MNRAS **000**, 1–5 (2020) Preprint 29 September 2020

Nth Dimensional Integral Convergence for Montecarlo Methodologies

A. F. Gómez-Muñoz,[1★]
[1]*Instituto de Física, Facultad de Ciencias Exáctas y Naturales, Universidad de Antioquia (UdeA), Medellín, Colombia*

29 September 2020

ABSTRACT
Montecarlo methodologies are quite general and often used in science to solve optimization, numerical integration and sampling problems. One of its main applications is the estimation of the mean value of some distribution; which improves as more samples are taken. In this article we first show how a better convergence could improve the results of the Ising Model and the determination of a potential in a galactic disk. Then we propose 3 convergence methodologies (weak, medium, and strong) to determine a mean value estimate. Afterwards, a routine for Nth dimensional Montecarlo integration is proposed for each methodology, and all of them are applied over a double Gaussian. We then conclude based in the results for several trials. The weak convergence is found to have fundamental flaws, but is good a first approximation. On the other hand, the medium and strong convergence are found consistent with multivariate Gaussian like function. Finally, we highlight the CLT and the concept of confidence interval as fundamental cores of the strong convergence methodology constructed.

Key words: ising model – heat bath algorithm – montecarlo integration – montecarlo convergence – CLT applied – mean value estimation.

1 INTRODUCTION

Montecarlo methods are a broad class of computational algorithms that are based in random sampling to obtain numerical results, which in principle permit to solve any problem with a probabilistic interpretation. Therefore, this methodologies are quite general and often used in science to solve optimization, numerical integration and sampling problems. Some examples of the application of Montecarlo methods are found in: the calculation of Quantum Mechanical observables, evolution of statistical physical systems, Galactic Evolution, business risk analysis, Artificial Intelligence applied to games, estimate the impact of pollution, and others Roese (2014); Weinzierl (2000).

In some applications as the heat bath algorithm or direct integration the results are often obtained directly from an estimate of a mean value of some distribution, and its error is directly obtained from the estimate deviation of such mean value. Naturally, the problem of how many samples of the distribution in consideration are needed to obtain a given precision with a certain probability arise; this is precisely what we seek to answer.

For that, we first expose succinctly the Heat Bath Algorithm and a Nth dimensional Montecarlo integration methodology. Then we show how the convergence could be a problem for the results of the specific heat capacity in the Ising Model and in the calculation of a galactic disk potential. In order to improve the results of this examples and more general ones, three types of convergence are proposed for a mean value estimate. A weak convergence that consist of an heuristic approach which use the difference between two estimates of the mean value. A medium convergence that simply associates the estimate of the deviation of the mean value with the error of the respective estimate, without taking into consideration the nature of the distribution and hence doesn't considering the probability of the estimate of being in an interval. Finally, a strong convergence is considered, which with the use of the central limit value theorem permit us to obtain the estimate of the integral, E, in consideration (or more generally mean value) with a certain error e and probability p, that is, the number of samples N is chosen such that E is between $(E-e, E+e)$ with probability of at least p.

Three algorithms are then constructed for the convergences proposed in the case of the Nth dimensional Montecarlo integration. This algorithms are then analyzed in a 2-dimensional Gaussian. As a results general conclusions about the algorithms are made. The weak one show fundamental flaws when N is not sufficiently high, still it can be useful as a fist approximation. The medium and the strong convergence can lead to consistent results but with fundamental differences. In the strong convergence, the requirement of a confidence level for $(E-e, E+e)$ likely affects the number of samples needed and the Central Limit Theorem plays fundamental role.

★ E-mail: andres.gomez27@udea.edu.co (UdeA)

2 THEORETICAL FRAMEWORK AND GENERAL ALGORITHMIC PROCEDURE

2.1 Ising model

The Ising model is easy to define, but its behavior is wonderfully rich. It is one of few exactly solvable models where we can actually compute thermodynamic quantities and interpret them, it exhibits symmetry breaks, it has a temperature where it suffer a two phase transition, and more (see Chang (2019)). This model consist in considering N particle, each fixed in a point of a lattice, so that in each site i we have a variable σ_i which only takes on the values +1 and -1. Ultimately we want to work with the full infinite lattice to model the thermodynamic limit. One approach to doing this is to first work with finite lattices and then try to take an "infinite volume limit" in which the finite lattice approaches the full infinite lattice in some sense; for example by gradually increasing a square grid.
In the present work we consider this case, so that the interaction experienced by each particle is due to its neighborhoods, such that the the energy of the particle i, E_i, can be written as 1, where J is the exchange integral, taken as 1.

