

Probability

A General Definitions

The entity under investigation is a *random variable* x , which has a set of possible *outcomes* $\mathcal{S} \equiv \{x_1, x_2, \dots\}$. The outcomes may be *discrete* as in the case of a coin toss, $\mathcal{S}_{\text{coin}} = \{\text{head}, \text{tail}\}$, or a die throw, $\mathcal{S}_{\text{dice}} = \{1, 2, 3, 4, 5, 6\}$, or *continuous* as for the velocity of a particle in a gas, $\mathcal{S}_{\vec{v}} = \{-\infty < v_x, v_y, v_z < \infty\}$, or the energy of an electron in a metal at zero temperature, $\mathcal{S}_\epsilon = \{0 \leq \epsilon \leq \epsilon_F\}$. An *event* is any subset of outcomes $E \subset \mathcal{S}$, and is assigned a *probability* $p(E)$, e.g. $p_{\text{dice}}(\{1\}) = 1/6$, or $p_{\text{dice}}(\{1, 3\}) = 1/3$. From an axiomatic point of view, the probabilities must satisfy the following conditions:

- (i) *Positivity*: $p(E) \geq 0$, i.e. all probabilities must be real and non-negative.
- (ii) *Additivity*: $p(A \text{ or } B) = p(A) + p(B)$, if A and B are disconnected events.
- (iii) *Normalization*: $p(\mathcal{S}) = 1$, i.e. the random variable must have some outcome in \mathcal{S} .

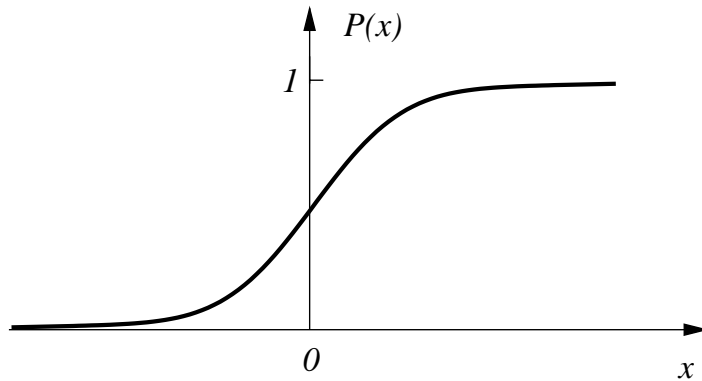
From a practical point of view, we would like to know how to assign probability values to various outcomes. There are two possible approaches:

- (1) *Objective* probabilities are obtained *experimentally* from the relative frequency of the occurrence of an outcome in many tests of the random variable. If the random process is repeated N times, and the event A occurs N_A times, then

$$p(A) = \lim_{N \rightarrow \infty} \frac{N_A}{N}.$$

For example, a series of $N = 100, 200, 300$ throws of a dice may result in $N_1 = 19, 30, 48$ occurrences of 1. The ratios .19, .15, .16 provide an increasingly more reliable estimate of the probability $p_{\text{dice}}(\{1\})$.

- (2) *Subjective* probabilities provide a *theoretical* estimate based on the uncertainties related to lack of precise knowledge of outcomes. For example, the assessment $p_{\text{dice}}(\{1\}) = 1/6$, is based on the knowledge that there are six possible outcomes to a dice throw, and that in the absence of any prior reason to believe that the dice is biased, all six are equally likely. All assignments of probability in Statistical Mechanics are subjectively based. The consequences of such subjective assignments of probability have to be checked against measurements, and they may need to be modified as more information about the outcomes becomes available.



1. A typical cumulative probability function.

B One Random Variable

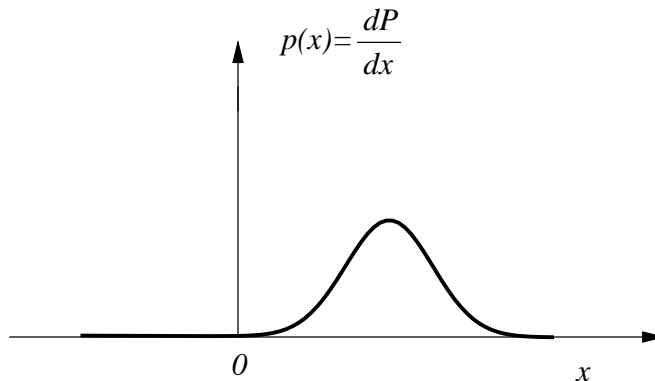
As the properties of a discrete random variable are rather well known, here we focus on continuous random variables, which are more relevant to our purposes. Consider a random variable x , whose outcomes are real numbers, i.e. $\mathcal{S}_x = \{-\infty < x < \infty\}$.

- The *cumulative probability function* (CPF) $P(x)$, is the probability of an outcome with *any value* less than x , i.e. $P(x) = \text{prob.}(E \subset [-\infty, x])$. $P(x)$ must be a monotonically increasing function of x , with $P(-\infty) = 0$ and $P(+\infty) = 1$.

