

## Chapter 2

### INFORMATION THEORY

Suppose that the probability of interaction between individuals (or extended families or other organizational nodes) linked in a network depends jointly on their *geographic* and *social* locations, which we characterize as multidimensional vector quantities  $\mathbf{X}$  and  $\mathbf{Z}$  respectively. These measures might be determined, for example, from survey data on individuals or inferred from environmental index data on groups. Thus for individual nodes  $j$  and  $k$  we assume their probability of interaction  $P_{j,k}$  is given by

$$P_{j,k} = P_{j,k}(\mathbf{X}_j, \mathbf{X}_k, \mathbf{Z}_j, \mathbf{Z}_k)$$

where  $0 \leq P_{j,k} \leq 1$ .

It may be possible to reduce  $P_{j,k}$  to a function of the differences  $\mathbf{X} = \mathbf{X}_j - \mathbf{X}_k$  and  $\mathbf{Z} = \mathbf{Z}_j - \mathbf{Z}_k$ , or, perhaps, by using a multivariate method such as principal components analysis, even to functions of their ‘length’  $x = |\mathbf{X}|$  and  $z = |\mathbf{Z}|$ , so that

$$P_{j,k} = P_{j,k}(x, z).$$

One way to proceed is to impose a generalized distance  $r^2 \equiv x^2 + z^2$ , and explore the effects of various probability distributions which are functions of  $r$ . This is, in fact, done at some length in Chapter 4. Here, rather, we finesse the argument and transform out of the space defined by  $x$  and  $z$  into the probability space itself, defining a metric according to

$$L_{j,k} \equiv \log(1/P_{j,k})$$

(2.1)

where  $\log$  is the logarithm to some base number. Note that it is the probability distribution based on the generalized distance  $r$  which induces the transformation between ‘real’ space and probability space.

If it can be assumed for all nodes  $j, k, l$  within a sufficiently small network patch, that

$$P_{j,l} \geq P_{j,k} P_{k,l},$$

so that

$$\frac{1}{P_{j,l}} \leq \frac{1}{P_{j,k}} \frac{1}{P_{k,l}}$$

and the strong ‘triangle’ inequality

$$\log(1/P_{j,l}) \leq \log(1/P_{j,k}) + \log(1/P_{k,l})$$

holds, then  $L$  is a pseudometric, and various standard attacks are possible.

That somewhat draconian condition can, however, be considerably weakened as follows:

Let  $\Delta L_j$  be the average ‘distance’ in probability space from the node  $j$  to all other nodes, that is,

$$\Delta L_j \equiv \sum_k P_{j,k} \log(1/P_{j,k})$$

(2.2)

Suppose some fairly elaborate ‘message,’ not otherwise characterized, is sent along the sociogeographic network, and a *traveling wave* condition is imposed, so that, for some time period  $\Delta t$ , the relation

$$\frac{\Delta L_j}{\Delta t} \approx C,$$

(2.3)

holds.  $C$  is, then, *the mean fixed rate at which the message is sent from the network to an embedded individual.*

Wallace et al, (1996) show – not unexpectedly and probably not originally – that the traveling wave assumption, on a fractal manifold, is equivalent to the Aharony-Stauffer conjecture, which directly relates the fractal dimension of the interior of an affected network to that of its growth surface. This gives a simple and explicit expression for  $C$ , usually an arduous calculation. The detailed derivation is left as an exercise.

By expanding equation (2.3) we obtain, taking  $\Delta t \equiv 1$ ,

$$-\sum_k P_{j,k} \log(P_{j,k}) \approx C \quad (2.4)$$

where  $C$  is a transmission rate constant characteristic of the particular sociogeographic network. We further assume that  $\sum_k P_{j,k} \equiv 1$ , i.e., the network is ‘tight’ in the sense that each node interacts with it as a whole with unit probability. Hence  $P_{j,k}$  is a legitimate probability distribution.

These are deep waters: For any probability distribution,  $0 \leq P_j \leq 1$  such that  $\sum_j P_j = 1$  the quantity

$$H = -\sum_j P_j \log(P_j) \quad (2.5)$$

is the distribution’s *Shannon uncertainty*, a fundamental quantity of classical information theory.

Neglecting details explored below, the transfer of uncertainty represents the transmission of information: The Shannon Coding Theorem, the first important result of information theory, states that for any rate  $R < C$ , where  $C$  represents the capacity of the information channel, it is possible to find a ‘coding scheme’ such that a sufficiently long message can be sent with arbitrarily small error.

This is surely one of the most striking conclusions of 20th Century applied mathematics.

