An Analysis of Monte Carlo Integration using various sampling methods

Mehmet Bahadir Gülhan (14125250)
m.b.guelhan@gmail.com
Nicolas Schrama (11881437)
nicolas.schrama@student.uva.nl

November 22, 2021

Abstract

In this paper, we study effect of sampling on convergence in Monte Carlo integration by calculating the area of the Mandelbrot set using four different sampling methods: pure random sampling, Latin hypercube sampling, orthogonal sampling, and orthogonal sampling with bootstrapping. Moving from pure random sampling to Latin hypercube sampling to orthogonal sampling, the spread of the sampled points becomes more even, resulting in lower variance. This effect is expressed in our results through a faster, more stable convergence to the population mean. Moreover, the radius of the confidence interval decreases faster with the sample size. Additionally, accurate results are achieved at relatively low sample sizes. Finally, we implement bootstrapping to reduce the sample variance. However, we do not observe a decrease of the radius of the confidence interval in our results.

Key words: Stochastic Simulation, Monte Carlo Integration, Sampling methods, Bootstrapping, pure random, latin hypercube, orthogonal

1 Introduction

Stochastic simulations can be used for different applications. For example, no analytical solution is available to some mathematical problems. The next best thing to an analytical solution is an approximation of the solution. Stochastic simulation methods can be useful in such cases.

In this paper we focus on a particular stochastic simulation method, namely the Monte Carlo method. We study the Monte Carlo method by applying Monte Carlo simulations to calculating the area of the Mandelbrot set. Specifically, we aim to answer the following question. *Which sampling method is best for the numerical integration of the Mandelbrot set?* To that end, we study Monte Carlo integration of the Mandelbrot set using pure random sampling, Latin hypercube sampling (LHS), and orthogonal sampling. These sampling methods are compared based on their rate of convergence to the best available estimates of the area of the Mandelbrot set.

The rate of convergence depends on two factors: the number of samples needed for the solution to converge to the true value and the number of iterations used to determine whether a generated sample belongs to the Mandelbrot set. The latter will be expanded upon in section 2. Finally, we implement a bootstrapping method to test whether it improves the accuracy of the Monte Carlo integration.

We set up two sets of experiments. The first set of experiments of Monte Carlo integration is performed given a fixed number of iterations to determine whether a sample is in the Mandelbrot set and a varying number of samples generated per sampling method. In the second set of experiments, the number of samples generated is fixed and the number of iterations varies.

2 Theoretical Background

The Mandelbrot set - famous for its aesthetic qualities, as shown in Figure 1 - consists of all points in the complex plane for which the iteration $z = z^2 + c$ does not diverge [1]. Here, z is initialized to zero. As a result, an exact determination of its area requires infinitely many repetitions. Prior research has focused on statistical techniques to get an estimate that is as close as possible within the bound of computational power.

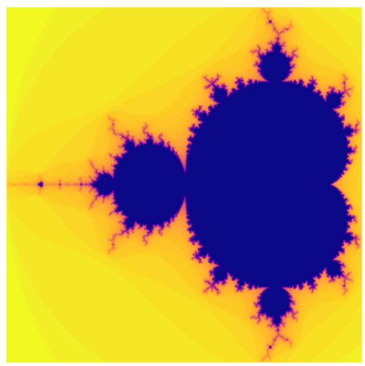

Figure 1: The Mandelbrot set

Estimates use a few basic facts about the Mandelbrot set. First, it is bounded by a circle of radius two, located around the origin of the complex plane, limiting the maximum area to approximately 12.6 [1]. The true size is smaller, however: horizontally, the Mandelbrot spans from x = -2 to x = approximately 0.47. Vertically, it spans approximately from -0.47 to 0.47. This results in a maximum value of approximately 5.5, though the true size is smaller than that.

In the literature, two approaches to estimate the area are predominant [2]. First, pixel counting yields an estimate of 1.50659177 +- 0.00000008 [3]. Second, statistical sampling yields an estimate of 1.506484 +- 0.000004 [1].

The statistical sampling approach is based on Monte Carlo integration [4]. Based on the strong law of large numbers, this method is based on the statistical property that the sample average converges to the population mean as the sample size goes to infinity. As a result, the solution becomes more accurate as the sample size increases. In the case of the Mandelbrot set, this entails sampling many random points in an area larger than the Mandelbrot set, e.g., taking the bounds described above, and using the fraction of randon points belonging to the Mandelbrot set to estimate the area.

