The relationship between information theory and thermodynamics: the mathematical basis

APPENDIX 1:
Maximum Entropy Principle
Minimum Cross Entropy Principle
Exergy
&
Statistical Mechanics

James J. Kay

jjkay@uwaterloo.ca

www.jameskay.ca

Appendix 1 of: Kay, J.J. 1984. *Self-Organization in Living Systems.* Ph.D. thesis. Systems Design Engineering, University of Waterloo, Waterloo, Ontario. 458 pp.

(For discussion of the implications of these mathematical relationships, see Chapter 6 of *Self-Organization in Living Systems*.)

PART 1 - JAYNES' FORMALISM

1.1 THE MAXIMUM ENTROPY PRINCIPLE

Given a set of n discrete possible observables $\{x_i\}$ (outcomes of an experiment, events, messages etc.) and the expectation value of m variables $\{f_k[x_i]\}$ related to what is being observed, what is the probability distribution $\{p_i\}$ (or $\{p[x_i]\}$) which describes the information we have about the observables in the least biased, most honest way, and allows us to make the best prediction about the values of some other variables $g_k[x_i]$?

More formally given[1]:

$$\langle f_k[x_i] \rangle = p_i \, f_k[x_i] \quad 1 \leq k \leq m \tag{1}$$

what $\{p_i\}$ should we choose if we wish to make the best unbiased use of the information (in effect minimize the expected square of the error in estimating $\langle g_k[x_i] \rangle$)?

Jaynes (1957, 1958) showed that the answer is the distribution which satisfies (1) while maximizing the Shannon entropy:

$$S = -p_i \ln p_i = \langle -\ln p_i \rangle \tag{2}$$

Thus the name Maximum Entropy Principle [M.E.P.].

The distribution[2] which satisfies the M.E.P. is:

$$p_i = (Z[\lambda_k])^{-1} \exp[-\lambda_k f_k[x_i]] \tag{3}$$

where

$$Z[\lambda_k] \equiv \sum_{i=1}^{n} \exp[-\lambda_k f_k[x_i]] \tag{4}$$

and $\{\lambda_k\}$ are the Lagrange multipliers chosen to satisfy (1).

---

[1] Summation convention in effect for repeated indices. That is $p_i \, f_k[x_i]$ is taken to mean:

$\sum_{i=1}^{n} p_i \, f_k[x_i]$ and $\lambda_k f_k[x_i]$ is $\sum_{k=1}^{m} \lambda_k f_k[x_i]$.

[2] It is assumed that the distribution is complete.

Some useful relationships are:

$$<f_k[x_i]> = -\frac{\partial \ln Z}{\partial \lambda_k} \qquad l \leq k \leq m \tag{5}$$

$$S[\{<f_k[x_i]>\}] = \log Z + \lambda_k <f_k[x_i]> \tag{6}$$

and

$$\lambda_k = \frac{\partial S}{\partial < f_k[x_i] >} \qquad l \leq k \leq m \tag{7}$$

(6) shows that the maximum value of S is a function only of the known information.

The ramifications of the M.E.P. for the interpretation of the p_i are revolutionary. The p_i can no longer be interpreted strictly as the number of times an outcome occurs in N trials. Rather it is an assessment of the likelihood of an event occurring in one trial relative to the other possible outcomes. The philosophical issues involved and their practical implications are discussed by Jaynes (1978) and Hobson (1971). The important papers of Jaynes have been collected by Rosenkrantz (1983).

1.1.1 STATISTICAL MECHANICS

The various distributions (Maxwell, Gibbs, Bose-Einstein, Fermi-Dirac) can all be derived rigorously using M.E.P. This is an improvement over the traditional ergodic derivation which required considerable "hand waving". As examples the Fermi-Dirac and Gibbs distributions are derived.

