



# An Introduction to Information Theory

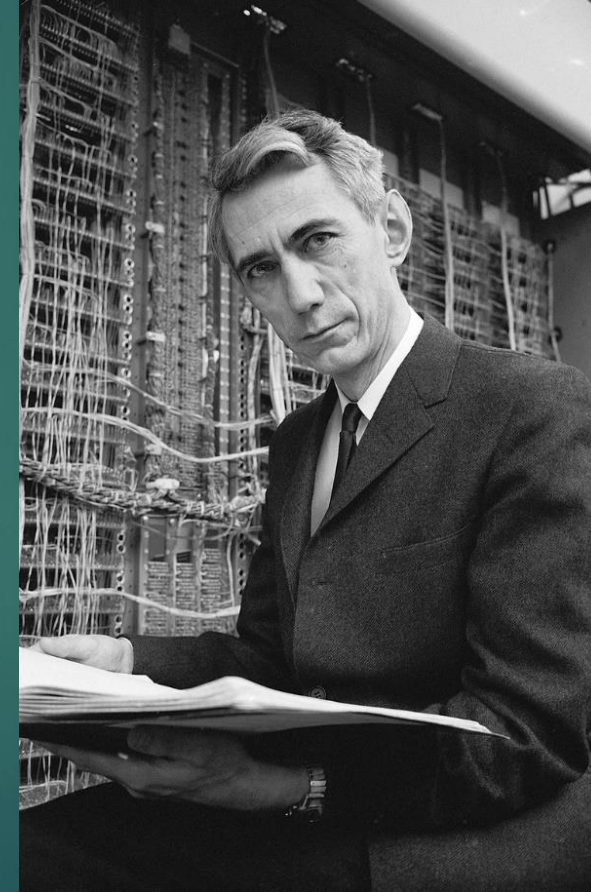
FARZAD FARNOUD

DATA SCIENCE INSTITUTE

3/25/2019

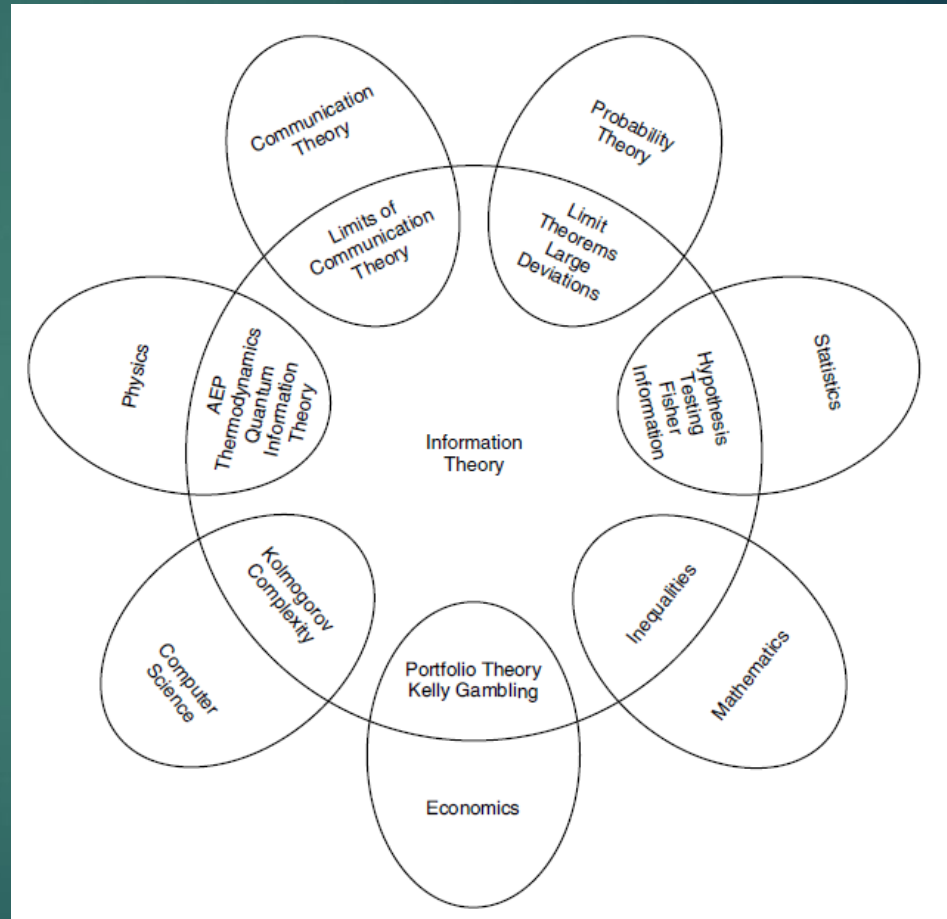
# Information Theory

- ▶ Developed by Claude Shannon, motivated by problems in communications
  - ▶ A Mathematical Theory of Communication," *The Bell System Tech J*, 1948. Cited  $\geq 100,000$  times
- ▶ Provides a way to **quantify information** suitable for engineering applications
- ▶ Relies on **probability, stochastic processes**
- ▶ Applications in **communications, data storage, statistics, machine learning**



