Statistical Inference, Statistical Mechanics and the Relationship to Information Theory

Sanjoy K. Mitter

September 15, 2004

Notes for Course 6.291, Fall 2004

1 Some Probabilistic Aspects of Entropy

1.1 Entropy as a Measure of Uncertainty

When Shannon had invented his quantity and consulted von Neumann on what 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."

Ludwig Boltzmann who first gave a probabilistic interpretation of Entropy, coined the famous formula:

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

which is engraved on his tombstone in Vienna.

This formula states that the entropy S of an observed Macrostate is proportional to the Logarithmic probability of its occurrence where k is Boltzmann's constant. To make this more precise and also explain how Boltzmann arrived at this formula, let E be a finite set and let μ be a Probability measure on E. Assume, $\mu(x) > 0$, $\forall x \in E$. In the Maxwell-Boltzmann picture, $E =$ set of all possible energy levels for a systems of particles and μ corresponds to a specific histogram of energies describing some macrostate of the system.

Let us assume that μ is a multiple of $\frac{1}{n}$, that is μ is a histogram for n trials or equivalently a **macrostate** for n-particles. On the microscopic level, the system is then described by a sequence $\omega \in E^n$, the microstate, associating with each particle its energy level.

Boltzmann's idea is then:

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

In more mathematical terms, for $\omega \in E^n$, let

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

be the empirical distribution which is the associated macrostate describing how the particles are distributed over the energy levels. Then

$$N_n(\mu) = \Big|\{\omega \in E^n | L_n^\omega = \mu\}\Big| = \frac{n!}{\prod\limits_{X\in E}\Big(n\mu(x)\Big)!} \tag{1.3}$$

(multi-nomial coefficient).

In view of the n-dependence of this quantity one should approximate μ by a sequence μ_n of n-particle microstates and define the entropy (uncertainty) $H(\mu)$ of μ when $n \to \infty$, limit of the mean uncertainty of μ_n per particle.

Theorem 1.1 *Let μ and μ_n be probability measures on E such that $\mu_n \to \mu$ and $n\mu_n(x) \in Z$, $\forall x \in E$. Then the limit*

$$\lim_{n\to\infty} \frac{1}{n}\log N_n(\mu) \quad \textit{exists and is equal to} \quad H(\mu) = -\sum_{x\in E} \mu(x)\log\mu(x) \tag{1.4}$$

Entropy thus describes the hidden multiplicity of "true" microstates consistent with the observed μ. In this sense, it is a measure of the complexity inherent in μ.

In summary, in the Boltzmann picture, μ is a histogram resulting from a random phenomenon on the microscopic level and $H(\mu)$ corresponds to the observer's uncertainty about what is really going on at the micro level.

It should be noted that in the Boltzmann picture, and more generally in Statistical Mechanics, the process of observation is never explicitly modeled. Nevertheless, the presence of an observer is postulated and entropy is a measure of uncertainty of the observer relative to a microphenomenon.

1.2 Entropy as a Measure of Information: The Shannon View

We now consider the problem of measuring the "Uncertainty Content" of a probability measure μ. Boltzmann's view is "backwards" to the microscopic origins of μ. Shannon's view is "forwards" taking μ as given and "randomizing" it by generating a random signal

X (i.e. a random sequence) with alphabet E and probability law μ. Given repeated realizations of the signal X, how large an effort do we need to identify X and hence μ? This corresponds to the **a priori uncertainty** that an observer has about μ (or a priori information contained in the source which generates the signal) the number of yes/no questions on the average that the identifier (observer) has to answer in identifying the signal. But, as observed by Shannon, this information is also the a-posterior information that the observer has after all questions have been answered. This leads to the following concept of information contained in a random signal:

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

Definition 1.1 *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 decodable in the sense that no codeword $f(x)$ is a prefix of another codeword $f(y)$.

Let $\ell(f(\mathbf{x}))$ denote the length of the codeword $f(x)$ and $E_\mu(\ell(f(x)))$ denote the expected length of the codeword. Let

$$I_p(\mu) = \inf \Big\{ E(\ell(f(x))) \,:\, f \ \text{ a binary prefix code for } E \Big\} \tag{1.5}$$

which we consider to be the information contained in the signal.

