Today's lecture: basics of probability and statistics; both discrete and continuous distributions.

Probability and statistics

Probability distributions

The behavior of systems at equilibrium is described by molecular statistical distribution functions. Therefore, we now briefly review some properties of statistics.

Let g be a discrete variable. Denote a probability distribution $\wp(g)$ as the probability that g will take a given value. This distribution is normalized such that

$$\sum_i \wp(g_i) = 1$$

where the sum proceeds over all allowable, distinct values of g. Here, $\wp(g)$ is dimensionless since it returns a probability.

On the other hand, let x be a continuous variable. Then, we define a probability distribution $\wp(x)$ such that:

$$\int \wp(x)dx = 1$$

Note that in order to be dimensionally consistent, $\wp(x)$ must have inverse dimensions of x, due to the presence of the dx term. Thus, $\wp(x)$ becomes a probability **density**. In this interpretation we think of the combined expression $\wp(x_0)dx$ as the probability (dimensionless) that x takes on a value in the range $x_0 - dx/2$ to $x_0 + dx/2$.

In simulation, we are often interested in measuring distribution functions. Typically this is accomplished using **histograms.** A histogram is simply a counting of the different numbers of times that we see a variable with different values. This is straightforward for discrete variables.

For continuous distributions, however, we must generate a discrete approximation to them by creating a histogram with bins of a finite width. For example, we may make a histogram of the energy with a bin width of 0.5 kcal/mol. The histogram bin for the range 0-0.5 kcal/mol would count the number of times we observed an energy in that range, similar for 0.5-1.0 kcal/mol, and so on and so forth.

Let the histogram bin width be Δx and let $H(x)$ be the number of times we observe x with values between $x - \Delta x/2$ and $x + \Delta x/2$. Then, our discrete distribution is

$$\widetilde{\wp}(x) = \frac{H(x)}{\Delta x \sum_{x'} H(x')}$$

Here, the use of the tilde indicates that the probability distribution is a discrete approximation to a continuous one. We can recover the continuous one in the limit:

$$\wp(x) = \lim_{\Delta x \to 0} \frac{H(x)}{\Delta x \sum_{x'} H(x')}$$

There is always a tradeoff in simulations: as the histogram bin size Δx grows smaller, our discrete approximate becomes increasingly continuous-like. However, the tradeoff is that the number of counts in any one histogram bin becomes very small, and statistical accuracy grows poor unless we extend the length of our simulation.

Multivariate distributions

The **joint probability** distribution for multiple variables indicates the probability that all variables will *simultaneously* attain a given value. For example, for two discrete variables g and f,

$$\wp(g = g_0, f = f_0)$$

gives the joint probability that a measurement will occur in which it is found that g has value g_0 and f has f_0.

If we consider continuous variables x and y, then

$$\wp(x_0, y_0)dxdy$$

is the joint probability that we see an event with x in the range $x_0 \pm dx/2$ and y in the range $y_0 \pm dy/2$. Notice that $\wp(x, y)$ has inverse dimensions of both x and y.

We can extract single-variable distributions from multiple variable ones by summing or integrating over possibilities for the other variables:

$$\wp(g) = \sum_f \wp(g, f)$$

$$\wp(x) = \int \wp(x, y)dy$$

Notice that $\wp(x)$ now has inverse dimensions of only x, as we would expect.

If two variables are completely **independent,** then their joint probability is simply the product of their individual probabilities:

$$\wp(x,y) = \wp(x)\wp(y) \qquad \text{(independent variables)}$$

The **conditional probability** is the probability of a variable value given that some other observation has been made. For example,

$$\wp(g|f) \quad \text{or} \quad \wp(g;f)$$

is the probability distribution for g given a particular observed value of f. This distribution can be related to the joint and single-variable distributions by:

$$\wp(g|f) = \frac{\wp(g,f)}{\wp(f)}$$

Identical expressions apply to continuous variables.

Distribution moments

We can compute the average value of a distribution variable. For example,

$$\langle x \rangle = \int x \wp(x) dx$$

The variance is given by:

$$\begin{aligned}\sigma_x^2 &= \langle (x - \langle x \rangle)^2 \rangle \\ &= \langle x^2 - 2x\langle x \rangle + \langle x \rangle^2 \rangle \\ &= \langle x^2 \rangle - 2\langle x \rangle \langle x \rangle + \langle x \rangle^2 \\ &= \langle x^2 \rangle - \langle x \rangle^2\end{aligned}$$

with

$$\langle x^2 \rangle = \int x^2 \wp(x) dx$$

What is the expectation value of some variable A that is a function of an observable x? To compute such a property, we integrate the distribution:

$$\langle A \rangle = \int A(x) \wp(x) dx$$

The variance in A is similarly found:

$$\sigma_A^2 = \langle (A - \langle A \rangle)^2 \rangle$$

$$= \langle A^2 \rangle - \langle A \rangle^2$$
$$= \int A(x)^2 \wp(x) dx - \left[\int A(x) \wp(x) dx \right]^2$$

If we compute $\theta = \langle x \rangle$ using only a finite number of independent observations n, such that

$$\theta = \frac{\sum x_i}{n}$$

we can also compute the expected variance in θ, σ_θ^2. This is called the **standard error of the mean**. As $n \to \infty$, we expect $\sigma_\theta^2 \to 0$. That is, as we take more and more samples, we expect our approximation to the underlying distribution average to grow more and more accurate. Under the assumption of independent observations,

$$\wp(\theta) = \prod \wp(x_i)$$

$$\sigma_\theta^2 = \langle \theta^2 \rangle - \langle \theta \rangle^2$$

Working through the math leads to,

$$\sigma_\theta^2 = \frac{\sigma_x^2}{n}$$

Thus, if we are trying to find the average of a distribution, our accuracy increases as the square root of the number of *independent* samples we take.

