RELATING THERMODYNAMICS TO INFORMATION THEORY: THE EQUALITY OF FREE ENERGY AND MUTUAL INFORMATION

Thesis by

David I. Feinstein

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1986

(Submitted May 21, 1986)

Acknowledgment

As a graduate student at Caltech I have been supported financially by several different sources. The Department of Applied Mathematics funded me as a teaching assistant during my first year. I was awarded an IBM fellowship for the next two years under the auspices of the Department of Computer Science. During the last two years I have been a research assistant in the laboratory of Professor Carver Mead; the funding for this period was provided by System Development Foundation. My thanks to all of these sources.

I wish also to thank Professor Carver Mead, Professor John Hopfield and John Platt for many interesting discussions on the subject of this thesis and other topics as well. These individuals have contributed substantially to my graduate education at Caltech. My education, (and the rest of me as well), owe a special note of thanks to my parents; their support has always been essential.

On the subject of this thesis, the less said the better! However, I gratefully acknowledge the assistance of my brother Jonathan. Without his aid, I probably would never have started writing. Also, I would never have finished writing, but for the sympathetic support of Carver Mead.

I would like to express my very sincere appreciation to Carver for all that he has done on my behalf. Carver has been a mentor to me. He has shown me, first–hand, the process whereby science happens. Meanwhile, he has taught me about exponentials and positive–sum games. These things are related, I think. Thank you, Carver, for all of these things.

Abstract

In this thesis we uncover a new relation which links thermodynamics and information theory. We consider time as a channel and the detailed state of a physical system as a message. As the system evolves with time, ever present noise insures that the "message" is corrupted. Thermodynamic free energy measures the approach of the system toward equilibrium. Information theoretical mutual information measures the loss of memory of initial state. We regard the free energy and the mutual information as operators which map probability distributions over state space to real numbers. In the limit of long times, we show how the free energy operator and the mutual information operator asymptotically attain a very simple relationship to one another. This relationship is founded on the common appearance of entropy in the two operators and on an identity between internal energy and conditional entropy. The use of conditional entropy is what distinguishes our approach from previous efforts to relate thermodynamics and information theory.

Table of Contents

ACKNOWLEDGMENT. iii

ABSTRACT. iv

LIST OF FIGURES. vi

CHAPTER I. 1

Thesis Introduction. 1

Statistical Mechanics. 2

Classical thermodynamics. 13

The principle of maximum entropy. 22

Information Theory. 25

The relation between thermodynamics and information theory. 33

CHAPTER II. 42

Relating free energy and mutual information. 42

Markov processes in the spectral representation. 59

Expanding $\vec{H}^{\Gamma}(n)\Gamma^{-n}$ around equilibrium The 2x2 case. 66

The rereferencing theorem. 71

$\vec{H}^{\Gamma}(n)\Gamma^{-n}$ in the general case. 75

A composition law for $\vec{H}^{\Gamma}(n)\Gamma^{-n}$. 81

CONCLUSION. 87

BIBLIOGRAPHY. 90

List of Figures

Figure 1. . . . . page 48

H(q) is the log of the number of messages which account for a total probability q. The graph suggests that the log of the number of reasonably probable messages is independent of the definition of reasonably probable for long messages.

Figure 2. 54

The horizonal line depicts the difference of state energies in units of kT. The curve is the difference of the components of $\vec{H}^{T}(n)\Gamma^{-n}$ versus time n. Asymptotically, the curve merges with the line. Thus the rereferenced conditional entropy is related to the rescaled internal energy.

Figure 3. 56

The coordinates of a point in the plane are the components of $\vec{H}^{T}(n)\Gamma^{-n}$. By sweeping n for any particular matrix Γ, we generate a curve. n= infinity is a point on this curve. Consider the locus of points, generated by taking n= infinity for all possible matrices. This locus is also the locus of state energy assignments for which the free energy of equilibrium is 0. Thus the rescaled internal energy and the rereferenced conditional entropy are identical at n= infinity.

Figures 4 and 5. 78

See that the limit of $\vec{H}^{T}(n)\Gamma^{-n}$ exists for all 3 components. The log scale shows that the approach to the limit is as predicted. (The slight ripple in the curves is caused by calculating with too few significant figures.)

Figures 6 and 7. 80

In this case, the limit of $\vec{H}^{T}(n)\Gamma^{-n}$ does not exist. The log scale shows that all three components diverge at the predicted rate.

Thesis Introduction

In this thesis we uncover a new relation which links information theory to thermodynamics. This relation is suggested by a structural simile which we draw between these two subjects. We find that free energy of thermodynamics, and mutual information of information theory are related to one another in a simple way. This relation follows from an asymptotic identity, which also is new, between conditional entropy and internal energy. These findings are best appreciated in a context which emphasizes the structure and organization of thermodynamics and information theory. Accordingly, we have devoted chapter I to a discussion of these two subjects; at the close of this chapter we outline the current understanding of the relation between them. Also, at the end of this chapter we sketch what we have done and how it augments this understanding. Chapter II contains the detailed presentation of our original work in this area.

At its most fundamental, thermodynamics introduces heat, temperature, entropy and the law dQ<=TdS. One can arrive at these precepts of thermodynamics from three different directions. The line of development currently in vogue is that of statistical mechanics. Statistical mechanics owes its ascendancy partly to the current popularity of quantum mechanics and partly to the deep insights that have come from its integration of thermodynamics with the rest of theoretical physics. In the nineteenth century, thermodynamics enjoyed a finely reasoned, classical development; this was the handiwork of the old masters. In some respects this classical line still provides the best explanation for the general success of thermodynamics and its sweeping applicability. Somewhat more recently the predictions of thermodynamics have been shown to be consistent with a

principle of maximum entropy. This principle constitutes a third and somewhat less well known line of development. We turn now to an extended discussion of these three lines of development.

Statistical Mechanics

The notion of a state space is employed throughout theoretical physics. In non–statistical physics, systems are described by specifying the point in state space corresponding to their condition. It would be very arduous to exactly specify the state of a macroscopic system having of order Avogadro's number degrees of freedom. The essential compromise of statistical mechanics is to relinquish exact specification of system state. In practice we effect this compromise by working only with parameters that we can macroscopically measure. If we confine our attention to the macroscopic level, then the development of thermodynamics from mechanics proceeds quite simply; we will sketch this development first. An excellent introduction to statistical physics at this level is found in [Reif]. When we consider systems in microscopic detail (but still in the classical limit), the task of relating the precepts of thermodynamics to the foundations of mechanics becomes quite challenging. The basic idea is to consider the probability distribution over state space, ρ, defined by an ensemble of macroscopically identical systems.[1] We introduce this deeper view of statistical mechanics second; here [Tolman] has been an invaluable source.

---

[1]Systems are macroscopically identical if they are independent of one another except that they share equivalent values of their macroscopic parameters.

Equilibrium: state is independent of past history.

A system that has evolved for a sufficiently long time ultimately approaches equilibrium. In equilibrium the state of a system is independent of all past history; it is decoupled from its initial configuration. A necessary (but not sufficient) condition for equilibrium is that the system be macroscopically stationary. Microscopically, equilibrium happens when the system has made enough transitions to have sampled a representative fraction of its accessible state space. Equilibrium is thus a statistical concept; it tends to defy a precise physical definition. Traditionally the definition of equilibrium has been something of a tautalogy: an equilibrium system is one which is consistent with the predictions of equilibrium thermodynamics.

Entropy measures state space volume.

A heat bath is an equilibrium system which is sufficiently large that it possesses a huge number of degrees of freedom,[2] typically of order Avogadro's number. Note that almost any macroscopic system which is in equilibrium can qualify as a heat bath. The essence of statical mechanics lies in estimating the volume of state space Ω which is accessible to the heat bath, given that the bath has an energy which is known to lie in some small interval around E. The typical result, which always emerges, is that Ω is roughly proportional to E^f, where f is of order the number of degrees of freedom of the heat bath. Since the exponent f is of order 10^{23}, the volume

---

[2]The number of degrees of freedom of a system is the dimension of the state space in which the system finds a complete description. Counting degrees of freedom can be a bit tricky; the number depends on which abstraction of physics one is using. In classical physics a particle has six degrees of freedom; three of these are position coordinates and three are velocity (or momentum) coordinates. In the rigid–body approximation of classical physics, a body has nine degrees of freedom; six of these are associated with position and velocity, and three more come from the abstraction of angular momentum. In quantum physics a particle can have more than six degrees of freedom; it can have spin, for example, and it can have even more exotic attributes as well.

of accessible state space is an exceedingly rapidly varying function of the energy. It is reasonable to employ logarithms in such circumstances. The entropy H is defined as the log of the volume of state space which is accessible to the heat bath; thus $H = \ln\Omega$. The parameter β is defined as the partial derivative of entropy with respect to energy; in taking this derivative the heat bath is assumed to be dynamically isolated, so E is varied because of heat exchanged and not because of work performed. β is the proportional rate of change of Ω with respect to E on account of heat flow; this partial derivative works out to be $\beta = f/E$. The reciprocal of β has units of energy; it is defined to be kT. $kT = E/f$ and is approximately the energy per degree of freedom. Thus thermodynamics introduces H which is dimensionless and β which has dimensions of reciprocal energy. These two useful quantities summarize the state of systems for which our information is otherwise incomplete.[3]

Entropy & heat. The 2nd law.

We can exactly calculate the change in entropy which accompanies the flow of heat into a system which is at equilibrium. Using the definition of β as the partial derivative of entropy with respect to energy on account of heat flow, it is easy to see that $dH = \beta\, dQ$ (or equivalently, $dS = dQ/T$). β is defined for a heat bath at equilibrium. What is the relationship between dH and dQ for a nonequilibrium system? The 2nd law of thermodynamics postulates that for any system, the change in entropy dH is never less than

---

[3]In the early days when thermodynamics was born there happened an unhappy confusion of units; the two primary quantities were taken to be the physical entropy, S, and the temperature, T, where $S = kH$ and $T = 1/(k\beta)$. This hapless choice forced physics to accept a new unit, the degree, which had no relation to any other commodity except through Boltzmann's constant k.

β dQ.[4] Thus dH>= β dQ (or equivalently, dS>= dQ/T). We would like to deduce this inequality from a more intuitive starting point; unfortunately, with the methods of statistical mechanics we cannot easily do so. This difficulty is a shortcoming of the statistical mechanical approach to thermodynamics.

The Boltzmann distribution.

The canonical question of thermodynamics couples a small system to a heat bath and asks how a fixed energy shared between them is apportioned. The small system and heat bath are thermally isolated from the rest of the universe so that their total energy is conserved. The heat bath is assumed to remain at some constant temperature T. The Boltzmann distribution answers the question: what is the probability P(ϵ) for the small system to be in a particular state which has energy ϵ? The derivation of the Boltzmann distribution P(ϵ) assumes the equipartition principle; this principle asserts that, in equilibrium, the composition of heat bath + small system is equally likely to be found at any of the points in state space that are consistent with them sharing together a total energy of E. If the small system is to have an energy ϵ and the total energy is to be E, then the energy of the heat bath must be E$-\epsilon$. The probability that the small system is in a state with energy ϵ, P(ϵ), is proportional to the volume of state space which is accessible to the heat bath when the latter has an energy near E$-\epsilon$. Thus P(ϵ) is proportional to $(E-\epsilon)^f$ which equals $E^f(1-\epsilon/E)^f$; so the dependence of P(ϵ) on ϵ goes like $(1-\epsilon/E)^f$. Now we recall that E, which is very nearly the energy of the heat bath, is equal to f/β; thus P(ϵ) is proportional to

---

[4] A system which is out of equilibrium is a system in which irreversible processes are happening. For these systems, the direction in which time flows is significant. The second law applies to changes in entropy which occur as a system moves forward through time, i.e., as the system ages normally.

$(1-\beta\epsilon/f)^f$. Since f is huge, this reduces to the familiar result: $exp(-\beta\epsilon)$. Thus the Boltzmann distribution follows readily from the methods of statistical mechanics; notice how few assumptions are required. This economy of assumption gives insight to the wide applicability of the Boltzmann distribution.

Heat & entropy in theoretical mechanics

In the world of Newtonian physics, heat appears when energy is lost to friction. Unfortunately, friction has no place in the elegant theoretical world of Lagrangian (or Hamiltonian) physics. Rather, what appears as friction is actually some small interaction whose only significance is to provide a coupling between otherwise orthogonal modes of a system. Heat manifests itself in such a system as the incoherent spread of energy from one mode into many.[5]

Consider, for example, the frictional heat that is generated when a moving block of material scrapes against a rough surface and slows to a halt. A classical analysis of this situation models the block as a collection of masses interconnected with springs. Suppose that before the block interacts with the rough surface, all of its component masses are moving with the same velocity, and the springs connecting the masses are unstretched. Thus the block is quiescent internally; all of its energy is due to the velocity of its

---

[5]In mechanics the detailed state of a system can be specified in either of two ways. For one of the ways the detailed state corresponds to a point in state space, where the coordinates of the point are the postitions and velocities of each particle in the system. Alternatively, for linear systems, the detailed state can also be regarded as corresponding to a point in eigen–state space where the coordinates of the point are the amplitudes and phases of each normal mode of the system. The two specifications carry equivalent information in that a constant, linear, nonsingular transformation carries one into the other. Thus, a probability distribution over the amplitudes and phases of the normal modes of a system corresponds uniquely to a probability distribution over the positions and velocities of the particles of the system. The entropies of these two probability distributions will agree with one another to within a constant additive factor having to do with the log of the jacobian of the transformation which connects one space to the other.

center of mass. We can model the rough surface as a collection of little rigid posts. As some of the masses of the block impact with these posts, internal vibrations of the block are established. The continued action of the rough surface ultimately leaches away nearly all the kinetic energy of the center of mass, converting it into energy of vibration about the center of mass. Thus the energy of the block gets distributed among all of its vibrational modes. Macroscopically, we see the block come to a quivering halt.

The example of the block shows us that heat and entropy are related. Initially, the distribution over mode amplitudes of the block is tightly confined to the zero frequency modes of uniform translation. As the block slows, the distribution over mode amplitudes widens; its entropy increases. A flow of heat (or, more properly, a conversion of mechanical energy to heat) accompanies this increase in entropy. The heat flow is the funneling of energy from the zero frequency modes to the multitude of higher frequency modes. Thus the flow of heat and the widening of the distribution over mode amplitudes are directly connected; in a sense, the increase of entropy describes the flow of heat.

The relative nature of entropy. Quantum physics & heat.

Entropy measures the volume of state space in which a system may be found. This volume depends critically on any constraints which the system is known to satisfy. For example, if the total energy of the system is known, then the system must be on the hyper–plane in state space which corresponds to that energy. If in addition to the energy, the momentum of the system is known, then the system is even more tightly constrained and the region of state space which is accessible to it becomes even smaller. Thus the entropy of a system depends on how much we know about the

system. As we measure more and more parameters of a system, its entropy becomes smaller and smaller.

The observation that the entropy of a system depends on our measurement skills is consistant with our interpretations of heat and work. Energy which is exchanged via a mechanism which we can observe is work, energy which is exchanged via a mechanism which we can not observe is heat. For example, consider the block made of masses and springs. Suppose that in addition to the zero frequency mode of uniform translation, we are also able to measure the amplitude of the next higher frequency mode. With these measurement skills we would see changes in the amplitudes of either of the lowest two modes as being work exchanged and we would interpret changes in the net excitation energy of all of the other modes together as being the flow of heat.

Is entropy ever absolute? Yes, in quantum mechanics entropy attains an absolute definition. As systems become larger, the energy spacing of their quantum states, ΔE, becomes smaller. Physical processes (transitions between quantum states) take some finite time to happen; call this minimum time Δt. When a physical process taking time Δt happens in a system which is large enough so that the ΔE of its quantum states is less than $h/\Delta t$, then[6] the uncertainty principle demands that quantum mechanical phase information be lost. This phase randomization is the quantum mechanical representation of heat. [Feynman – personal communication].[7] Thus, quantum mechanical uncertainty places limits on measurement skill and establishes an absolute lower limit for entropy.

---

[6] h is Planck's constant.

[7] This insight underlies the so–called "Master Equation" approach to thermodynamics. The master equation is what one gets by carefully averaging over phase in the standard dynamical equations of quantum mechanics [Prigogene].

Finding a well–behaved entropy. The Boltzmann H theorem.

The second law implies that the entropy of an isolated system (or, more properly, an ensemble of isolated systems) cannot decrease with time. As an isolated system evolves toward equilibrium, its entropy will increase monotonically. This time behavior of entropy holds even though the system is governed on a microscopic scale by dynamical laws which are invariant under time reversal. The Boltzmann H theorem attempts to show how microscopically reversible laws can imply macroscopically irreversible behavior.[8] The H theorem can be better appreciated if we try first on our own to define an entropy which evolves in the way we expect and which is computable in terms of microscopic quantities. Simply computing the integral over state space of $\rho \ln(\rho)$, where ρ is the state space probability density, does not do the job; we will shortly see that this "fine–grained" entropy does not evolve, it is a constant of the motion. The Boltzmann H theorem, as interpreted by Gibbs and modified by P. and T. Ehrenfest, succeeds in defining a quantity H which behaves correctly. Gibbs interprets H in the context of an ensemble of systems. He suggests that an ensemble flows through state space much as ink mixes with water when the two are stirred. The Ehrenfests inject Gibbs' interpretation into the definition of H. They introduce a "course graining" procedure which allows the calculation of H in terms of the probability density ρ.

Recall that the detailed configuration of a system corresponds to a point in state space. As the system evolves with time, its detailed configuration changes and the point in state space corresponding to the system moves. Now consider the time evolution of a small element of state

---

[8]The foregoing is the standard introduction to the H theorem which one finds in many texts. We think the H theorem is significant for a different reason; the H theorem shows how probability theory can apply to deterministic systems.

space volume. A way to keep track of the volume element is to follow the points which comprise its boundary as they move.[9] As time goes by, we see the volume element stretch into a long and complex filament which wraps round and round the state space. Note that in spite of this stretching we know, by Liouville's theorem, that the volume of the little element will be conserved. Suppose that all of the members of an ensemble of systems have initial configurations which correspond to points in this little volume element. The ensemble is described by a probability density which is initially quite simple. The density is uniform inside of the volume element and it is zero outside. After a while the volume element has become a convoluted filament; this filament still includes the configurations of all of the members of the ensemble. The probability density of the ensemble is uniform inside the filament and it is zero outside of the filament. Liouville's theorem implies that inside the filament this density has the same value as it had initially.

Now consider an ensemble of systems with a probability distribution over state space described by some density function ρ. Conceptually we can partition the state space with a grid of very fine (differential sized) volume elements. The density function ρ can be taken to be uniform within any one of these differential volume elements. Now we let the ensemble evolve. The differential volume elements all stretch into convoluted filamentary shapes; these shapes never actually intersect one another, but they do become mutually entwined in very complicated ways. Within any filament the density ρ retains the same value as it had initially. Since the fine grid

---

[9]Recall that an elementary property of state space is that two distinct points can never collide, because at the instant of collision and forever after they must follow the same trajectory and by time reversal symmetry they must have been on the same trajectory forever before as well. Thus, points inside a closed boundary can never escape to the outside, because they cannot cross the boundary.

partitioned the space initially, the unruly collection of filamentary shapes which the grid has become still does manage to partition the space. The constancy of the density ρ within a filament, coupled with the space partitioning property of the filaments implies a remarkable fact: the integral of ρ or of any function of ρ over the entire state space will remain constant over time. Thus the entropy, which is the integral of $\rho \ln(\rho)$ over the space, will remain constant in time.

Consider the behavior of the filaments. They become increasingly disordered and jumbled as time progresses. The course graining procedure of the Ehrenfests captures the essence of this filamentary behavior. At time zero partition the state space with a small (but not differential sized) grid. This grid is fixed once and for all; it does not change with time. Form a new distribution function P (capital ρ) which is constant within each cell and which in each cell is equal to the average of ρ over that cell. The quantity H of the Boltzmann H theorem is the entropy of the distribution P. We may suppose that at time zero the function ρ has been chosen so that the two functions ρ and P agree with one another very nearly. Now let time evolve. Each cell is invaded by a jumbled mixture of filaments; the filaments began life in other cells and so each carries a (generally) different, constant density. Thus each cell, which initially contained but a single density, now contains a jumbled mixture of densities. The function P is the average of these densities on a cell by cell basis. It is obvious (or anyway it is trivial to show) that the entropy of P, which we recall is H, will increase provided that the mixing of the filaments becomes ever more fine. A

hypothesis of the Boltzmann H theorem is that the mixing of the filaments does become ever more fine.[10]

How ensembles predict single systems. Ergodicity.

Notice that a probability distribution over state space only attains a predictive value in the context of a large ensemble of systems. Unfortunately, in statistical mechanics, we usually work with only a single system. At any instant of time our system occupies only a single point in state space; one cannot do statistics on a single point! The usual remedy for this deficiency is to replace the set of points in state space which the ensemble would have provided with the set of points occupied over time by the single system under investigation. The replacement of ensemble averages by time averages produces results which agree with experiment.

Experimental verification aside, nobody has ever been able to prove the validity of this replacement without introducing some hypothesis in addition to the known laws of physics. Such hypotheses have usually been called "ergodic hypotheses." The first of them was advanced by Boltzmann who also was the first to use the current terminology. Boltzmann conjectured that each surface of constant energy consists of a single trajectory. In other words, no matter what is the state of the system at a given time, it will pass (or has already passed) through any other state with the same value of the total energy. Using this hypothesis, it is possible to establish the coincidence of time averages with ensemble averages on surfaces of constant energy. Unfortunately, subsequent to Boltzmann, mathematicians have pointed out that this ergodic hypothesis is self contradictory; since a trajectory cannot have multiple points, it cannot fill a multidimensional

[10]The "theorem" in Boltzmann's H theorem is something of a misnomer since bonafide theorems don't ordinarily have hypotheses in them.

volume. The ergodic hypothesis is associated with the most profound questions of statistical mechanics. These questions have been much studied in the decades since Boltzmann introduced them, [Khinchin 1] is a readable review. They have not been resolved completely even yet.

