Basic Notions of Entropy and Entanglement

Frank Wilczek
Center for Theoretical Physics, MIT, Cambridge MA 02139 USA

March 3, 2014

Abstract

Entropy is a nineteenth-century invention, with origins in such practical subjects as thermodynamics and the kinetic theory of gases. In the twentieth century entropy came to have another, more abstract but more widely applicable interpretation in terms of (negative) information. Recently the quantum version of entropy, and its connection to the phenomenon of entanglement, has become a focus of much attention. This is a self-contained introduction to foundational ideas of the subject.

This is the first of three notes around centered around the concept of entropy in various forms: information-theoretic, thermodynamic, quantum, and black hole. This first note deals with foundations; the second deals mainly with black hole and geometric entropy; the third explores variational principles that flow naturally from the ideas.

A good reference for this first part is Barnett, "Quantum Information" [1], especially chapters 1, 8, and the appendices.

1. *Classical Entropy and Information*

   As with many scientific terms taken over from common language, the scientific meaning of *information* is related to, but narrower and more precise than, the everyday meaning. We think of information as relieving uncertainty, and this is the aspect emphasized in Shannon's scientific version. We seek a measure $I(p_j)$ of the relief of uncertainty gained, when we observe the actual result j of a stochastic event with probability distribution p_j. We would like for this measure to have the property that the information gain from successive observation of

independent events is equal to the sum of the two gains separately. Thus

$$I(p_j q_k) \;=\; I(p_j) \,+\, I(q_k) \tag{1}$$

Thus we are led to

$$I(p_j) \propto -\ln p_j \tag{2}$$

The choice of proportionality constant amounts to a choice of the base of logarithms. For physics purposes it is convenient to take equality in Eqn. (information)[1]. With that choice, a unit of information is called a *nat*. Increasingly ubiquitous, however, is the choice $I(p_j) = -\log_2 p$. In this convention, observation of the outcome of one fifty-fifty event conveys one unit of information; this is one *bit*. I will generally stick with ln, but with the understanding that the natural logarithm means the logarithm whose base seems natural in the context[2] .

Given a random event described by the probability distribution p_j, we define its *entropy* $S(p)$ to be the average information gained upon observation:

$$S(p) \;=\; -\sum_j p_j \ln p_j \tag{3}$$

For a two-state distribution, the entropy can be expressed as a function of a single number, the probability of (either) event:

$$S_2(p) = -p \ln p - (1-p)\ln(1-p) \tag{4}$$

This is maximized, at one nat, when $p = \frac{1}{2}$.

This notion of entropy can be used as a heuristic, but transparent and basically convincing, foundation for statistical mechanics, as follows. Suppose that we have a system whose energy levels are E_j. We observe that a wide variety of systems will, perhaps after some flow of energy, come into equilibrium with a large ambient "heat bath". We suppose that in equilibrium the probability of occupying some particular state j will depend only on E_j. We also assume that the randomizing influence of interactions with the heat bath will obliterate as much information about the system as can be obliterated, so that we should maximize the

[1] As will become clear shortly, the convention is closely related to the choice of units for temperature. If we used base 2, for instance, it would be natural to use the Maxwell-Boltzmann factor $2^{-E/\tilde{T}}$ in place of $e^{-E/T}$, which amounts to defining $\tilde{T} = T \ln 2$.

[2] In the same spirit, here is my definition of "naturality" in physics: Whatever nature does, is natural.

average information, or entropy, we would gain by actually determining which state occurs. Thus we are led to the problem of maximizing $S(\{p_j\})$ under the constraints that the average energy is some constant $\mathcal{E}$ and of course that the p_j sum to unity. That problem is most easily solved by introducing Lagrange multipliers λ and η for the two constraints. Then we get

$$\frac{\delta}{\delta p_k} \sum_j (-p_j \ln p_j + \lambda p_j E_j + \eta p_j) \;=\; -\ln p_j - 1 + \lambda E_j + \eta \;=\; 0 \quad (5)$$

so that

$$p_j \;=\; \frac{\eta}{e}\, e^{\lambda E} \quad (6)$$

We can use the two disposable constants η, λ to satisfy the two constraints

$$\begin{aligned} \sum_j p_j &= 1 \\ \sum_j p_j E_j &= \mathcal{E} \end{aligned} \quad (7)$$

