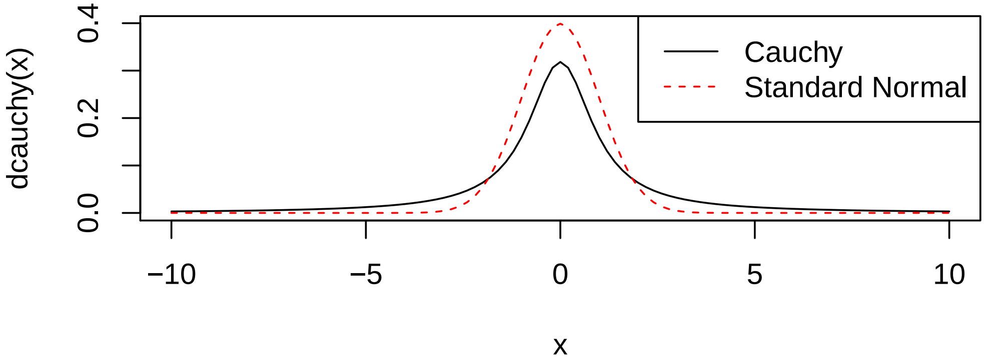
$$\frac{f(x)}{c\,g(x)} \leq 1, \quad \text{for some } c > 0 \text{ and all } x$$

While this restriction is not placed on importance sampling, we can get into trouble when the ratio $f(x)/g(x)$ can get very large (e. g., f has "fatter tails" than g).

In particular, we need the importance sampling estimator to have finite variance for the law of large numbers and the CLT to hold.

$$E_Y\left(h(Y)^2\frac{f(Y)^2}{g(Y)^2}\right) = \int_{-\infty}^{\infty} h(y)^2\frac{f(y)^2}{g(y)^2}g(y)\,dy = E_X\left(h(X)^2\frac{f(X)}{g(X)}\right) < \infty$$

Example: Targeting the Cauchy distribution with the Normal

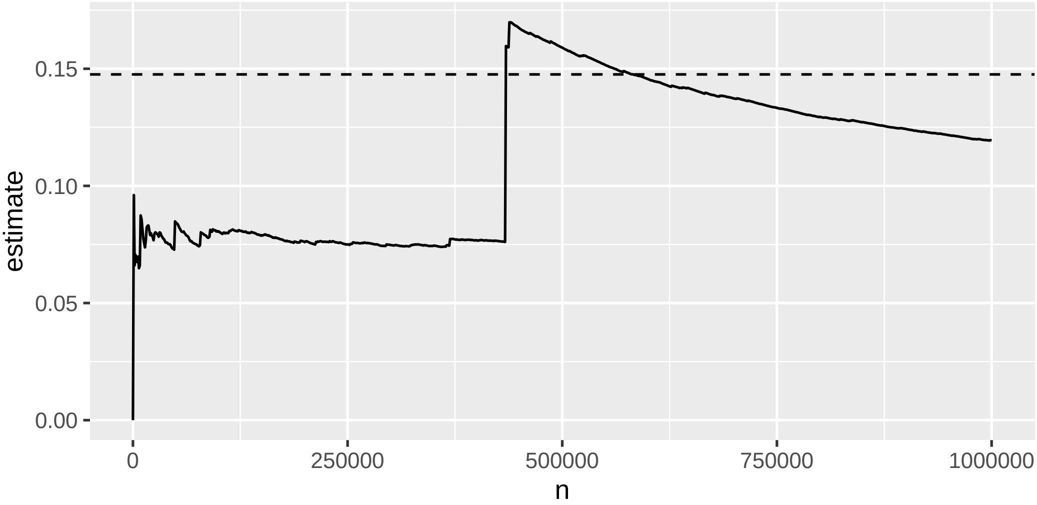


Cauchy from Normal

Let $C \sim \text{Cauchy}(0)$. Let's estimate $P(C \geq 2)$:

```
> k <- 1000000 # one million samples
> ys <- rnorm(k)
> iweights <- dcauchy(ys) / dnorm(ys)
> estimates <- cumsum(iweights * (ys >= 2)) / (1:k)
```

Plotting estimate vs. number of samples



Avoiding degenerate envelopes

Importance sampling for the Cauchy distribution will always be difficult due to the "fat tails".

Specifically, if the quantity

$$h(x)\frac{fx}{gx} < c, \quad \text{for some } c$$

then you should be ok.

