Problem1(e)

November 5, 2015

```
In [13]: from __future__ import division
         import numpy as np

         def simulate_genome(length, alpha):
             read_arr = np.random.choice([0,1], length, p=[1-alpha, alpha])
             reads = ''
             for x in read_arr:
                 reads +=str(x)
             return reads

         def countoccurences(word, read):
             return read.count(word)

         NN = 100000
         L = 100
         alpha = 0.5
         max_iterations = 1000
         word = '101'
         nn = 100000
         c = NN*L/nn

         counts = {}
         #for n in range(1, nn):
         counts[nn] = []
         for i in range(1, max_iterations):
             genome = simulate_genome(nn, alpha)
             read_start_positions = np.random.choice(nn-L-1, NN)
             occurences = 0
             for p in read_start_positions:
                 read = genome[p:p+L]
                 occurences += countoccurences(word, read)
             counts[nn].append(occurences)
```

1 Parameters used:

$$\alpha = 0.5$$
$$\beta = 0.5$$
$$N = 10000$$
$$n = 100000$$
$$L = 100$$
$$c = 100$$

$$\lim_{n\to\infty} \frac{Var(Y_n)}{n} = c^2 \frac{\alpha^2\beta(\alpha+\beta-3\alpha^2\beta+2\alpha\beta^2+2\alpha\beta+2\beta^2)}{(\alpha+\beta)^2}$$

$$\lim_{n\to\infty} \frac{Y_n}{n} = c\frac{\alpha^2\beta}{\alpha+\beta}$$

```
In [20]: a = alpha
         b = 1-alpha
         var = c**2 * ((a**2)*b *(a+b-3*(a**2)*b+2*a*(b**2)+2*a*b+2*b**2))/(nn*(a+b)**2)

In [25]: print 'Yn/n\t Simulated:{} \t Analytical:{}'.format(np.mean(counts[nn])/nn,c*alpha**2*(1-alpha

Yn/n         Simulated:9.84286574575          Analytical:12.5

In [26]: print 'Var(Yn)/n\t Simulated={}\t Analytical:{}'.format(np.var(counts[nn])/nn**2, var)

Var(Yn)/n         Simulated=0.0061746138811          Analytical:0.0234375
```

Thus, as seen from above, our analytical and simulated estimate of $\frac{Y_n}{n}$ are 12.83 and 9.83 respectively. I used 1000 iterations, so the bound can improve by increasing the iterations. Similarly the estimate for $\frac{Var(Y_n)}{n}$ is pretty close (simulated: 0.006, analytical: 0.02)