- The *probability density function* (PDF) is defined by $p(x) \equiv dP(x)/dx$. Hence, $p(x)dx = \text{prob.}(E \in [x, x + dx])$. As a probability density, it is *positive*, and normalized such that

$$\text{prob.}(\mathcal{S}) = \int_{-\infty}^{\infty} dx p(x) = 1 . \quad (1)$$

Note that since $p(x)$ is a *probability density*, it has no upper bound, i.e. $0 < p(x) < \infty$.



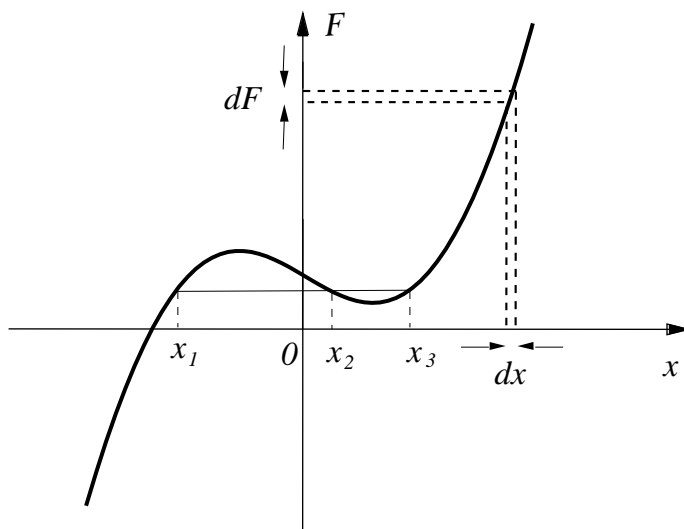
2. A typical probability density function.

- The expectation value of any function $F(x)$, of the random variable is

$$\langle F(x) \rangle = \int_{-\infty}^{\infty} dx p(x)F(x) . \quad (2)$$

The function $F(x)$ is itself a random variable, with an associated PDF of $p_F(f)df = \text{prob.}(F(x) \in [f, f + df])$. There may be multiple solutions x_i , to the equation $F(x) = f$, and

$$p_F(f)df = \sum_i p(x_i)dx_i, \implies p_F(f) = \sum_i p(x_i) \left| \frac{dx}{dF} \right|_{x=x_i} . \quad (3)$$



3. Obtaining the PDF for the function $F(x)$.

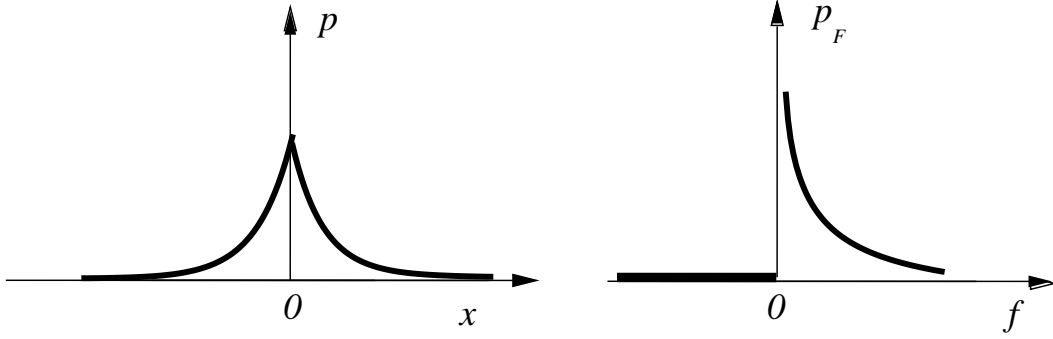
The factors of $|dx/dF|$ are the *Jacobians* associated with the change of variables from x to F . For example, consider $p(x) = \lambda \exp(-\lambda|x|)/2$, and the function $F(x) = x^2$. There are two solutions to $F(x) = f$, located at $x_{\pm} = \pm\sqrt{f}$, with corresponding Jacobians $|\pm f^{-1/2}/2|$. Hence,

$$P_F(f) = \frac{\lambda}{2} \exp(-\lambda\sqrt{f}) \left(\left| \frac{1}{2\sqrt{f}} \right| + \left| \frac{-1}{2\sqrt{f}} \right| \right) = \frac{\lambda \exp(-\lambda\sqrt{f})}{2\sqrt{f}},$$

for $f > 0$, and $p_F(f) = 0$ for $f < 0$. Note that $p_F(f)$ has an (integrable) divergence at $f = 0$.

