18-753: Information Theory and Coding
Lecture 1: Introduction and Probability Review
Notes build on those of Bobak Nazer at BU

Prof. Pulkit Grover

CMU

January 14, 2020

Syllabus

Syllabus: is here.

Bits and Information Theory

- More often than not, today's information is measured in bits.

Bits and Information Theory

- More often than not, today's information is measured in bits.
- Why?

Bits and Information Theory

- More often than not, today's information is measured in bits.
- Why?
- Is it ok to represent signals as different as audio and video in the same currency, bits?

Bits and Information Theory

- More often than not, today's information is measured in bits.
- Why?
- Is it ok to represent signals as different as audio and video in the same currency, bits?
- Also, why do I communicate over long distances with the exact same currency?

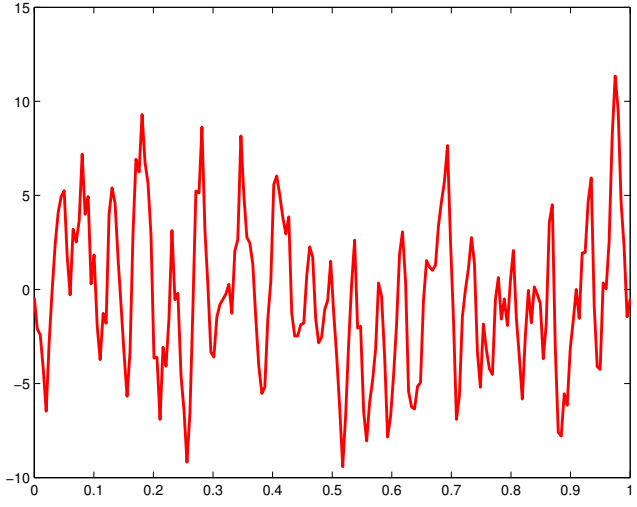
Video

- Video signals are made up of colors that vary in space and time.

- Even if we are happy with pixels on a screen how do we know that all these colors are optimally described by bits?

Audio

- Audio signals can be thought of as an amplitude that varies continuously with time.



- How can we optimally represent a continuous signal in something discrete like bits?

Just a Single Bit

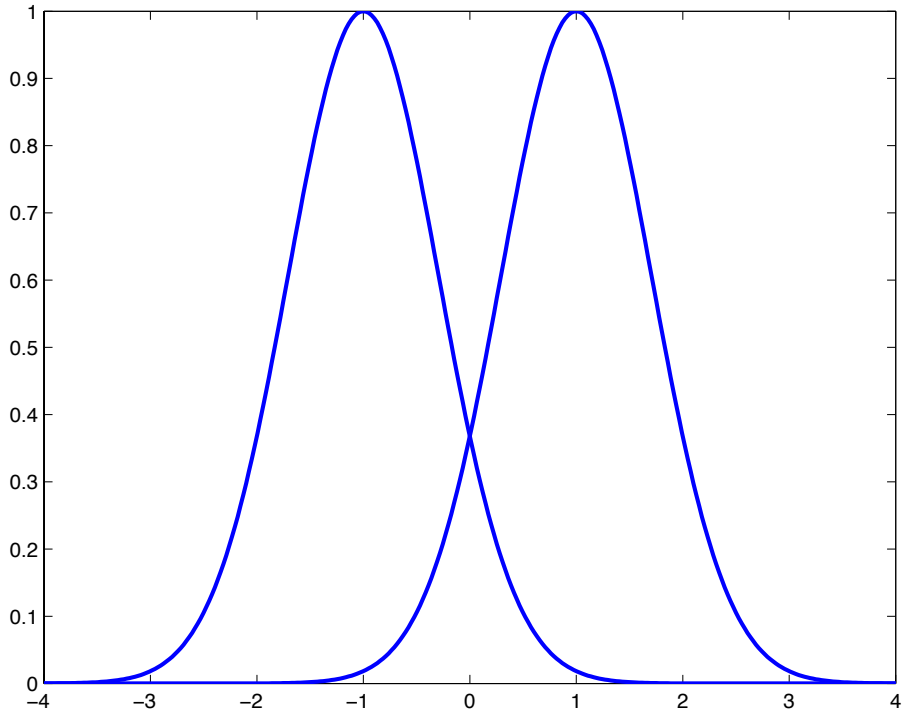
- These (and many other) signals can all be optimally represented by bits.
- Information theory explains why and how.
- For now, let's say we have a single bit that we really want to communicate to our friend. Let's say it represents whether or not it is going to snow tonight.

