Importance Sampling and Radon-Nikodym Derivatives

Steven R. Dunbar

Outline

1. Sampling with respect to 2 distributions
2. Rare Event Simulation
3. Absolute Continuity
4. Examples and Applications
5. Statistical Interpretation

More than one way to evaluate a statistic

A statistic for X with pdf $u(x)$ is

$$A = \mathbb{E}_u\left[F(X)\right] = \int F(x)u(x)\; \mathrm{d}x$$

Suppose $v(x)$ is another probability density such that

$$L(x) = \frac{u(x)}{v(x)}$$

is well-defined.

Then

$$A = \mathbb{E}_v\left[F(X)L(X)\right] = \int F(x)\frac{u(x)}{v(x)}v(x)\; \mathrm{d}x$$

Likelihood ratio

The likelihood ratio is

$$L(x) = \frac{u(x)}{v(x)}$$

so

$$A = \mathbb{E}_u\left[F(X)\right] = \mathbb{E}_v\left[F(X)L(X)\right]$$

To get A we can either:

- take samples with density $u(x)$ calculate with F; or
- take samples with density $v(x)$, calculate with $F \cdot L$.

Good choice of likelihood ratio

How should we seek $v(x)$ and corresponding $L(x)$?

Reduction of variance is a good criterion: Want

$$\mathrm{Var}_v\left[F(X)L(X)\right] = \mathbb{E}_v\left[(F(X)L(X))^2\right] - A^2$$

less than

$$\mathrm{Var}_u\left[F(X)\right] = \mathbb{E}_u\left[(F(X))^2\right] - A^2$$

Example: Large values for a normal r.v.

Suppose $X \sim N(0,1)$ and we wish to evaluate $\mathbb{P}[X > 4]$

R code for probability:

```
1 - pnorm(4)
```

The output

```
[1] 3.167124e-05
```

Sampling

Code for sampling

```
sample <- rnorm(10^5)
sum( sample > 4 )
sample[ which(sample > 4) ]
```

The output

```
[1] 6
[1] 4.110055 4.970314 4.026803 4.133050 4.014239
```

Most of the sample doesn't count, "wasted effort".
The samples with $X > 4$ are only a little larger than 4.

A better choice of sampling PDF

Draw samples from $N(4,1)$ instead. Put most of the weight in sampling near 4.

Likelihood ratio is

$$L(x) = \frac{u(x)}{v(x)} = \frac{k\mathrm{e}^{-x^2/2}}{k\mathrm{e}^{-(x-4)^2/2}} = \mathrm{e}^{4^2/2}\mathrm{e}^{-4x}.$$

New sampling

The v-method, with sample draw according to the new pdf:

$$\mathbb{P}_{N(0,1)}[X > 4] = \int_4^\infty L(X)v(x)\,\mathrm{d}x \approx \frac{1}{N}\sum_{v_k>4} L(v_k) = \frac{\mathrm{e}^{4^2/2}}{N}\sum_{v_k>4}\mathrm{e}^{-4v_k}$$

New sampling example

Code for sampling

```
vsample <- rnorm(10^5,4,1)
usevsample <- vsample[which(vsample > 4)]
Pv <- exp(4^2/2)*sum(exp(-4*usevsample))/10^5
Pv
```

The output

```
[1] 3.17203e-05
```

Intuition behind Importance Sampling

About half the samples are counted, but they are counted with a small weight.

A lot of small weight hits give a lower variance estimator than a few large weight hits.

Digression: Reduction of Variance

Variance for the naive sampling:

$$P = \mathbb{P}_{N(0,1)}\left[X > 4\right] = \int \mathbf{1}_{[X>4]}(x) \frac{\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}} \, \mathrm{d}x$$

$$\begin{aligned} \operatorname{Var}_u\left[F(X)\right] &= \mathbb{E}_u\left[(F(X))^2\right] - P^2 \\ &= \int (\mathbf{1}_{[X>4]}(x))^2 \frac{\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}} \, \mathrm{d}x \\ &= P - P^2 \\ &\approx P \end{aligned}$$

Variance is same order as P. (The standard deviation will be relatively large compared to P.)

