

Computer simulation of physical systems I

Task VI: Blocking analysis for error estimation on time averages

Molecular dynamics (MD) and Monte-Carlo (MC) computer simulations of physical systems generate raw data in the form of a *finite* “time” series of *correlated* data.¹ In the case of simulations of systems at thermal equilibrium, the first step of the data analysis consists in computing time averages. However, since these averages are performed over a finite time series of data, they are affected by a statistical error. The next step consists in estimating this error. Here we describe an elegant and computationally efficient method for evaluating the statistical error: the blocking analysis.

1 Evaluating time averages in simulations

At thermal equilibrium, the ensemble average of a physical quantity $A(\mathbf{x})$, where \mathbf{x} denotes a phase-space configuration, takes the form

$$I = \langle A \rangle = \frac{\int A(\mathbf{x})\omega(\mathbf{x}) d\mathbf{x}}{\int \omega(\mathbf{x}) d\mathbf{x}} \quad (1)$$

where $\omega(\mathbf{x})$ is the ensemble distribution function in phase space. The average $\langle A \rangle$ can be calculated through many independent simulations.

Assuming the system is ergodic, the ensemble average may be equated to a time average

$$\langle A \rangle \equiv \bar{A} = \lim_{T \rightarrow \infty} \frac{1}{T} \int A(\mathbf{x}(t)) dt. \quad (2)$$

Therefore, in practice, we perform a single long simulation and generate N time-correlated samples of the physical quantity we are interested in, A_1, A_2, \dots, A_N . Then, we compute the *finite* time average and use it as an estimate of the ensemble average [Eq. (1)]

$$I \approx \bar{A} = \frac{1}{N} \sum_{n=1}^N A_n. \quad (3)$$

At the same way one can evaluate the ensemble variance of the variable A as follows

$$\sigma_A^2 = \langle A^2 \rangle - \langle A \rangle^2 \approx \overline{A^2} - \bar{A}^2 = \frac{1}{N} \sum_{n=1}^N (A_n - \bar{A})^2. \quad (4)$$

Hereafter, we consider $\overline{A^2} - \bar{A}^2$ the actual value of σ_A^2 and we do not concern ourselves with its statistical error.

¹Hereafter we use the word time, even though, it does not necessarily denote a physical time.

2 Statistical error

Evaluating ensemble averages through Eq. (3) introduces a statistical error. This can be expressed through the variance:

$$\sigma_I^2 = \langle I^2 \rangle - \langle I \rangle^2, \quad (5)$$

where $\langle \dots \rangle$ indicates the ensemble average over many independent simulations. Inserting Eq. (3) into Eq. (5), we find

$$\sigma_I^2 = \frac{1}{N^2} \sum_{n,m}^N \langle A_n A_m \rangle - \langle A_n \rangle \langle A_m \rangle = \frac{\sigma_A^2}{N^2} \sum_{n,m}^N c_{AA}(n, m), \quad (6)$$

where we introduced the *correlation function*

$$c_{AA}(n, m) \equiv \frac{\langle [A_n - \langle A_n \rangle][A_m - \langle A_m \rangle] \rangle}{\langle A_n^2 \rangle - \langle A_n \rangle^2} = \frac{\langle A_n A_m \rangle - \langle A_n \rangle \langle A_m \rangle}{\sigma_A^2}. \quad (7)$$

For systems at thermal equilibrium, averages are invariant under time translations and $c_{AA}(n, m)$ depends only on time differences

$$c_{AA}(n, m) \equiv c_{AA}(k) \quad \text{with} \quad k = |n - m|. \quad (8)$$

Note that $c_{AA}(k = 0) = 1$ and that in the limit of large k $c_{AA}(k) \rightarrow 0$ because the data at large distances in the series eventually become independent, i.e. $\langle A_n A_m \rangle = \langle A_n \rangle \langle A_m \rangle$.

