Prof. Dr. G. M. Pastor
Institut für Theoretische Physik
Fachbereich Naturwissenschaften
Universität Kassel

Statistical Physics II (WS 2008/2009)

The purpose of this lecture is to cover some more advanced, yet fundamental subjects in statistical physics which are usually not presented in the introductory lectures on thermodynamics and statistical mechanics. The introductory courses usually follow the historical development, starting from the first and second laws of thermodynamics. The concepts of statistical mechanics are introduced later in order to provide a microscopic justification of the thermodynamic principles, as was done by L. Boltzmann more than 100 years ago. However, it is presently more meaningful to abandon the historical pathway and to start from the statistical perspective, based on our knowledge of quantum and classical mechanics, and to derive thermodynamics as a byproduct. This will be our standpoint while reviewing the basic concepts at the beginning of this course.

In the following we shall review the basic concepts in statistical physics including statistical ensembles and the maximum entropy principle. This should allow readers that are not quite familiar with statistical physics to profit from the more advanced topics. A basic background in classical and quantum mechanics will be assumed. Applications of statistical mechanics to Fermi and Bose systems should follow. This includes in particular a discussion of Landau's Fermi-liquid theory and Bose-Einstein condensation. In the second part of the course we shall discuss the theory of phase transitions and critical phenomena. In contrast to the first part, which concerns the statistical physics of particles, our purpose here is to move progressively into the statistical physics of fields.

1 Basic statistical concepts

A full mechanical description of the dynamics of a macroscopic system is both hopeless and not very meaningful, since we lack precise information on the exact initial conditions of all degrees of freedom and on the exact form of the Hamiltonian governing the system and its environment. Moreover, for describing the equilibrium properties a precise knowledge of all its constituents is not necessary at all. What we actually need to know is only the probability of finding the macroscopic system in each one of its possible microscopic states. It is the goal of statistical mechanics to provide such an inherently probabilistic description of macroscopic systems. This motivates a brief discussion of some basic concepts of probability theory.

1.1 Random variables and probabilities

Variables wich outcome cannot be predicted with certainty are usually known as *random variables*. We consider a random variable x with a *set of possible outcomes* S.

S may be discrete (e.g., $S = \{x_1, x_2, \ldots\}$) or continuous (e.g., $S \equiv \mathbb{R}$).

An *event* E is a subset of outcomes $E \subseteq S$, for example, $E = \{\text{even result of a dice throw}\} = \{2, 4, 6\}$. To each event $E \subseteq S$ we assign a *probabilty* $P(E)$ with the following three fundamental properties:

i) Positive definiteness: $P(E) \geq 0$.

ii) Additivity: $P(E_1) + P(E_2) = P(E_1 \cup E_2)$ if $E_1 \cap E_2$ and $E_1, E_2 \subseteq S$.
E_1 and E_2 are said to be disconnected events.

iii) Normalization: $P(S) = 1$.

Probabilities may be assigned in two ways:

i) Experimentally as

$$P(E) = \lim_{N \to \infty} \frac{N_E}{N},$$

where N_E is the number of occurences of event E after N "throws" or outcomes.

ii) Theoretically by means of an estimation based on the determination of the set of outcomes S and some hypothesis about the relative probabilities for a complete set of events. For instance, knowing that $S = \{1, 2, 3, 4, 5, 6\}$ for a dice and assuming equal probabilities $P(i) = 1/6 \quad \forall\, i \in S$, we conclude that $P(\text{even}) = 3 \cdot 1/6 = 1/2$. Due to the lack of knowledge of the precise mechanical properties of the *dice* ($\equiv$ system) and the *way of throwing it* ($\equiv$ Hamiltonian of the environment), and in the absence of any reason to believe that the dice is biased, we assume that all six possibilities of the elements in S ($\equiv$ states) are equally probable.
All probability assignments in statistical mechanics are theoretical or "subjective" (as opposed to "objective" in i) and their validity needs to be verified by contrasting them to experiment.

Excercise 1.1:
A metastable nucleus decays through β emission. What is the probability density $p(\varphi, \theta)$ for the electron to be emitted with a polar angle θ (relative to the z axis) and azimuthal angle φ? Imagine the nucleus at the coordinate origin.

$p(\varphi, \theta)\, d\varphi\, d\theta = \text{prob}\,(\text{emission with angle } \theta \in [\theta, \theta + d\theta] \text{ and angle } \varphi \in [\varphi, \varphi + d\varphi])\,.$

1.2 A single continuous random variable

We consider a continuous random variable $x \in \mathbb{R}$. The *cumulative probability function* (CPF) is defined as the probability $P(x)$ for any outcome *smaller* than x:

$$P(x) = \text{prob}\,(E \subset [-\infty, x])\,.$$

The basic properties of a CPF are

i) $P(-\infty) = 0$.

ii) $P(x)$ is monotonically increasing, i.e,

$$P(x+\Delta) = P(x) + P(\Delta) \geq P(x),$$

since any probability satisfies additivity and positiveness.

iii) Finally, the normalization condition implies $P(+\infty) = 1$.

The *probability density function* (PDF) is defined as $p(x) = \dfrac{dP}{dx}$. Consequently,

$$p(x)\,dx = \text{prob}\left(E \subset [x, x+dx]\right).$$

Notice that, in contrast to probabilities satisfying $P(x) \leq 1$, there is no upper bound for $p(x)$.

If x is a random variable, any function $F(x)$ of x is also a random variable with its own PDF, which is given by

$$p_F(f)\,df = \text{prob}\left(F(x) \in [f, f+df]\right).$$

Let x_i with $i = 1, \ldots \nu$ be the solutions of $F(x_i) = f$, we have

$$\begin{aligned} p_F(f)\,df &= \sum_{i=1}^{\nu} p(x_i)\,dx_i \\ \Rightarrow \quad p_F(f) &= \sum_{i=1}^{\nu} p(x_i) \left|\frac{dx}{dF}\right|_{x=x_i=F^{-1}(f).} \end{aligned} \tag{1.1}$$

Notice that $|dx/dF|$ is the Jacobian for the change of variables. Eq. (1.1) may be written as

$$p_F(f) = \sum_{i=1}^{\nu} p\Big(x_i = F^{-1}(f)\Big) \; \frac{1}{\left|\frac{dF}{dx}\right|_{x=x_i=F^{-1}(f).}}$$

Such a change of variables often leads to divergencies, which remain of course integrable, as one may easily verify by changing variables back to x.

Excercise 1.2:
Consider the Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\; e^{-x^2/2\,\sigma^2}$$

with $x \in \mathbb{R}$ and $f = F(x) = x^2$. Show that $p_F(f) = \dfrac{1}{\sqrt{2\,\pi\,\sigma^2}}\; \dfrac{e^{-f/2\,\sigma^2}}{\sqrt{f}}$ for $f > 0$ and $p_F(f) = 0$ for $f < 0$. Verify the normalization of $p_F(f)$.

The *expectation value* of a function $F(x)$ of the random variable x is given by

$$\langle\, F(x)\,\rangle = \int_{-\infty}^{+\infty} F(x)\,p(x)\,dx.$$

Particularly important expectation values are the *moments of the PDF*

$$\langle\, x^n \,\rangle = \int x^n \, p(x)\, dx$$

and the *characteristic function* $\chi(k)$

$$\chi(k) = \langle\, e^{-i\,k\,x} \,\rangle = \int e^{-i\,k\,x}\, p(x)\, dx,$$

which is just the *Fourier transform* of the PDF. The characteristic function is the generator of the moments:

$$\chi(k) = \left\langle \sum_{n=0}^{+\infty} \frac{(-i\,k)^n}{n!}\, x^n \right\rangle = \sum_{n=0}^{+\infty} \frac{(-i\,k)^n}{n!}\, \langle\, x^n \,\rangle,$$

from which we obtain

$$i^n \left.\frac{d^n \chi(k)}{dk^n}\right|_{k=0} = \langle\, x^n \,\rangle.$$

The PDF can be recovered from $\chi(k)$ by the inverse transformation

$$p(x) = \frac{1}{2\,\pi} \int e^{i\,k\,x}\, \chi(k)\, dk.$$

One can also easily obtain the moments around any other point x_0 from

$$e^{i\,k\,x_0}\, \chi(k) = \langle\, e^{-i\,k\,(x-x_0)} \,\rangle = \sum_{n=0}^{+\infty} \frac{(-i\,k)^n}{n!}\, \langle\, (x-x_0)^n \,\rangle.$$

The logarithm of the characteristic function $\chi(k)$ is known as the *cumulant generating function*

$$\ln \chi(k) = \sum_{n=1}^{+\infty} \frac{(-i\,k)^n}{n!}\, \langle\, x^n \,\rangle_c. \tag{1.2}$$

The cumulants are defined implicitly from the previous series expansion of $\ln \chi(k)$. Note that $\ln \chi(k=0) = \ln 1 = 0$.

The cumulants and the moments are of course related. One can obtain the cumulants from

$$\chi(k) = 1 + \underbrace{\sum_{n=1}^{+\infty} \frac{(-i\,k)^n}{n!}\, \langle\, x^n \,\rangle}_{\varepsilon} \tag{1.3}$$

and

$$\ln \chi(k) = \ln(1+\varepsilon) = \sum_{l=1}^{+\infty} (-1)^{l+1}\, \frac{\varepsilon^l}{l}. \tag{1.4}$$

Using the definition of the cumulants (1.2) and replacing ε from (1.3) in (1.4) we have

$$\sum_{n=1}^{+\infty} \frac{(-i\,k)^n}{n!}\, \langle\, x^n \,\rangle_c = \sum_{l=1}^{+\infty} \frac{(-1)^{\,l+1}}{l} \left(\sum_{m=1}^{+\infty} \frac{(-i\,k)^m}{m!}\, \langle\, x^m \,\rangle \right)^l.$$

This leads to

$$\langle x \rangle_c = \langle x \rangle \qquad \text{mean}$$

$$\langle x^2 \rangle_c = \langle x^2 \rangle - \langle x \rangle^2 \qquad \text{variance}$$

$$\langle x^3 \rangle_c = \langle x^3 \rangle - 3 \langle x^2 \rangle \langle x \rangle + 2 \langle x \rangle^3 \qquad \text{skewness}$$

$$\langle x^4 \rangle_c = \langle x^4 \rangle - 4 \langle x^3 \rangle \langle x \rangle - 3 \langle x^2 \rangle^2 + 12 \langle x^2 \rangle \langle x \rangle^2 - 6 \langle x \rangle^4 \qquad \text{curtosis (or kurtosis).}$$

A PDF can be described indistinctively in terms of its cumulants or of its moments.

1.3 Computation of moments in terms of cumulants

Theorem: The m-th moment $\langle x^m \rangle$ is obtained by considering all possible subdivisions of m points in p_n groups or connected clusters of n points each. Of course $\sum_n p_n n = m$. Each possible subdivision contributes with the product of the cumulants $\langle x^n \rangle_c$ associated to the connected cluster having n points.

Example:

$$\begin{aligned}
\langle x \rangle &= (\bullet) &&= \langle x \rangle_c \\
\langle x^2 \rangle &= (\bullet)(\bullet) + (\bullet\bullet) &&= \langle x \rangle_c^2 + \langle x^2 \rangle_c \\
\langle x^3 \rangle &= (\bullet)(\bullet)(\bullet) + 3(\bullet)(\bullet\bullet) + (\bullet\ \bullet\ \bullet) &&= \langle x \rangle_c^3 + 3 \langle x \rangle_c \langle x^2 \rangle_c + \langle x^3 \rangle_c
\end{aligned}$$

Excercise 1.3:
Obtain the expression for $\langle x^4 \rangle$ in terms of $\langle x^l \rangle_c$ and $l = 1$–4. Obtain $\langle x^l \rangle_c$ in terms of $\langle x^l \rangle$ for $l \leq 4$.

The theorem can be demonstrated by noting that

$$\begin{aligned}
\chi(k) = \sum_{m=0}^{+\infty} \frac{(-i\,k)^m}{m!} \langle x^m \rangle &= \exp\{\ln \chi(k)\} \\
&= \exp\left\{ \sum_{n=1}^{+\infty} \frac{(-i\,k)^n}{n!} \langle x^n \rangle_c \right\} \\
&= \prod_{n=1}^{+\infty} \exp\left\{ \frac{(-i\,k)^n}{n!} \langle x^n \rangle_c \right\} \\
&= \prod_{n=1}^{+\infty} \left[\sum_{p_n=0}^{+\infty} \frac{(-i\,k)^{n\,p_n}}{(n!)^{p_n}} \frac{\langle x^n \rangle_c^{p_n}}{p_n!} \right].
\end{aligned}$$

Matching the coefficients of the powers of $(i\,k)^m$ with all the possibilities yielding $\sum_n p_n = m$ we have

$$\frac{\langle x^m \rangle}{m!} = \sum_{\{p_n\}} \prod_n \frac{\langle x^n \rangle_c^{p_n}}{p_n!\,(n!)^{p_n}} \tag{1.5}$$

where the sum runs over all the possibilities of forming subgroups with $\sum_{p_n} n = m$. After rewriting Eq. (1.5) as

$$\langle\, x^m \,\rangle = \sum_{\{p_n\}} \prod_n \left(\frac{m!}{p_n!\,(n!)^{p_n}} \,\langle\, x^n \,\rangle_c^{p_n} \right)$$

we can identify the different variables and factors as follows:

n: number of points in one cluster.

p_n: number of clusters with the same number n of points inside.

$m!$: number of permutations of all the m points.

$(n!)^{p_n}$: permutations of the points within each cluster.

$p_n!$: number of permutations of the clusters with n points among them.

$\dfrac{m!}{p_n!\,(n!)^{p_n}}$: number of ways of splitting m points in $\{p_n\}$ subgroups with n points each.

1.4 Cumulant expansion in classical statistical mechanics

The characteristic function

$$\chi(k) = \int_{-\infty}^{+\infty} e^{-i\,k\,x}\, p(x)\, dx$$

can be extended analytically to complex values of k, since $p(x) \to 0$ for $x \to \pm\infty$ more rapidly than $e^{\alpha\, x}$, or simply because $p(x)$ vanishes beyond some finite bounds. One can then set $i\,k = \beta \in \mathbb{R} \geq 0$ in order to obtain

$$\chi(k = -i\,\beta) = \langle\, e^{-\beta\, x} \,\rangle,$$

and according to the definition of cumulants

$$\ln\langle\, e^{-\beta\, x} \,\rangle = \sum_{n=1}^{+\infty} \frac{(-\beta)^n}{n!} \,\langle\, x^n \,\rangle_c.$$

Replacing x by the interaction energy U of a classical gas we have

$$\ln\langle\, e^{-\beta\, U} \,\rangle = \sum_{n=1}^{+\infty} \frac{(-\beta)^n}{n!} \,\langle\, U^n \,\rangle_c.$$

This can be related to the free energy of the interacting gas since

$$-\beta\, F\,(T, V, N) = \ln Z = \ln Z_0 + \ln\langle\, e^{-\beta\, U} \,\rangle^0 \tag{1.6}$$

$$= \ln Z_0 + \sum_{l=1}^{\infty} \frac{(-\beta)^l}{l!} \,\langle\, U^l \,\rangle_c^0, \tag{1.7}$$

where $\langle\ \ \rangle^0$ indicates average over the probability distribution of the non-interacting gas and Z_0 is the partition function of the non-interacting gas. In order to demonstrate Eq.

