An introduction to information theory and entropy

Tom Carter

http://astarte.csustan.edu/~ tom/SFI-CSSS

Complex Systems Summer School

Santa Fe

June, 2011

Contents

Measuring complexity 5
Some probability ideas 9
Basics of information theory 15
Some entropy theory 22
The Gibbs inequality 28
A simple physical example (gases) 36
Shannon's communication theory 47
Application to Biology (genomes) 63
Some other measures 79
Some additional material
Examples using Bayes' Theorem 87
Analog channels 103
A Maximum Entropy Principle 108
Application: Economics I 111
Application: Economics II 117
Application to Physics (lasers) 124
Kullback-Leibler information measure 129
References 135

The quotes ↞

- ◎ Science, wisdom, and counting
- ◎ Being different – or random
- ◎ Surprise, information, and miracles
- ◎ Information (and hope)
- ◎ H (or S) for Entropy
- ◎ Thermodynamics
- ◎ Language, and putting things together
- ◎ Tools

To topics ←

Science, wisdom, and counting

“Science is organized knowledge. Wisdom is organized life.”

- Immanuel Kant

“My own suspicion is that the universe is not only stranger than we suppose, but stranger than we can suppose.”

- John Haldane

“Not everything that can be counted counts, and not everything that counts can be counted.”

- Albert Einstein (1879-1955)

“The laws of probability, so true in general, so fallacious in particular .”

- Edward Gibbon

Measuring complexity ←

- Workers in the field of complexity face a classic problem: how can we tell that the system we are looking at is actually a complex system? (i.e., should we even be studying this system? :-)

  Of course, in practice, we will study the systems that interest us, for whatever reasons, so the problem identified above tends not to be a real problem. On the other hand, having chosen a system to study, we might well ask "How complex is this system?"

  In this more general context, we probably want at least to be able to compare two systems, and be able to say that system A is more complex than system B. Eventually, we probably would like to have some sort of numerical rating scale.

- Various approaches to this task have been proposed, among them:

  1. Human observation and (subjective) rating

  2. Number of parts or distinct elements (what counts as a distinct part?)

  3. Dimension (measured how?)

  4. Number of parameters controlling the system

  5. Minimal description (in which language?)

  6. Information content (how do we define/measure information?)

  7. Minimal generator/constructor (what machines/methods can we use?)

  8. Minimum energy/time to construct (how would evolution count?)

- Most (if not all) of these measures will actually be measures associated with a *model* of a phenomenon. Two observers (of the same phenomenon?) may develop or use very different models, and thus disagree in their assessments of the complexity. For example, in a very simple case, counting the number of parts is likely to depend on the scale at which the phenomenon is viewed (counting atoms is different from counting molecules, cells, organs, etc.).

  We shouldn't expect to be able to come up with a single universal measure of complexity. The best we are likely to have is a measuring system useful by a particular observer, in a particular context, for a particular purpose.

  My first focus will be on measures related to how surprising or unexpected an observation or event is. This approach has been described as *information theory*.

Being different – or random ↞

“The man who follows the crowd will usually get no further than the crowd. The man who walks alone is likely to find himself in places no one has ever been before. Creativity in living is not without its attendant difficulties, for peculiarity breeds contempt. And the unfortunate thing about being ahead of your time is that when people finally realize you were right, they’ll say it was obvious all along. You have two choices in life: You can dissolve into the mainstream, or you can be distinct. To be distinct is to be different. To be different, you must strive to be what no one else but you can be. ”

-Alan Ashley-Pitt

“Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.”

- John von Neumann (1903-1957)

Some probability ideas ←

- At various times in what follows, I may float between two notions of the *probability* of an event happening. The two general notions are:

  1. A *frequentist* version of probability:

     In this version, we assume we have a set of possible events, each of which we assume occurs some number of times. Thus, if there are N distinct possible events $(x_1, x_2, \ldots, x_N)$, no two of which can occur simultaneously, and the events occur with frequencies $(n_1, n_2, \ldots, n_N)$, we say that the probability of event x_i is given by

     $$P(x_i) = \frac{n_i}{\sum_{j=1}^{N} n_j}$$

     This definition has the nice property that

     $$\sum_{i=1}^{N} P(x_i) = 1$$

2. An *observer relative* version of probability:

   In this version, we take a statement of *probability* to be an assertion about the belief that a specific observer has of the occurrence of a specific event.

   Note that in this version of *probability*, it is possible that two different observers may assign different probabilities to the same event.

   Furthermore, the *probability* of an event, for me, is likely to change as I learn more about the event, or the context of the event.

3. In some (possibly many) cases, we may be able to find a reasonable correspondence between these two views of probability. In particular, we may sometimes be able to understand the *observer relative* version of the probability of an event to be an approximation to the *frequentist* version, and to view new knowledge as providing us a better estimate of the relative frequencies.

- I won't go through much, but some probability basics, where a and b are events:
  $P(not\ a) = 1 - P(a)$.
  $P(a\ or\ b) = P(a) + P(b) - P(a$ and $b)$.
  We will often denote $P(a$ and $b)$ by $P(a, b)$. If $P(a, b) = 0$, we say a and b are mutually exclusive.

- Conditional probability:

  $P(a|b)$ is the probability of a, given that we know b. The joint probability of both a and b is given by:

  $$P(a, b) = P(a|b)P(b).$$

  Since $P(a, b) = P(b, a)$, we have Bayes' Theorem:

  $$P(a|b)P(b) = P(b|a)P(a),$$

  or

  $$P(a|b) = \frac{P(b|a)P(a)}{P(b)}.$$

- If two events a and b are such that

$$P(a|b) = P(a),$$

we say that the events a and b are *independent*. Note that from Bayes' Theorem, we will also have that

$$P(b|a) = P(b),$$

and furthermore,

$$P(a, b) = P(a|b)P(b) = P(a)P(b).$$

This last equation is often taken as the definition of *independence*.

- We have in essence begun here the development of a mathematized methodology for drawing inferences about the world from uncertain knowledge. We could say that our observation of the coin showing heads gives us *information* about the world. We will develop a formal mathematical definition of the *information* content of an event which occurs with a certain probability.

Surprise, information, and miracles ↞

“The opposite of a correct statement is a false statement. The opposite of a profound truth may well be another profound truth.”

- Niels Bohr (1885-1962)

“I heard someone tried the monkeys-on-typewriters bit trying for the plays of W. Shakespeare, but all they got was the collected works of Francis Bacon.”

- Bill Hirst

“There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”

- Albert Einstein (1879-1955)

Basics of information theory

- We would like to develop a usable measure of the *information* we get from observing the occurrence of an event having probability p . Our first reduction will be to ignore any particular features of the event, and only observe whether or not it happened. Thus we will think of an event as the observance of a symbol whose probability of occurring is p. We will thus be defining the *information* in terms of the probability p.

  The approach we will be taking here is axiomatic: on the next page is a list of the four fundamental axioms we will use. Note that we can apply this axiomatic system in any context in which we have available a set of non-negative real numbers. A specific special case of interest is *probabilities* (i.e., real numbers between 0 and 1), which motivated the selection of axioms . . .

- We will want our *information* measure $I(p)$ to have several properties (note that along with the axiom is motivation for choosing the axiom):

  1. Information is a non-negative quantity: $I(p) \geq 0$.

  2. If an event has probability 1, we get no information from the occurrence of the event: $I(1) = 0$.

  3. If two independent events occur (whose joint probability is the product of their individual probabilities), then the information we get from observing the events is the sum of the two informations: $I(p_1 * p_2) = I(p_1) + I(p_2)$. (This is the critical property . . .)

  4. We will want our *information* measure to be a continuous (and, in fact, monotonic) function of the probability (slight changes in probability should result in slight changes in *information*).

- We can therefore derive the following:

  1. $I(p^2) = I(p * p) = I(p) + I(p) = 2 * I(p)$

  2. Thus, further, $I(p^n) = n * I(p)$
     (by induction . . .)

  3. $I(p) = I((p^{1/m})^m) = m * I(p^{1/m})$, so
     $I(p^{1/m}) = \frac{1}{m} * I(P)$ and thus in general

     $$I(p^{n/m}) = \frac{n}{m} * I(p)$$

  4. And thus, by continuity, we get, for
     $0 < p \leq 1$, and $a > 0$ a real number:

     $$I(p^a) = a * I(p)$$

- From this, we can derive the nice property:

  $$I(p) = -\log_b(p) = \log_b(1/p)$$

  for some base b.

- Summarizing: from the four properties,

  1. $I(p) \geq 0$

  2. $I(p_1 * p_2) = I(p_1) + I(p_2)$

  3. $I(p)$ is monotonic and continuous in p

  4. $I(1) = 0$

  we can derive that

  $$I(p) = \log_b(1/p) = -\log_b(p),$$

  for some positive constant b. The base b determines the units we are using.

  We can change the units by changing the base, using the formulas, for $b_1, b_2, x > 0$,

  $$x = b_1^{\log_{b_1}(x)}$$

  and therefore

  $$\log_{b_2}(x) = \log_{b_2}(b_1^{\log_{b_1}(x)}) = (\log_{b_2}(b_1))(\log_{b_1}(x)).$$

- Thus, using different bases for the logarithm results in *information* measures which are just constant multiples of each other, corresponding with measurements in different units:

  1. $\log_2$ units are *bits* (from 'binary')

  2. $\log_3$ units are *trits*(from 'trinary')

  3. $\log_e$ units are *nats* (from 'natural logarithm') (We'll use $\ln(x)$ for $\log_e(x)$)

  4. $\log_{10}$ units are *Hartleys*, after an early worker in the field.

- Unless we want to emphasize the units, we need not bother to specifiy the base for the logarithm, and will write $\log(p)$. Typically, we will think in terms of $\log_2(p)$.

- For example, flipping a fair coin once will give us events h and t each with probability 1/2, and thus a single flip of a coin gives us $-\log_2(1/2) = 1$ bit of information (whether it comes up h or t).

  Flipping a fair coin n times (or, equivalently, flipping n fair coins) gives us $-\log_2((1/2)^n) = \log_2(2^n) = n * \log_2(2) = n$ bits of information.

  We could enumerate a sequence of 25 flips as, for example:

  $$hthhtthhhhthttthhhhthtt$$

  or, using 1 for h and 0 for t, the 25 bits

  1011001011101000101110100.

  We thus get the nice fact that n flips of a fair coin gives us n bits of information, and takes n binary digits to specify. That these two are the same reassures us that we have done a good job in our definition of our *information* measure . . .

Information (and hope) ↞

“In Cyberspace, the First Amendment is a local ordinance.”

- John Perry Barlow

“Groundless hope, like unconditional love, is the only kind worth having.”

- John Perry Barlow

“The most interesting facts are those which can be used several times, those which have a chance of recurring. . . . Which, then, are the facts that have a chance of recurring? In the first place, simple facts.”

H. Poincare, 1908

Some entropy theory ←

- Suppose now that we have n symbols $\{a_1, a_2, \ldots, a_n\}$, and some source is providing us with a stream of these symbols. Suppose further that the source emits the symbols with probabilities $\{p_1, p_2, \ldots, p_n\}$, respectively. For now, we also assume that the symbols are emitted independently (successive symbols do not depend in any way on past symbols).

  What is the average amount of *information* we get from each symbol we see in the stream?

- What we really want here is a weighted average. If we observe the symbol a_i, we will get be getting $\log(1/p_i)$ *information* from that particular observation. In a long run (say N) of observations, we will see (approximately) $N * p_i$ occurrences of symbol a_i (in the frequentist sense, that's what it means to say that the probability of seeing a_i is p_i). Thus, in the N (independent) observations, we will get total information I of

$$I = \sum_{i=1}^{n} (N * p_i) * \log(1/p_i).$$

But then, the average information we get per symbol observed will be

$$\begin{aligned} I/N &= (1/N) \sum_{i=1}^{n} (N * p_i) * \log(1/p_i) \\ &= \sum_{i=1}^{n} p_i * \log(1/p_i) \end{aligned}$$

Note that $\lim_{x \to 0} x * \log(1/x) = 0$, so we can, for our purposes, define $p_i * \log(1/p_i)$ to be 0 when $p_i = 0$.

