

# Chapter 2

## Review of Probability and Statistics

### 2.1 Basic Definitions

Probability is one of those familiar concepts that turn out to be difficult to define formally. The commonly accepted definition, is the axiomatic one due to Kolmogorov, that provides the minimal set of properties that a probability must satisfy, but does not say anything about what it represents. In fact, we know perfectly well what probability should represent: a measure of our expectation of the outcomes of a given experiment. Unfortunately, this is not enough to build a theory (for instance, how do we define the concept of expectation?).

The axiomatic definition of probability is obtained from elementary concepts in the following way.

- Let us call the set  $\Omega$  of the possible outcomes of a given experiment the **sample space** for that experiment. Examples: the set  $\Omega = \{1, 2, 3, 4, 5, 6\}$  of the results of a dice roll; the set  $\Omega = \mathbf{R}^3$  of the possible velocities of a molecule in a gas. A quantity, such as the velocity of the molecule, whose sample space is continuous, is called a **random variable**.
- A set  $A$  of possible results of an experiment is called an **event**. An event is by definition a subset of  $\Omega$ ; it must be mentioned, however, that in the case of a continuous sample space, not all the subsets  $A \subset \Omega$  are admissible; we shall not discuss further this topic. See e.g. [W. Feller, *An introduction to probability theory and its applications* (Wiley and Sons, 1968), Vol. II, Sect. I.11]. Two events  $A$  and  $B$  are said to be **mutually exclusive** if their intersection is empty  $A \cap B = \emptyset$ .
- The **probability** of an event  $a$  is a number  $P(A)$ ,  $0 \leq P(A) \leq 1$ , with the properties  $P(A \cup B) = P(A) + P(B)$ , in the case  $A$  and  $B$  are mutually exclusive, and  $P(\Omega) = 1$ .

Not bad with regard to economy, for a definition. It can be verified that the two axioms lead to a probability that has all the properties suggested by our intuition (e.g. that  $P(\emptyset) = 0$ ). Notice that the expression  $P(A \cup B)$  is nothing but a stenographic notation

to indicate the probability that the experiment produces an outcome belonging to either  $A$  or  $B$ .

In the case of a random variable  $x$ , events can be obtained, in general, as union and intersection of intervals of  $\mathbf{R}$ . We can then define a **probability density function** (PDF)  $\rho(x)$  through the formula

$$P([a, b]) = \int_a^b \rho(x) dx. \quad (2.1)$$

Once a certain system of units is selected, a physical event will correspond, typically, to an interval of values of the physical quantity expressed in those units. In other units, the corresponding interval would of course be different. The PDF of the same physical quantity in different units will be different as well. For instance, a physical event  $A \equiv [a, b]$  in the variable  $x$ , will transform into  $[2a, 2b]$ , in the variable  $y = 2x$ . From Eq. (2.1), the probability  $P(A)$  will be expressed therefore in terms of the PDF for the variables  $x$  and  $y$ :

$$P(A) = \int_a^b \rho_x(z) dz = \int_{2a}^{2b} \rho_y(z) dz \quad (2.2)$$

(subscript  $x$  and  $y$  label the random variables;  $z$  are their values). Considering infinitesimal intervals, would give us in Eq. (2.2):  $\rho_x(z) dz = 2\rho_y(2z) dz$ , from which we obtain the formula for the change of variables in PDF's:

$$\rho_x(x) = \rho_y(y(x)) \left| \frac{dy(x)}{dx} \right|. \quad (2.3)$$

We can associate probabilities to isolated points in a continuous sample space. A non-zero probability for such events, will produce a Dirac-delta component in the corresponding PDF. For instance:

$$\rho(x) = a\delta(x) + b\theta(x) \exp(-x), \quad x \in \mathbf{R}, \quad (2.4)$$

and we have of course  $P(0) = a$ . In order to have  $P(\Omega) = 1$ , we must impose  $b = 1 - a$ .

### 2.1.1 Joint, Conditional and Marginal Probabilities

Given two events  $A$  and  $B$ , we call the **joint probability** of  $A$  and  $B$ , the probability of the intersection  $P(A \cap B)$ . Important example: in the case  $A$  and  $B$  are, respectively, a vertical and a horizontal slice of a domain  $\Omega \subset \mathbf{R}^2$ ,  $A = [x, x + dx] \otimes \Omega_y(x)$ , and  $B = \Omega_x(y) \otimes [y, y + dy]$ , with  $\Omega_y(x)$  the set of the  $y$  contained in  $\Omega$  at horizontal position  $x$ , and  $\Omega_x(y)$  the set of the  $x$  contained in  $\Omega$  at vertical position  $y$ . We define

the **joint PDF**  $\rho_{xy}(x, y)$  of the values  $x$  and  $y$  of the random variables  $x$  and  $y$ , from the relation

$$P(A \cap B) = P([x, x + dx] \otimes [y, y + dy]) := \rho_{xy}(x, y) dx dy. \quad (2.5)$$

(In the following we shall put subscripts on the PDF's only when strictly necessary). Of course, we can define objects that are half-way between probability and PDF (e.g. integrating  $\rho(x, y)$  with respect to  $x$  in an interval  $A$ ; thus  $\rho(A, y)$  is a probability with respect to  $A$ , and a PDF with respect to  $y$ ).

**Note** The formula for the change of variable Eq. (2.3) can be generalized to higher dimension. If  $\mathbf{y} = \mathbf{y}(\mathbf{x})$ :  $\rho_{\mathbf{x}}(\mathbf{x}) = \rho_{\mathbf{y}}(\mathbf{y}(\mathbf{x})) |J_{\mathbf{y}}(\mathbf{x})|^{-1}$ , where  $J_{\mathbf{y}}(\mathbf{x}) = \det(\partial_{\mathbf{x}} \mathbf{y}(\mathbf{x}))$  is the Jacobian determinant of the transformation. Notice that we use here the vector notation  $\mathbf{x} = (x_1, x_2, \dots)$  and  $\mathbf{y} = (y_1, y_2, \dots)$  to indicate two sets of vector components that represent the same vector. For instance,  $\mathbf{x}$  and  $\mathbf{y}$  may represent polar and cartesian coordinates in the plane,  $\mathbf{x} = (r, \phi)$ ,  $\mathbf{y}(\mathbf{x}) = (r \cos(\phi), r \sin(\phi))$ , so that  $\rho_{r,\phi}(r, \phi) = r \rho_{y_1, y_2}(r \cos(\phi), r \sin(\phi))$ . ■

We introduce next the concept of **conditional probability** of an event  $A$  conditional to an event  $B \neq \emptyset$ :

$$P(A|B) = P(A \cap B)/P(B). \quad (2.6)$$

The conditional probability  $P(A|B)$  can thus be seen as the probability of  $A$  obtained considering as sample space  $B$ . We can generalize again to PDF's:

$$\rho(x|y) = \rho(x, y)/\rho(y); \quad P([x, x + dx]|y) = \rho(x|y) dx. \quad (2.7)$$

(prove this last result from the definition Eq. (2.5)).

Two events are said to be **statistically independent**, if  $P(A \cap B) = P(A)P(B)$ . The concept of conditional probability allows us to get an intuitive grasp of this definition. If  $A$  and  $B$  are statistically independent (and are not empty), we have in fact

$$P(A|B) = P(A); \quad P(B|A) = P(B). \quad (2.8)$$

In other words, conditioning by a statistically independent event, does not modify the probability. Similarly, in the case of a PDF: the condition  $\rho(x|y) = \rho(x)$  tells us that the occurrence of  $x$ , as outcome of an experiment, is independent of the fact that, simultaneously,  $y$  takes a particular value. If  $\rho(x, y) = \rho(x)\rho(y) \forall x, y$ , we shall say that the random variables  $x$  and  $y$  are statistically independent.

At last, we notice the following relation, that descends directly from the definition Eq. (2.5):

$$\rho(x) = \int_{\Omega_y(x)} \rho(x, y) dy; \quad \rho(y) = \int_{\Omega_x(y)} \rho(x, y) dx. \quad (2.9)$$

In this context, the two PDF's  $\rho(x)$  and  $\rho(y)$  are called **marginal**.

### 2.1.2 The Concept of Average

We define the average of a function  $f$  of a random variable  $x$ , through the relation:

$$\langle f \rangle = \int \rho(x)f(x)dx, \quad (2.10)$$

which, in the case of a discrete random variable, becomes, using Eq. (2.4):

$$\langle f \rangle = \sum_i P(x_i)f(x_i).$$

We can take averages, starting from a conditional PDF, and the result is called a **conditional average**:

$$\langle f|y \rangle = \int \rho(x)f(x|y)dx, \quad (2.11)$$

which, being  $y$  still random, will be itself a random quantity.

A little review of notation and terminology. Typically,  $\mu_x \equiv \langle x \rangle$  is called the mean of the PDF, and  $\sigma^2 = \langle (x - \mu_x)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2$ , the variance of the PDF ( $\sigma_x$  is called standard deviation or RMS of  $x$ ). The average  $\langle x^n \rangle$  is called  $n$ th **moment** of the PDF. The quantity  $\langle fg \rangle$  is called **correlation** of  $f$  and  $g$ . Important case:

$$C_{xy} = \langle xy \rangle = \int \rho(x, y)xydx dy. \quad (2.12)$$

Notice that statistical independence of  $x$  and  $y$  implies  $C_{xy} = 0$ , but the vice versa is not true (find a counterexample).

An interesting object, of which to take averages, is the **indicator function** of an event  $A$ :

$$\delta_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases} \quad (2.13)$$

We find:

$$\langle \delta_A \rangle = \int_A \rho(x)dx = P(A). \quad (2.14)$$

The indicator function of a random variable  $x$ , is just the Dirac delta:

$$\delta_{\bar{x}}(x) \equiv \delta(x - \bar{x}) \Rightarrow \langle \delta_{\bar{x}} \rangle = \rho(\bar{x}). \quad (2.15)$$

Equations (2.13–2.14), and in some sense also Eq. (2.15), are the starting point for the statistical determination of probabilities (and PDF's) from the concept of frequency.

### 2.1.3 Characteristic Function

Another quantity of which to take averages, is the exponential of a random variable: the so called **characteristic function**

$$Z(j) = \langle \exp(ijx) \rangle = \int dx \rho(x) \exp(ijx), \quad (2.16)$$

that is the Fourier transform of the PDF  $\rho(x)$ .

We notice a number of important properties.

The PDF  $\rho(x)$  and the characteristic function  $Z(j)$  are, respectively, the average of the indicator function  $\delta_x$ , and of the exponential  $\exp(ijx)$ , where the second can be written trivially as the Fourier transform of the Dirac delta:

$$\exp(ijx) = \int dx' \delta_x(x') \exp(ijx').$$

We can invert the Fourier transform in Eq. (2.16), to obtain again the PDF:

$$\rho(x) = \int \frac{dj}{2\pi} Z(j) \exp(-ijx),$$

which corresponds to taking the average of the Fourier integral representation of the Dirac delta:

$$\delta_x(x') \equiv \delta(x - x') = \int \frac{dj}{2\pi} \exp(ij(x' - x)).$$

The most important property of the characteristic function, however, is that its derivatives at  $j \rightarrow 0$ , are just the moments of  $\rho$  (provided they exist):

$$\lim_{j \rightarrow 0} \frac{d^n Z(j)}{dj^n} = \lim_{j \rightarrow 0} \int dx (ix)^n \rho(x) \exp(ijx) = i^n \langle x^n \rangle \quad (2.17)$$

(notice that, from normalization of  $\rho$ , we shall always have  $Z(0) = 1$ ). Existence of all the moments of a PDF, will imply analyticity of  $Z(j)$  in  $j = 0$ , with the moments of  $\rho$  providing the coefficients of the Taylor expansion of  $Z$ .

## 2.2 Entropy and Information

We have seen that the probability of an event, parameterizes our expectation that particular event occurs or not. A question that is worth to ask is how much information on the outcome of an experiment, is gained by knowing the distribution of its possible results. It is rather clear that, given a coin toss experiment, we know more about its outcome if the probability distribution is  $P(head) = 1$ ,  $P(tail) = 0$ , rather than

if  $P(head) = P(tail) = 1/2$ . Shannon found a natural way to parameterize this information content, introducing an entropy, that is neither that of thermodynamics, nor that of statistical mechanics, but that has a lot to do with both. As we shall see, the entropy introduced by Shannon, does not parameterize the content of information of a distribution, instead, that of “ignorance”, i.e. the amount of additional information, that should be provided to know with certainty the result of an experiment.

