16.548
Coding, Information Theory (and Advanced Modulation)

Prof. Jay Weitzen

Ball 411

Jay_weitzen@uml.edu

Notes Coverage

- Course Introduction
- Definition of Information and Entropy
- Review of Conditional Probability

Class Coverage

- Fundamentals of Information Theory (4 weeks)
- Block Coding (3 weeks)
- Advanced Coding and modulation as a way of achieving the Shannon Capacity bound: Convolutional coding, trellis modulation, and turbo modulation, space time coding (7 weeks)

Course Web Site

- http://faculty.uml.edu/jweitzen/16.548
  - Class notes, assignments, other materials on web site
  - Please check at least twice per week
  - Lectures will be streamed, see course website

Prerequisites (What you need to know to thrive in this class)

- 16.363 or 16.584 (A Probability class)
- Some Programming (C, VB, Matlab)
- Digital Communication Theory

Grading Policy

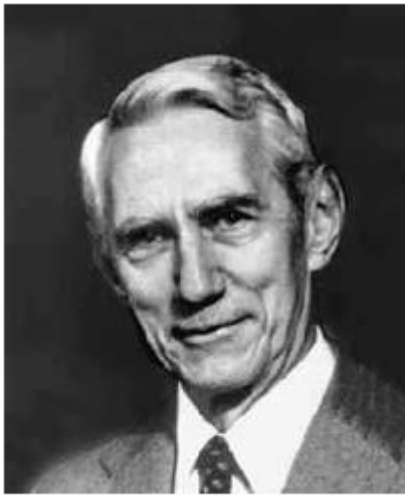
- 4 Mini-Projects (25% each project)
  - Lempel ziv compressor
  - Cyclic Redundancy Check
  - Convolutional Coder/Decoder soft decision
  - Trellis Modulator/Demodulator

Course Information and Text Books

- Coding and Information Theory by Wells, plus his notes from University of Idaho
- Digital Communication by Sklar, or Proakis Book
- Shannon's original Paper (1948)
- Other material on Web site

Claude Shannon Founds Science of Information theory in 1948

In his 1948 paper, ``A Mathematical Theory of Communication," Claude E. Shannon formulated the theory of data compression. Shannon established that there is a fundamental limit to lossless data compression. This limit, called the entropy rate, is denoted by *H*. The exact value of *H* depends on the information source --- more specifically, the statistical nature of the source. It is possible to compress the source, in a lossless manner, with compression rate close to *H*. It is mathematically impossible to do better than *H*.

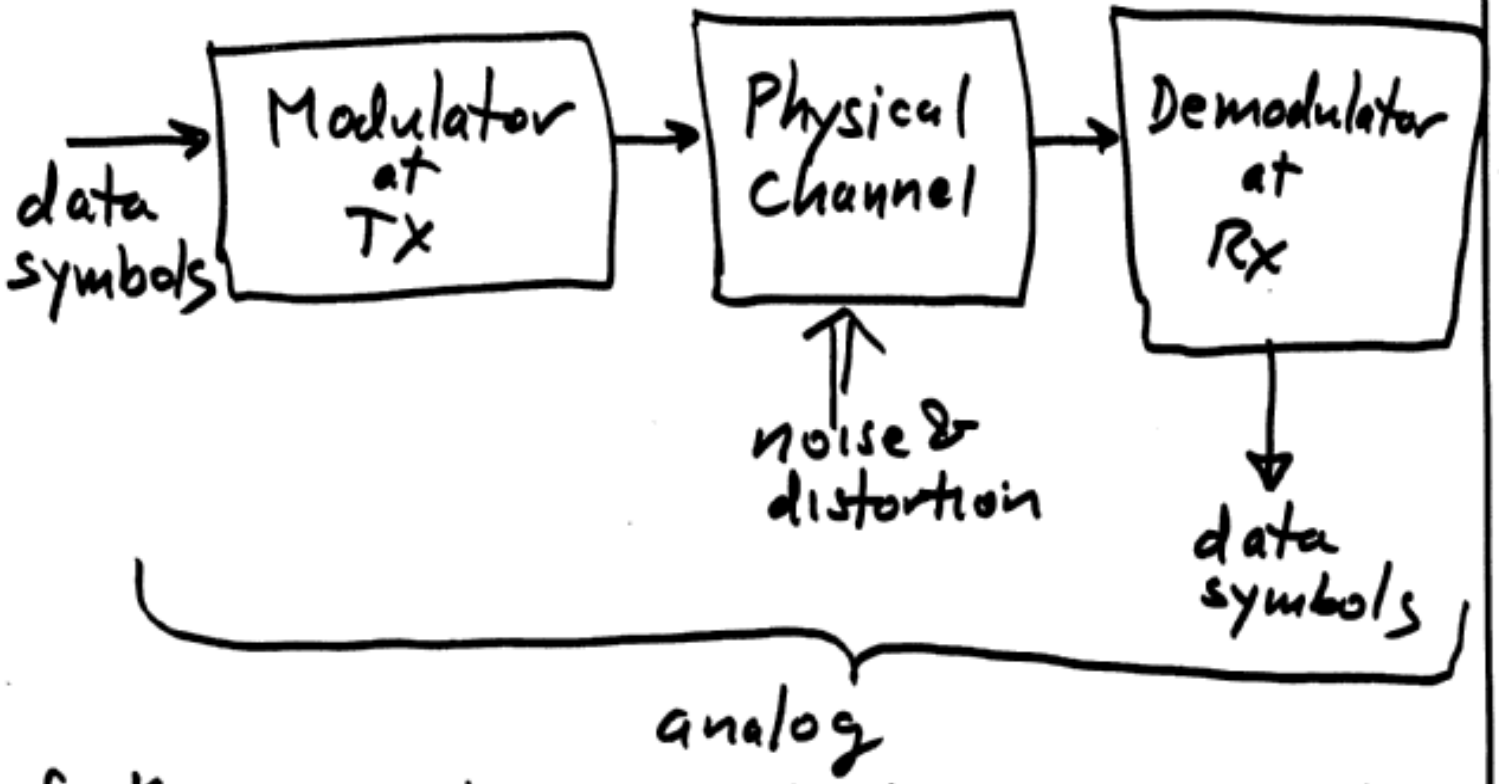


Claude E. Shannon

University of Idaho

(3)

