Journal of Scientific Exploration, Vol. 13, No. 4, pp. 639–664, 1999 0892-3310/99

Introductory Remarks on Large Deviation Statistics

ANTON AMANN

Universitätsklinik für Anästhesie und Allgemeine Intensivmedizin
Leopold-Franzens-Universität Innsbruck
Anichstr. 35, A-6020 Innsbruck, Austria

HARALD ATMANSPACHER

Institut für Grenzgebiete der Psychologie
Wilhelmstr. 3a, D-79098 Freiburg, Germany
and
Max-Planck-Institut für Extraterrestrische Physik
Giessenbachstrasse, D-85740 Garching, Germany

Abstract — The physical concept of entropy as it is used in thermodynamics is related to the mathematical formulation of a Shannon entropy. Usually only the Shannon entropy of equilibrium distributions such as a canonical distribution is considered. Large deviations statistics goes beyond that framework. Entropies are considered for arbitrary distributions or physical states, and they describe, *e.g.*, "how fast" non-equilibrium distributions and states "die out" with increasing number of degrees of freedom or increasing number of particles. Hence the concept of an entropy acquires a new meaning, referring to the statistical fluctuations in collectives of empirical events. In the particular case of experiments with independent and identically distributed (i.i.d.) events, Shannon entropy can be shown to play its usual role (Sanov's theorem). Jaynes' maximum entropy principle, important in statistical physics, is a consequence of Sanov's theorem and thereby obtains a precise interpretation. In the general case of non-i.i.d. events, all sorts of (even non-convex) entropies can arise. As illustrative examples, large deviation statistics of phase transitions and multifractals are addressed.

Keywords: large deviations — Shannon entropy — Jaynes' maximum entropy principle — multifractals

1. What Is a Large Deviation?

1.1. Heuristic Introduction

The term "large deviation" basically refers to deviations of an empirical average over some distribution from the corresponding theoretical expectation. Assume, *e.g.*, that seven throws of a dice lead to the result (4, 6, 6, 2, 1, 6, 5). Then the empirical average is given by

$$\frac{1}{7}(4 + 6 + 6 + 2 + 1 + 6 + 5) = 4.2857 \qquad (1)$$

and the empirical distribution is given by

$$\left(\frac{1}{7}, \frac{1}{7}, 0, \frac{1}{7}, \frac{1}{7}, \frac{3}{7}\right), \tag{2}$$

since 1 arises with empirical probability 1/7, … , and 6 arises with empirical probability 3/7. The mean theoretical distribution which we expect to arise for a large or infinite number of throws is

$$\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right), \tag{3}$$

with a corresponding mean value of

$$\frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5. \tag{4}$$

Deviations from the mean theoretical distribution and its mean value abound for a small number of throws (in the case of a dice) or for a small number of particles (in statistical physics). When the number of throws increases, large deviations typically die out soon, *e.g.*, if the throws are strictly random. How "fast" do deviations from a mean value disappear with increasing number of random throws? A typical large deviation result is the following. Let ω be an arbitrary series of N random throws and let $X_j(\omega)$ be the result of throw number j. Then the corresponding mean value is given as

$$\mu = \frac{1}{N} \sum_{j=1}^{N} X_j(\omega) \tag{5}$$

for the individual time series ω of random throws.

Call $Q^N[a, b]$ the probability (computed by consideration of all possible individual time series) to find the mean (5) in the interval $[a, b]$. By the weak law of large numbers, one knows that

$$\lim_{N\to\infty} Q^N[a, b] = 1, \tag{6}$$

if the most probable mean value, say μ_0, is contained in the interval $[a, b]$, *i.e.*, $a < \mu_0 < b$. Alternatively,

$$\lim_{N\to\infty} Q^N[a, b] = 0, \tag{7}$$

if the most probable mean value μ_0 is *not* contained in the interval $[a, b]$. If $\mu \neq \mu_0$, the time series ω shows large deviations, *i.e.*,

$$\left| \frac{1}{N} \sum_{j=1}^{N} X_j(\omega) - \mu_0 \right| > 0. \tag{8}$$

Under appropriate conditions (*e.g.*, if the throws are strictly random), such large deviations "die out" for a large number N of throws.

In large deviations statistics, this heuristic reasoning is made precise. In particular, it is shown that

$$Q^N([a, b]) \sim \exp(-NI_{[a,b]}), \tag{9}$$

where the scalar $I_{[a,b]}$ fulfills

$$I_{[a,b]} = 0 \tag{10}$$

if μ_0 is in the interval $[a, b]$, and where

$$I_{[a,b]} > 0 \tag{11}$$

if μ_0 is *not* in the interval $[a, b]$. In the latter case, $Q^N([a, b])$ goes to zero *exponentially* with increasing number N of throws. The constants $I_{[a,b]}$, which are different for different intervals $[a, b]$, can be summarized in one function $I : \mathbf{R} \to \mathbf{R}$:

$$I_{[a,b]} = \inf_{x \in [a,b]} I(x). \tag{12}$$

This function I (which must be shown to exist) is a *large deviation entropy*. It can be used to characterize the behavior of a system in space and time in terms of fluctuations around a mean value. Depending on whether the variables are independent or non-independent, random or partially non-random, *etc.*, one gets different characterizing entropies. The entropy changes, for example, if the behavior of a system changes from random behavior to (partial) correlations between different events.

1.2. Fluctuations Around Mean Values

In this section the preceding introductory remarks will be backed up more formally. Another presentation emphasizing the physical background can be found in an article by Lanford (1973).

Let us start with a "classical" example and consider an ensemble of N particles of identical structure with M possible energy levels with energies $\epsilon_1, \epsilon_2, \ldots, \epsilon_M$. In this example the state space Ω_N of N particles is taken as the Cartesian product of the N state spaces of the particles, *i.e.*, $\Omega_N = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_M\}^N$. States of the ensemble as a whole in Ω_N will be denoted by ω. Then $X_j(\omega) \in \{\epsilon_1, \epsilon_2, \ldots, \epsilon_M\}$, $j = 1, 2, \ldots, N$, is the actual energy value of particle j in the ensemble state ω. While the number M of possible energy values is fixed, we keep the number N of particles variable and study how the *fluctuations* of the specific energy

$$u(\omega) := \frac{\sum_{j=1}^{N} X_j(\omega)}{N} \tag{13}$$

in an ensemble change with increasing number of particles. Typically it is presupposed that, *e.g.*, the energy shell and/or certain expectation values are fixed for the total system. Moreover, the energies $X_j(\omega)$ and $X_k(\omega)$ of different particles are assumed to be independent and identically distributed with probability $\mu = \{\mu_1, \mu_2, \ldots, \mu_M\}$. The particular case

$$\mu_i = \frac{1}{M}, \qquad i = 1, 2, \ldots, M, \tag{14}$$

of an equidistribution of energies is typically considered in statistical mechanics, when one derives the (Boltzmann) population distribution over energy levels.

The product probability measures are denoted by μ^N. How is the specific energy (13) distributed (for some fixed N), and how does this distribution change with increasing number N of particles? (Note that these questions correspond to questions as to how pure states in quantum mechanics are distributed and how their distribution changes with increasing number of degrees of freedom.) Heuristically we expect that

$$\frac{\sum_{j=1}^{N} X_j(\omega)}{N} \approx \bar{\epsilon} = \sum_{k=1}^{M} \mu_k \epsilon_k \tag{15}$$

and that the probability to find *large deviations*, *i.e.*, states ω of the N-particle ensemble such that

$$\left| \sum_{j=1}^{N} X_j(\omega) - N\bar{\epsilon} \right| > N\eta \tag{16}$$

goes to zero with increasing number N of particles. Here η is a fixed positive scalar. Large deviations theory claims that the respective probability to find large deviations behaves as

$$\mu^N \left\{ \omega : \left| \sum_{j=1}^{N} X_j(\omega) - N\bar{\epsilon} \right| > N\eta \right\} \sim e^{-NI_\eta}. \tag{17}$$

Here, the positive scalar I_η is a large deviation entropy. It is related but in general not identical to the usual thermodynamic entropy.