# Information Theory

- ▶ Provides a way to quantify **information** independent of representation
- ▶ Quantifies **mutual information**, the amount of information one signal has about another
- ▶ Limits on the **shortest representation** of information without losing accuracy
- ▶ Trade-off between **accuracy and representation length**
- ▶ Limits on the amount of information that can be **communicated**



Beyond communication and data storage  
(Elements of Inf Theory, Cover and Thomas)

# Quantifying information

- ▶ Which statement carries more information?
  - ▶ Tomorrow, the sun will rise in the east.
    - ▶  $P = 1$  no information transferred.
  - ▶ Tomorrow, it will rain in Seattle.
    - ▶  $P = 158/365 = .43$ , rather likely, could guess either way
  - ▶ Tomorrow, it will rain in Phoenix.
    - ▶  $P = 36/365 = .1$ , rather unlikely, significant info
  - ▶ Tomorrow, Betsy DeVos will call you and explain the central limit theorem.
    - ▶  $P = 0$  – this would be a major story!
- ▶ Conclusion: Mathematical definition of information content is tied to (only) probability

# Properties of an information measure

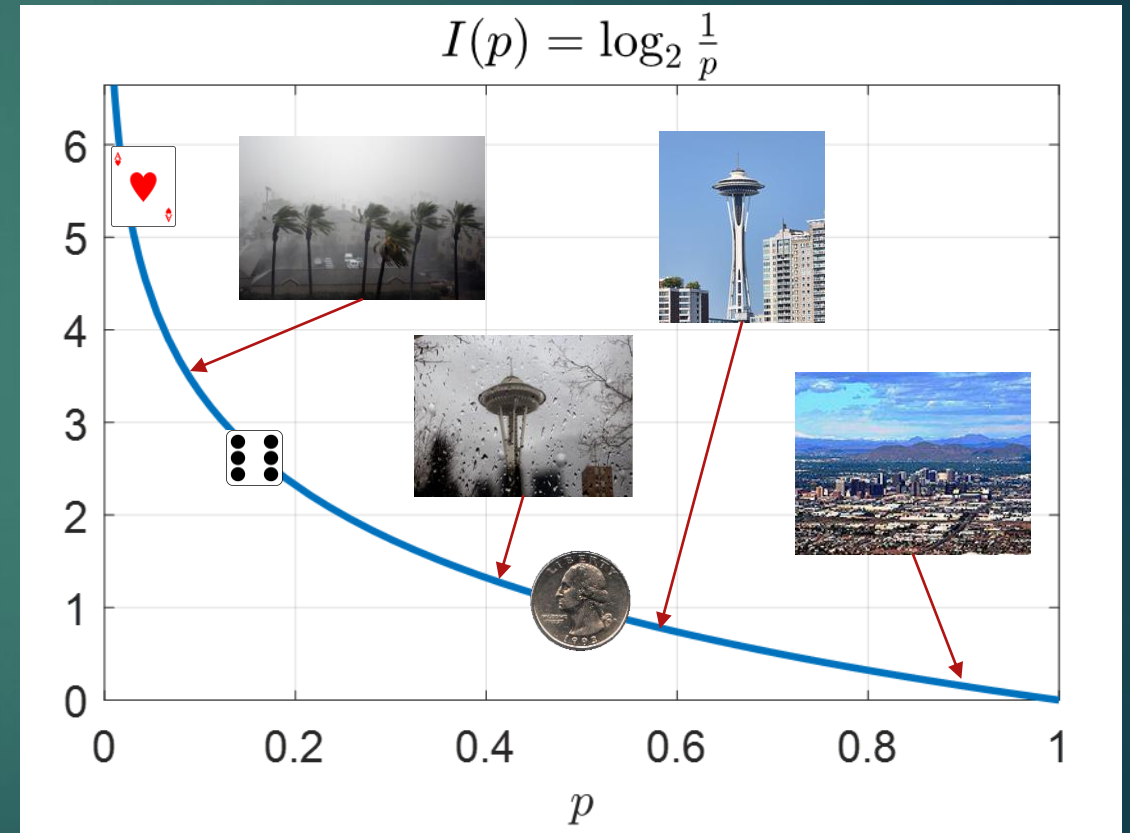
- ▶  $I(x)$ : the information in statement  $x$
- ▶ Desired properties:
  - ▶  $I(x) \geq 0$
  - ▶ Decreasing function of probability
  - ▶ If  $p(x) \rightarrow 0$  then  $I(x) \rightarrow \infty$
  - ▶ If  $x$  and  $y$  are results of independent events, then  $I(x \text{ and } y) = I(x) + I(y)$ 
    - ▶  $\Pr(\text{Virginia beats Florida State} \ \& \ \text{Duke beats UNC}) = \Pr(\text{Virginia beats Florida State}) \times \Pr(\text{Duke beats UNC})$
    - ▶  $I(\text{Virginia beats Florida State} \ \& \ \text{Duke beats UNC}) = I(\text{Virginia beats Florida State}) + I(\text{Duke beats UNC})$

