Home Assignment 2

MASM11 - Monte Carlo and empirical methods for stochastic inference

2023-02-21

Mårten Augustsson, Arvid Gramer

Part 1: Self-avoiding random walks

1. Product of random walk lengths

Convince yourself that for all $n \geq 1$ and $m \geq 1$, $c_{n+m}(d) \leq c_n(d)c_m(d)$.

This inequality makes sense due to the nature of self-avoiding random walks. If the walk were not self-avoiding, the number of possible next steps at each position would be $2d$, and for each of those steps there would be $2d$ possible choices, and so on. Thus, the number of possible paths in n steps is $(2d)^n$, meaning that

$$c_{n+m}(d) = (2d)^{n+m} = (2d)^n(2d)^m = c_n(d)c_m(d) \tag{1}$$

and we would have exact equality.

Now, let the random walks be self-avoiding. Then the walk of length $n+m$ must, after reaching step n, in each subsequent step from $n+1$ to $n+m$ avoid all previous positions from times 1 to n. By comparison, multiplying $c_n(d)$ and $c_m(d)$ means doing a self-avoiding walk of length n and, once step n is reached, spawning a new self-avoiding walk of length m from the end point of the first walk. This new walk needs to avoid returning to its own previous positions, but is allowed to visit the n points of the initial walk. This means that although the walk of length $n+m$ and the combination of two walks (lengths n and m) take the same number of steps, the previous one has many more occupied positions to account for, and is thus much more restricted than the other.

As a very simple example, consider $d = 2$ and $n = m = 2$: From the starting position, there are four possible steps, and from the second position there are three possible steps - all except back to the starting position. Then, $c_n(d) = c_m(d) = c_2(2) = 3 \cdot 4 = 12$. From the third position, there are still three possible steps. From the fourth, however, we encounter the possibility of returning to a previously visited point (apart from the immediately preceding one), if the four steps form a square back to the initial starting point. This may happen in eight different ways, making the number of possible steps $c_{n+m}(d) = c_4(2) = 4 \cdot 3 \cdot 3 \cdot 3 - 8 = 108 - 8 = 100$. In contrast, $c_n(d)c_m(d) = c_2(2)c_2(2) = (3 \cdot 4)(3 \cdot 4) = 144$ and we see that the proposed inequality holds for this example.

2. Connective constant

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

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

exists.

To prove this, we use our result from problem 1), namely that

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

where $c_n(d)$ is the number of possible walks of length n in d dimensions. Taking the logarithm of both sides of the inequality, we get

$$\log(c_{n+m}(d)) \leq \log(c_n(d)c_m(d)) \tag{2}$$

$$\log(c_{n+m}(d)) \leq \log(c_n(d)) + \log(c_m(d)) \tag{3}$$

since the logarithm is a strictly increasing function. Thus, $\log c_n(d)$ is subadditive. According to Fekete's lemma, this means that the limit $\lim_{n\to\infty} \frac{\log(c_n(d))}{n}$ exists and is equal to $\inf_{n\geq 1} \frac{\log(c_n(d))}{n}$. Now, we let this limit be equal to $\log(\mu)$ and take the exponential function of both sides. We have

$$\log(\mu) = \lim_{n\to\infty} \frac{\log(c_n(d))}{n} \tag{4}$$

$$e^{\log\mu} = \lim_{n\to\infty} e^{\log(c_n(d))/n} \tag{5}$$

$$e^{\log\mu} = \lim_{n\to\infty} (e^{\log(c_n(d))})^{1/n} \tag{6}$$

$$\mu = \lim_{n\to\infty} c_n(d)^{1/n} \tag{7}$$

since $\log(\mu)$ exists according to Fekete's lemma, $\mu = e^{\log(\mu)}$ must also exist - even if $\log\mu = -\infty$ this only means that $\mu = 0$ - and the proof is finished.

3. Naive approach to SAW

A first (naive) approach is to estimate $c_n(2)$ 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 $X_0 = 0$ 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?