Classical thermodynamics

The classical approach to thermodynamics starts with the first and second laws and from them constructs an elegant chain of reasoning along which the entire subject is developed. The first of the classical laws is conservation of energy. This law, which really is more a definition than a law, defines heat and work as the two forms in which energy can occur. The second law states that the natural direction in which heat flows is always from warmer bodies to cooler bodies. This law is supported by years and years of accumulated experience. From these two laws and one ingenious construct, the reversible cyclic engine, the old masters were able to define absolute temperature and to deduce the existence of entropy. They then defined a useful quantity, the free energy, and used this quantity to characterize the nature of equilibrium. Thus they deduced the whole subject from two laws. [Fermi] is an excellent exposition of this approach to thermodynamics. [Callen] contains a more modern treatment, but one that is still very much in the classical tradition.

Energy can be neither created nor destroyed: $dU=dQ+dW$.

This law can be regarded as the definition of heat: the amount of heat dQ which flows into a system is always that exact quantity which makes up the difference between the change in internal energy dU and the mechanical work dW for which we can account. In practice dQ is determined

experimentally by calibration against a phenomenological scale of temperature. One assumes that the unknown heat dQ is equal to that work dW which produces an equivalent change in temperature.

The possible transformations of energy & the second law.

Conservation of energy, the first law of thermodynamics, places no limitations on the possibility of transforming energy from one form into another. Both empirically and theoretically there appear to be no limitations on the transformation of work into heat; mechanical work can always be converted totally into heat by means of friction. There are very definite limitations however, to the possibility of transforming heat into work. Heat flows spontaneously from warmer bodies to cooler bodies when the bodies are in contact.[11] Clausius postulates that it is impossible to find a transformation whose only final result is to transfer heat from a body at a given temperature to a body at a higher temperature. Lord Kelvin postulates that it is impossible to find a transformation whose only final result is to transform into work heat extracted from a source which is at the same temperature throughout.

Either of these postulates can be taken as the classical version of the second law of thermodynamics; we can show that the two are equivalent. This equivalence is proved by showing that if the Clausius postulate were not valid, then neither would be the Kelvin postulate, and vice versa. If the Kelvin postulate were not valid, then we could perform a transformation whose only final result would be to transform completely into work a definite amount of heat taken from a single source at the temperature t_1. But we could then convert this work by means of friction into heat, with which we

[11]This behavior defines an empirical scale of temperature according to which we can compare the relative hotness of things.

raise the temperature of some other body. If this other body initially was at a higher temperature t_2, then the only final result of this process would be the transfer of heat from a body at a given temperature to a body at a higher temperature. This would be a violation of the Clausius postulate. On the other hand, suppose that the Clausius postulate were invalid. Then we could transfer some heat Q_2 from a body at temperature t_1 to a body at the higher temperature t_2 in such a way that no other change in the state of the system occurred. But then, with the aid of a heat engine (to be discussed shortly), we could absorb this same heat Q_2 and extract work as we cooled back down to the temperature t_1. Since the source at the temperature t_2 receives and gives up the same amount of heat, it suffers no net change. But this would violate the Kelvin postulate, since we have succeeded in transforming into work, heat extracted from a source which is at the same temperature t_1 throughout.

Work from heat via Carnot cycle. The efficiency η

If we have two sources of heat at different temperatures, then we can transform heat into work via an elegant process known as a Carnot cycle. This reversible process consists of an alternating sequence of isothermal and adiabatic transformations cleverly arranged so that the engine performing the transformations ends the cycle in the same macroscopic state as when it began. The first isothermal transformation absorbs an amount of heat Q_2 from a source at temperature t_2, while the second isothermal transformation surrenders an amount of heat Q_1 to a source at a lower termperature t_1. The purpose of the first adiabatic transformation is to cool the engine from the temperature t_2 down to t_1; since no heat flows during this phase, some work is performed. Similarly, the second adiabatic transformation warms the engine back up to t_2; again no heat flows, but in this case some work is

absorbed. Since the engine begins and ends in the same state, it must be that the total work performed during one cycle is $W = Q_2 - Q_1$. The efficiency of the Carnot cycle is defined as the ratio of the work performed to the heat extracted from the high temperature source. Thus the efficiency $\eta = W/Q_2 = 1 - Q_1/Q_2$. Whatever limitations attend the transformation of heat into work must show up as limitations on the ratio Q_1/Q_2.

$$\eta(\text{irreversible}) <= \eta(\text{Carnot}) = \eta(\text{reversible}).$$

If all the transformations comprising the Carnot cycle are reversed then we have a refrigerator. The net effect of a reversed Carnot cycle is to absorb the work W instead of producing it; also, Q_1 is absorbed at temperature t_1 and Q_2 is surrendered at temperature t_2. Using the Kelvin postulate and the idea of a reverse Carnot cycle, it is possible to prove that of all cyclic engines operating between the temperatures t_1 and t_2, the reversible ones all have the same efficiency and this efficiency exceeds that of any nonreversible engine. The old masters prove this fundamental result by devising an ingeneous "null" process whereby an arbitrary heat engine and a reversed Carnot engine exactly cancel out one another's effect on the heat source at the higher temperature t_2.

Specifically, N reverse cycles of the Carnot engine follow N' cycles of the arbitrary engine where N and N' are chosen so that $N'Q_2' = NQ_2$; in this defining relation Q_2' is the unsigned heat absorbed per cycle by the arbitrary engine from the source at the higher temperature t_2, and Q_2 is the unsigned heat surrendered to this source by a reverse Carnot cycle. The Kelvin postulate then implies that the total work W_{total} accomplished by this combination of engines must be nonpositive, since the entire process exchanges net heat only with a source at a single temperature t_1. Since $Q_{2,total} = 0$ by construction, conservation of energy implies that $W_{total} =$

$-Q_{1,total}$. Thus the Kelvin postulate implies $Q_{1,total} >= 0$. But $Q_{1,total} = N'Q_1' - NQ_1$, since the process consists of N' cycles of the arbitrary engine surrendering the unsigned heat Q_1' per cycle followed by N reverse Carnot cycles, each absorbing the unsigned heat Q_1. Thus $N'Q_1' - NQ_1 >= 0$. Substitute in this last relation the expression $N' = N^*Q_2/Q_2'$, obtained from the defining relation of N and N'. The result is $N(Q_2Q_1'/Q_2' - Q_1) >= 0$. In this last relation we can divide by the factor NQ_2 without altering the sense of the inequality since, by hypothesis, $N > 0$ and $Q_2 > 0$ on account of it being an unsigned quantity. Thus we obtain the fundamental result: $Q_1'/Q_2' >= Q_1/Q_2$. The fundamental result implies directly that $\eta' <= \eta$; thus the efficiency of the arbitrary engine can never be greater than the efficiency of the Carnot engine. Finally, consider the case where the arbitrary engine is itself reversible. In this case we can interchange the roles of the two engines in our construction and obtain an inequality opposite in sense to that which we had previously. Both inequalities must hold and so we conclude that the arbitrary reversible engine has the same efficiency as the Carnot engine.

The absolute temperature. $T_2/T_1 = Q_2/Q_1$ of a reversible cyclic engine.

The fundamental theorem shows that the ratio Q_2/Q_1 is the same for all reversible engines operating between the empirical temperatures t_1 and t_2. Thus $Q_2/Q_1 = f(t_1,t_2)$. We now deduce a key property of the function f via another tricky construction of the classical line – this time a "null" process involving three heat sources. Imagine two reversible cyclic engines R_1 and R_2. R_1 operates between the temperatures t_0 and t_1, thus $f(t_0,t_1) = Q_1/Q_0$. R_2 operates between t_0 and t_2, thus $f(t_0,t_2) = Q_2/Q_0$. Dividing we obtain $Q_2/Q_1 = f(t_0,t_2)/f(t_0,t_1)$. Notice that we have conveniently arranged things so that both engines exchange the same heat Q_0 with the body at t_0. Now the

classic trick: consider the reversible process consisting of a direct cycle of R_2 and a reverse cyclic of R_1. This compound process exchanges no net heat with the t_0 source; it absorbs Q_2 from the source at t_2, and expells Q_1 to the source at t_1. Thus, from the definition of the function f, $Q_2/Q_1 = f(t_1,t_2)$. Equating the two expressions for Q_2/Q_1, we obtain $f(t_1,t_2) = f(t_0,t_2)/f(t_0,t_1)$. Since t_0 is arbitrary, we conclude that $f(t_1,t_2) = T(t_2)/T(t_1)$, where T is some function which depends upon the choice of empirical temperature scale. The scale of temperature is arbitrary; a very convenient choice is to use T itself instead of t. T is called the absolute thermodynamic temperature. Notice that T is determined to within a constant multiplicative factor; we are thus free to choose the units of the new temperature scale; conventionally the difference between the boiling and freezing temperature of water at one atmosphere of pressure is taken to be 100 degrees. It is possible to show that this absolute thermodynamic scale of temperature coincides with the empirical temperature as determined by a gas thermometer.

Entropy.

The discovery of the state function entropy is the crowning achievement of classical thermodynamics.

Sum of $Q_i/T_i <= 0$.

Consider a system running a cycle which exchanges heat with several different sources. Suppose the system exchanges the signed heat Q_i with the source at temperature T_i; Q_i is positive if the system absorbs the heat from the source i, otherwise it is negative. Now introduce one last source at temperature T and a bevy of Carnot engines C_i, where for each i, C_i runs between the source at temperature T_i and the source at temperature T. We

adjust each C_i so that it absorbs from the source i the heat $-Q_i$. Thus after a complex cycle consisting of one cycle of the system and one cycle of each of the Carnot engines, we find that no net heat has been exchanged with any of the sources i. However, the source at temperature T has surrendered an amount of heat Q equal to the sum over the other sources i of TQ_i/T_i. Thus the net effect of the complex cycle has been to transform into work an amount of heat Q received from a source at a uniform temperature T. The Kelvin postulate requires that Q<= 0. Thus, for any cyclic process the sum of Q_i/T_i is always<= 0.

Integral dQ/T= 0 around any reversible cyclic transformation.

In deriving the result that the sum of Q_i/T_i<= 0 for an arbitrary cyclic system, we assumed that the system exchanged heat with a finite number of sources. Instead, the system might exchange heat with a continuous distribution of sources; then the sum over the sources becomes an integral around the cycle and the heat received by the system from any single source at a temperature T becomes the infinitesimal dQ. Thus, for an arbitrary system exchanging heat with a continuous distribution of sources, we know that the integral of dQ/T around a cycle is<= 0. Notice that if the system is reversible, then by running it in reverse we conclude that the integral of −dQ/T around the cycle is<= 0. Thus we conclude that the integral of dQ/T around a reversible cycle is identically zero.

State function S: dS= dQ/T for reversible dQ.

Consider now the integral of dQ/T along some reversible transformtion which takes the system from a standard initial state O to some final state A. Let the value of this integral be S. We could make a complete cycle and net a zero result by continuing the integral along any reversible

transformation from A back to O. Thus the integral along any reversible transformation from A to O must be −S. Evidently for a fixed initial state O, the integral S depends only on the final state A; thus S= S(A). S is a state function; it is called the entropy. More generally, the integral of dQ/T from A to B along any reversible transformation is S(B)−S(A). Differentiating the integral relation, we see that dS= dQ/T along any reversible infinitesimal transformation.

dS>= dQ/T for general dQ.

Suppose we take our system from some state A to some other state B via an irreversible transformation I, and back to A again via a reversible transformation R. I and R together form an irreversible cycle. We know that the integral of dQ/T around this cycle is <=0. But this integral consists of two pieces: the integral along I and the integral along R. The integral from B to A along the reversible transformation R, by definition just gives the entropy of A relative to B, S(A)−S(B). The integral around the entire cycle, which we know is <=0, equals the integral along the irreversible transformation I plus S(A)−S(B). Thus the integral along the irreversible transformation I <= S((B)−S(A). The differential form of this result is that dS>= dQ/T for an arbitrary infinitesimal transformation involving a heat flow dQ from a heat bath at temperature T.

Thermodynamic potentials. The free energy.

The work L performed by a purely mechanical system is always equal to minus the variation of its energy ΔU. Thus L= −ΔU. For thermodynamic systems there is no such simple relationship between the work performed and the variation in energy, because the energy can be exchanged between the system and its environment in the form of heat. The

first law of thermodynamics correctly accounts the relationship between heat, work and energy. This law takes the form $L = -\Delta U + Q$. Suppose that a system is in thermal contact with its environment (to be modeled as a heat bath), which remains at a constant temperature T as the system is transformed from an initial state A to a final state B. We know that the integral of dQ/T is less than the change in entropy associated with the transformation from A to B. Because T is assumed constant, we even know that the integral of dQ from A to B is $<= T[S(B)-S(A)]$. We thus obtain an upper bound on the amount of heat which the system can receive from the environment. Combining this bound with the first law, we conclude that the work L perfomed by the system during the transformation from A to B is $<= U(A)-U(B) + T[S(A)-S(B)]$. This motivates the definition of the state function F where $F = U-TS$. F is called the free energy and evidently the work $L<=F(A)-F(B) = -\Delta F$. Compare the thermodynamic result $L<= -\Delta F$ with the corresponding identity from mechanics $L = -\Delta U$; this is what motivates the name free energy, the work performed is bounded above by minus the change in the energy that is free.

Equilibrium & the minimum of free energy.

Consider a system S, which can exchange heat but not work with its environment. Systems such as S are said to be dynamically isolated. For any transformation of S, we know that $L = 0$; if the environment of S is at a constant temperature, then we can conclude that $0<= F(A)-F(B)$ and hence that $F(B)<= F(A)$. So we see that the free energy of a dynamically isolated system is always decreasing, or at least is always nonincreasing. A consequence of this fact is that, if the free energy is a minimum, then the system is in a state of stable equilibrium.

The principle of maximum entropy

The combined use of probability theory and the Boltzmann distribution makes possible a natural and mathematically clean formulation of thermodynamics.[12] However, probability theory alone can not generate the Boltzmann distribution; arriving at this distribution requires some sort of additional assumption. The principle of maximum entropy provides such an assumption in a simple and usable way. A drawback of this line of development is its blindness to some of the really fundamental issues of statistical mechanics; issues like ergodicity and generally the question of the extent to which an ensemble average represents the behavior of any one particular system. On the other hand, the maximum entropy principle has a strong foundation in statistics. In fact, if maximum entropy based inference should fail, then one can draw some very powerful conclusions. The original references on maximum entropy are the pair of papers [Jaynes 1] and [Jaynes 2]. More recently, the text [Tribus] does a very credible job of developing thermodynamics from the hypothesis of maximum entropy.

Statistical estimation & maximum entropy.

The generic problem which the principle of maximum entropy addresses is that of estimating some parameter of a probability distribution when this distribution is only partially specified. The problem is ill–posed; its solution requires some extra principle of statistical estimation, such as one of "minimum bias," or equivalently, "maximum uncertainty." A great advance provided by information theory lies in the discovery that there is a unique, unambiguous criterion for the "amount of uncertainty" represented by a discrete probability distribution. Shannon has proved that entropy can be

---

[12] We will explore this formulation a little later in the thesis.

uniquely characterized as that quantity which is positive, which increases with increasing uncertainty and which is additive for independent sources of uncertainty. The generic problem becomes one of finding a probability assignment which maximizes entropy while agreeing with whatever constraints are implied by the partial specification of the distribution.

The Boltzmann distribution from maximum entropy + constraints.

The canonical example of the application of maximum entropy to thermodynamics is its derivation of the Boltzmann distribution. The problem is to find the probability distribution over state space which has maximum entropy, and which satisfies two constraints: the sum over the probabilities so deduced should be unity, and the expected energy of the distribution should agree with a particular given value U. Lagrange's method of undetermined multipliers is used to solve this problem of constrained extremization. β is the Lagrange multiplier associated with constraint of fixed expected energy. $\ln$ Z is the Lagrange multiplier associated with the constraint of a unity sum over the assigned probabilities. The result is the familiar Boltzmann distribution.

A general identity associated with this method of constrained extremization is that the partial derivative of the extremized quantity (entropy) with respect to the value of the constraint (U) equals the value of the Lagrange multiplier (β). In taking this partial derivative, the variations are confined to those quantities over which the extremization has been taken. In the case at hand, this identity says that the partial derivative of entropy with respect to (mean) energy is β. In this partial derivative, the probability assignments are varied, not the state energies. Since changes in U on account of probability variations do constitute heat flow, we see that β as a Lagrange multiplier in the maximization of entropy is exactly the same as

the β defined in statistical mechanics. In each case β is the partial derivative of H with respect to U, and in each case the variation in U is to be accomplished by means of heat flow alone.[13]

β's are equal in equilibrium.

A nice feature of the definition of β with maximum entropy is that it allows us to readily deduce that the β's of two systems in thermal contact must agree if the two are to be mutually in equilibrium. We suppose that entropy is additive (extensive), so that the entropy of two systems in thermal contact is just the sum of their separate entropies. This extensivity of entropy will be true if the interaction of the two systems occurs only through mechanisms that already have been operating in each system alone, i.e., no surface interactions. Now we consider maximizing the combined entropy of the two systems. Suppose that the β of system one is less than the β of system two. Then by taking a little energy in the form of heat, dQ, from system one, and adding that energy as heat to system two, we would increase the joint entropy of the two systems by the amount $dQ(\beta_2-\beta_1)$. We could effect this flow of heat by appropriately diddling the state occupation probabilities of the two systems. Similarly, if the β of system one is larger then the β of system two, then by moving heat in the opposite direction we could again increase the total entropy. So we see that in order for the entropy of the combination to be at a maximum, it must be that the β's throughout agree uniformly.

---

[13]We note in passing that this definition of heat flow is somewhat over–restrictive; confining variations solely to the state occupation probabilities isn't the only way to insure that heat flow alone happens. In principle, the state energies can be allowed to vary too, provided that the average of their variations, weighted by the probabilities of state occupation, remains zero.

Any failure of maximum entropy uncovers new physics.

The maximum entropy principle leads to the broadest distribution that is consonant with the initial data. It follows that any sharp predictions of the principle must be consistant with the vast majority of states to which any appreciable weight is assigned. In a similar vein, it is reasonable to assume that experimentally reproducible results must be consistant with the vast majority of states that are compatible with the conditions of an experiment. Suppose that maximum entropy inference, based on knowledge of experimentally imposed conditions, makes a prediction which is refuted by further experiments. Then there must be a serious discrepancy between the fraction of states in the maximum entropy distribution that are consistant with the prediction, and the fraction of physically allowed states that are consistant with the prediction. A similar discrepancy will be revealed if a phenomenon is found which is experimentally reproducible, but which is not predictable by maximum entropy inference.

> "In either case there must exist new physical states, or new constraints on the physically accessible states, not contained in the presently known laws of physics. Thus if it can be shown that the class of phenomena predictable by maximum entropy differs in any way from the class of experimentally reproducible phenomena, that fact would demonstrate the existence of new laws of physics, not presently known."[14]

Information Theory

In 1948 Claude Shannon published a seminal article in which he developed a mathematical theory of communication. The fundamental problem of communication, according to Shannon, "is that of reproducing at

---

[14][Jaynes II, pg 172]

one point either exactly or approximately a message selected at another point." The message is generated by an information source; this source selects the message from a large set of possible messages. The message is transmitted from one point to another via an imperfect channel; on account of noise, the message received at the output of the channel may not exactly correspond to the message that was transmitted. Information theory, as Shannon's theory has come to be called, quantifies and compares the rate of information production of a source, with the information carrying capacity of a channel.[15] [Shannon] remains, in many respects, the best exposition of information theory. [McEliece] is a modern textbook on information theory which contains, among other things, a compendium of all the inequalities around which the subject is built. [Pierce] is a readable and thoughtful text which develops information theory and traces its connection to other disciplines. [Khinchin 2] discusses the mathematical foundations of information theory; his treatment of entropy is especially significant.

Insight into entropy is the essence of information theory.

Shannon proposes that the amount of information in a message depends on how much "choice" is involved in the selection of the message. The selection process chooses the message from a set of possible messages in a random way. Our uncertainty about the outcome of the selection process measures the amount of information in the message. Shannon proves that any internally consistant measure of choice or uncertainty must necessarily be based on entropy. If all the messages in the set of possible messages are equally likely, then the entropy is just the logarithm of the number of

---

[15]Shannon develops two versions of his theory: one version for discrete sources and channels, and another where sources and channels are continuous. In this thesis, we will deal only with the discrete version of Shannon's theory.

messages. If all the messages are not equally likely, then the entropy is proportional to the logarithm of the number of "reasonably probable" messages. The number of reasonably probable messages would appear to be a somewhat subjective quantity. Surprisingly, when the set of possible messages is sufficiently large, this seemingly qualitative definition of entropy actually does manage to specify a precise quantity. The notion of a set of "reasonably probable" alternatives, where the size of the set somehow manages to be independent of one's precise interpretation of the words "reasonably probable" is subtle and difficult to appreciate.[16] This notion motivates a powerful understanding of entropy which Shannon uses very effectively; it is perhaps his most significant contribution.

The entropy of an information source

We can think of a discrete source as generating a message, symbol by symbol. One model of a discrete source might be that the successive symbols of a message are chosen at random from some probability distribution over the set of possible symbols. A slightly more sophisticated model for a source would take into account the probabilities of pairs of symbols. Here the source would be modeled as a Markoff process, so that the probability distribution governing the ith symbol is conditional upon the $(i-1)^{st}$ symbol. The point at issue is the statistical structure of the source. We can generate a more and more accurate statistical approximation to any source, if we let the probability distribution of a symbol depend on more and more of the preceeding symbols.

