II.D Many Random Variables

With more than one random variable, the set of outcomes is an N-dimensional space, $\mathcal{S}_{\mathbf{x}} = \{-\infty < x_1, x_2, \cdots, x_N < \infty\}$. For example, describing the location and velocity of a gas particle requires six coordinates.

• *The joint PDF* $p(\mathbf{x})$, is the probability density of an outcome in a volume element $d^N\mathbf{x} = \prod_{i=1}^{N} dx_i$ around the point $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$. The joint PDF is normalized such that

$$p_{\mathbf{x}}(\mathcal{S}) = \int d^N\mathbf{x}\; p(\mathbf{x}) = 1 \; . \tag{II.32}$$

If, and only if, the N random variables are *independent*, the joint PDF is the product of individual PDFs,

$$p(\mathbf{x}) = \prod_{i=1}^{N} p_i(x_i) \quad . \tag{II.33}$$

• *The unconditional PDF* describes the behavior of a subset of random variables, independent of the values of the others. For example, if we are interested only in the location of a gas particle, an unconditional PDF can be constructed by integrating over all velocities at a given location, $p(\vec{x}) = \int d^3\vec{v}\; p(\vec{x}, \vec{v})$; more generally

$$p(x_1, \cdots, x_m) = \int \prod_{i=m+1}^{N} dx_i \; p(x_1, \cdots, x_N) \; . \tag{II.34}$$

• *The conditional PDF* describes the behavior of a subset of random variables, for specified values of the others. For example, the PDF for the velocity of a particle at a particular location $\vec{x}$, denoted by $p(\vec{v} \mid \vec{x})$, is proportional to the joint PDF $p(\vec{v} \mid \vec{x}) = p(\vec{x}, \vec{v})/\mathcal{N}$. The constant of proportionality, obtained by normalizing $p(\vec{v} \mid \vec{x})$, is

$$\mathcal{N} = \int d^3\vec{v}\; p(\vec{x}, \vec{v}) = p(\vec{x}), \tag{II.35}$$

the unconditional PDF for a particle at $\vec{x}$. In general, the unconditional PDFs are obtained from *Bayes' Theorem* as

$$p(x_1, \cdots, x_m \mid x_{m+1}, \cdots, x_N) = \frac{p(x_1, \cdots, x_N)}{p(x_{m+1}, \cdots, x_N)} \quad . \tag{II.36}$$

Note that if the random variables are independent, the unconditional PDF is equal to the conditional PDF.

• *The expectation value* of a function $F(\mathbf{x})$, is obtained as before from

$$\langle F(\mathbf{x})\rangle = \int d^N\mathbf{x}\; p(\mathbf{x})F(\mathbf{x}) \; . \tag{II.37}$$

• *The joint characteristic function*, is obtained from the N-dimensional Fourier transformation of the joint PDF,

$$\tilde{p}(\mathbf{k}) = \left\langle \exp\left(-i\sum_{j=1}^{N} k_j x_j\right)\right\rangle \quad . \tag{II.38}$$

The *joint moments* and *joint cumulants* are generated by $\tilde{p}(\mathbf{k})$ and $\ln\tilde{p}(\mathbf{k})$ respectively, as

$$\begin{aligned} \langle x_1^{n_1}x_2^{n_2}\cdots x_N^{n_N}\rangle &= \left[\frac{\partial}{\partial(-ik_1)}\right]^{n_1}\left[\frac{\partial}{\partial(-ik_2)}\right]^{n_2}\cdots\left[\frac{\partial}{\partial(-ik_N)}\right]^{n_N}\tilde{p}(\mathbf{k}=\mathbf{0}) \quad , \\ \langle x_1^{n_1}x_2^{n_2}\cdots x_N^{n_N}\rangle_c &= \left[\frac{\partial}{\partial(-ik_1)}\right]^{n_1}\left[\frac{\partial}{\partial(-ik_2)}\right]^{n_2}\cdots\left[\frac{\partial}{\partial(-ik_N)}\right]^{n_N}\ln\tilde{p}(\mathbf{k}=\mathbf{0}) \quad . \end{aligned} \tag{II.39}$$

The previously described graphical relation between joint moments (all clusters of labelled points), and joint cumulant (connected clusters) is still applicable. For example

$$\begin{aligned} \langle x_\alpha x_\beta\rangle &= \langle x_\alpha\rangle_c\langle x_\beta\rangle_c + \langle x_\alpha x_\beta\rangle_c \quad , \quad \text{and} \\ \langle x_\alpha^2 x_\beta\rangle &= \langle x_\alpha\rangle_c^2\langle x_\beta\rangle_c + \langle x_\alpha^2\rangle_c\langle x_\beta\rangle_c + 2\langle x_\alpha x_\beta\rangle_c\langle x_\alpha\rangle_c + \langle x_\alpha^2 x_\beta\rangle_c \quad . \end{aligned} \tag{II.40}$$

The *connected correlation*, $\langle x_\alpha x_\beta\rangle_c$, is zero if x_α and x_β are independent random variables.

