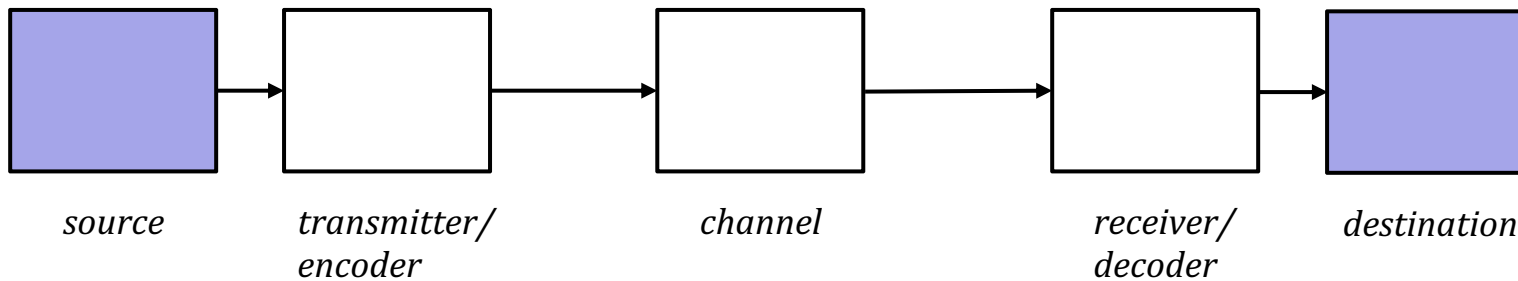


02. Shannon Information: Classical Information Theory

1. Key Concepts
2. Shannon Entropy S_{Shan}
3. Bits
4. What does S_{Shan} Measure?

1. Key Concepts

- Classical communication protocol (Shannon 1948)



Claude Shannon
(1916-2001)



Goal: "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point"

Data (stuff transmitted): values of a random variable

Def. 1 (*Random variable*). A **random variable** X consists of

- (a) a set of possible values $\{x_1, \dots, x_\ell\}$; and
- (b) a probability distribution $\{p_1, \dots, p_\ell\}$ that assigns a probability p_i to each value x_i .

- Note 1: Data $\{x_1, \dots, x_\ell\}$ can represent:
 - Letters that are used to construct a message.
 - Messages.
- Note 2: The *meaning* of the transmitted data is irrelevant to the goal.

"Frequently the messages have *meaning*... These semantic aspects of communication are irrelevant to the engineering problem."

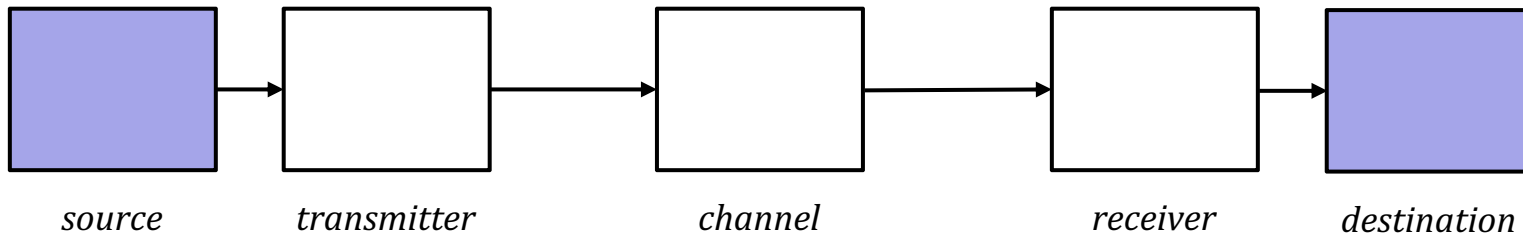


- Note 3: The probability distribution $\{p_1, \dots, p_\ell\}$ is an *essential* ingredient.

Def. 2 (*Average value*). The **average value** $\langle X \rangle$ of a random variable X is the weighted sum:

$$\langle X \rangle = \sum_{i=1}^{\ell} p_i x_i$$

↗
*Can't do classical
information theory
without it!*



Def. 3 (*Shannon information*). **Shannon information** is what is produced by a source that is required to be reproducible at a receiver if the transmission is to be counted a success.

