Estimating the standard error in an ensemble average

Estimating the error in an ensemble average calculated from an MD simulation is challenging as one does not know whether all the important regions of phase space have been sampled.[1-2] It is only possible to determine a lower bound for the error by assessing, in relative terms, the quality of sampling for those regions of configurational space that have been visited. The standard error of the average can be estimated by analyzing the distribution of the fluctuations in the data forming the time series using the two-sample Kolmogorov–Smirnov (KS) statistic. The KS statistic determines the equality between two sample distributions from the maximum vertical distance between their normalized cumulative distributions.[3-4] The KS–statistic has a number of advantages over other metrics used to assess the convergence of data derived from simulations: (1) it makes no assumptions about the nature of the distribution e.g. normality, (2) it involves no fitted parameters, (3) it is most sensitive to differences in the means and less sensitive to outliers. The main limitation is that values obtained from time series generated by molecular dynamics simulations are correlated. As a consequence, it is not possible to use a simple hypothesis test to determine the uncertainty. Instead, the standard error in the average of the time series was estimated by multiplying the KS–statistic, obtained by comparing the 1st and 2nd halves of the time series, with the standard deviation of the entire time series. The robustness of the standard error prediction based on this approach was quantified using a simple heuristic. For this, a predefined target error (E_{target}) for the average of the time series was compared to the estimate of the standard error considering increasingly larger portions of the time series. If the standard error is below E_{target} for the majority of the time series, it is highly likely that the target error has been met. Conversely, if standard error never drops below E_{target} (or does so only briefly), it is likely that E_{target} has not been met. This can be quantified by a convergence robustness score (CR_{score}) defined as the ratio of the portion of the time series continuously below E_{target} and portion above (or which crosses sporadically) E_{target}. The CR_{score} indicates the confidence in the error estimate.

Note the KS statistic was also used to identify parts of the time series not representative of an equilibrium ensemble, for example, those parts heavily biased by the initial conditions. The equilibration time (t_{eq}) was estimated by progressively excluding data from the beginning of the time series while analyzing the standard error. If the standard error decreases systematically as the initial part of the time-series is omitted, it is likely that initial configurations are not representative of the systems at equilibrium and that configurations corresponding to $t < t_{eq}$ should not be included in the calculation of the ensemble average.

For more details including validation see Stroet[5].

References

1. Grossfield, A.; Zuckerman, D. M., Quantifying uncertainty and sampling quality in biomolecular simulations. *Annu. Rep. Comput. Chem.* **2009**, *5*, 23-48.
2. Zuckerman, D. M., Equilibrium sampling in biomolecular simulation. *Annu. Rev. Biophys.* **2011**, *40*, 41.
3. Massey Jr, F. J., The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **1951**, *46*, 68-78.
4. Smirnov, N., Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Stat.* **1948**, 279-281.
5. Stroet, M. Improving the accuracy of molecular dynamics simulations: parameterisation of interaction potentials for small molecules. The University of Queensland, **2018**, doi:10.14264/uql.2018.432.