(1.6) let us recall that

$$Z = \frac{1}{N!} \prod_i \int \frac{d^3q_i\, d^3p_i}{h^{3N}}\, e^{-\beta \sum_i p_i^2/2m}\, e^{-\beta U(q_1 \ldots q_N)}$$
$$= Z_0\, \frac{1}{N!} \prod_i \int \frac{d^3q_i\, d^3p_i}{h^{3N}}\, \frac{e^{-\beta \sum_i p_i^2/2m}}{Z_0}\, e^{-\beta U(q_1 \ldots q_N)},$$

where

$$Z_0 = \frac{1}{N!} \int \prod_i \frac{d^3p_i\, d^3q_i}{h^{3N}}\, e^{-\beta \sum_i p_i^2/2m} = \left(\frac{V}{\lambda^3}\right)^N \frac{1}{N!}$$

with $\lambda = h/\sqrt{2\,\pi\, m\, k_B\, T}$.[1] The factor $e^{-\beta \sum_i p_i^2/2m}/Z_0$ represents the probability distribution of the non-interacting gas, which is of course independent of the coordinates q_i. The fact that the non-interacting gas is uniform renders the calculation of the cumulants $\langle\, U^n \,\rangle_c^0$ entering Eq. (1.7) very simple:

$$\langle\, U^n \,\rangle_c^0 = \int \prod_{i=1}^N \left(\frac{d^3q_i}{V}\right) U(q_1 \ldots q_N)^n$$

(uniform distribution of independent particles). The expectation values of observables $O = O(p_1, \ldots p_n,\, q_1, \ldots q_n)$ can be calculated perturbatively from

$$\langle\, O \,\rangle = \frac{1}{Z}\, \frac{1}{N!} \int \prod_i \frac{dp_i\, dq_i}{h^{3N}}\, e^{-\beta \sum_i p_i^2/2m}\, e^{-\beta U(q_1 \ldots q_n)} O$$
$$= \frac{\langle\, O\, e^{-\beta U(q_1 \ldots q_n)} \,\rangle^0}{\langle\, e^{-\beta U} \,\rangle^0} = i\, \frac{\partial}{\partial k}\, \ln\langle\, e^{-i\,k\,O - \beta\, U} \,\rangle^0 \Big|_{k=0}.$$

Noting that

$$\ln\langle\, e^{-i\,k\,O - \beta\, U} \,\rangle^0 = \sum_{l\,l'} \frac{(-i\,k)^{l'}\, (-\beta)^l}{l'!\; l!}\, \langle\, O^{l'} * U^l \,\rangle_c^0$$

---

[1]

$$Z_0 = \frac{1}{N!} \int \prod_i \frac{d^3q_i\, d^3p_i}{h^{3N}}\, e^{-\beta \sum_i p_i^2/2m} = \frac{1}{N!} \prod_i \left(\int dq_i^3 \int dp_i^3\, e^{-\beta\, p_i^2/2m} \right)$$
$$= \frac{1}{N!}\, V^N \left(\frac{4\,\pi}{h^3} \int_0^{+\infty} p^2\, dp\, e^{-\beta\, p^2/2m} \right)^N = \frac{1}{N!}\, V^N \left[4\,\pi\, \frac{\sqrt{\pi}}{4}\, \left(\frac{\beta}{2\,m}\right)^{-3/2} \right]^N$$
$$\int_0^{+\infty} x^2\, e^{-\alpha\, x^2}\, dx = -\frac{\partial}{\partial \alpha} \left(\int_0^{+\infty} e^{-\alpha\, x^2}\, dx \right) = -\frac{\partial}{\partial \alpha} \left(\frac{\sqrt{\pi}}{2\,\sqrt{\alpha}} \right) = \frac{\sqrt{\pi}}{2}\, \frac{1\, \alpha^{-3/2}}{2} = \frac{\sqrt{\pi}}{4}\, \alpha^{-3/2}$$
$$Z_0 = \frac{1}{N!}\, V^N \left(\frac{\sqrt{\pi\, 2\, m\, k_B\, T}}{h} \right)^3 = \frac{1}{N!}\, V^N \,/\, \lambda^{3N}$$

with $\lambda = \dfrac{h}{\sqrt{2\,\pi\, m\, k_B\, T}}$.

we finally obtain

$$\langle\, O \,\rangle = \sum_{l=0}^{\infty} \frac{(-\beta)^l}{l!}\, \langle\, O * U^l \,\rangle_c^0,$$

which provides a high-temperature perturbation expansion of $\langle\, O \,\rangle$.

1.5 The Gaussian or normal distribution

$$p(x) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\, e^{-\frac{(x-\lambda)^2}{2\,\sigma^2}}.$$

The characteristic function has also a Gaussian form:

$$\chi(k) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \int dx\; e^{-\frac{(x-\lambda)^2}{2\,\sigma^2} - i\,k\,x}.$$

Setting $\xi = x - \lambda$ and rewriting the exponent as

$$\begin{aligned}\frac{(x-\lambda)^2}{2\,\sigma^2} + i\,k\,x &= \frac{\xi^2}{2\,\sigma^2} + i\,k\,\xi + i\,k\,\lambda \\ &= \frac{(\xi + i\,\sigma^2\,k)^2}{2\,\sigma^2} + \frac{\sigma^2\,k^2}{2} + i\,k\,\lambda\end{aligned}$$

one obtains

$$\chi(k) = e^{-i\,k\,\lambda - \frac{k^2\sigma^2}{2}}\; \underbrace{\frac{1}{\sqrt{2\,\pi\,\sigma^2}} \int_{-\infty}^{+\infty} d\xi\; e^{-\frac{(\xi + i\,\sigma^2\,k)^2}{2\,\sigma^2}}}_{=1}.$$

The cumulant generating function is simply given by

$$\boxed{\ln \chi(k) = -i\,k\,\lambda - \frac{k^2\,\sigma^2}{2}}$$

which implies

$$\begin{aligned}\langle\, x \,\rangle_c &= \langle\, x \,\rangle = \lambda, \\ \langle\, x^2 \,\rangle_c &= \sigma^2,\end{aligned}$$

and

$$\langle\, x^n \,\rangle_c = 0 \quad \text{for } n \geq 3.$$

This makes the calculations using the cluster expansion particularly simple, since the graphical expansions involve only one- and two-point clusters.

1.6 Many random variables

For more than one variable $\vec{x} = (x_1, \ldots x_N) \in \mathbb{R}^n$ the set of outcomes $S \subseteq \mathbb{R}^n$. The *joint probability distribution function (PDF)* is defined by

$$p(\vec{x}) \prod_{i=1}^{N} dx_i = \text{prob}\left\{\text{event } \vec{x}\,' \text{ in } x_i < x_i' < x_i + dx_i \;\; \forall\, i\right\},$$

which satisfies the normalization condition

$$\int p(\vec{x})\; d^N x = 1.$$

If, and only if, the variables are *independent* we have $p(\vec{x}) = \prod_{i=1}^{N} p_i(x_i)$, where $p_i(x)$ is the PDF of the random variable x_i.

The *unconditional probability density* for a subset of random variables $x_1, \ldots x_m$ is given by

$$p(x_1 \ldots x_m) = \int \prod_{i=m+1}^{N} dx_i\; p(x_1 \ldots x_m\, x_{m+1} \ldots x_N).$$

It describes the behavior of these $x_1 \ldots x_m$ variables irrespective of all the others. For instance $p(\vec{x}) = \int d^3r\; p(\vec{x}, \vec{v})$ gives the *particle density* (i.e., probability distribution for the position) irrespective of the velocity $\vec{v}$.

The *conditional PDF* $p(x_1 \ldots x_m \,|\, x_{m+1} \ldots x_N)$ describes the behavior of some variables $x_1 \ldots x_m$, subject to the constraint that the other variables $x_{m+1} \ldots x_N$ have specified values. For example, one may search for the velocity distribution at a given point $\vec{x}$, which we denote by $p(\vec{v} \,|\, \vec{x})$.

The joint probability is given by

$$p(x_1 \ldots x_m,\, x_{m+1} \ldots x_N) = p(x_{m+1} \ldots x_N)\; p(x_1 \ldots x_m \,|\, x_{m+1} \ldots x_N),$$

where $p(x_{m+1} \ldots x_N)$ is the probability density for $x_{m+1} \ldots x_N$ irrespectively of the other variables $x_1 \ldots x_m$, and $p(x_1 \ldots x_m \,|\, x_{m+1} \ldots x_N)$ is the probability of $x_1 \ldots x_m$ given the values $x_{m+1} \ldots x_N$. Thus

$$p(x_1 \ldots x_m \,|\, x_{m+1} \ldots x_N) = \frac{p(x_1 \ldots x_N)}{p(x_{m+1} \ldots x_N)},$$

where $p(x_1 \ldots x_N)$ is the number of events $x_1 \ldots x_m$, $x_{m+1} \ldots x_N$ (divided by the number N of trials) and $p(x_{m+1} \ldots x_N)$ is the number of events $x_{m+1} \ldots x_N$ (divided by the number N of trials). The *expectation values* are calculated as usual from

$$\langle\, F(\vec{x}) \,\rangle = \int d^N x\; p(\vec{x})\; F(\vec{x}).$$

The joint characteristic function is given by the Fourier transform

$$\chi(\vec{k}) = \int d\vec{x}\; e^{-i\,\vec{k}\cdot\vec{x}}\; p(\vec{x})$$

$$= \langle\, e^{-i \sum_{j=1}^{N} k_j\, x_j} \,\rangle.$$

The *joint moments* and *joint cumulants* are obtained as

$$\langle\, x_1^{n_1} \dots x_N^{n_N} \,\rangle = \frac{\partial^{n_1}}{\partial(-i\,k_1)} \cdots \frac{\partial^{n_N}}{\partial(-i\,k_N)}\, \chi(\vec{k}) \,\Big|_{\vec{k}=0}$$

and

$$\langle\, x_1^{n_1} * \dots * x_N^{n_N} \,\rangle = \frac{\partial^{n_1}}{\partial(-i\,k_1)} \cdots \frac{\partial^{n_N}}{\partial(-i\,k_N)}\, \ln\chi(\vec{k}) \,\Big|_{\vec{k}=0}.$$

The *graphical relation* between moments and cumulants, that was demonstrated for one variable, also applies to N variables. For instance,

$$\langle\, x_1\, x_2 \,\rangle = \Big(\bullet_{\,1}\Big)\Big(\bullet_{\,2}\Big) + \Big(\bullet_{\,1}\ \bullet_{\,2}\Big) = \langle\, x_1 \,\rangle_c \langle\, x_2 \,\rangle_c + \langle\, x_1 * x_2 \,\rangle_c$$

or

$$\begin{aligned}\langle\, x_1^2\, x_2 \,\rangle &= \Big(\bullet_{\,1}\Big)\Big(\bullet_{\,2}\Big)\Big(\bullet_{\,1}\Big) + \Big(\bullet_{\,1}\ \bullet_{\,1}\Big)\Big(\bullet_{\,2}\Big) + 2\Big(\bullet_{\,1}\ \bullet_{\,2}\Big)\Big(\bullet_{\,1}\Big) + \Big(\bullet_{\,1}\ \bullet_{\,2}\ \bullet_{\,1}\Big) \\ &= \langle\, x_1 \,\rangle_c^2 \langle\, x_2 \,\rangle_c + \langle\, x_1^2 \,\rangle_c \langle\, x_2 \,\rangle_c + 2\,\langle\, x_1 * x_2 \,\rangle_c \langle\, x_1 \,\rangle + \langle\, x_1^2 * x_2 \,\rangle_c.\end{aligned}$$

Joint cumulants of independent random variables

It is easy to see that $\langle\, x_\alpha * x_\beta \,\rangle_c = 0$ if x_α and x_β are *independent random variables*. Let the PDF be of the form

$$p(\vec{x}) = p_1(x_1 \dots x_m)\; p_2(x_{m+1} \dots x_N).$$

Then

$$\begin{aligned}\chi(\vec{k}) &= \int d\vec{x}\; e^{-i\,\vec{k}\cdot\vec{x}}\; p(\vec{x}) \\ &= \langle\, e^{-i\sum_{j=1}^{m} k_j\, x_j} \,\rangle_1\; \langle\, e^{-i\sum_{j=m+1}^{N} k_j\, x_j} \,\rangle_2 \\ &= \chi_1(\vec{k}_1)\; \chi_2(\vec{k}_2).\end{aligned}$$

The joint moment

$$\langle\, x_\alpha\, x_\beta \,\rangle = \frac{\partial\chi_1}{\partial(-i\,k_\alpha)}\, \frac{\partial\chi_2}{\partial(-i\,k_\beta)} = \langle\, x_\alpha \,\rangle_1 \langle\, x_\beta \,\rangle_2$$

for $1 \le \alpha \le m$ and $m+1 \le \beta \le N$. It follows that

$$\ln\chi(\vec{k}) = \ln\chi_1(\vec{k}_1) + \ln\chi_2(\vec{k}_2)$$

and consequently

$$\frac{\partial}{\partial k_\alpha}\, \frac{\partial}{\partial k_\beta}\, \ln\chi(\vec{k}) = 0$$

if $1 \le \alpha \le m$ and $m+1 \le \beta \le N$. The joint cumulant $\langle\, x_\alpha * x_\beta \,\rangle_c$ is also known as the *connected* correlation.

1.7 Gaussian distribution

The generalization to N variables of the normal distribution has the form

$$p(\vec{x}) = \frac{1}{\sqrt{(2\,\pi)^N\ \det(\Sigma)}}\ \exp\Big\{\,-\,\frac{1}{2}\,\sum_{m\,n}\,\Big(\Sigma^{-1}\Big)_{mn}\,(x_m-\lambda_m)\ (x_n-\lambda_n)\,\Big\},$$

where Σ is a positive-definite symmetric matrix and Σ^{-1} refers to its inverse. Note that Σ^{-1} is also positive definite. In other words, the argument of the exponential is an arbitrary positive-definite quadratic form.

The characteristic function is given by

$$\chi(\vec{k}) = \int d\vec{x}\ e^{-i\,\vec{k}\cdot\vec{x}}\ e^{-\frac{1}{2}\,(\vec{x}-\vec{\lambda})\cdot\Sigma^{-1}\,(\vec{x}-\vec{\lambda})}\ \frac{1}{\sqrt{(2\,\pi)^N\ \det(\Sigma)}},$$

where we have introduced $\vec{k} = (k_1, \ldots k_N)$ and $\vec{\lambda} = (\lambda_1, \ldots \lambda_N)$.

One may easily verify the normalization of $p(\vec{x})$ and compute $\chi(\vec{k})$ by changing variables to $\vec{y} = \vec{x} - \vec{\lambda}$, so that the Gaussian distribution is centered at the origin of the coordinate system, and by performing an orthogonal transformation U such that $U^t\,\Sigma^{-1}\,U = \sigma_m^{-2}\,\delta_{mn}$ is diagonal (Σ^{-1} is symmetric). The Jacobian of the orthogonal transformation being equal to 1 ($\det U = 1$) and denoting the eigenvalues of Σ^{-1} by $1/\sigma_m^2 > 0$ we have

$$\Big(U^t\,\Sigma^{-1}\,U\Big)_{mn} = \delta_{mn}\,\frac{1}{\sigma_m^2} \quad \text{with} \quad U^t\,U = \mathbb{1}$$

and

$$\Big(U^t\,\Sigma\,U\Big)_{mn} = \delta_{mn}\,\sigma_m^2.$$

Setting $U\vec{\xi} = \vec{y} = (\vec{x} - \vec{\lambda})$ we have

$$\chi(\vec{k}) = \frac{e^{-i\,\vec{k}\cdot\vec{\lambda}}}{\sqrt{(2\,\pi)^N\ \det(\Sigma)}}\ \int d\vec{\xi}\,e^{-i\,\vec{k}\cdot U\vec{\xi}}\,\exp\Big\{\,-\,\frac{1}{2}\,\underbrace{\Big(U\vec{\xi}\,\Big)\cdot\Sigma^{-1}\,U\vec{\xi}}_{\vec{\xi}\cdot U^t\,\Sigma^{-1}\,U\vec{\xi}}\,\Big\}.$$

If we set for a moment $\vec{k} = 0$ to verify the normalization, we see that the integral splits in a product of N one-dimensional Gaussians each yielding an integral $\sqrt{2\,\pi\,\sigma_m^2}$, so that $\prod_m \Big(\sqrt{2\,\pi\,\sigma_m^2}\Big) = \sqrt{(2\,\pi)^N\ \det(\Sigma)}$. The joint PDF is therefore properly normalized.