The Info Channel



Info theory is where probability theory goes to work for a living

University of Idaho

4

Information ≠ Data

This is Important

data: how information is represented

info: "what" is represented in data

- tells us something that we did not already know and could not reliably predict.
- contains a certain element of surprise

Source Modeling

II. Source Modeling

Available at Amazon.com

Imagine that you go to the library. This library has a large selection of books --- say there are 100 million books in this library. Each book in this library is very thick --- say, for example, that each book has 100 million characters (or letters) in them. When you get to this library, you will, in some random manner, select a book to check out. This chosen book is the information source to be compressed. The compressed book is then stored on your zip disk to take home, or transmitted directly over the internet into your home, or whatever the case may be.

Mathematically, the book you select is denoted by

$$\mathcal{X} = (X_1, X_2, X_3, X_4, \ldots).$$

where $\mathcal{X}$ represents the whole book, X_1 represents the first character in the book, X_2 represents the second character, and so on. Even though in reality the length of the book is finite, mathematically we assume that it has infinite length. The reasoning is that the book is so long we can just imagine that it goes on forever. Furthermore, the mathematics turn out to be surprisingly simpler if we assume an infinite length book. To simplify things a little, let us assume that all the characters in all the books are either a lower-case letter (`a' through `z') or a SPACE (e. e. cummings style of writing shall we say). The source alphabet, $\mathcal{A}$, is defined to be the set of all 27 possible values of the characters:

$$\mathcal{A} = \{a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, SPACE\}.$$

Now put yourself in the shoes of the engineer who designs the compression algorithm. She does not know in advance which book you will select. All she knows is that you will be selecting a book from this library. From her perspective, the characters in the book $(X_i,\ i = 1, 2, \ldots)$ are random variables which take values on the alphabet $\mathcal{A}$. The whole book, $\mathcal{X}$ is just an infinite sequence of random variables -- that is $\mathcal{X}$ is a random process. There are several ways in which this engineer can model the statistical properties of the book.

Zero order models

A. Zero-Order Model: Each character is statistically independent of all other characters and the 27 possible values in the alphabet $\mathcal{A}$ are equally likely to occur. If this model is accurate, then a typical opening of a book would look like this (all of these examples came directly from Shannon's 1948 paper):

xfoml rxkhrjffjuj zlpwcfwkcyj ffjeyvkcqsghyd qpaamkbzaacibzlhjqd

This does not look like the writing of an intelligent being. In fact, it resembles the writing of a ``monkey sitting at a typewriter.''

It has been said, that if you get enough monkeys, and sit them down at enough typewriters, eventually they will complete the works of Shakespeare

First Order Model

B. First-Order Model: We know that in the English language some letters occur more frequently than others. For example, the letters `a' and `e' are more common than `q' and `z'. Thus, in this model, the character are still independent of one another, but the probability distribution of the characters are according to the first-order statistical distribution of English text. A typical text for this model looks like this:

ocroh hli rgwr nmielwis eu ll nbnesebya th eei alhenhttpa oobttva nah brl

Higher Order Models

C. Second-Order Model: The previous two models assumed statistical independence from one character to the next. This does not accurately reflect the nature of the English language. For exa#ple, some letters in thi# sent#nce are missi#g. However, we are still able to figure out what those letters should have been by looking at the context. This implies that there are some dependency between the characters. Naturally, characters which are in close proximity are more dependent than those that are far from each other. In this model, the present character X_i depends on the previous character X_{i-1} but it is conditionally independent of all previous characters $(X_1, X_2, \ldots, X_{i-2})$. According to this model, the probability distribution of the character X_i varies according to what the previous character X_{i-1} is. For example, the letter 'u' rarely occurs (probability=0.022). However, given that the previous character is 'q', the probability of a 'u' in the present character is much higher (probability=0.995). For a complete description, see the second-order statistical distribution of English text. A typical text for this model would look like this:

on ie antsoutinys are t inctore st be s deamy achin d ilonasive tucoowe at teasonare fuso tizin andy tobe seace ctisbe

D. Third-Order Model: This is an extension of the previous model. Here, the present character X_i depends on the previous two characters (X_{i-2}, X_{i-1}) but it is conditionally independent of all previous characters before those: $(X_1, X_2, \ldots, X_{i-3})$. In this model, the distribution of X_i varies according to what (X_{i-2}, X_{i-1}) are. See the third-order statistical distribution of English text. A typical text for this model would look like this:

in no ist lat whey cratict froure birs grocid pondenome of demonstures of the reptagin is regoactiona of cre

The resemblance to ordinary English text increases quite noticeably at each of the above steps.

E. General Model: In this model, the book $\mathcal{X}$ is an arbitrary stationary random process. The statistical properties of this model are too complex to be deemed practical. This model is interesting only from a theoretical point of view.

