Eric Fischer
UID: 303 759 361
emfischer712@ucla.edu
Stat 202C - Project 1

Importance Sampling and the Effective Number of Samples

Part A: Convergence Rates of 3 Reference Probabilities used in Importance Sampling
Before running the experiment, I hypothesized Alternative 3 would be more effective than Alternative 2 primarily due to the increased radius of its sample space, determined by its larger variance. Alternative 3 has a standard deviation of 4 in comparison to a standard deviation of 1 in Alternative 2. Examining the graphical representation of the sample space, one can intuitively see how a larger radius would correspond to a faster convergence toward the true target distribution. The larger variance of Alternative 3 causes a larger overlap with the target distribution, which means the values of theta for Alternative 3 will more quickly converge to the true theta values.

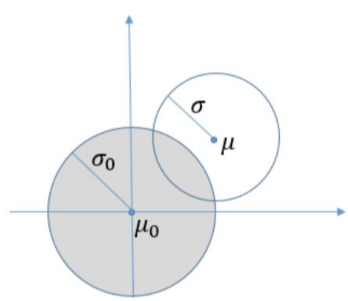
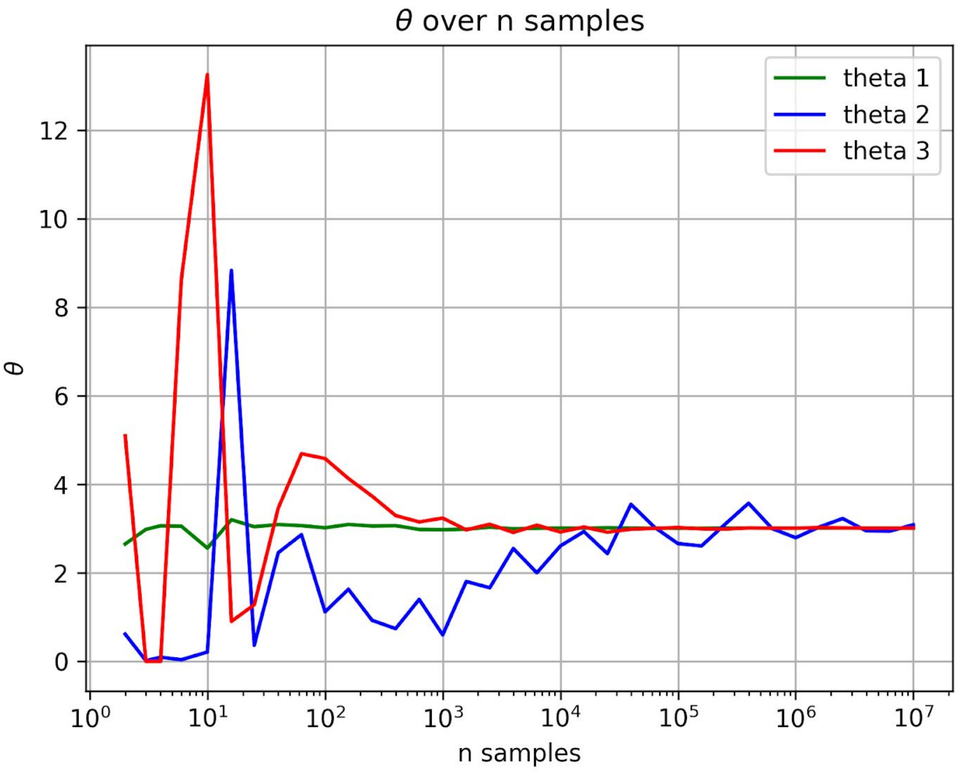