## 1. The Shannon Coding Theorem

Messages from a source, seen as symbols  $x_j$  from some alphabet, each having probabilities  $P_j$  associated with a random variable  $X$ , are ‘encoded’ into the language of a ‘transmission channel’, a random variable  $Y$  with symbols  $y_k$ , having probabilities  $P_k$ , possibly with error. Someone receiving the symbol  $y_k$  then retranslates it (without error) into some  $x_k$ , which may or may not be the same as the  $x_j$  that was sent.

More formally, the message sent along the channel is characterized by a random variable  $X$  having the distribution

$$P(X = x_j) = P_j, j = 1, \dots, M.$$

The channel through which the message is sent is characterized by a second random variable  $Y$  having the distribution

$$P(Y = y_k) = P_k, k = 1, \dots, L.$$

Let the joint probability distribution of  $X$  and  $Y$  be defined as

$$P(X = x_j, Y = y_k) = P(x_j, y_k) = P_{j,k}$$

and the conditional probability of  $Y$  given  $X$  as

$$P(Y = y_k | X = x_j) = P(y_k | x_j).$$

Then the Shannon uncertainty of  $X$  and  $Y$  independently and the joint uncertainty of  $X$  and  $Y$  together are defined respectively as

$$H(X) = - \sum_{j=1}^M P_j \log(P_j)$$

$$H(Y) = - \sum_{k=1}^L P_k \log(P_k)$$

$$H(X, Y) = - \sum_{j=1}^M \sum_{k=1}^L P_{j,k} \log(P_{j,k}).$$

(2.6)

The *conditional uncertainty* of  $Y$  given  $X$  is defined as

$$H(Y|X) = - \sum_{j=1}^M \sum_{k=1}^L P_{j,k} \log[P(y_k|x_j)]$$

(2.7)

For any two stochastic variates  $X$  and  $Y$ ,  $H(Y) \geq H(Y|X)$ , as knowledge of  $X$  generally gives some knowledge of  $Y$ . Equality occurs only in the case of stochastic independence.

Since  $P(x_j, y_k) = P(x_j)P(y_k|x_j)$ , we have

$$H(X|Y) = H(X, Y) - H(Y)$$

The information transmitted by translating the variable  $X$  into the channel transmission variable  $Y$  – possibly with error – and then retranslating without error the transmitted  $Y$  back into  $X$  is defined as

$$I(X|Y) \equiv H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y)$$

(2.8)

See, for example, Ash (1990), Khinchine (1957) or Cover and Thomas (1991) for details. The essential point is that if there is no uncertainty in  $X$  given the channel  $Y$ , then there is no loss of information through transmission.

In general this will not be true, and herein lies the essence of the theory.

Given a fixed vocabulary for the transmitted variable  $X$ , and a fixed vocabulary and probability distribution for the channel  $Y$ , we may vary the probability distribution of  $X$  in such a way as to maximize the information sent. The capacity of the channel is defined as

$$C \equiv \max_{P(X)} I(X|Y)$$

(2.9)

subject to the subsidiary condition that  $\sum P(X) = 1$ .

The critical trick of the Shannon Coding Theorem for sending a message with arbitrarily small error along the channel  $Y$  at any rate  $R < C$  is to encode it in longer and longer ‘typical’ sequences of the variable  $X$ ; that is, those sequences whose distribution of symbols approximates the probability distribution  $P(X)$  above which maximizes  $C$ .

If  $S(n)$  is the number of such ‘typical’ sequences of length  $n$ , then

$$\log[S(n)] \approx nH(X)$$

where  $H(X)$  is the uncertainty of the stochastic variable defined above. Some consideration shows that  $S(n)$  is much less than the total number of possible messages of length  $n$ . Thus, as  $n \rightarrow \infty$ , only a vanishingly small fraction of all possible messages is meaningful in this sense. This observation, after some considerable development, is what allows the Coding Theorem to work so well. In sum, the prescription is to encode messages in typical sequences, which are sent at very nearly the capacity of the channel. As the encoded messages become longer and longer, their maximum possible rate of transmission without error approaches channel capacity as a limit. Again, Ash (1990), Khinchine (1957) and Cover and Thomas (1991) provide details.

## 2. More heuristics: a ‘tuning theorem’

Telephone lines, optical wave guides and the tenuous plasma through which a planetary probe transmits data to earth may all be viewed in traditional information-theoretic terms as a *noisy channel* around which we must structure a message so as to attain an optimal error-free transmission rate.

Telephone lines, wave guides and interplanetary plasmas are, relatively speaking, fixed on the timescale of most messages, as are most sociogeographic networks. Indeed, the capacity of a channel, according to equation (2.9), is defined by varying the probability distribution of the ‘message’ process  $X$  so as to maximize  $I(X|Y)$ .