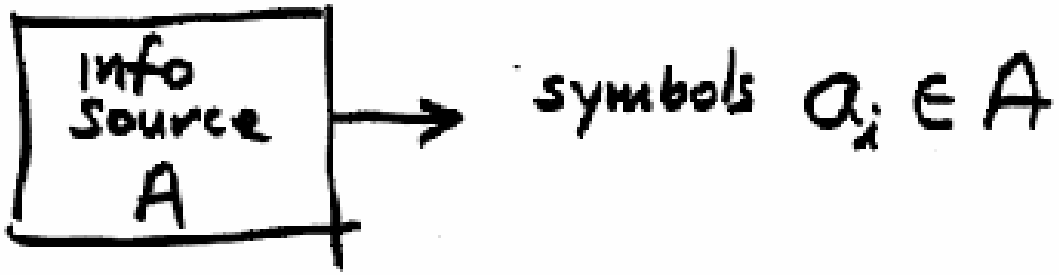
University of Idaho

8

Shannon Theory

- Claude Shannon 1948

Start at the source



symbol alphabet $A = \{a_0, a_1, \ldots, a_{M-1}\}$

discrete source

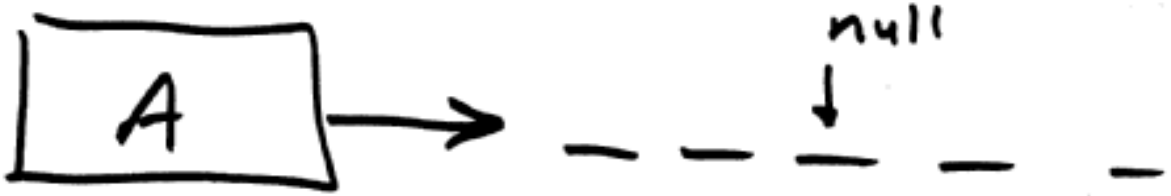
each symbol a_i has some probability $Pr[a_i] = p_i$

University of Idaho

9

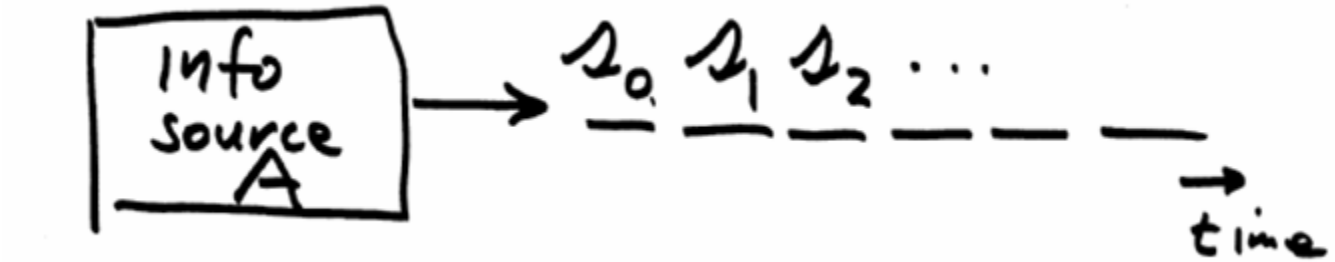
let $P_A = \{p_0, p_1, \ldots, p_{M-1}\}$

where $p_i = Pr[a_i]$



synchronous source: a new symbol is output every T seconds

asynchronous source: sometimes the async. source doesn't output anything; let one $a_0 \in A \triangleq$ "null"



$A = \{a_0, a_1, \ldots, a_{M-1}\}$

Discrete source

Λ ⇒ symbol

Λ_t ⇒ t = time index

$\Lambda_t \in A$

async. discrete source, "null" $\in A$

each $a_i \in A$ has a probability p_i that the info source will emit a_i at any given time t.

$$P_A = \{ p_0, p_1, \cdots, p_{M-1} \}$$

& $p_i = Pr[a_i]$

Prob. that $s_t = a_i$

$$\sum_{i=0}^{M-1} p_i = 1$$

Convenient notation

$$p_i \Longleftrightarrow a_i$$

$$\sum_{i=0}^{M-1} p_i \equiv \sum_{\forall a_i \in A} p_i = 1$$

$\forall$ = "all"

notation: given $A = \{a_0, a_1, \ldots, a_{M-1}\}$

"cardinality of A" $\triangleq |A| = M$

University of Idaho

4

time →

$\rightarrow s_0 s_1 \cdots s_t \cdots$

$s_i \in A \qquad a_i, a_j \in A$

joint events:

"event" is anything that happens

$Pr[s_0 = a_i, s_1 = a_j]$ = joint probability

$\triangleq p_{i,j}$

University of Idaho

5

suppose

$$p_{i,j} = p_i \cdot p_j$$

for every pair i, j

$\therefore \forall\, a_i, a_j \in A$

further suppose this is true for all time indexes t

Then A is called a stationary discrete memoryless source (DMS)

Zero'th Order Model

(6)

data

$x_0\ x_1 \cdots x_t$

$x_t \in A$

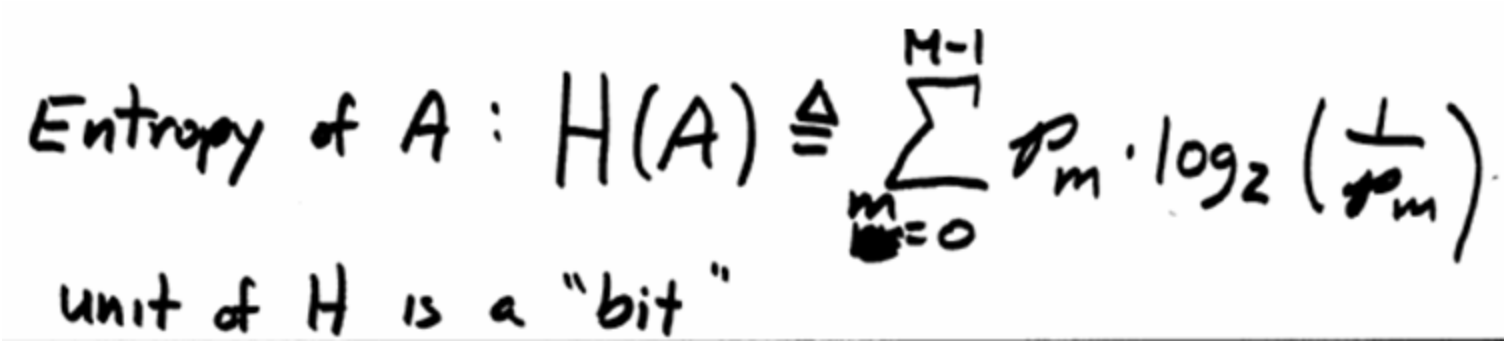
$A = \{a_0, a_1, \ldots, a_{M-1}\}$

$|A| = M$

What is the average amount of information carried in each symbol x_t?

