4. Random variables

In the previous chapter, we have seen that the essence of statistical physics consist in replacing the integration of the dynamics to compute averages over time – which is what the experiments report – by a statistical average of the values taken by the observable in phase space, weighted with an appropriate distribution. The present chapter aims at recalling some basic tools and concepts that are useful when dealing with random variables. We shall also take this opportunity to present the Legendre transform, which has nothing to do with random processes, but which happens to be a major tool of statistical physics and thermodynamics.

4.1 Random variables and Generating Functions

Let's consider a random variable X with real value and its probability density $\rho(X)$. One defines the n-th order moment of X as

$$\mu_n := \langle X^n \rangle = \int_{-\infty}^{\infty} x^n \rho(x)\, dx. \tag{4.1}$$

Définition 4.1.1 — Moment-generating function. The moment-generating function of X is defined as

$$M_X(k) := \left\langle e^{kX} \right\rangle, \quad k \in \mathbb{R}, \tag{4.2}$$

wherever this expected value exists.

The reason for defining this function, also called the characteristic function, is that it can be used to find all the moments of the distribution. Indeed, writing the series expansion of e^{kX}, one gets

$$M_X(k) = \left\langle e^{kX} \right\rangle = 1 + k\langle X \rangle + \frac{k^2 \langle X^2 \rangle}{2!} + \frac{k^3 \langle X^3 \rangle}{3!} + \cdots + \frac{k^n \langle X^n \rangle}{n!} + \cdots \tag{4.3}$$

$$= 1 + k\mu_1 + \frac{k^2 \mu_2}{2!} + \frac{k^3 \mu_3}{3!} + \cdots + \frac{k^n \mu_n}{n!} + \cdots, \tag{4.4}$$

Hence

$$M_X(k) = \sum_{n=1}^{\infty} \mu_n \frac{k^n}{n!}, \quad \text{with} \quad \mu_n = \left. \frac{d^n M_X(k)}{dk^n} \right|_{k=0} \tag{4.5}$$

One also defines the centered momenta, as $\hat{\mu}_n := \langle (X - \mu_1)^n \rangle$, which characterize the distribution, after centering it on the average of X.

Définition 4.1.2 — Cumulant-generating function. The cumulants of a random variable X are defined using the cumulant-generating function, which is the logarithm of the moment-generating function : $K_X(k) = \log(M_X(k)) = \log \langle e^{kX} \rangle$.

The cumulants κ_n are again obtained from a power series expansion of the cumulant generating function :

$$K_X(k) = \sum_{n=1}^{\infty} \kappa_n \frac{k^n}{n!}, \quad \text{with} \quad \kappa_n = \left. \frac{d^n K_X(k)}{dk^n} \right|_{k=0} \tag{4.6}$$

The first cumulant is the expected value ; the second and third cumulants are respectively the second and third centered momenta (the second centered moment is the variance) ; but the higher cumulants are neither momenta nor centered momenta, but rather more complicated polynomial functions of the momenta. For instance $\hat{\mu}_4 = \kappa_4 + 3\kappa_2^2$.

Proposition 4.1.1

- If two distributions have the same moment or cumulant generating function, then they are identical at almost all points.
- The cumulant generating function, if it exists, is infinitely differentiable and convex, and passes through the origin.
- The cumulant-generating function exists if and only if the tails of the distribution are majored by an exponential decay.

For statistically independent random variables X and Y,

$$K_{X+Y}(k) = \log \left\langle e^{k(X+Y)} \right\rangle = \log \left(\left\langle e^{kX} \right\rangle \left\langle e^{kY} \right\rangle \right) = \log \left\langle e^{kX} \right\rangle + \log \left\langle e^{kY} \right\rangle = K_X(k) + K_Y(k), \tag{4.7}$$

so that each cumulant of a sum of independent random variables is the sum of the corresponding cumulants

4.2 Entropy of a distribution, joint and marginal distributions

Let X be a random variable with probability density $\rho(X)$

Définition 4.2.1 — The Kullback-Leibler divergence. The Kullback-Leibler divergence of the probability density $\rho(X)$ with respect to the probability density $h(x)$ is a functional defined as

$$D[\rho||h] = \int dx \, \rho(x) \log \left(\frac{\rho(x)}{h(x)} \right) \tag{4.8}$$

It measures how different is the probability density $\rho(x)$ from $h(x)$. In particular it is always positive and it is zero if and only if $\rho(x) = h(x)$.

