2 Probability

2.1 Averages and probabilities

Mean and variance:

$$\langle x \rangle = \int dx x p(x)$$
$$\sigma^2 = \int dx (x - \langle x \rangle)^2 p(x)$$

The probability for the purpose of statistical mechanics can be understood as the frequency of occurrence. For instance, if the value $x = 2$ is measured n_i times out of n_t of total attempts, the probability is $p_i = n_i/n_t$. This means that if $x = 2$ is assigned the probability of 10^{-2}, one would measure $x = 2$ on average 100 times out of each set of 10^4 measurements.

When the variable x is continuous, one specifies the probability density $p(x)$. It implies that the frequency (or fraction) of x falling in the interval $x \leq x \leq x + \Delta x$ is

$$\Delta p = \Delta n/n_t = p(x)\Delta x. \tag{2.1}$$

The sum of fractions for all possible intervals is unity and the probability density satisfies the normalization condition

$$\sum_i \Delta p_i \xrightarrow[dx \to 0]{} \int dx p(x) = 1. \tag{2.2}$$

It can be specified either by its functional form or by its moments, of which the mean $\langle x \rangle$ and variance σ^2 are most commonly considered. Those two are the only ones needed for $p(x)$ specifying the Gaussian distribution discussed below.

The general definition of the moment of nth order is

$$\langle x^n \rangle = \int dx x^n p(x). \tag{2.3}$$

In addition, one defines central moments of nth order

$$\langle(\delta x)^n\rangle = \int dx (x - \langle x\rangle)^n p(x). \tag{2.4}$$

Variance, $\sigma^2 = \langle(\delta x)^2\rangle$, is the second-order central moment. The square root of the variance is known as the standard deviation. It also comes under the name of root-mean-squared deviation

$$\Delta x_{\text{rms}} = \sqrt{\langle(x - \langle x\rangle)^2\rangle}. \tag{2.5}$$

When several stochastic (random) variables are involved, one can specify the joint probability density $p(x_1, x_2, \ldots, x_m)$, which gives the probability of measuring $x_1, \ldots, x_m$ simultaneously. When the measurements are independent, one has

$$p(x_1, \ldots, x_m) = p_1(x_1) \times p_2(x_2) \times \cdots \times p_m(x_m). \tag{2.6}$$

When this approximation does not apply, the two variables are said to be correlated. The most common correlations are binary. The binary correlation function $g(x_1, x_2)$ shows how much the joint probability deviates from the assumption of independent stochastic variables

$$p(x_1, x_2) = p_1(x_1) p_2(x_2) g(x_1, x_2). \tag{2.7}$$

One has $g(x_1, x_2) = 1$ for statistically independent variables. The following notation is used to describe correlations between two variables x_1 and x_2

$$\langle x_1 x_2\rangle = \int dx_1 dx_2 x_1 x_2 p(x_1, x_2). \tag{2.8}$$

If x_1 and x_2 are uncorrelated (statistically independent), one obtains

$$\langle x_1 x_2\rangle = \langle x_1\rangle\langle x_2\rangle. \tag{2.9}$$

The correlation of deviations from averages $\delta x_i = x_i - \langle x_i\rangle$ (the covariance) is zero in this case

$$\langle \delta x_1 \delta x_2\rangle = 0. \tag{2.10}$$

2.1.1 Central limit theorem

A number of general rules can be formulated for composite variables or running averages

$$X_N = N^{-1} \sum_{i=1}^{N} x_i. \tag{2.11}$$

Such types of variables are most common in science. They apply either to results of repeated measurements, when N is the number of data points, or, more commonly, to measurements done on macroscopic objects. In the latter case, N is the number of particles and x_i is a property assigned to each single particle, for instance the particle energy.

One can start by looking at the variance of X_N assuming that variables $\delta x_i = x_i - \langle x_i \rangle$ are statistically independent: $\langle \delta x_i \delta x_j \rangle = 0$ when $i \neq j$. What follows from this assumption is a connection between the variance σ_X^2 of the composite variable X_N and the variance σ_x^2 of the individual variable x describing each separate particle

$$\sigma_X^2 = \sigma_x^2/N. \tag{2.12}$$

This is an important result known as the law of large numbers. It states that the result of measurements is improved when more measurements are done, and the standard deviation of those repeated measurements scales down as $1/\sqrt{N}$. Similarly, the reason a well-defined energy can be assigned to a macroscopic sample made of many small parts is because fluctuations of the sample energy scale down as $1/\sqrt{N}$, where N is the number of atoms or molecules making up the sample. Of course, the closest neighbors in that sample will still correlate, but there will always be sufficient number of far away molecules to render fluctuations of any macroscopic property negligible.