We want to know the number of messages of length n, N(n), which such a source might produce. Clearly this number will be variable since the

---

[16]We will explore this idea more deeply in the body of the thesis.

message generation process involves chance. Shannon directs our attention to long messages for which we expect the statistical variability to be proportionately less significant. He proves the remarkable fact that, in the limit as n goes to infinity, the log of the number of "reasonably probable messages", $\log N(n)$, is independent of the precise definition of "reasonably probable." He shows that, as n becomes large, the measure $(1/n)\log N(n)$ approaches a fixed limit H, where H depends only on the statistical properties of the source.

The quantity H measures the information content of the source. H is known as the entropy of the source. Typically H will have dimensions like bits per symbol. Fortunately, H can be calculated without resorting to the difficult technique of counting N(n). In the simple case where the source produces symbols as though they are independent random draws of some probability distribution, Shannon shows that H is just the entropy of the probability distribution. H can also be calculated for the more complicated case, where the probability distribution of the i^{th} symbol is conditional upon some number of preceeding symbols. Here Shannon introduces a new kind of entropy: conditional entropy. The conditional entropy is just the entropy of a conditional probability distribution. H is the expected value of the conditional entropy of the (conditional) probability distribution which governs the generation of symbols by the source.[17]

The capacity of a channel & mutual information.

A message is transmitted across a channel one symbol at a time. Different symbols may take different amounts of time to transmit. The capacity C of a discrete noiseless channel is defined as C= (1/n)Log N(n),

---

[17]Later we will focus much more closely on conditional entropy.

where N(n) is the number of possible messages of duration n; C is to be evaluated in the limit where n is taken to infinity. This limit is interesting because the number of different messages of duration n invariably grows exponentially with n.

A noisy channel is one which corrupts symbols so that the received message does not necessarily reflect the message that is transmitted. A reasonably general model of a channel with noise is the so called discrete memoryless channel.[18] Discrete memoryless channels are those for which the probability that any transmitted symbol x is corrupted, so that it is received as some other symbol y, depends only on x and y, and not on the symbols preceeding x which already have passed through the channel. The behavior of a discrete memoryless channel is thus completely specified by the set of conditional probabilities p(y|x).

Channel noise is significant only insofar as it makes it impossible for us to distinguish, on the basis of the received signal alone, between similar but distinct transmitted messages. Shannon suggests that the relevant measure of the information carrying capacity of a noisy channel is given by $(1/n)\log N(n)$; here N(n) is the number of reasonably probable distinct transmittable messages of length n which can be reliably distinguished at the output of the channel. This information measure, which has come to be known as the mutual information, I, can be expressed as a difference of entropies. The entropy H(y) is the logarithm of the number of reasonably probable messages that can happen at the output of the channel. The conditional entropy H(y|x) is the logarithm of the number of reasonably probable output messages to which a single typical input message may give rise. The mutual information is given by their difference; thus I= H(y)–

[18]Anyway, it's the most realistic model for which anything can be accomplished analytically.

H(y|x). We can understand this expression for I by noticing that the number of a distinguishable messages N(n) can be estimated as: (# output messages) / (#output messages that may arise from a single input message). I is just the logarithm of this quotient.

In general, the mutual information depends on both the symbol corruption probabilities of the channel p(y|x) and on the statistical composition of the messages we transmit p(x). The maximum of the mutual information I is defined as the capacity C of the noisy channel. In finding this maximum we are to search over the space of all possible statistical sources of messages. Thus C= $\max_{\text{over } p(x)}$ of I.

Comparing source entropy & channel capacity: the fundamental theorem.

The justification, ultimately, for Shannon's definition of the channel capacity C and the source entropy H is that they can be meaningfully intercompared. Shannon proves a fundamental theorem: when H is less than C it is possible to transmit long messages across the channel and have them be received with a negligible probability of error. When H is greater than C, such error–free reception is not possible, even in principle. Shannon's theorem rests upon two observations. The first is that for a channel with capacity C, there exists a set of about 2^{nC} messages of length n which can be sent across the channel and be reliably distinguished upon reception. The second observation is that a source with entropy H will produce no more than about 2^{nH} distinct messages of length n. The theorem is really just the statement that error–free transmission is possible only when the set of messages produced by the source is smaller in number than the set which can cross the channel and remain distinguishable.

Even when error–free performance is allowed by Shannon's theorem, it is not easy to attain. In general, the set of messages produced by the

source will not be the same as the set of messages which can be reliably distinguished after transmission across the channel. In order to use the channel effectively, we must transmit only messages of the distinguishable set. On the other hand, our whole purpose for using the channel is to communicate arbitrary messages. The resolution to this dilemma is to encode the messages produced by the source so that they appear to be messages of the distinguishable set. In other words, we must concoct a transformation which carries each element of the set of possible source messages into a unique element of the set of distinguishable messages. This transformation should have an inverse so that upon reception, the original message of the source can be recovered. Shannon's theorem is the statement that when H<C such a transformation exists; conversely, when H>C such a transformation does not exist. Finding the transformation in any particular case is usually extemely difficult; this is the province of coding theory.

The fundamental theorem & coding theory.

Coding theory is a difficult subject. Designing codes which are tailored to optimally handle the corruption probabilities of any particular channel is beyond the current state of the art in coding theory. Instead, attention has focused on the construction of so–called error correcting codes. A typical example of an error correcting code is the (7,4) Hamming code which forms words of 7 binary symbols apiece. Each (7,4) codeword consists of 4 bits of source information (=4 binary symbols if the source has an entropy of 1 bit per symbol) concatenated with 3 binary symbols of generalized parity. The (7,4) code enables us to recognize and correct any one symbol error in a codeword. Notice that the (7,4) codeword packs only 4 bits of information into 7 binary symbols; thus the effective entropy rate of the source is reduced to 4/7 =.57 bits of information per symbol. In the

years since Shannon developed the theory of communication, coding theory has grown into a rich and active discipline. Several ingenious and elegant algorithms are now known which implement a few kinds of error correcting codes. To date all of the known codes are based on the algebra of finite fields. These codes enjoy widespread use in diverse applications. The search for more and better codes continues, but progress is slow; coding theory remains a difficult subject. See [McEliece] for a thorough introduction to coding theory.

The data processing theorem.

Consider a communication setup in which the signal is transmitted sequentially through two independent channels. The signal suffers some degradation as it passes through the first channel, and then it suffers additional degradation as it passes through the second channel. Suppose that the mutual information between the source and the output of the first channel is I_1, and that the mutual information between the source and the output of the second channel is I_2. A fundamental result of information theory is that I_2 can never exceed I_1. Thus the information content of a signal is never enhanced by transmission through an additional channel. This result, which is known as the data processing theorem, applies in any situation where data is processed and where the most direct connection between the processing equipment and the source of the data is the data itself. In these situations, the mutual information which connects the data to its source is always degraded (or at best is unchanged) by the processing which the data receives.

The relation between thermodynamics and information theory

Since its inception in 1948, information theory has stirred the imagination of physicists. The feeling has remained that, in some way, information theory and physics must share a profound connection. Nonetheless, very little has been accomplished in the way of connecting the two subjects at a deep theoretical level; to date, [Szilard] and [Brillouin] are the best known attempts in this direction. Both of these authors concentrate on the problem of Maxwell's demon. The problem or paradox of Maxwell's demon has been the battle ground where theoretical physics and abstract information collide. More recently, researchers working on the physics of computation have met up with the demon; [Bennett] contains a summary of this work. Historical popularity notwithstanding, Maxwell's demon has not been a fruitful avenue of investigation for those wishing to find a connection between information theory and physics. In this thesis we take a different tack and try to establish a structural relation between the two subjects. This approach leads us to an identity involving mutual information and free energy. The statement and proof of this identity forms the core of this thesis.

Maxwell's demon: the canonical crucible for mixing information & physics.

Historically, the problem of Maxwell's demon has been the point of departure for any discussion which combines physics with a theory of information. The sorting demon was born in 1871 in Maxwell's Theory of

Heat as "a being whose faculties are so sharpened that he can follow every molecule" and is thus

> "able to do what is at present impossible to us. Let us suppose that a vessel is divided into two portions, A and B by a division in which there is a small hole, and that a being who can see the individual molecules opens and closes this hole, so as to allow only the swifter molecules pass from A to B, and only the slower ones to pass from B to A. He will thus, without expenditure of work, raise the temperature of B and lower that of A, in contradiction to the second law of thermodynamics."

Generations of physicists have considered this paradox; there have been various attempts to discredit the demon. One line of attack proceeds by analyzing various prototype demons. The results suggest (but do not prove) that failure of the demon is always inherent in the physical attributes which comprise the demon.[19] A fundamentally different kind of explanation for the demon was first raised by Szilard. He investigated the connection between the information which the demon must acquire about the detailed motion of the gas and the change in entropy of the physical system which this information makes possible.

The inherent imperfections of a physical demon.

Various simple demon prototypes have been proposed. Common to all the prototypes has been the use of some device having an asymmetric response function. The idea here is to extract energy from thermal noise by rigging some sort of asymmetric widget (the demon) which does work when random thermal agitation moves it in one direction, and which is unresponsive to thermal agitation which would tend to move it in the opposite direction. Detailed analysis of each of these mechanisms shows that

---

[19]There have been other, less significant attempts to discredit the demon. [Brillouin] contains a nice review.

none of them is viable for long term ("perpetual") operation of the demon. The similar manner by which each mechanism fails suggests that a deep physical principle is at work.

Smoluchowski has analyzed a one–way valve which controls the flow of a gas between two vessels. Brillouin has analyzed an electrical rectifier connected in series with an inductor; the series combination is driven by a noise source such as a resistor. Feynman has analyzed a ratchet and pawl arranged so that the rotation of the ratchet lifts a weight; the ratchet is also connected to a set of vanes which are bombarded by the molecules of a gas. All three individuals conclude that the demon mechanism gets warmer and warmer with continued operation and that this heating ultimately nullifies the demon's ability to convert random thermal agitation into stored energy. In each case, heating of the demon causes it to function less than perfectly. Thus the one way valve leaks slightly, the rectifier conducts slightly when it is reverse biased, and the pawl occasionally slips and lets the ratchet turn the wrong way.

The intriguing thing about these examples is that in each case heating of the demon and its subsequent failure appears to be an inherent aspect of the design of the demon. Consider, for example, the simple one–way valve, which consists of a thin plate, which in the resting position forms a seal against an orifice. Pressure fluctuations of the right kind deflect the plate and flow past it. Pressure fluctuations of the wrong kind merely seal the plate more firmly against the orifice and are unable to flow past. The plate must return to the resting position after a right kind of fluctuation has passed; thus a restoring force is necessary. Also the plate and the other parts of the valve cannot all be constructed of perfectly elastic parts. If the parts were elastic, then after the passage of a favorable fluctuation, the

restoring force would cause the plate to bounce against the orifice and to keep bouncing. Some kind of a damping or deadening mechanism is necessary to stop the bouncing; this mechanism converts the kinetic energy of the plate, as it returns to the resting position, into heat. Thus the heating of the valve is an essential aspect of its one–way operation.

Can this heating go on forever? No! The plate and the rest of the valve, all at some temperature T, also have a fluctuating (brownian) motion. This motion is such that, every once in a while, by accident, the plate pushes itself away from the orifice just at the moment when a wrong kind of pressure fluctuation is trying to go backwards through the valve. The valve fails to block the wrong way fluctuation; as things become hotter this type of failure occurs more and more often. This failure through heating happens also to the rectifier and to the ratchet and pawl. In each case a damping mechanism is necessary; the damping mechanism allows the demon to settle back to its resting configuration after it has acted to trap a fluctuation. The damping heats the demon and the efficiency of the demon falls as it becomes hotter and hotter.

The implicit cost of information.

In 1929, Szilard published a remarkable paper on the demon which uncovered, for the first time, a connection between information and entropy. Szilard considers a simplified version of Maxwell's demon which operates with only a single gas molecule. The molecule lives in a cylinder which is closed at both ends; the volume of the cylinder can be divided in two (without expending energy) by sliding in a partition at the middle. Szilard's demon extracts work from this apparatus by running a simple cycle. First, the demon installs the partition in the middle of the cylinder. Next, the demon ascertains in which half of the cylinder the molecule is trapped. Finally, the

demon extracts work by slowly expanding the volume accessible to the molecule; this expansion is achieved by sliding the partition, as though it were a piston, toward the end of the cylinder away from the molecule. The demon can then remove the partition from whichever end of the cylinder it has reached and repeat the cycle. Operating in this fashion, the demon gradually converts heat, in the form of the kinetic energy of the molecule, into work.

Szilard studied this paradox and unearthed a fundamental discrepancy at its core. He observed that the entropy measured for the single molecule system would depend on the fund of information available to the measurer. If, for example, the measurer knows in which half–cylinder the molecule resides, then the quoted entropy will be one–half as large as the entropy when measured by an individual who is not so informed. The reason is that the informed measurer sees (or measures, or knows) that the molecule occupies a volume which is half as big as the volume determined by an uninformed measurer. Like most physicists, Szilard desired to save the 2^{nd} law from the demon; his own analysis suggested however, that information about a system can be equivalent to a reduction in entropy of that system. Szilard reached the only conclusion which accommodates both of these concerns: somehow the gathering of the information itself must already cause an increase in entropy somewhere in the universe; moreover, this increase must be at least as large as the decrease which the information effects.

The discrepancy in entropy on account of the demon's knowledge hints that the crucial step to investigate is the one whereby the demon learns the location of the molecule. In 1956, Brillouin published an extensive study on the problem of physical measurement which corroborated Szilard's conclusion and expanded on it. Brillouin succeeded where Szilard had not, because of

an *essential* ingredient which Shannon had provided in the intervening years: the association of information content with uncertainty, as measured by entropy. Using this association, Brillouin found that any experiment which obtains information about a physical system produces, on average, an increase of entropy in the system or in its surroundings. The average entropy increase, is always at least as great as the amount of information obtained. When Szilard's demon learns in which half of the cylinder the molecule resides, he obtains one bit of information or equivalently $\ell n 2$ nats of information; in physical units this corresponds to an entropy of $k\ell n 2$. Thus Brillouin's principle says that the entropy of the universe increases by at least $k\ell n 2$ for every bit of information which the demon learns.

Recently, Bennett and others have studied the thermodynamics of computation. These studies uncover a connection between logical irreversibility and thermodynamic irreversibility. Apparently, only the performance of an operation which is logically irreversible necessarily dissipates free energy; the performance of an operation which is logically reversible can be achieved in a thermodynamically reversible manner. Bennett states that the process of measurement can always be accomplished in a manner that is logically and thermodynamically reversible. He concludes[20] that the step which prevents Maxwell's demon from breaking the 2nd law is not the making of a measurement, but rather the logically irreversible act of erasing the record of one measurement to make room for the next.

---

[20]Professor Mead disputes this conclusion; he observes that Bennett's proof fails to account the state of the demon's decision–making apparatus during the measurement process.

Information Theory & Thermodynamics Share a Common Structure.

In the past, as we have summarized, the attempts to relate thermodynamics and information theory have mostly amounted to detailed analyses of Maxwell demon–type mechanisms. The conclusions of these analyses are interesting, but they are also in conflict with one another. It is fair to say that thermodynamics and information theory are much better understood in isolation than they are in combination. We have a different strategy for relating these two subjects. We proceed from a structural simile: thermodynamics is to free energy as information theory is to mutual information.[21] Consider thermodynamic free energy and information theoretical mutual information. Both of these measures are of central importance to their respective subjects. Free energy is minimized by a special distribution – the Boltzmann distribution of equilibrium. Mutual information is maximized by a special distribution – the distribution which achieves channel capacity. Both are measures of state space volume; both involve entropy. The 2nd law of thermodynamics stipulates that the free energy of an isolated system will always tend to decrease. The data processing theorem of information theory proves that the mutual information of a signal will always be decreased by additional processing. These likenesses suggest that free energy and mutual

---

[21]While writing this section we found it most helpful to see what a dictionary had to say about words which relate things to other things. Four words seem especially relevant –

homologous: corresponding in structure and evolutionary origin, as the flippers of a seal and the arms of a man.

analogous: similar in function but not evolutionary origin.

metaphor: a figure of speech in which a word denoting one subject or idea is used in place of another to suggest a likeness between them (as in "the ship plows the sea.")

simile: a figure of speech in which two dissimilar things are compared by the use of like or as (as in "cheeks like roses)."

information are analogous measures. Is it possible that free energy and mutual information share even a closer bond than this structural analogy? Yes! In this thesis we show that in a certain limit and for the right class of systems the free energy and the mutual information become identical measures. This identity comes from a new asymptotic equality between thermodynamic internal energy and information theoretical conditional entropy.

To obtain these results, we need a viewpoint which allows definition of both thermodynamic and information theoretical quantities simultaneously. Regard time as a channel and the detailed state of a physical system as a message; the state at time zero is the transmitted message, and the state at time t is the received message. In this context, the free energy of the physical system at time t, and the mutual information which links the initial state of the system to the state at the later time t, can be calculated and compared. Since thermodynamics is concerned primarily with equilibrium, we might expect it to overlap information theory only in the limit as the time interval t is taken to infinity. Indeed, we can easily see that the free energy and the mutual information agree with one another in the asymptotic limit of large t. In this limit the mutual information approaches zero because, as the time interval t becomes very long, the state of the system at time t becomes nearly independent of its initial state, and so the mutual information coupling the two becomes negligible. In the limit of large t, physical systems approach equilibrium. The free energy approaches zero in this limit because, in an isolated system at equilibrium, there is no energy which is free and

available for use.[22] Thus, trivially, the free energy and the mutual information both approach zero as t becomes large.

We have found that the asymptotic relation between the free energy and mutual information measures is actually much deeper than the trivial statement that zero equals zero. Regard free energy and mutual information not as numbers, but as operators which map state vectors to numbers. These two fundamental operators can be recast so that they share quite similar forms. We "rescale" the free energy so that it is expressed in units of $-kT$. Also we "rereference" the mutual information so that it deals only with the state vector at time t, and no longer makes explicit reference to the state vector at time zero. The difference between the rescaled free energy operator and the rereferenced mutual information operator is a special kind of operator; it is a linear operator. This linear operator compares the (rescaled) internal energy to the (rereferenced) conditional entropy. In the asymptotic limit of long times, we prove that every component of this linear operator vanishes; thus, asymptotically, the rescaled free energy and the rereferenced mutual information become identical operators.[23]

---

[22]This explanation is somewhat deceptive since it hides the fact that we are really just defining the zero of energy. The free energy depends on the internal energy, which, like any other measure of potential energy, is only ever determined to within an additive constant. Defining the free energy to be zero at equilibrium determines this constant.

[23]Note that an operator equality is richer than a single equation between scalars. The operator equality applies to all possible state vectors and so implies several independent scalar equations.

Relating free energy and mutual information

In this chapter we state and prove our main result: the equality of the free energy and mutual information operators. The mathematics of Markov processes is a language common to both thermodynamics and information theory. In the first part of this chapter we sketch the essentials of both subjects in this language.[1] Mathematically, a Markov process consists of a probability state vector and a dynamical law of evolution which operates on the state vector. The dynamical law can be represented as a matrix Γ; element Γ_{ij} specifies the probability that a system in state j will transit to state i in the next time period.

Free Energy & Thermodynamics.

As an example of a simple thermodynamic process, we consider an ensemble of non–interacting spin 1/2 particles.[2] Each particle in this physical system has two states: a or b, corresponding to the two possible directions of its spin. A heat bath at temperature T agitates the system, causing the particles to flip–flop independently between the two states. If each particle has a magnetic moment (which aligns with its spin) and if we apply an external magnetic field, then the two states are at different energy levels; label these E_a and E_b. For convenience, collect these two energies

---

[1]For a less cryptic review, skip ahead and read the first two pages or so of the section on Markov processes.

[2]Spin 1/2 is a quantum mechanical concept. A measurement of a component (along any specified direction) of angular momentum of a spin 1/2 particle can have only two discrete possible outcomes: $+h/4\pi$ or $-h/4\pi$. Thus the spin points either parallel or antiparallel to the specified direction.

into a row vector $\vec{E}^T$. The system's internal energy U is defined as the mathematical expectation of the system's energy at time n:[3]

$$U[\vec{P}(n)] = E_a P_a(n) + E_b P_b(n) = \vec{E}^T \vec{P}(n) \ , \qquad (1)$$

where $\vec{P}(n)$ is the system's probability state vector, (a column vector) with $P_a(n)$ representing the probability that any particular particle of the system is in state a at time n, and $P_b(n)$ representing the probability that the particle is in state b at time n.

In general, the internal energy of the system changes if either $\vec{E}^T$ changes or $\vec{P}$ changes. Changes in $\vec{E}^T$ happen when we adjust the energy levels of the states of a system. This kind of change requires that we or the system expend work. If, for example, we instantaneously increase the strength of the magnetic field surrounding our spin system, we find that the energy levels of the two states become farther apart and that during the process we exchange work with the system. Changes in $\vec{P}$ happen because the system evolves under the combined influence of the heat bath and the vector of state energies $\vec{E}^T$. This kind of change corresponds to shifts in the fraction of spins pointing up or down in our ensemble of spin 1/2 particles. The internal energy of the ensemble changes as $\vec{P}$ changes, but no useable work is performed. Instead, heat flows between the ensemble and the heat bath. By differentiating (1) and identifying terms, we obtain expressions for work and heat; these definitions comprise the first law of thermodynamics.

$$\begin{aligned} dU &= (d\vec{E}^T)\vec{P} + \vec{E}^T(d\vec{P}) \\ &= dW + dQ \qquad \text{where,} \\ dW &= (d\vec{E}^T)\vec{P} \quad \text{is the work performed } \textit{on the system}, \text{ and} \\ dQ &= \vec{E}^T(d\vec{P}) \quad \text{is the heat flow } \textit{into the system}. \end{aligned}$$

---

[3]Actually, U is the internal energy *per particle*. Particle number is assumed constant; all extensive quantities will be normalized per particle.