FERMI-DIRAC

Fermions are 1/2 spin particles. Consequently the occupancy (j) of the ith orbital can be 0 or 1. p_{ij} is the probability of the ith orbital having occupancy j. The constraints are:

a) the probability of the ith orbital being occupied or unoccupied is 1.

b) $<j> = a$ (averaged over all orbitals) (8)

c) $<je_i> = b$ (e_i is the ith orbital's energy level)

The distribution is obtained from maximizing <-ln p_{ij}> subject to (8). Because the entropy of ith level is independant from the kth, <-ln p_{ij}> can be maximized by maximizing <-ln p_{ij}> for the ith level. Using (3) with f_i = j, f_2 = je_i:

$$p_{ij} = (Z_i)^{-1} \exp[-j (\lambda_1 + \lambda_2 e_i)] \qquad (9)$$

where from (4)

$$Z_i = \exp [-0 (\lambda_1 + \lambda_2 e_i)] + \exp[-1 (\lambda_1 + \lambda_2 e_i)]$$

$$Z_i = 1 + \exp [-(\lambda_1 + \lambda_2 e_i)] \qquad (10)$$

from (7):

$$\lambda_1 = \frac{\partial S}{\partial < j >} = -\mu/kT \quad \text{(definition of chemical potential)}$$

$$\lambda_2 = \frac{\partial S}{\partial < je_i >} = 1/kT \quad \text{(definition of temperature)} \qquad (11)$$

where μ=chemical potential, T=temperature and k=Boltzmann's constant.

Thus the expected occupancy of the ith orbital:

$$<j>_i = 0\, p_{i0} + p_{i1}$$

$$<j>_i = \exp [\mu/kT - e_i/kT] / (1 + \exp [\mu/kT - e_i/kT])$$

$$<j>_i = 1 / (\exp[(e_i-\mu)/kT] + 1) \qquad (12)$$

which is the Fermi-Dirac distribution. (See Kittel, 1969). Many of the properties of solid state device are calculated using this distribution.

GIBBS' DISTRIBUTION

Suppose we have a system in equilibrium with a reservoir at temperature T, chemical potential μ, pressure p, and general intensive variable λ_k. What is the probability p[N,i] of finding the system to be in a state with values N (number of particles), e_i (energy of ith state), f_k[i] (the kth general extensive variable of the ith state). It is assumed that the expectation values of the extensive variables are given.

a) $\langle e_i \rangle = U$

b) $\langle N \rangle = \overline{N}$ (the mean value of N)

c) $\langle f_k[i] \rangle = \bar{f}_k \qquad 1 \le k \le m$ (13)

Then M.E.P. gives from (3):

$p[N,i] = Z^{-1} \exp -[\lambda_k f_k[i] + \lambda_{m+1} e_i + \lambda_{m+2} N]$

where, using (7) and the thermodynamic definitions of temperature, chemical potential, and generalized intensive variables (See (11) and (13)), $\lambda_{m+1} = 1/kT$, $\lambda_{m+2} = -\mu/kT$ and the generalized intensive variable λ_k is the kth Lagrange multiplier. (It is interesting to note that the relationship (7) between intensive variables and the expectation values of the extensive variables arises naturally as a consequence of M.E.P.). It can be shown that

$$Z = \exp [pv/kT]$$

(<u>PROOF</u>:

$$S = \langle -\ln p_i \rangle = - \sum_{i=1}^{n} p[N,i] \left(-(\lambda_k f_k[i] + e_i/kT - \mu N/kT) - \ln Z\right)$$

using 13: $S = (\lambda_k \langle f_k \rangle + U/kT - \mu \langle N \rangle /kT) + \ln Z$

$$\ln Z = S - (\lambda_k \langle f_k \rangle + U/kT - \mu \langle N \rangle /kT)$$

But a thermodynamic relation is that[3]:

$$U = kTS + \mu \langle N \rangle - kT\, \lambda_k \langle f_k \rangle - PV \qquad (14)$$

Thus $S = (U/kT - \mu \langle N \rangle /kT + \lambda_k \langle f_k \rangle) + PV/kT$

Therefore $\ln Z = PV/kT$ Q.E.D.)