• *The joint Gaussian distribution* is the generalization of eq.(II.15) to N random variables, as

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N\det[C]}}\exp\left[-\frac{1}{2}\sum_{mn}\left(C^{-1}\right)_{mn}(x_m-\lambda_m)(x_n-\lambda_n)\right] \quad , \tag{II.41}$$

where C is a symmetric matrix, and C^{-1} is its inverse,. The simplest way to get the normalization factor is to make a linear transformation from the variables $y_j = x_j - \lambda_j$, using the unitary matrix that diagonalizes C. This reduces the normalization to that of the product of N Gaussians whose variances are determined by the eigenvalues of C. The product of the eigenvalues is the determinant $\det[C]$. (This also indicates that the matrix C must be positive definite.) The corresponding joint characteristic function is obtained by similar manipulations, and is given by

$$\tilde{p}(\mathbf{k}) = \exp\left[-ik_m\lambda_m - \frac{1}{2}C_{mn}k_mk_n\right] \quad , \tag{II.42}$$

where the summation convention is used.

The joint cumulants of the Gaussian are then obtained from $\ln \tilde{p}(\mathbf{k})$ as

$$\langle x_m \rangle_c = \lambda_m \quad , \quad \langle x_m x_n \rangle_c = C_{mn} \quad , \tag{II.43}$$

with all higher cumulants equal to zero. In the special case of $\{\lambda_m\} = 0$, all odd *moments* of the distribution are zero, while the general rules for relating moments to cumulants indicate that any even moment is obtained by summing over all ways of grouping the involved random variables into pairs, e.g.

$$\langle x_a x_b x_c x_d \rangle = C_{ab}C_{cd} + C_{ac}C_{bd} + C_{ad}C_{bc}. \tag{II.44}$$

This result is sometimes referred to as *Wick's theorem*.

II.E Sums of Random Variables & the Central Limit Theorem

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

$$p_X(x) = \int d^N\mathbf{x}\; p(\mathbf{x}) \delta\left(x - \sum x_i\right) = \int \prod_{i=1}^{N-1} dx_i \; p\left(x_1, \cdots, x_{N-1}, x - x_1 \cdots - x_{N-1}\right), \tag{II.45}$$

and the corresponding characteristic function (using eq.(II.38)) is given by

$$\tilde{p}_X(k) = \left\langle \exp\left(-ik \sum_{j=1}^{N} x_j\right) \right\rangle = \tilde{p}\left(k_1 = k_2 = \cdots = k_N = k\right). \tag{II.46}$$

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

$$\ln \tilde{p}\left(k_1 = k_2 = \cdots = k_N = k\right) = -ik \sum_{i_1=1}^{N} \langle x_{i_1} \rangle_c + \frac{(-ik)^2}{2} \sum_{i_1, i_2}^{N} \langle x_{i_1} x_{i_2} \rangle_c + \cdots, \tag{II.47}$$

as

$$\langle X \rangle_c = \sum_{i=1}^{N} \langle x_i \rangle_c \quad , \quad \langle X^2 \rangle_c = \sum_{i,j}^{N} \langle x_i x_j \rangle_c \quad , \quad \cdots. \tag{II.48}$$

If the random variables are independent, $p(\mathbf{x}) = \prod p_i(x_i)$, and $\tilde{p}_X(k) = \prod \tilde{p}_i(k)$. The cross–cumulants in eq.(II.48) vanish, and the n^{th} cumulant of X is simply the sum of the individual cumulants, $\langle X^n \rangle_c = \sum_{i=1}^{N} \langle x_i^n \rangle_c$. When all the N random variables

are independently taken from the same distribution $p(x)$, this implies $\langle X^n \rangle_c = N \langle x^n \rangle_c$, 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 $y = (X - N \langle x \rangle_c)/\sqrt{N}$, has zero mean, and cumulants that scale as $\langle y^n \rangle_c \propto N^{1-m/2}$. As $N \to \infty$, only the second cumulant survives, and the PDF for y converges to the normal distribution ,

$$\lim_{N\to\infty} p\left(y = \frac{\sum_{i=1}^{N} x_i - N \langle x \rangle_c}{\sqrt{N}} \right) = \frac{1}{\sqrt{2\pi \langle x^2 \rangle_c}} \exp\left(-\frac{y^2}{2 \langle x^2 \rangle_c} \right). \tag{II.49}$$

(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 a most important 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 $\sum_{i_1,\cdots,i_m}^{N} \langle x_{i_1} \cdots x_{i_m} \rangle_c \ll \mathcal{O}(N^{m/2})$, is sufficient for the validity of eq.(II.49).

