1 Introduction

Of broad interest to molecular modelers is the computation of expected values of the form

$$\mathbb{E}_{\pi_X}[f(X)] \tag{1}$$

π_X is probability distribution X. Unfortunately computations of this form are generally intractable for all but the simplest of problems. While certain systems may admit unique numerical schemes, broadly speaking black box methods reign as the standard in the field. Broadly speaking, successful methods aim to build a stochastic process $\{X_t\}_{t\in I}$ where the unique stationary distribution is given by π_X. Methods of this type are split into two main categories, (1) Markov chain Monte Carlo (MCMC) simulation and (2) molecular dynamics (MD). The former builds its chain with no assertion that the underlying "dynamics" is realistic, rather it relies on proposal distributions. These range from simple atomic perturbations to complex machine learning proposals with Boltzmann generators. The later relies on symplectic integrators built from trotter expansions of the Liouville operator, which through statistical mechanical theory are guaranteed to reproduce π_X.

Theoretically speaking, there is nothing wrong with these methods whatsoever. Under mild assumptions they can be proven to converge to π_X in the large sample limit $|I| \to \infty$. However, infinite samples can't be achieved in practice. Rather, two practical issues arise.

(i) A transient due to the initial state X_0 being sampled from some distribution $\alpha_X \neq \pi_X$ causes bias in approximations of eq. 1, resulting in a systematic deviation simulation results. When successive simulations depend on one another[1] the bias can compound and lead to a waste of computation.

(ii) The process produced by standard methods is not an independent and identically distributed (i.i.d.) random process. Therefore there exists an *autocorrelation* time for which successive states X_t and X_{t_l} depend on one another statistically. This reduces the effective number of samples to be less than the length of the chain $|I|$.

2 Methods

2.1 Equilibration Detection

Consider a sequence of states $\{X_t\}_{t\in I}$, the *marginal confidence rule* (MCR) of White [1] is given by

$$d^* = \underset{n\gg d\geq 0}{\arg\min}\left[\frac{z_{\alpha/2}\hat{\sigma}_d(\{X_t\})}{\sqrt{n-d}}\right] \tag{2}$$

where $z_{\alpha/2}$ is associated z-score of a two sided $100(1-\alpha)\%$ confidence interval and $\hat{\sigma}_d(\{X_t\})$ is the truncated empirical standard deviation[2]. They are computed like

$$\mathcal{N}_{std.}(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha \implies z_{\alpha/2} = \Phi^{-1}\left(1-\tfrac{\alpha}{2}\right), \tag{3}$$

where Φ^{-1} is the inverse CDF of the standard normal[3] and

$$s_{d:n} = \left(\frac{1}{n-d}\sum_{t=d}^{n}\Big(X_t - \overline{X}_{d:n}\Big)^2\right)^{0.5}, \quad \text{where} \quad \overline{X}_{d:n} = \frac{1}{n-d}\sum_{t=d}^{n}X_t. \tag{4}$$

For fixed α, the above objective is equivalent to

$$d^* = \underset{n\gg d\geq 0}{\arg\min}\left[\frac{1}{(n-d)^2}\sum_{t=d}^{n}\Big(X_t - \overline{X}_{d:n}\Big)^2\right] \tag{5}$$

This then builds an algorithm of the form

[1] thermodynamic integration, collective variable discovery, Bayesian optimization

[2] It seems like White does not apply the -1 correction to the std. estimates in his paper. Maybe its the base zero convention?

[3] For a 95% confidence interval $z_{0.025} = \Phi^{-1}(0.975) \approx 1.96$.

Algorithm 1 Marginal Confidence Rule Equilibration Check

while true **do**
  Run simulation for `substeps` steps.
  Compute $d^* = \arg\min_{n \gg d \geq 0} \left[\frac{1}{(n-d)^2} \sum_{t=d}^{n} \left(X_t - \overline{X}_{d:n} \right)^2 \right]$.
  if $\frac{|I|}{\text{tolerance}} > d^*$ **then**
    return current state X_t.
  end if
end while

2.2 Autocorrelation Determination

Consider a sequence of states $\{X_t\}_{t \in I}$, the *autocorrelation time* can be computed using the unnormalized autocorrelation function

$$C_l = \mathbb{E}[X_t X_{t+l}] - \mathbb{E}[X_t]^2 \approx \mathcal{F}^{-1}[\mathcal{F}[X_t] \cdot \mathcal{F}[X_t]^*] \tag{6}$$

where $\mathcal{F}$ is a Fourier transform of the sequence. This can be normalized like

$$\rho_l = \frac{C_l}{C_0}. \tag{7}$$

The integrated autocorrelation time is defined like

$$\tau_{\text{inf}} = \frac{1}{2} \sum_{l=-\infty}^{\infty} \rho_l. \tag{8}$$

This may be related to the statistical inefficiency s and the effective sample size like

$$s = \frac{\tau_{\text{inf}}}{2}, \quad N_{\text{eff}} = \frac{N}{2\tau_{\text{inf}}}. \tag{9}$$

The effective sample size determines approximately how many samples to skip when trying to derive an i.i.d. chain from $\{X_t\}$.