Définition 4.2.2 — Entropy. The entropy of the probability density $\rho(X)$ is defined as

$$S[\rho] = -\int dx \rho(x) \log(\rho(x)) \tag{4.9}$$

It provides a characterization of the shape of the probability density. The larger the entropy the larger is the width of the distribution.

R Please mathematician friends forgive the fact that the above definition includes taking the log of a dimensional quantity. One way of "solving" this problem, would be to see $S(\rho) = -D(\rho||\mu)$, with μ, the uniform measure. This is however not correct because the uniform measure cannot be normalized on unbounded domains... Assuming our friends keep their eyes closed, one could write $\mu = \frac{1}{\infty}$ and obtain

$$S[\rho] = \int dx \rho(x) log(\infty) - D(\rho||\mu).$$

From this we see that S is indeed maximal when ρ is uniform, actually infinite if the domain is unbounded, and that S decreases when going away from the uniform measure, namely when the width of ρ decreases. Also S[ρ] can be both positive and negative. If one feels uncomfortable with such little arrangements, he should come back to the case of discrete sets of configurations, where everything is well defined (see the discussion about the limiting density of discrete points on wikipedia.)

Up to now, we have only introduced random variables with values in $\mathbb{R}$. In order to compute averages like the ones introduced in the previous part (equations 2.9), we'll need to extend the above formalism to random variables in $\mathbb{R}^N$, where N is the number of degrees of freedom of the system of interest.

The joint probability density $\rho(\vec{X}) = \rho(X_1, X_2, \cdots X_N)$ is a prodigiously rich information about the system, since it describes the probability of all possible micro-states. In many cases, we'll see that one is often interested in a reduced information, namely the probability density of certain degrees of freedom, letting the others be anything.

Définition 4.2.3 — Marginal distributions. For simplicity and for visualization ease (see figure 4.1), let's consider a random variable $\vec{X} = (X_1, X_2)$.

The probability that $X_1 \in [X_1; X_1 + dX_1]$ and $X_2 \in [X_2; X_2 + dX_2]$ is :

$$\mathbb{P}\left(X_1 \in [X_1; X_1 + dX_1], X_2 \in [X_2; X_2 + dX_2]\right) = \rho_2(X_1; X_2) dX_1 dX_2, \tag{4.10}$$

where $\rho_2(X_1; X_2)$ is the joint-distribution of (X_1, X_2).

The probability that $X_1 \in [X_1; X_1 + dX_1]$, $\forall X_2$ is :

$$\mathbb{P}\left(X_1 \in [X_1; X_1 + dX_1], \forall X_2\right) = \left(\int \rho_2(X_1; X_2) dX_2\right) dX_1 = \rho_1(X_1) dX_1, \tag{4.11}$$

where $\rho_1(X_1)$ is called the marginal distribution of X_1. By symmetry one defines $\rho_1(X_2)$, the marginal distribution of X_2. Note that marginal distributions are naturally normalized.

The marginal distribution should not be confused with the conditional one. The probability that $X_2 \in [X_2; X_2 + dX_2]$, knowing X_1 is :

$$\mathbb{P}\left(X_2 \in [X_2; X_2 + dX_2], \text{knowing} X_1\right) = \rho(X_2|X_1) dX_2. \tag{4.12}$$