$$E_i = H_i = -J \sum_{\langle j \rangle} \sigma_i \sigma_j = \sigma_i h \tag{1}$$

2.2 Montecarlo Simulations and The Heat Bath Algorithm

There are several Montecarlo methods that make use of Markov Chains in order to model the dynamics of a grid of spins. One of them is the heath bath algorithm Binder (2018). This one consist in selecting randomly a particle in each iteration, so that thanks to the interaction with its neighbors a transition probability is defined and use it. The successive states that takes the grid of spins once it arrives to the equilibrium distribution are used to estimate the the energy and from the dissipation theorem $c_v = \frac{1}{N k_B T} < (E - < E >)^2 >$.

Then, from the knowledge that the probability of being in an up or down state is given by 2, where $E_{+/-}$ represent the energy (as given by 2.1) such that the particle has either a positive or negative spin, and seeking to satisfy the detailed balance condition, which implies that the Markov process has a stationary distribution, we can propose the transition probability as in expression 3.

$$\pi_{+,-} = \frac{\exp(-\beta E_{+,-})}{\exp(-\beta E_{+,-}) + \exp(-\beta E_{-,+})} \tag{2}$$

$$p_{ij} = min(1, \frac{\pi_j}{\pi_i}) = \frac{1}{1 + e^{-2\beta h \sigma_i}} \tag{3}$$

So, in summary:

1. Random chose of particle k.
2. Apply the rejection method for the probability p_{ij} and in case it is accepted change the spin σ_i to 1.
3. Calculate the macrostate and save the values in order to obtain later a mean value; do it until a good convergence is obtained.

The next sections are mainly based in Feigelson. (2010) and Weinzierl (2000).

2.3 Nth Dimensional Montecarlo Integration

Given an arbitrary function $f(x_1, ..., x_N) = f(\boldsymbol{x})$, we want to calculate $I = \int f(\boldsymbol{x}) d^N \boldsymbol{x}$ using a random sampling. One way to do is supposing a random variable $\boldsymbol{X}$ which follows a uniform distribution in the integration domain (e.g. $U(\boldsymbol{a}, \boldsymbol{b}) = \prod_{i=1}^{N} U(a_i, b_i)$), so that we can write (with $\mathcal{V}$ being the integration volume):

$$I = \int f(\boldsymbol{x}) d^N \boldsymbol{x} = \lim_{N \to \infty} \frac{\mathcal{V}}{N} \sum_{i=0}^{N} f(\boldsymbol{x_i}) = \mathcal{V} E[f(\boldsymbol{X})] \tag{4}$$

From the fact that the elements $f(\boldsymbol{x_i})$ in equation 4 are interchangeable and the definition of limit it is clear that we can propose a procedure to integrate f within a certain error once we sample enough values of $\boldsymbol{X}$. Equivalently, from the law of large numbers it is known that the average obtained from a large number of trials should tend to become closer to the expected value as more trials are performed. So, we propose the following estimate for I:

$$\overline{\mathcal{V} f(\boldsymbol{X})}_N = \frac{\mathcal{V}}{N} \sum_{i=0}^{N} f(\boldsymbol{x_i}) = \overline{\mathcal{V} f}_N \tag{5}$$

Now that we have a way to estimate $E[f(\boldsymbol{X})]$ through the random variable $\overline{f}_N$, we could naturally use the deviation of the last one to characterize how dispersed are our results. Nevertheless, there is not guarantee that there will be a high probability of having our random variable between $(\overline{f} - \sigma, \overline{f} + \sigma)$. In order to attack this problem we could first note that each $f(\boldsymbol{X})$ value is identically and independently sample, so that we it can be written that:

$$Var[\mathcal{V}\overline{f}] = Var\left[\frac{\mathcal{V}}{N} \sum_{i=0}^{N} f(\boldsymbol{x_i})\right] = \frac{\mathcal{V}^2}{N^2} \sum_{i=0}^{N} Var[f(\boldsymbol{x_i})] = \frac{\mathcal{V}^2}{N} \sigma_f^2 \tag{6}$$

where σ_f^2 is estimated by:

$$S^2 = \frac{\mathcal{V}^2}{N-1} \sum_{i=0}^{N} (f(\boldsymbol{x_i}) - \overline{f})^2 \tag{7}$$

2.4 Confidence Interval

Now, even with the useful result given by 6 that shows us that the method converge to the expected value at rate of $\frac{1}{\sqrt{N}}$ we still have not establish a interval in which we expect our estimate to be with certain probability; such interval is called confidence interval and its probability is refereed as confidence level.