In order to compute $\chi(\vec{k})$ one can use the result for one-dimensional Gaussians for $\vec{k}\,' = U^t\,\vec{k}$ noting that $(U^t\,\Sigma^{-1}\,U)_{mn} = \dfrac{\delta_{mn}}{\sigma_m^2}$. In this way one has

$$\begin{aligned}\chi(\vec{k}) &= e^{-i\,\vec{k}\cdot\vec{\lambda}}\ \prod_m e^{-\frac{k_m'^2\,\sigma_m^2}{2}} \\ &= e^{-i\,\vec{k}\cdot\vec{\lambda}}\,\exp\Big\{\,-\,\frac{1}{2}\sum_m\,k_m'^2\,\sigma_m^2\,\Big\}.\end{aligned}$$

Using that

$$\sum_m k_m'^2\,\sigma_m^2 = \sum_m k_m'\,\sigma_m^2\,k_m' = \vec{k}'\cdot U^t\,\Sigma\,U\,\vec{k}' =$$

$$= U^t\,\vec{k}\cdot U^t\,\Sigma\,\vec{k} = \vec{k}\cdot\Sigma\,\vec{k} = \sum_{mn}\Sigma_{mn}\,k_m\,k_n,$$

we finally obtain

$$\chi(\vec{k}) = e^{-i\,\vec{k}\cdot\vec{\lambda}-\frac{1}{2}\,\vec{k}\cdot\Sigma\,\vec{k}}$$

$$= \exp\Big\{-i\sum_m k_m\,\lambda_m - \frac{1}{2}\sum_{mn}\Sigma_{mn}\,k_m\,k_n\Big\}.$$

Consequently,

$$\ln\chi(\vec{k}) = -i\sum_m k_m\,\lambda_m - \frac{1}{2}\sum_{mn}\Sigma_{mn}\,k_m\,k_n,$$

which implies

$$\langle\, x_m\,\rangle_c = \lambda_m,$$

$$\langle\, x_m * x_n\,\rangle_c = \Sigma_{mn},$$

and all higher cumulants vanish.

In the special case of vanishing mean values, i.e., $\lambda_m = 0 \quad \forall\, m$, we have that all odd cumulants vanish. Thus, all odd moments vanish and any even moment is given by the sum of the products of cumulants obtained from all possible ways of forming pairs of variables. For instance,

$$\langle\, x_\alpha\,x_\beta\,x_\gamma\,x_\delta\,\rangle = \langle\, x_\alpha * x_\beta\,\rangle_c\,\langle\, x_\gamma * x_\delta\,\rangle_c +$$

$$+\,\langle\, x_\alpha * x_\gamma\,\rangle_c\,\langle\, x_\beta * x_\delta\,\rangle_c$$

$$+\,\langle\, x_\alpha * x_\delta\,\rangle_c\,\langle\, x_\beta * x_\gamma\,\rangle_c.$$

This is analogous to Wick's theorem in many-body Green's function theory.

1.8 The central limit theorem

We consider the average $\bar{x} = \frac{1}{N}\sum_{\nu=1}^N x_\nu$ of N random variables $x_1,\ldots x_N$ having the joint PDF $p(x_1,\ldots x_N)$. The PDF for the random variable $\bar{x}$ is

$$p_{\bar{x}}(x) = \int dx_1\ldots dx_N\;p(x_1,\ldots x_N)\;\delta\left(x - \frac{1}{N}\sum_{\nu=1}^N x_\nu\right)$$

and the characteristic function is

$$\begin{aligned}\chi_{\bar{x}}(k) &= \int dx\; p_{\bar{x}}(x)\; e^{-i\,k\,x} \\ &= \int dx\; e^{-i\,k\,x} \int dx_1 \ldots dx_N\; p(x_1, \ldots x_N)\; \delta\left(x - \frac{1}{N}\sum_{\nu} x_\nu\right) \\ &= \int dx_1 \ldots dx_N\; p(x_1, \ldots x_N)\; e^{-i\,\frac{k}{N}\,\sum_{\nu=1}^{N} x_\nu} \\ &= \chi_p\left(k_1 = \frac{k}{N}, \ldots k_N = \frac{k}{N}\right),\end{aligned}$$

where $\chi_p(\vec{k})$ is the characteristic function of $p(x_1 \ldots x_N)$ [$\chi_p : \mathbb{R}^N \to \mathbb{R}$].

Let us now assume that the x_ν are independent variables having all the same PDF $p_1(x)$, i.e.,

$$p(x_1, \ldots x_N) = \prod_{\nu=1}^{N} p_1(x_\nu),$$

where the index 1 in $p_1(x_\nu)$ indicates that only one variable is involved. The characteristic function for $\bar{x} = \dfrac{1}{N}\sum x_\nu$ takes then the form

$$\chi_{\bar{x}}(k) = \prod_{\nu=1}^{N}\left(\int dx_\nu\; p_1(x_\nu)\; e^{-i\,\frac{k}{N}\,x_\nu}\right) = \left[\chi_1\left(\frac{k}{N}\right)\right]^N,$$

where χ_1 is the characteristic function of $p_1(x)$. The cumulant generating function for $\bar{x}$ reads

$$\ln \chi_{\bar{x}}(k) = N\; \ln \chi_1\left(\frac{k}{N}\right). \tag{1.8}$$

We can now expand χ_1 or $\ln \chi_1$ for small k/N, i.e., large N, in terms of the cumulants of the probability distribution for one variable:

$$\ln \chi_1\left(\frac{k}{N}\right) = -i\,\frac{k}{N}\,\langle\, x\,\rangle_c - \frac{1}{2}\left(\frac{k}{N}\right)^2 \langle\, x^2\,\rangle_c + O(N^{-3}). \tag{1.9}$$

Combining Eqs. (1.8) and (1.9) we have

$$\ln \chi_{\bar{x}}(k) = -i\,k\,\langle\, x\,\rangle_c - \frac{1}{2}\,k^2\,\frac{\langle\, x^2\,\rangle_c}{N} + O(N^{-3}).$$

The average of $\bar{x}$ is, as expected, equal to $\langle\, x\,\rangle$ and the variance is given by $\langle\, \bar{x}^2\,\rangle_c = \dfrac{\langle\, x^2\,\rangle_c}{N}$, which vanishes for $N \to \infty$. Transforming back to $p_{\bar{x}}(x)$ we have asymptotically

$$p_{\bar{x}}(x) \cong \frac{1}{\sqrt{2\,\pi\,\sigma_{\bar{x}}^2}}\,\exp\left\{-\frac{1}{2\,\sigma_{\bar{x}}^2}\,(x - \langle\, x\,\rangle)^2\right\}$$

with $\sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{N}$. The distribution of the average follows a Gaussian distribution for $N \to \infty$. Notice that this holds independently of the form of the PDF, i.e., for any $p_1(x)$ of

the random variable x. The details of the physical processes behind $p_1(x)$ are therefore irrelevant. This important result is known as the *central limit theorem*. The same holds for the sum of N independent random variables $X = \sum_\nu x_\nu$. In this case the average $\langle X \rangle = N \langle x \rangle$ and $\langle X^2 \rangle_c = N \langle x^2 \rangle_c$, so that also in this case the fluctuations around the mean as measured by the standard deviation $T_{\bar{x}} = \sqrt{\langle X^2 \rangle_c} = \sqrt{N}\,\tau$ scale with $\sqrt{N}$. The relative fluctuation $\dfrac{\sigma_{\bar{x}}}{\langle \bar{x} \rangle} \sim \dfrac{1}{\sqrt{N}} \to 0$ for $N \to \infty$.

Excercise 1.4:
In many specialized populations (e.g., in the ensemble of marathon runners) there are very little fluctuations among the performances of the best individuals. Is this a consequence of optimal performance selection?

i) Let $\{r_\nu\}$ be a set of n random numbers $0 \le r_\nu \le 1$ with a probability density distribution $p(r)$ ($r \in [0,1]$). Calculate in terms of $p(r)$ the probability $p_n(x)\,dx$ that the largest value of r_ν is in the range $[x, x+dx]$. $p(x)$ is the probability density for the random variable $x = \max\{r_1, \ldots r_n\}$. Verify the normalization of $p_n(x)$ for an arbitrary $p(r)$.

ii) Assume r is uniformly distributed in $[0,1]$. Calculate $\langle x \rangle$ and $\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2$ as a function of n and analyze the corresponding limits for large n.

Excercise 1.5:
Benford's law [F. Benford, "The law of anomalous numbers", Proc. Amer. Phil. Soc. **78**, 551 (1938)] is a phenomenological law also called first digit law or leading digit phenomenon. The law states that the first digit of data appearing in listings and statistic tables (e.g., population, addresses, cost data, stock prices, death rates, etc.) follows a particular non-uniform distribution. The averages and probable error recorded in Benford's original paper are

First digit	1	2	3	4	5	6	7	8	9
Occurrence	30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7
in %	±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.2

It has been said that the validity of this law may be empirically verified (qualitatively) by noting that the first pages of the tables of logarithms are much more worn out than the later pages [Newcomb, Amer. J. Math. **4**, 39 (1881)]. However, this is not very helpful, since 21st century students hardly ever had a table of logarithms in their hands. The phenomenon is counter-intuitive since naively one would have expected that the probability of each digit 1–9 would have been $\frac{1}{9} \equiv 11.1\%$. It is important (actually crucial) to know that Benford's law applies to data that are not dimensionless, so that the numerical values depend on the units. If there is a universal probability distribution for the data $p(x)$ over several orders of magnitude, it must be invariant under a change of scale. Find a formula for the probability $p(i)$ that the first digit is i which fits Benford's data. While Benford's law applies unquestionably to many situations a reasonable explanation has only been given recently [T. P. Hill, Amer. Sci. **86**, 358 (1998)].

Try first to find your own solution and then follow separately each of the hints given below.

Hint 1
Consider a given data characterized by the random variable x, and let $p(x)$ be the probability density of finding the value x. Assume $p(x)$ is invariant under a change of scale. This means, for example, that the $p(x)$ of the price of goods must be the same function if you express the price in US$, Euro or Argentine pesos! Mathematically this means $p(kx) = f(k)\,p(x)$ where $f(k)$ is some function of k.

i) Find $f(k)$.
ii) Take advantage of the relation and find $p(x)$.
iii) Normalize $p(x)$ assuming $10^{q_1} \leq x \leq 10^{q_2}$ with $q_i \in \mathbb{Z}$, i.e., x runs over several orders of magnitude but has natural cutoffs.
iv) Find an expression for $p(i)$, i.e., for the probability that the first digit is i, in terms of $p(x)$.
v) Solve this expression and obtain $p(i)$ in a simple form.
vi) Verify the normalization of $p(i)$.
vii) Compare it with the experimental data of Benford.

Hint 2
Benford's law applies not only to scale-invariant data, but also to data obtained from a variety of different sources. Consider numbers that are obtained as the result of multiplications of a large number of random numbers. For example stock prices in Wall Street change, say, every minute or hour ν by a factor r_ν, which has some unspecified distribution $p(r)$. The random variable we are interested in is $x = \prod_{\nu=1}^{N} r_\nu$, where $r_\nu \geq 0$ is a random variable with some probability distribution.

i) Consider the random variable $l = \ln x$. Can you use the central limit theorem to infer the PDF $p(l)$ for l?
ii) Write down $p(l)$ assuming $\bar{l} = \langle\, l \,\rangle$ and $\sigma^2 = \langle\, l^2 \,\rangle - \langle\, l \,\rangle^2$.
iii) Obtain the PDF $p(x)$ of the variable x.
iv) Find an expression for $p_i = \{\text{prob. of the 1st digit being equal to } i\}$ in terms of $p(x)$. Assume $10^{q_1} \leq x \leq 10^{q_2}$.
v) Express p_i in terms of $p(l)$.
vi) Solve p_i ignoring multiplicative constants independent of i by assuming that $p(l) \simeq p(q)$ for $q + \ln_{10} i \leq l \leq q + \ln(i+1)$.
vii) Find the normalization constant for p_i and compare with Benford's data and with the solution from hint 1.

1.9 Information content of a probability distribution or probability density function

Consider a random variable with a discrete set of outcomes $S = \{x_i,\ i = 1, \ldots M\}$ having the probabilities p_i. Suppose we construct a message $x_1 \ldots x_N$ with N independent outcomes of the random variable x_i. We intend to quantify the possible information content of such a message as a function of the probability distribution $\{p_i,\ i = 1, \ldots M\}$. Analyzing the number of different possible messages will allow us to infer how much of the apparent information content of the message is already contained in the probability distribution.

For instance, if the probability distribution is $\{p_1 = 1 \text{ and } p_i = 0 \ \ \forall\, i > 1\}$ there is just one possible message $(x_1 \dots x_1)$ and actually no information can be conveyed. All the information is in the probability distribution (PD). In the other extreme case, where x_i is uniformly distributed, the PD carries no information at all.

Let us first consider the case where the values of x_i in the message $x_1 \dots x_N$ can be chosen at will. Since there are M possibilities for each x_i, the number of different messages is $g = M^N$. The number of bits K necessary to transmit such a message, or, if you want, the number of bits to distinguish one message from the other, is $K = \ln_2 g = \ln_2 M^N = N \ln_2 M$ (since $2^K = M^N$). On the other hand, if the x_i are taken from the probability distribution p_i, the possible choices of x_i are limited. For instance, if $p_1 \gg p_2$ it is unlikely to construct a message with more occurences of x_2 than x_1. In the limit of large number of message elements N, the number of occurrences of x_i in the message approaches asymptotically $N_i = p_i\, N$. In fact the probability of finding $|N_i - N\, p_i| > \sqrt{N_i}$ becomes exponentially small as $N \to \infty$. Taking into account the restriction that the message contains N_i occurrences of x_i, the number of possible messages is reduced to

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

This corresponds to the number of possible ways of arranging the $N_1, \dots N_M$ occurrences of $x_1 \dots x_M$. To specify the message we therefore need

$$\begin{aligned} K = \ln_2 g &= \ln_2 N! - \sum_{i=1}^{M} \ln_2 N_i! \\ &\cong N \ln_2 N - N - \sum_{i=1}^{M} (N_i \ln_2 N_i - N_i) \\ &= -N \sum_{i=1}^{M} \frac{N_i}{N} \ln_2 \frac{N_i}{N} = -N \sum_{i=1}^{M} p_i \ln_2 p_i \end{aligned}$$

bits of information. As expected, we recover here the two limiting cases discussed above: $\ln_2 g = 0$ for $p_i = 1$ and $p_j = 0\ \forall\, j \neq i$ and $\ln_2 g = N \ln_2 M$ for $p_i = 1/M\ \forall\, i$ (uniform distribution). For any non-uniform probability distribution the information content of the message $\ln_2 g$ is smaller than $N \ln_2 M$, which is the information content in the absence of any information on the relative probabilities p_i. One assigns this difference to the information carried by the probability distribution $\{p_i\}$. The information content of $\{p_i\}$ is thus defined as

$$I\{p_i\} = \ln_2 M + \sum_{i=1}^{M} p_i \ln_2 p_i,$$

i.e., the reduction of information per unit message-length or transmitted token.