Suppose there is some message  $X$  so critical that its probability distribution must remain fixed. The trick is to fix the distribution  $P(x)$  but *modify the channel* – i.e. tune it – so as to maximize  $I(X|Y)$ . The *dual* channel capacity  $C^*$  can be defined as

$$C^* \equiv \max_{P(Y), P(Y|X)} I(X|Y)$$

(2.10)

But

$$C^* = \max_{P(Y), P(Y|X)} I(Y|X)$$

since

$$I(X|Y) = H(X) + H(Y) - H(X, Y) = I(Y|X).$$

Thus, in a purely formal mathematical sense, *the message transmits the channel*, and there will indeed be, according to the Coding Theorem, a channel distribution  $P(Y)$  which maximizes  $C^*$ .

One may do better than this, however, by modifying the channel matrix  $P(Y|X)$ . Since

$$P(y_j) = \sum_{i=1}^M P(x_i) P(y_j|x_i),$$

$P(Y)$  is entirely defined by the channel matrix  $P(Y|X)$  for fixed  $P(X)$  and

$$C^* = \max_{P(Y), P(Y|X)} I(Y|X) = \max_{P(Y|X)} I(Y|X).$$

Calculating  $C^*$  requires maximizing the complicated expression

$$I(X|Y) = H(X) + H(Y) - H(X, Y)$$

which contains products of terms and their logs, subject to constraints that the sums of probabilities are 1 and each probability is itself between 0 and 1. Maximization is done by varying the channel matrix terms  $P(y_j|x_i)$  within the constraints. This is a difficult problem in nonlinear optimization. See Parker et al. (2003) for a comprehensive treatment, using traditional Lagrange multiplier methods. However, for the special case  $M = L$ ,  $C^*$  may be found by inspection:

If  $M = L$ , then choose

$$P(y_j|x_i) = \delta_{j,i}$$

where  $\delta_{i,j}$  is 1 if  $i = j$  and 0 otherwise. For this special case

$$C^* \equiv H(X)$$

with  $P(y_k) = P(x_k)$  for all  $k$ . *Information is thus transmitted without error when the channel becomes ‘typical’ with respect to the fixed message distribution  $P(X)$ .*

If  $M < L$  matters reduce to this case, but for  $L < M$  information must be lost, leading to ‘Rate Distortion’ arguments explored more fully below.

Thus modifying the channel may be a far more efficient means of ensuring transmission of an important message than encoding that message in a ‘natural’ language which maximizes the rate of transmission of information on a fixed channel.

We have examined the two limits in which either the distributions of  $P(Y)$  or of  $P(X)$  are kept fixed. The first provides the usual Shannon Coding Theorem, and the second, hopefully, a tuning theorem variant. It seems likely, however, than for many important systems  $P(X)$  and  $P(Y)$  will ‘interpenetrate,’ to use Richard Levins’ terminology. That is,  $P(X)$  and  $P(Y)$  will affect each other in characteristic ways, so that some form of mutual tuning may be the most effective strategy.

### 3. The Shannon-McMillan Theorem

Not all statements – sequences of the random variable  $X$  – are equivalent. According to the structure of the underlying language of which the message is a particular expression, some messages are more ‘meaningful’ than others, that is, in accord with the grammar and syntax of the language. The other principal result from information theory, the Shannon-McMillan or Asymptotic Equipartition Theorem, describes how messages themselves are to be classified.

Suppose a long sequence of symbols is chosen, using the output of the random variable  $X$  above, so that an output sequence of length  $n$ , with the form

$$x_n = (\alpha_0, \alpha_1, \dots, \alpha_{n-1})$$

has joint and conditional probabilities

$$P(X_0 = \alpha_0, X_1 = \alpha_1, \dots, X_{n-1} = \alpha_{n-1})$$

$$P(X_n = \alpha_n | X_0 = \alpha_0, \dots, X_{n-1} = \alpha_{n-1}).$$

(2.11)

Using these probabilities we may calculate the conditional uncertainty



$$H(X_n|X_0, X_1, \dots, X_{n-1}).$$

The uncertainty of the *information source*,  $H[\mathbf{X}]$ , is defined as

$$H[\mathbf{X}] \equiv \lim_{n \rightarrow \infty} H(X_n|X_0, X_1, \dots, X_{n-1}). \quad (2.12)$$

In general

$$H(X_n|X_0, X_1, \dots, X_{n-1}) \leq H(X_n).$$

Only if the random variables  $X_j$  are all stochastically independent does equality hold. If there is a maximum  $n$  such that, for all  $m > 0$

$$H(X_{n+m}|X_0, \dots, X_{n+m-1}) = H(X_n|X_0, \dots, X_{n-1}),$$

then the source is said to be of *order*  $n$ . It is easy to show that

$$H[\mathbf{X}] = \lim_{n \rightarrow \infty} \frac{H(X_0, \dots, X_n)}{n+1}.$$

In general the outputs of the  $X_j, j = 0, 1, \dots, n$  are *dependent*. That is, the output of the communication process at step  $n$  depends on previous steps. Such serial correlation, in fact, is the very structure which enables most of what follows in this book.