Since the random set of sampled points determines the solution obtained from Monte Carlo integration, the sampling method used to obtain is influential on (the accuracy of) the result. In the simplest case, points are randomly drawn from a uniform distribution, i.e., pure random sampling. Stratification strategies result in a more even spread of the points over the sampling space, hence reducing variance and increasing convergence to the population mean.

In LHS [5], the sample space is divided into equally sized intervals along every dimension. The number of intervals is equal to the total sample size. Only one sample may be drawn per interval per dimension. This results in a negative correlation between sampled points.

Furthermore, in orthogonal sampling [6], the sample space is first divided into subspaces. Then, all sample points are selected to ensure that the set of all sampled points is a Latin hypercube sample and that each subspace is sampled with the same density. This method minimizes the correlation between the sampling dimensions. As a result, the number of unsampled areas in the sample space is minimized, hence requiring fewer samples.

Finally, other methods can be employed to increase convergence. In bootstrapping [7], a total sample n is first drawn. Then, r subsamples of size n/r are drawn from the total sample. As a result, points that appear once in the total sample may appear in multiple subsamples. Finally, the sample mean results from the average of all subsamples.

2.1 Experimental setup

In this paper, we conduct two experiments on four sampling varieties: pure random sampling, LHS, orthogonal sampling, and orthogonal sampling with bootstrapping. Per sampling method, we first keep constant the number of iterations used to determine whether a point is in the Mandelbrot set, while varying the number of sampled points. Second, we vary the number of iterations per point but keep the number of sampled points constant. To obtain a better estimate and to construct confidence intervals, we repeat experiments. Tables 1 and 2 contain for each sampling variety the number of iterations per point, the number of sampled points, and the number of experiments. Though more experiments and sampled points increase the accuracy of the estimate, this research was conducted in limited time and with limited computational power.

Furthermore, we use ten subsamples for bootstrapping.

Sampling method	# iterations per point	# sampled points	# experiments
Pure Random, Latin Hypercube	10000	50, 100, 250, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000	25
Orthogonal, Orthogonal + Bootstrapping	10000	100, 900, 2500, 4900, 10000	25

Table 1: Experimental setup for experiments keeping constant the number of iterations

Sampling method	# iterations per point	# sampled points	# experiments
Pure Random, Latin Hypercube	50, 100, 250, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000	10000	25
Orthogonal, Orthogonal + Bootstrapping	100, 900, 2500, 4900, 10000	10000	25

Table 2: Experimental setup for experiments keeping constant the number of sampled points

Finally, the Random Number Generators used for drawing random samples are Python's random package and numpy.random. The random number generator algorithm used in these packages is the Mersenne Twister algorithm [8].

3 Results

3.1 Pure Random Sampling

First, we estimate the area of the Mandelbrot set using pure random sampling. Figure 2 presents the estimates for different values of the number of sampled points.

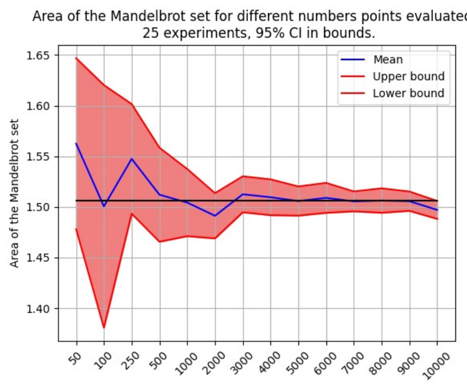

Figure 2: Estimates of the area of the Mandelbrot set for different numbers of sampled points according to pure random sampling. The sample average of 25 experiments is shown in blue, the estimate from the literature (1.506) is in black. The 95% confidence interval is given in red.

Figure 2 illustrates a few statistical properties. First, the variance between experiments decreases as the number of sampled points increases. The reason is that the sample standard deviation is inversely correlated with the sample size. Second, as the sample size increases, the sample average converges to the population mean. This is showcased by a more stable sample average that is also closer to the population mean.

Finally, the improvement in the accuracy decreases as the sample size decreases. This is shown in Figure 3.