Definition of Entropy

Shannon used the ideas of randomness and entropy from the study of thermodynamics to estimate the randomness (e.g. information content or entropy) of a process



Entropy is a measure of predictability or randomness

Entropy in a nut-shell

Low Entropy

..the values (locations of soup) sampled entirely from within the soup bowl

High Entropy

..the values (locations of soup) unpredictable… almost uniformly sampled throughout our dining room

University of Idaho

9

back to entropy:

$$H(A) = \sum_{\forall a_m \in A} p_m \cdot \log_2\left(\frac{1}{p_m}\right) \quad \text{bits}$$

example:

$A = \{a\} \qquad Pr[a] = 1$

$$H(A) = Pr[a] \cdot \log_2\left[\frac{1}{Pr[a]}\right]$$

$$= 1 \cdot \log_2(1) = 0$$

Absolute certainty implies 0 information

University of Idaho

10

Example

$A = \{0, 1\}$ $\quad p_0 = \frac{1}{2} \Rightarrow p_1 = \frac{1}{2}$

$$H(A) = \frac{1}{2}\log_2\left(\frac{1}{1/2}\right) + \frac{1}{2}\log_2\left(\frac{1}{1/2}\right)$$

$$= 2 \cdot \frac{1}{2} \cdot \log_2(2) = 1 \text{ bit}$$

$$\log_2(2) = 1$$

Randomness has high information content

Quick Review: Working with Logarithms

unit conversions

$$\log_2(x) = \frac{\ln(x)}{\ln(2)} = \frac{\log_{10}(x)}{\log_{10}(2)}$$

$$\log(x \cdot y) = \log(x) + \log(y)$$

$$\log(x^y) = y \cdot \log(x)$$

$$\log(1) = 0 \qquad \log(0) = -\infty$$

University of Idaho

8

$$\log\left(\frac{x}{y}\right) = \log\left(x \cdot y^{-1}\right)$$

$$= \log(x) + \log\left(y^{-1}\right)$$

$$= \log(x) - \log(y)$$

$$\lim_{x \to 0} x \cdot \log(x) \to 0 \implies$$

$$\lim_{x \to 0} x \cdot \log\left(\frac{1}{x}\right) \to 0$$

$$\frac{1}{x} = x^{-1}$$

example

$A = \{ \text{start}, 0, 1, \text{stop} \}$

$|A| = 4$

let $P_A = \{0.05, 0.45, 0.45, 0.05\}$

$$H(A) = 0.05 \log_2\left(\frac{1}{.05}\right) + .45 \log_2\left(\frac{1}{.45}\right) + 0.45 \log_2\left(\frac{1}{.45}\right) + .05 \log_2\left(\frac{1}{.05}\right)$$

$$= 1.468995$$

Entropy of English Alphabet

B. First-Order Model: The characters are statistically independent. Let m be the size of the alphabet and let p_i be the probability of the i-th letter in the alphabet. The entropy rate is

$$H = -\sum_{i=1}^{m} p_i \log_2 p_i \quad \text{bits/character.}$$

Using the first-order distribution, the entropy rate of English text would have been 4.07 bits/character had this been the correct model.

University of Idaho

Discrete Memoryless Source

$0 \leq H(A) \leq ?$

$A = \{a_0, a_1, \ldots, a_{M-1}\} \quad |A| = M$

$P_A = \{p_0, p_1, \ldots, p_{M-1}\} \quad Pr[a_i, a_j] = p_{i,j}$

$= p_i \cdot p_j$

$$H(A) = \sum_{a_i \in A} p_i \log_2\left(\frac{1}{p_i}\right)$$

What P_A maximizes $H(A)$?

maximizing $H(A)$ implies:

$$\frac{\partial H(A)}{\partial p_i} = 0 \quad ; \quad \frac{\partial^2 H(A)}{\partial p_i^2} < 0$$

subject to a constraint

$$\sum_{a_i \in A} p_i = 1$$

Handy Things to know:

$$f(x) = x \cdot \ln(x) \qquad \log_2(x) = \frac{\ln(x)}{\ln(2)}$$

$$\frac{df}{dx} = \frac{d}{dx}\left[x \ln(x)\right] = \ln(x) + x \cdot \frac{1}{x}$$

$$= \ln(x) + 1$$

$$H(A) = \sum_{q_i \in A} p_i \log_2\left(\frac{1}{p_i}\right)$$

$$= \frac{1}{\ln(2)} \sum_{q_i \in A} p_i \ln\left(\frac{1}{p_i}\right)$$

University of Idaho

④

$$H(A) = \frac{-1}{\ln(2)} \sum_{q_i \in A} p_i \ln(p_i)$$

if $i \neq j$

$$\frac{\partial p_j}{\partial p_i} = 0 \qquad \frac{\partial p_i}{\partial p_i} = 1$$

$$\frac{\partial H(A)}{\partial p_i} = \frac{-1}{\ln(2)} \left[\ln(p_i) + 1 \right]$$

University of Idaho

(5)

What about constraint $\sum_{q_i \in A} p_i = 1$?

