J. Math. Soc. Japan
Vol. 67, No. 4 (2015) pp. 1845–1857
doi: 10.2969/jmsj/06741845

Entropy and its many Avatars

By Srinivasa R. S. Varadhan

Dedicated to the memory of Professor Kiyosi Itô

(Received Mar. 24, 2015)
(Revised May 9, 2015)

Abstract. Entropy was first introduced in 1865 by Rudolf Clausius in his study of the connection between work and heat. A mathematical definition was given by Boltzmann as the logarithm of the number of micro states that corresponds to a macro state. It plays important roles in statistical mechanics, in the theory of large deviations in probability, as an invariant in ergodic theory and as a useful tool in communication theory. This article explores some of the connections between these different contexts.

1. Introduction.

The concept of entropy appears in many different contexts, in many different forms and is used in many different ways. The earliest appearance of the term seems to be in the work of Rudolf Clausius [**3**] in connection with his development of classical theory of thermodynamics. Its increase was viewed as a loss due to inefficiency in the conversion of heat to work. It was a property defined in the bulk. Boltzmann [**2**] later gave a microscopic definition of entropy. If S is a macro state that consists of a set $\Omega = \{s\}$ of micro states then the Boltzmann entropy is defined as

$$H(S) = k \log |\Omega|$$

where $|\Omega|$ is the size of Ω measured as the number of states $s \in \Omega$ in the discrete case or the volume of Ω in the continuous case. This plays an important role in statistical mechanics on which the modern theory of thermodynamics is based.

2. Entropy and Information theory.

In Shannon's theory of communication [**15**] the notion of entropy plays a central role. If $\boldsymbol{p} = \{p_1, p_2, \ldots, p_k\}$ is a probability distribution on a set of k points, i.e. $p_i \geq 0$ and $\sum_i p_i = 1$, then Shannon's entropy is

$$h(\boldsymbol{p}) = -\sum_{i=1}^{k} p_i \log p_i.$$

2010 *Mathematics Subject Classification.* Primary 60-02; Secondary 37A35, 94A17.
Key Words and Phrases. entropy, information, large deviations.
The research was partially supported by NSF grant DMS-1208334.

There are many axiomatic derivations of this (see [**1**] for instance) based on properties that a good entropy function, defined for all probability distributions on finite sets of points should have. However it is more interesting to see the connection with Boltzmann's definition. One can think of a large collection of N objects. Each object can be one of k possible types. The micro states are the set of functions

$$t : \{1, 2, \dots, N\} \to \{1, 2, \dots, k\}$$

specifying the type of the object i as $t(i)$. The macro state $\boldsymbol{p} = \{p_1, \dots, p_k\}$ corresponds to the collection $\Omega_{\boldsymbol{p}}$ of micro states $\{t\}$ where the proportion of objects of type j is roughly p_j. Then by the multinomial formula, we have

$$|\Omega_{\boldsymbol{p}}| \simeq \frac{N!}{(Np_1)!(Np_2)! \cdots (Np_k)!}.$$

One can use the approximation for factorials provided by Stirling's formula

$$\log N! = \left(N + \frac{1}{2}\right) \log N - N + \frac{1}{2} \log(2\pi) + o(1)$$

to conclude that

$$\log |\Omega_{\boldsymbol{p}}| = -N \sum_{i=1}^{k} p_i \log p_i + o(N) = Nh(\boldsymbol{p}) + o(N)$$

establishing a direct connection between Shannon's entropy and Boltzmann's entropy.