Here, however, the processes are all assumed stationary in time, that is, the serial correlations do not change in time, and the system is *memoryless*.

A very broad class of such self-correlated, memoryless, information sources, the so-called *ergodic* sources for which the long-run relative frequency of a sequence converges stochastically to the probability assigned to it, have a particularly interesting property:

It is possible, in the limit of large  $n$ , to divide all sequences of outputs of an ergodic information source into two distinct sets,  $S_1$  and  $S_2$ , having, respectively, very high and very low probabilities of occurrence, with the source uncertainty providing the splitting criterion. In particular the Shannon-McMillan Theorem states that, for a (long) sequence having  $n$  (serially correlated) elements, the number of ‘meaningful’ sequences,  $N(n)$  – those belonging to set  $S_1$  – will satisfy the relation

$$\frac{\log[N(n)]}{n} \approx H[\mathbf{X}].$$

(2.13)

More formally,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\log[N(n)]}{n} &= H[\mathbf{X}] \\ &= \lim_{n \rightarrow \infty} H(X_n | X_0, \dots, X_{n-1}) \\ &= \lim_{n \rightarrow \infty} \frac{H(X_0, \dots, X_n)}{n+1}. \end{aligned}$$

(2.14)

The Shannon Coding theorem, by means of an analogous splitting argument, shows that for any rate  $R < C$ , where  $C$  is the channel capacity, a message may be sent without error, using the probability distribution for  $X$  which maximizes  $I(X|Y)$  as the coding scheme. Using the internal structures of the information source permits *limiting attention only to meaningful sequences of symbols*. This restriction can greatly raise the maximum possible rate at which information can be transmitted with arbitrarily small error: if there are  $M$  possible symbols and the uncertainty of the source is  $H[\mathbf{X}]$ , then the effective capacity of the channel  $C$ , using this ‘source coding,’ becomes (Ash, 1990)

$$C_E = C \frac{\log(M)}{H[\mathbf{X}]}.$$

(2.15)

As  $H[\mathbf{X}] \leq \log(M)$ , with equality only for stochastically independent, uniformly distributed random variables,

$$C_E \geq C.$$

(2.16)

Note that, for a given channel capacity, the condition

$$H[\mathbf{X}] \leq C$$

always holds.

Source uncertainty has a very important heuristic interpretation. As Ash (1990) puts it,

...[W]e may regard a portion of text in a particular language as being produced by an information source. The probabilities  $P[X_n = \alpha_n | X_0 = \alpha_0, \dots, X_{n-1} = \alpha_{n-1}]$  may be estimated from the available data about the language; in this way we can estimate the uncertainty associated with the language. A large uncertainty means, by the [Shannon-McMillan Theorem], a large number of 'meaningful' sequences. Thus given two languages with uncertainties  $H_1$  and  $H_2$  respectively, if  $H_1 > H_2$ , then in the absence of noise it is easier to communicate in the first language; more can be said in the same amount of time. On the other hand, it will be easier to reconstruct a scrambled portion of text in the second language, since fewer of the possible sequences of length  $n$  are meaningful.

It is possible to significantly generalize this heuristic picture in such a way as to characterize the interaction between different 'languages,' something at the core of the development.

#### 4. The Rate Distortion Theorem

The Shannon-McMillan Theorem can be expressed as the 'zero error limit' of something called the Rate Distortion Theorem (Dembo and Zeitouni, 1998; Cover and Thomas, 1991), which defines a splitting criterion that identifies high probability pairs of sequences. We follow closely the treatment of Cover and Thomas (1991).

The origin of the problem is the question of representing one information source by a simpler one in such a way that the least information is lost. For example we might have a continuous variate between 0 and 100, and wish to represent it in terms of a small set of integers in a way that minimizes the inevitable distortion that process creates. Typically, for example, an analog

audio signal will be replaced by a ‘digital’ one. The problem is to do this in a way which least distorts the *reconstructed* audio waveform.

Suppose the original memoryless, ergodic information source  $Y$  with output from a particular alphabet generates sequences of the form

$$y^n = y_1, \dots, y_n.$$

These are ‘digitized,’ in some sense, producing a chain of ‘digitized values’

$$b^n = b_1, \dots, b_n,$$

where the  $b$ -alphabet is much more restricted than the  $y$ -alphabet.