The reason that this approach should work is that, reasonably, the proportion of the simulated walks that are self-avoiding should on average be equal to the probability that a random walk is self-avoiding. This, in turn, should be the proportion of total possible walks that are self-avoiding. The number of self-avoiding walks N_{SA} should follow a binomial distribution with parameters N and $\pi = \mathbb{P}(\text{walk is self-avoiding})$, making its expectation $\mathbb{E}[N_{SA}] = N\pi$ according to the properties of the binomial distribution. We also know that the number of possible non-restricted random walks should be the product of the possible directions at each of the n steps, which is $2d$ (see problem 1). The probability that a random walk is self-avoiding should then be $c_n(d)/(2d)^n$. Using this, we get

$$\mathbb{E}\left[\frac{N_{SA}}{N}\right] = \frac{1}{N}\mathbb{E}[N_{SA}] = \frac{1}{N}N \cdot \mathbb{P}(\text{walk is self-avoiding}) = \frac{c_n(d)}{(2d)^n} \tag{8}$$

$$\mathbb{E}\left[\frac{N_{SA}\cdot(2d)^n}{N}\right] = (2d)^n\mathbb{E}\left[\frac{N_{SA}}{N}\right] = (2d)^n\frac{c_n(d)}{(2d)^n} = c_n(d) \tag{9}$$

and will therefore use $\frac{N_{SA}\cdot(2d)^n}{N}$ as an unbiased estimator of $c_n(d)$. The variance of the estimate can be

Number of steps n	1	2	3	4	5	10	15	20
Proportion of SAW:s $N_{SA}/1000$	1.0	0.7486	0.568	0.3898	0.283	0.0418	0.006	0.0015
Estimate of $c_n(2)$	4	11.98	36.35	99.79	289.8	43 830	$6.442 \cdot 10^6$	$1.649 \cdot 10^9$
Lower CI	4	11.84	35.73	97.34	280.8	39 717	$4.817 \cdot 10^6$	$8.152 \cdot 10^8$
Upper CI	4	12.11	36.97	102.2	298.8	47 944	$8.068 \cdot 10^6$	$2.483 \cdot 10^9$

Table 1: Estimates of $c_n(2)$ with the naive approach using general random walks. Confidence intervals are 95%.

calculated by first calculating

$$\mathbb{V}[N_{SA}] = N \cdot \frac{c_n(d)}{(2d)^n} \cdot \left(1 - \frac{c_n(d)}{(2d)^n}\right) \tag{10}$$

since N_{SA} follows a binomial distribution. Then,

$$\begin{aligned}
\mathbb{V}\left[\frac{N_{SA} \cdot (2d)^n}{N}\right] &= \left(\frac{(2d)^n}{N}\right)^2 \mathbb{V}[N_{SA}] && (11)\\
&= \frac{(2d)^{2n}}{N^2} N \cdot \frac{c_n(d)}{(2d)^n} \cdot \left(1 - \frac{c_n(d)}{(2d)^n}\right) && (12)\\
&= \frac{(2d)^n}{N} \cdot c_n(d) \cdot \left(1 - \frac{c_n(d)}{(2d)^n}\right) && (13)\\
&= \frac{1}{N} c_n(d) \cdot ((2d)^n - c_n(d)) && (14)
\end{aligned}$$

Finally, replacing $c_n(d)$ with its estimate, we get an estimated variance that is

$$\hat{\mathbb{V}}\left[\frac{N_{SA} \cdot (2d)^n}{N}\right] = \frac{1}{N} \frac{N_{SA} \cdot (2d)^n}{N} \left((2d)^n - \frac{N_{SA} \cdot (2d)^n}{N}\right) \tag{15}$$

and can be used to construct confidence intervals using the central limit theorem. This theorem is applicable because we simulate a large number of independent random walks with the same distribution, and use its sums.

Implementing this for $N = 10000$, $d = 2$ and various step numbers n, we get the estimates and 95% confidence intervals presented in table 1.

