Mobile Communication Systems

Part I: Fundamental Limits

Professor Robert J. Piechocki

Merchant Venturers Building, room MVB 4.22
email: r.j.piechocki@bristol.ac.uk; tel: 45655

2020-21 Course structure

- TB1 (weeks 1-11)
- Weeks 1-5: Fundamental Limits of Communications
- Weeks 7-11: Mobile Comms Technologies
- Synch / Example / Reserve classes
- 2 Courseworks

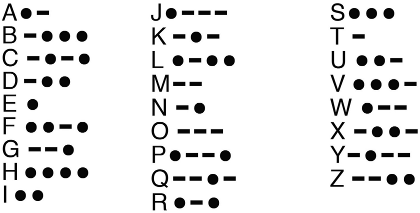
Lecture plan, Part I

1. Probability (quick revision)
2. Uncertainty, Entropy and Mutual Information
3. Capacity of communications channels
4. Differential entropy, Gaussian channels
5. Concatenated Channels, Markov Chains, Data processing inequality
6. Wireless Channels, MIMO Channels
7. Data compression, Kraft inequality, Huffman algorithm, Lempel-Ziv (optional material, non-examinable)

Literature

- Thomas M. Cover, Joy A. Thomas, *Elements of Information Theory*, Wiley 2006. *(Ebook available for free from UoB catalogue)*
- David J. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press 2003. *(available for free from: http://www.inference.phy.cam.ac.uk/mackay/)*
- John Proakis, *Digital Communications* (chapter 6), McGraw-Hill, 2008.
- James V Stone, *Information Theory - A Tutorial Introduction*, Sebtel Press, 2015.

A motivating example



A	●▬
B	▬●●●
C	▬●▬●
D	▬●●
E	●
F	●●▬●
G	▬▬●
H	●●●●
I	●●
J	●▬▬▬
K	▬●▬
L	●▬●●
M	▬▬
N	▬●
O	▬▬▬
P	●▬▬●
Q	▬▬●▬
R	●▬●
S	●●●
T	▬
U	●●▬
V	●●●▬
W	●▬▬
X	▬●●▬
Y	▬●▬▬
Z	▬▬●●

Morse code assigns short codes (codewords) to most frequently used letters in English language - why?

Motivating example - cont'd

Let's develop our own code!

Our new language has only 4 letters: $\alpha, \beta, \gamma, \delta$, with the following probabilities:

	$P\{X = x\}$
α	$\frac{1}{2}$
β	$\frac{1}{4}$
γ	$\frac{1}{8}$
δ	$\frac{1}{8}$

- The most obvious binary code would be:

	$P\{X = x\}$	$C0$
α	$\frac{1}{2}$	00
β	$\frac{1}{4}$	01
γ	$\frac{1}{8}$	10
δ	$\frac{1}{8}$	11

Hence on average we would use 2 bits. Can we do any better?

Motivating example - cont'd

How about $C1$:

	$P\{X = x\}$	$C0$	$C1$
α	$\frac{1}{2}$	00	0
β	$\frac{1}{4}$	01	10
γ	$\frac{1}{8}$	10	110
δ	$\frac{1}{8}$	11	111

Let's calculate how many bits on average do we use?

$$L = \sum_{i=1}^{n} p_i l(x_i) = \frac{1}{2}1 + \frac{1}{4}2 + \frac{1}{8}3 + \frac{1}{8}3 = 1.75$$

- i.e. we have just saved 0.25 bits on average without loosing any *information.*
- What's more, our code is still uniquely decodable e.g:
  - 110010111 can only be decoded as $\gamma\alpha\beta\delta$

Fundamental Limits

Information Theory answers fundamental questions:

- How much data (bits/second) can be sent reliably over a noisy communications channel?
- What is the ultimate data compression for an information source?
- How accurately can we represent an object (e.g., audio waveform or image) as a function of the number of bits used (the rate)?

These questions are answered by the three main theorems of Information theory: Channel Coding, Source Coding, and Rate-Distortion theorems (and their converses).

Fundamental Capacity Limits

- 5G Systems can transmit at up to 10 Gbit/s. Is that the fundamental limit? How about 6G Systems?
- Fibre optic cable can support 100 Gbit/s as standard. Is that all?
- Given 10MHz bandwidth, what is the maximum achievable data rate?
- How much capacity can 6G 128x128 Massive MIMO System Achieve?