(Timpson 2013)

- *Not just data!*
- *Not just data associated with a classical communication protocol?*
- *Material stuff?*
- *Abstract stuff?*

How is it measured?

Def. 4 (*Shannon information content*). Let X be a random variable with possible values $\{x_1, \dots, x_\ell\}$ and probability distribution $\{p_1, \dots, p_\ell\}$. The **Shannon information content** of the value x_i of X is given by $-\log_2 p_i$.

Shannon's intuition: The less probable a received piece of data x_i , the more surprising it is, and hence the more Shannon information it contains.

- *Assumption 1: The probability of a piece of data x_i is inversely proportional to its "surprise-content".*
- *Assumption 2: The surprise-content of a piece of data x_i is directly proportional to its Shannon information content.*

- So: The greater p_i is, the less Shannon information contained in x_i .
 - The greater p_i , the less surprising x_i is (Assumption 1).
 - So the less Shannon information x_i "contains" (Assumption 2).

Why the $\log_2$?

- *The log function is a way of comparing values $\{p_1, \dots, p_\ell\}$ that may range across a very large scale.*
- *The base 2 involves a conventional choice...*

2. Shannon Entropy S_{Shan}

Def. 5 (*Shannon entropy*). Let X be a random variable with possible values $\{x_1, \dots, x_\ell\}$ and probability distribution $\{p_1, \dots, p_\ell\}$. The **Shannon entropy** $S_{\text{Shan}}(X)$ of X is given by:

$$S_{\text{Shan}}(X) = -\sum_{i=1}^{\ell} p_i \log_2 p_i$$

- Note 1: $S_{\text{Shan}}(X)$ is the *average value* of the Shannon info content $-\log_2 p_i$ of x_i .

$S_{\text{Shan}}(X)$ is the average Shannon information gained upon determining a value of X .

2. Shannon Entropy S_{Shan}

Def. 5 (*Shannon entropy*). Let X be a random variable with possible values $\{x_1, \dots, x_\ell\}$ and probability distribution $\{p_1, \dots, p_\ell\}$. The **Shannon entropy** $S_{\text{Shan}}(X)$ of X is given by:

$$S_{\text{Shan}}(X) = -\sum_{i=1}^{\ell} p_i \log_2 p_i$$

- Note 2:

- $S_{\text{Shan}}(X)$ is zero when one p_i is 1 and the rest are 0.

$$\log_2 x = y \Rightarrow x = 2^y$$

$$\log_2 1 = y \Rightarrow 1 = 2^y \Rightarrow y = 0$$

*No surprises, no
Shannon info!*



"Thus only when we are certain of the outcome does $S_{\text{Shan}}(X)$ vanish. Otherwise $S_{\text{Shan}}(X)$ is positive."



- $S_{\text{Shan}}(X)$ takes its maximum value $\log_2 \ell$ when all the p_i 's are equal: $p_1 = p_2 = \dots = p_\ell = 1/\ell$.

*All values of X are
equally "surprising".*



"This is also intuitively the most uncertain situation."



Aside:

Claim. $S_{\text{Shan}}(X) = -\sum_{i=1}^{\ell} p_i \log_2 p_i$ takes its maximum value $\log_2 \ell$ when all the p_i 's are equal: $p_1 = p_2 = \dots = p_{\ell} = 1/\ell$.

Proof: $S_{\text{Shan}}(p_i)$ takes its max value for those p_i^* that satisfy $(d/dp_i)S_{\text{Shan}}(p_i^*) = 0$.

$$\begin{aligned}(d/dp_i)S_{\text{Shan}}(p_i) &= -\sum_{i=1}^{\ell} (d/dp_i)(p_i \log_2 p_i) \\ &= -\sum_{i=1}^{\ell} [\log_2 p_i + p_i(1/p_i)] \\ &= -\sum_{i=1}^{\ell} (\log_2 p_i + 1)\end{aligned}$$

- Now find values p_i^* that solve:

$$dS_{\text{Shan}}(p_i^*) = \sum_{i=1}^{\ell} (\log_2 p_i^* + 1) dp_i = 0 \quad \leftarrow \text{subject to constraint } \sum_i p_i = 1, \text{ or } \sum_i dp_i = 0$$