Introducing the definition of the *entropy of a probability distribution*

$$S\{p_i\} = -\sum_{i=1}^{M} p_i \ln p_i = -\langle \ln p_i \rangle \geq 0 \qquad (0 \leq p_i \leq 1)$$

we have

$$I\{p_i\} = (S_{\max} - S\{p_i\}) \,/\, \ln_2,$$

where

$$S_{\max} = -\sum_{i=1}^{M} \frac{1}{M} \ln \frac{1}{M} = \ln M$$

is the maximum of $S\{p_i\}$ corresponding to $p_i = 1/M$. A distribution with maximum entropy carries the least possible information. Therefore, S gives a measure of diversity of the distribution. S is actually the logarithm of the number of possible microscopically different states (messages) that can be constructed with elements satisfying $N_i/N = p_i$. For the distribution $p_i = \delta_{ij}$ (for some j) there is only one possible microscopic configuration or message $(x_j\, x_j \ldots x_j)$. In this case, and only in this case, we have $S = 0$.

The entropy does not depend on the values of the random variables. Any one-to-one mapping $x_i \to f_i = F(x_i)$ leaves the entropy unchanged since $p(x_i) = p(f_i)$. This implies in particular that the (non-equilibrium) entropy of system of interacting particles (e.g., an interacting electron gas) with occupation probabilities $n_{\vec{k}}$ for each quasi-particle state $\vec{k}$ is the same as the entropy of a non-interacting *having the same* $n_{\vec{k}}$. The actual equilibrium entropy at a given temperature T will of course be different, since in this case the entropy corresponds to the maximum value of $S\{n_{\vec{k}}\}$ compatible with the constraint of fixed average energy $\langle\, E \,\rangle$.

In contrast, any many-to-one mapping will reduce the entropy of the probability distribution, since it reduces its diversity or, in other words, it increases the definiteness or the information content. For example, given p_1 and p_2, the mapping

$$\left.\begin{matrix} x_1 \\ x_2 \end{matrix}\right\} \to f$$

gives

$$p(f) = p_1 + p_2.$$

The resulting change in the entropy reads

$$\begin{aligned} \Delta S &= -p(f)\, \ln p(f) + (p_1\, \ln p_1 + p_2\, \ln p_2) \\ &= S_f - S_{1\,2} \\ &= p_1\, \ln \frac{p_1}{p_1 + p_2} + p_2\, \ln \frac{p_2}{p_1 + p_2}, \end{aligned}$$

which is negative, provided that $p_2 \neq 0$ (or $p_1 \neq 0$). Conversely, removing a constraint in a probability distribution systematically increases S.

The entropy S can also be used to infer subjective (theoretical) estimates of probability distributions. For instance, in the absence of any information of p_i the *best unbiased estimate* of p_i is that all possible outcomes M are equally probable, i.e., $p_i = 1/M$. This distribution maximizes the entropy, the diversity of the distribution and the number of possible microscopic states for the given M available states or outcomes. One may also say that this choice of p_i minimizes the information content of $\{p_i\}$.

If additional information is available the unbiased estimate for p_i is obtained by maximizing S subject to the constraints imposed by the available information. As an example let

us assume that we know the average value $\langle\, F(x)\,\rangle = f$ of some function of the random variable x. In this case we obtain the unbiased p_i from the extremum of

$$L\,\{p_i\} = \underbrace{S\,\{p_i\}}_{-\sum_{i=1}^{M} p_i\,\ln p_i} - \alpha\left(\sum_{i=1}^{M} p_i - 1\right) - \beta\left(\sum_{i=1}^{M} p_i\,F(x_i) - f\right),$$

where α and β are Lagrange multipliers. Straightforward derivation yields

$$\frac{\partial L}{\partial p_i} = -\ln p_i - 1 - \alpha - \beta F(x_i) = 0$$

$$\Rightarrow \quad p_i = e^{-(1+\alpha)}\; e^{-\beta F(x_i)}\;\alpha\; e^{-\beta F(x_i)}$$

$$\Rightarrow \quad p_i = \frac{e^{-\beta F(x_i)}}{\sum_{i=1}^{M}\; e^{-\beta F(x_i)}},$$

where β is such that $f = \langle\, F(x)\,\rangle = \dfrac{\sum_{i=1}^{M}\; e^{-\beta F(x_i)}\; F(x_i)}{\sum_i\; e^{-\beta F(x_i)}}$.

Excercise 1.6:
Find the unbiased probability $p(x_i)$ for a random variable x_i $(i = 1, \ldots M)$ knowing the first n moments of $p(x_i)$ (i.e., $\langle\, x^n\,\rangle = M_n$). i) Show that $p(x_i) \propto \exp\left(\sum_{\nu=0}^{n} a_\nu\, x_i^\nu\right)$ with certain coefficients a_ν. ii) Consider the partition function $Z = \sum_{i=1}^{M} \exp\left\{\sum_{\nu=0}^{n} a_n\, x_i^n\right\}$ and show that the coefficients are given by the equations $\dfrac{\partial \ln Z}{\partial a_\nu} = M_\nu$.

In analogy with the discrete case we can define the entropy of a continuous probability density distribution $p(\vec{x})$ as

$$S = -\langle\, \ln p(\vec{x})\,\rangle = -\int p(\vec{x})\;\ln p(\vec{x})\,\mathrm{d}\vec{x}.$$

However, notice that this definition does not have some of the nice properties of $S = -\sum_i\, p_i\,\ln p_i$ for discrete random variables. For instance, for a uniform distribution in the interval $[a, b]$, i.e.,

$$p(x) = \begin{cases} 1/(b-a) & \text{for } a \le x \le b \\ 0 & \text{elsewhere,} \end{cases}$$

we have

$$S = -\int_a^b \frac{1}{b-a}\;\ln\left(\frac{1}{b-a}\right) = \ln(b-a).$$

For large intervals this is positive and diverges logarithmically [$S \to +\infty$ for $(b-a) \to$

$+\infty]$. In the opposite limit of a very narrow PDF around some point x_0, we have

$$p(x) = \begin{cases} 1/\varepsilon & \text{for } x_0 - \varepsilon/2 < x < x_0 + \varepsilon/2 \\ 0 & \text{elsewhere,} \end{cases}$$

and $S \to -\infty$ for $\varepsilon \to 0$.

Notice that S can take negative values for very sharp $p(x) \simeq \delta(x)$ since $p(x)$ is not bounded. These situations, however, never appear in the description of macroscopic systems. It is interesting to observe that S decreases as the diversity of the distribution decreases (e.g., $S[\delta(x)] \to -\infty$) as in the case of discrete random variables.

In order to avoid this problem, or rather to understand the origin of the divergence in $\langle \ln p(x) \rangle$ for $p(x) = \sum_i \delta(x - x_i)$, it is useful to derive the expression for S for continuous PDF starting from the expression for discrete variables

$$S = -\sum_i p_i \ln p_i$$

and introducing a finite lower bound or threshold Δ below which two outcomes (differing by less than Δ) are considered to be equivalent. With this coarse graining the logarithm of the number of possible messages (i.e., the diversity of the probability distribution) is given by

$$S = -\sum_i \tilde{P}(x_i < x < x_i + \Delta) \ln \left[\tilde{P}(x_i < x < x_i + \Delta)\right]$$

where $x_i = i\,\Delta$ ($x_{i+1} = x_i + \Delta$) and $p_i = \tilde{P}(x_i < x < x_i + \Delta)$ is the probability for x to lie in the interval $[x_i, x_i + \Delta]$. Using that

$$\tilde{P}(x_i < x < x_i + \Delta) = P(x_i + \Delta) - P(x_i) = \int_{x_i}^{x_i+\Delta} p(x)\, dx,$$

where $P(x)$ refers to the cumulative probability function and $p(x) = \dfrac{dP}{dx}$ to the probability density function, we have

$$S = -\sum_i \left(\int_{x_i}^{x_i+\Delta} p(x)\, dx \right) \ln \left(\int_{x_i}^{x_i+\Delta} p(x)\, dx \right) \geq 0.$$

If the spectrum of outcomes is discrete or shows very narrow peaks (narrower than Δ) we can still compute S and recover the limit of discrete random variables. However, if $p(x)$ is smooth we can write

$$\begin{aligned} S &\cong -\sum_i \left(\int_{x_i}^{x_i+\Delta} p(x)\, dx \right) \ln \left[p(x_i)\, \Delta \right] \\ &\cong -\sum_i \left(\int_{x_i}^{x_i+\Delta} p(x)\, dx \right) \left[\ln p(x_i) + \ln \Delta \right] \\ &\cong -\int p(x) \ln \left[p(x) \right] dx - \ln \Delta. \end{aligned}$$

The term $\ln \Delta$ cancels the divergence of $S = -\langle \ln p(x) \rangle$ for $p(x) \to \delta(x)$.

Another problem of the definition

$$S = -\int p(x)\, \ln\left[p(x)\right]\, dx$$

is that it is not invariant under any bijective mappings $f = F(x)$. In fact one has

$$p(f)\, df = p(x)\, dx \quad \Rightarrow \quad p(f) = p(x)\, \frac{1}{\left|\frac{dF}{dx}\right|}$$

and therefore

$$\begin{aligned} S\left[p(f)\right] &= -\int p(f)\, \ln p(f)\, df = -\int p(x)\, \left[\ln p(x) - \ln\left|\frac{dF}{dx}\right|\right] dx \\ &= -\int p(x)\, \ln p(x)\, dx + \int p(x)\, \ln\left|\frac{dF}{dx}\right| dx \\ &= S\left[p(x)\right] + \left\langle \ln\left|\frac{dF}{dx}\right| \right\rangle. \end{aligned}$$

In the case of many random variables we define

$$S = -\int p(\vec{x})\, \ln p(\vec{x})\, d\vec{x}$$

and consequently

$$S\left[p(\vec{f})\right] = S\left[p(\vec{x})\right] + \left\langle \ln\left|J(\vec{f})\right| \right\rangle,$$

where $J(\vec{f}) = \left|\frac{\partial f_i}{\partial x_j}\right|$ is the Jacobian of the variable transformation $f_i = f_i(x_1, \ldots x_N)$. The entropy is thus invariant under canonical transformations in classical mechanics and unitary transformation in quantum mechanics, which have Jacobian equal to 1.

Excercise 1.7: Loaded dice
A dice is loaded such that $p_6 = n\, p_1$ (e.g., $n = 2$). In other words, six occurs n times more often than 1.

i) Find the unbiased probabilities for the six faces of the dice.

ii) What is the information content of the probability distribution function p_i ($i = 1$–6) as a function of n?

---

2 Stochastic dynamics

So far we have discussed the fundamental properties of probability distributions for both discrete and continuous random variables from a static point of view. The purpose of this chapter is to study how probability distributions evolve in time. After introducing some useful definitions of probability theory, including now the time as explicit variable, we derive the main equation governing the dynamics of probability distributions: the master equation. In this context we discuss the concept of transition probabilities, transition rates, and the notion of Markov process in which memory or inertial effects are neglected. The master equation, which controls the dynamics of Markov processes, is one of the most important equations in statistical physics. This is due to its simplicity and wide range of applicability, for example in chemical reactions, relaxation dynamics of complex systems such as glasses or spin glasses, the process of protein folding, population dynamics of biological systems, data processing and data transfer in complex computer networks, etc.

Once the basic equations are derived we present some of the methods for its practical solution in the context of complex energy landscapes. This includes the important case of Markov chains, either by reducing the problem to eigenvalue equations using an exponential ansatz, or by the kinetic Monte Carlo algorithm. The master equation is, however, useless without a sound theory of transition rates. This motivates the study of transition state theory, first, in the framework of classical mechanics, and then in quantum mechanics, under the simplifying assumption of separability of the Hamiltonian with respect to the reaction coordinate.

2.1 Time-dependent probability distributions

We consider a system described by a random variable x, which can be a single variable or more generally a vector $[x \equiv (x_1, x_2 \ \ldots \ x_n)]$. The following notions are introduced:

$$P_1(x_1, t) = \textit{probability density} \text{ for the stochastic variable } x \text{ to have the value } x_1 \text{ at time } t.$$

$$P_n(x_1, t_1;\, x_2, t_2;\, \ldots\, x_n, t_n) = \textit{joint probability density} \text{ that } x \text{ has the values } x_1 \text{ at } t_1,\, x_2 \text{ at } t_2, \ldots \text{ and } x_n \text{ at } t_n.$$

The joint probabilities can be reduced as

$$P_{n-1}(x_1\, t_1;\, \ldots\, ; x_{n-1}, t_{n-1}) = \int dx_n\, P_n(x_1\, t_1;\, \ldots\, ; x_{n-1}\, t_{n-1};\, x_n, t_n). \tag{2.1}$$

This relation, known as the *Chapman-Kolmogorov equation*, is central to probability theory. Note that the actual value of the time variable t_n is irrelevant, as well as the actual variable (here x_n) with respect to which one integrates. If x is a discrete variable, the integrals are replaced by sums over the set of outcomes $\{x_1, \ \ldots \ x_k\}$.

Of course, the P_n satisfy the usual properties of a probability density: $P_n \geq 0$, additiveness and normalization. For instance,

$$\int dx\, P_1(x, t) = 1 \quad \forall\, t.$$

A process is said to be *stationary* if

$$P_n(x_1, t_1; \ldots x_n\, t_n) = P_n(x_1, t_1 + \tau; \ldots x_n, t_n + \tau)$$

for all n and $\tau \geq 0$. Consequently, for a stationary process the probability density $P_1(x)$ is independent of time:

$$P_1(x, t) = P_1(x).$$

All physical processes *in equilibrium* are stationary.

For the *conditional probabilities* we introduce the notation

$$P_{1|1}(x_1\, t_1 \,|\, x_2\, t_2) = \begin{array}{l}\text{Probability density for the random variable} \\ x \text{ to have the value } x_2 \text{ at time } t_2 \text{ given that it} \\ \text{had the value } x_1 \text{ at time } t_1 \text{ (usually } t_1 \leq t_2).\end{array}$$

$P_{1|1}$ is given by the relation

$$P_2(x_1\, t_1;\, x_2\, t_2) = P_1(x_1\, t_1)\; P_{1|1}(x_1\, t_1 \,|\, x_2\, t_2).$$

The reduction of P_2 according to the Champan-Kolmogorov equation (2.1) yields the important relation

$$P_1(x_2\, t_2) = \int P_1(x_1, t_1)\; P_{1|1}(x_1, t_1 \,|\, x_2\, t_2)\, dx_1.$$

Note that the fixed value of t_1 at which $P_1(x_1, t_1)$ and $P_{1|1}(x_1\, t_1,\, x_2\, t_2)$ are evaluated is irrelevant.

Conditional probabilities are of course normalized:

$$\int P_{1|1}(x_1\, t_1 \,|\, x_2\, t_2)\, dx_2 = 1 \quad \forall\, x_1, t_1 \text{ and } t_2.$$

This can be easily demonstrated (if necessary) by integrating the previous relation.