Communication over Noisy Channels

- Let's say we have to transmit our bit over the air using wireless communication.
- We have a transmit antenna and our friend has a receive antenna.
- We'll send out a negative amplitude (say -1) on our antenna when the bit is 0 and a positive amplitude (say 1) when the bit is 1.
- Unfortunately, there are other signals in the air (natural and man-made), or in the receiver, so the receiver sees a noisy version of our transmission.
- If there are lots of these little effects, then the central limit theorem tells us we can just model them as a single Gaussian random variable.

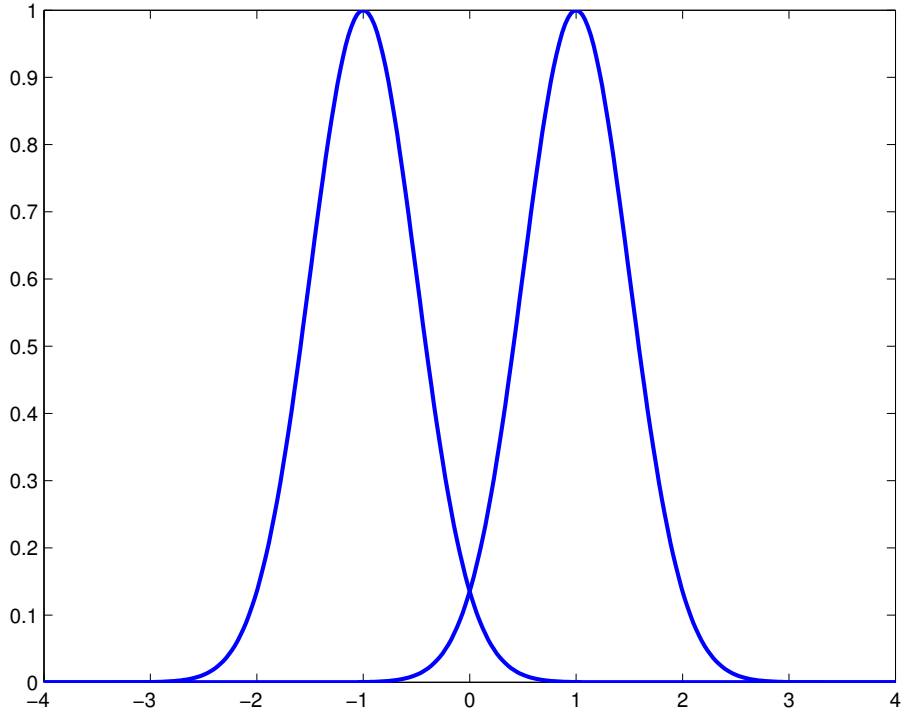
Received Signal

- Here is the probability distribution function for the received signal. No longer a clean -1 or 1.
- You can prove that the best thing to do now is just decide that it's -1 if the signal is below 0 and 1 otherwise.
- But that leaves us with a probability of error.