With $-\frac{1}{T} \equiv \lambda$, we find that we get the Maxwell-Boltzmann distribution

$$p_j \;=\; \frac{e^{-E_j/T}}{\sum\limits_k e^{-E_k/T}} \quad (8)$$

It is revealing that the inverse temperature appears as a Lagrange multiplier dual to energy. (Inverse temperature is in many ways more fundamental than temperature. For example, by using inverse temperature we would avoid the embarrassment that negative temperature is hotter than infinite temperature, as opposed to colder than zero temperature!)

It is convenient to introduce the inverse temperature variable $\beta \equiv \frac{1}{T}$ and the partition function

$$Z(\beta) \;\equiv\; \sum_k e^{-E_k/T} \;=\; \sum_k e^{-\beta E_k} \quad (9)$$

We find for the energy and the information-theoretic entropy

$$\begin{aligned} \mathcal{E} &= -\frac{\partial \ln Z}{\partial \beta} \\ S_{\text{inf.}} &= \sum_k (\beta E_k + \ln Z) \frac{e^{-\beta E_k}}{Z} \;=\; \beta \mathcal{E} - \overline{\ln Z} \end{aligned} \quad (10)$$

or

$$\overline{\ln Z} \;=\; -\beta(\mathcal{E} - TS_{\text{inf.}}) \tag{11}$$

where the overline denotes averaging over the Boltzmann distribution. Now in statistical mechanics we learn that $\overline{\ln Z}$ is related to the Helmholtz free energy $\mathcal{F} = \mathcal{E} - T\mathcal{S}$ as

$$\overline{\ln Z} \;=\; -\beta\mathcal{F} \tag{12}$$

And so we conclude that the thermodynamic and information-theoretic entropy are equal

$$\mathcal{S} \;=\; S_{\text{inf.}} \tag{13}$$

A much more entertaining and directly physical argument for the close connection of information entropy and thermodynamic entropy arises from consideration of Szilard's one molecule thought-engine, which strips Maxwell's demon down to its bare essence. We imagine a box of volume V that contains a single molecule, in equilibrium with a heat bath at temperature T. (See Figure 1.) A partition can be inserted midway in the box, and can move frictionlessly toward either end, where it can be removed. One also has weights attached to pulleys on either side, either one of which can be engaged by attachment to an extension of the partition, or disengaged onto a fixed pin. Now if when the partition is inserted the appropriate pulley is engaged, so that expansion of the one molecule "gas" lifts a weight, we can extract work from the heat bath! We can also let the gas fully expand, remove the partition, and start the cycle over again. This process appears, on the face of it, to contradict the laws of thermodynamics.

The work done by the gas is

$$W \;=\; \int_{V/2}^{V} P dV \;=\; T \ln 2 \tag{14}$$

using the ideal gas law $PV = T$. Thus the entropy of the heat bath *decreases* by

$$\Delta S_{\text{bath}} \;=\; -\frac{W}{T} \;=\; -\ln 2 \tag{15}$$

To reconcile this result with the second law of thermodynamics, we need to find some additional change in entropy that compensates this. A human experimenter who is somehow locating the molecule and

deciding which pulley to engage is too complicated and difficult to model, so let us replace her with a simple automaton and an indicator gauge. The indicator gauge settles into one or another of two settings, depending on the location of the molecule, and the automaton, in response, engages the correct pulley. And now we see that the state of the total system has settled into one of two equally likely alternatives, namely the possible readings of the indicator gauge. This state-outcome contributes exactly one nat:

$$\Delta S_{\text{pointer}} \;=\; -\ln\frac{1}{2} \;=\; \ln 2 \tag{16}$$

Thus the total entropy does not change. This is what we expect for a reversible process – and we might indeed reverse the process, by slowly pushing the partition back to the center, and then removing it while at the same time driving the indicator – if that indicator is frictionless and reversible! – back to its initial setting.

