4. Introduction:

Nowadays, manipulating daily loads of info: images, text, sounds = files = 0.1 MB
On which we are performing:

- compression
- smaller file
- transmission over noisy channel
- error correction

These two problems were already formalized and approximately solved by
Shannon '48 "A mathematical theory of communication".

Why should we care about this while not being telecommunication engineer?

- biological: DNA → RNA → proteins
- visual: receptors → retina → optimal nerve → brain
- social: language, etc.

-only some combination of sounds make up words

Theoretically:

- Information theory enlightens statistical mechanics.
- Faulty interdisciplinary of stat mech of disordered systems.
- Constraint satisfaction problems for error correction. LDA


2. The Meaning of Entropy:

A definition: S_mic = k ln Ω

S_c = k Σ p(c) ln p(c) with p(c) = \frac{1}{Ω} e^{-β H(c)}

Here it is:

Shannon entropy per distribution \rho = \{ p(x) \} x \in X

\mathcal{H}(\rho) = -\Sigma_{x \in X} p(x) \log_2 p(x)

\text{prop: } \mathcal{H}(\rho) = 0 \iff p(x) = \delta_{x=x_0} \text{ maximally concentrated on one value}

\mathcal{H}(\rho) \in [0, \log_2 |X|] \text{ and } \mathcal{H}(\rho) = \log_2 (|X|) \Rightarrow p(x) = \frac{1}{|X|} \forall x

"The entropy measures the lack of information on a realization of \rho"

Yes, there are many functions growing between uniform and concentrated,
so why should it be \mathcal{H}(\rho) with the logarithm?

Definition for binomial (p): \mathcal{H}(p) = -p \log_2 p - (1-p) \log_2 (1-p)

dB guessing game:

- Amy chooses x \in \{ \text{YES, NO} \} with p = \mathcal{H}(p)
- Sheiloh has to guess r asking YES/NO questions as fast as possible.

A strategy:

1. \{ B \} → 1/2
2. \{ A, B \} → 1/2
Formalization: Let strategy = tree of questions, \( \ell_a(T) = \# \) questions to find \( x \in X \) with strategy \( T \).

To define the best strategy, we compute
\[
\mathbb{E}(T) = \sum_{x \in X} p(x) \ell_a(T) = \min_{T^*} \mathbb{E}(T),
\]
which is bounded by \( H(p) \leq \mathbb{E}(T) \leq H(p) + 1 \).

Kraft inequality: \( \forall T, \sum_{x \in X} 2^{-\ell_a(T)} \leq \Delta \),
\[
\ell_{\max} = \max_{x \in X} \ell_a(T).
\]

Kullback-Leibler divergence: \( \rho, q \), Gibbs laws on \( X \),
\[
D(p || q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)} \geq 0 \quad (\text{due to the concavity of the log}).
\]

Proof:
\[
D(p || q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)} = -H(p) + \log_2 \Delta + \mathbb{E}(T) \geq 0.
\]

Consider \( q(x) = \frac{2^{-\ell_a}}{\sum_{x \in X} 2^{-\ell_a}} \). with \( \sum_{x \in X} 2^{-\ell_a} \leq 1 \) by Kraft.

\[
D(p || q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)} = -H(p) + \log_2 \Delta + \mathbb{E}(T) \geq 0.
\]

If \( \{\ell_a\}_x \) integers s.t. \( \sum_{x \in X} 2^{-\ell_a} \leq 1 \), then \( T^* \) such that \( \ell_a(T^*) = \ell_a \).

If \( \ell_a \leq \ell_b < \ell_2 \) in the leftmost node of depth \( \ell_1 \),
\( \ell_2 \) in the remaining leftmost node of depth \( \ell_2 \).

We would like to take \( \ell_a = \left\lceil \log_2 \frac{A}{p(x)} \right\rceil \) obey Kraft inequality,
\[
\mathbb{E}(T) \leq \sum_{x \in X} p(x) \left[ \log_2 p(x) + 1 \right] \leq H(p) + 1.
\]

2C data compression:
Consider a file of strings of symbols in \( X \): \( x_1, x_2, x_3 \), \( \cdots \) as short as possible.

It is actually the problem we have solved:
\[
\begin{array}{c}
100101 \cdots \\
\end{array}
\]

Let's formalize that with \( w \): \( x_1 \to w_1 \),

...
If \( x_i \)'s are iid with prob \( p(x) \), need \( nH(p) \) bits to compress \( x_1, \ldots, x_n \).

"Entropy is the best rate of compression possible."

