Computational Statistical Mechanics

# Assignment 1: Analysis of Data

Stephanie Grimmel

April 21, 2017

## 1 Introduction

In the following report the results of error analysises carried out on data set as they could have been obtained from Monte Carlo or Molecular Dynamics simulations are presented. The generation of the data is not part of this paper, but it was provided.
The work presented here is part of the course "Computational Statistical Mechanics" given by professor Andrea Pelissetto at "Sapienza - Università di Roma" in spring 2017.

## 2 Verification of Proper Thermalization

In order to verify that the data is thermalized, the results $U_i(t)$ were plotted against the Monte Carlo time. An exemplary plot is shown in figure 1. Since the data does not show a trend towards de- or increasing values for all four different sets but only fluctuations, it is legitimate to assume thermalization and include all the data points in the following analysis.

## 3 Analysis under the Assumption of Independence

Below, the equations for obtaining the estimated averages $\overline{U_i}$ and variances $VarU_i$ of a given data set are provided

$$\overline{U_i} = \frac{1}{N}\sum_{k=1}^{N} U_i(k) \tag{1}$$

$$VarU_i = \frac{N}{N-1}\left[\frac{1}{N}\sum_{k=1}^{N} U_i(k)^2 - \left(\frac{1}{N}\sum_{k=1}^{N} U_i(k)\right)^2\right] \tag{2}$$

$N$ is the number of data points $U_i(k)$ in a set. The factor $\frac{N}{N-1}$ in the expression for the variance, equation 2, although only having a small effect on the result for large $N$, removes the bias from the estimator.
The error $\sigma_i$ for thermalized (i.e. time-translation invariant) date sets is obtained from the variance as follows:

$$\sigma_i = \sqrt{\frac{VarU_i}{N}\left(1 + 2\sum_{\tau=1}^{\infty}\frac{C_i(\tau)}{VarU_i}\right)} \tag{3}$$

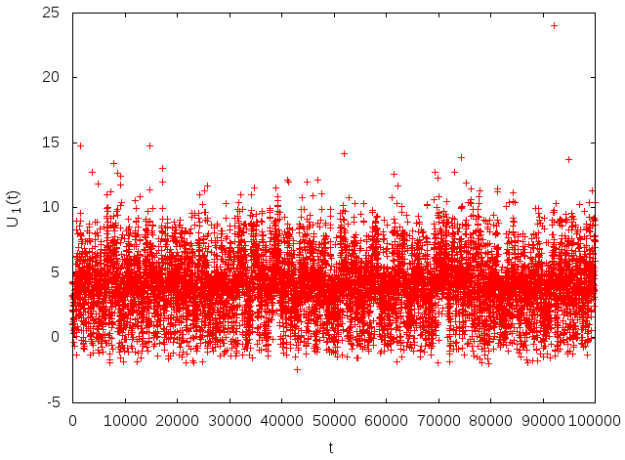



Figure 1: Measured values of $U_1$ at different times. Every 10th result is plotted.

The calculation of $C_i(\tau)$, the autocorrelation function, will be discussed in section 5. Here, it is assumed that the data points are independent which is equivalent to setting the correlation to zero. Thus the following formula is used:

$$\sigma_i = \sqrt{\frac{VarU_i}{N}} \tag{4}$$

The averages and errors obtained via this approach are listed in table 1.

