PHY 4105: Quantum Information Theory
Lecture 3

Anil Shaji
School of Physics, IISER Thiruvananthapuram
(Dated: August 8, 2013)

A. Classical information theory

buen día, niños, hoy vamos a hablar sobre la información clásica
Good morning students, today we are going to talk about classical information

Despite the differences in form, sounds, letters and shape we recognize that the information content in the statements above is the same. We need to quantify this qualitative observation. Obviously the number of words or number of letters in the phrases above is not a very good quantifier of the information content since these measures depend a lot on language used, the cultural context in which the language originated and lot of other sociological factors. So we are forced to make a choice and fix a language with reference to which we will attempt to quantify information. True to spirit of physics we remove the human element as far as possible from the choice of language and ask the question what is the simplest possible language one can think of?

A language with only one alphabet is not useful since you can really only see one thing with it. But with just two alphabets we can say anything. Coupled with a convention regarding block size, we can map any letter in any alphabet to a sequence of '0's and '1's (We are fixing the letters to be 0 and 1). So one naive way of quantifying information would be to count the number of 0's and 1's in any sentence provided the sentence has been appropriately translated into the binary language. We will see soon how we will be able to do better than this and the ability to do this quantification in a precise and acceptable way, it will turn out, depends on the weak law of large numbers that we talked about.

So the fundamental unit of classical information is a bit: A bit has two alternatives, 0 or 1. Information now appears in the form a sequence of 0's and 1's as

$$01000101001001111010 1.$$

Let N be the number of bits in a block or in a standard sentence/phrase we will consider. Let $\mathfrak{N}$ be the number of sequences that one construct with N bits,

$$\mathfrak{N} = (\text{Number of sequences of length } N) = 2^N.$$

We quantify the information that can be stored in N bits as (coming back to our original idea of quantifying in terms of the length of the sequence)

$$I = \log \mathfrak{N} = N.$$

We take the logarithm to make sure that information is additive.

It should be noted that we can as well start with a language with D letters. The basic unit of information will then be a "Dit" with D alternatives. Then

$$\mathfrak{N} = D^N = 2^{N \log D},$$

and

$$I = \begin{cases} \log_D \mathfrak{N} & - \quad N \text{ dits} \\ \log_2 \mathfrak{N} & - \quad N \log_2 D \text{ bits} \end{cases}$$

We see that a certain number of dits is equivalent to a certain number of bits. So there is nothing really we are going to get by using dits instead of bits.

Convention:

$$\log \equiv \log_2, \qquad \ln \equiv \log_e \qquad \text{(nats)}.$$

We have $\log x = \log e \ln x$.

The case we considered above, where we identified the information in a sequence with the length of the sequence in bits, is a special case in which each of the letters in the alphabet is equally probably. There is nothing distinguishing a 0 and 1 in that sense and so irrespective of the value of the bit, the information content is the same and is given by $I/N = 1$. When the symbols that appear in the sequence do not appear with the same probability then the information content varies from letter to letter. Let X be a random variable that represents the letter appearing in each location in a sequence and x be the value of the random variable with $p(x)$ denoting the probability for each value (letter) x. The information content per letter is again I_N/N as $N \to \infty$.

As N gets large, we know from the weak law of large numbers that the sequences that appear are "typical sequences" in which the value x_j appears n_j times with $n_j = Np_j$ $(j = 1, \ldots, D)$. In other words the probability is concentrated on such typical sequences. The probability for any such typical sequence is

$$\begin{aligned} P(\text{typical}) &= p_1^{n_1} \cdots p_D^{n_D} \\ &= p_1^{Np_1} \cdots p_D^{np_D} \\ &= 2^{\log(p_1^{Np_1})} \ldots 2^{\log(p_D^{Np_D})} \\ &= 2^{N(p_1 \log p_1 + \cdots + p_D \log p_D)} = 2^{-NH(\vec{p})}, \end{aligned}$$

where

$$H(\vec{p}) \equiv -\sum_{i=1}^{D} p_i \log p_i.$$

$H(\vec{p})$ is the *Shannon entropy* of the probability distribution $\vec{p} = (p_1, p_2, \ldots p_D)$ and is equivalently denoted as $H(X)$ (Shannon entropy of the random variable X).