# Self-information

- ▶ There is a unique function satisfying these conditions

$$I(x) = \log \frac{1}{p(x)}$$




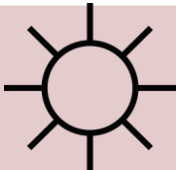
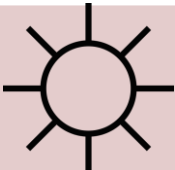


- ▶ The base of the log is arbitrary and determines the unit
- ▶ Base 2 gives the information in bits (term coined by Shannon)



# Independence from representation

- ▶ Our measure of information does not depend on representation

Mar. 24	25	26	27	28	29	30
Cloudy	Rainy	Cloudy	Sunny	Sunny	Cloudy	Rainy

Mar. 24	25	26	27	28	29	30
						

- ▶ Both tables carry the same (amount of) information

# Entropy: average information

- ▶ Information is defined in the context of a random event with uncertain outcomes
- ▶ A property of random variables and random processes

- ▶ The entropy of a random variable  $X$  is

$$H(X) = E[I(X)] = E \left[ \log \frac{1}{p(X)} \right] = \sum p(x) \log \frac{1}{p(x)}$$

- ▶ Entropy: the amount of information generated by a source, on average.



# Entropy: average information

- ▶ Entropy of rolling a die:

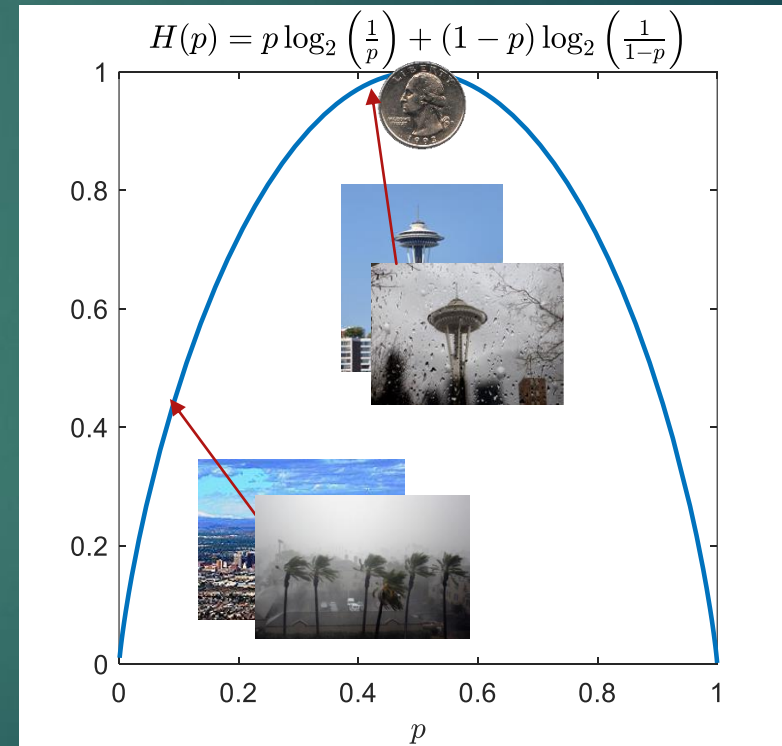
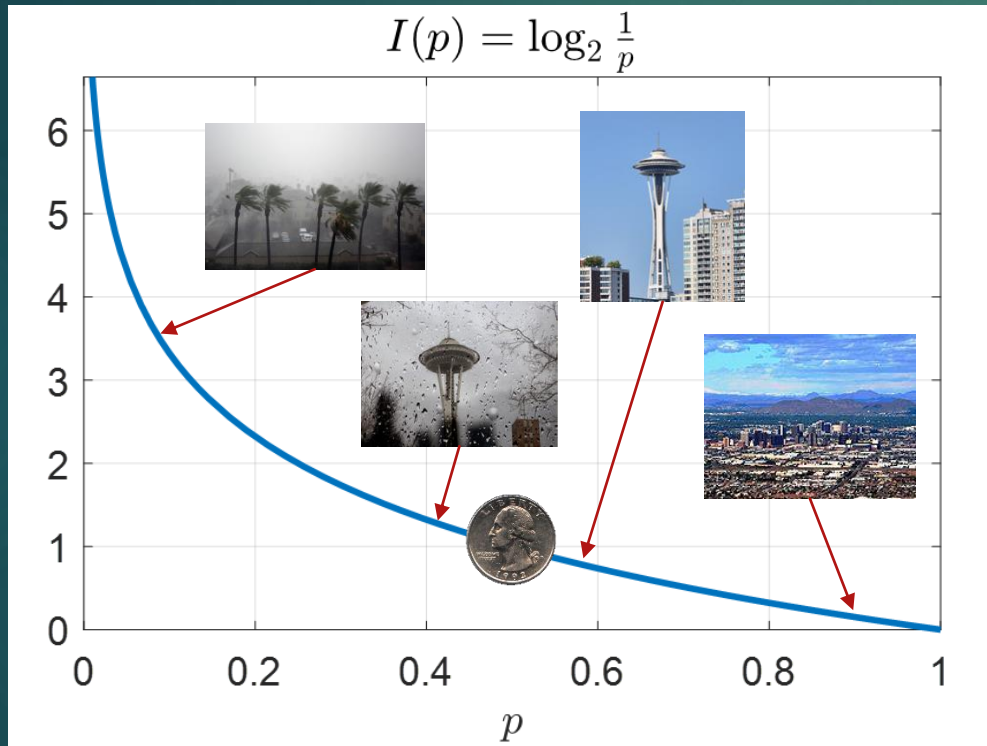
$$\sum_{i=1}^6 p(i) \log \frac{1}{p(i)} = 6 \times \frac{1}{6} \log \frac{1}{1/6} = \log 6 = 2.58 \text{ bits}$$

