Analytical expressions for the blocking error of exponentially correlated Gaussian data

Markus Deserno
Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
(Dated: January 31, 2019)

Blocking is a method for efficiently arriving at the error of the mean for time-correlated data. These brief notes derive an analytical expression for the estimated error as a function of blocking order, useful for instance to test implementations of blocking programs.

Assume we want to experimentally determine some quantity with good accuracy. Then we'll probably take a series of measurements and take the average. The outcome of each single measurement is a random variable, so if we measure N times, we have N random variables, call them $\{X_1, X_2, \ldots, X_N\}$. Let us further assume that these variables are *identically distributed* and have finite first and second moment:

$$\langle X_i \rangle = \mu \qquad , \qquad \langle X_i^2 \rangle = \mu^2 + \sigma^2 \,, \tag{1}$$

where $\langle \cdots \rangle$ denotes an *ensemble average* over the distribution underlying the random measurement process.

We would like to know μ accurately. Its standard estimator m is given by the average of the random variables:

$$m \;:=\; \frac{1}{N} \sum_{i=1}^{N} X_i \,. \tag{2}$$

Evidently, $\langle m \rangle = \mu$, so m is an *unbiased* estimator. But we also would like to know how accurately we were able to estimate μ. For this we need to know something about the variance of m, which is given by

$$\begin{aligned} \mathrm{Var}(m) \;&:=\; \langle m^2 \rangle - \langle m \rangle^2 & (3) \\ &=\; \frac{1}{N^2} \sum_{i,j=1}^{N} \Big[\langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle \Big] \,. & (4) \end{aligned}$$

If the single measurements are *uncorrelated*, then

$$\begin{aligned} \langle X_i X_j \rangle \;&\stackrel{\text{uncorr}}{=}\; \langle X_i^2 \rangle \, \delta_{ij} + \langle X_i \rangle^2 (1 - \delta_{ij}) \\ &=\; (\sigma^2 + \mu^2) \delta_{ij} + \mu^2 (1 - \delta_{ij}) \\ &=\; \sigma^2 \delta_{ij} + \mu^2 \,. \end{aligned} \tag{5}$$

and we obtain

$$\mathrm{Var}(m) \;\stackrel{\text{uncorr}}{=}\; \frac{\sigma^2}{N} \,, \tag{6}$$

showing that our estimator m improves if we make more measurements (provided σ is finite!).

However, generally our individual measurements are *not* uncorrelated. Let us thus assume that we have a nonvanishing *covariance* given by

$$C_{i,j} \;:=\; \mathrm{Cov}(X_i, X_j) \;:=\; \langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle \tag{7}$$

and a corresponding correlation coefficient

$$\gamma_{i,j} \;:=\; \frac{C_{i,j}}{\sigma^2} \,. \tag{8}$$

In most cases it is reasonable to assume that $C_{i,j}$ and $\gamma_{i,j}$ only depend on the (absolute) difference $|i-j|$ of the indices, namely, if the X_i correspond to *successive measurements* in a system that leaves some memory. In this case the autocorrelation functions C_t or γ_t, where $t = |i-j|$ is that difference, contain all relevant information.

What is now the variance of m? A simple calculation gives

$$\begin{aligned} \mathrm{Var}(m) \;&=\; \frac{1}{N^2} \sum_{i=1}^{N} C_{i,i} + \frac{2}{N^2} \sum_{i>j=1}^{N} C_{i,j} \\ &=\; \frac{\sigma^2}{N} + \frac{2}{N^2} \sum_{t=1}^{N-1} (N-t)\, C_t \\ &=\; \frac{\sigma^2}{N} \left\{ 1 + 2 \sum_{t=1}^{N-1} \Big(1 - \frac{t}{N}\Big) \gamma_t \right\} . \end{aligned} \tag{9}$$

In the second step we used the fact that in an $N \times N$ covariance matrix $C_{i,j}$ there are $2(N-t)$ entries which have a separation $t = |i-j|$, and in case of time translational symmetry they all have the *same* value C_t.

If the correlation function γ_t decays "sufficiently" rapidly, the expression $\sum_{t=1}^{N} t\, \gamma_t$ will quickly converge to some finite value. The last term in Eqn. (9) is thus of order $1/N$, and we can therefore write

$$\mathrm{Var}(m) \;=\; \frac{\sigma^2}{N} \left\{ 1 \,+\, 2\,T \,+\, \mathcal{O}\Big(\frac{1}{N}\Big) \right\} , \tag{10a}$$

where

$$T \;:=\; \sum_{t=1}^{N-1} \gamma_t \tag{10b}$$

is a measure of the correlation strength. We may view this as a "total correlation time" that is well defined *irrespective of the functional form of* γ_t.

A very common case is that the correlation function decays exponentially:

$$\gamma_t = \mathrm{e}^{-t/\tau} \equiv c^t \quad \text{with} \quad c = \mathrm{e}^{-1/\tau} \,, \tag{11}$$

where τ is the "conventional" correlation time. In this case, we can evaluate Eqn. (9) analytically, because the sums are