- *Moments* of the PDF are expectation values for powers of the random variable. The n^{th} moment is

$$m_n \equiv \langle x^n \rangle = \int dx p(x) x^n. \quad (4)$$



4. Probability density functions for x , and $F = x^2$.

- *The characteristic function* is the generator of moments of the distribution. It is simply the Fourier transform of the PDF, defined by

$$\tilde{p}(k) = \langle e^{-ikx} \rangle = \int dx p(x) e^{-ikx}. \quad (5)$$

The PDF can be recovered from the characteristic function through the inverse Fourier transform

$$p(x) = \frac{1}{2\pi} \int dk \tilde{p}(k) e^{+ikx}. \quad (6)$$

Moments of the distribution are obtained by expanding $\tilde{p}(k)$ in powers of k ,

$$\tilde{p}(k) = \left\langle \sum_{n=0}^{\infty} \frac{(-ik)^n}{n!} x^n \right\rangle = \sum_{n=0}^{\infty} \frac{(-ik)^n}{n!} \langle x^n \rangle. \quad (7)$$

Moments of the PDF around any point x_0 can also be generated by expanding

$$e^{ikx_0} \tilde{p}(k) = \langle e^{-ik(x-x_0)} \rangle = \sum_{n=0}^{\infty} \frac{(-ik)^n}{n!} \langle (x-x_0)^n \rangle. \quad (8)$$

- *The cumulant generating function* is the logarithm of the characteristic function. Its expansion generates the *cumulants* of the distribution defined through

$$\ln \tilde{p}(k) = \sum_{n=1}^{\infty} \frac{(-ik)^n}{n!} \langle x^n \rangle_c. \quad (9)$$

Relations between moments and cumulants can be obtained by expanding the logarithm of $\tilde{p}(k)$ in eq.(7), and using

$$\ln(1 + \epsilon) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{\epsilon^n}{n}. \quad (10)$$

The first four cumulants are called the *mean*, *variance*, *skewness*, and *curtosis* of the distribution respectively, and are obtained from the moments as

$$\begin{aligned}
\langle x \rangle_c &= \langle x \rangle, \\
\langle x^2 \rangle_c &= \langle x^2 \rangle - \langle x \rangle^2, \\
\langle x^3 \rangle_c &= \langle x^3 \rangle - 3 \langle x^2 \rangle \langle x \rangle + 2 \langle x \rangle^3, \\
\langle x^4 \rangle_c &= \langle x^4 \rangle - 4 \langle x^3 \rangle \langle x \rangle - 3 \langle x^2 \rangle^2 + 12 \langle x^2 \rangle \langle x \rangle^2 - 6 \langle x \rangle^4.
\end{aligned} \tag{11}$$

The cumulants provide a useful and compact way of describing a PDF.

An important theorem allows easy computation of moments in terms of the cumulants: Represent the n^{th} cumulant graphically as a *connected cluster* of n points. The m^{th} moment is then obtained by summing all possible subdivisions of m points into groupings of smaller (connected or disconnected) clusters. The contribution of each subdivision to the sum is the product of the connected cumulants that it represents. Using this result the first four moments are computed graphically as

$$\begin{aligned}
\langle x \rangle &= \bullet \\
\langle x^2 \rangle &= \text{---}\bullet\text{---}\bullet + \bullet\bullet \\
\langle x^3 \rangle &= \text{---}\bullet\text{---}\bullet\text{---}\bullet + 3 \text{---}\bullet\text{---}\bullet + \bullet\bullet\bullet \\
\langle x^4 \rangle &= \text{---}\bullet\text{---}\bullet\text{---}\bullet + 4 \text{---}\bullet\text{---}\bullet\text{---}\bullet + 3 \text{---}\bullet\text{---}\bullet + 6 \text{---}\bullet\text{---}\bullet + \bullet\bullet\bullet\bullet
\end{aligned}$$

5. Graphical computation of the first four moments.

The corresponding algebraic expressions are

$$\begin{aligned}
\langle x \rangle &= \langle x \rangle_c, \\
\langle x^2 \rangle &= \langle x^2 \rangle_c + \langle x \rangle_c^2, \\
\langle x^3 \rangle &= \langle x^3 \rangle_c + 3 \langle x^2 \rangle_c \langle x \rangle_c + \langle x \rangle_c^3, \\
\langle x^4 \rangle &= \langle x^4 \rangle_c + 4 \langle x^3 \rangle_c \langle x \rangle_c + 3 \langle x^2 \rangle_c^2 + 6 \langle x^2 \rangle_c \langle x \rangle_c^2 + \langle x \rangle_c^4.
\end{aligned} \tag{12}$$

This theorem, which is the starting point for various diagrammatic computations in statistical mechanics and field theory, is easily proved by equating the expressions in eqs. (7) and (9) for $\tilde{p}(k)$

$$\sum_{m=0}^{\infty} \frac{(-ik)^m}{m!} \langle x^m \rangle = \exp \left[\sum_{n=1}^{\infty} \frac{(-ik)^n}{n!} \langle x^n \rangle_c \right] = \prod_n \sum_{p_n} \left[\frac{(-ik)^{np_n}}{p_n!} \left(\frac{\langle x^n \rangle_c}{n!} \right)^{p_n} \right]. \tag{13}$$

Matching the powers of $(-ik)^m$ on the two sides of the above expression leads to

$$\langle x^m \rangle = \sum_{\{p_n\}}' m! \prod_n \frac{1}{p_n! (n!)^{p_n}} \langle x^n \rangle_c^{p_n} . \quad (14)$$

The sum is restricted such that $\sum np_n = m$, and leads to the graphical interpretation given above, as the numerical factor is simply the number of ways of breaking m points into $\{p_n\}$ clusters of n points.