- ▶ Entropy is a measure of uncertainty/predictability
- ▶ Entropy is non-negative (since self-information is non-negative)
- ▶ For a random variable  $X$  that takes  $M$  values,  
$$H(X) \leq \log M$$

# Binary Entropy

- ▶ Experiment with two outcomes with probabilities  $p$  and  $1 - p$

$$H(p) = p \log_2 \frac{1}{p} + (1 - p) \log_2 \frac{1}{1 - p}$$



- ▶ **Predictability:** Weather in Phoenix is more predictable than Seattle

# Why “Entropy”?

- ▶ *My greatest concern was what to call it. I thought of calling it ‘information,’ but the word was overly used, so I decided to call it ‘uncertainty.’ When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, ‘You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage.’*

Claude Shannon, *Scientific American* (1971), volume 225, page 180.

# Data representation

- ▶ We store data as a sequence of bits using a **code**
  - ▶ ASCII for representing English text
    - ▶  $A \rightarrow 01000001, B \rightarrow 01000010, \dots$
  - ▶ Bitmap for images
  - ▶ Storing a genome:
    - ▶  $A \rightarrow 00, G \rightarrow 01, C \rightarrow 10, T \rightarrow 11$
- ▶ The average number of bits per symbol is the **average code length**
- ▶ For a random variable that can take  $M$  values, need  $\leq \lceil \log M \rceil$  bits
- ▶ The entropy is also bounded by  $\log M$

# Data compression

- ▶ Can we do better than  $\log M$ , without losing information?
- ▶ Which is easier to store?
  - ▶ Weather in Phoenix: `RSSSSSRSSSSSSSSSSSSSSSSSSSSSSSSSSSRSSSS...`
  - ▶ Weather in Seattle: `RSRSSRRSRSRSRSSSSSRSSRSRRRRSSR...`
- ▶ Rothko vs Pollock



# Data compression

- ▶ What is the average length of the shortest representation of a random variable (source of information)?
- ▶ Example: A genome with non-uniform symbol probabilities:

	<b>A</b>	<b>C</b>	<b>G</b>	<b>T</b>
<b>Probability</b>	1/2	1/4	1/8	1/8
<b>Code</b>	00	01	10	11

- ▶ The average code length is 2 bits/symbol

# Data compression

- ▶ What if we choose representation with length equal to self-information,  $\log 1/p_i$ ?

	A	C	G	T
Probability	1/2	1/4	1/8	1/8
Code	0	10	110	111
Information	$\log 2 = 1$	$\log 4 = 2$	$\log 8 = 3$	$\log 8 = 3$

- ▶ Average code length:

$$\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = \frac{7}{4} = H(X)$$

- ▶ If the length of the representation for each symbol is equal to its self-information, the average code length equals entropy

# Data compression

- ▶ Shannon coding: represent a symbol with probability of  $p$  with a sequence of length  $\lceil \log(1/p) \rceil$ 
  - ▶  $\lceil \log(1/p) \rceil < \log(1/p) + 1$
  - ▶ Achieves average code length  $< H(X) + 1$
- ▶ Shannon showed that it's not possible to do better than entropy
- ▶ Shannon's source coding theorem: the average code length  $L$  of the optimum code satisfies:

