Understanding the Source Coding Theorem: A talk on Shannon's Entropy

Kwing Hei Li

University of Cambridge

khl61@cam.ac.uk

October 21, 2020

Question(1)

Figure: **ICMA photos: Creative Commons Attribution-Share Alike 2.0 Generic license**

Encoding a fair coin toss

Question(1)

Figure: **ICMA photos: Creative Commons Attribution-Share Alike 2.0 Generic license**

Encoding a fair coin toss

vs

Encoding a lot of fair coin tosses

Question(2)

Figure: **ICMA photos: Creative Commons Attribution-Share Alike 2.0 Generic license**

Encoding a **biased** coin toss (99% heads, 1% tails)

Question(2)

Figure: ICMA photos: Creative Commons Attribution-Share Alike 2.0 Generic license

Encoding a **biased** coin toss (99% heads, 1% tails)

vs

Encoding a lot of **biased** coin tosses

Question(3)

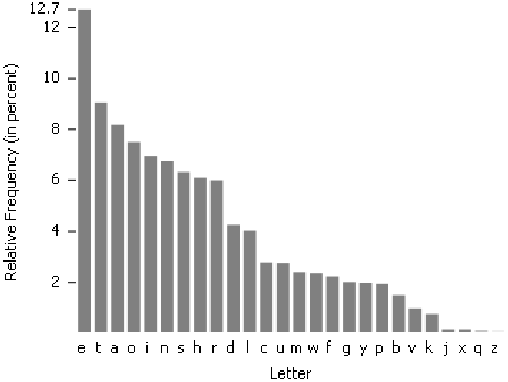

Figure: **Iain.dalton: Creative Commons Attribution-ShareAlike 3.0 License**

Encoding an English letter (Note $26 < 32 = 2^5$)

Question(3)

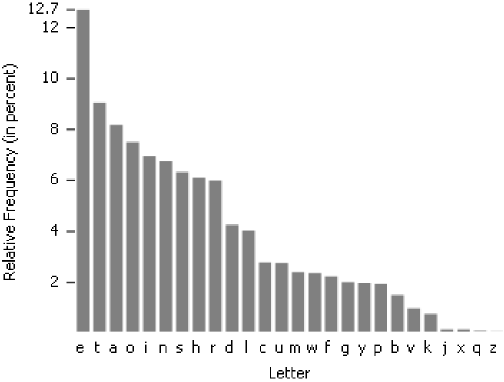

Figure: **Iain.dalton: Creative Commons Attribution-ShareAlike 3.0 License**

Encoding an English letter (Note $26 < 32 = 2^5$)

vs

Encoding a novel

Surprisal

Definition of Surprisal

Given event x with probability $P(x)$, surprisal of x, $I(x) = -\log P(x)$

Surprisal

Definition of Surprisal

Given event x with probability $P(x)$, surprisal of x, $I(x) = -\log P(x)$

- Measures information content of an event

Surprisal

Definition of Surprisal

Given event x with probability $P(x)$, surprisal of x, $I(x) = -\log P(x)$

- Measures information content of an event
- Consider
  1. "Hei Li ate a chocolate cake during his Compsci Talk"
  2. "Hei Li did not eat a chocolate cake during his Compsci Talk"

Shannon's Entropy

Definition of Entropy

Given discrete r.v. X, with possible outcomes x_1,...,x_n which occur with probability $P(x_1)$,...,$P(x_n)$, entropy of X, $H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$

Shannon's Entropy

Definition of Entropy

Given discrete r.v. X, with possible outcomes x_1,...,x_n which occur with probability $P(x_1)$,...,$P(x_n)$, entropy of X, $H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$

- Entropy is expected value of surprisal over all outcomes.

