Probabilistic aspects of entropy*

Hans-Otto Georgii
Mathematisches Institut der Universität München
Theresienstr. 39, D-80333 München, Germany.

September 7, 2000

> *When Shannon had invented his quantity and consulted von Neumann how to call it, von Neumann replied: "Call it entropy. It is already in use under that name and besides, it will give you a great edge in debates because nobody knows what entropy is anyway."* [7]

Abstract

We give an overview of some probabilistic facets of entropy, recalling how entropy shows up naturally in various different situations ranging from information theory and hypothesis testing over large deviations and the central limit theorem to interacting random fields and the equivalence of ensembles.

AMS 1991 subject classifications. 28D20, 60F05, 60F10, 60J10, 60K35, 82B05, 82C20, 94A17, 94A24.

Key words and phrases. Maxwell–Boltzmann statistics, source coding, asymptotic equipartition, hypothesis testing, large deviations, maximum entropy principle, Markov chains, central limit theorem, conditional limit theorem, Gibbs measure, interacting particle system, equivalence of ensembles.

1 Entropy as a measure of uncertainty

As is well-known, it was Ludwig Boltzmann who first gave a probabilistic interpretation of thermodynamic entropy. He coined the famous formula

$$S = k \log W \tag{1.1}$$

which is engraved on his tombstone in Vienna: the entropy S of an observed macroscopic state is nothing else than the logarithmic probability for its occurrence, up to some scalar factor k (the Boltzmann constant) which is physically significant but can be ignored from a mathematical point of view. I will not enter here into a discussion of the history and physical significance of this formula; this is the subject of other contributions to this volume. Here I will simply recall its most elementary probabilistic interpretation.

*Opening lecture given at the "International Symposium on Entropy" hosted by the Max Planck Institute for Physics of Complex Systems, Dresden, Germany, 26–28 June 2000.

Let E be a finite set and μ a probability measure on E. [†] In the Maxwell–Boltzmann picture, E is the set of all possible energy levels for a system of particles, and μ corresponds to a specific histogram of energies describing some macrostate of the system. Assume for a moment that each $\mu(x)$, $x \in E$, is a multiple of $1/n$, i.e., μ is a histogram for n trials or, equivalently, a macrostate for a system of n particles. On the microscopic level, the system is then described by a sequence $\omega \in E^n$, the microstate, associating to each particle its energy level. Boltzmann's idea is now the following:

> *The entropy of a macrostate μ corresponds to the degree of uncertainty about the actual microstate ω when only μ is known, and can thus be measured by $\log N_n(\mu)$, the logarithmic number of microstates leading to μ.*

Explicitly, for a given microstate $\omega \in E^n$ let[‡]

$$L_n^\omega = \frac{1}{n} \sum_{i=1}^{n} \delta_{\omega_i} \tag{1.2}$$

be the associated macrostate describing how the particles are distributed over the energy levels. L_n^ω is called the *empirical distribution* (or histogram) of $\omega \in E^n$. Then

$$N_n(\mu) \equiv |\{\omega \in E^n : L_n^\omega = \mu\}| = \frac{n!}{\prod_{x \in E} (n\,\mu(x))!} \,,$$

the multinomial coefficient. In view of the n-dependence of this quantity, one should approximate a given μ by a sequence μ_n of n-particle macrostates and define the uncertainty $H(\mu)$ of μ as the $n \to \infty$ limit of the "mean uncertainty of μ_n per particle". Using Stirling's formula, we arrive in this way at the well-known expression for the entropy:

(1.3) Entropy as degree of ignorance: *Let μ and μ_n be probability measures on E such that $\mu_n \to \mu$ and $n\,\mu_n(x) \in \mathbb{Z}$ for all $x \in E$. Then the limit $\lim_{n\to\infty} \frac{1}{n} \log N_n(\mu_n)$ exists and is equal to*

$$H(\mu) = -\sum_{x \in E} \mu(x) \log \mu(x) \,.$$

A proof including error bounds is given in Lemma 2.3 of [5].

Though we have used the terms uncertainty and ignorance, the entropy $H(\mu)$ should not be considered as a subjective quantity. It simply counts the number of possibilities to obtain the histogram μ, and thus describes the *hidden multiplicity* of "true" microstates consistent with the observed μ. It is therefore a measure of the *complexity* inherent in μ.

To summarize: In Boltzmann's picture, μ is a histogram resulting from a random phenomenon on the microscopic level, and $H(\mu)$ corresponds to the observer's uncertainty of what is really going on there.

[†] Here and throughout we assume for simplicity that $\mu(x) > 0$ for all x.

[‡] We write δ_x for the Dirac measure at $x \in E$.

2 Entropy as a measure of information

We will now approach the problem of measuring the "uncertainty content" of a probability measure μ from a different side suggested by Shannon [35]. Whereas Boltzmann's view is backwards to the microscopic origins of μ, Shannon's view is ahead, taking μ as given and "randomizing" it by generating a random signal with alphabet E and law μ. His question is: How large is the receiver's effort to recover μ from the signal? This effort can be measured by the number of yes-or-no questions to be answered on the average in order to identify the signal (and thereby μ, after many independent repetitions). So it corresponds to the receiver's *a priori* uncertainty about μ. But, as observed by Shannon, this effort measures also the degree of information the receiver gets *a posteriori* when all necessary yes-or-no questions are answered. This leads to the following concept of information:

> *The information contained in a random signal with prescribed distribution is equal to the expected number of bits necessary to encode the signal.*

Specifically, a binary prefix code for E is a mapping $f : E \to \bigcup_{\ell \geq 1} \{0,1\}^\ell$ from E into the set of all finite zero-one sequences which is decipherable, in that no codeword $f(x)$ is a prefix of another codeword $f(y)$. (Such an f can be described by a binary decision tree, the leaves of which correspond to the codewords.) Let $\#f(x)$ denote the length of the codeword $f(x)$, and $\mu(\#f)$ the expectation of the random variable $\#f$ under μ. A natural candidate for the information contained in the signal is then the minimal expected length

$$I_{\mathrm{p}}(\mu) = \inf \Big\{ \mu(\#f) : f \text{ binary prefix code for } E \Big\}$$

of a binary prefix code for E. This quantity is already closely related to $H(\mu)$, but the relationship becomes nicer if one assumes that the random signal forms a memoryless source, in that the random letters from E are repeated independently, and one encodes signal words of length n (which are distributed according to the product measure μ^n). In this setting, $I_{\mathrm{p}}(\mu^n)/n$ is the information per signal letter, and in the limit $n \to \infty$ one obtains the

(2.1) Source coding theorem for prefix codes: *The information contained in a memoryless source with distribution μ is*

$$\lim_{n\to\infty} I_{\mathrm{p}}(\mu^n)/n = -\sum_{x\in E} \mu(x) \log_2 \mu(x) \equiv H_2(\mu) = \frac{1}{\log 2}\, H(\mu)\ .$$

For a proof of a refined version see Theorem 4.1 of [5], for example.