- This brings us to a fundamental definition. This definition is essentially due to Shannon in 1948, in the seminal papers in the field of information theory.

  As we have observed, we have defined *information* strictly in terms of the probabilities of events. Therefore, let us suppose that we have a set of probabilities (a probability distribution) $P = \{p_1, p_2, \ldots, p_n\}$. We define the *entropy* of the distribution P by:

$$H(P) = \sum_{i=1}^{n} p_i * \log(1/p_i).$$

  I'll mention here the obvious generalization, if we have a continuous rather than discrete probability distribution $P(x)$:

$$H(P) = \int P(x) * \log(1/P(x))dx.$$

- Another worthwhile way to think about this is in terms of *expected value*. Given a discrete probability distribution $P = \{p_1, p_2, \ldots, p_n\}$, with $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$, or a continuous distribution $P(x)$ with $P(x) \geq 0$ and $\int P(x)dx = 1$, we can define the *expected value* of an associated discrete set $F = \{f_1, f_2, \ldots, f_n\}$ or function $F(x)$ by:

$$< F >= \sum_{i=1}^{n} f_i p_i$$

or

$$< F(x) >= \int F(x)P(x)dx.$$

With these definitions, we have that:

$$H(P) =< I(p) > .$$

In other words, the *entropy* of a probability distribution is just the *expected value* of the *information* of the distribution.

Several questions probably come to mind at this point:

- What properties does the function $H(P)$ have? For example, does it have a maximum, and if so where?

- Is *entropy* a reasonable name for this? In particular, the name *entropy* is already in use in thermodynamics. How are these uses of the term related to each other?

- What can we do with this new tool?

- Let me start with an easy one. Why use the letter H for entropy? What follows is a slight variation of a footnote, p. 105, in the book *Spikes* by Rieke, et al. :-)

H (or S) for Entropy ↞

"The enthalpy is [often] written U. V is the volume, and Z is the partition function. P and Q are the position and momentum of a particle. R is the gas constant, and of course T is temperature. W is the number of ways of configuring our system (the number of states), and we have to keep X and Y in case we need more variables. Going back to the first half of the alphabet, A, F, and G are all different kinds of free energies (the last named for Gibbs). B is a virial coefficient or a magnetic field. I will be used as a symbol for information; J and L are angular momenta. K is Kelvin, which is the proper unit of T. M is magnetization, and N is a number, possibly Avogadro's, and O is too easily confused with 0. This leaves S . . ." and H. In *Spikes* they also eliminate H (e.g., as the Hamiltonian). I, on the other hand, along with Shannon and others, prefer to honor Hartley. Thus, H for entropy . . .

The Gibbs inequality ←

- First, note that the function $\ln(x)$ has derivative $1/x$. From this, we find that the tangent to $\ln(x)$ at $x = 1$ is the line $y = x - 1$. Further, since $\ln(x)$ is concave down, we have, for $x > 0$, that

$$\ln(x) \leq x - 1,$$

with equality only when $x = 1$.

Now, given two probability distributions, $P = \{p_1, p_2, \ldots, p_n\}$ and $Q = \{q_1, q_2, \ldots, q_n\}$, where $p_i, q_i \geq 0$ and $\sum_i p_i = \sum_i q_i = 1$, we have

$$\begin{aligned}\sum_{i=1}^{n} p_i \ln\left(\frac{q_i}{p_i}\right) &\leq \sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i} - 1\right) = \sum_{i=1}^{n} (q_i - p_i) \\ &= \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i = 1 - 1 = 0,\end{aligned}$$

with equality only when $p_i = q_i$ for all i. It is easy to see that the inequality actually holds for any base, not just e.

- We can use the Gibbs inequality to find the probability distribution which maximizes the entropy function. Suppose $P = \{p_1, p_2, \ldots, p_n\}$ is a probability distribution. We have

$$\begin{aligned}
H(P) - \log(n) &= \sum_{i=1}^{n} p_i \log(1/p_i) - \log(n) \\
&= \sum_{i=1}^{n} p_i \log(1/p_i) - \log(n) \sum_{i=1}^{n} p_i \\
&= \sum_{i=1}^{n} p_i \log(1/p_i) - \sum_{i=1}^{n} p_i \log(n) \\
&= \sum_{i=1}^{n} p_i (\log(1/p_i) - \log(n)) \\
&= \sum_{i=1}^{n} p_i (\log(1/p_i) + \log(1/n)) \\
&= \sum_{i=1}^{n} p_i \log\left(\frac{1/n}{p_i}\right) \\
&\leq 0,
\end{aligned}$$

  with equality only when $p_i = \frac{1}{n}$ for all i. The last step is the application of the Gibbs inequality.

- What this means is that

$$0 \leq H(P) \leq \log(n).$$

  We have $H(P) = 0$ when exactly one of the p_i's is one and all the rest are zero. We have $H(P) = \log(n)$ only when all of the events have the same probability $\frac{1}{n}$.

  That is, the maximum of the entropy function is the log() of the number of possible events, and occurs when all the events are equally likely.

- An example illustrating this result: How much information can a student get from a single grade? First, the maximum information occurs if all grades have equal probability (e.g., in a pass/fail class, on average half should pass if we want to maximize the information given by the grade).

The maximum information the student gets from a grade will be:

Pass/Fail : 1 bit.

A, B, C, D, F : 2.3 bits.

A, A-, B+, . . ., D-, F : 3.6 bits.

Thus, using +/- grading gives the students about 1.3 more bits of information per grade than without +/-, and about 2.6 bits per grade more than pass/fail.

- If a source provides us with a sequence chosen from 4 symbols (say A, C, G, T), then the maximum average information per symbol is 2 bits. If the source provides blocks of 3 of these symbols, then the maximum average information is 6 bits per block (or, to use different units, 4.159 nats per block).

We ought to note several things.

- First, these definitions of *information* and *entropy* may not match with some other uses of the terms.

  For example, if we know that a source will, with equal probability, transmit either the complete text of Hamlet or the complete text of Macbeth (and nothing else), then receiving the complete text of Hamlet provides us with precisely 1 bit of information.

  Suppose a book contains ascii characters. If the book is to provide us with *information* at the maximum rate, then each ascii character will occur with equal probability – it will be a random sequence of characters.

- Second, it is important to recognize that our definitions of *information* and *entropy* depend only on the probability distribution. In general, it won't make sense for us to talk about the *information* or the *entropy* of a source without specifying the probability distribution.

  Beyond that, it can certainly happen that two different observers of the same data stream have different models of the source, and thus associate different probability distributions to the source. The two observers will then assign different values to the *information* and *entropy* associated with the source.

  This observation (almost :-) accords with our intuition: two people listening to the same lecture can get very different information from the lecture. For example, without appropriate background, one person might not understand

anything at all, and therefore have as probability model a completely random source, and therefore get much more information than the listener who understands quite a bit, and can therefore anticipate much of what goes on, and therefore assigns non-equal probabilities to successive words . . .

Thermodynamics ↞

"A theory is the more impressive the greater the simplicity of its premises is, the more different kinds of things it relates, and the more extended its area of applicability. Therefore the deep impression which classical thermodynamics made upon me. It is the only physical theory of universal content which I am convinced that, within the framework of the applicability of its basic concepts, it will never be overthrown (for the special attention of those who are skeptics on principle)."

- A. Einstein, 1946

"Thermodynamics would hardly exist as a profitable discipline if it were not that the natural limit to the size of so many types of instruments which we now make in the laboratory falls in the region in which the measurements are still smooth."

- P. W. Bridgman, 1941

A simple physical example (gases) ←

- Let us work briefly with a simple model for an idealized gas. Let us assume that the gas is made up of N point particles, and that at some time t_0 all the particles are contained within a (cubical) volume V. Assume that through some mechanism, we can determine the location of each particle sufficiently well as to be able to locate it within a box with sides 1/100 of the sides of the containing volume V. There are 10^6 of these small boxes within V.

- We can now develop a (frequentist) probability model for this system. For each of the 10^6 small boxes, we can assign a probability p_i of finding any specific gas particle in that small box by

counting the number of particles n_i in the box, and dividing by N. That is, $p_i = \frac{n_i}{N}$. From this probability distribution, we can calculate an entropy:

$$\begin{aligned} H(P) &= \sum_{i=1}^{10^6} p_i * \log(1/p_i) \\ &= \sum_{i=1}^{10^6} \frac{n_i}{N} * \log(N/n_i) \end{aligned}$$

If the particles are evenly distributed among the 10^6 boxes, then we will have that each $n_i = N/10^6$, and in this case the entropy will be:

$$\begin{aligned} H(evenly) &= \sum_{i=1}^{10^6} \frac{N/10^6}{N} * \log\left(\frac{N}{N/10^6}\right) \\ &= \sum_{i=1}^{10^6} \frac{1}{10^6} * \log(10^6) \\ &= \log(10^6). \end{aligned}$$

There are several ways to think about this example.

- First, notice that the calculated entropy of the system depends in a strong way on the relative scale of measurement. For example, if the particles are evenly distributed, and we increase our accuracy of measurement by a factor of 10 (i.e., if each small box is 1/1000 of the side of V), then the calculated maximum entropy will be $\log(10^9)$ instead of $\log(10^6)$.

  For physical systems, we know that quantum limits (e.g., Heisenberg uncertainty relations) will give us a bound on the accuracy of our measurements, and thus a more or less natural scale for doing entropy calculations. On the other hand, for macroscopic systems, we are likely to find that we can only make relative rather than absolute entropy calculations.

- Second, we have simplified our model of the gas particles to the extent that they have only one property, their position. If we want to talk about the *state* of a particle, all we can do is specify the small box the particle is in at time t_0. There are thus $Q = 10^6$ possible *states* for a particle, and the maximum entropy for the system is $\log(Q)$. This may look familiar for equilibrium statistical mechanics . . .

- Third, suppose we generalize our model slightly, and allow the particles to move about within V. A *configuration* of the system is then simply a list of 10^6 numbers b_i with $1 \leq b_i \leq N$ (i.e., a list of the numbers of particles in each of the boxes). Suppose that the motions of the particles are such that for each particle, there is an equal probability that it will move into any given new small box during

one (macroscopic) time step. How likely is it that at some later time we will find the system in a "high" entropy configuration? How likely is it that if we start the system in a "low" entropy configuration, it will stay in a "low" entropy configuration for an appreciable length of time? If the system is not currently in a "maximum" entropy configuration, how likely is it that the entropy will increase in succeeding time steps (rather than stay the same or decrease)?

Let's do a few computations using combinations:

$$\binom{n}{m} = \frac{n!}{m! * (n-m)!},$$

and Stirling's approximation:

$$n! \approx \sqrt{2\pi}\ n^n e^{-n} \sqrt{n}.$$

Let us start here:

There are 10^6 configurations with all the particles sitting in exactly one small box, and the entropy of each of those configurations is:

$$H(all\ in\ one) = \sum_{i=1}^{10^6} p_i * \log(1/p_i) = 0,$$

since exactly one p_i is 1 and the rest are 0. These are obviously minimum entropy configurations.

