4.4 From probability to certainty

4.4.1 Sums of random variables

Consider the sum $S = \sum_{i=1}^{N} x_i$, where x_i are random variables with a joint PDF of $p(\mathbf{x})$. The PDF for S is

$$p_S(x) = \int d^N\mathbf{x}\; p(\mathbf{x})\delta\left(x - \sum x_i\right) = \int \prod_{i=1}^{N-1} dx_i\; p\left(x_1, \cdots, x_{N-1}, x - x_1 \cdots - x_{N-1}\right), \tag{4.4.1}$$

and the corresponding characteristic function (using Eq. (4.3.7)) is given by

$$\tilde{p}_S(k) = \left\langle \exp\left(-ik\sum_{j=1}^{N} x_j\right)\right\rangle = \tilde{p}\left(k_1 = k_2 = \cdots = k_N = k\right). \tag{4.4.2}$$

Cumulants of the sum are obtained by expanding $\ln \tilde{p}_S(k)$,

$$\ln \tilde{p}\left(k_1 = k_2 = \cdots = k_N = k\right) = -ik\sum_{i_1=1}^{N} \left\langle x_{i_1}\right\rangle_c + \frac{(-ik)^2}{2}\sum_{i_1,i_2}^{N} \left\langle x_{i_1}x_{i_2}\right\rangle_c + \cdots, \tag{4.4.3}$$

as

$$\left\langle S\right\rangle_c = \sum_{i=1}^{N} \left\langle x_i\right\rangle_c \quad , \quad \left\langle S^2\right\rangle_c = \sum_{i,j}^{N} \left\langle x_i x_j\right\rangle_c \quad , \quad \cdots. \tag{4.4.4}$$

If the random variables are independent, $p(\mathbf{x}) = \prod p_i(x_i)$, and $\tilde{p}_S(k) = \prod \tilde{p}_i(k)$. The cross–cumulants in Eq. (4.4.4) vanish, and the n^{th} cumulant of S is simply the sum of the individual cumulants, $\left\langle S^n\right\rangle_c = \sum_{i=1}^{N} \left\langle x_i^n\right\rangle_c$. When all the N random variables are independently taken from the same distribution[1] $p(x)$, this implies $\left\langle S^n\right\rangle_c = N\left\langle x^n\right\rangle_c$, generalizing the result obtained previously for the binomial distribution. For large values of N, the average value of the sum is proportional to N, while fluctuations around the mean, as measured by the standard deviation, grow only as $\sqrt{N}$. The random variable $y = (S - N\left\langle x\right\rangle_c)/\sqrt{N}$, has zero mean, and cumulants that scale as $\left\langle y^n\right\rangle_c \propto N^{1-n/2}$. As $N \to \infty$, only the second cumulant survives, and the PDF for y converges to the normal distribution,

$$\lim_{N\to\infty} p\left(y = \frac{\sum_{i=1}^{N} x_i - N\left\langle x\right\rangle_c}{\sqrt{N}}\right) = \frac{1}{\sqrt{2\pi\left\langle x^2\right\rangle_c}}\exp\left(-\frac{y^2}{2\left\langle x^2\right\rangle_c}\right). \tag{4.4.5}$$

(Note that the Gaussian distribution is the only distribution with only first and second cumulants.)

The convergence of the PDF for the sum of many random variables to a normal distribution is an essential result in the context of statistical mechanics where such sums are frequently encountered. The *central limit theorem* states a more general form of this result: It is not necessary for the random variables to be independent, as the condition $\sum_{i_1,\cdots,i_m}^{N} \left\langle x_{i_1}\cdots x_{i_m}\right\rangle_c \ll \mathcal{O}(N^{m/2})$, is sufficient for the validity of Eq. (4.4.5).

[1] Such variables are referred to as IIDs for identical, independently distributed.