It is also useful to introduce the *joint conditional probability densities*

$$P_{k|l}(x_1\, t_1\, \ldots\, x_k\, t_k \,|\, x_1'\, t_1'\, \ldots\, x_l'\, t_l') = \begin{array}{l}\text{probability density that the random variable} \\ x \text{ has the values } x_1' \text{ at } t_1', \, x_2' \text{ at } t_2', \ldots \text{ and } x_l' \\ \text{at } t_l' \text{ given that it had the values } x_1 \text{ at } t_1, \ldots \\ \text{and } x_k \text{ at } t_k.\end{array}$$

$P_{k|l}$ describes correlations between the dynamics of the random variable at different times. They are important if the variable x has *memory effects*, also known as *inertial effects*.

The dynamics and the form of $P_{k|l}$ simplify considerably if the random variable has memory only of its immediate past, i.e., when the probability of finding the system in a given configuration x_n at time t_n does not depend on the previous history, except for its last configuration x_{n-1} at time t_{n-1}. This is actually a good approximation when the processes, which dynamics one is interested in, are *rare events*. By this we mean that the typical time $\Delta t = t_n - t_{n-1}$ involved in changes of state is much larger than the thermalization times, so that memory effects are washed away. This is actually the case in a number of important slow relaxation processes, such as a thermally activated chemical

reaction across an energy barrier $\Delta E \gg k_b T$, where vibrational periods are much shorter than transition times. Other physical situations, where the previous history can be safely neglected, are the following:

- $A \rightleftharpoons B$ unimolecular reaction
- Large molecule relaxations (e.g., protein folding)
- Proton transfer in proteins
- Magnetic relaxation of nanoparticle ensembles, where the precession of the nanoparticle moments is much faster than the collective rearrangements of magnetic order of the nanostructure as a whole.
- Domain wall motion, where the precession time of atomic spin is much shorter than the typical times of domain-wall motion.

Let us consider a *stochastic process* or *stochastic path*, which can be mathematically defined as a time-indexed collection of random variables:

$$(x_1 t_1;\, x_2 t_2;\, \ldots\, x_{n-1} t_{n-1}) \quad \text{with} \quad t_1 < t_2 < \ldots < t_{n-1}.$$

The probability for the state x_n at a subsequent time $t_n > t_{n-1}$ is formally given by

$$P_{n-1|1}(x_1 t_1;\, x_2 t_2;\, \ldots\, x_{n-1} t_{n-1} | x_n t_n).$$

We say that a process is *Marcovian* when the conditional probability for x to take the value x_n at t_n is fully determined only by the value of x_{n-1} at t_{n-1}. Formally, a Markov process is definend by the condition

$$P_{n-1|1}(x_1 t_1\, \ldots\, x_{n-1} t_{n-1} | x_n t_n) = P_{1|1}(x_{n-1} t_{n-1} | x_n t_n).$$

The fundamental conditional probability

$$P_{1|1}(x_1 t_1 | x_2 t_2)$$

is known as the *transition probability*. A Markov process is completely defined by only two functions

$$P_1(x,t) \quad \text{and} \quad P_{1|1}(x,t|x',t'),$$

from which the probability for any stochastic process can be derived. For example,

$$P(x_1 t_1;\, x_2 t_2;\, x_3 t_3) = P_1(x_1 t_1)\; P_{1|1}(x_1 t_1 | x_2 t_2)\; P_{1|1}(x_2 t_2 | x_3 t_3).$$

We can now eliminate $x_2 t_2$ on the left-hand side by using the Chapman-Kolmogorov relation

$$P(x_1 t_1;\, x_3 t_3) = P_1(x_1 t_1) \int P_{1|1}(x_1 t_1 | x_2 t_2)\; P_{1|1}(x_2 t_2 | x_3 t_3)\, dx_2$$

which implies

$$\frac{P(x_1 t_1;\, x_3 t_3)}{P_1(x_1 t_1)} = P_{1|1}(x_1 t_1 | x_3 t_3) = \int P_{1|1}(x_1 t_1 | x_2 t_2)\; P_{1|1}(x_2 t_2 | x_3 t_3)\, dx_2.$$

This is known as *Chapman-Kolmogorov equation* for Markov processes. It allows to split the transition probability for $x_1 t_1 \to x_3 t_3$ into two successive steps $x_1 t_1 \to x_2 t_2$ and

$x_2t_2 \to x_3t_3$ involving all possible intermediate states x_2t_2. The Markov character of the dynamics is reflected by the fact that the probability of the two successive steps is given by the product of the probabilities of the individual steps. In a Markov process the successive transitions are statistically independent, i.e., the transition probability for $x_2t_2 \to x_3t_3$ is not affected by the nature of the previous transition $x_1t_1 \to x_2t_2$. We can iterate the Chapman-Kolmogorov relation at will by introducing an arbitrary number k of intermediate times $t_1 \le t_1' \le \ldots \le t_k' < t_3$ and write

$$P_{1|1}(x_1t_1|x_3t_3) = \int dx_1' \ldots dx_k'\; P_{1|1}(x_1t_1|x_1't_1')\; P_{1|1}(x_1't_1'|x_2't_2')\; \ldots\; P_{1|1}(x_k't_k'|x_3t_3).$$

This allows us to split a complex evolution into the succession of elementary (statistically independent) processes, for example, by allowing to sum over different competing pathways, or eventually by going all over to a continuous functional-integral formulation.

In order to illustrate the functional-integral formulation of the stochastic dynamics we express the transition probability from the state x_k at time t_k to a state x_{k+1} at an arbitrarily close time $t_{k+1} = t_k + \Delta t$ in terms of the change of an auxiliary function $\Delta S(x_k, t_k;\, x_{k+1}, t_{k+1})$, which can be regarded as an action:

$$P_{1|1}(x_kt_k|x_{k+1}t_{k+1}) = e^{-\Delta S(x_k,t_k;\,x_{k+1},t_k+\Delta t)}.$$

In the limit of small time steps $\Delta t \to 0$ we can regard x_k as a function of t and associate $x_k = x(t)$ and $x_{k+1}' = x(t+\Delta t)$, so that ΔS becomes a function of $x(t)$, t, $x(t+\Delta t)$ and Δt. Assuming that $x(t)$ is a differentiable function of t, ΔS can be considered to be a function of $x(t)$, t, $\dot{x}(t)$ and Δt. Taking into account that $P_{1|1}(x_k, t_k, x_k, t_k) = 1$ we must have $\Delta S = 0$ for $\Delta t = 0$ and therefore

$$\Delta S = \Delta S\Big(x(t),\, x(t+\Delta t),\, t,\, \Delta t\Big) = s\Big(x(t),\, \dot{x}(t),\, t\Big)\Delta t \qquad (\Delta t > 0),$$

where $s\Big(x(t), \dot{x}(t), t\Big)$ reprensents the rate of change of S. Consequently, we can write $P_{1|1}$ as a functional integral of the form

$$P_{1|1}(x_1t_1|x_3t_3) = \int D[x(t)]\; e^{-\int_{t_1}^{t_3} s(x,\dot{x},t)\,dt}$$

with the constraints $x(t_1) = x_1$ and $x(t_3) = x_3$.

2.2 The Master equation

Let us return to the relation

$$P(x_1t_1;\, x_3t_3) = P_1(x_1t_1) \int P_{1|1}(x_1t_1|x_2t_2)\; P_{1|1}(x_2t_2|x_3t_3)\, dx_2$$

and integrate it with respect to x_1. In this way one recovers the known expression

$$P_1(x_3t_3) = \int P_1(x_1t_1)\; P_{1|1}(x_1t_1|x_3t_3)\, dx_1 \qquad \forall\, t_1$$

which can be rewritten as

$$P_1(x,t) = \int P_1(x',t')\; P_{1|1}(x't'|x\,t)\, dx' \qquad \forall\, t'. \tag{2.2}$$

In order to derive the equation for the time dependence of $P(x,t)$, we consider its change after a time increment Δt. This is given by

$$P(x,t+\Delta t) - P(x,t) = \int P_1(x',t') \left[P_{1|1}(x't'|x,t+\Delta t) - P_{1|1}(x',t'|x,t) \right] dx',$$

where the conditional probability $P_{1|1}(x't'|x,t+\Delta t)$ can be expanded for small Δt. First of all, for $\Delta t = 0$ and $t' = t$ we have

$$P_{1|1}(x't|x\,t) = \delta(x-x') \quad \forall\, t.$$

In order to infer the form of $P_{1|1}(x',t|x,t+\Delta t)$ for $\Delta t > 0$ one must consider the cases $x = x'$ and $x \neq x'$ separately. For $x \neq x'$ it is useful to introduce the transition probability per unit time $k(x',x,t)$ for the system to change from the state x' to the state x. In terms of k the conditional probability can be expanded as

$$P_{1|1}(x',t|x,t+\Delta t) = k(x',x,t)\,\Delta t + O(\Delta t^2).$$

The quantity $k(x',x,t)$ is known as *transition rate* or *rate constant* for the transition from x' to x, which in the most general case k depends on t. The conditional probability for $x = x'$, i.e., the probability density that no transition occurs, can be obtained from the normalization condition

$$\int P_{1|1}(x',t'|x\,t)\,dx = 1$$

applied to the first order approximation of $P_{1|1}(x't|x,t+\Delta t)$. The total probability that the system changes its state from x' to any other state x is equal to

$$\Delta t \int k(x',x)\,dx + O(\Delta t^2), \tag{2.3}$$

which is smaller than 1 since $\Delta t \to 0$. Here and in the following we drop for simplicity the time dependence of k. From Eq. (2.3) it follows that the probability that the random variable x remains equal to x' is equal to $1 - \Delta t \int k(x',x)\,dx$. Summarizing, the conditional probability density that the random variable has the value x at time $t+\Delta t$, given that it had the value x' at time t, reads

$$P_{1|1}(x',t|x,t+\Delta t) = \delta(x-x') \left(1 - \int dy\; k(x,y)\,\Delta t \right) + \Delta t\; k(x',x) + O(\Delta t^2).$$

Knowing that a Markov process is fully defined by $P(x,t)$ and $P_{1|1}(x't'|x\,t)$, one concludes that for short times $\Delta t \to 0$ the dynamics is controlled by the transition rate $k(x',x,t)$.

Let us now return to the fundamental relation (2.2) and determine the time derivative of $P_1(x,t)$:

$$\begin{aligned}
P_1(x,t) &= \int P_1(x',t')\; P_{1|1}(x't'|x,t)\,dx' \\
\frac{dP_1(x,t)}{dt} &= \int P_1(x',t')\,\frac{d}{dt}\; P_{1|1}(x't'|x,t)\,dx' \\
&= \int P_1(x',t) \left[k(x',x) - \delta(x-x') \int dy\; k(x,y) \right] dx' \\
&= \underbrace{\int P_1(x',t)\,k(x',x)\,dx'}_{\substack{\text{Total probability of going} \\ \text{from any } x' \text{ to } x.}} - \underbrace{P_1(x,t) \int k(x,y)\,dy.}_{\substack{\text{Total probability of changing} \\ \text{from } x \text{ to some } y \neq x.}}
\end{aligned}$$

In conclusion we obtain the *master equation*

$$\frac{dP_1(x,t)}{dt} = \int \left[P_1(x',t)\, k(x',x) - P_1(x,t)\, k(x,x')\right] dx', \tag{2.4}$$

which governs the dynamics of Markov processes. The physics of the problem is contained in the transition rates $k(x',x)$, which give the transition probability per unit time that the system changes its state to x at time $t+\Delta t$ given that it was in the state x' at time t.

2.3 Fokker-Planck and diffusion equations

For continuous random variables x, for which the elementary changes of state along the Markov dynamics take place by small jumps $\xi = x - x'$, we can derive a differential equation for $P(x,t)$ known as Fokker-Planck equation. If the transitions occur by small jumps ξ, this means that the transition rate $k(x',x)$ for $x = x' + \xi$ decreases rapidly with increasing $|\xi| = |x - x'|$. It is then useful to change the variables of k by replacing $k(x',x)$ by $k(x',\xi) = k(x-\xi,\xi)$ for the transition $x' \to x = x' + \xi$. Note that the rate constant for the reverse process $x \to x' = x - \xi$ is now given by $k(x,-\xi)$. We can then write the master equation

$$\frac{\partial P(x,t)}{\partial t} = \int dx'\; k(x',x)\, P(x',t) - P(x,t) \int dx'\; k(x,x')$$

as

$$\frac{\partial P(x,t)}{\partial t} = \int d\xi\; k(x-\xi,\xi)\, P(x-\xi,t) - P(x,t) \int d\xi\; k(x,-\xi). \tag{2.5}$$

We now expand the product $k(x-\xi,\xi')\, P(x-\xi,t)$ appearing in the first integral in powers of ξ for $\xi \to 0$ (i.e., $|\xi| \ll |x|$) and any value of the size of the jump ξ', which is the second variable of k:

$$\begin{aligned} k(x-\xi,\xi')\, P(x-\xi,t) = k(x,\xi')\, P(x,t) - \xi\, \frac{\partial}{\partial x}\left[k(x,\xi')\, P(x,t)\right] \\ + \frac{1}{2}\, \xi^2\, \frac{\partial^2}{\partial x^2}\left[k(x,\xi')\, P(x,t)\right] + O(\xi^3). \end{aligned} \tag{2.6}$$

For many variables, i.e., $x \equiv \vec{x}$ we have

$$\begin{aligned} k(\vec{x}-\vec{\xi},\vec{\xi}')\, P(\vec{x}-\vec{\xi},t) = k(\vec{x},\vec{\xi}')\, P(\vec{x},t) - \vec{\xi}\cdot\vec{\nabla}_{\vec{x}}\left[k(\vec{x},\vec{\xi}')\, P(\vec{x},t)\right] \\ + \frac{1}{2}\, \sum_{ij}\, \xi_i\, \xi_j\, \frac{\partial^2}{\partial x_i\, \partial x_j}\left[k(\vec{x},\vec{\xi}')\, P(\vec{x},t)\right] + O(\xi^3). \end{aligned}$$

Replacing Eq. (2.6) in the master equation (2.5) and neglecting the higher-order terms beyond ξ^2 we obtain

$$\begin{aligned}\frac{\partial P(x,t)}{\partial t} &= \int d\xi\; k(x,\xi)\, P(x,t) - \int \xi\; \frac{\partial}{\partial x}\left[k(x,\xi)\, P(x,t)\right] d\xi \\ &\quad + \frac{1}{2}\int \xi^2\; \frac{\partial^2}{\partial x^2}\left[k(x,\xi)\, P(x,t)\right] d\xi - P(x,t)\int d\xi\; k(x,-\xi) \\ &= -\int \xi\; \frac{\partial}{\partial x}\left[k(x,\xi)\, P(x,t)\right] d\xi + \frac{1}{2}\int \xi^2\; \frac{\partial^2}{\partial x^2}\left[k(x,\xi)\, P(x,t)\right] d\xi \\ &= -\frac{\partial}{\partial x}\left\{P(x,t)\int \xi\; k(x,\xi)\, d\xi\right\} + \frac{1}{2}\;\frac{\partial^2}{\partial x^2}\left\{P(x,t)\int \xi^2\; k(x,\xi)\, d\xi\right\}.\end{aligned}$$

Introducing the n-th order jump moments

$$\alpha_n(x) = \int \xi^n\, k(x,\xi)\, d\xi$$

the master equation takes the form

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}\Big(\alpha_1(x)\, P(x,t)\Big) + \frac{1}{2}\;\frac{\partial^2}{\partial x^2}\Big(\alpha_2(x)\, P(x,t)\Big), \tag{2.7}$$

which is known as *Fokker-Planck equation*. It is important to recall that in order that the Fokker-Planck equation is applicable, the random variable must be truly continuous, and the transitions must occur by infinitesimally small steps. However, very often the physically relevant changes of state or rare events, to which the Markovian hypothesis applies, involve discrete, non-infinitesimal changes of the random variables. This is for example the case when a molecule changes its conformation between two local energy minima across an energy barrier.

