Review: density matrices and quantum channels

Density matrices represent probability distributions over quantum states, and they necessarily arise in the description of quantum subsystems (or open quantum systems that interact with an environment).

$$\rho = \sum_{k=1}^{m} p_k |\psi_k\rangle\langle\psi_k| \quad , \quad p_k \geq 0 \ , \ k = 1, ..., m \quad , \quad \sum_{k=1}^{m} p_k = 1$$

Given a joint state $|\psi_{AB}\rangle$ on a Hilbert space $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$, the reduced state ρ_A on subsystem A is:

$$\rho_A = \sum_{a_i, a_j} \rho_{ij}^A |a_i\rangle\langle a_j| \quad , \quad \rho_{ij}^A = \sum_{b_k} \langle a_i b_k|\psi\rangle\langle\psi|a_j b_k\rangle$$

This mathematical operation is called the partial trace. It can be expressed more compactly:

$$\rho_A = \mathrm{tr}_B\left(\rho_{AB}\right) = \sum_{b_k} \langle b_k|\rho_{AB}|b_k\rangle$$

Where $\{b_k\}$ is a basis for $\mathcal{H}_B$ and each $\langle b_k|\rho_{AB}|b_k\rangle$is an operator acting on $\mathcal{H}_A$.

Review: density matrices and quantum channels

Given a joint state $|\psi_{AB}\rangle$ on a Hilbert space $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$, the Schmidt decomposition is a canonical form that reveals the bipartite entanglement between A and B in terms of a tensor product basis:

$$|\Psi_{AB}\rangle = \sum_{i,j} \alpha_{ij} |\psi_i^A\rangle \otimes |\psi_j^B\rangle$$

Generic State

Singular Value Decomposition of matrix α_{ij}

$$|\Psi_{AB}\rangle = \sum_{i=1}^{\chi} \sqrt{p_i} |\phi_i^A\rangle \otimes |\phi_i^B\rangle$$

Schmidt Decomposition

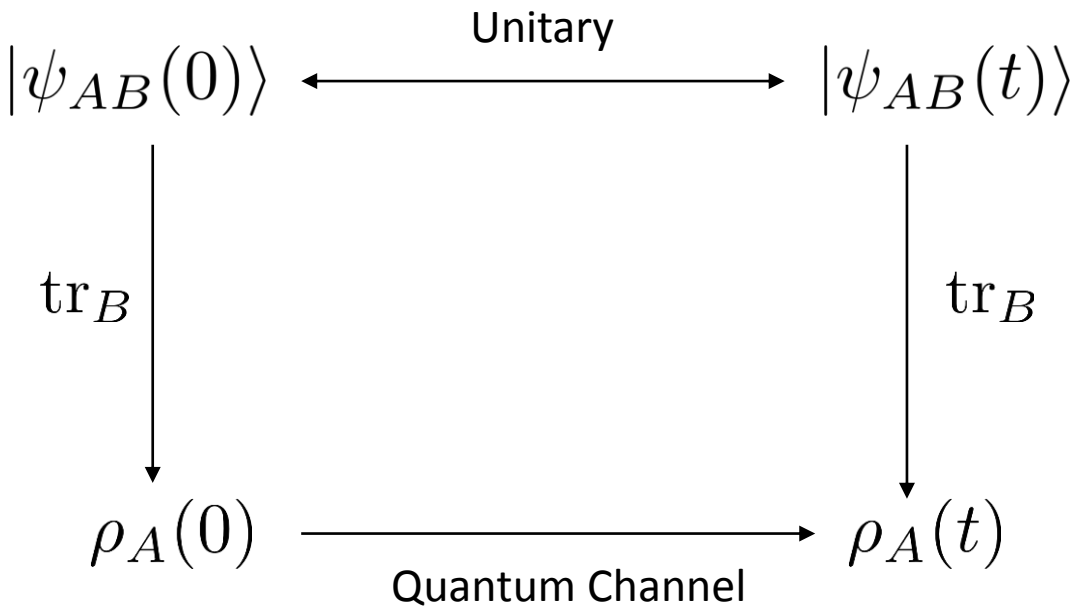
The Schmidt rank χ , which corresponds to the number of non-zero singular values of the matrix α_{ij}, provides a quantitative measure of the entanglement between A and B.