Another iconic result of the probability theory used in science across disciplines is the central limit theorem. It states that as the number N defining the composite variable X_N increases, the distribution of X_N tends to a Gaussian distribution. We have

$$p(X_N) \underset{N\gg 1}{\rightarrow} \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-(X_N - \langle X_N \rangle)^2/(2\sigma_X^2)}. \tag{2.13}$$

It is obvious that as N increases, $\langle X_N \rangle$ becomes $\langle x \rangle$ and the distribution width shrinks as $1/\sqrt{N}$. The result is an infinitely narrow distribution known as the delta-function

$$p(X_N) \underset{N\to\infty}{\rightarrow} \delta(X_N - \langle x \rangle). \tag{2.14}$$

This equation implies that one reports a single value $\langle x \rangle$ for a property per particle (e.g., energy per particle) when it is measured from a macroscopic sample. This result is the mathematical basis for our reliance on

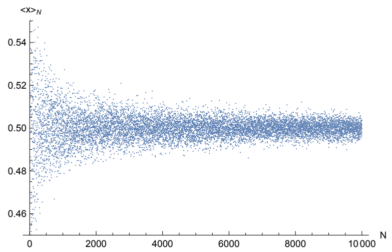

Figure 2.1: Running average X_N (Eq. (2.11)) vs N.

macroscopic measurements for defining properties per particle (or per mole), which, neglecting quantum effects, are limited only by instrumental resolution and not by thermal fluctuations inherent to any system of molecular dimension.

2.1.2 Uniform distribution

The uniforms distribution is formed by random numbers in the x-axis interval $[0, a]$. The probability density is uniform, $p(x) = p$, and has to be normalized on the interval

$$\int_0^a dxp(x) = p \times a = 1, \quad p = a^{-1}. \tag{2.15}$$

The mean value of x is

$$\langle x \rangle = \int_0^a dxxp(x) = p\frac{x^2}{2}\bigg|_0^a = \frac{a}{2}. \tag{2.16}$$

The variance is

$$\sigma^2 = \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2 = \frac{a^2}{3} - \frac{a^2}{4} = \frac{a^2}{12}. \tag{2.17}$$

Figure 2.1 illustrates the convergence of the running average X_N

$$X_N = \frac{1}{N}\sum_i r_i \tag{2.18}$$

calculated for the random number r_i in the interval $[0, 1]$. The average converges to the predicted value $X_N \to \langle r \rangle = 1/2$ as a function of the number of trials N. The spread of the points around the average is the variance σ_X^2 (Eq. (2.12)). It converges to the anticipated limit $\sigma_X = 1/\sqrt{12N} \approx 0.29/\sqrt{N}$ much slower than the average itself.

The linear scaling of the variance with $1/N$ predicted by Eq. (2.12) is illustrated in Fig. 2.2 which shows the variance of X_N vs $1/N$. Each point

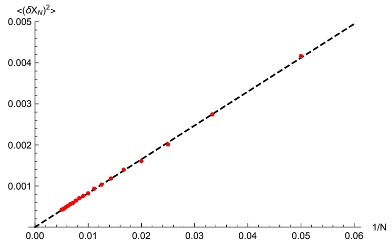

Figure 2.2: Variance of X_N (Eq. (2.18)) vs $1/N$.

in the plot was calculated from 10,000 trials of collecting X_N for a given value of N. As expected, the variance decays to zero as N^{-1} with the slope equal to 1/12 (Eq. (2.12)). The linear fit through the points (the dashed line in the figure) produces the expected slope.