An alternative coding scheme leading to a similar result is block coding with small error probability. A binary n-block code of length ℓ with error level $\alpha > 0$ is a mapping $f : E^n \to \{0,1\}^\ell$ together with a decoder $\varphi : \{0,1\}^\ell \to E^n$ such that $\mu^n(\varphi \circ f \neq \mathrm{id}) \leq \alpha$. Let

$$I_{\mathrm{b}}(\mu^n, \alpha) = \inf \Big\{ \ell : \exists\ n\text{-block code of length } \ell \text{ at level } \alpha \Big\}$$

be the minimal length of a binary n-block code with error level α. The following result then gives another justification of entropy.

(2.2) Source coding theorem for block codes: *The information contained in a memoryless source with distribution μ is*

$$\lim_{n\to\infty} I_{\rm b}(\mu^n, \alpha)/n = H_2(\mu) \ ,$$

independently of the error level $\alpha > 0$.

The proof of this result (see e.g. Theorem 1.1 of [5]) relies on an intermediate result which follows immediately from the weak law of large numbers. It reveals yet another role of entropy and is therefore interesting in its own right:

(2.3) Asymptotic equipartition property: *For all $\delta > 0$,*

$$\mu^n\Big(\omega \in E^n : |\tfrac{1}{n} \log \mu^n(\omega) + H(\mu)| \le \delta\Big) \underset{n\to\infty}{\longrightarrow} 1 \ .$$

In other words, most ω have probability $\mu^n(\omega) \approx e^{-nH(\mu)}$. This may be viewed as a random version of Boltzmann's formula (1.1).

To conclude this section, let us mention that the entropy $H(\mu)$ admits several axiomatic characterizations which underline its significance as a measure of uncertainty and information; cf. e.g. the discussion on pp. 25–27 of [5]. However, compared with the previous genuine results these characterizations should rather be considered as *a posteriori* justifications.

3 Relative entropy as a measure of discrimination

Let E still be a finite set, and consider two distinct probability measures μ_0 and μ_1 on E. Suppose we do not know which of these probability measures properly describes the random phenomenon we have in mind (which might again be a random signal with alphabet E). We then ask the following question:

> *How easy is it to distinguish the two candidates μ_0 and μ_1 on the basis of independent observations?*

This is a standard problem of statistics, and the standard procedure is to perform a test of the hypothesis μ_0 against the alternative μ_1 with error level α. In fact, if we want to use n independent observations then we have to test the product measure μ_0^n against the product measure μ_1^n. Such a test is defined by a "rejection region" $R \subset E^n$; if the observed outcome belongs to R one decides in favor of the alternative μ_1, otherwise one accepts the hypothesis μ_0. There are two possible errors: rejecting the hypothesis μ_0 although it is true (first kind), and accepting μ_0 though it is false (second kind). The common practice is to keep the error probability of first kind under a prescribed level α and to choose R such that the error probability of the second kind becomes minimal. The minimum value is

$$\rho_n(\alpha; \mu_0, \mu_1) = \inf\Big\{\mu_1^n(R^c) : R \subset E^n, \ \mu_0^n(R) \le \alpha\Big\} \ .$$

Consequently, it is natural to say that μ_0 and μ_1 are the easier to distinguish the smaller $\rho_n(\alpha; \mu_0, \mu_1)$ turns out to be. More precisely:

The degree to which μ_1 can be distinguished from μ_0 on the basis of independent observations can be measured by the rate of decay of $\rho_n(\alpha;\mu_0,\mu_1)$ as $n \to \infty$.

An application of the weak law of large numbers completely similar to that in the source coding theorem (2.2) gives:

(3.1) Lemma of C. Stein: *The measure for discriminating μ_1 from μ_0 is*

$$-\lim_{n\to\infty}\frac{1}{n}\log\rho_n(\alpha;\mu_0,\mu_1)=\sum_{x\in E}\mu_0(x)\log\frac{\mu_0(x)}{\mu_1(x)}\equiv D(\mu_0\,|\,\mu_1)\;,$$

independently of the choice of $\alpha\in]0,1[$.

This result was first published in [2]; see also Corollary 1.2 of [5] or Lemma 3.4.7 of [6]. $D(\mu_0\,|\,\mu_1)$ is known as the *relative entropy*, *Kullback-Leibler information*, *I-divergence*, or *information gain*. If μ_1 is the equidistribution on E then $D(\mu_0\,|\,\mu_1)=\log|E|-H(\mu_0)$. Hence relative entropy is a generalization of entropy to the case of a non-uniform reference measure, at least up to the sign. (In view of the difference in sign one might prefer calling $D(\mu_0\,|\,\mu_1)$ the *negative* relative entropy. Nevertheless, we stick to the terminology above which has become standard in probability theory.)

Stein's lemma asserts that the relative entropy $D(\,\cdot\,|\,\cdot\,)$ measures the extent to which two probability measures differ. Although $D(\,\cdot\,|\,\cdot\,)$ is not a metric (neither being symmetric nor satisfying the triangle inequality), it can be used to introduce some kind of geometry for probability measures, and in particular some kind of projection of a probability measure on a convex set of such measures [3]. As we will see in a moment, these so-called I-projections play a central role in the asymptotic analysis of the empirical distributions (1.2). But first, as some motivation, let us mention a refinement of Stein's lemma for which the error probability of the first kind is not held fixed but decays exponentially at a given rate. The answer is in terms of L_n^ω and reads as follows.

(3.2) Hoeffding's theorem: *Let $0<a<D(\mu_1\,|\,\mu_0)$, and consider the test of μ_0 against μ_1 on n observations with the rejection region $R_n=\{\omega\in E^n : D(L_n^\omega\,|\,\mu_0)>a\}$. Then the error probability of the first kind decays exponentially with rate a, i.e.,*

$$\lim_{n\to\infty}\frac{1}{n}\log\mu_0^n(R_n)=-a\;,$$

and the error probability of the second kind satisfies the exponential bound

$$\mu_1^n(R_n^c)\le\exp\Big[-n\min_{\nu:\ D(\nu\,|\,\mu_0)\le a}D(\nu\,|\,\mu_1)\Big]$$

with optimal exponent.