We point out that the concept of information depends a lot on the kind of experiment. In the case of the coin toss experiment, the question may seem obvious; not so in the case of a random variable. The fact is that we cannot measure a random variable with infinite precision; what we are able to “measure” are only events. We need in some way to discretize  $\Omega$ . We define a **partition** of  $\Omega$ , as a collection of mutually exclusive events  $A_i$ , such that their union is  $\Omega$  itself (see Fig. 2.1):

$$\mathcal{P} = \{A_i, i = 1, \dots; A_i \cap A_j = \emptyset \text{ if } i \neq j; \cup_i A_i = \Omega\}. \quad (2.18)$$

We define the **Shannon entropy** of the distribution  $P$ , referred to the partition  $\mathcal{P}$  of  $\Omega$ , as:

$$S[P] = - \sum_i P(A_i) \ln P(A_i) = -\langle \ln P \rangle, \quad (2.19)$$

where the notation  $S[P]$  is a shorthand for  $S(\{P_i, i = 1, \dots\})$ . We see at once that the definition works smoothly in the case of the coin toss experiment, in which, trivially,  $\mathcal{P} \equiv \{head, tail\}$ . If  $P(head) = 1$  (and  $P(tail) = 0$ ) or  $P(tail) = 1$  (and  $P(head) = 0$ ), as in the case of a loaded coin, we will find in fact  $S[P] = 0$  (we know the result: minimum ignorance, implying minimum entropy, as  $0 \leq P \leq 1 \Rightarrow \ln P \geq 0 \Rightarrow S[P] \geq 0$ ). On the other extreme,  $P(head) = P(tail) = 1/2$ , we have instead  $S[P] = \ln 2$  (maximum ignorance).

Actually, things work *very* well.

The key point is that entropy is in the form

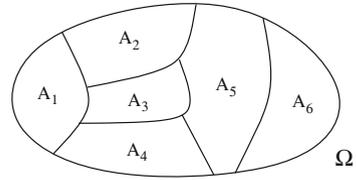
$$S[P] = \sum_i g(P_i),$$

with  $g(P)$  a **convex** function of the argument. In other words, the following relation is satisfied:

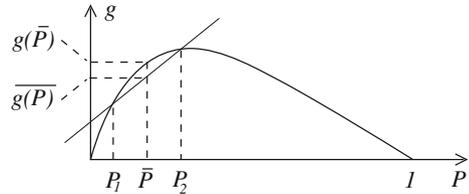
$$\overline{g(P_i)} \leq g(\overline{P}), \quad (2.20)$$

where  $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$  is the arithmetic mean of the  $x_i$ 's. The situation is illustrated in Fig. 2.2, in the case of two events  $A_1$  and  $A_2$  with probabilities respectively  $P_1$  and  $P_2$ . Their contribution to the entropy is  $g(P_1) + g(P_2) = 2g(\overline{P})$ . If these probabilities were equal, all the others,  $P_i, i = 3, \dots$ , maintaining their original values, we would have  $P(A_1) = P(A_2) = (P_1 + P_2)/2 = \overline{P}$ . In that case, the contribution to entropy would be  $2g(\overline{P}) \geq 2g(\overline{P})$ . From here, we can conclude that the entropy  $S[P]$  will become maximum, when all the events in the partition  $\mathcal{P}$  are equiprobable.

**Fig. 2.1** Partition of a two-dimensional domain  $\Omega$  (sample space)



**Fig. 2.2** Graphic representation of Eq. (2.20) in the case  $N = 2$



The information that must be supplied will be larger, again, if the number of outcomes of the experiment increases. This is what happens if we refine the partition, e.g. if we split event  $A_6$  in Fig. 2.1, into two disjoint events  $A_{61}$  and  $A_{62}$ . We see that, also in this case, entropy behaves as expected, increasing in response to refinement. In fact, the contribution to entropy, from  $A_{61}$  and  $A_{62}$ , will remain the same as before, only if  $P(A_{61}) = P(A_6)$  and  $P(A_{62}) = 0$ , or vice versa. Otherwise, due to convexity of  $g$ :  $g(A_6) < g(A_{61}) + g(A_{62})$ .

Another property that we should expect from entropy, if it is really a measure of an information content, is that its value, referred to two independent systems, should be the sum of the entropies referred to each system individually. Thus, if systems  $A$  and  $B$  are described by physical variables, with domains  $\Omega_A$  and  $\Omega_B$ , and all the elements of the partitions  $\mathcal{P}_A = \{A_1, A_2, \dots\}$ ,  $\mathcal{P}_B = \{B_1, B_2, \dots\}$ , are statistically independent,  $P(A_i, B_j) = P_A(A_i)P_B(B_j)$ , we should have:  $S[P] = S[P_A] + S[P_B]$ . In fact, from Eq. (2.19):

$$\begin{aligned}
 S[P] &= - \sum_{ij} P(A_i, B_j) \ln(P(A_i, B_j)) \\
 &= - \sum_{ij} [P_A(A_i)P_B(B_j) \ln P_B(B_j) + P_B(B_j)P_A(A_i) \ln P_A(A_i)] \\
 &= - \sum_i [P_B(B_i) \ln P_B(B_i) + P_A(A_i) \ln P_A(A_i)] = S[P_B] + S[P_A]. \quad (2.21)
 \end{aligned}$$

We see that the logarithm is important.

On the backward route from Shannon entropy to thermodynamic entropy, we shall see that the additivity of the entropy, is a somewhat reasonable requirement, in the case of weakly interacting system, exactly as additivity of the internal energy. Convexity lies in some sense at a deeper level, and is connected with the

possibility of defining a concept of thermodynamic equilibrium. A question that has been considered over the years, is whether alternative definitions of entropies, “without logarithms”, are possible, and could lead to a self-consistent thermodynamic theory of strongly interacting systems.

Coming back to the original definition of entropy Eq. (2.19), we may ask whether we could associate an entropy with a PDF, rather than with a probability distribution over a partition of  $\Omega$ . The answer is trivially yes, once we indicate, with  $\Delta x(x)$ , the partition element, in which the value of the random variable  $x$  is located. From Eqs. (2.1) and (2.19):

$$S[P] = - \int dx \rho(x) \ln[\rho(x)\Delta x(x)],$$

which, in the case of a uniform partition  $\Delta x(x) = \Delta x$ , becomes

$$S[P] = - \int dx \rho(x) \ln \rho(x) - \ln \Delta x. \quad (2.22)$$

The entropy, in the case of a uniform partition, coincides therefore, to within a constant, with the “PDF entropy”  $S_\rho = - \int dx \rho(x) \ln \rho(x)$ . As we shall see in the following, the uniform partition is the natural choice, in most situations of interest for thermodynamics systems.

### 2.3 Sums of Random Variables: The Thermodynamic Limit

Quantities such as a random variable average, or the probability of an event, could be estimated as limits over a large number of trials of the corresponding sample average and frequencies. In fact, there exists an empirical interpretation of probability theory, in which probabilities are understood as limits of frequencies. The sample average of a random variable  $y$  is simply the arithmetic mean of the results of the measurement of the quantity  $y$  in a sequence of repeated identical experiments:

$$\langle y \rangle_N = \frac{1}{N} \sum_{k=1}^N y_k. \quad (2.23)$$

Similarly, the frequency of an event  $A$  can be expressed as the sample average of the indicator function, defined in Eq. (2.13):

$$\tilde{P}_N(A) = \langle \delta_A \rangle_N. \quad (2.24)$$

Of course, the sample average  $\langle y \rangle_N$  is still a random variable, as its value will change, from a sequence of  $N$  experiments to the next. If the outcomes of different experiments can be considered as statistical independent, and if the PDF  $\rho(y)$  has finite first and

second moment, the **law of large number** will hold:

$$\langle \langle y \rangle_N \rangle = \mu_y, \quad \sigma_{\langle y \rangle_N}^2 = \frac{\sigma_y^2}{N}, \quad (2.25)$$

which expresses, formally, the notion that averages can be estimated as limits of sample averages. In the same hypotheses, the **central limit theorem** will hold, which tells us that the deviations  $\langle y \rangle_N - \mu_y$  are normally distributed.

The link between statistics and thermodynamics lies in the fact that the macroscopic state of a system is described by variables that are, typically, sums of contributions by microscopic constituents:

$$\tilde{Y}_N = \sum_{k=1}^N y_k, \quad (2.26)$$

where  $N$  could be of the order of the Avogadro number  $N_A \simeq 6 \cdot 10^{23} \text{ mol}^{-1}$ . The quantity  $y_k$  may be e.g. the kinetic energy of the molecules in an ideal gas, and  $\tilde{Y}_N$  the internal energy of the gas. The fact is that, in many instances (e.g. an ideal gas at thermodynamic equilibrium), the contributions from the microscopic constituents can be considered as independent identically distributed (i.i.d.) random variables. The macroscopic variable  $Y_N$  will then satisfy a relation in the form of the law of large numbers Eq. (2.25):

$$\mu_{\tilde{Y}_N} = N\mu_y; \quad \sigma_{\tilde{Y}_N} = N^{1/2}\sigma_y. \quad (2.27)$$

Let us check this result. The relation  $\mu_{Y_N} = N\mu_y$  descends immediately from linearity of the average operation. To obtain the second of Eq. (2.27), we write

$$\sigma_{\tilde{Y}_N}^2 = \langle [\sum_{i=1}^N (y_i - \mu_y)]^2 \rangle = \sum_{ij=1}^N \langle (y_i - \mu_y)(y_j - \mu_y) \rangle, \quad (2.28)$$

and observe that, since the different  $y_i$ 's are statistically independent, we must have, for  $j \neq i$   $\langle (y_i - \mu_y)(y_j - \mu_y) \rangle = \langle (y_i - \mu_y) \rangle \langle (y_j - \mu_y) \rangle = 0$ , and hence:

$$\sigma_{\tilde{Y}_N}^2 = \sum_{i=1}^N \langle (y_i - \mu_y)^2 \rangle = N\sigma_y^2.$$

Equation (2.27) leads us to identify three relevant scales for the variable  $Y$ :

- A microscopic scale, fixed by  $\mu_y$  and  $\sigma_y$ .
- A fluctuation scale  $\sigma_Y = N^{1/2}\sigma_y$ .
- A macroscopic scale, at which the condition of thermodynamic limit is considered satisfied. For  $\mu_y \neq 0$ , this macroscopic scale can be identified simply with  $\mu_{Y_N} = N\mu_y$ .

We see that, for  $N \rightarrow \infty$  (and provided  $\mu_y \neq 0$ ), fluctuations will become negligible on the scale of the mean  $\mu_{\tilde{Y}_N} = N\mu_y$ . The distribution  $\rho(\tilde{Y}_N)$  will then become so narrow (on the scale of interest), that we can identify the instantaneous value of the variable  $\tilde{Y}_N$ , with the most probable value  $Y_N$ , that coincides in turn with the mean  $\mu_{\tilde{Y}_N}$ . This condition is commonly referred to, as a **thermodynamic limit** for the variable  $Y$ .