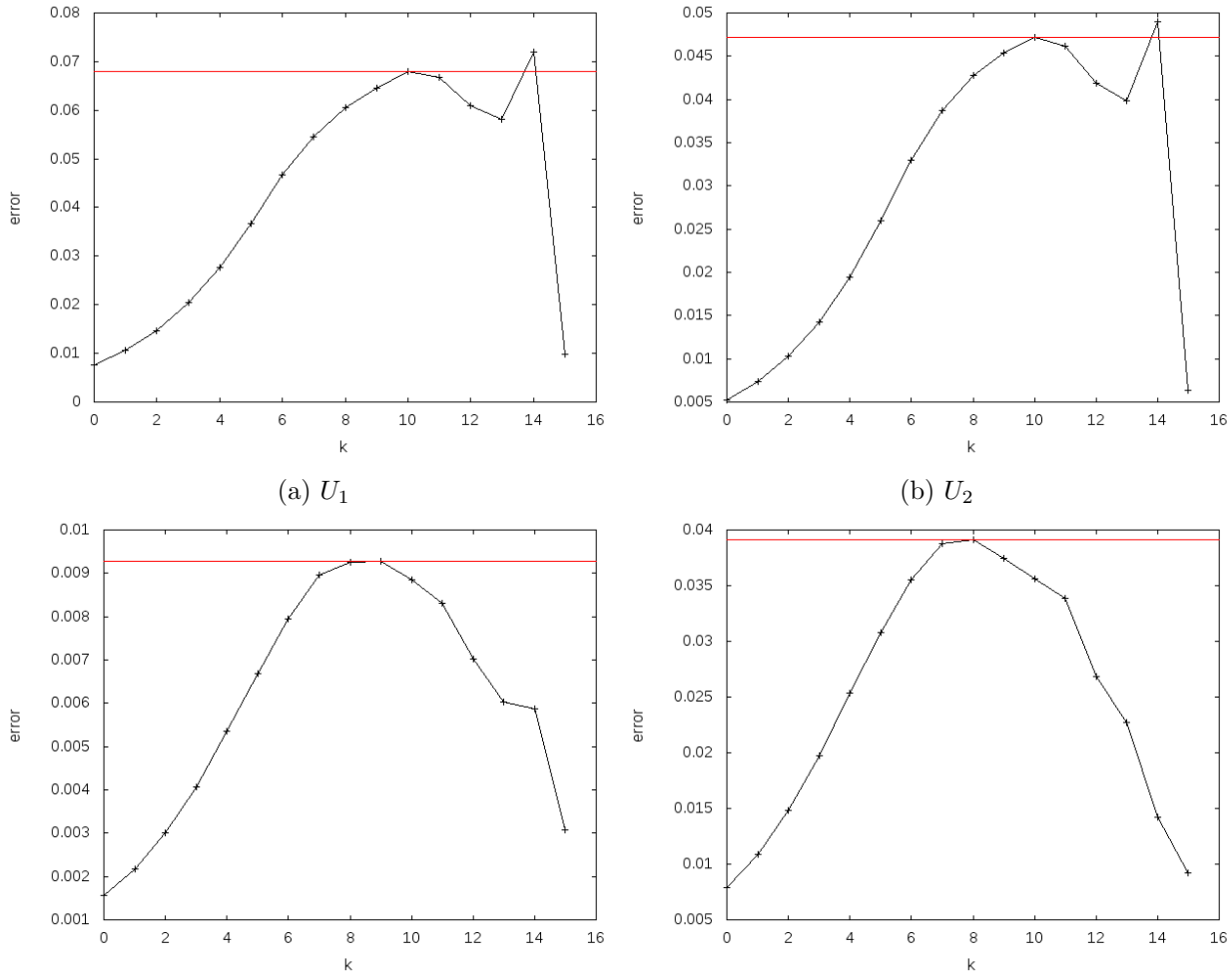
## 4 Blocking Analysis

However, usually, Monte Carlo techniques do not generate independent results in every step. One first approach in order to take into account correlation is the so-called blocking analysis. Here, pairwise blocking is performed, i.e. a new set of data is generated by averaging over neighboring points, with the result being used again as the input for the next blocking step.

$$U_i^{(1)}(t) = \frac{1}{2}\left[U_i(2t-1) + U_i(2t)\right] \tag{5}$$

$$U_i^{(k)} = \frac{1}{2}\left[U_i^{(k-1)}(2t-1) + U_i^{(k-1)}(2t)\right] \tag{6}$$