- Can add an arbitrary multiple of the constraint, $\alpha \sum_i dp_i$ and still get zero:

$$dS_{\text{Shan}} = \sum_{i=1}^{\ell} (\log_2 p_i^* + \alpha) dp_i = 0 \quad \text{or} \quad \log_2 p_i^* = -\alpha$$

- So:

$$p_i^* = 2^{-\alpha} \Rightarrow \sum_{i=1}^{\ell} p_i^* = \ell 2^{-\alpha} = 1 \Rightarrow \alpha = \log_2(\ell) \Rightarrow p_i^* = 1/\ell$$

- And: $S_{\text{Shan}}(p_i^*) = -\sum_{i=1}^{\ell} (1/\ell) \log_2(1/\ell) = -\ell(1/\ell) \log_2(1/\ell) = \log_2 \ell$

Note 3: Why the $\log_2$ in S_{Shan} ?

Claim. $S_{\text{Shan}}(X) = -\sum_i p_i \log_2 p_i$ is the unique function $H(X) : \{\text{probability distributions on } X\} \rightarrow \mathbb{R}$, that satisfies:

1. *Continuity.* $H(p_1, \dots, p_\ell)$ is continuous.
2. *Monotonicity.* Info increases with ℓ for uniform distributions: If $m > \ell$, then $H(Q) > H(P)$, for any $P = \{1/\ell, \dots, 1/\ell\}$ and $Q = \{1/m, \dots, 1/m\}$.
3. *Branching.* $H(p_1, \dots, p_\ell)$ is independent of how the process is divided into parts.
4. *Bit normalization.* The average Shannon info gained for two equally likely pieces of data is one bit: $H(1/2, 1/2) = 1$.

Bit renormalization requires $\log_2$

- Suppose: $X = \{x_1, x_2\}$, and $P = \{1/2, 1/2\}$.

- Then:

$$\begin{aligned} H(X) &= -(1/2 \log 1/2 + 1/2 \log 1/2) \\ &= \log 2 \\ &= 1 \quad \leftarrow \text{bit renormalization} \end{aligned}$$

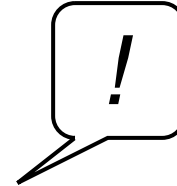
- And: $\log 2 = 1$ if and only if log is to base 2. $\leftarrow \log_2 x = y \Rightarrow x = 2^y$

Why call this "entropy"?

"Nobody really knows what entropy really is, so in a debate you will always have the advantage."



von Neumann



3. Bits



"The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called binary digits, or more briefly *bits*... A device with two stable positions, such as a relay or a flip-flop circuit, can store one bit of information."

Four senses of "bit"!

- (1) *Unit of measure sense*. A unit of Shannon information: One bit = average amount of Shannon information in a *binary random variable* (i.e., a random variable with just two possible values).
- (2) *Value sense*. The value of a binary random variable: 0 or 1. A "binary digit".
- (3) *Physical system sense*. A 2-state classical physical system. (A "relay or a flip-flop circuit".)
- (4) *Physical state sense*. A state of a 2-state classical system. (The "on" state or the "off" state.)

It's important to be consistent in our usage, particularly when it comes to comparing bits with "qubits"!

4. What does S_{Shan} Measure?

How can the average Shannon information content of a random variable X be understood?

Uncertainty Interpretation

The average Shannon information content of X is a measure of how uncertain X is.

- The greater $S_{\text{Shan}}(X)$, the greater the average Shannon info gained upon measuring X , the greater the uncertainty of its measured value.
- This interpretation is based on the following assumptions:

- Assumption 1: The probability of a piece of data x_i is inversely proportional to its "surprise-content".
- Assumption 2: The surprise-content of a piece of data x_i is directly proportional to its Shannon information content.
- Assumption 3: The surprise-content of a piece of data x_i is directly proportional to its uncertainty.

Message Compression Interpretation: The average Shannon info content of X is a measure of how much X can be compressed.

- *Interpretation of X* :

- $\{x_1, \dots, x_\ell\}$ = letters from which we construct N -letter messages.
- $\{p_1, \dots, p_\ell\}$ = probabilities for each letter to occur in a message.
- So: $p_i N$ = number of occurrences of x_i in a message.