Typically we take measurements at different points in time in a molecular simulation. If we take them too frequently, the measurements will not be independent but will be highly **correlated**. The above expression can be generalized to:

$$\sigma_\theta^2 = \frac{\sigma_x^2}{n} \left(1 + \frac{2\tau_x}{\Delta t} \right)$$

Here, τ_x is the **relaxation time** or **correlation time** for the variable x, and Δt is the spacing in time at which we take measurements. This equation shows that the standard error of the mean can be much larger than what we would expect for independent measurements if the time intervals are not appreciably larger than the relaxation time.

Microstate probability distributions in classical systems

Basic concepts

In the classical description, we describe a system by the positions and momenta of all of the atomic nuclei:

$$\mathbf{r}^N = (x_1, y_1, z_1, x_2, \ldots, y_N, z_N)$$

$$\mathbf{p}^N = \left(p_{x,1}, p_{y,1}, p_{z,1}, p_{x,2}, \ldots, p_{y,N}, p_{z,N}\right)$$

A **microstate** is just one "configuration" of the system. In a classical system, one microstate is characterized by a list of the $3N$ positions $\mathbf{r}^N$ and $3N$ momenta $\mathbf{p}^N$, for a total of $6N$ pieces of information. For a microstate m we might use the notation $(\mathbf{p}_m^N, \mathbf{r}_m^N)$ to indicate specific values of these variables.

At equilibrium, the bulk properties of a system are time-invariant. We might construct a **microscopic probability distribution function** that describes the probability with which we might see a given classical microstate m at any one instance in time. We have that

$$\wp(\mathbf{p}_m^N, \mathbf{r}_m^N) d\mathbf{p}^N d\mathbf{r}^N$$

gives the probability of microstate m. That is, $\wp(\mathbf{p}_m^N, \mathbf{r}_m^N) d\mathbf{p}_1 \ldots d\mathbf{p}_N d\mathbf{r}_1 \ldots d\mathbf{r}_N$ is proportional to the continuous joint probability that the system is at a microstate that lies between $p_{1,x} - dp_{1,x}/2$ and $p_{1,x} + dp_{1,x}/2$, $p_{1,y} - dp_{1,y}/2$ and $p_{1,y} + dp_{1,y}/2$, and so on and so forth. In other words, the probability corresponds to a microstate within a differential element $d\mathbf{p}_1 \ldots d\mathbf{p}_N d\mathbf{r}_1 \ldots d\mathbf{r}_N$ centered around $(\mathbf{p}_m^N, \mathbf{r}_m^N)$.

The particular form of the distribution function $\wp(\mathbf{p}^N, \mathbf{r}^N)$ depends on the conditions at which a system is maintained, i.e., the particular **statistical mechanical ensemble**. Systems can either be at:

constant E or constant T (exchanging E with a heat bath / energy reservoir)

constant V or constant P (exchanging V with a volume reservoir)

constant N or constant μ (exchanging N with a particle bath)

We will discuss these ensembles in more depth as we progress through different simulation methods. As an example, in the *canonical ensemble* (constant T, V, N), the probability for a specific structure / configuration $\mathbf{r}_m^N$ is proportional to the Boltzmann factor:

$$\wp(\mathbf{r}^N) \propto e^{-\frac{U(\mathbf{r}^N)}{k_B T}} = e^{-\beta U(\mathbf{r}^N)}$$

where $U(\mathbf{r}^N)$ gives the potential energy of that configuration, which in simulation is determined by the force field.

Very often we are interested in the distribution of some structural quantity or order parameter, such as the distance between two molecules. These distributions can be obtained formally with delta functions. Consider the distance r_{12} between particles 1 and 2:

$$\wp(r_{12}) = \int \wp(\mathbf{r}^N)\delta[r_{12} - |\mathbf{r}_1 - \mathbf{r}_2|]d\mathbf{r}^N$$

Other references for statistical mechanics

Though a detailed review of statistical mechanical concepts is beyond this course, it closely follows material from my graduate level text book:

> *Thermodynamics and Statistical Mechanics: An Integrated Approach*
> M. Scott Shell, Cambridge University Press, 2015
> The following chapters are particularly relevant to this course:
> Chapters 3, 4, 5, 7, 16, 17, 18, 19, and 22

The following two classic statistical mechanics texts are also highly recommended:

> *Statistical Mechanics*
> Donald A. McQuarrie, University Science Books, 2000 (2nd edition)
> This is a classic statistical mechanics text, and McQuarrie does an excellent job of laying out clear, concise explanations of the subject material. It serves well as both an introduction to the subject and a reference for specific models and theoretical techniques. In this course, we will cover material in parts of Chapters 1-3, 5, 7, and 9. This book is also frequently used as a primary text in ChE 210B.

> *An Introduction to Statistical Thermodynamics*
> Terrell Hill, Dover Books, 1987
> This very inexpensive paperback is a tour-de-force in laying out the foundations and early theoretical advancements of statistical mechanics. Hill takes care to discuss many of the subtleties that other texts glance over, and provides detailed derivations for major theories. The density of material in the book often necessitates careful study and re-reading at first, but it is an essential reference for anyone involved in research broadly related to molecular thermodynamics. In the course, we will cover material in parts of Chapters 1-4 and 6.