As we can see, the method gives clear estimates for very low numbers of steps n, and the estimates are also quite reasonable since theoretically $c_n(2)$ should be 4 for $n = 1$, 12 for $n = 2$, 36 for $n = 3$ and 100 for $n = 4$. The estimates quickly gain variance as n increases, however, and for large n it is also much harder to explicitly calculate $c_n(2)$. We had to use sample size $N = 10000$ to be able to estimate for $n = 20$ at all. For that many steps, random walks that are self avoiding are very rare, so when using for example $N = 1000$ the amount of self avoiding random walks of length 20 was often zero. Clearly, this is a very inefficient method for large n since so many samples are discarded due to not being self-avoiding. The method is much improved in the next problem.

Number of steps n	1	2	3	4	5	10	20	50	100
Estimate of $c_n(2)$	4	12	36	100.40	287.50	45 182	$1.0013 \cdot 10^9$	$5.302 \cdot 10^{21}$	$8.194 \cdot 10^{42}$

Table 2: Estimates of $c_n(2)$ with sequential importance sampling, using N = 10000

4. SIS approach to SAW

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

A refinement of the previous method is to try to make every random walk self avoiding. This is done by, in each step, only permitting moves to previously unvisited positions. If there are several free positions, we pick one of them with (discrete) uniform distribution. This ensures that none of the sampled walks crosses its own path. We may, however, get stuck in a position where there are no free neighbours, in which case that walk will end before taking all its n steps.

If we then, for a large number of particles, save the number of free neighbours we can extract an estimate of $c_n(d)$ as the following. For each particle take the product of how many free neighbours that particle has had throughout its trajectory up until step n. This product is an estimate $c_n(d)_i$ of the number of SAW:s of length n, since by simple combinatorics the number of possible paths should be the product of the number of possible directions in each step. This is of course complicated by the fact that the number of possible directions in a given step is highly dependent on which directions have been taken in previous steps. Therefore, there will be a deviation between the estimate and the true value $c_n(d)$. Taking the average $\sum_{i=1}^{N} c_n(d)_i/N$ of N samples, however, should make the estimate closer to $c_n(d)$.

Estimates of $c_n(2)$ for various walk lengths n are presented in table 2, and here seem perfectly accurate for $n = 1, 2, 3$. For comparability we keep N = 10000 samples. For $n \geq 4$ there still seems to be some deviation from the true value (judging by the deviation of 0.40 for $n = 4$. The expected value, variance and confidence interval of these estimates are not easily calculable as they were in problem 3. We note that here, we more easily than before get estimates of higher n since each walk is already self-avoiding. However, there is a problem with degenerating weights since many walks (approximately 78%) sooner or later end up in a position they cannot continue from since they are surrounded by previously visited positions. This makes their weights zero after that step since these weights are the repeated product of the number of possible directions at each step and this number is zero when the walk is stuck. The weight is therefore concentrated into the walks which are never stuck, which become fewer and fewer as n increases. Thus, there are very few samples helping in the estimate of $c_n(2)$ for high n, making these estimates less accurate. This weakness will be amended in the next step.

Number of steps n	1	2	3	4	5	10	20	50	100
Estimate of $c_n(2)$	4	12	36	99.07	281.96	45 542	$9.888 \cdot 10^8$	$5.520 \cdot 10^{21}$	$9.397 \cdot 10^{42}$

Table 3: Estimates of $c_n(2)$ with sequential importance sampling with resampling

5. SISR approach to SAW

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 $n = 1, 2, 3, ...$ Conclusion?

A further refined estimate is to utilise resampling. The idea is that if a particle is at a position with many unvisited neighbours, there are many alternative future paths stemming from that position. What we then do is that, after each step, we sample N "new" particles from the N present ones (with replacement) weighted on the amount of free neighbours for each particle. This means that a particle which is close to many visited positions has less probability of being resampled than one out in the free, since there are fewer possible paths originating from its position. This also means that particles which get stuck will have weight zero in the resampling, and the other particles will be the ones to continue their trajectory in many copies and thus continue helping in the estimation of $c_n(2)$ for higher n.

Adding this resampling into the algorithm, and again using N = 10000 particles, we get the estimates shown in table 3. The estimates are similar to those in table 2, except for a rather large difference for $n = 100$. The estimates for $n = 1, 2, 3$ are perfectly accurate, while there is still a deviation for $n = 4$ (actually slightly larger than with the SIS method) and reasonably for the larger n as well.