When (what happens the first time for $k = 7$) a set contains an uneven number of blocks, the last one is neglected in future steps.
The errors obtained for $U_i^{(k)}$ depending on the number of blocking steps $k$ are shown in figure 2. It is expected that from a certain block size onwards the values associated to the different blocks (i.e. the averages over the underlying data points) can be considered independent, because they will only contain a negligible share of data points that are closely correlated to some included in the neighboring blocks. From that point onwards the error obtained via formula 4 should no longer depend on $k$, i.e. a plateau is expected to appear in each of the graphs presented in figure 2. However, it should also be considered that with increasing $k$ the number of blocks decreases exponentially. Thus for high $k$, there is only a small number of blocks left resulting in fluctuations due to statistical errors. This explains why for high $k$ the errors plotted in figure 2 decrease or fluctuate. Although it is not in all cases entirely clear that the plateau has been reached, the slopes of the curves still decrease reasonably before fluctuations take over. Thus, it seems legitimate to




(a) $U_1$ (b) $U_2$

(c) $U_3$ (d) $U_4$

Figure 2: $k$-dependence of the errors $\sigma_i$ during the blocking analysis. The red lines indicate the error that was determined as the optimal estimate through visual analysis.

| $U_i$ | $\overline{U_i}$ | $\sigma_{i,ind}$ | $\sigma_{i,block}$ |
|---|---|---|---|
| $U_1$ | 3.984 | 0.008 | 0.068 |
| $U_2$ | 6.406 | 0.005 | 0.047 |
| $U_3$ | 2.015 | 0.002 | 0.009 |
| $U_4$ | 1.509 | 0.008 | 0.039 |

Table 1: Error estimates obtained under the assumption of independence $\sigma_{i,ind}$ and via a blocking analysis $\sigma_{i,block}$. Additionally, the averages $\overline{U_i}$ are given.

assume that the real error is close to the one graphically determined here. It is obtained after 8 to 10 blocking steps. The error estimates received via the blocking procedure are given in table 1. To allow a clearer observation of a plateau and thus a more precise estimation of the error, more data points would have to be generated. The errors obtained through the blocking analysis are up to 9 times higher than those calculated under the assumption of independence, which is a sign that the data points are indeed correlated.

## 5 Autocorrelation Analysis

In the following, correlation will be considered explicitly for the computation of the error of the sample mean via equation 3 by computing the autocorrelation functions and the integrated autocorrelation time $\tau_{int}$.
The correlation between measurements at $n$ and $m$ is defined as follows:

$$C(n,m) = \langle (U_i(n) - \langle U_i \rangle)\,(U_i(m) - \langle U_i \rangle) \rangle \tag{7}$$

In equilibrium $C$ no longer depends on the values of $n$ and $m$ themselves but only on their difference $k$. It can be shown that the $k$-dependent autocorrelation function can be computed as follows:

$$C(k) = \frac{1}{N-k} \sum_{j=1}^{N-k} (U_i(j+k) - \overline{U_i})(U_i(j) - \overline{U_i}) \tag{8}$$

With $k = 0$ equation 8 obviously becomes equivalent to the estimator for the variance as given in formula 2.
From the autocorrelation functions the integrated autocorrelation time $\tau_{int}$ can be computed:

$$\tau_{int} = \frac{1}{2} + \sum_{k=1}^{\infty} \frac{C(k)}{C(0)} \tag{9}$$

Entering this in equation 3 yields the following equation for estimating the error:

$$\sigma_i = \sqrt{\frac{C(0)}{N} 2\tau_{int}} \tag{10}$$

However, for an actual estimation of $\sigma_i$ the sum in equation 9 has to be cut. There is not only just a finite number of data points available, but also the results obtained for the autocorrelation functions through equation 8 become unreliable due to increasing statistical errors for high $k$. Thus, $\tau_{int}(k_{max})$ is defined and calculated for increasing values of $k_{max}$.

$$\tau_{int}(k_{max}) = \frac{1}{2} + \sum_{k=1}^{k_{max}} \frac{C(k)}{C(0)} \tag{11}$$

| $U_i$ | $\sigma_{i,block}$ | $\tau_{int}(k_{max})$ | $\sigma_{i,auto}$ | deviation / % |
|---|---|---|---|---|
| $U_1$ | 0.068 | 38.2 | 0.066 | -3.2 |
| $U_2$ | 0.047 | 39.6 | 0.047 | -1.5 |
| $U_3$ | 0.009 | 20.3 | 0.010 | 7.5 |
| $U_4$ | 0.039 | 13.5 | 0.041 | 5.4 |

Table 2: Error estimates obtained via a blocking analysis $\sigma_{i,block}$ and estimates for the autocorrelation times $\tau_{int}(k_{max})$ as well as the corresponding errors, $\sigma_{i,auto}$ . Additionally, the deviations of $\sigma_{i,auto}$ from $\sigma_{i,block}$ are given.

