4F5: Advanced Communications and Coding

Handout 2: The Typical Set, Compression, Mutual Information

Ramji Venkataramanan

Signal Processing and Communications Lab
Department of Engineering
ramji.v@eng.cam.ac.uk

Michaelmas Term 2015

The Typical Set

Recall the AEP:
If $X_1, X_2, \ldots$ are i.i.d. $\sim P$, then for any $\epsilon > 0$

$$Pr\left(\left| -\frac{1}{n} \log P(X_1, X_2, \ldots, X_n) - H(X) \right| < \epsilon \right) \overset{n\to\infty}{\longrightarrow} 1.$$

The *typical set* $A_{\epsilon,n}$ with respect to P is the set of sequences $(x_1, \ldots, x_n) \in \mathcal{X}^n$ with the property

$$2^{-n(H(X)+\epsilon)} \leq P(x_1, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$

"Sequences whose probability is concentrated around $2^{-nH(X)}$"

- Note the dependence on n and ϵ
- A sequence belonging to the typical set is called an ϵ-typical sequence.

Properties of the Typical Set

Notation: We will use X^n to denote the vector $X_1, X_2, \ldots, X_n$

Property 1

If $X^n = (X_1, \ldots, X_n)$ is generated i.i.d. $\sim P$, then

$$Pr(X^n \in A_{\epsilon,n}) \overset{n\to\infty}{\longrightarrow} 1$$

Proof: From the definition of $A_{\epsilon,n}$, note that

$$\begin{aligned} X^n \in A_{\epsilon,n} &\Leftrightarrow 2^{-n(H(X)+\epsilon)} \le P(X^n) \le 2^{-n(H(X)-\epsilon)} \qquad (1) \\ &\Leftrightarrow H(X) - \epsilon \le -\tfrac{1}{n}\log P(X^n) \le H(X) + \epsilon \end{aligned}$$

AEP says that

$$Pr(X^n \in A_{\epsilon,n}) = Pr\Big(H(X) - \epsilon \le -\tfrac{1}{n}\log P(X^n) \le H(X) + \epsilon\Big) \overset{n\to\infty}{\longrightarrow} 1.$$

□

Property 2

$$|A_{\epsilon,n}| \;\le\; 2^{n(H(X)+\epsilon)}$$

($|A_{\epsilon,n}|$ is the number of elements in the set $A_{\epsilon,n}$)

Proof:

$$\begin{aligned} 1 &= \sum_{x^n \in \mathcal{X}^n} P(x^n) \\ &\ge \sum_{x^n \in A_{\epsilon,n}} P(x^n) \\ &\overset{(a)}{\ge} \sum_{x^n \in A_{\epsilon,n}} 2^{-n(H(X)+\epsilon)} \\ &= 2^{-n(H(X)+\epsilon)}\, |A_{\epsilon,n}| \end{aligned}$$

Hence $|A_{n,\epsilon}| \;\le\; 2^{n(H(X)+\epsilon)}$. (Inequality (a) follows from the definition of the typical set.) □

Property 3

For sufficiently large n, $|A_{\epsilon,n}| \geq (1-\epsilon)\, 2^{n(H(X)-\epsilon)}$

Proof: From Property 1, $Pr(X^n \in A_{\epsilon,n}) \to 1$ as $n \to \infty$. This means that for any $\epsilon > 0$, for sufficiently large n we have $Pr(X^n \in A_{\epsilon,n}) > 1 - \epsilon$. Thus, for sufficiently large n:

$$\begin{aligned} 1-\epsilon &< Pr(X^n \in A_{\epsilon,n}) \\ &= \sum_{x^n \in A_{\epsilon,n}} P(x^n) \\ &\stackrel{(b)}{\leq} \sum_{x^n \in A_{\epsilon,n}} 2^{-n(H(X)-\epsilon)} \\ &= 2^{-n(H(X)-\epsilon)}\, |A_{\epsilon,n}| \end{aligned}$$

Hence $|A_{n,\epsilon}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$. (Inequality (b) follows from the definition of the typical set.) □

Properties of the Typical Set

Summary: For large n,

- Suppose you generate $X_1, \ldots, X_n$ i.i.d. $\sim P$. With high probability, the sequence you obtain will be typical, i.e., its probability is close to $2^{-nH(X)}$.
- The number of typical sequences is close to $2^{nH(X)}$.

