THE CONCEPTS OF STATE AND ENTROPY IN QUANTUM MECHANICS

by

G. DAVID FORNEY, JR.

B.S.E., Princeton University
(1961)

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June, 1963

Signature of Author ______________________________
Department of Electrical Engineering, May 17, 1963

Certified by ______________________________
Thesis Supervisor

Accepted by ______________________________
Chairman, Departmental Committee on Graduate Students

THE CONCEPTS OF STATE AND ENTROPY IN QUANTUM MECHANICS

by

G. DAVID FORNEY, JR.

Submitted to the Department of Electrical Engineering on May 17, 1963 in partial fulfillment of the requirements for the degree of Master of Science.

ABSTRACT

The central result of this thesis is a theorem which states that the entropy of a quantum mechanical system is an upper bound to the average amount of information obtainable about that system by any possible measurement.

This theorem suggests an operational definition of entropy as the maximum information obtainable about a system by any permissible measurement. While this definition is equivalent to the conventional one when all measurements are permitted, it is, unlike the conventional, equally applicable to situations in which only crude measurements of the type actually feasible are considered. It therefore promises a general elucidation of the second law of thermodynamics.

The theorem also allows a statement of the limits imposed in principle by the quantum mechanical laws upon all communication through physical media. The capacity of a quantum mechanical medium is shown to be finite when a finite average energy constraint is imposed, even in the absence of noise.

Thesis Supervisor: Robert G. Gallager
Title: Assistant Professor of Electrical Engineering

TABLE OF CONTENTS

PREFACE

This thesis began as an investigation into the fundamental limits on communication imposed by the laws of quantum mechanics, even in the absence of what is ordinarily called noise. Answers to most of the questions one can ask about a single transmission turned out to be derivable from a single theorem, which appears in the section on "Entropy" and is the core of this thesis.

The limits thus obtained are of theoretical interest, but will probably never be remotely approached in practice. We have therefore contented ourselves with minimal development of this result.

However, this theorem, when finally proved, immediately suggested an operational definition of the entropy of physics. Because of its practically unlimited generality and conceptual clarity, this definition may prove quite interesting for statistical mechanics and thermodynamics, especially if its promise of an elucidation of the second law of thermodynamics for non-equilibrium situations is realized. Again this development seems primarily of theoretical interest, but then the second law itself, except perhaps in the design of heat engines, has been valued not for its practical consequences but for its theoretical richness. Because this development seemed of more fundamental interest than the results concerning the limits on communication, this paper has been organized with emphasis on the former, even though there has not yet been time to follow out its consequences.

I have benefited from many conversations during the course of this work. In particular I should like to thank Professors R. G. Gallager of the M.I.T. Electrical

Engineering Department and K. A. Johnson of Physics for their comments, time and interest. I gratefully acknowledge also the financial support provided by an M.I.T. Lincoln Laboratory Staff Associateship.

THE CONCEPT OF "STATE"

Since the word "state" has served a variety of purposes in its scientific career, it is perhaps best to begin with an explanation of how we shall use it, especially since our use is probably unfamiliar, though not novel.

Generally the state of a physical system has been thought to have been an attribute of the system itself, independent of how much experimental knowledge was available about it. In classical physics, the state of a system with N degrees of freedom could be summarized at any time by the values of 2N coordinates; knowledge of the state of the system allowed certain prediction of the result of any measurement. In quantum mechanics, the state of a system is represented by a wave function, or by a complex vector in Hilbert space;[1,2] knowledge of the state of the system then allows predictions about the results of measurements which are in general probabilistic.

Thermodynamics and statistical mechanics use state in a somewhat broader sense, in which the state of the previous paragraph is a special case, sometimes called a pure state.[2,3,4] A statistical state describes one of an ensemble of identically prepared systems, sometimes called a Gibbsian ensemble, and allows prediction of the average behavior of a number of such identically prepared systems. There is frequently ambiguity about whether a statistical state is a description of the entire ensemble, of one of the systems in the ensemble, or of our knowledge about the system. For instance, the probability that any member of an ensemble will deviate more than a certain amount from the average is often discussed, as though each of the members of the ensemble were "actually" in a pure state. On the other hand, additional side information about a system will generally change the mathematical representation we give its state, as though the state were a summary of our knowledge.

We will take the state of a system to be equivalent to the totality of our knowledge about the system: what kind of system it is, where it is approximately located, what the results of past measurements on it have been, what its relation is to the rest of the universe, etc. We will therefore never find it necessary to speak of the "actual" state of some external reality, but only of the relationship between observed data and our knowledge.

Our knowledge may be sufficient to let us make predictions about the results of measurements on the system or the observed behavior of the system, or it may not; the business of science is to extend the class of states for which the former is true by collecting and systematizing data about the behavior of archetypical systems.

The validity of predictions is sometimes debatable, depending upon the assumptions that have been introduced. Strictly speaking, of course, as philosophers since Hume have noted, the making of any prediction can never be logically justified, since one can never demonstrate that the future will be like the past. However, at least with respect to sufficiently uncomplicated systems, the experience of mankind has invariably been the following. If we make a large number of like measurements, by which we mean experiments carried out in exactly the same way at different times and places, on systems in like states, by which we mean systems about which our knowledge at the time of the measurement was the same, and list the results, then the relative frequency of appearance of any particular result in any random selection from this list will be independent of the selection was made, to an approximation which improves as the number of results in the selection increases. We call this experience the local homogeneity of space and time, and extrapolate it to the future in this way: if the relative frequency of result b in the past has been approximately p_b, then we say the probability of result b in the future is p_b. Such a prediction is entirely valid or else science is hopeless.

It is more common, however, for us to be confronted with a system which is not exactly like systems we have previously observed, but sufficiently similar so that we assume that predictions derived on the basis of previous observations of similar systems can be extended to cover the new system. The formulation of general principles which guide such assumptions is a matter of enormous practical importance; however, the validity of predictions made in this way is always subject to experimental verification.

States can thus be classified as novel, predictable, or assumed predictable, depending on whether the probabilities of results of measurements are unknown, known as a result of exhaustive experimentation on identically prepared systems, or assumed to be known on the basis of partial information and general assignment principles. Predictable states rarely exist in practice, but are nonetheless the idealizations upon which scientific theories must be based. These mathematical theories must then be accompanied by more or less shadowy principles governing the assignment of a particular mathematical representation to an actual system; if such assignment leads to incorrect predictions, either the assignment principles or the theory itself must be modified. Henceforth we shall be primarily concerned with predictable states, which means that we shall be able to use the terminology and formalism of quantum statistical mechanics. However, the reader should not forget the distinction between the state, which is the totality of our knowledge about a system, and the mathematical representation of a state, which is assigned on the basis of more or less arbitrary principles.

Two systems cannot be in the same state, for then they would be indistinguishable and we could not speak of two systems. Two predictable states will be called equivalent with respect to a certain set of measurements, however, if the probabilities of the different results of any measurement in the set do not depend on which state actually exists. In a mathematical theory which gives rules for making predictions, we clearly have the freedom of assigning equivalent states the same mathematical representation if convenient.

Suppose we have a system about which the totality of our knowledge, (i.e. the state) is the following: it is in state one with probability p, or it is in state two with probability $1 - p$, where states one and two are not equivalent and p is not equal to zero or one. Any state equivalent to such a state we call decomposable, always with respect to a certain set of measurements; a state which is not decomposable we call pure. Since each component state may itself be decomposable, a predictable state can generally be decomposed into a large number of component states, which we call hypotheses, each with its own probability; nor will this decomposition be unique. (This composition of hypotheses, each with its own probability, has nothing to do with the expression of a pure state or wave function as a superposition of other pure states, each with a complex amplitude, which one encounters in quantum mechanics.)