The basis vectors $\{|\phi_i^A\rangle\}$ for $\mathcal{H}_A$ in the Schmidt decomposition are exactly the eigenvectors of ρ_A , with similar statements holding for subsystem B.

Review: density matrices and quantum channels

Maps that take valid density matrices to valid density matrices are called **quantum channels**.

Quantum channels are not necessarily unitary, but they can always be understood in terms of unitaries acting on a larger joint system, followed by a partial trace.



$$\mathcal{E}(\rho_{\mathrm{sys}}(0)) = \mathrm{tr}_{\mathrm{env}}\left[U\left(\rho_{\mathrm{sys}}(0)\otimes|0_E\rangle\langle 0_E|\right)U^\dagger\right] = \sum_k \langle k_E|U|0_E\rangle\rho_{\mathrm{sys}}(0)\langle 0_E|U^\dagger|k_E\rangle = \sum_k E_k\rho_{\mathrm{sys}}E_k^\dagger$$

Review: density matrices and quantum channels

We have shown that any quantum channel has a **Kraus operator-sum representation:**

$$\mathcal{E}(\rho) = \sum_k E_k \rho E_k^\dagger \qquad\qquad \sum_k E_k^\dagger E_k = I$$

Measurement channel: the Kraus operators are projectors $\{\Pi_k\}$ with $\sum_{k=1}^{m} \Pi_k = I$.

Depolarizing channel: do nothing with probability 1 - p, and apply a random Pauli with probability p. Equivalent to do nothing with prob 1 – p, and replace qubit with maximally mixed state with prob p.

Dephasing channel: with probability p, the environment learns whether the qubit is $|0\rangle$ or $|1\rangle$. After many applications the channel converges to a measurement in the 0 / 1 basis.

Amplitude-damping channel: if the qubit is in the $|1\rangle$ state then it decays to a $|0\rangle$. Commonly applies to qubits for which $|0\rangle$ and $|1\rangle$ correspond to discrete energy levels.

Classical Information Theory

What is information? As with many abstract nouns, this could be a difficult philosophical question.

One way to resolve such questions (“what is space?”) is to instead return to ordinary language for guidance (“do we have enough space in this room?”), and develop math to assist ordinary questions.

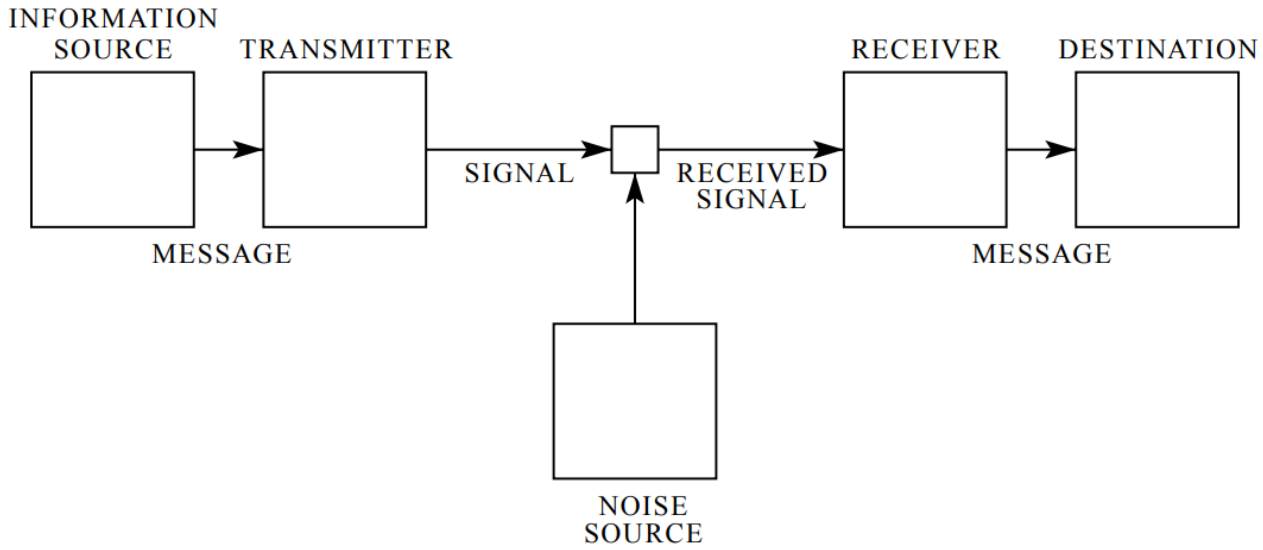
Knowingly or unknowingly, this is the strategy that enabled Claude Shannon to define and study information in his 1948 work, “A mathematical theory of communication.”