Shannon's Entropy

Definition of Entropy

Given discrete r.v. X, with possible outcomes $x_1, \ldots, x_n$ which occur with probability $P(x_1), \ldots, P(x_n)$, entropy of X, $H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$

- Entropy is expected value of surprisal over all outcomes.
- Informally, entropy $=$
  - amount of uncertainty of r.v. has before it is resolved
  - amount of information r.v. provides after it is resolved

Example of Shannon's Entropy

- Suppose $P(X = i) = \frac{1}{6}$ for $i = 1,...,6$ (fair die)

Example of Shannon's Entropy

- Suppose $P(X = i) = \frac{1}{6}$ for $i = 1,...,6$ (fair die)
- $H(X) = -\sum_{i=1}^{6} \frac{1}{6} \log \frac{1}{6}$
  ≈ 2.58496 bits

Example of Shannon's Entropy

- Suppose $P(X = i) = \frac{1}{6}$ for $i = 1,...,6$ (fair die)
- $H(X) = -\sum_{i=1}^{6} \frac{1}{6} \log \frac{1}{6}$
  ≈ 2.58496 bits
- Suppose $P(Y = i) = \begin{cases} \frac{1}{25} & \text{for } i = 1, ..., 5 \\ \frac{4}{5} & \text{for } i = 6 \end{cases}$ (biased die)

Example of Shannon's Entropy

- Suppose $P(X = i) = \frac{1}{6}$ for $i = 1,...,6$ (fair die)
- $H(X) = -\sum_{i=1}^{6} \frac{1}{6} \log \frac{1}{6}$
  ≈ 2.58496 bits
- Suppose $P(Y = i) = \begin{cases} \frac{1}{25} & \text{for } i = 1, ..., 5 \\ \frac{4}{5} & \text{for } i = 6 \end{cases}$ (biased die)
- $H(Y) = -(\sum_{i=1}^{5} \frac{1}{25} \log \frac{1}{25} + \frac{4}{5} \log \frac{4}{5})$
  ≈ 1.1863137 bits

Example of Shannon's Entropy

- Suppose $P(X = i) = \frac{1}{6}$ for $i = 1,...,6$ (fair die)
- $H(X) = -\sum_{i=1}^{6} \frac{1}{6} \log \frac{1}{6}$
  ≈ 2.58496 bits
- Suppose $P(Y = i) = \begin{cases} \frac{1}{25} & \text{for } i = 1, ..., 5 \\ \frac{4}{5} & \text{for } i = 6 \end{cases}$ (biased die)
- $H(Y) = -(\sum_{i=1}^{5} \frac{1}{25} \log \frac{1}{25} + \frac{4}{5} \log \frac{4}{5})$
  ≈ 1.1863137 bits
- High uncertainty = high information content = high entropy
  "Information is the resolution of uncertainty" - Shannon

Informal Source Coding Theorem

Informal Source Coding Theorem

N i.i.d. r.v.s each with entropy $H(X)$ can be compressed into more than N $H(X)$ bits with negligible risk of information loss, as N tends to infinity

Informal Source Coding Theorem

Informal Source Coding Theorem

N i.i.d. r.v.s each with entropy $H(X)$ can be compressed into more than $N\ H(X)$ bits with negligible risk of information loss, as N tends to infinity

Converse of Informal Source Coding Theorem

If N i.i.d. r.v.s each with entropy $H(X)$ are compressed into fewer than $N\ H(X)$ bits, it is virtually certain that information will be lost

Source Coding Theorem (A more formal definition)

Source Coding Theorem

Given discrete r.v. X and $\epsilon > 0$, $\exists$ positive integer N such that $\forall$ integers $n > N$, $\exists$ an encoder which takes n i.i.d. repetition of the source, X_1,...,X_n and maps it to $n(H(X) + \epsilon)$ binary bits such that the source outcomes are recoverable from the bits with probability of at least $1 - \epsilon$

The Source Coding Theorem Challenge

- Alice gives Bob real number $\epsilon > 0$ and a discrete r.v. X

The Source Coding Theorem Challenge

- Alice gives Bob real number $\epsilon > 0$ and a discrete r.v. X
- Bob gives Alice positive integer N

The Source Coding Theorem Challenge

- Alice gives Bob real number $\epsilon > 0$ and a discrete r.v. X
- Bob gives Alice positive integer N
- Alice gives Bob positive integer $n > N$