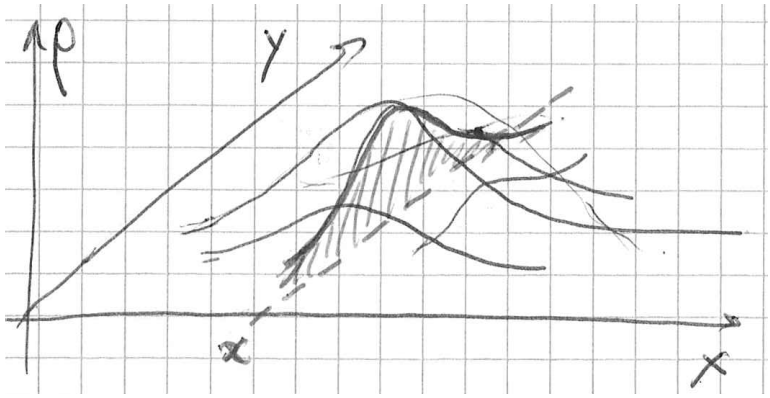

FIGURE 4.1 – Joint (the surface $\rho_2(x,y)$), marginal (integrated area $\rho_1(x)$) and conditional distribution (the weighted top curve $\rho(y|x)$)

where $\rho(X_2|X_1)$ is called the conditional probability of X_2, knowing X_1. By symmetry one defines $\rho(X_1|X_2)$ the conditional probability of X_1, knowing X_2. One clearly has :

$$\rho_2(X_1,X_2) = \rho(X_2|X_1)\rho_1(X_1) = \rho(X_1|X_2)\rho_1(X_2), \tag{4.13}$$

from which one easily checks that the conditional probabilities are also normalized.

R If the variables X_1 and X_2 are independent, then knowing anything about X_1 will not condition the distribution of X_2. Hence for two independent variable $\rho(X_2|X_1) = \rho_1(X_2)$, and $\rho_2(X_1,X_2) = \rho_1(X_1)\rho_1(X_2)$.

R If $\rho_2(X_1,X_2) = \rho_1(X_1)\rho_1(X_2)$ and $h_2(X_1,X_2) = h_1(X_1)h_1(X_2)$, then the Kullback-Leibler divergence is additive :

$$D(\rho_2(X_1,X_2)||h_2(X_1,X_2)) = D(\rho_1(X_2)||h_1(X_2)) + D(\rho_1(X_1)||h_1(X_1))$$

.

4.3 Gaussian variables

Let X be a random variable with a Gaussian distribution $G(\mu,\sigma)$:

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2}. \tag{4.14}$$

The prefactor $\frac{1}{\sqrt{2\pi\sigma^2}}$ ensures that the distribution is normalized : $\int_{-\infty}^{\infty} \rho(x)dx = 1$.

Two very useful relations are the so-called Hubbard-Stratonovich transformations :

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{2\sigma^2} \pm kx\right] dx = \exp\left(\frac{\sigma^2 k^2}{2}\right) \tag{4.15}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x \exp\left[-\frac{x^2}{2\sigma^2} \pm kx\right] dx = \pm\sigma^2 k \exp\left(\frac{\sigma^2 k^2}{2}\right) \tag{4.16}$$

R Note that the argument in the exponential on the right-hand side is always coming with a positive sign if x is real. The only way to obtain a negative sign is to apply the same transformation with x being a pure imaginary number. It is actually in this context that the transformation was introduced; and you may be surprised to learn that it was introduced to transform the simple-looking right-hand side into the less simple looking left-hand side. But this is another story....

The moment generating function exists and is equal to (direct consequence of the Hubbard Stratonovitch transformation)

$M_X(k) = \left\langle e^{kX} \right\rangle = e^{\mu k} e^{\frac{1}{2}\sigma^2 k^2}$

The cumulant generating function is thus simply the quadratic polynomial

$K_X(k) = \log M(k) = \mu k + \frac{1}{2}\sigma^2 k^2$

Thus only the first two cumulants, namely the mean $\langle X \rangle = \mu$, and the variance $\kappa_2 = \hat{\mu}_2 = \text{var}(X) = \left\langle (X - \langle X \rangle)^2 \right\rangle = \sigma^2$ are nonzero. This is actually an elegant way of computing the mean and the variance.

The entropy of the distribution $S(\sigma) = \frac{1}{2} + \log(\sqrt{2\pi\sigma^2})$ indeed increases with the distribution width as characterized by its variance.