On the other hand, if the indicator is "irreversible" and does not automatically return to its initial setting, we will not get back to the initial state, and we cannot literally repeat the cycle. In this situation, if we have a store of indicators, their values, after use, will constitute a memory. It might have been in any of 2^N states, but assumes (after N uses) just one. This reduction of the state-space, by observation, must be assigned a physical entropy equal to its information theoretic entropy, in order that the second law remain valid. Conversely, the act of erasing the memory, to restore blank slate ignorance, is irreversible, and is associated with net *positive* entropy $N \ln 2$. Presumably this means, that to accomplish the erasure in the presence of a heat bath at temperature T, we need to do work $NT \ln 2$.

2. *Classical Entanglement*

   We[3] say that two subsystems of a given system are *entangled* if they are not independent. Quantum entanglement can occur at the level of wave functions or operators, and has some special features, but there is nothing intrinsically quantum-mechanical about the basic concept. In particular, if we define composite stochastic events that have two aspects depending on variables a_j, b_k, it need not be the case that the joint probability distribution factorizes as

$$p_{AB}(a_j, b_k) \overset{\text{independent}}{=} p_A(a_j) p_B(b_k) \tag{17}$$

[3] At least, the more bombastic among us.

where the separate (also called marginal) distributions are defined as

$$\begin{aligned} p_A(a_j) &\equiv \sum_k p_{AB}(a_j, b_k) \\ p_B(b_k) &\equiv \sum_j p_{AB}(a_j, b_k) \end{aligned} \tag{18}$$

As indicated in Eqn. (17), if the joint distribution does factorize we say the variables are *independent*. Otherwise they are entangled.

I should also mention that the concept of wave function entanglement, while *profoundly* strange, is hardly new or unfamiliar to practitioners of the art. The common wave functions of atomic and molecular physics live in many-dimensional configuration spaces, contain spin variables, etc. and quite generally do not factorize. Almost every practical use of quantum mechanics tests the existence of entanglement, and its mathematical realization in the (tensor product) formalism.

Entropy is a good diagnostic for entanglement. We will prove momentarily that the entropy of the joint distribution is equal to or greater than the sum of the entropies of the separate distributions:

$$S(A) + S(B) \geq S(AB) \tag{19}$$

with equality if and only if the distributions are independent. The quantity

$$S(A:B) \equiv S(A) + S(B) - S(AB) \geq 0 \tag{20}$$

is called the *mutual information* between A and B. It plays an important role in information theory. (See, for example, the excellent book [2].)

3. *Inequalities*

There are several notable inequalities concerning entropy and related quantities.

One concerns the relative entropy $S(p \parallel q)$ of two probability distributions on the same sample space, as follows:

$$S(p \parallel q) \equiv \sum_k p_k \ln \frac{p_k}{q_k} \geq 0 \tag{21}$$

Indeed, this quantity goes to infinity at the boundaries and is manifestly bounded below and differentiable, so at the minimum we must

have

$$\frac{\delta}{\delta q_j}(\sum_k p_k \ln \frac{p_k}{q_k} - \lambda(\sum_l q_l - 1)) \;=\; \frac{p_j}{q_j} - \lambda \;=\; 0 \qquad (22)$$

where λ is a Lagrange multiplier. This implies $p_j = q_j$, since both are normalized probability distributions.

The mutual information inequality Eqn. (20) is a special case of the relative entropy inequality, corresponding to $p = p_{AB}, q = p_A p_B$, as you'll readily demonstrate.

Other inequalities follow most readily from concavity arguments. The basic entropy building-block $-p \ln p$ is, as a function of p, concave in the relevant region $0 \leq p \leq 1$. (See Figure 2.) One sees that if we evaluate several samples of this function, the average of the evaluations is greater than the evaluation of the average. Thus we have several probability distributions $p^{(\mu)}$, then the average distribution has greater entropy:

$$S(\lambda_1 p^{(1)} + ... + \lambda_n p^{(n)}) \;\geq\; \lambda_1 S(p^{(1)}) + ... + \lambda_n S(p^{(n)}) \qquad (23)$$

where of course $\lambda_j \geq 0, \sum_j \lambda_j = 1$. We may re-phrase this in a way that will be useful later, as follows. Suppose that

$$\begin{aligned} p'_j &= \sum_k \lambda_{jk} p_k \\ \sum_j \lambda_{jk} &= 1 \\ \sum_k \lambda_{jk} &= 1 \\ \lambda_{jk} &\geq 0 \end{aligned} \qquad (24)$$

Then

$$S(p') \;\geq\; S(p) \qquad (25)$$

It is also natural to define *conditional entropy*, as follows.

