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.

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

Classical Information Theory

So far we have given a reasonable, but non-rigorous argument for the nonnegativity of relative entropy.

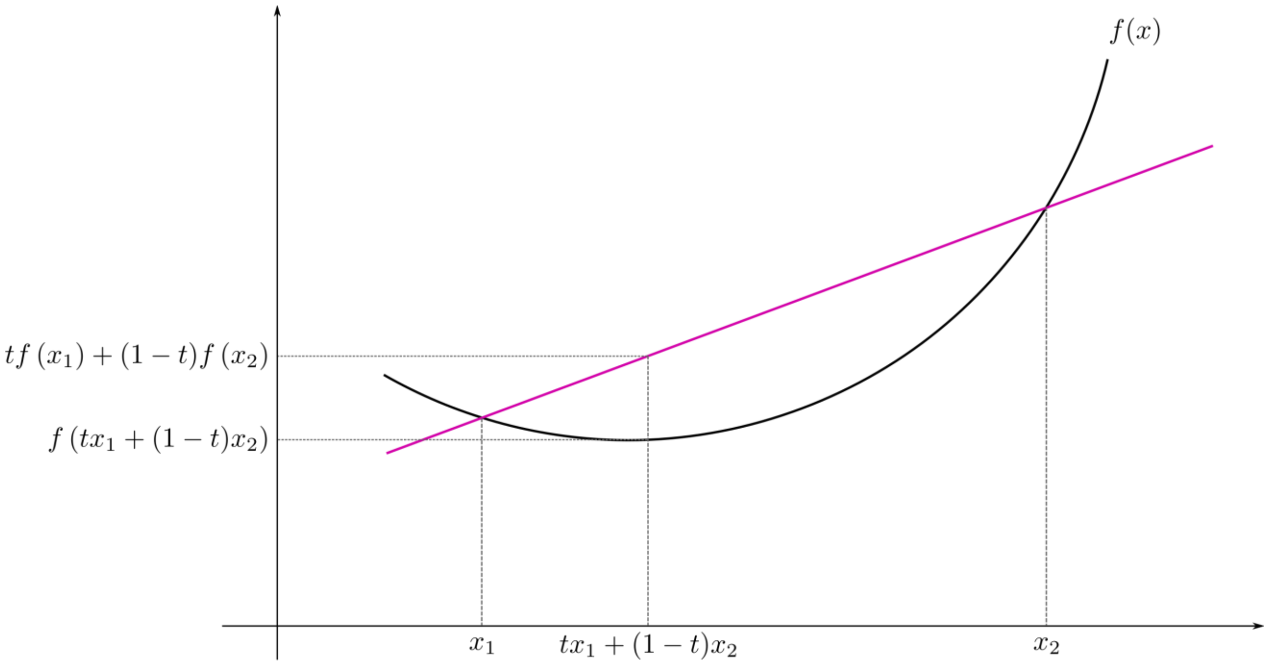
$$S(P\|Q) = \sum_{k=1}^{m} p_k \left(\log p_k - \log q_k\right) \geq 0$$

Now we will give a proper proof of this fact, and in the process introduce more sophisticated tools for proving more general entropy inequalities using properties of convex functions.

A real-valued convex function over a real domain is one which be visualized as an "upward facing" curve. More formally, the function is convex if the line segment between two points

$$(x_1, f(x_1)), (x_2, f(x_2))$$

will lie above the graph of the function.