What is remarkable is that the normal distribution is the only absolutely continuous distribution whose cumulants beyond the first two (i.e., other than the mean and variance) are zero. It is also the continuous distribution with the maximum entropy for a specified mean and variance.

4.4 Central Limit Theorem

The Gaussian distribution discussed above is key to statistical physics because it enters into the simplest version of the Central Limit Theorem (CLT). This theorem establishes that, in most situations, when independent random variables are added, their properly normalized sum tends toward a Gaussian distribution even if the original variables themselves are not normally distributed.

Let $(X_1, \ldots, X_N)$ be a set of independent and identically distributed random variables drawn from distributions of expected values given by μ and *finite* variances σ^2. Consider the normalized sum $S_N := \frac{1}{N}\sum X_i$ of these random variables, also called the empirical mean of X. Because the X_i are random variable, S_N also is.

For all N, the linearity of the average and the variance ensures that $\langle S_N \rangle = \mu$ and $Var(S_N) = \sigma^2/N$. In practice, one does not have access to $\langle S_N \rangle$ but only to a finite number of realizations and one would like to know how they are distributed. Intuitively, the larger N, the closer to μ should S_N be.

Proposition 4.4.1 — Law of Large Numbers. The law of large numbers indeed states that

$$\lim_{N\to\infty} S_N = \mu, \text{with probability one.} \tag{4.17}$$

Proposition 4.4.2 — Central Limit Theorem. The CLT precises the scaling in N in the following way : $Y_N = \frac{S_N - \mu}{\sigma/\sqrt{N}}$ converges in law to the Gaussian distribution G(0,1).

The usefulness of the theorem is that the distribution of S_N approaches Gaussianity regardless of the shape of the distribution of the individual X_i, provided that its variance is finite.

The central limit theorem has a number of variants. In its common form, the random variables must be independent and identically distributed. In variants, convergence of the mean to the normal

distribution also occurs for non-identical distributions or for non-independent observations, given that they comply with certain reasonable conditions, which guarantee that the contribution of any individual random variable to the variance is arbitrarily small, for sufficiently large N, and that the correlations are not long ranged.

4.5 Legendre-Fenchel transfrom

Définition 4.5.1 — Legendre-Fenchel transform. We define the Legendre-Fenchel transfrom of a function $\lambda(k)$ as

$$I(s) = \sup_k [sk - \lambda(k)]. \tag{4.18}$$

One particular and easy case to consider, although not the general case as we shall see below, is when $\lambda(k)$ is a strictly convex function, $\frac{d^2\lambda}{dk^2} > 0$. It is therefore differentiable and, deriving the element to be maximized, one finds $s = \frac{d\lambda}{dk}\big|_k$. Since $\frac{d\lambda}{dk}\big|_k$ is a monotonic and growing function of k, it is invertible and k(s) is the locus of the unique maximum. The Legendre-Fenchel transform of $\lambda(k)$ is then the less general Legendre transform :

$$I(s) = sk(s) - \lambda(k(s))) \quad \text{with} \quad s(k) = \frac{d\lambda}{dk}, \quad \text{and } k(s) \text{ its inverse.} \tag{4.19}$$

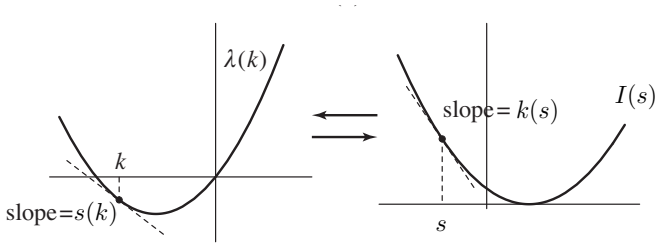

FIGURE 4.2 – Illustration of the Legendre transform.

R If $\lambda(k)$ is strictly concave, the same construction holds, provided that one defines $I(s) = \inf_x [sk - \lambda(k)]$.

Proposition 4.5.1 — Properties of the Legendre transform.
In the specific case, when $\lambda(k)$ is strictly convex, $\lambda(k)$ is the Legendre transform of $I(s)$.