Figure 1: Graphical representation of the sample space

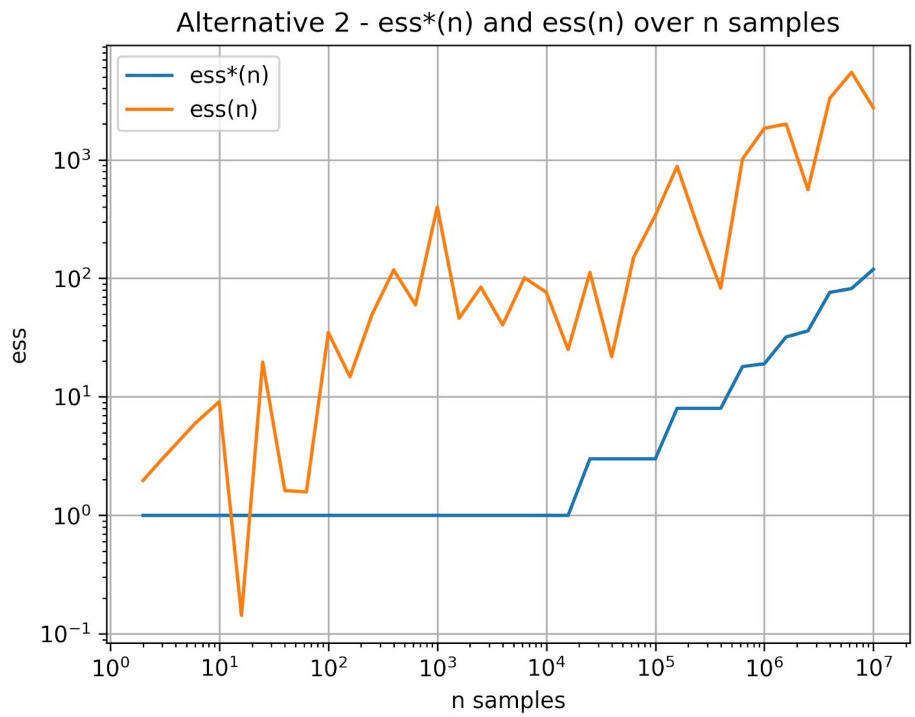
Comparing the convergence rates for the 3 reference probabilities used here for importance sampling, it is clear that by drawing n samples directly from the target distribution (theta 1, green), the theta values will converge quickest. Between Alternatives 2 (theta 2, blue) and 3 (theta 3, red), just as hypothesized, the theta values converge much faster for Alternative 3. The theta values for Alternative 3 converge after roughly 10^4 samples are drawn, while the theta values for Alternative 2 do not converge until roughly 10^7 samples are drawn.