Normally, the confidence level α is designated before examining the data and its cut-off values are given by equation 8. Nevertheless, here we take the advantage from the fact that σ_f^2 can be made as little as we want such that for a sufficiently small range we can find the appropriate N to associate the confidence level that we desire. In order to accomplish this we need to know in addition how $\overline{f}_N$ are distributed. For that we use the Central Limit Theorem.

$$\frac{\alpha}{2} = \int_0^{x_{left}} f(\boldsymbol{x}) d^N \boldsymbol{x} \qquad 1 - \frac{\alpha}{2} = \int_0^{x_{right}} f(\boldsymbol{x}) d^N \boldsymbol{x} \tag{8}$$

2.5 Central Limit Theorem (CLT)

The CLT states that if $S_1, S_2, ..., S_n$ are random samples each of size N, with $\bar{X}$ as the sample mean, taken from a population with overall mean μ and finite variance σ^2, then the variable $Z = \left(\frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}}\right)$ as $N \to \infty$ tend to follows standard normal distribution. In other words, we know that $Z_2 = \left(\frac{\overline{f}_N - E[f]}{\sigma_f/\sqrt{N}}\right)$ will begin to follow a standard normal. With this consideration we are now able to propose a convergence methodologies with a confidence interval.

3 PROCEDURE, RESULTS AND DISCUSSION

3.1 The Convergence Problem: The Specific Heat in the Ising Model and a Galactic Potential Example

First, with the procedure explained in 2.2 the thermodynamic variable c_v is calculated. The results are observed in figure 1. It is valuable to add that the curves obtained in each case does not depend of the seed used (in the random number generation) for the thermalization process. Different seeds lead to quite similar results; this is in agreement with the fact that the system is ergodic and the fact that we end up mapping representative configurations. The convergence problem is observed in figure 1, since the discontinuities along the c_v-axis for N = 64 and N = 32 can be resolved by demanding a certain error from $\overline{E^2}(T)$ and $\overline{E}(T)$.

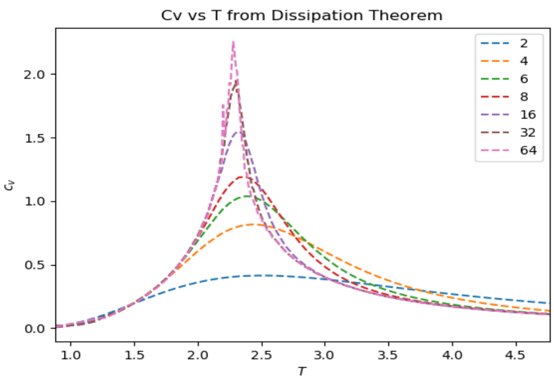

Figure 1: It illustrates the results $c_v(T)$ for the square sides labeled at the right and with periodic conditions. For the construction of each $c_v(T)$ 640x10^6 configurations were took into account.

Another example in which the convergence could be a problem is mapping the potential of a disk galaxy given by 9 (see Binney & Tremaine (1987)). The result of the potential along the Rz plane integrating with Montecarlo is shown in figure 2. Again it has slight discontinuities which can be naturally resolved by limiting the error of each value of the potential. In addition, the farthest regions converge more rapidly and therefore instead of sampling a fixed number of points to integrate for each (R, z) value, an optimal number could be used (by specifying error and confidence level); procedure which could also speed up the final result.

$$\Phi(R,z) = 4G \int_0^{\infty} da \sin^{-1}\left(\frac{2}{\sqrt{+} + \sqrt{-}}\right) \frac{d}{da} \int_a^{\infty} dR' \frac{R'\Sigma(R')}{\sqrt{R'^2 - a^2}} \tag{9}$$

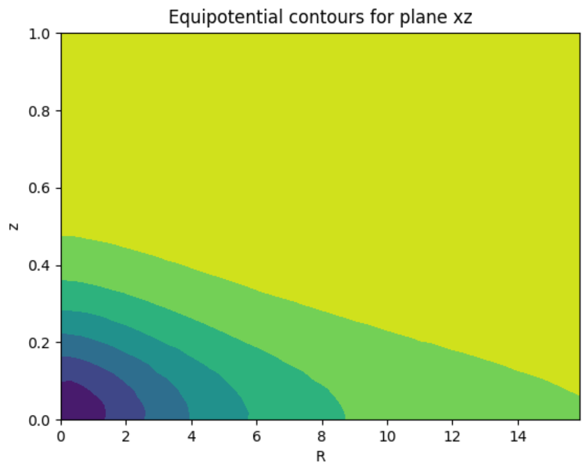

Figure 2: Potential contour in the Rz plane.