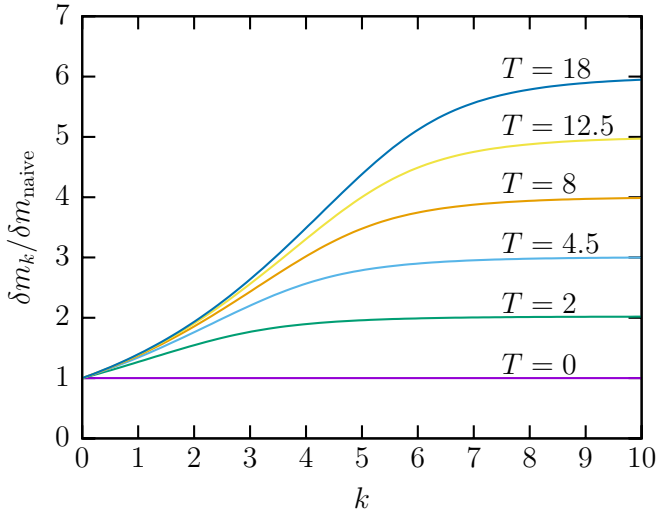

FIG. 1: Blocking error relative to its naive (correlation-free) estimate as a function of blocking order k for a sequence of exponentially correlated random numbers with a correlation time T. Six examples of correlation times are shown, as indicated in the figure. Notice that the nonzero correlation times are of the form $M^2/2$, leading to the (approximate) asymptote M.

merely variations of finite geometric series:

$$\sum_{t=1}^{N-1} c^t = \sum_{t=0}^{N-1} c^t \; - 1 \; = \; \frac{1-c^N}{1-c} - 1 \; = \; \frac{c-c^N}{1-c} \; , \tag{12a}$$

$$\begin{aligned} \sum_{t=1}^{N-1} t\, c^t = \sum_{t=1}^{N-1} \Big(c\frac{\partial}{\partial c}\Big) c^t \; &= \; \Big(c\frac{\partial}{\partial c}\Big) \sum_{t=1}^{N-1} c^t \\ &= \frac{c}{(1-c)^2}\Big[- Nc^{N-1} + Nc^N + 1 - c^N\Big] \; . \end{aligned} \tag{12b}$$

Plugging this into Eqn. (9), we get

$$\begin{aligned} \mathrm{Var}(m) &= \frac{\sigma^2}{N}\Bigg\{1 + 2\Bigg[\frac{c-c^N}{1-c} - \frac{1}{N}\,\frac{c}{(1-c)^2}\times \\ &\qquad \Big(- Nc^{N-1} + Nc^N + 1 - c^N\Big)\Bigg]\Bigg\} \\ &= \frac{\sigma^2}{N}\left\{\frac{1+c}{1-c} - \frac{2c}{N}\,\frac{1-c^N}{(1-c)^2}\right\} . \end{aligned} \tag{13}$$

For sufficiently large N, the second part in the curly parentheses vanishes, and the variance approaches

$$\lim_{N\to\infty} \frac{\mathrm{Var}(m)}{\sigma^2/N} = \frac{1+c}{1-c} = \coth\frac{1}{2\tau} = \begin{cases} 1 & : \; \tau \ll 1 \\ 2\tau & : \; \tau \gg 1 \end{cases} . \tag{14}$$

Hence, in the presence of a nonvanishing correlation time, the error of the mean is larger than the correlation free result, asymptotically by a factor of $\sqrt{2\tau}$.

It hence seems that in order to calculate the true error, we need to estimate the correlation function, or at least the correlation time τ. Unfortunately, though, getting an unbiased estimator of γ_t is tricky. This is why an alternative method, called *data blocking*, has become a popular workaround for this problem—see H. Flyvbjerg and H. G. Petersen, *Error estimates on averages of correlated data* J. Chem. Phys. **91**, 461 (1989). The general idea is that we can take our original data set and *pre-average* the data points into blocks of length b, which leaves N/b such blocks. It is easy to check that *if* the data are uncorrelated, then the mean and variance of the blocked data is the same as that of the unblocked data. Pre-averaging does clearly not change the mean, but it also does not change the error. However, if the data are correlated, than blocking *increases* the error of the mean, because it reduces the number of (blocked) data points from which to calculate the error of the mean, while not correspondingly reducing their individual variances (due to the correlations). However, once the blocks become sufficiently large, they effectively become uncorrelated, and any further blocking will no longer increase the error. This, therefore, gives insight into both the true error and the underlying correlation time.

Let us make an explicit example: If we have a sequence (X_i) of N exponentially correlated random variables (with a finite variance), we could decide to subdivide them into blocks of length b and pre-average those. An estimator for the mean of any such block is hence $m_b = (X_1 + X_2 + \cdots + X_b)/b$. From what we have just calculated, we see that its variance is given by Eqn. (13), where N is replaced by b. Averaging the N/b blocks, *and pretending they are independent*, then leads to the following estimator for the error of the mean, relying on blocks of length b:

$$\frac{\delta m(b)}{\delta m_{\rm naive}} \; = \; \sqrt{\frac{1+c}{1-c} - \frac{2c/b}{(1-c)^2}(1-c^b)} \; . \tag{15}$$

More specifically, blocking is typically done by taking the data and combining successive pairs of data points, resulting in a new sequence with half as many data points, *and then iterating this process*. If k enumerates the number of such pair-combination-steps, then at order k the block length is $b = 2^k$, and the blocking error at order k is given by

$$\frac{\delta m(k)}{\delta m_{\rm naive}} \; = \; \sqrt{\frac{1+c}{1-c} - \frac{2^{1-k}c}{(1-c)^2}(1-c^{2^k})} \; . \tag{16}$$

Fig. 1 illustrates what this expression looks like for a selection of correlation times. Notice it always starts at 1 and then monotonically increases to the true error from Eqn. (14), at which it saturates.

Observe that if we have more complicated time correlations, we could simply repeat the calculation that led to Eqn. (13). For instance, if the correlation function is a sum of exponentials with decay times τ_i and weights p_i, then the right hand side of Eqn. (13) simply gets replaced by a linear combination of such terms—one for each τ_i, having a weight p_i. Since essentially any correlation function could be Laplace-expanded in this way, this amounts to a complete and general solution of the problem. However, it might nevertheless be more convenient in specific cases to simply do the two sums over γ_t and $t\,\gamma_t$ explicitly for a given decay law.