There is also the notion of conditional entropy. Let $X : \Omega \to \{1, 2, \dots, k\}$ be a random variable on $(\Omega, \mathcal{F}, P)$ and $\Sigma \subset \mathcal{F}$ be a sub σ-field. The conditional probability $p(i, \omega) = P[X = i|\Sigma]$ is the probability that $X = i$ given Σ. The conditional entropy is given by

$$h(\boldsymbol{p}(\omega)) = -\sum_{i=1}^{k} p(i, \omega) \log p(i, \omega)$$

and the average conditional entropy given by

$$E^P[h(\boldsymbol{p}(\omega))].$$

The unconditional probability is $\boldsymbol{p} = E[\boldsymbol{p}(\omega)]$ and the convexity of the function $x \log x$ and Jensen's inequality imply that

$$h(\boldsymbol{p}) \geq E[h(\boldsymbol{p}(\omega)].$$

If F and G are two finite sets and $r(x, y) = p(x)q(y|x)$ is a probability distribution $\boldsymbol{r}$ on $F \times G$, with marginal $\boldsymbol{p} = p(\cdot)$ and conditional $\boldsymbol{q}_x = q(\cdot|x)$, it is easy to see that

$$h(\boldsymbol{r}) = h(\boldsymbol{p}) + E^{\boldsymbol{p}}[h(\boldsymbol{q}_x)].$$

Let us suppose that we have a stationary stochastic process $\{X_n\}$ with values from a finite set F. Denoting the distribution of the process by P, for any finite n, we have the joint distribution $\boldsymbol{p}_n$ of $X_1, \ldots, X_n$ on F^n and the corresponding entropy

$$H_n(P) = h(\boldsymbol{p}_n) = -\sum p_n(x_1, \ldots, x_n) \log p_n(x_1, \ldots, x_n)$$

where $p_n(x_1, \ldots, x_n) = P[X_1 = x_1, \ldots, X_n = x_n]$. It follows from the properties of the entropy function $h(\boldsymbol{p})$, that

$$H_{n+1}(P) \geq H_n(P); \qquad H_{n+m}(P) \leq H_n(P) + H_m(P)$$

and in fact

$$H_{n+1}(P) - H_n(P) \leq H_n(P) - H_{n-1}(P).$$

Therefore the limit

$$\lim_{n\to\infty} H_{n+1} - H_n(P) = \lim_{n\to\infty} \frac{H_n(P)}{N} = H(P)$$

exists and defines the entropy rate of the process. It is not difficult to see that

$$H(P) = E^P[h(\boldsymbol{p}(\omega))] = E^P\left[-\sum_{x\in F} p(x|\omega)\log p(x|\omega)\right]$$

where Σ is the σ-field of past history generated by $X_0, X_{-1}, \ldots$ and

$$p(x|\omega) = P[X_1 = x|\Sigma].$$

If P is a product measure with marginal $\boldsymbol{p}$, then

$$H(P) = h(\boldsymbol{p}).$$

If it is Markov with transition probability $q(y|x)$ and invariant distribution $\boldsymbol{p}$ then

$$H(P) = -\sum_{x,y} p(x)q(y|x)\log q(y|x).$$

Entropy plays an important role in measuring ones ability to compress data by coding. We want to code an incoming data stream distributed according to P in the alphabet F of size k into words from an alphabet G of size r. Although the number of possible incoming words of length n is k^n, a small fraction of them will carry most of the probability. Their number is roughly $e^{nH(P)}$, and they can be coded in to $e^{m \log r}$ words of length

$m = n\,H(P)/\log r$, because $e^{n(H(P)+o(1))} = e^{m \log r}$.

What makes it possible is the following theorem of Shannon, Breiman and McMillan: let P be the distribution of a stationary ergodic stochastic process. Given any $\epsilon > 0$, for sufficiently large n, we can find a set of $e^{nH(P)+o(n)}$ words of length n, each having, under the n dimensional joint distribution of P, probability $e^{-nH(P)+o(n)}$ and carrying a total probability of at least $1-\epsilon$.

