

# Math 350 - Homework 9 - Solutions

1. Suppose we wanted to estimate  $\theta$ , where

$$\theta = \int_0^1 e^{x^2} dx.$$

(a) Show that  $Y_1 = e^{U^2}(1 + e^{1-2U})/2$  is an unbiased estimator of  $\theta$ ;

(b) Show that  $Y_1$  is a better estimator than  $Y_2 = (\exp(U_1^2) + \exp(U_2^2))/2$ , where  $U_1$  and  $U_2$  are independent random numbers.

That  $Y_1$  is an unbiased estimator of  $\theta$  becomes clear once we write it in the form

$$Y_1 = (e^{U^2} + e^{(1-U)^2})/2.$$

In fact, changing variables  $u = 1 - x$  turns the integral representing  $E[\exp(1 - U)^2]$  into that of  $E[\exp U^2]$ .

Now let  $\sigma^2 = \text{Var}(\exp(U^2)) = \int_0^1 \exp(2x^2) dx - (\int_0^1 \exp(x^2) dx)^2$ . Then

$$\text{Var}(Y_1) = \frac{\sigma^2}{2}, \quad \text{Var}(Y_2) = \frac{1}{2}(\sigma^2 + c)$$

where

$$c = \text{Cov}(e^{U^2}, e^{(1-U)^2}) = \int_0^1 \exp(x^2 + (1-x)^2) dx - \left(\int_0^1 \exp x^2 dx\right)^2.$$

So

$$\frac{\text{Var}(Y_1)}{\text{Var}(Y_2)} = 1 + \frac{c}{\sigma^2}.$$

To compare these two variances, I estimate the integrals numerically. (In a simulation study we could estimate the variances with the simulated data.) I do this below using Matlab by naive Riemann sum approximation:

```
x=0:0.001:1;
I1=sum(exp(2*x.^2-2*x+1))*0.001;
I2=sum(exp(2*x.^2))*0.001;
I3=sum(exp(x.^2))*0.001;
1+(I1-I3^2)/(I2-I3^2)
ans =
```

0.2320

This shows that  $\text{Var}(Y_1) < \text{Var}(Y_2)/4$ . In other words, by using the antithetic estimator we achieve a reduction in the width of confidence intervals by a factor of at least 2 (this is the factor relating the standard deviations) compared to what it would be using twice as many independent values of  $\exp(U^2)$ .

2. Explain how antithetic variables can be used in obtaining a simulation estimate of the quantity

$$\theta = \int_0^1 \int_0^1 e^{(x+y)^2} dx dy.$$

Is it clear in this case that using antithetic variables is more efficient than generating a new pair of random variables?

We can write  $\theta = E[h(U_1, U_2)]$ , where  $h(x_1, x_2) = \exp(x + y)^2$  is monotone increasing in each variable. So by the corollary on page 209 we should expect that using antithetic variables is more efficient. To get an idea of how much more efficient, I approximate the variance and covariance integrals as in the first problem. (Again, these quantities would be estimated using simulation data in a simulation study.)

So let  $Y_1 = (h(U_1, U_2) + h(1 - U_1, 1 - U_2))/2$  and  $Y_2 = (h(U_1, U_2) + h(V_1, V_2))/2$ , where  $U_1, U_2, V_1, V_2$  are independent random numbers. We need to estimate

$$\frac{\text{Var}(Y_1)}{\text{Var}(Y_2)} = 1 + \frac{c}{v}$$

where  $c = \text{Cov}(h(U_1, U_2), h(1 - U_1, 1 - U_2))$  and  $v = \text{Var}(h(U_1, U_2))$ . Let

$$\begin{aligned} I_1 &= \int_0^1 \int_0^1 \exp((x + y)^2 + (2 - x - y)^2) dx dy \\ I_2 &= \int_0^1 \int_0^1 \exp(x + y)^2 dx dy \\ I_3 &= \int_0^1 \int_0^1 \exp(2(x + y)^2) dx dy. \end{aligned}$$

Then  $c = I_1 - I_2^2$  and  $v = I_3 - I_2^2$ . The following gives a numerical approximation of these quantities.

```
x=0:0.001:1;
s=length(x);
y=x;
x=repmat(x',1,s);
y=repmat(y,s,1);
w=x+y;
I1=sum(sum(exp(w.^2+(2-w).^2)))*(0.001)^2;
I2=sum(sum(exp(w.^2)))*(0.001)^2;
I3=sum(sum(exp(2*w.^2)))*(0.001)^2;
1+(I1-I2^2)/(I3-I2^2)
```

ans =

0.6423

Thus  $\text{Var}(Y_1) \approx 0.64\text{Var}(Y_2)$ .