The computed autocorrelation functions $C(k)$ and corresponding integrated autocorrelation times $\tau_{int}(k_{max})$ for the given data are shown in figure 3.
The maximum integrated autocorrelation time before fluctuations become dominant was determined via visual investigation. The errors calculated via equation 10 are given and compared to those obtained via the blocking analysis in table 2.
The estimates are not exactly equal, which is due to the difficulties in finding a correct estimate before correlations take over and the graphical determination of the best estimate that was not entirely clear, especially for the autocorrelation analysis. However, the deviation is not higher than 8 % in any case, which is acceptable because for the error in general no high accuracy is needed.

## 6 Jackknife Analysis

Finally, a Jackknife analysis is carried out on the ratios $R_i$, with

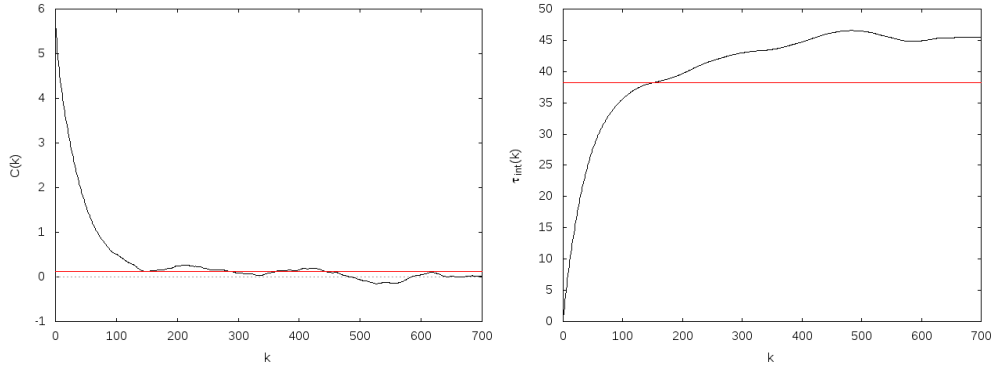
$$R_i = \frac{\langle U_i \rangle}{\langle U_1 \rangle} \tag{12}$$

and $i = 2, 3, 4$. The analysis starts from blocked variables, which each represent the average over 2000 data points. It is legitimate to neglect autocorrelation, because in section 4 it was shown that from 8 to 10 blocking steps onwards , what corresponds to blocks with 256 to 1024 elements, the blocks can be treated as sufficiently independent. Another criteria for a acceptable block size states that it should be greater or equal than 10 times $\tau_{int}$. According to the results obtained in section 5 this criteria is fulfilled, too.
The Jackknife method has the advantage that it is easily applicable on functions of mean values, and not only the averages themselves, while taking into account possible correlations between the different observables that were measured, here corresponding to nominator and denominator. The Jackknife average $\overline{U_{i,\alpha}^{JK}}$ is defined as the average taken over all blocks except block $\alpha$.

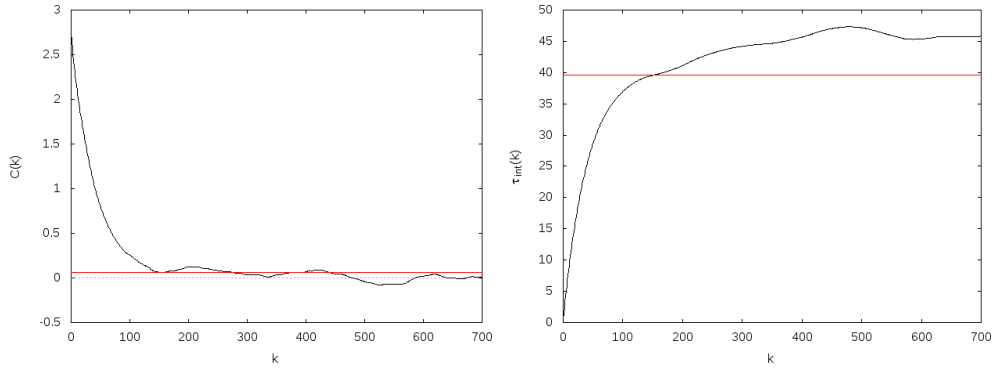
$$\overline{U_{i,\alpha}^{JK}} = \frac{1}{M-1} \sum_{\beta \neq \alpha} U_{i,\beta} \tag{13}$$