Indeed :

$$\frac{dI}{ds} = k(s) + s\frac{dk}{ds} - \frac{d\lambda}{dk}\frac{dk}{ds} = k(s), \tag{4.20}$$

hence

$$\lambda(k) = ks(k) - I(s(k)), \quad \text{with} \quad k(s) = \frac{dI}{ds}, \quad \text{and } s(k) \text{ its inverse} \tag{4.21}$$

There is thus a systematic correspondance : $I'(s) = k <=> \lambda'(k) = s$, as illustrated on figure 4.2

It is also easy to check that :

$$\frac{d^2\lambda}{dk^2}\frac{d^2I}{ds^2} = 1, \qquad I(0) = -\lambda_{\min}, \qquad \lambda(0) = -I_{\min}. \tag{4.22}$$

R These propositions do not hold in the general case of the Legendre-Fenchel transform. Consider the case where $\lambda(k)$ is not differentiable in one point (see figure 4.3). There is no more a one to one correspondance between $I(s)$ and $\lambda(k)$: all function with the same convex envelope $I^{**}(s)$ share the same $\lambda(k)$ with a singularity for k corresponding to the slope of the straight part in the convexe envelope.

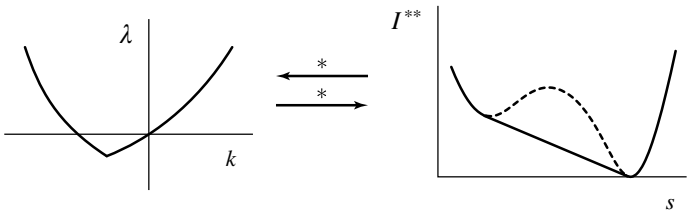

FIGURE 4.3 – Illustration of the Legendre-Fenchel transform in the case of a non differentiable $\lambda(k)$

We shall see now why the Legendre-Fenchel transform, and its specialized version, the Legendre transform are so present in statistical physics and thermodynamics.

4.6 Large deviations, Gärtner-Ellis and Cramér's theorems

In the context of statistical physics, we shall be interested in the statistics of some physical quantities, averaged over the N components of the system. Let S_N be such a random variable indexed by the integer N. One is interested in the statistical properties of this variable, namely its probability distribution $P(S_N = s)$ when N is very large.

Définition 4.6.1 — Large deviation principle. One says that $P(S_N = s)$ satisfies a large deviation principle with rate $I(s)$ if the following limit exists :

$$I(s) = -\lim_{N\to\infty}\frac{1}{N}\ln P(S_N = s), \quad \text{in other words} \quad P(S_N = s) \approx \exp[-N\,I(s)]. \tag{4.23}$$

The function $I(s)$ is called the rate function or the large deviation function. Be aware that the same wording sometime refers to $I(s)$, sometime to $-I(s)$.

Figure 4.4 provides an illustration of such a rate function, and the associated $P(S_N = s)$ for different N. One sees how the distributions becomes sharper and sharper around its most probable value, which corresponds to the minimum and zero of the rate function. Essentially, if the variable S_N satisfies a large deviation principle, the fluctuations outside the most probable value decreases exponentially fast to zero with N and the rate function tells us how exponentially fast. This is of course of primary importance to guarantee that most probable value are good estimates of the macroscopic variable of interest.

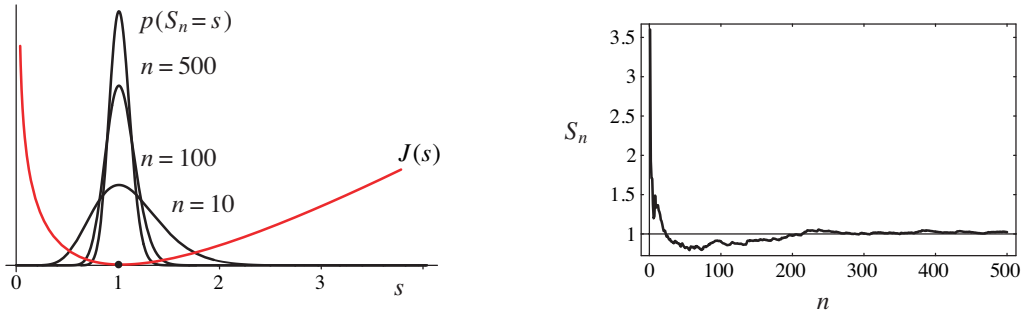

FIGURE 4.4 – Illustration of a Large deviation function and the associated variation of $P(S_N = s)$, with N and typical realization of S_N converging to its most probable value.