If a system is in a state which is not predictable, we may only say whether hypotheses are consistent with it or not; there is no way of assigning probabilities to component hypotheses since one cannot determine to which states the novel state is equivalent.

MEASUREMENTS

The measurements made by physicists include an enormous variety of operations ending with readings or a set of readings. Systematization of this variety requires groupings of some kind. Two measurements are clearly equivalent if for some ordering of their outputs, the probability of the i^{th} output of one is equal to the i^{th} output of the other for any state of the measured system and for all i. Further, measurement B is said to dominate measurement A if they are not equivalent and one can define a set of transition probabilities p_{ij} such that for any predictable state the probability of observing result a_j of A is $\sum_i p_{ij}$ Pr (b_i), where Pr (b_i) is the probability of observing result b_i of B. A measurement dominated by no other is called complete.

In the classical mechanics all complete measurements were equivalent, and their outputs generally fell in a continuous range. Quantum mechanics recognizes an uncountably infinite set of non-equivalent complete measurements, each with a countable number of discrete outputs. Actual measurements are considered more or less crude approximations to these idealized measurements; the probabilities of actual results are calculated by defining the complete measurement which dominates the actual measurement and the appropriate transition probabilities. Although the complete measurements are idealizations, quantum mechanics assumes that any of them can be approximated arbitrarily closely by some real experimental operation, and therefore admits the consideration of complete measurements in thought experiments.

After an approximation to a complete measurement A has been made and the result a_j observed, the state of the system is experimentally found to be such that the repetition of measurement A will yield a_j with certainty. This does not say anything about what the results of other complete measurements might be; such measurements might well depend on the state of the system prior to the first measurement. It is a remarkable fact that this is not so; all states in which the result of a complete measurement A can be predicted with certainty as a_j are apparently equivalent. It follows that such a state is a pure state; therefore, the performance of any measurement and observation of the result leaves the system in a pure state, independent of the prior state. What is equally remarkable is that any states can be decomposed into such pure states or are equivalent to them;[4] therefore these are the only pure states. This decomposition is not unique; it is always possible, however, to find a decomposition into pure states all of which are associated with the same complete measurement. Any such measurement (there may be many) is called the characteristic measurement of the state.

Finally, quantum mechanics sometimes finds it convenient to consider an ideal measurement whose outputs lie in a continuous range, such as the infinitely precise position measurement. Such measurements are useful for calculating probability distributions for the results of actual, approximate position measurements; however, in order to calculate the probability of a particular output of an actual position measurement, one must always turn to the complete discrete measurement which dominates it. We will therefore not find it necessary to consider these continuous ideal measurements in thought experiments.

MATHEMATICAL PRELIMINARIES

In order to proceed into the mathematical formulation of quantum mechanics, some mathematical background is required.

All predictions, it is postulated, can be expressed in terms of traces of square infinite complex matrices. The distinction between infinite and finite matrices is not trivial, and care is required in extending results true for finite matrices to infinite matrices.

The trace is the sum of the diagonal elements of a matrix. Its important properties are that $\mathrm{Tr}(A + B) = \mathrm{Tr}\,A + \mathrm{Tr}\,B$, $\mathrm{Tr}\,\lambda A = \lambda\,\mathrm{Tr}\,A$ (λ a constant), and, for most cases, $\mathrm{Tr}\,AB = \mathrm{Tr}\,BA$.

A consequence of the above postulate is that no prediction is changed if all the matrices appearing in the pertinent law undergo the same unitary transformation, since the trace of a matrix is invariant under a unitary transformation, otherwise called a change of representation. A unitary transformation may itself be represented by a complex matrix of the same dimensionality as the matrices being transformed. A necessary and sufficient condition that a matrix U_{jk} represent a unitary transform is that $\sum_j U_{jk} U^*_{jk'} = \sum_j U_{kj} U^*_{k'j} = \delta_{kk'}$; that is, that $UU^* = U^*U = I$, where U^* is the conjugate transpose of U and I is the identity matrix $I_{jk} = \delta_{jk}$. $\mathrm{Tr}UMU^* = \mathrm{Tr}U^*UM = \mathrm{Tr}M$, which shows the invariance of the trace.

A matrix M is Hermitian if $M_{ij} = M^*_{ji}$. A Hermitian matrix M satisfying certain conditions of boundedness can be put into diagonal form by some unitary transformation U; that is, $UMU^* = D$, where $D_{jk} = \lambda_k \delta_{jk}$. The elements appearing on the diagonal of D, the λ_k, are called the eigenvalues of M and are real. Since the trace is invariant under a unitary transformation, $\mathrm{Tr}M = \mathrm{Tr}D = \sum_k \lambda_k$, the sum of the eigenvalues of M.

If all eigenvalues are positive, the matrix is called positive definite; if all are non-negative, non-negative definite. It is easily shown that the diagonal elements of a positive (non-negative) matrix are positive (non-negative) in any representation.

We can define a matrix f(D) which is a function of a diagonal matrix D by the relation $\left[f(D) \right]_{jk} = f(\lambda_k)\delta_{jk}$. If the values of the function are complex numbers, and $M = UDU^*$, then we define $f(M) = U f(D) U^*$.

An especially interesting class of matrices is the set of projection operators, defined as all diagonalizable matrices P for which $P^2 = P$. Clearly the eigenvalues $\lambda_k^2 = \lambda_k$, and thus are all either 0 or 1. This implies that all projection operators are Hermitian and non-negative definite. TrP is the rank of the projection operator; if the rank of P is one we call it a one-dimensional projection operator or simply a projector.

Two projection operators for which $P_1 P_2 = 0$ are called orthogonal; since the projection operators are non-negative definite, $\mathrm{Tr}\, P_1 P_2 = 0$ implies that P_1 and P_2

are orthogonal. If P_1 and P_2 are projection operators, $(P_1 + P_2)$ is a projection operator if and only if P_1 and P_2 are orthogonal; for if $(P_1 + P_2)^2 \equiv P_1^2 + P_2^2 + P_1P_2 + P_2P_1 = P_1 + P_2$, $P_1P_2 = -P_2P_1$, and $TrP_1P_2 = TrP_2P_1 = -TrP_2P_1 = 0$. A set of projectors such that $\sum_k P_k = I$ is called a complete set of projectors; since I is itself a projection operator ($I^2 = I$), it follows that any two members of such a set are orthogonal.

Probabilities are central to any formulation of quantum mechanics and absolutely fundamental to this one. At this point we need merely note that if all possible events of a certain class are divided into disjoint subsets, the probabilities of each subset will be non-negative and real, and the probabilities of all subsets will sum to one.

MATHEMATICAL EXPRESSION OF THE QUANTUM MECHANICAL LAWS

We now have enough mathematics to put some physical flesh on our fundamental structure. This will involve giving some mathematical expression to the concepts of state, measurement, and the results of a measurement.

A complete measurement M is represented by a complete set of projectors P_j and a set of different results λ_j; conversely, each complete set of projectors represents a complete measurement. Sometimes we will write $M = \sum_j \lambda_j P_j$. The representation in which all the P_j are diagonal will be called the M representation. The matrix M is Hermitian if the λ_j are real and non-negative definite if the λ_j are non-negative.