A coding theorem due to Feinstein tells us how the rate at which one can transmit messages with a small probability of error through a noisy channel also involves entropy. A noisy stationary channel $\nu(\boldsymbol{x}, d\boldsymbol{y})$ is specified by the distribution of the output sequence $\boldsymbol{y} = \{y_j\}$ given that the input sequence was $\boldsymbol{x} = \{x_i\}$ and satisfies $\nu(\boldsymbol{x}, A) = \nu(T\boldsymbol{x}, TA)$ where T is the shift in the space of sequences. Let us suppose for simplicity that $\nu(\boldsymbol{x}, A)$ is such if A is measurable with respect to the σ-field generated by $y_1, \ldots, y_k$ then $\nu(\boldsymbol{x}, A)$ depends only on $x_1, x_2, \ldots x_k$. If N is the length of an input signal, we want to find k_N input sequences of length N and K_N mutually disjoint subsets in output space $\{y_1, \ldots, y_N\}$ such that $\nu(\boldsymbol{x}_i, A_i) \geq 1-\epsilon$. The maximum number k_N that we can find determines the capacity C of the channel

$$C = \lim_{N\to\infty} \frac{1}{N} \log k_N.$$

Feinstein's theorem states that under certain assumptions on the channel ν, C is given by

$$\sup_{\alpha}[H(\alpha) + H(\beta) - H(\gamma)].$$

Here the sup is taken over all stationary processes α and for any stationary process α, $\gamma(d\boldsymbol{x}, d\boldsymbol{y}) = \alpha(d\boldsymbol{x})\nu(\boldsymbol{x}, d\boldsymbol{y})$ and β is the marginal of $\boldsymbol{y}$. [**7**] and [**10**] are good resources for this material.

3. Entropy and dynamical systems.

A dynamical system is a measure space (Ω, Σ, P) with a measure preserving invertible transformation $T : \Omega \to \Omega$. Two dynamical systems (Ω, Σ, P, T) and $(\Omega', \Sigma', P', T')$ are isomorphic if there is a one to one map S of the measure space (Ω, Σ, P) onto (Ω', Σ', P') that intertwines T and T', i.e. $ST = T'S$. The measure preserving transformation generates a unitary map of $L_2(\Omega, \Sigma, P)$ and its spectral type is also invariant under isomorphisms. Kolmogorov [**11**] defined the entropy of a dynamical system in the following manner. Given any finite partition $\mathcal{P}$, i.e. a measurable map $f : \Omega \to F$ of Ω into a finite set F, a stationary stochastic process P_f can be generated by $X_n = f(T^n\omega)$. Its entropy can be calculated as $H(P_f)$ and the Kolmogorov entropy of (Ω, Σ, P, T) is defined as

$$H = \sup_{f}(H(P_f))$$

where the supremum is calculated over all maps f into finite sets.He proved that entropy is an invariant and that it is not determined by the spectrum. That left open the possibility that entropy could be a complete invariant, at least within a large class of dynamical systems. Sinai [**16**] proved a weak form of the converse. Ornstein later proved [**13**] that the equality of entropy was sufficient to establish an isomorphism under fairly general assumptions on the two systems. In particular if P_1 and P_2 are product measures on two product spaces F_1^∞ and F_2^∞ with marginals $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$, there exists a one to one measure preserving translation invariant mapping of $(F_1^\infty, \Sigma_1, P_1)$ onto $(F_2^\infty, \Sigma_2, P_2)$ if and only if

$$h(\boldsymbol{p}_1) = -\sum_{a\in F_1} p_1(a)\log p_1(a) = -\sum_{a\in F_2} p_2(a)\log p_2(a) = h(\boldsymbol{p}_2).$$

4. Relative entropy and large deviations.

Whereas the entropy $h(\boldsymbol{p})$ can be defined for discrete probability distributions, it is not clear how to define it more generally. A natural definition for a distribution given by a density $f(x)$ on $\mathbb{R}$ or $\mathbb{R}^d$ can be $h(f) = \int f(x)\log f(x)dx$. The Lebesgue measure is only defined up to a multiplicative constant which creates a mild ambiguity in $h(f)$. There is also the difference of the sign in front of the integral, when compared with Shannon's definition. If α is a singular distribution on $\mathbb{R}$ or a measure on some abstract space, it is not clear what one should do. It is important to note that there has always been a second measure in the background, i.e. the counting measure in the discrete case and the Lebesgue measure when we have a density. Relative entropy involves two probability measures α and β on the same measurable space (X, Σ) and it is given by