If the system is isotropic, $k(x,\xi) = k(x,-\xi)$, i.e., left and right jumps are equally probable, the odd-order jump moments and in particular $\alpha_1(x)$ vanish. Moreover, if the system is homogeneous, all values of the random variable are equivalent and therefore $k(x,\xi)$ is independent of x. In this case α_2 is also independent of x and the Fokker-Planck equation takes the form

$$\frac{\partial P(x,t)}{\partial t} = D\;\frac{\partial^2}{\partial x^2}\, P(x,t), \tag{2.8}$$

where $D = \alpha_2/2 > 0$ is the *diffusion coefficient*. The solution of the diffusion equation can be easily obtained by applying the usual methods for linear differential equations. One can write

$$P(x,t) = \int_{-\infty}^{+\infty} e^{i\,k\,x}\, \tilde{p}(k,t)\, dk \quad \Leftrightarrow \quad \tilde{p}(k,t) = \frac{1}{2\,\pi}\int_{-\infty}^{+\infty} e^{-i\,k\,x}\, p(x,t)\, dx$$

and replace it in the diffusion equation to obtain

$$\frac{\partial \tilde{p}(k,t)}{\partial t} = -D\,k^2\, \tilde{p}(k,t).$$

Straightforward integration yields

$$\tilde{p}(k,t) = \tilde{p}(k,0)\; e^{-D\,k^2\,t},$$

so that, after Fourier back transformation, the probability distribution is given by

$$P(x,t) = \int_{-\infty}^{+\infty} e^{i\,k\,x}\,\tilde{p}(k,0)\;e^{-D\,k^2\,t}\,dk.$$

As an example, let us consider the distribution $P(x,0) = \delta(x)$ as initial condition; which corresponds to the particle at the origin. In this case $\tilde{p}(k,0) = 1/2\,\pi$ and

$$\begin{aligned} P(x,t) &= \frac{1}{2\,\pi}\int_{-\infty}^{+\infty} e^{i\,k\,x}\;e^{-D\,k^2\,t}\,dk \\ &= \frac{1}{2\,\pi}\int_{-\infty}^{+\infty} e^{-(D\,k^2\,t - i\,k\,x)}\,dk. \end{aligned}$$

Completing the square of a binomial as

$$\begin{aligned} D\,k^2\,t - i\,k\,x &= Dt\left[k^2 - i\,2\,k\,\frac{x}{2Dt} + \left(\frac{ix}{2Dt}\right)^2 - \left(\frac{ix}{2Dt}\right)^2\right] \\ &= Dt\left(k - \frac{ix}{2Dt}\right)^2 + \frac{x^2}{4Dt}, \end{aligned}$$

and integrating the Gauss function in the complex plane

$$P(x,t) = e^{-\frac{x^2}{4Dt}}\,\frac{1}{2\,\pi}\underbrace{\int_{-\infty}^{+\infty} e^{-Dt\left(k-\frac{ix}{2Dt}\right)^2}\,dk}_{\sqrt{2\pi}\,\sigma=\sqrt{\frac{\pi}{Dt}}=\sqrt{\frac{\pi}{\alpha}}},$$

one obtains

$$P(x,t) = \underbrace{\frac{1}{\sqrt{4\pi\,Dt}}}_{\sqrt{2\pi\,\sigma^2}}\;e^{-\frac{x^2}{4Dt}}.$$

The solution is therefore a Gaussian distribution with a time-independent mean value $x = 0$ and a square mean deviation $\sigma = \sqrt{2\,D\,t}$ that increases monotonously with t.

Excercise 2.1:
Starting from the corresponding equations (2.4), (2.7) and (2.8), show that the master, Fokker-Planck and diffusion equations preserve the norm of the probability distribution $P(x,t)$. If necessary, state any conditions on $k(x,x')$ or $\alpha_n(x)$ for the norm to be independent of t.

2.4 Discretization of complex energy landscapes

In order to motivate the physical background behind some of the theoretical methods to be discussed below it is useful to make a small digression and to discuss how the energy landscape of complex systems can be simplified in a way that allows one to focus on the relevant dynamical processes of physical interest.

We consider a system described by a configurational coordinate $\vec{x}$. This can be, for example, a molecule or cluster where $\vec{x}$ refers to the coordinates of the atoms, a magnetic

nanostructure composed of an ensemble of nanospins where $\vec{x}$ stands for the ensemble of polar and azimuthal angles defining the direction of the magnetization of each particle, or a magnetic material undergoing a magnetization reversal process (e.g., a domain wall motion) where $\vec{x}$ refers to the size and orientation of the local magnetic moment at each atom. In the following we assume that we know the energy $V(\vec{x})$ of the system as a function of the configuration of $\vec{x}$ and that we are equally able to calculate the gradient $\overrightarrow{\nabla V}$ and if necessary the Hessian matrix at each point. Except for very small systems consisting of a few degrees of freedom, it is clear that a full description of the dynamics in the continuum is hopeless. It is therefore important to identify the degrees of freedom that are really relevant for the dynamical process of interest and, in a way, integrate out the other degrees of freedom which may vary too rapidly to be considered explicitly.

Suppose you want to describe a chemical reaction

$$A + B \rightleftharpoons C.$$

It is clearly uninteresting to follow the position of each atom through every possible reactive trajectory. One would rather like to identify and understand the physically relevant reaction path or paths. Which are the parts of the molecules that actually interact? How does the reaction depend on temperature? What are the intermediate transition states? Are there several competing reaction paths? Following atomic vibrations in detail is obviously irrelevant for this, although the entropy associated to vibrations is likely to condition the actual pathway or rate. The same holds for a number of magnetic relaxation processes where the precession of individual spins is too fast compared to domain-wall motion or to the changes in the relative orientation of the nanoparticle moments in a nanostructure. In these cases, it is neither possible nor meaningful to consider the short-time dynamics explicitly.

The solution to the problem is to focus on the stationary states of the system, namely, the local minima of $V(\vec{x})$ and the transition states or first-order saddle-points connecting them. The local minima are actually the regions of configurational space where the system spends most of the time and where the thermalizations occuring between the transition between the minima occur. The first-order saddle-points define the bottlenecks through which the relatively rare elementary transition processes take place. This Markovian approach based on stationary points is expected to work as long as the temperature or average energy is smaller than the energy barriers, so that the residing time in each minimum allows for a thermalization which washes away any memory effects. The idea is to discretize the energy surface $V(\vec{x})$ by the following transformation

$$\tilde{V}(\vec{x}) = \min\{V(\vec{x})\}, \tag{2.9}$$

where $\min\{\ldots\}$ refers to the local minimization following the gradient at the point $\vec{x}$. The ensemble of points in configuration space that lead to a given minimum is called the *catchment volume* of the given minimum.

The idea of attaching all points that fall to the same minimum to this minimum seems to leave away the transition states, which belong to the surfaces separating the different catchment volumes. Transition states are, however, important and need to be recovered for the calculation of transition rates and in order to understand the dynamics. In the context of global optimization methods the transformation (2.9) of $V(\vec{x})$ leads to the bassin hopping method. Notice, for example, that a transition $x_1 \to x_2$ is highly improbable if

the true energy difference $\Delta V = V(x_2) - V(x_1)$ is taken into account ($k_B T \ll \Delta V$), but that it becomes extremely favourable in terms of $\Delta\tilde{V}$. Quenching the configuration to the nearest minimum allows to explore large volumes in configurational space and to efficiently localize the lowest energy configurations. A huge amount of computer time is saved by increasing the acceptance rate of configurations located on the other side of potentially large energy barriers, which leads to true configurational changes. In practice, the most effective approach appears to be to relax the structure to the new minimum configuration (Li and Scheraga, see Wales p. 342 and refs. 669 + 670 therein). In this way the cooling down to the new minimum is accelerated.

The bassin hopping global optimization algorithm can be summarized as follows:

i) Given a configuration at a minimum $\vec{X}_{\rm old}$ with energy $V_{\rm old}$, propose a rearrangement Markov step by perturbing the configuration and *minimizing to obtain the new minimum energy* $V_{\rm new}$. Large displacements are used here, since the idea is to hop between different attraction bassins. Typical maximum displacements are about 30% of the NN distance with an acceptance rate of about 0.5.
ii) Accept the move if $V_{\rm new} < V_{\rm old}$ or if $e^{-(V_{\rm new}-V_{\rm old}/k_B T)}$ is larger than a random number in the interval [0,1]. In this context the temperature T is the only parameter of the simulation. The maximum step size can be adjusted dynamically to a fixed acceptance rate.
iii) Usually the structure is relaxed to the current (accepted) minimum.

The idea of attaching the same minimum energy to the whole catchment area of a minimum suggests to split the configuration space in these zones and to calculate the thermodynamic properties or even the long-time dynamic properties in terms of the contributions from the different local minima. For thermodynamics we can split the configuration space in the catchment bassins A_i associated to each minimum i. The density of states can be written as

$$\Omega(E) = \sum_i \Omega_i(E)$$

and the partition function

$$Z(T) = \int dE\; e^{-\beta E}\, \Omega(E) = \sum_i Z_i(T).$$

The microcanonical and canonical probabilities of finding the system in the catchment bassin i are given by

$$P_i(E) = \frac{\Omega_i(E)}{\Omega(E)} = \frac{\Omega_i(E)}{\sum_i \Omega_i(E)}$$

and

$$P_i(T) = \frac{Z_i(T)}{Z(T)}.$$

It should be, however, noted that the properties are not always the weighted average of the properties corresponding to each minimum. The entropy $S(E) = k_B \ln \Omega(E)$ is an interesting example in this context. The entropy associated to minimum i is given by

$S_i = k_B \ln \Omega_i(E)$ and the microcanonic average over all minima reads

$$\sum_{V_i<E} P_i(E)\, S_i(E) = k_B \sum_{V_i<E} P_i(E)\, \ln \Omega_i(E).$$

Using that

$$\begin{aligned} \sum_i P_i(E)\, \ln \Omega_i(E) &= \ln \Omega(E) + \sum_i P_i(E)\, \ln \left[\frac{\Omega_i(E)}{\Omega(E)}\right] \qquad \left(\sum_i P_i = 1\right) \\ &= \ln \Omega(E) + \sum_i P_i(E)\, \ln P_i(E) \end{aligned}$$

we have

$$S(E) = k_B \ln \Omega(E) = \sum_i P_i(E)\, S_i(E) - k_B \sum_i P_i(E)\, \ln P_i(E). \tag{2.10}$$

The first term on the right-hand side of Eq. (2.10) represents the entropy within each bassin, while the second term represents the entropy associated to the different occupations of the bassins. The latter term takes into account the entropy of mixing throughout the different minima. It vanishes when only one bassin is occupied.

2.5 Marcovian dynamics based on stationary points

A knowledge of the local minima and the development of $V(\vec{x})$ in its environment allows to determine equilibrium properties including possible anharmonicity effects. If we want to get insight in the dynamics we need to take into account the transition states that connect the various minima with each other.

The dynamics of an elementary transition between two nearby minima i and j is characterized by the rate constant for the transtion $A \to B$. This rate constant can be determined in the framework of transition state theory and is given by

$$k_{ji}(T) = \frac{k_B T}{2\,\pi\,\hbar}\, \frac{Q^*(T)}{Q_i(T)}, \tag{2.11}$$

where $Q^*(T)$ is the canonical partition function at the transition state, and $Q_i(T)$ is the partition function at the initial minimum i. $Q_i(T)$ can be determined by expanding the potential at minimum i in the harmonic approximation. $Q^*(T)$ is obtained in the same way as $Q_i(T)$ under the constraint that the system is at the hypersurface separating the two minima, which goes through the saddle point. The basic assumption of transition state theory is that in the time evolution of the system there are no recrossings of the hypersurface separating the two minima. This implies that the flux from "left to right", i.e., from i to j through the dividing surface S is equal to the number of transitions from i to j. In other words, the possibility that a trajectory starting at minimum i crosses the dividing surface from the i-side to the j-side and then returns to minimum i is neglected. A clear derivation of $k_{ji}(T)$ is given by Miller [Acc. Chem. Res. **26** (1993)] and shall be reviewed briefly later on. For the moment we will content ourselves with the fact that $k_{ji}(T)$ can be determined from local properties of $V(\vec{x})$ and with its physical interpretation. Indeed, Eq. (2.11) shows that $k_{ji}(T)$ depends on the probability that the system finds a channel to get through the hypersurface S dividing the minima, which is proportional to $Q^*(T)$,

relative to the probability of finding the system in minimum i, which is proportional to $Q_i(T)$.

Besides these elementary transitions we must take into account the probability of finding the system in each local minima and the transitions between different local minima. This problem can be handled by the master equation by assuming that each minimum corresponds to one of the possible values of the random variable x. In the following we discuss some important properties of the Markov processes taking Markov chains as particularly relevant example. The general solution of the master equation is discussed at a second step under the assumption of detailed balance.

2.6 Markov chains

One of the simplest examples of a Markov process is the Markov chain, which can be characterized by two main assumptions: i) The random variable takes discrete values $x_1,\ x_2, \ldots x_M$, which presupposes a discrete set of outcomes, and ii) the transitions occur at a discrete set of times $t = s\tau$, where τ is some characteristic time interval and $s \in \mathbb{N}$. For the moment we can picture τ as a measure of the typical vibrational or precession period in the system. Later on we shall either solve the master equation as a function of time as a continuous variable or replace τ by the actual waiting time between transitions in the kinetic Monte Carlo method.

Given the initial probability distribution $P_1(x_i, 0)$ at time $t = 0$ $(i = 1, \ldots M)$, the probability distribution at time τ (time step 1) is given by

$$P_1(x_i, \tau) = \sum_{j=1}^{M} P_1(x_j, 0)\, P_{1|1}(x_j, 0|x_i, \tau), \tag{2.12}$$

where, as usual, $P_{1|1}(x_j, 0|x_i, \tau)$ gives the probability for a transition to the state x_i at a time τ given that the system is at state x_j at time 0. We will simplify the notation by introducing

$$P_i(s) = P_1(x_i, t = s\tau)$$

and

$$Q_{ij} = P_{1|1}(y_j, 0|y_i, \tau) = P_{1|1}\Big(y_j, s\tau|y_i, (s+1)\tau\Big),$$

where the last equality expresses the assumed time independence of Q_{ij}. Note that Q_{ij} is the *transition probability from j to i*. Let us recall that the usual normalization of conditional probabilities involves a sum over the final states and therefore reads

$$1 = \sum_i P_{1|1}(y_j, t|y_i, t') \quad \Rightarrow \quad \sum_i Q_{ij} = 1 \quad \forall\, j.$$

The sum of probabilities over the final states i is 1 irrespectively of the initial state j.