A pure state which describes a system immediately after an M measurement has yielded λ_j is represented by the corresponding projector P_j. The probability of observing λ_k in state P_j is $\delta_{jk} = TrP_jP_k$. (For those familiar with the Dirac notation, if $|\lambda_j\rangle$ is the eigenstate corresponding to the eigenvalue λ_j, $P_j = |\lambda_j\rangle\langle\lambda_j|$.)

The fundamental law governing the results of measurements can now be stated. For any complete measurement $M' = \sum_k \lambda'_k P'_k$, when the state is P_j, the probability of observing λ'_k is $\mathrm{Tr} P'_k P_j$. The fact that this relatively simple representation of states and measurements combined with such a simple law can predict all small-scale phenomena accurately accounts for much of the elegance of quantum mechanics.

For a general state ρ whose characteristic measurement is M, and for which the probability of observing λ_j is $p_j = \sum_k p_k \mathrm{Tr}\, P_j P_k = \mathrm{Tr}\, P_j (\sum_k p_k P_k)$, we define $\rho = \sum_k p_k P_k$ so that $\Pr(\lambda_j) = \mathrm{Tr} P_j \rho$. It follows then that $\Pr(\lambda'_k) = \mathrm{Tr} P'_k \rho$. for any measurement. The matrix ρ is known as the density matrix.[4] ρ is diagonal in the M representation where its j^{th} diagonal element $\rho_{jj} = p_j$. Since probabilities are real, non-negative, and sum to 1, ρ is Hermitian, non-negative definite, and has unit trace.

For a measurement N which is not complete, we can calculate the probability q_k of the k^{th} result η_k from the probabilities ρ_{jj} of the results of any complete measurement which dominates it and the appropriately defined transition probabilities p_{kj}; $q_k = \sum_j \rho_{jj} p_{kj} = \sum_j p_{kj} \mathrm{Tr} P_j \rho = \mathrm{Tr}\, \rho (\sum_j p_{kj} P_j)$. Defining the matrix $Q_k = \sum_j p_{kj} P_j$, we can then write the probability of observing λ_k as $\mathrm{Tr} Q_k \rho$. Sometimes we will write $N = \sum_k \eta_k Q_k$.

The mean value of any complete measurement which has numbers for results can be defined as $\overline{M} = \sum_j \lambda_j \Pr(\lambda_j) = \sum_j \lambda_j \mathrm{Tr}\, P_j \rho = \mathrm{Tr}\, M\rho$; similarly $\overline{N} = \mathrm{Tr}\, N\rho$.

The continuous ideal measurements mentioned above are represented by non-diagonalizable Hermitian matrices. One can be strict and insist that they have no eigenvalues, or loose and say they have a continuous range of eigenvalues. In either case, one recognizes that to predict their outputs one uses a continuous probability distribution rather than a discrete set of probabilities. Even when X is a non-diagonalizable matrix, we can in general compute X^2, X^3, etc., and thus for any state ρ find $\overline{X} = \mathrm{Tr}\,\rho X$, $\overline{X^2} = \mathrm{Tr}\,\rho X^2$, etc., whenever these sums are convergent. The existence of all moments is sufficient to determine the probability distribution for x, although it would be misleading to claim that this is the easiest way to do so, the Schrodinger wave mechanics being ideally suited to this purpose.

Although there is assumed to be a physical measurement corresponding to every representation, it is seldom obvious just which laboratory measurement corresponds to which representation. For measurements of electron spin, however, we can make such a correspondence easily. It is experimentally found that when an electron moving in the x direction passes through a magnetic field which is homogeneous in the y direction but inhomogeneous in z, the electron is deflected either up or down; all electrons of the same initial velocity take one or the other of the two paths. Further, no matter what the original state of an electron which has been deflected up, a second measurement of the same type but at an angle θ with the first will deflect the electron along its up axis with probability $\cos^2 \theta/2$, down with $\sin^2 \theta/2$; in particular, if $\theta = 0$, the electron will certainly be deflected up again. This electronic behavior in an inhomogeneous magnetic field is assumed due to an internal spin of the electron, which couples with the field, and the experiment described above, known as the Stern-Gerlach experiment, is assumed a good approximation to a complete measurement of the spin in the z-direction. The spin state of an electron is represented by a 2×2 density matrix (there being only two possible results of any electron spin measurement); each representation of this matrix corresponds to a different orientation of the measuring apparatus. In particular, in the representation corresponding to the z-spin measurement, the state of the electron is either $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ or $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ after a z-spin measurement, depending on whether the electron was observed to go up or down. A real orthogonal transformation $U = \begin{bmatrix} \cos\theta/2 & -\sin\theta/2 \\ \sin\theta/2 & \cos\theta/2 \end{bmatrix}$ of the matrix changes it to a representation corresponding to a measurement of the spin in a direction in the y-z plane at an angle θ with the z-axis:

$$U\rho_1 U^* = \begin{bmatrix} \cos\theta/2 & -\sin\theta/2 \\ \sin\theta/2 & \cos\theta/2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \cos\theta/2 & \sin\theta/2 \\ -\sin\theta/2 & \cos\theta/2 \end{bmatrix} = \begin{bmatrix} \cos^2\theta/2 & 1/2\sin\theta \\ 1/2\sin\theta & \sin^2\theta/2 \end{bmatrix}$$

with the consequence that the probabilities are $\cos^2\theta/2$ and $\sin^2\theta/2$ for such a measurement. Complex unitary transformations correspond to rotations into the x-direction as well.

Another archetypical system is the harmonic oscillator, often used in quantum mechanical treatments of the electromagnetic field. A harmonic oscillator is specified by its frequency ω; its energy measurement is complete and has the eigenvalues $(n - 1/2)\hbar\omega$, $n = 1, 2, \ldots$, and therefore its density matrix is infinite dimensional. In the energy representation the infinitely precise position measurement is represented by the undiagonalizable matrix X with elements $[X]_{nn'} = i\sqrt{\frac{\hbar}{2M\omega}}\left(\sqrt{n}\,\delta_{n,n'+1} - \sqrt{n'}\,\delta_{n+1,n'}\right)$. If the system is in the energy eigenstate corresponding to the eigenvalue $1/2\,\hbar\omega$ $(n = 1)$, for example, $\overline{X^m} = [X^m]_{11}$, the first matrix element of X^m. It is readily verified that all odd moments are zero, while the even moments are $(\hbar/4M\omega)^{m/2}\,m!/(m/2)!$. These are the moments of a gaussian distribution with zero mean and variance $\hbar/2M\omega$, so that the probability distribution $p(x) = \sqrt{\frac{M\omega}{\pi\hbar}}\,\exp\left[-\frac{M\omega x^2}{\hbar}\right]$.

THE MEASURE OF INFORMATION[5]

The idea that the information content of a message is a measurable quantity has been a stimulating one. In their excitement, writers have applied the notions of information theory to the most diverse phenomena, with success approximately proportional to their understanding of these notions and the rigor of their application.

A frequent source of confusion has been the failure to distinguish between the information contained in a single message and the information contained in a set of messages about another set. The former is the intuitive concept; the latter is the fundamental concept of information theory. There has also been considerable confusion between the entropy of information theory and the entropy of physics. We will try below to make the distinctions clear.