Hoeffding's original paper is [16]; see also p. 44 of [5] or Theorem 3.5.4 of [6]. It is remarkable that the asymptotically optimal tests R_n do not depend on the alternative μ_1. One should note that K. Pearson's well-known χ^2-test for the parameter of a multinomial distribution (see e.g. [31]) uses a rejection region similar to R_n, the relative entropy $D(L_n^\omega\,|\,\mu_0)$ being replaced by a quadratic approximation.

Hoeffding's theorem is in fact an immediate consequence of a much more fundamental result, the theorem of Sanov. This elucidates the role of relative entropy for

the asymptotic behavior of the empirical distributions L_n^ω. The basic observation is the identity

$$\mu^n(\omega) = \exp\Big[- n\Big(D(L_n^\omega \,|\, \mu) + H(L_n^\omega)\Big)\Big] \tag{3.3}$$

which holds for any probability measure μ on E and any $\omega \in E^n$. In view of our first assertion (1.3), it follows that

$$\frac{1}{n} \log \mu^n(\omega \in E^n : L_n^\omega = \nu_n) \to -D(\nu \,|\, \mu)$$

whenever $\nu_n \to \nu$ such that $n\nu_n(x) \in \mathbb{Z}$ for all x and n. This can be viewed as a version of Boltzmann's formula (1.1) and leads directly to the following theorem due to Sanov [34], cf. also p. 43 of [5] or Theorem 2.1.10 of [6].

(3.4) Sanov's large deviation theorem: *Let μ be any probability measure on E and $\mathcal{C}$ a class of probability measures on E with dense (relative) interior, i.e., $\mathcal{C} \subset \text{cl int } \mathcal{C}$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log \mu^n(\omega \in E^n : L_n^\omega \in \mathcal{C}) = - \inf_{\nu\in\mathcal{C}} D(\nu \,|\, \mu) \; .$$

Sanov's theorem provides just a glimpse into large deviation theory in which (relative) entropies of various kinds play a central role. (More on this can be found in [6] and the contributions of den Hollander and Varadhan to this volume.) Its meaning can be summarized as follows:

> *Among all realizations with histogram in $\mathcal{C}$, the most probable are those having a histogram closest to μ in the sense of relative entropy.*

We will return to this point later in (5.3). Needless to say, Sanov's theorem can be extended to quite general state spaces E, see [4] or Theorem 6.2.10 of [6].

4 Entropy maximization under constraints

The second law of thermodynamics asserts that a physical system in equilibrium has maximal entropy among all states with the same energy. Translating this into a probabilistic language and replacing entropy by the more general relative entropy, we are led to the following question:

> *Let $\mathcal{C}$ be a class of probability measures on some measurable space $(E, \mathcal{E})$ and μ a fixed reference measure on $(E, \mathcal{E})$. What are then the probability measures in $\mathcal{C}$ minimizing $D(\,\cdot\,|\,\mu)$?*

The universal significance of such minimizers has been put forward by Jaynes [19, 20]. As noticed above, they come up also in the context of Sanov's theorem (3.4). In the present more general setting, the relative entropy can be defined by $D(\nu \,|\, \mu) = \sup_{\mathcal{P}} D(\nu_{\mathcal{P}} \,|\, \mu_{\mathcal{P}})$, where the supremum extends over all finite $\mathcal{E}$-measurable partitions $\mathcal{P}$ and $\nu_{\mathcal{P}}$ stands for the restriction of ν to $\mathcal{P}$. Equivalently, $D(\nu \,|\, \mu) = \nu(\log f)$ if ν is absolutely continuous with respect to μ with density f, and $D(\nu \,|\, \mu) = \infty$ otherwise; see Corollary (15.7) of [13], for example. (For a third expression see (4.1) below.) The first definition shows

in particular that $D(\,\cdot\,|\,\mu)$ is lower semicontinuous in the so-called τ-topology generated by the mappings $\nu \to \nu(A)$ with $A \in \mathcal{E}$. Consequently, a minimizer does exist whenever $\mathcal{C}$ is closed in this topology. If $\mathcal{C}$ is also convex, the minimizer is uniquely determined due to the strict convexity of $D(\,\cdot\,|\,\mu)$, and is then called the I-projection of μ on $\mathcal{C}$. We consider here only the most classical case when $\mathcal{C}$ is defined by an integral constraint. That is, writing $\nu(g)$ for the integral of some bounded measurable function $g : E \to \mathbb{R}^d$ with respect to ν, we assume that $\mathcal{C} = \{\nu : \nu(g) = a\}$ for suitable $a \in \mathbb{R}^d$. In other words, we consider the constrained variational problem

$$D(\nu\,|\,\mu) \stackrel{!}{=} \min\,, \quad \nu(g) = a\;.$$

In this case one can use a convex Lagrange multiplier calculus as follows.

For any bounded measurable function $f : E \to \mathbb{R}$ let $P(f) = \log \mu(e^f)$ be the log-Laplace functional of μ. One then has the variational formula

$$D(\nu\,|\,\mu) = \sup_f [\,\nu(f) - P(f)\,]\;, \tag{4.1}$$

meaning that $D(\,\cdot\,|\,\mu)$ and P are convex conjugates (i.e., Legendre-Fenchel transforms) of each other; cf. Lemma 6.2.13 of [6]. Let

$$J_g(a) = \inf_{\nu:\ \nu(g)=a} D(\nu\,|\,\mu)$$

be the "entropy distance" of $\{\nu : \nu(g) = a\}$ from μ. A little convex analysis then shows that

$$J_g(a) = \sup_{t\in\mathbb{R}^d} [t \cdot a - P(t \cdot g)]\;, \tag{4.2}$$

i.e., J_g is a partial convex conjugate of P (or, in other terms, the Cramér transform of the distribution $\mu \circ g^{-1}$ of g under μ). Moreover, if g is non-degenerate (in the sense that $\mu \circ g^{-1}$ is not supported on a hyperplane) then J_g is differentiable on the interior $I_g = \mathrm{int}\{J_g < \infty\}$ of its essential domain, and one arrives at the following

(4.3) Gibbs-Jaynes principle: *For any non-degenerate $g : E \to \mathbb{R}^d$, $a \in I_g$ and $t = \nabla J_g(a)$, the probability measure*

$$\mu_t(dx) = Z_t^{-1}\, e^{t\cdot g(x)}\, \mu(dx)$$

on $(E, \mathcal{E})$ is the unique minimizer of $D(\,\cdot\,|\,\mu)$ on $\{\nu :\ \nu(g) = a\}$. Here $Z_t = e^{P(t\cdot g)}$ is the normalizing constant.