In the particular case of independent data, A_1, A_2, \dots, A_N , we have

$$\langle A_n A_m \rangle = \langle A_n \rangle \langle A_m \rangle = \langle A_n \rangle^2 \quad \text{for} \quad n \neq m, \quad (9)$$

$$\langle A_n A_m \rangle = \langle A_n^2 \rangle \quad \text{for} \quad n = m. \quad (10)$$

The correlation function then reduces to a delta function:

$$c_{AA}(n, m) = \frac{\langle A_n A_m \rangle - \langle A_n \rangle \langle A_m \rangle}{\sigma_A^2} = \delta_{nm}. \quad (11)$$

In this case, the statistical error reads

$$\sigma_I^2 = \frac{\sigma_A^2}{N}. \quad (12)$$

It is convenient to introduce the definition of *correlation time*

$$\tau = \frac{1}{2} \sum_{k=-\infty}^{\infty} c_{AA}(k) \quad (13)$$

Property 2.1. For an exponential correlation function, $c_{AA}(k) = \exp(k/\tilde{\tau})$, the correlation time τ is equal to $\tilde{\tau}$ (for $\tilde{\tau} \gg 1$).

Proof. From the definition of integrated correlation time we have

$$\tau = \frac{1}{2} \sum_{k=-\infty}^{+\infty} e^{-|k|/\tilde{\tau}} = \frac{1}{2} \left[\sum_{k=0}^{+\infty} e^{-|k|/\tilde{\tau}} + \sum_{k=0}^{-\infty} e^{-|k|/\tilde{\tau}} - 1 \right] = \frac{1}{1 - e^{1/\tilde{\tau}}} - \frac{1}{2} \underset{\tilde{\tau} \gg 1}{=} \tilde{\tau} \quad (14)$$

where we have used the formula for a geometric series

$$\sum_{n=0}^{+\infty} q^n = \frac{1}{1-q} \quad \text{for } q < 1. \quad (15)$$

□

Let's rewrite Eq. (6) keeping in mind the time translation invariance

$$\sigma_I^2 = \frac{\sigma_A^2}{N^2} \sum_{n,m}^N c_{AA}(n-m). \quad (16)$$

We see that the double sum only depends on $l = n - m$, where $l_{\min} < l < l_{\max}$.

$n \backslash m$	1	2	3	4	...	N
1	0	1	2	3	...	$N-1$
2	-1	0	1	2	...	\vdots
3	-2	-1	0	1	...	\vdots
4	-3	-2	-1	0	...	\vdots
\vdots	\vdots	\vdots	\vdots	\vdots	\ddots	\vdots
N	$1-N$	0

$l = n - m$	frequency
0	N
1	$N-1$
-1	$N-1$
2	$N-2$
-2	$N-2$
\dots	\dots
l	$N - l $

As shown in the graphical example above, $l_{\min} = 1 - N = -(N-1)$ for $n = 1, m = N$ and $l_{\max} = N - 1$ for $n = N, m = 1$. In Eq. (16), terms with index $l = n - m$ appear with a frequency of $(N - |l|)$. Thus, we can rewrite

$$\sum_{n=1}^N \sum_{m=1}^N c_{AA}(n-m) \rightarrow \sum_{l=l_{\min}}^{l_{\max}} c_{AA}(l)(N - |l|) \quad (17)$$

Eventually, these considerations lead to the expression

$$\begin{aligned} \sigma_I^2 &= \frac{\sigma_A^2}{N^2} \sum_{l=l_{\min}}^{l_{\max}} c_{AA}(l)(N - |l|) = \frac{\sigma_A^2}{N} \sum_{l=l_{\min}}^{l_{\max}} c_{AA}(l)(1 - \frac{|l|}{N}) \\ &\stackrel{N \rightarrow \infty}{=} \frac{\sigma_A^2}{N} \sum_{l=-\infty}^{+\infty} c_{AA}(l) = \frac{\sigma_A^2}{N} 2\tau \end{aligned} \quad (18)$$

where we have made the assumption of large N and used the definition of correlation time [Eq. (13)].

