05. Entropy in Classical Information Theory

1. Motivation

- *Recall: Gibbs' approach to statistical mechanics*
  - $\{\Gamma, \rho\}$ = an ensemble of classical states
  - Γ = a phase space of multi-particle microstates x
  - ρ = a Gibbs probability distribution defined on Γ
  - $S_{\text{Gibbs}}(\rho)$ = ensemble average of $-\ln\rho$

- *Shannon's approach to classical information*
  - Generalize the notion of a classical phase space Γ of microstates x to a *random variable* X with possible values x.
  - View ρ as a probability distribution that assigns probabilities to the possible values x of X.
  - *View $-\ln\rho$ as a measure of "information"*

Intuition: The greater the probability $\rho(x)$, the more certain that the value of X is x, and the less information associated with this result.

Claude Shannon
(1916-2001)

- *Ex*: Let $X = \{x_1, ..., x_\ell\}$ = set of ℓ messages.

The amount of information gained from the reception of a message depends on how *likely* it is.

Intuition: The less likely a message is, the more info gained upon its reception!

Def. 1 (*Shannon entropy*). Let X be a random variable with possible values $\{x_1, ..., x_\ell\}$ and probability distribution $\{p_1, ..., p_\ell\}$. The **Shannon entropy** $S_{\text{Shan}}(X)$ of X is given by

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

- Compare with S_{Boltz}:

$$S_{\text{Boltz}}(\Gamma_M) = -Nk\sum_{i=1}^{\ell} p_i \ln p_i + \text{const.}$$

p_i are probabilities defined on single-particle microstates

- Continuous version of S_{Shan}:

$$S_{\text{Shan}}(X) = -\int_X \rho(x)\log_2\rho(x)dx$$

X takes a continuum of values

- Compare with S_{Gibbs}:

$$S_{\text{Gibbs}}(\rho) = -k\int_\Gamma \rho(x)\ln\rho(x)dx$$

$\rho(x)$ are probabilities defined on multi-particle microstates

Why the $\log_2$ in S_{Shan}?

- *Short answer: Classical info is measured in units of "bits".*
- *Long answer....*

Claim (Shannon 1949). $S_{\text{Shan}}(X) = -\sum_i p_i \log_2 p_i$ is the unique function $H(X) : \{$*probability distributions on* $X\} \to \mathbb{R}$, that satisfies:

- *Continuity*. $H(p_1, ..., p_\ell)$ is continuous.
- *Additivity*. $H(p_1 q_1, ..., p_\ell q_\ell) = H(P) + H(Q)$, for probability distributions P, Q.
- *Monoticity*. Info increases with ℓ for uniform distributions: If $m > \ell$, then $H(Q) > H(P)$, for any $P = \{1/\ell, ..., 1/\ell\}$ and $Q = \{1/m, ..., 1/m\}$.
- *Branching*. $H(p_1, ..., p_\ell)$ is independent of how the process is divided into parts.
- *Bit normalization*. The average info gained for two equally likely messages is one bit: $H(½, ½) = 1$.

<u>*Bit renormalization requires*</u> <u>$\log_2$</u>

- <u>*Suppose*</u>: $X = \{x_1, x_2\}$, and $P = \{½, ½\}$.
- <u>*Then*</u>:

$$\begin{aligned} H(X) &= -(½\log ½ + ½\log ½) \\ &= \log 2 \\ &= 1 \end{aligned}$$

bit renormalization

- <u>*And*</u>: $\log 2 = 1$ *if and only if* log is to base 2.

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

Why call this "entropy"?

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

von Neumann

!

2. S_{Shan} as a Measure of Uncertainty

- Let X be a random variable with possible values $\{x_1, ..., x_\ell\}$ and probability distribution $\{p_1, ..., p_\ell\}$.

Def. 3. The **expected value** $E(X)$ of X is given by $E(X) = \sum_{i=1}^{\ell} p_i x_i$

Def. 4. The **information gained** if X is measured to have the value x_i is given by $-\log_2 p_i$.

- Then the expected value of $-\log_2 p_i$ is $S_{\text{Shan}}(X)$:

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

- *$S_{\text{Shan}}(X)$ is the expected information gained upon measuring X.*
- *The greater $S_{\text{Shan}}(X)$, the greater the info gained upon measuring X, and the greater the uncertainty of its measured value.*