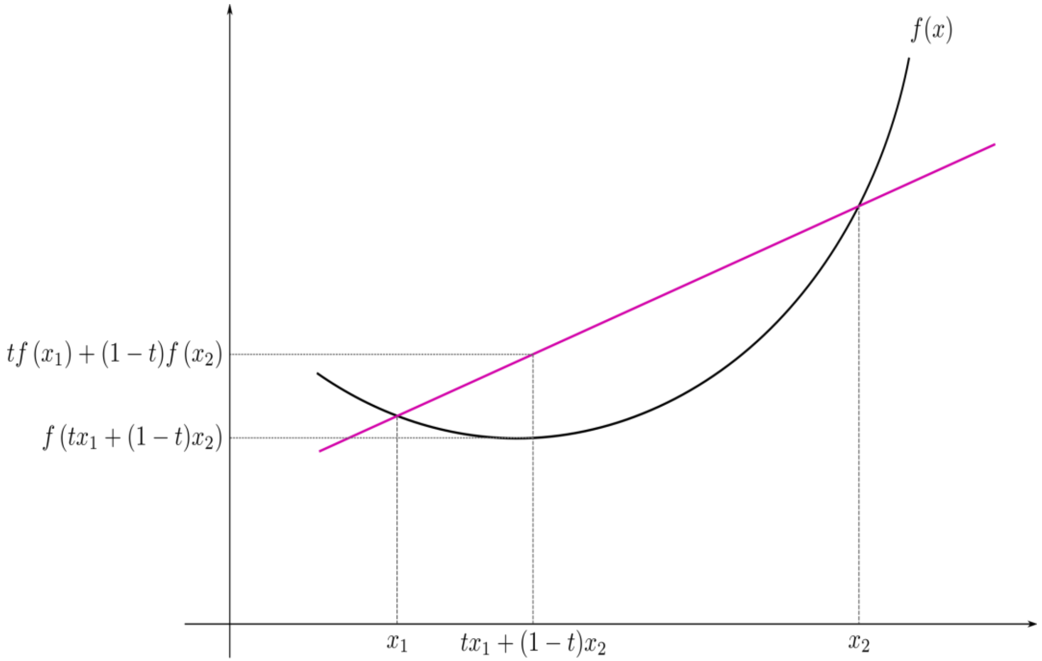


Classical Information Theory

A real-valued convex function *f* over a real domain $X \subseteq \mathbb{R}$ is defined as one with the property

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

for all $x_1, x_2 \in X$ and all $t \in [0,1]$.



More generally, a convex function can be defined over any domain that is a convex set, which is a set closed under convex combinations:

$$x_1, x_2 \in X \implies tx_1 + (1-t)x_2 \in X$$

We have previously seen that the set of probability distributions is closed under convex combinations. Is the same true for density matrices, is $tA + (1-t)B$ a density matrix if *A* and *B* are, and $t \in [0,1]$?

Classical Information Theory

A convex function is a real-valued function $f : X \to \mathbb{R}$, where *X* is a convex set, which satisfies

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2) \quad , \quad \forall\, x_1, x_2 \in X \;,\; t \in [0,1]$$

Theorem (Jensen's inequality): If x is a random variable and f is convex, then

$$f(\mathbb{E}(x)) \leq \mathbb{E}(f(x))$$

Proof: $$f(\mathbb{E}(x)) = f\left(\sum_x x p_x\right) \leq \sum_x p_x f(x) = \mathbb{E}(f(x))$$

Jensen's inequality is powerful because the LHS is often simpler to calculate than the RHS.

Classical Information Theory

Now we will apply Jensen's inequality to show that the relative entropy is nonnegative.

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

Let X be a random variable which takes value $X = p_k/q_k$ with probability q_k, then

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

Now since $f(x) = x \log x$ is a convex function (as can be checked from the fact that its 2nd derivative is nonnegative), we can apply Jensen's inequality

$$S(P\|Q) = \mathbb{E}(X \log X) \geq \mathbb{E}(X) \log \mathbb{E}(X) = 0$$

With this, we have properly proven the subadditivity of Shannon entropy $(I(A:B) \geq 0)$.

Classical Information Theory

Is Shannon entropy a convex function?

The first step is to see that it does have a convex set as a domain, since the set of probability distributions is closed under convex combinations.

However, the function – x log x is not convex, rather it is concave. And the sum of concave functions is concave, and so the Shannon entropy is a concave function. This means it obeys a reversed version of Jensen’s inequality:

$$S(tp_1 + (1-t)p_2) \geq tS(p_1) + (1-t)S(p_2)$$

This corresponds to the fact that the entropy of a mixture is no smaller than the mixture of the entropies.

Classical Information Theory

The relative entropy is not only useful for distinguishing probability distributions, it is also useful for deriving further inequalities (as seen in the proof that $I(A:B) \geq 0$).

$$S(P\|Q) = \sum_{k=1}^{m} p_k \left(\log p_k - \log q_k\right) = \sum_{k=1}^{m} p_k \log\left(\frac{p_k}{q_k}\right) \geq 0$$

$$\implies I(A:B) = \sum_{a,b}^{m} p(a,b) \log\left(\frac{p(a,b)}{p(a)p(b)}\right) \geq 0$$

Suppose we wish to use the relative entropy to distinguish a hypothesis Q(x,y) from the true distribution P(x,y), but our observations only allow us to observe x. In this case we compare P(x) to the marginal Q(x). It is harder to disprove our initial hypothesis if we only have access to x, and so we expect

$$S(P_{X,Y}\|Q_{X,Y}) \geq S(P_X\|Q_X)$$

This property is called monotonicity of relative entropy, and it means that restricting our attention to a subsystem X can never increase the distinguishability of the distributions P and Q.

