Computational Statistical Mechanics - Homework 1

Fabrizio Camerin
AtoSiM Track
April 18, 2016

1) *Data are thermalized*

In order to verify whether the data given where effectively thermalized, a plot of the values (one every 75) of the first observable U_1 has been made. Indeed, the data are dispersed within a certain range which is the same for the first and the last values. Consequently, there is no need to avoid the calculation of average quantities for low values of Monte Carlo (MC) time. Below, the distribution of the data for U_1 (for the other four observables something similar is expected) is shown.

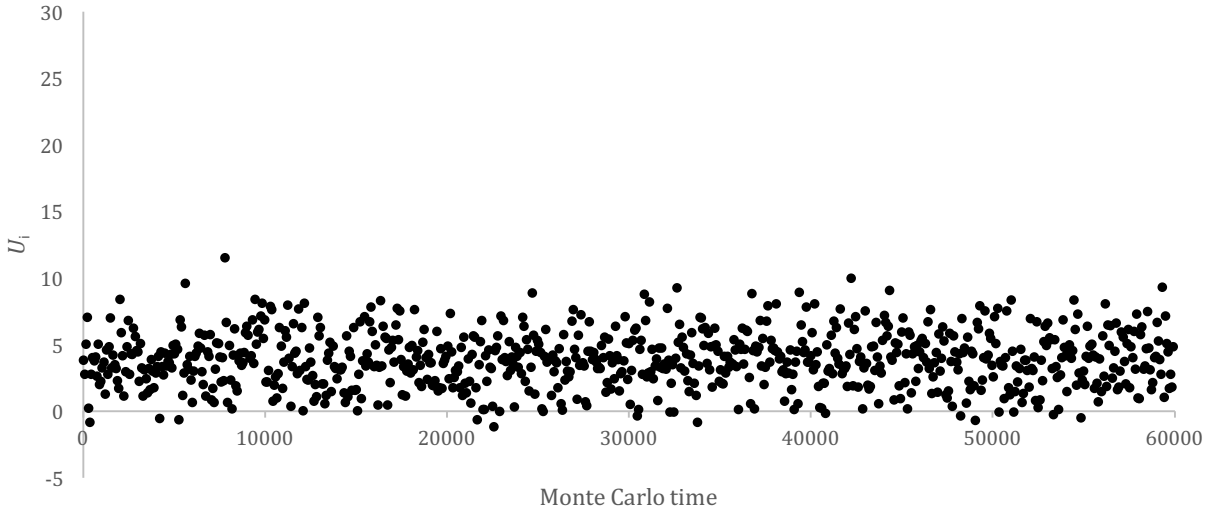


Figure 1.1 - Data are thermalized

2) *Neglecting correlations*

The computation of the averages on the observables neglecting correlations is equivalent to assume that the data we are treating are independent. The formula used to compute the averages $\langle U_i \rangle$ and the errors σ_i are shown in the following.

$$\langle U_i \rangle = \frac{1}{N}\sum_{k=1}^{N} U_i(k)$$

$$\sigma_i = \sqrt{\frac{VarU_i}{N}}$$

with N the number of data available (60000) and $VarU_i = \frac{N}{N-1}\left[\frac{1}{N}\sum_{k=1}^{N} U_i^2(k) - \left(\frac{1}{N}\sum_{k=1}^{N} U_i(k)\right)^2\right]$.

Here, it is taken into account that the estimate of the variance is a biased estimator of the variance itself. Effectively, the correction only introduces the pre-factor $\frac{N}{N-1}$ to the formula of the variance, leading basically to the same result we would have got without taking it into account.
The results are summarized in Table 2.1.

Table 2.1 – Averages and errors neglecting correlations

U_i	$\langle U_i \rangle$	σ_i
U_1	3.897	0.008
U_2	6.808	0.006
U_3	2.029	0.002
U_4	1.256	0.012
U_5	-0.074	0.005

3) *Blocking analysis*

The blocking analysis is a method that allows to take into account the correlations between data.
The core of the method is that averages between pairs of data are computed in an iterative way: while these quantities do not vary during the analysis, the error computed on them depends on the blocking of the data and particularly it increases as a function of the iterations K (blocking length) performed, before reaching a plateau. This behavior is due to the fact that data, at first, are still correlated each other, but then, they become essentially independent. In fact, it can be demonstrated that there is no correlation when

$$VarU_i \approx 2VarU_i^{(1)} \approx \cdots \approx 2^K VarU_i^{(K)} \approx \cdots$$

where the "almost equal" sign has to be replaced by a "lower than" sign when data are, on the contrary, not independent.
Nevertheless, after a certain K, the error starts to oscillate or, in some cases, to decrease: those data cannot be considered reliable anymore given that the amount of blocked data is too small.
Figure 3.1 reports, for the five observables, $2^k VarU_i^{(k)}$ with respect to the blocking length. Once the correct value of K is chosen, the error can be directly computed using

$$\sigma_i = \sqrt{\frac{2^K VarU_i^{(K)}}{N}}$$

In Table 3.1, the errors are reported.