Other examples of Information theory applications

- Speech can be compressed 8-16 times with no perceptible loss in quality: 64,000 bps; 8,000 bps; 2400 bps - what is information content of speech?
- The uncompressed bit rate for the 8K/12bit/60 Hz format is approximately 72 Gbps. How much can High Efficiency Video Coding (HEVC), aka H.265, compress?

History of Information theory

The main content of this course was defined by Claude Shannon in his 1948 paper *"The Mathematical Theory of Communication"* (published in a 1949 book of the same title). Since that paper appeared:

- Theorems have been proved and improved.
- Practical methods have been invented and implemented:
  - source coding: Huffman codes (compact), Lempel-Ziv (compress, gzip)
  - channel coding: error-correcting codes (Hamming, Reed-Solomon, convolutional, trellis, turbo)
  - rate-distortion: vocoders, minidiscs, MP3, JPEG, MPEG

Implementations in most cases have caught up with theory.

Probability Review

- A probability ensemble is a triple $(x, \mathcal{X}, P_X)$, where the outcome x is the value of a random variable, which takes on one of a set of possible values $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$. We will denote it by:

$$P(x) = P_X(x) = P\{X = x\}$$

We have: $P(X = x_i) = p_i,\ p_i \geq 0$ and $\sum_{i=1}^{n} P(X = x_i) = 1$

Examples of pdfs:

- Uniform $P(x) = \frac{1}{n}, \qquad x \in \mathcal{X} = \{1, 2, \ldots, \mathrm{n}\}$
- Binomial $P(x) = \binom{n}{x} p^x (1-p)^{n-x}$,
- Geometric $P(x) = (1-p)^{x-1} p, \qquad x = 1, 2, 3 \ldots$

- A joint ensemble XY is an ensemble in which each outcome is an ordered pair x, y, with $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ and $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$. We call the $P\left(X = x_i, Y = y_j\right)$ the joint probability of X and Y.

- Marginal probability is given by the following summation:

$$P\left(X = x_i\right) = \sum_{j=1}^{m} P\left(X = x_i, y_j\right)$$

- The Conditional probability is defined as:

$$P\left(X = x_i \,|\, Y = y_j\right) = \frac{P\left(X = x_i, Y = y_j\right)}{P\left(Y = y_j\right)}$$

When the range of X is understood, say $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, we may denote $P(x_i)$ by p_i.

Joint probability distribution functions for two or more random variables can also be represented concisely:

$$P(x, y) = P_{XY}(x, y) = P\{X = x, Y = y\}$$

$$P(x, y, z) = P_{XYZ}(x, y, z) = P\{X = x, Y = y, Z = z\}$$

Conditional probabilities can also be abbreviated:
$P(y\,|x) = p_{Y|X}(y\,|x) = P\{Y = y\,|X = x\} = \frac{P\{X=x, Y=y\}}{P\{X=x\}} = \frac{P(x,y)}{P(x)}$

The definition of the conditional probability implies the chain rule for probabilities:

$$P(x, y) = P(y \,|x\,) \, P(x) = P(x \,|y\,) \, P(y)$$

$$P(x_1, x_2, \ldots, x_n) = P(x_1 \,|x_2, \ldots, x_n\,) \, P(x_2, \ldots, x_n)$$

Simple examples

Table: Joint probabilities (pmf) $P(x, y)$

	Y		
X	0.1	0.05	0.05
	0.2	0.1	0.05
	0.15	0.2	0.1

Table: Joint probabilities $P(x, y)$ and marginal probabilities

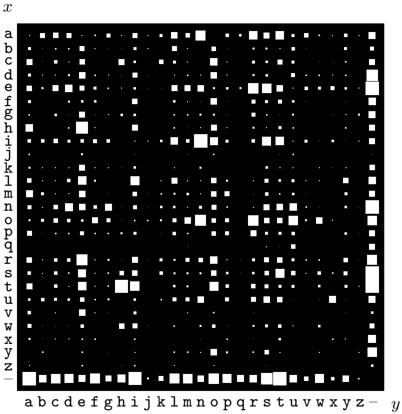
	Y			$P(x)$
X	0.1	0.05	0.05	0.2
	0.2	0.1	0.05	0.35
	0.15	0.2	0.1	0.45

Table: Conditional probability $P(y|x)$

$P(y\|x)$	Y			$P(x)$
X	0.5	0.25	0.25	0.2
	0.5714	0.2857	0.1429	0.35
	0.3333	0.4444	0.2222	0.45

