Mathematical Statistics **Home assignment 2: 2024–02–06**
Centre for Mathematical Sciences **MASM11/FMSN50 Monte Carlo and Empirical Methods**
Lund University **for Stochastic Inference 7.5 hp**

Detailed solutions, making sure that you describe your algorithms in text, and not only as code, must be submitted **before Tuesday 20 Feb, 13:00:00**. You are strongly encouraged to work in groups of two.

Report submissions are accepted in PDF format only.

Also submit your MATLAB-files (or implementation in other language), with a file named `proj2.m` that can be used to run your analysis (Remember to submit **all** of the files you us e to the create your solution). Submit the projects in CANVAS.

Discussion between groups is permitted, as long as your report reflects your own work.

Part1: Self-avoiding walks in $\mathbb{Z}^d$

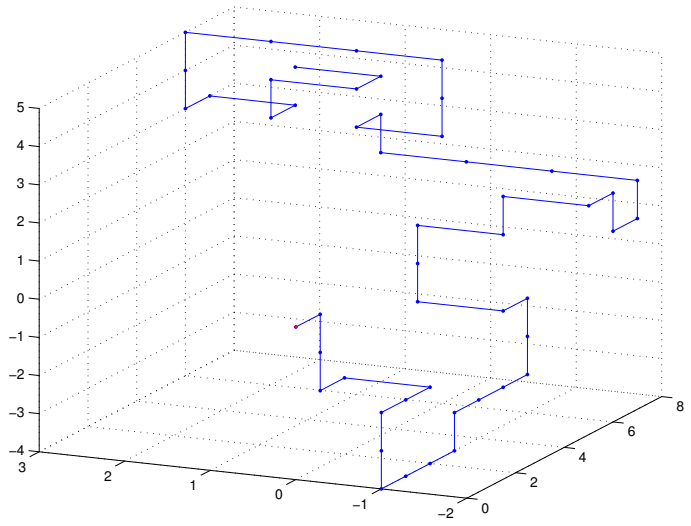

Figure 1: A self-avoiding walk of length 50 in $\mathbb{Z}^3$ starting at $(0,0,0)$.

A *self-avoiding walk* (SAW) is a sequence of moves on a lattice that does not visit the same point more than once. In $\mathbb{Z}^d$, the set $\mathsf{S}_n \subseteq \mathbb{Z}^d$ of possible such walks of length n is formally given by

$$\mathsf{S}_n(d) = \{x_{0:n} \in \mathbb{Z}^{d(n+1)} : x_0 = \mathbf{0}, |x_k - x_{k-1}| = 1, x_\ell \neq x_k, \forall 0 \leq \ell < k \leq n\}.$$

To compute the number $c_n(d) = |\mathsf{S}_n(d)|$ of possible such walks is, when n is large, considered to be a very challenging problem in enumerative combinatorics. The aim of this home assignment is to solve this problem using sequential Monte Carlo (SMC) methods.

First two theoretical problems.

1. Convince yourself that for all $n \geq 1$ and $m \geq 1$,

$$c_{n+m}(d) \leq c_n(d)c_m(d). \tag{1}$$

2. A sequence $(a_n)_{n\geq 1}$ is called *subadditive* if $a_{m+n} \leq a_m + a_n$. *Fekete's*[1] *lemma* states that for every subadditive sequence $(a_n)_{n\geq 1}$, the limit $\lim_{n\to\infty} a_n/n$ exists and is equal to $\inf_{n\geq 1} a_n/n$ (which may be equal to $-\infty$). Use Fekete's lemma to prove that the limit

$$\mu_d = \lim_{n\to\infty} c_n(d)^{1/n} \tag{2}$$

exists.

[1] Mihály Fekete (1886–1957) was an Israeli-Hungarian mathematician.