Generalized versions of this result can be found in [3, 4] or Example (9.42) of [37].

In Statistical Mechanics, the measures μ_t of the form above are called *Gibbs distributions*, and the preceding result (or a suitable extension) justifies that these are indeed the equilibrium distributions of physical systems satisfying a finite number of conservation laws. In Mathematical Statistics, such classes of probability measures are called *exponential families*. Here are some familiar examples from probability theory.

(4.4) Example: Let $E = \mathbb{R}^d$, μ the standard normal distribution on E, and for any positive definite symmetric matrix C let μ_C be the centered normal distribution with covariance matrix $\mathrm{Cov}(\mu_C) = C$. Then μ_C minimizes the relative entropy $D(\,\cdot\,|\,\mu)$ among

all centered distributions ν with covariance matrix C. Equivalently, μ_C maximizes the differential entropy

$$H(\nu) = \begin{cases} -\int dx\ f(x)\ \log\ f(x) & \text{if } \nu \text{ has Lebesgue-density } f\ , \\ -\infty & \text{otherwise} \end{cases}$$

in the same class of distributions, and $H(\mu_C) = \frac{d}{2} \log[2\pi e (\det C)^{1/d}]$.

(4.5) Example: Let $E = [0, \infty[$, $a > 0$ and μ_a be the exponential distribution with parameter a. Then $\nu = \mu_a$ minimizes $D(\nu \,|\, \mu_1)$ resp. maximizes $H(\nu)$ under the condition $\nu(\mathrm{id}) = 1/a$.

(4.6) Example: Let $E = \mathbb{N}$, $a > 0$ and μ_a be the Poisson distribution with parameter a. Then $\nu = \mu_a$ minimizes $D(\nu \,|\, \mu_1)$ under the condition $\nu(\mathrm{id}) = a$, and $D(\mu_a \,|\, \mu_1) = 1 - a + a \log a$.

(4.7) Example: Let $E = C[0,1]$, $a \in E$ and μ_a the image of the Wiener measure $\mu_0 = \mu$ under the shift $x \to x + a$ of E. Then $\nu = \mu_a$ minimizes $D(\nu \,|\, \mu)$ under the condition $\nu(\mathrm{id}) = a$, and $D(\mu_a \,|\, \mu) = \frac{1}{2} \int_0^1 \dot{a}(t)^2\, dt$ if a is absolutely continuous with derivative $\dot{a}$, and $D(\mu_a \,|\, \mu) = \infty$ otherwise; see e.g. Section II.1.1 of [11].

5 Asymptotics governed by entropy

We will now turn to the dynamical aspects of the second law of thermodynamics. As before, we will not enter into a physical discussion of this fundamental law. Rather we will show by examples that the principle of increasing entropy (or decreasing relative entropy) stands also behind a number of well-known facts of probability theory.

Our first example is the so-called ergodic theorem for Markov chains. Let E be a finite set and $P_t = e^{tG}$, $t \geq 0$, the transition semigroup for a continuous-time Markov chain on E. The generator G is assumed to be irreducible. It is well-known that there is then a unique invariant distribution μ (satisfying $\mu P_t = \mu$ for all $t \geq 0$ and, equivalently, $\mu G = 0$). Let ν be any initial distribution, and $\nu_t = \nu P_t$ be the distribution at time t. Consider the relative entropy $D(\nu_t \,|\, \mu)$ as a function of time $t \geq 0$. A short computation (using the identities $\mu G = 0$ and $G1 = 0$) then gives the following result:

(5.1) Entropy production of Markov chains: *For any $t \geq 0$ we have*

$$\begin{aligned} \frac{d}{dt}\, D(\nu_t \,|\, \mu) &= -\sum_{x,y \in E:\, x \neq y} \nu_t(y)\, \bar{G}(y,x)\ \varphi\left(\frac{\nu_t(x)\, \mu(y)}{\mu(x)\, \nu_t(y)} \right) \\ &= -a(\nu_t)\ D(\tilde{\nu}_t \,|\, \bar{\nu}_t)\ \leq\ 0\ , \end{aligned}$$

and in particular $\frac{d}{dt}\, D(\nu_t \,|\, \mu) < 0$ when $\nu_t \neq \mu$.

In the above, $\bar{G}(y,x) = \mu(x)\, G(x,y)/\mu(y)$ is the generator for the time-reversed chain, $\varphi(s) = 1 - s + s \log s \geq 0$ for $s \geq 0$, $a(\nu) = -\sum_{x \in E} \nu(x)\, G(x,x) > 0$, and the probability measures $\tilde{\nu}$ and $\bar{\nu}$ on $E \times E$ are defined by $\tilde{\nu}(x,y) = \nu(x)\, G(x,y)(1 - \delta_{x,y})/a(\nu)$ and $\bar{\nu}(x,y) = \nu(y)\, \bar{G}(y,x)(1 - \delta_{x,y})/a(\nu)$, $x, y \in E$. The second statement follows from the

fact that 1 is the unique zero of φ, and G is irreducible. A detailed proof can be found in Chapter I of Spitzer [36]. The discrete time analogue was apparently discovered repeatedly by various authors; it appears e.g. in [32] and on p. 98 of [23].

The entropy production formula above states that the relative entropy $D(\cdot\,|\,\mu)$ is a strict Lyapunov function for the fixed-time distributions ν_t of the Markov chain. Hence $\nu_t \to \mu$ as $t \to \infty$. This is the well-known ergodic theorem for Markov chains, and the preceding argument shows that this convergence result fits precisely into the physical picture of convergence to equilibrium.

