Error analysis for `PAFI` code

Thomas D Swinburne
CNRS / CINaM, Aix-Marseille University, France
`swinburne@cinam.univ-mrs.fr`

October 2, 2019

The `PAFI` code distributes N_E workers, here indexed by $j \in [1, N_E]$, each producing a time series of N_S scalar projected forces, indexed by $i \in [1, N_S]$, giving in principle $N_E \times N_S$ data points $f_j(t_i)$. Details of the algorithm can be found in [1] and [2].

Each worker averages over the local time series, giving N_E time averaged projected forces

$$\langle f \rangle_j \equiv \sum_{i=1}^{N_S} \frac{f_j(t_i)}{N_S}, \tag{1}$$

where $\langle \ldots \rangle_j$ defines a time average on worker j. This data is then collated across all workers and averaged again to give the central observable

$$\langle\langle f \rangle\rangle \equiv \sum_{j=1}^{N_E} \frac{\langle f \rangle_j}{N_E} = \sum_{i=1}^{N_S} \frac{f_j(t_i)}{N_S N_E}, \tag{2}$$

Importantly, as averaging is a linear operation $\langle\langle f \rangle\rangle$ is statistically identical to a single average over a time series of length $N_E \times N_S$. Our goal is to estimate variance of the generating distribution for $\langle\langle f \rangle\rangle$, which is complicated by the twofold averaging procedure.

To proceed, we note that when the thermalization and averaging windows used to produce the N_S samples for each $\langle f \rangle_j$ are much longer than the autocorrelation time of the raw data $f_j(t_i)$, the time averages will be independent random variables, satisfiying a central limit theorem in the series length N_S. This can be easily shown by modelling the raw time series data as a Gaussian random variable. In this decorrelated limit the time averages will therefore be independent samples from a Gaussian distribution-

$$\langle f \rangle_j \sim \mathcal{N}(f_0, \sigma_0/\sqrt{N_S}), \tag{3}$$

where $\sim$ signifies 'distributed as' and $\mathcal{N}$ is a Gaussian of mean f_0 and variance σ_0^2/N_S, emphasizing the central limit scaling in N_S. This scaling can be directly tested in `PAFI` or by exploiting the linearity of the averaging process, as shown below.

A crucial implication of (3) is that the simple ensemble variance of the *time averaged* data $\{\langle f \rangle_j\}$ has a *finite* variance even in the limit $N_E \to \infty$:

$$\lim_{N_E \to \infty} \sum_{j=1}^{N_E} \frac{\langle f \rangle_j}{N_E} = \lim_{N_E \to \infty} \langle\langle f \rangle\rangle = f_0, \quad \lim_{N_E \to \infty} \sum_{j=1}^{N_E} \frac{\langle f \rangle_j^2}{N_E} = f_0^2 + \frac{\sigma_0^2}{N_S}. \tag{4}$$

As a result, the ensemble variance of the $\{\langle f \rangle_j\}$ does not estimate the variance in $\langle\langle f \rangle\rangle$. To construct a convergent estimator, we note that the decorrelation assumption (3) can be directly tested by exploiting

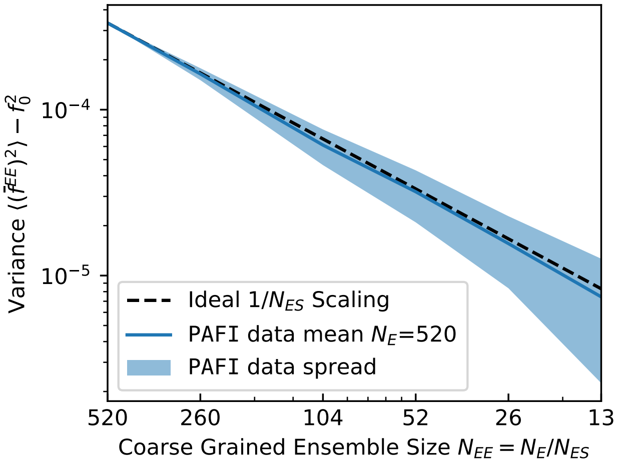

Figure 1: Example test of the decorrelation assumptions in (3) using a set of `PAFI` sampling data across different systems. With increasing reconstructed sample length $N_{ES} = N_E/N_{EE}$ we see a lower variance in agreement with (6), until we reach the small ensemble error with large sample length

the linearity of the averaging process to combining individual worker averages. If we find a factorization $N_E = N_{ES} \times N_{EE}$, we can now artificially construct a new ensemble of size N_{EE} with a time series of length $N_{ES} \times N_S$ formed from N_{ES} individual time averages. This gives 'new' time series averages $\langle \ldots \rangle_j^{EE}$

$$\langle f \rangle_j^{EE} = \sum_{l=(j+1)N_{ES}}^{l=(j+1)N_{ES}} \frac{\langle f \rangle_l}{N_{ES}}, \tag{5}$$

where clearly the partitioning of the $\{\langle f \rangle_l\}$ to form the $\{\langle f \rangle_j^{EE}\}$ is arbitrary. Under the decorrelation assumption (3) these new time series averages will be distributed as

$$\langle f \rangle_j^{EE} \sim \mathcal{N}(f_0, \sigma_0/\sqrt{N_S N_{ES}}). \tag{6}$$

In figure (1) we plot numerical estimations of the ensemble variance for different factorizations, finding (6) to hold very well when the effective ensemble size $N_{EE} = N_E/N_{ES}$ is sufficiently large to suppress statistical error. As a result, when the decorrelation assumption can be shown to hold, we can infer that $\langle\langle f \rangle\rangle$ will be an independent sample from the Gaussian distribution

$$\langle\langle f \rangle\rangle \sim \mathcal{N}(f_0, \sigma_0/\sqrt{N_S N_E}). \tag{7}$$

Our final estimators for the mean and variance of $\langle\langle f \rangle\rangle$ therefore read

$$\hat{\mu}_{\langle\langle f \rangle\rangle} = \sum_{j=1}^{N_E} \frac{\langle f \rangle_j}{N_E}, \quad \hat{\sigma}^2_{\langle\langle f \rangle\rangle} = \sum_{j=1}^{N_E} \frac{\left(\langle f \rangle_j - \hat{\mu}_{\langle\langle f \rangle\rangle}\right)^2}{N_E^2}. \tag{8}$$

The final error estimation for `PAFI` therefore utilises the central limit theorem with respect to N_S to predict the width of the underlying distribution for the 'average of averages' $\langle\langle f \rangle\rangle$.

References

[1] T. D. Swinburne and M.-C. Marinica, "Unsupervised calculation of free energy barrier in large crystalline systems" Phys. Rev. Lett., 2018. `https://doi.org/10.1103/PhysRevLett.120.135503`

[2] T. D. Swinburne and M.-C. Marinica, `PAFI` code (2019). `https://github.com/tomswinburne/pafi`