M-LAB
computational physics_

# Toolbox: Binning analysis

The binning analysis is a tool to estimate statistical errors of means computed from sequential data $X_i$. If the data of the sequence were uncorrelated, the standard error of the mean would be a reliable error quantification. However, it is often the case that subsequent data are correlated, i.e. the data series exhibits a non-vanishing autocorrelation time. In this case, the standard error would underestimate the actual statistical error. The binning analysis introduced in Ref. [1] accounts for the presence of autocorrelation to give reliable error estimates also in that case; here, we give a compressed summary and recommend reading Ref. [1] for a very clear and detailed explanation of the matter.

The idea of the binning analysis for a sequence of data $\{X_i\}_{i=1}^{M}$ is to construct a series of coarse-grained sequences

$$X_l^{(k)} = \frac{1}{k} \sum_{i=lk}^{(l+1)k-1} X_i \, . \tag{1}$$

with $M_k = \lfloor M/k \rfloor$ elements. The mean of each of these sequences remains the same, but with increasing $n$, consecutive elements of the sequences become less correlated. Therefore, the corresponding standard error estimate of the mean

$$\Delta_X^{(k)} = \sqrt{\frac{1}{M_k(M_k-1)} \sum_{l=1}^{M_k} \left( X_l^{(k)} - \langle\langle X^{(k)} \rangle\rangle_{M_k} \right)^2} \tag{2}$$

increases and eventually converges to an *accurate error estimate*

$$\Delta_X = \lim_{k\to\infty} \Delta_X^{(k)} \, . \tag{3}$$

In Eq. (2) above we use the notation $\langle\langle X \rangle\rangle_M = \frac{1}{M} \sum_{i=1}^{M} X_i$ for the empirical mean.

From this sequence of error estimates one can extract the *autocorrelation time* of the original sequence as

$$\tau = \frac{1}{2} \left[ \left( \frac{\Delta_X}{\Delta_X^{(1)}} \right)^2 - 1 \right] . \tag{4}$$

You will deploy such a binning analysis in this lab course for the numerical data created in the Markov chain Monte Carlo or Molecular Dynamics simulations. Note, however, that such a binning analysis can be performed on any sequence (or time series) of data samples to detect its intrinsic autocorrelation effects independent of the way this data was produced – be it in numerical experiments or in real-life situations (think of any time series such as traffic measurements on the Zülpicher Strasse, orders of cappuccinos at your favorite coffee place, or the daily stock market data of some company).

# Implementation

Write a function that performs a binning analysis for a given sequence of observables $\mathbf{X} = (X_i)_{i=1,\dots,M}$. The following pseudo-code sketches the procedure assuming that statistical and reshaping functions are available (as it is the case in `julia` and `python/numpy`):

**Algorithm 1** Pseudocode for binning analysis

**function** BINNINGANALYSIS($\mathbf{X}$, $k_{\text{max}}$)
  $M$ = size($\mathbf{X}$)
  **for** $k$ in 1:$k_{\text{max}}$ **do**
    $M_k = \lfloor M/k \rfloor$
    $\mathbf{X}^{(k)} \leftarrow$mean(reshape($\mathbf{X}$[1:$kM_k$], (k,:)), axis=0)
                ▷ Compute coarse grained sequence
    error-est[k] $\leftarrow$ std($\mathbf{X}^{(k)}$)/$\sqrt{M_k}$  ▷ Error estimate of current sequence
  **return** error_est

# References

[1] V. Ambegaokar and M. Troyer. Estimating errors reliably in Monte Carlo simulations of the Ehrenfest model. *Am. J. Phys.*, 78:150–157, 2010.