**Note** We can apply the concepts developed so far, to evaluate the density fluctuations in a gas in thermodynamic equilibrium. No external forces are present, so that the gas is spatially homogeneous. To fix the ideas, imagine that the gas is contained in a volume  $V$ , and indicate with  $N$  the total number of molecules. We choose, as macroscopic variable, the coarse grained density  $\tilde{n}_{V_a}(\mathbf{x}_a, t) = \tilde{N}_a/V_a$ , where  $\tilde{N}_a$  is the instantaneous molecule count in the volume  $V_a$  centered around  $\mathbf{x}_a$ . Indicating with  $\mathbf{x}_i(t)$  the instantaneous position of the  $i$ th molecule:

$$\tilde{n}_{V_a}(\mathbf{x}_a, t) = \frac{1}{V_a} \sum_{i=1}^N \delta_{V_a}(\mathbf{x}_i(t)).$$

From the property of the indicator function, Eq. (2.14):  $\langle \delta_{V_a} \rangle = P(V_a)$ , the probability to find the generic molecule in  $V_a$ . For a homogeneous gas,  $P(V_a) = V_a/V$  and we find the obvious result, that the local density of the gas coincides with the mean density  $\bar{n} = N/V$ :

$$n_{V_a}(\mathbf{x}_a, t) = \bar{n}.$$

To calculate the fluctuation, we must determine  $\sigma_{\delta_{V_a}}^2$ . We have

$$\langle \delta_{V_a}^2 \rangle = \langle \delta_{V_a} \rangle = V_a/V,$$

and therefore

$$\sigma_{\delta_{V_a}}^2 = \langle \delta_{V_a}^2 \rangle - \langle \delta_{V_a} \rangle^2 = V_a/V(1 - V_a/V).$$

For  $V_a$  small, we can disregard the  $(V_a/V)^2$ , and we find for the fluctuation amplitude:

$$\sigma_{\tilde{N}_a}^2 = N\sigma_{\delta_{V_a}}^2 \simeq N_a \Rightarrow \frac{\delta N_a}{N_a} = O(N_a^{-1/2}). \quad (2.29)$$

The fluctuation density, in a volume  $V_a$ , will be  $\delta n_{V_a} = \delta N_a/V_a \sim (V_a \bar{n})^{-1/2} \bar{n}$ . ■

It is worth stressing the physical content of the conditions of i.i.d. statistics, and existence of the first moments,  $\mu_y$  and  $\sigma_y^2$ , in the derivation of Eq. (2.27). We see that these conditions express, in different ways, the need of a separation between microscopic and macroscopic scales.

Let us analyze first the condition of finiteness of the lowest order moments. As discussed in the appendix, if  $\mu_y$  were infinite, the ratio  $Y_N/N \equiv \langle y \rangle_N$  would grow indefinitely, as larger and larger values of  $y$  are picked up by the molecules in the sample. (Similar arguments would apply as regards  $\sigma_y^2$  and  $\sigma_{Y_N}^2$ ). Contrary to the finite  $\mu_y$  case, in which all molecules contribute the same to  $Y_N$ , in the infinite  $\mu_y$  case, only a few molecules would contribute to  $Y_N$ , with  $y_k \sim Y_N$ . The separation of scale between microscopic contributions and macroscopic world, would clearly be broken.

Let us pass to the condition of independence of the microscopic contributions. We see that this condition is essential in order for  $\sigma_{Y_N}$  to grow sublinearly in  $N$  and become negligible, for  $N$  large, compared with  $\mu_{Y_N}$ . If, say, all molecules were equally correlated, we would have a quadratic  $\sigma_Y^2$ , and, in order to neglect fluctuations, we would have to be sure from the start that  $\sigma_y \ll \mu_y$ . The typical situation is that the molecules are correlated below a certain scale  $\lambda$ , while, at larger separations, they are basically independent. This would mean that, in place of a set of  $N$  independent molecules, we would have a bunch of independent molecule clumps, each of size  $\sim \lambda$ . In order for a thermodynamic limit argument, along the line of the one leading from Eq. (2.28) to Eq. (2.27), we need that the number of clumps is large. This requires, in turn, that  $\lambda$  is a microscopic quantity. We shall see in Sect. 3.3 that the condition of independence of the molecules at macroscopic scales, is basically a condition of fluctuations in the system being confined to microscopic scales.

The condition that the different  $y_i$ 's are identically distributed, is just a different way of saying that the system must be macroscopic. An obvious condition for this is in fact that the spatial scale of variation of  $Y_N$  is itself macroscopic, which in turn guarantees the existence of portions of the system sufficiently large to be considered macroscopic, but not too much for the microscopic statistics inside, to be spatially inhomogeneous. An example better illustrates the situation. Think again  $Y_N$  as the density of a gas, coarse-grained at scale  $V_a$  (see note above), but relax the hypothesis of spatial homogeneity. It is clear that, to have a local meaning, the volume  $V_a$  must be smaller than the scale of variation for the density. At the same time, however,  $V_a$  must be large and the number of molecules in it to be large and fluctuations be negligible.

We conclude with the observation that there are systems that are better described by quantities that are not simple sums of microscopic contributions. An example is the interaction component of the internal energy, in a system of interacting particle. In the case of binary interactions:  $U = \sum_{i>j} U_{ij}$ , with  $U_{ij}$  the interaction energy of molecules  $i$  and  $j$ . In analogy with the problem of correlations, the situation can be brought back to that of an additive variable, in the form of Eq. (2.26), provided the interaction is short-ranged, meaning that the interaction length is microscopic. We would then have a bunch of microscopic clumps of molecules, such that the potential energy could be written as a sum over contributions:  $U = \sum_i U_i^{clump}$ , with  $U_i^{clump}$  the sum of the interaction energies of the pairs in the  $i$ th clump. Again, we need that the clump be microscopic, in order for a large number of them to be present, and a thermodynamic limit, along the lines of Eq. (2.27), to be possible.

## 2.4 Entropy and Thermodynamic Equilibrium

The most important characteristic of thermodynamic entropy, is growth in the approach to thermodynamic equilibrium, We could even arrive to define thermodynamic equilibrium, as that state of an isolated system in which entropy is maximum. After all, the statement that thermodynamic entropy is maximum, is perfectly equivalent to saying that quantities such as temperature, pressure, etc. are spatially uniform in the system.

It turns out that we can cook, out of the Shannon entropy, a perfectly reasonable thermodynamic entropy, that will itself be maximum at thermodynamic equilibrium.

Let us go back to the problem of a gas in a volume  $V$ , considered in the previous section. At thermodynamic equilibrium, and in the absence of external forces, the gas density will be uniform, and the probability to find a molecule, picked up at random, in a volume  $V_a$ , will be simply  $P(V_a) = V_a/V$ . We can partition  $V$  in volumes  $V_a$ , and calculate the Shannon entropy of the resulting distribution:

$$S[P] = - \sum_a P(V_a) \ln P(V_a). \quad (2.30)$$

We see immediately that, if the partition is uniform, the distribution  $P(V_a)$ , corresponding to thermodynamic equilibrium, will be the equiprobable distribution. This guarantees that the Shannon entropy is maximum at equilibrium, as required. We stress the probabilistic nature of this result. Nevertheless, since the probability  $P(V_a)$  is associated with the mean density  $\bar{n}$ ,  $S[P]$  will describe the density profile that is actually observed in equilibrium in the gas. Substituting  $P(V_a) = V_a/V$  into Eq. (2.30), we obtain

$$S[P] = \ln V/V_a. \quad (2.31)$$

Within constant factors, this is the right form of the spatial part of the entropy for an ideal gas (see Eq. (3.33)).

We could go one step further, and use the Shannon entropy to parameterize how improbable a certain configuration  $\tilde{n}_{V_a}(\mathbf{x}, t)$  is, working with frequencies in place of probabilities:

$$S[\tilde{P}_N] = - \sum_a \tilde{P}_N(V_a) \ln \tilde{P}_N(V_a); \quad \tilde{P}_N(V_a) = \tilde{N}_a/N. \quad (2.32)$$

This entropy is associated directly with the fluctuating densities  $\tilde{n}_{V_a}$ , as can be seen from the relation  $\tilde{P}_N(V_a) = (\tilde{n}_{V_a}/\bar{n})P(V_a)$ . We see that any fluctuation  $\tilde{n}_{V_a} - \bar{n}$  is associated with a value of the instantaneous entropy  $S[\tilde{P}]$  that is smaller than the equilibrium value  $S[P]$ . We stress the different origin of the differences  $S[P] - S[\tilde{P}]$ , and  $S[P] - S[P^{NE}]$ , with  $P^{NE}$  a generic non-equilibrium distribution. The first describes a microscopic effect (fluctuations); the second a macroscopic deviation from equilibrium of the system. This is revealed by the fact that the difference  $P - \tilde{P}$  will go to zero in the thermodynamic limit, while  $P^{NE} - P$ , that represents a macroscopic

condition (think e.g. of a configuration in which the gas is prepared in such a way that it occupies only a portion of  $V$ ), will remain finite in the limit.

We need at this point to make an observation. In order to obtain a reasonable definition of entropy, the partition  $\{V_a\}$  had to be uniform. We could have chosen, in principle, a different partition, in which certain regions of  $V$  were subdivided more finely than others. It is clear that a non-uniform partition would correspond to place different weight on information in different regions in  $V$ , with the most finely subdivided regions, the ones considered most important. The resulting entropy, however, would not have been maximum at equilibrium.

The important question is whether the uniform partition choice we have made, could have been guessed without knowledge of the character of the equilibrium state. After all, all points in the volume are equivalent for molecules interacting with short-range forces (once they are sufficiently away from the wall). Similarly, Galileian invariance suggests that uniform partition of velocity space should be the right choice, if we were to calculate the entropy associated with the velocity distribution of the molecules in the gas.

These statements can be made in fact more general: as it will be discussed in Sect. 5.2, the choice of a uniform partition has a lot to do with energy conservation, and with the geometry of the microscopic phase space of thermodynamic systems.

## 2.5 Stochastic Processes

The microscopic degrees of freedom of a thermodynamic system (think e.g. of the molecules in a gas) will be characterized, typically, by a random, unpredictable dynamics. Similar random behavior could be expected, at macroscopic scales, in non-equilibrium conditions. An example is turbulence. Quantities, such as the velocity of a molecule, or the fluid velocity in a turbulent could then be represented, as arrangement of random variables that depend on time (the velocity of the molecule), on space (a snapshot of the fluid velocity field in a turbulent flow) or even on space-time (the evolution of that turbulent flow). We introduce a minimum of terminology.

- We refer to a sequence of random variables, indexed in time, as a **stochastic process**. Indexing could be performed either on discrete or continuous time (discrete or continuous stochastic process).
- We speak of a **random field**, when the random variables are indexed in space or in space-time.

These definitions force us to reconsider concepts such as that of sample space, result of a measurement, and probability. To fix the ideas, let us focus, for the moment, on the case of a stochastic process, the generalization to the case of a random field being rather natural.

We notice at once that, while in the case of a random variable the result of an experiment was a single number, in the case of a stochastic process, the result is a “history” of the process, i.e., a function of time. In the case, say, of a noise generator, this could be the record of the signal during the time the generator was turned on.

- We shall call **realization** of the stochastic process, the particular function that is generated in the given experiment.

We see that, while in the case of a random variable,  $\Omega$  was some real interval, in the case of a stochastic process, it will be a function space. Thus, while in the case of a random variable, the PDF  $\rho(x)$  was a real function, in the case of stochastic process, the PDF will be a functional. In the case of a discrete stochastic process  $\psi$ , a realization would be a sequence  $\{\psi(t_n), n = 1, 2, \dots\}$ , and the PDF would be essentially a function of a vector (possibly with infinite components)  $\rho(\{\psi(t_n), n = 1, 2, \dots\})$ . In the case of a continuous stochastic process, the argument of  $\rho$  would be a function of real variable  $\psi = \psi(t)$ . We shall use the square bracket notation  $\rho[\psi]$  to indicate simultaneous dependence of  $\rho$  on all values of  $\psi(t)$  in the domain of  $t$ . An example:

$$\rho[\psi] = \mathcal{N} \exp\left(-\int_0^T |\psi(t)|^2 dt\right). \quad (2.33)$$

It is difficult, in general, to work with objects such as Eq. (2.33). In most situations, we must content ourselves with a reduced description, such as the one provided e.g. by the joint PDF

$$\rho(\{\psi_i, t_i\}) \equiv \rho(\{\psi(t_i) = \psi_i, i = 1, 2, \dots\}), \quad (2.34)$$

where the  $t_i$ 's represent generic sampling times of the continuous stochastic process.