How is all this relevant to communication ?

We will soon answer this. First, data compression.

Compression

GOAL: To compress a source producing symbols $X_1, X_2, \ldots$ that are i.i.d. $\sim P$

- For concreteness, consider English text:

$$\mathcal{X} = \{a \ b \ \ldots \ z \ , \ . \ space \ ; \ @ \ \#\} \qquad |\mathcal{X}| = 32$$

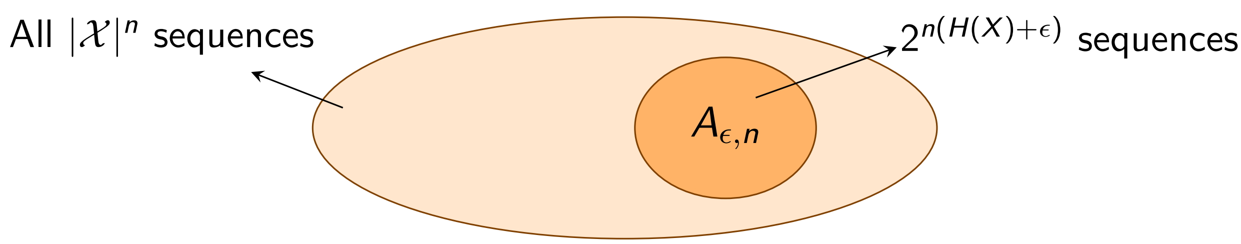
  (English text is not really i.i.d., but for now assume it is)
- Assume that we know the source entropy $H(X)$.
- $H(X)$ can be estimated by measuring the frequency of each symbol. E.g. by measuring the frequencies of a, b etc. separately, the entropy estimate for English text is ≈ 4 bits.

Naïve Representation

- List all the $|\mathcal{X}|^n$ possible length n sequences.
- *Index* these as $\{0, 1, \ldots, |\mathcal{X}|^n - 1\}$ using $\lceil \log |\mathcal{X}|^n \rceil$ bits

Number of bits/ sequence $= n \log |\mathcal{X}|$ (5n for English)

Compression via the Typical Set



Compression scheme:

- There are at most $2^{n(H(X)+\epsilon)}$ ϵ-typical sequences. ($2^{n(4+\epsilon)}$ for our example)
- Index each sequence in $A_{\epsilon,n}$ using $\lceil \log 2^{n(H(X)+\epsilon)} \rceil$ bits. Prefix each of these by a flag bit 0.

$$\text{Bits/typical seq.} = \lceil n(H(X)+\epsilon) \rceil + 1 \leq n(H(X)+\epsilon) + 2$$

- Index each sequence *not* in $A_{\epsilon,n}$ using $\lceil \log |\mathcal{X}|^n \rceil$ bits . Prefix each of these by a flag bit 1.

$$\text{Bits/non-typical seq.} = \lceil n \log |\mathcal{X}| \rceil + 1 \leq n \log |\mathcal{X}| + 2$$

This code assigns a *unique* codeword to each sequence in $|\mathcal{X}|^n$

Average code length

Let $\ell(X^n)$ be length of the codeword assigned to sequence X^n.

$$
\begin{aligned}
\mathbb{E}[\ell(X^n)] &= \sum_{x^n} P(x^n)\ell(x^n) \\
&= \sum_{x^n \in A_{\epsilon,n}} P(x^n)\ell(x^n) + \sum_{x^n \notin A_{\epsilon,n}} P(x^n)\ell(x^n) \\
&\leq \sum_{x^n \in A_{\epsilon,n}} P(x^n)(n(H(X)+\epsilon)+2) \; + \sum_{x^n \notin A_{\epsilon,n}} P(x^n)(n\log|\mathcal{X}|+2) \\
&\leq 1 \cdot n(H(X)+\epsilon) \; + \; \epsilon \cdot n\log|\mathcal{X}| + 2 \\
&= n(H(X)+\epsilon) + \epsilon n\log|\mathcal{X}| + 2 \\
&= n(H(X)+\epsilon')
\end{aligned}
$$

where $\epsilon' = \epsilon + \epsilon\log|\mathcal{X}| + \frac{2}{n}$.