The conditional probability $p(a_j|b_k)$ is, by definition, the probability that a_j will occur, given that b_k has occurred. Since the probability that both occur is $p(a_j, b_k)$, we have

$$p(a_j|b_k)p(b_k) \;=\; p(a_j, b_k) \qquad (26)$$

From these definitions, we can derive the celebrated, trivial yet profound theorem of Bayes

$$p(b_k|a_j) \;=\; \frac{p(b_k)p(a_j|b_k)}{p(a_j)} \tag{27}$$

This theorem is used in statistical inference: We have hypotheses that hold with "prior" probabilities b_k, and data described by the events a_j, and would like to know what the relative probabilities of the hypotheses look like, after the data has come in – that is, the $p(b_k|a_j)$. Bayes' theorem allows us to get to those from (presumably) calculable consequences $p(a_j|b_k)$ of our hypotheses.

Similarly, the conditional entropy $S(A|B)$ is defined to be the average information we get by observing a, given that b has already been observed. Thus

$$S(A|B) \;=\; -\sum_k p(b_k)(-\sum_j p(a_j|b_k)\ln p(a_j|b_k)) \tag{28}$$

Upon expanding out the conditional probabilities, one finds the satisfying result

$$S(A|B) \;=\; S(AB) - S(B) \tag{29}$$

This can be regarded as the entropic version of Bayes' theorem.

4. *Quantum Entropy*

A useful quantum version of entropy was defined by von Neumann, and it is generally referred to as von Neumann entropy. It is defined, for any (not necessarily normalized) positive definite Hermitean operator ρ as

$$S(\rho) \;=\; -\frac{\mathrm{Tr}\,\rho\ln\rho}{\mathrm{Tr}\,\rho} \tag{30}$$

In most applications, ρ is the density matrix of some quantum-mechanical system.

This is a natural definition, for several reasons. First, it reduces to the familiar definition of entropy, discussed previously, when ρ is diagonal and we regard its entries as defining a probability distribution. Second, we can use it, as we used the classic entropy, to provide a foundation for statistical mechanics. Thus we seek to maximize the (von Neumann) entropy subject to a fixed average energy. Using Lagrange multipliers, we vary

$$-\mathrm{Tr}\,\rho\ln\rho + \lambda(\mathrm{Tr}\rho H - \mathcal{E}) + \eta(\mathrm{Tr}\rho - 1) \tag{31}$$

Now when we put $\rho \rightarrow \rho + \delta\rho$, we have to be concerned that the two summands might not commute. Fortunately, that potential complication does not affect this particular calculation, for the following reason. Only the first term is nonlinear in ρ. Now imagine expanding the function $\rho \ln \rho$ as a power series (around some regular point), varying term by term, and focusing on the terms linear in $\delta\rho$. Because we are inside a trace, we can cyclically permute, and bring $\delta\rho$ to the right in every case – in other words, we can do calculus as if ρ were simply a numerical function. So the variation gives

$$\mathrm{Tr}(-\ln\rho - 1 + \lambda H + \eta)\delta\rho \;=\; 0 \tag{32}$$

Since this is supposed to hold for all $\delta\rho$, we find that every matrix element of the quantity in parentheses must vanish, and following the same steps as we did earlier, in the classical case, we get

$$\rho \;=\; \frac{e^{-\beta H}}{\mathrm{Tr}\, e^{-\beta H}} \tag{33}$$

This is indeed the standard thermal density matrix. Furthermore, if we use this ρ to evaluate thermodynamic entropy, we find that the expectation value of the thermodynamic entropy is given by its von Neumann entropy $-\mathrm{Tr}\,\rho \ln \rho$.