In practice, for the evaluation of the statistical error, we need to calculate the variance of the physical quantity of interest, σ_A^2 , and the correlation time. To sum up, for long enough simulations

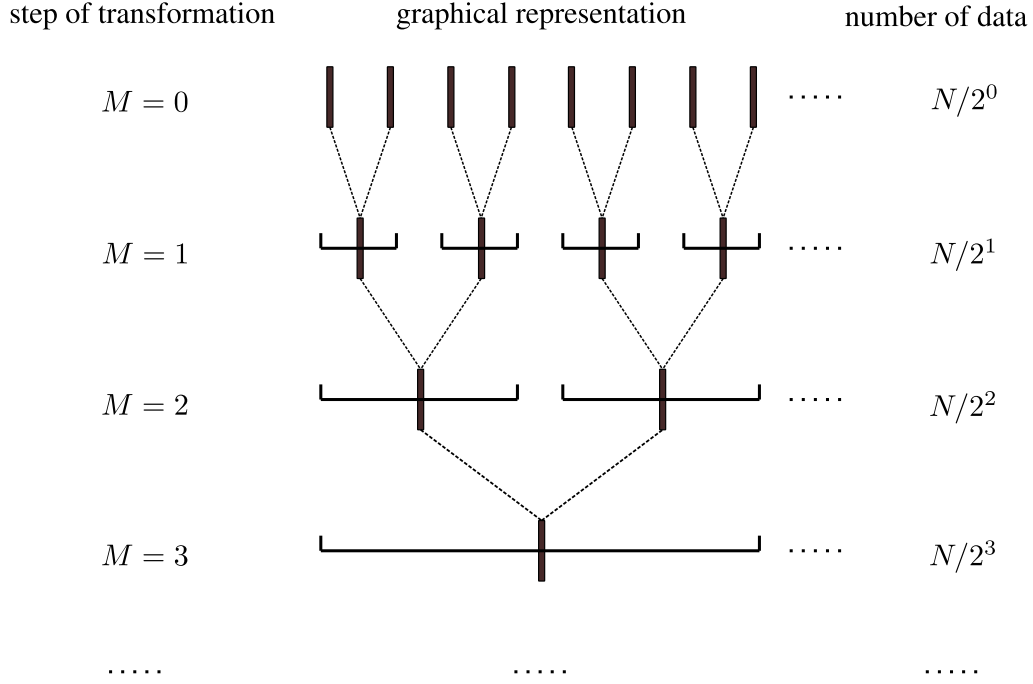


Figure 1: Schematic representation of the blocking procedure. Starting from a N -sample series, new series of data are generated through the transformation (22).

these quantities can be obtained as follows

$$\sigma_A^2 \approx \overline{A^2} - \overline{A}^2, \quad (19)$$

$$2\tau \approx \sum_{n=-\infty}^{+\infty} \overline{c_{AA}(n)}, \quad (20)$$

$$\sigma_I^2 \approx \frac{\sigma_A^2}{N} 2\tau. \quad (21)$$

This method for estimating the statistical error is based on the calculation of the time correlation function and is a demanding computational task. A simpler and more elegant technique is discussed in the following.

3 Blocking analysis

In this section, we describe a way of evaluating the statistical error σ_I^2 which is computationally more efficient than the previous procedure. In fact, it elegantly avoids the expensive computation of time correlation functions and allows for an estimation of the correlation time τ and the statistical error σ_I^2 by studying the behavior of the “block” averages.