The classical problem outlined above has been solved by Cramér (1937). The solution is reviewed here in a simplified way. Define the "Massieu potential"

$$\phi(y) = \lim_{N\to\infty} \frac{1}{N} \ln \left(\int_{\Omega_N} \exp \left\{ y \sum_{j=1}^{N} X_j(\omega) \right\} d\mu^N \right) \tag{18}$$

$$= \ln \int_{\Omega_1} \exp \{ y X_1 \} d\mu, \qquad y \in \mathbf{R}, \tag{19}$$

using the independence and identical distribution of the energies X_j. Denote now the Legendre transform of the Massieu potential $\phi = \phi(y)$ by I,

$$I(e) \stackrel{\text{def.}}{=} \sup_{y \in \mathbf{R}} \{ye - \phi(y)\}, \tag{20}$$

and call it an entropy. Let furthermore

$$Q^N([d_1, d_2]) := \mu^N \left\{ \omega \in \Omega_N : \frac{1}{N} \sum_{i=1}^{N} X_j(\omega) \in [d_1, d_2] \right\} \tag{21}$$

denote the distribution of the mean $1/N \sum_{j=1}^{N} X_j$. Then Cramér's result is that

$$Q^N([e_1, e_2]) \sim \exp \left\{ -N \inf_{e \in [e_1, e_2]} I(e) \right\} \tag{22}$$

If the variables X_j are energies of independent particles, the Massieu potential $\phi(y)$ corresponds to the usual thermodynamic Massieu potential

$$\phi_{\text{therm}}(\beta) = -k\beta f = \frac{k}{N} \ln Z_N = k \ln \sum_{i=1}^{M} e^{-\beta \epsilon_i}, \tag{23}$$

where f is the specific free energy and k is Boltzmann's constant. The only differences between the thermodynamic Massieu potential $\phi_{\text{therm}}(\beta)$ and the Massieu potential $\phi(y)$ used in large deviation statistics are their sign and normalization conventions.

In other situations, the mathematical Massieu potential $\phi(y)$ may be a Gibbs free energy (modulo β), or an $f(\alpha)$-spectrum in the multifractal formalism (Bohr and Tél, 1988; Oono, 1989), or the spectrum of Lyapunov exponents of a dynamical system (Peinke *et al.*, 1992). It also arises in the contexts of dynamical systems (Deuschel and Stroock, 1989) and the quantum mechanical measurement problem (Amann, 1994).

Cramér (1937) derived an important result for the probabilities $Q^N([e_1, e_2])$ and $Q^N(]e_1, e_2[)$ to find the specific energy (13) in some arbitrarily chosen closed or open interval $[e_1, e_2]$:

$$\begin{aligned} \limsup_{N \to \infty} \frac{1}{N} \ln Q^N([e_1, e_2]) &\leq -\left(\inf_{e \in [e_1, e_2]} I(e) \right), \\ \liminf_{N \to \infty} \frac{1}{N} \ln Q^N(]e_1, e_2[) &\geq -\left(\inf_{e \in]e_1, e_2[} I(e) \right). \end{aligned} \tag{24}$$

If the entropy I is continuous at e_1 and e_2, one may simply write

$$\lim_{N \to \infty} \frac{1}{N} \ln Q^N([e_1, e_2]) = -\left(\inf_{e \in [e_1, e_2]} I(e) \right), \tag{25}$$

and similarly for the open interval. These are mathematically more precise formulations of

$$Q^N[e_1, e_2] \quad \sim \quad \exp \left\{ -N \inf_{e \in [e_1, e_2]} I(e) \right\} \tag{26}$$

As mentioned above, the theoretical "reference" distribution μ for N different particles is typically chosen to be the equidistribution $(1/M, 1/M, \ldots, 1/M)$. Since any distribution is acceptable from a mathematical point of view, Cramér's result can be applied to arbitrary "reference" distributions on the real numbers $\mathbf{R}$ or on the finite set of energies $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_M\}$. In particular:

- If the random variables X_j are normally distributed with variance σ^2, $X_1 \sim N(0, \sigma^2)$, then the entropy is given by $I(x) = x^2/(2\sigma^2)$.
- If the random variables X_j are exponentially distributed, $X_1 \sim \exp(\theta)$, then the entropy is given by $I(x) = \theta x - 1 - \ln(\theta x)$ for $x > 0$ and $I(x) = \infty$ otherwise.

Since the entropy function is the Legendre transform of a convex Massieu potential, it is itself convex. This convexity of the entropy function is nice from a mathematical point of view but unpleasant for certain physical applications. There is no problem when the random variables X_j are independent and identically distributed, a situation which gives rise to a convex entropy function. As soon as phase transitions arise, however, the i.i.d. assumption is no longer satisfied (*cf.* Section 4) and it is desirable to have non-convex entropy functions (*cf.* also Figure 2 in Section 2.1).

In thermodynamics, the Massieu potential depends on the inverse temperature $\beta = 1/kT$ and is defined as $\Phi(\beta) := -k\beta F$, where F is the free energy (not the *specific* free energy f as in Equation (23)). The conjugate variable with respect to $k\beta$ is

$$-\frac{\partial\Phi(\beta)}{\partial(k\beta)} = \frac{\partial(k\beta F)}{\partial(k\beta)} = F + k\beta\frac{\partial F}{\partial(k\beta)} = F + k\beta\left(\frac{\partial F}{\partial T}\right)_V\left(\frac{\partial T}{\partial(k\beta)}\right) \tag{27}$$

$$= F + k\beta(-S)\left(-\left(\frac{1}{k\beta}\right)^2\right) = F + \frac{S}{k\beta} = (U - TS) + TS = U \tag{28}$$

i.e., the inner energy U. Hence a Legendre transform of the Massieu potential provides the entropy S,

$$L(\Phi) = k\beta(-U) - \Phi = -k\beta U + k\beta F = k\beta(F - U) = -S. \tag{29}$$

Consequently, the Legendre transform $I = I(u)$, depending on the *specific* energy u, as defined in Equation (20) corresponds to the negative *specific* thermodynamic entropy s/k up to appropriate (additive) normalization. The entropy I is (typically) convex whereas the entropy S is (typically) concave.

1.3. Formal Framework

The general scenario of large deviations can be compactly expressed by the following formal points.

- *Definition:* A function $I : \Xi \to [0, \infty]$ on a complete separable metric space Ξ is an *entropy function* if I is lower semicontinuous and has com-

pact level sets. Probability distributions $\{Q^N\}$, $N = 1, 2, \ldots$, on a complete separable metric space Ξ are defined to have a large deviation property with respect to an entropy function I from Ξ into $[0, \infty]$, if there exists an increasing sequence $\{a_1, a_2, \ldots\}$ of positive numbers which tend to infinity such that

$$\limsup_{N\to\infty} \frac{1}{a_N} \ln Q^N(K) \leq -\left(\inf_{x\in K} I(x)\right),$$
$$\liminf_{N\to\infty} \frac{1}{a_N} \ln Q^N(G) \geq -\left(\inf_{x\in G} I(x)\right), \tag{30}$$

where K is an arbitrary closed and G an arbitrary open set.

- *Comment:* The large deviation approach is very general, since Ξ is an arbitrary complete separable metric space. Often the space Ξ is simply $\mathbf{R}^d$ or even just the real numbers $\mathbf{R}$.
- *Proposition:* For independent, identically distributed (i.i.d) variables $X_i : \Omega \to \mathbf{R}$, the entropy I describing the distributions Q^N defined in Equation (21) is given as the Legendre transform of the Massieu potential ϕ defined in Equation (18).
- The transition from the mathematical formulation to physics is often provided by *Varadhan's lemma* (Varadhan, 1966): Assume that the probability distributions $\{Q^N\}$, $N = 1, 2, \ldots$, on the complete separable metric space Ξ have a large deviation property with respect to the entropy function I and the constants $\{a_n\}$. Assume furthermore that F is a real-valued function on Ξ. Then the following result holds:
  (a) Assume that $\sup_{x\in\Xi} F(x)$ is finite. Then $\sup_{x\in\Xi}\{F(x) - I(x)\}$ is finite and

$$\lim_{N\to\infty} \frac{1}{a_N} \ln \int_\Xi e^{a_N F(x)} Q^N(dx) = \sup_{x\in\Xi}\{F(x) - I(x)\}. \tag{31}$$

  (b) More generally, assume that F satisfies

$$\lim_{L\to\infty} \limsup_{N\to\infty} \frac{1}{a_N} \ln \int_{\{F\geq L\}} e^{a_N F(x)}\, Q^N(dx) = -\infty. \tag{32}$$

  Then the limit (31) holds and is finite. In particular, if F is bounded above on the union of the supports of the $\{Q^N\}$, then (32) is satisfied and thus the limit (31) holds and is finite.
- *Consequence of Varadhan's lemma:* The sequence of probability measures

$$Q^{N,F}\{A\} := \frac{\int_A \exp(a_N F(x)) Q^N(dx)}{\int_\Xi \exp(a_N F(x)) Q^N(dx)}, \qquad A \subseteq \Xi, n = 1, 2, \ldots, \tag{33}$$

has a large deviation property with respect to the entropy function

$$I_F(x) := (I(x) - F(x)) - \inf_{x \in \Xi}\{I(x) - F(x)\}. \tag{34}$$

Here A is an arbitrary measurable subset of Ξ, *e.g.,* an interval $[a, b]$ if $\Xi = \mathbf{R}$.