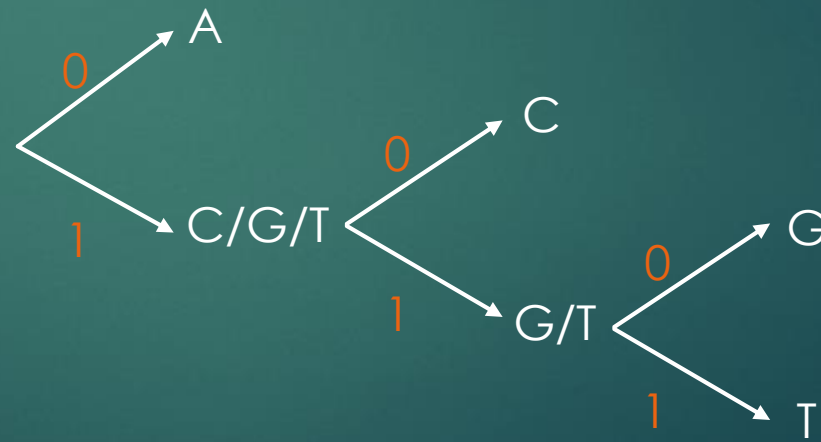
$$H(X) \leq L < H(X) + 1$$



# Huffman codes

- ▶ Shannon codes, while close to entropy, are not necessarily optimal
- ▶ To achieve optimality, each bit must divide the probability space to two nearly equal halves

	A	C	G	T
Prob	1/2	1/4	1/8	1/8
Code	0	10	110	111



# Huffman codes

- ▶ Shannon and others, including Huffman's professor, Fano, tried to find an optimal algorithm but were not successful
- ▶ Fano gave students a choice of final exam or a term paper solving given problems
- ▶ Huffman invented an algorithm for finding optimal codes
  - ▶ Huffman's algorithm builds the tree in a bottom-up approach, grouping smallest probabilities to create super-nodes
- ▶ The average code length for the Huffman code is still at least as large as the entropy

# Relative Entropy

- ▶ Suppose the true distribution of a source  $X$  is given by  $p$
- ▶ Not knowing this true distribution, we construct a code based on a distribution  $q$
- ▶ What is the inefficiency caused by this mismatch?
- ▶ Average code length with the true and assumed distributions:

$$\sum_x p(x) \log \frac{1}{p(x)}, \quad \sum_x p(x) \log \frac{1}{q(x)}$$

- ▶ The difference is the *relative entropy* (aka Kullback-Leibler divergence)

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

# Relative Entropy

- ▶ Relative entropy is used as a measure of difference between distributions
- ▶  $D(p||q) = 0$  if and only if  $p = q$
- ▶ Relative entropy is used as loss function in machine learning
  - ▶ Suppose we are interested in estimating an unknown distribution  $p$
  - ▶ We choose a simple class of distributions  $Q$
  - ▶ We find  $q \in Q$  that minimizes  $D(p||q)$
- ▶ This results in a distribution  $q$  that does not under-estimate  $p$ 
  - ▶ Avoids assigning zero probability where  $p(x) > 0$

# Relative Entropy

- ▶ Could also choose to minimize  $D(q||p)$  → different answer
  - ▶ Tries to not over-estimate  $p$

$$D(q||p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

- ▶ Avoids assigning probability where  $p(x) = 0$

# Cross-entropy

- ▶ Recall:

$$D(p||q) = \sum_x p(x) \log \frac{1}{q(x)} - \sum_x p(x) \log \frac{1}{p(x)}$$

- ▶  $q$  only appears in the first term, called cross-entropy

$$H(p||q) = \sum_x p(x) \log \frac{1}{q(x)}$$

- ▶ Minimizing relative entropy is the same as minimizing cross-entropy

# Joint entropy

- ▶ For two random variables  $X$  and  $Y$ , their joint entropy is

$$H(X, Y) = E \log \frac{1}{p(X, Y)} = \sum p(x, y) \log \frac{1}{p(x, y)}$$

- ▶  $X$  and  $Y$  are independent if and only if

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

- ▶ Example:  $X = \text{Ber}(1/2), Y = \text{Ber}(1/2), Z = X + Y$

$$\begin{aligned} H(X) = H(Y) = \log 2 = 1, & \quad H(Z) = 1.5 \\ H(X, Y) = H(X) + H(Y) = 2 = H(X, Z) \neq H(X) + H(Z) \end{aligned}$$

X	Y	Z	P
0	0	0	1/4
1	0	1	1/4
0	1	1	1/4
1	1	2	1/4

# Conditional entropy

- ▶ Conditional entropy of  $X$  given  $Z$

$$H(X|Z) = \sum_z p(z)H(X|Z = z) = \sum_z p(z) \sum_x p(x|z) \log \frac{1}{p(x|z)}$$

