21 Aug 2024

Plan

1. Summary of the role of H[X], H[Y|X] and I[X:Y] in the growth of typical sequences.
2. Properties of entropic quantities, relative entropy
3. The Loomis-Whitney inequality
4. Application of these classical ineqalities to derive ineqalities about Von Neumann entropy

Reference: For applications of entropy in combinatorics, please see the following:

David Galvin: Three tutorial lectures on entropy and counting, https://arxiv.org/abs/1406.7872

So far ...

We have understood H[X], H[Y|X] and I[X:Y] in terms of the growth of typical sets. In particular, for visualizing typical sets arising out k samples drawn from the joint distribution {p(a,b)} on the set of elements of A x B, we have a bipartite graph.

*******   THE BIPARTITE GRAPH   *******

On the left are the typical sequences drawn according to the marginal distribution of the random variable X.
There are about $2^{k H[X]}$ such typical sequences.

On the right are the typical sequences drawn according to the marginal distribution of Y.
There are about $2^{k H[Y]}$ such typical sequences.

We connect a sequence x-bar to a sequence y-bar by an edge if (x-bar, y-bar) are jointly typical.

Each typical x-bar has about $2^{k H[Y|X]}$ edges incident on it. Each typical y-bar has about $2^{k H[X|Y]}$ edges incident on it.

So the density of the graph (edges present/edge
the maximum is

$$2^{k H[X]} * 2^{k H[Y|X]} / 2^{k H[X]} * 2^{k H[Y]}$$

=

2^{-k I[X:Y]}

The number of codewords we can pack is the inverse of the
For channel coding, we are given {p(b|a)}. We then adjust
the distribution of X, so that the edge density is small as
possible. For this reason, the capacity is given by an expression
of the form

max_X I[X:Y]

****

Properties of entropic quantities

H[X] = E[log 1/p(X)]

By applying Gibbs inequalty taking P to be the distribution
of X and Q to be the uniform distribution on the support of
X. We see

H[X] <= log n, where n is the number of elements in the
support of X.

H[Y|X] = sum_a p(a) sum_b p(b|a) log 1/p(b|a)
= sum_a p(a,b) log 1/p(b|a)
= E[log 1/p(b|a)]

H[XY] = E[log 1/p(X,Y)]
= E[log 1/p(X)] + E[log 1/p(Y|X)]
= H[X] + H[Y|X]

I[X:Y] = E[log p(X,Y)/p(X)q(Y)]
= D({p(a,b)} || {p(a) q(b)})
>= 0 (by Gibbs inequality)

= sum_a p(a) D(P_{Y|X=a} || Q)

measure how for the distribution is from the product distribution

In particular, from I[X:Y] >= 0, we conclude that

    H[Y|X] <= H[Y]

(conditioning reduces entropy)

Another way of saying this is that the entropy function is convave.

Three application

Application 1: The Loomis-Whitney inequality

Consider N points in R^3.

Suppose we project these points onto the two coordinate planes (along directions, x, y, z), and get N_x, N_y, N_z points.

Loomis-Whitney: $N_x N_y N_z >= N^2$

We formalize the following intuition. We pick on of the N points uniformly at random. Say the point is (X,Y,Z).

$H[(X,Y,Z)] = \log N$.
$H[(X,Y)] <= \log N_z$
$H[(Y,Z)] <= \log N_x$
$H[(X,Z)] <= \log N_y$

But every piece of information in (X,Y,Z) is available from two sources. So, $\log N_x + \log N_y + \log N_z >= 2 \log N$.

Application 2: The Shannon entropy of the diagonal of a density matrix is at least the Von Neumann entropy of the original density matrix.

$S(\rho_D) >= S(\rho)$.

Proof.

$\mathrm{diag}(\rho_D) = (p_1, p_2, ..., p_N)^T$
$\mathrm{diag}(\rho) = (\lambda_1, \lambda_2, ..., \lambda_N)^T$

Then, $\mathrm{diag}(\rho_D) = M\, \mathrm{daig}(\rho)$, where M is a doubly stochastic matrix.

But a doubly stochastic matrix is a convex combination of permutation matrics.

$M = \sum_\sigma q(\sigma) M_\sigma$

(The 1 in column i of M_σ is in row $\sigma(i)$.)

Let X be the random variable with distribution $\mathrm{diag}(\rho)$.

Pick a random permutation sigma with probability $q(\sigma)$, and consider the random variable $Z=\sigma(X)$.

$S(\mathrm{diag}(\rho_D)) = H[Z] >= H[Z | \sigma] = H[X] = S(\rho)$.

Application 3: $S(\rho)$ is concave. That is,

$\rho = \alpha \rho_1 + (1-\alpha(\rho2)$
$S(\rho) >= \alpha S(\rho_1) + (1-\alpha) S(\rho_2)$.

Work in the eigen basis of rho. Combine application 2 with classical concavity.

What does relative entropy measure?

We saw that if the I[X:Y] (which is a relative entropy) is small, then the density of edges in our bipartite graph is small.

Theorem: Let X be a random variable taking values in a set A with distribution Q. Consider X-bar = (X_1,X_2, ..., X_k) drawn independently according to Q. Thus, X-bar takes values in A^k. Let F be as subset of A^k, and let P be the average of the average emperical distributions of the strings in F. That is,

P(a) = sum_{x-bar} [Q^k(x-bar)/Q^k(F)] (1/k)N(a|x-bar)

Then,

Q^k(F) <= 2^{-k D(P||Q)}

[Intuition, if P differs from Q too much, the F will have to be tiny.]

Prof.
Let P_F be the distribution on A^k given by Q^k(x-bar)/Q^k(F).
Then,

Q^k(F) = 2^{-D(P_F||Q^k)}

Exercise: D(P_F||Q^k) >= k D(P || Q).

[
D(P_F || Q^k) = -H[P'] - sum_{a-bar} P_F(a-bar) log 1/Q^k(a-bar)

First term: Let X-bar be distributed according to P'. Then
H[P'] = H[X-bar] = k H[X_J | J] <= k H[X_J] = k H(P).
So, -H[P'] >= -k H(P)
Second term:
- sum_{a-bar} P_F(a-bar) log 1/Q^k(a-bar)
= - k sum_a P(a) log 1/Q(a)

So, D(P_F || Q^k) >= k D(P||Q).
]

In particular, if the actual distribution is Q, then the probability that the samples drawn from it will look P-typical with probability at most 2^{-k D(P || Q)}.

This is one of the important reasons why D(P || Q) appears in the study of various statistical problems.

[Postscript: In these three lectures, we covered only half of what we had originally intended. The quantum part, perhaps the main reason the audience showed up, got left out. I am very sorry about that. However, what I planned to present is a subset of what is there in Witten's notes.

Edward Witten, A Mini-Introduction To Information Theory, https://arxiv.org/abs/1805.11965

I hope the classical information theory we discussed helps in some way while reading these notes. -- Jaikumar]