Handy trick

$H(A) + 0 = H(A)$

let λ be any constant

Then $\lambda\left(\sum_{q_i \in A} p_i - 1\right) \equiv 0$

Lagrange multipliers

$$H(A) = \frac{-1}{\ln(2)} \sum_{q_i \in A} p_i \ln(p_i) + \lambda\left(\sum_{q_i \in A} p_i - 1\right)$$

University of Idaho

(6)

$$\frac{\partial H(A)}{\partial p_i} \left(\ln(p_i) + 1 \right) + \lambda = 0$$

Kind of Intuitive, but hard to prove

$\Rightarrow \ln(p_i) = \lambda - (2) - 1 = \text{constant}$

independent of which a_i we're talking

$\therefore$ to maximize $H(A)$

$$p_0 = p_1 = p_2 = \cdots = p_{M-1} \Rightarrow p_i = \frac{1}{|A|} = \frac{1}{M}$$

University of Idaho

7

$$\therefore \max H(A) = \sum_{a_i \in A} \frac{1}{M} \log_2\left(\frac{1}{1/M}\right)$$

$$|A| = M$$

$$= M \cdot \frac{1}{M} \cdot \log_2(M) = \log_2 M$$

- any DMS with $|A| = M$

$$0 \leq H(A) \leq \log_2(M)$$

Information content is bounded by certainty (0) and uncertainty

University of Idaho

8

Discussion on set theory

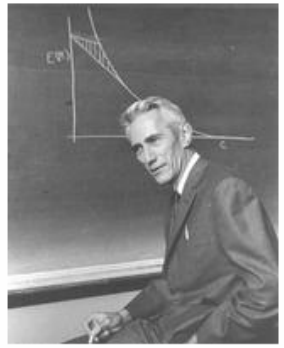
- When things are very complicated or very nonlinear, then regular math (i.e. calculus, algebra, etc.) can be hard to use.
- modern math is <u>founded</u> on set theory
  ∴ if we can not say it in set theory, we cannot say it in math at all.

Bounds on Entropy

The entropy rate of a source is a number which depends only on the statistical nature of the source. If the source has a simple model, then this number can be easily calculated. Here, we consider an arbitrary source:

$$\mathcal{X} = (X_1, X_2, X_3, X_4, \ldots),$$

while paying special attention to the case where $\mathcal{X}$ is English text.



Shannon in 1948

A. Zero-Order Model: The characters are statistically independent of each other and every letter of the alphabet, $\mathcal{A}$, are equally likely to occur. Let m be the size of the alphabet. In this case, the entropy rate is given by

$$H = \log_2 m \quad \text{bits/character}$$

For English text, the alphabet size is $m=27$. Thus, if this had been an accurate model for English text, then the entropy rate would have been $H=\log_2 27=4.75$ bits/character.

B. First-Order Model: The characters are statistically independent. Let m be the size of the alphabet and let p_i be the probability of the i-th letter in the alphabet. The entropy rate is

$$H = -\sum_{i=1}^{m} p_i \log_2 p_i \quad \text{bits/character.}$$

Using the first-order distribution, the entropy rate of English text would have been 4.07 bits/character had this been the correct model.

Math 495 Micro-Teaching

Quick Review: JOINT DENSITY OF RANDOM VARIABLES

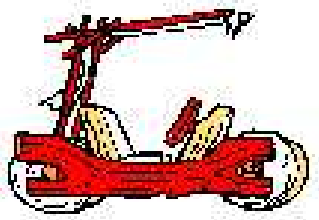
David Sherman

Bedrock, USA

In this presentation, we'll discuss the *joint density* of two random variables. This is a mathematical tool for representing the interdependence of two events.

First, we need some random variables.

Lots of those in Bedrock.



Let X be the number of days Fred Flintstone is late to work in a given week. Then X is a random variable; here is its density function:

N	1	2	3
F(N)	.5	.3	.2

Amazingly, another resident of Bedrock is late with exactly the same distribution. It's...

Fred's boss, Mr. Slate!

N	1	2	3
F(N)	.5	.3	.2

Remember this means that P(X=3) = .2.

Let Y be the number of days when Slate is late. Suppose we want to record BOTH X and Y for a given week. How likely are different pairs?

We're talking about the *joint density* of X and Y, and we record this information as a function of two variables, like this:

	1	2	3
1	.35	.1	.05
2	.15	.1	.05
3	0	.1	.1

This means that

P(X=3 and Y=2) = .05.

We label it f(3,2).

N	1	2	3
F(N)	.5	.3	.2

	1	2	3
1	.35	.1	.05
2	.15	.1	.05
3	0	.1	.1
	.2		

The first observation to make is that this joint probability function contains all the information from the density functions for X and Y (which are the same here).

For example, to recover P(X=3), we can add f(3,1)+f(3,2)+f(3,3).

The individual probability functions recovered in this way are called *marginal*.

Another observation here is that Slate is never late three days in a week when Fred is only late once.

N	1	2	3
F(N)	.5	.3	.2

Since he rides to work with Fred (at least until the directing career works out), Barney Rubble is late to work with the same probability function too. What do you think the joint probability function for Fred and Barney looks like?

	1	2	3
1	.5	0	0
2	0	.3	0
3	0	0	.2

It's diagonal!

This should make sense, since in any week Fred and Barney are late the same number of days.

This is, in some sense, a maximum amount of interaction: if you know one, you know the other. P(Barney late |Fred late)= 1

N	1	2	3
F(N)	.5	.3	.2

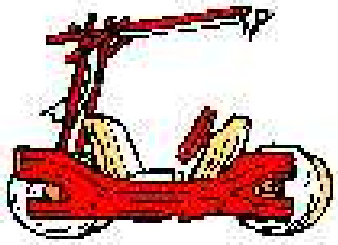
A little-known fact: there is actually another famous person who is late to work like this.