6. Estimation of parameters

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!

Since we have the asymptotic relation that

$$c_n(2) \sim A_2 \mu_2^n n^{\gamma_2 - 1} \tag{16}$$

$$\iff \lim_{n\to\infty} \frac{A_2 \mu_2^n n^{\gamma_2 - 1}}{c_n(2)} = 1 \tag{17}$$

$$\implies \lim_{n\to\infty} \log\left(A_2 \mu_2^n n^{\gamma_2 - 1}\right) - \log c_n(2) = 0 \tag{18}$$

$$\implies \lim_{n\to\infty} \log c_n(2) = \lim_{n\to\infty} \log A_2 + \log \mu_2 \cdot n + (\gamma_2 - 1)\log n \tag{19}$$

we should for reasonably large n be able to use the estimate of $c_n(2)$ to estimate A_2, μ_2 and γ_2 through a linear regression using the two input variables n and $\log n$. Their respective coefficients from the fitting of the linear model should then be estimates of $\log \mu_2$ and $\gamma_2 - 1$, while the constant term of the linear model should be an estimate of A_2. Doing this using the SISR estimate of $c_n(2)$ presented in problem 5, with

Parameter	A_d	μ_d	γ_d
Lowest estimate	1.2976	2.6327	1.1693
Highest estimate	1.7161	2.6475	1.3761
Range	0.4185	0.0148	0.2068
Mean estimate	1.4235	2.6389	1.2873
Standard deviation	0.1178	$7.115 \cdot 10^{-3}$	0.054
Coefficient of variation	8.27%	0.27%	4.21%

Table 4: Statistics on the ten estimates of the parameters, each using $N = 1000$ and $n = 100$ in dimension $d = 2$.

$N = 1000$ and $n = 100$, our first estimates of the parameters are then given by

$$\widehat{\log A_2} = 0.3906 \tag{20}$$

$$\implies \hat{A}_2 = e^{0.3906} = 1.4779 \tag{21}$$

$$\widehat{\log \mu_2} = 0.9736 \tag{22}$$

$$\implies \hat{\mu}_2 = e^{0.9736} = 2.6475 \tag{23}$$

$$\hat{\gamma}_2 - 1 = 0.2583 \tag{24}$$

$$\implies \hat{\gamma}_2 = 1.2583 \tag{25}$$

These estimates all align well with the theory. For A_d, we only show a lower bound of 1 (problem 8) for dimensions $d \geq 5$. However, the bound seems to hold in this case too even if d is only 2, since our estimate is 1.4779. For μ_d, we have theoretical bounds for all d, namely $d \leq \mu_d \leq 2d - 1$, which we prove in problem 7. For $d = 2$, we then have $2 \leq \mu_d \leq 3$, which agrees with our estimate 2.6475. Finally, γ_2 should theoretically be exactly $43/32 = 1.34375$, and our estimate is 1.2583.

However, this estimate is not stable for all parameters. Repeating the SISR process ten times and each time re-estimating the parameters using linear regression, we find that the estimate of the connective constant μ_d is the most stable, varying between 2.6327 and 2.6475 with a mean of 2.6389 and a standard deviation of $7.115 \cdot 10^{-3}$, about 0.27% of the mean. The other two parameters have greater variation: For A_d, the estimates vary between 1.2976 and 1.7161, with mean 1.4235 and standard deviation 0.1178 (8.27% of the mean). Finally, for γ_d, the estimates vary between 1.1693 and 1.3761, with mean 1.2873 and standard deviation 0.054 (4.21% of the mean). These result probably stems from the difference in impact of the different parameters. If a parameter contributes greatly to the final result, variations in it will change the result greatly and thus be easier to measure. Since μ_2^n changes a lot if μ_2 changes, it will be more clear in the regression, which makes the numerical result more stable, than for example the linear A_2. All these data about the parameter estimates are summarized in table 4

7. Bounds of the connective constant

Verify that the following general bound should hold: $d \leq \mu_d \leq 2d - 1$.