ϵ' can be made arbitrarily small by picking ϵ small enough and then n sufficiently large.

Fundamental Limit of Compression

We have just shown that we can represent sequences X^n using $nH(X)$ bits on the average.

More precisely . . .

Let X^n be i.i.d. $\sim P$. Fix any $\epsilon > 0$. For n sufficiently large, there exists a code that maps sequences x^n of length n into binary strings such that the mapping is *one-to-one* and

$$\mathbb{E}\left[\frac{1}{n}\ell(X^n)\right] \leq H(X) + \epsilon.$$

In fact, more is true – you cannot do any better than $H(X)$, i.e.,

The expected length of *any* uniquely decodable code satisfies

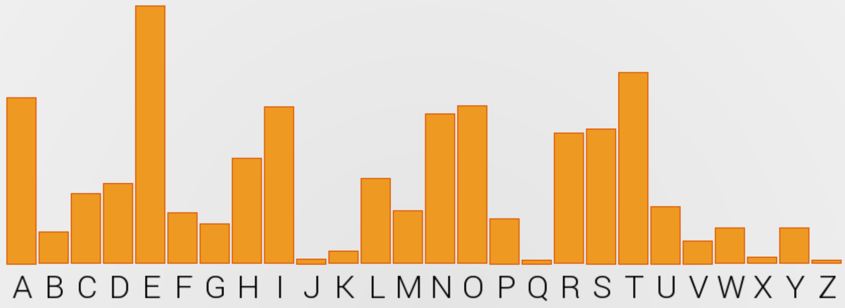
$$\mathbb{E}\left[\frac{1}{n}\ell(X^n)\right] \geq H(X)$$

(For a proof of this, see [Cover & Thomas, Chapter 5]; also in 3F1 notes.)

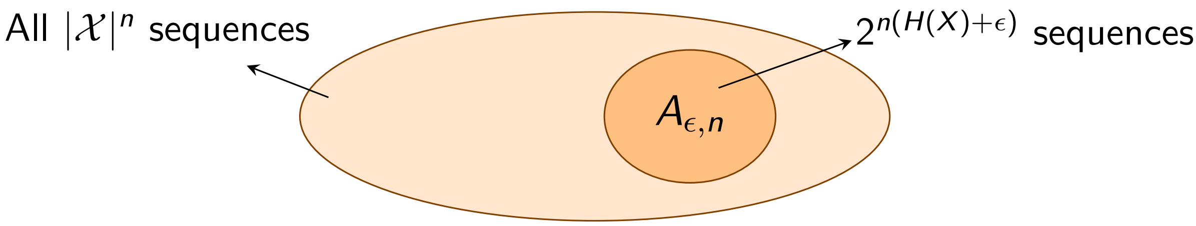
Entropy is the fundamental limit of lossless compression

Histogram of English text

☞ Norvig (2013)



This is a histogram of letter occurrences in English text, obtained by ☞ Norvig (2013) using millions of books from Google's n-gram corpus.



- The typical set is *very* small subset of the set of all sequences, but contains almost all the probability!
- This is the reason that even with an i.i.d. model, English text can be compressed to $\approx$ 4 bits/sample.
- Can compress even more if we consider correlations in the text. E.g. q always followed by u.
- What kind of source cannot be compressed at all ?

Is this scheme *practical*?

- No. To find the codeword for any x^n, we need to go through a table of 2^{nH} entries – *computationally complex*!
- Practical schemes like Huffman coding, Lempel-Ziv achieve rates close to the entropy with much lower complexity. (3F1)

Relative entropy

The *relative entropy* or the Kullback-Leibler (KL) divergence between two pmfs P and Q is

$$D(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

(Note: P and Q are defined on the same alphabet $\mathcal{X}$)

- Measure of distance between distributions P and Q
- Not a true distance. For e.g., $D(P||Q) \neq D(Q||P)$.
- $D(P||Q) \geq 0$ with equality if and only if $P = Q$.
  Proof : First use $\log a = \frac{\ln a}{\ln 2}$. Then use the fact that $\ln a \leq (a-1)$ with equality iff $a = 1$.