C Some Important Probability Distributions

The properties of three commonly encountered probability distributions are examined in this section.

(1) *The normal (Gaussian) distribution* describes a continuous real random variable x , with

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x-\lambda)^2}{2\sigma^2} \right] . \quad (15)$$

The corresponding characteristic function also has a Gaussian form,

$$\tilde{p}(k) = \int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x-\lambda)^2}{2\sigma^2} - ikx \right] = \exp \left[-ik\lambda - \frac{k^2\sigma^2}{2} \right] . \quad (16)$$

Cumulants of the distribution can be identified from $\ln \tilde{p}(k) = -ik\lambda - k^2\sigma^2/2$, using eq.(9), as

$$\langle x \rangle_c = \lambda \quad , \quad \langle x^2 \rangle_c = \sigma^2 \quad , \quad \langle x^3 \rangle_c = \langle x^4 \rangle_c = \dots = 0 \quad . \quad (17)$$

The normal distribution is thus completely specified by its two first cumulants. This makes the computation of moments using the cluster expansion (eqs.(12)) particularly simple, and

$$\begin{aligned} \langle x \rangle &= \lambda , \\ \langle x^2 \rangle &= \sigma^2 + \lambda^2 , \\ \langle x^3 \rangle &= 3\sigma^2\lambda + \lambda^3 , \\ \langle x^4 \rangle &= 3\sigma^4 + 6\sigma^2\lambda^2 + \lambda^4 , \quad \dots \quad . \end{aligned} \quad (18)$$

The normal distribution serves as the starting point for most perturbative computations in field theory. The vanishing of higher cumulants implies that all graphical computations involve only products of one point, and two point (known as propagators) clusters.

(2) *The binomial distribution:* Consider a random variable with two outcomes A and B (e.g. a coin toss) of relative probabilities p_A and $p_B = 1 - p_A$. The probability that in N trials the event A occurs exactly N_A times (e.g. 5 heads in 12 coin tosses), is given by the binomial distribution

$$p_N(N_A) = \binom{N}{N_A} p_A^{N_A} p_B^{N-N_A} . \quad (19)$$

The prefactor,

$$\binom{N}{N_A} = \frac{N!}{N_A!(N-N_A)!} , \quad (20)$$

is just the coefficient obtained in the binomial expansion of $(p_A + p_B)^N$, and gives the number of possible orderings of N_A events A and $N_B = N - N_A$ events B . The characteristic function for this discrete distribution is given by

$$\tilde{p}_N(k) = \langle e^{-ikN_A} \rangle = \sum_{N_A=0}^N \frac{N!}{N_A!(N-N_A)!} p_A^{N_A} p_B^{N-N_A} e^{-ikN_A} = (p_A e^{-ik} + p_B)^N . \quad (21)$$

The resulting cumulant generating function is

$$\ln \tilde{p}_N(k) = N \ln (p_A e^{-ik} + p_B) = N \ln \tilde{p}_1(k), \quad (22)$$

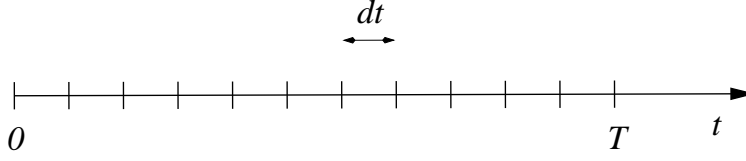
where $\ln \tilde{p}_1(k)$ is the cumulant generating function for a single trial. Hence, the cumulants after N trials are simply N times the cumulants in a single trial. In each trial, the allowed values of N_A are 0 and 1 with respective probabilities p_B and p_A , leading to $\langle N_A^\ell \rangle = p_A$, for all ℓ . After N trials the first two cumulants are

$$\langle N_A \rangle_c = N p_A , \quad \langle N_A^2 \rangle_c = N (p_A - p_A^2) = N p_A p_B . \quad (23)$$

A measure of fluctuations around the mean is provided by the *standard deviation*, which is the square root of the variance. While the mean of the binomial distribution scales as N , its standard deviation only grows as \sqrt{N} . Hence, the *relative uncertainty* becomes smaller for large N .

The binomial distribution is straightforwardly generalized to a *multinomial* distribution, when the several outcomes $\{A, B, \dots, M\}$ occur with probabilities $\{p_A, p_B, \dots, p_M\}$. The probability of finding outcomes $\{N_A, N_B, \dots, N_M\}$ in a total of $N = N_A + N_B + \dots + N_M$ trials is

$$p_N(\{N_A, N_B, \dots, N_M\}) = \frac{N!}{N_A! N_B! \dots N_M!} p_A^{N_A} p_B^{N_B} \dots p_M^{N_M} . \quad (24)$$



6. Subdividing the time interval into small segments of size dt .

(3) *The Poisson distribution:* The classical example of a Poisson process is radioactive decay. Observing a piece of radioactive material over a time interval T shows that:

- (a) The probability of one and only one event (decay) in the interval $[t, t + dt]$ is proportional to dt as $dt \rightarrow 0$,
- (b) The probabilities of events at different intervals are independent of each other.