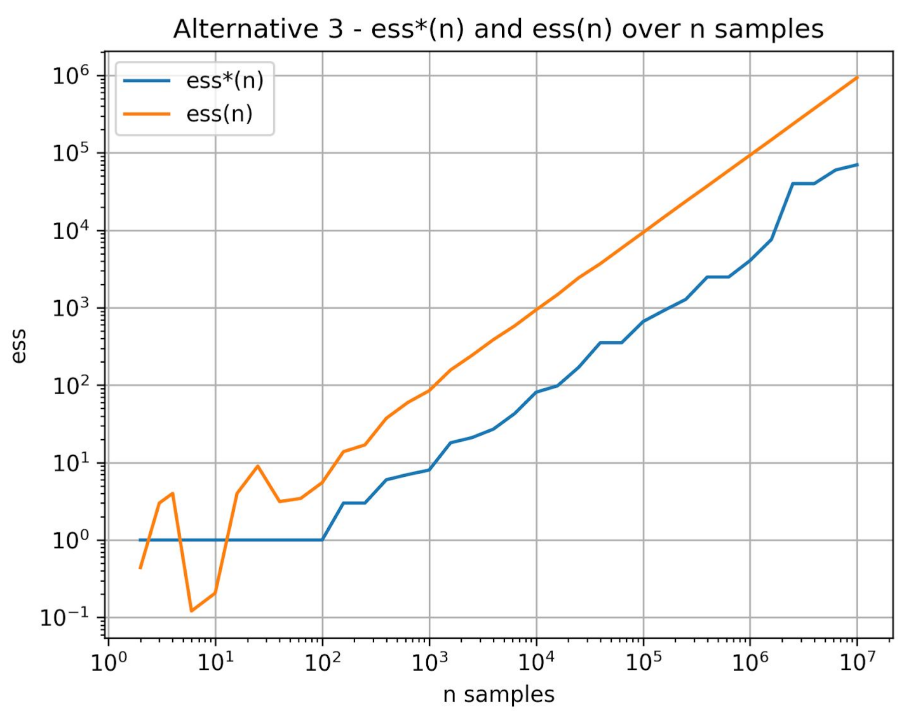


Part B: Effective Samples Sizes for the 3 Reference Probabilities

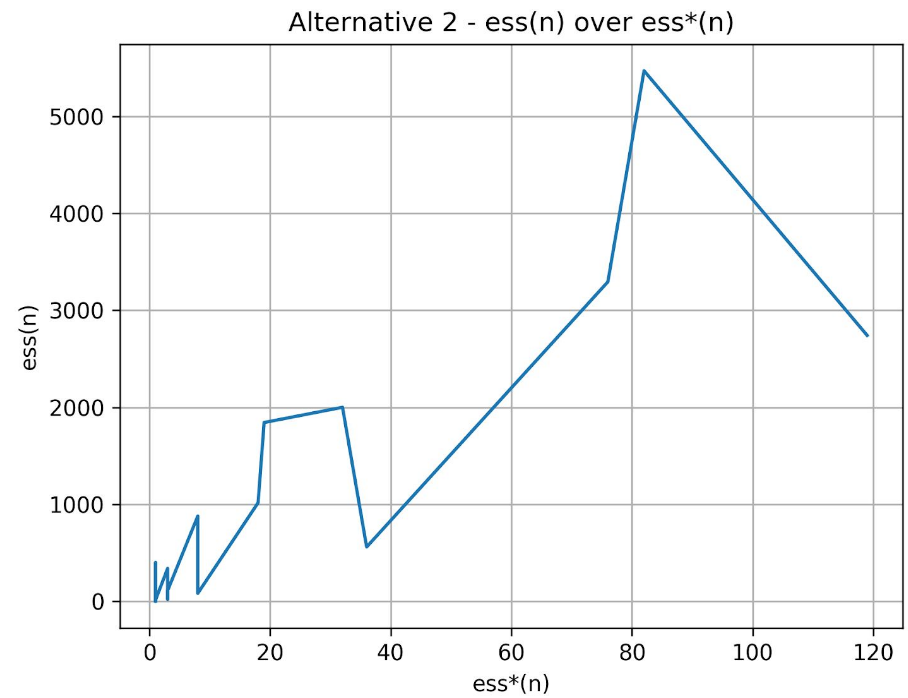
Using ess*(n_1) = n_1 as the truth, as samples in Alternative 1 are each "effective" samples as they are drawn directly from the target distribution, we can compare the effective sample sizes for Alternatives 2 and 3. Plotted below are ess(n_2) and ess*(n_2) over sample size and ess(n_2) over ess*(n_2), and also ess(n_3) and ess*(n_3) over sample size and ess(n_3) over ess*(n_3). The true ess*(n_2) and ess*(n_3) are the sample sizes required for the estimated errors to reach the same level as in Alternative 1.

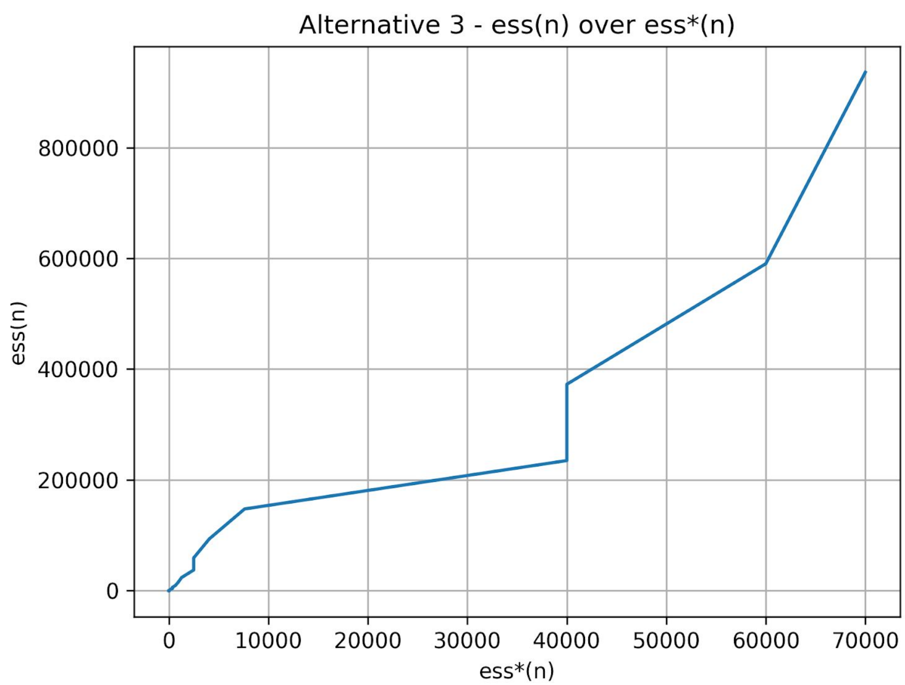
We can see firstly in the plots below that the suggested estimate in **BZ**, eq. (2.6) (labeled as ess(n_2) and ess(n_3)) is not very good. It consistently has higher values for the effective sample size than the true effective sample sizes ess*(n_2) and ess*(n_3), which again are the real sample sizes required for the estimated errors to reach the same level as in Alternative 1. So the estimator in **BZ** suggests too large of quantities for an effective sample size, i.e. quantities we do not need to achieve the same amount of estimated errors and variance from a sample as we would with Alternative 1.