We suppose two ensembles of objects, the X ensemble with members x_i and the Y ensemble with members y_j, and a probability distribution $p(x_i, y_j)$, the probability of the simultaneous occurrence of both x_i and y_j. We then define $p(x_i) = \sum_j p(x_i, y_j)$ and $p(y_j) = \sum_i p(x_i, y_j)$; $p(x_i)$ is the probability of the occurrence of x_i with any y_j. If $p(x_i, y_j) = p(x_i)\, p(y_j)$ for all x_i, y_j, the ensembles are said to be independent. $\sum_i \sum_j p(x_i, y_j) = 1$, the probability of some occurrence. Finally, we define $p(x_i \mid y_j) = \dfrac{p(x_i, y_j)}{p(y_j)}$ and $p(y_j \mid x_i) = \dfrac{p(x_i, y_j)}{p(x_i)}$; $p(y_j \mid x_i)$ is the conditional probability of y_j when x_i is known to have occurred. For three ensembles and a probability $p(x_i, y_j, z_k)$, we can make similar definitions of such quantities as $p(x_i, z_k)$, $p(x_i, y_j \mid z_k)$, and $p(x_i \mid y_j, z_k)$, where the sense can always be determined by substituting "the probability of" for p, "given" for the vertical bar, and "and" for the comma.

Let us then seek a function $I(x_i; y_j)$ to represent the information provided by a simple object (message) y_j about another object x_i: the functions $I(x_i; y_j \mid z_k)$ and $I(x_i; y_j\, z_k)$ will be used to represent the information provided by y_j about x_i where z_k was previously known, and the information provided about x_i by y_j and z_k occurring simultaneously, respectively.

It can be shown that if $I(x_i;y_j)$ is required to be a function of $p(x_i)$ and $p(x_i|y_j)$, $I(x_i;y_j|z_k)$ to be the same function of $p(x_i|z_k)$ and $p(x_i|y_j,z_k)$, and $I(x_i;y_jz_k) = I(x_i;y_j) + I(x_i;z_k|y_j)$, then $I(x_i;y_j) = F\left[p(x_i)\right] - F\left[p(x_i|y_j)\right]$. If we further require that for two independent pairs of ensembles — that is, two pairs of ensembles for which $p(x_i,y_j,\zeta_k,\eta_\ell) = p(x_i,y_j)\,p(\zeta_k,\eta_\ell)$ — $I(x_i,\zeta_k;y_j,\eta_\ell) = I(x_i;y_j) + I(\zeta_k;\eta_\ell)$, then the function F must be the negative logarithm. These intuitively agreeable postulates lead then to $I(x_i;y_j) = \log \dfrac{p(x_i|y_j)}{p(x_i)}$, where the base of the logarithm is arbitrary. Commonly the base 2 is used, and the (dimensionless) units of I called bits; we will use the base e, whose units are nats, and will emphasize this choice by writing

$$I(x_i;y_j) = \ln \frac{p(x_i|\,y_j)}{p(x_i)} .$$

Because $I(x_i;y_j) = \ln \dfrac{p(x_i,y_j)}{p(x_i)p(y_j)} = I(y_j;x_i)$, it is often called the mutual information between x_i and y_j to emphasize the symmetry. When x_i and y_j are independent of each other, the information between them is zero. Positive information results when the occurrence of y_j increases the probability of x_i, and negative when it decreases the probability of x_i. At the extreme, when the occurrence of y_j guarantees or excludes the occurrence of x_i, the information takes on its extreme values of $\ln \dfrac{1}{p(x_i)}$ and minus infinity. The former value, the information given by a message y_j which specifies x_i uniquely, is sometimes called the self-information of x_i.

Hitherto we have considered the mutual information between members of two ensembles. Interesting results do not occur, however, until we consider the mutual information $I(X;Y)$ between two ensembles. $I(X;Y)$ is defined as the average mutual information between all pairs of members of the ensembles, or $I(X;Y) = \sum_i \sum_j p(x_i,y_j) \cdot \ln \dfrac{p(x_i|\,y_j)}{p(x_i)}$. Like $I(x_i;y_j)$ it is symmetrical in X and Y; unlike $I(x_i;y_j)$ it is never negative, as can be seen by use of the inequality $\ln x \geq 1 - 1/x$.

The context in which $I(X;Y)$ assumes its greatest significance in communication theory is in relation to a channel. A channel in the communications sense is nothing more nor less than an ensemble X, called inputs, and an ensemble Y, called outputs, with a matrix of conditional probabilities $p(y_j | x_i)$. Clearly for any set of $p(x_i)$, $I(X;Y)$ can be calculated; the maximum value of $I(X;Y)$ over all allowable sets of $p(x_i)$ is called the capacity of a channel.

A noiseless channel is one for which there are at least as many outputs as inputs and for a suitable arrangement of the outputs, $p(y_j | x_i) = \delta_{ij}$. In other words, the inputs are a one-to-one function of the outputs and vice versa. For such a channel $I(X;Y) = \sum_i \sum_j \delta_{ij} \, p(x_i) \ln \frac{\delta_{ij}}{p(x_i)} = - \sum_i p(x_i) \ln p(x_i) = H(X)$. This function $H(X)$ is called the entropy or uncertainty about the ensemble X; it is clearly equal to the average of the self-information of the X ensemble. It, too, is never negative; its minimum value of zero is assumed when the probability of the occurrence of one of the x_i is one; that is, when x_i can be predicted with certainty. If we define the uncertainty about X when Y is known as $H(X| Y) = \sum_i \sum_j p(x_i, y_j) \ln p(x_i|y_j)$, another non-negative quantity, then $I(X;Y) = H(X) - H(X| Y)$. $H(X)$ is therefore the maximum average information which can be given by the output of any channel about the X ensemble when its probability distribution is $p(x_i)$, or, equivalently, the minimum average information necessary to determine X completely.

The fundamental result of information theory is the coding theorem. If the same channel is used many times in succession, it is said to be memoryless if the transition probabilities for any one use do not depend on the past history of the channel. For a memoryless channel (as well as a broad class of general channels) with capacity C, if a sequence of messages is picked from X ensemble with $H(X) \leq C$, then the coding theorem states that if the number of messages which can be coded together is sufficiently large, there exist codes for which an arbitrarily low probability of transmission error can be achieved.

ENTROPY

The concept of entropy was originally introduced in thermodynamics to quantify the notion of irreversibility. It is observed that almost all natural processes flow naturally in one direction, not the other: two gases will mix spontaneously, but not unmix; heat will diffuse from a warmer body to a colder, but not from a colder to a warmer. Thermodynamics concerns itself almost entirely with equilibria; when one or more systems originally in equilibrium are disturbed by some change in external constraints, and then let reach a new equilibrium, it proves possible to define the entropy of an equilibrium state in such a way that the final entropy of the total system will always be greater than or equal to the original entropy; further, if the entropy is unchanged, the process is reversible, and vice versa. The second law of thermodynamics is an expression of this result.

Such a powerful concept begs for extension to non-equilibrium situations. There have been numerous attempts both in classical and in quantum statistical mechanics[2,3] to define a quantity as a function of the instantaneous state of the system which would invariably be monotone-non-decreasing in time. These attempts have not been without success, but, as far as the author can determine, they have neither broadened the conceptual power of the second law nor led to further interesting new results.