Using this result yields:

$$p[N,i] = \exp -[e_i/kT + PV/kT - \mu N/kT + \lambda_k f_k[i]\,] \qquad (15)$$

---

[3] See Brzustowski and Golem 1977, equation 9 or Edgerton, 1982, equation 4.27. Note that traditionally $S = k\langle -\ln p_i \rangle$ in thermodynamics.

This is Gibb's distribution. Z is the grand sum. This is the basic distribution of equilibrium thermodynamics.

These results are very striking if one is used to deriving them using classic thermodynamic arguments. The following questions are immediately brought to mind by these results.

1) Why should the maximum entropy formalism give these physical results? What is the physical (as versus statistical) interpretation of the M.E.P.?
2) What is the relationship between the second law and the M.E.P.?
3) How does one physically interpret the p_i?
4) How does one interpret the entropy measure, especially the physical meaning of other distributions which give an entropy less than the maximum entropy?
5) What is the relationship between information and thermodynamics entropy?

Before discussing an even more compelling result some more information theory must be discussed.

PART 2 - THE KULLBACK MEASURE OF INFORMATION

THE MEASURE:

Kullback (1957) developed a measure he termed the directed divergence:

$$\mathrm{I}\,[P,Q] \equiv \sum_{i=1}^{n} p_i \ln \left[\frac{p_i}{q_i} \right] \qquad (1)$$

where Q is the prior distribution and P the posterior distribution. This measure has also been called the Kullback-Liebler Discrepancy Measure and the cross-entropy. It should not be confused with the mutual information I [A,B].

Suppose some new information is obtained which causes an observer to change his estimation of the probability distributions associated with an experiment from Q to P. Then I [P,Q] measures the average change in uncertainty about the outcome of the experiment, which results from gaining the new information. This is in contrast to H[Q] - H[P] (H[X] = Shannon's entropy) which measures the change in average uncertainty due to gaining the new information. Mathematically the difference between $\mathcal{I}$ [P,Q] and H[Q] - H[P] is that the former measures the average of differences while the latter measures the difference of averages.

Some of the properties of the $\mathcal{I}$ are:

1) $\mathcal{I}[P,Q] \geq 0$, equality holds if and only if $p_i = q_i \; \forall \, i$

(This is known as Shannon's inequality).

2) Additivity and Recursivity (Aczel 1975, p.201 (7.2.7) and p. 203 (7.2.19)).

SOME DIFFERENCES OF $\mathcal{I}$ AND H:

Some of $\mathcal{I}$'s differences from H[Q] - H[P] are:

3) $\mathcal{I}[P,Q] - (H[Q] - H[P]) = \sum_{i=1}^{n} (q_i - p_i) \ln q_i$

PROOF: $\sum_{i=1}^{n} (p_i \ln p_i - p_i \ln q_i + q_i \ln q_i - p_i \ln p_i) = \sum_{i=1}^{n} (q_i - p_i) \ln q_i$

This represents the difference between $<(\ln p_i - \ln q_i)>$ and $<\ln p_i> - <\ln q_i>$.

4) Suppose the information gained pre-ordains the kth outcome of the experiment:

$p_i = \delta_{ik}$

Then $\mathcal{I}[P,Q] = -\ln q_k$.

This indicates how much information is gained, if knowledge about the experiment is obtained which indicates that only one possible outcome can occur. This is different from the information gained on average (H[Q]) from performing the experiment. This is intuitively correct because the information gained from realizing that the experiment is constrained to one outcome is different from the expectation value of the information gained from performing the experiment.

5) $\mathcal{I}[P,Q] - \mathcal{I}[Q,P]$ is indefinite.

Thus it is impossible to tell from examining $\mathcal{I}$ [P,Q] and $\mathcal{I}$ [Q,P] which distribution has the highest entropy. Thus $\mathcal{I}$ [P,Q] can only be used to detect changes in information between distributions, it cannot be used in general to indicate the direction of the changes in uncertainty. Only H[Q] - H[P] can indicate this.