The constant μ_d in (2) is called the *connective constant* and depends on the particular lattice chosen. One may of course consider SAW:s on other lattices than $\mathbb{Z}^d$, e.g. the honeycomb lattice in 2D. For this case it is proven that $\mu = \sqrt{2+\sqrt{2}}$. The number μ_d could be interpreted as the geometric mean of the number of un-visited neighbours (sequentially looking one step ahead) along a self-avoiding path. In the light of (2) it is conjectured for all d (actually proven for $d \geq 5$) that

$$c_n(d) \sim \begin{cases} A_d \mu_d^n n^{\gamma_d - 1} & d = 1, 2, 3, \ d \geq 5 \\ A_d \mu_d^n \log(n)^{1/4} & d = 4 \end{cases} \quad \text{as} \quad n \to \infty, \tag{3}$$

where in the power law correction $n^{\gamma_d - 1}$, γ_d does not depend on the lattice under consideration only the dimension d. It is known that $\gamma_1 = 1, \ \gamma_2 = 43/32$ and that $\gamma_d = 1$ for $d \geq 5$. More on self-avoiding walks can be found in Slade (2011) (See also Duminil-Copin and Smirnov, 2012).

We now aim at estimating $c_n(d)$ for $n = 1, 2, 3, \ldots$ using SMC methods. Moreover, we are particularly interested in estimating the connective constant μ_d via the relation (3) by investigating how $c_n(d)$ depends on n for large n's. As mentioned in the lectures, estimates of the $c_n(d)$'s can be obtained by considering the sequence $(f_n)_{n \geq 1}$ of uniform distributions on the sets $(\mathsf{S}_n(d))_{n \geq 1}$, i.e., letting

$$f_n(x_{0:n}) = \frac{\mathbb{1}_{\mathsf{S}_n(d)}(x_{0:n})}{c_n(d)}, \quad x_{0:n} \in \mathbb{Z}^{d(n+1)},$$

where $\mathbb{1}$ denotes the indicator function[2], and estimating sequentially the $c_n(d)$'s using SMC.

Note: For problem 3–6 use $d = 2$.

3. A first (naive) approach is to estimate the $c_n(2)$'s using the *sequential importance sampling* (SIS) algorithm with instrumental distribution g_n being that of a standard *random walk* $(X_k)_{k=0}^n$ in $\mathbb{Z}^2$, where $X_0 = \mathbf{0}$ and each X_{k+1} is drawn uniformly among the four neighbours of X_k. In fact, this method simply amounts (why?) to simulating a large number N of random walks in $\mathbb{Z}^2$, counting the number N_{SA} of self-avoiding ones, and estimating $c_n(2)$ using the observed ratio N_{SA}/N. Implement this approach and use it for estimating $c_n(2)$ for $n = 1, 2, 3, \ldots$ Conclusion?

4. In order to improve the naive approach, let g_n be the distribution of a *self-avoiding random walk* $(X_k)_{k=0}^n$ in $\mathbb{Z}^2$ starting in the origin. This means that

   (i) $X_0 = \mathbf{0}$ and

   (ii) given $X_{0:k}$, the next point X_{k+1} is drawn uniformly among the free neighbours $\mathsf{N}(X_{0:k})$ of X_k, where

   $$\mathsf{N}(x_{0:k}) = \{x \in \mathbb{Z}^2 : |x_k - x| = 1, x \neq x_\ell, \forall 0 \leq \ell < k\};$$

   if $\mathsf{N}(X_{0:k}) = \varnothing$, then X_{k+1} is set to X_k.

   Implement the SIS algorithm based on the instrumental distribution g_n and use it for estimating $c_n(2)$ for $n = 1, 2, 3, \ldots$ Conclusion?

5. Implement the *sequential importance sampling with resampling* (SISR) algorithm based on the instrumental distribution g_n in Problem 4 and use it for estimating $c_n(2)$ for $n = 1, 2, 3, \ldots$ Conclusion?