Thermodynamic equilibrium can be characterized by Boltzmann's equation which relates negative logs of equilibrium probabilities to energy differences (in units of kT). Recall that Γ is the transition matrix describing, for each state i, the possible states to which a system in state i can jump, and the probabilities of these jumps. The components of the equilibrium probability vector $P_a(eq)$, $P_b(eq)$ are those which balance the probabilities of jumping into and out of each state, so that the net flow ("into" minus "out of") is zero. Thus the transition matrix Γ determines the equilibrium probabilities and hence the ratio $P_b(eq)/P_a(eq)$. Now, since Boltzmann's equation for a two-state system equates the ratio $P_b(eq)/P_a(eq)$ to the negative exponential of $(E_b-E_a)/kT$, we see that this energy difference is implicitly specified by Γ.

Thermodynamic equilibrium can be more elegantly described as the state space probability distribution (a vector $\vec{P}(eq)$) which minimizes the system's free energy. Free energy is denoted $\mathcal{F}[\vec{P}]$, and is defined as $\mathcal{F}[\vec{P}]= U[\vec{P}]-TS[\vec{P}]$, where $U[\vec{P}]$ is the internal energy defined by equation (1), and $S[\vec{P}]$ is the conventional thermodynamic entropy with dimensions Joules/°K.[4] By way of brief intuitive review, one might say that the equation $\mathcal{F}=U-TS$ reckons the free energy $\mathcal{F}$ as the total energy internal to the system, U, less an amount of energy that is tied up in the system as heat and is unavailable for use. This unuseable energy, which is estimated by the term TS, is the unique contribution of thermodynamics.

Boltzmann was the first to deduce the remarkable fact that the thermodynamic entropy S may be identified with the quantity $kH[\vec{P}]$, where k is (Boltzmann's) constant and $H[\vec{P}]$ is the mathematical entropy of the probability distribution $\vec{P}$:

---

[4]Entropy is extensive and is normalized per particle; so is the free energy.

$$H[\vec{P}] = -P_a \ln P_a - P_b \ln P_b \quad . \tag{2}$$

We can interpret the entropy as defined by Equation (2) as Shannon suggests; it measures our uncertainty about which way the spin of a particular particle points. But, as Szilard suggests, our uncertainty (or lack of same) can have physical consequence. [Bennett] shows how, with clever manipulation of a magnetic field, we can extract useful work from a spin by randomizing its orientation. Essentially, we expand the volume of state space available to the spin. The success or failure of this procedure depends on how well we know the state of the spin initially. We can only expand the volume of state space accessible to a spin when the spin doesn't already occupy that volume. Thus, the greater the entropy in (2), the more energy in the system will be tied up and unavailable for use. Using (2), the expression for free energy may be written as

$$\mathcal{F}[\vec{P}] = \vec{E}^T \vec{P} - kT\ H[\vec{P}] \quad . \tag{3}$$

Mutual Information & Information Theory.

Information theory models the transmission of messages over a noisy channel. At one end of the channel is the source and at the other end the receiver. The source transmits a message consisting of a sequence of symbols, much as grammatical text consists of a sequence of letters. We trace the evolution of the message through the channel in units of time: at time 0, the message is transmitted at the source, and at some later time, n, it arrives at the receiver. Now consider a typical message sequence of length M. At time 0 the elements of M are generated at the source as independent random draws from the probability distribution $\vec{P}(0)$,[5] where $P_1(0)$ is the

---

[5]Actually, information theory deals readily with messages that have a much more complicated statistical structure; for our purposes, it is sufficiently general to have messages which are composed of independently, randomly drawn symbols.

probability of drawing symbol i to be the resident of an arbitrary element of the M sequence. As the message traverses the channel, its elements are subject to distortion. If at time *0* the resident of element m is symbol i, then the probability that it will be corrupted to symbol j by time period *1* is Γ_{ji} . Similarly, the probability that the original symbol i will be corrupted to symbol j by time period n, at the receiver, is just $(\Gamma^n)_{ji}$. The probability of any particular element being corrupted is independent of the probabilities of any other elements being corrupted. Recall that the probability distribution of symbols at the source is given by $\vec{P}(0)$. Due to corruption, the probability distribution of symbols at the receiver will not equal $\vec{P}(0)$; instead, it will be $\vec{P}(n)$, where $\vec{P}(n) = \Gamma^n \vec{P}(0)$.

As a relevant physical example of a channel, consider an ensemble of M spin 1/2 particles. Assume that the particles are separated from one another so that at time 0 we can prepare each spin so that it is in whatever state we desire. The particular configuration of the entire ensemble of spins constitutes the "message" at the source; each spin contains one symbol's worth of the message. The ensemble is now allowed to evolve with time in the presence of a heat bath (and also, perhaps, an external magnetic field). The configuration of the spins at time n constitutes the received message. Because of the heat bath, the received message only partially resembles the transmitted message. We are interested in the amount of information which survives the heat bath; this information, which Shannon termed the mutual information, connects the configuration of the spins at time n with their configuration at time 0. The mutual information depends on both entropy and conditional entropy for its definition. We now discuss these two kinds of entropy and how they are combined to form mutual information.

Shannon defines the information content of a message to be the log of the number of messages in the set from which the particular message is chosen. To estimate this number, he invites us to consider the set of "reasonably probable" messages. The set of reasonably probable messages is typically very large. If messages of M symbols are drawn from an alphabet of two symbols, then the set of reasonably probable messages could contain as many as 2^M elements. We might expect that the size of the set of reasonably probable messages should depend upon the exact definition of "reasonably probable." Let's investigate more deeply the relation between the size of the reasonably probable set and the definition of reasonably probable. Imagine an enumeration of all 2^M possible messages. Each message in this list has a particular probability of being generated at the source end of our channel. Suppose that the messages are listed in in descending order of probability from the single most probable message to the least probable one. Starting with the most probable message, we go down the list and keep track of the sum of the probabilities of the successive messages. Suppose that at the N^{th} message we have accumulated a total probability q. What is q? q is the probability that a message generated by the source will be an element of the set of the N(q) most probable messages.

In Figure 1, we plot the quantity $H(q) = (1/M)\log_2 N(q)$ versus q for various values of M for a two symbol alphabet with symbol probabilities *0.8* and *0.2*. Notice that as M becomes large and for q not too near 0 or 1 the graph of H(q) becomes increasingly flat and hovers near the value $\log_2(e)$ H[.8,.2].[6] We have discovered graphically that which Shannon first deduced: the log of the size of the set of reasonably probable messages,

---

[6] H[.8,.2] is the entropy function of equation (2) applied to a state vector containing the probabilities .8 and .2. The logarithm prefactor in this expression converts the ln's of H[,] to base 2.

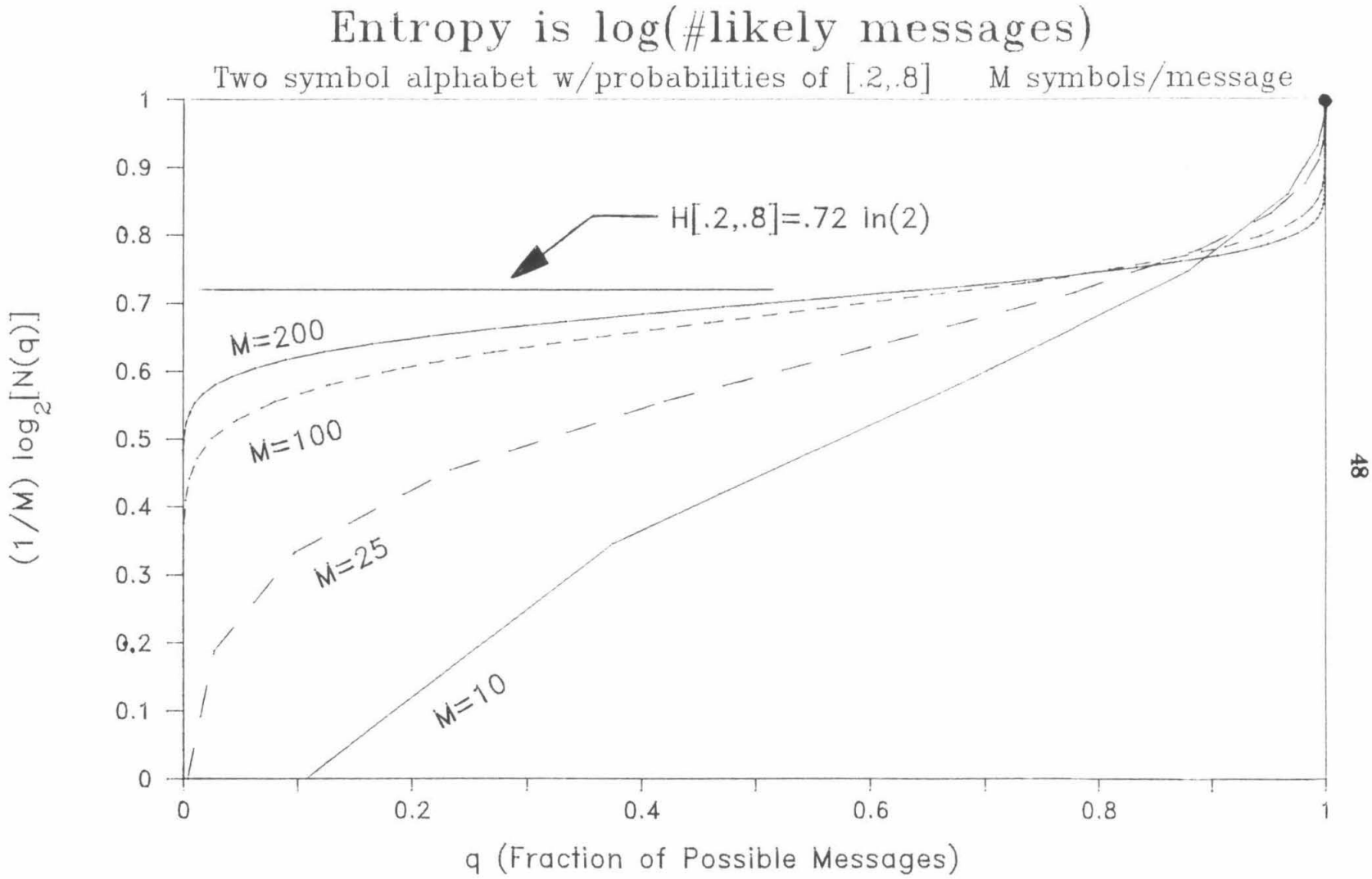

Figure 1. The meaning of entropy.

scaled by the number of symbols per message, yields a quantity which is independent of the definition of "reasonably probable" for sufficiently long messages. Moreover, this scaled log–of–set–size attains a value which is equal to the entropy of the probability distribution of the symbols. Thus, to discriminate a particular message from among all the other $2^{MH(q)}$ reasonably probable messages, one needs M $\log_2(e)$ $H[\hat{P}]$ bits of information, where $H[\hat{P}]$ is the entropy defined by equation (2) taken over the distribution of symbol usage frequencies $\hat{P}$. In the remainder of this paper we shall drop the $\log_2(e)$ factor, dealing with information in "nats" instead of bits; also, following Shannon, we shall normalize the message information by the number of symbols in the message. In summary, Shannon equated the entropy of the distribution of symbol frequencies in a message to the information content, in nats per symbol, of the message.

Shannon realized that while entropy is sufficient to characterize information content, it is not sufficient to characterize information transmission. Therefore he defined a new quantity, conditional entropy, which measures the information lost to channel noise during transmission of each element of a message. Channel noise corrupts a message element by randomly transmuting the symbol in that element to some other symbol. Suppose the original resident of a particular element is symbol j. Symbol j faces a probability $(\mathbf{\Gamma}^n)_{ij}$ of being turned into symbol i during an n step journey through the channel. The set of transition probabilities which govern the fate of symbol j actually comprise the j^{th} column of $\mathbf{\Gamma}^n$. It is useful to think of the matrix $\mathbf{\Gamma}^n$ as a collection of columns; each column j is a probability distribution describing the likelihoods of the different alternatives which symbol j may become if it is mangled in transmission. Information is lost when a message element is transmitted because of uncertainty about

what the identity of the element will be upon reception.[7] If the original resident of a message element is symbol j, then the information lost is exactly equal to the mathematical entropy of the probability distribution which is the j^{th} column of Γ^n.

Let $\vec{H}^T(n)$ be the vector whose components are these column entropies.[8] Thus,

$$\vec{H}^T(n) = (H[\text{column 1 of } \Gamma^n], H[\text{column 2 of } \Gamma^n]) , \quad (4)$$

where H[.] is the mathematical entropy function defined in equation (2). Each component of $\vec{H}^T(n)$ is the information lost during transmission of a particular symbol. Conditional entropy, denoted by $H(n|0)$,[9] summarizes the channel's overall information loss per message element as the weighted average of $\vec{H}^T(n)$'s components, where the weights reflect the relative usage probabilities of different symbols at the source, $\vec{P}(0)$:

$$\begin{aligned} H(n|0) &= H_a(n)P_a(0) + H_b(n)P_b(0) \\ &= \vec{H}^T(n)\vec{P}(0) \quad . \end{aligned} \quad (5)$$

We have seen that entropy measures the information content of a message and that conditional entropy measures the information lost to noise during transit of the message through the channel. Shannon showed that the difference of these two is a measure of the amount of information surviving transmission through the channel; in a sense, this is the amount of

---

[7] We emphasize uncertainty; a channel which always complements symbol a to symbol b and vice–versa transmits messages perfectly, if perniciously.

[8] It is interesting to note that the vector of conditional entropies summarizes most of what one needs to know about a channel in order to study information transmission and distortion.

[9] $H(n|0)$ is vocalized as "the conditional entropy of the distribution at time n, *given* the distribution at time 0."

information that is mutually shared by the source and the receiver. Thus the mutual information, $\mathcal{I}$, across a channel is defined as

$$\mathcal{I}[\vec{P}(n),\ \vec{P}(0)] = H[\vec{P}(n)] - \vec{H}^{T}(n)\vec{P}(0) \quad . \tag{6}$$

Equating the Free Energy and Mutual Information Operators.

In the previous two subsections we have introduced thermodynamics and information theory from the viewpoint of Markov processes. We have suggested that each subject can be organized around a defining measure whose behavior characterizes the system: free energy for thermodynamics, mutual information for information theory.

We now present a summary of our main result. We view the two quantities free energy and mutual information as operators which map state space probability distributions into real numbers. Then we demonstrate that in the appropriate limit of long times, these two operators are asymptotically equivalent.

Consider Equations (3) and (6):

$$\mathcal{F}[\vec{P}(n)] = \vec{E}^{T}\vec{P}(n) - kT\ H[\vec{P}(n)] \qquad \text{and} \tag{3}$$

$$\mathcal{I}[P(n),\ P(0)] = H[\vec{P}(n)] - \vec{H}^{T}(n)\vec{P}(0) \ . \tag{6}$$

Structurally these equations are quite similar. Both involve the function $H[\vec{P}(n)]$, the entropy of the system's state space probability vector. Both involve linear operators acting on state vectors: the thermodynamic operator $\vec{E}^{T}$, which we call the internal energy operator, and the information theoretical $\vec{H}^{T}(n)$, which we call the conditional entropy operator.[10] As yet, however, the

---

[10]For our purposes, a linear operator is just the transpose of a vector. Linear operators act on probability state vectors by a simple inner product; the result is simply the sum of the components of the state vector weighted by the components of the operator. Equivalently, and more intuitively, the result is the average of the components of the operator weighted by the probabilities in the state vector.

two operators are not alike in detail. We now eliminate two easily remedied algebraic differences; afterwards, we explore what remains.

The first algebraic difference is that the equations have different units, since the free energy $\mathcal{F}$ is expressed in Joules, while mutual information is expressed in nats, which are dimensionless.[11] The second difference is that the internal energy and conditional entropy operators are acting on state vectors $\hat{\mathrm{P}}$ corresponding to different time periods; conditional entropy requires the state vector from time *0*, while internal energy uses the state vector at time *n*. These differences may be eliminated as follows. To bring the respective units of equations (3) and (6) into agreement, we measure the free energy $\mathcal{F}$ in units of $-\mathrm{kT}$. Now the entropy terms $\mathrm{H}[\hat{\mathrm{P}}(n)]$ of the two equations agree exactly. Notice that the internal energy operator has become $\hat{\mathrm{E}}^{\mathrm{T}}/\mathrm{kT}$; we call this the "rescaled" internal energy operator. Next, using the identity $\hat{\mathrm{P}}(0) = \mathbf{\Gamma}^{-n}\hat{\mathrm{P}}(n)$ rewrite the conditional entropy operator as $(\hat{\mathrm{H}}^{\mathrm{T}}(n)\mathbf{\Gamma}^{-n})$ so that it acts on $\hat{\mathrm{P}}(n)$ instead of $\hat{\mathrm{P}}(0)$. Notice that $(\hat{\mathrm{H}}^{\mathrm{T}}(n)\mathbf{\Gamma}^{-n})$ is still an operator; we will call it the rereferenced conditional entropy operator. Notice that we could have accomplished this alignment of reference times by reexpressing internal energy as $(\hat{\mathrm{E}}^{\mathrm{T}}/\mathrm{kT})\mathbf{\Gamma}^{n}$, which would have established time *0* as the common point of reference, rather than time *n*. Both choices of reference time alignment are useful; we proceed here with the first method of realignment because this method leads to a more spectacular and stronger form of our main result.[12]

---

[11] One might compare nats to bits as one relates radians to degrees; in each case the former are without dimension.

[12] To some extent, the choice to be made here depends on one's purpose. Our major enthusiasm is to enhance thermodynamics by using ideas from information theory. Accordingly, we heed the common usage of physics, which generally establishes the present (time n) as the preferred point of reference.

Equations (3) and (6) have now become

$$-\mathcal{J}[\hat{P}(n)]/kT = H[\hat{P}(n)] - (\hat{E}^T/kT)\ \hat{P}(n) \quad \text{and} \tag{7}$$

$$\mathcal{J}[\hat{P}(n)] = H[\hat{P}(n)] - (\hat{H}^T(n)\Gamma^{-n})\ \hat{P}(n) \quad . \tag{8}$$

Notice how similar are the right hand sides of Equations (7) and (8) with respect to their dependence on $\hat{P}(n)$. In fact, as n becomes sufficiently large, these two equations become identical. We now demonstrate this asymptotic identity by exploring graphically the relationship over time between the two linear operators $\hat{E}^T/kT$ and $\hat{H}^T(n)\Gamma^{-n}$. Start by picking, at random, a $2x2$ Markov matrix Γ. Any Γ will do just as long as it has nonnegative elements and columns which sum to unity. Since we have picked Γ at random, we don't immediately know the vector of state energies, $\hat{E}^T/kT$. Thus, we find next the equilibrium state vector, $\hat{P}(eq)$, of Γ, in order that we may determine the state energies by Boltzmann's equation. Now, since from physics we know that only differences in energy effect system dynamics, we calculate the difference of the components of the rescaled internal energy operator. This difference, which is just $(E_b - E_a)/kT$, is equal to $ln(P_a(eq)/P_b(eq))$ using Boltzmann's equation. Finally, analogously, we calculate the difference of the components of the rereferenced conditional entropy operator for values of the time n, ranging from say 1 to 40.[13] This last step is tedious but simple: for each n determine Γ^n, $\hat{H}^T(n)$ and the difference of the two components of $\hat{H}^T(n)\Gamma^{-n}$.

Figure (2) illustrates graphically the relationship that emerges from these calculations. The straight line depicts the component difference of the rescaled internal energy operator. This line is horizontal because we have assumed that Γ is constant which implies that the state energies are

---

[13] How large n need be taken depends on how close Γ's non−unity eigenvalue is to 1. For most random Γ's, 40 will usually be adequate.

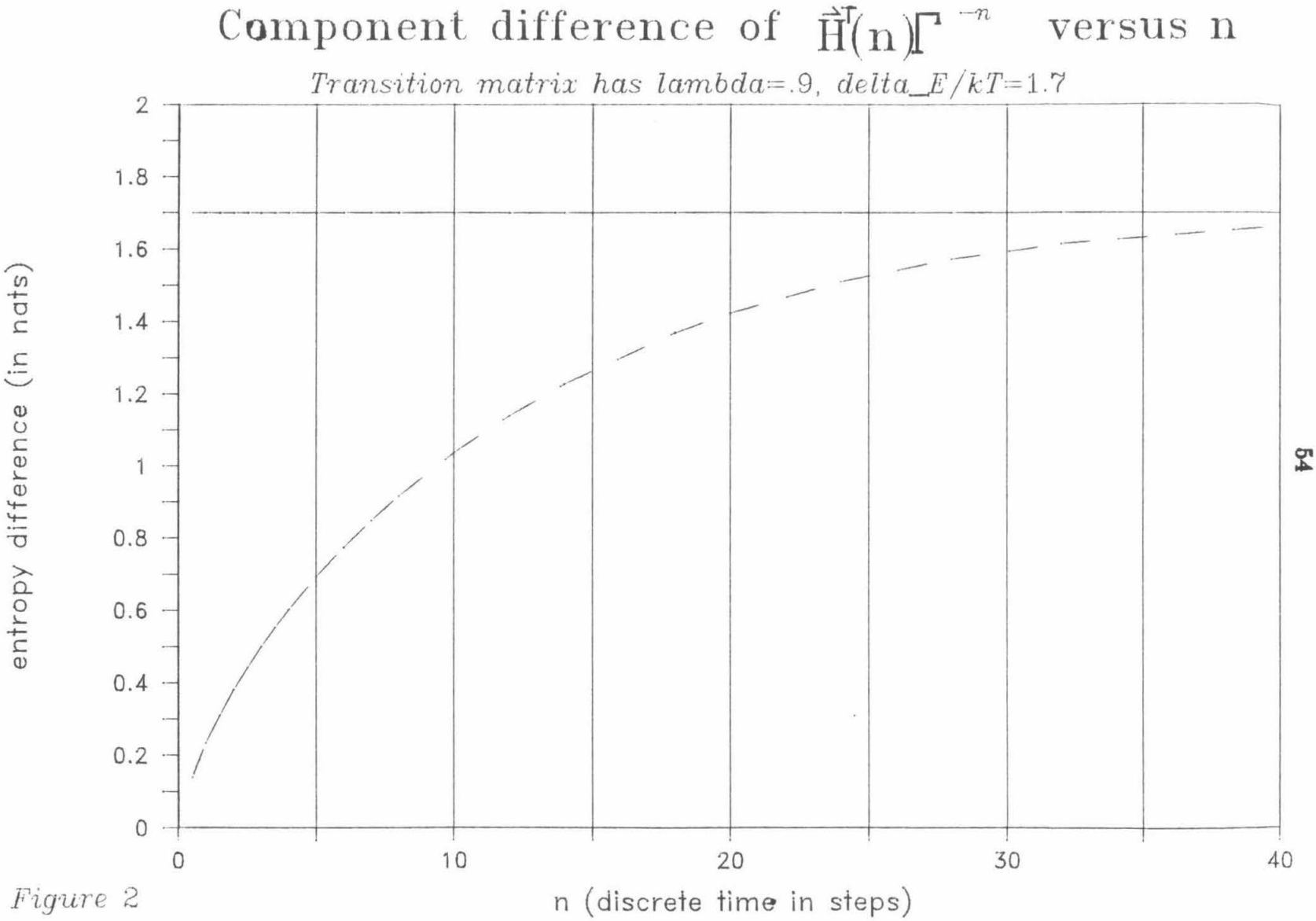
Component difference of $\vec{H}(n)\Gamma^{-n}$ versus n
Transition matrix has lambda=.9, delta_E/kT=1.7
entropy difference (in nats)
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
n (discrete time in steps)

Figure 2

unchanging. The points on the smoothly interpolated curve depict the component difference of the rereferenced conditional entropy operator at a variety of time values *n*. The main thing to notice is that asymptotically, as *n* goes to infinity, the curve merges with the line. This merging of curve with line is a general result; later on in this chapter we show that it occurs for all 2x2 Markov matrices Γ, and that something similar happens for suitably restricted systems of arbitrary size. Thus, modulo the question of absolute energies, the two linear operators become identical as *n* goes to infinity. What is the significance of this identity? Tracing back through the argument, we see that in Equations (7) and (8), these linear operators were the only terms that were not obviously identical. Their graphical asymptotic equality implies that the two defining measures, free energy and mutual information, become identical operators (to within an additive constant) as *n* goes to infinity.