- *Definition:* Consider an $\mathbf{R}^d$-valued random variable Y on the (complete separable metric) space Ξ. Then the *distribution* of Y with respect to a probability measure Q is defined as the probability measure Q_Y:

$$Q_Y(B) := Q\{Y^{-1}(B)\}, \tag{35}$$

where B is an arbitrary measurable subset of Ξ.

- *Theorem:* Let w be a real-valued random-variable on $\mathbf{R}^d$. Then

$$\int_{\Xi} w(Y(\omega))\, Q(d\omega) = \int_{\mathbf{R}^d} w(y)\, Q_Y(dy). \tag{36}$$

This result allows us to translate many problems of statistical physics into the formalism of large deviation theory and to apply Varadhan's theorem. Note that in the large deviation setting and in Varadhan's theorem *one* particular space is used whereas in statistical physics the underlying spaces Ω_N change with the number of particles. The variable y formally corresponds to (minus) the inverse temperature.

1.4. Applying Varadhan's Lemma

Let Q^N, $N = 1, 2, \ldots,$ be the distributions of some real variable x, *e.g.*, the mean energy of N-particle models, and assume that the Q^N fulfill a large deviation principle with respect to the entropy I. Setting $F(x) = yx$, Varadhan's lemma implies that the respective Massieu potential can be computed as

$$\phi(y) = \lim_{N \to \infty} \frac{1}{a_N} \ln \int_{\mathbf{R}} e^{y\, a_N x}\, Q^N(dx) = \sup_{x \in \mathbf{R}}\{xy - I(x)\}. \tag{37}$$

Hence the Massieu potential ϕ is the Legendre transform of the entropy I.

Assume, for example, that Y_N is some "mean variable" of an N-particle system. In an N-spin system, Y_N could be the mean magnetization

$$Y_N(\omega) = \frac{1}{N} \sum_{i=1}^{N} S_i(\omega) \tag{38}$$

of the spin variables $S_i(\omega)$ (with, *e.g.*, values in $\{-1, +1\}$ with probabilities $\mu(-1) = \mu(+1) = 1/2$). We take the product measure μ^N as the distribution of the N spins and consider the distributions Q^N of the mean magnetization, defined as in Equation (35),

$$Q^N(B) := \mu^N\{Y_N^{-1}(B)\}, \tag{39}$$

for arbitrary measurable subsets B of the real numbers, *e.g.*, intervals

$B = [a, b]$. The distributions Q^N are distributions on the real numbers $\mathbf{R}$. Instead of speaking of spins with values in $\{+1, -1\}$, we could also call this a dice with two different possible values $\{+1, -1\}$. Energy considerations do not play a role for the definition of the distributions Q^N up to this point.

Assume now that there is a physical energy and that the mean value $e = e(\omega)$ of the N-spin system is a function h of the mean magnetization, *i.e.*,

$$e(\omega) = h(Y_N(\omega)) = h\left(\frac{1}{N}\sum_{i=1}^{N} S_i(\omega)\right), \tag{40}$$

the total energy being $E(\omega) = Ne(\omega) = Nh(Y_N(\omega))$. The canonical distribution of the system states ω with inverse temperature β is given as

$$\mu^N_{\mathrm{can},\beta}(A) = \frac{\int_{A\subseteq\Xi} e^{-\beta E(\omega)}\mu^N(d\omega)}{\int_{\Xi} e^{-\beta E(\omega)}\mu^N(d\omega)} = \frac{\int_{A\subseteq\Xi} e^{-\beta N h(Y_N(\omega))}\mu^N(d\omega)}{\int_{\Xi} e^{-\beta N h(Y_N(\omega))}\mu^N(d\omega)}. \tag{41}$$

As a consequence of Equation (36) we have

$$\int_{Y_N^{-1}(B)} e^{-\beta N h(Y_N(\omega))}\mu^N(d\omega) = \int_B e^{-\beta N h(m)} Q^N(dm), \tag{42}$$

for arbitrary measurable subsets B of the reals $\mathbf{R}$, *e.g.*, intervals $B = [a, b]$. Therefore the canonical distributions $Q^N_{\mathrm{can},\beta}$ of the mean magnetization m are in a simple way related to the distributions Q^N of the mean magnetization, namely by

$$Q^N_{\mathrm{can},\beta}(dm) := \frac{e^{-\beta N h(m)} Q^N(dm)}{\int_{\mathbf{R}} e^{-\beta N h(m)} Q^N(dm)}. \tag{43}$$

At this stage, Varadhan's lemma can be applied: If the distributions Q^N, $N = 1, 2, \ldots$, have a large deviation property with respect to the entropy function $I : \mathbf{R} \to \mathbf{R}$, then the corresponding *canonical* distributions $Q^N_{\mathrm{can},\beta}$, $N = 1, 2, \ldots$ have a large deviation property with respect to the entropy function

$$I_{(-\beta h)}(y) = (I(y) + \beta h(y)) - \inf_{y\in\mathbf{R}}\{I(y) + \beta h(y)\}. \tag{44}$$

Therefore the entropy function I gives rise to entropy functions $I^\beta := I_{(-\beta h)}$ depending on the inverse temperature β. I^β describes "how fast" particular energies die out with increasing number of particles. For infinite temperature ($\beta = 0$), the entropy I^0 coincides with the original entropy I. An entropy $I^\beta := I_{(-\beta h)}$ for some given β is *not* an entropy in the sense of phenomenological thermodynamics, but describes the canonical distributions with changing number N of particles *at temperature* β. The entropies I^β show clearly that large deviation statistics provides an approach more detailed than usual statistical mechanics (Ellis, 1985).

2. Relations Between Large Deviation Entropy and Other Entropies

2.1. Shannon Entropy and Jaynes' Principle of Maximum Entropy

In large deviations statistics, entropy describes fluctuations (*e.g.*, around a mean energy). It is not restricted to independent random observables or independent particles. Moreover, in large deviations theory entropy is introduced as a *statistical* quantity. There are two questions which come up immediately if one tries to understand large deviation entropy in a broader context:

- What is the connection between the large deviation entropy $I = I(x)$ of Equation (12) and conventional entropies such as the *Shannon entropy*?
- What is the role of Jaynes' *principle of maximum entropy* within large deviations statistics?

The concept of entropy even in its conventional sense is used in quite a variety of versions (*cf.* Wehrl, 1978). A particularly important one among them is the Shannon entropy for classical systems. It is the entropy of a probability distribution $(p_1, p_2, \ldots, p_L)$, $\sum_{i=1}^{L} p_i = 1$, whereas the entropy in equation (12) depends on some real variable x, where x is, *e.g.*, the mean value of dice throws or the mean energy in a thermodynamic system or the mean magnetization, *etc.*

This mean value is the mean over some empirical probability distribution and refers to some number N of experiments (throws of dice, particles in a thermodynamic system, *etc.*). A simple example is the empirical probability distribution (1/6, 0, 2/6, 1/6, 1/6, 1/6) for the results 1, 2, 3, 4, 5, 6 in $N = 12$ throws of a dice. This particular probability distribution gives rise to a mean value of 3.667 (by contrast to the theoretical mean value of 3.5). The large deviation entropy $I(3.667)$ indicates "how fast" the probability to observe the mean value 3.667 goes to zero with increasing N.

The Shannon entropy of the empirical probability distribution

$$(p_1, p_2, \ldots, p_L) = (1/6, 0, 2/6, 1/6, 1/6, 1/6) \tag{45}$$

is 1.5607 whereas the maximum possible Shannon entropy is log 6 = 1.7918. What is the relation between the Shannon entropy of the distribution (45) and the large deviation entropy $I(3.667)$ of the mean value 3.667? And what is the meaning of the difference between the particular value 1.5607 of the Shannon entropy and the maximum possible Shannon entropy 1.7918? Both questions will be addressed in this and the subsequent section. It will be shown how the principle of maximum entropy can be derived and refined within the framework of large deviations statistics. In the particular case of independent and identically distributed (i.i.d.) events the underlying mathematical result is Sanov's theorem.

