Shannon entropy

Daniel Bosk

Department of Information and Communication Systems,
Mid Sweden University, Sundsvall.

6th April 2020

1 Shannon entropy
- History
- Definition of Shannon Entropy
- Properties for Shannon entropy
- Conditional entropy
- Information gain

- Created 1948 by Shannon's paper 'A Mathematical Theory of Communication' [Sha48].
- He starts using the term 'entropy' as a measure for information.
  - In physics entropy measures the disorder of molecules.
  - Shannon's entropy measures disorder of information.
- He used this theory to analyse communication.
  - What are the theoretical limits for different channels?
  - How much redundancy is needed for certain noise?

- Created 1948 by Shannon's paper 'A Mathematical Theory of Communication' [Sha48].
- He starts using the term 'entropy' as a measure for information.
  - In physics entropy measures the disorder of molecules.
  - Shannon's entropy measures disorder of information.
- He used this theory to analyse communication.
  - What are the theoretical limits for different channels?
  - How much redundancy is needed for certain noise?

- Created 1948 by Shannon's paper 'A Mathematical Theory of Communication' [Sha48].
- He starts using the term 'entropy' as a measure for information.
    - In physics entropy measures the disorder of molecules.
    - Shannon's entropy measures disorder of information.
- He used this theory to analyse communication.
    - What are the theoretical limits for different channels?
    - How much redundancy is needed for certain noise?

- This theory is interesting on the physical layer of networking.
- It's also interesting for security.
    - Field of Information Theoretic Security
    - 'Efficiency' of passwords
    - Measure identifiability
    - ...

- This theory is interesting on the physical layer of networking.
- It's also interesting for security.
  - Field of Information Theoretic Security
  - 'Efficiency' of passwords
  - Measure identifiability
  - ...

Definition (Shannon entropy)

- Stochastic variable $\mathbf{X}$ assumes values from X.
- Shannon entropy $H(\mathbf{X})$ defined as

$$H(\mathbf{X}) = -K \sum_{x \in X} \Pr(\mathbf{X} = x) \log \Pr(\mathbf{X} = x),$$

- Usually $K = \frac{1}{\log 2}$ to give entropy in unit bits (bit).

Shannon entropy can be seen as . . .

- . . . how much choice in each event.
- . . . the uncertainty of each event.
- . . . how many bits to store each event.
- . . . how much information it produces.

Example (Toss a coin)

- Stochastic variable **S** takes values from $S = \{h, t\}$.
- We have $\Pr(\mathbf{S} = h) = \Pr(\mathbf{S} = t) = \frac{1}{2}$.
- This gives $H(\mathbf{S})$ as follows:

$$H(\mathbf{S}) = -\left(\Pr(\mathbf{S} = h)\log\Pr(\mathbf{S} = h) + \Pr(\mathbf{S} = t)\log\Pr(\mathbf{S} = t)\right)$$
$$= -2 \times \frac{1}{2}\log\frac{1}{2} = \log 2 = 1.$$

Example (Roll a die)

- Stochastic variable **D** takes values from $D = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$.
- We have $\Pr(\mathbf{D} = d) = \frac{1}{6}$ for all $d \in D$.
- The entropy $H(\mathbf{D})$ is as follows:

$$\begin{aligned} H(\mathbf{D}) &= -\sum_{d \in D} \Pr(\mathbf{D} = d) \log \Pr(\mathbf{D} = d) \\ &= -6 \times \frac{1}{6} \log \frac{1}{6} = \log 6 \approx 2.585. \end{aligned}$$

Remark

- If we didn't know already, we now know that a roll of a die . . .
  - contains more 'choice' than a coin toss.
  - is more uncertain to predict than a coin toss.
  - requires more bits to store than a coin toss.
  - produces more information than a coin toss.
- What if we modify the die a bit?