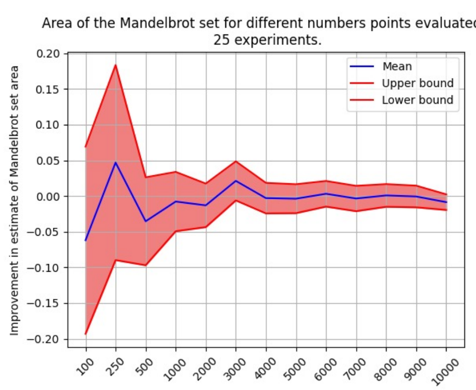

Figure 3: The improvement in the accuracy of the estimate of the area of the Mandelbrot set as the sample size increases.

Figure 4 presents the estimate of the area of the Mandelbrot set for different values of the number iterations used to determine whether a sampled point belongs to the Mandelbrot set.

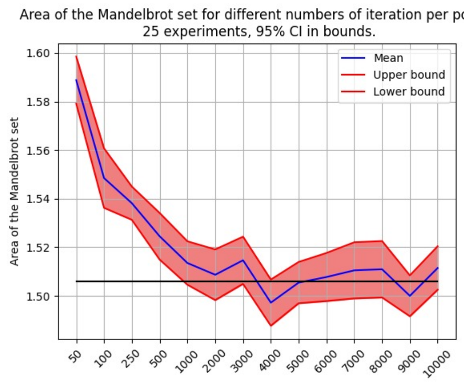

Figure 4: Estimates of the area of the Mandelbrot set for different numbers of iterations to determine whether a sampled points belongs to the Mandelbrot set according to pure random sampling. The sample average of 25 experiments is shown in blue, the estimate from the literature (1.506) is in black. The 95% confidence interval is given in red.

Intuitively, as the number of iterations per point increases, the algorithm finds that some points that do not diverge within x iterations do after more iterations. Hence, as the number of iterations increases, points wrongfully deemed stable are deleted from the Mandelbrot set. Hence, the estimate of the area of the Mandelbrot set decreases as the number of iterations increases. This is also illustrated by Figure 4. Furthermore, as the number of iterations increases, the sample size is uninfluenced. Hence, there is no effect on the variance or the 95% confidence interval. This is illustrated in Figure 4 by the stable width of the 95% confidence interval.

Finally, for completeness, we vary both the number of sampled points and the number of iterations, using fifty experiments. The results are presented in Figure 5.

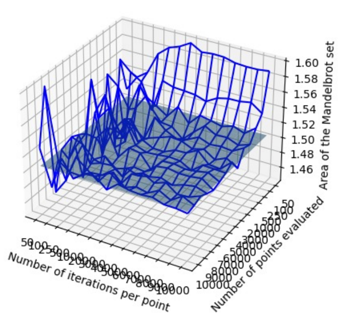

Figure 5: Estimates of the area of the Mandelbrot set for different numbers of sampled points and for different numbers of iterations to determine whether a sampled points belongs to the Mandelbrot set according to pure random sampling. The sample average of 50 experiments is shown by the blue wireframe, the estimate from the literature (1.506) is given the blue surface. The 95% confidence interval is omitted for readability.

Figure 5 shows that the speed of convergence to the population mean decreases rapidly as the number of points evaluated increases, independent of the number of iterations.

3.2 Latin Hypercube Sampling

In this section, we discuss the results of the experiments using LHS. In Figure 6, the experiment with the LHS method with different sample sizes is shown. The LHS method already shows an improvement in convergence, with the confidence region becoming small for large sample sizes. The LHS method is also shown to yield more accurate results as the sample size increases. The convergence of the sample mean to the population mean is also more stable than in the case of pure random sampling.

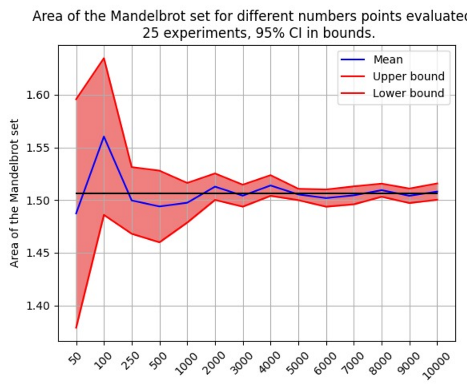

Figure 6: Estimates of the area of the Mandelbrot set for different numbers of sampled points according to LHS. The sample average of 25 experiments is shown in blue, the estimate from the literature (1.506) is in black. The 95% confidence interval is given in red.