$$h(\beta : \alpha) = \int \log \frac{d\beta}{d\alpha}(x)\beta(dx) = \int \frac{d\beta}{d\alpha}(x)\log\frac{d\beta}{d\alpha}(x)\alpha(dx)$$

provided $\beta \ll \alpha$ and the Radom–Nikodym derivative $f(x) = (d\beta/d\alpha)(x)$ has the property

$$\int |\log f(x)|\beta(dx) = \int f(x)|\log f(x)|\alpha(dx) < \infty.$$

Otherwise $h(\beta : \alpha)$ is taken as $+\infty$. It is important to observe that on $X_- = \{x : f(x) < 1\}$ $\int_{X_-} |\log f(x)|f(x)\alpha(dx)$ is always finite, in fact bounded by e^{-1} and the divergence can arise only from large values of $f(x)$ or $\log^+ f(x)$. Shannon entropy and the relative entropy $h(\boldsymbol{p};\boldsymbol{q})$ where $\boldsymbol{q}$ is the uniform distribution on k points with $q_i = 1/k$ are related

$$h(\boldsymbol{p};\boldsymbol{q}) = \log k + \sum_i p_i \log p_i = h(\boldsymbol{q}) - h(\boldsymbol{p}).$$

Relative entropy or Kullback–Leibler number $h(\beta : \alpha)$ satisfies $h \geq 0$ and $h = 0$ if and only if $\alpha = \beta$. It can be thought of as measuring the distance from α to β. However it is not symmetric. If we have n independent observations from α then the empirical distribution $(1/n)\sum_{i=1}^n \delta_{x_i}$ is close to α with probability nearly 1 and the probability that it is close to a different distribution β is very small as $n \to \infty$. It is exponentially small in n, i.e. $\exp[-cn+o(n)]$ and the constant c depends on β and is in fact $c = h(\beta : \alpha)$. This was proved by Sanov [**14**].

The proper definition of Large Deviations for a sequence P_n of probability measures on a complete separable metric space X is the existence of a rate function $I(x) \geq 0$ such that for each $\ell < \infty$, the set $K_\ell = \{x : I(x) \leq \ell\}$ is compact and for any C closed

$$\limsup_{n\to\infty} \frac{1}{n} \log P_n(C) \leq - \inf_{x\in C} I(x) \tag{4.1}$$

and for any G open

$$\liminf_{n\to\infty} \frac{1}{n} \log P_n(G) \geq - \inf_{x\in G} I(x). \tag{4.2}$$

Such estimates are often obtained locally, i.e. one shows that

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \frac{1}{n} \log P_n(B(x,\delta)) = \lim_{\delta\to 0} \liminf_{n\to\infty} \frac{1}{n} \log P_n(B(x,\delta)) = -I(x).$$

From such an estimate the lower bound (4.2) would follow but the upper bound (4.1) can be shown only for compact sets by the standard covering argument. To obtain (4.1) for all closed sets one needs a companion estimate of the following form. Given any $\ell < \infty$ there is a compact set K_ℓ such that for any closed set $C \subset K_\ell^c$

$$\limsup_{n\to\infty} \frac{1}{n} \log P_n(C) \leq -\ell.$$

If we have a large deviation result for P_n with rate function $I(x)$ on some Polish space X and $f : X \to Y$ is a continuous map then $Q_n = P_n f^{-1}$ on Y satisfies a large deviation property with rate function

$$J(y) = \inf_{x:f(x)=y} I(x).$$

In this language, Sanov's theorem takes the following form. Let P_n be the distribution of the empirical distribution $(1/n)\sum_{i=1}^n \delta_{x_i}$ induced by the product measure $\Pi\,\alpha$ on X^∞ on the space $\mathcal{M}(X)$ of probability distributions on X, endowed with the topology of weak convergence. Then P_n satisfies a large deviation property with rate function $I(\beta) = h(\beta : \alpha)$.