The Source Coding Theorem Challenge

- Alice gives Bob real number $\epsilon > 0$ and a discrete r.v. X
- Bob gives Alice positive integer N
- Alice gives Bob positive integer $n > N$
- Bob constructs an encoder that
  1. takes in n i.i.d. repetition of X
  2. outputs $n(H(X) + \epsilon)$ binary bits
  3. inputs are recoverable with at least probability $1 - \epsilon$

The Source Coding Theorem Challenge

- Alice gives Bob real number $\epsilon > 0$ and a discrete r.v. X
- Bob gives Alice positive integer N
- Alice gives Bob positive integer $n > N$
- Bob constructs an encoder that
  1. takes in n i.i.d. repetition of X
  2. outputs $n(H(X) + \epsilon)$ binary bits
  3. inputs are recoverable with at least probability $1 - \epsilon$
- The Source Coding Theorem states that Bob always wins this challenge!

Example of The Source Coding Theorem Challenge(1)

- Alice gives Bob real number $\epsilon = \frac{1}{30} > 0$ and a discrete r.v. X where $\begin{cases} P(X = A) = \frac{1}{2} \\ P(X = B) = \frac{1}{3} \\ P(X = C) = \frac{1}{6} \end{cases}$
  Note $H(X) \approx 1.459$ bits

Example of The Source Coding Theorem Challenge(1)

- Alice gives Bob real number $\epsilon = \frac{1}{30} > 0$ and a discrete r.v. X where $\begin{cases} P(X = A) = \frac{1}{2} \\ P(X = B) = \frac{1}{3} \\ P(X = C) = \frac{1}{6} \end{cases}$
  Note $H(X) \approx 1.459$ bits
- Bob gives Alice positive integer $N = 1$

Example of The Source Coding Theorem Challenge(1)

- Alice gives Bob real number $\epsilon = \frac{1}{30} > 0$ and a discrete r.v. X where $\begin{cases} P(X = A) = \frac{1}{2} \\ P(X = B) = \frac{1}{3} \\ P(X = C) = \frac{1}{6} \end{cases}$
  Note $H(X) \approx 1.459$ bits
- Bob gives Alice positive integer $N = 1$
- Alice gives Bob positive integer $n = 2 > N$

Example of The Source Coding Theorem Challenge(2)

- Bob constructs an encoder as follows:

$$\begin{cases} AA \rightarrow 000 \\ AB \rightarrow 001 \\ AC \rightarrow 010 \\ \dots \\ CB \rightarrow 111 \\ CC \rightarrow \text{encoder explodes} \end{cases}$$

Example of The Source Coding Theorem Challenge(2)

- Bob constructs an encoder as follows:

$$\begin{cases} AA \to 000 \\ AB \to 001 \\ AC \to 010 \\ \dots \\ CB \to 111 \\ CC \to \text{encoder explodes} \end{cases}$$

- This encoder
  1. takes in 2 i.i.d. repetition of X
  2. outputs $n(H(X) + \epsilon) \approx 2.98$ binary bits
  3. inputs are recoverable with at least probability $\frac{29}{30}$ (This encoder actually works with probability $\frac{35}{36}$)

Limits to data compression

- Always possible to compress data with code rate more than entropy of source with negligible risk of information loss

Limits to data compression

- Always possible to compress data with code rate more than entropy of source with negligible risk of information loss
- Impossible to compress data with code rate lower than entropy of source, without losing information

Limits to data compression

- Always possible to compress data with code rate more than entropy of source with negligible risk of information loss
- Impossible to compress data with code rate lower than entropy of source, without losing information
- Lossless data compression methods, e.g. Huffman codes

Operational meaning of Shannon's entropy

- Informally, operational definition describes how you can measure an abstract concept through empirical observations or a set of procedures

Operational meaning of Shannon's entropy

- Informally, operational definition describes how you can measure an abstract concept through empirical observations or a set of procedures
- Theorem provides operational definition to Shannon's entropy

Operational meaning of Shannon's entropy

- Informally, operational definition describes how you can measure an abstract concept through empirical observations or a set of procedures
- Theorem provides operational definition to Shannon's entropy
- Shannon's entropy = limit of how well you can compress the source