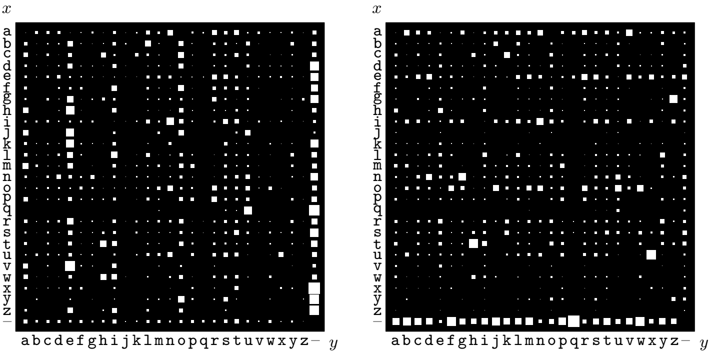
English language example

The probability distribution over 27 × 27 bigrams xy (ordered pairs) in the English Language



English language example

Conditional probability distribution. (a) $P(y|x)$: each row shows the distribution of the second letter Y given the first letter X in the bigram XY. (b) $P(x|y)$ each column shows conditional distribution of the first letter given the second letter



(a) $P(y\,|\,x)$

(b) $P(x\,|\,y)$

Independence

Two random variables X, Y are independent (denoted as $X \perp Y$) if and only if:

$$P(x, y) = P(x) P(y)$$

Also, iff $X \perp Y$

$$P(x|y) = \frac{P(x, y)}{P(y)} = \frac{P(x) P(y)}{P(y)} = P(x)$$

Bayes Theorem

Let X and Y be two arbitrary RV with $P\left(x\right) \neq 0$ and $P\left(y\right) \neq 0$. Then:

$$P\left(x\,|y\,\right) = \frac{P\left(y\,|x\,\right)P\left(x\right)}{P\left(y\right)}$$

Which can also be expressed as:

$$P\left(x\,|y\,\right) = \frac{P\left(y\,|x\,\right)P\left(x\right)}{\sum_{j=1}^{n} P\left(y\,|X = x_j\,\right)P\left(X = x_j\right)}$$

Where:
$P(y) = \sum_{j=1}^{n} P\left(y, X = x_j\right) = \sum_{j=1}^{n} P\left(y\,|X = x_j\,\right)P\left(X = x_j\right)$.
This result is very useful in evaluating causal relationships between events (an outcome and a cause). We can evaluate *a posteriori* probability $P\left(x\,|y\,\right)$ in terms of *a priori* probability $P\left(x\right)$ and *likelihood* $P\left(y\,|x\,\right)$.

Bayes Theorem

Example

A simple binary communications channel carries messages by using two signals 0 and 1. We assume that for 40% of the time 1 is transmitted; the probability that a transmitted 0 is received correctly is 0.95; transmitted 1 is received correctly is 0.9.

- Determine: (a) the probability that 0 is received, (b) given that 0 is received, the probability that 0 was transmitted.

i.e.

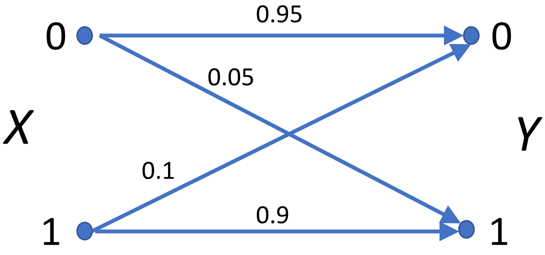
$P(X = 1) = 0.4$

$P(X = 0) = 0.6$

$P(Y = 0|X = 0) = 0.95$

$P(Y = 1|X = 1) = 0.9$

Bayes Theorem example



$P(X=1)=0.4,\ P(X=0)=0.6,$
$P(Y=0|X=0)=0.95, P(Y=1|X=0)=0.05$
$P(Y=1|X=1)=0.9, P(Y=0|X=1)=0.1$

Bayes Theorem example

- Calculate $P(Y = 0) = ?$

$P(Y = 0) = P(Y = 0|X = 1)P(X = 1) + P(Y = 0|X = 0)P(X = 0) = 0.1 \cdot 0.4 + 0.95 \cdot 0.6 = 0.61$

- Calculate $P(X = 0|Y = 0) = ?$

From Bayes theorem:
$P(X|Y) = \frac{P(Y=0|X=0)P(X=0)}{P(Y=0)} = \frac{0.95 \cdot 0.6}{0.61} = 0.9344$

Example: COVID test

A new coronavirus test is marketed. The test has *Sensitivity* of 98% and *Specificity* of 99%.