II.F Rules for Large Numbers

To describe equilibrium properties of macroscopic bodies, statistical mechanics has to deal with the very large number N, of microscopic degrees of freedom. Actually, taking the *thermodynamic limit* of $N \to \infty$ leads to a number of simplifications, some of which are described in this section.

There are typically three types of N dependence encountered in the thermodynamic limit:

(a) *Intensive* quantities, such as temperature T, and generalized forces, e.g. pressure P, and magnetic field $\vec{B}$, are independent of N, i.e. $\mathcal{O}(N^0)$.

(b) *Extensive* quantities, such as energy E, entropy S, and generalized displacements, e.g. volume V, and magnetization $\vec{M}$, are proportional to N, i.e. $\mathcal{O}(N^1)$.

(c) *Exponential* dependence, i.e. $\mathcal{O}\big(\exp(N\phi)\big)$, is encountered in enumerating discrete micro-states, or computing available volumes in phase space.

Other asymptotic dependencies are certainly not ruled out a priori. For example, the Coulomb energy of N ions at fixed density scales as $Q^2/R \sim N^{5/3}$. Such dependencies are rarely encountered in every day physics. The Coulomb interaction of ions is quickly

screened by counter-ions, resulting in an extensive overall energy. (This is not the case in astrophysical problems since the gravitational energy can not be screened. For example the entropy of a black hole is proportional to the square of its mass.)

In statistical mechanics we frequently encounter sums or integrals of exponential variables. Performing such sums in the thermodynamic limit is considerably simplified due to the following results.

(1) *Summation of Exponential Quantities*

Consider the sum

$$\mathcal{S} = \sum_{i=1}^{\mathcal{N}} \mathcal{E}_i \quad , \tag{II.50}$$

where each term is positive, with an exponential dependence on N,

$$0 \leq \mathcal{E}_i \sim \mathcal{O}\big(\exp(N\phi_i)\big), \tag{II.51}$$

and the number of terms $\mathcal{N}$, is proportional to some power of N. Such a sum can be approximated by its largest term $\mathcal{E}_{\max}$, in the following sense. Since for each term in the sum, $0 \leq \mathcal{E}_i \leq \mathcal{E}_{\max}$,

$$\mathcal{E}_{\max} \leq \mathcal{S} \leq \mathcal{N}\mathcal{E}_{\max} \ . \tag{II.52}$$

An intensive quantity can be constructed from $\ln \mathcal{S}/N$, which is bounded by

$$\frac{\ln \mathcal{E}_{\max}}{N} \leq \frac{\ln \mathcal{S}}{N} \leq \frac{\ln \mathcal{E}_{\max}}{N} + \frac{\ln \mathcal{N}}{N} \ . \tag{II.53}$$

For $\mathcal{N} \propto N^p$, the ratio $\ln \mathcal{N}/N$ vanishes in the large N limit, and

$$\lim_{N\to\infty} \frac{\ln \mathcal{S}}{N} = \frac{\ln \mathcal{E}_{\max}}{N} = \phi_{\max} \quad . \tag{II.54}$$

(2) *Saddle Point Integration*

Similarly, an integral of the form

$$\mathcal{I} = \int dx \exp\big(N\phi(x)\big) \quad , \tag{II.55}$$

can be approximated by the maximum value of the integrand, obtained at a point $x_{\max}$ which maximizes the exponent $\phi(x)$. Expanding around this point,

$$\mathcal{I} = \int dx \exp\left\{ N\left[\phi(x_{\max}) - \frac{1}{2}|\phi''(x_{\max})|(x - x_{\max})^2 + \cdots \right] \right\}. \tag{II.56}$$

Note that at the maximum, the first derivative $\phi'(x_{\max})$, is zero, while the second derivative $\phi''(x_{\max})$, is negative. Terminating the series at the quadratic order results in

$$\mathcal{I} \approx e^{N\phi(x_{\max})} \int dx \exp\left[-\frac{N}{2}|\phi''(x_{\max})|(x-x_{\max})^2\right] \approx \sqrt{\frac{2\pi}{N|\phi''(x_{\max})|}} e^{N\phi(x_{\max})}, \quad \text{(II.57)}$$

where the range of integration has been extended to $[-\infty, \infty]$. The latter is justified since the integrand is negligibly small outside the neighborhood of $x_{\max}$.