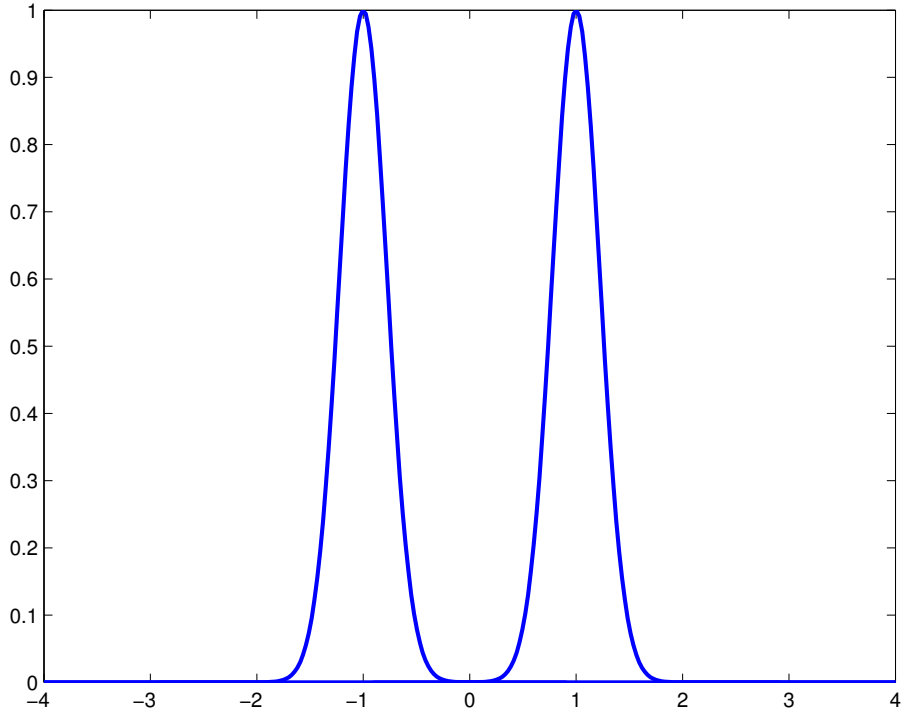
Received Signal

- One thing we can do is boost our transmit power.
- The received signal will look less and less noisy.



Received Signal

- One thing we can do is boost our transmit power.
- The received signal will look less and less noisy.



Repetition Coding

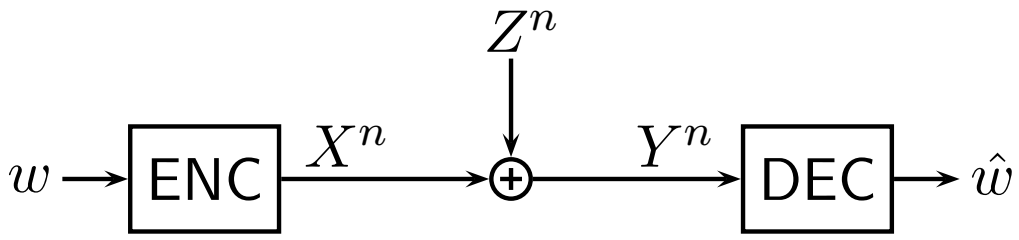
- What if we can't arbitrarily increase our transmit power?
- We can just repeat our bit many times! For example, if we have a 0, just send $-1, -1, -1, \ldots, -1$ and take a majority vote.
- Now we can get the probability of error to fall with the number of repetitions.
- But the rate of incoming bits quickly goes to zero. Can we do better?
- I call the main result that allows us to do this as "Shannon Whispering". You don't need to speak louder, you don't need to repeat yourself many times, you just ...

Repetition Coding

- What if we can't arbitrarily increase our transmit power?
- We can just repeat our bit many times! For example, if we have a 0, just send $-1, -1, -1, \ldots, -1$ and take a majority vote.
- Now we can get the probability of error to fall with the number of repetitions.
- But the rate of incoming bits quickly goes to zero. Can we do better?
- I call the main result that allows us to do this as "Shannon Whispering". You don't need to speak louder, you don't need to repeat yourself many times, you just ... need to have a lot to say,

Repetition Coding

- What if we can't arbitrarily increase our transmit power?
- We can just repeat our bit many times! For example, if we have a 0, just send $-1, -1, -1, \ldots, -1$ and take a majority vote.
- Now we can get the probability of error to fall with the number of repetitions.
- But the rate of incoming bits quickly goes to zero. Can we do better?
- I call the main result that allows us to do this as "Shannon Whispering". You don't need to speak louder, you don't need to repeat yourself many times, you just ... need to have a lot to say, and "code" the information.
- (Need to draw on the board)



- We know the capacity for an Gaussian channel:

$$C = \frac{1}{2}\log\left(1 + \text{SNR}\right) \text{ bits per channel use}$$

- Proved by Claude Shannon in 1948.
- What does this mean?

The Benefits of Blocklength

- We can't predict what the noise is going to be in a single channel use.
- But we do know that in the long run the noise is going to behave a certain way.
- For example, if a given channel flips bits with probability 0.1 then in the long run approximately $\frac{1}{10}$ of the bits will be flipped.