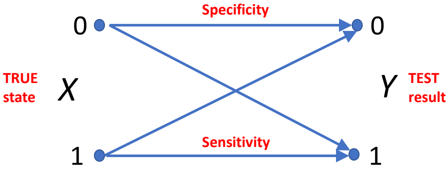
- Is it a good test?
- What is the probability that a person has the virus given the positive test result?

Assume prevalence of the virus in the population of 1%.

Example: COVID test as Communication channel

We will model the uncertainty of the test as a communication channel and use the Bayes Theorem.

- Specificity is a true negative rate (TNR) i.e. P(Y=0|X=0)
- Sensitivity is a true positive rate (TPR) i.e. P(Y=1|X=1)



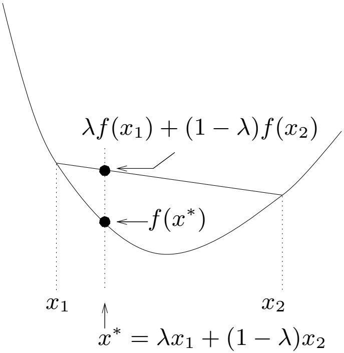
The probability that a person has the virus given the positive test result is $P(X=1|Y=1)$
$= \frac{P(Y|X)P(X)}{P(Y)} = \frac{0.98 \cdot 0.01}{0.0197} = 0.4975$

Convex functions

Convexity plays a crucial role in IT. The are many examples of convex functions: $x^2, \exp(x), -\log(x)$.

Convex functions are sometimes referred to as cup functions because their graphs can hold water (or smile functions $\smile$). Two of the above convex functions do not satisfy this condition, but that may be a technicality. The graphical definition of convexity is that the graph of the function lies below any chord connecting two points on the graph.

Convex functions


$\lambda f(x_1) + (1-\lambda) f(x_2)$
$f(x^*)$
x_1
x_2
$x^* = \lambda x_1 + (1-\lambda) x_2$

Convex functions

Definition

A function $f(x)$ is convex on an interval $a < x < b$ if

$$f((1-\lambda)x_1 + \lambda x_2) \leq (1-\lambda) f(x_1) + \lambda f(x_2)$$

for every $x_1; x_2$ such that $a < x_1; x_2 < b$ and λ such that $0 < \lambda < 1$. If

$$f((1-\lambda)x_1 + \lambda x_2) < (1-\lambda) f(x_1) + \lambda f(x_2)$$

then $f(x)$ is strictly convex.

Convex functions

A function $f(x)$ is concave if $-f(x)$ is convex.

Fact

If $f''(x)$ exists and $f''(x) \geq 0$ then $f(x)$ is convex; if $f''(x) > 0$ then $f(x)$ is strictly convex. We can think of $\bar{x} = (1-\lambda)x_1 + \lambda x_2$ as an average of the values $x_1; x_2$ since $p_1 = (1-\lambda)$ and $p_2 = \lambda$, are positive numbers that sum to 1. Then the condition for convexity is

$$f(\bar{x}) \leq \overline{f(x)}$$

This leads to a probabilistic definition of convexity: $f(x)$ evaluated at the average input is $\leq$ the average of $f(x)$ on its inputs.

Jensen's Inequality

We have just established the following important result:

Theorem

Jensen's Inequality. For any function $f(x)$ that is convex on the range of a random variable X:

$$f(E[X]) \leq E[f(X)] \Leftrightarrow E[f(X)] \geq f(E[X])$$

If $f(x)$ is strictly convex, then equality holds if and only if X is a deterministic random variable, that is, $Pr(X = x_0) = 1$ for some x_0.

Jensen's Inequality

Example

Consider $f(x) = x^2$, which is convex over the entire set of real numbers. By Jensen's inequality, for any real-valued random variable X

$$E\left[X^2\right] \geq (E[X])^2$$

with equality only when $X \equiv x_0$. It is well known that the difference between these two quantities is the variance of X: $Var(X) = \sigma_X^2 = E\left[X^2\right] - (E[X])^2$

Acknowledgments and further reading

- This section of the lecture is based on the book by MacKay - i.e. Chapter 2.
- Several figures in this section have been borrowed from MacKay's book.