The notion of relative entropy can be pushed to the process level [**6**]. Suppose P and Q are two stationary stochastic processes, i.e. two measures on the product space X^∞, then we saw earlier that we can consider the conditional distributions $p(dx|\omega)$ and $q(dx|\omega)$ of x_1 given the past history $\{x_i; i \leq 0\}$. The relative entropy $H(Q : P)$ can perhaps be defined as

$$H(Q;P) = E^Q[h(q(\cdot|\omega), p(\cdot|\omega))].$$

One would expect this to control the large deviation probability that the empirical distribution of not just the one dimensional marginals as in Sanov's theorem, but the probability that all the finite dimensional distributions look like they came from Q. It should be roughly $\exp[-nH(Q : P) + o(n)]$. But there is a serious problem. $p(\cdot|\omega)$ and $q(\cdot|\omega)$ are only defined almost everywhere with respect to P and Q respectively. But in general P and Q are orthogonal and one can not define $h(q(\cdot|\omega), p(\cdot|\omega))$ a.e. Q in order to be able to integrate. But often $p(\cdot|\omega)$ is globally defined. Like when P is a product measure p is independent of ω or when it is Markov with a well defined transition probability. In such situations $H(Q : P)$ is well defined and one can prove a large deviation result for the "empirical process". Technically the empirical process is a random stationary process obtained by placing equal probability of $1/n$ at each of the n points of the periodic orbit obtained by extending $(x_1, x_2, \ldots, x_n) \in X^n$ periodically in both directions.

In the end any large deviation from ergodicity is eventually a contraction of $H(Q : P)$. To illustrate this, Cramer's rate function [**4**] for sums of independent random variables with a common distribution α is given by

$$I(a) = \sup_\theta[\theta a - \log M(\theta)] \tag{4.3}$$

where

$$M(\theta) = \int \exp[\theta\, y] d\alpha(y).$$

One can verify that with $P = \Pi\, \alpha$

$$\inf_{Q:\int x_1 dQ = a} H(Q : P) = I(a)$$

and that the infimum is attained at $Q = \Pi\, \beta$ with $d\beta = (1/M(c))e^{c\,y}d\alpha$, c being the value of θ where the supremum is attained in (4.3).

If P is a Markov Chain on a finite state space X with transition probabilities $p(x, y)$ and invariant distribution $\alpha(x)$, one can ask for the analog of Sanov's theorem for the empirical distribution. From the contraction principle

$$I(\beta) = \inf_{Q \in \mathcal{M}(\beta)} H(Q : P)$$

where $\mathcal{M}(\beta)$ are stationary processes with marginal β. The infimum can be limited to Markov Chains with transition probability q and invariant distribution β so that

$$I(\beta) = \inf_{q:\beta q = \beta} \sum_{x,y} \beta(x)\, q(x, y) \log \frac{q(x, y)}{p(x, y)}.$$

One can deal with continuous time processes as well. One looks at the conditional distributions on $[0, T]$ given the past and calculates $H_T(Q : P)$ as the average relative entropy on $[0, T]$ of the conditional of Q with respect to the conditional of P. Assuming it is well defined one can show that $H_T(Q : P) = TH$ for some constant H and $H = H(Q : P)$ now controls the large deviation rates.

5. Entropy and duality.

In analysis the duality between the two functions $|x|^p/p$ and $|y|^q/q$ plays an important role. If $1/p + 1/q = 1$,

$$\frac{|x|^p}{p} = \sup_y \left[xy - \frac{|y|^q}{q} \right]$$

and we have the Hölder inequality

$$\left| \int fg d\mu \right| \leq \|f\|_p \|g\|_q.$$

The two functions $x \log x - x$ and e^y are duals

$$e^y = \sup_{x>0} [xy - (x \log x - x)]$$

and for $x \geq 0$

$$x \log x - x = \sup_y [xy - e^y].$$

This leads to the following generalization of Jensen's inequality

$$\int f(x) d\beta(x) \leq h(\beta : \alpha) + \log \int \exp[f(x)] d\alpha(x).$$

One can take f to be $c\chi_A(x)$ and obtain by optimizing over $c > 0$,

$$\beta(A) \leq \frac{h(\beta : \alpha) + 2}{\log(1/\alpha(A))}. \tag{5.1}$$

This can be thought of as the entropy analog of

$$\int_A |f(x)| d\alpha \leq [\alpha(A)]^{1/p} \|f\|_q.$$

When the dimension n of the spaces gets large β and α are getting nearly orthogonal. When dealing with $f = d\beta/d\alpha$ one has better control on the entropy $\int f \log f d\alpha$ that typically grows linearly, than on the L_p norms $\|f\|_p$ for $p > 1$ that grow exponentially with n. In some sense, the entropy $\int f \log f d\alpha$ should be thought of as $\|f\|_{1+0}$.

