4F5: Advanced Communications and Coding

Handout 1: Introduction, Entropy, Typicality, AEP

Ramji Venkataramanan

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

Michaelmas Term 2015

Course Information

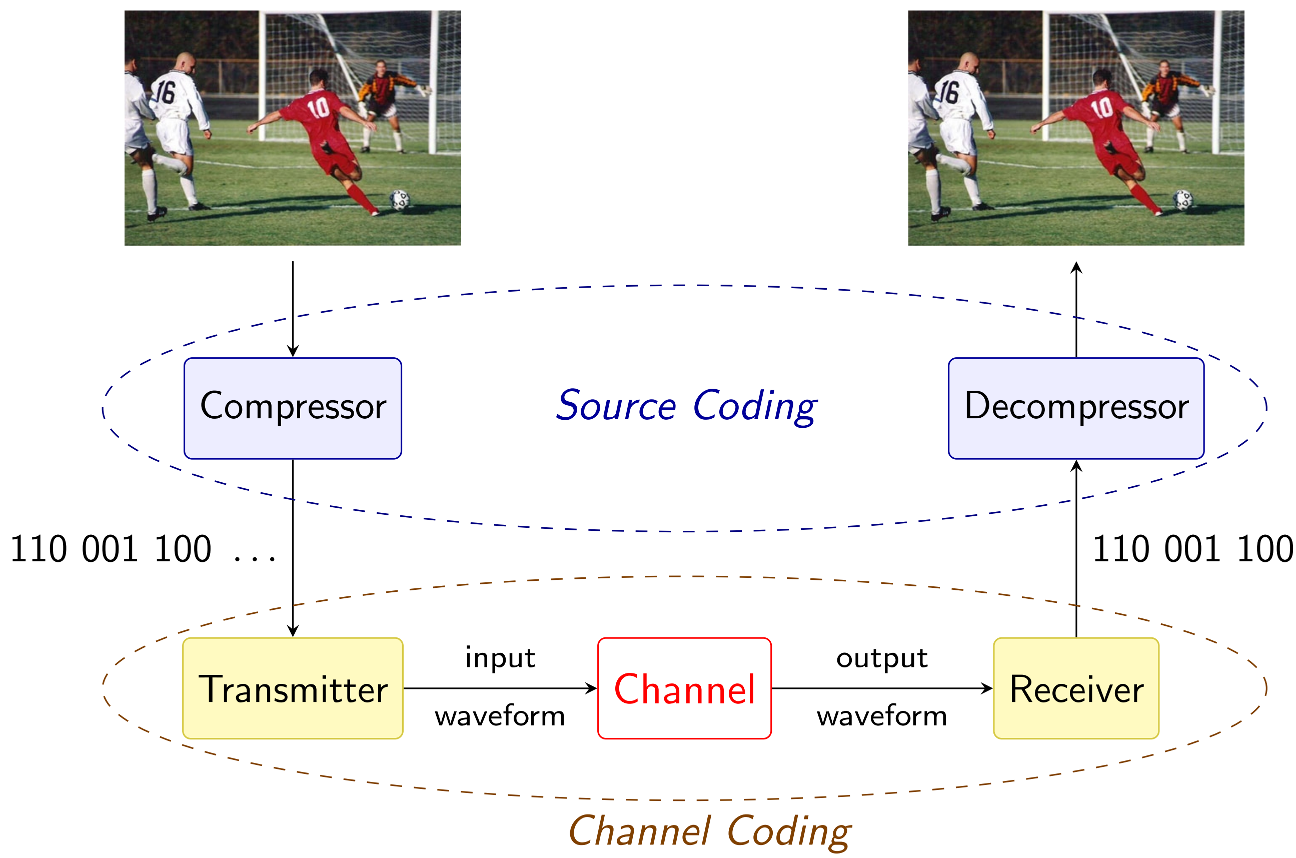
1. 16 Lectures in three parts:
   - Information Theory (5L, Ramji)
   - Coding (6L, Jossy Sayir)
   - Modulation and Wireless Communication (5L, Ramji)
2. Main pre-requisite: good background in probability
   - 1B Paper 7, 3F1 highly recommended
   - 3F4 useful, but not required
3. Handouts, examples sheets, announcements on Moodle
   `https://www.vle.cam.ac.uk`
4. Drop-in 'supervision' hours (Ramji): Tuesdays 2:00-3:30pm in BE3-12: e-mail me if you cannot make these times

Questions and active participation in lectures encouraged!