$b^n$  is, in turn, *deterministically retranslated* into a reproduction of the original signal  $y^n$ . That is, each  $b^n$  is mapped on to a unique  $n$ -length  $y$ -sequence in the alphabet of the information source  $Y$ :

$$b^n \rightarrow \hat{y}^n = \hat{y}_1, \dots, \hat{y}_n.$$

Note, however, that many  $y^n$  sequences may be mapped onto the *same* re-translation sequence  $\hat{y}^n$ , so that information will, in general, be lost.

The central problem is to explicitly minimize that loss.

The retranslation process defines a new memoryless, ergodic information source,  $\hat{Y}$ .

The next step is to define a *distortion measure*,  $d(y, \hat{y})$ , which compares the original to the retranslated path. For example the *Hamming distortion* is

$$d(y, \hat{y}) = 1, y \neq \hat{y}$$

$$d(y, \hat{y}) = 0, y = \hat{y}.$$

(2.17)

For continuous variates the *Squared error distortion* is

$$d(y, \hat{y}) = (y - \hat{y})^2.$$

(2.18)

Possibilities abound.

The distortion between paths  $y^n$  and  $\hat{y}^n$  is defined as

$$d(y^n, \hat{y}^n) = \frac{1}{n} \sum_{j=1}^n d(y_j, \hat{y}_j).$$

(2.19)

Suppose that with each path  $y^n$  and  $b^n$ -path retranslation into the  $y$ -language and denoted  $\hat{y}^n$ , there are associated individual, joint, and conditional probability distributions

$$p(y^n), p(\hat{y}^n), p(y^n | \hat{y}^n).$$

The *average distortion* is defined as

$$D = \sum_{y^n} p(y^n) d(y^n, \hat{y}^n).$$

(2.20)

It is possible, using the distributions given above, to define the information transmitted from the incoming  $Y$  to the outgoing  $\hat{Y}$  process in the usual manner, using the Shannon source uncertainty of the strings:

$$I(Y, \hat{Y}) \equiv H(Y) - H(Y | \hat{Y}) = H(Y) + H(\hat{Y}) - H(Y, \hat{Y}).$$

If there is no uncertainty in  $Y$  given the retranslation  $\hat{Y}$ , then no information is lost.

In general, this will not be true.

The *information rate distortion function*  $R(D)$  for a source  $Y$  with a distortion measure  $d(y, \hat{y})$  is defined as

$$R(D) = \min_{p(y, \hat{y}); \sum_{(y, \hat{y})} p(y)p(\hat{y})d(y, \hat{y}) \leq D} I(Y, \hat{Y}). \quad (2.21)$$

The minimization is over all conditional distributions  $p(y|\hat{y})$  for which the joint distribution  $p(y, \hat{y}) = p(y)p(y|\hat{y})$  satisfies the average distortion constraint (i.e. average distortion  $\leq D$ ).

The *Rate Distortion Theorem* states that  $R(D)$  is the maximum achievable rate of information transmission which does not exceed the distortion  $D$ . Cover and Thomas (1991) or Dembo and Zeitouni (1998) provide details, and Parker et al. (2003) formalize a comprehensive attack.

More to the point, however, is the following: Pairs of sequences  $(y^n, \hat{y}^n)$  can be defined as *distortion typical*; that is, for a given average distortion  $D$ , defined in terms of a particular measure, pairs of sequences can be divided into two sets, a high probability one containing a relatively small number of (matched) pairs with  $d(y^n, \hat{y}^n) \leq D$ , and a low probability one containing most pairs. As  $n \rightarrow \infty$ , the smaller set approaches unit probability, and, for those pairs,

$$p(y^n) \geq p(\hat{y}^n|y^n) \exp[-nI(Y, \hat{Y})]. \quad (2.22)$$

Thus, roughly speaking,  $I(Y, \hat{Y})$  embodies the splitting criterion between high and low probability pairs of paths.

For the theory of interacting information sources, then,  $I(Y, \hat{Y})$  can play the role of  $H$  in the dynamic treatment that follows.

The rate distortion function of eq. 2.21 can actually be calculated in many cases by using a Lagrange multiplier method – see Section 13.7 of Cover and Thomas (1991).

At various points in the development we will suggest using  $s \equiv d(\hat{x}, x)$  as a metric in a geometry of information sources, e.g. when simple ergodicity fails, and  $H(x) \neq H(\hat{x})$  for high probability paths  $\hat{x}$  and  $x$ . See eq. (3.2).

## 5. Large Deviations

The use of information source uncertainty above as a splitting criterion between high and low probability sequences (or pairs of them) displays the fundamental characteristic of a growing body of work in applied probability often termed the ‘Large Deviations Program,’ (LDP) which seeks to unite information theory, statistical mechanics and the theory of fluctuations under a single umbrella. It serves as a convenient starting point for further developments.