(In thermodynamics $\mathcal{I}$ [P,Q] is proportional to the exergy between two systems, that is the available work. However $\mathcal{I}$ [P,Q] does not tell us which way the energy will flow. This can only be determined from examining ΔS, the change is entropy. Similarly H[Q] - H[P] must be examined to tell us what direction the information is flowing in.)

6) I [P,Q] is not symmetric (i.e. is sensitive to changes in indices).

Suppose P={a,b,c}, Q={a,c,b}

$\mathrm{I}[P,Q] = (c-b) \ln b/c \neq 0$.

whereas H[P] - H[Q] = 0.

SHANNON'S ENTROPY AS A SPECIAL CASE OF CROSS-ENTROPY

7) Hobson and Cheng (1973) argue that the Shannon entropy is a special case of the more general information measure I.

They define the Kullback uncertainty:

$$U_k [P:P^o, P^m] = \mathrm{I} [P^m, P^o] - \mathrm{I} [P, P^o] \quad (2)$$

where P is the posterior, P^o the prior, and P^m the distribution representing the maximum information which an observer could have while still being consistent with the fundamental physical constraints of the situation. $\mathrm{I}[P^m, P^o]$ measures the maximum information which can be gained by an observer about an experiment. $\mathrm{I}[P, P^o]$ represents the actual gain in information. (This gain is not due to determining the outcome of the experiment, but rather from discovering some new constraint on the experiment.) Thus U_k represents the still missing information about the experiment. U_k reduces to the Shannon entropy in the case of $p_i^m=\delta_{ij}$, j is arbitrary; P^o = the uniform distribution. Thus they argue that the Shannon entropy is special case of the Kullback uncertainty.

The central argument used to justify this claim is that the Shannon entropy is not generalizable to continuous distributions in a consistent way whereas the Kullback cross-entropy measure is (Hobson & Cheng 1973, Johnson 1979).

One problem with I is that it is not additive over arbitrary distributions:

$$\mathrm{I}[P^2, P^0] \neq \mathrm{I}[P^2, P^1] + \mathrm{I}[P^1, P^0] \quad (3)$$

whereas the Shannon entropy is:

$$(H[P^0] - H[P^2]) = (H[P^1] - H[P^2]) + (H[P^0] - H[P^1])$$

8) However Hobson and Cheng (1972) show that:

$$I[P^2, P^0] = I[P^2, P^1] + I[P^1, P^0] \quad (4)$$

where P^0 is arbitrary, P^1 is the maximum entropy distribution for a set of constraints f_1 and P^2 is the maximum entropy distribution for the set of constraints $f_1 \cap f_2$. Thus if our knowledge of a situation is being refined (in accordance with the maximum entropy principle) from $P^0 \rightarrow P^1 \rightarrow P^2$ then I is additive.

THE MINIMUM CROSS-ENTROPY PRINCIPLE

Johnson (1979), Shore and Johnson (1980,1981) and Shore (1982) have explored the use of the cross-entropy (I) and have shown rigorously (and quite readably) that "Jayne's principle of maximum entropy and Kullback's principle of minimum cross-entropy (minimum directed divergence) provide correct general methods of inductive inference when given new information in the form of expected values" (p.34, Shore and Johnson 1980)

9) This means that given a distribution $P^o[x]$ and some new information in the form of constraints:

$$\int dx\, P'[X]\, C_k[X] \geq 0 \quad k=1,\ldots,m \quad (5)$$

then the new distribution $P'[X]$, which incorporates this information in the least biased way and which is arrived at in a way which does not lead to any inconsistencies or contradictions, is the one obtained from minimizing:

$$\int dx\, P'[X] \ln \left[\frac{P'[X]}{P^o[X]} \right]. \quad (6)$$

This is the MINIMUM CROSS-ENTROPY PRINCIPLE (Shore and Johnson, 1980).