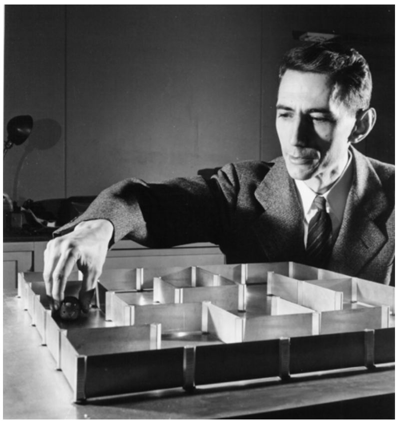
Useful References

- T. M. Cover and J. A. Thomas,
  Elements of Information Theory,
  Wiley Series in Telecommunications, 2nd Edition, 2006.
- D. J. C. MacKay,
  Information theory, inference, and learning algorithms,
  Cambridge University Press, 2003. (free online version)
- R. G. Gallager,
  Principles of Digital Communications,
  Cambridge University Press, 2008.
- T. Richardson, R. Urbanke
  Modern Coding Theory,
  Cambridge University Press, 2008. (free online version)
- D. Tse and P. Viswanath,
  Fundamentals of Wireless Communication,
  Cambridge University Press, 2005. (free online version)

An End-to-End Communication System



Two Fundamental Limits



Claude Shannon (1948: A Mathematical Theory of Communication)

Posed and answered two fundamental questions:

- Given a source of data, how much can compress it?
- Given a channel, at what rate can you transmit data?

How do you *model* sources and channels?
Using probability distributions.

Probability Review

A *random variable* (rv) X:

- Is a function that maps outcome of experiment to value in set $\mathcal{X}$. This definition is not completely rigorous, but suffices.
- Can be discrete (e.g. $\mathcal{X} = \{0, 1\}$) or continuous (e.g. $\mathcal{X} = \mathbb{R}$)

Discrete Random Variables

- Characterised by a probability mass function (pmf): $P_X(x) = Pr(X = x)$. $x \in \mathcal{X}$ is a *realisation* of the rv X
- Cumulative distribution function (cdf) : $F_X(a) = Pr(X \leq a) = \sum_{x \leq a} P_X(x)$
- Expected value: $\mathbb{E}[X] = \sum_a a P_X(a)$
- A function $g(X)$ of an rv X is also an rv
- We will often take expectations of functions of rvs, e.g. $\mathbb{E}[g(X)] = \sum_a g(a) P_X(a)$

We sometimes drop the subscript and write $P(x)$ — need to be careful!

Jointly distributed discrete rvs X, Y:

- Joint pmf $P_{XY}(x,y), \; x \in \mathcal{X}, y \in \mathcal{Y}$
- Conditional distribution of Y *given* X:

$$P_{Y|X}(y|x) = \frac{P_{XY}(x,y)}{P_X(x)} \quad \text{for } x \text{ such that } P_X(x) > 0.$$

- Two key properties: For rvs X, Y, Z:
  1. Product rule:

$$\begin{aligned} P_{XYZ} &= P_X P_{Y|X} P_{Z|YX} \\ &= P_Y P_{X|Y} P_{Z|XY} \\ &= P_Z P_{X|Z} P_{Y|XZ} \qquad \text{etc.} \end{aligned}$$

  2. Sum rule (marginalisation):

$$P_{XY}(x,y) = \sum_z P_{XYZ}(x,y,z)$$

$$P_X(x) = \sum_{y,z} P_{XYZ}(x,y,z) = \sum_y P_{XY}(x,y) \qquad \text{etc.}$$

Continuous random variables X, Y:

- Joint *density* function $f_{XY}(x,y), \; x \in \mathbb{R}, y \in \mathbb{R}$
- $Pr(a \leq X \leq b, \; c \leq Y \leq d) = \int_{x=a}^{b} \int_{y=c}^{d} f_{XY}(x,y)dxdy$
- Important example: (X, Y) *jointly Gaussian* rvs.
  In this case, f_{XY} is fully specified by the mean vector $\underline{\mu}$ and covariance matrix Σ
- Conditional density, product and sum rule analogous to discrete case, with density replacing the pmf and integrals instead of sums