The cases we'll consider in this class will be safe, but keep this in mind when using this a final project, e.g.

Efficiency of our estimators

Suppose we need to estimate $\theta = \mathrm{E}\,(h(X))$ and we have identified **two distributions that we can sample from**:

$$\mathrm{E}\,(h(X)) = \mathrm{E}\left(h(Y)\frac{f(Y)}{g(Y)}\right) = \mathrm{E}\left(h(W)\frac{f(W)}{d(W)}\right)$$

(where g is the PDF of Y and d is the PDF of W). How do we pick one or the other to use?

Notice that in large samples:

$$\hat{\theta}_1 \sim N\left(\theta, \mathrm{Var}\left(\hat{\theta}_1\right)\right), \quad \hat{\theta}_2 \sim N\left(\theta, \mathrm{Var}\left(\hat{\theta}_2\right)\right)$$

They only differ in the variance terms!

Variance terms

Recall that for **IID data**, the variance of the sample mean is

$$\mathrm{Var}\left(\hat{\theta}_1\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} h(Y_i)\frac{f(Y_i)}{g(Y_i)}\right) = \frac{1}{n}\mathrm{Var}\left(h(Y)\frac{f(Y)}{g(Y)}\right)$$

$$\begin{aligned}\mathrm{Var}\left(h(Y)\frac{f(Y)}{g(Y)}\right) &= \mathrm{E}\left(h(Y)^2\frac{f(Y)^2}{g(Y)^2}\right) - \left[\mathrm{E}\left(h(Y)\frac{f(Y)}{g(Y)}\right)\right]^2 \\ &= \int h(y)^2\frac{f(y)^2}{g(y)^2}\, g(y)\, dy - \theta^2\end{aligned}$$

Further decomposing variance

Notice that for any function $t(x)^2 = |t(x)|\,|t(x)|$.

$$\begin{aligned}
\mathrm{Var}\left(h(Y)\frac{f(Y)}{g(Y)}\right) &= \int h(y)^2 \frac{f(y)^2}{g(y)^2}\, g(y)\, dy - \theta^2 \\
&= \int |h(y)||f(y)| \frac{|h(y)||f(y)|}{g(y)}\, dy - \theta^2 \\
&= \int |h(y)| f(y) \frac{|h(y)| f(y)}{g(y)}\, dy - \theta^2
\end{aligned}$$

Picking a $g(y)$ to minimize variance

Observe that that because $|h(x)|f(x) > 0$, the following is a proper density:

$$g(y) = \frac{|h(y)|f(y)}{\int_{-\infty}^{\infty} |h(y)|f(y)\,dy}$$

$$\begin{aligned}
\text{Var}\left(h(Y)\frac{f(Y)}{g(Y)}\right) &= \int |h(y)|f(y)\frac{|h(y)|f(y)}{g(y)}\,dy - \theta^2 \\
&= \int |h(y)|f(y)\left[\int |h(y)|f(y)\,dy\right]\,dy - \theta^2 \\
&= \left[\int |h(y)|f(y)\,dy\right]\left[\int |h(y)|f(y)\,dy\right] - \theta^2 \\
&= \left[\int |h(y)|f(y)\,dy\right]^2 - \theta^2
\end{aligned}$$

and if $h(y) > 0$ for all y where $f(y) > 0$, then the **variance would be zero!**

Using minimum variance g

So we should pick

$$g(y) = \frac{|h(y)|f(y)}{\int_{-\infty}^{\infty} |h(y)|f(y)\,dy}$$

Why is this not helpful? We would need to know $\int |h(x)|f(x)\,dx$, effectively θ.