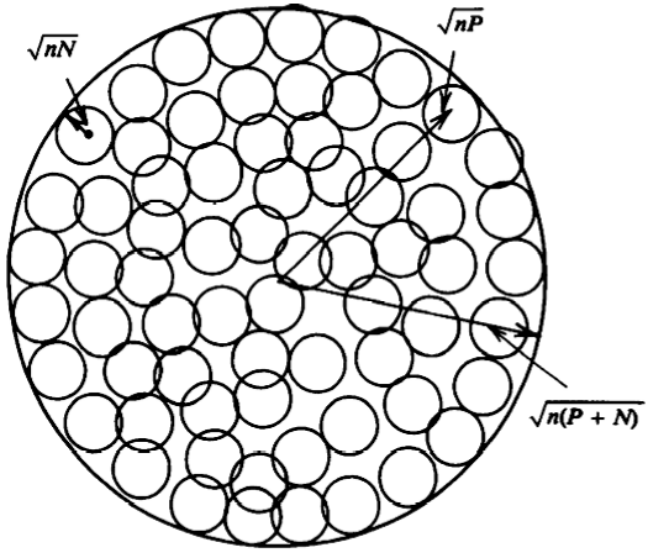


Figure 10.2. Sphere packing for the Gaussian channel.

(Cover and Thomas,
Elements of Information Theory)

Capacity

- So if we are willing to allow for some delay we can communicate reliably with positive rate!

- Capacity is actually very simple to calculate using the mutual information, $C = \max_{p(x)} I(X;Y)$.

What is Information Theory?

- A powerful mathematical framework that allows us to determine the fundamental limits of information compression, storage, processing, communication, and use.
- Provides the theoretical underpinnings as to why today's networks are completely digital.
- Unlike many other classes, we will strive to understand "why" through full proofs.
- As initially formulated, information theory ignores semantics of the message. We will explicitly discuss applications on how it is being extended to incorporate semantics.

Organizational Details

- This is 18-753: Information Theory and Coding.
- Designed and taught by Pulkit Grover, Sanghamitra Dutta, Praveen Venkatesh.
- Pre-requisites: Fluency in probability and mathematical maturity.
- Course Ingredients: $2-3$ Homeworks, class participation, and a course project.
- The class focuses on formal statements, formal proofs. The main question we are asking is: how do we arrive at informational measures for different applications? Applications, such as neuroscience and FATE of AI, will be introduced, but not in depth.
- Very few lectures will use slides.
- Textbook: Cover & Thomas, Elements of Info Theory, 2nd ed. Available online: http://onlinelibrary.wiley.com/book/10.1002/047174882X (need to login through CMU library)
- Office Hours: TBD.

- Consider the following (hypothetical) interactions between two students in CMU.
  - A: Have you ever been to the Museum of Natural History?
    B: Yes.

- Consider the following (hypothetical) interactions between two students in CMU.
  - A: Have you ever been to the Museum of Natural History?
    B: Yes.
  - A: Have you ever been to the moon?
    B: No.

- Consider the following (hypothetical) interactions between two students in CMU.
  - A: Have you ever been to the Museum of Natural History?
    B: Yes.
  - A: Have you ever been to the moon?
    B: No.
- Both questions had two possible answers. Which interaction conveyed more information?

- Consider the following (hypothetical) interactions between two students in CMU.
  - A: Have you ever been to the Museum of Natural History?
    B: Yes.
  - A: Have you ever been to the moon?
    B: No.
- Both questions had two possible answers. Which interaction conveyed more information?
- The "amount of information" in an event appears to be related to how likely the event is.

- Consider the following (hypothetical) interactions between two students in CMU.
  - A: Have you ever been to the Museum of Natural History?
    B: Yes.
  - A: Have you ever been to the moon?
    B: No.
- Both questions had two possible answers. Which interaction conveyed more information?
- The "amount of information" in an event appears to be related to how likely the event is.
- Maybe also on the utility of the answer?

