SAPIENZA UNIVERSITÀ DI ROMA

Computational Statistical Mechanics Homework 1: Error Analysis

Roberto Verdel Aranda

18.04.2016

ABSTRACT

In this report the error analysis of numerical data generated in a computer simulation is presented. The approaches that were followed are: independence assumption, blocking analysis, autocorrelation analysis and the Jackknife method. The data analyzed were downloaded from `http://elearning2.uniroma1.it/course/view.php?id=2878`.

1 INDEPENDENT DATA

Five data sets, denoted by U_i, with $i = 1, ..., 5$, were analyzed in the first part of this assignment. In this section, the estimate of the averages $\langle U_i \rangle$, and their corresponding errors are computed under the assumption that data are independent. The estimator of the mean value is given by the *sample mean*

$$\overline{U_i} \equiv \overline{U_i(t)} = \frac{1}{N} \sum_{n=1}^{N} U_i(t_n). \tag{1.1}$$

In the case of independent data, the error is simply

$$\sigma_i = \sqrt{\frac{\mathrm{Var}(U_i)}{N}}, \tag{1.2}$$

where the variance can be approximated by the unbiased estimator

$$[\mathrm{Var}(U_i)]_{\mathrm{unbiased}} = \frac{N}{N-1}\Big[\frac{1}{N}\sum_{n=1}^{N} U_i^2(t_n) - \overline{U_i}^2\Big]. \tag{1.3}$$

The results obtained appliying these formulas are shown in Table 1.1.

Table 1.1: Average and error of the data sets assuming statistical independence.

Quantity	**Average**	**Error**
U_1	3.897	0.008
U_2	6.808	0.006
U_3	2.029	0.002
U_4	1.253	0.012
U_5	-0.074	0.005

2 BLOCKING ANALYSIS

Here we employ the blocking technique. The blocked variables are defined as

$$\begin{aligned} U_i^{(1)}(t) &= \frac{1}{2}[U_i(2t-1) + U_i(2t)], \\ U_i^{(k)}(t) &= \frac{1}{2}[U_i^{(k-1)}(2t-1) + U_i^{(k-1)}(2t)], \end{aligned} \tag{2.1}$$

Note that every time we block the data, the original number of data points is reduced by a factor of $1/2^k$, where k indicates the number of blocking operations. For instance, if we have N values of the quantity $U_i(t)$, after the first blocking operation we will get $N/2$ new blocked data points. If we repeat the procedure once more, the size of the new data set will be $N/4$, and so on. If N is odd we can just discard the last point in the data set, provided that the number of measurements is large enough.
If the averages of the blocked data were uncorrelated, we would have that

$$\mathrm{Var}(U_i) \approx 2^1 \mathrm{Var}(U_i^{(1)}) \approx \cdots . \tag{2.2}$$

However, if this is not the case, the following chain of inequalities holds

$$2^k \mathrm{Var}(U_i^{(k)}) > 2^{(k-1)} \mathrm{Var}(U_i^{(k-1)}) > \cdots > \mathrm{Var}(U_i). \tag{2.3}$$

Thus the idea of the method is to look for a block size sufficiently large, such that the averages of the blocks generated are more or less independent. At that point, we will observe that the error stabilizes, namely

$$2^{(j-1)} \mathrm{Var}(U_i^{(j-1)}) \approx 2^j \mathrm{Var}(U_i^{(j)}) \approx \cdots , \tag{2.4}$$

for some j. This is reflected in the appearance of a plateau in the plot of $2^k \mathrm{Var}(U_i^{(k)})$ vs. k. If the value of this plateau is $\sigma^\star$, then the statistical error we are interested in can be approximated as

$$\sigma^2 = \frac{\sigma^\star}{N}. \tag{2.5}$$

Of course, as the number of blocking operations rises, the number of samples will decrease. Therefore, some fluctuations may appear after the plateau, that is for large values of k. The plots of the variance as a function of the number of block operations are shown in Fig. 2.1. We can see that in some of these graphs it is possible to identify a plateau (data U_3 and U_4), which will define the estimator of the error we wish to compute (red dashed line). However, in some