Mutual Information

Consider two random variables X and Y with joint pmf P_{XY}. The *mutual information* between X and Y is defined as

$$I(X;Y) = H(X) - H(X|Y) \quad \text{bits.}$$

"Reduction in the uncertainty of X when you observe Y"

Property 1

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

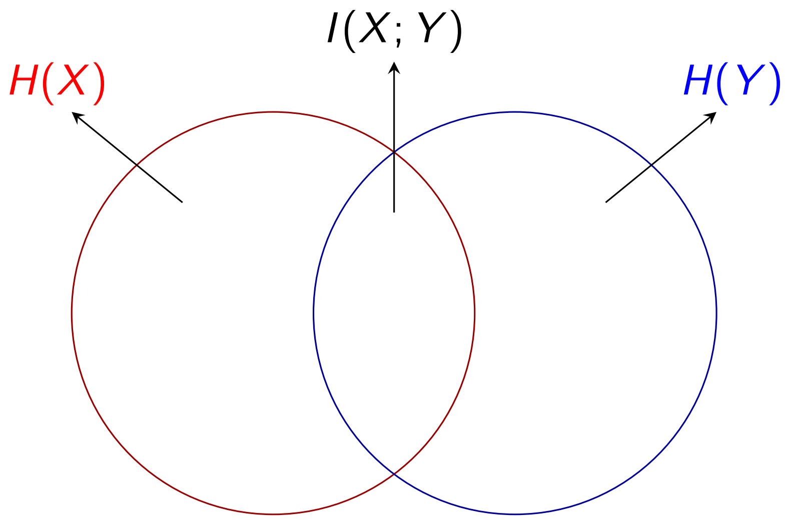
"X says as much about Y as Y says about X"

Proof : From the chain rule of entropy,

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

In the definition of $I(X;Y)$, use $H(X|Y) = H(X,Y) - H(Y)$. □

Venn Diagram



The two circles together represent $H(X, Y)$

Questions

1. What is $I(X;Y)$ when X and Y are independent? Ans: 0
2. What is $I(X;Y)$ when $Y = X$? Ans: $H(X)$
3. What is $I(X;Y)$ when $Y = f(X)$? Ans: $H(Y) = H(f(X))$

Example

X is the event that tomorrow is cloudy; Y is the event that it will rain tomorrow. Joint pmf P_{XY}:

	Rain	No Rain
Cloudy	3/8	3/8
Not cloudy	1/16	3/16

In Handout 1, we calculated:

$$H(X,Y) = 1.764, \qquad H(X) = 0.811, \qquad H(Y|X) = 0.953$$

To compute $I(X;Y)$, we need to compute $H(Y)$ (or $H(X|Y)$).

$$P(Y = \text{ rainy}) = \frac{3}{8} + \frac{1}{16} = \frac{7}{16}, \quad P(Y = \text{ not rainy}) = \frac{9}{16}$$
$$H(Y) = 0.989$$
$$I(X;Y) = H(Y) - H(Y|X) = 0.036$$

Verify that you get the same answer by computing $H(X|Y)$ and using $I(X;Y) = H(X) - H(X|Y)$.

Property 2 of Mutual Information

$$I(X;Y) = D\left(P_{XY} || P_X P_Y\right)$$

"The relative entropy between the joint pmf and the product of the marginals"

Proof:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= -\sum_x P_X(x) \log P_X(x) + \sum_{x,y} P_{XY}(x,y) \log P_{X|Y}(x|y) \\ &= \sum_{x,y} P_{XY}(x,y) \log \frac{P_{X|Y}(x|y)}{P_X(x)} \\ &= \sum_{x,y} P_{XY}(x,y) \log \frac{P_{X|Y}(x|y) P_Y(y)}{P_X(x) P_Y(y)} \\ &= D\left(P_{XY} || P_X P_Y\right). \end{aligned}$$

$\square$

Property 3

$$I(X;Y) \geq 0$$

Proof: Follows from Property 2 because $D(P||Q) \geq 0$ for any pair of pmfs P, Q.

Implication:

$$H(X|Y) \leq H(X), \quad H(Y|X) \leq H(Y)$$

" Knowing another random variable Y can only reduce the average uncertainty in X"

Preview:

- Let X be the input to a communication channel, and Y the output.
- We will show that $I(X;Y)$ is key to understanding of how much information can be transmitted over the channel.

"Reduction in the uncertainty of the channel input X when you observe the output Y"

You can now do Questions 1 – 10 on Examples Paper 1.