We can begin to place information theory in the context of the LDP as follows (Dembo and Zeitouni, 1998, p.2):

Let  $X_1, X_2, \dots, X_n$  be a sequence of independent, standard Normal, real-valued random variables and let

$$S_n = \frac{1}{n} \sum_{j=1}^n X_j.$$

(2.23)

Since  $S_n$  is again a Normal random variable with zero mean and variance  $1/n$ , for all  $\delta > 0$

$$\lim_{n \rightarrow \infty} P(|S_n| \geq \delta) = 0,$$

(2.24)

where  $P$  is the probability that the absolute value of  $S_n$  is greater or equal to  $\delta$ . Some manipulation, however, gives

$$P(|S_n| \geq \delta) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} \exp(-x^2/2) dx,$$

(2.25)

so that

$$\lim_{n \rightarrow \infty} \frac{\log P(|S_n| \geq \delta)}{n} = -\delta^2/2$$

(2.26)

This can be rewritten for large  $n$  as

$$P(|S_n| \geq \delta) \approx \exp(-n\delta^2/2).$$

(2.27)

That is, for large  $n$ , the probability of a large deviation in  $S_n$  follows something much like equation (2.13), i.e. that meaningful paths of length  $n$  all have approximately the same probability  $P(n) \propto \exp(-nH[\mathbf{X}])$ .

Our questions about ‘meaningful paths’ appear suddenly as formally isomorphic to the central argument of the LDP which encompasses statistical mechanics, fluctuation theory, and information theory into a single structure (Dembo and Zeitouni, 1998).

A cardinal tenet of large deviation theory is that the ‘rate function’  $-\delta^2/2$  in equation (2.26) can, under proper circumstances, be expressed as a mathematical ‘entropy’ having the standard form

$$-\sum_k p_k \log p_k,$$

(2.28)

for some set of probabilities  $p_k$ . This striking result goes under various names at various levels of approximation – Sanov’s Theorem, Cramer’s Theorem, the Gartner-Ellis Theorem, the Shannon-McMillan Theorem, and so on (Dembo and Zeitouni, 1998).



## 6. Fluctuations

The standard treatment of ‘fluctuations’ (Onsager and Machlup, 1953; Fredlin and Wentzell, 1998) in physical systems is the principal foundation for much current study of stochastic resonance and related phenomena and also serves as a useful reference point.

The macroscopic behavior of a complicated physical system in time is assumed to be described by the phenomenological Onsager relations giving large-scale fluxes as

$$\sum_i R_{i,j} dK_j/dt = \partial S/\partial K_i, \quad (2.29)$$

where the  $R_{i,j}$  are appropriate constants,  $S$  is the system entropy and the  $K_i$  are the generalized coordinates which parametrize the system’s free energy.

Entropy is defined from free energy  $F$  by a Legendre transform – more of which follows below:

$$S \equiv F - \sum_j K_j \partial F/\partial K_j,$$

where the  $K_j$  are appropriate system parameters.

Neglecting volume problems (which will become quite important later), free energy can be defined from the system’s partition function  $Z$  as

$$F(K) = \log[Z(K)].$$

The partition function  $Z$ , in turn, is defined from the system Hamiltonian – defining the energy states – as

$$Z(K) = \sum_j \exp[-K E_j],$$

where  $K$  is an inverse temperature or other parameter and the  $E_j$  are the energy states.

Inverting the Onsager relations gives

$$dK_i/dt = \sum_j L_{i,j} \partial S/\partial K_j = L_i(K_1, \dots, K_m, t) \equiv L_i(K, t).$$

(2.30)

The terms  $\partial S/\partial K_i$  are macroscopic driving ‘forces’ dependent on the entropy gradient.

Let a white Brownian ‘noise’  $\epsilon(t)$  perturb the system, so that

$$\begin{aligned} dK_i/dt &= \sum_j L_{i,j} \partial S/\partial K_j + \epsilon(t) \\ &= L_i(K, t) + \epsilon(t), \end{aligned}$$

(2.31)

where the time averages of  $\epsilon$  are  $\langle \epsilon(t) \rangle = 0$  and  $\langle \epsilon(t)\epsilon(0) \rangle = D\delta(t)$ .  $\delta(t)$  is the Dirac delta function, and we take  $K$  as a vector in the  $K_i$ .