There are two types of correction to the above result. Firstly, there are higher order terms in the expansion of $\phi(x)$ around $x_{\max}$. These corrections can be looked at perturbatively, and lead to a series in powers of $1/N$. Secondly, there may be additional local maxima for the function. A maximum at $x'_{\max}$, leads to a similar Gaussian integral that can be added to eq.(II.57). Clearly such contributions are smaller by $\mathcal{O}\big(\exp\{-N[\phi(x_{\max})-\phi(x'_{\max})]\}\big)$. Since all these corrections vanish in the thermodynamic limit,

$$\lim_{N\to\infty} \frac{\ln \mathcal{I}}{N} = \lim_{N\to\infty} \left[\phi(x_{\max}) - \frac{1}{2N}\ln\big(\frac{N|\phi''(x_{\max})|}{2\pi}\big) + \mathcal{O}(\frac{1}{N^2})\right] = \phi(x_{\max}) \quad . \quad \text{(II.58)}$$

The *saddle point* method for evaluating integrals is the extension of the above result to more general integrands, and integration paths in the complex plane. (The appropriate extremum in the complex plane is a saddle point.) The simplified version presented above is sufficient for the purposes of this course.

• *Stirling's approximation* for $N!$ at large N can be obtained by saddle point integration. In order to get an integral representation of $N!$, start with the result

$$\int_0^\infty dx e^{-\alpha x} = \frac{1}{\alpha}. \quad \text{(II.59)}$$

Repeated differentiation of both sides of the above equation with respect to α leads to

$$\int_0^\infty dx x^N e^{-\alpha x} = \frac{N!}{\alpha^{N+1}}. \quad \text{(II.60)}$$

Although the above result only applies to integer N, it is possible to define by analytical continuation a function,

$$\Gamma(N+1) \equiv N! = \int_0^\infty dx x^N e^{-x}, \quad \text{(II.61)}$$

for all N. While the integral in eq.(II.61) is not exactly in the form of eq.(II.55), it can still be evaluated by a similar method. The integrand can be written as $\exp\big(N\phi(x)\big)$, with $\phi(x) = \ln x - x/N$. The exponent has a maximum at $x_{\max} = N$, with $\phi(x_{\max}) = \ln N - 1$, and $\phi''(x_{\max}) = -1/N^2$. Expanding the integrand in eq.(II.61) around this point yields,

$$N! \approx \int dx \exp\big(N \ln N - N - \frac{1}{2N}(x-N)^2\big) \approx N^N e^{-N}\sqrt{2\pi N}, \tag{II.62}$$

where the integral is evaluated by extending its limits to $[-\infty, \infty]$. Stirling's formula is obtained by taking the logarithm of eq.(II.62) as,

$$\ln N! = N \ln N - N + \frac{1}{2}\ln(2\pi N) + \mathcal{O}(\frac{1}{N}). \tag{II.63}$$

II.G Information, Entropy, and Estimation

• *Information:* Consider a random variable with a discrete set of outcomes $\mathcal{S} = \{x_i\}$, occurring with probabilities $\{p(i)\}$, for $i = 1, \cdots, M$. In the context of information theory, there is a precise meaning to the *information content* of a probability distribution: Let us construct a message from N independent outcomes of the random variable. Since there are M possibilities for each character in this message, it has an apparent information content of $N \ln_2 M$ bits; i.e. this many binary bits of information have to be transmitted to convey the message precisely. On the other hand, the probabilities $\{p(i)\}$ limit the types of messages that are likely. For example, if $p_2 \gg p_1$, it is very unlikely to construct a message with more x_1 than x_2. In particular, in the limit of large N, we expect the message to contain "roughly" $\{N_i = Np_i\}$ occurrences of each symbol.† The number of typical messages thus corresponds to the number of ways of rearranging the $\{N_i\}$ occurrences of $\{x_i\}$, and is given by the multinomial coefficient

$$g = \frac{N!}{\prod_{i=1}^{M} N_i!}. \tag{II.64}$$

This is much smaller than the total number of messages M^n. To specify one out of g possible sequences requires

$$\ln_2 g \approx -N \sum_{i=1}^{M} p_i \ln_2 p_i \qquad (\text{for } N \to \infty), \tag{II.65}$$