We transform the initial series of N data, A_1, A_2, \dots, A_N , into half as large a data set, $A_{1,1}, A_{1,2}, \dots, A_{1,N_1}$, through the following transformation:

$$A_{1,i} = \frac{1}{2}(A_{2i-1} + A_{2i}) \quad (22)$$

$$N_1 = \frac{1}{2}N \quad (23)$$

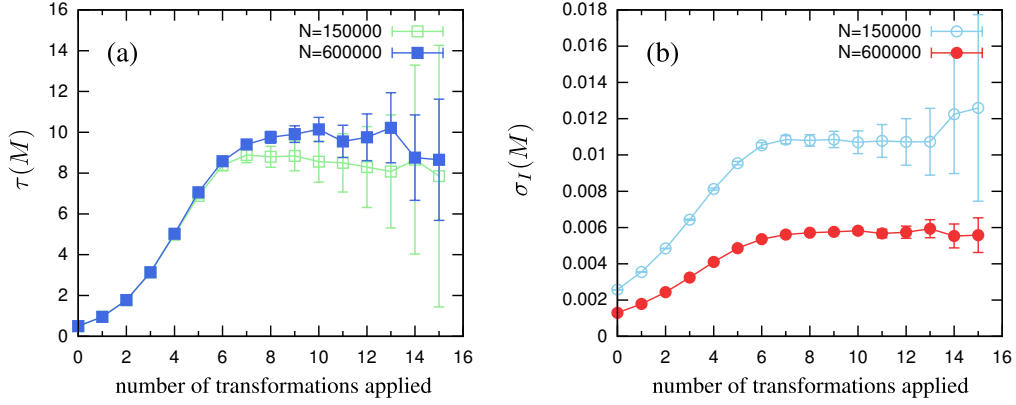


Figure 2: (a) Correlation time, calculated by means of Eq. (26), and (b) statistical error on time averages, Eq. (28), of exponentially correlated Gaussian random numbers as a function of the number of block transformations applied. These quantities are reported for a sequence of 150'000 and of 600'000 data, respectively.

It is immediately verified that the transformation preserves the average, $I_1 = I$. The transformation is recursively applied to the new data as schematically represented in Fig. 1. At the M -th step of transformation, there are $N(M) = N/2^M$ number of transformed samples $\{A_{M,i}\}$, each of them resulting from the average of a block of 2^M samples of the initial data set. At each step we evaluate the variance associated to the new data set:

$$\sigma_A^2(M) = \frac{1}{N(M)} \sum_{i=1}^{N(M)} (A_{M,i} - \langle A \rangle)^2. \quad (24)$$

In particular, note that $\sigma_A^2(0) = \sigma_A^2$.

We define $\tau(M)$ through the following relation:

$$\sigma_A^2(M) = \frac{\sigma_A^2}{2^M} 2\tau(M) \quad (25)$$

When $2^M \gg \tau$ each block is sufficiently large to be considered as an independent simulation, and $\tau(M)$ corresponds to the correlation time τ by virtue of Eq. (18) applied to independent simulations of duration 2^M . In particular, for $M = 0$ one finds $\tau(M) = \frac{1}{2}$. In practice, one evaluates $\tau(M)$ vs M at fixed N :

$$\tau(M) = \frac{\sigma_A^2(M) 2^M}{\sigma_A^2} \quad (26)$$

and determines the value of τ from the plateau reached at large M [cf. Fig. 2(a)]. The statistical error σ_I^2 can then be evaluated through Eq. (18) considering the full data set.

Equivalently, $\sigma_I^2(M)$ can be expressed as a function of M through Eq. (18):

$$\sigma_I^2(M) = \frac{\sigma_A^2}{N} 2\tau(M) = \frac{2^M \sigma_A^2(M)}{N} \quad (27)$$

In this way, the statistical error is obtained from the plateau value of $\sigma_I^2(M)$ [cf. Fig. 2(b)]. When the plateau has been reached, Eq. (27) reads:

$$\sigma_I^2 = \frac{\sigma_A^2(M)}{N(M)}, \quad (28)$$

which corresponds to the variance of a set of independent data. This implies that the transformed data $\{A_{M,i}\}$ have become uncorrelated when the plateau is reached. In practice, one could thus continue the blocking transformation until $\sigma_I^2(M)$ reaches a plateau value, which then corresponds to the desired statistical error σ_I^2 .