A popular way of introducing entropy is information theory. Let us consider "experiments" P and Q with collectives of events $\{\eta_1, \eta_2, \ldots, \eta_L\}$ and $\{\sigma_1, \sigma_2, \ldots, \sigma_M\}$ and corresponding probabilities $\{p_1, p_2, \ldots, p_L\}$ and $\{q_1,$

$q_2, \ldots, q_M\}$, respectively. In order to quantify the information contained in these experiments by probability distributions one can use a "measure of information" $H(P) = H(p_1, p_2, \ldots, p_L)$ or $H(Q) = H(q_1, q_2, \ldots, q_M)$. There are some conditions which such a measure of information H should satisfy:

1. H is symmetric, *i.e.*, $H(p_{I(1)}, p_{I(2)}, \ldots, p_{I(L)}) = H(p_1, p_2, \ldots, p_L)$ for any permutation I;
2. H is extendable, *i.e.*, $H(p_1, p_2, \ldots, p_L) = H(p_1, p_2, \ldots, p_L, 0)$;
3. H is additive for independent experiments P and Q, *i.e.*, $H(P \wedge Q) = H(P) + H(Q)$, where $P \wedge Q$ is the experiment with the collective of events $\{(\eta_1, \sigma_1), (\eta_1, \sigma_2), \ldots, (\eta_2, \sigma_1), \ldots, (\eta_L, \sigma_M)\}$;
4. H is subadditive for non-independent experiments P and Q, *i.e.*, $H(P \wedge Q) \leq H(P) + H(Q)$;
5. H vanishes, *i.e.*, $\lim_{p_1 \to 1} H(p_1, p_2, \ldots, p_L) = 0$, if the outcome of an experiment is (almost) certain.

These conditions determine the measure of information H, up to some positive constant k, as the Shannon entropy. The *theorem of Aczél, Forte, and Ng* (1974) states (see also Aczél and Daróczy, 1975): Assume that H is a measure of information satisfying conditions (1)–(5). Then there exists a positive constant k such that

$$H(p_1, p_2, \ldots, p_L) = -k \sum_{i=1}^{L} p_i \ln p_i. \tag{46}$$

The concept of a measure of "information" is not the only possible interpretation of an entropy. As an alternative, Jaynes (1957a,b) has called $H(P)$ a measure for the "amount of uncertainty" represented by the distribution P. A broad distribution represents more uncertainty than a sharply peaked one. Jaynes proposed to describe a maximum amount of uncertainty by a maximum amount of entropy. Let us look at his principle of maximum entropy as a first application of entropy concepts.

Suppose that all available information about some system with the collective of events $\{\sigma_1, \sigma_2, \ldots, \sigma_L\}$ is incorporated in the m different expectation values of m observables a_ν,

$$\langle a_\nu \rangle = \sum_{i=1}^{L} p_i \, a_\nu(\sigma_i), \qquad \nu = 1, 2, \ldots, m. \tag{47}$$

What can be said about expectation values of other observables $b(\sigma_i)$ and about the unknown probabilities p_i themselves? Jaynes proposed to estimate the probabilities p_i, $i = 1, 2, \ldots, L$, as the "most uncertain" probability distribution compatible with the expectation values in (47), *i.e.*, the probability distribution $P^{\max} = (p_1^{\max}, \ldots, p_L^{\max})$ with maximum Shannon entropy H subject to the constraints in (47).

The principle of maximum entropy distinguishes a probability distribution $P^{\max}$ which has very favorable properties summarized in the following

theorem. Information measures other than Shannon entropy, such as $(-\sum p_i^2)$, are much more difficult to handle.

Theorem: The maximum entropy distribution subject to the constraints has the form

$$p_j^{\max} = \frac{1}{Z} \exp\left\{-\sum_{\nu=1}^{m} \lambda_\nu a_\nu(\sigma_j)\right\}, \qquad j = 1, 2, \ldots, L, \tag{48}$$

where the "partition function" Z is the normalization constant

$$Z = \sum_{i=1}^{L} \exp\left\{-\sum_{\nu=1}^{m} \lambda_\nu a_\nu(\sigma_i)\right\}. \tag{49}$$

Here, the coefficients λ_ν (one such coefficient for each constraint) are Lagrange multipliers depending on the expectation values in (47). In order to determine them, the partition function Z in (49) is used. As a typical example for a maximum entropy distribution, consider the Boltzmann distribution

$$p_i^{\max} = \frac{1}{Z} \exp\{-\beta\epsilon(\sigma_i)\}, \qquad i = 1, 2, \ldots, L, \tag{50}$$

where $\epsilon(\sigma_i)$ is the "energy" of the event (*e.g.*, molecular state) σ_i. The Lagrange multiplier β is usually interpreted as an inverse temperature $\beta = 1/kT$. Its particular value depends on the "mean energy" u in the constraint

$$u = \sum_{i=1}^{L} p_i \epsilon(\sigma_i). \tag{51}$$

For the maximum entropy principle, it is important to include all available information into the constraints (47). But even if this is not done, one can sometimes conclude from maximum entropy considerations what should have been included. An example is the mean magnetization m of a Curie–Weiss model consisting of N spins without external magnetic field below Curie temperature (Ellis, 1985). We shall discuss this model in the large deviation setting, since it will turn out in Section 2.2 below that large deviation statistics is a more refined way to treat maximum entropy estimates, and that the crucial arguments cannot easily be grasped in Jaynes' setting alone.

The Hamiltonian of the Curie–Weiss model without external field is given by

$$H_N = -\frac{J}{2N} \sum_{i,j=1}^{N} S_i S_j, \tag{52}$$

where the spins are $S_i = 1$ or $S_i = -1$. If one takes only the mean energy u as in (51) as a constraint, one arrives at the canonical distribution (50). The respective distribution density $p_N^\beta = p_N^\beta(m)$ of the mean magnetization m turns out to have two peaks as sketched in Figure 1. This sort of distribution tells us that we actually have *two different* regimes, one for negative and one for posi-

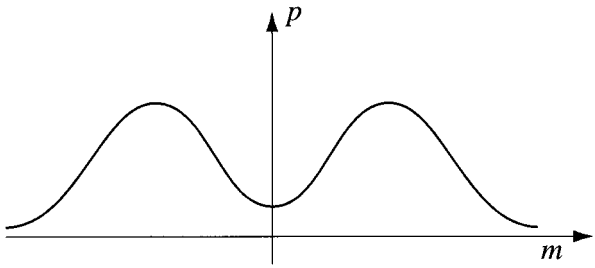

Fig. 1. Schematic representation of a typical distribution of the specific magnetization of the Curie–Weiss model below the Curie temperature.

tive permanent magnetization. If we would study an underlying spin dynamics (such as the Glauber dynamics used in numerical Monte Carlo simulations), it would turn out that for finite N a change from negative to positive magnetization occurs only with low but non-zero probability. Hence for finite N the system is still ergodic, though this is not necessarily visible on a short time scale. The larger the number N of spins, the smaller the probability for changes between negative and positive magnetizations. For infinite N, this probability becomes zero and ergodicity is strictly broken.

This is to say that effectively one has only negative or only positive mean magnetization (depending on the initial state and after some transient time, if the mean magnetization of the initial state is close to zero). The mean value $\int_{-\infty}^{+\infty} m p_N^\beta(m) dm$ of the distribution of m is zero, but nevertheless mean magnetization zero arises with low probability. Increasing the number N of spins leads to a more and more pronounced two-peak distribution density of m with smaller and smaller peaks, and finally there are two delta peaks left in the limit of infinitely many spins. Hence the probability $\int_{-\delta}^{+\delta} p_N^\beta(m) dm$ to find a mean magnetization in the small interval $[-\delta, +\delta]$ around $m = 0$ goes rapidly to zero for increasing N. Actually

$$\int_{-\delta}^{+\delta} p_N^\beta(m) dm \sim \exp\{-N s_0\}, \tag{53}$$

where s_0 is some positive scalar. In this way maximum entropy calculations lead to the conclusion that another macroscopic observable besides mean energy has been "overlooked," namely the mean magnetization. Consequently an additional constraint for the mean magnetization must be introduced.

The specific (mathematical) Massieu potential for a spin system admitting two values, $S_i \in \{-1, +1\}$, is

$$\phi(t) = \ln \int_{\Omega} e^{tS_1} d\mu(\omega) = \ln\left(\frac{1}{2}e^{t} + \frac{1}{2}e^{-t}\right) = \ln(\cosh(t)), \tag{54}$$

since the spin variable S_1 takes the values +1 and –1 with probability 1/2 each. The Legendre transform of this Massieu potential can be computed as

$$I_{\text{spin}}(z) = \frac{1-z}{2}\ln(1-z) + \frac{1+z}{2}\ln(1+z), \qquad |z| < 1, \tag{55}$$

$$I_{\text{spin}}(z) = \infty, \qquad |z| \geq 1. \tag{56}$$

The distributions Q^N of the mean magnetization $\sum_{j=1}^{N} S_j / N$ fulfill a large deviation result with respect to this entropy I. This result allows us to compute the specific (physical, but large-deviation normalized) Massieu potential for the Curie–Weiss model using Varadhan's lemma:

$$\Phi_{\text{CW}}(\beta) = \lim_{N\to\infty} \ln \int_{\Omega_N} \exp\left(-\beta H_N(\omega)\right) d\mu^N(\omega) \tag{57}$$

$$= \lim_{N\to\infty} \ln \int_{\Omega_N} \exp\left\{N h_{\beta,B}\left(\frac{\sum_{j=1}^{N} S_j}{N}\right)(\omega)\right\} d\mu^N(\omega) \tag{58}$$

$$= \lim_{N\to\infty} \frac{1}{N} \ln \int_{\mathbf{R}} e^{N h_{\beta,B}(x)} Q^N(dx) \tag{59}$$

$$= \sup_{z\in\mathbf{R}} \{h_{\beta,B}(z) - I_{\text{spin}}(z)\}, \tag{60}$$

where the function $h_{\beta,B}$ is defined as

$$h_{\beta,B} := \beta B z + \frac{1}{2}\beta J z^2. \tag{61}$$

Looking for the supremum in Equation (60) leads to the Curie–Weiss equation

$$h'_{\beta,B}(z) = \beta(B + Jz) \stackrel{!}{=} I'_{\text{spin}}(z) = \frac{1}{2}\ln\frac{1+z}{1-z}, \tag{62}$$

yielding one or two solutions with a maximum in Equation (60), depending on whether or not $\beta > J^{-1}$ (*i.e.*, for situations below or above the critical point).