Independence

Discrete random variables $X_1, X_2, \ldots, X_n$ are statistically *independent* if

$$P_{X_1 \ldots X_n}(x_1, \ldots, x_n) = P_{X_1}(x_1)\, P_{X_2}(x_2) \ldots P_{X_n}(x_n) \quad \forall (x_1, \ldots, x_n).$$

Recall the product rule: we can *always* write

$$\begin{aligned} &P_{X_1 \ldots X_n}(x_1, \ldots, x_n) \\ &= P_{X_1}(x_1)\, P_{X_2|X_1}(x_2|x_1) \ldots P_{X_n|X_{n-1} \ldots X_1}(x_n|x_{n-1}, \ldots, x_1) \end{aligned}$$

Thus when $X_1, \ldots, X_n$ are independent, we have

$$P_{X_i|\{X_j\}_{j \neq i}} = P_{X_i}$$

We will often consider independent and *identically distributed* (i.i.d.) random variables, i.e., $P_{X_1} = P_{X_2} = \ldots = P_{X_n} = P$

Review your notes from 1B Paper 7 and 3F1!

Weak Law of Large Numbers (WLLN)

Roughly: "*Empirical average converges to the mean*"

Formal statement

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite mean μ. Let $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, for any $\epsilon > 0$

$$\lim_{n \to \infty} Pr(|S_n - \mu| < \epsilon) = 1.$$

Much of information theory is a (very clever) application of WLLN!

Entropy

The entropy of a discrete random variable X with pmf P is

$$H(X) = \sum_x P(x) \log \frac{1}{P(x)} \quad \text{bits}$$

- "log" in this course will mean $\log_2$
- For x such that $P(x) = 0$, "$0 \log \frac{1}{0}$" is the limiting value 0
- Can be written as $\mathbb{E}[\log \frac{1}{P(X)}]$
- $H(X)$ is the **uncertainty** associated with the rv X.

Example

1. Let rv X represent the event of England winning the World Cup. Let
   $X = 1$ with probability 0.2, and $X = 0$ with probability 0.8
2. Let the rv Y represent the event of rain tomorrow.
   $Y = 1$ with probability 0.4, and $Y = 0$ with probability 0.6

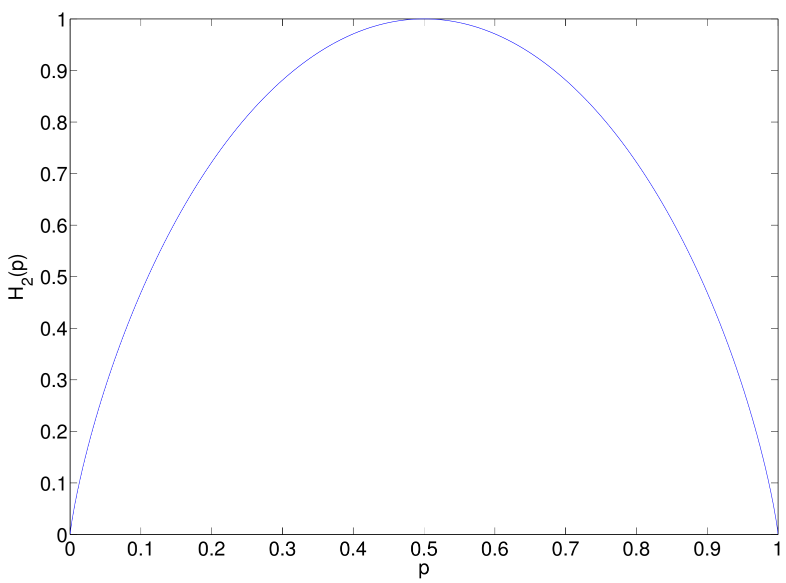
Which event has greater entropy?