For the experiment with different iteration sizes, as seen in Figure 7, the sample mean also converges to the population mean faster than for pure random sampling. As the iteration size increases, the sample mean is also seen to be more stable around the population mean. This can be explained by the fact that the LHS method generates more evenly spread samples, hence the calculations of the area are more accurate.

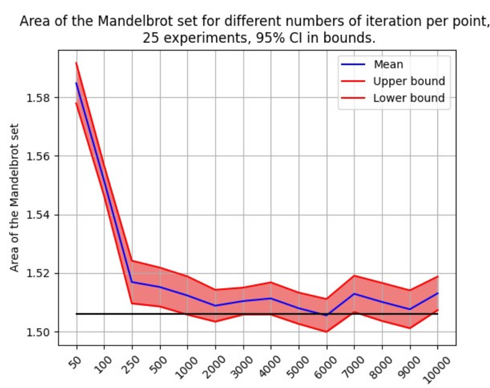

Figure 7: Estimates of the area of the Mandelbrot set for different numbers of iterations to determine whether a sampled points belongs to the Mandelbrot set according to pure LHS. The sample average of 25 experiments is shown in blue, the estimate from the literature (1.506) is in black. The 95% confidence interval is given in red.

3.3 Orthogonal Sampling

The results of the experiments using orthogonal sampling are shown in Figures 8 and 9. The points that are sampled using orthogonal sampling are even more evenly spread out than in the LHS method. For the experiment where we look at different sample sizes, this leads to a stable convergence of the sample mean to the population mean, as illustrated by Figure 8. As the sample size grows, the confidence bounds tighten around the sample mean. This means calculating the area of the Mandelbrot set using Monte Carlo simulations with orthogonal sampling yields accurate results for relatively low sample sizes of up to 10000.

Figure 9 shows that the rate of convergence as the iterations increase closely resembles that of the LHS sampling experiment results, however with a smaller confidence region.

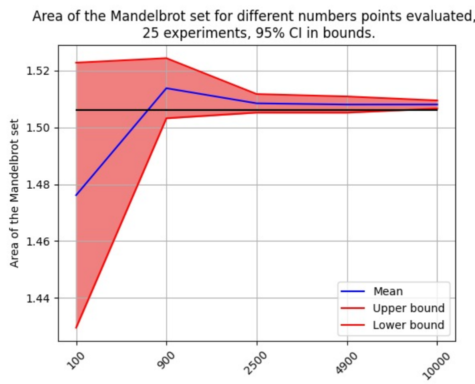

Figure 8: Estimates of the area of the Mandelbrot set for different numbers of sampled points according to orthogonal sampling. The sample average of 25 experiments is shown in blue, the estimate from the literature (1.506) is in black. The 95% confidence interval is given in red.

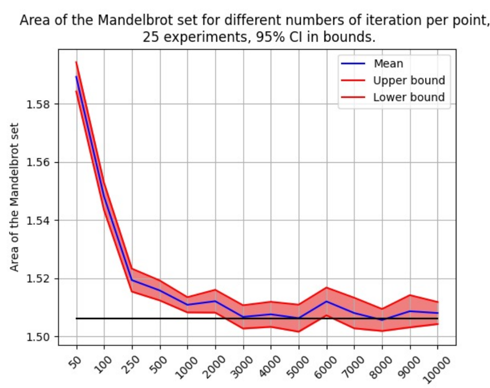

Figure 9: Estimates of the area of the Mandelbrot set for different numbers of iterations to determine whether a sampled points belongs to the Mandelbrot set according to orthogonal sampling. The sample average of 25 experiments is shown in blue, the estimate from the literature (1.506) is in black. The 95% confidence interval is given in red.

3.4 Bootstrapping

Finally, we present the results of the bootstrapping method. We expected the implementation of the bootstrapping to further improve on the convergence of the orthogonal sampling experiments. In particular, we expected the standard deviation, and thus the confidence region, to decrease. As can be seen in figures 10 and 11 this is not the case, however. The convergence of the sample mean appears to be slower and the sample mean is less stable around the population mean. The confidence region is also wider. An explanation for this finding may be that the number of subsamples r we considered is too small. In both experiments we opted to make $r = 10$ subsamples of size n/r. Possibly, increasing r would yield better results.