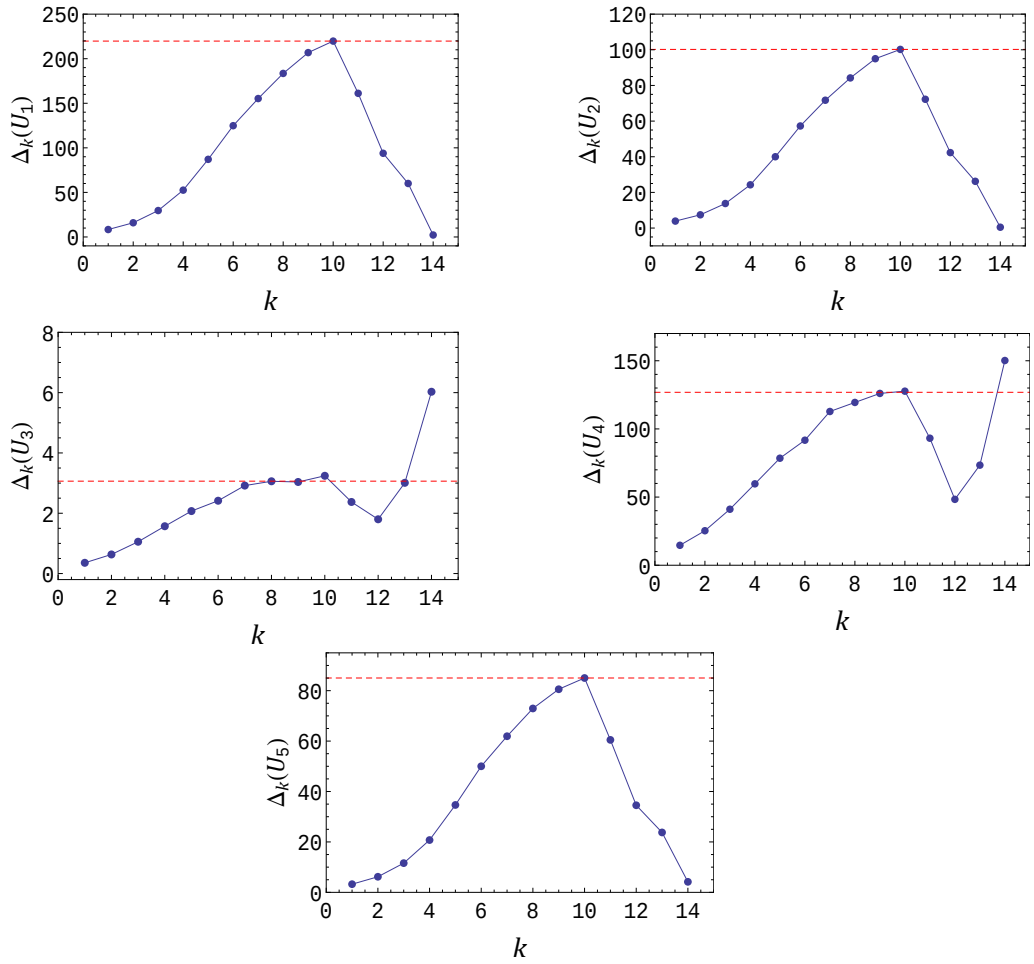

Figure 2.1: Variance of the quantities $\langle U_i \rangle$ as a function of the number of block operations k. The red dashed lines determine the values of $\sigma^\star$ (see Eq. (2.5) and Table 3.1.) The notation $\Delta_k(U_i)$ stands for $2^k \mathrm{Var}(U_i^{(k)})$.

other cases (data sets U_1, U_2, and U_5) there is no formation of a plateau. This indicates that for the determination of such quantities the simulation was probably too short. Nevertheless, in those cases we have considered the highest value obtained as an estimator for the variance we are interested in. The errors that we get by taking the squared root of Eq. (2.5) are summarized in Table 3.1, where they are compared to the results obtained using the autocorrelation analysis. The discussion of such compirison is made in the following section.