SPOCK!

(Pretty embarrassing for a Vulcan.)

Before you try to guess what the joint density function for Fred and Spock is, remember that Spock lives millions of miles (and years) from Fred, so we wouldn't expect these variables to influence each other at all.

In fact, they're independent….



N	1	2	3
F(N)	.5	.3	.2

	1	2	3
1	.25	.15	.1
2	.15	.09	.06
3	.1	.06	.04

Since we know the variables X and Z (for Spock) are independent, we can calculate each of the joint probabilities by multiplying.

For example, f(2,3) = P(X=2 and Z=3)

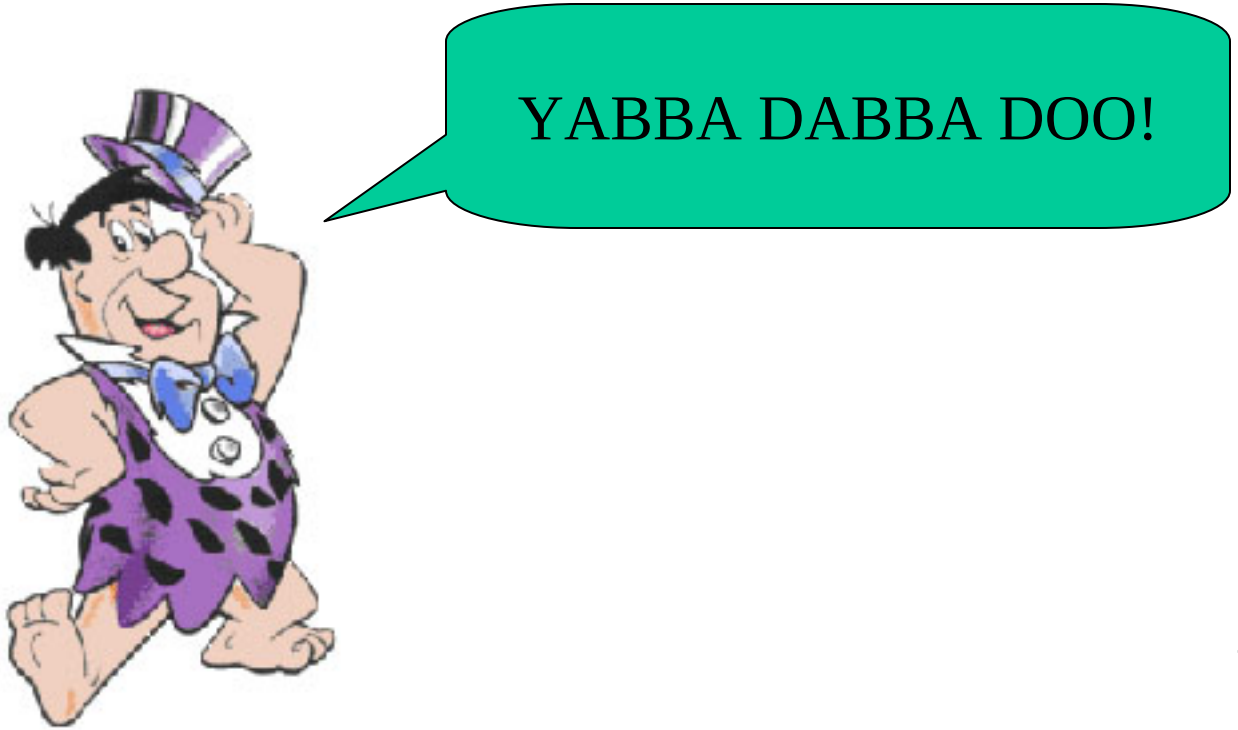
= P(X=2)P(Z=3) = (.3)(.2) = .06.

This represents a minimal amount of interaction. P(spock|fred)=P(spock)

Dependence of two events means that knowledge of one gives information about the other.

Now we've seen that the joint density of two variables is able to reveal that two events are independent (and), completely dependent (and), or somewhere in the middle (and).

Later in the course we will learn ways to quantify dependence. Stay tuned….



University of Idaho

10

union, intersect, difference

$A \cup B$ $A \cap B$ $A - B$ $B - A$

ex.

$A = \{0, 1, 2, 3, 4\}$

$B = \{0, 2, 4, 6, 8\}$

$A \cup B = \{0, 1, 2, 3, 4, 6, 8\}$

$A \cap B = \{0, 2, 4\}$

Set difference

$$A - B = \{x \in A \mid x \notin B\}$$

$$B - A = \{x \in B \mid x \notin A\}$$

$$A = \{0, 1, 2, 3, 4\}$$

$$B = \{0, 2, 4, 6, 8\}$$

$$A - B = \{1, 3\} \; ; \; B - A = \{6, 8\}$$

Compound symbol alphabet

$$C \triangleq \left\{ c_{i,j} \,\middle|\, c_{i,j} = \langle a_i, b_j \rangle, \; \begin{array}{l} \forall\, a_i \in A \\ \forall\, b_j \in B \end{array} \right\}$$

$$Pr[c_{i,j}] = Pr[a_i, b_j]$$

Example: $A = \{a_0, a_1, a_2\}$; $B = \{b_0, b_1\}$

	b_0	b_1
a_0	$c_{0,0}$	$c_{0,1}$
a_1	$c_{1,0}$	$c_{1,1}$
a_2	$c_{2,0}$	$c_{2,1}$

$c_{i,j}$

	b_0	b_1
a_0	$p_{0,0}$	$p_{0,1}$
a_1	$p_{1,0}$	$p_{1,1}$
a_2	$p_{2,0}$	$p_{2,1}$

← P_C

$p_{i,j}$

$p_{i,j} = Pr[a_i, b_j]$ = Prob. of symbol $C_{i,j}$

$$\sum_{C_{i,j} \in C} p_{i,j} = 1 = \sum_{a_i \in A} \sum_{b_j \in B} p_{i,j}$$

	b_0	b_1	
a_0	$p_{0,0}$	$p_{0,1}$	$\Sigma \rightarrow Pr[a_0]$
a_1	$p_{1,0}$	$p_{1,1}$	$\rightarrow Pr[a_1]$
a_2	$p_{2,0}$	$p_{2,1}$	
$Pr[b_0]$			