The entropy of a system is apparently related to its disorderliness or randomness. Many writers have pointed out that the entropy of a system is in some sense a measure of our lack of information about the system, or of our uncertainty about the values of the parameters of the system. We have found it possible to make precise this relation, in a way that suggests an unambiguous operational definition of entropy which is sufficiently general for use in either classical or quantum mechanical applications. It promises to provide an interesting and conceptually limpid extension of the second law to non-equilibrium situations, but like the second law itself will probably lead to no new experimental predictions.

Von Neumann[2] extended standard thermodynamical techniques to find what the entropy of a quantum mechanical system with density matrix ρ must be: $-k \operatorname{Tr} \rho \ln \rho$, where k is Boltzmann's constant, whose appearance is due to the historical accident that temperature was originally measured not in energy units but in units of its own. $\rho \ln \rho$ is of course the matrix which is diagonal in the representation in which ρ is diagonal and has in that representation the diagonal elements $\rho_{ii} \ln \rho_{ii}$.

This definition, in common with the definitions that have been made for classical statistical mechanics, has several difficulties, notably the failure of the quantity so defined ever to increase with time. Alternative definitions reflecting the crudity of actual measurements have been introduced;[2,3] these quantities maintain (with small and improbable fluctuations) an upward trend toward equilibrium with the passage of time.

Another difficulty, not so commonly noted, is that two equivalent states may be assigned different entropies. Assignment of a matrix to a state is unambiguous in general only if all complete measurements are conceived as possible (see 'The Incompletely Specified State,' below). If the class of measurements is sufficiently restricted, two different matrices may become equivalent with respect to the remaining set of measurements. For example, the density matrix of an electron is in general infinite-dimensional, but if only spin measurements are allowed, then each of the infinite-dimensional matrices becomes equivalent to one of the 2×2 matrices described earlier. In this case the entropy of the system would seem to depend on how one chose to represent the state as a matrix — although ad hoc rules are commonly introduced to make this choice unambiguous.

The alternative to such an ad hoc rule is to recognize that the entropy of a system depends on the set of permissible measurements, and define entropy to make this dependence explicit. Such a procedure may seem offensive to those who conceive entropy to be a parameter of the system itself, independent of what measurements we may make on it, but in fact even in the present formulation the entropy of an electron is different depending on whether all types of measurements or only spin measurements may be made upon it. Stated differently, we want two equivalent states to have the same entropy; but equivalence is always defined with respect to a set of measurements; therefore the entropy of a system depends not only on the state of the system but also on the class of permissible measurements.

Take the state ρ, a superposition of states (hypotheses) ρ_i, each having probability p_i. Any measurement will give us some information about which ρ_i was actually the case prior to the measurement; if the results of the measurement are y_k, then $I(M;\rho) = \sum_i \sum_k \Pr(\rho_i, \eta_k) \ln \dfrac{\Pr(\rho_i | \eta_k)}{\Pr(\rho_i)} = \sum_i \sum_k p_i \Pr(\eta_k | \rho_i) \ln \dfrac{\Pr(\eta_k|\rho_i)}{(\sum \Pr(\eta_k|\rho_i) p_i}$

$$= \sum_i \sum_k p_i \Pr(\eta_k|\rho_i) \ln \frac{\Pr(\eta_k|\rho_i)}{\Pr(\eta_k|\rho)} .$$

Now take an arbitrary state ρ'. Unless it is a pure state, it is equivalent to another state ρ of the type above, and indeed to an infinite number of such states for each of which the decomposition is different. Since all these states are equivalent, it is a matter of indifference whether we choose to describe our knowledge about the system as ρ' or as one of these superpositions; each has as much operative reality as any other as long as all are equivalent to ρ'. If we choose one particular decomposition, the question of which of the component states (hypotheses) is actually present is a meaningful one, regardless of whether these component states were mentioned in our original description of ρ'.

It is reasonable then to take as a measure of our lack of information about the system the following: the maximum information obtainable by any permissible measurement about any set of hypotheses consistent with the state of the system. We define this quantity as the physical entropy of the system. It will subsequently be shown that this is consistent with the usual definition of entropy in simple cases.

This definition applies only to predictable states, as does the conventional definition. The entropy of a state assumed to be predictable is then assumed to be the entropy of the predictable state to which it is assumed equivalent. The entropy of a novel state is not defined; therefore one cannot speak of the entropy of the universe.

A pure state is not subject to decomposition; there is no set of hypotheses other than itself equivalent to it, and hence no information obtainable about it; its entropy is therefore zero; a pure state is a state of maximal knowledge.

Let us now calculate the entropy explicitly for a general state when all measurements are permitted. We will need the following Lemma: for any representation of any set of density matrices ρ^i, $\sum_j \rho_{jj} \ln \rho_{jj} - \mathrm{Tr}\, \rho \ln \rho - \sum_i \sum_j p_i \rho^i_{jj} \ln \rho^i_{jj} + \sum_i p_i \mathrm{Tr}\, \rho^i \ln \rho^i \geq 0.$ where the p_i are probabilities, $\sum_i p_i = 1$,

ρ^i_{jj} is the j^{th} diagonal element of ρ^i, and

$$\rho = \sum_i p_i \rho^i, \quad \rho_{jj} = \sum_i p_i \rho^i_{jj}.$$

Since the set of all density matrices is convex, this is an assertion of the convexity of the function $\sum_j \rho_{jj} \ln \rho_{jj} - \mathrm{Tr}\, \rho \ln \rho$ on the set of all density matrices.

The proof is straightforward but requires some additional notation. Call the representation above the M representation. Let the diagonal elements of ρ in the representation in which it is diagonal (the ρ representation) be ρ_{kk}, and of the ρ^i in their respective diagonal representations (the ρ^i representations), $\rho^i_{\ell\ell}$. Then

$$-\mathrm{Tr}\, \rho \ln \rho = -\sum_k \rho_{kk} \ln \rho_{kk};$$

$$\sum_i p_i \,\mathrm{Tr}\, \rho^i \ln \rho^i = \sum_i \sum_\ell p_i\, \rho^i_{\ell\ell} \ln \rho^i_{\ell\ell} .$$

There is a unitary transformation U which transforms matrices from the ρ to the M representation; let its matrix elements be U_{jk}. There are also unitary transformations V^i which transform matrices from each of the ρ^i representations to the ρ representation; let their elements be $V^i_{k\ell}$. Let the elements of ρ^i in the ρ representation be $\rho^i_{kk'}$. Then

$$\rho^i_{kk'} = \sum_\ell V^i_{k\ell} V^{i*}_{k\ell} \rho^i_{\ell\ell}$$

$$\rho^i_{jj} = \sum_k \sum_{k'} \sum_\ell U_{jk} U^*_{jk'} V^i_{k\ell} V^{i*}_{k'\ell} \rho^i_{\ell\ell}$$

$$\rho_{kk'} = \sum_i \sum_\ell p_i V^i_{k\ell} V^{i*}_{k'\ell} \rho^i_{\ell\ell}$$

$$\rho_{jj} = \sum_i \sum_k \sum_{k'} \sum_\ell p_i\, U_{jk} U^*_{jk'} V^i_{k\ell} V^{i*}_{k'\ell} \rho^i_{\ell\ell} .$$