The probability of observing exactly M decays in the interval T is given by the Poisson distribution. It is obtained as a limit of the binomial distribution by subdividing the interval into $N = T/dt \gg 1$ segments of size dt . In each segment, an event occurs with probability $p = \alpha dt$, and there is no event with probability $q = 1 - \alpha dt$. As the probability of more than one event in dt is too small to consider, the process is equivalent to a binomial one. Using eq.(21), the characteristic function is given by

$$\tilde{p}(k) = (pe^{-ik} + q)^n = \lim_{dt \rightarrow 0} [1 + \alpha dt (e^{-ik} - 1)]^{T/dt} = \exp [\alpha (e^{-ik} - 1)T] \quad . \quad (25)$$

The Poisson PDF is obtained from the inverse Fourier transform in eq.(6) as

$$p(x) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} \exp [\alpha (e^{-ik} - 1)T + ikx] = e^{-\alpha T} \int_{-\infty}^{\infty} \frac{dk}{2\pi} e^{ikx} \sum_{M=0}^{\infty} \frac{(\alpha T)^M}{M!} e^{-ikM} \quad , \quad (26)$$

using the power series for the exponential. The integral over k is

$$\int_{-\infty}^{\infty} \frac{dk}{2\pi} e^{ik(x-M)} = \delta(x - M) \quad , \quad (27)$$

leading to

$$p_{\alpha T}(x) = \sum_{m=0}^{\infty} e^{-\alpha T} \frac{(\alpha T)^m}{m!} \delta(x - m) \quad . \quad (28)$$

The probability of M events is thus $p_{\alpha T}(M) = e^{-\alpha T} (\alpha T)^M / M!$. The cumulants of the distribution are obtained from the expansion

$$\ln \tilde{p}_{\alpha T}(k) = \alpha T (e^{-ik} - 1) = \alpha T \sum_{n=1}^{\infty} \frac{(-ik)^n}{n!} \quad , \quad \implies \quad \langle M^n \rangle_c = \alpha T \quad . \quad (29)$$

All cumulants have the same value, and the moments are obtained from eqs.(12) as

$$\langle M \rangle = (\alpha T), \quad \langle M^2 \rangle = (\alpha T)^2 + (\alpha T), \quad \langle M^3 \rangle = (\alpha T)^3 + 3(\alpha T)^2 + (\alpha T). \quad (30)$$

Example: Assuming that stars are randomly distributed in the galaxy (clearly unjustified) with a density n , what is the probability that the nearest star is at a distance R ?

Since, the probability of finding a star in a small volume dV is ndV , and they are assumed to be independent, the number of stars in a volume V is described by a Poisson process as in eq.(28), with $\alpha = n$. The probability $p(R)$, of encountering the first star at a distance R is the product of the probabilities $p_{nV}(0)$, of finding zero stars in the volume $V = 4\pi R^3/3$ around the origin, and $p_{ndV}(1)$, of finding one star in the shell of volume $dV = 4\pi R^2 dR$ at a distance R . Both $p_{nV}(0)$ and $p_{ndV}(1)$ can be calculated from eq.(28), and

$$\begin{aligned} p(R)dR &= p_{nV}(0) p_{ndV}(1) = e^{-4\pi R^3 n/3} e^{-4\pi R^2 ndR} 4\pi R^2 ndR, \\ \implies p(R) &= 4\pi R^2 n \exp\left(-\frac{4\pi}{3} R^3 n\right). \end{aligned} \quad (31)$$

D Many Random Variables

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

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

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

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

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

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

$$p(x_1, \dots, x_m) = \int \prod_{i=m+1}^N dx_i p(x_1, \dots, x_N). \quad (34)$$

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

$$\mathcal{N} = \int d^3\vec{v} p(\vec{x}, \vec{v}) = p(\vec{x}), \quad (35)$$

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

$$p(x_1, \dots, x_m | x_{m+1}, \dots, x_N) = \frac{p(x_1, \dots, x_N)}{p(x_{m+1}, \dots, x_N)} . \quad (36)$$

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

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

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

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

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

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

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

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

$$\begin{aligned} \langle x_1 x_2 \rangle &= \begin{array}{c} \bullet \bullet \\ i \quad 2 \end{array} + \begin{array}{c} \bullet \bullet \\ \overline{i \quad 2} \end{array} \\ \langle x_1^2 x_2 \rangle &= \begin{array}{c} \bullet \bullet \\ \bullet \bullet \\ i \quad i \end{array} + \begin{array}{c} \bullet \bullet \\ \bullet \bullet \\ \overline{i \quad i} \end{array} + 2 \begin{array}{c} \bullet \bullet \\ \bullet \bullet \\ i \quad 2 \end{array} + \begin{array}{c} \bullet \bullet \\ \bullet \bullet \\ \overline{i \quad i} \end{array} \end{aligned}$$

we obtain

$$\begin{aligned}\langle x_1 x_2 \rangle &= \langle x_1 \rangle_c \langle x_2 \rangle_c + \langle x_1 * x_2 \rangle_c \quad , \quad \text{and} \\ \langle x_1^2 x_2 \rangle &= \langle x_1 \rangle_c^2 \langle x_2 \rangle_c + \langle x_1^2 \rangle_c \langle x_2 \rangle_c + 2 \langle x_1 * x_2 \rangle_c \langle x_1 \rangle_c + \langle x_1^2 * x_2 \rangle_c \quad .\end{aligned}\tag{40}$$

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

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

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

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

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

where the summation convention is used.