Figure (2) exhibits our main result if one is content to leave energies relative to one another. It is interesting to ask what the appropriate definition of the zero of energy would be in order for the free energy and mutual information operators to become absolutely identical as *n* goes to infinity. The answer is simple and intuitive: offset the energies E_a and E_b by an amount which causes the free energy of the equilibrium state to be zero.[14] This offset allows us to meaningfully compare, component by component, the linear operators $\vec{E}^{\Gamma}/kT$ and $\vec{H}^{\Gamma}(n)\Gamma^{-n}$. Figure (3) graphically accomplishes this comparison. In Figure (3), we interpret the cartesian coordinates of a point as specifying the two components of an operator. Consider an arbitrary point $\vec{R}$ on the curve which is labeled "$F[\vec{P}(eq)]=0$;"

---

[14]This happens when the energy of each state i has become $-kT \ln(P_i(eq))$. Thus the partition function Z which is the sum of negative exponentials of E_i/kT is unity.

Comparing rereferenced conditional entropy to rescaled internal energy

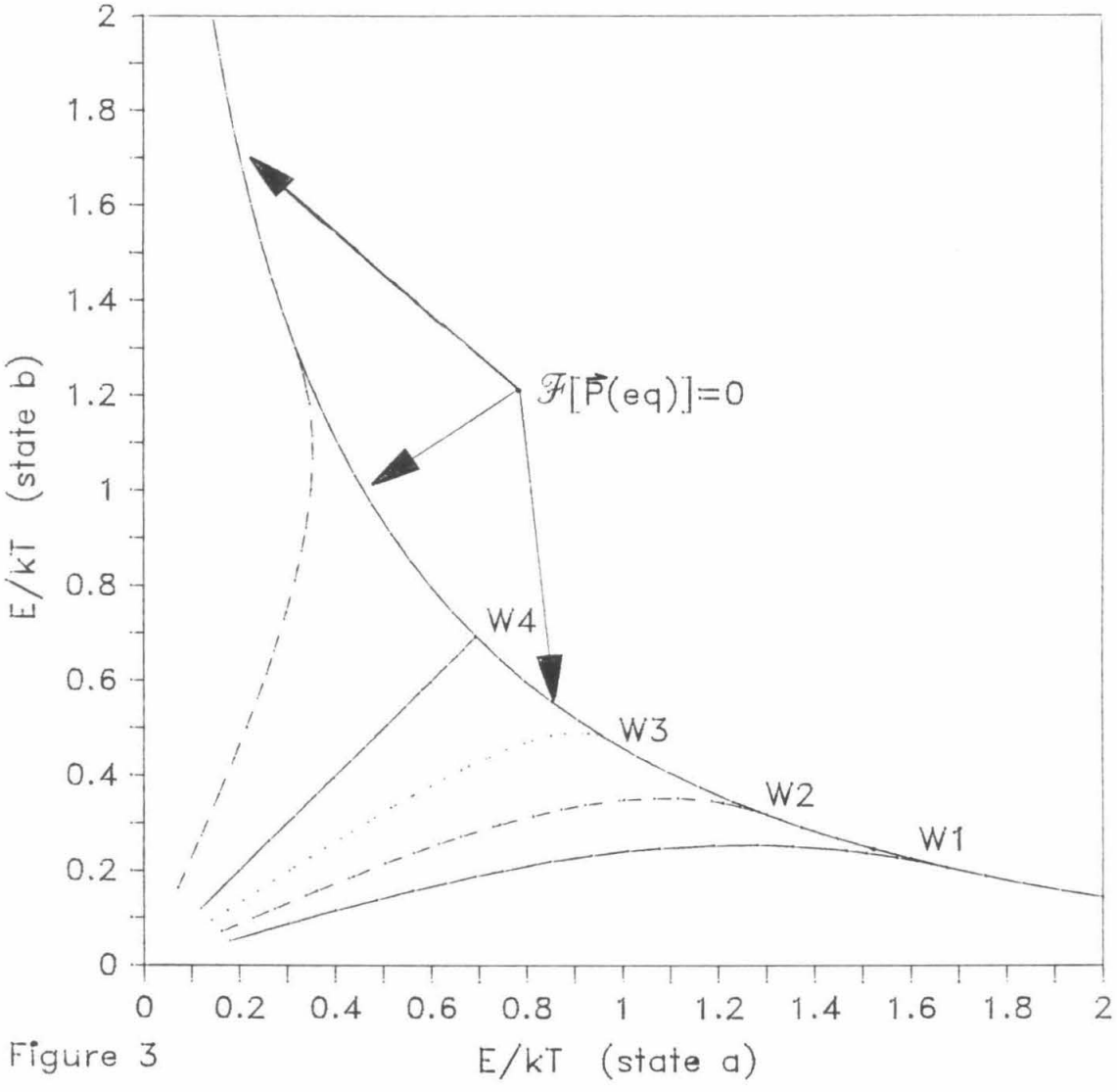

Figure 3

the coordinates of $\tilde{R}$ are just the rescaled state energies (E_a/kT, E_b/kT) of some two–state system which has zero free energy at equilibrium. Thus the rescaled internal energy operator $\tilde{E}^T/kT$ of any two–state system specifies some point on the curve labeled "$F[\tilde{P}(eq)]=0$" after we adjust both state energies by the same additive constant, so that the free energy at equilibrium $\mathcal{F}[\tilde{P}(eq)]$ is zero. The remaining curves in Figure (3), which are labeled "W1" through "W4," depict the rereferenced conditional entropy operator $\tilde{H}^T(n)\Gamma^{-n}$ as a function of time n for four representative matrices, one matrix per curve. Each curve results from a smooth interpolation of the set of points that is generated by evaluating $\tilde{H}^T(n)\Gamma^{-n}$ for different n.[15] The essential feature to notice in Figure (3) is that each curve has an accumulation point as n goes to infinity which always lies on the curve labeled $F[\tilde{P}(eq)]=0$. Thus, Figure (3) lets us see how the individual components of the rereferenced conditional entropy asymptotically approach the corresponding components of the rescaled internal energy as time n goes to infinity.

It seems reasonable to hope that even for small values of n, where the rereferenced conditional entropy does not closely match the rescaled internal energy, still there might be a physical interpretation of $\tilde{H}^T(n)\Gamma^{-n}$ which is based on energy. Suppose that for n time periods we observe the fluctuating state of a physical thermodynamic system such as the two–state system described previously. Then we try to infer the energies of the states of this

---

[15]Conceptually, the curve is generated as described in the text; actually it is generated by smoothly sweeping the eigenvalue of the matrix Γ from nearly 1 down to 0. Thus the curve shows a continuous time analog of the results for descrete systems. Negative eigenvalues have no continuous time analog. However, had the eigenvalue been swept from 1 down to below 0, then each curve would have continued smoothly past the accumulation point, never crossing the $F[\underline{P}(eq)]=0$ curve, until it exited the first quadrant.

system as best we can, using our observations.[16] Ultimately, we will be able to deduce the equilibrium probabilities of occupation and hence the energies of all the states, although this may require times that are long, even when compared to the time required by the mathematical Markov model to reach equilibrium. Initially, we can tell very little, since the statistical imprint left by the different state energies on our data will not yet have sufficient definition to be visible above the noise. In between, we can presumably deduce an intermediate amount about the state energy levels.

Figures (2) and (3) suggest that we can view $\vec{H}^T(n)\Gamma^{-n}$ as a vector of "revealed state energies."[17] Qualitatively $\vec{H}^T(n)\Gamma^{-n}$ behaves much as our inferred energies should; initially its components are the same, suggesting that for short observation times, noise fluctuations conceal any differences between the energies of the states. Later its components approach $\vec{E}^T/kT$, which is consonant with the fact that to the patient observer and data analyst ultimately all is revealed. More deeply, we might have expected a relation between conditional entropy and state energy on physical grounds. Intuitively, we expect high–energy states to be less accessible than low–energy states, since high–energy states are generally less populated than low–energy states. It is also intuitively reasonable to estimate state accessability by counting the number of reasonably probable ways of entering

---

[16]It is crucial here to distinguish the physical system from its mathematical model which is the Markov process. Though it is true that *on average* the system evolves according to a Markov process, it is not true that we ever observe the state vector of this mathematical process. The state vector of the mathematical process records the outcome of a hypothetical experiment involving a large ensemble of systems. We observe only a single system; moreover our observations do not take the form of probabilities, rather they are a record of the sequence of states occupied by our system during the interval that we observe it.

[17]By "revealed state energies" we mean the energies of the states as revealed to an astute observer who calculates them by applying sophisticated statistical estimators to his data.

or leaving a state. Thus it is plausible for conditional entropy, which performs counts such as these, to be related to state energy.

Markov processes in the spectral representation

In addition to being physically interesting, the behavior of $\vec{H}^{\Gamma}(n)\Gamma^{-n}$ as a function of n is curious mathematically. Correctly evaluating $\vec{H}^{\Gamma}(n)\Gamma^{-n}$ in the limit as n goes to infinity is tricky; this is a singular limit. Even for the 2x2 case, slogging through with only naive algebra is remarkably difficult. There is a better way: first, decompose the matrix Γ into its spectral representation so that it is expressed as a sum of orthogonal projectors; with Γ in this form, we are able to evaluate the limit easily and elegantly. An added benefit of using the spectral representation is that it affords a direct insight into the operation of the limit. With the spectral representation, we shall see how new time scales are generated in systems with more than two states, and the manner in which these time scales can cause the limit to fail to exist. Now we turn to a discussion of Markov processes and the spectral representation of stochastic matrices. We work through the 2 by 2 case in detail, and then state the results for the N by N case. Afterwards, we use this representation to evaluate $\vec{H}^{\Gamma}(n)\Gamma^{-n}$ in the limit as n goes to infinity.

Modeling spin 1/2 as a two–state Markov process.

Consider the system of spin 1/2 particles. The spin state of any given particle fluctuates with time on account of thermal agitation. We can model the spin as a two state Markov process. Suppose that when the spin is oriented so that it is in state b, it is at a higher energy than when it is in state a. Let p_{up} be the probability that a particle with spin state a transits to state b in some fixed interval of time, and let p_{dn} be the

probability that a particle in spin state b transits to state a in the same interval of time. Correspondingly, $1-p_{up}$ denotes the probability of a particle remaining in state a, and $1-p_{dn}$ the probability of a particle remaining in state b.

We can now write

$$P_a(n+1)= (1-p_{up})P_a(n) + p_{dn}P_b(n) \qquad (1)$$

and $P_b(n+1)= p_{up}P_a(n) + (1-p_{dn})P_b(n)$

as the total probabilities of a particle residing in states a or b at time period $(n+1)$, conditional on the probabilities of time period n. These equations are intuitive; each accounts for all the possible ways that a particle may find itself in a given state.

We may collect equations (1) into matrix form:

$$\vec{P}(n+1)= \begin{pmatrix} 1-p_{up} & p_{dn} \\ p_{up} & 1-p_{dn} \end{pmatrix} \vec{P}(n) = \Gamma\vec{P}(n) \quad . \qquad (2)$$

The matrix Γ possesses a number of remarkable properties. The most interesting, from our point of view, is that each column of Γ sums independently to unity. Matrices with solely non−negative elements and unity column sums are known as stochastic matrices. Physically, a typical column i of Γ accounts for all the possible ways of either departing state i, or remaining in state i. Hence the stochasticity of Γ embodies mathematically, conservation of probability. In Equation (2) write $\vec{P}(n)$ as $\Gamma\vec{P}(n-1)$ and iterate to obtain

$$\vec{P}(n) = \Gamma^n\vec{P}(0). \qquad (3)$$

As currently written, Equation (3) is not analytically convenient. The usual alternative to (3) expresses Γ (and hence Γ^n) in diagonal (or perhaps only Jordan normal) form by transforming the coordinate basis of the state vector. Here, unfortunately, such a transformation will not suffice. We are

interested in calculating the conditional entropy. This calculation combines matrix elements in a novel way that can not be expressed in terms of elementary functions of a matrix. On account of this novelty, the calculation of the conditional entropy does not commute with the operation of diagonalization.

The spectral representation of a 2 by 2 stochastic matrix.

The spectral representation offers an alternative method of simplifying (3). This representation is especially effective for calculating high powers of Γ and for evaluating the conditional entropy in this long time limit. Felicitously, in the case of two–state processes (hence 2x2 transition matrices), this representation can always be expressed in terms of only two projectors which we obtain through the following steps: first, we determine Γ's equilibrium state vector, $\vec{P}(eq)$, and from $\vec{P}(eq)$ deduce one of Γ's projectors. Second, we construct a second projector represented as the difference between the identity operator and the first projector. This second projector turns out to be orthogonal to the first; therefore, the two in combination span Γ's two–dimensional range. Factoring each projector into outer–product form allows us to deduce Γ's eigenvalues, and, in combination with the projectors, Γ itself in the spectral representation.

System equilibrium is defined as a state vector $\vec{P}(eq)$ with components $P_a(eq)$, $P_b(eq)$ which satisfies the condition $\Gamma\vec{P}(eq) = \vec{P}(eq)$. For all physically relevant Γ's, $\vec{P}(eq)$ exists and is unique.[18] Equilibrium is important because, given any initial vector $\vec{P}(0)$, $\Gamma^n\vec{P}(0)$ will eventually converge to $\vec{P}(eq)$ as n increases towards infinity (i.e. after a sufficiently long time). In other words,

---

[18]This will be true provided that Γ models a physical process with a unique ground state. More formally, in the standard terminology of Markov processes, we are assuming that Γ is irreducible.

for any *initial* $\hat{P}(0)$, defining $\Gamma^{inf} = lim$ (as n goes to infinity) of Γ^{n}, we can write

$$\hat{P}(eq) = \Gamma^{inf}\hat{P}(0) \quad . \tag{4}$$

What are $\hat{P}(eq)$ and Γ^{inf}? Consider an arbitrary probability state vector with components $1-\theta$, θ. Since Γ^{inf} maps all such vectors into $\hat{P}(eq)$, it must be that

$$\begin{pmatrix} P_a(eq) \\ P_b(eq) \end{pmatrix} = \begin{pmatrix} w & x \\ y & z \end{pmatrix} \begin{pmatrix} 1 & -\theta \\ \theta \end{pmatrix} \qquad \text{for all } \theta, \text{ with } 0<\theta<1$$

where w, x, y, and z represent the four elements of Γ^{inf}. Hence

$$\begin{pmatrix} P_a(eq) \\ P_b(eq) \end{pmatrix} = \begin{pmatrix} (x-w)\theta+w \\ (z-y)\theta+y \end{pmatrix} \qquad \text{for all } \theta, \text{ with } 0<\theta<1.$$

Clearly $x=w$ and $z=y$, and $w=P_a(eq)$ and $y=P_b(eq)$. We conclude that

$$\Gamma^{inf} = \begin{pmatrix} P_a(eq) & P_a(eq) \\ P_b(eq) & P_b(eq) \end{pmatrix} .$$

Notice that Γ^{inf} factors, allowing expression as an outer product:

$$\Gamma^{inf} = \begin{pmatrix} P_a(eq) \\ P_b(eq) \end{pmatrix} \begin{pmatrix} 1 & 1 \end{pmatrix} = \hat{P}(eq)\hat{\Sigma}^{T}, \qquad \hat{\Sigma} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{5}$$

A projector is a matrix whose square is itself; intuitively, a projector matrix projects vectors into a subspace, but leaves vectors already in the subspace unaltered. Using equation (5), the outer product representation, we can simply show that Γ^{inf} is a projector:

$$\begin{aligned} \Gamma^{inf}\,\Gamma^{inf} &= (\,\hat{P}(eq)\hat{\Sigma}^{T}\,)\;(\,\hat{P}(eq)\hat{\Sigma}^{T}\,) \\ &= \hat{P}(eq)\;(\,\hat{\Sigma}^{T}\hat{P}(eq)\,)\,\hat{\Sigma}^{T} \\ &= \hat{P}(eq)\;(1)\;\hat{\Sigma}^{T} = \Gamma^{inf} \quad . \end{aligned} \tag{6}$$

Equation (6) uses two nifty concepts: 1.) the reassociation made possible by the outer product representation, and 2.) the fact that $\vec{\Sigma}^T$ sums elements of a vector, and, by the definition of probability measure, sums probability vectors to unity.

It will be useful to know the components of $\vec{P}(eq)$. In fact, $\vec{P}(eq)$ turns out to be an eigenvector of Γ, with eigenvalue *1*. Why is this true? $\vec{P}(n)$ represents the system state at time n, and Γ applied to $\vec{P}(n)$ evolves the system one time period, to $\vec{P}(n+1)$. Since $\vec{P}(eq)$ is the point of system equilibrium, Γ applied to $\vec{P}(eq)$ must leave $\vec{P}(eq)$ unchanged: $\Gamma\vec{P}(eq) = \vec{P}(eq)$. We can determine the components of $\vec{P}(eq)$ by writing out the equation $\Gamma\vec{P}(eq) = \vec{P}(eq)$ in component form.

$$\begin{pmatrix} 1-p_{up} & p_{dn} \\ p_{up} & 1-p_{dn} \end{pmatrix} \begin{pmatrix} P_a(eq) \\ P_b(eq) \end{pmatrix} = \begin{pmatrix} P_a(eq) \\ P_b(eq) \end{pmatrix} \tag{7}$$

As is the way of these things, the two equations implied by (7) are linearly dependent; it is easy to verify that both equations are satisfied when $P_a(eq)p_{up} = P_b(eq)p_{dn}$.[19] Combining this equation together with one specifying unity total probability: $P_a(eq) + P_b(eq) = 1$ we obtain

$$P_a(eq) = \frac{p_{dn}}{p_{up}+p_{dn}} \quad \text{and} \quad P_b(eq) = \frac{p_{up}}{p_{up}+p_{dn}} \; . \tag{8}$$

---

[19]Notice that this relation determines equilibrium for the two-state system by directly equating the flow from a to b with the flow from b to a. Thus the eigen-relation for the unity eigenvalue of Γ determines the ratio $P_a(eq)/P_b(eq)$ as we asserted in the beginning of this chapter in the section on thermodynamics.

We have now identified one of Γ's projectors, Γ^{inf}, and have written Γ^{inf} as the outer product of $\vec{P}(eq)$ with $\vec{\Sigma}^T$. In addition, we have observed that $\vec{P}(eq)$ is the eigenvector of Γ with unity eigenvalue. It happens that $I-\Gamma^{inf}$ is also a projector: $(I-\Gamma^{inf})^2 = I-2\Gamma^{inf}+(\Gamma^{inf})^2 = I-\Gamma^{inf}$. $I-\Gamma^{inf}$ is orthogonal to Γ^{inf}; it projects to a subspace which lies in the null space of Γ^{inf}: $\Gamma^{inf}(I-\Gamma^{inf}) = \Gamma^{inf}-(\Gamma^{inf})^2 = 0$. In fact, $I-\Gamma^{inf}$ has an outer product representation:

$$I-\Gamma^{inf} = \begin{pmatrix} 1-P_a(eq) & -P_a(eq) \\ -P_b(eq) & 1-P_b(eq) \end{pmatrix} = \begin{pmatrix} P_b(eq) & -P_a(eq) \\ -P_b(eq) & P_a(eq) \end{pmatrix}$$

$$= \begin{pmatrix} 1 \\ -1 \end{pmatrix} \begin{pmatrix} P_b(eq) & -P_a(eq) \end{pmatrix} . \qquad (9)$$

Compare the outer product form of $I-\Gamma^{inf}$ in (9) with the outer product form of Γ^{inf} itself, (5). Γ^{inf} had an eigenvector of Γ as the left component of its outer product factorization. Analogously, we might hope to have found another eigenvector of Γ as the left component in the outer product factorization of our new projector $I-\Gamma^{inf}$. Is the vector with components *(1, −1)* an eigenvector? Yes,

$$\Gamma \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1-p_{up} & p_{dn} \\ p_{up} & 1-p_{dn} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1-p_{up}-p_{dn} \\ p_{up}+p_{dn}-1 \end{pmatrix} \qquad (10)$$

$$= (1-p_{up}-p_{dn}) \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \lambda \begin{pmatrix} 1 \\ -1 \end{pmatrix} \qquad \text{where } \lambda = 1-p_{up}-p_{dn}.$$

Equation (10) buys us a lot, since it directly provides us with Γ's other eigenvalue, λ; evidently $I-\Gamma^{inf}$ projects to a subspace associated with λ.