Weak Law of Large Numbers (WLLN)

Weak Law of Large Numbers

Let X_1, X_2,... be an infinite sequence of i.i.d r.v. with expected value $\mathbb{E}(X_i) = \mu$ for all $i = 1, 2, ...$
Let sample average $\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i$
Then for all $\epsilon > 0$, $\lim_{n \to \infty} P(| \overline{X_n} - \mu | < \epsilon) = 1$

Weak Law of Large Numbers (WLLN)

Weak Law of Large Numbers

Let X_1, X_2,... be an infinite sequence of i.i.d r.v. with expected value $\mathbb{E}(X_i) = \mu$ for all $i = 1, 2, ...$
Let sample average $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Then for all $\epsilon > 0$, $\lim_{n\to\infty} P(|\overline{X_n} - \mu| < \epsilon) = 1$

- Informally, the more you sample a r.v., the closer the mean gets to the expected value

Weak Law of Large Numbers (WLLN)

Weak Law of Large Numbers

Let X_1, X_2,... be an infinite sequence of i.i.d r.v. with expected value $\mathbb{E}(X_i) = \mu$ for all $i = 1, 2, \ldots$
Let sample average $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Then for all $\epsilon > 0$, $\lim_{n\to\infty} P(\mid \overline{X_n} - \mu \mid < \epsilon) = 1$

- Informally, the more you sample a r.v., the closer the mean gets to the expected value
- Example: Consider a game where you lose \$ 6 if you roll 6 but you win \$ 1 otherwise
  Expected value of game $=$ \$-$\frac{1}{6}$
  If you play many times $\approx$ losing \$$\frac{1}{6}$ every time you play it!

Asymptotic Equipartition Property (AEP)

- Consider discrete r.v. X with possible outcomes $x_1,...,x_m$ which occur with probability $P(x_1),...,P(x_m)$

Asymptotic Equipartition Property (AEP)

- Consider discrete r.v. X with possible outcomes $x_1,...,x_m$ which occur with probability $P(x_1),...,P(x_m)$
- Let $Y \sim -\log P(X)$.

Asymptotic Equipartition Property (AEP)

- Consider discrete r.v. X with possible outcomes x_1,...,x_m which occur with probability $P(x_1)$,...,$P(x_m)$
- Let $Y \sim -\log P(X)$.
- Y is surprisal of outcome of X !
  So $\mathbb{E}(Y) = -\sum_{i=1}^{m} P(x_i) \log P(x_i)$
  $= H(X)$

Asymptotic Equipartition Property (AEP)

- Consider discrete r.v. X with possible outcomes $x_1,...,x_m$ which occur with probability $P(x_1),...,P(x_m)$
- Let $Y \sim -\log P(X)$.
- Y is surprisal of outcome of X !
  So $\mathbb{E}(Y) = -\sum_{i=1}^{m} P(x_i) \log P(x_i)$
  $= H(X)$
- Therefore given n i.i.d r.v. $X_1, ..., X_n \sim X$,
  We have
  $= \frac{1}{n} \sum_{i=1}^{n} -\log P(X_i)$
  $= \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad (Y_i \sim Y)$
  $\to H(X)$ as $n \to \infty \qquad$ (WLLN)

Asymptotic Equipartition Property (AEP)

- Consider discrete r.v. X with possible outcomes x_1,...,x_m which occur with probability $P(x_1)$,...,$P(x_m)$
- Let $Y \sim -\log P(X)$.
- Y is surprisal of outcome of X !
  So $\mathbb{E}(Y) = -\sum_{i=1}^{m} P(x_i) \log P(x_i)$
  $= H(X)$
- Therefore given n i.i.d r.v. $X_1, ..., X_n \sim X$,
  We have
  $= \frac{1}{n} \sum_{i=1}^{n} -\log P(X_i)$
  $= \frac{1}{n} \sum_{i=1}^{n} Y_i$ $\qquad (Y_i \sim Y)$
  $\rightarrow H(X)$ as $n \rightarrow \infty$ $\qquad$ (WLLN)
- (The AEP is just a special case of WLLN where mean of Y_i's tends to entropy of X!)