Following Luchinsky (1997), if the probability that the system starts at some initial macroscopic parameter state  $K_0$  at time  $t = 0$  and gets to the state  $K(t)$  at time  $t$  is  $P(K, t)$ , then a somewhat subtle development (e.g. Feller, 1971) gives the forward Fokker-Planck equation for  $P$ :

$$\partial P(K, t)/\partial t = -\nabla \cdot (L(K, t)P(K, t)) + (D/2)\nabla^2 P(K, t).$$

(2.32)

In the limit of weak noise intensity this can be solved using the WKB, i.e. the eikonal, approximation, as follows: take

$$P(K, t) = z(K, t) \exp(-s(K, t)/D).$$

(2.33)

$z(K, t)$  is a prefactor and  $s(K, t)$  is a classical action satisfying the Hamilton-Jacobi equation, which can be solved by integrating the Hamiltonian equations of motion. The equation reexpresses  $P(K, t)$  in the usual parametrized negative exponential format.

Let  $p \equiv \nabla s$ . Substituting equation (2.33) in equation (2.32) and collecting terms of similar order in  $D$  gives

$$dK/dt = p + L, dp/dt = -\partial L/\partial K p$$

$$-\partial s/\partial t \equiv h(K, p, t) = pL(K, t) + \frac{p^2}{2},$$

with  $h(K, t)$  the ‘Hamiltonian’ for appropriate boundary conditions.

Again following Luchinsky (1997), these ‘Hamiltonian’ equations have two different types of solution, depending on  $p$ . For  $p = 0$ ,  $dK/dt = L(K, t)$  which describes the system in the absence of noise. We expect that with finite noise intensity the system will give rise to a distribution about this deterministic path. Solutions for which  $p \neq 0$  correspond to *optimal paths* along which the system will move with overwhelming probability.

This is a formulation of fluctuation theory which has particular attraction for physicists, few of whom can resist the nearly magical appearance of a Hamiltonian. These results can, however, again be directly derived as a special case of a Large Deviation Principle based on ‘generalized ‘entropies’ mathematically similar to Shannon’s uncertainty from information theory, bypassing the ‘Hamiltonian’ formulation entirely (Dembo and Zeitouni, 1998).

For languages, of course, there is no possibility of a Hamiltonian, but the generalized entropy or splitting criterion treatment still works. The trick will be to do with entropies what is most often done with Hamiltonians:

Here we will be concerned, not with a random Brownian distortion of simple physical systems, but with a complex ‘behavioral’ structure, in the largest sense, composed of quasi-independent ‘actors’ for which

[1] the usual Onsager relations of equations (2.29) and (2.30) may be too simple,

[2] the ‘noise’ may not be either small or random, and, most critically,

[3] *the meaningful/optimal paths have extremely structured serial correlation, amounting to a grammar and syntax, precisely the fact which allows definition of an information source* and enables the use of the very sparse equipartition of the Shannon-McMillan and Rate Distortion Theorems. The sparseness and equipartition, in fact, permit solution of the problems we will address.

In sum, to again paraphrase Luchinsky (1997), large fluctuations, although infrequent, are fundamental in a broad range of processes, and it was recognized by Onsager and Machlup (1953) that insight into the problem could be gained from studying the distribution of fluctuational paths along which the system moves to a given state. This distribution is a fundamental characteristic of the fluctuational dynamics, and its understanding leads toward control of fluctuations. Fluctuational motion from the vicinity of a stable state may occur along different paths. For large fluctuations, the distribution of these paths peaks sharply along an optimal, most probable, path. In the theory of large fluctuations, the pattern of optimal paths plays a role similar to that of the phase portrait in nonlinear dynamics.

In this development ‘meaningful’ paths play the role of ‘optimal’ paths in the theory of large fluctuations, but without benefit of a ‘Hamiltonian.’

## 7. The fundamental homology

Section 5 above gives something of the flavor of the LDP which tries to unify statistical mechanics, large fluctuations and information theory. This opens a methodological Pandora’s Box: the LDP provides justification for a massive transfer of superstructure from statistical mechanics to information theory, including real-space renormalization for address of phase transition, thermodynamics and an equation of state, generalized Onsager relations, and so on. From fluctuation theory and nonlinear dynamics come phase space, domains of attraction and related matters.

Several particulars distinguish this approach.

First is a draconian simplification which seeks to employ information theory concepts only as they directly relate to the basic limit theorems of the subject. That is, message uncertainty and information source uncertainty are interesting only because they obey the Coding, Source Coding, Rate Distortion, and related theorems. ‘Information Theory’ treatments which do not sufficiently center on these theorems are, from this view, far off the mark. Thus most discussion of ‘complexity,’ ‘entropy maximization,’ different definitions of ‘entropy,’ and so forth, just does not appear on the horizon. In the words of William of Occam, “Entities ought not be multiplied without necessity.”