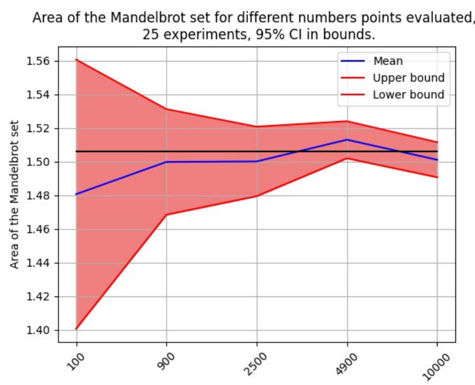

Figure 10: Estimates of the area of the Mandelbrot set for different numbers of sampled points according to orthogonal sampling with bootstrapping. The sample average of 25 experiments is shown in blue, the estimate from the literature (1.506) is in black. The 95% confidence interval is given in red.

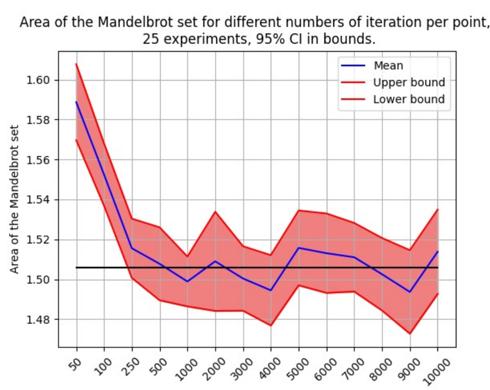

Figure 11: Estimates of the area of the Mandelbrot set for different numbers of iterations to determine whether a sampled points belongs to the Mandelbrot set according to orthogonal sampling with bootstrapping. The sample average of 25 experiments is shown in blue, the estimate from the literature (1.506) is in black. The 95% confidence interval is given in red.

4 Discussion

In this paper, we study Monte Carlo integration by calculating the area of the Mandelbrot set using four different sampling methods: pure random sampling, Latin hypercube sampling, orthogonal sampling, and orthogonal sampling with bootstrapping. Specifically, we compare these sampling methods based on their rate of convergence to the best available estimates of the area of the Mandelbrot set.

We set up two sets of experiments. The first set of experiments of Monte Carlo integration is performed given a fixed number of iterations to determine whether a sample is in the Mandelbrot set. In the second set of experiments, the number of samples generated is fixed and the number of iterations varies.

Across sampling methods, we consistently observe a few statistical properties. First, the variance between experiments decreases as the number of sampled points increases. Second, as the sample size increases, the sample average converges to the population mean. Third, the improvement in the accuracy decreases as the sample size decreases. On the other hand, increasing the number of iterations per point improves accuracy by better classifying which points belong to the Mandelbrot set. However, no statistical forces are at play since the sample size is unaffected. Hence, there is no effect on the variance or the 95% confidence interval.

Furthermore, moving from pure random sampling to Latin hypercube sampling to orthogonal sampling, the spread of the sampled points becomes more even, resulting in lower variance. This effect is expressed in our results through a faster, more stable convergence to the population mean. Moreover, the radius of the confidence interval decreases faster. Additionally, accurate results are achieved at relatively low sample sizes.

Finally, we implement bootstrapping to reduce the sample variance. However, we do not find that bootstrapping decreases the radius of the confidence interval. Possibly, the number of subsamples in our application is inappropriate.

References

[1] Mitchel, K. (2001): A Statistical Investigation of the Area of the Mandelbrot set. Fractalus. `https://www.fractalus.com/kerry/articles/area/mandelbrot-area.html`

[2] Weisstein, Eric W. "Mandelbrot Set." From MathWorld–A Wolfram Web Resource. `https://mathworld.wolfram.com/MandelbrotSet.html`

[3] Lesmoir-Godon, N., Rood, W., & Ralph Edney (2000). Introducing Fractal Geometry. Icon Books UK & Totem Books USA, p.97.

[4] Munafo, R. (2003). Area of the Mandelbrot set. Mandelbrot Set Glossary and Encyclopedia `http://www.mrob.com/pub/muency/areaofthemandelbrotset.html`

[5] Tang, B. (1994). A theorem for selecting oa-based latin hypercubes using a distance criterion. *Communications in Statistics - Theory and Methods*, 23(7), 2047–2058.

[6] Leary, S., Bhaskar, A., & Keane, A. (2003). Optimal orthogonal-array-based latin hypercubes. *Journal of Applied Statistics*, 30(5), 585–598.

[7] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1).

[8] Mersenne Twister (MT19937). (z.d.). NumPy.
https://numpy.org/doc/stable/reference/random/bit_generators/mt19937.html