We see that, along with Γ^{inf}, $I-\Gamma^{inf}$ holds a special place in the scheme of things. Together, these two can do all that Γ does. In fact $\Gamma = \Gamma^{inf}+\lambda(I-\Gamma^{inf})$, as we can readily show:

$$\Gamma^{inf} + \lambda(I-\Gamma^{inf}) = \lambda I + (1-\lambda)\Gamma^{inf}$$

$$= (1-p_{up}-p_{dn}) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + (p_{up}+p_{dn}) \begin{pmatrix} P_a(eq) & P_a(eq) \\ P_b(eq) & P_b(eq) \end{pmatrix}$$

$$= \begin{pmatrix} 1-(p_{up}+p_{dn}) & 0 \\ 0 & 1-(p_{up}+p_{dn}) \end{pmatrix} + \begin{pmatrix} p_{dn} & p_{dn} \\ p_{up} & p_{up} \end{pmatrix} \qquad (11)$$

$$= \begin{pmatrix} 1-p_{up} & p_{dn} \\ p_{up} & 1-p_{dn} \end{pmatrix}$$

The spectral representation of general stochastic matrices.

The above discussion, which, for simplicity, has been confined to $2x2$ matrices, illustrates certain properties which are more generally true. Explicitly, every M by M matrix has a spectral representation. In general, the projectors associated with distinct eigenvalues are orthogonal, and the sum of these eigenprojectors is the identity matrix, **I**. If the matrix is diagonalizable, then it can be expressed as the weighted sum of its eigen-projectors; in this sum the weights are the eigenvalues associated with each projector.[20] If the matrix is not diagonalizable, then the situation is somewhat more complicated; the decomposition of the matrix involves eigen-nilpotents as well as eigenprojectors. A matrix **N** is nilpotent if $\mathbf{N}^m=\mathbf{0}$ for some positive number m. The nilpotents can appear only in connection with repeated eigenvalues. See [Kato] for a thorough exposition of the general case. For a general M state Markov process with transition matrix Γ, define Γ^{inf} to be the projector associated with the unique unity eigenvalue, and

[20] A matrix will be diagonalizable if all of its eigenvalues have an algebraic multiplicity of 1; this prospect is overwelmingly likely to be true in any physical situation.

$\vec{P}(eq)$ to be the eigenvector associated with this eigenvalue.[21] Set $\vec{\Sigma}^T = (1, \ldots, 1)_M$. Let λ_i be other distinct eigenvalues, $|\lambda_i| < 1$. Each λ_i is associated with an eigenprojector Γ_i, and possibly also with an eigennilpotent N_i. Then

$$\begin{aligned} \Gamma = \Gamma^{inf} &+ \lambda_1\Gamma_1 + \lambda_2\Gamma_2 + \ldots + \Gamma_p\lambda_p \qquad p<M \\ &+ N_1 + N_2 + \ldots + N_p \end{aligned} \tag{12}$$

where $\Gamma^{inf} = \vec{P}(eq)\vec{\Sigma}^T$,

$\Gamma_i\Gamma_j=\delta_{ij}\Gamma_i$, $\Gamma_iN_j=N_j\Gamma_i=\delta_{ij}N_j$ & $N_iN_j=0$ if i,j unequal.

Expanding $\vec{H}^T(n)\Gamma^{-n}$ Around Equilibrium The 2x2 Case

We are now ready to calculate the rereferenced conditional entropy $\vec{H}^T(n)\Gamma^{-n}$ in the limit as n is taken to infinity. In this section, we show this calculation for the case of a two–state system. The calculation proceeds in three steps. First, we shall find Γ^{-n}, an operation that is made very easy by the spectral representation. Second, we shall calculate the conditional entropy operator $\vec{H}^T(n)$. $\vec{H}^T(n)$ is messy for arbitrary n, however, for large n it may be expanded in a Taylor series about its equilibrium value. The spectral representation allows us to readily calculate the first couple of terms of this Taylor series. In the third and final step, we shall use the spectral representation yet again to recast our expansion for $\vec{H}^T(n)$ to a simple and beautiful form from which calculation of the limit $\vec{H}^T(n)\Gamma^{-n}$ becomes especially easy.

---

[21] We have been assuming that the eigenvalue 1 is simple. [Gantmacher] proves that all the eigenvalues with unity modulus of a stochastic matrix are at worst semisimple and thus have no nilpotents associated with them.

Calculating powers of Γ.

Let us begin by calculating powers of Γ. In terms of projectors, Γ to the first power is $\Gamma^{inf} + \lambda(I-\Gamma^{inf})$. Using the orthogonality of these projectors, see that Γ^2 is $\Gamma^{inf} + \lambda^2(I-\Gamma^{inf})$. It is apparent that only the eigenvalue λ has been affected by the operation of squaring Γ. Continuing, we calculate Γ^n inductively as the product of Γ^{n-1} and Γ; with each multiply the orthogonal projectors produce no cross terms, so the only net effect is that the power of λ has been increased by one. Thus, Γ^n may be determined from $\Gamma^n=\Gamma^{inf} + \lambda^n(I-\Gamma^{inf})$. What about Γ^{-n}? The form of Γ^n for positive powers strongly suggests that

$$\Gamma^{-n} = \Gamma^{inf} + \lambda^{-n}(I-\Gamma^{inf}) \quad . \tag{1}$$

Check this by multiplying the right hand side by the projector form of Γ^n; as always the projectors produce no cross terms, and the net effect is that the power to which λ appears becomes zero. Since both sides of Equation (1) become the identity matrix I when they are multiplied by Γ^n, it must be that Equation (1) correctly expresses inverse powers of Γ.

Notice that Γ^{-n} diverges as n goes to infinity. This behavior is to be expected; since $|\lambda|<1$, we know that $|\lambda^{-1}|>1$, and hence that $|\lambda^{-n}|$ grows exponentially with n. Thus the eigenvalue of Γ^{-n} grows exponentially with n.[22] From a state–space perspective, the cause of this divergence is intuitive. Γ^{-n} diverges because always it must be able to invert the mapping which is Γ^n. As n gets large, the mapping which is Γ^{-n} must be able to magnify a tiny volume of state space which is centered on the equilibrium point $\hat{P}(eq)$ so that the mapped image of this volume fills the

---

[22]These remarks concerning the algebraic divergence of Γ^{-n}, and the remarks that follow considering the situation in state space apply equally well to general systems with arbitrary numbers of states. In the general case all of Γ's nonunity eigenvalues contribute to the divergence of Γ^{-n} and the eigenvalue of Γ with the smallest modulus dominates the divergence.

entire state space.[23] What is the effect of Γ^{-n} when it multiplies a linear operator such as the conditional entropy operator $\vec{H}^T(n)$? Corresponding to the equilibrium point $\vec{P}(eq)$ in state space there is the point $\vec{\Sigma}^T$ in the adjoint space.[24] Just as (for large n) Γ^{-n} maps a tiny volume around $\vec{P}(eq)$ onto the entire state space, so also does it map a tiny volume around $\vec{\Sigma}^T$ onto the entire adjoint space. The fact that $\vec{\Sigma}^T$ is the fixed point of Γ^{-n} is the central mechanism which allows the large n limit of $\vec{H}^T(n)\Gamma^{-n}$ to exist. As n goes to infinity $\vec{H}^T(n)$ becomes proportional to $\vec{\Sigma}^T$.

Finding $\vec{H}^T(n)$ for large n

Now we investigate the large n behavior of $\vec{H}^T(n)$. We advance this investigation by obtaining the conditional entropy operator $\vec{H}^T(n)$ in a form that is analytically tractable when n is large. Recall that the components of $\vec{H}^T(n)$ are the column entropies of Γ^n:

$$\vec{H}^T(n) = (H[\text{column 1 of } \Gamma^n], H[\text{column 2 of } \Gamma^n]) \quad .$$

What does a typical column j of Γ^n look like when n is large? Since $\Gamma^n = \Gamma^{inf} + \lambda^n(I-\Gamma^{inf})$ and the columns of Γ^{inf} are all alike and are equal to $\vec{P}(eq)$, we can see that column j of Γ^n takes the form: $\vec{P}(eq) + \delta\vec{C}$ where $\delta=\lambda^n$ and $\delta\vec{C}$ is a vector whose components are the elements of the j^{th} column of the projector $I-\Gamma^{inf}$. By hypothesis, n is large, thus δ is small, and it is reasonable to expand the entropy of the j^{th} column of Γ^n,

---

[23] Since any multiple of $\underline{P}(eq)$ is also a fixed point of Γ^{-n}, we might more properly say that Γ^{-n} magnifies a cylindrical volume of state space which is infinitely long but infinitesimally slender so that the mapped image of this cylinder fills the entire state space. The cylinder is concentric about a line which contains both the origin of the state space and the point $\underline{P}(eq)$.

[24] Recall that linear operators are transposes of vectors and so have a correspondence to points in a space which is known formally as the adjoint of the state space.

$H[\vec{P}(eq) + \delta\vec{C}]$, in a power series in δ around $\delta=0$. Performing this expansion,[25] we obtain

$$H[\vec{P}(eq) + \delta\vec{C}] = -\sum_{i=1}^{2}(P_i(eq)+\delta C_i)ln(P_i(eq)+\delta C_i)$$

$$= H[\vec{P}(eq)] - \sum_{i=1}^{2}(1+\ln P_i(eq))\delta C_i + O(||\delta\vec{C}||^2)$$

$$= H[\vec{P}(eq)] - (\vec{\Sigma}^T-\vec{E}^T/kT)\delta\vec{C} + O(||\delta\vec{C}||^2)$$

where the last equality follows on account of the definition $\vec{\Sigma}^T=(1, 1)$, and on account of the Boltzmann relation $E_i/kT= -lnP_i(eq)$, which assumes $\vec{E}^T$ is constrained so the free energy of equilibrium is zero. Recall that $\Gamma^{inf}= \vec{P}(eq)\vec{\Sigma}^T$; the orthogonality of the projectors in a projection decomposition assures us that the (inner) product of $\vec{\Sigma}^T$ with a column from any projector other than Γ^{inf} must always be zero; thus $\vec{\Sigma}^T\vec{C}=0$.[26] Note that the $O(||\delta\vec{C}||^2)$ term contains sums of squares of components of $\delta\vec{C}$; that this is an unsimplifiable mess becomes evident when one considers that in general (for systems with more than 2 states), $\delta\vec{C}$ is itself a sum of vectors. Fortunately, further expansion of this term will not be necessary to establish the limit in which we are interested.

The limit of $\vec{H}^T(n)\Gamma^{-n}$.

We have now assembled all the ingredients which are essential in the Taylor series expansion of the conditional entropy operator $\vec{H}^T(n)$. We know

---

[25] Remember $d(xlnx)/dx$ is $1+lnx$, and so $(x+\delta)ln(x+\delta)$ is $xlnx + (1+lnx)\delta + O(\delta^2)$. Also note that the summation index i enumerates the different elements of the fixed column vectors $\underline{P}(eq)$ and $\delta\underline{C}$.

[26] Actually the constraint on $\underline{E}$ that there be zero free energy in equilibrium isn't necessary, since all physically equivalent $\underline{E}$'s are the same up to an additive multiple of $\underline{\Sigma}$, which in any event makes no difference here.

the components of $\vec{H}^T(n)$ are column entropies of $\mathbf{\Gamma}^n$. We have shown that for large n, all these column entropies take the form $H[\hat{P}(eq)]+(\vec{E}^T/kT)\lambda^n\vec{C}$. Thus,

$$\vec{H}^T(n)= H[\hat{P}(eq)]\vec{\Sigma}^T + (\vec{E}^T/kT)\lambda^n(\mathbf{I}-\mathbf{\Gamma}^{inf}) + \vec{\mathcal{O}}^T(|\lambda|^{2n})$$

where $\vec{\mathcal{O}}^T(|\lambda|^{2n})$ stands for a vector transpose all of whose components are $O(|\lambda|^{2n})$. $\vec{\mathcal{O}}^T(|\lambda|^{2n})$ is the error we incur when we truncate the Taylor series after two terms. This equation for $\vec{H}^T(n)$ possesses an intriguing factorization if we recast the leading order term, $H[\hat{P}(eq)]\vec{\Sigma}^T$, in an alternate form. To find this alternative, notice that $H[\hat{P}(eq)]$, the entropy of equilibrium, is just the same as the inner product $(\vec{E}^T/kT)\hat{P}(eq)$, if we agree as before to reference energies from a zero free energy at equilibrium so that $E_i/kT= -ln P_i(eq)$. Then, $H[\hat{P}(eq)]\vec{\Sigma}^T$ becomes $(\vec{E}^T/kT)\hat{P}(eq)\vec{\Sigma}^T$, which we can reduce immediately to $(\vec{E}^T/kT)\mathbf{\Gamma}^{inf}$. Substituting $(\vec{E}^T/kT)\mathbf{\Gamma}^{inf}$ for $H[\hat{P}(eq)]\vec{\Sigma}^T$ allows us to express $\vec{H}^T(n)$ very elegantly:

$$\begin{aligned}\vec{H}^T(n) &= (\vec{E}^T/kT)\ (\mathbf{\Gamma}^{inf} + \lambda^n(\mathbf{I}-\mathbf{\Gamma}^{inf})) + \vec{\mathcal{O}}^T(|\lambda|^{2n}) \\ &= (\vec{E}^T/kT)\mathbf{\Gamma}^n + \vec{\mathcal{O}}^T(|\lambda|^{2n}) \quad . \end{aligned} \tag{3}$$

In spite of its compactness, note that equation (3) does give the asymptotic form of $\vec{H}^T(n)$ accurate to second order, since the error in (3) diminishes as the square of $|\lambda|^n$. This quadratic dependence of the error on $|\lambda|^n$ is generally what we would expect from two terms of a Taylor series; the surprising thing here is the way the two terms cooperate to produce an expression which has the matrix $\mathbf{\Gamma}^n$ as a factor. If we imagine that $\mathbf{\Gamma}^n$ propagates transposes of vectors forward through time, just as it does for state vectors, then equation (3) has a curious interpretation; it suggests that the conditional entropy operator at time n is the result of propagating the transpose vector of state energies forward through n time steps, at least asymptotically.

Let us now show that $\vec{H}^T(n)\Gamma^{-n}$ approaches $\vec{E}^T/kT$ in the limit as n goes to infinity and thereby complete the chain of reasoning which asymptotically links mutual information to free energy. Multiplying both sides of equation (3) by Γ^{-n}, we can see that the difference $\vec{H}^T(n)\Gamma^{-n} - \vec{E}^T/kT$ approaches $\vec{\mathcal{O}}^T(|\lambda|^{2n})\Gamma^{-n}$. Now $\vec{\mathcal{O}}^T(|\lambda|^{2n})$ has no special direction relative to the projectors of Γ, so multiplication by Γ^{-n} magnifies it by a factor of order $|\lambda|^{-n}$.[27] The result of this multiplication is some vector $\vec{\mathcal{O}}^T(|\lambda|^n)$ which still manages to go to zero as n goes to infinity. Thus, we have established that $\vec{H}^T(n)\Gamma^{-n}$ does indeed approach $\vec{E}^T/kT$ asymptotically as n goes to infinity, at least for the two–state case.

The rereferencing theorem

The existence of the limit of $\vec{H}^T(n)\Gamma^{-n}$ (for a general system) turns out to be independent of many of the properties of the conditional entropy operator $\vec{H}^T(n)$. The existence of this limit depends only on the spectrum of the stochastic matrix Γ, and on the fact that $\vec{H}^T(n)$ is a function of the columns of Γ. In this section, we state and prove a theorem about the rereferencing of linear operators which are generated from the columns of a stochastic matrix. In order that this section should be as self contained as possible, we explain the notation with several definitions prior to stating and proving the theorem.

Definition: $\vec{\Sigma}^T = (1, 1, \ldots, 1)_N$.

---

[27] Most relevant here is the direction of $\vec{\mathcal{O}}^T$ with respect to $\vec{\Sigma}^T$. Split $\vec{\mathcal{O}}^T$ into a sum of two vectors, one parallel to $\vec{\Sigma}^T$ and the other perpendicular to $\vec{\Sigma}^T$. The parallel piece remains fixed under the mapping Γ^{-n}, while the piece perpendicular to $\vec{\Sigma}^T$ gets magnified by $|\lambda|^{-n}$.

Definition: *The class of matrices having a unique equilibrium state.*

Γ is an NxN stochastic matrix if

a) $\Gamma >= 0$ and $\vec{\Sigma}^T \Gamma = \vec{\Sigma}^T$.

Γ is a regular NxN stochastic matrix if

a) holds and

b) $|\lambda|=1$ implies $\lambda=1$, λ an eigenvalue of Γ.

Γ is a fully regular NxN stochastic matrix if

a) holds and

b) holds and

c) $\lambda=1$ is an eigenvalue with algebraic multiplicity 1.

Definition: *The equilibrium vector of* Γ.

$\vec{P}(eq)$ is the equilibrium vector of a fully regular stochastic matrix Γ if

$\Gamma\vec{P}(eq) = \vec{P}(eq)$ and $\vec{\Sigma}^T \vec{P}(eq) = 1$.

Definition: *Functions of probability vectors.*

f[ ] is a function on a probability vector[28] if

$f[\vec{P}] = f(P_1, P_2, \ldots, P_N)$, where f is a scalar valued function of the N numbers $P_1, P_2, \ldots, P_N$ which are the components of the probability vector $\vec{P}$.

$\vec{D}^T f[\vec{P}]$ is the linear operator which is the derivative of the function f[ ] evaluated at $\vec{P}$ if

$(\vec{D}^T f[\vec{P}])_i = \partial f(P_1, P_2, \ldots, P_N)/\partial P_i$.

---

[28] $\vec{P}$ is a probability vector if $\vec{P} >= 0$ and $\vec{\Sigma}^T \vec{P} = 1$.

Definition: *Building an operator by applying a function to the columns of* Γ^n.

$\vec{f}(n)$ is the f–operator associated with Γ if

$(\vec{f}(n))_i = f[$ column i of $\Gamma^n]$, where $f[\ ]$ is some given function.

Theorem: *Rereferencing linear operators.*

Given a fully regular NxN stochastic matrix Γ with the equilibrium vector $\vec{P}(eq)$, and a function $f[\vec{P}]$ which is continuously differentiable in the neighborhood of $\vec{P} = \vec{P}(eq)$, then

$$|\lambda_1|^2 < |\lambda_p| \quad \text{implies}$$

$$lim_{n \to infinity}\ \vec{f}(n)\Gamma^{-n} = \{f_{eq} - \vec{Df}_{eq}\ \vec{P}(eq)\}\vec{\Sigma}^T + \vec{Df}_{eq}\ ,$$

where $\vec{f}(n)$ is the f–operator associated with Γ, $f_{eq} = f[\vec{P}(eq)]$, $\vec{Df}_{eq} = \vec{Df}[\vec{P}(eq)]$, and λ_i $(i=1, \ldots, p)$ are the distinct non–unity eigenvalues of Γ, arranged so that $|\lambda_1| >= |\lambda_2| >= \ldots >= |\lambda_p|$ $(p <= N-1)$.

Proof:

The proof follows directly from an expansion of $\vec{f}(n)$ for large n. To perform this expansion, we need Γ^n in an accessible form. If the spectral representation of Γ is

$$\Gamma = \Gamma^{inf} + \sum_{i=1}^{p} \lambda_i \Gamma_i + N_i\ ,$$

then the spectral representation of Γ^n is

$$\Gamma^n = \Gamma^{inf} + \sum_{i=1}^{p} \lambda_i^n (\Gamma_i + n^{m[i]-1} N_i')\ .$$

In the first expression, each eigenvalue λ_i is associated with eigenprojector Γ_i, and a possible eigennilpotent N_i. In the second expression, each eigenvalue λ_i^n is associated with eigenprojector Γ_i, and a possible eigennilpotent

$\mathbf{N}_i'$.[29] *In* both of these expressions, $\mathbf{\Gamma}^{inf}$ is the projector associated with the unity eigenvalue. By hypothesis, the unity eigenvalue is not degenerate. Therefore, $\mathbf{\Gamma}^{inf}$ has rank 1. Also, we know that $\mathbf{\Gamma}^{inf}\vec{P}(eq) = \vec{P}(eq)$. Thus, every column of $\mathbf{\Gamma}^{inf}$ is identical to $\vec{P}(eq)$, and we conclude that $\mathbf{\Gamma}^{inf} = \vec{P}(eq)\vec{\Sigma}^T$.