- Consider the following (hypothetical) interactions between two students in CMU.
  - A: Have you ever been to the Museum of Natural History?
    B: Yes.
  - A: Have you ever been to the moon?
    B: No.
- Both questions had two possible answers. Which interaction conveyed more information?
- The "amount of information" in an event appears to be related to how likely the event is.
- Maybe also on the utility of the answer?
- Classical information theory takes the former view (just based on likelihood). In this course, we will start with the classical view, and then go significantly beyond that.

My personal motivation for going beyond

I was bothered for a very long time, and still am, that information theory is largely a theory of communication/compression. A theory of information should be much broader.

- My PhD thesis examined how information theory can solve a distributed control problem (the Witsenhausen counterexample, 1967). This is an application of information theory to a problem where *end-use of information matters*: it is defined by the optimization goal (which is *not* communication).
  – also, this is a *scalar* problem.

My personal motivation for going beyond

I was bothered for a very long time, and still am, that information theory is largely a theory of communication/compression. A theory of information should be much broader.

- My PhD thesis examined how information theory can solve a distributed control problem (the Witsenhausen counterexample, 1967). This is an application of information theory to a problem where *end-use of information matters*: it is defined by the optimization goal (which is *not* communication).
  – also, this is a *scalar* problem.
- I then examined a problem of communication and computing: minimizing total (comm+compute) energy in a communication system. Shannon capacity does not fall out as the answer.

My personal motivation for going beyond

I was bothered for a very long time, and still am, that information theory is largely a theory of communication/compression. A theory of information should be much broader.

- My PhD thesis examined how information theory can solve a distributed control problem (the Witsenhausen counterexample, 1967). This is an application of information theory to a problem where *end-use of information matters*: it is defined by the optimization goal (which is *not* communication).
  – also, this is a *scalar* problem.
- I then examined a problem of communication and computing: minimizing total (comm+compute) energy in a communication system. Shannon capacity does not fall out as the answer.
- Since then, I have examined error correction in distributed computing, measures of information flow in neural circuits, and measures of fairness in machine learning.

My personal motivation for going beyond

I was bothered for a very long time, and still am, that information theory is largely a theory of communication/compression. A theory of information should be much broader.

- My PhD thesis examined how information theory can solve a distributed control problem (the Witsenhausen counterexample, 1967). This is an application of information theory to a problem where *end-use of information matters*: it is defined by the optimization goal (which is *not* communication).
  – also, this is a *scalar* problem.
- I then examined a problem of communication and computing: minimizing total (comm+compute) energy in a communication system. Shannon capacity does not fall out as the answer.
- Since then, I have examined error correction in distributed computing, measures of information flow in neural circuits, and measures of fairness in machine learning.
- In no case is Shannon capacity the answer. But in every case, a careful, theoretical approach, yields answers and insights.

Classic applications: Compression and Communication

- Information Theory is the science of measuring of information.

Classic applications: Compression and Communication

- Information Theory is the science of measuring of information.
- This science has had a profound impact on sensing, compression, storage, extraction, processing, and communication of information.
  - Compressing data such as audio, images, movies, text, sensor measurements, etc. (Example: We will see the principle behind the 'zip' compression algorithm in this course.)
  - Communicating data over noisy channels such as wires, wireless links, memory (e.g. hard disks), etc.

Classic applications: Compression and Communication

- Information Theory is the science of measuring of information.
- This science has had a profound impact on sensing, compression, storage, extraction, processing, and communication of information.
  - Compressing data such as audio, images, movies, text, sensor measurements, etc. (Example: We will see the principle behind the 'zip' compression algorithm in this course.)
  - Communicating data over noisy channels such as wires, wireless links, memory (e.g. hard disks), etc.
- Specifically, we will be interested in determining the **fundamental limits** of compression and communication. This will shed light on how to engineer near-optimal systems.