Table 3.1 – Errors from blocking analysis

U_i	σ_i
U_1	0.060
U_2	0.041
U_3	0.007
U_4	0.046
U_5	0.037

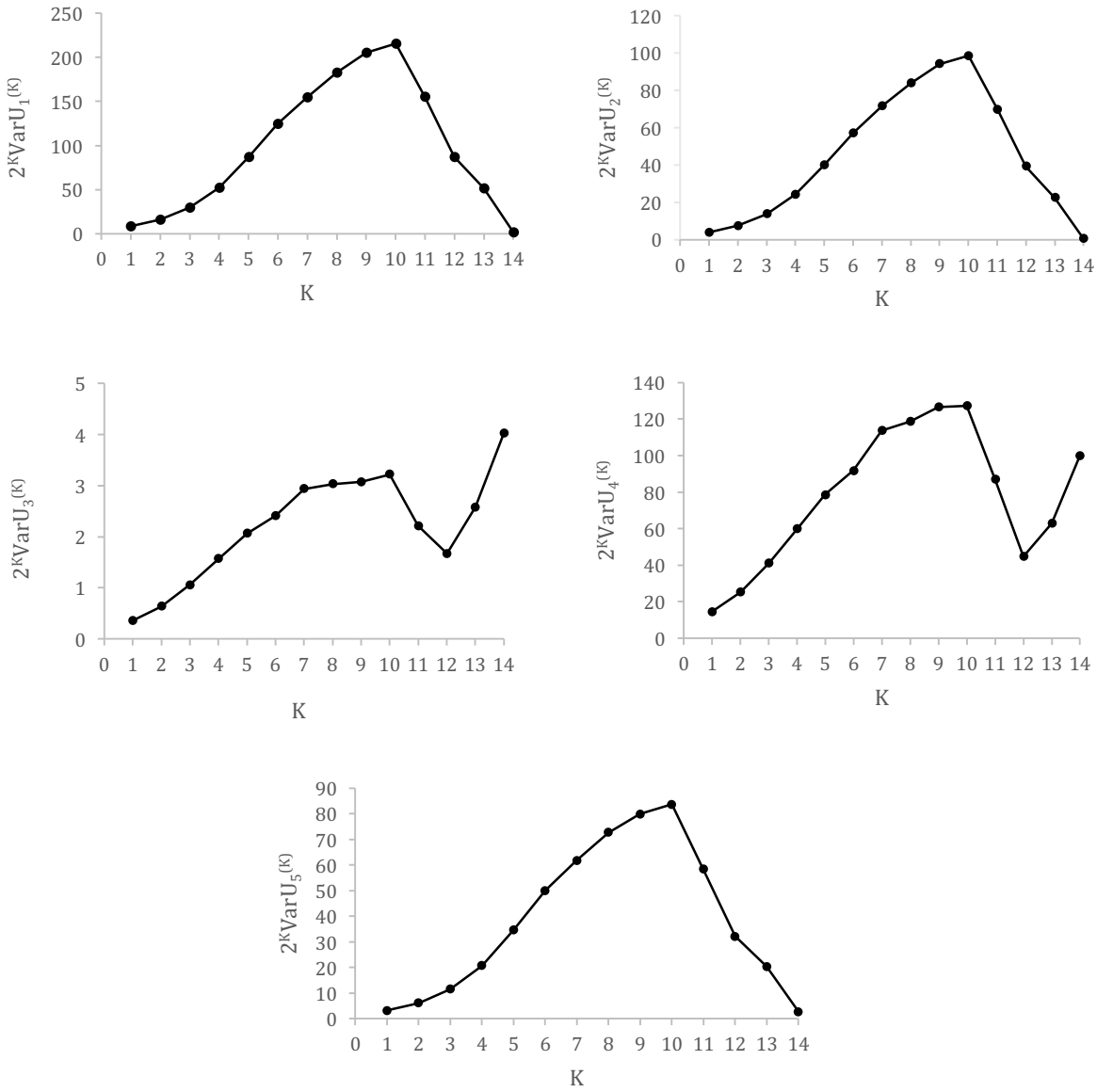


Figure 3.1 - These graphs, one for each observable, allow to determine the error. The x-axes reports the blocking length.