Digression: Reduction of Variance

Naive Sample Variance

```
P <- sum((sample > 4)^2)/10^5
P-P^2
```

Results

```
[1] 5.99964e-05
```

Digression: Reduction of Variance

Importance sampling: $L(x) = \mathrm{e}^{4^2/2}\mathrm{e}^{-4x}$

Variance for importance sampling:

$$\begin{aligned}\operatorname{Var}_v\left[F(X)\right] &= \mathbb{E}_u\left[(F(X)L(X))^2\right] - P^2 \\ &= \int (\mathbf{1}_{[X>4]}(x)\mathrm{e}^{4^2/2}\mathrm{e}^{-4x})^2 \frac{\mathrm{e}^{-(x-4)^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x - P^2\end{aligned}$$

Digression: Reduction of Variance

Importance Sample Variance

```
secmomv <- sum((exp(4^2/2)*exp(-4*usevsample))^2
secmomv - Pv^2
```

Results

```
[1] 4.550243e-09
```

Digression: Reduction of Variance

What is best normal distribution $N(b, 1)$ for importance sampling of $P = \mathbb{P}_{N(0,1)}[X > 4]$?

After some calculation with $L(x)$:

$$\operatorname{Var}_v\left[\mathbf{1}_{[X>4]}\right] = \mathrm{e}^{b^2}(1 - \Phi(b+4)) - P^2$$

($\Phi(x)$ is the cdf of the $N(0,1)$ distribution.)

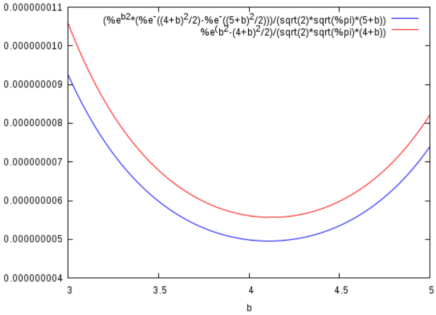
Digression: Reduction of Variance

Unfortunately, $(1 - \Phi(b+4))$ is 0 to precision-levels for $3 \le b \le 5$ leading to severe numerical error in $e^{b^2}(1 - \Phi(b+4)) - P^2$.

Use standard asymptotic estimates for the normal cdf tail:

$$\frac{1}{\sqrt{2\pi}(b+5)} \left(e^{-(b+4)^2/2} - e^{-(b+5)^2/2} \right) \le 1 - \Phi(x) \le \frac{1}{\sqrt{2\pi}(b+4)} \left(e^{-(b+4)^2/2} \right)$$

Digression: Reduction of Variance



Using the estimates with $L(x)$ provides algebraic simplification which reduces numerical error.

Using a $N(b, 1)$ with $b \approx 4.1$ gives minimum variance.

New Measures from Old

If $\mathbb{Q}$ is a probability, can define a *new* probability measure

$$d\mathbb{P} = L(x) \; d\mathbb{Q}$$

provided

- $L(x) \geq 0$ (almost surely w.r.t. $\mathbb{Q}$).
- $\int_\Omega L(x) \; d\mathbb{Q} = 1$.

Then

$$\mathbb{E}_{\mathbb{P}}\left[F(X)\right] = \mathbb{E}_{\mathbb{Q}}\left[F(X)L(X)\right]$$

Relation between measures

Ask the reverse question: Given measures $\mathbb{P}$ and $\mathbb{Q}$, is there an $L(x)$ relating them?

Necessary condition: $\mathbb{Q}[B] = 0 \implies \mathbb{P}[B] = 0$ because

$$\mathbb{P}[B] = \mathbb{E}_{\mathbb{P}}\left[\mathbf{1}_B(X)\right] = \mathbb{E}_{\mathbb{Q}}\left[\mathbf{1}_B(X) L(X)\right] = 0$$

"Impossible under $\mathbb{Q}$ is also impossible under $\mathbb{P}$."

Absolute Continuity

If $\mathbb{Q}[B] = 0 \implies \mathbb{P}[B] = 0$ we say $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{Q}$ and write $\mathbb{P} \ll \mathbb{Q}$.

"Impossible under $\mathbb{Q}$ is also impossible under $\mathbb{P}$."

Radon-Nikodym Theorem

The R-N Theorem says the necessary condition of *absolute continuity is also sufficient*:

Radon-Nikodym Theorem

- $\mathbb{P}$ and $\mathbb{Q}$ are probability measures on common σ-algebra $\mathcal{F}$; and
- $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{Q}$

then there is a function

$$L(x) = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}.$$

so that

$$\mathbb{E}_{\mathbb{P}}\left[F(X)\right] = \mathbb{E}_{\mathbb{Q}}\left[F(X)L(X)\right].$$

Completely Singular

If there is a B with $\mathbb{P}[B] = 0$ and $\mathbb{Q}[B] = 1$ we say $\mathbb{Q}$ is completely singular w.r.t. $\mathbb{P}$.