Instead of the semantic content of messages, Shannon (who built a neighborhood telegraph as a teenager) was focused on practical problems like compressing and transmitting messages.

His idea is that the information content of a message reflects the number of possibilities from which that message could have been chosen. The amount of information I receive when you send me the signal “a” reflects how surprised I am to receive “a” from the set of all messages you could have sent.

Classical Information Theory

Shannon: “The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called binary digits, or more briefly bits. Change from the base *a* to base *b* merely requires multiplication by $\log_b a$.”



Following the choice to use bits, we are ready to define a noiseless classical channel: if the sender sends 0 the receiver receives 0, and similarly if the sender sends 1 the receiver receives 1.

Classical Information Theory

Shannon unified different types of information (e.g. written English and electronic signals) by considering them as stochastic processes which produce a discrete sequence of symbols from a discrete set.

Suppose we wish to consider a stochastic process that generates written English messages. Shannon's zeroth order approximation to English takes ever letter (including space) to be equiprobable:

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD

His first-order approximation to English takes letters independently, but with their true frequencies:

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

The second-order approximation to English takes pairs of letters with their true frequencies:

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

Classical Information Theory

If we think of the letters within a message as interacting degrees of freedom in a statistical mechanics theory, then correlations between letters correspond to short range interactions in 1 dimension.

Instead of letters we could also consider alphabets of words, and so on. (“coarse graining”). But eventually, as a first approximation, we consider the “ideal gas limit” with no correlations between symbols.

So we consider a sequence of random variables which are independent and identically distributed (iid) e.g.

aababbaaaab · · ·

Where in each position *a* occurs with probability *p*, and *b* occurs with probability 1 – *p*.

Again, this simplification is motivated by the practical problem Shannon wanted to solve. If you build a communication system, you may take advantage of some correlations to compress messages (e.g. he mentions abbreviations in telegrams). But beyond a certain point this becomes impractical, and your system needs to handle messages that are effectively iid.

Classical Information Theory

Now suppose we wish to transmit messages (iid sequences) from an alphabet with 4 letters, {a,b,c,d}.

If our channel is capable of sending 4 bits per second, we could encode each message with two bits, and send two letters per second.

But suppose the distribution of letter frequencies is non-uniform. If our messages mostly consist of a and b, and only rarely of c and d, then it should be possible to send information through our channel more quickly.

We can think of the receiver, who knows the distribution of letters but not what the next symbol will be, as receiving answers to a series of yes/no (binary) questions that allow them to ascertain the next letter.

From this, the rate at which the receiver receives information is related to the expected number of questions it needs to ask at each step in order to determine the next letter.

Classical Information Theory

Suppose the frequencies for our alphabet with 4 letters, {a,b,c,d} are as follows:

$$p_a = 1/2 \quad , \quad p_b = 1/4 \quad , \quad p_c = 1/8 \quad , \quad p_d = 1/8$$

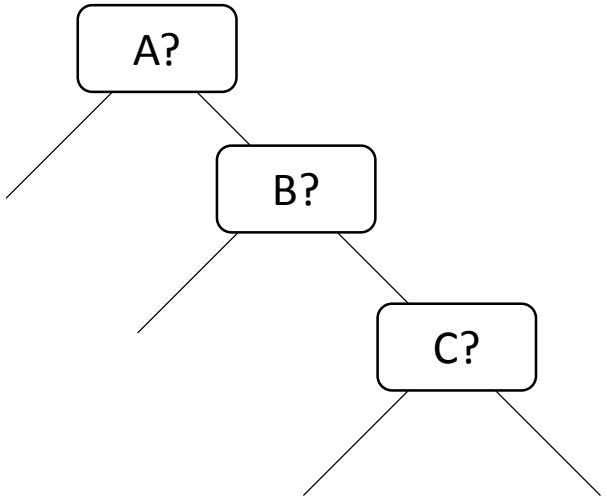
Knowing this distribution, what series of questions would you ask to determine the next letter?
(with the goal of minimizing the expected number of questions)