The second matter is somewhat more complicated: Rojdestvenski and Cottam (2000, p.44), following Wallace and Wallace (1998), see the linkage between information theory and statistical mechanics as a characteristic

...[homological] mapping... between... unrelated... problems that share the same mathematical basis... [whose] similarities in mathematical formalisms...become powerful tools for [solving]... traditional problems.

The possible relation of information theory to biological and social process, both of which can involve agency, appears very sharply constrained, involving:

(1) a ‘linguistic’ equipartition of sets of probable paths consistent with the Shannon-McMillan, Rate Distortion, or related theorems which serves as the formal connection with nonlinear mechanics and fluctuation theory, and

(2) a homological correspondence between information source uncertainty and statistical mechanical free energy density, not statistical mechanical entropy.

In this latter regard, the definition of the free energy density of a parametrized physical system is

$$F(K_1, \dots, K_m) = \lim_{V \rightarrow \infty} \frac{\log[Z(K_1, \dots, K_m, V)]}{V}, \quad (2.34)$$

where the  $K_j$  are parameters,  $V$  is the system volume, and  $Z$  is, again, the partition function.

For an ergodic information source the equivalent relation associates the source uncertainty with the number of ‘meaningful’ statements  $N(n)$  of length  $n$ , in the limit,

$$H[\mathbf{X}] = \lim_{n \rightarrow \infty} \frac{\log[N(n)]}{n}.$$

This can be parametrized in various manners to obtain the crucial expression on which all else is built:

$$H[K_1, \dots, K_m, \mathbf{X}] = \lim_{n \rightarrow \infty} \frac{\log[N(K_1, \dots, K_m, n)]}{n}. \quad (2.35)$$

At first glance, Shannon uncertainty has the algebraic form of the entropy of a physical system,  $\propto \sum_k P_k \log(P_k)$ , where the  $P_k$  constitute a probability distribution. This is deceptive. In the absence of a ‘distinguishing two-form’ which defines a ‘symplectic geometry’, that is, in the absence of a second order Hamiltonian defining energy, Shannon uncertainty cannot be the ‘entropy’ of a system, even if it has the same mathematical form. See Arnold (1989) for an explanation of the ‘symplectic’ jargon, but the basic point is that the concept

of entropy is directly derived from ideas of work and heat, while Shannon uncertainty has its origin in the process of sending a message.

While it is sometimes possible to impose a distinguishing two-form on a contact manifold, to symplectify it by artificially constructing an analog to a Hamiltonian (Arnold, 1989), such logical convolutions are not really needed. In any event, when such a ‘duality’ is invoked, Shannon uncertainty does not become the analog of thermodynamic entropy: resolution of a famous paradox in physics requires an identification of Shannon uncertainty with free energy density. As Elitzur (1996, p. 179) puts it

Recall ... the lesson of Maxwell’s Demon: Information, when applied under the appropriate circumstances, can save work.

Bennett (1988, p. 230), as quoted by Elitzur (1996) states

...[T]he value of a message is the amount of mathematical or other work plausibly done by its originator, which the receiver is saved from having to repeat.

Similarly, Feynman (1996) provides a formal example, showing that, for a certain class of microscopic systems, transmission of information can be interpreted as exchange of free energy.

This, then, is the essential ‘homology’ linking information theory to the technology of statistical mechanics and related disciplines. Only for very simple systems – e.g. Bennett’s microscopically reversible computing machinery – can the homology be an identity. In general this will not be the case, as individual ‘agency’ increasingly imposes behavioral regularities which are not simply mechanistic: Thus the ‘Hamiltonian’ goes away, but an ‘entropy’ treatment using source uncertainties, remains possible.

That is, for mesoscale or ‘behavioral’ systems, infinite-volume-based or Hamiltonian-driven thermodynamic treatments familiar from physics are inappropriate since either the usual forms of the ergodic theorem break down (e.g. Bar-Yam, 1997), or there is simply no underlying scalar function to maximize or minimize. It is possible to regain something much like the ergodic theorem for such phenomena by imposing the grammar and syntax inherent in the Shannon-McMillan or the Rate Distortion Theorems through the limit relations defining the splitting between high and low probability sequences,

$$H[\mathbf{X}] = \lim_{n \rightarrow \infty} \frac{\log[N(n)]}{n}.$$

In the context of an appropriate parametrization, a kind of thermodynamic formalism can, then, also be imposed, but the results will usually have little relation to ordinary thermodynamics, particularly for the usual energetically open systems of most interest.

The next task is to reexamine cognitive process from an information theory perspective.

Consciousness

A Mathematical Treatment of the Global Neuronal  
Workspace Model

Wallace, R.

2005, XIII, 116 p., Hardcover

ISBN: 978-0-387-25242-1