Note: $\mathbb{P}[B^C] = 1$ and $\mathbb{Q}[B^C] = 0$ so then $\mathbb{P}$ is completely singular w.r.t. $\mathbb{Q}$ and we write

$$\mathbb{P} \perp \mathbb{Q}$$

Simple Example 1

- $X \sim U([0,1])$ with probability $\mathbb{P}$
- $Y \sim N(0,1)$ with probability $\mathbb{Q}$

Then

- $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{Q}$ ($\mathbb{Q}[B] = 0 \implies \mathbb{P}[B] = 0$)
- $\mathbb{Q}$ is *not* absolutely continuous w.r.t. $\mathbb{P}$

($\mathbb{P}[1,2] = 0$ but $\mathbb{Q}[1,2] \approx 0.136$)

Simple Example 2

- $X \in \mathbb{R}^2$, $X \sim N((0,0), I_{2,2})$ with probability $\mathbb{P}$
- $Y = X/\|X\|$, $Y \sim U(S^1)$ with probability $\mathbb{Q}$

Then

- $\mathbb{P}[S^1] = 0$ and $\mathbb{Q}[S^1] = 1$
- $\mathbb{P} \perp \mathbb{Q}$.

Sophisticated Example

- $X \sim U([0,1])$ with probability $\mathbb{P}$
- Y has the Cantor distribution on $[0,1]$, with probability $\mathbb{Q}$

Then

- $\mathbb{Q}$ has a continuous c.d.f. (Devil's staircase), but the p.d.f. does not exist in any easy sense.
- $\mathbb{P} \perp \mathbb{Q}$.

This example shows that *absolute continuity* and *completely singular* are extensions of the "calculus" definitions for functions from real analysis.

Example of an R-N derivative

Previous sampling example:

- $X \sim N(0,1)$, with probability $\mathbb{P}$, p.d.f. $\mathrm{e}^{-x^2/2}$
- $Y \sim N(4,1)$, with probability $\mathbb{Q}$, p.d.f $\mathrm{e}^{-(x-4)^2/2}$

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} = \mathrm{e}^{4^2/2}\mathrm{e}^{-4x}$$

Bigger example of an R-N derivative

- $X \in \mathbb{R}^n$, $X \sim N(\mathbf{0}, C_{n,n})$, $H = C^{-1}$.
- $Y \in \mathbb{R}^n$, $Y \sim N(\mu, C_{n,n})$, $H = C^{-1}$.

$$
\begin{aligned}
u(x) &= k_n \mathrm{e}^{-\mathbf{x}^T H \mathbf{x}/2} \\
v(x) &= k_n \mathrm{e}^{-(\mathbf{x}^T - \mu) H (\mathbf{x} - \mu)/2} \\
&= k_n \mathrm{e}^{-\mu^T H \mu/2} \mathrm{e}^{\mu^T H \mathbf{x}} \mathrm{e}^{-\mathbf{x}^T H \mathbf{x}/2}
\end{aligned}
$$

Continuation, Bigger Example

Let

$$\begin{aligned} \mu^T H &= \nu^T \\ H\mu &= \nu \\ \mu &= C\nu \end{aligned}$$

so

$$\frac{d\mathbb{P}}{d\mathbb{Q}} = e^{\nu^T \mu/2} e^{-\nu^T x}$$

Deciding the Distribution

Have a single sample point: X. Is:

- $X \sim \mathbb{P}$ (null hypothesis); or
- $X \sim \mathbb{Q}$ (alternate hypothesis)

Test: Criterion is set B:

- If $x \in B$ say $\mathbb{Q}$
- If $x \in B^C$ say $\mathbb{P}$

Types of errors

- Type I error is saying $\mathbb{Q}$ when truth is $\mathbb{P}$ (reject null and accept alternative when null is true)
- Type II error is saying $\mathbb{P}$ when truth is $\mathbb{Q}$ (accept null and reject alternative when null is false)

(Usually we believe) Type I errors are worse than Type II

Confidence in the test

The confidence in the procedure is $\mathbb{P}\left[B^C\right] = 1 - \mathbb{P}\left[B\right]$, probability of accepting null, when the null is true.

The power in the procedure is $\mathbb{Q}\left[B\right]$ of rejecting null, when the null is false.

If $\mathbb{Q} \perp \mathbb{P}$, there is a test with 100% confidence and 100% power.

Conversely, If there is a test with 100% confidence and 100% power, then $\mathbb{Q} \perp \mathbb{P}$,

Neyman-Pearson Lemma

A test is efficient if there is no way to increase the confidence without decreasing the power.

Neyman-Pearson Lemma

If B is efficient, then there is L_0 such that

$$B = \left\{ x : \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} > L_0 \right\}$$