We can then write Eq. (2.12) as

$$P_i(1) = \sum_j Q_{ij}\, P_j(0)$$

or, introducing the column vector $\vec{P}(s) = [P_1(s),\, P_2(s)\, \ldots\, P_M(s)]$, as

$$\vec{P}(1) = Q\vec{P}(0),$$

where Q is the $M \times M$ matrix Q_{ij}. Note that $Q_{ij} \neq Q_{ji}$. After s time steps we have

$$\vec{P}(s) = Q^s\vec{P}(0) \qquad (Q \text{ indep. of } s).$$

It is easy to see that $\sum_i Q_{ij} = 1\ \forall\, j$ implies that the normalization is preserved, since

$$\sum_i P_i(1) = \sum_i \sum_j Q_{ij}\, P_j(0) = \sum_j P_j(0) = 1.$$

2.6.1 Some properties of Q

The matrix Q_{ij} is, according to our convention, a *left stochastic matrix*. Mathematically, this is defined as a square matrix whose columns consist of positive real numbers summing up to 1 (i.e., $\sum_i Q_{ij} = 1\ \forall\, j$). The columns of the left stochastic matrix are also called *stochastic vectors*, having positive components that sum up to 1. The set S of all *non-singular* stochastic matrices forms a group, namely, the *stochastic group*. The fundamental group properties are easily demonstrated:

1) Closure: Let Q and P be stochastic matrices, then
$$\sum_i (QP)_{ij} = \sum_{ik} Q_{ik}\, P_{kj} = \sum_k \underbrace{\left(\sum_i Q_{ik}\right)}_{=1} P_{kj} = 1.$$

2) Existence of neutral element: $\mathbb{1} \in S$.

3) Existence of inverse: Let Q be a non-singular stochastic matrix and Q^{-1} denote its inverse: $(QQ^{-1}) = \mathbb{1} \quad \Rightarrow \quad \sum_k Q_{ik}\, Q^{-1}_{kj} = \delta_{ij}$. It follows that
$$\sum_i \left(\sum_k Q_{ik}\, Q^{-1}_{kj}\right) = 1 \quad \Rightarrow \quad \sum_k \underbrace{\left(\sum_i Q_{ik}\right)}_{=1} Q^{-1}_{kj} = \sum_k Q^{-1}_{kj} = 1.$$
Consequently, $Q^{-1} \in S$.

2.6.2 The limit of large s: Equilibrium

The behaviour of the probability distribution for large times ($s \to \infty$) depends on the structure of the transition matrix Q. We are particularly interested in *regular* or *ergodic* transition matrices Q. These are characterized by the property that for some power k all elements of Q^k are non-negative, i.e.,

$$(Q^k)_{ij} > 0\ \forall\, ij \quad \text{for some } k \geq 1. \tag{2.13}$$

This is equivalent to requesting that for every pair of states i and j there is a path that connects i and j through a finite number k of transitions:

$$(Q^k)_{ij} = \sum_{l_1} \sum_{l_2} \dots \sum_{l_{k-1}} Q_{il_1} Q_{l_1 l_2} Q_{l_2 l_3} \dots Q_{l_{k-1} j} \tag{2.14}$$

with all matrix elements of Q strictly positive.

If $(Q^k)_{ij} > 0 \; \forall \, ij$ for some k we can apply the Perron-Frobenius theorem which states that the largest eigenvalue of Q is equal to 1, and that this eigenvalue is *non-degenerate*. This implies that all other eigenvalues λ have $|\lambda| < 1$. The eigenvector corresponding to $\lambda = 1$ satisfies

$$Q\vec{\Pi} = \vec{\Pi}$$

and describes the *equilibrium state*. The Perron-Frobenius theorem also ensures that $\Pi_j > 0 \; \forall \, j$, which means that the probability of finding the system in any of its states is finite. This justifies denoting the underlying transition matrix Q as ergodic. Moreover, one can show that

$$\lim_{s \to \infty} Q^s \, \vec{P}(0) = \vec{\Pi}$$

for any starting configuration $\vec{P}(0)$.

For the mathematical demonstration of this important statement we expand $\vec{P}(0) = \sum_{i=1}^{M} \alpha_i \, \vec{\nu}_i$, where $\vec{\nu}_i$ is the ith vector of the Jordan basis of the matrix Q ($\vec{\nu}_1 = \vec{\Pi}$). Note that Q need not be diagonalizable, but that it can always be brought into the Jordan form. Applying $Q^s \, \vec{P}(0)$ and taking the limit of $s \to +\infty$, only the non-degenerate eigenvector Π with the largest eigenvalue survives, since $|\lambda| < 1$ for all $\vec{\nu}_i \neq \vec{\Pi}$. Consequently, if the transition matrix is ergodic, the equilibrium state $\vec{\Pi}$ is unique and independent of the initial state.

Transition matrices that are regular (or ergodic) are also said to be *irreducible*. Actually, it is the Markov chain and the directed graph associated to Q that are irreducible, when all the possible states of the system are connected under Q. In the directed graph derived from Q there is a connection from state j to i whenever $Q_{ij} \neq 0$. The irreducibility of a directed graph means that all states i can be reached starting from any state j [see Eqs. (2.13) and (2.14)]. In our physical problem of the dynamics of a system in an energy landscape, this implies that the ensemble of local minima and intermediate transition states forms a connected set.

2.6.3 Reducible transition matrices and absorbing states

Examples of reducible or non-ergodic transition matrices are matrices that are lower or upper diagonal, or that can be brought to this form by a simple reordering of the states. Another situation in which Q is reducible is when Q can be brought to a block diagonal form. In this case one has more than one equilibrium state. The probability distribution in the limit for $s \to +\infty$ depends on the initial state, since no transitions between the blocks are possible.

In this context it is interesting to introduce the concept of *absorbing states*, which are states from which the system cannot escape (e.g., extinction in the case of a population). Let us consider, for instance, the matrix

$$Q = \begin{pmatrix} 1/2 & 1/3 & 0 \\ 1/2 & 1/3 & 0 \\ 0 & 1/3 & 1 \end{pmatrix}$$

for which the state x_3 is absorbing. In fact, $Q_{32} = 1/2$, but $Q_{i3} = 0 \; \forall \, i \neq 3$. One never gets out of the state 3. In this example there is a unique equilibrium state $\vec{\Pi} = (0, 0, 1)$. However, the system is obviously not ergodic.

2.7 Master equation dynamics

The master equation we derived for a continuous random variable x reads

$$P_1(x,t) = \int \left[P_1(x',t)\, k(x',x) - P(x,t)\, k(x,x') \right] dx'.$$

We focus now on a discrete random variable with a discrete set of outcomes x_1, x_2, $\ldots$ x_M. Each value of x_i corresponds, for example, to one of the minima of our energy surface. We then have

$$\frac{\partial P_1}{\partial t}\,(x_i,t) = \sum_{j=1}^{M} P_1(x_j,t)\, k(x_j,x_i) - P_1(x_i,t)\, k(x_i,x_j),$$

where $k(x_j, x_i)$ is the transition rate for going from state x_j to state x_i. Let us recall that $k(x_j, x_i)$ represents the probability per unit time of making a transition to x_i at time $t + \Delta t$ given that the system is in the state x_j at time t. Let us simplify the usual notation as $P_i(t) = P_1(x_i, t)$ and $k_{ij} = k(x_j, x_i)$ and rewrite the master equation as

$$\frac{dP_i(t)}{dt} = \sum_{j=1}^{M} \left[k_{ij}\, P_j(t) - k_{ji}\, P_i(t) \right]. \tag{2.15}$$

Notice that the term with $j = i$ in the sum cancels out, although k_{ii} is not defined. The "process" that consists in making no transition does not affect $P_i(t)$ and therefore does not contribute to dP_i/dt. The following simple rearrangements allow us to bring Eq. (2.15) in a compact form:

$$\frac{dP_i}{dt} = \sum_{\substack{j=1 \\ j\neq i}}^{M} (k_{ij}\, P_j) - P_i \left(\sum_{l\neq i}^{M} k_{li} \right)$$

or equivalently,

$$\frac{dP_i}{dt} = \sum_{j=1}^{M} \underbrace{\left[(1-\delta_{ij})\, k_{ij} - \delta_{ij} \left(\sum_{l\neq i}^{M} k_{li} \right) \right]}_{W_{ij}} P_j,$$

where we identify

$$W_{ii} = -\sum_{l \neq i} k_{li}$$

as the total rate out of state i and

$$W_{ij} = k_{ij}$$

as the rate from j to i ($i \neq j$). With this definition of the *transition matrix* W, the master equation can be written as

$$\frac{dP_i}{dt} = \sum_{j=1}^{M} W_{ij}\, P_j,$$

or in vector form as

$$\frac{d\vec{P}}{dt} = W\vec{P}.$$

If the matrix k_{ij} is regular or ergodic, the matrix W is also ergodic, i.e., it cannot be decomposed into blocks.

Notice that

$$\begin{aligned}\sum_i W_{ij} &= \sum_i \left[(1-\delta_{ij})\, k_{ij} - \delta_{ij}\left(\sum_{l\neq i} k_{li}\right)\right] \\ &= \sum_{i\neq j} k_{ij} - \sum_{l\neq j} k_{lj} = 0 \quad \forall\, j.\end{aligned}$$

Therefore, the matrix $Q = \mathbb{1} + \alpha W$ for some small $\alpha > 0$ has all the properties of a stochastic matrix. First, we have $W_{ij} = k_{ij} \geq 0\ \forall\, i \neq j$. Second, concerning the diagonal terms, we can always choose a sufficiently small α, or scale the unit of time, so that $\sum_{l\neq j} k_{lj} \leq 1\ \forall\, j$ in order that $Q_{jj} \geq 0$ for all j.

We can then apply the Perron-Frobenius theorem and conclude that W has a unique eigenvalue equal to zero and that all other eigenvalues are negative. One can write $Q = \mathbb{1} + \alpha W$ with $\alpha > 0$ (usually $0 < \alpha < 1$). If q_i are the eigenvalues of Q, we have $-1 \leq q_i \leq 1$ with $q_1 = 1$ being the largest eigenvalue (Perron-Frobenius theorem). Thus, the eigenvalues λ_i of W satisfy $-2 \leq q_i - 1 \leq 0 \Rightarrow\ -2/\alpha \leq \lambda_i \leq 0$, with $\lambda_1 = 0$ a non-degenerate eigenvalue and all other eigenvalues being strictly negative.

Let us first consider the equilibrium configuration $\vec{\Pi} = \vec{P}_{\rm eq}$. It is given by the condition

$$\frac{d\vec{P}_{\rm eq}}{dt} = 0 \quad \Leftrightarrow \quad W\vec{P}_{\rm eq} = 0.$$

$\vec{\Pi} = \vec{P}_{\rm eq}$ is the eigenvector with vanishing eigenvalue. The Perron-Frobenius theorem assets that this eigenvalue always exists and that all the components $\Pi_i = P_i^{\rm eq}$ of the *unique* eigenvector are strictly positive. In terms of the rate constants k_{ij} the equilibrium condition reads

$$\sum_j \left(k_{ij}\, \Pi_j - k_{ji}\, \Pi_i\right) = 0 \quad \forall\, i. \tag{2.16}$$

In most cases of interest the equilibrium is established between every possible pair of states of the system, for example, between every two nearby minima. This is the so-called *detailed balance condition*

$$k_{ij}\,\Pi_j = k_{ji}\,\Pi_i \qquad \text{for all } i \text{ and } j.$$

In other words, each term in the sum vanishes separately. This is a stronger condition than the general equilibrium condition (2.16). It excludes the possibility of three-state equilibrium. However, in the problems of physical interest we have that each pair of minima equilibrates through one or more connecting transition states. In particular the rate constants k_{ij} derived from transition-state theory satisfy detailed balance.

Detailed balance can be used to bring W to a symmetric form and to solve the master equation analytically. We shall assume detailed balance in the following and define

$$\begin{aligned}\tilde{W}_{ij} &= \sqrt{\frac{\Pi_j}{\Pi_i}}\,W_{ij} \qquad \Pi_i > 0\ \forall\, i \text{ (Perron-Frobenius)} \\ &= \sqrt{\frac{\Pi_j}{\Pi_i}}\left[(1-\delta_{ij})\,k_{ij} - \delta_{ij}\left(\sum_{l\neq i} k_{li}\right)\right].\end{aligned}$$

Since $k_{ij}\,\Pi_j = k_{ji}\,\Pi_i$, we have $k_{ij}\sqrt{\frac{\Pi_j}{\Pi_i}}\sqrt{\frac{\Pi_j}{\Pi_i}} = k_{ji} \;\Leftrightarrow\; k_{ij}\sqrt{\frac{\Pi_j}{\Pi_i}} = k_{ji}\sqrt{\frac{\Pi_i}{\Pi_j}}$ and therefore $\tilde{W}_{ij} = \tilde{W}_{ji}$.

Introducing $\tilde{P}_i(t) = P_i(t)/\sqrt{\Pi_i}$ we have

$$\frac{dP_i}{dt} = \sum_j W_{ij}\,P_j \quad \Leftrightarrow \quad \frac{d\tilde{P}_i}{dt} = \sum_j \frac{1}{\sqrt{\Pi_i}}\,W_{ij}\,\sqrt{\Pi_j}\,\frac{1}{\sqrt{\Pi_j}}\,P_j$$

$$\Leftrightarrow \quad \boxed{\frac{d\tilde{P}_i(t)}{dt} = \sum_j \tilde{W}_{ij}\,\tilde{P}_j(t).}$$

Since $\tilde{W}$ is real and symmetric, it is diagonalizable. Let $\vec{\nu}_\alpha$ be the αth eigenvector of $\tilde{W}$ with eigenvalue λ_α, and let us expand the vector $\vec{\tilde{P}}$ in the form

$$\vec{\tilde{P}}(t) = \sum_\alpha p_\alpha(t)\,\vec{\nu}_\alpha \tag{2.17}$$

with $p_\alpha \in \mathbb{R}$. From Eq. (2.17) it follows that

$$\frac{d\vec{\tilde{P}}}{dt} = \sum_\alpha p_\alpha(t)\,W\,\vec{\nu}_\alpha = \sum_\alpha p_\alpha(t)\,\lambda_\alpha\,\vec{\nu}_\alpha.$$

Comparison with the straightforward differentiation of Eq. (2.17) implies

$$\sum_\alpha \frac{dp_\alpha}{dt}\,\vec{\nu}_\alpha = \sum_\alpha p_\alpha(t)\,\lambda_\alpha\,\vec{\nu}_\alpha.$$

This system of first-order linear differential equations is solved by the usual exponential ansatz

$$p_\alpha(t) = p_\alpha(0)\,e^{\lambda_\alpha\,t}.$$

Replacing in Eq. (2.17) one obtains

$$\vec{\tilde{P}}(t) = \sum_{\alpha} p_\alpha(0)\, \vec{\nu}_\alpha\, e^{\lambda_\alpha\, t}$$

with $\lambda_1 = 0$ and $\lambda_\alpha < 0\ \forall\, \alpha \geq 2$. Notice that the coefficients $p_\alpha(0)$ are given by the initial condition $\vec{\tilde{P}}(0)$. They are obtained by expanding the latter in terms of the eigenvectors $\vec{\nu}_\alpha$, namely, $p_\alpha(0) = \vec{\tilde{P}}(0) \cdot \vec{\nu}_\alpha$. If one wishes to abandon the compact vector notation in favor of the more explicit component form, one may write $\vec{\nu}_\alpha = (\nu_{1\alpha},\, \nu_{2\alpha},\, \ldots\, \nu_{M\alpha})$ and

$$\tilde{P}_i(t) = \sum_{\alpha} p_\alpha(0)\, \nu_{i\alpha}\, e^{\lambda_\alpha\, t}.$$

Finally, the solution of the Master equation reads

$$P_i(t) = \sqrt{\Pi_i} \sum_{\alpha} p_\alpha(0)\, \nu_{i\alpha}\, e^{\lambda_\alpha\, t}.$$

Since all eigenvalues are negative except $\lambda_1 = 0$, it is clear that for $t \to \infty$ we have $\vec{P}(t) \to \vec{\nu}_1 = \vec{\Pi}$ for any initial condition. In this context it is important to remark that $p_1(0) > 0$ for all initial conditions $P_i(0)$. In fact, the probabilities $P_i(0) \geq 0$, which implies $\tilde{P}_i(0) \geq 0$ for all i. Since the components of the equilibrium state $\Pi_i = \nu_{i1}$ are all strictly positive (Perron-Frobenius theorem), we always have $p_1(0) = \vec{\tilde{P}}(0) \cdot \vec{\nu}_1 > 0$. In other words, ergodicity implies that all initial states have a non-vanishing projection on the equilibrium state. Let us finally point out that the eigenvalues of W and $\tilde{W}$ are the same since

$$\sum_j \tilde{W}_{ij}\, \tilde{P}_j = \lambda\, \tilde{P}_i \ \Rightarrow\ \sum_j \tilde{W}_{ij}\, P_j/\sqrt{\Pi_j} = \lambda\, P_i/\sqrt{\Pi_i} \ \Rightarrow\ \sum_j W_{ij}\, P_j = \lambda\, P_j.$$

The analytical solution of the master equation requires the diagonalization of $\tilde{W}$. It has the advantage that, once $\tilde{W}$ is diagonalized, all possible initial conditions can be propagated at a low computational cost. If the size of the problem is too large, i.e., for a too large number M of local minima, one can take advantage of the sparse character of W and use propagation methods. First of all the Lanczos method can be applied to the matrix $\tilde{W}$ in order to estimate the eigenvalue with the largest $|\lambda_c|$. At a second step the relevant spectrum for long times, i.e., $|\lambda_c| \to 0$, can also be computed with the Lanczos method after introducing an appropriate spectral shift so that the smallest eigenvalue has the largest absolute value. If the spectrum of eigenvalues of $\tilde{W}$ is wide, very fast and very slow processes are present. In this case it is advisable to introduce a variable time-integration step size to account for fast initial relaxation followed by slower processes until equilibrium is reached.