Theoretically, this may be proven by noting that regardless of dimension d, it is possible to construct a random walk which only takes steps in the increasing directions of each dimension (i.e., upwards and to the right in the two-dimensional case). It is evident that this random walk is self-avoiding and thus a subset of $\mathbf{S}_n(d)$, since it may never return "backwards" to a previous position. Let $\mathbf{S}_n^*(d)$ be the set of all such increasing random walks. It is also evident that the length of this set is d^n, since at each point there are always d possible paths to choose from. We may then write that

$$\begin{aligned} \mathbf{S}_n^*(d) &\subseteq \mathbf{S}_n(d) && (26)\\ \iff c_n^*(d) &\leq c_n(d) && (27)\\ \iff d^n &\leq c_n(d) && (28)\\ \iff (d^n)^{1/n} &\leq (c_n(d))^{1/n} && (29)\\ \iff d &\leq (c_n(d))^{1/n} && (30) \end{aligned}$$

and on this we may apply our result from problem 2, taking the limit as n goes to infinity:

$$d \leq \lim_{n\to\infty} (c_n(d))^{1/n} = \mu_d \tag{31}$$

and the lower bound is proven. To prove the upper bound, we instead construct a random walk which may return to previously visited locations but not to the one visited in the previous step. That is, the restriction is simply $x_k \neq x_{k-2}$ rather than $x_k \neq x_l$, $\forall 0 \leq l < k \leq n$. This means that at any given step besides the first, there are $2d-1$ possible directions to take, while at the first there are $2d$ directions. This gives us $2d(2d-1)^{n-1}$ possible paths in n steps. We call the set of possible paths for this random walk $\tilde{\mathbf{S}}_n(d)$. Since a self-avoiding random walk follows all the restrictions of this new random walk (but also follows more restrictions), we have that

$$\begin{aligned} \mathbf{S}_n(d) &\subseteq \tilde{\mathbf{S}}_n(d) && (32)\\ \iff c_n(d) &\leq \tilde{c}_n(d) && (33)\\ \iff c_n(d) &\leq 2d(2d-1)^{n-1} && (34)\\ \iff (c_n(d))^{1/n} &\leq (2d(2d-1)^{n-1})^{1/n} && (35)\\ \iff (c_n(d))^{1/n} &\leq (2d-1)^{(n-1)/n}(2d)^{1/n} && (36) \end{aligned}$$

and finally, taking the limit just as when proving the lower bound, we have

$$\begin{aligned} \lim_{n\to\infty} (c_n(d))^{1/n} &\leq \lim_{n\to\infty} (2d-1)^{(n-1)/n}(2d)^{1/n} && (37)\\ \implies \mu_d &\leq (2d-1)^1(2d)^0 = 2d-1 && (38) \end{aligned}$$

and both bounds have then been proved.

8. Bound for A_d

Verify that the following general bound should hold: $A_d \geq 1$ for $d \geq 5$.

We know that for dimensions $d \geq 5$, the number of possible walks $c_n(d)$ approaches proportionality with $A_d \mu_d^n n^{\gamma_d - 1}$ as n approaches infinity. That is, the ratio of the two expressions approaches one. Furthermore, the factor $n^{\gamma_d - 1}$ is only relevant for $d = 1, 2, 3$ since for $d \geq 5$, $\gamma_d = 1 \implies n^{\gamma_d - 1} = n^0 = 1$. This means that we may write the limit of the ratio as

$$\lim_{n\to\infty} \frac{c_n(d)}{A_d \mu_d^n} = 1 \tag{39}$$

Next, we use our result from problem 1, namely that

$$c_{m+n}(d) \leq c_m(d) c_n(d) \tag{40}$$

$$\implies \frac{c_{m+n}(d)}{A_d \mu_d^{m+n}} \leq \frac{c_m(d) c_n(d)}{A_d \mu_d^{m+n}} \tag{41}$$

where the second equality follows from the fact that A_d and μ_d must both be positive. The number of possible walks $c_n(d)$ must be positive, and so must the connective constant μ_d, since it is a measure of the average number of possible directions for the next step. If A_d were not positive, the limit of $\frac{c_n(d)}{A_d \mu_d^n}$ would not be either, which would be a contradiction. Thus, A_d must also be positive. We then take the limits of both sides of the inequality as m and n both go to infinity:

$$\lim_{n,m\to\infty} \frac{c_{m+n}(d)}{A_d \mu_d^{m+n}} \leq \lim_{n,m\to\infty} \frac{c_m(d) c_n(d)}{A_d \mu_d^{m+n}} \tag{42}$$

$$\lim_{n,m\to\infty} \frac{c_{m+n}(d)}{A_d \mu_d^{m+n}} \leq \lim_{n,m\to\infty} A_d \frac{c_m(d) c_n(d)}{A_d^2 \mu_d^{m+n}} \tag{43}$$

$$\lim_{n,m\to\infty} \frac{c_{m+n}(d)}{A_d \mu_d^{m+n}} \leq \lim_{n,m\to\infty} A_d \frac{c_m(d)}{A_d \mu_d^m} \frac{c_n(d)}{A_d \mu_d^n} \tag{44}$$

$$\lim_{n,m\to\infty} \frac{c_{m+n}(d)}{A_d \mu_d^{m+n}} \leq A_d \lim_{m\to\infty} \frac{c_m(d)}{A_d \mu_d^m} \lim_{n\to\infty} \frac{c_n(d)}{A_d \mu_d^n} \tag{45}$$

$$1 \leq A_d \cdot 1 \cdot 1 \tag{46}$$

$$1 \leq A_d \tag{47}$$

where the final steps utilize the limit of the ratio once again. The proof is thus concluded.

9. Parameter estimation for higher dimensions

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)

Implementing the same code as we used in problems 5 and 6, but changing the dimension d, we get new estimates of our three parameters. We chose to try for $d = 3$ because it is the easiest and closest to 2, but also for $d = 5$ in order to be able to properly use the theoretical results from problems 7 and 8. As in problem 6, we did the SISR and estimation through regression 10 times for each d and calculated simple descriptive statistics. The statistics for $d = 3$ are found in table 5 and the statistics for $d = 5$ in table 6.

Clearly, the estimates become much more stable as the dimension increases, with the coefficients of variation (standard deviation divided by mean) becoming less than one percent for all three parameters at $d = 5$. The connective constant estimate is still clearly more stable than the other two parameters, however.

Parameter	A_d	μ_d	γ_d
Lowest estimate	1.2080	4.6808	1.1061
Highest estimate	1.3077	4.6939	1.1735
Range	0.0997	0.0131	0.0674
Mean estimate	1.2585	4.6866	1.1389
Standard deviation	0.0311	$5.311 \cdot 10^{-3}$	0.0203
Coefficient of variation	2.47%	0.11%	1.79%

Table 5: Statistics on the ten estimates of the parameters, each using $N = 1000$ and $n = 100$ in dimension $d = 3$

Parameter	A_d	μ_d	γ_d
Lowest estimate	1.1171	8.8291	1.0229
Highest estimate	1.1510	8.8386	1.0427
Range	0.0339	0.0095	0.0198
Mean estimate	1.1370	8.8353	1.0291
Standard deviation	$9.299 \cdot 10^{-3}$	$3.167 \cdot 10^{-3}$	$6.175 \cdot 10^{-3}$
Coefficient of variation	0.82%	0.04%	0.60%

Table 6: Statistics on the ten estimates of the parameters, each using $N = 1000$ and $n = 100$ in dimension $d = 5$

As for the theoretical bounds, we first note that all estimates of A_d, regardless of dimension, are larger than one, in keeping with the result from problem 8. γ_d should theoretically be 1 for $d \geq 5$, and the estimates are very close to this. However, they are all slightly above one and never below it, indicating a degree of bias in the estimate. Thirdly, the connective constant also falls in the theoretical intervals stipulated in problem 7, being between 3 and 5 for $d = 3$ and between 5 and 9 for $d = 5$. In fact, the estimates are clearly much closer to the upper bound than to the lower one. Whether this is in line with theory or due to bias in the estimation is beyond the scope of this assignment.