A basic question we may want to ask is the evolution of a stochastic process, given a certain past history. We may want to consider, therefore, also conditional PDF's such as

$$\rho(\{\psi_i, t_i\}|\{\bar{\psi}_j, \tau_j\}) \equiv \rho(\{\psi(t_i) = \psi_i\}|\{\psi(\tau_j) = \bar{\psi}_j\}), \quad (2.35)$$

where  $t_i > \tau_j \forall i, j$ . More in general, we may consider the PDF resulting from an ensemble of trajectories stemming from initial conditions distributed with a generic PDF  $\rho_0(\{\bar{\psi}_j, \tau_j\})$ :

$$\rho(\{\psi_i, t_i\}) = \int \prod_j d\bar{\psi}_j \rho(\{\psi_i, t_i\}|\{\bar{\psi}_j, \tau_j\}) \rho_0(\{\bar{\psi}_j, \tau_j\}) \quad (2.36)$$

We stress that the PDF  $\tilde{\rho}_0$ , that acts as initial condition for  $\rho$ , is arbitrary. This seems to suggest that, in order to define a stochastic process, we should impose always some kind of "initial condition". That is to say, the PDF in Eq. (2.34) will contain in general information on the way the process is turned on. The problem of isolating the intrinsic properties of the stochastic process can be solved if memory of the initial conditions decays with time:

$$\lim_{\tau_j \rightarrow -\infty} \rho(\{\psi_i, t_i\}|\{\bar{\psi}_j, \tau_j\}) = \rho(\{\psi_i, t_i\}). \quad (2.37)$$

This is a condition of statistical independence of the values of the stochastic process at large time separations, that is actually rather difficult not to be satisfied. We have basically two possibilities:

- The system is deterministic. An example is the “ping-pong”, a discrete time process that oscillates indefinitely between two states 0 and 1. Thus, even for  $t \rightarrow \infty$ :  $P(\psi, t|\psi_0, 0) = \delta_{\psi\psi_0}$ , if  $t$  is even, while  $P(\psi, t|\psi_0, 0) = 1 - \delta_{\psi\psi_0}$ , if  $t$  is odd.
- The ambient space of the process is decomposable in subdomains that are mutually inaccessible, in the sense that, if  $\psi$  was initially in one of them, it will remain there forever. An example of such a situation, is a molecule that moves in a volume divided in two parts by an impermeable wall.

A stochastic process will be said to be in a condition of **stationary statistics**, if the PDF's in Eq. (2.34) are invariant by time translations:

$$\rho(\{\psi_i, t_i + T\}) = \rho(\{\psi_i, t_i\}), \quad \forall T. \quad (2.38)$$

This means that the PDF's  $\rho(\{\psi_i, t_i\})$  depend solely on the time differences  $t_i - t_j$ , and that the PDF  $\rho(\psi, t) = \rho(\psi)$  is itself independent of the time  $t$ . If the limit in the LHS of Eq. (2.37) exists (this basically rules out deterministic systems such as the ping-pong), one speaks of **equilibrium statistics**. The limit

$$\rho(\{\psi_i, t_i|\bar{\psi}\}) \equiv \rho(\{\psi_i, t_i|\bar{\psi}, \tau \rightarrow -\infty\})$$

is called an **equilibrium PDF**. (Notice that a single initial condition  $\bar{\psi}$  is sufficient to identify the portion of ambient space where a given realization evolves). If Eq. (2.37) is satisfied, which implies that the stochastic process ambient space is undecomposable, the equilibrium statistics will also be unique. This means that any initial condition  $\rho_0(\psi, t_0)$  will evolve, after a sufficiently long time, into the single equilibrium PDF  $\rho(\psi)$ . If the ambient space were decomposable, the same statement would remain true, but only locally in each of the subdomains in which the ambient space is subdivided. If  $\rho_0$  had support in just one of such subdomains, which can be identified by one of its points  $\bar{\psi}$ :  $\rho_0(\psi, t_0) \rightarrow \rho(\psi|\bar{\psi})$ . Notice finally, that a PDF could be stationary, and, at the same time, not be an equilibrium PDF. An example is again provided by the ping-pong: the PDF  $\rho(1) = \rho(0) = 1/2$  is obviously stationary, but it is not true that, after a long enough time, any initial PDF would collapse to the stationary PDF.

### 2.5.1 The Ergodic Property

Exactly as moments can be used to describe the statistics of random variables, **correlation functions** in the form

$$\langle \psi(t_1)\psi(t_2)\dots \rangle, \quad (2.39)$$

can be used to describe the statistics of a stochastic process  $\psi(t)$ . The two-time (subtracted) correlation function

$$C(t, t') = \langle \psi(t)\psi(t') \rangle - \langle \psi(t) \rangle \langle \psi(t') \rangle, \quad (2.40)$$

in particular, allows to estimate the degree of statistical independence of the values at different times of the stochastic process. We have in fact, by definition:

$$\langle \psi(t)\psi(t') \rangle = \int d\psi \int d\psi' \rho(\psi, t; \psi', t'),$$

and, in the case of independent values of  $\psi$ , i.e. if  $\rho(\psi, t; \psi', t') = \rho(\psi, t)\rho(\psi', t')$ :

$$\langle \psi(t)\psi(t') \rangle = \langle \psi(t) \rangle \langle \psi(t') \rangle.$$

We can define a **correlation time**, that tells us at what time separation, correlations can be considered negligible. The definition commonly adopted is

$$\tau_\psi(t) = \frac{1}{C(t, t)} \int_0^\infty C(t, t + \tau) d\tau \quad (2.41)$$

(as it is to say, area under  $C(t, t + \tau)$  equal to basis – i.e.  $\tau_\psi(t)$  – times height – i.e.  $C(t, t)$ ). In terms of a single realization of the stochastic process,  $\tau_\psi(t)$  tells us how long we must wait until  $\psi(t + \tau)$  differs much from  $\psi(t)$ . In other words, the correlation time gives us the time scale of variation of the process.

The averages that appear in the equations that have been written up to now, could be estimated by means of sample averages. We wonder whether, in the case of a time-independent process, such averages could be replaced by time averages in the form

$$\langle \psi \rangle_T = \frac{1}{T} \int_0^T \psi(t) dt. \quad (2.42)$$

In other words, whether, instead of estimating averages out of sample averages over multiple realization of the same process, we could perform this operation, working with only one realization of that process.

An example, in which the sample and time averages are trivially equivalent, is that of the stochastic process, obtained putting in sequence the independent measurements of a random variable  $y$ . It is rather clear that carrying out a sample average using separate realizations of such process, would be pointlessly costly.

- The condition that averages  $\langle \cdot \rangle$  could be approximated, for long enough sample times, by time averages, is called an **ergodic property** for the process.

The first condition for an ergodic property is clearly that the statistics of the process be stationary on the sampling time of the average.

The second condition is that Eq. (2.37) be satisfied. This implies in particular that the ambient space of the stochastic process is indecomposable. If this condition were not satisfied, any PDF in the form, say,

$$\rho(\psi) = \int d\bar{\psi} \rho_0(\bar{\psi}) \rho(\psi, t|\bar{\psi}, \tau \rightarrow -\infty)$$

could be an equilibrium PDF for the process, and time averages starting from the subdomain of  $\psi$ , containing  $\bar{\psi}$ , could estimate averages only for  $\rho(\psi, t|\bar{\psi}, \tau)$ . More in general, Eq. (2.37) implies statistical independence of  $\psi$  at infinite time separation. This guarantees that we can imagine the sampling interval  $T$  as a sequence of many intervals, such that the values of the quantity, in any two of them, are approximately independent. Actually, in order for ergodicity to be satisfied, it is enough that the correlations of the quantity to be averaged decay to zero for  $t \rightarrow \infty$ .

The third condition, that parallels the one in the law of large numbers, is that average and variance of the quantity in exam are finite.

We illustrate the result, in the case the quantity to average is simply  $\psi$ , and the correlation time  $\tau_\psi$  is finite. We proceed as in Sect. 2.3, comparing  $\langle\psi\rangle$  with  $\langle\psi\rangle_T$ , and calculating  $\sigma_{\langle\psi\rangle_T}^2$ . Exchanging average and time integral in  $\langle\langle\psi\rangle_T\rangle$ , we obtain at once

$$\langle\langle\psi\rangle_T\rangle = \langle\langle\psi\rangle\rangle_T = \langle\psi\rangle.$$

Let us indicate, by  $\tilde{\psi}(t) = \psi(t) - \langle\psi\rangle$ , the deviation from the mean, and write

$$\sigma_{\langle\psi\rangle_T}^2 = \frac{1}{T^2} \int_0^T dt \int_0^T dt' \langle\tilde{\psi}(t)\tilde{\psi}(t')\rangle = \frac{1}{T^2} \int_0^T dt \int_0^T dt' C(t-t').$$

Now, for  $T \gg \tau_\psi$ , we can approximate the integral in  $dt'$ , for almost all values of  $t$ , with an integral between  $\pm\infty$  (at  $t' = 0$  and at  $t' = T$ , if  $|t-t'| \gg \tau_\psi$ ,  $C(t-t')$  will be in any case already zero). Exploiting Eq. (2.41), we find:

$$\sigma_{\langle\psi\rangle_T}^2 \approx \frac{1}{T^2} \int_0^T dt \int_{-\infty}^{+\infty} d\tau C(\tau) = \frac{2\tau_\psi}{T} C(0) \equiv \frac{2\tau_\psi}{T} \sigma_\psi^2. \quad (2.43)$$

The variance  $\sigma_{\langle\psi\rangle_T}^2$  decreases with the sampling time  $T$ , that plays here the same role as the sample size  $N$  in the law of large numbers Eq. (2.25).

The line of reasoning that we have followed so far, could be adapted to the random field case, introducing space correlations in the form

$$C(\mathbf{x}, \mathbf{y}) = \langle\psi(\mathbf{x})\psi(\mathbf{y})\rangle - \langle\psi(\mathbf{x})\rangle\langle\psi(\mathbf{y})\rangle.$$

The concept of statistical stationarity translates in that of spatially uniform statistics, as a requirement of independence on spatial translation. This means, in particular

$C(\mathbf{x}, \mathbf{y}) \rightarrow C(\mathbf{x} - \mathbf{y})$ . In many instances, the statistics is also isotropic, meaning that  $C(\mathbf{r}) \rightarrow C(r)$ . In this case, definition of a orientation-independent **correlation length** is possible:

$$\lambda_\psi = \left( \frac{1}{C(0)} \int_0^\infty d^3r C(r) \right)^{1/3}.$$

In the case of spatially uniform statistics, we can define the concept of spatial average; for instance, in the case of  $\psi$ :

$$\langle \psi(\mathbf{x}) \rangle_V = \frac{1}{V} \int_V d^3y \psi(\mathbf{y}),$$

where  $V$  is centered around  $\mathbf{x}$ . An ergodic property will be satisfied in space, if the statistics is uniform on the sampling volume and  $\lambda_\psi$  is small. A simple generalization of the calculation leading to Eq. (2.43), gives in fact

$$\sigma_{\langle \psi \rangle_V}^2 \approx \frac{\lambda_\psi^3 \sigma_\psi^2}{V}.$$

This result brings us back to the considerations in Sect. 2.3, on the necessity that correlations between molecules decay at microscopic scales, in order for a thermodynamic limit to be satisfied. The molecule count in a volume, in that case, could be seen as an integral over a density in that same volume. Correlations among molecules, would translate into correlation among density fluctuations, and the condition that correlations among molecules, occur only at microscopic scales, corresponds to the condition that  $\lambda_n$  itself be microscopic.

## 2.5.2 Markov Processes

We have dealt so far with the “kinematics” of a stochastic process. We would like to pass to the “dynamics”, i.e. to the task of writing evolution equations for time-dependent probabilities and PDF’s. The simplest object that we have considered, is the conditional PDF  $\rho(\psi, t | \bar{\psi}, \tau)$ .

We concentrate, for the moment, on the case of stochastic processes that take discrete values  $\psi_n = n\Delta\psi$  (called **states** of the system), at discrete times  $t_k = k\Delta t$ . Given a stochastic process  $\psi(t_k)$ , we utilize the notation

$$P(\{n_i; i = 0, 1, \dots\}) \equiv P(\{\psi(t_i) = \psi_{n_i}, i = 0, 1, \dots\}).$$

Similar notations could be introduced in the case of conditional probabilities. We are interested, in particular, in the quantity  $P(n_{i+1} | \{n_k; k = 0, 1, \dots, i\})$ , that is

connected with the probability of one realization (starting at discrete time  $k = 0$ ), by the standard relation:

$$P(n_{i+1}|\{n_k; k = 0, 1, \dots, i\}) = \frac{P(\{n_k; k = 0, 1, \dots, i+1\})}{P(\{n_k; k = 0, 1, \dots, i\})}.$$

We shall consider only two key cases:

- Totally uncorrelated processes (e.g. the process resulting from repeated independent measurements of a random variable):

$$P(n_{i+1}|\{n_k; k = 0, 1, \dots, i\}) = P(n_{i+1}).$$

- Systems in which, knowledge of the state of the system at a certain time, makes all information on the previous states of the system, irrelevant (example: the random walker that decides each new step with a throw of dice. Its position at any given step will depend only on its position one step earlier, and on the result of the last throw of dice). In formulae:

$$P(n_{i+1}|\{n_k; k = 0, 1, \dots, i\}) = P(n_{i+1}|n_i). \quad (2.44)$$

A process with such characteristics, is called a **Markov process**.

