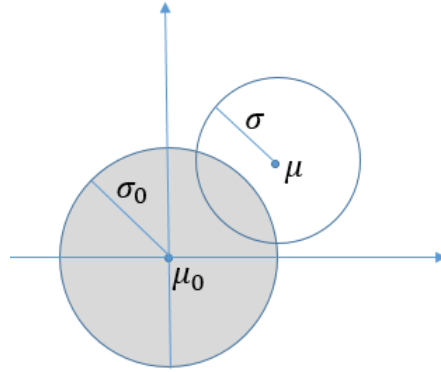


## Stat 202C Project no.1 (15 points)

Due date: **April 11** Monday, **Upload your report to CCLE.**

### Problem 1: Importance sampling and the effective number of samples



In a 2D plane, suppose the target distribution  $\pi(x, y)$  is a symmetric Gaussian with mean  $\mu = (2, 2)$  and standard deviation  $\sigma = 1$ . Suppose we use an approximate distribution  $p(x, y)$  as the trial density which is an Gaussian with mean  $\mu_0 = (0, 0)$  with standard deviation  $\sigma_0$ . So

$$\pi(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}[(x-2)^2 + (y-2)^2]} \quad p(x, y) = \frac{1}{2\pi\sigma_0} e^{-\frac{1}{2\sigma_0^2}[x^2 + y^2]}$$

We estimate the quantity  $\theta = \int (x + y) \pi(x, y) dx dy$ .

Alright, we know  $\theta = 4$ , we just want to study the effectiveness of importance sampling.

- Step 1, Compute  $\widehat{\theta}_1$  : estimate  $\theta$  by drawing  $n_1$  samples directly from  $\pi(x, y)$ .
- Step 2, Compute  $\widehat{\theta}_2$  : estimate  $\theta$  by drawing  $n_2$  samples from  $p(x, y)$  with  $\sigma_0 = 1$ .
- Step 3, Compute  $\widehat{\theta}_3$  : estimate  $\theta$  by drawing  $n_3$  samples from  $p(x, y)$  with  $\sigma_0 = 4$ .

i) Plot  $\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3$  over  $n$  (increasing  $n$  so that they converge) in one figure to compare the convergence rates. Before running the experiment, try to guess whether step 3 is more effective than step 2. [you may use a log plot at a few points  $n=10, 100, 1000, 10000, \dots$ ]

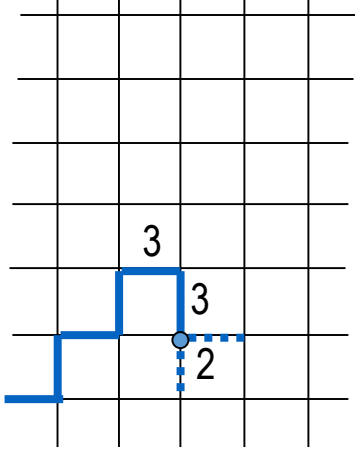
ii) Estimating the number of “effective samples”. We suggested an estimator

$$ess(n) = \frac{n}{1 + Var_p[\omega]}$$

but we are not sure how good it is. Since the samples in step 1 are all “effective” samples directly drawn from the target distribution, we use  $ess^*(n_1) = n_1$  as the truth and compare the effective sample sizes for step 2 and step 3, i.e. the true  $ess^*(n_2)$  and  $ess^*(n_3)$  are the numbers when the estimated errors reach the same level as in step 1. Plot  $ess(n_2)$  over  $ess^*(n_2)$ , and  $ess(n_3)$  over  $ess^*(n_3)$ . **Discuss your results.**

## Problem 2: Estimating the number of Self-Avoiding-Walks in an (n+1) x (n+1) grid.

Suppose we always start from position (0, 0), i.e. lower-left corner. We design a trial probability  $p(r)$  for a SAW  $r$ . Then we sample a number of  $M$  SAWs from  $p(r)$ , and the estimation is calculated below.

$K = \sum_{r \in \Omega_{n^2}} 1 = \sum_{r \in \Omega_{n^2}} \frac{1}{p(r)} p(r)$ $= E\left[\frac{1}{p(r)}\right]$ $\approx \frac{1}{M} \sum_{i=1}^M \frac{1}{p(r_i)}$ $p(r) = \prod_{j=1}^m \frac{1}{k(j)}$	
--	--

At each step, the trial probability  $p(r)$  can choose to stop (terminate the path) or walk to the left/right/up/down as long as it does not intersect itself. Each option is associated with a probability and these probabilities sum to 1 at each point.

1, What is the total number  $K$  of SAWs for  $n=10$  [try  $M=10^7$  to  $10^8$ ] ? To clarify: a square is considered a  $2 \times 2$  grid with  $n=1$ . Plot  $K$  against  $M$  (in log-log plot) and monitor whether the Sequential Importance Sampling (SIS) process has converged. Try to compare at least 3 different designs for  $p(r)$  and see which is more efficient. Make sure your results have converged.

2, What is the total number of SAWs that start from (0,0) and end at (n,n)?

Here you can still use the same sampling procedure above, but only record the SAWs which successfully reach (n, n). The truth for this number is what we discussed:  $1.5687 \times 10^{24}$ .

3, For each experiment in 1 and 2, plot the distribution of the lengths of the SAWs in a histogram (Think: Do you need to weight the SAWs in calculating the histogram?), and visualize the longest SAW that you find in print.

### Presentation:

**Your grade will be based on the quality of results and analysis of different designs.**