Finally, we compare the estimates of μ_d with the theoretical asymptotic bound

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

which become

$$\mu_2 \sim 4 - 1 - \frac{1}{4} - \frac{3}{16} - \frac{16}{64} + \mathcal{O}(\frac{1}{16}) \tag{48}$$

$$\approx 2.3125 + \mathcal{O}(\frac{1}{16}) \tag{49}$$

$$\mu_3 \sim 6 - 1 - \frac{1}{6} - \frac{3}{36} - \frac{16}{216} + \mathcal{O}(\frac{1}{81}) \tag{50}$$

$$\approx 4.6759 + \mathcal{O}(\frac{1}{81}) \tag{51}$$

$$\mu_5 \sim 10 - 1 - \frac{1}{10} - \frac{3}{400} - \frac{16}{8000} + \mathcal{O}(\frac{1}{625}) \tag{52}$$

$$\approx 8.8905 + \mathcal{O}(\frac{1}{625}) \tag{53}$$

and we see that these differ slightly from the average estimates found in the tables. In the case of $d = 2$ the difference is largest at approximately 0.3264 (+14% of the theoretical value), whereas it becomes smaller for $d = 3$ (difference +0.0107 or +0.23% of the theoretical value) and $d = 5$ (difference -0.0552 or -0.62% of the theoretical value). This makes a degree of sense since the bound is asymptotic for large d and a large difference for $d = 2$ is reasonable. It is, however, slightly strange that the difference is larger (even relatively speaking) for $d = 5$ than for $d = 3$, but this could possibly be due to chance.

Part 2: Filter estimation of noisy population measurements

10. a) Point estimates

Estimate the filter expectation $\tau_k = \mathbb{E}[X_k|Y_{0:k}]$ for $k = 0, 1, 2, ..., 50$.

To do this, we once again use a sequential importance sampling with resampling. Using the sample size $N = 1000$, we create a vector of 1000 random variables $x_{0i}, i = 1, 2, ..., 1000$ from a uniform $\mathcal{U}(0.6, 0.99)$ distribution, all at generation 0. Next, to each variable we assign a weight w_{0i} which is equal to the density

$$f_{Y_0|X_0=x}(y_0) = \begin{cases} \frac{1}{1.2x-0.7x} & 0.7x \leq y_0 \leq 1.2x \\ 0 & \text{otherwise.} \end{cases} \tag{54}$$

since $Y_0|X_0 = x \sim \mathcal{U}(0.7x, 1.2x)$. y_0 is the first element from the vector of measurements $\mathbf{y}$ we have been given in the data material *population_2023.mat*. Then, we form our initial estimate $\hat{X}_0$ of the true population by taking

$$\hat{X}_0 = \frac{\sum_{i=1}^{1000} x_{1i} w_{1i}}{\sum_{i=1}^{1000} w_{1i}} \tag{55}$$

Finally, before moving on to generation 1 and beyond, we do a resampling of our observations x_{0i}, randomly selecting 1000 observations (with replacement) out of the 1000 we have but with probabilities proportional to the weights of the observations.

This concludes step 0 of the sampling. For steps 1 through 50, we repeat the same process with only one difference, namely the distribution of X. Where X_0 followed a $\mathcal{U}(0.6, 0.99)$, we have for other $k = 1, 2, ..., 50$