Binary Entropy Function

Bernoulli RVs:

X is called a Bernoulli(p) random variable if takes value 1 with probability p and 0 with probability $1 - p$. Its entropy is

$$H_2(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$$



As p approaches $\frac{1}{2}$, more uncertainty about the outcomes of X

Exercise

Suppose that we have a horse race with 4 horses. Assume that the probabilities of winning for the 4 horses are $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$. What is the entropy of the race? Answer: $\frac{7}{4}$ bits

Properties of Entropy

For a discrete random variable X taking values in $\mathcal{X}$:

1. $H(X) \geq 0$ (because $\frac{1}{P(x)} \geq 1$ implies $\log \frac{1}{P(x)} \geq 0$)
2. If we denote the alphabet size by $|\mathcal{X}|$, then $H(X) \leq \log|\mathcal{X}|$
   Proof: Use the inequality $\ln x \leq (x-1)$ for $x \geq 0$. Also note that $\log x = \frac{\ln x}{\ln 2}$. (In 3F1 examples paper)
3. Among all random variables taking values in $\mathcal{X}$, the equiprobable distribution $(\frac{1}{|\mathcal{X}|}, \ldots, \frac{1}{|\mathcal{X}|})$ has the maximum entropy, equal to $\log|\mathcal{X}|$.

Joint and Conditional Entropy

The *joint* entropy of discrete rvs X, Y with joint pmf P_{XY} is

$$H(X,Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_{XY}(x,y)}$$

The *conditional* entropy of Y given X is

$$H(Y|X) = \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_{Y|X}(y|x)}$$

- $H(Y|X)$ is the *average* uncertainty in Y given X:

$$H(Y|X) = \sum_{x} P_X(x) \underbrace{\sum_{y} P_{Y|X}(y|x) \log \frac{1}{P_{Y|X}(y|x)}}_{H(Y|X=x)}$$

- $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
  (verify using product and sum rule of probability)

Example

Let X be the event that tomorrow is cloudy; Y be 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

$$H(X,Y) = \frac{3}{8}\log\frac{8}{3} + \frac{3}{8}\log\frac{8}{3} + \frac{1}{16}\log 16 + \frac{3}{16}\log\frac{16}{3}$$
$$= 1.764 \text{ bits}$$

$P(X = \text{ cloudy}) = \frac{3}{8} + \frac{3}{8} = \frac{3}{4}$, $P(X = \text{ not cloudy}) = \frac{1}{4}$

$H(X) = 0.811$ bits

$H(Y|X) = H(X,Y) - H(X) = 0.953$ bits

Exercise: Compute $H(Y|X)$ directly; Compute $H(Y)$, $H(X|Y)$

Joint Entropy of Multiple RVs

The *joint* entropy of $X_1, \ldots, X_n$ with joint pmf $P_{X_1 \ldots X_n}$ is

$$H(X_1, X_2, \ldots, X_n) = \sum_{x_1,\ldots,x_n} P_{X_1\ldots X_n}(x_1,\ldots,x_n) \log \frac{1}{P_{X_1\ldots X_n}(x_1,\ldots,x_n)}$$

Chain Rule of Joint Entropy:
The joint entropy can be decomposed as

$$H(X_1, X_2 \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1)$$

where the *conditional* entropy

$$H(X_i|X_{i-1},\ldots,X_1) = \sum_{x_1,\ldots,x_i} P_{X_1,\ldots,X_i}(x_1,\ldots,x_i) \log \frac{1}{P_{X_i|X_1,\ldots,X_{i-1}}(x_i|x_1,\ldots,x_{i-1})}$$

(The chain rule is a generalisation of $H(X,Y) = H(X) + H(Y|X)$.)

Proof of Chain Rule

Recall that

$$P(x_1,\ldots,x_n) = P(x_1)P(x_2|x_1)\ldots P(x_n|x_{n-1},\ldots,x_1) = \prod_{i=1}^{n} P(x_i|x_{i-1},\ldots,x_1)$$

(For brevity, we drop the subscripts on $P_{X_1\ldots X_n}$)

$$H(X_1,X_2\ldots,X_n) = -\sum_{x_1,\ldots,x_n} P(x_1,\ldots,x_n)\log P(x_1,\ldots,x_n)$$

$$= -\sum_{x_1,\ldots,x_n} P(x_1,\ldots,x_n)\log\prod_{i=1}^{n} P(x_i|x_{i-1},\ldots,x_1)$$

$$= -\sum_{x_1,\ldots,x_n}\sum_{i=1}^{n} P(x_1,\ldots,x_n)\log P(x_i|x_{i-1},\ldots,x_1)$$

$$= -\sum_{i=1}^{n}\sum_{x_1,\ldots,x_n} P(x_1,\ldots,x_n)\log P(x_i|x_{i-1},\ldots,x_1)$$

$$= -\sum_{i=1}^{n}\sum_{x_1,\ldots,x_i} P(x_1,\ldots,x_i)\log P(x_i|x_{i-1},\ldots,x_1) = \sum_{i=1}^{n} H(X_i|X_{i-1},\ldots,X_1)$$

Joint Entropy of Independent RVs

If $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$H(X_1,X_2,\ldots,X_n) = \sum_{i=1}^{n} H(X_i)$$

Proof.