Moreover, a large-deviation result for the distributions of the mean magnetization with respect to the canonical state at inverse temperature β

$$Q_{N,h_{\beta,B}}([a,b]) := \frac{\int_a^b e^{N h_{\beta,B}} Q^N(dz)}{\int_{\mathbf{R}} e^{N h_{\beta,B}} Q^N(dz)} \tag{63}$$

can be derived using Equation (33). The respective entropy is given as

$$I_{\text{CW},\beta} = I_{\text{spin}} - h_{\beta,B} - \inf_{z\in\mathbf{R}} \{I_{\text{spin}} - h_{\beta,B}\}. \tag{64}$$

If the temperature is below the Curie point, this leads to a non-convex entropy $I_{\mathrm{CW},\beta}$ with *two minima* $-m_\beta$ and $+m_\beta$ (Ellis, 1985) for negative and positive permanent magnetization (for vanishing external field $B = 0$ see Figure 2).

Such a non-convex entropy tells us "how fast" a phase separation appears with increasing number N of spins (at some fixed inverse temperature β), *i.e.*, "how fast" the Curie–Weiss magnet concentrates on the usual equilibrium magnetizations $-m_\beta$ and $+m_\beta$. For infinitely many spins, there are only two equilibrium magnetizations, whereas for a finite number of spins other mean magnetizations such as $m = 0$ arise in the canonical state below the critical Curie temperature. This is quite an interesting result, though its derivation is technically simple (by use of Varadhan's lemma) due to the mean-field structure of the Curie–Weiss Hamiltonian. We would be interested to see a similar result for something like the van der Waals gas, showing how the two phases (liquid and gas) separate with increasing number of particles. Incidentally, the separation of background noise and foreground patterns in images (Atmanspacher, Wiedenmann, and Amann, 1995) or the separation of random and non-random components in time series (Atmanspacher and Scheingraber, 1999) can be based on related ideas.

2.2. Deriving Jaynes' Principle from Large Deviations Statistics

The maximum entropy principle is of crucial importance in statistical physics. It allows us, for example, to derive the canonical distribution for a given temperature. What is the relation between this maximum entropy principle on the one hand and large deviations statistics on the other? Can the maximum entropy principle be derived in an appropriate large deviation setting?

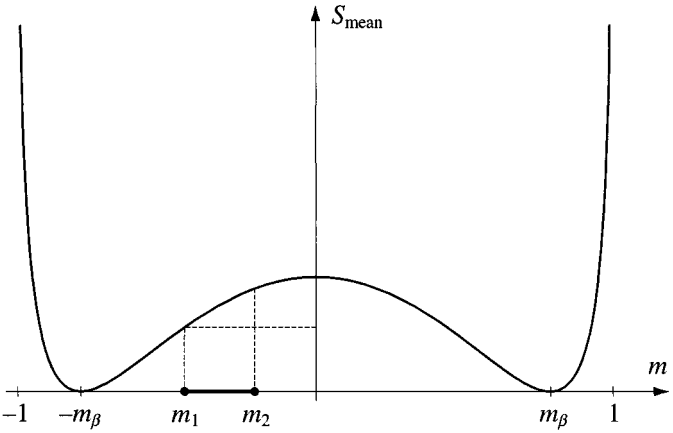

Fig. 2. Schematic representation of an entropy function S_{mean} in the sense of large deviations theory, describing the Curie–Weiss model below the Curie temperature.

And, if yes, does large deviations statistics provide results which lead beyond the maximum entropy principle?

In Section 1.2 the general scheme of large deviations was introduced, but in our examples only large deviation entropies with respect to some mean value (such as the mean energy or the mean magnetization) were discussed. The corresponding large deviation principles are called *level I* large deviation principles. Now we proceed a step further to *level II* large deviations and look at entropies with respect to some probability distribution $(p_1, p_2, \ldots, p_M)$ *as a whole*. (Note that there is an extension of large deviations statistics to *level III* (*cf.* Oono, 1989) which remains disregarded in this introductory paper.)

We shall use the same standard example of N particles as in Section 1.2. Given an ensemble $\omega \in \Omega_N$, a corresponding *empirical distribution* (or occupational distribution) $L_N(\omega, \cdot)$ is defined by

$$\mathbf{R} \ni A \to L_N(\omega, A) = \frac{1}{N} \sum_{j=1}^{N} \delta_{X_j(\omega)}(A). \tag{65}$$

For convenience, A is some arbitrary measurable subset of the real numbers $\mathbf{R}$; $\delta_{X_j(\omega)}(A)$ is 1 if $X_j(\omega)$ is contained in A and 0 if not. If, for example, the ensemble $\omega \in \Omega_5$ is given as

$$\omega = (\epsilon_1, \epsilon_3, \epsilon_1, \epsilon_1, \epsilon_4), \tag{66}$$

then the corresponding empirical distribution on $(\epsilon_1, \epsilon_2, \ldots, \epsilon_M)$ is (3/5, 0, 1/5, 1/5, 0, 0, ..., 0), and the corresponding empirical distribution on the real numbers (as defined in (65)) is given as

$$L = \frac{3}{5}\delta_{\epsilon_1} + \frac{1}{5}\delta_{\epsilon_3} + \frac{1}{5}\delta_{\epsilon_4}. \tag{67}$$

For a given number N of particles one gets an empirical distribution for every $\omega \in \Omega_N$ and therefore a *distribution* Q^N *of empirical distributions*. Level II large deviations theory addresses the question which of these empirical distributions "survive" and which of them "die out" with increasing N. The conceptual connection with Jaynes' maximum entropy principle is as follows:

- The entropy $H(p_1, p_2, \ldots, p_M)$ of some empirical probability distribution $P = (p_1, p_2, \ldots, p_M)$ describes the *decay rate* of P with increasing number of particles N.
- Only the empirical probability distribution with maximum entropy (subject to the constraints in (47)) survives in the limit $N \to \infty$.

In Jaynes' setting, *only* the empirical distribution with maximal entropy is considered. In a level II large deviation setting, the entropy is *statistically* relevant for *any* empirical distribution. It describes the decay of this empirical distribution with increasing N.

Sometimes entropies more general than the Shannon entropy have to be used to describe situations where the random variables (particles) are not independent. If, furthermore, the entropy function H is very flat (*e.g.*, at a phase transition), then empirical distributions other than the Boltzmann distribution can play an important role even for large N. In such a case, the maximum entropy principle fails to give a correct description of the situation.

The mathematical result connecting large deviations statistics and the maximum entropy principle is Sanov's theorem. We do not present its most general version here but only a special case that is sufficient to understand our example of N independent particles with energies $\Sigma := (\epsilon_1, \ldots, \epsilon_M)$. Moreover, we avoid mathematical details such as the correct formulation of the large deviation lower and upper bounds for open and closed sets (of empirical distributions) and the correct (weak or strong) topologies of the space of empirical distributions (Deuschel and Stroock, 1989).