Additionally, we note that the value $\sigma_I(M = 0)$ corresponds to the error calculated for the original data $\{A_i\}$ as if they were independent. As M becomes larger, $\sigma_I(M)$ increases indicating the occurrence of correlations and eventually reaches the plateau value. By virtue of Eq. (18), the ratio between the plateau value and $\sigma_I(M = 0)$ is related to the correlation time:

$$\frac{\sigma_I(M^{\text{plateau}})}{\sigma_I(M = 0)} = \sqrt{2\tau}. \quad (29)$$

In case the plateau is not clearly discernable, the statistical error cannot clearly be determined because the duration of the simulation N is not sufficiently large with respect to the correlation time τ . In other words, the acquisition of data should be continued. As illustrated in Fig. 2(b), this has two effects. First, the plateau becomes better distinguishable. Second the overall statistical error is reduced by the scaling of σ_I with $N^{-1/2}$ [Eq. (18)].

4 Errors in the determination of τ and σ_I

To facilitate the recognition of the plateau regime, it is convenient to determine estimates for the errors of $\tau(M)$ and $\sigma_I(M)$. As can be seen from Eqs. (26) and (27), both quantities depend on the error by which we determine $\sigma_A^2(M)$ [Eq. (24)]. We estimate the error on $\sigma_A^2(M)$ by assuming that $(A_{M,i} - \langle A \rangle)^2$ are independent random variates as this occurs for sufficiently large M . Since $\sigma_A^2(M)$ corresponds to an average of random variates, the targeted variance (Var) reads:

$$\text{Var}\{\sigma_A^2(M)\} = \frac{1}{N(M)} \text{Var}\{(A_{M,i} - \langle A \rangle)^2\}. \quad (30)$$

We thus need to evaluate

$$\text{Var}\{(A_{M,i} - \langle A \rangle)^2\} = \langle (A_{M,i} - \langle A \rangle)^4 \rangle - \langle (A_{M,i} - \langle A \rangle)^2 \rangle^2 = \langle (A_{M,i} - \langle A \rangle)^4 \rangle - \sigma_A^4(M). \quad (31)$$

Invoking the central limit theorem, the random variates $(A_{M,i} - \langle A \rangle)^2$ are distributed according to a normal distribution. Hence, we find:

$$\langle (A_{M,i} - \langle A \rangle)^4 \rangle = 3\langle (A_{M,i} - \langle A \rangle)^2 \rangle^2 = 3\sigma_A^4(M), \quad (32)$$

and thus

$$\text{Var}\{(A_{M,i} - \langle A \rangle)^2\} = 2\sigma_A^4(M), \quad (33)$$

and

$$\text{Var}\{\sigma_A^2(M)\} = \frac{2}{N(M)} \sigma_A^4(M). \quad (34)$$

We thus determine $\sigma_A^2(M)$ with the following standard deviation (SD):

$$\text{SD}\{\sigma_A^2(M)\} = \sqrt{\frac{2}{N(M)}} \sigma_A^2(M). \quad (35)$$

By consequence, this gives:

$$\text{SD}\{\tau(M)\} = \sqrt{\frac{2}{N(M)}} \tau(M), \quad (36)$$

$$\text{SD}\{\sigma_I^2(M)\} = \sqrt{\frac{2}{N(M)}} \sigma_I^2(M). \quad (37)$$

From the expansion of the square root to first order, one then derives:

$$\text{SD}\{\sigma_I(M)\} = \frac{1}{\sqrt{2N(M)}} \sigma_I(M). \quad (38)$$

We note that as M becomes larger, the number $N(M)$ of available statistical data becomes smaller. Eventually, this leads to the large errors observed in Fig. 2 at large M and delimites the length of the plateau.

5 Generation of exponentially correlated Gaussian deviates

We generate a sequence of Gaussian random numbers, A_1, A_2, \dots, A_N , in a controlled manner in order to introduce an exponential correlation with a given correlation time.

