A2 Systematic deviations due to random errors

When $f(x)$ has a non-negligible curvature, systematic deviations may occur in f as a result of random deviations in x, even when the latter are symmetrically distributed. Suppose you have a batch of spheres with approximately – but not exactly – equal radii. You measure the radii and find an approximately normal distribution with $r = 1.0 \pm 0.1$ mm. For the volume you thus find (in too many decimals): $V = \frac{4}{3}\pi r^3 = 4.19$ mm^3. However, if you work out the third power of r to higher order, you find:

$$(r \pm \Delta r)^3 = r^3 \pm 3r^2\Delta r + 3r(\Delta r)^2 \pm (\Delta r)^3.$$

Assuming the distribution function for Δr to be symmetric, you must conclude that the third term is always positive and thus gives a contribution to the expected value of f:

$$E[r^3] = r^3 + 3r\,\mathrm{var}\,(r).$$

If $E[f(x)] \neq f(E[x])$ we have a *systematic deviation* or *bias*. In our example this extra contribution to the volume is 0.13 mm^3 and the expected volume is 4.32 mm^3. Without this correction, the predicted volume has a bias of -0.13. This is ten times smaller than the standard deviation itself and is therefore not very important. But there are cases when this kind of bias must be corrected.

The general equation results from the second term in a Taylor expansion:

$$f(x) = f(a) + (x-a)f'(a) + \frac{1}{2}(x-a)^2 f''(a) + \cdots \qquad \text{(A2.1)}$$

$$E[f] = f(E[x]) + \frac{1}{2}\frac{d^2 f}{dx^2}\,\mathrm{var}\,(x) + \cdots \qquad \text{(A2.2)}$$

A special case: sampling exponential functions

There is at least one type of fairly common application where evaluation of the bias is essential: computing the average over an exponential function

of a statistically distributed observable. For example, computation of the thermodynamic potential μ of a molecular species in a molecular simulation (molecular dynamics or Monte Carlo) by the *particle insertion method* requires many random trial insertions of a particle. If the computed interaction energy of the inserted particle with its environment of the i-th insertion is E_i, the *excess* thermodynamic potential (in excess of the ideal gas value) is approximated by

$$\mu^{\text{exc}} = \beta^{-1} \ln \left[\frac{1}{N} \sum_{i=1}^{N} e^{-\beta E_i} \right], \tag{A2.3}$$

where $\beta = 1/(k_B T)$ with k_B = Boltzmann's constant and T the absolute temperature. The same kind of averaging occurs in other types of free-energy determinations from simulations. The reader is referred to Berendsen (2007)[1] for details on the physics of these methods.

The essential statistics in problems of this kind can be formulated as *averaging over an exponential function* of a randomly sampled variable x, with distribution function $f(x)$. We are interested in the logarithm of such an average:

$$y = -\frac{1}{\beta} \ln \langle e^{-\beta x} \rangle, \tag{A2.4}$$

where

$$\langle e^{-\beta x} \rangle = E[e^{-\beta x}] = \int_{-\infty}^{\infty} f(x)\, e^{-\beta x}\, dx. \tag{A2.5}$$

The parameter β functions as a scaling for x: given a fixed probability distribution for x, the larger β, the more severe the statistical problems on averaging appear to be. The problem is that occasional large negative values for x contribute heavily to the average. We can get some insight by expanding y in powers of β; such an expansion is called a *cumulant expansion*. For simplicity we take $\langle x \rangle = 0$, so that all moments of the distribution of x are central moments. Adding an arbitrary value a to every x simply results in adding a to the result y. The cumulant expansion is

$$y = -\frac{\beta}{2!} \langle x^2 \rangle + \frac{\beta^2}{3!} \langle x^3 \rangle - \frac{\beta^3}{4!} (\langle x^4 \rangle - 3 \langle x^2 \rangle^2) + \mathcal{O}(\beta^4). \tag{A2.6}$$

For a normal distribution only the first term survives, resulting in

$$y = -\beta/2, \tag{A2.7}$$

as can be checked by direct integration of (A2.5). This is a bias due to the width of the normal distribution. If we know for sure that the distribution

[1] See reference list on page 123.

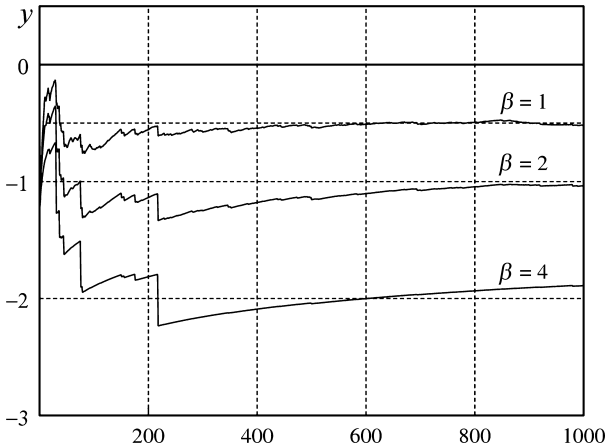

Figure A2.1 Cumulative average of $y = -\beta^{-1} \ln\langle \exp(-\beta x)\rangle$ over n samples drawn from a normal distribution (average 0, variance $\sigma^2 = 1$). The theoretical limits are -0.5β, indicated by dotted lines (from Berendsen, 2007).

function of x is normal, y can be accurately determined from (A2.7), but it is difficult to determine y from random sampling of x. To show this, Fig. A2.1 gives the values of y obtained from running averages for 1000 samples of x from a normal distribution and for three values of β. It turns out that 1000 samples are barely sufficient to find convergence for $\beta = 2$, but for $\beta = 4$ this number by no means suffices.