- Claim: Given a probability distribution $P = \{p_1, \dots, p_\ell\}$,

$$\left(\begin{array}{l} \text{The average number of} \\ \text{distinct } N\text{-letter messages} \end{array} \right) = \frac{N!}{(p_1 N)! \cdots (p_\ell N)!}$$

← Number of ways to arrange N letters into ℓ bins with capacities $p_1 N, \dots, p_\ell N$.

Check: Let $P_a = (1, 0, \dots, 0)$ and $P_b = \left(1 - \frac{1}{N}, \frac{1}{N}, 0, \dots, 0\right)$

$$\left(\begin{array}{l} \text{ave \# distinct } N\text{-letter} \\ \text{messages for } P_a \end{array} \right) = \frac{N!}{N!} = 1$$

← Only one distinct N -letter message with N occurrences of x_1 .

$$\left(\begin{array}{l} \text{ave \# distinct } N\text{-letter} \\ \text{messages for } P_b \end{array} \right) = \frac{N!}{(N-1)!} = \frac{N(N-1)(N-2) \cdots 1}{(N-1)(N-2) \cdots 1} = N$$

← N distinct messages with one occurrence of x_2 : one when it's first, one when it's second, etc.

Message Compression Interpretation: The average Shannon info content of X is a measure of how much X can be compressed.

- *Interpretation of X* :

- $\{x_1, \dots, x_\ell\}$ = letters from which we construct N -letter messages.
- $\{p_1, \dots, p_\ell\}$ = probabilities for each letter to occur in a message.
- So: $p_i N$ = number of occurrences of x_i in a message.

- Claim: Given a probability distribution $P = \{p_1, \dots, p_\ell\}$,

$$\left(\begin{array}{l} \text{The average number of} \\ \text{distinct } N\text{-letter messages} \end{array} \right) = \frac{N!}{(p_1 N)! \cdots (p_\ell N)!}$$

← Number of ways to arrange N letters into ℓ bins with capacities $p_1 N, \dots, p_\ell N$.

- So:

$$\log_2 \left(\begin{array}{l} \text{The average number of} \\ \text{distinct } N\text{-letter messages} \end{array} \right) = \log_2 \left(\frac{N!}{(p_1 N)! \cdots (p_\ell N)!} \right)$$

Let's simplify the RHS...

$$\begin{aligned}
\log_2 \left(\frac{N!}{(p_1 N)! \cdots (p_\ell N)!} \right) &= \log_2(N!) - \{\log_2(p_1 N)! + \cdots + \log_2(p_\ell N)!\} \\
&\approx (N \log_2 N - N) - \{(p_1 N \log_2 p_1 N - p_1 N) + \cdots + (p_\ell N \log_2 p_\ell N - p_\ell N)\} \\
&= N\{\log_2 N - 1 - p_1 \log_2 p_1 - p_1 \log_2 N + p_1 - \cdots - p_\ell \log_2 p_\ell - p_\ell \log_2 N + p_\ell\} \\
&= -N \sum_{i=1}^{\ell} p_i \log_2 p_i \\
&= NS_{\text{Shan}}(X)
\end{aligned}$$


Stirling's approx:
 $\log_2 n! \approx n \log_2 n - n$

- Thus: $\log_2 \left(\begin{array}{l} \text{The average number of} \\ \text{distinct } N\text{-letter messages} \end{array} \right) = NS_{\text{Shan}}(X)$

- So: $\left(\begin{array}{l} \text{The average number of} \\ \text{distinct } N\text{-letter messages} \end{array} \right) = 2^{NS_{\text{Shan}}(X)}$ $\log_2 x = y \Rightarrow x = 2^y$

- So: On average, there are $2^{NS_{\text{Shan}}(X)}$ messages with N letters.
- This means, *at the message level*, we can encode them using only $NS_{\text{Shan}}(X)$ bits.

Check: 2 possible messages require 1 bit: 0, 1.
 4 possible messages require 2 bits: 00, 01, 10, 11.
etc.

- Note: *At the letter level*, how many bits are needed to encode a message of N letters drawn from an ℓ -letter alphabet?