2.8 The stationary point network

In order to understand the dynamics of a complex system it is important to determine the connectivity of the network of local minima. This information is of course contained in the matrix of transition rates k_{mn}. It is clear that systems with similar distributions of local minima and barrier energies can have very different dynamical properties depending on how the minima are linked together. A practical strategy to sample the stationary points (minima and saddles) is discussed in the following.

2.8.1 Sampling stationary points

For the dynamics we need a set or network of minima that is connected through saddle points. One can follow the following scheme:

i) Starting from a given minimum one searches for the "closest" saddle point (SP) starting along the direction of the eigenvector of the Hessian matrix with the lowest frequency. It is along this direction that one can reasonably expect to find the lowest energy barrier or the most important relaxation (i.e., change of configuration) for a given barrier height.

ii) Once the SP is found, one follows the gradient starting from the SP along the two opposite directions defined by the eigenvector with negative frequency ($\hat{e}_1$ and $-\hat{e}_1$). This provides the *minimum energy path* (MEP) which contains the physically relevant information on the relaxation mechanism, the path length and barrier height (compare with catastrophe theory). In this way one usually finds two different minima. However, in some cases (for example when the landscape looks like a mexican hat) the MEP starting along $\vec{e}_1$ and $-\vec{e}_1$ might lead to the same initial minimum. These rather rare situations are of course irrelevant for the relaxation.

iii) There are several possibilities after step ii):

   a) One or both of the minima found are already known. For example, one of the ends of the MEP is likely to be the minimum from which we started. If one of the minima is new, this is added to the list of minima. Then, either we keep on searching SPs from the minimum where we were at i), or we move to another minimum. This decision will actually be taken at step iv).
   b) Both minima M_1 and M_2 found from the SP are new. In this case one adds them to the list. At the end one checks if they get connected to the network. If not, one may try to connect them by force with double-ended techniques such as the nudged elastic band method.
   c) One or both of the minima are already known, but they are *not* the starting minimum at i). In this case one attaches the link to the list of transitions, provided the saddle point is new.

iv) To follow the search and in order to explore the landscape systematically one may search along the two directions of the normal mode with lowest frequency at the minimum and then follow systematically with increasing vibrational frequency. A full search of all normal modes is neither feasible nor really necessary. One may then conceive different criteria for stopping the search on one minimum and move to another one. These are dictated by the physical problem and the available resources. For example, one may stop once the barriers found are beyond a given threshold, hoping that the other modes will lead to higher barriers and lower rates, or one may search preferentially in a given "direction", i.e., one may bias the search in the spirit of forward flux approaches. For instance, if the dynamics of a magnetization reversal process is investigated, one may search preferentially in directions where the projection of the total magnetization along the applied field increases. Calculating projections along a given direction or distance in configurational space provides a means of quantifying a physically motivated bias.

As in any simulation, an exhaustive search of stationary points is not necessary to predict the results with a given accuracy. One actually expects that the dynamics of the phenomenon under study can be understood by identifying the most relevant relaxation mechanisms. However, it is also clear that as the temperature increases, incomplete searches will systematically underestimate the relaxation rates. A partial remedy for this can be to reweight the superposition sums [*Energy Landscapes*, D. J. Wales, Cambridge University Press, Cambridge/UK, 2003]. In any case the sampling strategy has to cope with the compromise between exploring some part of the landscape more exhaustively or to explore larger regions more superficially. The appropriate choice depends of course on the specific problem under study.

2.9 Kinetic Monte Carlo method

The kinetic Monte Carlo (KMC) scheme provides an alternative solution of the master equation. In the approaches discussed so far, which were based on the eigenvalues and eigenvectors of $\tilde{W}$ or on the numerical propagation of $\vec{P}$ using $\dot{\vec{P}} = W\vec{P}$, one needs to determine the relevant set of starting points and rate constants prior to the simulation. The KMC method allows one to perform propagations by determining the rate constants and the visited minima "on the flight," thereby avoiding the exploration of large domains of the energy landscape, which might not be visited when the simulation starts with the initial conditions of actual experimental or theoretical interest.

Consider a system in a minimum i. The total rate or probability per unit time for the system to jump to another minimum is

$$k = \sum_{j \neq i} k_{ji},$$

provided that all possible transitions from minimum i to another minimum j are independent processes. This presupposes that the resident time at minimum i is large enough, so that any memory or inertial effects are washed out. In other words, we assume that the system is thermalized at minimum i, which is quite reasonable for the rare processes we are interested in.

The transition out of minimum i is a *Poisson process* or *continuous-time Markov process*. It can occur at any time, just the probability for its occurrence per unit time k is known. What is then the probability that the transition to some other minimum occurs between the times t and $t + \Delta t$? This defines the so-called *waiting time probability density* $P(t)$ and can be calculated as follows:

$$P\left\{\text{decay within}\,[t, t+\Delta t]\right\} = P(t)\,\Delta t = (1-p)^N p,$$

where $N = \dfrac{t}{\Delta t}$ and $p = k\,\Delta t = \dfrac{kt}{N}$. Then we have

$$P(t)\,\Delta t = \left(1 - \frac{kt}{N}\right)^N k\,\Delta t \xrightarrow{N\to\infty} e^{-kt}\,k\,\Delta t.$$

Notice that $P(t)$ is properly normalized, since

$$\int_0^{+\infty} P(t)\,dt = k\int_0^{+\infty} e^{-kt}\,dt = 1.$$

The average waiting time τ is therefore given by

$$\tau = \int_0^{+\infty} t\, P(t) = \int_0^{+\infty} t\, k\, e^{-kt}\, dt = \frac{1}{k} \int_0^{+\infty} x\, e^{-x}\, dx = \frac{1}{k}.$$

The KMC method simulates the succession of Markov processes as follows:

i) Starting from a given minimum i, calculate the rates k_{ji} for all significant escape routes out of i. In practice, one computes the total rate $k = \sum_j k_{ji}$ until it converges within a predefined reasonable accuracy. This yields the average waiting time $\tau = 1/k$. Notice that omitting a significant relaxation channel, i.e., omitting some $k_{ji} > 0$, systematically implies an underestimation of k and thus an overestimation of τ.

ii) Choose randomly one of the escape channels j according to the probability $p_j = \dfrac{k_{ji}}{\sum_j k_{ji}} = \tau\, k_{ji}$.

iii) Draw randomly a waiting time $t \geq 0$ according to the probability density function for the waiting time $P(t) = k\, e^{-kt}$.

iv) Move the system to the minimum j drawn in step iii) and advance the time of the simulation by the time t drawn from the Poisson distribution in iii). Finally, go back to i).

How to obtain random numbers with a probability distribution $P(t) = k\, e^{-kt}$? We seek for a function $t = t(x)$ such that $P(t) = k\, e^{-kt}$ when x is uniformly distributed in the interval [0,1]. For this we must have

$$P(t)\, dt = p(x)\, dx = dx,$$

where we have used that $p(x) = 1$ for $0 \leq x \leq 1$. Substituting one obtains

$$k\, e^{-kt}\, dt = dx \quad \Leftrightarrow \quad \frac{dt}{dx} = \frac{1}{k}\, e^{+kt}.$$

This equation can be easily solved by the method of separation of variables:

$$\int k\, e^{-kt}\, dt = \int dx \quad \Rightarrow \quad -e^{-kt} = x - x_0 \quad \Rightarrow \quad -kt = \ln(x_0 - x)$$

$$t = -\frac{1}{k} \ln(x_0 - x) \quad \begin{cases} x = 0 \text{ for } t = 0 \quad \Rightarrow \quad x_0 = 1 \\ x \to 1 \quad \Rightarrow \quad t \to +\infty \end{cases}$$

Finally, one obtains

$$t = -\frac{1}{k} \ln(1 - x) = \frac{1}{k} \ln \frac{1}{1 - x}.$$

Since $1 - x$ is uniformly distributed, one can also use $t = -(1/k) \ln x$.

The KMC algorithm provides a formally exact solution of the master equation provided that all the transition processes are taken into account. As in the exponential solution, the limitation comes from the detail with which the landscape is described. One expects

that the dynamics will not be distorted significantly if some rare processes (with small k) are neglected unless these rare processes are very numerous. However, this situation can be encountered in glassy systems or in other highly connected networks of low-energy local minima.

It is interesting to compare the transition rates k_{ij} corresponding to the different saddle points i surrounding a given minimum j in the framework of transition-state theory. These are given by

$$k_{ij} = \frac{k_B T}{2\pi\hbar} \frac{Q^*_{ij}(T)}{Q(j)(T)},$$

where Q^*_{ij} is the partition function at the saddle point connecting j to i. Thus, for a given starting minimum j, $k_{ij} \propto Q^*_{ij}(T)$ and the probability of escaping from j through the saddle point leading to minimum i (channel i) is given by

$$p_i = \frac{Q^*_{ij}(T)}{\sum_j Q^*_{ij}(T)}.$$

It is tempting to explore the possibility of avoiding the calculation of "all" (in practice many) relaxation channels and their rates before doing a move. Instead, one could search for a saddle point starting from j along a random direction, find the neighboring minimum i and calculate the rate k_{ij}. One could argue that the resulting probabilities of escaping through channel i would roughly correspond to $p_i \propto Q^*_{ij}(T)$, since $Q^*_{ij}(T)$ gives a measure of the catchment area of the saddle point i. The rate k_{ij} is, however, much smaller than the total escape rate $k = \sum_i k_{ij}$. Consequently, the escape time $\tau_i = 1/k_{ij}$ is much longer than the actual escape time $\tau = 1/k$. It is therefore *not* reasonable to advance the time by $\tau_i = 1/k_{ij}$. In fact, other faster mechanisms could largely dominate the dynamics in this time scale. Without a knowledge of the dominant channels, it is possible that the system escapes with a large probability in some other minimum different from the one found by starting in a randomly chosen direction. One would then have to take a small time step Δt, and make the jump to i on the basis of the rate k_{ij} using that for $\Delta t \ll 1/k_{ij}$ the probability of jumping is $k_{ij}\,\Delta t$. The consequence would be a very small acceptance rate for the jump to i. One would then search another direction until $\sum_i k_{ij}$ converges. The method is no better than the original KMC, actually worse, since in KMC one would search systematically in the directions that are likely to yield the largest k_{ij} (low frequency eigenmodes). In addition, one can bias the search on physical grounds taking into account the magnetization relaxation or in general the distance to the product, if known.

2.10 Time dependence of the configurational entropy

The configurational or mixing entropy S resulting from the distribution of the system throughout its stationary states is given by

$$S = -k_B \sum_i P_i \ln P_i, \tag{2.18}$$

where the sum runs over all local minima. This represents the logarithm of the number of accessible states for a given probability distribution P_i. Eq. (2.18) holds both at

equilibrium and out of equilibrium, at any time t. The time dependence of S is given by

$$\frac{dS}{dt} = -k_B \sum_i \left(\frac{dP_i}{dt} \ln P_i - \frac{dP_i}{dt} \right).$$

Since $\sum_i P_i = 1 \; \forall \, t$, the sum of the last terms vanishes. It follows that

$$\frac{dS}{dt} = -k_B \sum_i \frac{dP_i}{dt} \ln P_i. \tag{2.19}$$

According to the master equation, i.e., assuming a Markovian dynamics without memory effects, we have

$$\frac{dP_i}{dt} = \sum_j \left(k_{ij} P_j - k_{ji} P_i \right). \tag{2.20}$$

Replacing Eq. (2.20) in (2.19) we obtain

$$\frac{dS}{dt} = -k_B \sum_{ij} \left(k_{ij} P_j - k_{ji} P_i \right) \ln P_i$$

or equivalently

$$\frac{dS}{dt} = -\frac{k_B}{2} \sum_{ij} \left(k_{ij} P_j - k_{ji} P_i \right) \left(\ln P_i - \ln P_j \right).$$

In general $k_{ij} \neq k_{ji}$ and consequently $\dfrac{dS}{dt}$ does not have a defined sign. Therefore, S needs not to increase monotonously with time. This should be clear since the assumption of Markovian dynamics does not exclude, for example, the presence of absorbing states for which the equilibrium configuration $\vec{\pi}$ has zero entropy ($\pi_i = \delta_{i0}$, where 0 is the absorbing state). This is not in contradiction with Boltzmann's H-theorem, since the configurational or mixing entropy does not take into account the change in the entropy associated to the other degrees of freedom. In particular the change in phase space due to the change in kinetic energy is ignored in Eq. (2.18).

However, in a microscopic quantum theory we know from Fermi's golden rule that the transition rate

$$k_{ij} = \left(\frac{2\pi}{\hbar} \right) |V_{ij}|^2 \, \delta(E_i - E_j)$$

is symmetric ($k_{ij} = k_{ji}$, since $V_{ij} = V_{ji}^*$). In this case ($k_{ij} = k_{ji}$) we have

$$\frac{dS}{dt} = -\frac{k_B}{2} \sum_{ij} k_{ij} \left(P_j - P_i \right) \left(\ln P_i - \ln P_j \right).$$

Since $\ln(x)$ is a monotonously increasing function, $(\ln P_i - \ln P_j)$ has always the opposite sign as $(P_j - P_i)$. Consequently $\dfrac{dS}{dt} \geq 0 \; \forall \, t$ and $\forall \, P_i(t)$. The assumption of symmetric rates ($k_{ij} = k_{ji}$) and of Markovian dynamics (master equation) implies that the time evolution always leads to an increase of the entropy of the system.

2.11 Transition state theory

We follow here an article by W. H. Miller [Acc. Chem. Res. **26**, 174 (1993)] to which the reader should refer for further details.

Table of Contents