Sanov's theorem (Sanov, 1957): Consider independent, identically distributed random variables and their corresponding distributions Q^N of empirical distributions on measurable sets D ($\subseteq M_1(\Sigma)$) defined by

$$Q^N(D) := \mu^N \{\omega \in \Omega_N : L_N(\omega, \cdot) \in D\}. \tag{68}$$

Here $M_1(\Sigma)$ is the set of all probability measures on the set Σ of energies. Then the following level II large deviation principle holds:

$$\lim_{N\to\infty} \frac{1}{N} \ln Q^N(D) = -\left(\inf_{\nu\in D} I_\mu^{(2)}(\nu)\right), \tag{69}$$

where the level II entropy $I_\mu^{(2)}$ is defined as

$$I_\mu^{(2)}(\nu) := \sum_{i=1}^{M} \nu_i \ln\left(\frac{\nu_i}{\mu_i}\right). \tag{70}$$

The large deviation entropy $I_\mu^{(2)}$ is the relative entropy of the empirical distribution ν with respect to the measure μ. The measure μ corresponds to the reference measure of maximum entropy considerations. If μ is the equipartition measure $(1/M, 1/M, \ldots, 1/M)$, the large deviation entropy $I_\mu^{(2)}$ is related to the Shannon entropy $H(\nu)$ by:

$$I_\mu^{(2)}(\nu) = \ln M - H(\nu). \tag{71}$$

Since $H(\nu)$ and $I_\mu^{(2)}(\nu)$ have opposite sign, a maximum entropy in Jaynes' setting turns into a minimum entropy in the large deviation setting.

The level I entropy $I^{(1)} = I^{(1)}(u)$ can be derived from the level II entropy $I_\mu^{(2)}$ by a so-called *contraction principle:* Denote by $\Phi(u) := \{\nu : \int_{\mathbf{R}} x\nu(dx) = u\}$ the set of all empirical distributions with mean energy u, *i.e.*, the set of all empirical distributions satisfying the constraint (51). Then

$$I^{(1)}(u) = \inf_{\nu \in \Phi(u)} I_\mu^{(2)}(\nu). \tag{72}$$

Consequently, the level I entropy $I^{(1)} = I^{(1)}(u)$ can be computed as

$$I^{(1)}(u) = \ln M - \sup_{\nu \text{ has energy } u} H(\nu_1, \nu_2, \ldots, \nu_M). \tag{73}$$

Note that the infimum of $I_\mu^{(2)}(\nu)$ turns into the supremum of $H(\nu)$ since I and H are of opposite sign.

This result — based on Sanov's theorem — gives us a first hint for the understanding of the maximum entropy principle (Ellis, 1985, p. 44, and lemma III.4.5). Suppose that we want to estimate a distribution ν for which the only information supplied is the mean energy u. Then Equation (73) suggests that the (unique) distribution $\nu_{\max,u}$ with maximum Shannon entropy (under the constraint that the energy is u) might be the distribution looked for. A more detailed understanding of $\nu_{\max,u}$ will be presented below.

Contraction principles can be formulated in a more general context. The mathematical basis is contained in the following two lemmas:

Lemma 1 (see Dembo and Zeitouni, 1993, lemma 4.1.4): A family of probability measures Q^N (*e.g.,* distributions of the mean energy or distributions of empirical distributions) on a regular topological space can have at most one entropy function associated with its large deviation property.

Lemma 2 (see Dembo and Zeitouni, 1993, theorem 4.2.1): Let X and Y be Hausdorff topological spaces and $f : X \to Y$ a continuous function. Consider a good rate function (*i.e.*, entropy) $I : X \to [0, \infty]$.

(a) For each $y \in Y$, define

$$\tilde{I}(y) := \inf\{I(x) : x \in X, y = f(x)\}. \tag{74}$$

$\tilde{I}$ is a good rate function (*i.e.*, entropy) on Y, where as usual the infimum over the empty set is taken as ∞.

(b) If the entropy I controls the large deviation behavior associated with a family of probability distributions Q^N on X, then $\tilde{I}$ controls the large deviation behavior associated with the family of probability distributions $\{Q^N \circ f^{-1}\}$ on Y.

Adapted to our situation, the space X is the space of empirical distributions (on which the Shannon entropy and the related large deviation level II entropy $I_\mu^{(2)}$ are defined), whereas Y is the space of possible mean energies, *i.e.*, $Y = \mathbf{R}$. The function f relates the mean energy to every empirical distribution,

$$f(\nu) := \int_{\mathbf{R}} x\nu(dx). \tag{75}$$

Furthermore, if Q^N is the distribution of empirical distributions, then $\{Q^N \circ f^{-1}\}$ is the distribution of mean energies. The contraction lemmas

allow us to construct non-convex entropy functions $\tilde{I}$ starting from a convex entropy function I.

Large deviations statistics is applicable even if the moments of a distribution (such as mean values) are not defined. In such cases one still has large deviations level II statistics of empirical distributions, but the contraction principle does not work since, *e.g.*, the mean in Equation (75) is infinite.

Let us now reconsider the connection between Sanov's theorem and the maximum entropy principle. Since Jaynes' constraints as in (51) are not incorporated in Sanov's theorem, we focus on empirical distributions with mean energy u, *i.e.*, empirical distributions ν contained in the set $\Phi(u) := \{\nu : \int_{\mathbf{R}} x\nu(dx) = u\}$. Restricting Sanov's large deviation result in (69) to $\Phi(u)$ then leads to

$$\lim_{N\to\infty} \frac{1}{N} \ln Q^N(D) = -\left(\inf_{\nu\in D} I_\mu^{(2)}(\nu) - \inf_{\nu\in\Phi(u)} I_\mu^{(2)}(\nu) \right), \qquad D \subseteq \Phi(u), \quad (76)$$

or, rewritten in terms of the Shannon entropy $H = H(\nu)$

$$\lim_{N\to\infty} \frac{1}{N} \ln Q^N(D) = -\left(\sup_{\nu\in\Phi(u)} H(\nu) - \sup_{\nu\in D} H(\nu) \right), \qquad D \subseteq \Phi(u). \quad (77)$$

Here D are appropriate sets of empirical distributions.

Proof: Define a continuous function F on the set $M_1(\Sigma)$ of all probability measures on the "set of energies" Σ which takes the positive value $R > 1$ on the energy shell $[u - \delta, u + \delta]$ and the value 0 outside the energy shell $[u - \delta_1, u + \delta_1]$, where $\delta < \delta_1$ are small positive numbers. Restrict the distributions Q^N to an energy shell by setting (see Equation (33)

$$Q_{N,\text{energy shell}}(A) := \frac{\int_A \exp(NF(\nu))Q_N(d\nu)}{\int_{M_1(\Sigma)} \exp(NF(\nu))Q_N(d\nu)}, \qquad A \subseteq M_1(\Sigma),\ n = 1, 2, \ldots. \quad (78)$$

Then by Varadhan's lemma these distributions $Q_{N,\text{energy shell}}$ fulfill a large deviation principle with respect to the entropy

$$I^{(2)}_{\text{energy shell}}(\nu) = (I^{(2)}(\nu) - F(\nu)) - \inf_{\nu\in M_1(\Sigma)} \{I^{(2)}(\nu) - F(\nu)\}. \quad (79)$$

If the constant R is large enough, we get a large deviation principle which is restricted to the empirical distributions in the considered energy shell. This is nothing but restricting the entropy $I^{(2)}$ in Sanov's theorem to the energy shell and normalizing it to zero. Equation (77) thus gives an energy rather than a distribution over an energy shell.

The result in (77) is the derivation of Jaynes' principle of maximum entropy in the scheme of large deviations statistics. The main differences between Jaynes' principle and large deviations statistics are:

- The maximum entropy principle focuses only on one particular empirical distribution, namely that with maximum entropy.
- Large deviations statistics accepts that additional empirical distributions (with the same mean energy u) exist and describes how these other empirical distributions "die out" with increasing number N of particles. The decay rate of an empirical distribution ν is given by the entropy $I^{(2)}(\nu)$. Hence the entropy of arbitrary empirical distributions (and not just of equilibrium distributions) has a well-defined statistical meaning.

This sort of explanation of the maximum-entropy principle assigns statistical meaning to probability distributions even if they do *not* have maximum entropy.

3. Small Fluctuations

The large deviation approach was originally designed to address the decay of large deviations with an increasing number of particles or dice throws *etc.* Now we shall discuss the so-called "small" fluctuations — proportional to the square root $\sqrt{N}$ of the number of particles — which survive even for a large or infinite number of particles. They are physically important; for instance, they can become very large at phase transitions. We shall see that the large deviation entropy also gives information about the behavior of small fluctuations and not just large fluctuations.

The fact that only small fluctuations survive for $N \to \infty$ (in the absence of phase transitions) is expressed in the *central limit theorem*. It says that for a distribution of independent and identically distributed random variables X_j, $j = 1, 2, \ldots$, one has

$$\frac{\sum_{j=1}^{N} X_j - N\,\mathcal{E}(X_1)}{\sqrt{N}} \xrightarrow{\text{distr}} N(0, \sigma^2) \qquad \text{for } N \to \infty, \tag{80}$$

which means that $P[N^{-1/2} \sum_{j=1}^{N} X_j - N\,\mathcal{E}(X_1)]$ converges weakly to $N(0, \sigma^2)$. Here $\mathcal{E}(X_1)$ is the expectation value of the random variable X_1 (identical to the expectation values of the other variables X_j), and the variance σ^2 is given by the variance of X_1, *i.e.*,

$$\sigma^2 := \mathcal{E}\big((X_1 - \mathcal{E}(X_1))^2\big). \tag{81}$$

The limit in (80) is taken with respect to the distribution of i.i.d. variables, and $N(0, \sigma^2)$ is the normal (Gaussian) distribution with variance σ^2.