Classical Information Theory

Suppose the frequencies for our alphabet with 4 letters, {a,b,c,d} are as follows:

$$p_a = 1/2 \quad , \quad p_b = 1/4 \quad , \quad p_c = 1/8 \quad , \quad p_d = 1/8$$

Knowing this distribution, what series of questions would you ask to determine the next letter? (with the goal of minimizing the expected number of questions)



Classical Information Theory

The expected number of questions needed to determine the next letter is

$$\frac{1}{2}\cdot 1+\frac{1}{4}\cdot 2+\frac{2}{8}\cdot 3=1.75 \ \text{ bits}$$

Repeating this exercise for a general distribution $\{p_1,...,p_m\}$, the expected number of questions is

$$S=\sum_{k=1}^{m} p_k \log\left(\frac{1}{p_k}\right)=-\sum_{k=1}^{m} p_k \log p_k$$

Which is called the **Shannon entropy** of the distribution $\{p_1,...,p_m\}$. Shannon calls $\log(1/p_k)$ the information associated with message *k* , so the entropy is the expected information (per message).

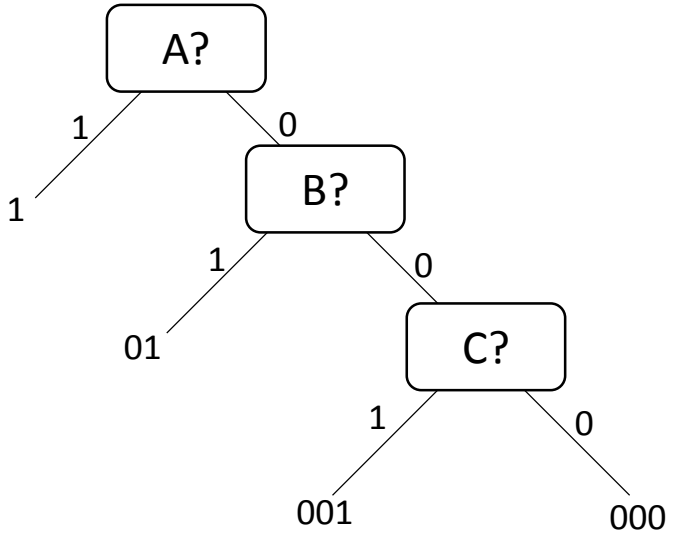
Classical Information Theory

So the expected information sent with each symbol of our message is:

$$\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{2}{8} \cdot 3 = 1.75 \ \text{bits}$$

To take advantage of this being less than 2 bits, we can use a variable length encoding scheme. Shannon proposed a scheme along with Fano, a professor at MIT, but the optimal variable-length coding scheme was found a few years later by a student in Fano's class named Huffman (which is still used in JPEG, etc).

These Huffman codes are optimal for *lossless* compression. For noiseless channels and in the asymptotic limit of a large number of messages, they come arbitrarily close to achieving the rate determined by the Shannon entropy.



Classical Information Theory

Another way of understanding the Shannon entropy is by counting the number of "typical" messages. Returning to the sequence in which a occurs with probability p and b with probability 1 – p,

aababbaaaab · · ·

If the sequence has length N, we expect N p occurrences of a and N (1- p) occurrences of b. The number of ways this can happen is:

$$\frac{N!}{(Np)!(N(1-p))!} \sim \frac{N^N}{(pN)^{pN}(N(1-p))^{N(1-p)}} = \frac{1}{p^{pN}(1-p)^{N(1-p)}} = 2^{NS}$$

Where S is the Shannon entropy: $S = -p\log p - (1-p)\log(1-p)$

The total number of typical messages of length N is 2^{NS}. This means that the information gained by observing one such sequence is NS bits.

Classical Information Theory

The binary entropy function corresponding to the biased coin distribution {p, 1-p} has the form

$$S = -p \log p - (1-p) \log(1-p)$$

Using Shannon's original plot of this function (note his H is our S), we see that it is peaked at p = ½. The maximum entropy (and hence the messages with the most information content, which are most difficult to predict) are drawn from the uniform distribution.

The messages that are easiest to predict are ones with p very close to 1 or 0, which only transmit information very slowly.