- ▶ The uncertainty left in  $X$  after we learn  $Z$

- ▶ Previous example:

$$H(X|Z) = \frac{1}{4} \times 0 + \frac{1}{2} \times 1 + \frac{1}{4} \times 0 = \frac{1}{2}, \quad H(Z|X) = 1$$

- ▶ Relationship between joint and conditional entropies

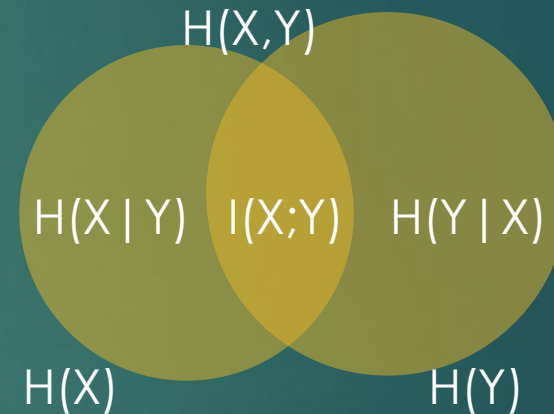
$$H(X, Z) = H(X) + H(Z|X)$$

X	Y	Z	P
0	0	0	1/4
1	0	1	1/4
0	1	1	1/4
1	1	2	1/4



# Mutual Information

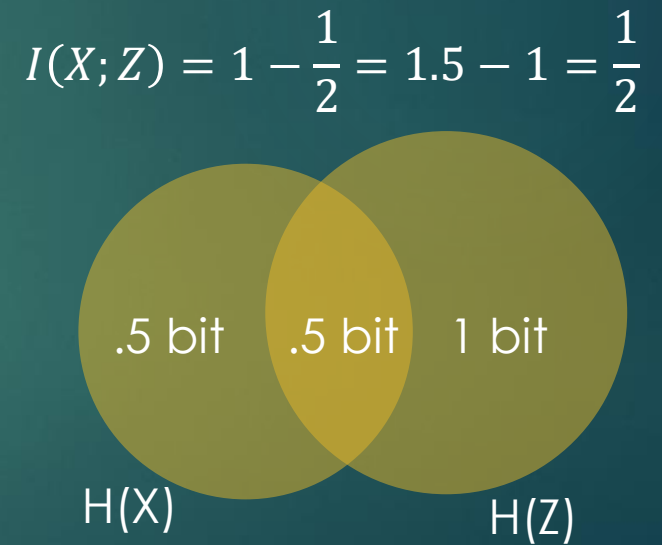
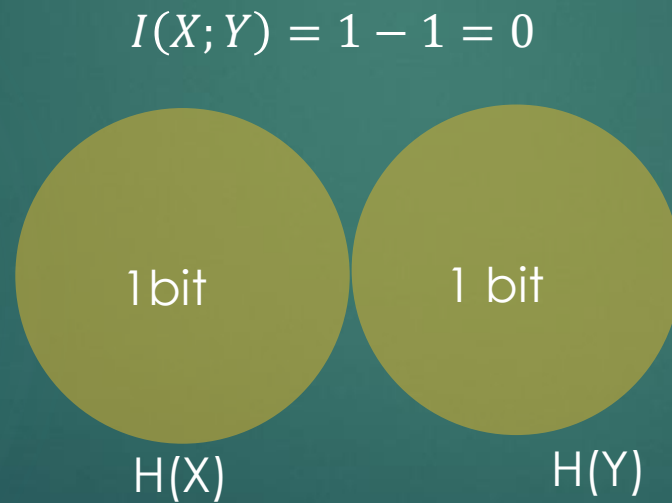
- ▶  $I(X; Y)$ : Mutual information between two random variables
- ▶ The reduction of uncertainty about  $X$  due to knowledge of  $Y$
- ▶  $I(X; Y) = H(X) - H(X|Y)$
- ▶  $\quad \quad = H(Y) - H(Y|X)$



# Mutual Information

- ▶  $I(X; Y) = H(X) - H(X|Y)$
- ▶ Example:

X	Y	Z=X+Y	P
0	0	0	1/4
1	0	1	1/4
0	1	1	1/4
1	1	2	1/4



# Entropy $\neq$ (Mutual) Information

- ▶ Example: cable news (high entropy, little mutual information to news)



