

ECE 587 / STA 563: Lecture 1 – Introduction

Information Theory
Duke University, Fall 2024

History: Written by Galen Reeves (-2023). Updated by Henry Pfister (2024-).

Last Modified: September 11, 2024

Outline of lecture:

1.1	What is information Theory	1
1.1.1	Brief history of communication	1
1.1.2	Contributions of Claude Shannon	2
1.1.3	Broader role of information theory	3
1.2	Who should take this course	4
1.3	What will we learn in this course	4
1.4	Resources	5
1.5	Probability Review*	6

1.1 What is information Theory

Information theory is the science of processing, transmitting, storing, and using information.

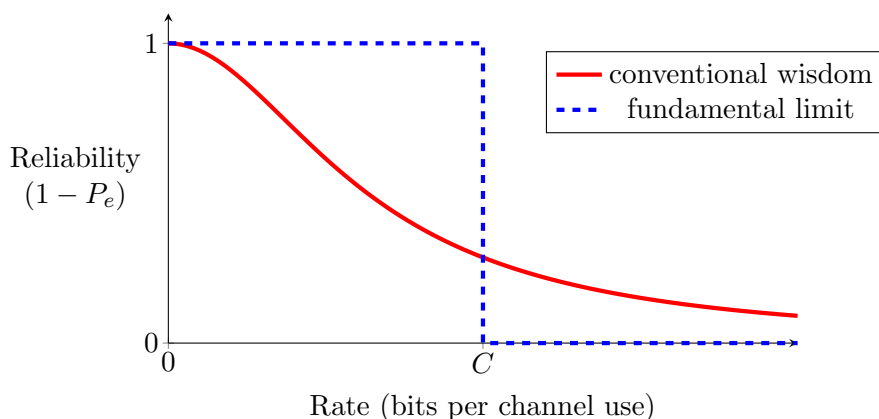
1.1.1 Brief history of communication

- History of communications: A partial timeline from http://en.wikipedia.org/wiki/History_of_communication
 - language (100,000 BCE)
 - cave paintings (30,000 BCE)
 - petroglyphs – carvings in a rock surface (10,000 BCE)
 - pictograms (6,000 BCE)
 - writing (3,000 BCE)
 - fires, beacons, smoke signals, communication drums
 - mail systems (Persia, 6th century BCE), pigeon post,
 - maritime flags (15th century CE)
 - telegraph (1838) use Morse code. Frequent letters have short codes; infrequent letters have long codes. (e.g. E is a dot and J is a dot followed by three dashes.)
 - telephone (1848)
 - radio (1896)
 - television (1927)
 - trans-atlantic telegraph cable (1858)
 - fiber-optic communication (1970s)
 - internet (1983)

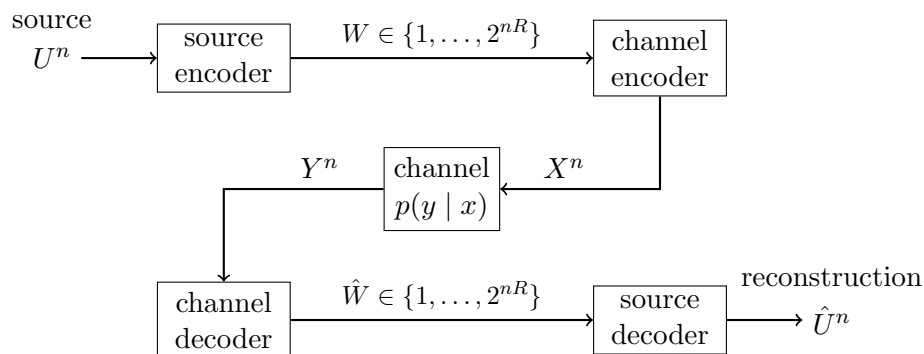
- * Napster (1999)
- * Facebook (2004)
- * YouTube (2005)
- * Twitter (2006)
- * Instagram (2010)
- * TikTok (2018)
- mobile telephony: from 1G (introduced in 1980s) through 5G (current)
- A turtle with a hard drive on its back...