Furthermore they show that the Maximum Entropy Principle is a special case of the Minimum Cross-Entropy Principle in the following way. Suppose we are trying to estimate the probabilities of finding a system in a state X. If we know that only n discrete states are possible we already have some information about the system. This expressed by $P_i^o = 1/n \; \forall \; i$. If we obtain more information in the form of an inequality, in (5), then the correct estimate of the probabilities of the system being in a state i is given by minimizing:

$I\,[p', P^o] = \; p_i' \ln[p_i'/p_i^o] = \; p_i' \ln p_i' - \ln n$

which is equivalent to maximizing the entropy

$H[P'] = - \; p_i' \ln p_i'$.

Shore and Johnson (1981) develops a number of properties of cross-entropy minimization. Of particular importance is their generalization of (4).

10) $I\,[P^k, P^o] \geq I\,[P^k, P'\,] + I\,[P', P^o]$ (triangle relations)

where P^o is the original distribution, P' is the new distribution derived from the minimum cross-entropy with P^o and which satisfies inequality constraints as in (5) and P^k is any distribution which satisfies the same inequality constraints as P'. Equality occurs if the constraints on the system (experiment) are strict equalities. (In fact this rarely happens in practice, as P^k must satisfy the same moment constraints as P'.)

Another important property is that:

11) $I\,[P^T, P^o] \geq I\,[P^T, P'\,] + I\,[P', P^o]$

where P^T is the "true" distribution associated with the system/experiment. This shows that the distribution P' is closer to the P^T than P^o is. Again equality occurs for strict equality constraints.

RECAPITULATION

Suppose a system is being observed. We wish to know what state the system is in. The probability of finding the system in state X is P[X]. This distribution reflects our state of knowledge about the system. P^T is the distribution obtained when we know everything we can know about the system.

The problem addressed by the Minimum Cross-Entropy Principle is this. Suppose we have initially a distribution P^o [X]. Now suppose we gain more information about the system, that is constraints on the state (X) of the system. How can we incorporate this new information into a probability distribution in a consistent way? (This problem is the problem of inductive inference).

(By consistent is meant that different ways of using the same information, should lead to the same results. This notion is formally defined in Shore and Johnson 1980 and leads to fifteen properties of the inference technique of minimizing cross-entropy (Shore & Johnson 1981). In particular iteratively incorporating information from different constraints sets will lead to the same distribution regardless of the order of the iterations).

The solution is to pick the distribution which minimizes the Kullback divergence or cross-entropy:

$$I[P', P^o] = \int dx\, p'[x] \ln[p'[x] / p^o[x]]$$

subject to the constraints

$$\int dx\, p'[x]\, f_k[x] \geq \langle f_k \rangle \quad k=1,...,m.$$

The solution is unique.

Furthermore:

$$I[P^T, P^o] \geq I[P^T, P'] + I[P', P^o]$$

(Equality occurs when all the constraints are strict equalities) Thus P' is closer to P^T than P^o is.

The amount of information remaining to be found (i.e. the uncertainty) is given by:

$$I[P^T, P^o] - I[P', P^o] \geq I[P^T, P'].$$

In the case of knowing that only n possible states exist, and nothing else, P_i^o = the uniform distribution. In this case, any new information in the form of constraints is incorporated by minimizing:

$$I[P', P^o] = p_i' \ln p_i' + \ln n$$

$$= H[1/n] - H[P']$$

This is equivalent to maximizing the Shannon entropy. Thus if the prior distribution is {1/N}, the minimum cross-entropy and maximum entropy principles are equivalent for discrete distributions.

If one looks at inductive inference as an iterative problem of discovering information about the constraints operating on the system's state, then the process is characterized by refining one's probability distribution in accordance with the minimum cross-entropy principle. If P^N is the distribution after the Nth iteration then we have the following: (This assumes equality constraints).

$I[P^T, P^o] = I[P^T, P^N] + I[P^N, P^{N-1}] + I[P^{N-1}, P^{N-2}] + ... + I[P^1, P^0]$.