We can now proceed with the expansion of $\vec{f}^T(n)$ for large n. The first term in this expansion is the row vector which is obtained by applying f[ ] to the columns of $\mathbf{\Gamma}^{inf}$. Since all the columns of $\mathbf{\Gamma}^{inf}$ are identical to $\vec{P}(eq)$, we see that the first term in the expansion is $f[\vec{P}(eq)]\vec{\Sigma}^T$. The second term in the expansion of $\vec{f}^T(n)$ is obtained by applying the derivative of $f[\vec{P}(eq)]$ to the columns of the difference $\mathbf{\Gamma}^n - \mathbf{\Gamma}^{inf}$. Thus, the second term in the expansion is $\vec{Df}^T[\vec{P}(eq)](\mathbf{\Gamma}^n - \mathbf{\Gamma}^{inf})$. If we truncate the expansion of $\vec{f}^T(n)$ after these first two terms, then we incur an error. We assume the magnitude of this error can be estimated from the first neglected term in the expansion. The third term in the expansion of $\vec{f}^T(n)$ is quadratic in the elements of the columns of $\mathbf{\Gamma}^n - \mathbf{\Gamma}^{inf}$. $||\mathbf{\Gamma}^n - \mathbf{\Gamma}^{inf}||$ is $O(n^{m[1]-1}|\lambda_1|^n)$, where m[1] does not exceed the algebraic multiplicity of the eigenvalue λ_1. Thus, the error incurred by truncating the expansion after the first two terms is $O(n^{2m[1]-2}|\lambda_1|^{2n})$.

$$\vec{f}^T(n) = f[\vec{P}(eq)]\vec{\Sigma}^T + \vec{Df}^T[\vec{P}(eq)](\mathbf{\Gamma}^n - \mathbf{\Gamma}^{inf}) + \vec{\mathcal{O}}^T(n^{2m[1]-2}|\lambda_1|^{2n}),$$

where $\vec{\mathcal{O}}^T(a)$ is a row vector with norm O(a). We assume the direction of $\vec{\mathcal{O}}^T(n^{2m[1]-2}|\lambda_1|^{2n})$ is arbitrary.[30] $\vec{f}^T(n)$ can be expressed more compactly if we employ the definitions used in the statement of our theorem. Replace $f[\vec{P}(eq)]$ with f_{eq} and $\vec{Df}^T[\vec{P}(eq)]$ with $\vec{Df}^T_{eq}$.

$$\vec{f}^T(n) = f_{eq}\vec{\Sigma}^T + \vec{Df}^T_{eq}(\mathbf{\Gamma}^n - \mathbf{\Gamma}^{inf}) + \vec{\mathcal{O}}^T(n^{2m[1]-2}|\lambda_1|^{2n}).$$

---

[29] We have pulled out the factor $n^{m[i]-1}\lambda_i^n$ so that the norm of $\mathbf{N}_i'$ remains $O(1)$ as n is taken to infinity.

[30] This is a worst case assumption because it implies that $||\mathbf{\Gamma}^{-n}\vec{\mathcal{O}}^T||$ is as large as it can possibly be. It is, however, as strong an assumption as the hypothesis of the theorem allows. On account of this assumption, the theorem establishes a sufficient (and not a necessary) condition for the limit of the rereferenced linear operator to exist.

We now complete the proof of our theorem.

$$\vec{f}^{T}(n)\Gamma^{-n} = f_{eq}\vec{\Sigma}^{T} + \vec{Df}^{T}_{eq}(I-\Gamma^{inf}) + \vec{0}^{T}(n^{2m[1]-2}|\lambda_1|^{2n})\Gamma^{-n}.$$

From the spectral representation[31] of Γ^{-n},

$$\Gamma^{-n} = \Gamma^{inf} + \sum_{i=1}^{p} \lambda_i^{-n}(\Gamma_i + n^{m[i]-1}N_i^n),$$

we see $||\Gamma^{-n}|| = O(n^{m[p]-1}|\lambda_p|^{-n})$, where m[p] does not exceed the algebraic multiplicity of the eigenvalue λ_p. Thus the error term in $\vec{f}^{T}(n)\Gamma^{-n}$ is $\vec{0}^{T}(n^{2m[1]+m[p]-3}|\lambda_1^2/\lambda_p|^n)$. Clearly, when $|\lambda_1^2/\lambda_p| < 1$, this error term is o(1). Therefore, if $|\lambda_1|^2 < |\lambda_p|$, then

$$lim_{n-->infinity}\ \vec{f}^{T}(n)\Gamma^{-n} = f_{eq}\vec{\Sigma}^{T} + \vec{Df}^{T}_{eq}(I-\Gamma^{inf}).$$

The proof of our theorem is completed if we substitute $\vec{P}(eq)\vec{\Sigma}^{T}$ for Γ^{inf} in this expression.

$\vec{H}^{T}(n)\Gamma^{-n}$ in the general case

Does $\vec{H}^{T}(n)\Gamma^{-n}$ approach $\vec{E}^{T}/kT$ in systems having more than two states? Since $\vec{H}^{T}(n)$ is a function of the columns of Γ^n, this question can be answered by an application of our rereferencing theorem. The theorem tells us immediately that the limit of $\vec{H}^{T}(n)\Gamma^{-n}$ will exist if $|\lambda_1|^2 < |\lambda_p|$, where λ_1 is the nonunity eigenvalue of Γ with the largest modulus, and $|\lambda_p|$ is the eigenvalue of Γ with the smallest modulus. When the limit exists, what is its value? This question requires us to connect the quantities which appear in the theorem with quantities which appear in the present circumstances. Evidently, $\vec{f}^{T}(n) = \vec{H}^{T}(n)$, f[ ] = H[ ] and $f_{eq} = H[\vec{P}(eq)]$. $\vec{Df}^{T}_{eq}$ is a row vector whose elements, in this case, are the partial derivatives of $H[\vec{P}]$ with respect to the components of $\vec{P}$, evaluated at $\vec{P} = \vec{P}(eq)$. The i^{th} component of $\vec{Df}^{T}_{eq}$

---

[31] In this expression Γ_i is the eigenprojector and N_i^n is the eigennilpotent associated with the eigenvalue λ_i^{-n}. $||N_i^n|| = O(1)$ (n −−>infinity).

is $\partial H[\vec{P}(eq)]/\partial P_i(eq) = -(1+\ln P_i(eq)) = -(\vec{\Sigma}^T - \vec{E}^T/kT)_i$, and so $\vec{Df}^T_{eq} = -(\vec{\Sigma}^T - \vec{E}^T/kT)$. The rereferencing theorem tells us that, in the limit as n goes to infinity, $\vec{f}^T(n)\Gamma^{-n} = \{f_{eq} - \vec{Df}^T_{eq}\,\vec{P}(eq)\}\vec{\Sigma}^T + \vec{Df}^T_{eq}$. Notice that $\vec{Df}^T_{eq}\,\vec{P}(eq) = -(\vec{\Sigma}^T - \vec{E}^T/kT)\vec{P}(eq) = -1 + H[\vec{P}(eq)]$. Thus, $\{f_{eq} - \vec{Df}^T_{eq}\,\vec{P}(eq)\}\vec{\Sigma}^T + \vec{Df}^T_{eq} = \{1+0\}\vec{\Sigma}^T - (\vec{\Sigma}^T - \vec{E}^T/kT) = \vec{E}^T/kT$. We conclude that $|\lambda_1|^2 < |\lambda_p|$ implies $\vec{H}^T(n)\Gamma^{-n} = \vec{E}^T/kT$, in the limit as n goes to infinity. Actually, the proof of the theorem tells us somewhat more than this. We know that

$$\vec{H}^T(n)\Gamma^{-n} = \vec{E}^T/kT + \vec{\mathcal{O}}^T(c(n)|\lambda_1^2/\lambda_p|^n), \text{ where}$$

$$c(n) = n^{2m[1]+m[p]-3}.$$

<u>$|\lambda_1|^2 < |\lambda_p|$ compresses the time scales of a system.</u>

Physically, the limitation $|\lambda_1^2| < |\lambda_p|$ is rather severe; it specifically includes only those systems having decay time scales which span less than a factor of two. This interpretation follows directly from a reasonable definition of decay time scale. Notice that $|\lambda_i|^n = exp(n\ ln|\lambda_i|)$; compare this with the standard form of exponential decay: $exp(-n/\tau_i)$, where τ_i is the one over e time of decay of mode i. We see that $\tau_i = -1/ln|\lambda_i|$. Taking logs of both sides of the inequality $|\lambda_1^2| < |\lambda_p|$ and then negating we obtain $-2ln|\lambda_1| > ln|\lambda_p|$, whence comes our result:

$$\tau_1 < 2\,\tau_p \quad .$$

When interpreting this inequality, remember that the τ_j are in decreasing order. The decay time of the equilibrium mode which is τ_0, is infinite.[32] τ_1 is the next largest decay time; it is follwed by τ_2, and so on down to τ_p, which is the smallest of the decay times.

The notion of decay time is especially useful when our theorem does apply. Recall that $\vec{H}^T(n)\Gamma^{-n} = \vec{E}^T/kT + \vec{\mathcal{O}}^T(c(n)|\lambda_1^2/\lambda_p|^n)$. The residual term

---

[32] Formally this follows from $\tau_0 = -1/ln|\lambda_0|$, if we take a limit where λ_0 approaches 1 from below.

$\vec{\mathcal{O}}^T(c(n)|\lambda_1{}^2/\lambda_p|^n)$ decays as $|\lambda_1{}^2/\lambda_p|^n$. This corresponds to a decay time τ_H, where:[33] $\tau_H = \tau_1\tau_p\ /(2\tau_p-\tau_1)$. τ_H is a nonlinear combination of the smallest and largest (but finite) decay times of the system. It is easy to show that $\tau_H>\tau_p$;[34] this means that τ_p is still the shortest time scale of the system. However, τ_H is unbounded above, so that as τ_1 approaches $2\tau_p$, τ_H becomes infinite. Thus the rereferenced conditional entropy can take a longer time to approach its equilibrium value than any of the modes of the system take to decay. In fact it is precisely as τ_H becomes infinite that the theorem breaks down and $\vec{H}^T(n)\Gamma^{-n}$ fails to approach $\vec{E}^T/kT$.

A three–state example.

We close this section with a graphical presentation showing the behavior of the rereferenced conditional entropy as a function of time for two similar (but distinct) three–state systems. The first two graphs (Figures 4 and 5) deal with a three–state system having a transition matrix which satisfies the conditions of our theorem. This transition matrix has the eigenvalues 1, 0.9, 0.85; since $.9^2=.81$ which is less than .85, we expect the rereferenced conditional entropy of this system to be well–behaved. Figure 4 shows that in this case each component of $\vec{H}^T(n)\Gamma^{-n}$ does indeed approach the corresponding component of $\vec{E}^T/kT$ as n becomes large. Figure 5 plots the components of the difference $|\vec{H}^T(n)\Gamma^{-n}-\vec{E}^T/kT|$ versus n on a log scale for the same system. The theory predicts that this difference should agree very nearly with the residual term $\vec{\mathcal{O}}^T(|\lambda_1{}^2/\lambda_p|^n)$. All of the components of $\vec{\mathcal{O}}^T(|\lambda_1{}^2/\lambda_p|^n)$ decay to zero as $(.81/.85)^n$; this corresponds to a 1/e decay time

---

[33] $\tau_H = (-2\ln|\lambda_1|+\ln|\lambda_p|)^{-1} = (2/\tau_1-1/\tau_p)^{-1} = \tau_1\tau_p/(2\tau_p-\tau_1)$

[34] Consider the expression $c=ab/(2b-a)$ where $b<a<2b$. Suppose that $a=(1+\theta)b$ with $0<\theta<1$. Then $c=b(1+\theta)/(1-\theta)$. The graph of $(1+\theta)/(1-\theta)$ versus θ increases smoothly from the value 1 when θ is zero to infinity as θ approaches 1.

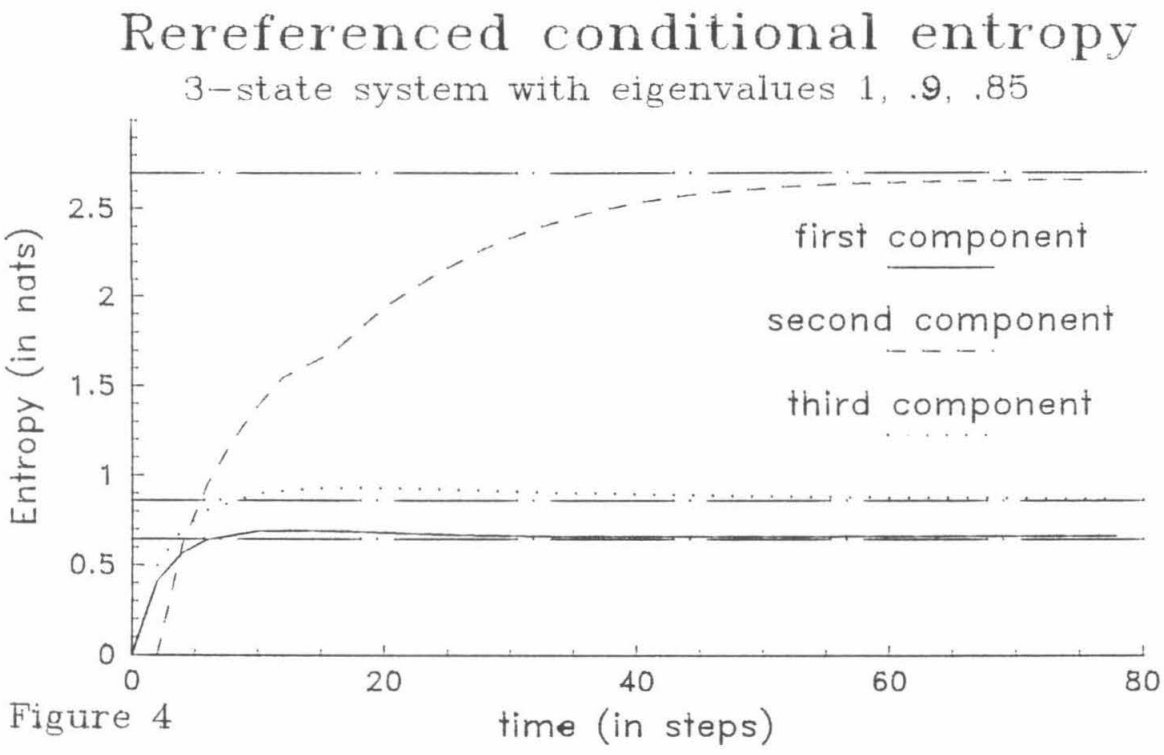

Figure 4

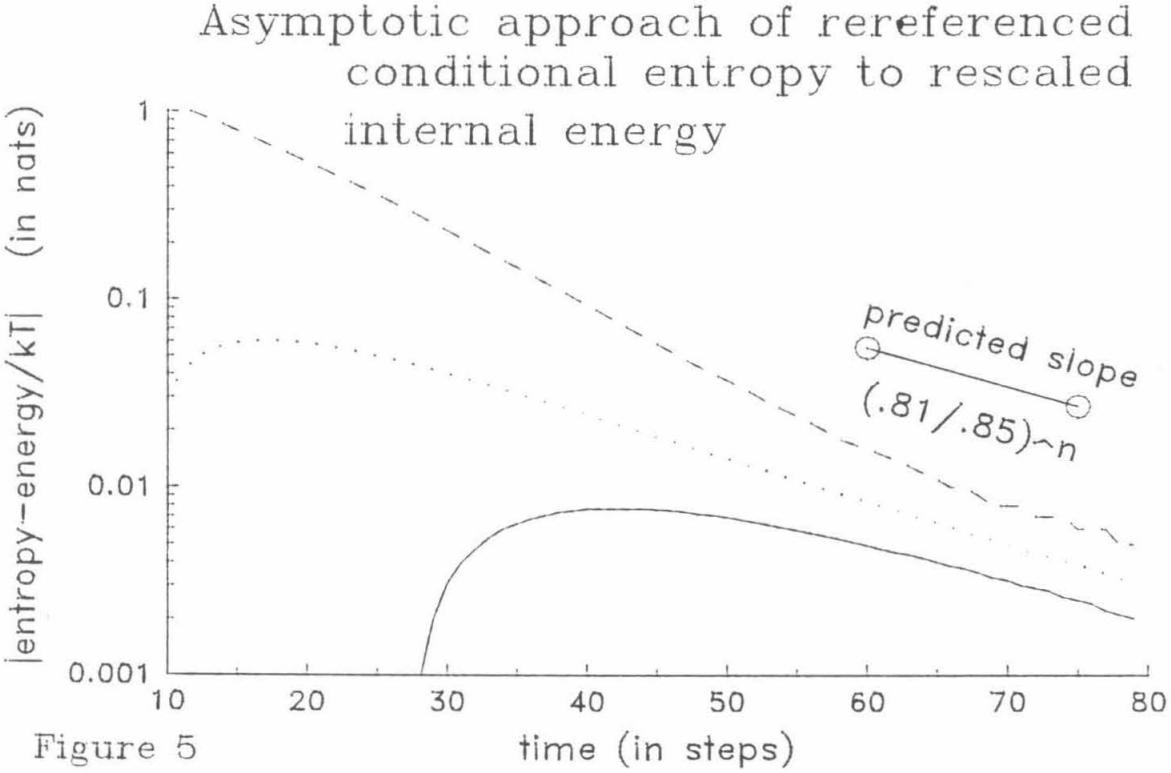

Figure 5

τ_H *of 20.7* time steps which is what we observe in figure 5. Note that the system of Figures 4 and 5 has eigenvalues corresponding to decay times τ_1, τ_2 of *9.5* and *6.2* respectively. τ_H is more than twice as big as the largest of these times; this lends support to our observation that the rereferenced conditional entropy operates on a time scale which is different from, and conceivably much larger than, the natural time scales of the system.

Figures 6 and 7 display the rereferenced conditional entropy on linear and log scales versus time for another three state system. The transition matrix of the system portrayed in Figures 6 and 7 is different in only one respect from the transition matrix used in Figures 4 and 5; the eigenvalue *0.85* has been changed to the value *0.77*. In every other respect, the transtion matrices of Figures 4 and 5 and of Figures 6 and 7 are identical; they have exactly the same eigenprojectors, and two of these projectors are weighted by the common eigenvalues *1* and *0.9*. The change of one eigenvalue from *0.85* to *0.77* is crucial to the rereferenced conditional entropy operator because it leaves unsatisfied the existence condition which this operator requires in order to be well behaved as n goes to infinity. $\lambda_1^2=.81$ is no longer less than $\lambda_p=.77$; Figure 6 shows the consequences of this inequality failure. Notice that for small n the curves of Figure 6 behave in a somewhat similar manner to those of Figure 4, then the divergence hits and they move off toward infinity. Can we account for the curves rate of divergence? Figure 7, which is a semilog plot of the components of $|\vec{H}^T(n)\Gamma^{-n}-\vec{E}^T/kT|$, shows us that all the components grow exponentially, increasing by a factor of e every *20* or so time steps. This agrees with the calculated e folding time $\tau_H=19.7$ which one gets by assuming that the (diverging) "residual term" $\vec{\mathcal{O}}^T(|\lambda_1^2/\lambda_p|^n)$ still dominates the large n behavior of $|\vec{H}^T(n)\Gamma^{-n}-\vec{E}^T/kT|$.

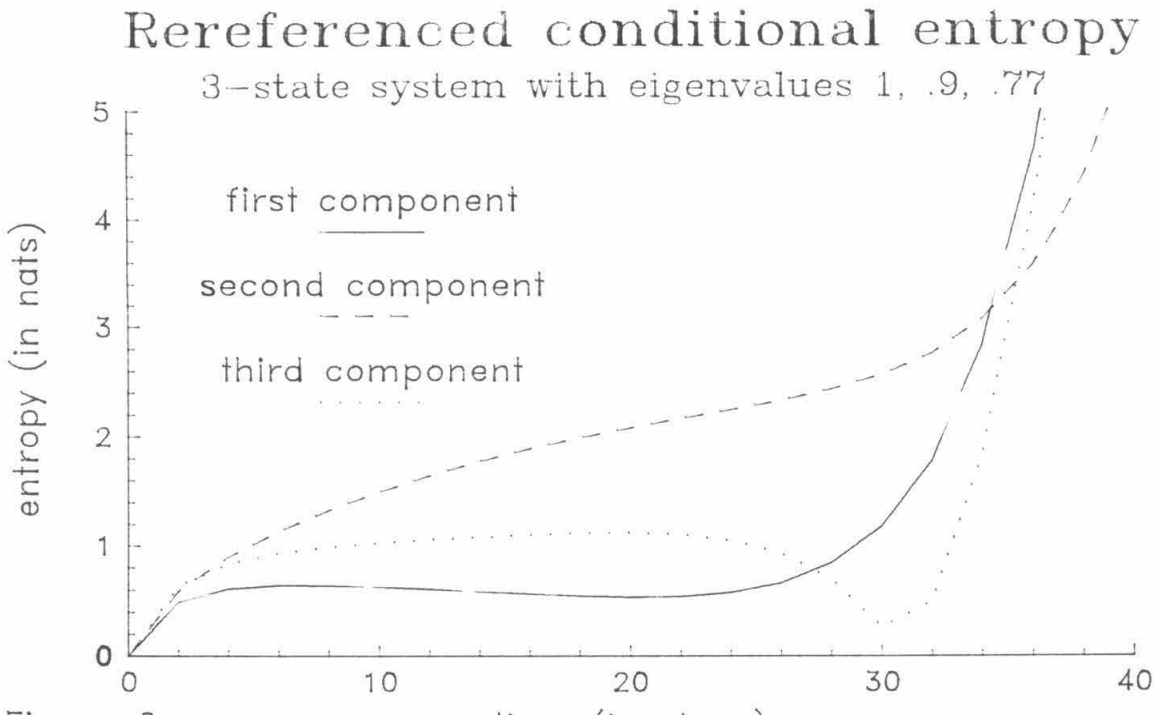

Figure 6

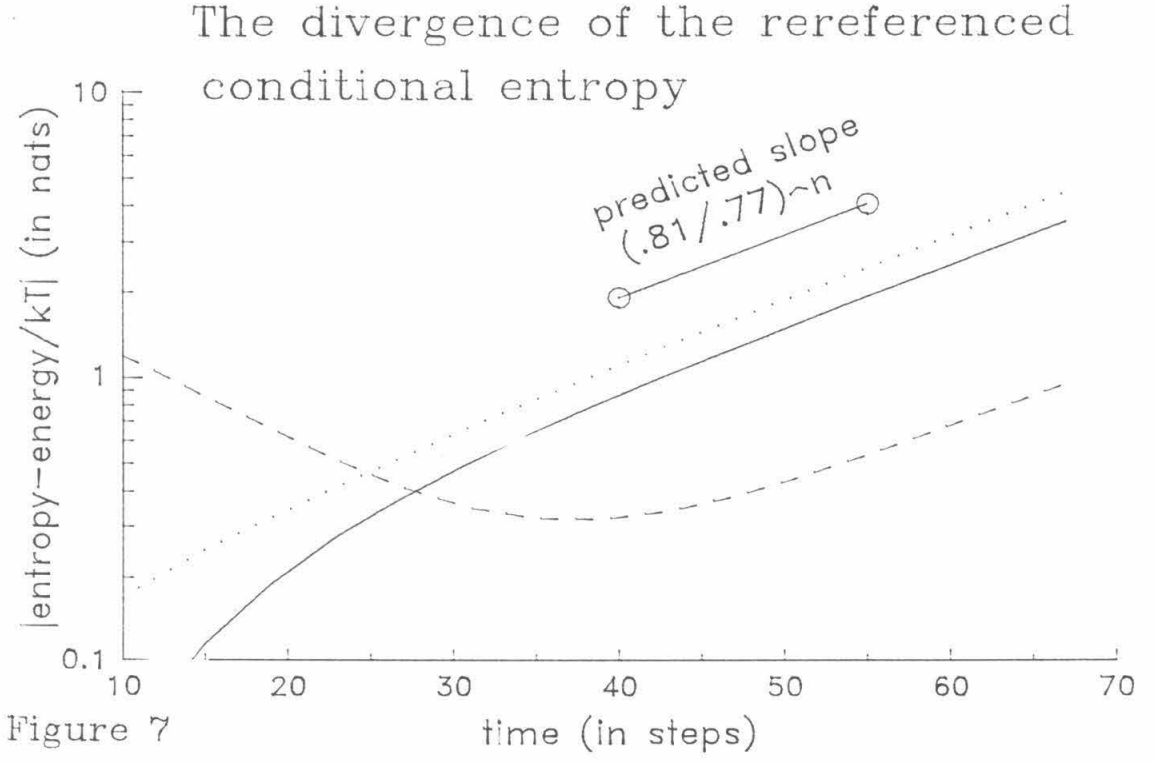

Figure 7

A composition law for $\tilde{H}^{\Gamma}(n)\Gamma^{-n}$

We have shown that, *for a general system,* $\tilde{H}^{\Gamma}(n)\Gamma^{-n}$ possesses an equilibrium limit only when the time constants of the system span less than an octave. This restriction will almost certainly be violated if we construct a system out of several independent subsystems. Consider, for example, a four–state system which consists of two independent spin 1/2 particles. The time constants of this system *never* span less than a factor of two. Systems which are built of independent pieces have a special kind of structure. For such systems, we can relax somewhat the conditions of our theorem.[35] In this section, we prove that $\tilde{H}^{\Gamma}(n)\Gamma^{-n}$ always possesses an equilibrium limit for a system consisting of two independent spin 1/2 particles. The proof suggests that $\tilde{H}^{\Gamma}(n)\Gamma^{-n}$ possesses an equilibrium limit for a system composed of several independent components if and only if this limit exists for each of the components taken separately.[36]