6. Log Sobolev inequality.

If we have a Markov process on X, with a reversible invariant distribution $\mu(dx)$, then we have the semigroup

$$(T_t f)(x) = \int f(y) p(t, x, dy)$$

which is a one parameter semigroup of self adjoint contractions on $L_2(X, \mu)$. Assuming some regularity we have the Dirichlet form

$$\mathcal{D}(f) = \lim_{t \to 0} \frac{1}{t} \langle f - T_t f, f \rangle = \lim_{t \to 0} \frac{1}{2t} \big[\|f\|_2^2 - \|T_t f\|_2^2 \big].$$

The log Sobolev inequality is an inequality of the following form. If $f \geq 0$ and $\|f\|_1 = 1$

$$\int f \log f d\mu \leq c \mathcal{D}(\sqrt{f}).$$

If we start the Markov process initially from a distribution with density f relative to μ, the density then evolves like $f(t) = T_t f$. A simple calculation shows that if L is differential operator $\nabla \cdot A \nabla$ on some $\mathbb{R}^n$

$$-\frac{d}{dt} \int f(t) \log f(t) d\mu = \int [\log f(t) L f(t)] d\mu = \int \frac{\langle \nabla f(t), A \nabla f(t) \rangle}{f(t)} d\mu = 4\mathcal{D}(\sqrt{f}).$$

If the operator L is nonlocal, we still have the inequality

$$-\frac{d}{dt} \int f(t) \log f(t) d\mu \geq 4\mathcal{D}(\sqrt{f}).$$

There are many useful estimates on the maps T_t from $L_p(X, \mu)$ to $L_q(X, \mu)$ that follow from a log Sobolev inequality. See [**8**] and [**5**] for an exposition.

7. Gibbs states.

In statistical mechanics one studies probability distributions on $\mathbb{R}^N$ defined through an interaction. Let us for simplicity consider the one dimensional case. A probability distribution μ_N on $\mathbb{R}^N$ is defined by

$$d\mu_N = \frac{1}{Z_N} \exp \Big[- \sum_{i=1}^{N} F(x_i, x_{i+1}, \ldots, x_{i+k-1}) - \sum_{i=1}^{N} V(x_i) \Big] \Pi dx_i$$

where

$$Z_N = \int_{\mathbb{R}^N} \exp \Big[- \sum_{i=1}^{N} F(x_i, x_{i+1}, \ldots, x_{i+k-1}) - \sum_{i=1}^{N} V(x_i) \Big] \Pi dx_i.$$

[We can assume that $\int e^{-V(x)}dx = 1$ and $1, \ldots, N$ are arranged periodically so that $N+i$ is identified with i.]

The first step is to calculate the limit

$$\lim_{N\to\infty} \frac{1}{N} \log Z_N = \psi(F).$$

We can express the sum $\sum_{i=1}^{N} F(x_i, x_{i+1}, \ldots, x_{i+k-1})$ as $N \int FdQ$ in terms of the empirical process Q with a large deviation rate of $H(Q : P)$ relative to the product measure $P = \Pi_i e^{-V(x_i)} dx_i$. The local contribution to the integral Z_N from Q is $\exp[-N[\int FdQ + H(Q : P)]]$. It is not hard to see that under suitable conditions

$$\psi(F) = -\inf_Q \left[\int FdQ + H(Q : P) \right]$$

where the infimum is taken over all stationary processes. If the infimum is attained at a unique Q then it defines a Gibbs state Q

$$dQ = c \exp\left[-\sum_i F(x_i, x_{i+1}, \ldots, x_{i+k-1}) - \sum_i V(x_i) \right] \Pi_i dx_i.$$

$\psi(F)$ referred to as “Free Energy” can be used to calculate $\int GdQ$ as the derivative at $\lambda = 0$

$$\frac{d}{d\lambda} \psi(F + \lambda G)\bigg|_{\lambda=0}.$$

See [**12**] for a detailed exposition.