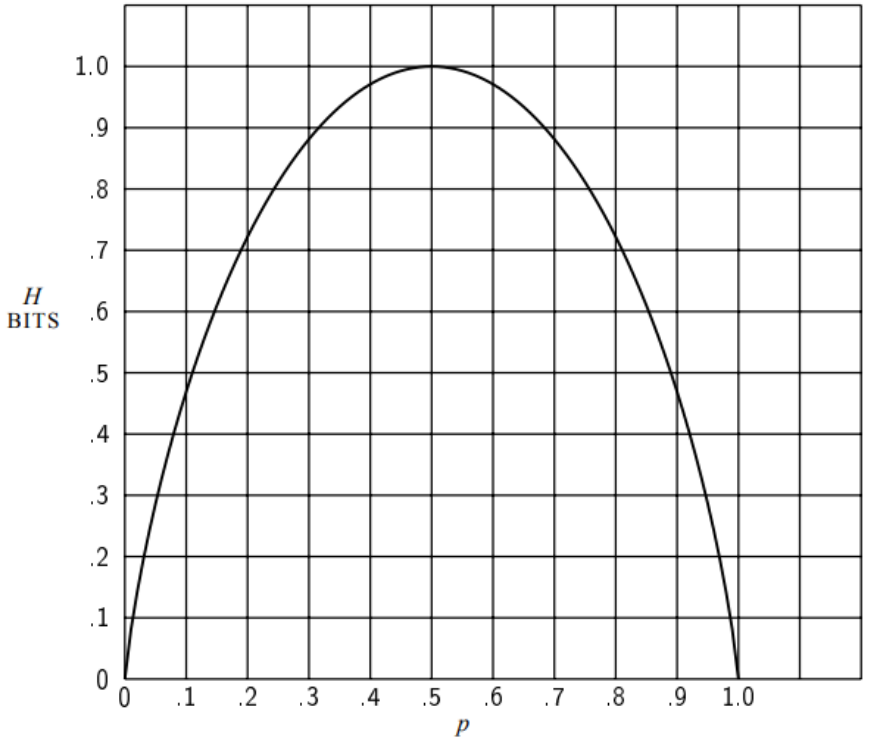


Fig. 7—Entropy in the case of two possibilities with probabilities p and $(1-p)$.

Classical Information Theory

The idea that the Shannon entropy counts the number of "typical" messages also generalizes to alphabets with more than two symbols. Suppose the letters have probabilities $\{p_1, ..., p_m\}$.

The number of ways to see Np_k occurrences of each event k in a sequence of length N is

$$\frac{N!}{\prod_{k=1}^{m}(p_k N)!} \sim \frac{N^N}{\prod_{k=1}^{m}(p_k N)^{p_k N}} = 2^{NS}$$

Where $S = -\sum_{k=1}^{m} p_k \log p_k$ is the Shannon entropy of the distribution. Since all such typical sequences are equally likely, we note that they each occur with probability 2^{-NS}.

Classical Information Theory

The extremes of the binary entropy function are also representative of the general behavior of the entropy. In the case that the probability distribution only has support on one event, we have

$$S = -\sum_k p_k \log p_k = 1 \cdot \log 1 = 0$$

And the maximum occurs when the distribution is uniform:

$$S = -\sum_k p_k \log p_k = -\sum_{k=1}^{m} (1/m) \cdot \log(1/m) = \log m$$

Finally, as Shannon puts it: “Any change toward equalization of the probabilities increases the entropy.” Therefore a low entropy corresponds to a relatively concentrated distribution, and a large entropy corresponds to a more delocalized distribution that is closer to uniform.

Classical Information Theory

For general communication scenarios, we need two parties Alice and Bob. Alice sends a random variable A, and Bob receives a random variable B. To communicate successfully A and B must be correlated.

The joint entropy S(A,B) is the entropy of the joint distribution:

$$S(A,B) = -\sum_{a,b} p(a,b) \log p(a,b)$$

The information needed to predict the joint event (a,b) is no larger than the sum of the information needed to predict a, and that needed to predict b.

$$S(A,B) \leq S(A) + S(B)$$

With equality holding iff A and B are independent ($p(a,b) = p(a)p(b)$). The gap between the two sides of this inequality is called the mutual information:

$$I(A:B) = S(A) + S(B) - S(A,B)$$

Classical Information Theory

The quantity $p(a,b)$ is the probability that Alice sends *a* and Bob hears *b*. If Bob hears a particular b, then his estimate of the probability that Alice sent *a* is the conditional probability:

$$p(a|b) = \frac{p(a,b)}{p(b)}$$

The entropy of this conditional distribution represents the information contained in Alice's signal even after Bob has observed B = b,

$$S(A|B=b) = -\sum_a p(a|b)\log p(a|b)$$

The **conditional entropy** is the expected entropy in Alice's signal after Bob observes B,

$$S(A|B) = \sum_{a,b} p(b)S(A|B=b) = S(A,B) - S(B)$$

Classical Information Theory

So far we have introduced the **joint entropy** (the Shannon entropy of the joint distribution), the **conditional entropy** (the expected entropy of the conditional distribution), and the **mutual information**.