The theory of large deviations aims at establishing when a large deviation principle exists for a given random variable S_N, and at deriving the expression of the associated rate function.

In some rare case, one can compute explicitly $P(S_N = s)$ and derive the rate function from its asymptotic development at large N. In general, however, $P(S_N = s)$ cannot be computed. This is where the large deviation theory becomes useful. A key theorem of the theory is the Gärtner-Ellis theorem :

Proposition 4.6.1 — Gärtner-Ellis Theorem. One defines the *scaled* cumulant generating function of S_N by the limit

$$\lambda(k) = \lim_{N\to\infty} \frac{1}{N} K_{NS_N}(k) = \lim_{N\to\infty} \frac{1}{N} \log \left\langle e^{kNS_N} \right\rangle . \tag{4.24}$$

If $\lambda(k)$ exists and is differentiable for all k, then S_N satisfies a large deviation principle, with a rate function I(s) given by the Legendre–Fenchel transformation :

$$I(s) = \sup_k [ks - \lambda(k)] \tag{4.25}$$

It is instructive to consider the following heuristic argument for the expression of the rate function. Assuming the existence of the large deviation principle, one has in the limit of large N

$$P(S_N = s) \approx \exp[-NI(s)], \tag{4.26}$$

and

$$\left\langle e^{kNS_N} \right\rangle \approx \int e^{N[ks-I(s)]} ds \approx e^{N \sup_s [ks-I(s)]}, \tag{4.27}$$

where the last equality results from the celebrated *saddle-point, or Laplace's approximation* (see below). One thus finds that $\lambda(k)$ is the Legendre transform of $I(s)$ and therefore, if $I(s)$ is strictly convex, that $I(s)$ is the Legendre transform of $\lambda(k)$

$$\lambda(k) = \sup_s [ks - I(s)], \quad \text{and} \quad I(s) = \sup_k [ks - \lambda(k)]. \tag{4.28}$$

We thus see that the Gärtner-Ellis Theorem is essentially a consequence of the large deviation principle combined with the Laplace's approximation and that the Legendre-Fenchel transform appears into this theory as a natural consequence of Laplace's approximation.

R Here we have used the Laplace's approximation in its crudest form. At the next order one has, in the limit of large N, for a function g with its maximum in s^* :

$$\int e^{Ng(s)}ds \approx e^{Ng(s^*)}\int e^{-\frac{N}{2}g''(s^*)(s-s^*)^2}ds = e^{Ng(s^*)}\left(\frac{2\pi}{N|g''(s^*)|}\right)^{1/2} \tag{4.29}$$

Nota Bene : The rate functions obtained from the Gärtner-Ellis Theorem are necessarily strictly convex. Unfortunately rate functions are not all strictly convex ; some may even have local minima in addition to global ones... Those rate function can thus not be found using the Gärtner-Ellis theorem...

The Gärtner-Ellis theorem best applies to an extension of the Central Limit Theorem, called the Cramér's theorem.

Proposition 4.6.2 — Cramérs Theorem. Like for the central limit theorem, let $(X_1,\ldots,X_N)$ be a set of (i.i.d) random variables drawn from distributions of expected values given by μ and *finite* variances σ^2 and consider the normalized sum $S_N := \frac{1}{N}\sum X_i$ of these random variables. The theorem states that S_N satisfies a large deviation principle and that the rate function

$$I(s) = \sup_k [ks - K_X(k)], \tag{4.30}$$

where $K_X(k)$ is simply the cumulant generating function of the X_i's.

The proof is the simple result of the fact that for (i.i.d), $K_{NS_N}(k) = NK_X(k)$, and therefore that $\lambda(k) = K_X(k)$.