We see that it is possible to derive local evolution equations for Markov processes, based solely on knowledge of the **transition probability**  $P(n_{i+1}|n_i)$ . The role of building block played by the transition probability, is particularly apparent in the case of the random walker, as  $P(n_{i+1}|n_i)$  tells us precisely with which probability the walker goes, in one step, from location  $n_i$  to  $n_{i+1}$ .

The starting point is the following obvious relation, which descends from Eq. (2.44):

$$P(n_{i+1}, n_i, n_0) = P(n_{i+1}|n_i, n_0)P(n_i|n_0)P(n_0) = P(n_{i+1}|n_i)P(n_i|n_0)P(n_0).$$

From here we get

$$P(n_{i+1}, n_i|n_0) = P(n_{i+1}|n_i)P(n_i|n_0).$$

Summing over the intermediate states, we obtain the following relation, called the **Chapman-Kolmogorov equation**:

$$P(n_{i+1}|n_0) = \sum_{n_i} P(n_{i+1}|n_i)P(n_i|n_0). \quad (2.45)$$

In plain words: the probability of finding the system at a given time in a certain state, is obtained, summing on the probabilities of the possible states in which the system could lie at the previous time, times the transition probability from the past to the current state.

The starting point in the derivation of the Chapman-Kolmogorov equation (2.45), was  $P(n_{i+1}, n_i, n_0)$ . We could have used a generic probability  $P(n_k, n_i, n_0)$ , with  $k > i > 0$  generic, and, proceeding in the same way, we would have obtained

$$P(n_k|n_0) = \sum_{n_i} P(n_k|n_i)P(n_i|n_0).$$

Similarly, we can multiply both sides of the Chapman-Kolmogorov equation by an arbitrary probability  $P(n_0)$ , sum over the initial state  $n_0$ , and obtain the evolution equation for the time dependent probability

$$P(n_{i+1}) = \sum_{n_0} P(n_i|n_0)P(n_0), \quad (2.46)$$

that has solution, once the initial condition  $P(n_0)$  is given.

The most complete statistical description we could reach, of a discrete stochastic process such as the one considered, is provided by the joint probability  $P(\{n_i, i = 1, 2, \dots\}|n_0)$ . Knowledge of the transition probability  $P(n_{i+1}|n_i)$  is again enough to obtain such a description. By repeated application of the relation

$$\begin{aligned} P(\{n_i, i = 1, 2, \dots, k+1|n_0\}) \\ &= P(\{n_{k+1}|n_i, i = 0, 2, \dots, k\})P(\{n_i, i = 1, 2, \dots, k\}|n_0) \\ &= P(\{n_{k+1}|n_k\})P(\{n_i, i = 1, 2, \dots, k\}|n_0), \end{aligned}$$

we obtain the following so-called **Markov chain expression**:

$$P(n_i, n_{i-1}, \dots, n_1|n_0) = P(n_i|n_{i-1})P(n_{i-1}|n_{i-2}) \dots P(n_1|n_0). \quad (2.47)$$

If the transition probability is independent of time,

$$P(n_{i+1} = m|n_i = q) \equiv P(q \rightarrow m)$$

(in which case, the Markov process is said to be itself time-independent), and  $P(n_0)$  is a stationary solution of the Chapman-Kolmogorov equation, multiplication of Eq. (2.47) by  $P(n_k)$  will lead to an expression for  $P(\{n_i, i = 0, 1, \dots\})$  that is invariant by time translation, i.e. to stationary statistics.

## 2.6 Coarse-Graining and Continuous Limit

The passage from a microscopic description of a physical system to a macroscopic one, can be envisioned as a coarse-graining operation, in which microscopic degrees of freedom of the system are averaged away. Thermodynamic variables such as pressure, temperature and density can then be seen as local (volume) averages of microscopic quantities.

By and large, the coarse-graining process can be visualized as a sequence of three steps:

- Identification of a **microscale** of the system, below which the details of the dynamics are unknown (and supposed irrelevant). In the case of a gas, this could be the scale of the individual molecules.
- Introduction of a **mesoscale**  $l$ , where the dynamics is “less complicated” than at the microscale. This could mean many things. In the case of the gas, again, it could be the scale at which a thermodynamic limit begins to be possible (i.e. fluctuations are small). In this case, the mesoscale is precisely the smallest scale at which a description of the system in terms of macroscopic variables (e.g. the density) becomes possible.
- Identification of the **macroscale**  $L$  of the system, as the scale of variation for the macroscopic variables.

In the example of the gas, considered in Sect. 2.3, the microscopic description was provided by knowledge of the exact position of all molecules, which could be expressed through the fine-grained expression for the density field

$$\tilde{n}(\mathbf{x}, t) = \sum_{i=1}^N \delta(\mathbf{x} - \mathbf{x}_i(t)), \quad (2.48)$$

with  $\mathbf{x}_i(t)$  the position of the  $i$ th molecule. The mesoscopic description of the same system corresponds to coarse-graining the density field  $\tilde{n}(\mathbf{x}, t)$  at the scale  $l$  identified by the volume  $V_a$ :

$$\tilde{n}_{V_a}(\mathbf{x}_a, t) = \langle \tilde{n}(\mathbf{x}_a, t) \rangle_{V_a} \equiv N_a(\mathbf{x}_a, t)/V_a.$$

The important point is that, if the coarse graining scale is much larger than that of the microscopic constituents, the coarse-grained quantity will take a (more or less) continuous character. This is a **continuous limit** for the coarse-grained quantity, that corresponds to our notion of macroscopic variables as continuous fields at the macroscopic scales of interest for the system.

Similar coarse-graining procedures can be carried out with respect to time, if, in place of a random field, we consider a stochastic process.

Once coarse graining has been carried out, and a continuous limit has been taken, it may become necessary to discretize again the variables, e.g. to solve numerically the evolution equations. In certain sense, the operation is akin to follow backward the coarse graining procedure, passing from a mesoscopic description of the system to a fictitious microscopic one. The fact that different microscopic dynamics lead to the same macroscopic behaviors is interesting, as it suggests us that only a small subset of features at the microscale could actually be relevant at the macroscopic level. It could then be expedient to seek from the start a model of the microscopic dynamics that is as simple as possible.

Unfortunately, the process of building a discrete model that has an appropriate continuous limit is far from trivial. In fact, it is not guaranteed at all that a microscopic

dynamics will admit in general a continuous limit. This is why we stressed earlier the fact that continuity had to be understood in a “more or less” sense. There is a general problem of lack of regularity in the realizations of continuous stochastic processes and random fields, that, as we shall see in the next section, disappears only if one decides to deal with average quantities. An example of such lack of regularity is provided already in the case of the random walker.

Indicate with  $\Delta t$  and  $\Delta x$  the time constant and the step length of the walker (both assumed constant and position independent). Focus on the case of an unbiased, spatially uniform and stationary random walk process, such that the jump probability can be written in the form

$$P(x \rightarrow x + \Delta x) = P(x \rightarrow x - \Delta x) = P_J, \quad P(x \rightarrow x) = 1 - 2P_J, \quad (2.49)$$

with  $P_J$  a constant. We want to take the continuous limit of this process by considering macroscopic times  $t$  and separations  $x(t) - x(0)$ , such that  $t \gg \Delta t$  and  $x(t) - x(0) \gg \Delta x$ . From Eq. (2.49), we obtain immediately:

$$\langle |x(t) - x(0)|^2 \rangle = 2P_J \sum_{i=1}^{t/\Delta t} \Delta x^2 = 2 \frac{P_J \Delta x^2}{\Delta t} t. \quad (2.50)$$

The two quantities  $\Delta x$  and  $\Delta t$  are seen as zero at the macroscopic scales fixed by  $x(t) - x(0)$  and  $t$ . In order for Eq. (2.50) to make sense, however, it is necessary that the quantity  $\kappa = P_J \Delta x^2 / \Delta t$ , remains finite as  $\Delta t$  and  $\Delta x$  are sent to zero at fixed  $P_J$ . In other words, in order for a continuous limit to exist, it is necessary that  $\Delta x = O(\Delta t^{1/2})$ . We can verify that the realizations of the continuous random walk process obtained, taking this limit, are continuous. In fact, Eq. (2.50) tells us that the typical displacement  $|x(t) - x(0)|$  goes to zero as  $t^{1/2}$ , so that the typical realization of the process is continuous.

But they are not differentiable.

If we try to define a velocity scale as limit of a finite difference ratio, we obtain in fact an infinite result:

$$\dot{x} \sim \lim_{t \rightarrow 0} \frac{\langle |x(t) - x(0)|^2 \rangle^{1/2}}{t} \propto \lim_{t \rightarrow 0} t^{-1/2} = \infty.$$

As we shall in the next section, in order to find a well-behaved quantity in the limit considered, we must consider average quantities such as the PDF  $\rho(x, t)$ . In alternative, coarse graining of the stochastic process must be carried out:

$$x(t) \rightarrow x_\tau(t) = \int dt' G_\tau(t - t') x(t'), \quad (2.51)$$

where  $G$  has support  $\sim \tau$ , with  $\int dt G_\tau(t) = 1$  and  $G_\tau(t) > 0$ . The regularity properties of  $\dot{x}_\tau(t)$  would be therefore those of  $G_\tau$ . (The simplest choice  $G_\tau(t) = (1/(2\tau))\delta_{[-\tau, \tau]}(t)$ , would be good to cure discontinuities in  $\dot{x}$ , but would not be enough to guarantee differentiability of  $\dot{x}_\tau$ ).

**Note** We provide an example of the way in which coarse-graining can be used to simplify a complicated microscopic dynamics. Consider a discrete random walker, whose steps have variable length and are correlated in time. Due to time correlations, the walker dynamics will be non-Markovian. Imagine for simplicity that correlations exist only between successive steps. Indicating by  $\Delta n_i$  the step at discrete times  $t_i = i\Delta t$ :

$$\langle \Delta n_i \Delta n_{i\pm 1} \rangle = c\sigma_{\Delta n}^2, \quad 0 < c < 1; \quad \langle \Delta n_i \Delta n_{i\pm k} \rangle = 0, \quad k \geq 2.$$

Let us coarse-grain time at a scale  $\overline{\Delta t}$  that contains many microscale intervals  $\Delta t$ . We shall have

$$\overline{\Delta n_i} = \sum_{j=1}^m \Delta n_{i+j},$$

from which, exploiting  $c > 0$ :

$$\sigma_{\overline{\Delta n}}^2 \geq m\sigma_{\Delta n}^2,$$

while

$$\langle \overline{\Delta n_i} \overline{\Delta n_{i\pm 1}} \rangle = \langle \Delta n_i \Delta n_{i\pm 1} \rangle = c\sigma_{\Delta n}^2$$

(only the last contribution  $\Delta n_i$  in an interval  $\overline{\Delta t}$ , and the first one in the next interval, are correlated). We thus see that the correlation strength, parameterized by the ratio  $\langle \overline{\Delta n_i} \overline{\Delta n_{i\pm 1}} \rangle / \sigma_{\overline{\Delta n}}^2$ , becomes negligible for large  $m$ . In other words, coarse-graining in time brings us back to an independent step situation, such that  $n_i$  becomes a Markov process. ■

## 2.7 Master Equation and Its Applications

As we have said, the regularity problems in the continuous limit of a Markov process, will disappear, if we work with averages rather than with realizations. In particular, we can carry out a continuous limit on both arguments of the probability  $P(n, k) \equiv P(\psi(t_k) = \psi_n)$ ;  $\psi_n = n\Delta\psi$ ;  $t_k = k\Delta t$ , and be able to write evolution equations for such probability (and associated PDF's) in differential, or integro-differential form.