$I[P^N, P^{N-1}]$ corresponds to the information gained by the Nth iteration. $I[P^T, P^N]$ corresponds to the remaining information to be gained, i.e. the remaining uncertainty.

Once the "true" distribution has been obtained no more information can be gained without physically changing the constraints on the system so as to further decrease our uncertainty. (In fact the scientific method essentially consists of sufficiently constraining a system so it can only be in one state. A constraint is then modified (an expectation value changed) so as to put the system in new state. This is done to test an hypothesis which relates constraint changes to state changes. However the problem considered here, that is discovering the constraints which naturally occur, is a very different problem. It is a problem which involves discovering the constraints which determine the states in which a system can actually exist. This problem and the associated methodology should also be recognized as a valid pursuit of scientists. In quantum physics our limit to constraining the system has been reached. Perhaps physicists' time would be better spent defining these limits rather than chasing the elusive "elementary" particles. Such particle discoveries represent constraining the system so that only one isolated mass-energy-spin etc. state can be observed. Each new particle simply indicates that a physicist has discovered a new way of constraining the system.)

Once the inference process is complete, that is the "true" distribution (P^T) has been found, then all that can be measured is the information we will gain on average from determining the system state at a specific point. (i.e. determining the outcome of one trial of an experiment). This uncertainty is measured by calculating

$I[P^M,U] - I[P^T,U]$

where P^M is the distribution which reflects our knowledge of the system at the point when we do the experiment, that is know the state of the system (usually $\delta[X-X^k]$, except in the case of limitations such as the Heisenberg uncertainty principle), and U is the uniform distribution. In the case of a discrete number of states

$I[P^M,U] - I[P^T,U] = H[P^T]$.

To summarize the cross-entropy is a general information measure of the difference between two distributions. Given information in the form of constraints on the system's states, the Minimum Cross-Entropy principle governs how to use this information. In the case of discrete distributions the Maximum Entropy Principle leads to the same results as the Minimum Cross-Entropy Principle. Thus in the context of the inference problem the Shannon entropy and Maximum Entropy Principle are special cases of the cross-entropy and the Minimum Cross-Entropy Principle.

However once all of the constraints have been determined and the "true" distribution is known, the cross-entropy alone can only tell us how much information can be gained from determining the specific state of a system at a point. If we wish to know how much uncertainty we have on average about what state the system will be found in we must use $I[P^M,U] - I[P^T,U]$. In the discrete case this is the Shannon entropy . Thus, in the case of determining the uncertainty associated with an experiment, the general measure is the differences of the cross-entropies. In the limited case of a discrete distribution this measure simplifies to the Shannon entropy.

POSTSCRIPT

While reviewing the literature on cross-entropy, K. Palaniappan and I were not happy with the implications of (3) and Property 10. Some further investigation and reflection lead to the following results. Our review of the literature and discussions with faculty here, who are knowledgeable in this area, lead us to believe that these results are original.

The fundamental definition of the Shannon entropy is $-<\ln p_i>$. The cross-entropy can be defined as:

$$<(\ln p_i - \ln q_i)> \qquad (7)$$

If we treat I(P,Q) in terms of expectation values then Property 10 becomes:

$$<(\ln p_i^k - \ln p_i^o)> \; \geq \; <(\ln p_i^k - \ln p_i')> + <(\ln p_i' - \ln p_i^o)> \qquad (8)$$

where the left hand side and the first term on the right are expectation values over the P^k and the last term is with respect to the P'. In this form, it is not intuitively disturbing that the cross-entropy is not additive as new information is obtained as the expectation values are over different distributions.