The rate function $I(s)$ inherits some interesting properties from those of the cumulant generating function $\lambda(k) = K_X(k)$. Remember that the cumulant generating function of a random variable is always convex (although not strictly convex).

- $\lambda(0) = 0 \Rightarrow I(s) \geq 0$; proof : $\lambda(0) = \sup_s[0k - I(s)] = -\inf[I(s)]$.
- $\lambda'(0) = \mu \Rightarrow I'(\mu) = 0 : \mu$ is the minimum of the rate function
- $\lambda''(0) = \sigma^2 \Rightarrow I''(\mu) = 1/\sigma^2$.

The Central Limit Theorem states that, for any well behaved distribution for the X_i's, $I(x)$ has a parabolic shape of width $\frac{\sigma}{\sqrt{N}}$ around the maximum $x = \mu$. It is clear that further away from the maximum, the parabolic shape is not guaranteed unless strictly speaking when $N \to \infty$. The above theorem provides a way to compute the deviations, far away from the maximum, hence the name "Large deviations" of the theory

■ **Exemple 4.1 — The case of Gaussian variables.** Let the X_i's obey a Gaussian distribution of mean μ and variance σ^2. We have seen that $K_X(k) = \log\langle e^{kX}\rangle = \mu k + \frac{1}{2}\sigma^2 k^2$. Hence

$$I(s) = sup_k[ks - K_X(k)] = k^*s - K_X(k^*), \quad \text{with} \quad \frac{dK_X}{dk}\Big|_{k^*} = s, \quad \text{or} \quad k^* = \frac{s-\mu}{\sigma^2} \tag{4.31}$$

$$I(s) = \frac{1}{2}\frac{(s-\mu)^2}{\sigma^2}. \tag{4.32}$$

In the case of Gaussian variables, the large deviation function is a parabola itself. ■

■ **Exemple 4.2 — The case of Poisson variables.** Let the X_i's be integer variables obeying a Poisson distribution $\rho(x) = \frac{\lambda^x e^{-\lambda}}{x!}$. One can show that $\langle X\rangle = \lambda$ and the variance $var(X) = \lambda$. The

cumulant generating function $K_X(k) = \log\left\langle e^{kX}\right\rangle = \lambda\left(e^k - 1\right)$. Hence

$$I(s) = sup_k[ks - K_X(k)] = k^* s - K_X(k^*), \quad \text{with} \quad \frac{dK_X}{dk}\Big|_{k^*} = s, \quad \text{or} \quad k^* = \log\left(\frac{s}{\lambda}\right) \tag{4.33}$$

$$I(s) = s\log\left(\frac{s}{\lambda}\right) - s + \lambda. \tag{4.34}$$

We here see that the large deviation function is certainly not a parabola. It is however easy to check that its maximum takes place at $s = \lambda$ and that the parabolic shape around that maximum has width $\sqrt{\lambda}$, as prescribed by the CLT (see figure 4.5). ∎

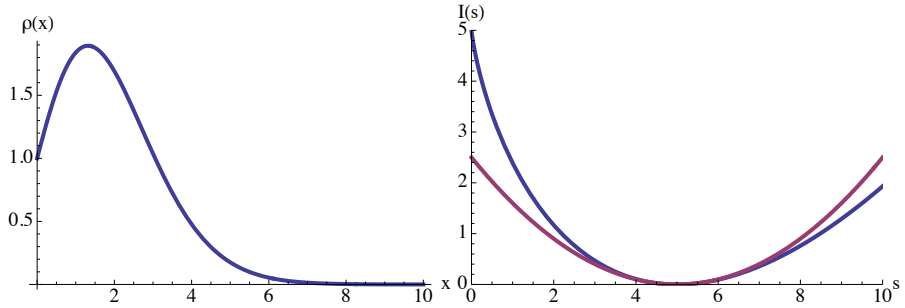

FIGURE 4.5 – Poisson distribution on the left and large deviation function for a sum of Poisson variables together with the parabola provided by the CLT (dashed) on the right ; (here $\lambda = 5$).