Our starting point is the Chapman-Kolmogorov equation (2.45), that we can rewrite in the form

$$\begin{aligned} P(m, i+1) - P(m, i) &= \sum_k P(k \rightarrow m)P(k, i) - P(m, i) \\ &= \sum_{k \neq m} [P(k \rightarrow m)P(k, i) - P(m \rightarrow k)P(m, i)], \end{aligned} \quad (2.52)$$

where we have utilized the relation  $\sum_k P(m \rightarrow k) = 1$ , and the step  $\sum_k \rightarrow \sum_{k \neq m}$  is justified by the fact that the term  $k = m$ , in the two addends in square brackets, cancel one another.

If the difference  $P(m, i + 1) - P(m, i)$  is small on the scale of  $P(m, i)$ , which is equivalent to say that the transition probability is small:  $P(k \rightarrow m) \ll 1$ , it is meaningful to introduce the transition rate (transition probability per unit time):

$$W(k \rightarrow m) = \frac{P(k \rightarrow m)}{\Delta t}.$$

The transition probability, in a time interval  $\overline{\Delta t} > \Delta t$ , will thus be

$$P(m, t + \overline{\Delta t} | k, t) = W(k \rightarrow m) \overline{\Delta t} + O((W \overline{\Delta t})^2),$$

where the first term to RHS is the probability of just one transition in the interval, while the second is the contribution from multiple transitions. The continuous limit corresponds to consider the coarse graining scale  $\overline{\Delta t}$  much larger than  $\Delta t$ , but, at the same time, small enough for  $O((W \overline{\Delta t})^2)$  contributions to  $P(m, t + \overline{\Delta t} | k, t)$  to be negligible. Equation (2.52) can be cast therefore in the differential form

$$\frac{\partial P(m, t)}{\partial t} = \sum_{k \neq m} [W(k \rightarrow m) P(k, t) - W(m \rightarrow k) P(m, t)]. \quad (2.53)$$

This is called the **master equation** for the process. The Markov process, described in Eq. (2.53), is continuous in time, but is still defined in a discrete state space. The condition to take the continuous limit also in  $\psi$ , is that the variations of  $P(m, t)$ , this time with respect to  $m$ , be small on the scale of  $P(m, t)$ . We can in this case introduce the PDF  $\rho(\psi_m, t) = P(m, t) / \Delta \psi$ , and the transition rate density

$$w(\psi_k \rightarrow \psi_m) = \frac{W(k \rightarrow m)}{\Delta \psi}.$$

The master equation will take the integro-differential form:

$$\frac{\partial \rho(\psi, t)}{\partial t} = \int d\psi' [w(\psi' \rightarrow \psi) \rho(\psi', t) - w(\psi \rightarrow \psi') \rho(\psi, t)]. \quad (2.54)$$

The non-local character of the above equation, reflects the fact that the transitions at the microscopic timescale  $\Delta t$ , occur at a macroscopic scale. In order for a clear separation to exist between macroscopic world and microscopic stochastic dynamics, it is necessary that the transitions occur at the microscopic scale fixed by  $\Delta \psi$ . We shall provide below, two important examples of systems with such clear separation of scales, whose evolution will be described by partial differential equations, rather than integro-differential equations for the probability.

### 2.7.1 The Random Walk

We have already seen examples of this process in the previous sections: an individual that at discrete times  $t_k = k\Delta t$ , with equal probabilities  $P_J$ , may jump either left or right one discrete step  $\Delta x$ , and with probability  $1 - 2P_J$ , may remain where it is. The states of the stochastic process are therefore the discrete positions  $x_n = n\Delta x$ .

We have seen in Sect. 2.6, that a continuous limit requires that the quantity  $\kappa = P_J(\Delta x)^2\Delta t$  remains finite for  $\Delta x, \Delta t \rightarrow 0$ , which imposes  $\Delta x = O((\Delta t)^{1/2})$ . If we plug into the master equation (2.52), we find

$$\frac{P(m, i+1) - P(m, i)}{\Delta t} = \kappa \frac{P(m+1, i) + P(m-1, i) - 2P(m, i)}{(\Delta x)^2}. \quad (2.55)$$

We recognize immediately, to RHS, the finite difference representation of  $\partial^2 P / \partial x^2$ . Taking the continuous limit, and introducing the PDF  $\rho(x_n, t_k) = P(n, k) / \Delta x$ , we obtain the **diffusion equation** (or **heat equation**):

$$\frac{\partial \rho(x, t)}{\partial t} = \kappa \frac{\partial^2 \rho(x, t)}{\partial x^2}, \quad (2.56)$$

in which  $\kappa$  takes the name of **diffusivity** of the random walker. The dynamics described by Eq. (2.56) corresponds to the progressive spreading of the initial walker distribution. The nature of the equation suggests the existence of similarity solutions in the form

$$G(x, t) = g(x^2/t).$$

Substituting into Eq. (2.56), we find in fact the Gaussian solution

$$G(x, t) = \frac{1}{\sqrt{4\pi\kappa t}} \exp\left(-\frac{x^2}{4\kappa t}\right), \quad (2.57)$$

corresponding to the initial condition  $G(x, 0) = \delta(x)$ . The solution  $G(x, t)$  is the Green function for the problem in an infinite domain. The evolution of the PDF  $\rho$ , out of the generic initial condition  $\rho(x, 0)$ , will be therefore

$$\rho(x, t) = \int dx' G(x - x', t) \rho(x', 0). \quad (2.58)$$

If the domain is finite, determination of Green function of the problem will be more complicated, but the stationary solutions to Eq. (2.56) are easy to find. In the case of a finite domain with impermeable walls,  $A = [x_1, x_2]$ , in particular, we find the spatially uniform solution  $\rho(x) = 1/(x_2 - x_1)$ . Notice that this was the maximum-entropy equilibrium distribution, described in Sect. 2.4. In fact, it is possible to prove

that the dynamics described by Eq. (2.56) corresponds always to growth of the PDF entropy of  $\rho(x, t)$  (see problem 4 at the end of this chapter).

Let us go back to analysis of the diffusion equation (2.56). We see that its structure is that of a continuity equation

$$\frac{\partial \rho(x, t)}{\partial t} + \frac{\partial J(x, t)}{\partial x} = 0, \quad (2.59)$$

in which  $J = -\kappa \partial \rho / \partial x$  has the meaning of a probability flux. We can associate the current  $J$  with a **current velocity**

$$u(x, t) = J(x, t) / \rho(x, t) = -\kappa \frac{\partial \ln \rho(x, t)}{\partial x},$$

that is the average velocity of the walkers that cross coordinate  $x$  at time  $t$ . It must be stressed that this quantity does not represent an average velocity of the individual walkers (this average is indeed zero), rather, it is the cumulative result of the fact that, if there are more walkers at one side of  $x$ , more walkers will cross  $x$  from that side. The current  $J$  has the form of a **flux-gradient** relation,  $J = -\kappa \partial \rho / \partial x$ , whose effect is to smooth out the inhomogeneities in  $\rho$ .

We can use Eq. (2.59) to write the evolution equation for the probability  $P_A(t) = \int_A dx \rho(x, t)$ , of finding the walker in a domain  $A = [x_1, x_2]$ , in the form

$$\dot{P}_A(t) = J(x_1, t) - J(x_2, t),$$

(the boundary conditions, appropriate to describe a domain with impermeable walls, will thus be the zero-gradient condition  $\partial_x \rho(x, t)|_{x=x_{1,2}} = 0$ ).

In general, evolution equations for the PDF of stochastic processes cannot be solved in closed form. A partial solution of such equation, can nevertheless be obtained, in some cases, in terms of moments.

We can convert an evolution equation for a PDF into one for the moments  $\langle x^n \rangle$ , by multiplying both sides of the equation, by the desired power of  $x$ , and integrating over  $x$ . Let us carry out the procedure explicitly in the case of Eq. (2.56). Indicate by  $f(x)$  the quantity, of whose average we want to determine the evolution. Multiplying the LHS of Eq. (2.56) by  $f(x)$ , and integrating over  $x$ , we find:

$$\int dx f(x) \frac{\partial \rho(x, t)}{\partial t} = \frac{\partial}{\partial t} \int dx f(x) \rho(x, t) = \frac{d}{dt} \langle f(x(t)) \rangle.$$

Doing the same thing with the RHS of that same equation, and eliminating derivatives on  $\rho$ , by repeated integration by parts, we obtain

$$\int dx f(x) \frac{\partial^2 \rho(x, t)}{\partial x^2} = \left[ f(x) \frac{\partial \rho(x, t)}{\partial x} \right]_a^b - \left[ \frac{\partial f(x)}{\partial x} \rho(x, t) \right]_a^b + \left\langle \frac{\partial^2 f}{\partial x^2} \right\rangle,$$

where  $a$  and  $b$  are the extrema of the walker domain. If the domain is infinite, the boundary terms can be eliminated,<sup>1</sup> and we are left with an equation in the form

$$\frac{d}{dt}(f) = \kappa(f'').$$

Taking  $f(x) = x$ , we find  $x(t) = x(0)$ , and taking  $f(x) = x^2$ , we obtain once more Eq. (2.50).

### 2.7.2 Extinction Process

In this case, the stochastic process describes the evolution of the number  $N$  of individuals in a sterile community. We count the individuals at discrete times  $t_k = k\Delta t$  so closely spaced, that it is very unlikely that more than one individual dies in the interval  $\Delta t$ . We suppose that age does not affect the death process, as it would occur if the individuals were nuclei undergoing radioactive decay. In the same way, we suppose the death events to be independent, so that the probability to observe a death in the population in the interval  $\Delta t$ , will be  $N$  times the probability of death of one individual in the same time interval. If we indicate by  $\Gamma_D$  the death probability per unit time of one individual, the transition probability for the system will be

$$P(N \rightarrow N - 1) = W(N \rightarrow N - 1)\Delta t = N\Gamma_D\Delta t.$$

No other transitions are possible. The master equation for the process will be therefore

$$\frac{\partial P(N, t)}{\partial t} = \Gamma_D[(N + 1)P(N + 1, t) - NP(N, t)] \simeq \Gamma_D \frac{\partial(NP(N, t))}{\partial N},$$

which could be converted into an equation for the PDF  $\rho(N, t) = P(N, t)/\Delta N \equiv P(N, t)/1 = P(N, t)$ :

$$\frac{\partial \rho(N, t)}{\partial t} = \Gamma_D \frac{\partial(N\rho(N, t))}{\partial N}. \quad (2.60)$$

Proceeding as in the random walk case, we can convert the evolution equation for the probability, into an evolution equation for the mean population and its fluctuation.

As in the case of the diffusion equation (2.56), the evolution equation (2.60) can be cast in the form of a continuity equation, such as (2.59), with probability flux  $J(N, t) = -N\Gamma_D\rho(N, t)$ . In this case, the current velocity  $u(N, t) = -N\Gamma_D$  is the speed at which a community of  $N$  individuals shrinks, due to the death

---

<sup>1</sup> The simpler way to be convinced of this fact is that the discrete walker will not be able, at finite  $t$ , to reach infinity; thus  $\rho = 0$  there. The principle extends to generic discrete Markov processes with local transitions.

process. The end result is extinction, corresponding to an equilibrium distribution  $P(N, t \rightarrow \infty) = \delta_{N,0}$ . This is to be contrasted with the maximally spread-out state, towards which tends the random walker distribution. We thus see that a random microscopic dynamics may equally well lead to a statistical equilibrium state, whose Shannon entropy is maximum (the random walker), or zero (the extinction process). The difference between the two processes is clearly the presence of bias (towards zero population) in the extinction process, that was absent in the case of the random walk. This is a delicate and important point. We shall see that energy conservation, in thermally isolated systems, plays a role analogous to the requirement of zero bias, to guarantee entropy growth in the relaxation to equilibrium.

## 2.8 The Wiener-Khinchin Theorem

A rather natural way to characterize a random field, especially in the case of spatially homogeneous statistics, is to consider its Fourier spectrum. The same operation could obviously be carried out with a stochastic process, working in time.