University of Idaho

15

algebra-speak

$$\sum_{b_j \in B} Pr[a_i, b_j] = Pr[a_i]$$

$$\sum_{a_i \in A} Pr[a_i, b_j] = Pr[b_j]$$

Marginal Density Functions

Conditional Probability

- In many cases, we have only partial knowledge of the outcome of an event
- Conditional probability is the situation in which the probability of one event is influenced by that of another event
- It is written as

  P[A/B] = "probability of event A given event B"
- **Definition:** The conditional probability of an event A with respect to event B is given by

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

Conditional Probability (cont'd)

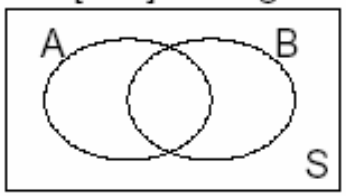
- The conditional probability of B given A is

  $P(B|A)$

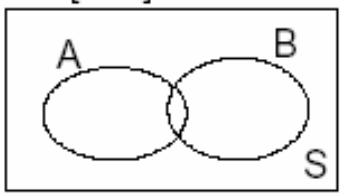
- **Note:** P[A] is often called the **a priori** probability

  P[A/B] is often called the **a posteriori** probability

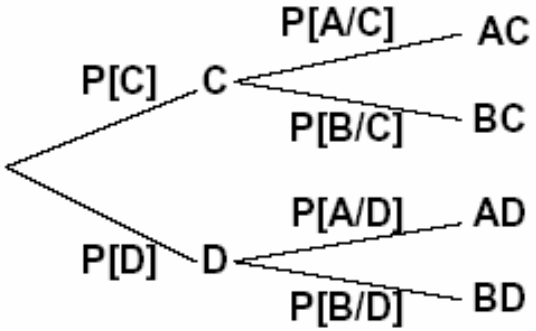
P[A/B] is large



P[A/B] is small



- The idea of conditional probability can often be drawn out in the form of a tree diagram (probability tree)



Definition of conditional probability

- If P(B) is not equal to zero, then the conditional probability of A relative to B, namely, the probability of A given B, is

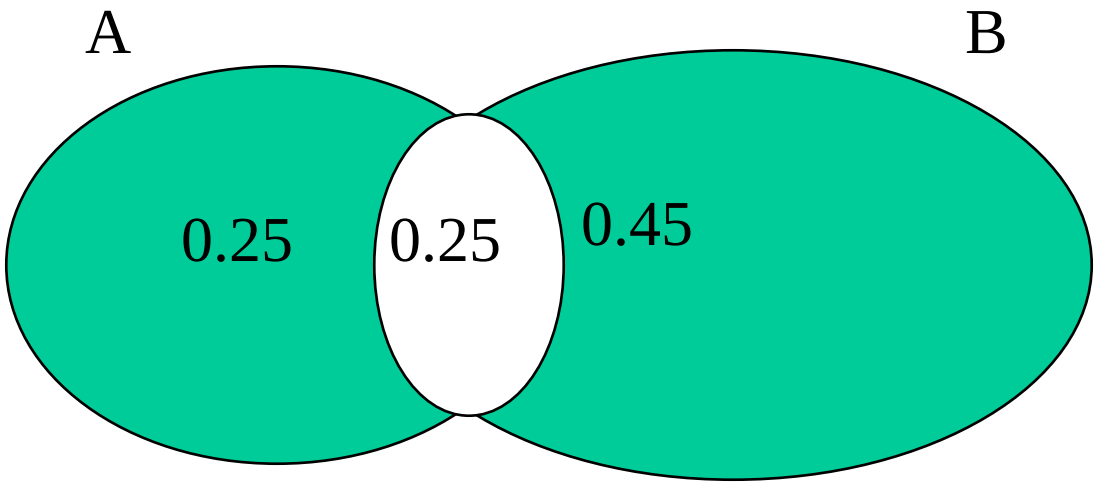
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(B) \bullet P(A|B)$$

or

$$P(A \cap B) = P(A) \bullet P(B|A)$$

Conditional Probability



P(A) = 0.25 + 0.25 = 0.50
P(B) = 0.45 + 0.25 = 0.70
P(A') = 1 - 0.50 =0.50
P(B')= 1-0.70 =0.30

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.25}{0.70} = 0.357$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.25}{0.50} = 0.5$$

Some Observations:

- In previous Slides P(Fred late and Spock late) were independent
  - Therefore P(Fred|Spock) =P(Fred)P(spock)/P(spock)=P(Fred)
- P(Fred late and Barney late) are totally dependent
  - P(Fred|Barney)=1

Law of Total Probability

If $B_1, B_2, \ldots\ldots,$ and B_k are mutually exclusive events of which one must occur, then for any event A

$$P(A) = P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2) + \ldots\ldots + P(B_k) \cdot P(A|B_k)$$