8. Interacting particle systems.

In dealing with limiting behavior of large interacting systems with conserved quantities and multiple equilibria the system starting from non equilibrium evolves slowly towards equilibrium with local equilibria being established first and then the local equilibria converging slowly to a global equilibrium. This can be made precise. Initially the relative entropy with respect to a global equilibrium is about the size of the system. Its total decrease can not be more than the initial entropy $H(0)$.

$$\int_0^\infty \mathcal{D}(\sqrt{f(t)})dt \le H(0) \le CN$$

where N is the size of the system. This provides an average decay rate for $\mathcal{D}(\sqrt{f(t)})$ which can be used to establish local equilibria, and study how the system evolves to global equilibria. This has been carried out for example in [**9**] for a class of models. The critical step is the replacement of averages of rapidly changing quantities by their expectations under the appropriate equilibrium determined by the averages of the conserved quantities. In the end it boils down to showing that certain quantities are negligible. By using Feynman–Kac formula and eigenvalue estimates obtained through the Dirichlet form we can manage to get the required quantities to be negligible in equilibrium with exponentially small error probabilities. One can then use the entropy bound (5.1) to control the probability of error in nonequilibrium.

9. Another example.

The totally asymmetric simple exclusion process on the one dimensional periodic lattice of size N consists of ρN particles arranged with at most one particle per site at N equally spaced points on the circle of arc length 1. The particles wait independently for a random exponentially distributed time and try to jump to the next anti clockwise site provided it is free. Otherwise it waits for the next chance. As $N \to \infty$ the current state can be condensed to a density profile

$$\rho(x)dx = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \delta_{i/N} \xi_i$$

where $\xi_i = 1$ if there is a particle at site i and 0 otherwise. As the system evolves, if we look at times Nt, the profiles $\rho(t, x)$ in the limit as $N \to \infty$ are supposed to evolve according to

$$\frac{\partial \rho}{\partial t} + [\rho(t, x)(1 - \rho(t, x))]_x = 0. \tag{9.1}$$

If we are given a smooth solution of (9.1) with $0 < a \leq \rho(t, x) \leq b < 1$, we can start the particle system with a random product distribution μ_0 with

$$\mu_0[\xi_i = 1] = \rho\left(0, \frac{i}{N}\right).$$

This will evolve in time Nt to a distribution $\mu_N(t)$ which will not in general be a product measure and is difficult to compute. But we can define $\lambda_N(t)$ as the product measure with

$$\lambda_N(t)[\xi_i = 1] = \rho\left(t, \frac{i}{N}\right).$$

How close are $\lambda_N(t)$ and $\mu_N(t)$? One can look at the relative entropy $H_N(t) = h(\mu_N(t); \lambda_N(t))$ and note that $H_N(0) = 0$. By a Gronwall type argument one can show that $H_N(t) = o(N)$. This is enough to conclude that the empirical density profile at time Nt under $\mu_N(t)$ is very close to $\rho(t, x)$ with probability close to 1. See [**17**] where this method is carried out for a different model.