3 Autocorrelation analysis

The error of the sample mean, Eq. (1.1), can also be expressed in terms of the *autocorrelation function* $c(m, n; x_0)$, which is defined as

$$c_i(m, n; x_0) = \langle (U_i(t_m) - \langle U \rangle)(U_i(t_n) - \langle U \rangle) \rangle, \tag{3.1}$$

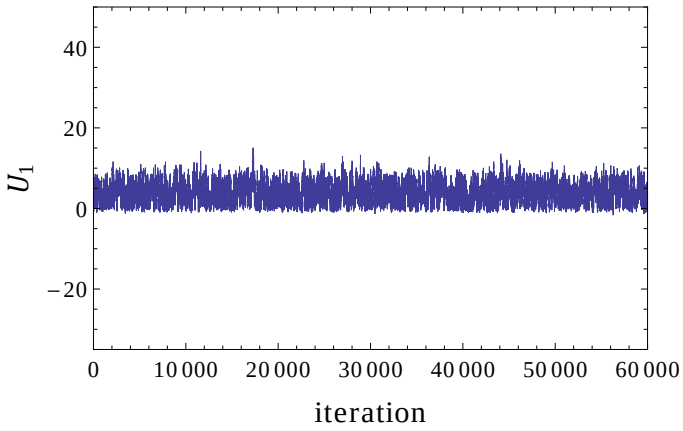

Figure 3.1: Plot of the data set U_1. The data fluctuate around the mean value, so the system under study has already thermalized. Similar results are obtained for the other data sets.

where x_0 is the initial condition. The error can then be written as

$$\sigma_i^2 = \frac{1}{(N+1)^2} \sum_{m=0}^{N} \sum_{n=0}^{N} c_i(m,n;x_0). \tag{3.2}$$

In particular, if we are dealing with equilibrium calculations, as it is case with the data analyzed (see Fig. 3.1), the autocorrelation function depends only on $|k| \equiv |m-n|$ (time-translation invariant). One can then prove that the expressions above become

$$c_i(k) = \frac{1}{N-k} \left[\sum_{n=1}^{N-k} (U_i(t_n) - \overline{U_i})(U_i(t_{n+k}) - \overline{U_i}) \right] \tag{3.3}$$

and,

$$\sigma_i^2 = \sigma_{i,\mathrm{ind}}^2 \cdot 2\tau_{i,\mathrm{int}}, \tag{3.4}$$

with $\sigma_{i,\mathrm{ind}}^2 = \mathrm{Var}(U_i)/N$, and $\tau_{i,\mathrm{int}}$ being the so-called *integrated autocorrelation time*:

$$\tau_{\mathrm{int}} = \frac{1}{2} + \sum_{k=1}^{\infty} \frac{c_i(k)}{c_i(0)}. \tag{3.5}$$

The latter expressions give reliable results for small values of k. Moreover, in order to compute the error we have to truncate the sum in Eq. (3.5). This is done by looking at the plot of the *partial sum* $A_i(n)$:

$$A_i(n) = \sum_{k=1}^{n} \frac{c_i(k)}{c_i(0)}. \tag{3.6}$$

The plots of the partial sums $A_i(n)$, are shown in Fig.3.3. We can observe that in all cases $A_i(n)$ reaches a "maximum" value $A_i^{\max} \equiv A_i(n_{\max})$, before some noise (oscillations) arise. This maximum value is taken as an approximation to the right end side of Eq. (3.5). The errors that are obtained in this way are shown in Table 3.1 along with the errors of the blocking method. In general, the agreement is very good. The greatest discrepancies are found in the cases of U_1, U_2 and U_5, in which the blocking error is larger than the one computed in this section. The reason