Consider a random field  $\psi(\mathbf{x})$ ,  $\mathbf{x} \in \mathbf{R}^d$ . We can define its Fourier transform:

$$\psi_{\mathbf{k}} = \int d^d x \psi(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}}; \quad \psi(\mathbf{x}) = \int \frac{d^d k}{(2\pi)^d} \psi_{\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{x}}. \quad (2.61)$$

From the mathematical point of view, it may be convenient to define the Fourier transform in terms of distributions; from the point of view of physics, a discretization and a finite domain for  $\psi(\mathbf{x})$ , will always be understood. Suppose that the statistics is spatially homogeneous, so that

$$\langle \psi(\mathbf{x}) \psi(\mathbf{x}') \rangle = C(\mathbf{x} - \mathbf{x}')$$

In this case, the following relation will hold:

$$\begin{aligned} \langle \psi_{\mathbf{k}} \psi_{\mathbf{k}'} \rangle &= \int d^d x \int d^d x' \langle \psi(\mathbf{x}) \psi(\mathbf{x}') \rangle \exp[-i(\mathbf{k} \cdot \mathbf{x} + \mathbf{k}' \cdot \mathbf{x}')] \\ &= \int d^d y C(\mathbf{y}) \exp(-i\mathbf{k} \cdot \mathbf{y}) \int d^d x' \exp[-i(\mathbf{k} + \mathbf{k}') \cdot \mathbf{x}'], \end{aligned}$$

We notice that the integral  $(2\pi)^{-d} \int d^d x' \exp[-i(\mathbf{k} + \mathbf{k}') \cdot \mathbf{x}']$  is just a representation of the Dirac delta  $\delta(\mathbf{k} + \mathbf{k}')$ . We find therefore

$$\langle \psi_{\mathbf{k}} \psi_{\mathbf{k}'} \rangle = (2\pi)^d \delta(\mathbf{k} + \mathbf{k}') C_{\mathbf{k}}, \quad (2.62)$$

with

$$C_{\mathbf{k}} = \int d^d x C(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}}, \quad (2.63)$$

the Fourier transform of the correlation function. This relation between correlations of Fourier components,  $\psi_{\mathbf{k}}$ , and the Fourier transform of the correlation,  $C_{\mathbf{k}}$ , is the content of the **Wiener-Khinchin theorem**. In several circumstances, it is possible to interpret  $C(0) = \langle \psi^2 \rangle$ , as an energy density. Then, the spectral decomposition provided by Eq. (2.62):

$$C(0) = \int \frac{d^d k}{(2\pi)^d} C_{\mathbf{k}}, \quad (2.64)$$

allows to identify  $C_{\mathbf{k}}$  as the **energy spectral density** of the random field (or of the signal).

The presence of the Dirac delta in Eq. (2.62), is the consequence of the infinite volume limit and of homogeneous statistics (homogeneous statistics implies that the random field does not decay to zero at infinity). In analogy to what was observed in the case of the random walker, additional singular behaviors may arise, due to non differentiability, or even lack of continuity of  $\psi$ . Such singular behaviors would imply divergence of quantities, such as  $\langle |\nabla^p \psi|^2 \rangle$ , with  $p \geq 1$  in case of lack of continuity,  $p \geq 2$  in case of non-differentiability, and so on. Writing in analogy with Eq. (2.64)

$$\langle |\nabla^p \psi|^2 \rangle \sim \int d^d k k^{2p} C_{\mathbf{k}}, \quad (2.65)$$

we see that such divergent behaviors are associated with a power-law scaling at large  $k$  of the energy spectrum,  $C_{\mathbf{k}} \sim k^{-d-2p+\zeta}$ , with  $\zeta \geq 0$ .

From the point of view of physics, the divergence in Eq. (2.65) means simply that the value of  $\langle |\nabla^p \psi|^2 \rangle$  is determined by the smallest scales  $a$  in the system, that provides an effective ‘‘ultraviolet’’ cutoff (at  $k \sim a^{-1}$ ) in the integral.

We thus reach an alternative characterization of microscopic fluctuations, as fluctuations, whose spectrum decays slowly (or does not decay at all) at large  $k$ . An example is the instantaneous fine-grained density  $\tilde{n}(\mathbf{x}, t) = \sum_i \delta(\mathbf{x} - \mathbf{x}_i(t))$ . Its fluctuating part  $\Delta \tilde{n}(\mathbf{x}, t) = \tilde{n}(\mathbf{x}, t) - \bar{n}$ , has correlation  $C(\mathbf{x} - \mathbf{y}) = \langle \Delta \tilde{n}(\mathbf{x}, t) \Delta \tilde{n}(\mathbf{y}, t) \rangle = \bar{n} \delta(\mathbf{x} - \mathbf{y})$  (see Sect. 3.3). Fourier transforming, we obtain trivially  $C_{\mathbf{k}} = \bar{n}$ , that in fact does not decay at large  $k$ .

We conclude illustrating how working with Fourier components, simplifies the coarse-graining operations illustrated in Sect. 2.6. The coarse-graining operation described in Eq. (2.51), was in fact in the form of a convolution. Hence, the Fourier transform of the coarse grained field, is just a filtered version of the original Fourier signal:

$$\psi_V(\mathbf{x}) = (G_V \psi)(\mathbf{x}); \quad \psi_{V,\mathbf{k}} = \int d^d x \psi_V(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} = G_{V,\mathbf{k}} \psi_{\mathbf{k}}, \quad (2.66)$$

where the condition  $\int d^d x G_V(\mathbf{x}) = 1$  gives us  $G_{V,\mathbf{k} \rightarrow 0} = 1$ , while, for  $kV^{1/d} \gg 1$ , we must have  $G_{V,\mathbf{k}} \rightarrow 0$ .

## 2.9 Problems

**Problems 1** The two random variable  $x$  and  $y$  are distributed in the square  $[0, 1] \otimes [0, 1]$  with the law  $\rho(x, y) = C(1 - x - y)\theta(1 - x - y)$ , with  $C$  a constant and  $\theta$  the Heaviside step function ( $\theta(z > 0) = 1$ ,  $\theta(z < 0) = 0$ ). Calculate the conditional PDF  $\rho(x|y > 0.5)$  and the conditional probability  $P(y > 0.5|x)$ .

**Problems 2** A rather poor target shooter, the few times that hits the target, hits it at a random position. Suppose the target to be a radius  $R$  circle. Calculate the PDF that a hit on target, occurs at radius  $r$ . Calculate the probability that a hit on target, falls at  $r < R/2$ .

**Problems 3** The velocity PDF for a molecule of a monoatomic ideal gas at thermodynamic equilibrium, is given by the Maxwell distribution

$$\rho_1(\mathbf{v}) = \left(\frac{m}{2\pi KT}\right)^{3/2} \exp\left(-\frac{m\mathbf{v}^2}{2KT}\right), \quad (2.67)$$

where  $m$  is the mass of the molecule,  $T$  is the temperature and  $K$  is the Boltzmann constant. Indicate by  $N$  the number of molecules in the gas, and by  $V$  the volume in which it is contained.

- Write down the one-molecule PDF  $\rho_1(\mathbf{x}, \mathbf{v})$ , and the joint PDF, for all the molecules in the gas  $\rho_N(\{\mathbf{r}_i, \mathbf{v}_i\})$ .
- Calculate the PDF entropy  $S = S(T, V, N) = -\langle \ln \rho_N \rangle$ . How does it compare with standard expressions for the thermodynamic entropy of a gas?
- Estimate numerically the internal energy  $E = \sum_{i=1}^N \frac{1}{2} m |\mathbf{v}_i|^2$ , and the amplitude of its fluctuations, given values of the parameters  $K \simeq 1.4 \cdot 10^{-23} \text{J K}^{-1}$ ,  $N \sim 10^{20}$ ,  $V \simeq 1 \text{m}^3$ ,  $T \sim 300 \text{K}$ ,  $m \simeq 7 \cdot 10^{-27} \text{Kg}$ .

**Problems 4** Adapt the derivation of the evolution equations for the moments  $x^n$ , described in Sect. 2.7.1, to calculate the evolution of the PDF entropy of the walker. Prove that Eq. (2.56) leads to an entropy that is a monotonously increasing function of time. Repeat the calculation for the extinction process, and verify that entropy growth, in that case, is not guaranteed. Provide an example of a situation of (temporary) entropy growth for an extinction process.

## 2.10 Further Reading

An elementary presentation of probability theory, with lots of useful exercises is contained in:

- S. Lipschutz, *Theory and Problems of Probability* (Schaum outline series, McGraw Hill, 1968. Freely available online)

The reference book for probability and basic stochastic process theory, however, is:

- W.Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, 1968)

More on the concept of entropy and on information theory:

- A. Papoulis, S.U. Pillai, *Probability, Random Variables and Stochastic Processes* (McGraw Hill, 2002)

Good reference books for stochastic processes:

- C. Gardiner, *Handbook of Stochastic Methods* (Springer, 1986)
- P.E. Kloeden, E. Platen, *Numerical Solution of Stochastic Differential Equations* (Springer, 1992)
- R.L. Stratonovich, *Topics in the Theory of Random Noise. Volume I* (Gordon and Breach, 1963)

I have taken the material for anomalous diffusion from:

- J.-P. Bouchaud, A. Georges, Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications. *Phys. Rep.* **195**, 127 (1990)

More on this also in:

- D. Sornette, *Critical Phenomena in Natural Sciences* (Springer, 2000)

## Appendix

### A.1 The Central Limit Theorem

We want to determine the form of the probability distribution for sums of independent identically distributed random variables (i.i.d.) in the form

$$X_N = \sum_{k=1}^N x_k.$$

A fundamental question is the existence of limit forms at large  $N$  for the PDF  $\rho(X_N)$ . Such limit distributions indeed exist, and their form depends solely on the behavior of the PDF  $\rho(x)$  at large values of the argument: the so-called **tails of the distribution**. In particular, the existence of the first moments of the distribution is crucial in the determination of the limit form for  $\rho(X_N)$ .

We recall that the  $n$ th moment of a distribution will exist,

$$\langle x^n \rangle = \int dx x^n \rho(x) < \infty,$$

provided  $\rho(x)$  goes to zero at infinity faster than  $x^{-1-n}$ .

We can distinguish, substantially, three cases:

- Both the variance  $\sigma_x^2$  and the mean  $\mu_x$  are finite, that is the situation, assumed in Sect. 2.3, for a thermodynamic limit to occur. In this case,  $\rho(X_N)$  will be a gaussian distribution, with mean  $\mu_{X_N} = N\mu_x$  and variance  $\sigma_{X_N}^2 = N\sigma_x^2$ .
- The mean  $\mu_x$  is finite, but the variance  $\sigma_x^2$  is infinite, meaning that, for  $x \rightarrow \infty$ ,  $\rho(x) \sim x^{-1-\alpha}$  (within logarithms), with  $1 < \alpha < 2$ . In this case, again,  $\mu_{X_N} = N\mu_x$ , but the deviations  $\hat{X}_N = X_N - \mu_{X_N}$  are distributed with a so-called Lévy law of index  $\alpha$ , whose asymptotic behavior at large  $\hat{X}_N$  is  $\rho(\hat{X}_N) \sim \hat{X}_N^{-1-\alpha}$ .
- Also the mean  $\mu_x$  is infinite. This is typically realized by an asymmetric distribution  $\rho(x)$ , with asymptotic behavior at large  $x$ :  $\rho(x) \sim x^{-1-\alpha}$ , with  $0 < \alpha < 1$  (for  $\alpha < 0$ , the PDF would not be normalized). In this case, the limit distribution for  $X_N$  would be a Lévy law of index  $\alpha$ , whose asymptotic behavior at large  $X_N$  is  $\rho(X_N) \sim X_N^{-1-\alpha}$ .

The form of the limit distribution, can be calculated, exploiting the important property that the characteristic function of a sum of random variables, is the product of the characteristic functions of the addends in the sum. The PDF of a sum of independent random variables, is in fact the convolution of the PDF's of the addends:

$$\rho_{x+y}(z) = \int dy \rho_x(z-y)\rho_y(y),$$

and the characteristic function of  $x+y$ , being the Fourier transform of a convolution, will be the product

$$Z_{x+y}(j) = Z_x(j)Z_y(j).$$

Thus, in the case of a sum of i.i.d. random variables:

$$Z_{X_N}(j) = (Z_x(j))^N. \tag{2.68}$$

This is the quantity on which we shall focus, to determine the asymptotic behavior, for  $N \rightarrow \infty$ , of the PDF  $\rho(X_N)$ .