3. Show that if  $X$  and  $Y$  have the same distribution then  $\text{Var}[(X + Y)/2] \leq \text{Var}(X)$ , and conclude that the use of antithetic variables can never increase variance (although it need not be as efficient as generating an independent set of random numbers).

Let  $\mu_X$  and  $\mu_Y$  denote the expected values of  $X$  and  $Y$ , respectively. Under the assumption that  $X$  and  $Y$  are equally distributed, these two expectations are the same, but let us assume that  $X$  and  $Y$  are general

for now. The following very useful fact is a form of the Cauchy-Schwartz inequality for positive-definite inner-products from linear algebra:

$$|E[(X - \mu_X)(Y - \mu_Y)]| \leq E[(X - \mu_X)^2]^{1/2} E[(Y - \mu_Y)^2]^{1/2}.$$

In particular, if  $X$  and  $Y$  are equally distributed, but not necessarily independent, then by the definition of covariance,

$$|\text{Cov}(X, Y)| \leq E[(X - \mu_X)^2]^{1/2} E[(Y - \mu_Y)^2]^{1/2} = \text{Var}(X)^{1/2} \text{Var}(Y)^{1/2} = \text{Var}(X).$$

Now, from the general fact that

$$\text{Var}\left(\frac{X+Y}{2}\right) = \frac{1}{4}[\text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)]$$

we conclude that, when  $X$  and  $Y$  are equally distributed, then

$$\text{Var}\left(\frac{X+Y}{2}\right) = \frac{1}{2}[\text{Var}(X) + \text{Cov}(X, Y)] \leq \frac{1}{2}[2\text{Var}(X)] = \text{Var}(X).$$

But this is what we wished to show.

When  $X$  and  $Y$  are antithetic, then by definition they are negatively correlated, so  $\text{Cov}(X, Y) \leq 0$ . (I am interpreting the property in the weak sense, that allows the covariance to be zero.) In this case we have a better estimate:

$$\text{Var}\left(\frac{X+Y}{2}\right) = \frac{1}{2}[\text{Var}(X) + \text{Cov}(X, Y)] \leq \frac{1}{2}\text{Var}(X).$$

The right-hand side is the variance of  $X + Y$  when  $X$  and  $Y$  are i.i.d.

4. *Suppose that  $X$  is an exponential random variable with mean 1. Give another random variable that is negatively correlated with  $X$  and that is also exponential with mean 1.*

An exponential random variable  $X$  with mean one can be expressed as  $X = -\log U$ , where  $U$  is uniform over  $(0, 1)$ . Then  $Y = -\log(1 - U)$  is also exponential with mean 1 and antithetic to  $X$ . Expressed differently,

$$Y = -\log(1 - e^{-X}).$$

5. (a) *If  $Z$  is a standard normal random variable, design a study using antithetic variables to estimate  $\theta = E[Z^3 e^Z]$ .*  
 (b) *Using the above, do the simulation to obtain an interval of length no greater than 0.1 that you can assert, with 95% confidence, contains the value of  $\theta$ .*

If  $Z$  is a standard normal random variable, then  $-Z$  is also standard normal and it is antithetic to  $Z$ . So we can use the estimator of  $\theta$  given by

$$X = \frac{Z^3 e^Z + (-Z)^3 e^{-Z}}{2} = Z^3 \frac{e^Z - e^{-Z}}{2}.$$

Thus we use the estimator  $\bar{X}_n = (X_1 + \dots + X_n)/n$ . In order that the length of the 95% confidence interval be less than 0.1 we need the estimator  $S_n/\sqrt{n}$  of the standard deviation of  $\bar{X}_n$  to be less than  $d = 0.1/2 \times 1.96 \approx 0.025$ .

In the following simulation, I will use Matlab's `randn` to obtain normally distributed random numbers. (We

could also start with uniform random numbers and simulate normally distributed ones using techniques learned in Chapter 5.)

```
Z = randn(1);
X = Z^3*(exp(Z)-exp(-Z))/2;
S2 = 0;
n = 1;
d = 0.025;
while sqrt(S2/n)>d | n<100
    X_old = X;
    Z = randn(1);
    h = Z^3*(exp(Z)-exp(-Z))/2;
    X = X+(h-X)/(n+1);
    S2 = (1-1/n)*S2+(n+1)*(X-X_old)^2;
    n = n+1;
end
X
n
```

One run of the program gave the estimated value  $X = 6.6028$ , achieved for  $n = 2826929$ , almost 3 million! (My estimated standard deviation was approximately 40 and we want  $40/\sqrt{n}$  approximately 0.025. This means that  $n$  is approximately  $(40/0.025)^2 = 2560000$ .)

Using the raw Monte Carlo simulation required  $n$  about 5.6 million.