It is readily verified that

$$\sum_j \rho_{jj} \ln \rho_{jj} = \sum_i \sum_j \sum_k \sum_{k'} \sum_\ell p_i U_{jk} U^*_{jk'} V^i_{k\ell} V^{i*}_{k'\ell} \rho^i_{\ell\ell} \ln \rho_{jj}$$

$$\sum_k \rho_{kk} \ln \rho_{kk} = \sum_i \sum_j \sum_k \sum_{k'} \sum_\ell p_i U_{jk} U^*_{jk'} V^i_{k\ell} V^{i*}_{k'\ell} \rho^i_{\ell\ell} \ln \rho_{kk}$$

$$\sum_i \sum_j p_i \rho^i_{jj} \ln \rho^i_{jj} = \sum_i \sum_j \sum_k \sum_{k'} \sum_\ell p_i U_{jk} U^*_{jk'} V^i_{k\ell} V^{i*}_{k'\ell} \rho^i_{\ell\ell} \ln \rho^i_{jj}$$

$$\sum_i \sum_\ell p_i \rho^i_{\ell\ell} \ln \rho^i_{\ell\ell} = \sum_i \sum_j \sum_k \sum_{k'} \sum_\ell p_i U_{jk} U^*_{jk'} V^i_{k\ell} V^{i*}_{k'\ell} \rho^i_{\ell\ell} \ln \rho^i_{\ell\ell}$$

In sum, $\sum_j \rho_{jj} \ln \rho_{jj} - \mathrm{Tr}\, \rho \ln \rho - \sum_i \sum_j p_i \rho^i_{jj} \ln \rho^i_{jj} + \sum_i p_i \mathrm{Tr}\, \rho^i \ln \rho^i =$

$$\sum_i \sum_j \sum_k \sum_{k'} \sum_\ell p_i U_{jk} U^*_{jk'} V^i_{k\ell} V^{i*}_{k'\ell} \rho^i_{\ell\ell} \ln \frac{\rho_{jj}\rho^i_{\ell\ell}}{\rho_{kk}\rho^i_{jj}} \geq$$

$$\sum_i \sum_j \sum_k \sum_{k'} \sum_\ell p_i U_{jk} U^*_{jk'} V^i_{k\ell} V^{i*}_{k'\ell} \rho^i_{\ell\ell} \left(1 - \frac{\rho_{kk}\rho^i_{jj}}{\rho_{jj}\rho^i_{\ell\ell}}\right) = 0,$$

with use of the inequality $\ln x \geq 1 - 1/x$, and the equalities

$$\sum_\ell V^i_{k\ell} V^{i*}_{k'\ell} = \sum_j U_{jk} U^*_{jk'} = \delta_{kk'}$$

$$\sum_\ell \rho^i_{\ell\ell} = \sum_k \rho_{kk} = 1.$$

Equality is achieved where the M, ρ, and ρ^i representations coincide, or when $p_i = 1$.

We shall also need the intuitively reasonable Lemma:[5] for any set of hypotheses x_i, any set of inputs to a channel y_j, and any set of outputs from the channel z_k, if $p(z_k|x_i) = \sum_j p(z_k|y_j)\, p(y_j|x_i)$, then $I(X;Y) \geq I(X;Z)$.

This says that if the Z ensemble is dependent on the X ensemble only through the Y ensemble, then it never can provide more information about the X ensemble than can the Y ensemble.

Proof: $$I(X;Y) - I(X;Z) = \sum_i \sum_j p(x_i)\, p(y_j|x_i) \ln \frac{p(x_i|y_j)}{p(x_i)}$$

$$-\sum_i \sum_k p(x_i)\, p(z_k|x_i) \ln \frac{p(x_i|z_k)}{p(x_i)}$$

$$= \sum_i \sum_j \sum_k p(x_i)\, p(z_k|y_j)\, p(y_j|x_i) \ln \frac{p(x_i|y_j)}{p(x_i|z_k)}$$

$$\geq \sum_i \sum_j \sum_k p(x_i)\, p(z_k|y_j)\, p(y_j|x_i) \left(1 - \frac{p(x_i|z_k)}{p(x_i|y_j)}\right.$$

$$= 1 - \sum_i \sum_j \sum_k \frac{p(x_i, y_j, z_k)\; p(y_j)\; p(x_i|z_k)}{p(x_i, y_j)}$$

$$= 1 - \sum_i \sum_j \sum_k p(z_k|y_j, x_i)\; p(y_j)\; p(x_i|z_k)$$

$$= 1 - \sum_j \sum_k p(z_k|y_j)\; p(y_j) = 0.$$

again with use of the inequality $\ln x \geq 1 - 1/x$.

We can now show

Theorem: the entropy of a state represented by the matrix ρ when all measurements are permitted is $-\mathrm{Tr}\, \rho \ln \rho$.

Proof: only complete measurements need be considered. For any other measurement is dominated by a complete measurement; that is, the probabilities of its outputs $p(\eta_k)$ are $\sum_j p_{kj}\, p(\lambda_j)$, where p_{kj} are the pertinent transition probabilities and $p(\lambda_j)$ are the probabilities of the different outputs of the complete measurement. It follows from $p(\eta_k|\rho^i) = \sum_j p_{kj}\, p(\lambda_j|\rho^i)$ and the second lemma that $I(\rho;\eta) \leq I(\rho;\lambda)$; therefore the measurement that yields maximum information is one of the class of complete measurements.

For any complete measurement M, $P(\lambda_j | \rho^i) = \rho^i_{jj}$, the j^{th} diagonal element of ρ^i in the M representation. Then

$$I(M;\rho) = \sum_i \sum_j p_i \rho^i_{jj} \ln \frac{\rho^i_{jj}}{\sum p_i \rho^i_{jj}} = -\sum_j \rho_{jj} \ln \rho_{jj} + \sum_i \sum_j p_i \rho^i_{jj} \ln \rho^i_{jj} ,$$

$$\leq -\mathrm{Tr}\, \rho \ln \rho + \sum_i p_i \mathrm{Tr}\, \rho^i \ln \rho^i$$

$$\leq -\mathrm{Tr}\, \rho \ln \rho$$

since $\mathrm{Tr}\, \rho^i \ln \rho^i$ is essentially negative.

It remains to investigate whether the equality ever holds. For the equality it is sufficient that the M and ρ^i representations coincide with the ρ representation and that the ρ^i be one-dimensional projectors. In other words, the measurement must be the characteristic measurement of ρ, and the hypotheses must be the various pure states of the characteristic measurement.

With the aid of this theorem we see that, except for a difference of units, our definition of entropy is the same as the conventional one for the case in which all complete measurements are permitted. It will thus exhibit the constancy in time which is characteristic of this situation; this constancy is due to the peculiar property of complete measurements that to a complete measurement at one time there always corresponds one at any other time which will yield the same information about a system.

We have not yet examined in detail the behavior of our entropy with time in the more interesting situations in which the limitations of actual measurements are taken into account, but we may expect the following general features.

First, we point out that the entropy as we have defined it is associated not with the time parameter of the system but with the subjective time of the observer. Because the laws of nature are symmetrical with respect to time inversion, there has been some question as to why entropy should increase in a preferred time direction. For an observer, on the other hand, the order in which data are observed or information is received defines a unique direction of time, which in turn in our view defines the direction of entropy increase.

Second, we note that our definition of entropy can be taken over without amendment to situations in which only crude measurements are allowed, which generality may be considered a conceptual virtue. For von Neumann's[2] "macroscopic measurements," our entropy is not equivalent to his "macroscopic entropy," nor for Tolman's[3] "coarse-grained probabilities," to his "$\bar{\bar{H}}$", but the difference is only a factor having to do with the number of states indistinguishably lumped together.