Even though we do not observe a plateau for all the five observable, the value corresponding to $K = 10$ (the highest one) is the one considered in order to get a correct estimate of the error.

4) *Autocorrelation analysis*

The aim of this method is essentially the same of the previous one.
Starting from the usual definition of the error, it can be demonstrated that

$$\sigma_i^2 = \frac{VarU_i}{N}\left(1 + 2\sum_{n=1}^{k}\frac{C_i(n)}{VarU_i}\right)$$

where

$$C_i(n) = \frac{1}{N-n}\left[\sum_{k=1}^{N-n}(U_i(k) - \langle U_i\rangle)\,(U_i(k+n) - \langle U_i\rangle)\right]$$

and $C_i(0) = VarU_i$.
The so called autocorrelation function $C_i(n)$ (with n that goes from 0 to $N-1$) takes into account the correlation between data. In particular, the formula above is derived for translationally invariant system, where the correlation depends only on the difference in time. It is worth to note that in case the autocorrelation function is null, the error is determined as if data were independent. Another way to see it, is by re-writing σ_i^2 as

$$\sigma_i^2 = 2\sigma_{i,independent}^2 t_{i,int}$$

where

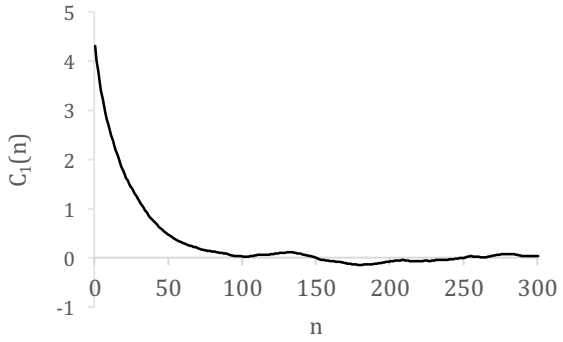
$$t_{i,int}(k) = \frac{1}{2} + \sum_{n=1}^{k}\frac{C_i(n)}{VarU_i}$$

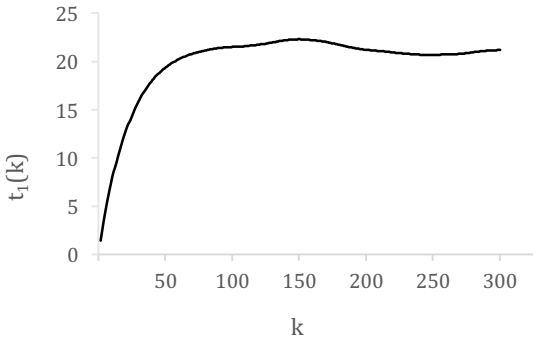
with t_{int} integrated autocorrelation time. Whenever $t_{i,int}$ is greater than 1, the error for correlated data is increased with respect to the one in which data are independent.
It is correct to expect that the autocorrelation function goes to 0 with increasing values of n, indicating that data are no longer dependent. However, while it is decreasing to 0, the function starts to oscillate due to the fact that the amount of terms included in the summation over k in $C_i(n)$ is lowering. Consequently, in order to get a good estimate of the error it is important to understand where to truncate the term $\sum_{n=1}^{k}\frac{C_i(n)}{VarU_i}$. In practice, this is done by calculating the autocorrelation time which represents (except for the 0.5 factor) the area underlying the autocorrelation function. The former is a function that increases and reach an approximately stable value at which we decide the summation has to be truncated. In this way the error in the estimate of σ_i, obtained using,

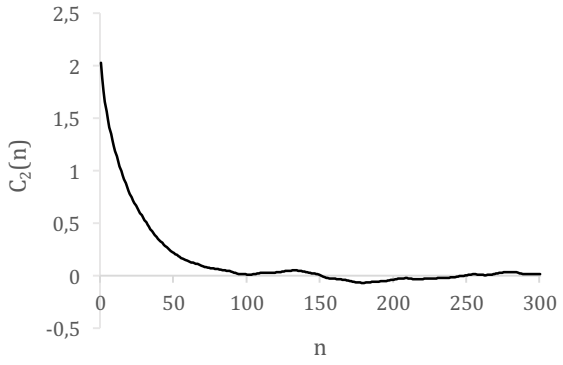
$$\sigma_i = \sqrt{\frac{VarU_i}{N}2t_{i,int}(k)}$$

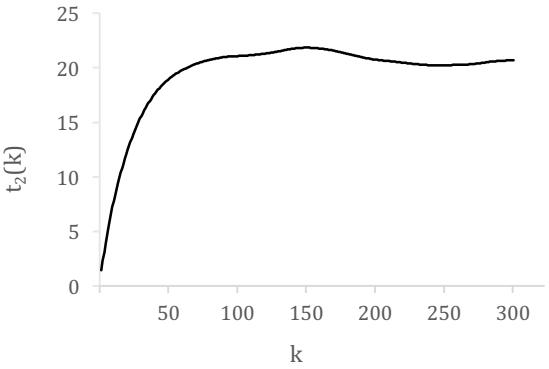
is not greater than the 10%.