Why is this helpful? We can try to pick $g(y) \approx c|h(y)|f(y)$

Example: Truncated Normal

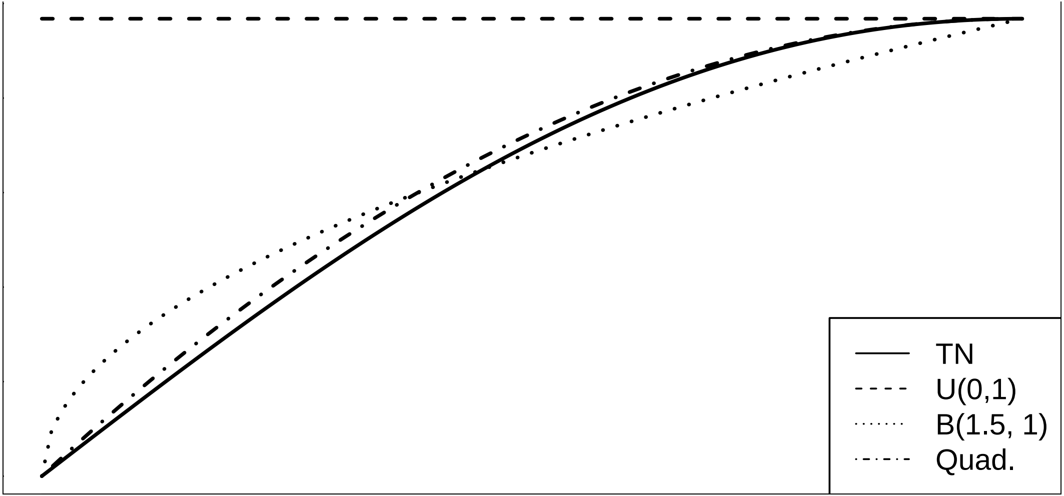
Again, we'll consider the **truncated Normal distribution** $Z \mid 0 \leq Z \leq 1$, $Z \sim N(0,1)$.

Specifically, we want to estimate $\mathrm{E}\left(X\right)$. We could pick many bounded distributions on $(0,1)$. Which would be best?

- Uniform(0,1)
- Beta(1.5, 1)
- Quadratic: $g(x) = (3/2)(2x - x^2)$

Goal: we want a density $\propto |x|\phi(x)$.

Graphing $|h(x)|f(x)/g(x)$



Using Beta(1.5, 1) and the Quadratic Density

```
> tn <- function(x) { dnorm(x) / (pnorm(1) - pnorm(0))}
> k <- 10000
> yg <- qx(runif(k))
> iwg <- tn(yg) / gx(yg) ## h(y) f(y) / g(y)
> mean(yg * iwg)

[1] 0.4597

> yr <- rbeta(k, 1.5, 1)
> iwr <- tn(yr) / dbeta(yr, 1.5, 1)
> mean(yr * iwr)

[1] 0.4587
```

Comparing Variance

```
> varg <- var(yg * iwg) # variance of single h(Y) f(Y) / g(Y) term
> varr <- var(yr * iwr)
```

Percentage decrease in variance:

```
> (varr - varg) / varr
```

```
[1] 0.9526
```

Relative CI width:

```
> diff(t.test(yg * iwg)$conf.int) / diff(t.test(yr * iwr)$conf.int)
```

```
[1] 0.2177
```

Importance Sampling Resampling

Generating samples

So far, we have been computing **expectations**. What if we need **samples**? (E.g., for evaluating **estimators** or **hypothesis tests**.)

Using **importance sampling re-sampling** (ISRS) we use the **importance weights** to pick samples.

Generate m samples Y_i. Pick a number J between 1 and m with probability*

$$\frac{1}{m}\frac{f(Y_i)}{g(Y_i)}, \quad i = 1, \ldots, m$$

and set $X = Y_J$ (notice: all Y are random and J is random).

Proof

The distribution of Y_J is the same as X. We need to show $P(Y_J \leq t) = P(X \leq x)$.
How can the event $\{Y_J \leq t\}$ occur?

$$\begin{aligned}
\Pr(Y_J \leq t) &= P\left([J = 1, Y_1 \leq t] \text{ or } [J = 2, Y_2 \leq t] \text{ or } \ldots\right) && \text{(def. } Y_J\text{)} \\
&= \sum_{i=1}^{m} \Pr(Y_i \leq t, J = i) && \text{(disjoint events)} \\
&= \sum_{i=1}^{m} \int_{-\infty}^{t} \frac{1}{m} \frac{f(y)}{g(y)} g(y)\, dy && \text{(joint dist. } Y_i,\ J\text{)} \\
&= \int_{-\infty}^{t} f(y)\, dy = \Pr(X \leq t) && f \text{ is density of } X
\end{aligned}$$

Normalizing the weights

Problem: the weights $m^{-1}f(Y_i)/g(Y_i)$ may not be less than 1 and may not sum to 1.

We can normalize them as:

$$\omega_i = \frac{m^{-1}f(Y_i)/g(Y_i)}{\sum_{i=1}^{m} m^{-1}f(Y_i)/g(Y_i)}$$

The resulting Y_J will not have exactly the same distribution as X, but when m is large, the difference can be very small. This bias is the tradeoff for accepting all samples.