Crude measurements sometimes differ from complete measurements in only being able to be performed near some time. For example, consider two electrons which interact and then proceed in different directions. While they interact one can think of measuring the total spin of the combined system of both electrons; while this measurement on the combined system remains always possible in principle, after the electrons are sufficiently separated it is not feasible to perform such a measurement, but rather only measurements on each system separately. In this sense the class of permissible measurements is always shrinking, as measurements become impossible by moving from the future to the past. We speculate that this phenomenon leads to an increase of entropy with time, though no calculations have been made.

One frequently finds it convenient to discard certain information. For instance, in the case of the interacting electrons mentioned in the previous paragraph, after the electrons are sufficiently separated one may wish to disregard the correlations between the states of the two due to their interaction. Or, if one is interested only in predicting the results of one measurement, say the energy measurement, one may at intervals discard the off-diagonal elements of the density matrix in the energy representation as irrelevant. Either way of "forgetting" information can easily be shown[6] never to result in a decrease in entropy.

The increase of entropy which occurs in thermodynamics when a system makes an irreversible transition between two equilibrium states is almost certainly due to discarding information in this way. After a certain time the only information about the initial state worth anything is the values of the constants of the motion, archetypically of the energy. The time at which such a situation obtains is almost by definition the time at which equilibrium is said to have been reached.

We may thus hope that this definition of entropy will lead to a unified treatment of all phenomena related to the second law of thermodynamics, including both classical and quantum systems, both ideal and actual measurements. For such a unification we must pay the following price: entropy becomes a function of our relevant knowledge rather than an attribute of a physical system; lack of information becomes the fundamental characteristic of a high-entropy situation, rather than a peripheral feature, while disorder, uncertainty, number of complexions, and so forth, become relegated to secondary roles.

THE "INCOMPLETELY SPECIFIED" STATE

Sometimes a state ρ is described by a set of data, such as the mean values of different measurements, which are consistent with a great many matrix representations for ρ. In order to choose one of these representations, an ad hoc principle is commonly introduced.[3,6,7,8] This will be shown not to be necessary.

There are two distinct situations in which this problem arises: all measurements may be possible, or some may be impermissible. We begin with consideration of the former.

First let us note that any state ρ must have a unique matrix representation if all measurements are allowed. For if two different matrices ρ_1 and ρ_2 each represented ρ, then there would be some representation in which the diagonal elements of ρ_1 and ρ_2 were different, and therefore ρ_1 and ρ_2 would give different predictions for the result of some measurement; but then ρ would not be equivalent to itself, so we have a contradiction. The freedom in choice of ρ suggested by the paucity of data about it does not in fact exist.

Recall that the state of a system is the totality of our knowledge about it. If a state is specified by X, Y, and Z, this means not just that X, Y, and Z are known, but that X, Y, and Z are known and nothing else. The latter formulation makes it clear that an "incompletely specified" state is as completely specified as a pure state.

In order to fix this idea, let us consider a common practical situation. A series of experiments has been performed on a batch of independent systems, about all of which our knowledge was the same before the experiments. This series of experiments has been sufficiently complete to allow us to complete the difficult task of assigning a matrix representation to the common state of all these systems; whatever members of the batch are left undisturbed after this series of measurements can each be given the same matrix representation (with allowance for time development).

Suppose, however, we are given a second batch, which we have reason to believe is very much like the first. That we are able, however, to separate the batches into "first" and "second" means that their states are not equivalent (our knowledge about each being different, however slightly), and we are therefore not justified in assigning the previously determined matrix representation to the members of the second batch. Since it is tedious to repeat the difficult process of matrix determination, we frequently assume the same representation for the second batch, invoking more or less general principles to justify the assumption. In sum, each of the states is completely specified and different; the ad hoc assumption is that they are equivalent and may therefore be represented by the same matrix. Such an assumption is often reasonable, but we must not be surprised, only disappointed, when the matrix we have assigned the second batch occasionally proves to be inconsistent with subeequent experiments.

In the case in which "X, Y, and Z are known" is actually a complete and accurate description of the state, however, there is a straightforward procedure for finding the matrix representation. Of all the matrices superficially consistent with the observed data, there is one, ρ, whose entropy is greatest. If we decompose this one into the pure states of its characteristic measurement, then by making the characteristic measurement we can obtain an amount of information equal to $-\mathrm{Tr}\ \rho \ln \rho$ about which of these pure states is the case. One cannot object that this decomposition may be impossible, for by definition nothing else is known to exclude it. Choice of any other ρ leads to a lesser entropy, and hence to an upper bound on the information which is not consistent with what is actually possible. Therefore one must choose that ρ for which $-\mathrm{Tr}\ \rho \ln \rho$ is maximum; that this ρ is unique follows from the uniqueness of the state description and the possibility of making all measurements.

It may seem that we have simply substituted one ad hoc assumption for another. This is true; it is of value, however, to state with precision what one's assumptions are. Arbitrary choice is involved when we assume that only a certain portion of the state description is relevant to predicting the results of experiments, or that two different states are equivalent; once these assumptions are made the matrix representation for ρ is unambiguous though possibly wrong.

The situation is different in the case in which not all measurements are allowed. Two different matrices may be equivalent for this restricted class of measurements. It follows that we are indeed allowed to choose any of these, depending on which is most convenient. In general, however, $-\mathrm{Tr}\, \rho \ln \rho$ will depend on which ρ is chosen, so that the use of the conventional definition of entropy leads to an ambiguous value. It was largely in order to avoid this difficulty that we formulated the unambiguous definition introduced above.

It has sometimes been found convenient to choose that ρ for which $-\mathrm{Tr}\, \rho \ln \rho$ was maximized; often this has been done without realization of the distinction between this case and the previous one, in which such a choice was the only one permissible. $-\mathrm{Tr}\, \rho \ln \rho$ for this ρ has then been called the entropy. Now it may happen that the characteristic measurement for such a ρ may fall in the class of measurements which is prohibited, in which case the maximum information achievable is less than $-\mathrm{Tr}\, \rho \ln \rho$. It is hard to see what operational significance the quantity $-\mathrm{Tr}\, \rho \ln \rho$ might have in such a situation.

A common specification for a state consists in a list of the mean values $\overline{M_i}$ of a set of measurements M_i. Stratonovich[9] has shown by a straightforward use of Lagrange multipliers that the ρ for which $-\mathrm{Tr}\, \rho \ln \rho$ is maximized in such a case is

$$\rho = \frac{\exp[-\sum_i \beta_i M_i]}{\operatorname{Tr}\exp[-\sum_i \beta_i M_i]}$$

For $\operatorname{Tr} \rho M_i = \overline{M_i}$, $\operatorname{Tr} \rho = 1$;

$$\delta(\operatorname{Tr} \rho \ln \rho) = \delta [\operatorname{Tr} \rho(\ln \rho + \gamma + \sum_i \beta_i M_i)]$$

$$= \operatorname{Tr} \delta\rho (\ln \rho + 1 + \gamma + \sum_i \beta_i M_i)$$

$$= 0, \text{ all } \delta\rho$$

Therefore $\ln \rho = -1 - \gamma - \sum_i \beta_i M_i$.

Evaluation of γ from $\operatorname{Tr} \rho = 1$ leads to the expression for ρ above. Note that the matrix $e^{-(A+B)} \neq e^{-A} e^{-B}$ in general.