Although the central limit theorem is a cornerstone of probability theory, it is often not realized that this theorem is also an instance of the principle of increasing entropy. (This is certainly due to the fact that the standard proofs do not use this observation.) To see this, let (X_i) be a sequence of i.i.d. centered random vectors in $\mathbb{R}^d$ with existing covariance matrix C, and consider the normalized sums $S_n^* = \sum_{i=1}^n X_i/\sqrt{n}$. By the very definition, S_n^* is again centered with covariance matrix C. But, as we have seen in Example (4.4), under these conditions the centered normal distribution μ_C with covariance matrix C has maximal differential entropy. This observation suggests that the relative entropy may again serve as a Lyapunov function. Unfortunately, a time-monotonicity of relative entropies seems to be unknown so far (though monotonicity along the powers of 2 follows from a subadditivity property). But the following statement is true.

(5.2) Entropic central limit theorem: *Let ν_n be the distribution of S_n^*. If ν_1 is such that $D(\nu_n\,|\,\mu_C) < \infty$ for some n, then $D(\nu_n\,|\,\mu_C) \to 0$ as $n \to \infty$.*

This theorem traces back to Linnik [27], whose result was put on firm grounds by Barron [1]. The multivariate version above is due to [21]. By an inequality of Pinsker, Csiszár, Kullback and Kemperman (cf. p. 133 of [11] or p. 58 of [5]), it follows that $\nu_n \to \mu_C$ in total variation norm (which is equal to the L^1-distance of their densities).

A similar result holds for sums of i.i.d. random elements X_i of a compact group G. Let μ_G denote the normalized Haar measure on G, and let ν_n be the distribution of $\sum_{i=1}^n X_i$, i.e., the n-fold convolution of the common distribution of the X_i. A recent result of Johnson and Suhov [22] then implies that $D(\nu_n\,|\,\mu_G) \downarrow 0$ as $n \uparrow \infty$, provided $D(\nu_n\,|\,\mu_G)$ is ever finite. Note that μ_G is the measure of maximal entropy (certainly if G is finite or a torus), and that the convergence here is again monotone in time.

Our third example is intimately connected to Sanov's theorem (3.4). Suppose again (for simplicity) that E is finite, and let μ be a probability measure on E. Let $\mathcal{C}$ be a closed convex class of probability measures on E such that $\operatorname{int}\mathcal{C} \neq \emptyset$. We consider the conditional probability

$$\mu_{\mathcal{C}}^n = \mu^n(\,\cdot\,|\{\omega \in E^n : L_n^\omega \in \mathcal{C}\})$$

under the product measure μ^n given that the empirical distribution belongs to the class $\mathcal{C}$. (By Sanov's theorem, this condition has positive probability when n is large enough.) Do these conditional probabilities converge to a limit? According to the interpretation of Sanov's theorem, the most probable realizations ω are those for which $D(L_n^\omega\,|\,\mu)$ is as small as possible under the constraint $L_n^\omega \in \mathcal{C}$. But we have seen above that there exists a unique probability measure $\mu_* \in \mathcal{C}$ minimizing $D(\cdot\,|\,\mu)$, namely the I-projection of μ on $\mathcal{C}$. This suggests that, for large n, $\mu_{\mathcal{C}}^n$ concentrates on configurations ω for which L_n^ω is close to μ_*. This and even more is true, as was shown by Csiszár (5.3).

(5.3) Csiszár's conditional limit theorem: *For closed convex $\mathcal{C}$ with non-empty interior,*

$$\mu^n(\,\cdot\,|\{\omega \in E^n : L_n^\omega \in \mathcal{C}\}) \to \mu_*^{\mathbb{N}} \quad \text{as } n \to \infty\,,$$

where μ_* is the I-projection from μ on $\mathcal{C}$.

Note that the limit is again determined by the maximum entropy principle. It is remarkable that this result follows from purely entropic considerations. Writing $\nu_{\mathcal{C},n} = \mu_{\mathcal{C}}^n(L_n^\cdot)$ for the mean conditional empirical distribution (which by symmetry coincides with the one-dimensional marginal of $\mu_{\mathcal{C}}^n$), Csiszár (5.3) observes that

$$\begin{aligned} -\frac{1}{n}\log\mu^n(\omega \in E^n : L_n^\omega \in \mathcal{C}) &= \frac{1}{n}\,D(\mu_{\mathcal{C}}^n \,|\, \mu^n) \\ &= \frac{1}{n}\,D(\mu_{\mathcal{C}}^n \,|\, (\nu_{\mathcal{C},n})^n) + D(\nu_{\mathcal{C},n} \,|\, \mu) \\ &\geq \frac{1}{n}\,D(\mu_{\mathcal{C}}^n \,|\, \mu_*^n) + D(\mu_* \,|\, \mu)\,. \end{aligned}$$

The inequality can be derived from the facts that $\nu_{\mathcal{C},n} \in \mathcal{C}$ by convexity and μ_* is the I-projection of μ on $\mathcal{C}$. Now, by Sanov's theorem, the left-hand side tends to $D(\mu_* \,|\, \mu)$, whence $\frac{1}{n}\,D(\mu_{\mathcal{C}}^n \,|\, \mu_*^n) \to 0$. In view of the superadditivity properties of relative entropy, it follows that for each $k \geq 1$ the projection of $\mu_{\mathcal{C}}^n$ onto E^k converges to μ_*^k, and one arrives at (5.3).

The preceding argument is completely general: Csiszár's original paper [4] deals with the case when E is an arbitrary measurable space. In fact, some modifications of the argument even allow to replace the empirical distribution L_n^ω by the so-called empirical process; this will be discussed below in (6.7).

6 Entropy density of stationary processes and fields

Although occasionally we already considered sequences of i.i.d. random variables, our main concern so far was the entropy and relative entropy of (the distribution of) a single random variable with values in E. In this last section we will recall how the ideas described so far extend to the set-up of stationary stochastic processes, or stationary random fields, and our emphasis here is on the non-independent case.

Let E be a fixed state space. For simplicity we assume again that E is finite. We consider the product space $\Omega = E^{\mathbb{Z}^d}$ for any dimension $d \geq 1$. For $d = 1$, Ω is the path space of an E-valued process, while for larger dimensions Ω is the configuration space of an E-valued random field on the integer lattice. In each case, the process or field is determined by a probability measure μ on Ω. We will assume throughout that all processes or fields are *stationary* resp. *translation invariant*, in the sense that μ is invariant under the shift-group $(\vartheta_x)_{x\in\mathbb{Z}^d}$ acting on Ω in the obvious way.