Let be G_n a sequence of independent Gaussian numbers with zero mean and unit variance, their probability distribution function reads

$$\omega(G_n = x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \quad (39)$$

Now, we define a new sequence of random numbers generated recursively via

$$\begin{aligned} A_0 &= G_0 \\ A_{n+1} &= f A_n + \sqrt{1 - f^2} G_{n+1} \\ f &= e^{-1/\tau} \end{aligned} \quad (40)$$

or in a compact form

$$A_n = f^n G_0 + \sqrt{1 - f^2} \sum_{l=1}^n G_l f^{n-l}. \quad (41)$$

Property 5.1. *The random numbers $\{A_n\}$ are Gaussian deviates with zero mean and unit variance*

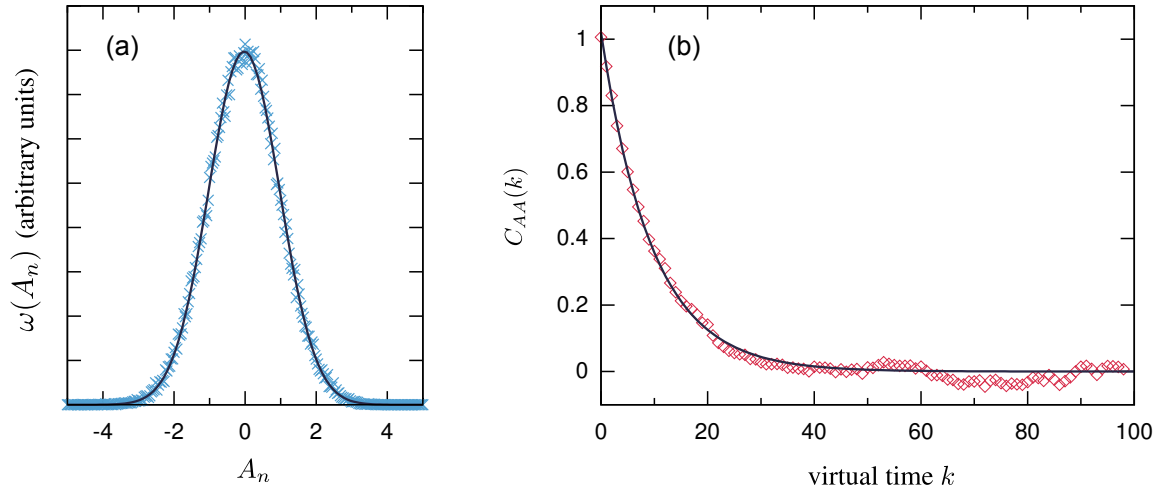


Figure 3: Random numbers generated with the algorithm of Eq. (40) are (a) Gaussian distributed and (b) show an exponential correlation.

Proof. A_n is a linear combination of independent random variables Gaussian distributed, $\{G_n\}$, therefore, also A_n is a Gaussian deviate. Indeed, we have started with $A_0 = G_0$, which belongs to a Gaussian distribution with zero mean and unit variance and by induction we have

$$\begin{aligned}\langle A_{n+1} \rangle &= f \underbrace{\langle A_n \rangle}_{=0} + \sqrt{1-f^2} \underbrace{\langle G_{n+1} \rangle}_{=0} = 0 \\ \langle A_{n+1}^2 \rangle &= f^2 \underbrace{\langle A_n^2 \rangle}_{=1} + (1-f^2) \underbrace{\langle G_{n+1}^2 \rangle}_{=1} = 1\end{aligned}$$

For the calculation of the variance, the cross-term $\langle A_n G_{n+1} \rangle$ vanishes because A_n and G_{n+1} are independent and thus uncorrelated. \square