Now consider pairs of small boxes. The number of configurations with all the particles evenly distributed between two boxes is:

$$\begin{aligned} \binom{10^6}{2} &= \frac{10^6!}{(2)!(10^6-2)!} \\ &= \frac{10^6 * (10^6-1)}{2} \\ &= 5 * 10^{11}, \end{aligned}$$

which is a (comparatively :-) large number. The entropy of each of these configurations is:

$H(two\ boxes) = 1/2*\log(2)+1/2*\log(2) = \log(2).$

We thus know that there are at least $5 * 10^{11} + 10^6$ configurations. If we start the system in a configuration with entropy 0, then the probability that at some later time it will be in a configuration with entropy $\geq \log(2)$ will be

$$\begin{aligned} &\geq \frac{5 * 10^{11}}{5 * 10^{11} + 10^6} = (1 - \frac{10^6}{5 * 10^{11} + 10^6}) \\ &\geq (1 - 10^{-5}). \end{aligned}$$

As an example at the other end, consider the number of configurations with the particles distributed almost equally, except that half the boxes are short by one particle, and the rest have an extra. The

number of such configurations is:

$$\begin{aligned}
\binom{10^6}{10^6/2} &= \frac{10^6!}{(10^6/2)!(10^6 - 10^6/2)!} \\
&= \frac{10^6!}{((10^6/2)!)^2}
\end{aligned}$$

$$\begin{aligned}
&\approx \frac{\sqrt{2\pi}(10^6)^{10^6} e^{-10^6}\sqrt{10^6}}{(\sqrt{2\pi}(10^6/2)^{10^6/2} e^{-(10^6/2)}\sqrt{10^6/2})^2} \\
&= \frac{\sqrt{2\pi}(10^6)^{10^6} e^{-10^6}\sqrt{10^6}}{2\pi(10^6/2)^{10^6} e^{-(10^6)} 10^6/2} \\
&= \frac{2^{10^6+1}\sqrt{10^6}}{\sqrt{2\pi}\sqrt{10^6}} \\
&\approx 2^{10^6} \\
&= (2^{10})^{10^5} \\
&\approx 10^{3*10^5}.
\end{aligned}$$

Each of these configurations has entropy essentially equal to $\log(10^6)$.

From this, we can conclude that if we start the system in a configuration with

entropy 0 (i.e., all particles in one box), the probability that later it will be in a higher entropy configuration will be $> (1 - 10^{-3*10^5})$.

Similar arguments (with similar results in terms of probabilities) can be made for starting in any configuration with entropy appreciably less than $\log(10^6)$ (the maximum). In other words, it is overwhelmingly probable that as time passes, macroscopically, the system will increase in entropy until it reaches the maximum.

In many respects, these general arguments can be thought of as a "proof" (or at least an explanation) of a version of the second law of thermodynamics: Given any macroscopic system which is free to change configurations, and given any configuration with entropy less than the maximum, there will be

overwhelmingly many more accessible configurations with higher entropy than lower entropy, and thus, with probability indistinguishable from 1, the system will (in macroscopic time steps) successively change to configurations with higher entropy until it reaches the maximum.

Language, and putting things together ↞

“An essential distinction between language and experience is that language separates out from the living matrix little bundles and freezes them; in doing this it produces something totally unlike experience, but nevertheless useful.”

- P. W. Bridgman, 1936

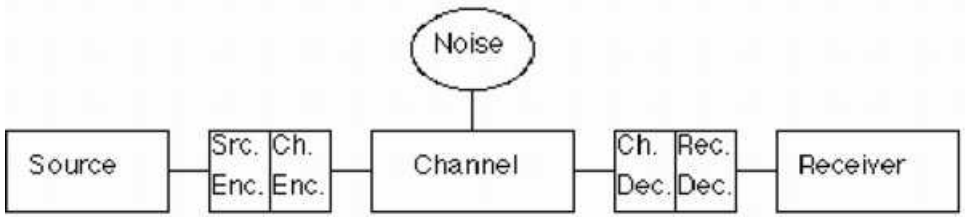
“One is led to a new notion of unbroken wholeness which denies the classical analyzability of the world into separately and independently existing parts. The inseparable quantum interconnectedness of the whole universe is the fundamental reality.”

- David Bohm

Shannon's communication theory ←

- In his classic 1948 papers, Claude Shannon laid the foundations for contemporary *information*, *coding*, and *communication* theory. He developed a general model for communication systems, and a set of theoretical tools for analyzing such systems.

  His basic model consists of three parts: a sender (or source), a channel, and a receiver (or sink). His general model also includes encoding and decoding elements, and noise within the channel.



Shannon's communication model

- In Shannon's discrete model, it is assumed that the source provides a stream of symbols selected from a finite alphabet $A = \{a_1, a_2, \ldots, a_n\}$, which are then encoded. The code is sent through the channel (and possibly disturbed by noise). At the other end of the channel, the receiver will decode, and derive information from the sequence of symbols.

  Let me mention at this point that sending information from *now* to *then* is equivalent to sending information from *here* to *there*, and thus Shannon's theory applies equally as well to information storage questions as to information transmission questions.

- One important question we can ask is, how efficiently can we encode information that we wish to send through the channel? For the moment, let's assume that the channel is noise-free, and that the receiver can accurately recover the channel symbols transmitted through the channel. What we need, then, is an efficient way to encode the stream of source symbols for transmission through the channel, and to be sure that the encoded stream can be uniquely decoded at the receiving end.

  If the alphabet of the channel (i.e., the set of symbols that can actually be carried by the channel) is $C = \{c_1, c_2, \ldots, c_r\}$, then an encoding of the source alphabet A is just a function $f : A \to C^*$ (where C^* is the set of all possible finite strings of symbols from C). For future calculations, let $l_i = |f(a_i)|, i = 1, 2, \ldots, n$ (i.e., l_i is the length of the string encoding the symbol $a_i \in A$).

- There is a nice inequality concerning the lengths of code strings for uniquely decodable (and/or instantaneous) codes, called the McMillan/Kraft inequality. There is a uniquely decodable code with lengths $l_1, l_2, \ldots, l_n$ if and only if

$$K = \sum_{i=1}^{n} \frac{1}{r^{l_i}} \leq 1.$$

The necessity of this inequality can be seen from looking at

$$K^n = \left[\sum_{i=1}^{n} \frac{1}{r^{l_i}} \right]^n .$$

We can rewrite this as

$$K^n = \sum_{k=n}^{nl} \frac{N_k}{r^k}$$

where l is the length of the longest code and N_k is the number of encodings of strings having encoded length k.

Note that N_k cannot be greater than r^k (the total number of strings of length k, whether they encode anything or not). From this we can see that

$$K^n \leq \sum_{k=n}^{nl} \frac{r^k}{r^k} = nl - n + 1 \leq nl.$$

From this we can conclude that $K \leq 1$ (as desired), since otherwise K^n would exceed nl for some (possibly large) n.

We can now prove a very important property of the entropy: the entropy gives a lower bound for the efficiency of an encoding scheme (in other words, a lower bound on the possible compression of a data stream).

With K defined as above, we can define a set of numbers Q_i (pseudo-probabilities) by

$$Q_i = \frac{r^{-l_i}}{K}.$$

We call these pseudo-probabilities because we have $0 < Q_i \leq 1$ for all i, and

$$\sum_{i=1}^{n} Q_i = 1.$$

If p_i is the probability of observing a_i in the data stream, then we can apply the Gibbs inequality to get

$$\sum_{i=1}^{n} p_i \log\left(\frac{Q_i}{p_i}\right) \leq 0,$$

or

$$\sum_{i=1}^{n} p_i \log\left(\frac{1}{p_i}\right) \leq \sum_{i=1}^{n} p_i \log\left(\frac{1}{Q_i}\right).$$

The left hand side is the entropy of the source, say $H(S)$. Recalling the definition of Q_i (and that $K \leq 1$) we find

$$H(S) \leq \sum_{i=1}^{n} p_i \left(\log(K) - log\left(r^{-l_i}\right)\right)$$

$$= \log(K) + \sum_{i=1}^{n} p_i l_i \log(r) \leq \log(r) \sum_{i=1}^{n} p_i l_i.$$

- From this, we can draw an important conclusion. If we let $L = \sum_{i=1}^{n} p_i l_i$, then L is just the average length of code words in the encoding. What we have shown is that

$$H(S) \leq L \log(r).$$

In other words, the entropy gives us a lower bound on average code length for any uniquely decodable symbol-by-symbol encoding of our data stream. Note that, for example, if we calculate entropy in bits and use binary ($r = 2$) encoding, then we have simply

$$H(S) \leq L.$$

Shannon went beyond this, and showed that the bound (appropriately recast) holds even if we use extended coding systems where we group symbols together (into "words") before doing our encoding. The generalized form of this inequality is called *Shannon's noiseless coding theorem*.

- In building encoding schemes for data streams (or, alternatively, in building data compression schemes), we will want to use our best understandings of the structure of the data stream – in other words, we will want to use our best probability model of the data stream. Shannon's theorem tells us that, since the entropy gives us a lower bound on our encoding efficiency, if we want to improve our schemes, we will have to develop successively better probability models.

  One way to think about a scientific theory is that a theory is just an efficient way of encoding (i.e., structuring) our knowledge about (some aspect of) the world. A good theory is one which reduces the (relative) entropy of our (probabilistic) understanding of the system (i.e., that decreases our average lack of knowledge about the system) . . .

- Shannon went on to generalize to the (more realistic) situation in which the channel itself is noisy. In other words, not only are we unsure about the data stream we will be transmitting through the channel, but the channel itself adds an additional layer of uncertainty/probability to our transmissions.

  Given a source of symbols and a channel with noise (in particular, given probability models for the source and the channel noise), we can talk about the *capacity* of the channel. The general model Shannon worked with involved two sets of symbols, the input symbols and the output symbols. Let us say the two sets of symbols are $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_m\}$. Note that we do not necessarily assume the same number of symbols in the two sets. Given the noise in the channel, when symbol b_j comes out of the channel, we can not be certain

which a_i was put in. The channel is characterized by the set of probabilities $\{P(a_i|b_j)\}$.

- We can then consider various related information and entropy measures. First, we can consider the information we get from observing a symbol b_j. Given a probability model of the source, we have an *a priori* estimate $P(a_i)$ that symbol a_i will be sent next. Upon observing b_j, we can revise our estimate to $P(a_i|b_j)$. The change in our information (the *mutual information*) will be given by:

$$\begin{aligned} I(a_i;b_j) &= \log\left(\frac{1}{P(a_i)}\right) - \log\left(\frac{1}{P(a_i|b_j)}\right) \\ &= \log\left(\frac{P(a_i|b_j)}{P(a_i)}\right) \end{aligned}$$

We have the properties:

$$\begin{aligned} I(a_i;b_j) &= I(b_j;a_i) \\ I(a_i;b_j) &= \log(P(a_i|b_j)) + I(a_i) \\ I(a_i;b_j) &\leq I(a_i) \end{aligned}$$

If a_i and b_j are independent (i.e., if $P(a_i, b_j) = P(a_i) * P(b_j)$), then $I(a_i; b_j) = 0$.

- What we actually want is to average the *mutual information* over all the symbols:

$$\begin{aligned} I(A; b_j) &= \sum_i P(a_i|b_j) * I(a_i; b_j) \\ &= \sum_i P(a_i|b_j) * \log\left(\frac{P(a_i|b_j)}{P(a_i)}\right) \\ I(a_i; B) &= \sum_j P(a_i|b_j) * \log\left(\frac{P(b_j|a_i)}{P(b_j)}\right), \end{aligned}$$

and from these,

$$\begin{aligned} I(A; B) &= \sum_i P(a_i) * I(a_i; B) \\ &= \sum_i \sum_j P(a_i, b_j) * \log\left(\frac{P(a_i, b_j)}{P(a_i)P(b_j)}\right) \\ &= I(B; A). \end{aligned}$$

We have the properties: $I(A; B) \geq 0$, and $I(A; B) = 0$ if and only if A and B are independent.

- We then have the definitions and properties:

$$
\begin{aligned}
H(A) &= \sum_{i=1}^{n} P(a_i) * \log(1/P(a_i)) \\
H(B) &= \sum_{j=1}^{m} P(b_j) * \log(1/P(b_j)) \\
H(A|B) &= \sum_{i=1}^{n} \sum_{j=1}^{m} P(a_i|b_j) * \log(1/P(a_i|b_j)) \\
H(A,B) &= \sum_{i=1}^{n} \sum_{j=1}^{m} P(a_i,b_j) * \log(1/P(a_i,b_j)) \\
H(A,B) &= H(A) + H(B|A) \\
&= H(B) + H(A|B),
\end{aligned}
$$

and furthermore:

$$
\begin{aligned}
I(A;B) &= H(A) + H(B) - H(A,B) \\
&= H(A) - H(A|B) \\
&= H(B) - H(B|A) \\
&\geq 0
\end{aligned}
$$

- If we are given a channel, we could ask what is the maximum possible information that can be transmitted through the channel. We could also ask what mix of the symbols $\{a_i\}$ we should use to achieve the maximum. In particular, using the definitions above, we can define the *Channel Capacity* C to be:

$$C = \max_{P(a)} \ I(A; B).$$

- We have the nice property that if we are using the channel at its capacity, then for each of the a_i,

$$I(a_i; B) = C,$$

and thus, we can maximize channel use by maximizing the use for each symbol independently.