In this setting it is natural to consider the entropy or relative entropy *per time* resp. *per lattice site*, rather than the (total) entropy or relative entropy. (In fact, $D(\nu|\mu)$ is infinite in all interesting cases.) The basic result on the existence of the entropy density is the following. In its statement, we write $\Lambda \uparrow \mathbb{Z}^d$ for the limit along an arbitrary increasing sequence of cubes exhausting Λ, μ_Λ for the projection of μ onto E^Λ, and ω_Λ for the restriction of $\omega \in \Omega$ to Λ.

(6.1) Shannon-McMillan-Breiman theorem: *For any stationary μ on Ω, there exists the entropy density*

$$\mathsf{h}(\mu) = \lim_{\Lambda\uparrow\mathbb{Z}^d} |\Lambda|^{-1}\, H(\mu_\Lambda)\;,$$

and for the integrands we have

$$-\lim_{\Lambda\uparrow\mathbb{Z}^d} |\Lambda|^{-1}\, \log \mu_\Lambda(\omega_\Lambda) = \mathsf{h}(\mu(\,\cdot\,|\,\mathcal{I})(\omega))$$

for μ-almost ω and in $L^1(\mu)$. Here $\mu(\,\cdot\,|\,\mathcal{I})(\omega)$ is a regular version of the conditional probability with respect to the σ-algebra $\mathcal{I}$ of shift-invariant events in Ω.

For a proof we refer to Section 15.2 of [13] (and the references therein), and Section I.3.1 of [11]. In the case of a homogeneous product measure $\mu = \alpha^{\mathbb{Z}^d}$ we have $\mathsf{h}(\mu) = H(\alpha)$.

In view of Boltzmann's interpretation (1.3) of entropy, $\mathsf{h}(\mu)$ is a measure of the lack of knowledge about the process or field per time resp. per site. Also, the L^1-convergence result of McMillan immediately implies an asymptotic equipartition property analogous to (2.3), whence $\mathsf{h}(\mu)$ is also the optimal rate of a block code, and thus the *information per signal* of the stationary source described by μ (provided we take the logarithm to the base 2).

What about the existence of a *relative* entropy per time or per site? Here we need to assume that the reference process has a nice dependence structure, which is also important in the context of the maximum entropy problem.

Let $f : \Omega \to \mathbb{R}$ be any function depending only on the coordinates in a finite subset Δ of $\mathbb{Z}^d$. Such a function will be called local. A probability measure μ on Ω is called a *Gibbs measure* for f if its conditional probabilities for observing a configuration ω_Λ in a finite region $\Lambda \subset \mathbb{Z}^d$, given a configuration ω_{Λ^c} outside of Λ, are almost surely given by the formula

$$\mu(\omega_\Lambda \,|\, \omega_{\Lambda^c}) = Z_\Lambda(\omega_{\Lambda^c})^{-1}\; \exp\Big[\sum_{x:\,(\Delta+x)\cap\Lambda\neq\emptyset} f(\vartheta_x\omega)\Big]\;,$$

where $Z_\Lambda(\omega_{\Lambda^c})$ is the normalization constant. Since f is local, each Gibbs measure μ is Markovian in the sense that the conditional probabilities above only depend on the restriction of ω_{Λ^c} to a bounded region around Λ. (This assumption of finite range could be weakened, but here is no place for this.) The main interest in Gibbs measures comes from its use for describing systems of interacting spins in equilibrium, and the analysis of phase transitions; a general account can be found in Georgii [13], for example. (To make the connection with the definition given there let the potential Φ be defined as in Lemma (16.10) of this reference.) In the present context, Gibbs measures simply show up because of their particular dependence properties. We now can state the following counterpart to (6.1).

(6.2) Ruelle-Föllmer theorem: *Suppose μ is a Gibbs measure for some local function f, and ν is translation invariant. Then the relative entropy density*

$$\mathsf{d}(\nu \,|\, \mu) = \lim_{\Lambda\uparrow\mathbb{Z}^d} |\Lambda|^{-1}\, D(\nu_\Lambda \,|\, \mu_\Lambda) \tag{6.3}$$

exists and is equal to $\mathsf{p}(f) - \mathsf{h}(\nu) - \nu(f)$*, where*

$$\mathsf{p}(f) = \max_{\nu}[\mathsf{h}(\nu) + \nu(f)] = \lim_{\Lambda \uparrow \mathbb{Z}^d} |\Lambda|^{-1} \log Z_\Lambda(\omega_{\Lambda^c}) \,, \tag{6.4}$$

the so-called pressure of f*, is the counterpart of the log-Laplace functional appearing in (4.3).*

The second identity in (6.4) is often called the *variational formula*; it dates back to Ruelle [33]. Föllmer [10] made the connection with relative entropy; for a detailed account see also Theorem (15.30) of [13] or Section I.3.3 of [11]. An example of a non-Gibbsian μ for which $\mathsf{d}(\cdot \,|\, \mu)$ fails to exist was constructed by Kieffer and Sokal, see pp. 1092–1095 of [9]. As in (6.1), there is again an $L^1(\nu)$ and ν-almost sure convergence behind (6.3) [10]. In the case $f = 0$ when the unique Gibbs measure μ is equal to $\alpha^{\mathbb{Z}^d}$ for the equidistribution α on E, the Ruelle-Föllmer theorem (6.2) reduces to (6.1).

Since $D(\nu \,|\, \mu) = 0$ if and only if $\nu = \mu$, the preceding result leads us to ask what one can conclude from the identity $\mathsf{d}(\nu \,|\, \mu) = 0$. The answer is the following celebrated variational characterization of Gibbs measures first derived by Lanford and Ruelle [25]. Simpler proofs were given later by Föllmer [10] and Preston, Theorem 7.1 of [30]; cf. also Section 15.4 of [13], or Theorem (I.3.39) of [11].

(6.5) Variational principle: *Suppose* ν *is stationary. Then* ν *is a Gibbs measure for* f *if and only if* $\mathsf{h}(\nu) + \nu(f)$ *is equal to its maximum value* $\mathsf{p}(f)$*.*

Physically speaking, this result means that the stationary Gibbs measures are the minimizers of the free energy density $\nu(-f) - \mathsf{h}(\nu)$, and therefore describe a physical system with interaction f in thermodynamic equilibrium.