Due to independence, $P_{X_i|X_{i-1},\ldots,X_1} = P_{X_i}$ for $i = 2,\ldots,n$. Use this to show that for all i

$$H(X_i|X_{i-1},\ldots,X_1) = H(X_i).$$

The result then follows from the chain rule. □

Typicality – a simple example

Consider an i.i.d. Bernoulli($\frac{1}{4}$) source. It produces symbols $X_1, X_2, \ldots$ according to

$$P(X_i = 1) = \tfrac{1}{4}, \quad P(X_i = 0) = \tfrac{3}{4} \quad \text{for } i = 1, 2, \ldots$$

One of the following sequences is a "real" output of the source.

1. 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2. 0 1 0 0 0 0 0 0 1 1 0 0 0 1 0 0
3. 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Which sequence is a real output?

1. 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2. 0 1 0 0 0 0 0 0 1 1 0 0 0 1 0 0
3. 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

- Each sequence has 16 bits.
- The probability of a sequence with k ones and $16 - k$ zeros is

$$\left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{16-k}$$

- Probability of source emitting first sequence $= \left(\frac{3}{4}\right)^{16}$
- Probability of source emitting second sequence $= \left(\frac{1}{4}\right)^4 \left(\frac{3}{4}\right)^{12}$

The *first* sequence is 3^4 times more likely that the second!

Typical sequences

1. 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2. 0 1 0 0 0 0 0 0 1 1 0 0 0 1 0 0

- Though less likely than the first sequence, the second sequence seems more "typical" of the $(\frac{1}{4}, \frac{3}{4})$ source.
- We will make this idea precise via the notion of a typical set.

We will show that if $X_1, \ldots, X_n$ are chosen $\sim$ i.i.d. Bernoulli(p), then for large n:

- With high probability, the fraction of ones in the observed sequence will be close to p
- Equivalently: with high probability, the observed sequence will have probability close to $p^{np}(1-p)^{n(1-p)}$
- Note that any number a can be written as $2^{\log a}$. Hence
$$p^{np}(1-p)^{n(1-p)} = (2^{\log p})^{np}\ (2^{\log(1-p)})^{n(1-p)} = 2^{-nH_2(p)}$$

An *operational* meaning of entropy: For large n, almost all sequences have probability close to $2^{-nH_2(p)}$!

Asymptotic Equipartition Property

- We will prove such a "concentration" result for i.i.d. discrete sources, specifying exactly what "with high probability" means
- The main tool: Asymptotic Equipartition Property (AEP)

AEP

If $X_1, X_2, \ldots$ are i.i.d. $\sim\ P_X$, then for any $\epsilon > 0$

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

Remarks:

- $\frac{-1}{n} \log P_X(X_1, X_2, \ldots, X_n)$ is a *random variable*
  (Note the capitals; A function of the rvs $(X_1, \ldots, X_n)$ is a rv)
- AEP says this rv converges to $H(X)$, a **constant**, as $n \to \infty$.

Proof of the AEP

Simple application of Weak Law of Large Numbers (WLLN). Let

$$Y_i = -\log P_X(X_i), \quad \text{for } i = 1, \ldots, n.$$

- *Functions of independent rvs are also independent rvs*
  $\Rightarrow \quad Y_1, \ldots, Y_n$ are i.i.d.
- WLLN for Y_i's says that for any $\epsilon > 0$

$$\lim_{n\to\infty} Pr\left(\left|\frac{1}{n}\sum_i Y_i - \mathbb{E}[Y_1]\right| < \epsilon\right) = 1. \tag{1}$$

- Note that

$$\sum_i Y_i = -\sum_i \log P_X(X_i) = -\log\left[P_X(X_1)P_X(X_2)\ldots P_X(X_n)\right]$$
$$\overset{\text{why?}}{=} -\log P_X(X_1, X_2, \ldots, X_n) \tag{2}$$

- Substitute (2) in (1), and note that $\mathbb{E}[Y_1] = H(X)$ to get the AEP.

$\square$