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

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

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

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

In the context of field theories, this result is referred to as *Wick's theorem*.

E Sums of Random Variables & the Central Limit Theorem

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

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

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

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

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

$$\ln \tilde{p}(k_1 = k_2 = \dots = k_N = k) = -ik \sum_{i=1}^N \langle x_{i1} \rangle_c + \frac{(-ik)^2}{2} \sum_{i_1, i_2}^N \langle x_{i_1} x_{i_2} \rangle_c + \dots, \quad (47)$$

as

$$\langle X \rangle_c = \sum_{i=1}^N \langle x_i \rangle_c \quad , \quad \langle X^2 \rangle_c = \sum_{i,j}^N \langle x_i x_j \rangle_c \quad , \quad \dots \quad (48)$$

If the random variables are independent, $p(\mathbf{x}) = \prod p_i(x_i)$, and $\tilde{p}_X(k) = \prod \tilde{p}_i(k)$. The cross-cumulants in eq.(48) vanish, and the n^{th} cumulant of X is simply the sum of the individual cumulants, $\langle X^n \rangle_c = \sum_{i=1}^N \langle x_i^n \rangle_c$. When all the N random variables are independently taken from the same distribution $p(x)$, this implies $\langle X^n \rangle_c = N \langle x^n \rangle_c$, generalizing the result obtained previously for the binomial distribution. For large values of N , the average value of the sum is proportional to N , while fluctuations around the mean, as measured by the standard deviation, grow only as \sqrt{N} . The random variable $y = (X - N \langle x \rangle_c) / \sqrt{N}$, has zero mean, and cumulants that scale as $\langle y^n \rangle_c \propto N^{1-n/2}$. As $N \rightarrow \infty$, only the second cumulant survives, and the PDF for y converges to the normal distribution,

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

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

The convergence of the PDF for the sum of many random variables to a normal distribution is a most important result in the context of statistical mechanics where such sums are frequently encountered. The *central limit theorem* states a more general form of this result: It is not necessary for the random variables to be independent, as the condition $\sum_{i_1, \dots, i_m}^N \langle x_{i_1} \cdots x_{i_m} \rangle_c \ll \mathcal{O}(N^{m/2})$, is sufficient for the validity of eq.(49).

Note that the above discussion implicitly assumes that the cumulants of the individual random variables (appearing in eq.(48)) are finite. What happens if this is not the case, i.e. when the variables are taken from a very wide PDF? The sum may still converge to a so-called *Levy distribution*. Consider a sum of N independent, identically distributed random variables, with the mean set to zero for convenience. The variance does not exist if the individual PDF falls off slowly at large values as $p_i(x) = p_1(x) \propto 1/|x|^{1+\alpha}$, with $0 < \alpha \leq 2$. ($\alpha > 0$ is required to make sure that the distribution is normalizable; while for $\alpha > 2$ the variance is finite.) The behavior of $p_1(x)$ at large x determines the behavior of $\tilde{p}_1(k)$ at small k , and simple power counting indicates that the expansion of $\tilde{p}_1(k)$ is singular, starting with $|k|^\alpha$. Based on this argument we conclude that

$$\ln \tilde{p}_X(k) = N \ln \tilde{p}_1(k) = N[-a|k|^\alpha + \text{higher order terms}]. \quad (50)$$

As before we can define a rescaled variable $y = X/N^{1/\alpha}$ to get rid of the N dependence of the leading term in the above equation, resulting in

$$\lim_{N \rightarrow \infty} \tilde{p}_y(k) = -a|k|^\alpha. \quad (51)$$

The higher order terms appearing in eq.(50) scale with negative powers of N and vanish as $N \rightarrow \infty$. The simplest example of a Levy distribution is obtained for $\alpha = 1$, and corresponds to $p_y = a/[\pi(y^2 + a^2)]$. (This is the Cauchy distribution discussed in problem set#3.) For other values of α the distribution does not have a simple closed form, but can be written as the asymptotic series

$$p_\alpha(y) = \frac{1}{\pi} \sum_{n=1}^{\infty} (-1)^{n+1} \sin\left(\frac{n\pi}{2}\alpha\right) \frac{\Gamma(1+n\alpha)}{n!} \frac{a^n}{y^{1+n\alpha}}. \quad (52)$$

Such distributions describe phenomena with large rare events, characterized here by a tail that falls off slowly as $p_\alpha(y \rightarrow \infty) \sim y^{-1-\alpha}$.