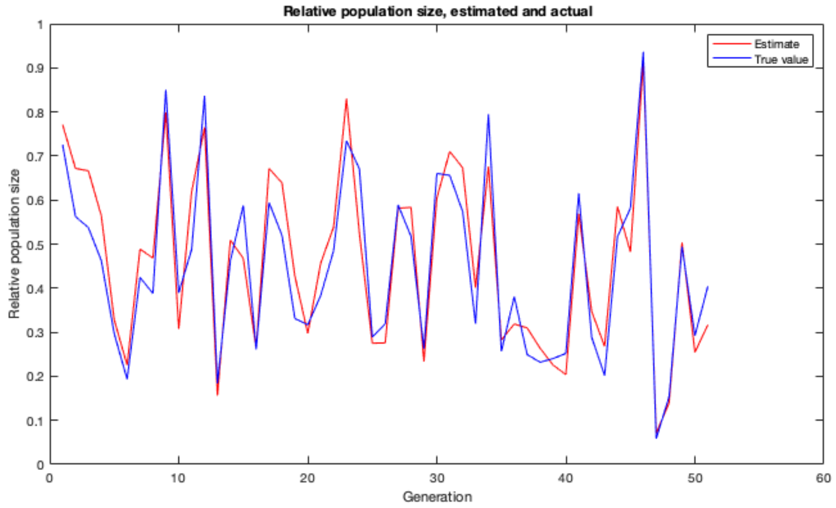

Figure 1: Estimated (red) and actual (blue) values of the relative population size X in generations 0-50.

that

$$X_k = B_k X_{k-1}(1 - X_{k-1}),\ B_k \sim \mathcal{U}(0.9, 3.9) \tag{56}$$

where X_{k-1} is taken from the resampled vector of observations.

This process gives us estimates of the relative population size X_k, $k = 0, 1, 2, ..., 50$ which are drawn in figure 1 together with the true values provided in *population_2023.mat*. As we can see, the estimates follow the true values rather closely, though at some points there are clear deviations.

10. b) Confidence intervals

Use the technique on slide 25 of Lecture 7 to make a point wise confidence interval for X_k for $k = 0, 1, 2, ..., 50$. Check if the true X_k is between the upper and lower limit for $k = 0, 1, 2, ..., 50$.

To create confidence intervals, we ran the same simulation as in problem 10a), but at each generation we also created a cumulative normalized weight sum to act as an empirical distribution function of the estimate $\hat{X}_k$. We then found the indexes of the 2.5th and 97.5th percentiles of this distribution, and took the observations x_k with these indexes (from a vector of ordered observations). Repeating this at every generation, we saved two vectors of length 51 representing the upper and lower interval bounds for each generation.

The resulting confidence intervals are plotted along with the estimated and actual values in figure 2. We note that the actual value is almost never outside the bounds of the confidence interval - there is no generation where this can be observed in the figure.

However, we also investigated numerically whether any actual values of X_k was outside the bounds of the confidence interval. We noted that the value was greater than the upper bound once (generation 50) and smaller than the lower bound once (generation 42). This means that $2/50 = 4\%$ of the observations were

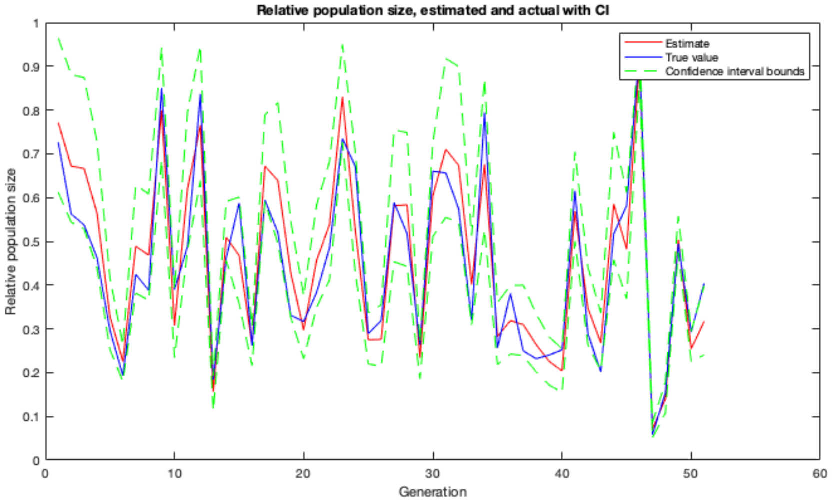

Figure 2: Estimated (red) and actual (blue) values of the relative population size, as well as upper and lower confidence interval bounds of the estimate (green dashed), in generations 0-50.

outside the 95% confidence interval, close to the theoretical proportion of 5%. The interval thus seems correct. Going back to figure 2, we note that the blue line is indeed slightly higher than the upper CI bound at generation 50, and slightly lower than the lower bound at generation 42.