- We also have Shannon's main theorem:

  For any channel, there exist ways of encoding input symbols such that we can simultaneously utilize the channel as closely as we wish to the capacity, and at the same time have an error rate as close to zero as we wish.

- This is actually quite a remarkable theorem. We might naively guess that in order to minimize the error rate, we would have to use more of the channel capacity for error detection/correction, and less for actual transmission of information. Shannon showed that it is possible to keep error rates low and still use the channel for information transmission at (or near) its capacity.

- Unfortunately, Shannon's proof has a a couple of downsides. The first is that the proof is non-constructive. It doesn't tell us how to construct the coding system to optimize channel use, but only tells us that such a code exists. The second is that in order to use the capacity with a low error rate, we may have to encode very large blocks of data. This means that if we are attempting to use the channel in real-time, there may be time lags while we are filling buffers. There is thus still much work possible in the search for efficient coding schemes.

  Among the things we can do is look at natural coding systems (such as, for example, the DNA coding system, or neural systems) and see how they use the capacity of their channel. It is not unreasonable to assume that evolution will have done a pretty good job of optimizing channel use . . .

Tools ↞

“It is a recurring experience of scientific progress that what was yesterday an object of study, of interest in its own right, becomes today something to be taken for granted, something understood and reliable, something known and familiar – a tool for further research and discovery.”

-J. R. Oppenheimer, 1953

“Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry.”

- Richard Feynman

Application to Biology (analyzing genomes) ←

- Let us apply some of these ideas to the (general) problem of analyzing genomes.

  We can start with an example such as the comparatively small genome of Escherichia coli, strain K-12, substrain MG1655, version M52. This example has the convenient features:

  1. It has been completely sequenced.

  2. The sequence is available for downloading (http://www.genome.wisc.edu/).

  3. Annotated versions are available for further work.

  4. It is large enough to be interesting (somewhat over 4 mega-bases, or 4

million nucleotides), but not so huge as to be completely unwieldy.

5. The labels on the printouts tend to make other people using the printer a little nervous :-)

6. Here's the beginning of the file:

```
>gb|U00096|U00096 Escherichia coli
   K-12 MG1655 complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCT
CTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAA
TTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAG
AGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGG
TGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
CCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACC
AAAGGTAACGAGGTAACAACCATGCGAGTGTTGAA
```

- In this exploratory project, my goal has been to apply the *information* and *entropy* ideas outlined above to genome analysis. Some of the results I have so far are tantalizing. For a while, I'll just walk you through some preliminary work. While I am not an expert in genomes/DNA, I am hoping that some of what I am doing can bring fresh eyes to the problems of analyzing genome sequences, without too many preconceptions. It is at least conceivable that my naiveté will be an advantage . . .

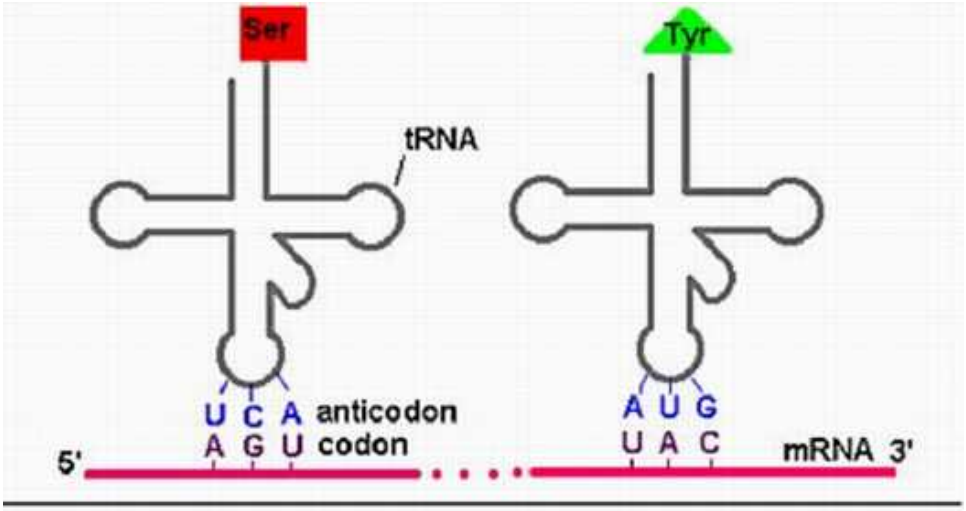
- My first step was to generate for myself a "random genome" of comparable size to compare things with. In this case, I simply used the Unix 'random' function to generate a file containing a random sequence of about 4 million A, C, G, T. In the actual genome, these letters stand for the nucleotides adenine, cytosine, guanine, and thymine.

  Other people working in this area have taken some other approaches to this process, such as randomly shuffling an actual genome (thus maintaining the relative proportions of A, C, G, and T). Part of the justification for this methodology is that actual (identified) coding sections of DNA tend to have a ratio of C+G to A+T different from one. I didn't worry about this issue (for various reasons).

- My next step was to start developing a (variety of) probability model(s) for the genome. The general idea that I am working on is to build some automated tools to locate "interesting" sections of a genome. Thinking of DNA as a coding system, we can hope that "important" stretches of DNA will have entropy different from other stretches. Of course, as noted above, the entropy measure depends in an essential way on the probability model attributed to the source. We will want to try to build a model that catches important aspects of what we find interesting or significant. We will want to use our knowledge of the systems in which DNA is embedded to guide the development of our models. On the other hand, we probably don't want to constrain the model too much. Remember that information and entropy are measures of *unexpectedness*. If we constrain our model too much, we won't leave any room for the unexpected!

- We know, for example, that simple repetitions have low entropy. But if the code being used is redundant (sometimes called *degenerate*), with multiple encodings for the same symbol (as is the case for DNA codons), what looks to one observer to be a random stream may be recognized by another observer (who knows the code) to be a simple repetition.

- The first element of my probability model(s) involves the observation that coding sequences for peptides and proteins are encoded via *codons*, that is, by sequences of blocks of triples of nucleotides. Thus, for example, the codon AGC on mRNA (messenger RNA) codes for the amino acid serine (or, if we happen to be reading in the reverse direction, it might code for alanine). On DNA, AGC codes for UCG or CGA on the mRNA, and thus could code for cysteine or arginine.



2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code

Amino acids specified by each codon sequence on mRNA.
A = adenine G = guanine C = cytosine
T = thymine U = uracil
Table from
http://www.accessexcellence.org

Key for the above table:

Ala: Alanine
Arg: Arginine
Asn: Asparagine
Asp: Aspartic acid
Cys: Cysteine
Gln: Glutamine
Glu: Glutamic acid
Gly: Glycine
His: Histidine
Ile: Isoleucine
Leu: Leucine
Lys: Lysine
Met: Methionine
Phe: Phenylalanine
Pro: Proline
Ser: Serine
Thr: Threonine
Trp: Tryptophane
Tyr: Tyrosine
Val: Valine

- For our first model, we will consider each three-nucleotide codon to be a distinct symbol. We can then take a chunk of genome and estimate the probability of occurence of each codon by simply counting and dividing by the length. At this level, we are assuming we have no knowledge of where codons start, and so in this model, we assume that "readout" could begin at any nucleotide. We thus use each three adjacent nucleotides.

  For example, given the DNA chunk:

  ```
  AGCTTTTCATTCTGACTGCAACGGGCAATATGTC
  ```

  we would count:

  ```
  AAT  1  AAC  1  ACG  1  ACT  1  AGC  1
  ATA  1  ATG  1  ATT  1  CAA  2  CAT  1
  CGG  1  CTG  2  CTT  1  GAC  1  GCA  2
  GCT  1  GGC  1  GGG  1  GTC  1  TAT  1
  TCA  1  TCT  1  TGA  1  TGC  1  TGT  1
  TTC  2  TTT  2
  ```

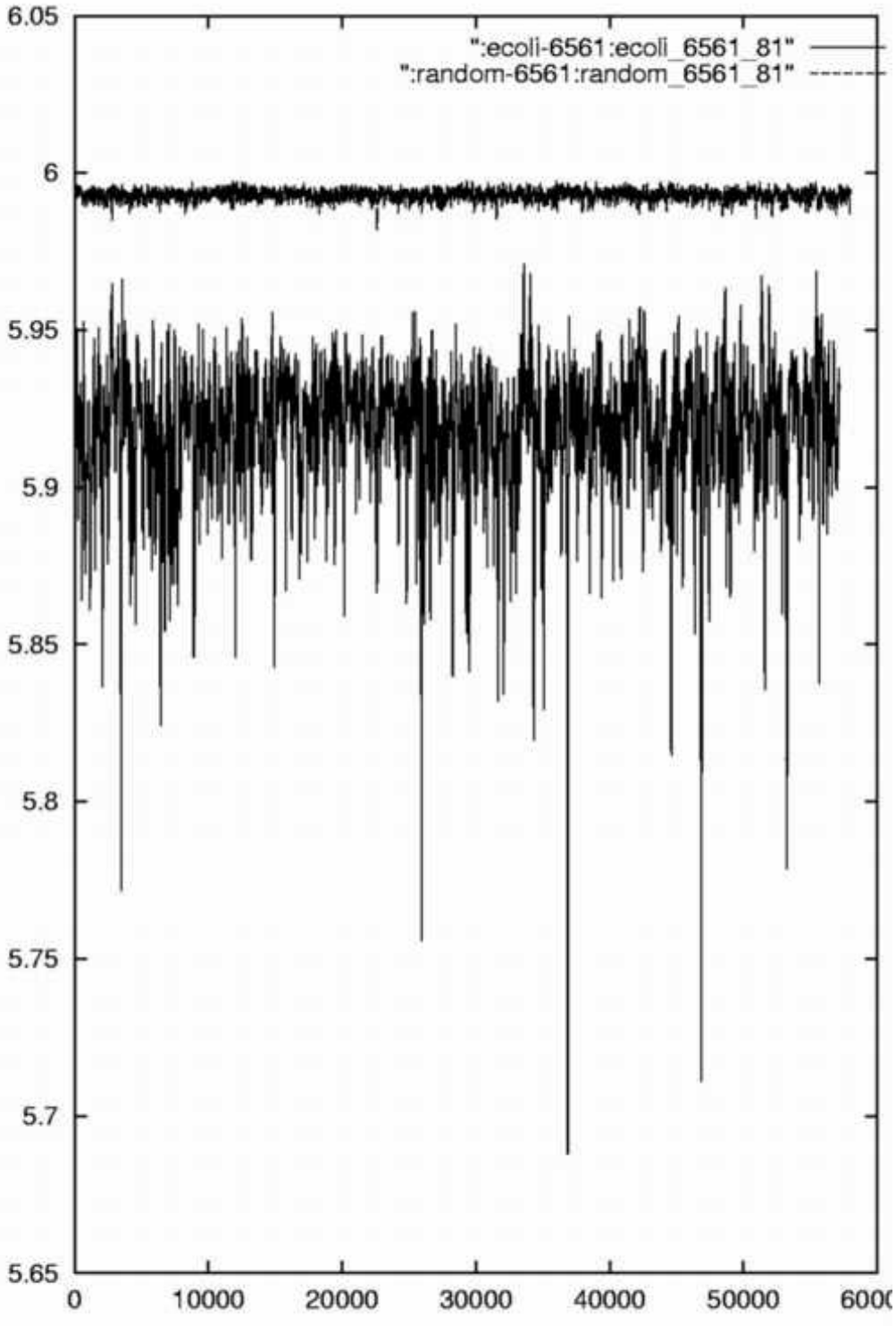
- We can then estimate the entropy of the chunk as:

$$\sum p_i * \log_2(1/p_i) = 4.7 \ bits.$$

The maximum possible entropy for this chunk would be:

$$\log_2(27) = 4.755 \ bits.$$

- We want to find "interesting" sections (and features) of a genome. As a starting place, we can slide a "window" over the genome, and estimate the entropy within the window. The plot below shows the entropy estimates for the E. coli genome, within a window of size 6561 ($= 3^8$). The window is slid in steps of size 81 ($= 3^4$). This results in 57,194 values, one for each placement of the window. For comparison, the values for a "random" genome are also shown.