Example (Roll of a modified die)

- Stochastic variable D' taking values from D.
- We now have $\Pr(\mathbf{D}' =$ ⚅$) = \frac{9}{10}$ and $\Pr(\mathbf{D}' = d) = \frac{1}{10} \times \frac{1}{5}$ for $d \neq$ ⚅.
- This yields

$$H(\mathbf{D}') = -\left(\frac{9}{10}\log\frac{9}{10} + \sum_{d \neq 6}\frac{1}{50}\log\frac{1}{50}\right)$$
$$= -\frac{9}{10}\log\frac{9}{10} - 5 \times \frac{1}{50}\log\frac{1}{50}$$
$$= -\frac{9}{10}\log\frac{9}{10} - \frac{1}{10}\log\frac{1}{50} \approx 0.701.$$

- Note that the log function is the logarithm in base 2 (i.e. $\log_2$).

Remark

- This die is much easier to predict.
- It produces much less information — less than a coin toss!
- Requires less data for storage etc.

Definition

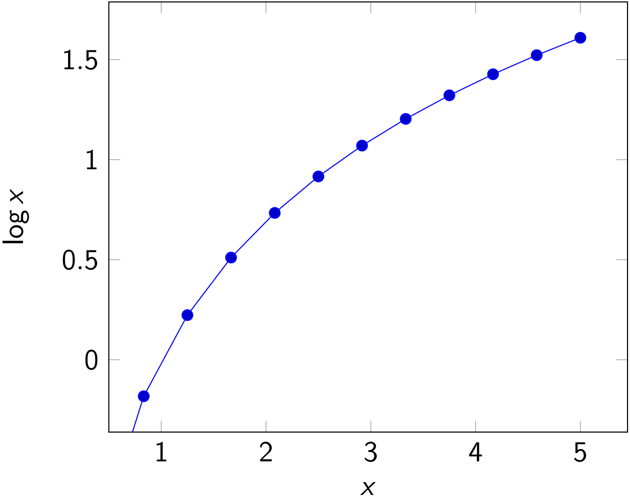
- Function $f\colon \mathbb{R} \to \mathbb{R}$ such that

$$tf(x) + (1-t)f(y) \leq f(tx + (1-t)y),$$

- Then f is *concave*.
- With strict inequality for $x \neq y$ we say that f is *strictly concave*.

Example

$\log\colon \mathbb{R} \to \mathbb{R}$ is strictly concave.


1.5
1
0.5
0
$\log x$
1
2
3
4
5
x

Theorem (Jensen's inequality)

- *Strictly concave function* $f\colon \mathbb{R} \to \mathbb{R}$.
- *Real numbers* $a_1, a_2, \ldots, a_n > 0$ *such that* $\sum_{i=1}^{n} a_i = 1$.
- *Then we have*

$$\sum_{i=1}^{n} a_i f(x_i) \leq f\left(\sum_{i=1}^{n} a_i x_i\right).$$

- *We have equality iff* $x_1 = x_2 = \cdots = x_n$.

Theorem

- *Stochastic variable* $\mathbf{X}$ *with probability distribution*

$$p_1, p_2, \ldots, p_n, \text{ where } p_i > 0 \text{ for } 1 \leq i \leq n.$$

- *Then* $H(\mathbf{X}) \leq \log n$.
- *Equality iff* $p_1 = p_2 = \cdots = p_n = 1/n$.

Proof.

The theorem follows directly from Jensen's inequality:

$$H(\mathbf{X}) = -\sum_{i=1}^{n} p_i \log p_i = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$$
$$\leq \log \sum_{i=1}^{n} p_i \frac{1}{p_i} = \log n.$$

With equality iff $p_1 = p_2 = \cdots = p_n$. Q.E.D.

Corollary

$H(\mathbf{X}) = 0$ iff $\Pr(\mathbf{X} = x) = 1$ for some $x \in X$ and $\Pr(\mathbf{X} = x') = 0$ for all $x \neq x' \in X$.

Proof.

- If $\Pr(\mathbf{X} = x) = 1$, then $n = 1$ and thus $H(\mathbf{X}) = \log n = 0$.
- If $H(\mathbf{X}) = 0$, then $H(\mathbf{X}) \leq \log n = 0$. Thus $n = 1$.

Q.E.D.