Typical Set (TS)

Definition

Let X_1, X_2,... be an infinite sequence of i.i.d r.v.s with same distribution as X

The typical set

$A_n^{\epsilon} = \{(x_1, ..., x_n) : | \ (\frac{1}{n}\sum_{i=1}^{n} \text{-} \log P(x_i)) - H(X) \ | < \epsilon\}$

where x_i is the outcome of X_i for $i = 1, 2, ...$

Typical Set (TS)

Definition

Let X_1, X_2,... be an infinite sequence of i.i.d r.v.s with same distribution as X
The typical set
$A_n^{\epsilon} = \{(x_1, ..., x_n) : | \ (\frac{1}{n} \sum_{i=1}^{n} \text{-} \log P(x_i)) - H(X) \ | < \epsilon\}$
where x_i is the outcome of X_i for $i = 1, 2, ...$

Properties of TS:

Typical Set (TS)

Definition

Let X_1, X_2,... be an infinite sequence of i.i.d r.v.s with same distribution as X

The typical set

$A_n^\epsilon = \{(x_1, ..., x_n) : | \, (\frac{1}{n} \sum_{i=1}^{n} \text{-} \log P(x_i)) - H(X) \, | < \epsilon\}$

where x_i is the outcome of X_i for $i = 1, 2, ...$

Properties of TS:

1. By AEP, since $\frac{1}{n} \sum_{i=1}^{n} \text{-} \log P(X_i) \to H(X)$ as $n \to \infty$, $P((x_1, ..., x_n) \in A_n^\epsilon) \to 1$ as $n \to \infty$

Typical Set (TS)

Definition

Let X_1, X_2,... be an infinite sequence of i.i.d r.v.s with same distribution as X

The typical set

$A_n^\epsilon = \{(x_1, ..., x_n) : \mid (\frac{1}{n}\sum_{i=1}^{n} \text{-}\log P(x_i)) - H(X) \mid < \epsilon\}$

where x_i is the outcome of X_i for $i = 1, 2, ...$

Properties of TS:

1. By AEP, since $\frac{1}{n}\sum_{i=1}^{n} \text{-}\log P(X_i) \to H(X)$ as $n \to \infty$, $P((x_1, ..., x_n) \in A_n^\epsilon) \to 1$ as $n \to \infty$
2. By the definition of TS, if $(x_1, ..., x_n) \in A_n^\epsilon$ then
   $\mid (\frac{1}{n}\sum_{i=1}^{n} \text{-}\log P(x_i)) - H(X) \mid < \epsilon$
   $\Rightarrow -\frac{1}{n}\log P(x_1, ..., x_n) < H(X) - \epsilon$
   $\Rightarrow \mathbf{2^{-n(H(X)+\epsilon)}} < P(x_1, ..., x_n)$

Checkpoint

- Started with WLLN

Checkpoint

- Started with WLLN
- Proved AEP with WLLN

Checkpoint

- Started with WLLN
- Proved AEP with WLLN
- We defined TS and noticed 2 properties

Checkpoint

- Started with WLLN
- Proved AEP with WLLN
- We defined TS and noticed 2 properties
    1. By AEP, the probability a sequence of outcomes exists in TS tends to 1 as n increases

Checkpoint

- Started with WLLN
- Proved AEP with WLLN
- We defined TS and noticed 2 properties
  1. By AEP, the probability a sequence of outcomes exists in TS tends to 1 as n increases
  2. By the definition of TS, if a sequence exists in the typical set, the probability it occurs is larger than $2^{-n(H(X)+\epsilon)}$

Chocolate Cake Example(1)

- Consider a chocolate cake that is cut into multiple pieces

Chocolate Cake Example(1)

- Consider a chocolate cake that is cut into multiple pieces
- Volume of the cake is smaller than x units
  Volume of each piece of the cake is larger than y units

Chocolate Cake Example(1)

- Consider a chocolate cake that is cut into multiple pieces
- Volume of the cake is smaller than x units
  Volume of each piece of the cake is larger than y units
- What is maximum number of pieces of chocolate cake I can possibly have?