Entropy of E. coli and random
window 6561, slide-step 81

- At this level, we can make the simple observation that the actual genome values are quite different from the comparative random string. The values for E. coli range from about 5.8 to about 5.96, while the random values are clustered quite closely above 5.99 (the maximum possible is $\log_2(64) = 6$).

- From here, there are various directions we could go. With a given window size and step size (e.g., 6561:81, as in the given plot), we can look at interesting features of the *entropy* estimates. For example, we could look at regions with high entropy, or low entropy. We could look at regions where there are abrupt changes in entropy, or regions where entropy stays relatively stable.

- We could change the window size, and/or step size. We could work to develop adaptive algorithms which zoom in on interesting regions, where "interesting" is determined by criteria such as the ones listed above.

- We could take known coding regions of genomes, and develop entropy "fingerprints" which we could then try to match.

- There are various "data massage" techniques we could use. For example, we could take the fourier transform of the entropy estimates, and explore that. Below is an example of such a fourier transform. Notice that it has some interesting "periodic" features which might be worth exploring. It is also interesting to note that the fourier

transform of the entropy of a "random" genome has the shape of approximately $1/f = 1/f^1$ (not unexpected . . .), whereas the E. coli data are closer to $1/f^{1.5}$.

- The discrete Fourier transform of a sequence $(a_j)_{j=0}^{q-1}$ is the sequence $(A_k)_{k=0}^{q-1}$ where

$$A_k = \frac{1}{\sqrt{q}} \sum_{j=0}^{q-1} a_j e^{\frac{2\pi i j k}{q}}$$

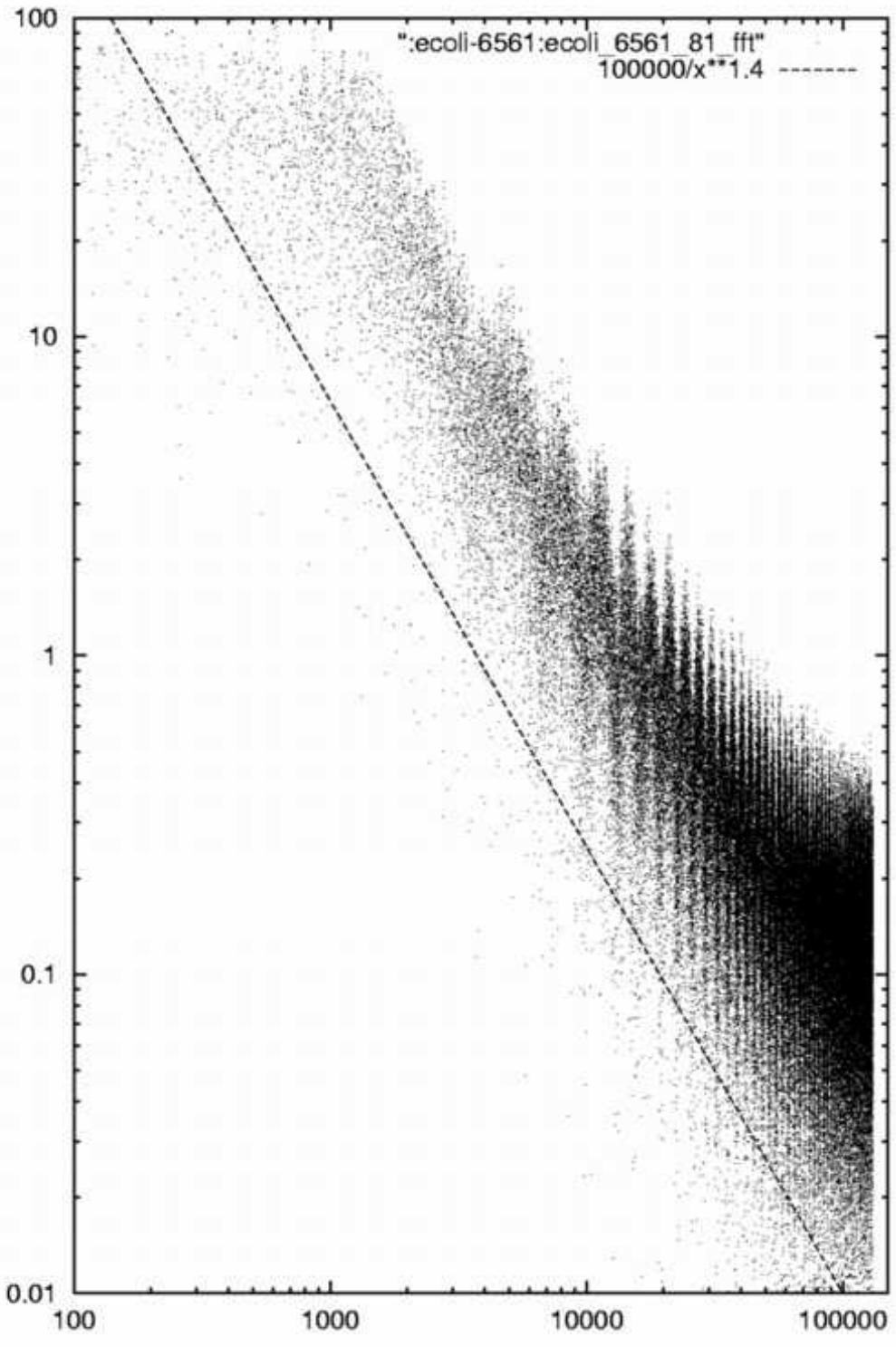
One way to think about this is that $(A_k) = F((a_j))$ where the linear transformation F is given by:

$$[F]_{j,k} = \frac{1}{\sqrt{q}} e^{\frac{2\pi i j k}{q}}$$

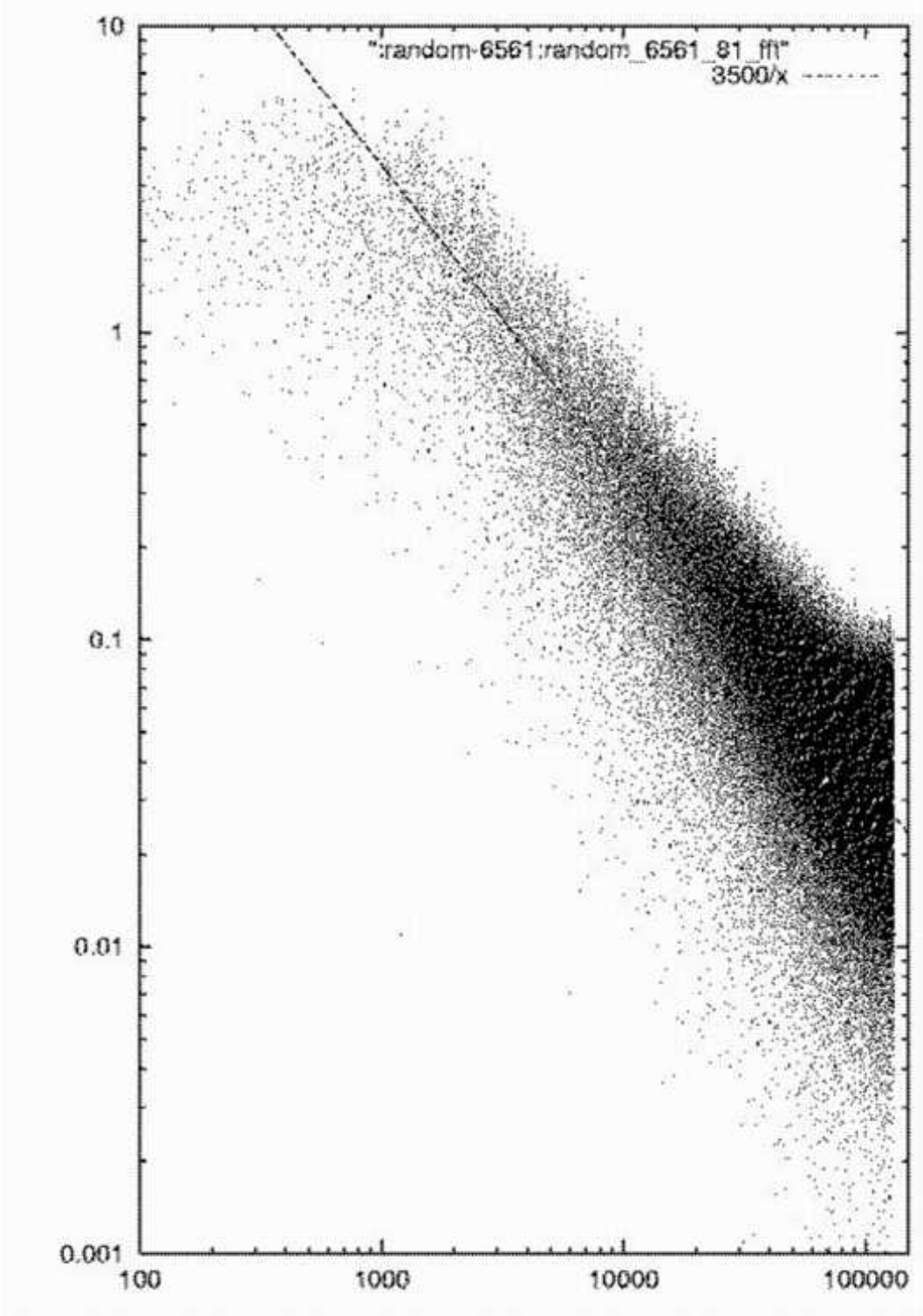
Note that the inverse of F is its conjugate transpose $F^\dagger$ – that is,

$$[F^{-1}]_{k,j} = \frac{1}{\sqrt{q}} e^{-\frac{2\pi i j k}{q}}.$$

The plots that follow are log-log plots of the norms $|A_k| = (A_k \bar{A}_k)^{1/2}$ (power spectra).



Fourier transform of E. coli window 6561, slide-step 81



Fourier transform of random window 6561, slide-step 81

Some other measures ←

- There have been various approaches to expanding on the idea of entropy as a measure of complexity. One useful generalization of entropy was developed by the Hungarian mathematician A. Rényi. His method involves looking at the moments of order q of a probability distribution $\{p_i\}$:

$$S_q = \frac{1}{q-1} \log \sum_i p_i^q$$

If we take the limit as $q \to 1$, we get:

$$S_1 = \sum_i p_i \log(1/p_i),$$

the entropy we have previously defined. We can then think of S_q as a generalized entropy for any real number q.

- Expanding on these generalized entropies, we can then define a generalized *dimension* associated with a data set. If we imagine the data set to be distributed among bins of diameter r, we can let p_i be the probability that a data item falls in the i'th bin (estimated by counting the data elements in the bin, and dividing by the total number of items). We can then, for each q, define a dimension:

$$D_q = \lim_{r \to 0} \frac{1}{q-1} \frac{\log \sum_i p_i^q}{\log(r)}.$$

- Why do we call this a generalized *dimension*?

  Consider D_0. First, we will adopt the (analyst's?) convention that $p_i^0 = 0$ when $p_i = 0$. Also, let N_r be the number of non-empty bins (i.e., the number of bins of diameter r it takes to cover the data set).

Then we have:

$$D_0 = \lim_{r \to 0} \frac{\log \sum_i p_i^0}{\log(1/r)} = \lim_{r \to 0} \frac{\log(N_r)}{\log(1/r)}$$

Thus, D_0 is the Hausdorff dimension D, which is frequently in the literature called the *fractal dimension* of the set.