$M$ is the number of blocks and $U_{i,\alpha}$ is the average of $U_i$ over block $\alpha$. Now one can average over all ratios obtained by always not including another block i.e. by averaging over the different choices for $\alpha$:

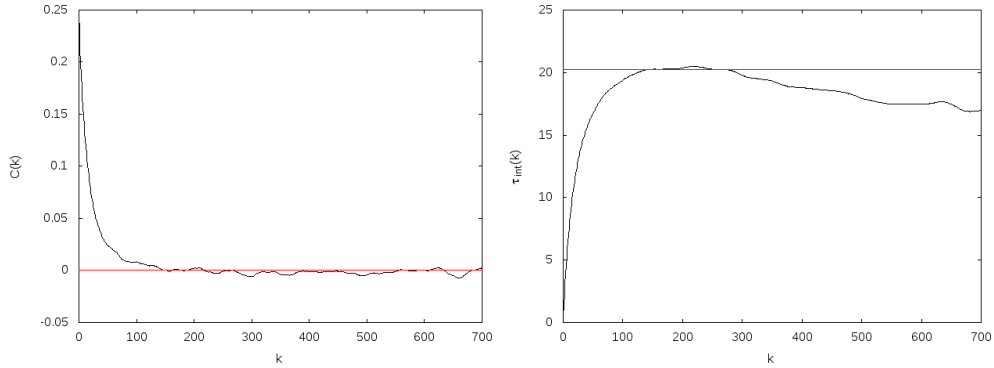
$$R_i^{JK} = \frac{1}{M} \sum_{\alpha=1}^{M} \frac{\overline{U_{i,\alpha}^{JK}}}{\overline{U_{1,\alpha}^{JK}}} \tag{14}$$




(a) $C(k)$ for $U_1$ (b) $\tau_{int}(k)$ for $U_1$

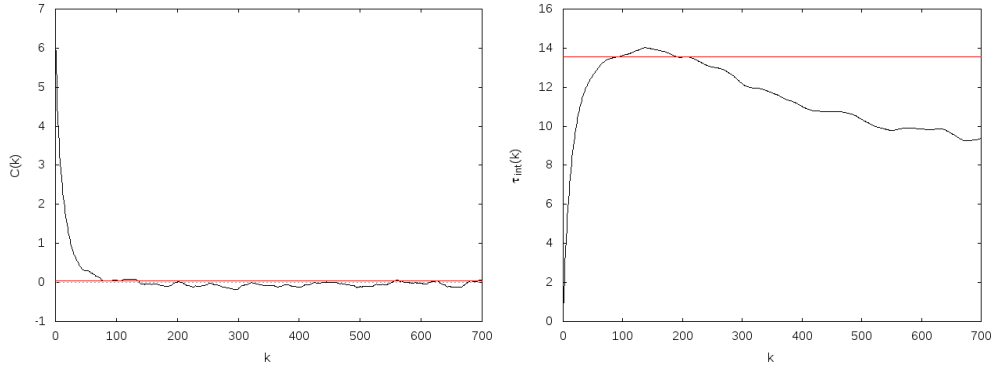



(c) $C(k)$ for $U_2$ (d) $\tau_{int}(k)$ for $U_2$




(e) $C(k)$ for $U_3$ (f) $\tau_{int}(k)$ for $U_3$

$U_3$




(g) $C(k)$ for $U_4$ (h) $\tau_{int}(k)$ for $U_4$

Figure 3: $k$-dependence of $C(k)$ and $\tau_{int}(k)$. The red lines indicate the values used for calculating the estimated error.