2D mutual information

An inference problem: q ens of random \( X \) from observation \( Y \).

formally: the pair of r.v \( (X,Y) \in (X \times X') \)
distributed according to \( p_{x'y'}(x,y) = p(x = x, Y = y) \)

\[
\begin{align*}
\rightarrow & \text{ marginal laws: } p_x(x) = \sum_y p_{x'y}(x,y), \quad p_y(y) = \sum_x p_{x'y}(x,y) \\
\rightarrow & \text{ conditional laws: } p_{x'y}(x|y) = \frac{p_{x'y}(x,y)}{p_y(y)}
\end{align*}
\]

From which we consider various interesting entropies:

* joint law: entropy \( H(X,Y) = -\sum_{x,y} p_{x'y}(x,y) \log_2 p_{x'y}(x,y) \)

* entropy of marginal: \( H(X) = -\sum_x p_x(x) \log_2 p_x(x) \)

* conditional ent.: \( H(X|Y) = -\sum_y p_y(y) \sum_x p_{x'y}(x|y) \log_2 p_{x'y}(x|y) \)

\[
= -\sum_{x,y} p_{x'y}(x,y) \log_2 p_{x'y}(x,y)
\]

\[= H(X) + H(Y) - H(X,Y) \rightarrow \text{prop: } H(X,Y) \leq H(X) + H(Y) \]

\[
= H(X) + H(Y) - H(X|Y) \rightarrow \text{prop: } H(X|Y) \leq H(Y)
\]

\[ \text{mutual info. } \quad I(X,Y) = D(\mathbb{P}_{x'y} \parallel \mathbb{P}_{x'y}) \rightarrow \text{prop: } I(X,Y) = 0 \Rightarrow X,Y \text{ indep.} \]

\[
= \sum_{x,y} p_{x'y}(x,y) \log \frac{p_{x'y}(x,y)}{p_x(x)p_y(y)}
\]

\[
= H(X) + H(Y) - H(X,Y) \rightarrow \text{prop: } H(X,Y) \leq H(X)
\]


EXERCISES:
- Show that Gibbs maximizes entropy under constraint \( E = U \)
- Prove that uniform distribution maximizes entropy under no constraint (KL div)
- SHANNON'S DECODER 
- Appendix A:

\[
H(p_1, \ldots, p_n) \text{ continuous in } p_i \\
H(\frac{1}{n}, \ldots, \frac{1}{n}) \text{ growing with } n \\
\Rightarrow H = -\sum_{i=1}^n p_i \log p_i
\]

\[
\begin{pmatrix}
\begin{array}{c}
p_1 \\
p_2 \\
p_3
\end{array}
\end{pmatrix}
\rightarrow
\begin{pmatrix}
\frac{p_1}{p_2p_3} \\
\frac{p_2}{p_1p_3} \\
\frac{p_3}{p_1p_2}
\end{pmatrix}
+ H
\]

3 COMMUNICATION OVER NOISY CHANNELS

3A: definitions: message \( \rightarrow \) noise \( \rightarrow \) corrupted message.

Examples:
- Binary Erasure Channel (BEC)
  \( \varepsilon \) = probability of erasure.

\[ \begin{array}{ccc}
\otimes & \varepsilon & O \\
\otimes & O & \varepsilon
\end{array} \]

- Binary Symmetric Channel (BSC)

\[ \begin{array}{ccc}
\otimes & \varepsilon & O \\
\otimes & O & \varepsilon
\end{array} \]
Capacity of a channel: 
\[ C = \max_{X} I(X; Y) \] 
with \( X \) input of a channel, \( Y \) output, \( p(x) \) probability law of input.

Exercises 1: Code = \( 4 - 2 \)
\[ C_{\text{BSC}} = 1 - \ln(p_f) \]

Encoding and decoding:
message \( \rightarrow \) encoded message \( \rightarrow \) noisy encoded message \( \rightarrow \) decoded message?

String of bits \( \rightarrow \) "longer string" 

Rate of a code: \( = \frac{\text{# bits of message}}{\text{# bits of encoded msg.}} \)

\( \lambda \) the larger the better to reduce the cost of message transmission.

3B. Naive coding = repetition

Encoder
\[
\begin{align*}
0 & \rightarrow 000 \quad \text{with the BSC } p_f < \frac{1}{2} \\
1 & \rightarrow 111
\end{align*}
\]

Reasonable Decoder

"majority rule"

\( \lambda/4 \) odd number of bits: 
001, 010, 100 \( \rightarrow 0 \)
110, 101, 011 \( \rightarrow 1 \)

How good is the naive coding? 
Probability of error without encoding = \( p_f \)
Probability of error with 3 repetitions = \( p_f^3 + 3p_f^2(1 - p_f) \)

\( \lambda \) rate = \( \frac{1}{3} \)

With encoding \( \rightarrow \) better than no encoding but \( \lambda > 0 \) as soon as \( p_f > 0 \)

rate = \( \frac{1}{3} \)

Is actually not that good. Can one do better?

3C. Shannon channel coding theorem

There exist encoding (of growing length) with \( \text{Pen} \rightarrow 0 \) for all rates \( R \) smaller than the capacity of the channel \( C(R < C) \).