Example: Drawing from a truncated distribution of Z

An example that can't be directly estimated using importance sampling alone:

$$\text{median}(Z \mid Z \geq 4.5)$$

We'll use the Exp(1) samples and their importance weights to estimate the conditional mean.

```
> ys <- rexp(k) + 4.5
> imp_weights <- dnorm(ys) / (dexp(ys - 4.5))
> omega <-  imp_weights / sum(imp_weights) ## the 1/m term gets canceled
> xs <- sample(ys, replace = TRUE, prob = omega)
> median(xs)
```

```
[1] 4.646
```

Conclusion: Standard Normal tails go to zero really fast! 50% of all Z larger than 4.5 are within .15 of 4.5.

Densities Known to a Constant

Unnomralized Densities

Recall from **importance sampling re-sampling** we used **normalized weights**:

$$\omega_i = \frac{f(Y_i)/g(Y_i)}{\sum_{j=1}^n f(Y_j)/g(Y_j)}$$

Suppose **did not** actually know f, but only knew

$$f^*(x) \propto f(x) \quad \text{i.e.,} f(x) = cf^*(x), \quad \text{s. t.} \int_{-\infty}^{\infty} f^*(x)\, dx = c$$

Notice that the weights would not change:

$$\omega_i = \frac{f(Y_i)/g(Y_i)}{\sum_{j=1}^n f(Y_j)/g(Y_j)} = \frac{cf^*(Y_i)/g(Y_i)}{\sum_{j=1}^n cf^*(Y_j)/g(Y_j)} = \frac{f^*(Y_i)/g(Y_i)}{\sum_{j=1}^n f^*(Y_j)/g(Y_j)}$$

Implication: we can use ISRS to draw from f^* (essentially, we'll estimate c).

Using ISRS

Previously we saw IRSR for **generating samples**. Here we will use if for **Monte Carlo estimation**. For regular IS, we computed:

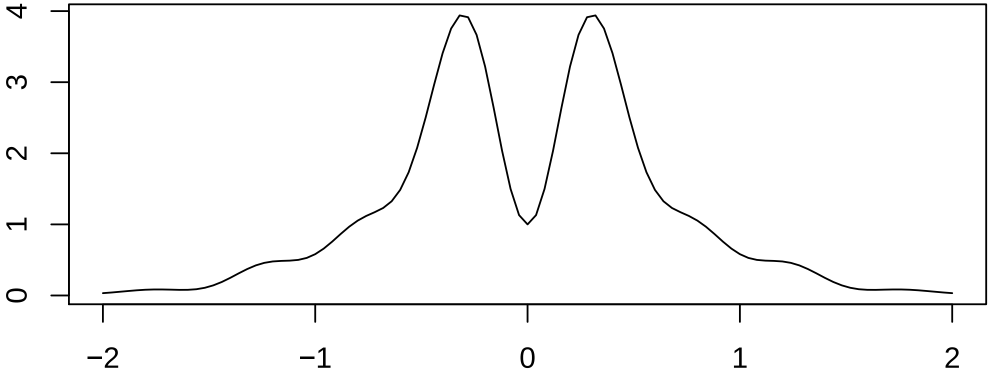
$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} h(Y_i) \frac{f(Y_i)}{g(Y_i)}$$

for ISRS, we use

$$\tilde{\theta} = \sum_{i=1}^{n} h(Y_i)\, \omega_i = \frac{1}{n} \sum_{i=1}^{n} h(Y_i)\, (n\omega_i)$$

Example: "Rabbit" distribution

$$f(x) \propto \exp(-x^2/2)\left[\sin(6x)^2 + 3\cos(x)^2\sin(4x)^2 + 1\right] = f^*(x), \quad -\infty < x < \infty$$



Estimating the variance of the rabbit distribution

Since the distribution is symmetric about 0, the mean is clearly 0, so the variance is:

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

Using a standard normal as the envelope:

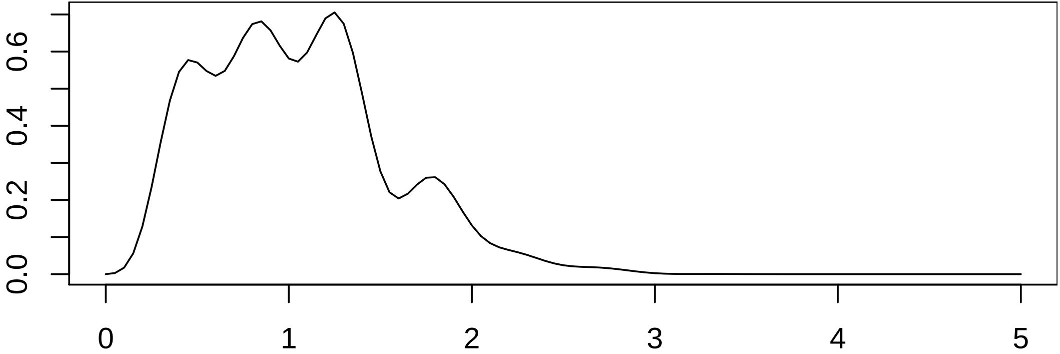
```
> k <- 10000
> ys <- rnorm(k)
> as <- fstar(ys) / dnorm(ys)
> omegas <- as / sum(as)
> reweighted_ys2 <- ys^2 * (k * omegas)
> mean(reweighted_ys2)
```

```
[1] 0.387
```

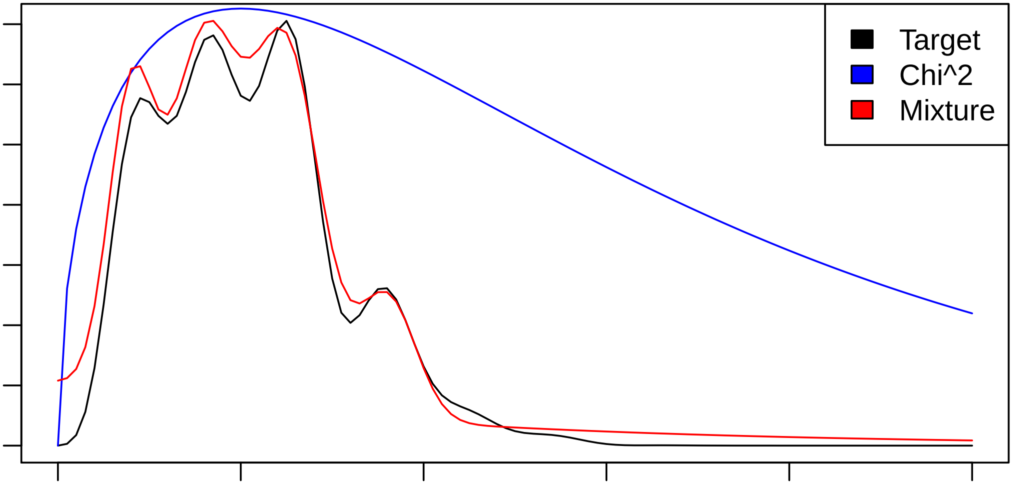
Rabbit distribution: Plotting $|h(x)|f(x)$

While achieving a **variance of zero is probably impossible**, we can tune our envelope by making it as **close to** $|h(x)|f(x)$ as possible.

Recall our goal of estimating $\mathrm{Var}(X) = \mathrm{E}\left(X^2\right)$ for the "rabbit distribution".



A few choices



$\chi^2(3)$ and Mixture of truncated Normals and Exponential

Candidate 1 is a χ^2 on 3 degrees of freedom.

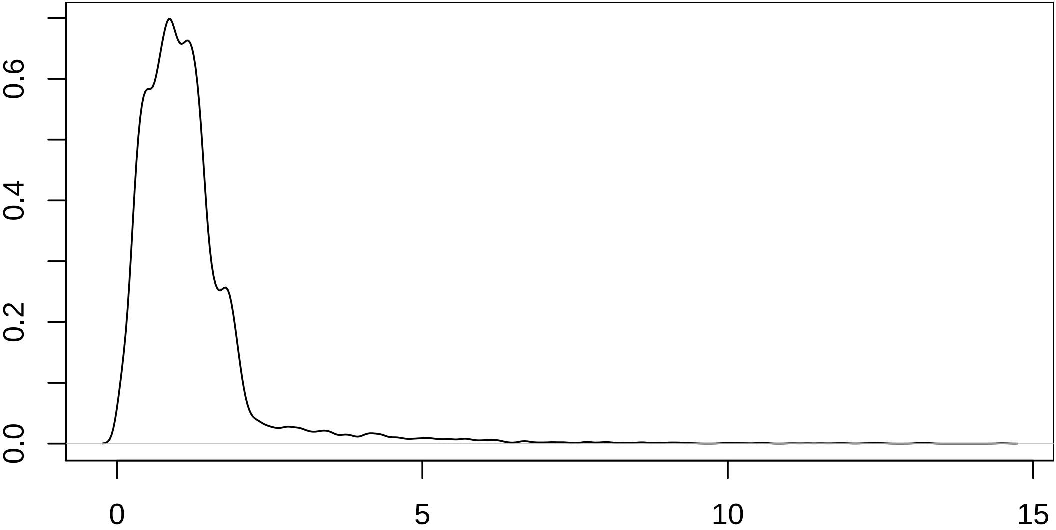
Candidate two is a mixture of truncated Normals and an Exponential:

$$0.15\, N_{[0,\infty)}\left(\frac{2}{5}, \frac{1}{8^2}\right) + 0.28\, N_{[0,\infty)}\left(\frac{4}{5}, \frac{3^2}{16^2}\right) +$$
$$0.28\, N_{[0,\infty)}\left(\frac{5}{4}, \frac{3^2}{16^2}\right) + 0.08\, N_{[0,\infty)}\left(\frac{9}{5}, \frac{5^2}{32^2}\right) +$$
$$0.21\, \mathrm{Exp}\left(\frac{1}{2}\right)$$

Mixture (unormalized) Density

```
function(x) {
    0.15 * (x >= 0) * dnorm(x, mean = 2/5, sd = 1/8) +
    0.28 * (x >= 0) *  dnorm(x, mean = 4/5, sd = 3/16) +
    0.28 * (x >= 0) *  dnorm(x, mean = 5/4, sd = 3/16) +
    0.08 * (x >= 0) *  dnorm(x, mean = 9/5, sd = 5/32) +
    0.21 * dexp(x, 1/2)
}
<bytecode: 0x55f8666698e8>
```

Validating Mixture Sampler



χ^2 estimator

```
> chi3 <- rchisq(k, df = 3)
> chi3_ratios <- fstar(chi3) / dchisq(chi3, df = 3)
> chi3_omegas <- chi3_ratios / sum(chi3_ratios)
> chi3_x2 <- chi3^2 *  (k * chi3_omegas) #
> (chi3_est <- mean(chi3_x2))

[1] 0.3904
```

Mixture estimator

```
> mixs <- rmix(k)
> mixs_ratios <- fstar(mixs) / dmix_star(mixs)
> mixs_omegas <- mixs_ratios / sum(mixs_ratios)
> mixs_x2 <- mixs^2 * (k * mixs_omegas)
> (mixs_est <- mean(mixs_x2))

[1] 0.3893
```

Stacking up the variances

We've already seen method using a **reweighted standard Normal**.

```
> var(reweighted_ys2)
[1] 0.05506
> var(chi3_x2)
[1] 0.2351
> var(mixs_x2)
[1] 0.01227
```

So the *Normal* envelope beats the χ^2 envelope, but the mixture beats both.

Other Variance Reduction Techniques

Additional techniques

Importance sampling is a very powerful tool, but it is not the only method of **reducing the variance of estimates**.

We will briefly look at three others:

- Antithetic variables
- Control variates
- Stratified sampling

Antithetic variables

So far, we have been generating, **independent, identically distributed** collections of random variables.

With **antithetic variables**, we introduce **dependence** in way to **decrease the variance** of the estimator.

The trick is to find a way to generate the dependence in a very particular way.

Thinking more about dependence and variance

Suppose we have an estimator for θ that is a sample mean of T_i: $\hat{\theta} = m^{-1} \sum_{i=1}^{m} T_i$ (e.g., $T_i = h(Y_i) f(Y_i)/g(Y_i)$).

The variance of this estimator is

$$\text{Var}(\hat{\theta}) = \frac{1}{m^2} \left[\sum_{i=1}^{m} Var(T_i) + \sum_{i \neq j} \text{Cov}(T_i, T_j) \right]$$

When T_i is independent of T_j is the covariance term is zero.

What if we could generate **negatively correlated T_i and T_j**?

Example: Uniform random variables

Suppose that we are going to use the **inversion method** to generate a variable X for which we want to estimate $E(X)$.