† More precisely, the probability of finding any N_i that is different from Np_i by more than $\pm\sqrt{N}$ becomes exponentially small in N, as $N \to \infty$.

bits of information. The last result is obtained by applying Stirling's approximation for $\ln N!$. It can also be obtained by noting that

$$1 = \left(\sum_i p_i \right)^N = \sum_{\{N_i\}} N! \prod_{i=1}^{M} \frac{p_i^{N_i}}{N_i!} \approx g \prod_{i=1}^{M} p_i^{N p_i}, \tag{II.66}$$

where the sum has been replaced by its largest term, as justified in the previous section.

Shannon's Theorem proves more rigorously that the minimum number of bits necessary to ensure that the percentage of errors in N trials vanishes in the $N \to \infty$ limit, is $\ln_2 g$. For any non-uniform distribution, this is less than the $N \ln_2 M$ bits needed in the absence of any information on relative probabilities. The difference per trial is thus attributed to the information content of the probability distribution, and is given by

$$I\left[\{p_i\}\right] = \ln_2 M + \sum_{i=1}^{M} p_i \ln_2 p_i \quad . \tag{II.67}$$

• *Entropy:* Eq.(II.64) is encountered frequently in statistical mechanics in the context of mixing M distinct components; its natural logarithm is related to the *entropy of mixing.* More generally, we can define an *entropy* for *any probability distribution* as

$$S = -\sum_{i=1}^{M} p(i) \ln p(i) = -\left\langle \ln p(i) \right\rangle \quad . \tag{II.68}$$

The above entropy takes a minimum value of zero for the delta–function distribution $p(i) = \delta_{i,j}$, and a maximum value of $\ln M$ for the uniform distribution, $p(i) = 1/M$. S is thus a measure of dispersity (disorder) of the distribution, and does not depend on the values of the random variables $\{x_i\}$. A one to one mapping to $f_i = F(x_i)$ leaves the entropy unchanged, while a many to one mapping makes the distribution more ordered and decrease S. For example, if the two values, x_1 and x_2, are mapped onto the same f, the change in entropy is

$$\Delta S(x_1, x_2 \to f) = \left[p_1 \ln \frac{p_1}{p_1 + p_2} + p_2 \ln \frac{p_2}{p_1 + p_2} \right] < 0. \tag{II.69}$$

• *Estimation:* The entropy S, can also be used to quantify subjective estimates of probabilities. In the absence of any information, the best *unbiased estimate* is that all M outcomes are equally likely. This is the distribution of maximum entropy. If additional information is available, the unbiased estimate is obtained by maximizing the entropy subject to the

constraints imposed by this information. For example, if it is known that $\langle F(x)\rangle = f$, we can maximize

$$S\left(\alpha,\beta,\{p_i\}\right) = -\sum_i p(i)\ln p(i) - \alpha\left(\sum_i p(i) - 1\right) - \beta\left(\sum_i p(i)F(x_i) - f\right), \quad \text{(II.70)}$$

where the Lagrange multipliers α and β are introduced to impose the constraints of normalization, and $\langle F(x)\rangle = f$, respectively. The result of the optimization is a distribution $p_i \propto \exp\big(-\beta F(x_i)\big)$, where the value of β is fixed by the constraint. This process can be generalized to an arbitrary number of conditions. It is easy to see that if the first n moments (and hence n cumulants) of a distribution are specified, the unbiased estimate is the exponential of an n^{th} order polynomial.

In analogy with eq.(II.68), we can define an entropy for a continuous random variable ($\mathcal{S}_x = \{-\infty < x < \infty\}$) as

$$S = -\int dx\, p(x)\ \ln p(x) = -\left\langle \ln p(x)\right\rangle \quad . \quad \text{(II.71)}$$

There are, however, problems with this definition, as for example S is not invariant under a one to one mapping. (After a change of variable to $f = F(x)$, the entropy is changed by $\langle |F'(x)|\rangle$.) Since the Jacobian of a canonical transformation is unity, canonically conjugate pairs offer a suitable choice of coordinates in classical statistical mechanics. The ambiguities are also removed if the continuous variable is discretized. This happens quite naturally in quantum statistical mechanics where it is usually possible to work with a discrete ladder of states. The appropriate volume for discretization of phase space is set by Planck's constant $\hbar$.