

Information Theory

Lecture Notes for Self Learning

Amir H. Ebrahimnezhad

University of Tehran Department of Physics.

July 20, 2023

Preface

In this lecture notes, I will try to establish a ground work for information theory, then would work of it's philosophy, and applications in mathematics and quantum physics. By reading this leacture note one would essentially learn information theory in it's complete form. This edition is updated until July 20, 2023. For any question or corrections please contact me *Thisis-meamir@outlook.com*.

Contents

1 Introduction to Information Theory

1.1 Some People and Some History

The concept of information, was hidden in Boltzmann view on entropy and statistical mechanics. Boltzmann earns the title of founding father, without doubt. Then it was mathematics that found the potential in the concept, gave it a formalism, which was sufficient enough to be used all over the science. Shannon proposed the basic definition which is still valid. Charles Babbage, Ada Lovelace, and Alan Turing, pioneered the idea of a computer and foundations of its mathematical description. Kurt Godel, though never talked about information explicitly, but it is undisputed by now that it is the key to understanding his Incompleteness Theorem.

In physics the idea was developed within theoreticians who recognized the correspondence between information theory and quantum mechanics. John Von Neumann brought the idea of Von-Neumann entropy. Leo Szilard reduced Maxwell's demon to its very essence, a single-bit decision, bringing out for the first time the equivalence of entropy and information. In a sense the theory of information was all along a key concept for our understandings of nature and physical theories.

1.2 Three Dimensions of Information

§ **Syntactic Information** refers to the quantity and formal structure of messages. It counts signs, circumscribes their repertoire and their mutual relations. Syntax defines rules and restrictions how to compose legitimate sequences of signs, often in a hierarchial fashion. The syntactic aspect of information is relevant for physics and chemistry, for structural sciences such as mathematics and informatics, and for technology related to communication and computation.

§ **Semantic Information:** It is important to keep in mind that meaning is excluded in syntactic information. The meaning of signs, and more generally, communication as an exchange of messages, enter the scene only with semantic information. Semantic is dedicated to relationship between symbols and their meaning. Obviously, to apply, semantic information requires a language to exist. In the context of semantics, the concepts of sender and reciever acquire the additional aspect of understanding, not contained in syntactic information. Only on

this level, it makes sense talking about the *truth*.

§ **Pragmatic Information** finally takes into account that senders may have intentions emitting messages, and receivers may or may not react as desired by these intentions. With pragmatics, norms come into the play, the dichotomy of true and false if complemented by that of good and evil.

It is a fascinating question to be addressed on passing, exactly where in the history of nature a phenomenon to be called "meaning" emerged for the first time. As far as we know till now, symbols, that is, objects encoding for others, did not exist in any sense before the advent of life on Earth. With the genetic code, a systematic relationship arose between one class of long-lived molecules (DNA and RNA), serving as memory, and another class of very reactive but short-lived molecules (proteins), capable of affecting and alternating the environment. In this context, a terminology should be introduced that will become indispensable in the discussion of self reference below: **Symbols**, irrespective of the acquired function in communication, continue forming part of objective reality, and as such can be referred to by other symbols. Symbols whose meaning is itself a symbol can be associated to a distinct, higher layer of language, called *meta-language*, each one forming a *metasymbol* and so on.

1.3 From Boltzmann's Entropy to Shannon's Information

Basically information is different from other physical quantities. Since it's hard to give any definition on how to measure it. Information is a mathematical concept, it pretains to the realm of structural sciences, not to the natural sciences. Yet it is so ubiquitous in physics and in particular, as thermodynamical entropy, becomes indirectly measurable, that the statement "*Information is physical*" appears to be well justified.

Ludwig Boltzmann's ingenious intuition when the concept of entropy is introduced as counting the number N of "complexions" of a physical system, a huge step forward from previous attempts to define it as a thermodynamic quantity. In convincing this term, Boltzmann thought of microscopically distinct states of matter, states that can be distinguished by some microscopic physical quantity, but are not necessarily discernible by macroscopic observation. Anticipating later generalizations, in particular by Shannon, we can define complexions as *distinguishable states of a system*, thus avoiding a reference to physical terms. All the subtleties of this definition are hidden in the word "distinguishable". **It implies a high degree of subjectivity.**

A quantity that allows to interpret Boltzmann's complexions in modern terms is the probability density in phase space. It is remarkable that Boltzmann himself was already aware of this problem and in fact based his original argument on the assumption of a *finite number of discrete states, corresponding, to energies $E_n = n\Delta E/N$.*