Three examples:

1. Consider the unit interval $[0, 1]$. Let $r_k = 1/2^k$. Then $N_{r_k} = 2^k$, and

$$D_0 = \lim_{k \to \infty} \frac{\log(2^k)}{\log(2^k)} = 1.$$

2. Consider the unit square $[0, 1]X[0, 1]$. Again, let $r_k = 1/2^k$. Then $N_{r_k} = 2^{2k}$, and

$$D_0 = \lim_{k \to \infty} \frac{\log(2^{2k})}{\log(2^k)} = 2.$$

3. Consider the *Cantor set*:

The construction of the Cantor set is suggested by the diagram. The Cantor set is what remains from the interval after we have removed middle thirds countably many times. It is an uncountable set, with measure ("length") 0. For this set we will let $r_k = 1/3^k$. Then $N_{r_k} = 2^k$, and

$$D_0 = \lim_{k\to\infty} \frac{\log(2^k)}{\log(3^k)} = \frac{\log(2)}{\log(3)} \approx 0.631.$$

The Cantor set is a traditional example of a *fractal*. It is self similar, and has $D_0 \approx 0.631$, which is strictly greater than its topological dimension ($=$ 0).

It is an important example since many nonlinear dynamical systems have trajectories which are locally the product of a Cantor set with a manifold (i.e., Poincaré sections are generalized Cantor sets).

An interesting example of this phenomenon occurs with the *logistics* equation:

$$x_{i+1} = k * x_i * (1 - x_i)$$

with $k > 4$. In this case (of which you rarely see pictures . . .), most starting points run off rapidly to $-\infty$, but there is a *strange repellor*(!) which is a Cantor set. It is a *repellor* since arbitrarily close to any point on the trajectory are points which run off to $-\infty$. One thing this means is that any finite precision simulation will not capture the *repellor* . . .

- We can make several observations about D_q:

  1. If $q_1 \leq q_2$, then $D_{q_1} \leq D_{q_2}$.

  2. If the set is strictly self-similar with equal probabilities $p_i = 1/N$, then we do not need to take the limit as $r \to 0$, and

$$\begin{aligned} D_q &= \frac{1}{q-1}\frac{\log(N * (1/N)^q)}{\log(r)} \\ &= \frac{\log(N)}{\log(1/r)} \\ &= D_0 \end{aligned}$$

     for all q. This is the case, for example, for the Cantor set.

  3. D_1 is usually called the *information dimension*:

$$D_1 = \lim_{r \to 0} \frac{\sum_i p_i * \log(1/p_i)}{\log(r)}$$

     The numerator is just the entropy of the probability distribution.

4. D_2 is usually called the *correlation dimension*:

$$D_2 = \lim_{r \to 0} \frac{\log \sum_i p_i^2}{\log(r)}$$

This dimension is related to the probability of finding two elements of the set within a distance r of each other.

←

Some additional material

What follows are some additional examples, and expanded discussion of some topics . . .

Examples using Bayes' Theorem ←

- A quick example:
  Suppose that you are asked by a friend to help them understand the results of a genetic screening test they have taken. They have been told that they have tested positive, and that the test is 99% accurate. What is the probability that they actually have the anomaly?

  You do some research, and find out that the test screens for a genetic anomaly that is believed to occur in one person out of 100,000 on average. The lab that does the tests guarantees that the test is 99% accurate. You push the question, and find that the lab says that one percent of the time, the test falsely reports the absence of the anomaly when it is there, and one percent of the time

the test falsely reports the presence of the anomaly when it is not there. The test has come back positive for your friend. How worried should they be? Given this much information, what can you calculate as the probability they actually have the anomaly?

In general, there are four possible situations for an individual being tested:

1. Test positive (Tp), and have the anomaly (Ha).

2. Test negative (Tn), and don't have the anomaly (Na).

3. Test positive (Tp), and don't have the anomaly (Na).

4. Test negative (Tn), and have the anomaly (Ha).

We would like to calculate for our friend the probability they actually have the anomaly (Ha), given that they have tested positive (Tp):

$$P(Ha|Tp).$$

We can do this using Bayes' Theorem.

We can calculate:

$$P(Ha|Tp) \;=\; \frac{P(Tp|Ha) * P(Ha)}{P(Tp)}.$$

We need to figure out the three items on the right side of the equation. We can do this by using the information given.

Suppose the screening test was done on 10,000,000 people. Out of these 10^7 people, we expect there to be $10^7/10^5 = 100$ people with the anomaly, and 9,999,900 people without the anomaly. According to the lab, we would expect the test results to be:

- Test positive (Tp), and have the anomaly (Ha):

$$0.99 * 100 = 99 \text{ people.}$$

- Test negative (Tn), and don't have the anomaly (Na):

$$0.99 * 9,999,900 = 9,899,901 \text{ people.}$$

- Test positive (Tp), and don't have the anomaly (Na):

$$0.01 * 9,999,900 = 99,999 \text{ people.}$$

- Test negative (Tn), and have the anomaly (Ha):

$$0.01 * 100 = 1 \text{ person.}$$

Now let's put the the pieces together:

$$
\begin{aligned}
P(Ha) &= \frac{1}{100,000} \\
&= 10^{-5} \\
P(Tp) &= \frac{99 + 99,999}{10^7} \\
&= \frac{100,098}{10^7} \\
&= 0.0100098 \\
P(Tp|Ha) &= 0.99
\end{aligned}
$$

Thus, our calculated probability that our friend actually has the anomaly is:

$$\begin{aligned} P(Ha|Tp) &= \frac{P(Tp|Ha) * P(Ha)}{P(Tp)} \\ &= \frac{0.99 * 10^{-5}}{0.0100098} \\ &= \frac{9.9 * 10^{-6}}{1.00098 * 10^{-2}} \\ &= 9.890307 * 10^{-4} \\ &< 10^{-3} \end{aligned}$$

In other words, our friend, who has tested *positive*, with a test that is 99% correct, has less that one chance in 1000 of actually having the anomaly!

- There are a variety of questions we could ask now, such as, "For this anomaly, how accurate would the test have to be for there to be a greater than 50% probability that someone who tests positive actually has the anomaly?"

  For this, we need fewer false positives than true positives. Thus, in the example, we would need fewer than 100 false positives out of the 9,999,900 people who do not have the anomaly. In other words, the proportion of those without the anomaly for whom the test would have to be correct would need to be greater than:

$$\frac{9,999,800}{9,999,900} = 99.999\%$$

- Another question we could ask is, "How prevalent would an anomaly have to be in order for a 99% accurate test (1% false positive and 1% false negative) to give a greater than 50% probability of actually having the anomaly when testing positive?"

  Again, we need fewer false positives than true positives. We would therefore need the actual occurrence to be greater than 1 in 100 (each false positive would be matched by at least one true positive, on average).

- Note that the current population of the US is about 280,000,000 and the current population of the world is about 6,200,000,000. Thus, we could expect an anomaly that affects 1 person in 100,000 to affect about 2,800 people in the US, and about 62,000 people worldwide, and one affecting one person in 100 would affect 2,800,000 people in the US, and 62,000,000 people worldwide . . .

- Another example: suppose the test were not so accurate? Suppose the test were 80% accurate (20% false positive and 20% false negative). Suppose that we are testing for a condition expected to affect 1 person in 100. What would be the probability that a person testing positive actually has the condition?

We can do the same sort of calculations.

Let's use 1000 people this time. Out of this sample, we would expect 10 to have the condition.

- Test positive (Tp), and have the condition (Ha):

$$0.80 * 10 = 8 \text{ people.}$$

- Test negative (Tn), and don't have the condition (Na):

$$0.80 * 990 = 792 \text{ people.}$$

- Test positive (Tp), and don't have the condition (Na):

$$0.20 * 990 = 198 \text{ people.}$$

- Test negative (Tn), and have the condition (Ha):

$$0.20 * 10 = 2 \text{ people.}$$

Now let's put the the pieces together:

$$
\begin{aligned}
P(Ha) &= \frac{1}{100} \\
&= 10^{-2} \\
P(Tp) &= \frac{8 + 198}{10^3} \\
&= \frac{206}{10^3} \\
&= 0.206 \\
P(Tp|Ha) &= 0.80
\end{aligned}
$$

Thus, our calculated probability that our friend actually has the anomaly is:

$$\begin{aligned} P(Ha|Tp) &= \frac{P(Tp|Ha) * P(Ha)}{P(Tp)} \\ &= \frac{0.80 * 10^{-2}}{0.206} \\ &= \frac{8 * 10^{-3}}{2.06 * 10^{-1}} \\ &= 3.883495 * 10^{-2} \\ &< .04 \end{aligned}$$

In other words, one who has tested *positive*, with a test that is 80% correct, has less that one chance in 25 of actually having this condition. (Imagine for a moment, for example, that this is a drug test being used on employees of some corporation . . .)

- We could ask the same kinds of questions we asked before:

  1. How accurate would the test have to be to get a better than 50% chance of actually having the condition when testing positive?

     (99%)

  2. For an 80% accurate test, how frequent would the condition have to be to get a better than 50% chance?

     (1 in 5)

- Some questions:

  1. Are these examples realistic? If not, why not?

  2. What sorts of things could we do to improve our results?

  3. Would it help to repeat the test? For example, if the probability of a false positive is 1 in 100, would that mean that the probability of two false positives on the same person would be 1 in 10,000 ($\frac{1}{100} * \frac{1}{100}$)? If not, why not?

  4. In the case of a medical condition such as a genetic anomaly, it is likely that the test would not be applied randomly, but would only be ordered if there were other symptoms suggesting the anomaly. How would this affect the results?

- Another example:
  Suppose that Tom, having had too much time on his hands while an undergraduate Philosophy major, through much practice at prestidigitation, got to the point where if he flipped a coin, his flips would have the probabilities:

$$P(h) = 0.7,\ P(t) = 0.3.$$

  Now suppose further that you are brought into a room with 10 people in it, including Tom, and on a table is a coin showing heads. You are told further that one of the 10 people was chosen at random, that the chosen person flipped the coin and put it on the table, and that research shows that the overall average for the 10 people each flipping coins many times is:

$$P(h) = 0.52,\ P(t) = 0.48.$$

  What is the probability that it was Tom who flipped the coin?

By Bayes' Theorem, we can calculate:

$$\begin{aligned} P(\text{Tom}|h) &= \frac{P(h|\text{Tom})P(\text{Tom})}{P(h)} = \frac{0.7 * 0.1}{0.52} \\ &= 0.1346. \end{aligned}$$

Note that this estimate revises our *a priori* estimate of the probability of Tom being the flipper up from 0.10.

This process (revising estimated probability) of course depends in a critical way on having *a priori* estimates in the first place . . .

Analog channels ←

- The part of Shannon's work we have looked at so far deals with discrete (or digital) signalling systems. There are related ideas for continuous (or analog) systems. What follows gives a brief hint of some of the ideas, without much detail.

- Suppose we have a signalling system using band-limited signals (i.e., the frequencies of the transmissions are restricted to lie within some specified range). Let us call the bandwidth W. Let us further assume we are transmitting signals of duration T. In order to reconstruct a given signal, we will need $2WT$ samples of the signal. Thus, if we are sending continuous signals, each signal can be represented by $2WT$ numbers x_i, taken at equal intervals.

We can associate with each signal an *energy*, given by:

$$E = \frac{1}{2W} \sum_{i=1}^{2WT} x_i^2.$$

The distance of the signal (from the origin) will be

$$r = \left(\sum x_i^2 \right)^{1/2} = (2WE)^{1/2}$$

We can define the *signal power* to be the average energy:

$$S = \frac{E}{T}.$$

Then the radius of the sphere of transmitted signals will be:

$$r = (2WST)^{1/2}.$$

Each signal will be disturbed by the noise in the channel. If we measure the power of the noise N added by the channel, the disturbed signal will lie in a sphere around the original signal of radius $(2WNT)^{1/2}$.