You proved that $U' = 1 - U$ has the same distribution as $U \sim U(0,1)$ and

$$\mathrm{Cov}(U, U') = E(U(1-U)) - (1/4) = E(U) - E(U^2) - 1/4 = 1/4 - 1/3 = -1/12$$

Since Q_X (the quantile function for X) is **monotonic**, $Q_X(U)$ and $Q_X(1-U)$ are also **negatively correlated**.

Example continued

Suppose we have

$$f(x) = \frac{x^2}{9\theta^3}, \quad 0 \le x \le 3\theta$$

For $\theta = 1/3$, let's estimate the mean (which we can compute as $3/4$):

$$f(x) = 3x^2 \Rightarrow F(t) = t^3 \Rightarrow Q(p) = p^{1/3}$$

In order to reduce the variance, we'll use the following identity:

$$\hat{\theta} = \frac{1}{m}\left[\sum_{i=1}^{m/2} Q(U_i) + \sum_{i=1}^{m/2} Q(1 - U_i)\right] = \frac{1}{m/2}\sum_{i=1}^{m/2}(Q(U_i) + Q(1 - U_i))/2$$

IID solution

```
> iids <- runif(10000)^(1/3)
> mean(iids)

[1] 0.7518

> (est_var_iid <- var(iids) / 10000)

[1] 3.721e-06
```

Antithetic variables

To keep the sample size the same, let's only generate 5000 uniforms, then supplement those with 5000 copies of $1 - U$.

```
> tmp <- runif(5000)
> antis <- (tmp^(1/3) + (1 - tmp)^(1/3)) / 2
> mean(antis)
```

```
[1] 0.7499
```

```
> (est_var_anti <- var(antis) / 5000)
```

```
[1] 4.76e-07
```

Percent variance reduction

```
> 1 - (est_var_anti / est_var_iid)
[1] 0.8721
```

An 86% reduction in variance (and we only had to generate 1/2 as many random variables).

Note: it is not always obvious how to generate the antithetic variable pairs, but when you can, they are a very powerful tool.

Control variates

Consider the following

- Want to estimate $\theta = \mathrm{E}\left(g(X)\right)$
- Happen to know $\mu = \mathrm{E}\left(f(X)\right)$
- and know $\mathrm{Cov}(g(X), f(X)) \neq 0$ (i.e., $g(X)$ and $f(x)$ are correlated).

We could consider an estimator (of a single observation) of the form:

$$\hat{\theta}_c = g(X_1) + c(f(X_1) - \mu)$$

Observe that

$$E(\hat{\theta}_c) = \theta + c(\mu - \mu) = \theta \quad \text{(unbiased)}$$

Variance

What is the variance of $\hat{\theta}_c$?

$$\mathrm{Var}(\hat{\theta}_c) = \mathrm{Var}(g(X_1)) + c^2\mathrm{Var}(f(X_1)) + 2c\mathrm{Cov}(g(X_1), f(X_1))$$

We can minimize this function by picking

$$c = -\frac{\mathrm{Cov}(g(X), f(X))}{\mathrm{Var}(f(X))}$$

We call $Y = f(X)$ the **control variate**. Antithetic variables are a **special case**.

We don't have time for an example, but see **sec. 5.5 of SCR**.

Stratified Sampling

As we have seen, it can be very useful to closely approximate the function $|h(x)|f(x)$ in picking an envelope distribution for importance sampling.

To **target regions of $|h(x)|f(x)$**,

- **break up the integral** into k regions, such that $A_i \cap A_j = \emptyset$ and $\bigcup A_i = (-\infty, \infty)$ (the A_i are disjoint).
- Estimate $\mathrm{E}\left(I(X \in A_i)h(X)\right)$ using m_i samples ($\hat{\theta}_i$)
- Combine the estimates:

$$\hat{\theta}_k = \frac{1}{\sum_{i=1}^{k} m_i} \sum_{i=1}^{k} m_i \hat{\theta}_i$$

If the variances of the individual portions are smaller than $\mathrm{Var}(\hat{\theta})$ on average, the overall variance will be smaller for $\hat{\theta}_k$ than $\hat{\theta}$.

Summary

- Importance sampling is a very flexible technique that is used widely.
- The key is picking a good envelope distribution that matches $|h(x)|f(x)$
- Control variates are a very powerful approach when possible, but requires more knowledge of the problem.
- These are often combined in practice.
- Stratified estimation is also relatively easy to implement and complements importance sampling.