# Channel Capacity

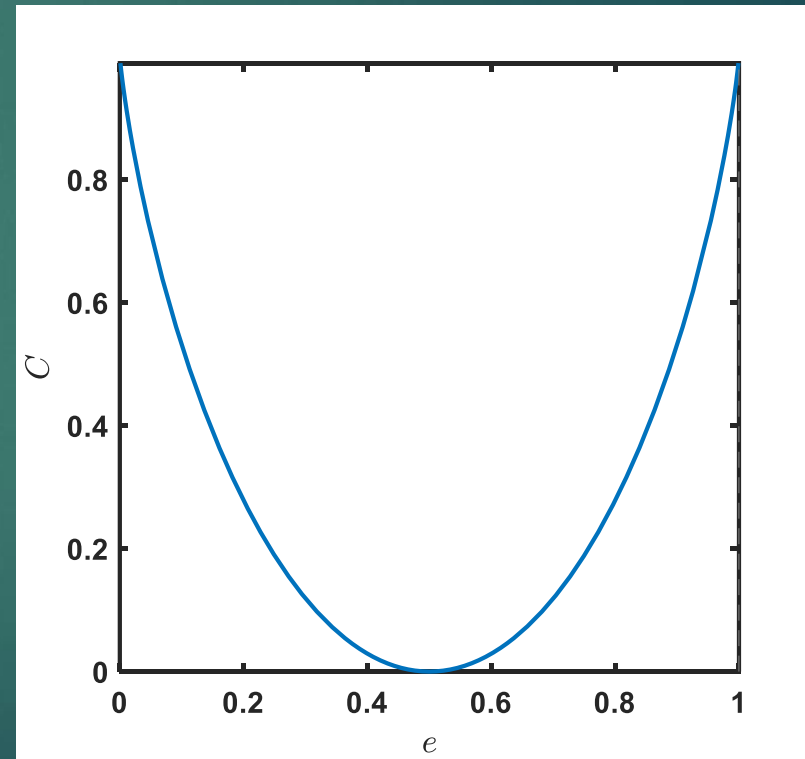
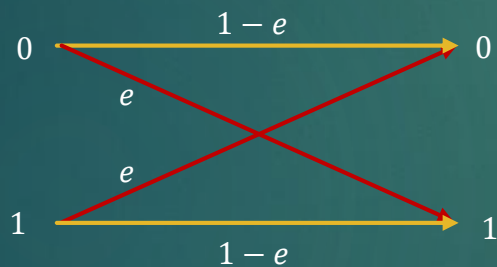
- ▶ Communication channel



- ▶ Due to noise, the input and output are only statistically related
- ▶ **Shannon Channel Coding Theorem:**
  - ▶ The maximum information rate that can be carried by a communication channel, is the **maximum mutual information** between its input and output

# Channel Capacity

- ▶ Binary symmetric channel  $\rightarrow$  Capacity =  $1 - H(e)$



# Data processing inequality

- ▶ Random variables  $X, Y, Z$  form a Markov chain if  $X$  and  $Z$  are conditionally independent given  $Y$ 
  - ▶ Denoted  $X \rightarrow Y \rightarrow Z$
- ▶ The data processing inequality: If  $X \rightarrow Y \rightarrow Z$ , then  $I(X; Z) \leq I(X; Y)$ .



- ▶ No processing, whether deterministic or random, can increase the amount of information that  $Y$  has about  $X$

# Sufficient statistics

- ▶ Consider
  - ▶  $\{p_\theta\}$ : a family of distributions indexed by  $\theta$
  - ▶  $X$ : a sample from this distribution
  - ▶  $T(X)$ : any statistic (function of the sample), e.g., sample mean
- ▶ Then  $\theta \rightarrow X \rightarrow T(X)$
- ▶  $I(\theta; T(X)) \leq I(\theta; X)$
- ▶ If  $I(\theta; T(X)) = I(\theta; X)$ , then  $T(X)$  is a **sufficient statistic**
  - ▶ The condition is equivalent to  $\theta \rightarrow T(X) \rightarrow X$
  - ▶  $X$  is independent of  $\theta$  given  $T(X)$
- ▶ The sufficient statistic contains all the information in  $X$  about  $\theta$

# Sufficient Statistics

- ▶  $X_i \sim \text{Bernoulli}(\theta)$ ,  $X = (X_1, \dots, X_n)$ ,  $S = \sum X_i$ 
  - ▶  $\theta \rightarrow X \rightarrow S$
  - ▶  $\theta \rightarrow S \rightarrow X$
  - ▶ Given the number of ones,  $X$  is independent of  $\theta$  since all sequences with  $S$  ones are equally probable, with probability  $1/\binom{n}{S}$
- ▶  $X_i \sim \text{Normal}(\theta, 1)$ ,  $X = (X_1, \dots, X_n)$ ,  $\bar{X} = \sum_i X_i / n$  is a sufficient statistic
- ▶  $X_i \sim \text{Uniform}[0, \theta]$ ,  $X = (X_1, \dots, X_n)$ ,  $M = \max X_i$  is a sufficient statistic
- ▶ Minimal sufficient statistic: a SS that is a function of every other SS