Thus the original sphere must be enlarged to a larger radius to enclose the disturbed signals. The new radius will be:

$$r = (2WT(S+N))^{1/2}.$$

In order to use the channel effectively and minimize error (misreading of signals), we will want to put the signals in the sphere, and separate them as much as possible (and have the distance between the signals at least twice what the noise contributes . . .). We thus want to divide the sphere up into sub-spheres of radius $= (2WNT)^{1/2}$. From this, we can get an upper bound on the number M of possible messages that we can reliably distinguish. We can use the formula for the volume of an n-dimensional sphere:

$$V(r,n) = \frac{\pi^{n/2} r^n}{\Gamma(n/2+1)}.$$

We have the bound:

$$\begin{aligned} M &\leq \frac{\pi^{WT}\left(2WT(S+N)\right)^{WT}}{\Gamma(WT+1)} \frac{\Gamma(WT+1)}{\pi^{WT}(2WTN)^{WT}} \\ &= \left(1+\frac{S}{N}\right)^{WT} \end{aligned}$$

The information sent is the log of the number of messages sent (assuming they are equally likely), and hence:

$$I = \log(M) = WT * \log\left(1+\frac{S}{N}\right),$$

and the rate at which information is sent will be:

$$W * \log\left(1+\frac{S}{N}\right).$$

We thus have the usual *signal/noise* formula for channel capacity . . .

- An amusing little side light: "Random" band-limited natural phenoma typically display a power spectrum that obeys a power law of the general form $\frac{1}{f^{\alpha}}$. On the other hand, from what we have seen, if we want to use a channel optimally, we should have essentially equal power at all frequencies in the band. This means that a possible way to engage in SETI (the search for extra-terrestrial intelligence) will be to look for bands in which there is white noise! White noise is likely to be the signature of (intelligent) optimal use of a channel . . .

A Maximum Entropy Principle ←

- Suppose we have a system for which we can measure certain macroscopic characteristics. Suppose further that the system is made up of many microscopic elements, and that the system is free to vary among various states. Given the discussion above, let us assume that with probability essentially equal to 1, the system will be observed in states with maximum entropy.

  We will then sometimes be able to gain understanding of the system by applying a *maximum information entropy* principle (MEP), and, using Lagrange multipliers, derive formulae for aspects of the system.

- Suppose we have a set of macroscopic measurable characteristics f_k, $k = 1, 2, \ldots, M$ (which we can think of as constraints on the system), which we assume are related to microscopic characteristics via:

$$\sum_i p_i * f_i^{(k)} = f_k.$$

Of course, we also have the constraints:

$$p_i \geq 0, \text{ and}$$

$$\sum_i p_i = 1.$$

We want to maximize the entropy, $\sum_i p_i \log(1/p_i)$, subject to these constraints. Using Lagrange multipliers λ_k (one for each constraint), we have the general solution:

$$p_i = \exp\left(-\lambda - \sum_k \lambda_k f_i^{(k)}\right).$$

If we define Z, called the partition function, by

$$Z(\lambda_1, \ldots, \lambda_M) = \sum_i \exp\left(-\sum_k \lambda_k f_i^{(k)}\right),$$

then we have $e^{\lambda} = Z$, or $\lambda = \ln(Z)$.

Application: Economics I (a Boltzmann Economy) ←

- Our first example here is a very simple economy. Suppose there is a fixed amount of money (M dollars), and a fixed number of agents (N) in the economy. Suppose that during each time step, each agent randomly selects another agent and transfers one dollar to the selected agent. An agent having no money doesn't go in debt. What will the long term (stable) distribution of money be?

  This is not a very realistic economy – there is no growth, only a redistribution of money (by a random process). For the sake of argument, we can imagine that every agent starts with approximately the same amount of money, although in the long run, the starting distribution shouldn't matter.

- For this example, we are interested in looking at the distribution of money in the economy, so we are looking at the probabilities $\{p_i\}$ that an agent has the amount of money i. We are hoping to develop a model for the collection $\{p_i\}$.

  If we let n_i be the number of agents who have i dollars, we have two constraints:

$$\sum_i n_i * i = M$$

  and

$$\sum_i n_i = N.$$

  Phrased differently (using $p_i = \frac{n_i}{N}$), this says

$$\sum_i p_i * i = \frac{M}{N}$$

  and

$$\sum_i p_i = 1.$$

- We now apply Lagrange multipliers:

$$L = \sum_i p_i \ln(1/p_i) \;-\; \lambda \left[\sum_i p_i * i - \frac{M}{N} \right] \;-\; \mu \left[\sum_i p_i - 1 \right],$$

from which we get

$$\frac{\partial L}{\partial p_i} = -[1 + \ln(p_i)] - \lambda i - \mu = 0.$$

We can solve this for p_i:

$$\ln(p_i) = -\lambda i - (1 + \mu)$$

and so

$$p_i = e^{-\lambda_0} e^{-\lambda i}$$

(where we have set $1 + \mu \equiv \lambda_0$).

- Putting in constraints, we have

$$\begin{aligned} 1 &= \sum_i p_i \\ &= \sum_i e^{-\lambda_0} e^{-\lambda i} \\ &= e^{-\lambda_0} \sum_{i=0}^{M} e^{-\lambda i}, \end{aligned}$$

and

$$\begin{aligned} \frac{M}{N} &= \sum_i p_i * i \\ &= \sum_i e^{-\lambda_0} e^{-\lambda i} * i \\ &= e^{-\lambda_0} \sum_{i=0}^{M} e^{-\lambda i} * i. \end{aligned}$$

We can approximate (for large M)

$$\sum_{i=0}^{M} e^{-\lambda i} \approx \int_0^M e^{-\lambda x} dx \approx \frac{1}{\lambda},$$

and

$$\sum_{i=0}^{M} e^{-\lambda i} * i \approx \int_0^M x e^{-\lambda x} dx \approx \frac{1}{\lambda^2}.$$

From these we have (approximately)

$$e^{\lambda_0} = \frac{1}{\lambda}$$

and

$$e^{\lambda_0}\frac{M}{N} = \frac{1}{\lambda^2}.$$

From this, we get

$$\lambda = \frac{N}{M} = e^{-\lambda_0},$$

and thus (letting $T = \frac{M}{N}$) we have:

$$\begin{aligned} p_i &= e^{-\lambda_0}e^{-\lambda i} \\ &= \frac{1}{T}e^{-\frac{i}{T}}. \end{aligned}$$

This is a Boltzmann-Gibbs distribution, where we can think of T (the average amount of money per agent) as the "temperature," and thus we have a "Boltzmann economy" . . .

Note: this distribution also solves the functional equation

$$p(m_1)p(m_2) = p(m_1 + m_2).$$

- This example, and related topics, are discussed in

  Statistical mechanics of money

  by Adrian Dragulescu and Victor M. Yakovenko,

  http://arxiv.org/abs/cond-mat/0001432

  and

  Statistical mechanics of money: How saving propensity affects its distribution

  by Anirban Chakraborti and Bikas K. Chakrabarti

  http://arxiv.org/abs/cond-mat/0004256

Application: Economics II (a power law) ←

- Suppose that a (simple) economy is made up of many agents a, each with wealth at time t in the amount of $w(a,t)$. (I'll leave it to you to come up with a reasonable definition of "wealth" – of course we will want to make sure that the definition of "wealth" is applied consistently across all the agents.) We can also look at the total wealth in the economy $W(t) = \sum_a w(a,t)$.

  For this example, we are interested in looking at the distribution of wealth in the economy, so we will assume there is some collection $\{w_i\}$ of possible values for the wealth an agent can have, and associated probabilities $\{p_i\}$ that an agent has wealth w_i. We are hoping to develop a model for the collection $\{p_i\}$.

- In order to apply the maximum entropy principle, we want to look at global (aggregate/macro) observables of the system that reflect (or are made up of) characteristics of (micro) elements of the system.

  For this example, we can look at the growth rate of the economy. A reasonable way to think about this is to let $R_i = w_i(t_1)/w_i(t_0)$ and $R = W(t_1)/W(t_0)$ (where t_0 and t_1 represent time steps of the economy). The growth rate will then be $\ln(R)$. We then have the two constraints on the p_i:

$$\sum_i p_i * \ln(R_i) = \ln(R)$$

  and

$$\sum_i p_i = 1.$$

- We now apply Lagrange multipliers:

$$L = \sum_i p_i \ln(1/p_i) \;-\; \lambda \left[\sum_i p_i \ln(R_i) - \ln(R) \right] \;-\; \mu \left[\sum_i p_i - 1 \right],$$

from which we get

$$\frac{\partial L}{\partial p_i} = -[1 + \ln(p_i)] - \lambda \ln(R_i) - \mu = 0.$$

We can solve this for p_i:

$$p_i = e^{-\lambda_0} e^{-\lambda \ln(R_i)} = e^{-\lambda_0} R_i^{-\lambda}$$

(where we have set $1 + \mu \equiv \lambda_0$).

Solving, we get $\lambda_0 = \ln(Z(\lambda))$, where $Z(\lambda) \equiv \sum_i R_i^{-\lambda}$ (the partition function) normalizes the probability distribution to sum to 1. From this we see the power law (for $\lambda > 1$):

$$p_i = \frac{R_i^{-\lambda}}{Z(\lambda)}.$$

- We might actually like to calculate specific values of λ, so we will do the process again in a continuous version. In this version, we will let $R = w(T)/w(0)$ be the relative wealth at time T. We want to find the probability density function $f(R)$, that is:

$$\max_{\{f\}} H(f) = -\int_1^\infty f(R)\ln(f(R))dR,$$

subject to

$$\begin{aligned}\int_1^\infty f(R)dR &= 1,\\ \int_1^\infty f(R)\ln(R)dR &= C\ln(R),\end{aligned}$$

where C is the average number of transactions per time step.

We need to apply the calculus of variations to maximize over a class of functions.

When we are solving an extremal problem of the form

$$\int F[x, f(x), f'(x)]dx,$$

we work to solve

$$\frac{\partial F}{\partial f(x)} - \frac{d}{dx}\left(\frac{\partial F}{\partial f'(x)}\right) = 0.$$

Our Lagrangian is of the form

$$\begin{aligned} L \equiv & - \int_1^\infty f(R)\ln(f(R))dr - \mu\left(\int_1^\infty f(R)dR - 1\right) \\ & - \lambda\left(\int_1^\infty f(R)\ln(R)dR - C * \ln(R)\right). \end{aligned}$$

Since this does not depend on $f'(x)$, we look at:

$$\frac{\partial[-f(R)\ln f(R) - \mu(f(R)-1) - \lambda(f(R)\ln R - R)]}{\partial f(R)}$$

$$= 0$$

from which we get

$$f(R) = e^{-(\lambda_0 - \lambda\ln(R))} = R^{-\lambda}e^{-\lambda_0},$$

where again $\lambda_0 \equiv 1 + \mu$.