Chocolate Cake Example(1)

- Consider a chocolate cake that is cut into multiple pieces
- Volume of the cake is smaller than x units
  Volume of each piece of the cake is larger than y units
- What is maximum number of pieces of chocolate cake I can possibly have?
- It is simply $\frac{x}{y}$!

Chocolate Cake Example(2)

- Consider a TS with many sequences as members

Chocolate Cake Example(2)

- Consider a TS with many sequences as members
- Probability a sequence exists in the set is less than 1(Trivial)
  Probability a sequence in the set occurring is larger than $2^{-n(H(X)+\epsilon)}$ (Definition of TS)

Chocolate Cake Example(2)

- Consider a TS with many sequences as members
- Probability a sequence exists in the set is less than 1(Trivial)
  Probability a sequence in the set occurring is larger than $2^{-n(H(X)+\epsilon)}$ (Definition of TS)
- What is maximum possible number of sequences in the typical set?

Chocolate Cake Example(3)

- What is maximum possible number of sequences in the typical set?

Chocolate Cake Example(3)

- What is maximum possible number of sequences in the typical set?
- It is simply $\frac{1}{2^{-n(H(X)+\epsilon)}} = 2^{n(H(X)+\epsilon)}$

Chocolate Cake Example(3)

- What is maximum possible number of sequences in the typical set?
- It is simply $\frac{1}{2^{-n(H(X)+\epsilon)}} = 2^{n(H(X)+\epsilon)}$
- Important observation: We simply need $n(H(X)+\epsilon)$ binary bits to encode every sequence in the typical set!

Proof of Achievability (Bob's Strategy)

1. Given $\epsilon > 0$ and discrete r.v. X, give Alice positive integer N such that for all positive integers $n > N$, probability that a sequence of outcomes produced by n i.i.d. samples of X lies in the TS A_n^{ϵ} is larger than $1 - \epsilon$

Proof of Achievability (Bob's Strategy)

1. Given $\epsilon > 0$ and discrete r.v. X, give Alice positive integer N such that for all positive integers $n > N$, probability that a sequence of outcomes produced by n i.i.d. samples of X lies in the TS A_n^ϵ is larger than $1 - \epsilon$
2. Receive positive integer $n > N$ from Alice and construct TS A_n^ϵ

Proof of Achievability (Bob's Strategy)

1. Given $\epsilon > 0$ and discrete r.v. X, give Alice positive integer N such that for all positive integers $n > N$, probability that a sequence of outcomes produced by n i.i.d. samples of X lies in the TS A_n^ϵ is larger than $1 - \epsilon$
2. Receive positive integer $n > N$ from Alice and construct TS A_n^ϵ
3. Sort the elements in the TS

Proof of Achievability (Bob's Strategy)

1. Given $\epsilon > 0$ and discrete r.v. X, give Alice positive integer N such that for all positive integers $n > N$, probability that a sequence of outcomes produced by n i.i.d. samples of X lies in the TS A_n^ϵ is larger than $1 - \epsilon$
2. Receive positive integer $n > N$ from Alice and construct TS A_n^ϵ
3. Sort the elements in the TS
4. Construct the encoder as follows:
   - if input is in TS, output index of sequence in TS using $n(H(X) + \epsilon)$ binary bits. (This happens with probability of at least $1 - \epsilon$)
   - otherwise explode

Summary

1. What?
   - What is the Source Coding Theorem?
2. Why?
   - Why is the Source Coding Theorem important?
3. How?
   - How do you prove the Source Coding Theorem?

Summary

1. What?
   - What is the Source Coding Theorem?
2. Why?
   - Why is the Source Coding Theorem important?
3. How?
   - How do you prove the Source Coding Theorem?

> "Claude Shannon, the founder of information theory, invented a way to measure 'the amount of information' in a message without defining the word 'information' itself, nor even addressing the question of the meaning of the message."
> - Information, The New Language of Science (2003)

The End

Figure: **Pixabay: Pixabay License**

"Did you find this talk to be a piece of (chocolate) cake?"