Classical Information Theory

The proof of the monotonicity of relative entropy is again relatively straightforward.

$$
\begin{aligned}
S(P_{X,Y}\|Q_{X,Y}) - S(P_X\|Q_X) &= \sum_{xy} p(x,y)\left(\log\left(\frac{p(x,y)}{q(x,y)}\right) - \log\left(\frac{p(x)}{q(x)}\right)\right) \\
&= \sum_{xy} p(x,y)\log\left(\frac{p(x,y)/p(x)}{q(x,y)/q(x)}\right) \\
&= \sum_{x} p(x)\sum_{y}\frac{p(x,y)}{p(x)}\log\left(\frac{p(x,y)/p(x)}{q(x,y)/q(x)}\right) \\
&= \sum_{x} p(x)\sum_{y} S(P(y|x)\|Q(y|x)) \\
&\geq 0
\end{aligned}
$$

Classical Information Theory

This **montonicity of relative entropy** also implies **monotonicity of mutual information**. For this property we consider a tripartite system P(x,y,z), and a hypothesis Q that forgets correlations between X and Y,Z:

$$Q(x,y,z) = P(x)P(y,z)$$

From the monotonicity of relative entropy we know that

$$S(P_{X,Y,Z} \| Q_{X,Y,Z}) \geq S(P_{X,Y} \| Q_{X,Y})$$

Combining this with the definition of mutual information yields the monotonicity of mutual information:

$$I(X:YZ) \geq I(X:Y)$$

This inequality has a simple interpretation: the information we gain about X by observing Y and Z is at least as much as the information we gain about X by observing Y alone.

Classical Information Theory

We can also express the monotonicity of mutual information in terms of entropies, where the inequality is known as strong subadditivity.

$$I(X:Y) = S_X + S_Y - S_{XY}$$

$$I(X:YZ) = S_X + S_{YZ} - S_{XYZ}$$

So the monotonicity of mutual information $I(X:YZ) \geq I(X:Y)$ is equivalent to:

$$S_{XY} + S_{YZ} \geq S_Y + S_{XYZ}$$

Which is the strong subadditivity of Shannon entropy.

Classical Information Theory

The monotonicity of mutual information $I(X:YZ) \geq I(X:Y)$ can be further quantified in terms of a quantity called the conditional mutual information. Let X, Y, and Z be random variables and define the CMI:

$$I(X:Y|Z) = \sum_z p_z \sum_{x,y} p(x,y|z) \log \left(\frac{p(x,y|z)}{p(x|z)p(y|z)} \right)$$

The point of this definition is for I(X:Y|Z) to quantify any additional correlation between X and Y when Z is in some sense known to both. By expanding the definition in terms of conditional entropies one can show:

$$I(X:Y|Z) = I(X:YZ) - I(X:Z)$$

Therefore the CMI quantifies the correlation between X and Y,Z that is not just due to correlation between X and Z. Because of this relation we have another equivalent expression to strong subadditivity:

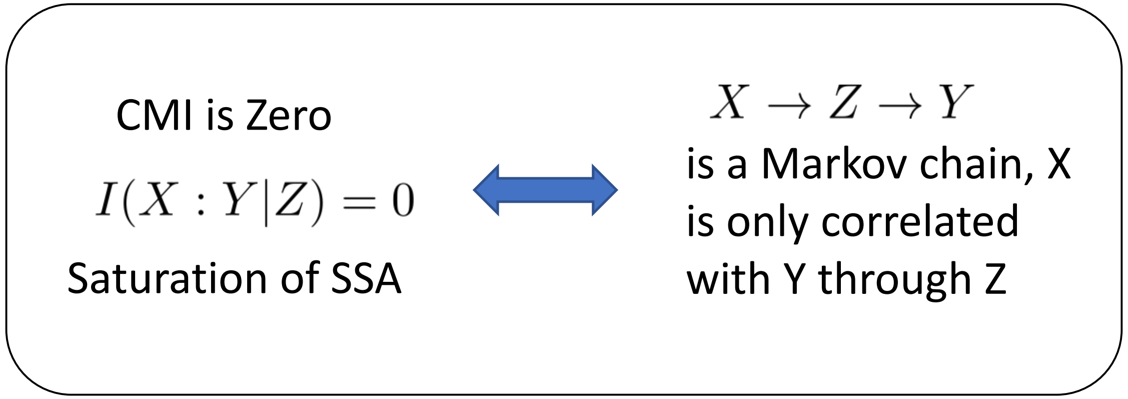
$$I(X:Y|Z) \geq 0$$

Classical Information Theory

The saturation of strong subadditivity $I(X:Y|Z)=0$ occurs iff $X\to Z\to Y$ forms a Markov chain:

$$p(x,y|z)=p(x|z)p(y|z)$$

In other words, $I(X:Y|Z)=0$ expresses the statement that the correlations between X and Y only come through Z. The connection to “Markov chain” in the context of random walks is to think about X,Y,Z as consecutive time steps of the random walk. But here we think of it as a static property of a distribution p(x,y,z).