We need one more definition in order to complete our cast of characters: the **relative entropy**, which measures the statistical difference between distributions.

Suppose we predict that the probability distribution describing a set of events is *q*, but in testing our hypothesis we observe data from the true distribution *p*. How do we distinguish q and p?

After making N observations, we will observe event *k* to occur a number of times that is nearly Np_k . But since we believe the distribution is Q, the probability we assign to the observed sequence is:

$$\mathcal{P} = \prod_{k=1}^{m} q_k^{(Np_k)} \cdot \frac{N!}{\prod_{k=1}^{m} (p_k N)!}$$

Classical Information Theory

After making N observations, we will observe event *k* to occur a number of times that is nearly Np_k .
But since we believe the distribution is Q, the probability we assign to the observed sequence is:

$$\mathcal{P} = \prod_{k=1}^{m} q_k^{(Np_k)} \cdot \frac{N!}{\prod_{k=1}^{m}(p_k N)!}$$

We've already analyzed the multinomial factor using Stirling's approximation,

$$\frac{N!}{\prod_{k=1}^{m}(p_k N)!} \sim 2^{-N\sum_{k=1}^{m} p_k \log p_k}$$

Therefore $$\mathcal{P} \sim \prod_{k=1}^{m} 2^{\log q_k^{(Np_k)}} \cdot 2^{-N\sum_{k=1}^{m} p_k \log p_k} = 2^{-N\sum_k p_k(\log p_k - \log q_k)} = 2^{-NS(P\|Q)}$$

Classical Information Theory

This leads to the definition of the **relative entropy**:

$$S(P\|Q) = \sum_{k=1}^{m} p_k \left(\log p_k - \log q_k\right)$$

Which has the operational meaning that we will surely reject our hypothesis of the distribution Q once

$$NS(P\|Q) \gg 1$$

Since at this point the probability of observing the sequence becomes exponentially small if we insist on using the prediction Q.

An immediate property following from the derivation is $S(P\|Q) \geq 0$, with equality iff P = Q.

The relative entropy is useful for distinguishing distributions, but note that it is not a proper distance metric because it is not symmetric; the symmetry is broken by the choice of P as the true distribution.

Classical Information Theory

Suppose we use the relative entropy to test a hypothesis that two random variables are independent. The events are drawn from a joint distribution $p(a,b)$, and our prediction is that they are independent:

$$q(a,b) = p(a)p(b)$$

The relative entropy of these distributions is

$$S(P\|Q) = S(A) + S(B) - S(A,B) = I(A:B) \geq 0$$

Therefore the mutual information is the relative entropy between a joint distribution (which is potentially correlated) and the uncorrelated distribution given by the product of the marginals.

Classical Information Theory

To summarize, Shannon's conceptual contribution was to recognize that information is reflected in the choice of a message from a large set of possible messages. When the sequence of messages is iid, the expected number of bits per message (and hence the optimal compression rate) is the **Shannon entropy**:

$$S = -\sum_{k=1}^{m} p_k \log p_k$$

For a joint system of random variables A, B, the **conditional entropy** measures the expected entropy remaining in variable A after variable B is observed:

$$S(A|B) = S(A,B) - S(B)$$

The **relative entropy** between two distributions P, Q is inversely proportional to the number of observations needed to reject the hypothesis Q when the true distribution is P.

$$S(P\|Q) = \sum_{k=1}^{m} p_k \left(\log p_k - \log q_k\right)$$

The **mutual information** I(A:B) between A and B is the relative entropy between the true joint distribution, and the hypothesis that A and B are uncorrelated. The nonnegativity of I(A:B) expresses **subadditivity** of entropy.

$$I(A:B) = S(A) + S(B) - S(A,B) \geq 0$$