Quantifying the repertoire of states a system can assume was the radically new aspect Boltzmann introduced with his entropy. Otherwise the quantity should comply with all general

conditions other thermodynamic potentials fulfill. In particular he required it to be additive. If a system 1 and 2 comprise N_1 and N_2 states, the total number of states is the product, $N = N_1 N_2$. This function replacing products by sums is the logarithm, hence the decision is to define entropy S as the logarithm of this number:

$$S = c \log(N) \quad (1)$$

The total number of distinguishable states is proportional to the volume Ω of the accessible phase space, the quotient being given by the size $\Delta\Omega$ of the smallest discernible phase-space cell:

$$S = c \ln\left(\frac{\Omega}{\Delta\Omega}\right) \quad (2)$$

Boltzmann's original version, assumes that there are no a priori preferences for anyone of the N microstates, attributing the same weight to all of them. In almost all applications of entropy, however the states or symbols counted in measuring an information content already correspond to sets of microstates that are considered as equivalent or indistinguishable. Depending on the size of these sets, the states no longer occur with the same frequency but can be assigned different probabilities. This freedom is indispensable for example in the context of abbreviations and substitution rules that replace sequences of more elementary signs by higher-level symbols.

Callout — For modern version we encounter subsets of the system which happen probabilistically more! thus the notation should not weight all of them equally, since encoding them would make a much efficient explanation.

As mentioned in the callout, an important reason for noticing varied frequency, is apart from the statistical physics, it is the design of codes. In this context, the role of macrostates is played by high-level symbols sets or alphabets. Fixing the number of signals per Morse code, as an example, representing a single Latin character, would require 5 signals per letter of the alphabet. But considering shorter signal packs as letters to we can optimize the code, concerning the total number of signals per average message, by assigning the shortest symbols (a dot or a dash) to the most frequent letter. This is how the Morse code has been devised.

If M is the total number of Microstates, M_j of which belong to the j th distinguishable class, $j = 1, \dots, J$, $\sum_{j=1}^J M_j = M$, then the probabilities:

$$p_j = \frac{M_j}{M} \quad (3)$$

for each class is defined. Of course the sum of the probabilities would be 1. The total number of distinct sequences consisting of M of these symbols is not $M!$, but has to be divided by all the numbers $M_j!$ of combination of microstates that can be formed within each class j ,

$$N = \frac{M!}{\prod_{j=1}^J M_j!} \quad (4)$$

The total corresponding information is then:

$$I_M = c \ln(N) = c \left(\ln(M!) - \sum_{j=1}^J \ln(M_j!) \right) \quad (5)$$

If the number of symbols is large, $M_j \gg 1$ for all classes j , this expression can be simplified using the Stirling's formula:

$$\begin{aligned} I_M &\approx c \left(M \ln(M) - M - \sum_{j=1}^J M_j \ln(M_j) + \sum_{j=1}^J M_j \right) \\ &= cM \left(\ln(M) - \sum_{j=1}^J \frac{M_j}{M} \ln(M_j) \right) \\ &= -cM \sum_{j=1}^J \frac{M_j}{M} \ln \left(\frac{M_j}{M} \right) \end{aligned}$$

comparing this with the definition of probabilities we get for the information per symbol:

$$I_1 = \frac{I_M}{M} = -c \sum_{j=1}^J p_j \ln(p_j) \quad (6)$$

This is the definition proposed by Claude E. Shannon, following a similar expression suggested by J. W. Gibbs in the context of statistical mechanics. Putting $p_j = 1/J$ would reduce the equation to the one Boltzmann proposed. For a deterministic case we have the probabilities distributed, the opposite of the Boltzmann case. In the Boltzmann case all the probabilities were equally likely (maximally mixed case) in the deterministic case, the probability of one class is 1 and others are all 0

$$p_j = \delta_j^{j_0} \quad (7)$$

This way the entropy would be zero, meaning that there's no way to attain more information since the outcome is always the same.

1.4 Sign: Entropy and Negentropy: Actual Versus Potential Information

The concepts of entropy and information have been used indiscriminately, the term "entropy" will be preferred in thermodynamic contexts and "information" where symbol strings and other discrete entities are concerned, always keeping in mind that this distinction is ambiguous and in fact unnecessary. Calling the quantity in eq[6] information, would appear inadequate also on the background of the colloquial use of the word. As the two limiting cases show that I reaches its maximum precisely when all states involved are equally probable, that is, if nothing is known about the system, while the minimum 0 is assumed when the state is exactly fixed. It is somewhat counterintuitive.

Here we define another distinction that helps keeping the meaning clear. What is in fact measured by entropy is the magnitude of state space accessible to the system or the number of signs available to compose a message of. It is therefore adequate to denominate it potential information, also it makes sense to define relative information thus we get the definition:

$$\Delta I_{\text{pot}} = c (\ln(N_{\text{final}}) - \ln(N_{\text{initial}})) \quad (8)$$

As complementary quantity, introduce actual information as measuring what is already known about the system. This would amount to:

$$\Delta I_{\text{act}} = -\Delta I_{\text{pot}} = c (\ln(N_{\text{initial}}) - \ln(N_{\text{final}})) \quad (9)$$

Callout — Actual information stands for what is known, for the knowledge an observer, a measurement, a theory, has on the state of the system. It measures the fraction of its state space that is not excluded by these constraints and is analogous to the concept of negentropy, thus coincides with the familiar sense of information.