CMI is Zero

$I(X:Y|Z)=0$

Saturation of SSA

$X\to Z\to Y$

is a Markov chain, X is only correlated with Y through Z

The conditional mutual information is a sophisticated and flexible tool for factorizing probability distributions. X and Y may be highly correlated, but still factorize through Z.

Classical Information Theory

Our first application of this notion of a Markov chain is to understand the behavior of the mutual information under the action of a stochastic map.

To motivate this example, let X be a random variable representing a message source, and Y a receiver. The ability to transmit information between X and Y depends on a large mutual information I(X:Y).

Now suppose Y transmits the message to a new receiver Z. Intuitively, this second transmission can only serve to degrade the signal further, and so we expect:

$$I(X:Y) \geq I(X:Z)$$

Indeed this is the case. The way to formalize it is to see that $X \to Y \to Z$ is a Markov chain, and so

$$I(X:Z|Y) = 0 \implies I(X:Y) = I(X:YZ) \geq I(X:Z)$$

Classical Information Theory

We can also make the previous observation more symmetric by applying a classical channel (i.e. a stochastic map i.e. a conditional probability distribution) to both X and Y to obtain new variables X', Y' according to

$$p(x'|x), p(y'|y)$$

Appling the previous inequality twice and noting the mutual information is symmetric in its arguments yields

$$I(X:Y) \geq I(X':Y')$$

This powerful relation is known as the **data processing inequality**. It says that sending two random variables through independent quantum channels cannot increase the correlation between them.

Classical Information Theory

The monotonicity of the mutual information under stochastic maps also applies more generally to relative entropy. The version of monotonicity we discussed so far was:

$$S(P_{X,Y}\|Q_{X,Y}) \geq S(P_X\|Q_X)$$

But this is really a statement about monotonicity under taking the partial trace (i.e. the marginal distribution). More generally, the relative entropy is nonincreasing under the action of any stochastic map.

Just as before, we model a classical (noise) channel according to a conditional probability distribution:

$$\mathcal{N}: \quad p(z|y).$$

The general form of monotonicity of relative entropy, which is also the general data processing inequality, is:

$$S(P\|Q) \geq S(\mathcal{N}(P)\|\mathcal{N}(Q))$$

One intuition for this is the idea that no amount of post processing in the form of the channel N can make the two distributions more distinguishable than they previously were (in an information theoretic sense).

Classical Information Theory

Recall that we defined the total variation distance between probability distributions p, q as:

$$d_{\mathrm{tvd}}(p,q) = \frac{1}{2}\sum_{x\in\Omega}|p_x - q_x| = \frac{1}{2}\|p-q\|_1$$

Where in the second expression p,q are unit vectors in the 1-norm (or equivalently diagonal density matrices) and the norm is just the usual 1-norm of vectors.

The reason we like this notion of “trace” distance is that it can be used to bound differences of expectations:

$$|\langle A\rangle_p - \langle A\rangle_q| = \left|\sum_x A_x(p_x - q_x)\right| \leq A_{\max}\sum_x |p_x - q_x| = 2A_{max}\|p-q\|_1$$

So far we have been focused on using the relative entropy to distinguish probability distributions, so a natural question is how the relative entropy relates to trace distance. On one side the answer is Pinsker’s inequality:

$$S(P\|Q) \geq \frac{1}{2\ln 2}\|p-q\|_1^2$$

Classical Information Theory

Pinsker's inequality relates the relative entropy to the trace distance:

$$S(P\|Q) \geq \frac{1}{2\ln 2}\|p-q\|_1^2$$

With Pinsker's inequality we can also relate the mutual information to another standard notion of correlation.