Consider a system which is composed of two, independent, spin 1/2 particles. Let the particles be labeled r and s respectively. Suppose that r is governed by a 2x2 stochastic matrix of transition probabilities $\Gamma(r)$. Similarly, s is governed by the 2x2 stochastic matrix $\Gamma(s)$. If we consider r and s jointly, then we have a single system with four states. We label these four states as follows:

---

[35] The rereferencing theorem establishes a condition which is sufficient to ensure that the limit of a rereferenced operator exists. The established condition is not a necessary condition, i.e., the converse of the theorem does not hold.

[36] *(late note added at final printing 5/29/86):* It appears certain that the argument of this section can be directly extended to cover the composition of two independent systems of arbitrary size. Simple induction on this enhanced argument yields the general composition law which we have stated.

state of particle r	state of particle s	state of composite system
1	1	1
1	2	2
2	1	3
2	2	4 .

Operators which factor the four–state system.

The transition matrix of the four–state system is Γ_{4x4}. Γ_{4x4} has a special structure; it is the product of a pair of factors, both of which have a noteworthy form. The forms of these factors motivate us to introduce two pairs of operators. These operators convert back and forth between the two, two–state systems and the single four–state system which is their composition.[37]

$$\Gamma_{4x4}= \begin{vmatrix} \Gamma_{11}(r)\Gamma_{11}(s) & \Gamma_{11}(r)\Gamma_{12}(s) & \Gamma_{12}(r)\Gamma_{11}(s) & \Gamma_{12}(r)\Gamma_{12}(s) \\ \Gamma_{11}(r)\Gamma_{21}(s) & 11 \quad 22 & 12 \quad 21 & 12 \quad 22 \\ \Gamma_{21}(r)\Gamma_{11}(s) & 21 \quad 12 & 22 \quad 11 & 22 \quad 12 \\ \Gamma_{21}(r)\Gamma_{21}(s) & 21 \quad 22 & 22 \quad 21 & 22 \quad 22 \end{vmatrix}$$

$$\Gamma_{4x4}= \begin{vmatrix} \Gamma_{11}(r)\ \Gamma(s) & \Gamma_{12}(r)\Gamma(s) \\ \Gamma_{21}(r)\ \Gamma(s) & \Gamma_{22}(r)\Gamma(s) \end{vmatrix}$$

$$\Gamma_{4x4}= \begin{vmatrix} \Gamma_{11}(r)\ \mathbf{I} & \Gamma_{12}(r)\mathbf{I} \\ \Gamma_{21}(r)\ \mathbf{I} & \Gamma_{22}(r)\ \mathbf{I} \end{vmatrix} \begin{vmatrix} \Gamma(s) & 0 \\ 0 & \Gamma(s) \end{vmatrix}$$

Each element of Γ_{4x4} is the product of an "r type" transition probability with an "s type" transition probability. This product comes about because each element of Γ_{4x4} specifies the probability of a pair of independent events. Notice that each factor of Γ_{4x4} is a 4x4 matrix which is built in a simple

---

[37] More precisely, these are imbedding operators. The four–state system is the tensor product of the two two–state systems. This section is really just a simple introduction to the algebra of tensor products.

way from one or the other of the 2x2 matrices $\Gamma(r)$ or $\Gamma(s)$. We now define two operators, **R** and **S**, which map 2x2 matrices to 4x4 matrices.

$$\mathbf{R}[\mathbf{A}]= \begin{pmatrix} A_{11}\mathbf{I} & A_{12}\mathbf{I} \\ A_{21}\mathbf{I} & A_{22}\mathbf{I} \end{pmatrix} \qquad \mathbf{S}[\mathbf{B}]= \begin{pmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}$$

A and **B** are arbitrary 2x2 matrices; **A** has elements A_{ij}. **I** is the 2x2 identity matrix. **0** is the 2x2 matrix of zeroes. It is easy to verify that $\mathbf{R}[\mathbf{A}]\mathbf{S}[\mathbf{B}]= \mathbf{S}[\mathbf{B}]\mathbf{R}[\mathbf{A}]$, thus **R** and **S** commute with one another. Also we see $(\mathbf{R}[\mathbf{A}])^n= \mathbf{R}[\mathbf{A}^n]$ and $(\mathbf{S}[\mathbf{B}])^n= \mathbf{S}[\mathbf{B}^n]$.

Two more operators prove useful as well. $\vec{R}^T$ and $\vec{S}^T$ map 2 element row vectors to four element row vectors.

$$\vec{R}[\vec{V}]= (V_1, V_1, V_2, V_2) \quad \text{and}$$

$$\vec{S}[\vec{V}]= (V_1, V_2, V_1, V_2),$$

where $\vec{V}^T= (V_1, V_2)$ is an arbitrary 2 element row vector. $\vec{R}[\vec{V}]$ and $\vec{S}[\vec{V}]$ satisfy several interesting identities involving $\mathbf{R}[\Gamma]$ and $\mathbf{S}[\Gamma]$, when Γ is any (2x2) matrix with columns that sum to unity:[38]

$$\vec{R}[\vec{V}]\mathbf{S}[\Gamma]= \vec{R}[\vec{V}], \qquad \vec{S}[\vec{V}]\mathbf{R}[\Gamma]= \vec{S}[\vec{V}]$$

$$\vec{R}[\vec{V}]\mathbf{R}[\Gamma]= \vec{R}[\vec{V}\Gamma] \text{ and} \qquad \vec{S}[\vec{V}]\mathbf{S}[\Gamma]= \vec{S}[\vec{V}\Gamma].$$

Thus, $\mathbf{R}[\Gamma]$ is an identity for $\vec{S}^T$, and $\mathbf{S}[\Gamma]$ is an identity for $\vec{R}^T$. We can describe the second pair of identities by saying $\vec{R}^T$ consolidates the argument of $\mathbf{R}[\Gamma]$, and $\vec{S}^T$ consolidates the argument of $\mathbf{S}[\Gamma]$. We verify these identities by direct inspection.

---

[38] In our application, Γ will always be a stochastic matrix or a positive or negative power of a stochastic matrix.

The conditional entropy of Γ_{4x4}:

$\vec{H}_4^T(n)$ is a row vector of column entropies of $(\Gamma_{4x4})^n$. Any column of Γ_{4x4} (or its n^{th} power) has the form of a probability distribution over two independent events. What is the entropy of such a probability distribution? The entropy of the four element probability vector $\vec{P}_4$, where

$$\vec{P}_4^T = (p_1(r)p_1(s),\ p_1(r)p_2(s),\ p_2(r)p_1(s),\ p_2(r)p_2(s)),$$

is just the sum of the entropies of the independent distributions of which $\vec{P}_4$ is composed.[39] Thus, the entropy of $\vec{P}_4$ equals $H[\vec{P}(r)] + H[\vec{P}(s)]$, where

$$\vec{P}^T(r) = ((p_1(r),\ .p_2(r)) \text{ and } \vec{P}^T(s) = ((p_1(s),\ .p_2(s)).$$

Each column of $(\Gamma_{4x4})^n$ is a probability distribution with the form of $\vec{P}_4$. Thus the entropy of each column of $(\Gamma_{4x4})^n$ is just the sum of the entropies of the distributions of which the column is composed. The first two columns of $(\Gamma_{4x4})^n$ involve transitions out of the "1" state of particle r. Thus the contribution of particle r to the entropy of both of these two columns is the same, and is just the entropy of the first column of the 2x2 transition matrix $\Gamma(r)$. This entropy is the first element of $\vec{H}_r^T(n)$, where $\vec{H}_r^T(n)$ is the conditional entropy of $(\Gamma(r))^n$. Similarly, the last two columns of $(\Gamma_{4x4})^n$ involve transitions out of the "2" state of particle r. Thus the contribution of particle r to the entropy of both of these two columns is the same, and is just the second element of $\vec{H}_r^T(n)$. Thus, particle r contributes $\vec{R}[\vec{H}_r^T(n)]$ to $\vec{H}_4^T(n)$. Recall that in the state assignments of the four–state system, particle s had the pattern: 1, 2, 1, 2. Particle s therefore

---

[39] Entropies sum for independent distributions essentially because the *log* of a product is the sum of the *logs*. Let's calculate the entropy of P_4. The first two elements of P_4 contribute $p_1(r)p_1(s)(lnp_1(r)+lnp_1(s)) + p_1(r)p_2(s)(lnp_1(r)+lnp_2(s))$ which factors and becomes $p_1(r)H[P(s)] + p_1(r)lnp_1(r)$. In a similar way, the last two elements of P_4 contribute $p_2(r)H[P(s)] + p_2(r)lnp_2(r)$. The sum of these two contributions is $H[P(s)] + H[P(r)]$.

contributes $\vec{S}[\vec{H}^{T}_{\theta}(n)]$ to the four–state conditional entropy $\vec{H}^{T}_{4}(n)$. Summing these two contributions we obtain:

$$\vec{H}^{T}_{4}(n) = \vec{R}[\vec{H}^{T}_{r}(n)] + \vec{S}[\vec{H}^{T}_{\theta}(n)] \quad .$$

$$\underline{\vec{H}^{T}_{4}(n)(\mathbf{\Gamma}_{4x4})^{-n}.}$$

If we assemble the identities detailed above, we can calculate the four–state rereferenced conditional entropy. We find that $\vec{H}^{T}_{4}(n)(\mathbf{\Gamma}_{4x4})^{-n}$ becomes:[40]

$$= (\vec{R}[\vec{H}^{T}_{r}(n)] + \vec{S}[\vec{H}^{T}_{\theta}(n)]) \; \mathbf{R}[\mathbf{\Gamma}_{r}^{-n}] \; \mathbf{S}[\mathbf{\Gamma}_{s}^{-n}]$$

$$= \vec{R}[\vec{H}^{T}_{r}(n)] \; \mathbf{S}[\mathbf{\Gamma}_{s}^{-n}] \; \mathbf{R}[\mathbf{\Gamma}_{r}^{-n}] + \vec{S}[\vec{H}^{T}_{\theta}(n)] \; \mathbf{R}[\mathbf{\Gamma}_{r}^{-n}] \; \mathbf{S}[\mathbf{\Gamma}_{s}^{-n}]$$

$$= \vec{R}[\vec{H}^{T}_{r}(n)] \; \mathbf{R}[\mathbf{\Gamma}_{r}^{-n}] + \vec{S}[\vec{H}^{T}_{\theta}(n)] \; \mathbf{S}[\mathbf{\Gamma}_{s}^{-n}]$$

$$= \vec{R}[\vec{H}^{T}_{r}(n)\mathbf{\Gamma}_{r}^{-n}] + \vec{S}[\vec{H}^{T}_{\theta}(n)\mathbf{\Gamma}_{s}^{-n}] \quad .$$

The first equality utilizes the expression for $\vec{H}^{T}_{4}(n)$ and the factorization of $(\mathbf{\Gamma}_{4x4})^{-n}$. The second equality requires the distributivity of matrix multiplication and the commutivity of the factors of $(\mathbf{\Gamma}_{4x4})^{-n}$. The third equality depends on $\mathbf{S}[\mathbf{\Gamma}]$ being an identity for $\vec{R}^{T}$, and on $\mathbf{R}[\mathbf{\Gamma}]$ being an identity for $\vec{S}^{T}$. Finally, in the fourth equality, $\vec{R}^{T}$ consolidates the argument of $\mathbf{R}[\mathbf{\Gamma}]$ and $\vec{S}^{T}$ consolidates the argument of $\mathbf{S}[\mathbf{\Gamma}]$.

Thus, $\vec{H}^{T}_{4}(n)(\mathbf{\Gamma}_{4x4})^{-n} = \vec{R}[\vec{H}^{T}_{r}(n)\mathbf{\Gamma}_{r}^{-n}] + \vec{S}[\vec{H}^{T}_{\theta}(n)\mathbf{\Gamma}_{s}^{-n}]$. This identity is very intuitive. It says, the rereferenced conditional entropy of the four–state system is the direct sum of the rereferenced conditional entropies of the two spins which compose the four–state system. The operators $\mathbf{R}$ and $\mathbf{S}$ merely serve to convert between the bases of the independent spin systems and their four–state composition. If the limit exists as n goes to infinity of $\vec{H}^{T}_{r}(n)\mathbf{\Gamma}_{r}^{-n}$ and $\vec{H}^{T}_{\theta}(n)\mathbf{\Gamma}_{s}^{-n}$, then obviously it exists for the four–state composition,

---

[40] We've streamlined the notation slightly by subscripting the r and s of Γ. Thus $\Gamma(r)=\Gamma_{r}$, etc.

$\hat{H}_4^T(n)(\Gamma_{4x4})^{-n}$. Note that our result is valid, not only in the limit as n becomes infinite, but also for finite n as well.[41]

[41]Professor Hopfield points out that physical quantities of independent systems do not change merely because we aggregate the systems in our notation. He reasons that, if the rereferenced conditional entropy is to have any physical significance, then, at the very least, it must satisfy some sort of composition law.

Conclusion

In this thesis we have viewed time as a channel and the state of a system as a message. As the system state evolves with time, the message gets degraded. Thermodynamics quantifies the advance of the system toward equilibrium with the free energy measure. Information theory quantifies the loss of memory of initial state with the mutual information measure. The free energy depends on the internal energy. The mutual information depends on the conditional entropy. The internal energy is a linear operator which maps the state vector at time t to a scalar with the dimensions of energy. The conditional entropy is a linear operator which maps the state vector at time zero to a dimensionless scalar. "Rescaling" the internal energy makes it dimensionless. "Rereferencing" the conditional entropy makes it refer to the state vector at time t, rather than the state vector at time zero. In this thesis, we have proved that the rescaled internal energy and the rereferenced conditional entropy become identical operators in the asymptotic limit of long times. This identity holds for the class of systems where the time constants of different modes span less than a factor of two.

This thesis contains several items which are original. In particular, our calculation of the long time limit of the rereferenced conditional entropy is new. This limit is nifty because it is a singular limit. Our statement of the relation between the internal energy and the conditional entropy is new. Our statement of the relation between the free energy and the mutual information is new. These relations imply that thermodynamics and information theory are structurally similar. Any new relation between thermodynamics and information theory is intrinsically interesting; a structural relation is valuable because it allows us to reason by analogy. If we strip away all this hype, then what is left of the contribution of this

thesis? In preparing this thesis, we have come to believe that Shannon's conditional entropy has a place in physics, either in thermodynamics or some other allied area. We think that the thesis makes this conjecture credible and we see this credibility as the contribution of the thesis. Conditional entropy may have a place in nonequilibrium thermodynamics; it may also have a place in that area of statistical physics which deals with things like fluctuation dissipation theorems. We now briefly elaborate these possibilities.

Conditional entropy and nonequilibrium thermodynamics.

All physical theories simplify reality by abstracting it; this is necessary because reality is terribly complicated. Physical theories are judged by their simplicity and by the accuracy of their predictions. Thermodynamics works in an abstraction which discards the complications of detailed dynamics; for equilibrium systems, heat, temperature and entropy effectively summarize what is left of dynamics. A major impediment to the development of a satisfactory theory of nonequilibrium thermodynamics has been the lack of an appropriate abstraction. Nonequilibrium thermodynamics needs to retain more of system dynamics than heat, temperature and entropy; still it should retain appreciably less of system dynamics than, say, the first order rate equations of chemical kinetics. In this context, conditional entropy, or perhaps $\vec{H}^T(n)$, seem especially attractive. The components of $\vec{H}^T(n)$ effectively summarize system dynamics; since these components measure volumes in state space, they should fit naturally into the framework of any theory which is built upon thermodynamics.

Conditional entropy and fluctuation dissipation theorems.

Physical systems usually are found in thermal environments where they are bombarded by noise. Such systems exhibit fluctuations. Also, such

systems exhibit dissipation; if we disturb them, then they respond, but in time the disturbance dies away. Einstein was the first to point out that the bombarding noise is the common cause of both the fluctuations and the dissipation.[42] Conditional entropy might offer another way to connect fluctuations and dissipation. Conditional entropy measures the volume of state space which is swept out on account of noise. Thus, conditional entropy connects fluctuations to state space. A fundamental result of the theory of dynamical systems connects dissipation with contraction of state space volume. Thus, with conditional entropy, we can hope to link fluctuations with dissipation via state space.

---

[42]Einstein studied brownian motion; he concluded that microscopic bombardment caused these fluctuations and that the same bombardment was also responsible for the dissipation which was observed. Subsequently, Nyquist studied voltage fluctuations across a resistor; he too concluded that the source of the fluctuations was also the source of the resistance. Later, Callen & Welton and then Kubo proved "fluctuation dissipation" theorems of increasing elegance.

Bibliography

[Bennett]: Bennett, Charles H. "The Thermodynamics of Computation– a Review," *International Journal of Theoretical Physics*, Vol. 21, No. 12, 1982.

[Brillouin]: Brillouin, Leon Science and Information Theory. New York: Academic Press, Inc., 1956.

[Callen]: Callen, Herbert B. Thermodynamics, New York: John Wiley & Sons, Inc., 1960.

[Fermi]: Fermi, Enrico Thermodynamics, New York: Dover Publications, Inc., 1956.

[Feynman]: Feynman, Richard P., Leighton and Sands The Feynman Lectures on Physics. Menlo Park, California: Addison–Wesley Publishing Company, 1963.

[GantMacher]: Gantmacher, F. R.The Theory of Matrices (volumes I and II). New York: Chelsea Publishing Company, 1959.

[Jaynes I]: Jaynes, E. T. "Information Theory and Statistical Mechanics," *Physical Review*, vol 106, No. 4, 1957.

[Jaynes II]: Jaynes, E. T. "Information Theory and Statistical Mechanics. II," *Physical Review*, vol. 108, No. 2, 1957.

[Kato]: Kato, Tosio Perturbation Theory for Linear Operators. New York: Springer–Verlag, Inc., 1966.

[Khinchin 1]: Khinchin, A. I. Mathematical Foundations of Statistical Mechanics. New York: Dover Publications, Inc., 1949.

[Khinchin 2]: Khinchin, A. I. Mathematical Foundations of Information Theory. New York: Dover Publications, Inc., 1957.

[McEliece]: McEliece, Robert J. The Theory of Information and Coding. Reading, Massachusetts: Addison–Wesley Publishing Company, (Advanced Book Program), 1977.

[Pierce]: Pierce, John R. An Introduction to Information Theory, New York: Dover Publications, Inc., 1980.

[Prigogine]: Prigogine, Ilya Introduction to Nonequilibrium Thermodynamics. New York: Wiley–Interscience, 1962.

[Reif]: Reif, F. Statistical Physics, New York: McGraw–Hill, Inc., 1964.

[Shannon]: Shannon, Claude E. and Weaver, W. The Mathematical Theory of Communication. Urbana Ill.: University of Illinois Press, 1949.

[Szilard]: Szilard, L., *Z. Physik* 53, 840, 1929.

[Tolman]: Tolman, Richard C. The Principles of Statistical Mechanics. New York: Dover Publications, Inc., 1979.

[Tribus]: Tribus, Myron; Thermostatics and Thermodynamics; D Van Nostrand Company, Inc., 1961.