1.1.2 Contributions of Claude Shannon

- In 1948 Claude Shannon surprised the communication theory community by answering two fundamental questions:
 - How much can data be compressed (on average)?
 - How fast can data be reliably transmitted (on average)?
- Fundamental tradeoff between *reliability* and *efficiency*



- One of Shannon's key insights was that information is *not* about semantics but about probabilities.
- Other key insights include focusing on long blocks of information and using *bits* (binary digits) to provide a universal currency for measuring information.



- **Source Coding Theorem:** The minimum rate R at which an information source (i.e., data stream) can be compressed losslessly is the *entropy rate* H of the source (e.g., measured in units of bits per source symbol)
 - **Channel Coding Theorem:** The maximum rate R that information can be transmitted reliably over an information channel is the *capacity* C of the channel (e.g., measured in bits per channel use)
 - **Source - Channel Separation Theorem:** An information source with entropy rate H can be transmitted reliably over an information channel with capacity C if and only if $H < C$. Furthermore, there is no loss in using separate source coding and channel coding.
- The Source Coding Theorem actually provides an operational definition for the entropy rate. Let X^n be a Bernoulli- p source (i.e., an i.i.d. binary sequence where the probability of a 1 is p). What are some implied upper bounds on the entropy rate of this sequence?
 - If we don't compress, then we can represent the source using 1-bit per symbol. Thus, the entropy rate is less than 1 bit per symbol. But, one can do better.
 - The resulting sequence will have a random number $K \in \{0, 1, \dots, n\}$ of ones. But, we can send $\lceil \log_2(n+1) \rceil$ bits to represent K and then send $\lceil \log_2 \binom{n}{K} \rceil$ bits to represent which length- n sequence with exactly K ones was observed.
 - How will K behave for large n ? How should we handle randomness in K ?
 - How will $\lceil \log_2 \binom{n}{K} \rceil$ behave for typical K and large n ? How can we use this to upper bound the entropy rate?
 - Shannon showed the fundamental limits of what was possible for *any* point-to-point communication system. At the time, some thought this was an esoteric theory with little practical value.
 - Over the last 70 years, researchers have figured out how to achieve these limits practically.
 - (1977) Lempel and Ziv develop optimal lossless compression data schemes. Ubiquitous in .zip compression formats. (See <https://www.hanshq.net/zip2.html>)
 - Lossy compression: JPEG, MPEG, Many other codes
 - (1962) Gallager proposes low-density parity check (LDPC) codes but the decoding complexity is too high; so they are forgotten for more than 30 years.
 - (1993) Berrou, Glavieux, and Thitimajshima invent Turbo codes. They are the first practical codes with good performance near capacity and are currently used in 3G and 4G mobile communications.
 - (1996) MacKay rediscovers LDPC codes and introduces irregularity. Using modern computational power, Richardson, Urbanke, and Shokrollahi optimize them to approach capacity.
 - (2008) Arikan discovers Polar Codes. These are first *deterministic codes* to provably achieve capacity.

1.1.3 Broader role of information theory

- Secrecy, privacy, cryptography
 - Wiretap channel with an eavesdropper
 - Covert communication
 - Information leakages
- Coding theory – design and analysis of efficient coding and encoding strategies
- Network information theory – multiple receivers and multiple senders
 - Source - channel separation need not hold
 - Slepian and Wolf describe fundamental limits of distributed encoding of correlated sources.
 - Models distributed computation (e.g./ federated learning) with communication constraints.
- Close connection with many fields (see Figure 1.1 in [CT]):
 - Probability Theory: Limit Theorems, Large deviations
 - Statistics: Hypothesis testing, Fisher Information
 - Economics: Portfolio Theory, Gambling
 - Computer Science: Kolmogorov Complexity
 - Physics: Thermodynamics, Quantum Information Theory
 - Biology, Neuroscience
- Some active research areas at the interface high-dimensional inference and machine learning
 - Information theoretic limits of estimation, testing, and learning.
 - Characterizing phase transitions and computational-to-statistical gaps
 - Probabilistic understanding of machine learning method
 - Use of information measures for training models and regularization
 - Privacy
 - Distributed learning
 - Estimation of information measures.
 - Tackling *high-dimensional* information sources.

1.2 Who should take this course

- Students in ECE, Statistics, Computer Science, Math, Physics, Econ, Neuroscience, other fields.
- Prerequisites: comfort with probability at the undergraduate level; measure theoretic probability is not assumed. Experience with convergence of sequences and linear algebra.
- Prerequisites are light but ideas are deep. Student feedback indicates course is difficult but material is enjoyable.