First: How many bits are needed to encode each letter in an ℓ -letter alphabet?

$\ell = \# \text{letters}$

$x = \# \text{bits per letter}$

2 letters

1 bit: 0, 1

4 letters

2 bits: 00, 01, 10, 11

8 letters

3 bits: 000, 001, 010, 011, 100, 101, 110, 111

So: $\ell = 2^x$, or $x = \log_2 \ell$

- Note: $\log_2 \ell$ bits per letter entails $N \log_2 \ell$ bits for a sequence of N letters.
- Thus: *If we know how probable each letter is*, then instead of requiring $N \log_2 \ell$ bits to encode our messages, we can get by with only $NS_{\text{Shan}}(X)$ bits.
- So:

$S_{\text{Shan}}(X)$ represents the maximum amount that messages drawn from a set of letters with a probability distribution defined on it can be compressed.

Ex: Let $X = \{A, B, C, D\}$ ($\ell = 4$)

For instance:

$A = 00, B = 01, C = 10, D = 11$

- Then: We need $\log_2 4 = 2$ bits per letter.
- So: We need $2N$ bits to encode a message with N letters.
- Now: Suppose the probabilities for each letter to occur in a typical N -letter message are the following:

$$p_A = 1/2, \quad p_B = 1/4, \quad p_C = p_D = 1/8$$

- Then: The minimum number of bits needed to encode all possible N -letter messages is:

$$NS_{\text{Shan}}(X) = -N\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{8}\log_2\frac{1}{8}\right) = 1.75N$$

- Thus: If we know how probable each letter is, instead of requiring $2N$ bits to encode all possible messages, we can get by with only $1.75N$.
- Note: If all letters are equally likely (the equilibrium distribution), then $p_A = p_B = p_C = p_D = 1/4$.
- And: $NS_{\text{Shan}}(X) = -N\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) = 2N$

Interpretive Issues:

(1) How should the probabilities p_i in $S_{\text{Shan}}(X) = -\sum_i p_i \log_2 p_i$ be interpreted?

(a) *Epistemic probabilities* = measures of degrees of belief

- *Objective (rational) degrees of belief?*
- *Subjective degrees of belief?*

(b) *Ontic probabilities* = properties of physical systems

- *Long run frequencies?*
- *Single-case propensities?*

Interpretive Issues:

(1) How should the probabilities p_i in $S_{\text{Shan}}(X) = -\sum_i p_i \log_2 p_i$ be interpreted?

- Emphasis is on uncertainty: The Shannon information content of the value x_i of a random variable X is a function of how uncertain it is, with respect to the receiver.
 - So: Perhaps the probabilities are *epistemic*.
 - In particular: p_i is a measure of the receiver's degree of belief in the accuracy of the value x_i .
- But: The probabilities are set by the nature of the source.
 - If the source is not probabilistic, then p_i can be interpreted epistemically.
 - If the source is inherently probabilistic, then p_i can be interpreted as the *ontic* probability that the source produces the value x_i .

(2) How is S_{Shan} related to other notions of entropy?

Shannon:	$S_{\text{Shan}}(X) = - \sum_{i=1}^{\ell} p_i \log_2 p_i$ $S_{\text{Shan}}(X) = - \int_X \rho(x) \log_2 \rho(x) dx$
Thermodynamic:	$S_{\text{TD}}(\sigma_2) = \int_{\sigma_1}^{\sigma_2} \frac{\delta Q_R}{T} + S_0$
Boltzmann:	$S_{\text{Boltz}}(\Gamma_M) = k \ln \Gamma_M $ $= -k \sum_{i=1}^{\ell} n_i \ln n_i + \text{const.}$ $= -Nk \sum_{i=1}^{\ell} p_i \ln p_i + \text{const.}$ $S_{\text{Boltz}}(\Gamma_M) = -Nk \int_{\Gamma_\mu} \rho_\mu(x_\mu) \ln \rho_\mu(x_\mu) dx_\mu$
Gibbs:	$S_{\text{Gibbs}}(\rho) = -k \int_{\Gamma} \rho(x) \ln \rho(x) dx$

Can the 2nd Law of Thermodynamics be given an information-theoretic foundation?
Can statistical mechanics be given an information-theoretic foundation?