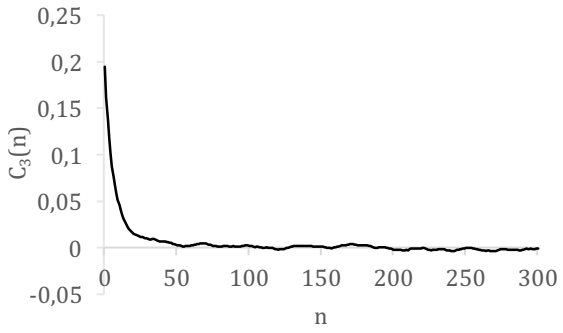
The figure below reproduces on the left hand side the autocorrelation function, and on the right hand side the integrated autocorrelation time. The final errors calculated using the autocorrelation analysis are reported in Table 4.1.

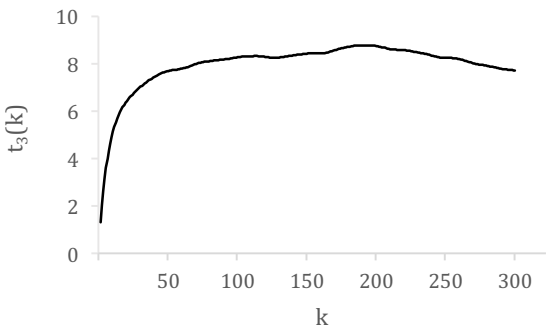


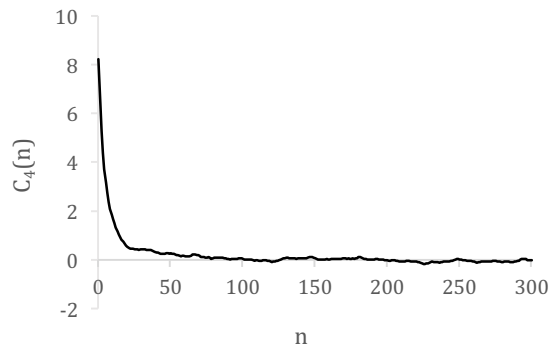


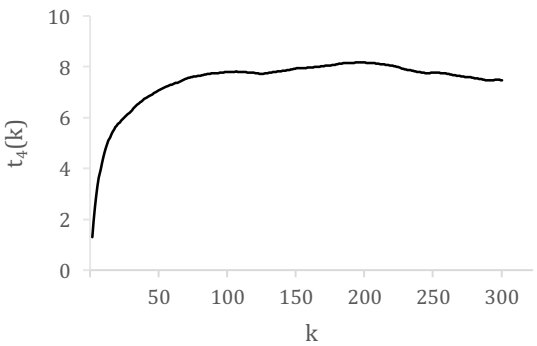












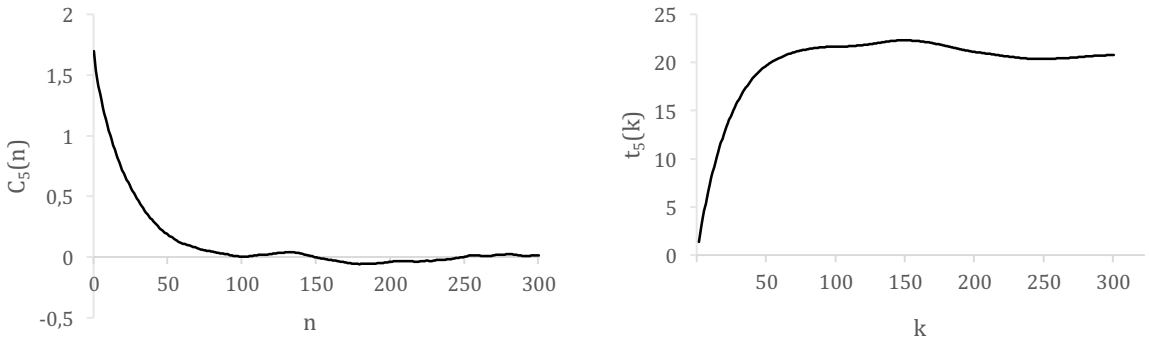


Figure 4.1 - On the left hand side the autocorrelation function as a function of n. On the right hand side the integrated autocorrelation time as a function of k.