Theorem (Martin-Löf, 1982): Let X_j, $j = 1, 2, \ldots$, satisfy a large deviations property with an entropy function I admitting a unique minimum z^*. Then the variance σ^2 in the central limit theorem is given as

$$\sigma^2 = \frac{1}{I''(z^*)}, \tag{82}$$

where I'' is the second derivative of the entropy function I.

For example, in the Curie–Weiss model above the critical temperature T_{crit}, the second derivative of the entropy $I_{\text{CW}}(z)$ at the minimum $z = 0$ can be computed as

$$I''_{\text{spin}}(z) - h''_{\beta,B}(z) = \frac{1}{1 - z^2} - \beta J \stackrel{z=0}{=} 1 - \beta J. \tag{83}$$

Hence the variance σ^2 of the small fluctuations is given as $\sigma^2 = (1 - \beta J)^{-1}$. In terms of the temperature T this gives

$$\sigma^2 = \frac{1}{1 - \beta J} = \frac{T}{T - T_{\text{crit}}} = t^{-1}, \tag{84}$$

where the usual notation $t = (T - T_{\text{crit}})/T$ has been used. If β converges to the critical value $\beta_{\text{crit}} = J^{-1}$, one therefore gets $\sigma \sim t^{-1/2}$. This result can be compared with the correlation length ξ in mean-field models which behaves as $\xi \sim t^{-1/2}$ at the critical point.

Martin-Löf's theorem states that for a flat entropy function one gets a very large variance σ^2. For a phase transition, in particular, the second derivative $I''(z^*)$ of the entropy function goes to zero and hence σ^2 diverges. In this case one must be extremely careful with applications of the maximum entropy principle. The large deviation result in (77) tells us that — due to the flat entropy function — even for large (but finite) N empirical distributions other than the equilibrium distributions may survive (in all practical applications N is never infinite). Martin-Löf's theorem, on the other hand, makes clear that the variance of the small fluctuations (proportional to $\sqrt{N}$) diverges and therefore small fluctuations play an important role.

At the phase transition itself, the normalization of the global observables must be chosen very carefully. The usual normalization as in the central limit theorem cannot be used any more. The second derivative $I''(z^*)$ vanishes in the minimum z^*, but higher derivatives of $I^{(1)}$ at z^* can have non-zero values. In the case of the Curie–Weiss model at the Curie temperature, the central limit theorem is replaced by the result that the distribution $\sum_{j=1}^{N} S_j / N^{3/4}$ converges to the non-Gaussian distribution $\exp(-x^{4/12}) / \int_{\mathbf{R}} \exp(-x^{4/12})$. The exponents 3/4 and 4/12 in this distribution are connected with the behavior of the large deviation entropy $I^{(1)}$ at its minimum. Therefore, the large deviation entropy does not only give information about large deviations, but also about small deviations and fluctuations at phase transitions.

4. Non-I.I.D. Situations

Most of our examples so far were concerned with independent and identically distributed (i.i.d.) random variables. As the intensive coupling between the

spin variables in the Curie–Weiss model demonstrates, however, random variables are *not* always independent, or their independence is only an idealization. Successive psychological experiments, for example, are typically correlated, and it may even be that this correlation is just the interesting point.

In this section we consider random variables which are *not* independent, replacing the sum $\sum_{j=1}^{N} X_j$ by an "overall" stochastic variable W_N. The respective large deviation principle opens the way to applications which are not related to standard statistical physics at all (*cf.* also Section 5). The product measure μ^N, which plays a crucial role in i.i.d. situations, is now replaced by an arbitrary probability measure P_N on some measure space Ξ_N.

As mentioned in Section 1.2, the entropy $I^{(1)}$ is the Legendre transform of the (specific) Massieu potential

$$\phi_N(t) = \frac{1}{a_N} \ln \left(\int_{\Omega_N} e^{\langle t | W_N(\omega) \rangle} P_N(\omega) \right). \tag{85}$$

For independent random variables X_j, the Massieu potential (85) is given by

$$\frac{1}{N} \ln \left(\int_{\Omega_N} e^{\langle t | \sum_{j=1}^{N} X_j(\omega) \rangle} \mu^N(\omega) \right) = \ln \left(\int_{\Omega} e^{\langle t | X_1(\omega) \rangle} \mu(\omega) \right). \tag{86}$$

If the X_j are independent and identically distributed, ϕ_N does not depend on N, no limit $N \to \infty$ is required, and the Massieu potential is continuous and differentiable. In the general case of non-i.i.d. situations, differentiability is not a necessary consequence of arbitrarily chosen stochastic variables W_N and arbitrarily chosen probability measures P_N, and $\lim_{N\to\infty} \phi_N$ does not necessarily exist.

If the limit $\phi(t) = \lim_{N\to\infty} \phi_N(t)$ exists nevertheless and is differentiable, i.i.d. conditions are not required to derive large deviation results. The following is a (very) special case of the *Ellis-Gärtner theorem* (Dembo and Zeitouni, 1993; theorem 2.3.6): Assume that the limit $\phi(t) = \lim_{N\to\infty} \phi_N(t)$ for an appropriate sequence a_N exists in $\mathbf{R}$ (often $a_N = N$ is sufficient) and is differentiable. Then the Legendre transform

$$I^{(1)}(z) := \sup_{t \in \mathbf{R}} \{tz - \phi(t)\} \tag{87}$$

serves as a large deviation entropy for the large deviation principle

$$\lim_{N\to\infty} \frac{1}{a_N} \ln Q^N([a, b]) = - \inf_{z \in [a,b]} I^{(1)}(z). \tag{88}$$

5. Multifractals

As an example, consider some measure ρ describing the chemical growth of an aggregate. For instance, $\rho = \rho(B)$ can be the aggregation probability at the surface of the aggregate where B is an arbitrary measurable subset of the surface. The support of the measure ρ, supp(ρ), is partitioned into boxes (2-, 3-,

or d-dimensional, depending on the situation) of size δ. These boxes are called B_m^δ, $m = 1, 2, \ldots, M_\delta$. For the large deviation result to be derived, we do not consider energies of particles but the random variables $\ln\{\rho(B_m^\delta)\}$. The normalization with respect to the number of particles N is replaced by a normalization with respect to $|\ln\delta|$. The corresponding (non-normalized) Massieu potential in the multifractal formalism is usually defined as (Falconer, 1990)

$$\tau_\delta(q) = \frac{1}{|\ln(\delta)|} \ln\left(\sum_{m=1}^{M_\delta} \rho(B_m^\delta)^q\right) \tag{89}$$

$$= \frac{1}{|\ln(\delta)|} \ln\left(\sum_{m=1}^{M_\delta} e^{q \ln\{\rho(B_m^\delta)\}}\right) \tag{90}$$

$$= \frac{1}{|\ln(\delta)|} \ln\left(M_\delta \int_{\Omega_\delta = \{1,2,\ldots,M_\delta\}} e^{q \ln\{\rho(B_m^\delta)\}} P_\delta(dj)\right) \tag{91}$$

$$= \frac{\ln M_\delta}{|\ln(\delta)|} + \frac{1}{|\ln(\delta)|} \ln\left(\int_{\Omega_\delta = \{1,2,\ldots,M_\delta\}} e^{q \ln\{\rho(B_m^\delta)\}} P_\delta(dj)\right). \tag{92}$$

The probability measure P_δ is constant on Ω_δ (equiprobability) and distinguishes the second term in (92) as an expression $\tilde{\tau}_\delta$ of large deviation type. The first term in (92) converges to the box-counting dimension of the support $\mathrm{supp}(\rho)$ of the measure ρ. Hence the Massieu potential $\tau(q) = \lim_{\delta\to 0} \tau_\delta$ can be formulated as

$$\tau(q) = \lim_{\delta\to 0} \tau_\delta(q) = \dim_F\{\mathrm{supp}(\rho)\} + \tilde{\tau}(q), \tag{93}$$

where $\dim_F\{\mathrm{supp}(\rho)\}$ is the box-counting dimension of the support of the measure ρ, which corrects the "wrong" normalization of the Massieu potential used. Here $\tilde{\tau}$ is the Massieu potential with the correct large-deviation conform normalization. If the Massieu potential $\tau(q) = \lim_{\delta\to 0} \tau_\delta(q)$ exists and is differentiable, a large-deviation result for the distribution Q_δ of $\ln\{\rho(B_m^\delta)\}/|\ln(\delta)|$ follows from the Gärtner-Ellis theorem in Section 4. This result is:

$$\lim_{\delta\to 0} \frac{\ln\{Q_\delta([a,b])\}}{|\ln(\delta)|} = -\inf_{\alpha\in[a,b]} I(\alpha), \tag{94}$$

where I is the Legendre transform of the Massieu potential $\tilde{\tau}$, *i.e.*, $I(\alpha) = \sup_{q\in\mathbf{R}}(q\alpha - \tilde{\tau}(q))$. The large deviation entropy I describes how fast (with decreasing δ) certain values of α die out.