6. Use your (SISR) estimates of the $c_n(2)$'s to obtain an estimate of A_2, μ_2 and γ_2 via the relation (3). Hint: Look at $\ln(c_n)$ and identify a linear regression in the transformed parameters. Which of the original parameters are most easy to estimate? Redo the estimation in several independent replicates to check how the estimates varies. Explain why!

[2]The indicator function is defined by: $\mathbb{1}_{\mathsf{A}}(x) = 1$ if $x \in \mathsf{A}$ and $\mathbb{1}_{\mathsf{A}}(x) = 0$ if $x \notin \mathsf{A}$.

Now we consider the problem for a general d:

7. Verify that the following general bound should hold

   $$d \leq \mu_d \leq 2d - 1.$$

8. Verify that the following general bound should hold

   $$A_d \geq 1 \text{ for } d \geq 5.$$

   Hint: Plugin (3) into (1) for appropriate choices of n and m.

9. Use the (SISR) approach estimate of A_d, μ_d and γ_d via the relation (3) for some $d \geq 3$ with the same technique as in problem 6. First compare with the bounds from problems 7-8. Finally compare with the asymptotic bound on μ_d for large d found in Graham (2014):

   $$\mu_d \sim 2d - 1 - 1/(2d) - 3/(2d)^2 - 16/(2d)^3 + \mathrm{O}(1/d^4)$$

   Conclusion?

Part2: Filter estimation of noisy population measurements

10. Let X_k be the relative population size in generation k of some organism. Relative population sizes are between zero and one where zero means extinction and one is the maximum size relative to the carrying capacity of the environment. However the closer we are to size one the more resources are consumed and then the next generation will be smaller. In a simplified relative population model we have the following dynamics.

    $$X_{k+1} = R_{k+1}X_k(1 - X_k), R_{k+1} \in U(A, B), \text{iid}, k = 0, 1, 2, \cdots$$

    where R is the stochastic reproduction rate due to fluctuating environmental conditions. Assume that $X_0 \in U(C, D)$. Due to problems of measuring the exact relative population we get a measurement

    $$Y_k | X_k = x \in U(Gx, Hx).$$

    This can be seen as a hidden Markov model where the relative population size X is the hidden Markov chain. We now want to estimate the filter expectation $\tau_k = E[X_k | Y_{0:k}]$ for $k = 0, 1, 2, \cdots, n$. You can do this by modifying the code on slide 24 of Lecture 7 (i.e. modifying the transition density and the observation density). The file `population_2024.mat` contains a simulation ($n = 100$) of the model with the measurements Y as well as the true values of X for comparison. Assume that the parameters are given as $A = 0.8,\ B = 3.8,\ C = 0.6,\ D = 0.99,\ G = 0.8$ and $H = 1.25$.

    a) Estimate the filter expectation $\tau_k = E[X_k | Y_{0:k}]$ for $k = 0, 1, 2, \cdots, 100$. Try with N=500, 1000 and 10000, where N is the number of particles. Which choice seems most reasonable?

    b) Use the technique on slide 25 of Lecture 7 to make a point wise confidence interval for X_k for $k = 0, 1, 2, \cdots, 100$. Check if the true X_k is between the upper and lower limit for $k = 0, 1, 2, \cdots, 100$. Try with N=500, 1000 and 10000. Which choice seems most reasonable?

References

Duminil-Copin, H., & Smirnov, S. (2012). The connective constant of the honeycomb lattice equals $\sqrt{2 + \sqrt{2}}$. *Annals of Mathematics*, **175**, 1653–1665. Available at:`http://annals.math.princeton.edu/wp-content/uploads/annals-v175-n3-p14-p.pdf`

Graham, BT. (2018) Borel-type bounds for the self-avoiding walk connective constant, Available at:`http://arxiv.org/pdf/0911.5163.pdf`

Slade, Gordon. (2011) "The self-avoiding walk: A brief survey." Surveys in Stochastic Processes,**181–199**. Available at:`https://www.math.ubc.ca/~slade/spa_proceedings.pdf`

Good luck!