1.3 What will we learn in this course

- We will cover chapters 1– 5, 7–10, 13 in [CT].
- We will also cover additional topics through course notes and other readings.
- Unlike many other classes, we strive to understand “why” through full proofs
- Overview is on course website.

1.4 Resources

See course website for full list of resources.

1.5 Probability Review*

Students who take this class often have a variety of backgrounds, including electrical engineering, statistics, computer science, math, physics, and other engineering and science disciplines. In order to be successful in this course it is necessary that you are *comfortable* working with probability at the undergraduate level. This includes conditioning, expectation, discrete and continuous randoms as well as familiarity with multivariate Gaussian distributions, Markov chains, and convergence of random variables. That said, this course is designed such that you do *not* need probability at the graduate level (i.e., measure-theoretic probability).

- **Probability space**

- sample space Ω of all possible outcomes
- event space \mathcal{F} of events defined on the sample space (e.g., $A, B \in \mathcal{F}$ are events)
- probability measure \mathbb{P} that satisfies three axioms

$$\mathbb{P}[\Omega] = 1, \quad \mathbb{P}[A] \geq 0, \quad \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] \quad \text{for all } A, B \in \mathcal{F} \text{ with } A \cap B = \emptyset$$

- Formally, a (real) random variable X is a function from Ω to \mathbb{R} which specifies the value of X when outcome ω occurs (e.g., $X(\omega) = \mathbf{1}_A(\omega)$ is the indicator rv of event A)

- **Example:** There is a 1% chance I have a certain disease. I take a test for this disease which is 90% accurate. i.e.

$$\mathbb{P}[\text{positive} \mid \text{disease}] = \mathbb{P}[\text{negative} \mid \text{no disease}] = 0.9$$

Given the test is positive, what is the probability I have the disease?

- Let A be the event I have the disease and B be the event that the test is positive.

$$\mathbb{P}[A] = 0.01, \quad \mathbb{P}[B \mid A] = \mathbb{P}[B^c \mid A^c] = 0.9$$

$$\begin{aligned} \mathbb{P}[A \mid B] &= \frac{\mathbb{P}[B \mid A]\mathbb{P}[A]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B \mid A]\mathbb{P}[A]}{\mathbb{P}[B \mid A]\mathbb{P}[A] + \mathbb{P}[B \mid A^c]\mathbb{P}[A^c]} \\ &= \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.1 \times 0.99} = \frac{9}{118} = \frac{1}{12} \end{aligned}$$

- Even though the test is highly accurate, the probability I have the disease given a positive outcome is relatively small. This is because the prior probability that I have the disease is very small so the most likely explanation of a positive result is that it's a false positive.

- **Notation**

- Random variables are denoted by uppercase: X, Y, Z
- Deterministic (i.e., non-random) values denoted by lower case: x, y, z
- Support (or alphabet) of random variable denoted by calligraphic font $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$

- **Discrete random variables:**

- The **probability mass function (pmf)** of a discrete random variable is X with support \mathcal{X} is given by

$$p_X(x) = \mathbb{P}[X = x] \quad \text{for all } x \in \mathcal{X}$$

To simplify notation, it is common to write $p(x)$ where the association with the random variable X is implied by the argument of the function.

- The joint pmf of random variables (X, Y) is given by

$$p_{X,Y}(x, y) = \mathbb{P}[X = x, Y = y] \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}$$

To simplify notation, it is common to write $p(x, y)$ where the association with the pair (X, Y) is implied by the argument of the function.

- The marginal distributions of X and Y respectively are given by

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y), \quad p_Y(y) = \sum_{x \in \mathcal{X}} p_{X,Y}(x, y)$$

- **Independence:**

- Events A and B are independent if and only if their joint probability equals the product of their probabilities

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$$

- Random variables X and Y are **independent** if and only if the joint probability is equal to the product of the marginals:

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}$$

- **Example:** Let X and Y be independent random variables supported on $\mathcal{X} = \mathcal{Y} = \{1, \dots, M\}$. What is the probability that X equals Y ?

$$\begin{aligned} \mathbb{P}[X = Y] &= \sum_{m=1}^M p_X(m)\mathbb{P}[X = Y \mid X = m] \\ &= \sum_{m=1}^M p_X(m)\mathbb{P}[Y = m] \\ &= \sum_{m=1}^M p_X(m)p_Y(m) \end{aligned}$$