Special case of rule of Total Probability

$$P(A) = P(B) \cdot P(A|B) + P(B') \cdot P(A|B')$$

Bayes Theorem

Useful Probability Relationships

$$P(A+B) = P(A) + P(B) - P(AB)$$

$$P(A|B) = \frac{P(AB)}{P(B)}$$

- This is known as **Conditional Probability**. It simply gives us the **probability that event A occurs GIVEN that B occurred**. Likewise, we have

$$P(B|A) = \frac{P(AB)}{P(A)} \left(= \frac{P(A|B)P(B)}{P(A)} \right)$$

- This is known as "**Bayes Rule**".
- Finally, A and B are **statistically independent** iff

$$P(AB) = P(A)P(B)$$

Useful Properties

- **Chain Rule for Probabilities**:

$$P(A_1 \cap A_2 \ldots \cap A_n) = P(A_1 \mid A_2 \ldots A_n) P(A_2 \mid A_3 \ldots A_n) \ldots P(A_{n-1} \mid A_n) P(A_n)$$

- Also if an event $B = \bigcup_{j=1}^{m} A_j$ with A_j being disjoint (I.e., the events A_j form a **partition** of B). Then

$$P(B) = \sum_{j=1}^{m} P(B \cap A_j) = \sum_{j=1}^{m} P(B \mid A_j) P(A_j)$$

Law of total Probability

Generalized Bayes' theorem

If $B_1, B_2, \ldots$ and B_k are mutually exclusive events of which one must occur, then

$$P(B_i / A) = \frac{P(B_i) \cdot P(A / B_i)}{P(B_1) \cdot P(A / B_1) + P(B_2) \cdot P(A / B_2) + \ldots\ldots + P(B_k) \cdot P(A / B_k)}$$

for $i = 1, 2, \ldots\ldots, k.$

Urn Problems

- Applications of Bayes Theorem
- Begin to think about concepts of Maximum likelihood and MAP detections, which we will use throughout codind theory

University of Idaho

16

Handy Things to Know: Bayes rule

$$Pr[a,b] = Pr[b|a] \cdot Pr[a]$$

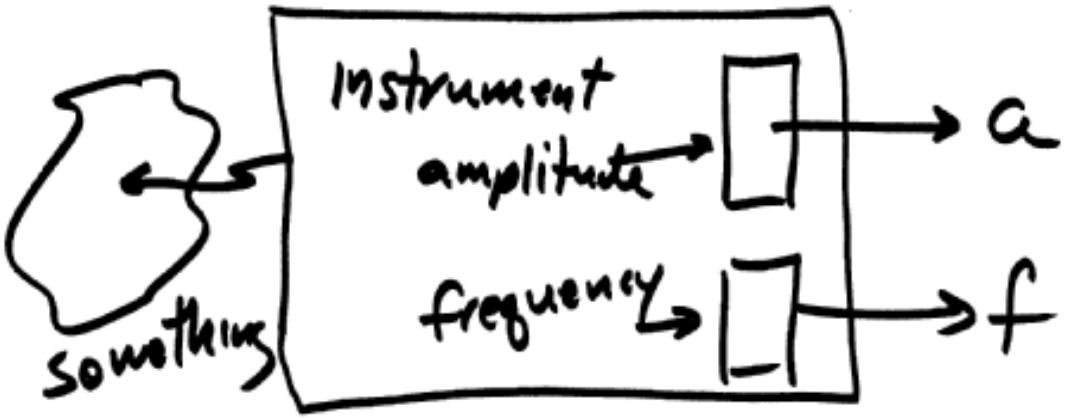
$$= Pr[a|b] \cdot Pr[b]$$

$$\sum_{b \in B} Pr[a_i,b] = \sum_{b \in B} Pr[b|a_i] \cdot Pr[a_i]$$

$$= Pr[a_i] \Rightarrow \sum_{b \in B} Pr[b|a] = 1$$

University of Idaho

9



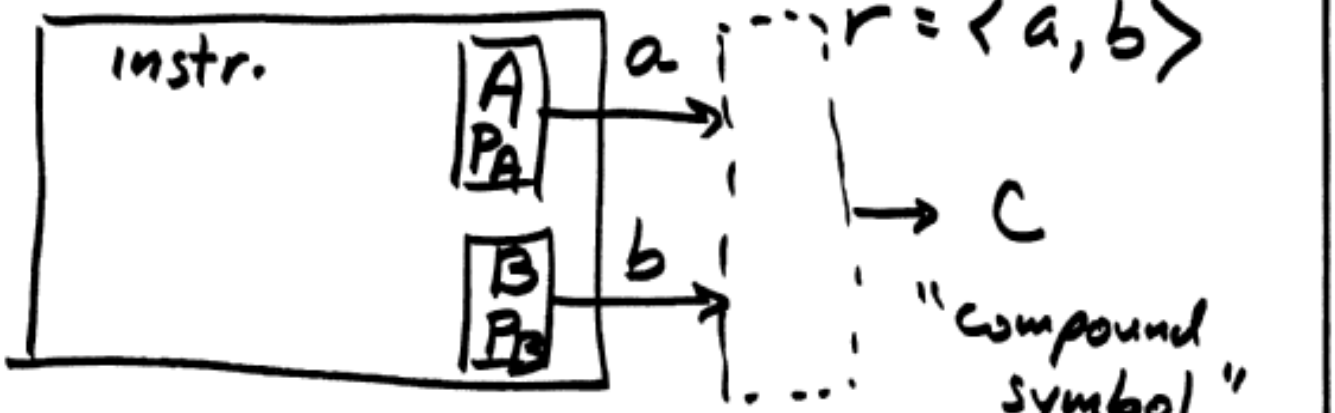
$r = \langle a, f \rangle$ ordered pair *

$r_1 = \langle a_1, f_1 \rangle$ $r_3 = \langle a_1, f_3 \rangle$

$r_2 = \langle a_2, f_2 \rangle$

University of Idaho

12



$\langle a, b \rangle \neq \langle b, a \rangle$

$C_{i,j} = \langle a_i, b_j \rangle$

$a_i \in A \; ; |A| = M_A$

$b_j \in B \; ; |B| = M_B$

End of Notes 1