2.1.3 Poisson distribution

Poisson distribution defines the probability of achieving a given number of discrete outcomes n given the expected mean value μ. The individual events are random and uncorrelated between each other. The resulting formula for the probability density is

$$P(n) = \frac{\mu^n}{n!} e^{-\mu}. \tag{2.19}$$

The probability density is normalized to unity and has an additional property of equal values for the mean and the variance

$$\begin{aligned} \sum_{n=0}^{\infty} P(n) &= 1, \\ \langle n \rangle = \sum_{n=0}^{\infty} nP(n) &= \mu, \\ \sigma^2 = \langle (n - \langle n \rangle)^2 \rangle = \sum_{n=0}^{\infty} (n - \langle n \rangle)^2 P(n) &= \mu. \end{aligned} \tag{2.20}$$

At $\mu \gg 1$, $P(n)$ tends to a Gaussian distribution. This can be proven by using Stirling's formula

$$\boxed{n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \approx \left(\frac{n}{e}\right)^n.} \tag{2.21}$$

The proof requires estimating the function

$$P(n) \approx \exp[-\mu + F(n)], \quad F(n) = n \ln \mu + n - n \ln n. \tag{2.22}$$

The function $F(n)$ can be expanded around its maximum $F'(n^*) = 0$

$$F'(n^*) = \ln \mu - \ln n^*, \quad n^* = \mu. \tag{2.23}$$

The series expansion is

$$F(n) = F(n^*) - \frac{1}{2\mu}(n-\mu)^2. \tag{2.24}$$

One arrives at the Gaussian distribution as an estimate of $P(n)$

$$P(n) = [2\pi\mu]^{-1/2} \exp\left[-\frac{(n-\mu)^2}{2\mu}\right]. \tag{2.25}$$

The Gaussian distribution produced from $P(n)$ at $\mu \gg 1$ carries the property of the Poisson distribution of having $\sigma^2 = \langle n \rangle = \mu$.

As an application of the Poisson distribution, consider the fluctuation of the number of proteins with the concentration $c \approx 1$ mM in the volume of $V = 1\ \mu\text{m}^3$. The arrival and departure of proteins from this volume is a random noise of uncorrelated arrivals and departures of individual proteins. The number of proteins in the volume N is a stochastic variable following the Poisson distribution with the average number

$$\langle N \rangle = c \times N_A \times 10^{-15}, \tag{2.26}$$

where the transformation $1\ \mu\text{m}^3 = 1\text{fL}$ was used in the equation. One obtains from this equation $\langle N \rangle = 6 \times 10^5$. This number of proteins is of the order typically found in a bacterial cell, which is consistent with the chosen volume and concentration. Given the large average number, one can directly write the distribution $P(N)$ by Eq. (2.25). The variance of the number of particles in the volume V is equal to $\langle N \rangle$.

2.1.4 Gaussian distribution

The Poisson distribution tends at $n \gg 1$ to a specific form of the Gaussian distribution function $p_G(x)$ (probability density) specified by $\langle x \rangle = \langle (\delta x)^2 \rangle$. This relation does not have to hold for a general form of the Gaussian distribution which is fully specified by two parameters, the mean (or the average value) $\langle x \rangle$ and the variance $\sigma^2 = \langle (\delta x)^2 \rangle$. The probability density is given by the following equation

$$p_G(x) = \left[2\pi\sigma^2\right]^{-1/2} \exp\left[-\frac{(x - \langle x \rangle)^2}{2\sigma^2}\right] \tag{2.27}$$

The delta function in Eq. (2.14) is the limit of the Gaussian function at the variance approaching zero, as is the case for macroscopic systems following the central limit theorem

$$p_G(x) \xrightarrow[\sigma\to 0]{} \delta(x - \langle x \rangle). \tag{2.28}$$

2.1.5 Maxwell distribution

The Maxwell distribution i a specific form of the Gaussian distribution of the vector field $\mathbf{v}$ considering the probability density of the vector magnitude $v = |\mathbf{v}|$. It is historically applied to the distribution of thermal velocities, but is not limited by this specific application. The Gaussian distribution of $\mathbf{v}$ is

$$p_G(\mathbf{v}) = \left[2\pi\sigma^2\right]^{-3/2} \exp\left[-\frac{(\mathbf{v} - \langle \mathbf{v} \rangle)^2}{2\sigma^2}\right]. \tag{2.29}$$

It is normalized as

$$\iiint_{-\infty}^{\infty} dv_x dv_y dv_z p_G(\mathbf{v}) = 1. \tag{2.30}$$

The Maxwell distribution is obtained for a specific case of $\langle \mathbf{v} \rangle = 0$ when the probability density for v becomes

$$p_M(v) = 4\pi v^2 p_G(\mathbf{v}) \tag{2.31}$$

Correspondingly, the normalization becomes

$$\int_0^{\infty} dv p_M(v) = 1. \tag{2.32}$$

We now apply these general results to the distribution of thermal velocities. We start with the 1D space of v_x projections.