| $i$ | $R_{i,est}$ | $\sigma_i^{JK}$ | $\sigma_{i,ind}$ | $\sigma_{i,WE}$ |
|---|---|---|---|---|
| 2 | 1.607 | 0.039 | 0.029 | 0.038 |
| 3 | 0.506 | 0.010 | 0.009 | 0.011 |
| 4 | 0.378 | 0.013 | 0.012 | 0.017 |

Table 3: Overview of the estimated ratios $R_{i,est}$, the errors obtained via a Jackknife analysis on blocked data $\sigma_i^{JK}$, the estimates received via the independent-error formula, $\sigma_{i,ind}$, and the errors computed with the worst-error formula, $\sigma_{i,WE}$.

Additionally, the estimator obtained by simply dividing the estimator for the average of $U_i$ over the estimated average for $U_1$ is calculated:

$$R_i^{tot} = \frac{\overline{U_i}}{\overline{U_1}} \tag{15}$$

By using these definitions the Jackknife estimate for the ratio can be obtained:

$$R_{i,est} = MR_i^{tot} - (M-1)R_i^{JK} \tag{16}$$

The error of the Jackknife estimate $\sigma_i^{JK}$ is obtained via the following relation:

$$\sigma_i^{JK} = \sqrt{(M-1)VarR_{i,est}} \tag{17}$$

$$= \sqrt{\frac{(M-1)}{M}\sum_{\alpha=1}^{M}\left(\frac{\overline{U_{i,\alpha}^{JK}}}{\overline{U_{1,\alpha}^{JK}}} - R_{i,est}\right)^2} \tag{18}$$

In addition to the Jackknife formalism also two other, very common techniques for estimating the errors of functions of averages were applied: The independent-error formula and the worst-error formular: Under the assumption that $U_1$ and $U_i$ are independent the error $\sigma_{i,ind}$ for the ratio is obtained via equation 19. Equation 20 is the worst-error approximation for the considered ratios.

$$\sigma_{i,ind} = \left|\frac{\overline{U_i}}{\overline{U_1}}\right| \left|\left(\frac{\sigma_{U_i}^2}{\overline{U_i}^2} + \frac{\sigma_{U_1}^2}{\overline{U_1}^2}\right)\right|^{\frac{1}{2}} \tag{19}$$

$$\sigma_{i,WE} = \left|\frac{\overline{U_i}}{\overline{U_1}}\left(\frac{\sigma_{U_i}}{|\overline{U_i}|} + \frac{\sigma_{U_1}}{|\overline{U_1}|}\right)\right| \tag{20}$$

As input data for the calculation of $\sigma_{i,ind}$ and $\sigma_{i,WE}$ the estimates obtained from the autocorrelation analysis were used. All results are summarized in table 3.
The Jackknife method is expected to yield the best estimate of the error. The worst-error formula gives a higher limit to the error whereas the result of the independent-error equation can yield lower or higher estimates. Here it always gives a lower result, which, while being close to the result of the Jackknife method for $R_3$ and $R_4$, differs significantly from the Jackknife result for $R_2$. On the first sight it seems counterintuitive that the worst-error formula gives a (although only slightly) lower result for the error in the case of $R_2$ than the Jackknife method. However, one should consider that the result of the worst-error formula again depends on the estimates made during the autocorrelation analysis. The freedom in the choice of $\tau(k_{max})$ is easily large enough to explain this unexpected result.

# 7 Conclusion

First, four sets of data, as they would arise from a Monte Carlo or Molecular Dynamics simulation, were analyzed separately. The errors computed in a blocking and in an autocorrelation analysis deviated from each other, but only slightly. Both methods clearly showed that the individual data points cannot be considered independent. The errors computed using the assumption of independence are by far too low.
Then a Jackknife analysis was performed and the results compared with those obtained by applying the independent-error and worst-error equation on the results of the autocorrelation analysis. Although the results are mostly similar, especially the error based on the assumption of independence differs significantly for one data set. Finally, one should not oversee that the two formulas depend on additional estimations since they require the errors of the single data sets as input which might explain some unexpected results.