Also, we can note in the plots below of ess(n) over ess*(n) for Alternatives 2 and 3 that this ess(n):ess*(n) ratio is roughly 9:1 for Alternative 3 but much, much larger for Alternative 2 in general. This indicates that Alternative 3 requires a smaller effective sample size for a given sample than Alternative 2 would require, which shows that Alternative 3 with a larger variance is a more effective sampling schema or reference probability for generating effective sample sizes.





Estimating the Number of Self-Avoiding Walks in an (n+1)*(n+1) Grid

Part A: Total Number of Self-Avoiding Walks for n = 10

To estimate the total number of self-avoiding walks (SAWs) for a grid size n = 10, we will use Monte Carlo integration. First we design a trial probability g(x) for a SAW that is easier to sample from than the true target distribution. We then sample M SAWs from g(x) and estimate the total count of SAWs by the mean of 1 / g(x) for SAWs x. The crux of the issue lies in how to design g(x). Here we will examine three different designs for g(x), each assuming the grid size n = 10 and each generating a total of $M = 10^7$ total samples.

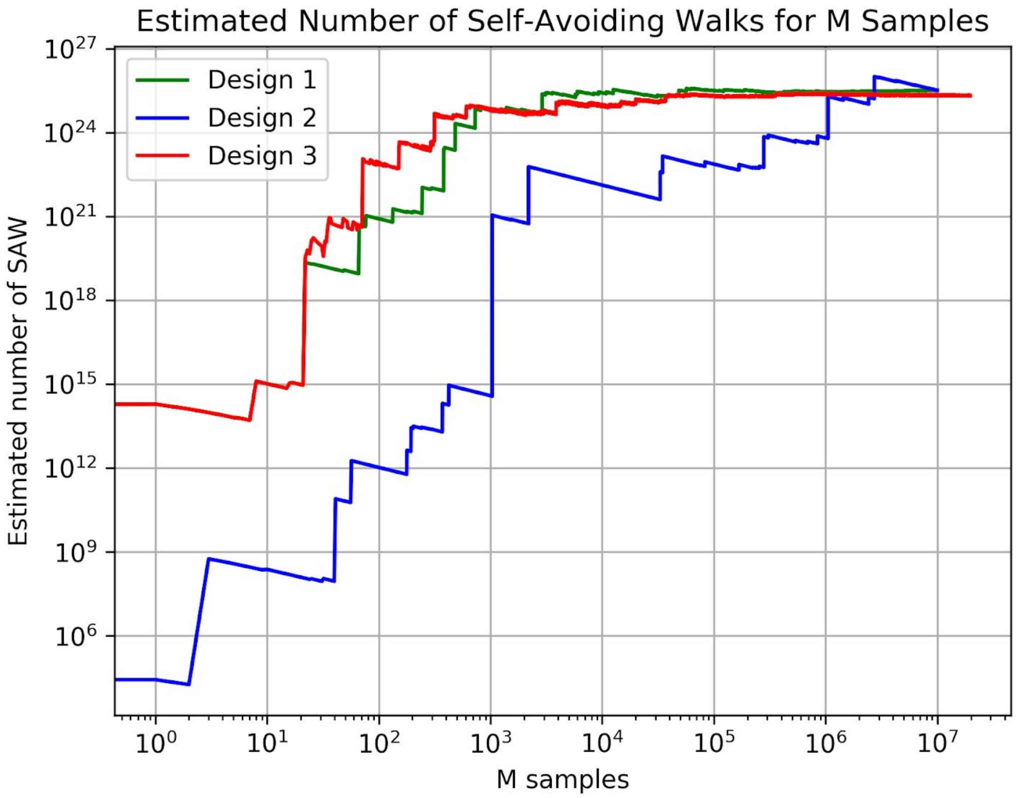
Three Designs for g(x):

Design 1: Here g(x) is equal to the product of 1 / k_j terms, where k_j represents the number of possible choices for the j^{th} move. At each step j, we sample uniformly from the k_j choices. In this design the distribution will resemble a Gaussian because we do not constrain the length of walk. The total estimated number K of SAWs for Design 1 was **3.1804×10^{25}**.