Elementary properties of the von Neumann entropy include:

- It is invariant under change of basis, or in other words under unitary transformations

$$\rho \;\rightarrow\; U\rho U^{-1} \tag{34}$$

- If ρ is the density matrix of a closed quantum dynamical system, the entropy will not change in time. (Indeed, the density matrix evolves unitarily.)
- If ρ is the density matrix of a pure state, the entropy vanishes.

The last two properties illustrate that some coarse-graining must be involved in passing from the von Neumann entropy to thermodynamic entropy, for an isolated system.

A relative entropy inequality, in the form

$$\mathrm{Tr}\,\rho\,(\frac{\ln\rho}{\mathrm{Tr}\rho} - \frac{\ln\sigma}{\mathrm{Tr}\sigma}) \;\geq\; 0 \tag{35}$$

is valid in the quantum case, as is a concavity inequality

$$S(\lambda_1\rho^{(1)} + ... + \lambda_n\rho^{(n)}) \;\geq\; \lambda_1 S(\rho^{(1)}) + ... + \lambda_n S(\rho^{(n)}) \tag{36}$$

One also has the "erasure" result, that setting all the off-diagonal entries in a density matrix to zero increases its entropy. Indeed, let the $|\phi_j\rangle$ be the eigenvectors of the new density matrix, associated with the eigenvalues p'_j, and the $|\psi_k\rangle$ the eigenvectors of the old density matrix, with associated with the eigenvalues (probabilities) p_k. We have

$$\begin{aligned} p'_j &= \rho_{jj} = \sum_k \langle\phi_j|\psi_k\rangle p_k \langle\psi_k|\phi_j\rangle = \sum_j \lambda_{jk} p_k \\ \lambda_{jk} &\equiv \sum_k |\langle\phi_j|\psi_k\rangle|^2 \end{aligned} \tag{37}$$

This gives us the set-up anticipated in Eqn.(24), and the result follows.

The composite of systems A, B will live on the Hilbert space $H_A \otimes H_B$. We are here using "composite" in a very broad sense, simply to mean that we have a division of the dynamical variables into distinct subsets. If the density matrix of the composite system is ρ_{AB}, we derive subsystem density matrices by taking traces over the complementary variables

$$\begin{aligned} \rho_A &= \mathrm{Tr}_B\, \rho_{AB} \\ \rho_B &= \mathrm{Tr}_A\, \rho_{AB} \end{aligned} \tag{38}$$

The joint density matrix ρ_{AB} is the quantum version of a joint probability distribution, and we the subsystem density matrices are the quantum version of marginal distributions. As a special case of the quantum relative entropy inequality, we have the inequality

$$S(A:B) \;=\; S(A) + S(B) - S(AB) \;\geq\; 0 \tag{39}$$

also in the quantum case.

We might expect that $S(A:B)$ is sensitive to the quantum entanglement of systems A and B, and provides a measure of such entanglement. As a minimal test of that intuition, let us consider the singlet state of a system of two spin-$\frac{1}{2}$ particles:

$$|\psi\rangle_{AB} \;=\; \frac{1}{\sqrt{2}}(|\uparrow\rangle \otimes |\downarrow\rangle \;-\; |\downarrow\rangle \otimes |\uparrow\rangle) \tag{40}$$

with the corresponding density matrix (in the basis $|\uparrow\uparrow\rangle, |\uparrow\downarrow\rangle, |\downarrow\uparrow\rangle, |\downarrow\downarrow\rangle$

$$\frac{1}{2}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{41}$$

Tracing over either subsystem, we get

$$\rho_A \;=\; \rho_B \;=\; \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{42}$$

with von Neumann entropy of 1 nat.

5. *Quantum Entanglement and Quantum Entropy*

So far, the properties of the quantum entropy have appeared as direct analogues of the properties of classical entropy. Now we will discuss a dramatic difference. For the classical entropy of a composite system we have

$$S_{\text{cl.}}(AB) \;\geq\; \text{Max}\left(S_{\text{cl.}}(A), S_{\text{cl.}}(B)\right) \tag{43}$$

– the composite system always contains more untapped information than either of its parts. In the quantum case, the analogue of Eqn.(43) fails dramatically, as we will now demonstrate. Thus it is not correct to think of the quantum entropy as a measure of information, at least not in any simple way.