The number of typical sequences is

$$\mathfrak{N} = \frac{N!}{n_1! \cdots n_D!} = 2^{NH(\vec{p})},$$

using

$$\begin{aligned}\ln \mathfrak{N} &= \ln N! - \sum_j \ln n_j \\ &\sim N\ln N - N - \sum_j (n_j \ln n_j - n_j) \\ &= N\ln N - \sum_j N p_j \ln N p_j \\ &= N\Big(-\sum_j p_j \ln p_j\Big).\end{aligned}$$

So approximately,

$$\log \mathfrak{N} = N\Big(-\sum_{j=1}^{D} p_j \log p_j\Big) = NH(\vec{p}).$$

Most of the probability is with the typical sequences and each typical sequence has the same probability.

$$I_N = \log \mathfrak{N} = NH(\vec{p}),$$

and the information per letter is

$$\frac{I_N}{N} = H(\vec{p}) = H(X).$$

Let us make things a bit more rigorous. Let us start with N iid random variables, $X_1, \ldots X_N$ (the subscript denotes which letter/trial and not the value of the letter).

$$p(x_1, \ldots, x_N) = p(x_1)\cdots p(x_N).$$

A sequence is ϵ-typical if

$$\Big| -\frac{1}{N}\log p(x_1, \ldots x_N) - H(\vec{p})\Big| \le \epsilon.$$

Equivalently

$$2^{-N(H(\vec{p})+\epsilon)} \le p(x_1, \ldots, x_N) \le 2^{-N(H(\vec{p})-\epsilon)}.$$

We denote the set of all ϵ-typical sequences by $T(N, \epsilon)$.

Now consider the variable

$$S \equiv -\frac{1}{N}\log p(x_1, \ldots, x_N) = \frac{1}{N}\sum_{l=1}^{N} -\log p(x_l).$$

We can think of this variable as the sample mean of $-\log p(x)$. The mean of S is

$$\langle s\rangle = \frac{1}{N}\sum_{l=1}^{N}\Big(-\sum_{x_1,\ldots x_N} p(x_1)\cdots p(x_N)\log p(x_l)\Big) = \frac{1}{N}N(-\sum_{x_j} p(x_j)\log p(x_j)) = H(\vec{p}).$$

and

$$\langle (\Delta s)^2\rangle = \frac{1}{N}\langle (\Delta(-\log p(x)))^2\rangle = \frac{1}{N}\sum_x p(x)\Big(-\log p(x) - H(\vec{p})\Big)^2.$$

The asymptotic equipartition theorem or Typical sequences theorem

(i) For any ϵ, $\delta > 0$, there exists N_0 such that for all $N > N_0$, the probability that a sequence is ϵ-typical is $\geq 1 - \delta$

Proof:

$$p\left(\left| -\frac{1}{N}\log p(x_1,\ldots,x_N) - H(\vec{p})\right| \leq \epsilon\right) = 1 - p\left(\left| -\frac{1}{N}\log p(x_1,\ldots,x_N) - H(\vec{p})\right| > \epsilon\right)$$
$$\leq 1 - \frac{\langle(\Delta(-\log p(x)))^2\rangle}{N\epsilon^2},$$

The inequality coming from the weak law of large numbers of the form,

$$p(|s - \langle x\rangle| > \epsilon) \leq \frac{\langle(\Delta s)^2\rangle}{\epsilon^2} = \frac{\langle(\Delta x)^2\rangle}{N\epsilon^2},$$

We choose

$$N_0 = \frac{\langle(\Delta(-\log p(x)))^2\rangle}{\delta\epsilon^2},$$

so that

$$p\left(\left| -\frac{1}{N}\log p(x_1,\ldots,x_N) - H(\vec{p})\right| \leq \epsilon\right) \geq 1 - \delta.$$

(ii) The number of ϵ-typical sequences, $[T(N,\epsilon)]$, satisfied

$$(1-\delta)2^{N(H(\vec{p})-\epsilon)} \leq [T(N,\epsilon)] \leq 2^{N(H(\vec{p})+\epsilon)}, \quad N \geq N_0.$$