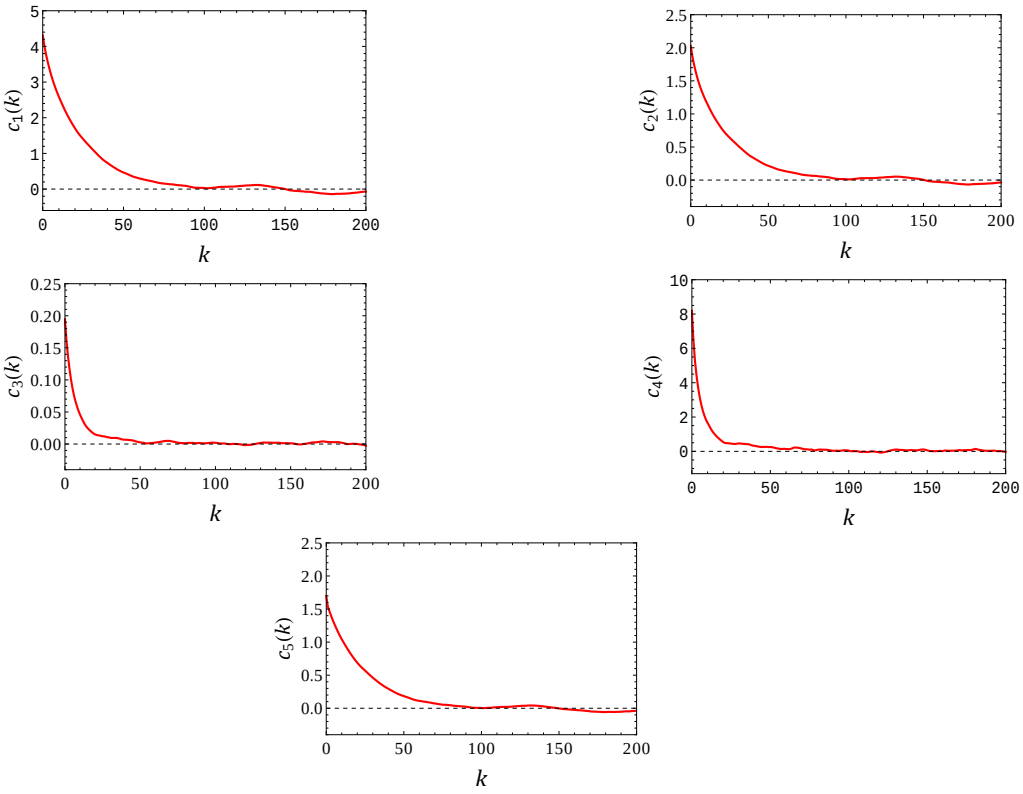

Figure 3.2: Autocorrelation functions.

for this might be related to a point stressed in the previous section, namely that for these data sets no plateau was found in the plots of the variance against the number of bolcking operations, meaning that perhaps the simulations were not large enough to analyze those data sets using the blocking approach.

Table 3.1: Estimates of the error on the samples means of U_i. σ_{Blocking} gives the error using blocking analysis, whereas $\sigma_{\tau_{\text{int}}}$ represents the estimate described in the present section.

Quantity	$\sigma_{\textbf{Blocking}}$	$\sigma_{\tau_{\textbf{int}}}$
U_1	0.061	0.056
U_2	0.041	0.038
U_3	0.007	0.007
U_4	0.046	0.047
U_5	0.038	0.035

4 The Jackknife method

Now we address a different problem. Suppose we have blocked the data sets in blocks of length 2500. The averages of the blocked variables can then be regarded as independent. Indeed, from

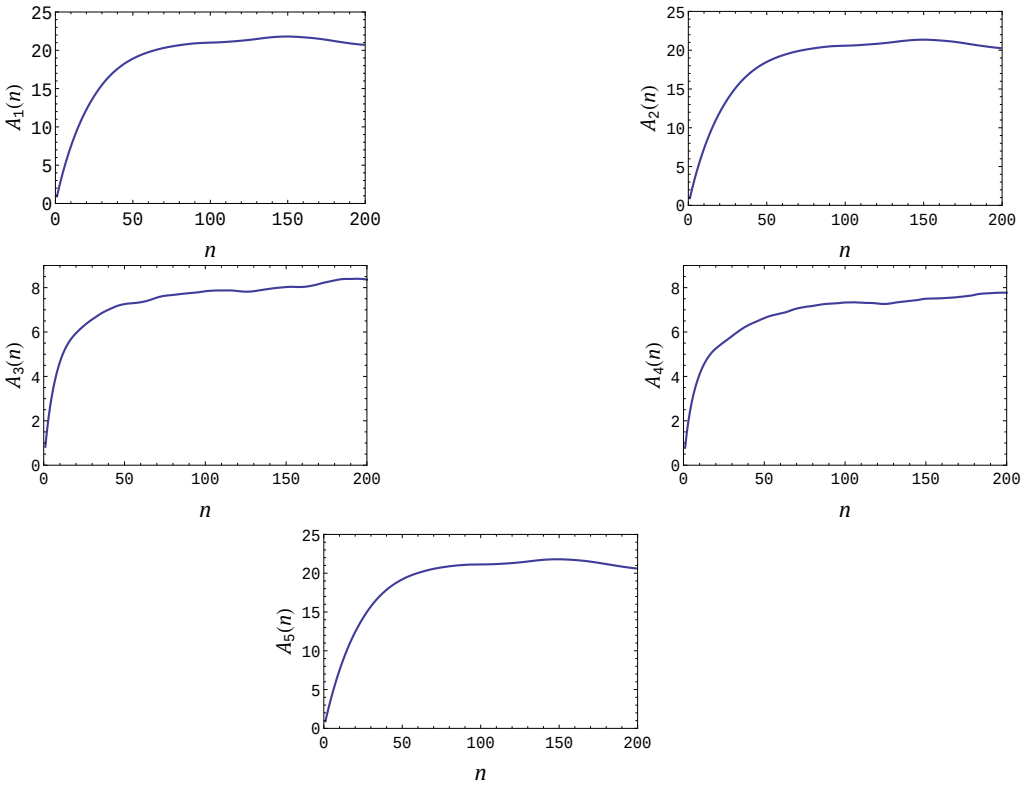

Figure 3.3: Partial sums.