The empirical mean value of the chain is computed with

$$\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} X_t \tag{10}$$

which has variance given by

$$\text{var}[\hat{\mu}] = \text{var}\left[\frac{1}{n} \sum_{t=1}^{n} X_t \right] = \text{cov}\left(\frac{1}{n} \sum_{t=1}^{n} X_t, \frac{1}{n} \sum_{s=1}^{n} X_s \right) \tag{11}$$

but the covariance is bilinear[4], like an inner product, meaning

$$= \frac{1}{n^2} \sum_{t,s=1}^{n} \text{cov}\left(X_t, X_s\right) = \frac{1}{n^2} \sum_{t,s=1}^{n} C_{t-s} = \frac{1}{n^2} \sum_{l=-(n-1)}^{n-1} (n - |l|) C_l = \frac{1}{n} \sum_{l=-(n-1)}^{n-1} (1 - \tfrac{|l|}{n}) C_l \tag{12}$$

where the last equality can be made by considering C_{t-s} as an N by N matrix and counting the length of the diagonals where the lag l is a constant. If the number of samples is large $n \to \infty$ then $|t|/n$ tends to zero giving

$$\text{var}[\hat{\mu}] \approx \frac{1}{n} \sum_{l=-\infty}^{\infty} C_l = \frac{2C_0 \tau_{\text{inf}}}{n}. \tag{13}$$

This implies the error in the mean can be determined by the integrated autocorrelation time.

[4] $\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. Take $X = \lambda(U + V)$, this implies $\text{cov}(X, Y) = \mathbb{E}[\lambda(U + V)Y] - \mathbb{E}[\lambda(U + V)]\mathbb{E}[Y] = \lambda\mathbb{E}[UY] + \lambda\mathbb{E}[VY] - \mathbb{E}[U]\mathbb{E}[Y] - \mathbb{E}[V]\mathbb{E}[Y]$, but this is equal to $\lambda\text{cov}(U, Y) + \lambda\text{cov}(V, Y)$. The covariance is symmetric, this is obvious from the formula. This means we also have linearity in the other slot. Overall showing it is indeed bilinear.

Estimation of this quantity through a blocking procedure has consistently been garbage for me. Rather I use the method of Alan Sokal [2]. Given finite data, with unknown mean, C_l is estimated like

$$\hat{C}_l = \frac{1}{n-|l|}\sum_{t=1}^{n-|l|}(X_t - \hat{\mu})(X_{t+l} - \hat{\mu}) \tag{14}$$

where then the normalized version is simply

$$\hat{\rho}_l = \frac{\hat{C}_l}{\hat{C}_0} \tag{15}$$

One might expect then that the integrated autocorrelation time is found with

$$\hat{\tau}_{\text{inf}} \overset{?}{=} \frac{1}{2}\sum_{l=-(n-1)}^{n-1}\hat{\rho}_l, \tag{16}$$

however due to the variance in this estimator it becomes quite unreliable when $|t| \gg \tau$. Therefore Alan recommends you window the signal in the following way.

$$\hat{\tau}_{\text{inf}} = \frac{1}{2}\sum_{l=-(n-1)}^{n-1}\lambda(l)\hat{\rho}_l \tag{17}$$

where

$$\lambda(t) \;=\; \begin{cases} 1, & |t| \leq M, \\ 0, & |t| > M. \end{cases} \tag{18}$$

This effectively zeros out the signal beyond some cutoff M. Alan recommends to chose M to be the smallest integer such that $M \geq c\hat{\tau}_{\text{inf}}$ where c is at least 5. For a purely exponential decay c would be $= 4$, therefore you can certainly result get decent results for most signals with $c = 10$. Alan says that this method "works well in practice, provided that a sufficient quantity of data is available ($n < 1000\tau$), however at the time of his writing there was a little knowledge on exact error determination of $\hat{\tau}_{\text{inf}}$ theoretically and empirically.

In practice one can use the fast Fourier transform to compute the empirical autocorrelation quickly. With this in mind his method produces the following algorithm for determining if a sufficient number of samples has been reached during the production period of a

Algorithm 2 Sokal Autocorrelation Production Check

while true **do**
  Run simulation for `substeps` steps
  Compute $n = \text{next\_pow\_two}(t)$.
  Compute $\tilde{X}_\omega = \text{FFT}_{2n}(X_t - \hat{\mu})$
  Compute $\hat{C}_l = \text{IFFT}_t(\tilde{X}_\omega \odot \tilde{X}^*_\omega)$
  while k < t **do**
    Compute $\hat{\tau}_k = -1 + 2\sum_{i=1}^{k}\hat{C}$
    $k = k + 1$
  end while
  $M = \arg\min_M [M < c\hat{\tau}_M \iff \text{True}]$
  if $\frac{t}{\text{tolerance}} > \hat{\tau}_M$ **then**
    return current chain $\{X_t\}$,$\hat{\tau}_M$.
  end if
end while

Essentially this algorithm estimates a sequence of $\hat{\tau}$s for differing window sizes M and then chooses the smallest window which satisfies $M \geq c\hat{\tau}$. Using that autocorrelation estimate the production stage completes when the chain is tolerance times longer than autocorrelation time. Effectively producing tolerance$/2$ effective samples. This parameter can easily be reformatting to ensure a particular confidence interval under the Gaussian assumption for a given observable.

References

(1) White, K. P. *Simulation* **1997**, *69*, 323–334.

(2) Sokal, A. In *Functional Integration: Basics and Applications*, DeWitt-Morette, C., Cartier, P., Folacci, A., Eds.; Springer US: Boston, MA, 1997, pp 131–192.