However if we calculate all the expectation values using the current best estimate of the "true" distribution, then it would be intuitively satisfying if the information gain was additive as new information is obtained. That is:

$$<(\ln p_i^k - \ln p_i^o)> = <(\ln p_i^k - \ln p_i')> + <(\ln p_i' - \ln p_i^o)> \qquad (9)$$

In fact this is the case for:

$$\sum_{i=1}^{n} p_i^k \ln[p_i^k / p_i^o] = \sum_{i=1}^{n} p_i^k \ln[p_i^k / \ln p_i'] + \sum_{i=1}^{n} p_i^k \ln[p_i' / \ln p_i^o] \qquad (10)$$

This result is independant of what constraints any of the distributions satisfy.

THE IMPLICATION OF THIS RESULT IS THAT THE CORRECT INFORMATION MEASURE TO USE IS GIVEN BY (9) RATHER THAN BY (1). The difference between them is in which distribution is used to calculate the expectation values. Our argument is that the best estimate of the "true" distribution should be used to calculate distances between distributions.

If the distributions P^k and P' satisfy the same equality constraints then equality holds in Property 10. Therefore (10) implies that:

$$\sum_{i=1}^{n} p_i^k \ln[p_i' / \ln p_i^o] = \sum_{i=1}^{n} p_i' \ln[p_i' / \ln p_i^o] \qquad (11)$$

This is a curious result and bears further consideration. It means that the distance between the Minimum Cross-Entropy distribution and the prior is independant of the distribution used to calculate the expectation value as long as the distribution also satisfies the same constraints as the Minimum Cross-Entropy distribution.

It should also be noted that equality holds in Property 10 if both P^k and P' satisfy the same equality constraints. However this is NOT an if and only if statement. For example the numerical example in Hobson and Cheng (1972), which is used to demonstrate the different cases in (3) and (4), is an example of distributions which do not satisfy the same constraints but for which (4) holds. Hobson and Cheng seemed to have missed this point which undermines their example.

PART 3 - EXERGY AND ESSERGY

Exergy[4] is a measure of the maximum work that can be extracted from a system if is is allowed to come to equilibrium with its environment.

$\text{Exergy} = U - kT^o S + P^o V - u^o <N> + kT^o \lambda_k^o <f_k>$

where superscript "o" denotes the environment value of the variable. (Brzustowski and Golem, 1978, Ahern, 1980). In effect exergy measures the quality of energy.

Let us assume that the Gibb's distribution describes the system when it is in equilibrium with its environment. What is the distance of the current state of the system from the environment state in terms of the cross entropy?

$$I\,[P,Q] = \sum_{i=1}^{n} p[i,N] \ln \frac{p[i,N]}{q[i,N]}$$

using equation 15, section 1.1:

$I\,[P,Q] = -S + p[i,N]\,(e_i/kT^o + P^o V/kT^o - u^o N/kT^o + \lambda_k^o f_k[i])$

using equation 14, section 1.1:

$+kT^o I\,[P,Q] = -kT^o S + U + P^o V - u^o <N> + kT^o \lambda_k^o <f_k>$

Therefore

$$\text{Exergy} = kT^o I\,[P,Q]$$

4 See http://www.jameskay.ca/about/exergy.html

The term I [P,Q] has been called essergy by Evans (1969). (More discussion can be found in Edgerton, 1982).

This means that the available work (exergy) is proportional to the average difference in information required to discriminate between the thermodynamic system and the environment.

REFERENCES and BIBLIOGRAPHY

Aczel, J. and Daroczy, Z. On Measures of Information and their Characterizations. Academic Press; 1975.

Ahern, J. E. The Exergy Method of Energy Systems Analysis. New York: J. Wiley & Sons; 1980.

Brzustowski, T.A., Golem, P.J.; "Second Law Analysis of Energy Processes Part 1: Exergy-An Introduction"; Transactions of the C.S.M.E., Vol.4, #4, pp. 209-208, 1978.

Costa de Beauregard, O.; Tribus, M. Information Theory and Thermodynamics. Helvetica Physica Acta; 1974; 47: 238-247.

Cyranski, John F. Analysis of the Maximum Entropy Principle "Debate". Foundations of Physics; 1978; 8(5/6): 493-506.