COMMUNICATION

Suppose one observer, called the transmitter, wants to communicate with another, called the receiver, and there is a physical system, called a medium, to which each has access. Then for each message he wishes to send, the transmitter may perturb the system in a different way, so that a subsequent measurement on the medium by the receiver will give some information about which message was sent.

Let the state of the medium after the transmitter has performed the perturbation characteristic of message i be represented by the matrix ρ^i, and let the probability of message i be p_i. Prior to any measurement, but with the knowledge that some perturbation has been made, the receiver describes his knowledge as $\rho = \sum_i p_i \rho^i$.

Suppose further that the transmitter must not exceed certain limitations on the overall ρ. A reasonable limitation is that the average value of the energy be less than some limit; in other words $\operatorname{Tr} \rho H \leq \mathbf{E}$, where H is the matrix of the energy measurement.

The medium is not a channel in the strict sense of communication theory, since the transition probabilities depend on the choices of perturbation and measurement. With this in mind, we define the capacity of a medium as the maximum information the transmitter can communicate to the receiver subject to whatever restrictions are imposed on ρ.

With our definition of entropy, then, the capacity of a medium is the entropy of that medium when only the limitations on ρ are known and only the measurements available to the receiver are permitted. The optimum set of messages is identical to the optimum decomposition of the state in which only the limitations on ρ are known.

As an example, we consider the case in which $\mathrm{Tr}\, \rho H \leq \mathbf{E}$, and the receiver is permitted all measurements. The state in which the average energy is $\mathbf{E}$ and nothing else is known is uniquely represented by the matrix $\rho = \dfrac{e^{-\beta H}}{\mathrm{Tr}\, e^{-\beta H}}$, where β is chosen so that $\dfrac{\mathrm{Tr}\, He^{-\beta H}}{\mathrm{Tr}\, e^{-\beta H}} = \mathbf{E}$. The entropy of this state when all measurements are permitted is $-\mathrm{Tr}\, \rho \ln \rho = \dfrac{\mathrm{Tr}\, \beta He^{-\beta H}}{\mathrm{Tr}\, e^{-\beta H}} + \ln \mathrm{Tr}\, e^{-\beta H} = \beta \mathbf{E} + \ln \mathrm{Tr}\, e^{-\beta H}$. This is the capacity of the medium; it is attained when the transmitter lets each pure state of the energy measurement correspond to one message, with the probability of the message corresponding to the energy value $\mathbf{E}_i$ being $\dfrac{e^{-\beta \mathbf{E}_i}}{\mathrm{Tr}\, e^{-\beta H}}$, and the receiver makes the energy measurement. Since this ρ is the canonical distribution for a black body whose temperature is $T = 1/\beta k$, it is as though the receiver measured the energy of what he thought was a black body at temperature T, only to find that the body had a perfectly definite energy which had been set by the transmitter.

We note that even in the absence of external disturbances, or noise, the capacity of a physical medium is finite if there is a finite energy limitation. As

$T \to 0$, $\beta \to \infty$, but in the limit only the lowest energy state can be used, so that the capacity of the medium goes to zero unless this energy level is degenerate. This contrasts with the possibility of infinite capacity with arbitrarily small energy which is consistent with the classical mechanics in the absence of noise.

For a harmonic oscillator of frequency ω, a simple calculation gives $\beta = 1/h\omega \ln \dfrac{E/h\omega + 1/2}{E/h\omega - 1/2}$ and the capacity as $\ln(1/2 + E/h\omega) + (E/h\omega - 1/2) \ln \dfrac{E/h\omega + 1/2}{E/h\omega - 1/2}$. As E approaches its minimum of $1/2\ h\omega$, C goes to zero; as E becomes large, C becomes approximately $\ln E/h\omega$.

As an example in which the measurements are restricted, we could consider the electron as a communications medium and limit the possible measurements to spin measurements. Since only two results are obtainable, the capacity of this medium is ln 2 (one bit), obtained when any two antiparallel spins are used with equal probabilities.

One usually wishes to make repeated use of the same medium. In this case one must consider the minimum time required for transmission of any one message, as well as correlations between messages. There seems to be no restriction in principle on how short a time is required to prepare a system or to make a measurement, even of energy;[10] the price paid is in increasingly strong coupling. In the absence of any stated limitations to the contrary, the transmitter can in principle completely reset the system in an arbitrarily short time, thereby making the channel memoryless.

Since the medium becomes a channel upon specification of the different input perturbations and of the output measurement, the validity of the coding theorem for media, which results from the substitution of "medium" for "channel" in the coding theorem for channels, then follows from the validity of the coding theorem for channels.

It must be noted, however, that the capacity of a medium when all measurements are allowed is a totally impractical concept. Both the transmitter and the receiver are required to be able to make complete measurements, the former to prepare the system in a pure state, the latter to find out which was prepared. Both must therefore not only have access to the whole system, but must be capable of extremely strong coupling to the system, in inverse proportion to the length of time allowed for the measurement.

The situation of greatest practical interest is that of communication by electromagnetic radiation through a medium which is a near vacuum, with the distance between transmitter and receiver far greater than the dimensions of either. In this case the transmitter is able to perturb the medium only slightly, and the receiver to couple to the medium very weakly. In particular, the transmitter is not able to absorb radiation from other sources, called noise; there result constraints not just on ρ but on each ρ_i individually. This so-called photon channel has been studied by a number of authors.[11,12,13]

Although these investigations may give rough, intuitive guidance to the design of laser communication systems, in the opinion of the author the interesting areas of investigation in quantum communication are two: first, the philosophically interesting questions of fundamental limits which we have studied above; second, the practically important problems of how one actually communicates with a laser. Development and comparison of a range of modulation and reception techniques are goals of prime importance; an insightful and useful quantum model of the laser radiation field seems to be the most pressing theoretical problem. Studies aimed half way between these poles of interest, typically computing asymptotic capacities of crudely modelled photon channels, seem neither pertinent to practical problems nor of independent philosophical interest.

BIBLIOGRAPHY

1. P. A. M. Dirac, The Principles of Quantum Mechanics, Clarendon Press, Oxford, 1948.

2. J. von Neumann, Mathematical Foundations of Quantum Mechanics, Princeton University Press, Princeton, 1955.

3. R. C. Tolman, The Principles of Statistical Mechanics, Clarendon Press, Oxford, 1938.

4. U. Fano, "Description of States in Quantum Mechanics by Density Matrix and Operator Techniques," Revs. Mod. Phys. 29, 74 (1957).

5. R. M. Fano, The Transmission of Information, Wiley, New York, 1961.

6. E. T. Jaynes, "Information Theory and Statistical Mechanics," Phys. Rev. 106, 620; II-108, 171 (1957).

7. W. M. Elsasser, "Quantum Measurements and Uncertainty Relations," Phys. Rev. 52, 987 (1937).

8. E. C. Kemble, "Quantum Mechanical Basis of Statistical Mechanics and Thermodynamics," Phys. Rev. 56, 1013, 1146 (1939).

9. R. L. Stratonovich, JETP USSR 28, 547 (1955).

10. Y. Aharonov and D. Bohm, "Time in Quantum Theory and the Uncertainty Relation for Time and Energy," Phys. Rev. 122, 1649 (1961).

11. J. P. Gordon, "Quantum Effects in Communication Channels," Proc. IRE 50, 1898 (1962).

12. H. A. Haus and J. A. Mullen, "Photon Noise," (private communication).

13. T. Hagfors, "Information Capacity and Quantum Effects in Propagation Circuits," (private communication).