We can use the first constraint to solve for e^{λ_0}:

$$e^{\lambda_0} = \int_1^\infty R^{-\lambda} dR = \left[\frac{R^{-\lambda+1}}{1-\lambda}\right]_1^\infty = \frac{1}{\lambda - 1},$$

assuming $\lambda > 1$. We therefore have a power law distribution for wealth of the form:

$$f(R) = (\lambda - 1)R^{-\lambda}.$$

To solve for λ, we use:

$$C * \ln(R) = (\lambda - 1)\int_1^\infty R^{-\lambda} \ln(R) dR.$$

Using integration by parts, we get

$$\begin{aligned} C * \ln(R) &= (\lambda - 1)\left[\ln(R)\frac{R^{1-\lambda}}{1-\lambda}\right]_1^\infty \\ &\quad -(\lambda - 1)\int_1^\infty \frac{R^{-\lambda}}{1-\lambda} dR \\ &= (\lambda - 1)\left[\ln(R)\frac{R^{1-\lambda}}{1-\lambda}\right]_1^\infty + \left[\frac{R^{1-\lambda}}{1-\lambda}\right]_1^\infty. \end{aligned}$$

By L'Hôpital's rule, the first term goes to zero as $R \to \infty$, so we are left with

$$C * \ln(R) = \left[\frac{R^{1-\lambda}}{1-\lambda}\right]_1^{\infty} = \frac{1}{\lambda - 1},$$

or, in other terms,

$$\lambda - 1 = C * \ln(R^{-1}).$$

For much more discussion of this example, see the paper *A Statistical Equilibrium Model of Wealth Distribution* by Mishael Milakovic, February, 2001, available on the web at:

http://astarte.csustan.edu/~ tom/SFI-CSSS/Wealth/wealth-Milakovic.pdf

Application to Physics (lasers) ←

- We can also apply this maximum entropy principle to physics examples. Here is how it looks applied to a single mode laser. For a laser, we will be interested in the intensity of the light emitted, and the coherence property of the light will be observed in the second moment of the intensity. The electric field strength of such a laser will have the form

$$E(x,t) = E(t)\sin(kx),$$

and E(t) can be decomposed in the form

$$E(t) = Be^{-i\omega t} + B^* e^{i\omega t}.$$

If we measure the intensity of the light over time intervals long compared to the frequency, but small compared to fluctuations of $B(t)$, the output will be

proportional to BB^* and to the loss rate, 2κ, of the laser:

$$I = 2\kappa BB^*.$$

The intensity squared will be

$$I^2 = 4\kappa^2 B^2 B^{*2}.$$

- If we assume that B and B^* are continuous random variables associated with a stationary process, then the information entropy of the system will be:

$$H = \int p(B, B^*) \log \left(\frac{1}{p(B, B^*)} \right) d^2B.$$

The two constraints on the system will be the averages of the intensity and the square of the intensity:

$$\begin{aligned} f_1 &= < 2\kappa B B^* >, \\ f_2 &= < 4\kappa^2 B^2 B^{*2} > . \end{aligned}$$

Then, of course, we will let

$$\begin{aligned} f^{(1)}_{B,B^*} &= 2\kappa B B^*, \\ f^{(2)}_{B,B^*} &= 4\kappa^2 B^2 B^{*2}. \end{aligned}$$

We can now use the method outlined above, finding the maximum entropy general solution derived via Lagrange multipliers for this system.

- Applying the general solution, we get:

  $p(B, B^*) = \exp\left[-\lambda - \lambda_1 2\kappa BB^* - \lambda_2 4\kappa^2 (BB^*)^2\right],$

  or, in other notation:

  $$p(B, B^*) = N * \exp(-\alpha|B|^2 - \beta|B|^4).$$

  This function in laser physics is typically derived by solving the Fokker-Planck equation belonging to the Langevin equation for the system.

- For quick reference, the typical generic Langevin equation looks like:

  $$\dot{\mathbf{q}} = K(\mathbf{q}) + \mathbf{F}(t)$$

  where $\mathbf{q}$ is a state vector, and the fluctuating forces $F_j(t)$ are typically assumed to have

  $$\begin{aligned} < F_j(t) > &= 0 \\ < F_j(t) F_{j'}(t') > &= Q_j \delta_{jj'} \delta(t - t'). \end{aligned}$$

- The associated generic Fokker-Planck equation for the distribution function $f(q,t)$ then looks like:

$$\frac{\partial f}{\partial t} = -\sum_j \frac{\partial}{\partial q_j}(K_j f) + \frac{1}{2}\sum_{jk} Q_{jk}\frac{\partial^2}{\partial q_j \partial q_k} f.$$

  The first term is called the drift term, and the second the diffusion term. This can typically be solved only for special cases ...

- For much more discussion of these topics, I can recommend the book *Information and Self-organization, A Macroscopic Approach to Complex Systems* by Hermann Haken, Springer-Verlag Berlin, New York, 1988.

Kullback-Leibler information measure ←

- Suppose we have a data set, and we would like to build a (statistical) model for the data set. How can we tell how good a job our model does in representing the statistical properties of the data set? One approach is to use ideas from Information Theory (and in particular the framework of the Gibbs inequality).

  So, suppose we have a data set for which the actual statistical distribution is given by $P = p(x)$. We propose a model $Q = q(x)$ for the data set (a traditional example would be to use a least-squares line fit for Q). We would like a measure which can tell us something about how well our model matches the actual distribution.

- One approach is to use the so-called Kullback-Leibler information measure:

$$\begin{aligned} KL(P;Q) &= \left\langle \log\left(\frac{p(x)}{q(x)}\right)\right\rangle_P \\ &= \int_{-\infty}^{\infty} \log\left(\frac{p(x)}{q(x)}\right) p(x)d(x) \end{aligned}$$

(in other words, the P-expected value of the difference of the logs). The KL measure has the nice properties that

$$\begin{aligned} KL(P;Q) &>= 0, and \\ KL(P;Q) &= 0 \iff p(x) = q(x)\ (a.e.) \end{aligned}$$

(I'll leave it to you to specialize to the discrete case . . .)

The KL measure is sometimes also called the *relative entropy*, although that term might better be used for $-KL(P;Q)$, in which case minimizing the KL measure would be the same as maximizing relative entropy. The notation in the literature is sometimes inconsistent on this point.

I should probably also mention that the KL measure is not a true metric (it is not symmetric in P and Q, nor does it satisfy the triangle inequality), but it can be a useful measure of the "distance" between two distributions.

One approach to understanding the KL measure is consider things relative to the entropy of the distribution P. Thinking in the discrete case, we have

$$\begin{aligned} 0 \;<= \;& KL(P;Q) \\ = \;& \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) \\ = \;& \sum_x p(x) \log\left(\frac{1}{q(x)}\right) - \sum_x p(x) \log\left(\frac{1}{p(x)}\right) \\ = \;& H(P;Q) - H(P) \end{aligned}$$

(where $H(P;Q)$ is what is sometimes called the "cross entropy" between P and Q). In other words, the entropy of the "true" distribution P ($H(P)$) is a lower bound for the cross entropy. As we saw

elsewhere, $H(P)$ is a lower bound on efficiency of encoding (a description of) the data set. The Kullback-Leibler measure can be thought of as the (added) inefficiency of encoding the data with respect to the distribution Q, rather than the "true" distribution P.

- Now, suppose that our data set is a sample from the distribution P, and we would like to estimate P. We can (with care . . .) sometimes use the KL measure to compare various candidate distributions even without knowing P itself.
  Considering the discrete case (i.e., a finite sample size), we have (as above)

$$\begin{aligned} KL(P;Q) &= \sum_x p(x) \log\left(\frac{1}{q(x)}\right) - H(P) \\ &= -\sum_x p(x) \log(q(x)) - H(P) \end{aligned}$$

Thus, we can minimize the KL measure by maximizing

$$\sum_x p(x) \log(q(x)) = \langle \log(q(x)) \rangle_P$$

which is often called the *expected log-likelihood*.

Now, if we are feeling lucky (or at least brave :-) we could try maximizing the *expected* log-likelihood by maximizing the *estimated* log-likelihood – i.e., by maximizing

$$\sum_x \log(q(x)).$$

There are a variety of subtleties in this. Some approaches involve estimating the bias involved in using the estimated log-likelihood instead of the expected log-likelihood. Perhaps another time or place there can be more discussion of these issues.

But, just for kicks, let's look at one specific example. Suppose we have reason

to believe that P is actually a normal distribution with mean m and variance 1. From a sample, we want to estimate m. We will want to compare various normal distributions

$$\begin{aligned} Q(\mu) &= q(x, \mu) \\ &= \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2}\right)}. \end{aligned}$$

The corresponding log-likelihood function will be

$$L(\mu) = -\frac{N}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^2.$$

In other words, maximizing the log-likelihood function is the same as minimizing the least-squares function

$$ls(\mu) = \sum_{i=1}^{N} (x_i - \mu)^2.$$

Oh, well. Enough of this for now . . .

References

[1] Bar-Yam, Yaneer, *Dynamics of Complex Systems (Studies in Nonlinearity)* , Westview Press, Boulder, 1997.

[2] Brillouin, L., *Science and information theory* Academic Press, New York, 1956.

[3] Brooks, Daniel R., and Wiley, E. O., *Evolution as Entropy*, Toward a Unified Theory of Biology, Second Edition, University of Chicago Press, Chicago, 1988.

[4] Campbell, Jeremy, *Grammatical Man*, Information, Entropy, Language, and Life, Simon and Schuster, New York, 1982.

[5] Cover, T. M., and Thomas J. A., *Elements of Information Theory,* John Wiley and Sons, New York, 1991.

[6] DeLillo, Don, *White Noise*, Viking/Penguin, New York, 1984.

[7] Feller, W., *An Introduction to Probability Theory and Its Applications*, Wiley, New York,1957.

[8] Feynman, Richard, *Feynman lectures on computation*, Addison-Wesley, Reading, 1996.

[9] Gatlin, L. L., *Information Theory and the Living System*, Columbia University Press, New York, 1972.

[10] Greven, A., Keller, G., Warnecke, G., *Entropy*, Princeton Univ. Press, Princeton, 2003.

[11] Haken, Hermann, *Information and Self-Organization, a Macroscopic Approach to Complex Systems*, Springer-Verlag, Berlin/New York, 1988.

[12] Hamming, R. W., Error detecting and error correcting codes, *Bell Syst. Tech. J.* **29** 147, 1950.

[13] Hamming, R. W., *Coding and information theory*, 2nd ed, Prentice-Hall, Englewood Cliffs, 1986.

[14] Hill, R., *A first course in coding theory* Clarendon Press, Oxford, 1986.

[15] Hodges, A., *Alan Turing: the enigma* Vintage, London, 1983.

[16] Hofstadter, Douglas R., *Metamagical Themas: Questing for the Essence of Mind and Pattern*, Basic Books, New York, 1985

[17] Jones, D. S., *Elementary information theory* Clarendon Press, Oxford, 1979.

[18] Knuth, Eldon L., *Introduction to Statistical Thermodynamics*, McGraw-Hill, New York, 1966.

[19] Landauer, R., Information is physical, *Phys. Today*, May 1991 23-29.

[20] Landauer, R., The physical nature of information, *Phys. Lett. A*, **217** 188, 1996.

[21] van Lint, J. H., *Coding Theory*, Springer-Verlag, New York/Berlin, 1982.

[22] Lipton, R. J., Using DNA to solve NP-complete problems, *Science*, **268** 542–545, Apr. 28, 1995.

[23] MacWilliams, F. J., and Sloane, N. J. A., *The theory of error correcting codes*, Elsevier Science, Amsterdam, 1977.

[24] Martin, N. F. G., and England, J. W., *Mathematical Theory of Entropy*, Addison-Wesley, Reading, 1981.

[25] Maxwell, J. C., *Theory of heat* Longmans, Green and Co, London, 1871.

[26] von Neumann, John, Probabilistic logic and the synthesis of reliable organisms from unreliable components, in *automata studies(Shanon,McCarthy eds)*, 1956 .

[27] Papadimitriou, C. H., *Computational Complexity*, Addison-Wesley, Reading, 1994.

[28] Pierce, John R., *An Introduction to Information Theory – Symbols, Signals and Noise*, (second revised edition), Dover Publications, New York, 1980.

[29] Roman, Steven, *Introduction to Coding and Information Theory*, Springer-Verlag, Berlin/New York, 1997.

[30] Sampson, Jeffrey R., *Adaptive Information Processing, an Introductory Survey*, Springer-Verlag, Berlin/New York, 1976.

[31] Schroeder, Manfred, *Fractals, Chaos, Power Laws, Minutes from an Infinite Paradise*, W. H. Freeman, New York, 1991.

[32] Shannon, C. E., A mathematical theory of communication *Bell Syst. Tech. J.* **27** 379; also p. 623, 1948.

[33] Slepian, D., ed., *Key papers in the development of information theory* IEEE Press, New York, 1974.

[34] Turing, A. M., On computable numbers, with an application to the Entscheidungsproblem, *Proc. Lond. Math. Soc. Ser. 2* **42**, 230 ; see also *Proc. Lond. Math. Soc. Ser. 2* **43**, 544, 1936.

[35] Zurek, W. H., Thermodynamic cost of computation, algorithmic complexity and the information metric, *Nature* **341** 119-124, 1989.

To top ←