Fig. 2.1, we can see that the error stabilizes after 10 or 11 blocking operations, which correspond to blocks of size 1034 and 2068, respectively. Next, we define the quantities

$$R_i = \frac{\langle U_i \rangle}{\langle U_1 \rangle}, \quad \text{with } i = 2,3,4,5. \tag{4.1}$$

We use the Jackknife method to get an estimate of R_i as well as its error. This is done as follows: first, we define the Jackknife averages

$$\begin{aligned} U_{i,\alpha}^{JK} &= \frac{1}{N-1} \sum_{\beta \neq \alpha}^{N} U_{i,\beta}, \\ U_{1,\alpha}^{JK} &= \frac{1}{N-1} \sum_{\beta \neq \alpha}^{N} U_{1,\beta}, \end{aligned} \tag{4.2}$$

with $\alpha = 1, \cdots, N$. Then, we form the ratios

$$R_\alpha^{JK} = U_{i,\alpha}^{JK} / U_{1,\alpha}^{JK}, \tag{4.3}$$

and get a first estimate, namely

$$R_{\text{est}}^{JK} = \frac{1}{N} \sum_{\alpha=1}^{N} R_\alpha^{JK}. \tag{4.4}$$

Note that R_α^{JK} differs from $R_{\text{est}} = \overline{U_i}/\overline{U1}$, in that the former involves only $N-1$ data points, whereas the latter is computed using N points. Lastly, we calculate the final estimate

$$\hat{R}_{\text{est}}^{JK} = N R_{\text{est}} - (N-1) R_{\text{est}}^{JK}. \tag{4.5}$$

The error is then computed through the following equation

$$\sigma_{JK} = \sqrt{(N-1)\mathrm{Var}(R^{JK})}, \tag{4.6}$$

where

$$\mathrm{Var}(R^{JK}) = \frac{1}{N}\sum_{\alpha=1}^{N}(\hat{R}^{JK}_{\mathrm{est}} - R^{JK}_{\alpha})^2. \tag{4.7}$$

Two other alternative methods that are, however, worse than the Jackknife technique, are the *independent error* approximation and the *worst-error formula*. The error on the estimation of (4.1) in these two approaches is given by

$$\sigma_{IE} = \left[\left(\frac{\overline{U_i}}{\overline{U_1}}\right)^2\left(\frac{\sigma_i^2}{\overline{U_i}^2} + \frac{\sigma_1^2}{\overline{U_1}^2}\right)\right]^{1/2}, \tag{4.8}$$

and

$$\sigma_{WE} = \left|\frac{\overline{U_i}}{\overline{U_1}}\right|\left|\frac{\sigma_i}{|\overline{U_i}|} + \frac{\sigma_1}{|\overline{U_1}|}\right|, \tag{4.9}$$

respectively. In practice, one should not use neither Eq. (4.8) nor Eq. (4.9), for they overestimate the error (the independent error formula can also underestimate it). Nonetheless, here we will employ these techniques in order to make a comparison with the results of the Jackknife method. Such results are shown in Table 4.1. For σ_i and σ_1 in the latter expressions we used the errors calculated with the autocorrelation analysis (see Table 3.1).

Table 4.1: Error made on the estimation of the quantities R_i. The third column (σ_{JK}) corresponds to the Jackknife method. The last two columns give the error computed according to Eqs. (4.8) and (4.9), respectively.

Quantity	$\hat{R}_i^{JK}$	$\sigma_{\mathbf{JK}}$	$\sigma_{\mathbf{IE}}$	$\sigma_{\mathbf{WE}}$
R_2	1.746	0.034	0.027	0.035
R_3	0.521	0.008	0.008	0.009
R_4	0.322	0.010	0.013	0.017
R_5	-0.019	0.009	0.009	0.009