Table 4.1 – Averages and errors from autocorrelation analysis

U_i	$t_{i,int}$	σ_i
U_1	22.2	0.056
U_2	21.3	0.038
U_3	8.3	0.007
U_4	7.8	0.046
U_5	21.8	0.035

It can be noticed that the errors computed in this way are comparable to the once calculated using the blocking method. Instead, they are different from the values obtained in 2), and, as expected, they have higher values.

5) *Jackknife, independent-error and worst error formula*

The last task of the homework was to calculate ratios of the form

$$R_i = \frac{\langle U_i \rangle}{\langle U_1 \rangle}$$

given that $i = 2,3,4,5$ and starting from a set of blocked variables with blocks of length 2500 (that can consequently be considered as independent).
In this case the quantities R_i are computed via the Jackknife algorithm, which is a method that intrinsically take into account correlations between numerator and denominator. In particular, the calculation of the errors is easier as we don't have to calculate co-variances between data.
The final formula that allows the application of this algorithm is

$$R_i \equiv J^{JK}_{i,est,final} = NJ_{i,est} - (N-1)J^{JK}_{i,est}$$

where J_{est} corresponds to the ratio of the averages obtained in the standard way

$$J_{i,est} = \frac{\frac{1}{N}\sum_{k=1}^{N} U_i(k)}{\frac{1}{N}\sum_{k=1}^{N} U_1(k)}$$

and

$$J_{i,est}^{JK} = \frac{1}{N}\sum_{m=1}^{N} J_m^{JK} = \frac{1}{N}\sum_{m=1}^{N} \frac{U_{i,m}^{JK}}{U_{1,m}^{JK}} = \frac{1}{N}\sum_{m=1}^{N} \frac{\frac{1}{N-1}\sum_{k=1,k\neq m}^{N} U_{i,m}(k)}{\frac{1}{N-1}\sum_{k=1,k\neq m}^{N} U_{1,m}(k)}$$

The error is instead computed using

$$\sigma_i = \sqrt{(N-1)VarR_i}$$

with

$$VarR_i = \frac{1}{N}\sum_{m=1}^{N} (J_m^{JK} - J_{i,est,final}^{JK})^2$$

The error obtained with the JK method can be compared with two other ways of computing errors of ratios that rely on the independent-error formula and the worst-error formula:

$$\sigma_i^{IE} = \left[\left(\frac{\langle U_i\rangle^2}{\langle U_1\rangle^2}\right)\left(\frac{{\sigma_{U_i}}^2}{\langle U_i\rangle^2} + \frac{{\sigma_{U_1}}^2}{\langle U_1\rangle^2}\right)\right]^{1/2}$$

$$\sigma_i^{WE} = \left|\frac{\langle U_i\rangle}{\langle U_1\rangle}\right|\left|\frac{\sigma_{U_i}}{|\langle U_i\rangle|} + \frac{\sigma_{U_1}}{|\langle U_1\rangle|}\right|$$

The independent error formula is built in such a way that, with respect to the correct formula to compute the error for ratio of averages (in our case estimated with JK method), there is no co-variance. Consequently, depending on whether this quantity was positive or negative, the computed error can be higher or lower: this is indeed what is found throughout the values reported in Table 5.1.
The worst-error formula, as recalled by its name, allows to predict the "worst" possible error and it is obtained maximizing the terms of the correct estimate of the error; for all the quantities R_i, the error computed in this way is the highest.

Table 5.1 – Values of $\boldsymbol{R_i}$ computed with standard averages and with the JK. Errors are calculated using JK, IE, WE.

$\boldsymbol{R_i}$	$\boldsymbol{R_{i,average}}$	$\boldsymbol{R_{i,JK}}$	$\boldsymbol{\sigma_{i,JK}}$	$\boldsymbol{\sigma_{i,IE}}$	$\boldsymbol{\sigma_{i,WE}}$
$\boldsymbol{R_2}$	1.747	1.746	0.034	0.027	0.035
$\boldsymbol{R_3}$	0.521	0.521	0.008	0.008	0.009
$\boldsymbol{R_4}$	0.323	0.322	0.010	0.013	0.017
$\boldsymbol{R_5}$	-0.019	-0.019	0.009	0.009	0.009