Let us assume that the random signal is generated by a memoryless source, in the sense that the random sequence corresponds to a sequence of independent random variables taking values in E and one encodes signals of length n, distributed according to the product measure μ^n. Hence

$$\frac{1}{n} I_p(\mu^n)$$

could be considered as the information contained in a single letter.

Theorem 1.2 *The information contained in a memoryless source with distribution μ is*

$$\lim_{n\to\infty} \frac{1}{n} I_p(\mu^n) = -\sum_{x \in E} \mu(x) \log_2(\mu(x)) \,:= H_2(\mu) = \frac{1}{\log_2} H(\mu) \ .$$

Definition 1.2 *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 $\phi \,:\, \{0,1\}^\ell \to E^n$, such that

$$\mu^n(\phi \cdot f \neq Id_{E_n}) \leq \alpha \ .$$

Define

$$I_p(\mu^n, \alpha) = \inf\Big\{\ell \,:\, \exists \;\; n\text{-block code of length } \ell \text{ at error level } \alpha\Big\} \tag{1.6}$$

Theorem 1.3 Source Coding Theorem for Block Codes

The information contained in a memoryless source with distribution μ is

$$\lim_{n\to\infty} \frac{1}{n} I_p(\mu^n, \alpha) = H_2(\mu)$$

independent of the error level α.

The proof of this result depends on

Proposition 1.1 (Asymptotic Equipartition Property)

For all $\delta > 0$,

$$\mu^n\Big(\omega \in E^n \Big| \Big|\frac{1}{n}\log \mu^n(\omega) + H(\mu)\Big| \le \delta\Big) \to 1 \tag{1.7}$$

as $n \to \infty$.

This says, most ω have probability

$$\mu^n(\omega) \simeq e^{-nH(\mu)} \ .$$

We could think of this as a random version of Boltzmann's formula (1.1).

1.3 Entropy arising from Large Deviation Considerations

Let us toss a fair coin n times and observe the number of heads. The number of ways of getting k heads is $\binom{n}{k}$. Using Stirling's Approximation for factorials, we get

$$\begin{aligned}\binom{n}{k} &\simeq \frac{\sqrt{2\pi}\ e^{-n} n^{n+\frac{1}{2}}}{\sqrt{2\pi}\ e^{-k} k^{k+\frac{1}{2}} \sqrt{2\pi}\ e^{-(n-k)} (n-k)^{(n-k+\frac{1}{2})}} \\ &= \frac{\sqrt{n}}{\sqrt{2\pi k(n-k)}} \exp\left[n\Big[-\frac{k}{n}\log\frac{k}{n} - \frac{n-k}{n}\log\frac{n-k}{n}\Big]\right] \end{aligned} \tag{1.8}$$

If we fix the proportions of heads and tails to be approximately p and $q = 1-p$, then we get

$$\binom{n}{k} \simeq \frac{1}{\sqrt{2\pi npq}} \exp\Big(nh(p,q)\Big) \tag{1.9}$$

where $h(p,q)$ is the **entropy** function

$$\Big(= -p\log p - q\log q\Big)$$

Therefore, the entropy is the factor in the exponential growth rate of the number of distinct outcomes in the space of coin tosses that corresponds to a given number of k heads.

The actual probability is

$$p_n(k) \simeq \frac{1}{\sqrt{2\pi npq}} \exp\Big(n[h(p,q) - \log 2]\Big) \tag{1.10}$$

Let us look at the situation when an "unfair" coin with probabilities α for heads and $(1-\alpha)$ for tails is tossed. Then

$$p_n(\alpha; k) \simeq \frac{1}{\sqrt{2\pi npq}} \exp\Big(n[h(p,q) + p\log\alpha + (1-p)\log(1-\alpha)]\Big) \tag{1.11}$$

The exponential constant

$$h(p,q) + p\log\alpha + (1-p)\log(1-\alpha) \tag{1.12}$$

is a sum of two terms depending on $p = (1-q) = \frac{k}{n}$.