So that we can now interpret the capacity as the best achievable rate with \( \text{Pen} \rightarrow 0 \).
The paradoxical statement that we could be sure of the signal despite the noise is resolved by the fact that this is an asymptotic statement (thermodynamic limit).

More formal definitions:

\( Z = (z_1, z_2) \)
\[ X = (x_1, x_2) \]

encoded msg.
\[ z \in \mathbb{Z}^4 \]
\[ x \in \{0, 1\}^2 \]

corrupted encoded msg.
\[ y \in \mathbb{Z}^4 \]
\[ x_{\text{out}} = (x_1, x_2) \]

encoded msg.
\[ z^{(y)} \]
\[ x^{(y)} \]

corrupted encoded msg.
\[ y^{(y)} \]
\[ x^{(y)} \]
encoding: \( x = \mathcal{F}(x) \)

code book: \( \mathcal{C} = \{ x^{(1)}, \ldots, x^{(r)} \} \)

channel: \( x^{(w)} \xrightarrow{\text{add}} y \)

Correlation: \( \hat{x}(y) \) estimation of \( x^{(w)} \)

decoding: \( \hat{x}(y) \mathcal{F}^{-1}(\hat{x}(y)) \)

From now on we will focus on the channel \( x \xrightarrow{\text{add}} y \xrightarrow{\hat{x}(y)} \). If the \( \hat{x}(y) \) are uniformly at random, the \( x \) are also \( u.a.m \).

**DECODING AS AN INFERENEE PROBLEM:**

\[ P_{\hat{x}|y}(x|y) = \frac{1}{y} \mathcal{N}(x|y, 1) \]

\[ P_{\hat{y}|x}(y|x) = \frac{1}{y} \mathcal{N}(y|x) \]

**BEC:**

\[ \begin{align*}
Q(0,1) &= \frac{1}{2} - E \quad Q(1,0) = E \\
Q(1,1) &= 0
\end{align*} \]

using Bayes theorem:

\[ P_{\hat{x}|y}(x|y) = \frac{1}{y} \mathcal{N}(x|y, 1) \]

\[ P_{\hat{x}(y)}(x|y) = \sum_{i=x^{(w)}} \mathcal{N}(x|y, 1) \]

A posterior probability

\[ P_{\hat{x}|y}(x|y) = \sum_{i=x^{(w)}} \mathcal{N}(x|y, 1) \]

**Symbol HAP:**

\[ \min \ E \left(d(\hat{x}, x)\right) \]

**Example with the BEC:**

\[ y = (0,0,0,1,0,0,0,0,0,0,0) \]

\[ P_{\hat{x}|y}(x|y) = \sum_{i=x^{(w)}} \mathcal{N}(x|y, 1) \]

**Shannon random code ensemble**

\[ \mathcal{E} = \{ x^{(1)}, \ldots, x^{(r)} \} \text{ with } \mathcal{F}(x) = \{ x^{(1)}, \ldots, x^{(r)} \} \text{ chosen at random, } x^{(w)} \] and \( 0,1 \) with probability \( \frac{1}{2} \).

**Remark:**

- How should we construct code books? Let's start with a simple construction proposed by Shannon.

**Correlation to the BEC:**

Analyses on the BEC:

Assume we received \( y \) that \( x^{(w)} \) has been sent. \( y = y^{(w)} \) on \( \text{in}(N,1-E) \) correctly transmitted.
Even the random codes achieve capacity. (Then we considered averages for codewords etc...)  

**Homework:** proof for the BSC (text on website)  

It we prove that there exist such codes, but we did not show that we could not do better. Ruse bound: if \( R > C \) \( \Rightarrow \) \( \text{Perr} > 0 \) coming from \( H(X|Y) > 0 \), Fano inequality.  

In practice, encoding and decoding the random code ensemble takes exponential time and memory. C has to be decoded by \( N^2 \) bits. The encoding is easy by looking up the table. But the decoding is NP hard as one need to look for the corresponding match in the table. The randomness limits any compression, we can only do exhaustive search in the table.  

4. **LOW DENSITY PARITY CHECKS CODES (LDPC).** Putting some structure.  

4A. Linear codes  
- \( \{0,1\}^n \) is a linear space over \( \mathbb{Z}_2 = \{0,1\} \) "modulo", addition mod 2, multiplication  
- The codebook \( C \subseteq \{0,1\}^n \) is said to be a linear code if \( C \) is a linear subspace of \( \{0,1\}^n \).  
  \[ z + y = (z_1 + y_1, z_2 + y_2, ..., z_n + y_n) \in C \]  
  \[ \text{mod} 2 \]  
  \[ z + y = 0 \in C \text{ (the origin belongs to the linear subspace).} \]  
\( \text{If } C \text{ is always a codeword of a linear code} \)  
\( \text{All codewords are equivalent (can always make gauge transformation to change position of origin).} \)