Classic applications: Compression and Communication

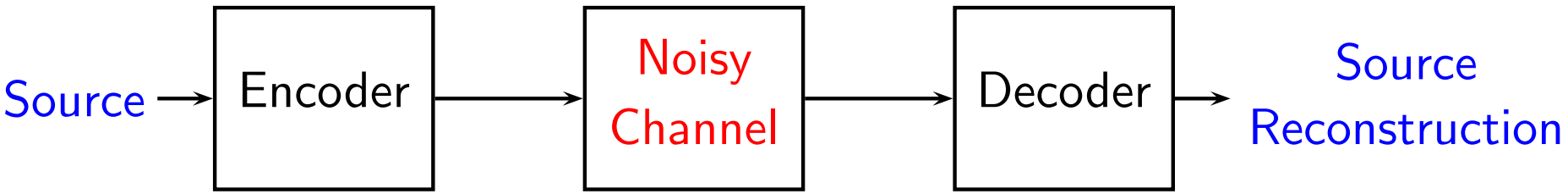
- Information Theory is the science of measuring of information.
- This science has had a profound impact on sensing, compression, storage, extraction, processing, and communication of information.
  - Compressing data such as audio, images, movies, text, sensor measurements, etc. (Example: We will see the principle behind the 'zip' compression algorithm in this course.)
  - Communicating data over noisy channels such as wires, wireless links, memory (e.g. hard disks), etc.
- Specifically, we will be interested in determining the **fundamental limits** of compression and communication. This will shed light on how to engineer near-optimal systems.
- We will use probability as a "language" to describe and derive these limits.

Classic applications: Compression and Communication

- Information Theory is the science of measuring of information.
- This science has had a profound impact on sensing, compression, storage, extraction, processing, and communication of information.
  - Compressing data such as audio, images, movies, text, sensor measurements, etc. (Example: We will see the principle behind the 'zip' compression algorithm in this course.)
  - Communicating data over noisy channels such as wires, wireless links, memory (e.g. hard disks), etc.
- Specifically, we will be interested in determining the **fundamental limits** of compression and communication. This will shed light on how to engineer near-optimal systems.
- We will use probability as a "language" to describe and derive these limits.
- Information Theory has strong connections to Statistics, Physics, Biology, Computer Science, and many other disciplines. Some of those connections/applications are the focus of this course.

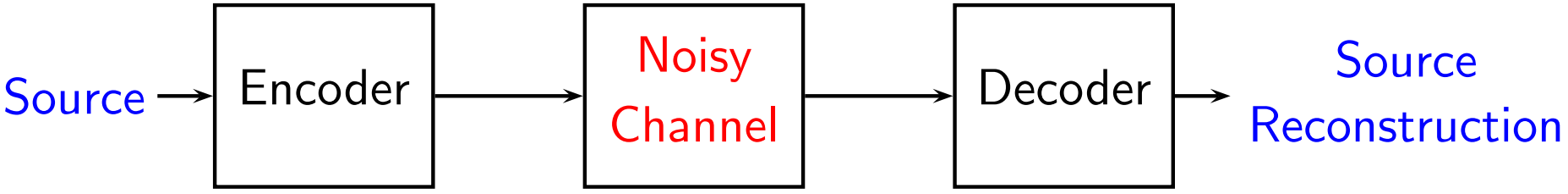
A General Communication Setting

- Information Source: Data we want to send (e.g. a movie).
- Noisy Channel: Communication medium (e.g. a wire).
- Encoder: Maps source into a channel codeword (signal).
- Decoder: Reconstructs source from channel output (signal).
- Fidelity Criterion: Measures quality of the source reconstruction.



A General Communication Setting

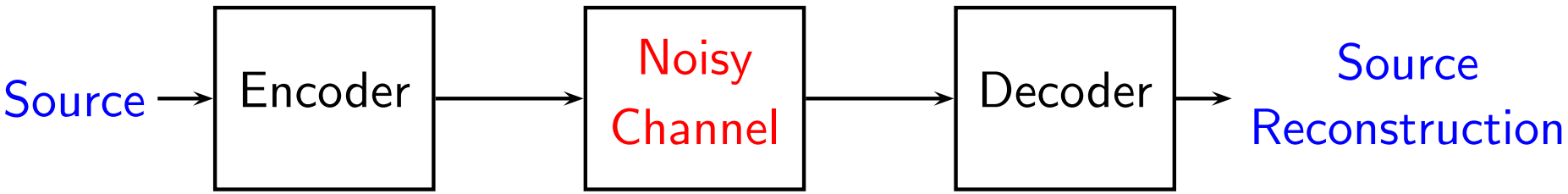
- Information Source: Data we want to send (e.g. a movie).
- Noisy Channel: Communication medium (e.g. a wire).
- Encoder: Maps source into a channel codeword (signal).
- Decoder: Reconstructs source from channel output (signal).
- Fidelity Criterion: Measures quality of the source reconstruction.