The average v_x for a liquid or gas is zero in equilibrium (no overall motion): $\langle v_x \rangle = 0$. To specify the Gaussian distribution, one needs only to provide the variance. This is achieved with the equipartition theorem stating that $m\langle v_x^2\rangle = k_\mathrm{B}T$. The variance thus becomes

$$\sigma_x^2 = \langle v_x^2\rangle = v_\mathrm{th}^2 = k_\mathrm{B}T/m. \tag{2.33}$$

and the probability density is

$$p(v_x) = \left[\frac{m}{2\pi k_\mathrm{B}T}\right]^{1/2} \exp\left[-\frac{mv_x^2}{2k_\mathrm{B}T}\right]. \tag{2.34}$$

All directions are equivalent for for a uniform system and one gets the probability density of velocities

$$p(\mathbf{v}) = p(v_x)p(v_y)p(v_z) = \left[\frac{m}{2\pi k_\mathrm{B}T}\right]^{3/2} \exp\left[-\frac{mv^2}{2k_\mathrm{B}T}\right], \tag{2.35}$$

where $v^2 = v_x^2 + v_y^2 + v_z^2$. Therefore, one obtains $\sigma^2 = \sigma_x^2 = k_\mathrm{B}T/m$ in Eq. (2.29).

The probability density of speeds $v = |\mathbf{v}|$ is found by summing up all probabilities characterized by the same value of the speed

$$p(v) = \sum_{\mathbf{v},|\mathbf{v}|=v} p(\mathbf{v}). \tag{2.36}$$

The sum represents all possible configurations of the vector $\mathbf{v}$ on the sphere with the radius v. The surface of this sphere is $4\pi v^2$ and one obtains

$$p_M(v) = 4\pi v^2 p(\mathbf{v}) = 4\pi v^2 \left[\frac{m}{2\pi k_\mathrm{B}T}\right]^{3/2} \exp\left[-\frac{mv^2}{2k_\mathrm{B}T}\right]. \tag{2.37}$$

This probability density is known as the Maxwell distribution of thermal speeds. The variances of $p(\mathbf{v})$ and $p_M(v)$ are equal and are given by the sum of individual x, y, z-variances

$$\sigma^2 = \sigma_x^2 + \sigma_y^2 + \sigma_z^2 = 3\sigma_x^2 = 3k_\mathrm{B}T/m. \tag{2.38}$$

The average speed is

$$\langle v\rangle = \int_0^\infty dv v p_M(v) = \sqrt{8k_\mathrm{B}T/(\pi m)}. \tag{2.39}$$

It is somewhat lower than the RMS speed ($8/\pi < 3$)

$$v_{\rm rms} = \sqrt{\langle v^2 \rangle} = \sqrt{3k_{\rm B}T/m}. \tag{2.40}$$

When both nominator and denominator in this equation are multiplied with $\sqrt{N_A}$, one gets the equation for $v_{\rm rms}$ in terms of the gas constant $R = N_A k_{\rm B}$ and the molar mass $M = mN_A$

$$v_{\rm rms} = \sqrt{\langle v^2 \rangle} = \sqrt{3RT/M}. \tag{2.41}$$

Both $\langle v \rangle$ and $v_{\rm rms}$ are higher than the velocity $v_{\rm max}$ at which the Maxwell distribution reaches its maximum

$$P'_M(v_{\rm max}) = 0, \quad v_{\rm max} = \sqrt{2k_{\rm B}T/m}. \tag{2.42}$$

One therefore obtains the sequence

$$v_{\rm max} < \langle v \rangle < v_{\rm rms}. \tag{2.43}$$

2.2 Least square problem

The least square formulation of fitting considers the best set of constants in the fitting function $y(t) = f(c,t)$ representing the data set (t_i, y_i), $i = 1, N$, where $c = c_1, \ldots, c_M$ is a set of fitting parameters. The principle of least squares states that the best estimate of the of a set of data is a function that minimizes the sum of squares of deviations of the data points from their estimates.

The best known example is the linear regression which is considered here explicitly. The function is the straight line

$$y = a + bt, \tag{2.44}$$

where we want to find the coefficients a and b best representing the data.

To solve the problem, one determines the residual vector of length N

$$r_i = y_i - a - bt_i$$

and finds the coefficients by the condition that the magnitude of $\mathbf{r}$ is minimized

$$\min_{a,b}\left[\mathbf{r}.\mathbf{r}\right] = \min_{a,b} \sum_{i=1}^{N} r_i^2.$$