# Fano's inequality

- ▶ We know a random variable  $Y$  and want to estimate  $X$
- ▶ How is the probability of error affected by  $H(X|Y)$ ?
  - ▶ Best case:  $X$  is a function of  $Y$ :  $H(X|Y) = 0$
  - ▶ Worst case:  $X$  and  $Y$  are independent:  $H(X|Y) = H(X)$
- ▶ Let the estimate be  $\hat{X} = g(Y)$ , a (possibly random) function of  $Y$
- ▶  $P_e = \Pr(\hat{X} \neq X)$ ,  $M$ : number of possible values of  $X$
- ▶ Fano's inequality:  $H(P_e) + P_e \log M \geq H(X|Y)$  and

$$P_e \geq \frac{H(X|Y) - 1}{\log M}$$

# Fano's inequality

- ▶ Special case:  $P_e = 0 \Rightarrow H(X|Y) = 0$ 
  - ▶  $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
  - ▶  $H(Y) \geq H(X)$
- ▶ On average, how many pairwise comparisons do we need to sort a list of size  $n$ 
  - ▶  $Y$ : the results of pairwise comparisons
  - ▶  $M$ : average number of comparisons
  - ▶ We need to identify one permutation among  $n!$
  - ▶  $M \geq H(Y) \geq H(X) = \log n! \simeq n \log n$
  - ▶ Independent of how we choose items to compare

# Entropy rate

- ▶ Consider the sequence:
  - ▶ 0000000111100000011111111000001111110000001111
  - ▶ What is the entropy per symbol?
  - ▶  $p_0 \simeq p_1 \simeq \frac{1}{2} \Rightarrow H \simeq 1 \text{ bits}$
  - ▶ We are ignoring the dependence between symbols
- ▶ Probability distribution for the next symbol depends on the previous symbol
  - ▶  $P(X_i = 1|X_{i-1} = 1) = 0.9$
  - ▶  $P(X_i = 0|X_{i-1} = 0) = 0.9$
- ▶ This is called a Markov chain
- ▶ What is the **entropy rate**  $h$ , amount of information in each symbol?

# Entropy Rate of Markov Chains

- ▶ What is the entropy rate of a two state Markov chain?

- ▶  $h = H(X_i|X_{i-1}) = \sum \Pr(X_{i-1} = x_{i-1})H(X_i|X_{i-1} = x_{i-1})$

- ▶ Example: two-state Markov chain

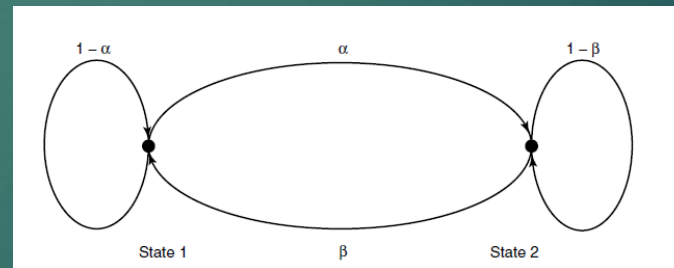
- ▶  $H(X_i|X_{i-1} = 0) = H(\alpha)$

- ▶  $H(X_i|X_{i-1} = 1) = H(\beta)$

- ▶  $\Pr(X_{i-1} = 0) = \frac{\beta}{\alpha + \beta}$

- ▶  $\Pr(X_{i-1} = 1) = \frac{\alpha}{\alpha + \beta}$

- ▶  $h = \frac{\alpha}{\alpha + \beta}H(\alpha) + \frac{\beta}{\alpha + \beta}H(\beta)$



Credit: Elements of Inf Theory, Cover and Thomas

# Entropy rate

- ▶ Markov chains can have memory larger than 1 symbol
- ▶ Some processes, such as English text can only be approximated as a Markov chain
- ▶ From Shannon's original paper:
  - ▶ **0<sup>th</sup> order:** XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD  
QPAAMKBZAACIBZLHJQD
  - ▶ **1<sup>st</sup> order:** OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA  
NAH BRL
  - ▶ **4<sup>th</sup> order:** THE GENERATED JOB PROVIDUAL BETTER TRAND THE DISPLAYED CODE,  
ABOVERY UPONDULTS WELL THE CODERST IN THESTICAL IT DO HOCK BOTHE MERG.
  - ▶ **2<sup>nd</sup> order word model:** THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT  
THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT  
THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED
- ▶ 0 order entropy =  $\log 27 = 4.76$  bits
- ▶ 4<sup>th</sup> order entropy = 2.8 bits

# Thank you



- ▶ References:

- ▶ “Elements of information theory,” Thomas Cover, Joy Thomas
- ▶ “Information Theory, Inference, and Learning Algorithms,” David MacKay