To bring out the difference, and for many other purposes, it is useful to develop the *Schmidt decomposition*. The Schmidt decomposition is a normal form for wave functions in a product Hilbert space. In general, we can write our wave function

$$\psi_{AB} \;=\; \sum_{ab} c_{ab}|a\rangle \otimes |b\rangle \tag{44}$$

in terms of its coefficients relative to orthonormal bases $|a\rangle, |b\rangle$. The Schmidt decomposition gives us a more compact form

$$\psi_{AB} \;=\; \sum_{ab} s_k|\phi_k\rangle \otimes |\psi_k\rangle \tag{45}$$

for a suitable choice of orthonormal vectors $|\phi_k\rangle, \psi_k\rangle$ in the two component Hilbert spaces. (These may not be complete bases, but of course they can be beefed up to give bases.)

Assuming Eqn. (45), we see that the ϕ_k are eigenvectors of the density matrix ρ_A, with eigenvalues $|s_k|^2$. This gives us a strategy to prove it! That is, we define the $|\psi_k\rangle$ to be the eigenvectors of ρ_A[4]. Writing out what this means in terms of the general expansion

$$|a\rangle \;=\; \sum_k d_{ak}|\phi_k\rangle \tag{46}$$

we have

$$\begin{aligned}(\rho_A)_{k'k} &= \sum_b c^*_{a'b}c_{ab}|a\rangle\langle a'| \\ &= \sum_{b,a',a} c^*_{a'b}d^*_{a'k'}c_{ab}d_{ak}\,|\phi_k\rangle\langle\phi_{k'}|\end{aligned} \tag{47}$$

Matching coefficients, we have

$$\sum_{b,a',a} c^*_{a'b}d^*_{a'k'}c_{ab}d_{ak} \;=\; \delta_{k'k}|s_k|^2 \tag{48}$$

for the non-zero, and therefore positive, eigenvalues $|s_k|^2$.

Going back to Eqn. (44), we have

$$\psi_{AB} \;=\; \sum_{ab} c_{ab}d_{ak}|\phi_k\rangle\otimes|b\rangle \tag{49}$$

so the candidate Schmidt decomposition involves

$$|\psi_k\rangle \;=\; \sum_{ab} s_k c_{ab}d_{ak}|b\rangle \tag{50}$$

and it only remains to verify that the $|\psi_k\rangle$ are orthonormal. But that is precisely the content of Eqn. (48).

Note that the phase of s_k is not determined; it can be changed by re-definition of $|\phi_k\rangle$ or $|\psi_k\rangle$.

Purification is another helpful concept in considering quantum entropy and entanglement. It is the observation that we can obtain any given density matrix ρ by tracing over the extra variables for a *pure state* in a composite system. Indeed, we need only diagonalize ρ:

$$\rho \;=\; \sum_k |e_k|^2|\phi_k\rangle\langle\phi_k| \tag{51}$$

[4]Including zero eigenvalue eigenvectors, if necessary, to make a complete basis.

and consider the pure state

$$\psi_{AB} \;=\; \sum_k e_k |\phi_k\rangle \otimes |\psi_k\rangle \tag{52}$$

where we introduce orthonormal states $|\psi_k\rangle$ in an auxiliary Hilbert space. Then manifestly $\rho = \rho_A$ for the pure state ψ_{AB}.

By combining the Schmidt decomposition and purification, we can draw two notable results concerning quantum entropy.

- First: The demise of Eqn. (43), as previously advertised. Indeed, the entropy of a pure state vanishes, so $S(AB) = 0$ in the purification construction – but ρ_A can represent any density matrix, regardless of its entropy $S(A)$.
- Second: If ρ_A and ρ_B are both derived from a pure state of the composite system AB, then $S(A) = S(B)$. This follows from the Schmidt decomposition, which shows that their non-zero eigenvalues are equal, including multiplicities.