10. Postscript.

Before I met Professor Itô I had studied Stochastic Differential Equations. As some one interested in Markov processes I had started with semigroups of operators and the functional analytic approach. But with SDE I felt I understood better what diffusions were. I first met Professor Itô in Ithaca, NY when my wife Vasu and I visited Cornell in the spring of 1972. I remember going out for dinner with Professor and Mrs Itô. Four years later in 1976 there was a conference on SDE at Kyoto University for a week which I attended and got to know him quite well. That was one of many conferences organized by Professor Balakrishnan of UCLA on SDE where we usually met. My close interaction with him was in 1980 when I visited Japan for a six week stay at Osaka University. Professor Ikeda was my host. There were weekly seminars at Osaka as well as Kyoto when most of probabilists from Tokyo to Fukuoka came and Professor Itô was an active participant. Later I attended a couple of Taniguchi Symposia that were organized by him. Another interesting occasion was when we both attended a Vilnius conference in probability and went later to Moscow. I remember a lunch for a small group of us that included both Professors Kolmogorov and Itô. The last time I saw him was when he was in a nursing home near Kyoto. A small group of us visited him, Mrs Itô was there and he was in good spirits.

Today everybody on Wall street knows about Itô's formula and once after my lecture on Itô's formula in a course on Stochastic processes I was going up to my office in an elevator. My notes were open and a student next to me asked if I was teaching Mathematical Finance. I told him no and that I was teaching probability. His remark 'Is there Itô's formula in probability too?'

References

[1] J. Axzel and Z. Daroczy, On Measures of Information and Their Characterizations, Academic Press, New York, 1975.

[2] L. Boltzmann, Über die Mechanische Bedeutung des Zweiten Hauptsatzes der Wärmetheorie, Wiener Berichte, **53** (1866), 195–220.

[3] R. Clausius, Théorie mécanique de la chaleur, 1ère partie, Paris: Lacroix, 1868.

[4] H. Cramer, On a new limit theorem in the theory of probability, Colloquium on the Theory of Probability, Hermann, Paris, 1937.

[5] J. D. Deuschel and D. W. Stroock, Large deviations, Pure and Appl. Math., **137**, Academic Press, Inc., Boston, MA, 1989, xiv+307 pp.

[6] M. D. Donsker and S. R. S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time, IV, Comm. Pure Appl. Math., **36** (1983), 183–212.

[7] A. Feinstein, A new basic theorem of information theory, IRE Trans. Information Theory PGIT-4 (1954), 2–22.

[8] L. Gross, Logarithmic Sobolev inequalities, Amer. J. Math., **97** (1975), 1061–1083.

[9] M. Z. Guo, G. C. Papanicolaou and S. R. S. Varadhan, Nonlinear diffusion limit for a system with nearest neighbor interactions, Comm. Math. Phys., **118** (1988), 31–59.

[10] A. I. Khinchin, On the fundamental theorems of information theory, Translated by Morris D. Friedman, 572 California St., Newtonville MA 02460, 1956, 84 pp.

[11] A. N. Kolmogorov, A new metric invariant of transitive dynamical systems and automorphisms of Lebesgue spaces, (Russian) Topology, ordinary differential equations, dynamical systems, Trudy Mat. Inst., Steklov., **169** (1985), 94–98, 254.

[12] O. Lanford, Entropy and equilibrium states in classical statistical mechanics, Statistical Mechanics

and Mathematical Problems, Lecture notes in Physics, **20**, Springer-Verlag, Berlin and New York, 1971, 1–113.

[13] D. S. Ornstein, Ergodic theory, randomness, and dynamical systems, James K. Whittemore Lectures in Mathematics given at Yale University, Yale Mathematical Monographs, No. 5. Yale University Press, New Haven, Conn.-London, 1974, vii+141 pp.

[14] I. N. Sanov, On the probability of large deviations of random magnitudes, (Russian) Mat. Sb. (N. S.), **42** (84) (1957), 11–44.

[15] C. E. Shannon, A mathematical theory of communication, Bell System Tech. J., **27** (1948), 379–423, 623–656.

[16] Y. G. Sinai, On a weak isomorphism of transformations with invariant measure, (Russian) Mat. Sb. (N.S.), **63** (105) (1964), 23–42.

[17] H. T. Yau, Relative entropy and hydrodynamics of Ginzburg-Landau models, Lett. Math. Phys., **22** (1991), 63–80.

Srinivasa R. S. Varadhan
Courant Institute of Mathematical Sciences
251 Mercer Street
New York, USA
E-mail: varadhan@cims.nyu.edu