It is now easy to obtain an analogue of the Gibbs-Jaynes principle (4.3). Let $g : \Omega \to \mathbb{R}^d$ be any vector-valued local function whose range $g(\Omega)$ is not contained in a hyperplane. Then for all $a \in \mathbb{R}^d$ we have in analogy to (4.2)

$$\mathsf{j}_g(a) \equiv - \sup_{\nu:\ \nu(g)=a} \mathsf{h}(\nu) = \sup_{t \in \mathbb{R}^d} [t \cdot a - \mathsf{p}(t \cdot g)] \,,$$

which together with (6.5) gives us the following result, cf. Section 4.3 of [14].

(6.6) Gibbs-Jaynes principle for the entropy density: *Suppose* $a \in \mathbb{R}^d$ *is such that* j_g *is finite on a neighborhood of* a*, and let* ν *be translation invariant. Then* $\mathsf{h}(\nu)$ *is maximal under the constraint* $\nu(g) = a$ *if and only if* ν *is a Gibbs measure for* $t_a \cdot g$*, where* $t_a = \nabla \mathsf{j}_g(a)$*.*

The next topic to be discussed is the convergence to stationary measures of maximal entropy density. The preceding Gibbs–Jaynes principle suggests that an analogue of Csiszár's conditional limit theorem (5.3) might hold in the present setting. This is indeed the case, as was proved by Deuschel-Stroock-Zessin [8], Georgii [14], and Lewis-Pfister-Sullivan [26] using suitable extensions of Sanov's theorem (3.4). We state the result only in the most interesting particular case.

(6.7) The equivalence of ensembles: *Let* $\mathcal{C} \subset \mathbb{R}^d$ *be closed and such that*

$$\inf \mathsf{j}_g(\mathcal{C}) = \inf \mathsf{j}_g(\operatorname{int} \mathcal{C}) = \mathsf{j}_g(a)$$

for a unique $a \in \mathcal{C}$ having the same property as in (6.6). For any cube Λ in $\mathbb{Z}^d$ let $\nu_{\Lambda,\mathcal{C}}$ be the uniform distribution on the set

$$\left\{\omega \in E^\Lambda : \ |\Lambda|^{-1} \sum_{x\in\Lambda} g(\vartheta_x^{per}\omega_\Lambda) \in \mathcal{C}\right\},$$

where ϑ_x^{per} is the periodic shift of E^Λ defined by viewing Λ as a torus. (The assumptions imply that this set is non-empty when Λ is large enough.) Then, as $\Lambda \uparrow \mathbb{Z}^d$, each (weak) limit point of the measures $\nu_{\Lambda,\mathcal{C}}$ is a Gibbs measure for $t_a \cdot g$.

In Statistical Mechanics, the equidistributions of the type $\nu_{\Lambda,\mathcal{C}}$ are called *microcanonical Gibbs distributions*, and "equivalence of ensembles" is the classical term for their asymptotic equivalence with the (grand canonical) Gibbs distributions considered before. A similar result holds also in the context of point processes, and thus applies to the classical physical models of interacting molecules [15].

Finally, we want to mention that the entropy approach (5.1) to the convergence of finite-state Markov chains can also be used for the time-evolution of translation invariant random fields. For simplicity let $E = \{0,1\}$ and thus $\Omega = \{0,1\}^{\mathbb{Z}^d}$. We define two types of continuous-time Markov processes on Ω which admit the Gibbs measures for a given f as reversible measures. These are defined by their pregenerator G acting on local functions g as

$$Gg(\omega) = \sum_{x\in\mathbb{Z}^d} c(x,\omega)[g(\omega^x) - g(\omega)]$$

or

$$Gg(\omega) = \sum_{x,y\in\mathbb{Z}^d} c(xy,\omega)[g(\omega^{xy}) - g(\omega)] \ ,$$

respectively. Here $\omega^x \in \Omega$ is defined by $\omega_x^x = 1 - \omega_x$, $\omega_y^x = \omega_y$ for $y \neq x$, and ω^{xy} is the configuration in which the values at x and y are interchanged. Under mild locality conditions on the rate function c the corresponding Markov processes are uniquely defined. They are called *spin-flip or Glauber processes* in the first case, and *exclusion or Kawasaki processes* in the second case. The Gibbs measures for f are reversible stationary measures for these processes as soon as the rate function satisfies the detailed balance condition that $c(x,\omega) = \exp[\sum_{z:\, x\in\Delta+z} f(\vartheta_z\omega)]$ does not depend on ω_x, resp. an analogous condition in the second case. The following theorem is due to Holley [17, 18]; for streamlined proofs and extensions see [29, 12, 38].

(6.8) Holley's theorem: *For any translation-invariant initial distribution ν on Ω, the negative free energy $\mathsf{h}(\nu_t) + \nu_t(f)$ is strictly increasing in t as long as the time-t distribution ν_t is no Gibbs measure for f. In particular, ν_t converges to the set of Gibbs measures for f.*

This result is just another instance of the principle of increasing entropy. For similar results in the non-reversible case see [24, 28] and the contribution of C. Maes to this volume.

Let me conclude by noting that the results and concepts of this section serve also as a *paradigm for ergodic theory.* The set Ω is then replaced by an arbitrary compact metric space with a μ-preserving continuous $\mathbb{Z}^d$-action $(\vartheta_x)_{x\in\mathbb{Z}^d}$. The events in a set $\Lambda \subset \mathbb{Z}^d$ are those generated by the transformations $(\vartheta_x)_{x\in\Lambda}$ from a generating partition of Ω.

The entropy density $\mathsf{h}(\mu)$ then becomes the well-known Kolmogorov-Sinai entropy of the dynamical system $(\mu, (\vartheta_x)_{x\in\mathbb{Z}^d})$. Again, $\mathsf{h}(\mu)$ can be viewed as a measure of the inherent randomness of the dynamical system, and its significance comes from the fact that it is invariant under isomorphisms of dynamical systems. Measures of maximal Kolmogorov-Sinai entropy play again a key role. It is quite remarkable that the variational formula (6.4) holds also in this general setting, provided the partition functions $Z_\Lambda(\omega_{\Lambda^c})$ are properly defined in terms of f and the topology of Ω. $\mathsf{p}(f)$ is then called the topological pressure, and $p(0)$ is the so-called topological entropy describing the randomness of the action $(\vartheta_x)_{x\in\mathbb{Z}^d}$ in purely topological terms. All this is discussed in more detail in the contributions by Keane and Young to this volume.

Acknowledgment. I am grateful to A. van Enter, O. Johnson, H. Spohn and R. Lang for a number of comments on a preliminary draft.