- **Example:** Provide an example of three random variables X, Y, Z that are pair-wise independent but not independent.
- **Independent and identically distributed (i.i.d.):** A sequence of random variables X_1, X_2, \dots is i.i.d. if the variables are independent and share a common marginal distribution $p(x)$, i.e.,

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i)$$

- **Expectation:**

- The expected value of a random variable X is given by

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p_X(x)$$

- The expected value of a function $f : \mathcal{X} \rightarrow \mathbb{R}$ is given by

$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x) p_X(x)$$

- Warning: the expectation $\mathbb{E}[f(X)]$ not a random variable. Instead, it is a functional of the *distribution* of X . Sometimes it is written as $\mathbb{E}_{p_X}[f]$ to make this relationship clear.
- The conditional expectation of X given a particular realization $\{Y = y\}$ is

$$\mathbb{E}[X | Y = y] = \sum_{x \in \mathcal{X}} x p_{X|Y}(x | y)$$

This is a deterministic (nonrandom) quantity that is a function of y .

- The conditional expectation of X given Y is

$$\mathbb{E}[X | Y] = \sum_{x \in \mathcal{X}} x p_{X|Y}(x | Y)$$

This this is a random variable because it is a function of the random variable Y .

- **Variance:** The variance of a random variable X is given by

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

and the conditional variance $\text{Var}(X | Y)$ is a random variable that, given $Y = y$, equals the variance of $X \sim p_{X|Y}(x | y)$.

$$\text{Var}(X | Y = y) = \mathbb{E}[(X - \mathbb{E}[X | Y = y])^2 | Y = y]$$

- **Example:** The law of total variance states that

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X | Y)] + \text{Var}(\mathbb{E}[X | Y])$$

- **Law of large numbers (LLN):** If a sequence of random variables X_1, X_2, \dots is i.i.d. with finite absolute first moment $\mathbb{E}[|X_1|] < \infty$, then the long-term average converges to the mean:

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mathbb{E}[X]$$

with probability one as $n \rightarrow \infty$.

- **Central limit theorem (CLT):** If a sequence of random variables X_1, X_2, \dots is i.i.d. with mean $\mu = \mathbb{E}[X_1]$ and finite variance $\sigma^2 = \text{Var}(X_1) < \infty$ then the fluctuation of the long-term average about the the mean has a Gaussian distribution

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \rightarrow \mathcal{N}(0, \sigma^2)$$

in distribution as $n \rightarrow \infty$.

- **Example:** Let X_1, X_2, \dots , be a sequence of i.i.d. Bernoulli(p) variables, $p \in (0, 1)$, and let $S_n = X_1 + \dots + X_n$ be the sum of the first n terms.

- S_n has a binomial distribution with parameters n and p , and probability mass function

$$p_{S_n}(s) = \binom{n}{s} p^s (1-p)^{n-s}, \quad s \in \{0, 1, \dots, n\}$$

and cumulative distribution function

$$F_{S_n}(s) = \mathbb{P}[S_n \leq s] = \sum_{k \leq s} p_{S_n}(k), \quad s \in (-\infty, \infty)$$

- By law of large numbers, $\frac{1}{n}S_n \rightarrow p$ as $n \rightarrow \infty$. This implies that the cdf converges to a step function

$$F_{\frac{1}{n}S_n}(t) = \mathbb{P}[n^{-1}S_n \leq t] \rightarrow \begin{cases} 1, & t > p \\ 0, & t < p \end{cases}$$

- By central limit theorem, $Z_n = (S_n - \mathbb{E}[S_n])/\sqrt{n \text{Var}(S_1)}$ converges in distribution to a $\mathcal{N}(0, 1)$ random variable as $n \rightarrow \infty$. This is equivalent to saying that the cdf converges to the cdf of a standard Gaussian variable:

$$F_{Z_n}(z) = \mathbb{P}\left[\frac{S_n - n\alpha}{\sqrt{n\alpha(1-\alpha)}} \leq z\right] \rightarrow \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

- **Warning:** Convergence in distribution can occur to any distribution and it does not always imply that the pmf $p_{Z_n}(\cdot)$ converges to the pdf of the Gaussian distribution.