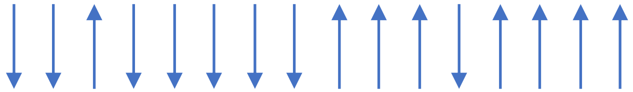
The covariance of two random variables X,Y is defined to be: $\mathrm{Cov}(X,Y) = \langle XY\rangle - \langle X\rangle\langle Y\rangle$

If the variables are independent, then p(x,y) = p(x)p(y) for all events x,y, and $\mathrm{Cov}(X,Y) = 0$.

In field theory or many-body physics the covariance is called a "correlation function", and knowledge of these covariances suffices to make experimental predictions.

Classical Information Theory

The following example illustrates the notion of a correlation function. Consider a statistical model of microscopic magnetic spins $s_i = \pm 1$ that want to align, but also fluctuate due to being at non-zero temperature.



Spins that are nearby have a high propensity to align, while as the distance $|i - j|$ between two spins s_i, s_j increases the probability that they remain aligned decays to zero.

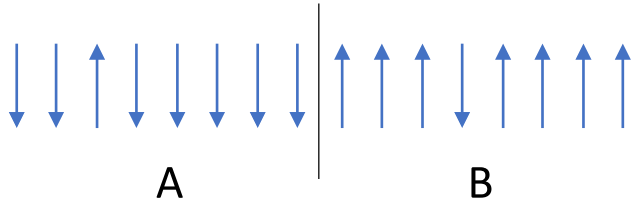
In this case, if there is no preferred direction between up and down then we have:

$$\langle S_i \rangle = \langle S_j \rangle = 0$$

And so the covariance between these spins is $\mathrm{Cov}(s_i, s_j) = \langle s_i s_j \rangle$ which is a value that ranges in +/-1, representing strong correlation (+1) to no correlation (0) to strong anti-correlation (-1).

Classical Information Theory

Now suppose we break up the joint distribution $p(s_1, ..., s_n)$ by defining two regions A and B:



If spin i is in region A, and spin j is in region B, then the covariance can be expressed as:

$$\mathrm{Cov}(s_i, s_j) = \langle s_i s_j \left(p_{AB} - p_A \otimes p_B\right) \rangle$$

This has the form of the difference of between an expectation value for two distributions. Therefore generalizing beyond the spins i and j, we can upper bound all the covariances between A and B in terms of a trace distance:

$$\Delta = \frac{1}{2} \| p_{AB} - p_A \otimes p_B \|_1$$

Classical Information Theory

We can now apply Pinsker's inequality to upper bound this 1-norm distance in terms of relative entropy:

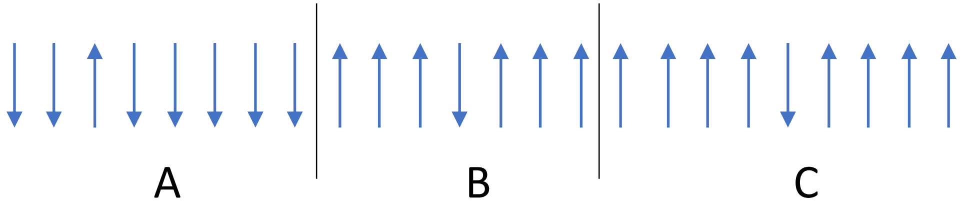
$$S(P_{AB} \| P_A P_B) \geq \frac{1}{2 \ln 2} \Delta^2$$

But this relative entropy is the definition of the mutual information, $I(A:B) = S(P_{AB} \| P_A P_B)$, so

$$I(A:B) \geq \frac{1}{4 \ln 2} \| p_{AB} - p_A \otimes p_B \|_1^2$$

Therefore an upper bound on mutual information between subregions is an upper bound on all possible covariances between pairs of observables in subsystem A and subsystem B.

However, knowing the mutual information is excessive for a specific problem involving just one spin in each region. And often in practice the bound above may be very coarse. We will obtain tighter relations later by considering conditional mutual information I(A:C|B).



Classical Information Theory

We used **Jensen's inequality** for **convex functions** to properly show the **nonnegativity of relative entropy**.

We then demonstrated the **monotonicity of the relative entropy** under partial trace, which is equivalently known as the **strong subadditivity of Shannon entropy**.

Next we interpreted the **monotonicity of the mutual information** as a **data processing inequality**: the information between X and Y cannot increase if Y is further post processed (by a classical channel) into Z.

The **monotonicity of the relative entropy** under arbitrary stochastic maps is the most general **data processing inequality**. It states that stochastic maps cannot make two distributions more distinguishable.

Finally we defined the covariance of random variables, related it to correlation functions in physics, and used **Pinsker's inequality** to show that the mutual information upper bounds correlation functions.