References

[1] Barron, A.R. (1986) Entropy and the central limit theorem, *Ann. Prob.* **14**, 336–342.

[2] Chernoff, H. (1956) Large sample theory: parametric case, *Ann. Math. Stat.* **23**, 493–507.

[3] Csiszár, I. (1975) I-divergence geometry of probability distributions and minimization problems, *Ann. Prob.* **3**, 146–158.

[4] Csiszár, I. (1984) Sanov property, generalized I-projection and a conditional limit theorem *Ann. Prob.* **12**, 768–793.

[5] Csiszár, I. and Körner, J. (1981) *Information Theory, Coding Theorems for Discrete Memoryless Systems*, Akdaémiai Kiadó, Budapest.

[6] Dembo, A. and Zeitouni, O. (1993) *Large Deviation Techniques and Applications*, Jones and Bartlett, Boston London.

[7] Denbigh, K. (1981) How subjective is entropy, in: Leff, H.S. and Rex, A.F. (eds.), *Maxwell's demon, entropy, information, computing*, Princeton University Press, Princeton, 1990, pp. 109–115.

[8] Deuschel, J.-D., Stroock, D.W. and Zessin, H. (1991) Microcanonical distributions for lattice gases, *Commun. Math. Phys.* **139**, 83–101.

[9] van Enter, A.C.D., Fernandez, R. and Sokal A.D. (1993) Regularity properties and pathologies of position-space renormalization-group transformations: scope and limitations of Gibbsian theory, *J. Stat. Phys.* **72**, 879–1167.

[10] Föllmer, H. (1973) On entropy and information gain in random fields, *Z. Wahrscheinlichkeitstheorie verw. Geb.* **26**, 207–217.

[11] Föllmer, H. (1988) *Random Fields and Diffusion Processes*, in: P.L. Hennequin (ed.), École d'Été de Probabilités de Saint Flour XV–XVII, 1985–87, Lecture Notes in Mathematics Vol. 1362, Springer, Berlin etc., pp. 101–203.

[12] Georgii, H.-O. (1979) *Canonical Gibbs Measures*, Lecture Notes in Mathematics Vol. 760, Springer, Berlin Heidelberg New York.

[13] Georgii, H.-O. (1988) *Gibbs Measures and Phase Transitions*, de Gruyter, Berlin New York.

[14] Georgii, H.-O. (1993) Large deviations and maximum entropy principle for interacting random fields on $\mathbb{Z}^d$, *Ann. Prob.* **21**, 1845–1875.

[15] Georgii, H.-O. (1995) The equivalence of ensembles for classical systems of particles, *J. Statist. Phys.* **80**, 1341–1378.

[16] Hoeffding, W. (1965) Asymptotically optimal tests for multinomial distributions, *Ann. Math. Stat.* **36**, 369–400.

[17] Holley, R. (1971) Pressure and Helmholtz free energy in a dynamic model of a lattice gas, *Proc. Sixth Berkeley Symp. Prob. math. Statist.* **3**, 565–578.

[18] Holley, R. (1971) Free energy in a Markovian model of a lattice spin system, *Commun. Math. Phys.* **23**, 87–99.

[19] Jaynes, E.T. (1957) Information theory and statistical mechanics, *Phys. Rev.* **106**, 620–630 & **108**, 171–190.

[20] Jaynes, E.T. (1982) On the rationale of maximum entropy methods, *Proc. IEEE* **70**, 939–952.

[21] Johnson, O. (2000) Entropy inequalities and the central limit theorem, *Stoch. Proc. Appl.* **88**, 291–304.

[22] Johnson, O. and Suhov, Y. (2000) Entropy and convergence on compact groups, *J. Theor. Prob.* **13**, 843–857.

[23] Kac, M. (1959) *Probability and related topics in physical sciences*, Lectures in Applied Mathematics. Proceedings of the Summer Seminar, Boulder, Colo., 1957, Vol. I, Interscience Publishers, London New York.

[24] Künsch, H. (1984) Nonreversible stationary measures for infinite interacting particle systems, *Z. Wahrscheinlichkeitstheorie verw. Geb.* **66**, 407–424.

[25] Lanford, O. and Ruelle, D. (1969) Observables at infinity and states with short range correlations in statistical mechanics, *Commun. Math. Phys.* **13**, 194–215.

[26] Lewis, J.T., Pfister, C.-E. and Sullivan, W.G. (1995) Entropy, concentration of probability and conditional limit theorems, *Markov Proc. Rel. Fields* **1**, 319–386.

[27] Linnik, Yu.V. (1959) An information-theoretic proof of the central limit theorem with the Lindeberg condition, *Theor. Prob. Appl.* **4**, 288–299.

[28] Maes, C., Redig F. and van Moffaert, A. (2000) On the definition of entropy production via examples, *J. Math. Phys.* **41**, 1528–1554.

[29] Moulin-Ollagnier, J. and Pinchon, D. (1977) Free energy in spin-flip processes is non-increasing, *Commun. Math. Phys.* **55**, 29–35.

[30] Preston, C.J. (1976) *Random Fields*, Lect. Notes in Math. **534**, Springer–Verlag, Berlin.

[31] Rao, C.R. (1973) *Linear Statistical Inference and its Applications*, Wiley, New York London Sydney.

[32] Renyi, A. (1970) *Probability Theory*, North-Holland, Amsterdam.

[33] Ruelle, D. (1967) A variational formulation of equilibrium statistical mechanics and the Gibbs phase rule, *Commun. Math. Phys.* **5**, 324–329.

[34] Sanov, I.N. (1957) On the probability of large deviations of random variables (in Russian), *Mat. Sbornik* **42**, 11–44; English translation in: *Selected Translations in Mathematical Statistics and Probability I* (1961), pp. 213–244.

[35] Shannon, C.E. (1948) A mathematical theory of communication, *Bell System Techn. J.* **27**, 379–423, 623–657. Reprinted in: D. Slepian (ed.), Key papers in the development of information theory, IEEE Press, New York 1974.

[36] Spitzer, F. (1971) *Random Fields and Interacting Particle Systems*, Notes on lectures given at the 1971 MAA Summer Seminar Williamstown, Massachusetts, Mathematical Association of America.

[37] Vajda, I. (1989) *Theory of Statistical Inference and Information*, Kluwer, Dordrecht.

[38] Wick, W.D. (1982) Monotonicity of the free energy in the stochastic Heisenberg model, *Commun. Math. Phys.* **83**, 107–122.