F Rules for Large Numbers

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

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

- (a) *Intensive* quantities, such as temperature T , and generalized forces, e.g. pressure P , and magnetic field \vec{B} , are independent of N , i.e. $\mathcal{O}(N^0)$.
- (b) *Extensive* quantities, such as energy E , entropy S , and generalized displacements, e.g. volume V , and magnetization \vec{M} , are proportional to N , i.e. $\mathcal{O}(N^1)$.
- (c) *Exponential* dependence, i.e. $\mathcal{O}(\exp(N\phi))$, is encountered in enumerating discrete micro-states, or computing available volumes in phase space.

Other asymptotic dependencies are certainly not ruled out a priori. For example, the Coulomb energy of N ions at fixed density scales as $Q^2/R \sim N^{5/3}$. Such dependencies are rarely encountered in every day physics. The Coulomb interaction of ions is quickly screened by counter-ions, resulting in an extensive overall energy. (This is not the case in astrophysical problems since the gravitational energy is not screened. For example the entropy of a black hole is proportional to the square of its mass.)

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

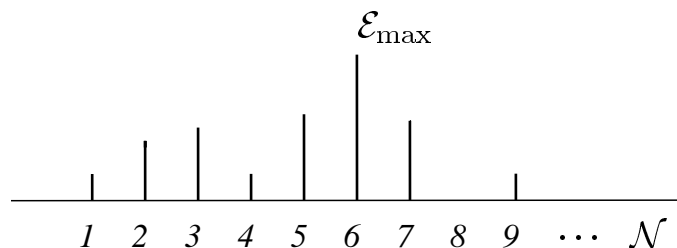
(1) *Summation of Exponential Quantities:* Consider the sum

$$\mathcal{S} = \sum_{i=1}^{\mathcal{N}} \mathcal{E}_i \quad , \quad (53)$$

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

$$0 \leq \mathcal{E}_i \sim \mathcal{O}(\exp(N\phi_i)), \quad (54)$$

and the number of terms \mathcal{N} , is proportional to some power of N .



7. A sum over \mathcal{N} exponentially large quantities is dominated by the largest term.

Such a sum can be approximated by its largest term \mathcal{E}_{\max} , in the following sense. Since for each term in the sum, $0 \leq \mathcal{E}_i \leq \mathcal{E}_{\max}$,

$$\mathcal{E}_{\max} \leq \mathcal{S} \leq \mathcal{N}\mathcal{E}_{\max} . \quad (55)$$

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

$$\frac{\ln \mathcal{E}_{\max}}{N} \leq \frac{\ln \mathcal{S}}{N} \leq \frac{\ln \mathcal{E}_{\max}}{N} + \frac{\ln \mathcal{N}}{N} . \quad (56)$$

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

$$\lim_{N \rightarrow \infty} \frac{\ln \mathcal{S}}{N} = \frac{\ln \mathcal{E}_{\max}}{N} = \phi_{\max} . \quad (57)$$

(2) *Saddle Point Integration:* Similarly, an integral of the form

$$\mathcal{I} = \int dx \exp(N\phi(x)) , \quad (58)$$

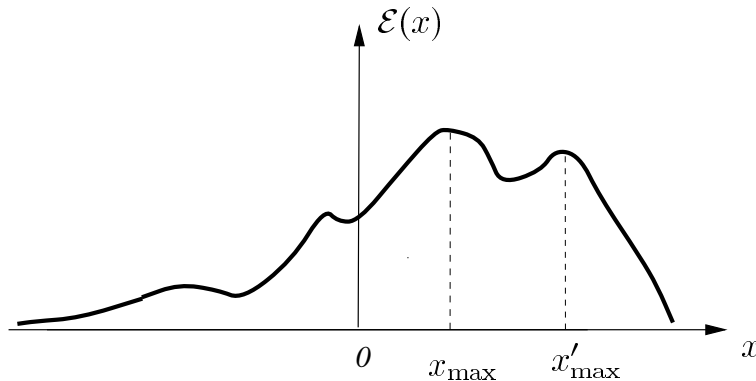
can be approximated by the maximum value of the integrand, obtained at a point x_{\max} which maximizes the exponent $\phi(x)$. Expanding the exponent around this point gives

$$\mathcal{I} = \int dx \exp \left\{ N \left[\phi(x_{\max}) - \frac{1}{2} |\phi''(x_{\max})| (x - x_{\max})^2 + \dots \right] \right\} . \quad (59)$$

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

$$\mathcal{I} \approx e^{N\phi(x_{\max})} \int dx \exp \left[-\frac{N}{2} |\phi''(x_{\max})| (x - x_{\max})^2 \right] \approx \sqrt{\frac{2\pi}{N|\phi''(x_{\max})|}} e^{N\phi(x_{\max})} , \quad (60)$$