Cyranski, John F. Quantum Measurement of a Communication with Nature. Foundations of Physics; 1978; 8(11/12): 805-821.

Edgerton, Robert H. Avalable Energy and Environmental Economics. D.C. Heath; 1982

Eriksson, K.E.; Lindgren, K.; Mansson, B.A. Structure, Context, Complexity, Organisation: Physical Aspects of Information and Value. Singapore: World Scientific; 1987; ISBN: 9971-50-023-4.

Evans, R. B. A Proof that Essergy is the Only Consistent Measure of Potential Work (for Chemical Systems). Hanover, New Hampshire: Thayer School of Engineering, Dartmouth College; 1969I have a copy.

Forte, B.; Sempi,C. Maximizing Conditional Entropies: a Derivation of Quantal Statistics. Rendiconti di Matematica(4): Serie VI; 1976; 9: 521-566.

Hobson, A. Concepts in Statistical Mechanics: Gordon and Breach; 1971; ISBN: 0-677-03240-4.

Hobson, A. and Cheng, B. A Comparision of the Shannon and Kullback Information Measures. J. of Statistical Physics. 1973; 7(4):301-310.

Jaynes, E. T. Foundations of Probability Theory and Statistical Mechanics. Bunge, M., Ed. Delaware Seminar in the Foundations of Physics: Springer-Verlag; 1967: 77-101.

Jaynes, E. T. Gibbs vs. Boltzman Entropies. Am. J. Phys.; 1965; 33: 391-398.

Jaynes, E. T. Information Theory and Statistical Mechanics. Physics Review; 1957; 108: 171-190.

Jaynes, E. T. Information Theory and Statistical Mechanics. Physics Review; 1957; 106: 620-630.

Jaynes, E. T. Information Theory in Statistical Physics. Ford, K.W., Ed. Brandeis Lectures in Modern Physics: Benjamin; 1963: 182-218.

Jaynes, E. T. The Minimum Entropy Production Principle. Ann. Rev. Phys. Chem.; 1980; 31: 579-601.

Jaynes, E. T. Where Do We Stand on Maximum Entropy. Levine, R.; Tribus, M., Eds. The Maximum Entropy Formalism: MIT Press; 1979: 15-118.

Johnson, R. Axiomatic Charaterizations of the Directed Divergences and their Linear Combinations. IEEE Transactions on Information Theory; 1979; IT-25(6): 709-716.

Kapur, J.N.; Kesavan, H.K. Entropy Optimization Principles with Applications. Toronto: Academic Press; 1992; ISBN: 0-12-397670-7.

Kittel, C. Thermal Physics. J. Wiley & Sons; 1969.

Levine, R.; Tribus, M., Eds. The Maximum Entropy Formalism: MIT Press; 1979.

Rosenkrantz, R.D., Compiler. E.T. Jaynes: Papers on Probabilty, Statistics, and Statistical Physics: D. Reidel; 1983.

Shore, J. Cross-Entropy Minimization Given Fully Decomposable Subset and Aggregate Constraints. IEEE Transactions on Information Theory; 1982; IT-28(6): 956-961.

Shore, J.; Johnson, R. Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. IEEE Transactions on Information Theory; 1980; IT-26(1): 26-37.

Shore, J. ; Johnson, R. Properties of Cross-Entropy Minimization. IEEE Transactions on Information Theory; 1981; IT-27(4): 472-482.

Tribus, Myron. Information Theory as a Basis for Thermostatics and Thermodynamics. J. Applied Mechanics; 1961; March: p.1-8.

Tribus, Myron; Rossi, R. On the Kullback Information Measure as a Basis for Information Theory: Comments on a Proposal bt Hobson & Cheng. J. of Statistical Physics; 1973; 9(4): 331-338.

Tribus, Myron; Shannon, R.T.; Evans, R. Why Thermodynamics is a Logical Consequence of Information Theory. A.I.Ch.E.J; 1965; 12(3): 244-248.