The first term $h(p,q)$, is a term which corresponds to the number of distinct outcomes in the space of coin tosses which corresponds to a given number of k heads is akin to the volume term and is independent of α. The second term is akin to an energy term and is a function of p and α. The most likely state for k corresponds to the value of p that maximizes the combination

$$h(p,q) - E(p,q) \quad \text{which is} \quad p = \alpha \ .$$

Deviations of p away from α are called large deviations. Their probabilities decay exponentially as:

$$\exp(-nI(\alpha; p)) \tag{1.13}$$

where

$$I(\alpha, p) = p\log\frac{p}{\alpha} + (1-p)\log\frac{1-p}{1-\alpha} := E(p,q) - h(p,q)$$

$I(\alpha; p)$ is called a **rate function**, and $I(\alpha; p) \geq 0$. Indeed $I(\alpha, p) > 0$, unless $p = \alpha$.

1.4 Relative Entropy as a Measure of Discrimination between Probability Measures

Let μ_0 and μ_1 be two distinct probability measures on a finite set E. Suppose we do not know which of the two probability measures describes the random phenomenon. We wish to answer the question:

> How easy is it to distinguish between the two probability measures on the basis of independent observations of the random phenomenon?

The standard procedure in Statistics to answer this question is to perform a test of the hypothesis μ_0 against the alternative μ_1 with error level α. Since we want to do this on the basis of n-independent observations, we have to test the product measure μ_0^n against μ_1^n. Such a test is defined by a rejection region $R \subset E^n$, namely, if the observation belongs to R one decides in favor of the alternative μ_1. Otherwise, one chooses μ_0.

There are two types of errors:

(i) Rejecting the hypothesis μ_0 although it is true (error of the first kind)

(ii) Accepting μ_0 although it is false (error of the second kind).

One way to proceed is to keep the error probability of the first kind below a prescribed level α and to choose R such that the error probability of the second kind is minimized.

We therefore have the optimization problem

$$\rho_n(\alpha \,:\, \mu_0, \mu_1) := \inf \left\{ \mu_1^n(E^n \backslash R) \middle| R \subset E^n, \mu_0^n(R) \leq \alpha \right\} \tag{1.14}$$

and we can state the principle:

> 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 ρ_n and $n \to \infty$.

Lemma 1.1 Stein's Lemma: *The measure for discriminating μ_1 from μ_0 is*

$$-\lim_{n\to\infty} \log \rho_n(\alpha; \mu_0, \mu_1) = \sum_{x \in E} \mu_0(x) \log \frac{\mu_0(x)}{\mu_1(x)} := D(\mu_0|\mu_1) \tag{1.15}$$

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

$D(\mu_0|\mu_1)$ is called the relative entropy of μ_0 w.r.t. μ_1 (also Kullback-Leibler Distance, Information Gain).

If μ_1 is the uniform distribution on E, then $D(\mu_0|\mu_1) = \log|E| - H(\mu_0)$. $D(\mu_0|\mu_1)$ is not symmetric and does not satisfy the triangle inequality

$$D(\mu_0|\mu_1) \geq 0 \quad \text{and} \quad 0 \Leftrightarrow \mu_0 = \mu_1 \ .$$

Hence it is not really a distance. Nevertheless it has interesting properties such as lower semicontinuity and convexity which have important consequences.

Suppose we do not require the error probability of the first kind to remain fixed below a certain level, but allow it to decay exponentially at a given rate. The following theorem stated in terms of the empirical distribution L_n^ω describes the behavior of the error probabilities.

Theorem 1.4 Hoeffding's Theorem

Let $0 < a < D(\mu_0|\mu_1)$ and consider the test of μ_0 against μ_1 based on n independent observations with the rejection

$$R_n = \left\{ \omega \in E^n \middle| D(L_n^\omega|\mu_0) > a \right\} .$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_0^n(\mathbb{R}_n) = -a \ , \tag{1.16}$$

that is, the error probability of the first kind decays exponentially at rate α, and

$$\mu_1^n(E^n \backslash \mathbb{R}_n) \leq \exp \Big[-n \min_{\{\nu|D(\nu|\mu_0)\leq a\}} D(\nu|\mu_1) \Big] \tag{1.17}$$

that is, the error probability of the second kind satisfies an exponential bound with optimal exponent.

Note that the asymptotically optimal tests R_n does not depend on the alternative μ_1.

Hoeffding's theorem can be obtained from the more fundamental theorem of Sanov.

Theorem 1.5 Sanov's Theorem *Let μ be any probability measure on E and let $\mathcal{C}$ be a class of probability measures on E such that $\mathcal{C} \subset c\ell \inf \mathcal{C}$. Then*

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