Lemma

- *Stochastic variables* $\mathbf{X}$ *and* $\mathbf{Y}$.
- *Then we have*

$$H(\mathbf{X},\mathbf{Y}) \leq H(\mathbf{X}) + H(\mathbf{Y}).$$

- *Equality iff* $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent.*

Definition (Conditional entropy)

- Define *conditional entropy* $H(\mathbf{Y} \mid \mathbf{X})$ as

$$H(\mathbf{Y} \mid \mathbf{X}) = -\sum_{y} \sum_{x} \Pr(\mathbf{Y} = y) \Pr(\mathbf{X} = x \mid y) \log \Pr(\mathbf{X} = x \mid y).$$

Remark

This is the uncertainty in $\mathbf{Y}$ which is not revealed by $\mathbf{X}$.

Definition (Conditional entropy)

- Define *conditional entropy* $H(\mathbf{Y} \mid \mathbf{X})$ as

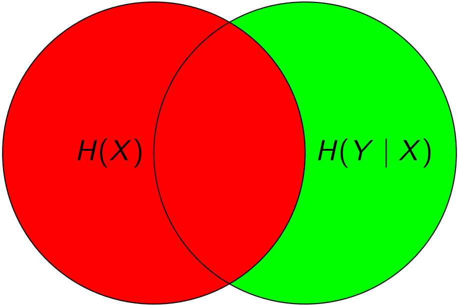
$$H(\mathbf{Y} \mid \mathbf{X}) = -\sum_{y} \sum_{x} \Pr(\mathbf{Y} = y) \Pr(\mathbf{X} = x \mid y) \log \Pr(\mathbf{X} = x \mid y).$$

Remark

This is the uncertainty in $\mathbf{Y}$ which is not revealed by $\mathbf{X}$.

Theorem

$$H(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y} \mid \mathbf{X})$$



Corollary

$H(\mathbf{X} \mid \mathbf{Y}) \leq H(\mathbf{X})$.

Corollary

$H(\mathbf{X} \mid \mathbf{Y}) = H(\mathbf{X})$ *iff* $\mathbf{X}$ *and* $\mathbf{Y}$ *independent.*

Definition

- Set U of possible outcomes.
- Probability of outcome $u \in U$ denoted p_u.
- We learn that some *unknown* outcome is in $A \subset U$.
- Then the *information gain* $G(A \mid U)$ is defined as

$$G(A \mid U) = \log \frac{1}{\Pr(A)} = -\log \Pr(A),$$

where $\Pr(A) = \sum_{i \in A} p_i$.

Example (Roll of dice again)

- Someone rolls and we should guess the result, $\frac{1}{6}$ chance.
- We learn that it was an even number, we gain

$$-\log\left(\frac{1}{6}+\frac{1}{6}+\frac{1}{6}\right) = -\log\frac{3}{6} = \log\frac{6}{3} = \log 2 = 1.$$

- The remaining uncertainty is 1.58 bit.

Remark

- $X' = \{$⚁, ⚃, ⚅$\}$
- $H(\mathbf{X}') = -\sum_{x \in X'} \Pr(\mathbf{X}' = x) \log \Pr(\mathbf{X}' = x)$
- I.e. $-3 \times \frac{1}{3} \log \frac{1}{3} \approx 1.58$.

Example (Roll of dice again)

- Someone rolls and we should guess the result, $\frac{1}{6}$ chance.
- We learn that it was an even number, we gain

$$-\log\left(\frac{1}{6}+\frac{1}{6}+\frac{1}{6}\right) = -\log\frac{3}{6} = \log\frac{6}{3} = \log 2 = 1.$$

- The remaining uncertainty is 1.58 bit.

Remark

- $X' = \{$⚁, ⚃, ⚅$\}$
- $H(\mathbf{X}') = -\sum_{x \in X'} \Pr(\mathbf{X}' = x) \log \Pr(\mathbf{X}' = x)$
- I.e. $-3 \times \frac{1}{3} \log \frac{1}{3} \approx 1.58$.

Example (Dice yet again)

- We learn the die show less than five, i.e. not ⚄ nor ⚅.
- This yields

$$-\log\left(4 \times \frac{1}{6}\right) = \log\frac{6}{4} \approx 0.58$$

References

[Sha48] C. E. Shannon. 'A Mathematical Theory of Communication'. In: *The Bell System Technical Journal* 27 (July 1948), pp. 379–423, 623–656.