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



8. Saddle point evaluation of an ‘exponential’ integral.

There are two types of corrections to the above result. Firstly, there are higher order terms in the expansion of $\phi(x)$ around x_{\max} . These corrections can be looked at perturbatively, and lead to a series in powers of $1/N$. Secondly, there may be additional local maxima for the function. A maximum at x'_{\max} , leads to a similar Gaussian integral that can be added to eq.(60). Clearly such contributions are smaller by $\mathcal{O}(\exp\{-N[\phi(x_{\max}) - \phi(x'_{\max})]\})$. Since all these corrections vanish in the thermodynamic limit,

$$\lim_{N \rightarrow \infty} \frac{\ln \mathcal{I}}{N} = \lim_{N \rightarrow \infty} \left[\phi(x_{\max}) - \frac{1}{2N} \ln \left(\frac{N|\phi''(x_{\max})|}{2\pi} \right) + \mathcal{O} \left(\frac{1}{N^2} \right) \right] = \phi(x_{\max}) \quad . \quad (61)$$

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

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

$$\int_0^{\infty} dx e^{-\alpha x} = \frac{1}{\alpha}. \quad (62)$$

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

$$\int_0^{\infty} dx x^N e^{-\alpha x} = \frac{N!}{\alpha^{N+1}}. \quad (63)$$

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

$$\Gamma(N+1) \equiv N! = \int_0^{\infty} dx x^N e^{-x}, \quad (64)$$

for all N . While the integral in eq.(64) is not exactly in the form of eq.(58), it can still be evaluated by a similar method. The integrand can be written as $\exp(N\phi(x))$, with $\phi(x) = \ln x - x/N$. The exponent has a maximum at $x_{\max} = N$, with $\phi(x_{\max}) = \ln N - 1$, and $\phi''(x_{\max}) = -1/N^2$. Expanding the integrand in eq.(64) around this point yields,

$$N! \approx \int dx \exp \left[N \ln N - N - \frac{1}{2N} (x - N)^2 \right] \approx N^N e^{-N} \sqrt{2\pi N}, \quad (65)$$

where the integral is evaluated by extending its limits to $[-\infty, \infty]$. Stirling's formula is obtained by taking the logarithm of eq.(65) as,

$$\ln N! = N \ln N - N + \frac{1}{2} \ln(2\pi N) + \mathcal{O} \left(\frac{1}{N} \right). \quad (66)$$

G Information, Entropy, and Estimation

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

$$g = \frac{N!}{\prod_{i=1}^M N_i!}. \quad (67)$$

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

$$\ln_2 g \approx -N \sum_{i=1}^M p_i \ln_2 p_i \quad (\text{for } N \rightarrow \infty), \quad (68)$$

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

$$1 = \left(\sum_i p_i \right)^N = \sum_{\{N_i\}} N! \prod_{i=1}^M \frac{p_i^{N_i}}{N_i!} \approx g \prod_{i=1}^M p_i^{Np_i}, \quad (69)$$

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

Shannon’s Theorem proves more rigorously that the minimum number of bits necessary to ensure that the percentage of errors in N trials vanishes in the $N \rightarrow \infty$ limit, is $\ln_2 g$. For any non-uniform distribution, this is less than the $N \ln_2 M$ bits needed in the absence

[†] More precisely, the probability of finding any N_i that is different from Np_i by more than $\mathcal{O}(\sqrt{N})$ becomes exponentially small in N , as $N \rightarrow \infty$.

of any information on relative probabilities. The difference per trial is thus attributed to the information content of the probability distribution, and is given by

$$I[\{p_i\}] = \ln_2 M + \sum_{i=1}^M p_i \ln_2 p_i \quad . \quad (70)$$

- *Entropy:* Eq.(67) is encountered frequently in statistical mechanics in the context of mixing M distinct components; its natural logarithm is related to the *entropy of mixing*. More generally, we can define an *entropy* for *any probability distribution* as

$$S = - \sum_{i=1}^M p(i) \ln p(i) = - \langle \ln p(i) \rangle \quad . \quad (71)$$

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

$$\Delta S(x_1, x_2 \rightarrow f) = \left[p_1 \ln \frac{p_1}{p_1 + p_2} + p_2 \ln \frac{p_2}{p_1 + p_2} \right] < 0. \quad (72)$$

- *Estimation:* The entropy S , can also be used to quantify subjective estimates of probabilities. In the absence of any information, the best *unbiased estimate* is that all M outcomes are equally likely. This is the distribution of maximum entropy. If additional information is available, the unbiased estimate is obtained by maximizing the entropy subject to the constraints imposed by this information. For example, if it is known that $\langle F(x) \rangle = f$, we can maximize

$$S(\alpha, \beta, \{p_i\}) = - \sum_i p(i) \ln p(i) - \alpha \left(\sum_i p(i) - 1 \right) - \beta \left(\sum_i p(i) F(x_i) - f \right), \quad (73)$$

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

moments (and hence n cumulants) of a distribution are specified, the unbiased estimate is the exponential of an n^{th} order polynomial.

In analogy with eq.(71), we can define an entropy for a continuous random variable ($\mathcal{S}_x = \{-\infty < x < \infty\}$) as

$$S = - \int dx p(x) \ln p(x) = - \langle \ln p(x) \rangle \quad . \quad (74)$$

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