### A.1.1 The Gaussian Case

If the mean and the variance of the random variable  $x$  are both finite, we know from the law of large numbers that  $\mu_{X_N} = N\mu_x$  and  $\sigma_{X_N}^2 = N\sigma_x^2$ . We scale out the dependence of the PDF  $\rho(X_N)$  on the parameters  $\mu_{X_N}$  and  $\sigma_{X_N}^2$ , by considering a rescaled version of the deviation from the mean  $\hat{X}_N = X_N - \mu_{X_N}$ :  $Y_N = \hat{X}_N/\sigma_{X_N}$ . We verify that the limit  $\rho = \lim_{N \rightarrow \infty} \rho_{Y_N}$  exists, and is indeed the Gaussian.

We have for the characteristic function of the rescaled variable  $Y_N$ :

$$\begin{aligned} Z_{Y_N}(j) &= \int dY_N \rho_{Y_N}(Y_N) (ijY_N) \\ &= \int d\hat{X}_N \rho_{\hat{X}_N}(\hat{X}_N) \exp(ij\hat{X}_N/\sigma_{X_N}) = Z_{\hat{X}_N}(j/\sigma_{X_N}), \end{aligned} \quad (2.69)$$

and, from Eq. (2.68):

$$Z_{Y_N}(j) = (Z_{\hat{x}}(j/\sigma_{X_N}))^N, \quad (2.70)$$

where  $\hat{x} = x - \mu_x$ . Taking the  $N \rightarrow \infty$  limit in  $\rho_{Y_N}$ , corresponds to taking the limit  $j/\sigma_{X_N} \rightarrow 0$  in  $Z_{\hat{x}}(j/\sigma_{X_N})$ . We can proceed by Taylor expansion. Since  $\langle \hat{x} \rangle = 0$ , the Taylor expansion of  $Z_{\hat{x}}$ , does not contain the linear term:

$$Z_{\hat{x}}(j) = 1 - \frac{1}{2}\sigma_x^2 j^2 + o(j^2). \quad (2.71)$$

Hence, substituting into Eq. (2.70), and using  $\sigma_{X_N}^2 = N\sigma_x^2$ :

$$Z(j) = \lim_{N \rightarrow \infty} Z_{Y_N}(j) = \lim_{N \rightarrow \infty} \left(1 - \frac{j^2}{2N} + o(N^{-1})\right)^N = \exp(-j^2/2). \quad (2.72)$$

Inverse Fourier transforming, we find that the limit of the distribution for  $Y_N$  is the Gaussian

$$\rho(Y_N) = \frac{1}{(2\pi)^{1/2}} \exp(-Y_N^2/2), \quad (2.73)$$

as claimed.

The relevance of this result to the random walk dynamics, described in Sect. 2.7.1, should be apparent. If we attach a time label  $t_k = k\Delta t$  to each random variable  $x_k$ , and take  $\mu_x = 0$ , the sum  $X_N$  will be precisely the displacement of a random walker, that, in the time  $t_N = N\Delta t$ , has performed  $N$  independent steps  $x_k$ . What Eq. (2.73) tells us, is that the density profile of a population of walkers, starting from a common initial position, will have, after a sufficiently long time (provided the domain is infinite), a Gaussian profile. This is in fact the result in Eq. (2.57), where the  $N \rightarrow \infty$  limit was obtained implicitly with the continuous limit  $t/\Delta t \rightarrow \infty$ .

We have proved that, for fixed  $Y$ , the PDF  $\rho_{Y_N}(Y)$  has the limit  $\rho(Y)$ , given in Eq. (2.73). The natural question then arises about the range of  $Y_N$  for which the result in Eq. (2.73), for large *but finite*  $N$ , holds. For sure,  $\rho_{Y_N}(Y)$  will begin to be sensitive to the properties of the tails of  $\rho_x$ , when  $X_N - \mu_{X_N} \sim N\sigma_x$ , i.e. when  $Y_N \sim N^{1/2}$ . For instance, if  $\rho(x) = 0$  for  $|x - \mu_x| > \Delta x$  (as in the case of the jump distribution of a random walker),  $\rho_{Y_N}$  will surely be zero for  $|Y_N| > N^{1/2}$ . In other words, the central limit result of Gaussian statistics, for a large but finite sum of i.i.d.

random variables, with finite mean and variance, will hold only far from the tails of the distribution.

We can provide a quantitative estimate of this effect, in the case the first correction to  $Z_{\hat{x}}$  is a non-zero third moment  $\langle \hat{x}^3 \rangle \neq 0$ :

$$Z_{\hat{x}}(j) = 1 - \frac{1}{2}\sigma_x^2 j^2 - \frac{i}{6}\langle \hat{x}^3 \rangle j^3 + o(j^3).$$

Substituting into Eq. (2.70), we are able to include the leading large  $N$  correction in Eq. (2.72):

$$Z_{Y_N}(j) = \left(1 + i\alpha s_3 N^{-1/2} j^3 + o(N^{-1/2})\right) \exp(-j^2/2),$$

where  $\alpha$  is a numerical coefficient, and  $s_3 = \langle \hat{x}^3 \rangle / \sigma_x^3$  is the normalized third moment (so-called skewness) of the distribution  $\rho(\hat{x})$ . Inverse Fourier transforming, we find the correction to Eq. (2.73):

$$\frac{\rho_{Y_N} - \rho}{\rho} \sim s_3 N^{-1/2} Y_N^3. \quad (2.74)$$

In order for the central limit to hold, we need that  $|Y_N| \ll s_3^{1/3} N^{1/6}$ , i.e.  $|X_N - \mu_{X_N}| \ll s_3^{1/3} N^{2/3}$ .

### A.1.2 Lévy Distributions

We pass to consider the case in which the PDF  $\rho(x)$  does not have its first or second moment. We have seen that this corresponds to a power law behavior in the tails of the distribution:  $\rho(x) \sim x^{-1-\alpha}$ , with  $0 < \alpha < 1$  in the first case,  $1 < \alpha < 2$  in the second. The presence of a power law in the tails of the distribution,  $\rho(x) \sim x^{-1-\alpha}$ , will be associated, in general, with non-existence of moments  $\langle x^n \rangle$  with  $n > \alpha$ . Thus, the Taylor expansion in  $j = 0$  of the characteristic function  $Z_x$ , will stop at  $\text{int}(\alpha)$ , and it is possible to see that the remnant in the Taylor expansion is in the form  $c|j|^\alpha$ , with  $c$  a numerical constant. If  $\alpha > 2$ , this non-analyticity will not modify the limiting form of  $Z_{\hat{x}}$ , given in Eq. (2.71). If, on the other hand, either  $0 < \alpha < 1$ , or  $1 < \alpha < 2$  and  $\mu_x = 0$ , this form must be replaced by

$$Z_x(j) \simeq 1 - c|j|^\alpha. \quad (2.75)$$

We thus see that, contrary to the finite  $\mu_x$  and  $\sigma_x^2$  case, the behavior of the tails of the PDF  $\rho_x$  is reflected in the behavior, for  $j \rightarrow 0$ , of the characteristic function  $Z_x$ .

To determine the PDF  $\rho_{X_N}$ , we resort again to the method of characteristic functions. We find immediately the result  $Z_{X_N}(j) = (1 - c|j|^\alpha)^N$ , which suggests us that,

in order to obtain a limit distribution, it is necessary to rescale  $X_N$ . From Eq. (2.69), we see that the proper rescaling is

$$Y_N = N^{-\alpha} X_N.$$

This leads to the asymptotic form of the characteristic function

$$Z(J) = \lim_{N \rightarrow \infty} Z_{Y_N}(j) = \lim_{N \rightarrow \infty} \left( 1 - \frac{c|j|^\alpha}{N} + o(N^{-1}) \right)^N = \exp(-c|J|^\alpha). \quad (2.76)$$

The limit distribution  $\rho(Y_N)$ , obtained inverse Fourier transforming Eq. (2.76), is called a **Lévy distribution** (or **stable distribution**) of order  $\alpha$ . Its most important property, revealed by the non-analyticity of  $Z$  in  $j = 0$ , is the power law behavior in the tails of  $\rho(Y_N)$ :  $\rho(Y_N) \sim Y_N^{-1-\alpha}$ , that is the same as for the individual variable  $x$ . The fact that  $Y_N$ , that is a sum of infinite-mean i.i.d. random variables, is still an infinite-mean random variable, is not surprising. The fact that the scaling behavior, in the tails of  $\rho_x$  and  $\rho = \lim_{N \rightarrow \infty} \rho_{Y_N}$ , is identical, reflects the different origin of the limit behavior of the  $\rho_{Y_N}$  in the Gaussian and in the Lévy case. While in the Gaussian case, the form of the limit distribution does not probe the tails of  $\rho_x$ , in any way other than existence of the first moments of the distribution, in the Lévy case, it is precisely the scaling in the tails that determines the form of the limit distribution. In fact, we can interpret the law of large number result,  $\langle x \rangle_N \rightarrow \mu_x$ , as a manifestation of the fact that each  $x_k$  contributes to  $X_N$  a term of the same order  $\sim \mu_x$ . This property must apparently be lost in the  $\mu_x \rightarrow \infty$  case. What happens is that the value of  $X_N$  is determined typically by the largest  $x_k$  in the sequence  $\{x_k, k = 1, \dots, N\}$ .

We can get a quantitative feeling of this phenomenon, from the observation that the values of  $x_k$ , in a typical sequence  $\{x_1, x_2, \dots, x_N\}$ , will be distributed with the PDF  $\rho_x$ . We see that, in order for a certain large value  $\bar{x}$  (or larger), in a typical sequence of  $N$ , to be observed, it is necessary that  $NP(x > \bar{x}) \gtrsim 1$ , i.e.:

$$\int_{\bar{x}}^{\infty} dx \rho_x(x) \sim x^{-\alpha} \gtrsim 1/N.$$

Thus, the largest value of  $x$  observed in a typical sequence  $\{x_1, x_2, \dots, x_N\}$  will be:

$$x_N^{\text{MAX}} \sim N^{1/\alpha}.$$

All the smaller  $x_k$ 's in the typical sequence, will be distributed with  $\rho_x$ . Hence, we can estimate

$$X_N \sim N \int_0^{x_N^{\text{MAX}}} dx x \rho_x(x) \sim N (X^{1/\alpha})^{1-\alpha} = N^{1/\alpha}.$$

The sum  $X_N$  scales with the largest typical contribution  $x_N^{\text{MAX}}$ , meaning that it is the largest contribution in the sequence  $\{x_k, k = 1, \dots, N\}$ , that dominates  $X_N$ .

As we have done in the Gaussian case, we can map the problem of summing i.i.d. random variables, to a random walk, by attaching a time label  $t_k = k\Delta t$  to each increment  $x_k$ . To have an unbiased random walk, we need  $\mu_x = 0$ , so that the interesting dynamics is the one originating from the  $1 < \alpha < 2$  regime of infinite variance—zero mean increments. The result is an infinite variance displacement  $X_N$ , in time  $t_N = N\Delta t$ . The resulting process goes under the name of **Lévy flight**, and could be used to describe migration of individuals, that have at their disposal means of locomotion of very diverse nature (think of human beings that, in a single day, can move by few meters, as well as embark on an intercontinental flight). The resulting evolution equation in the continuous limit, is an example of the non-local master equation (2.54), in which the propagation kernel is precisely the single increment distribution:  $w(x \rightarrow y) \propto \rho(x - y) \sim |x - y|^{-1-\alpha}$ .

Additional discussion of these issues can be found e.g. in [J.-P. Bouchaud and A. Georges, Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications, Phys. Rep. **195**, 127 (1990)] and in [W. Feller, An introduction to probability theory and its applications (Wiley and Sons, 1968), Vol. 1].



<http://www.springer.com/978-3-319-06187-0>

An Introduction to Thermodynamics and Statistical  
Physics

Olla, P.

2015, VII, 185 p. 22 illus., Hardcover

ISBN: 978-3-319-06187-0