Property 5.2. The correlation function of the stochastic variable A depends only on the difference of its arguments, as already seen in Eq. (8), (this correspond to a stationary stochastic process) and is given by

$$c_{AA}(k) = f^k = e^{-k/\bar{\tau}}$$

Proof. Since $\{A_i\}$ have zero mean and unit variance, its correlation function reduces to

$$c_{AA}(k) = \langle A_n A_{n+k} \rangle$$

thus,

$$\begin{aligned}\langle A_n A_{n+k} \rangle &= \left\langle A_n \left(f^{n+k} G_0 + \sqrt{1-f^2} \sum_{l=1}^{n+k} G_l f^{n+k-l} \right) \right\rangle \\ &= \left\langle A_n \left(f^k f^n G_0 + f^k \sqrt{1-f^2} \sum_{l=1}^n G_l f^{n-l} + \sqrt{1-f^2} \sum_{l=n+1}^{n+k} G_l f^{n+k-l} \right) \right\rangle\end{aligned}$$

$$\begin{aligned}
&= \left\langle A_n \left(f^k A_n + \sqrt{1-f^2} \sum_{l=n+1}^{n+k} G_l f^{n+k-l} \right) \right\rangle \\
&= f^k \langle A_n^2 \rangle + \underbrace{\sqrt{1-f^2} \langle A_n \rangle}_{=0} \underbrace{\left\langle \sum_{l=n+1}^{n+k} G_l f^{n+k-l} \right\rangle}_{=0} \\
&= f^k \langle A_n^2 \rangle = e^{-k/\tilde{\tau}}
\end{aligned}$$

We have used the fact that the Gaussian random numbers G_l are not correlated with A_n because $l > n$. \square

We show in Fig. 3 the numerical confirmation of properties 5.1 and 5.2.

6 Exercise

1. Generate N samples, $\{A_i\}$, with correlation time $\tilde{\tau}$ using the algorithm described in Sec. 5.
2. Verify numerically that $\{A_i\}$ are Gaussian deviates with

$$\langle A \rangle = 0 \quad \text{and} \quad \sigma_A^2 = 1.0.$$

3. Calculate $c_{AA}(k)$, fit the data with $e^{-|k|/\tau}$ and verify that $\tau = \tilde{\tau}$.
4. Integrate $c_{AA}(k)$, Eq. (13), and verify that the time obtained is equal to $\tilde{\tau}$ (property 2.1).
5. Evaluate the statistical error σ_I using Eq. (18).
6. Use the blocking method and plot $\sigma_I(M)$ as a function of the block transformation step. Evaluate the statistical error with the value of this function at its plateau, $\sigma_I^{\text{plateau}}$. Eventually, calculate the ratio

$$\frac{\sigma_I^{\text{plateau}}}{\sigma_I(0)}$$

and verify that is equal to $\sqrt{2\tau}$.

7. Use the blocking method and plot $\tau(M)$, Eq. (26), and evaluate the correlation time with the value of this function at its plateau, τ^{plateau} .
8. At fixed τ , generate different data sequences increasing the number of samples, N . Determine the minimum number of samples that you need for an accurate evaluation of the correlation time and the statistical error. What is the behavior of the statistical error as a function of N ?
9. Calculate σ_I with the blocking analysis for many data set with different correlation time (be sure that $\tau \ll 2^{M_{\max}}$ with $M_{\max} = \log_2 N$). Plot $\sigma_I(\tau)$ as a function of the correlation time and show that

$$\sigma_I(\tau) = s \sqrt{\frac{2\tau}{N}} \quad \text{where} \quad s = \sigma_A.$$

Codes

The following codes are provided:

stochastic.x generates a sequence of N samples, $\{A_n\}$, Gaussian distributed and exponentially correlated

histogram.x calculates the probability distribution function

correlation.x calculates the time correlation function

blocking.x performs the blocking analysis

References

- [1] Daan Frenkel and Berend Smit *Understanding Molecular Simulation: from algorithms to applications*. Computational Science Series, Vol. 1, 2nd Edition
- [2] Jos Thijssen *Computational Physics*. Cambridge University Press 2nd Edition
- [3] H. Flyvbjerg and H. G. Petersen *Error estimates on averages of correlated data*. Journal of Chemical Physics, Volume 91, pag. 461 (1989)
- [4] *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press <http://www.nrbook.com/a/bookcpdf.php>