Uncertainty Interpretation Comparison

Shannon	**Boltzmann**	**Gibbs**
X = random variable	Γ_μ = single-particle phase space	Γ = multi-particle phase space
$\{x_1, ..., x_\ell\}$ = values of X	$\{x_1, ..., x_\ell\}$ = single-particle microstates.	$x \in \Gamma$: multi-particle microstates.
$\{p_1, ..., p_\ell\}$ = probabilty distribution over values.	$\{p_1, ..., p_\ell\}$ = probabilty distribution on Γ_μ.	ρ = probabilty distribution on Γ.
p_i = probability that X has value x_i upon measurement.	p_i = probability that microstate x_i of particle is in ith cell of Γ_μ.	$\rho(x,t)$ = prob that microstate of system at time t is x.
$-\log_2 p_i$ = info gained upon measurement of X with outcome x_i.	$-\ln p_i$ = info gained upon finding a particle to be in microstate x_i in ith cell of Γ_μ.	$-\ln\rho(x,t)$ = info gained upon finding multi-particle system to be in microstate x at time t.
$S_{\text{Shan}}(X) = -\sum_i p_i \log_2 p_i$	$S_{\text{Boltz}}(\Gamma_{M_D}) = -Nk\sum_i p_i \ln p_i$	$S_{\text{Gibbs}}(\rho) = -\int_\Gamma \rho \ln\rho\, dx$
$S_{\text{Shan}}(X)$ = expected info gain upon measurement of X.	S_{Boltz}/N = expected info gain upon finding a single particle of an N-particle system in microstate x_i in ith cell of Γ_μ.	$S_{\text{Gibbs}}(\rho)$ = expected info gain upon finding multi-particle system to be in microstate x at time t.

3. S_{Shan} as Maximum Amount of Message Compression

- Let $X = \{x_1, ..., x_\ell\}$ be a set of letters from which we construct messages.
- Suppose the messages have N letters a piece.
- Let $\{p_1, ..., p_\ell\}$ be a probability distribution over X.

> *What this means*:
> - Each letter x_i has a probability of p_i of occuring in a message.
> - *In other words*: A typical message will contain $p_1 N$ occurrences of x_1, $p_2 N$ occurrences of x_2, *etc*.

- *Thus*:

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

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

- *So*:

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

Let's simplify the RHS...

$$\log_2\left(\frac{N!}{(p_1N)!\cdots(p_\ell N)!}\right) = \log_2(N!) - \{\log_2(p_1N)! + \cdots + \log_2(p_\ell N)!\}$$

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

$$\approx (N\log_2 N - N) - \{(p_1N\log_2 p_1N - p_1N) + \cdots + (p_\ell N\log_2 p_\ell N - p_\ell N)\}$$

$$= N\{\log_2 N - 1 - p_1\log_2 p_1 - p_1\log_2 N + p_1 - \cdots - p_\ell\log_2 p_\ell - p_\ell\log_2 N + p_\ell\}$$

$$= -N\textstyle\sum_i p_i\log_2 p_i$$

$$= NS_{\text{Shan}}(X)$$

- *Thus*:

$$\log_2\left(\begin{array}{l}\textit{The number of distinct}\\ \textit{typical messages}\end{array}\right) = NS_{\text{Shan}}(X)$$

- *So*:

$$\left(\begin{array}{l}\textit{The number of distinct}\\ \textit{typical messages}\end{array}\right) = 2^{NS_{\text{Shan}}(X)}$$

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

- *So*: There are only $2^{NS_{\text{Shan}}(X)}$ typical 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.

- *Now*: *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?

ℓ = #letters	x = #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 typical 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:

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

- *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*: $NS_{\text{Shan}}(X) = -N\left(\tfrac{1}{4}\log_2\tfrac{1}{4} + \tfrac{1}{4}\log_2\tfrac{1}{4} + \tfrac{1}{4}\log_2\tfrac{1}{4} + \tfrac{1}{4}\log_2\tfrac{1}{4}\right) = 2N$

Message Compression Interpretation Comparison

Shannon	**Boltzmann**
X = set of letters	Γ_μ = single-particle phase space
$\{x_1, ..., x_\ell\}$ = letters	$\{x_1, ..., x_\ell\}$ = single-particle microstates
N-letter message	N-particle microstate
N = # of letters in message	N = # single-particle microstates in a multi-particle microstate
$\{p_1, ..., p_\ell\}$ = probability distribution over letters	$\{p_1, ..., p_\ell\}$ = probabilty distribution over single-particle microstates
p_i = probability that letter x_i occurs in a message	p_i = prob that single-particle microstate x_i occurs in an N-particle microstate
Np_i = # of occurrences of letter x_i in typical message	Np_i = # occurrences of single-particle microstate x_i in typical N-particle microstate
$S_{\text{Shan}}(X) = -\sum_i p_i \log_2 p_i$	$S_{\text{Boltz}}(\Gamma_{M_D}) = -Nk\sum_i p_i \ln p_i$
NS_{Shan} = minimum number of base 2 numerals ("bits") needed to encode a message composed of N letters drawn from set $\{x_1, ..., x_\ell\}$.	$S_{\text{Boltz}} \sim NS_{\text{Shan}}$ = minimum number of base e numerals ("e-bits?") needed to encode a multi-particle microstate composed of N single-particle microstates drawn from $\{x_1, ..., x_\ell\}$.

Gibbs?

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 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?

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

Can statistical mechanics be given an information-theoretic foundation?

Can the 2nd Law be given an information-theoretic foundation?