Design 2: Here g(x) is equal to the g(x) of Design 1, but multiplied by $(1 - \varepsilon)^m$, where $\varepsilon = 0.1$ is an early termination probability at each step. This has the effect of lending shorter but more walks overall than design 1. In the log-log plot below of K SAWs over M samples, we can see that Design 2 obviously converges the slowest. This makes sense because Design 2 terminates some paths early, preventing the algorithm from finding paths it might otherwise find. The total estimated number K of SAWs for Design 2 was **3.2535×10^{25}**.

Design 3: This design is a modification of Design 1 to favor longer walks. For any walk longer than 50, Design 3 generates 5 more children from that branch of the walk, reweighting each of the children by $w_0 = w / 5$. This design converges the fastest. The total estimated number K of SAWs for Design 3 was **2.0898×10^{25}**.

Plotting the total number of SAWs K against $M = 10^7$ examples in the log-log plot below for each of the Designs 1, 2, and 3, we can analyze whether the sequential importance sampling process has converged. As expected, Design 2 has the slowest convergence after roughly $M = 10^7$ samples due to its early termination probability. Design 3 has an improved convergence rate over Design 1 due to the behavior of generating 5 children each time a SAW reaches a length of 50. Both Designs 1 and 3 converge after roughly $M = 10^3$ samples, but Design 3 converges sooner and is thus the optimal design in this study.



Part B: Total Number of Self-Avoiding Walks from (0, 0) to (n, n)

We can make a modification to the most optimal sampling procedure, Design 3, to investigate the total number of self-avoiding walks that start from (0, 0) and reach a specific point (n, n). Namely, we now only count the SAWs that successfully reach (n, n) during their walk instead of all the SAWs. Generating 10^6 samples, the true value for the total number of SAWs from (0, 0) to (n, n) is 1.5687×10^{24}.

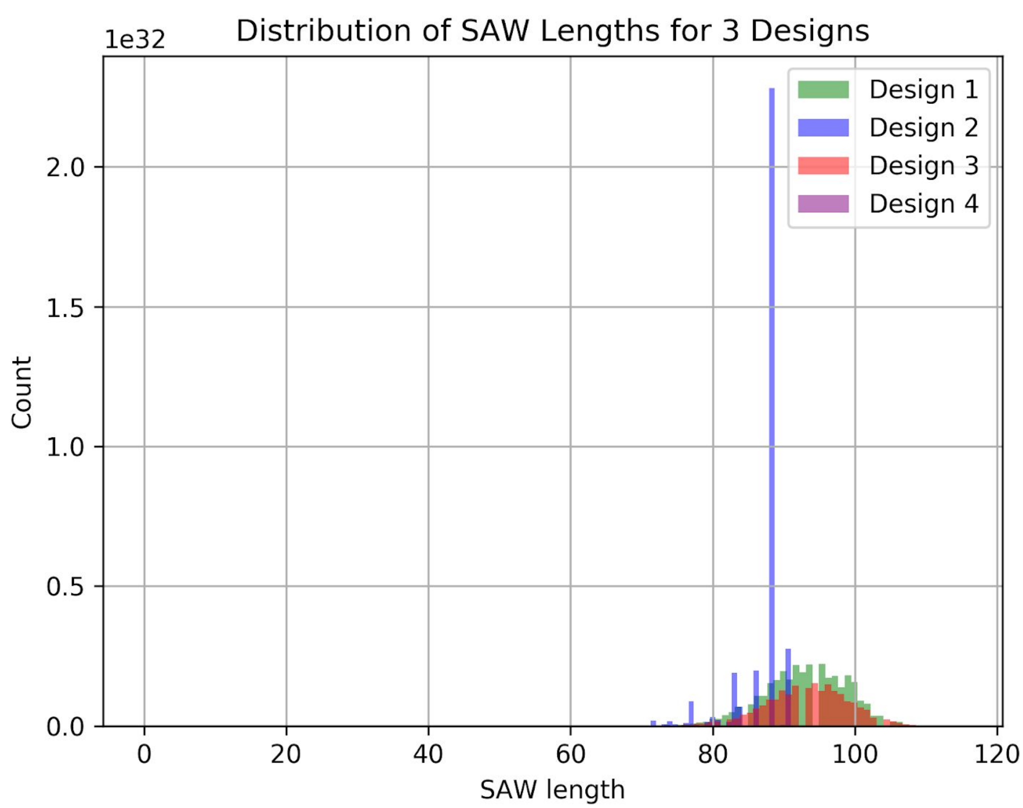
In an effort to approximate this true value as closely as possible, I experimented with generating 3, 5, or 7 children for paths that had grown to a checkpoint length of 25, 50, 70, 80, 90, or 100 and found that it generally helps to spawn *more* children at a given checkpoint length, and it also helps to increase the checkpoint length at which this happens. These conclusions are reasonable, as the longer a SAW path is, the more "special" it is and the more unlikely it is that we have already found SAWs that stem from this one. Of course, if the checkpoint length is too high, e.g. 200, then of course we will never even get the opportunity to generate children (because a path will never reach that length).