3.2 Weak Convergence

According to equation 4 for a sufficiently large N we have a number ϵ such that for two different arbitrary sample sets $(x_1^{(j)}, ..., x_N^{(j)}; j = 1, 2)$ we can write 10:

$$\left| I - \frac{\nu}{N}\sum_{i=0}^{N} f(\boldsymbol{x}_i^{(j)}) \right| \leq \epsilon \quad \left| \frac{\nu}{N}\sum_{i=0}^{N} f(\boldsymbol{x}_i^{(2)}) - \frac{\nu}{N}\sum_{i=0}^{N} f(\boldsymbol{x}_i^{(1)}) \right| \leq 2\epsilon \tag{10}$$

And therefore to evaluate this convergence we could propose to analyze the right inequality in 10. Nevertheless, this leads to an heuristic method, since strictly speaking it should be verified for any pair of sets; which is not possible. Despite of this, it could be used as a first approximation to analyze the results that are being obtained before proceeding with more complicated and computationally expensive methods. Whit this in mind we propose the following algorithm and test it with the integral 11:

1. Estimate I and σ_I^2 with N samples according to 5, 6, and 7.
2. If σ_I estimate is less than ϵ proceed, in not $N \equiv 2N$ and go to 1.
3. Estimate I and σ_I^2 with N new samples and evaluate criteria 10.
4. If criteria 10 is meet return (I, σ_I) guesses; if not $N \equiv 2N$ and go to 1.

$$f(x,y) = \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y} \int_{0,0}^{10,10} dxdye^{-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}} \tag{11}$$

The results obtained by demanding that ϵ in 10 is 0.05 vary considerably, table 1 illustrate the results obtained in different repetitions of the experiment. It can be observed that for most test the integral is obtained at least with the accuracy desired. Nevertheless, the σ_I estimated is in each case an underestimate as the relative error shows. In addition, the number of points N used in each estimate vary considerably. Test5 illustrates that when N is initialized low enough not sufficient representative points are obtained, and hence the estimates of (I, σ_I) can be underestimated.

	I	σ_I	|I - $\overline{fV}$|	N
test1	0.2371	0.0007	0.01	400
test2	0.2508	0.0007	0.0008	400
test3	0.256575	0.0004	0.006	1600
test4	0.2316	0.0004	0.02	800
test5	0.182	0.001	0.06	40

Table 1: Weak convergence results. The first row shows the values being estimated. Initially $N = 100$; except test5 ($N = 10$).

3.3 Medium Convergence

In order to account for the general variability and the underestimation of σ_I we propose as hypothesis that this could be solved by directly calculating $Ntrial$ estimates of I, once we have found one $\overline{f}_N$ for which the estimated error given by 6 and 7 is less than the one desired. In this way directly from $(\overline{Vf}_1, ..., \overline{Vf}_{Ntrial})$, each calculated with N points, σ_I is estimated. The following algorithm is the result of the considerations done:

1. Do steps 1 and 2 of weak convergence algorithm.
3. Calculate $(\overline{Vf}_1, ..., \overline{Vf}_{Ntrial})$ with $Ntrial = 100$.
4. Estimate σ_I with $(\overline{Vf}_1, ..., \overline{Vf}_{Ntrial})$ directly using 7. If the error ϵ is less than the desired, return (I, σ_I) estimates; if not goes again to 1.

Table 2 illustrates the results obtained in different repetitions by demanding $\epsilon = 0.05$. It is observed that the estimations have a greater accuracy and precision; I end up having a small relative error contained in the deviation estimated. In addition, the number of N does not vary considerably; off course this is also the result of the fact that N increase exponentially in the algorithm.

	I	σ_I	|I - $\overline{fV}$|	N
test1	0.25	0.03	0.004	1600
test2	0.25	0.02	0.001	1600
test3	0.25	0.04	0.001	1600
test4	0.25	0.04	0.0004	1600

Table 2: Medium convergence results. The first row shows the values being estimated. Initially $N = 100$.