The Araki-Lieb inequality, which generalizes the second of these results, affords addition insight into the nature of quantum entropy. We consider again a composite system AB, but no longer restrict to a pure state. We purify our mixed state of AB, to get a pure state Ψ_{ABC}. For the entropies defined by tracing different possibilities in that system, we have

$$\begin{aligned} S(A) &= S(BC) \\ S(B) &= S(AC) \\ S(C) &= S(AB) \end{aligned} \tag{53}$$

and therefore

$$\Big(S(A) + S(C) - S(AC) \geq 0\Big) \;\Rightarrow\; S(AB) \;\geq\; S(B) - S(A) \tag{54}$$

By symmetry between A and B, we infer

$$S(AB) \;\geq\; |S(A) - S(B)| \tag{55}$$

which is the Araki-Lieb inequality. We see that correlations between A and B can "soak up" no more information than the information in the smaller (that is, less information-rich) of them.

6. *More Advanced Inequalities*

Finally let us briefly discuss two additional inequalities, that might benefit from additional clarification or support further application.

Strong subadditivity applies to a composite system depending on three components A, B, C. It states

$$S(ABC) + S(C) \ \leq \ S(AC) + S(BC) \tag{56}$$

Alternatively, if we allow A and B to share some variables, we can re-phrase this as

$$S(A \cup B) + S(A \cap B) \ \leq S(A) + S(B) \tag{57}$$

The known proofs of this are intricate and uninformative; an incisive proof would be welcome, and might open up new directions.

We also have a remarkable supplement to the concavity inequality, in the form

$$-\sum_k \lambda_k \ln \lambda_k + \sum_k \lambda_k S(\rho_k) \ \geq \ S(\sum_k \lambda_k \rho_k) \ \geq \ \sum_k \lambda_k S(\rho_k) \tag{58}$$

with the usual understanding that the λ_k define a probability distribution and the ρ_k are normalized density matrices. The second of these is just the concavity inequality we discussed previously, but the first – in the other direction! – is qualitatively different.

We can prove it, following [1], in two stages.

First, let us suppose that the ρ_k are density matrices for pure states, $\rho_k = |\phi_k\rangle\langle\phi_k|$, where now the $|\phi_k\rangle$ need not be orthonormal. We can purify ρ using

$$\Psi_{AB} \ = \ \sum_k \sqrt{p_k} |\phi_k\rangle \otimes |\psi_k\rangle \tag{59}$$

where the $|\psi_k\rangle$ are orthonormal – clearly, tracing over the B system gives us ρ. On the other hand, if we work in the $|\psi_k\rangle$ basis, and throw-away off-diagonal terms, the modified density matrix reduces to simply p_k along the diagonal, and has entropy $-\sum_k p_k \ln p_k$. But as we saw, this erasure operation increases the entropy. Thus we have

$$-\sum_k p_k \ln p_k \ \geq \ S(B) \ = \ S(A) \tag{60}$$

That gives us what we want, in this special case.

In the general case, let us express each ρ_k in terms of its diagonalization, so

$$\begin{aligned} \rho_k &= \sum_j P_k^j |\phi_k^j\rangle\langle\phi_k^j| \\ \rho &= \sum_{jk} p_j P_k^j |\phi_k^j\rangle\langle\phi_k^j| \end{aligned} \tag{61}$$

According to our result for the special case, we have

$$-\sum_{jk}(p_j P_k^j)\ln(p_j P_k^j) \;\geq\; S(\rho) \tag{62}$$

But we can evaluate the left-hand side as

$$\begin{aligned} -\sum_{jk}(p_j P_k^j)\ln(p_j P_k^j) &= -\sum_{jk} p_j P_k^j \ln p_j - \sum_{jk} p_j P_k^j \ln P_k^j \\ &= -\sum_j p_j \ln p_j \;+\; \sum_j p_j S(\rho_j) \end{aligned} \tag{63}$$

This completes the proof.

It is quite unusual to have two-sided variational bounds on a quantity of physical interest. As we shall see, one can derive useful variational bounds from entropy inequalities. Even when they accurate numerically, however, variational estimates are usually uncontrolled analytically. Use of two-sided bounds might improve that situation.

References

[1] S. Barnett *Quantum Information* (Oxford University Press, 2009).

[2] D. MacKay *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, 2003).