Ultimately, the closest approximation to the true value of 1.5687 x 10^{24} reached was for generating 7 children once a path had reached a length of 90. For these parameters, the estimated number of SAWs was 1.2224 x 10^{24}.

Part C: Histogram of Lengths N of the Self-Avoiding Walks

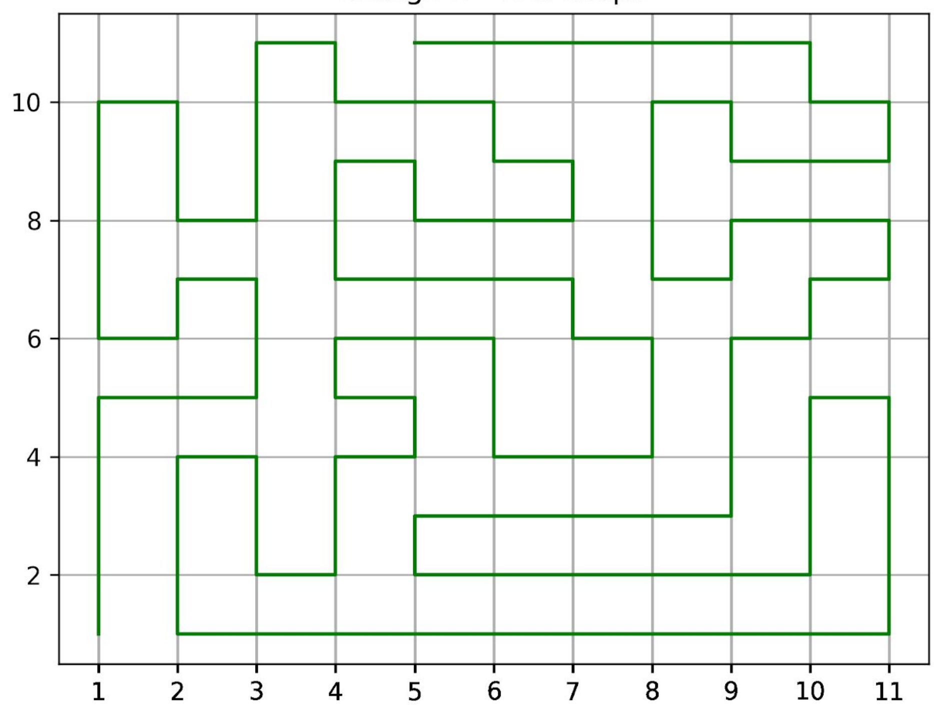
For each experiment in Part A and B, we can plot the distribution of lengths N of the SAWs in a histogram. We need to weight the SAWs in calculating the histogram, because to appropriately determine how many SAWs are in each bin corresponding to a specific walk length, we need to sum the weights of all SAWs with that same length and divide by the total number of SAWs.

As mentioned previously, the distribution of Design 1 resembles a Gaussian because we did not constrain the length of its walk. The distribution of Design 2 is highly concentrated at roughly a SAW length of 85, as the early termination probability eliminated longer walks. Design 3 favored longer walks by generating children, so it is intuitive that the distribution is not as concentrated as Designs 1 and certainly Design 2; it is more dispersed. The distribution of design 4 is similar to Design 3 in that we see more long walks, but it is even more skewed towards long walks and thus less concentrated than Design 3.