If the scaling assumption

$$\lim_{\delta\to 0} \frac{\ln\{\rho(B_x^\delta)\}}{|\ln(\delta)|} = \alpha \tag{95}$$

holds at some point x, one has

$$\rho(B_x^\delta) \sim \delta^\alpha, \tag{96}$$

where B_x^δ is the box of size δ containing the point x. Hence the exponent α describes the *scaling behavior* of the measure ρ around the point x, and Q_δ is the distribution of scaling exponents. These formal considerations on scaling exponents can be given a rigorous mathematical meaning in certain (non-representative) examples. The large deviation result (94), on the other hand, is always rigorous as long as the Massieu potential is differentiable. Under the assumption of this differentiability, it follows that

$$I(\alpha) = \sup_{q\in\mathbf{R}} \{q\alpha - \bar{\tau}(q)\} \tag{97}$$

$$= \sup_{q\in\mathbf{R}} \{q\alpha - (\tau(q) - \dim_F\{\mathrm{supp}(\rho)\})\} \tag{98}$$

$$= \dim_F\{\mathrm{supp}(\rho)\} + \sup_{q\in\mathbf{R}} \{q\alpha - \tau(q)\} \tag{99}$$

$$= \dim_F\{\mathrm{supp}(\rho)\} - \inf_{q\in\mathbf{R}} \{q(-\alpha) + \tau(q)\} \tag{100}$$

$$= \dim_F\{\mathrm{supp}(\rho)\} - f(-\alpha) \tag{101}$$

where the definition

$$f(\alpha) := \inf_{q\in\mathbf{R}} \{q(\alpha) + \tau(q)\} \tag{102}$$

has been used.

Halsey *et al.* (1986) introduced the function $f(\alpha)$ (a "spectrum of singularities") as the Legendre transform of τ rather than $\bar{\tau}$ (*cf.* also Mandelbrot, 1974). The relation between this $f(\alpha)$ spectrum and the large deviation entropy I is then given by (*cf.* Riedi, 1995, for a careful investigation):

$$f(\alpha) := \inf_{q\in\mathbf{R}} \{q\alpha + \tau(q)\} = \dim_F\{\mathrm{supp}(\rho)\} - I(-\alpha). \tag{103}$$

Here τ is a convex function, and $f = f(\alpha)$ is a concave function (Falconer, 1990). Note that the usual mathematical Legendre transform as previously defined transforms convex functions into convex functions. The corresponding mathematical Legendre transform for concave functions ψ is given as $\inf_{q\in\mathbf{R}}(q\alpha - \psi(q))$.

The maximum of $f(\alpha)$ gives the box-counting dimension of $\mathrm{supp}(\rho)$. In certain (non-representative) situations, it can be shown that the support $\mathrm{supp}(\rho)$ is the union of the sets S_α,

$$S_\alpha := \{x \in \mathrm{supp}(\rho) : \rho(B_x^\delta) \sim \delta^\alpha\} \tag{104}$$

and that $f(\alpha)$ is the *Hausdorff dimension* of the sets S_α. In general, $f(\alpha)$ is convex over a finite range of scaling exponents α. By contrast to *single fractals* whose scaling behavior is completely characterized by a single value of α,

multifractals have to be characterized by a function $f(\alpha)$. The theory of multifractals is a typical example for large deviations at level I.

Physical situations are always bound to finiteness, and therefore the mathematical background of the large deviations theory with limits as in (88) is not directly applicable to data from real measurements. In such cases, large deviation techniques have to be properly and carefully adapted to the specific situation considered. If a system is not "sufficiently" large or cannot be observed "sufficiently" long, the empirical distribution itself has to be considered as a stochastic object whose fluctuations are studied. An example for such a large deviations level II scaling approach is discussed by Atmanspacher and Scheingraber (1999). For more information on the use of large deviation techniques in the area of complex systems research see, *e.g.*, Aizawa (1989), Young and Crutchfield (1994), and Seppäläinen (1995).

References

Aczél, J. and Daróczy, C. (1975). *Measures of Information and Their Characterizations*. New York: Academic.

Aczél, J., Forte, B., and Ng, C. T. (1974). Why the Shannon and Hartley entropies are "natural." *Adv. Appl. Prob.*, 6, 131–146.

Aizawa, Y. (1989). Non-stationary chaos revisited from large deviation theory. *Prog. Theor. Phys. Suppl.*, 99, 149–164.

Amann, A. (1994). The quantum-mechanic al measurement process in the thermodynamic formalism. In: Busch, P., Lahti, P., and Mittelstaedt, P. (Eds.), *Symposium on the Foundations of Modern Physics 1993 — Quantum Measurement, Irreversibility, and the Physics of Information*. Singapore: World Scientific. pp. 3–19.

Atmanspacher, H. and Scheingraber, H. (1999): Investigating deviations from dynamical randomness with scaling indices. *Journal of Scientific Exploration,* in press 2000.

Atmanspacher, H., Wiedenmann, G., and Amann, A. (1995). Descartes revisited — the endo/exo-distinction and its relevance for the study of complex systems. *Complexity* 1, 3, 15–21.

Bohr, H. and Tél, T. (1988). The thermodynamics of fractals. In: Hao, B.-L. (Ed.) *Directions in Chaos, Vol. 2*. Singapore: World Scientific. pp. 194–237.

Cramér, H. (1937): Sur un nouveau theorème-limite de la théorie des probabilités. In *Actualités Scientifiques et Industrielles, 736. Colloque consacré à la théorie des probabilités* Paris: Hermann. pp. 5–23.

Dembo, A. and Zeitouni, O. (1993). *Large Deviation Techniques and Applications*. Boston: Jones and Bartlett.

Deuschel, J. -D. and Stroock, D. W. (1989). *Large Deviations*. San Diego: Academic Press.

Ellis, R. S. (1985). *Entropy, Large Deviations, and Statistical Mechanics*. Berlin: Springer.

Falconer, R. S. (1990). *Fractal Geometry. Mathematical Foundations and Applications*. Chichester: Wiley.

Halsey, T. C., Jensen, M. H., Kadanoff, L. P., Procaccia, I., and Shraiman, B. I. (1986). Fractal measures and their singularities: the characterization of strange sets. *Phys. Rev. A,* 33, 1141–1151.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Phys. Rev.,* 106, 620–630.

Jaynes, E. T. (1957b). Information theory and statistical mechanics II. *Phys. Rev.,* 108, 171–190.

Lanford, O. E. (1973). Entropy and equilibrium states in classical statistical mechanics. In: Lenar, A. (Ed.), *Statistical Mechanics and Mathematical Problems*. Berlin: Springer. pp. 1–113.

Mandelbrot, B. B. (1974). Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier. *J. Fluid Mech.*, 62, 331–358.

Martin-Löf, A. (1982). A Laplace approximation for sums of independent random variables. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 59, 101–116.

Oono, Y. (1989). Large deviation and statistical physics. *Prog. Theor. Phys. Suppl.*, 99, 165–205.

Peinke, J., Parisi, J., Rössler, O. E., and Stoop, R. (1992). *Encounter with Chaos. Self Organized Hierarchical Complexity in Semiconductor Experiments*. Berlin: Springer.
Riedi, R. (1995). An improved multifractal formalism and self-similar measures. *J. Math. Anal. Appl.,* 189, 462–490.
Sanov, I. N. (1957). On the probability of large deviations of random variables (in Russian). *Mat. Sb*., 42, 11–44.
Seppäläinen, T. (1995): Entropy, limit theorems, and variational principles for disordered lattice systems. *Commun. Math. Phys*., 171, 233–277.
Varadhan, S.R.S. (1966). Asymptotic probabilities and differential equations. *Comm. Pure Appl. Math*., 19, 261–286.
Wehrl, A. (1978). General properties of entropy. *Rev. Mod. Phys.,* 50, 221–260.
Young, K. and Crutchfield, J. P. (1994): Fluctuation spectroscopy. *Chaos, Solitons, & Fractals,* 4, 5–39.