3.4 Strong Convergence

Finally to talk about a confidence interval (see 2.4) for $\overline{f}_N$ the medium convergence proposal was modified. Instead of directly calculating the deviation from $(\overline{Vf}_1, ..., \overline{Vf}_{Ntrial})$ we use these values to construct a histogram, and from it we adjust a Gaussian to Z_2 (see section 2.5), so that we can easily evaluate if the estimate have a a probability p of being between $(\overline{Vf}_1 - \epsilon, \overline{Vf}_1 + \epsilon)$. The following algorithm is the result of the considerations done:

	I	σ_I	|I - $\overline{fV}$|	N
test1	0.25	0.02	0.0005	6400
test2	0.25	0.02	0.0002	6400
test3	0.25	0.02	0.0005	6400
test4	0.25	0.02	0.0016	6400

Table 3: Strong convergence results. The first row shows the values being estimated. Initially $N = 100$.

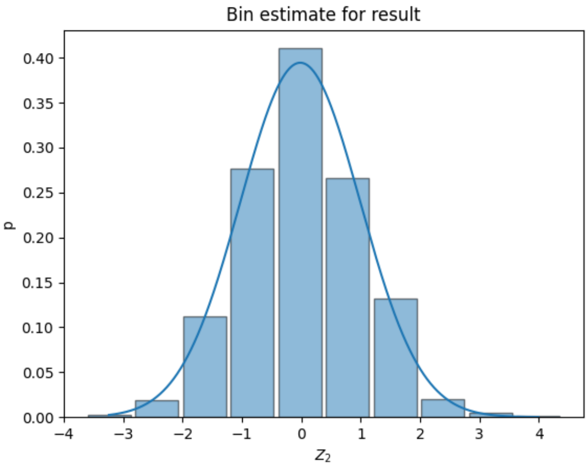

Figure 3: The graph illustrates a histogram of 100 estimates of I, each one obtained with 6400 points, through the normalized variables Z_2. It is observed that the Gaussian fit quite well, in addition it permitted to estimate σ_I with an error of $\approx 2\%$.

1. Do steps 1 - 3 of medium convergence algorithm.
4. Fit a Guassian for Z_2 and estimate (I, σ_I).
5. Calculate the range $(-x_{cut}, x_{cut})$ in which the p percent of the estimates are expected to be found.
6. Finally, if $x_{cut} \leq (\overline{Vf} + \epsilon)$ the estimations are given; otherwise go to step 1.

In the case in which we choose $\epsilon = 0.05$ and $p = 0.68$ the results were very similar to the ones in table 2, which means that for $N = 1600$ there is already a good convergence to a Gaussian. Also, in agreement with this, once we increase the p to 0.95, the I estimate become more precise, the number of samples needed increased (by a factor of 4), and the relative error was contained in the confidence interval; table 3 shows the results and figure 3 the histogram that was fitted for Z_2. In addition, it appears that a more precise estimation of σ_I was obtained, since in all the test the same value was obtained.

4 CONCLUSIONS

The weak convergence proposed can lead to biased estimations in the case in which N is small enough as well as to an underestimation of the error. Despite of this, as an heuristic method can give us a first idea of the value that the integral could take.

A general class to integrate Nth dimensional functions with a given error was written successfully, showing that medium and strong convergences are consistent with functions like a multivariate Gaussian, with the requirement of a confidence level greatly affecting the results. Despite the success, the limitations of the algorithm to integrate correctly a general function are still unclear, for instance this method will likely underperform in spectra like functions.

The CLT and the statistical concept of confidence interval show to be fundamental for the construction of a result with a given error and probability.

ACKNOWLEDGEMENTS

I thank my teacher Jose David Luis, because thanks to his classes I was able propose a solution for the convergence problem of a mean value estimate; in particular to the Nth dimensional integration of a function.

REFERENCES

Binder k., 2018, The Montecarlo Methods in Condensed Matter Physics
Binney J., Tremaine S., 1987, Galactic dynamics, Chapter 2.
Chang J., 2019, The Ising Model
Feigelson. D., 2010, Modern Statistical Methods for Astronomy, Chapter 3
Roese D. P. e. a., 2014, WIREs Comput Stat., 6, 386–392
Weinzierl S., 2000, Introduction to Monte Carlo methods