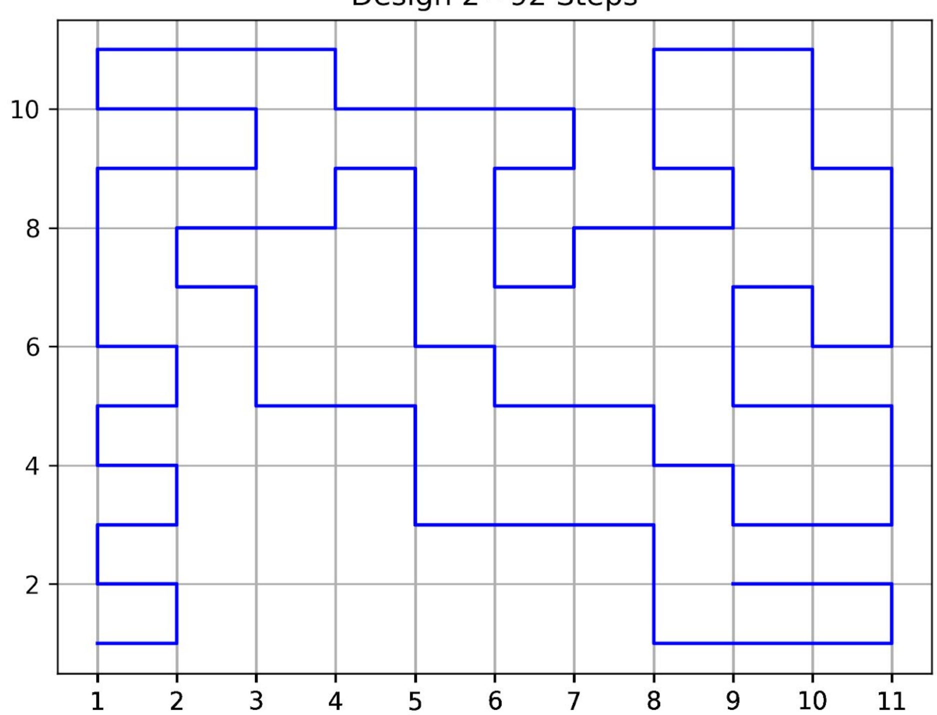


Also, below is a visualization of the longest SAWs for each of the Designs 1, 2, 3, and 4 from Parts A and B.

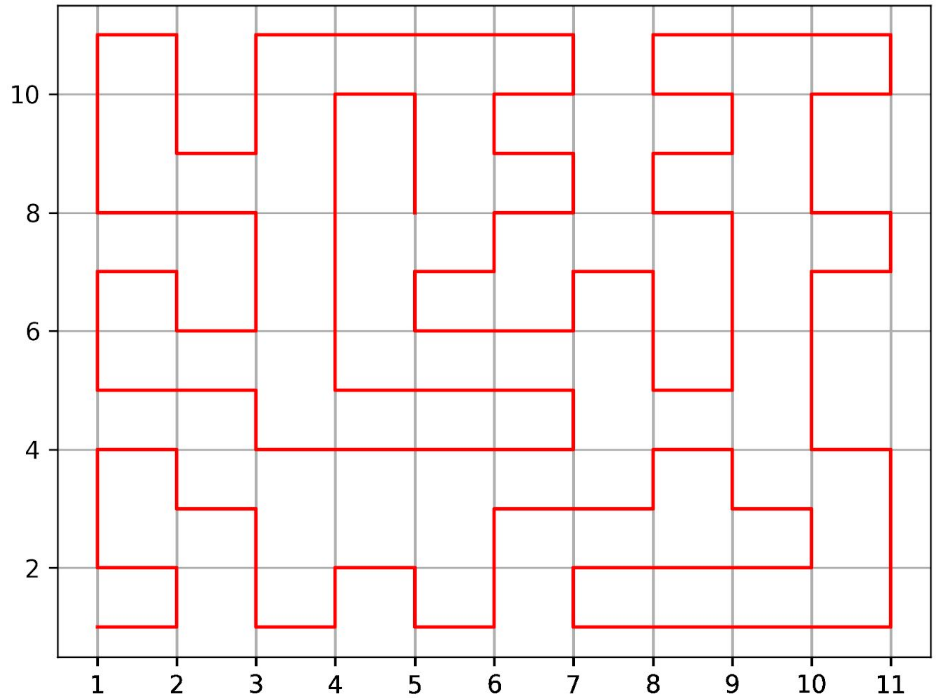
Design 1 - 115 Steps



Design 2 - 92 Steps



Design 3 - 116 Steps



Design 4 - 115 Steps