- **Goal:** Transmit at the highest rate possible while meeting the fidelity criterion.
- Example: Maximize frames/second while keeping mean-squared error below 1%.

A General Communication Setting

- Information Source: Data we want to send (e.g. a movie).
- Noisy Channel: Communication medium (e.g. a wire).
- Encoder: Maps source into a channel codeword (signal).
- Decoder: Reconstructs source from channel output (signal).
- Fidelity Criterion: Measures quality of the source reconstruction.



- **Goal:** Transmit at the highest rate possible while meeting the fidelity criterion.
- Example: Maximize frames/second while keeping mean-squared error below 1%.
- Look Ahead: We will see a theoretical justification for the layered protocol architecture of communication networks (combine optimal compression with optimal communication)

Bits: The currency of information for communication

- As we will see, bits are a "universal" currency of information for single sender, single receiver communication.
- Specifically, when we talk about sources, we often describe their size in bits. Example: A small JPEG is around 100kB.
- Also, when we talk about channels, we often mention what rate they can support. Example: A dial-up modem can send 14.4kB/sec.
- But this requires a sophisticated *source-channel separation* theorem to hold. That theorem holds for point-to-point communication, but breaks down for even small communication networks. It certainly breaks down for computing problems, and beyond. So, bits may not be the currency to express information in your problem.

Applications

The course is aimed to understand *informational measures*. In introductory information theory courses, these are classically examined only for compression and communication. This course, on the other hand, will focus on two other applications specifically:

- Information flow in neural circuits.
- Fairness, Accountability, Transparency, Explainability (FATE) in/of AI.

Applications

The course is aimed to understand *informational measures*. In introductory information theory courses, these are classically examined only for compression and communication. This course, on the other hand, will focus on two other applications specifically:

- Information flow in neural circuits.
- Fairness, Accountability, Transparency, Explainability (FATE) in/of AI.

There are many other applications where information theory is used, which we will a.s. not discuss in great detail:

- Large deviation theory.
- Deriving minimax lower bounds in statistics.
- Quantifying *relevance* of data features being sampled from different sensors.
- Distributed computing.

High Dimensions

- To compress and communicate data close to the fundamental limits, we will need to operate over long blocklengths.
- This is on its face an extremely complex problem: nearly impossible to "guess and check" solutions.
- Using probability, we will be able to reason about the existence (or non-existence) of good schemes. This will give us insight into how actually construct near-optimal schemes.
- Along the way, you will develop a lot of intuition for how high-dimensional random vectors behave.
- We will now review some basic elements of probability that we will need for the course.

Elements of a Probability Space $(\Omega, \mathcal{F}, \mathbb{P})$:

1. Sample space $\Omega = \{\omega_1, \omega_2, \ldots\}$ the set of possible outcomes.
2. Set of events $\mathcal{F} = \{E_1, E_2, \ldots\}$, where each event is a set of possible outcomes (from Ω). We say that the event E_i occurs if the outcome ω_j is an element of E_i.
3. Probability measure $\mathbb{P} : \mathcal{F} \to \mathbb{R}_+$, an assignments of probabilities to events, a function that satisfies
   - (i) $\mathbb{P}(\emptyset) = 0$.
   - (ii) $\mathbb{P}(\Omega) = 1$.
   - (iii) If $E_i \cap E_j = \emptyset$ (i.e. E_i and E_j are disjoint) for all $i \neq j$, then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i) \ .$$

Union of Events:

- $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2)$. (Venn Diagram)

Union of Events:

- $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2)$. (Venn Diagram)
- More generally, we have the inclusion-exclusion principle:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathbb{P}(E_i) - \sum_{i<j} \mathbb{P}(E_i \cap E_j) + \sum_{i<j<k} \mathbb{P}(E_i \cap E_j \cap E_k)$$

$$\cdots + (-1)^{\ell+1} \sum_{i_1<i_2<\cdots<i_\ell} \mathbb{P}\left(\bigcap_{m=1}^{\ell} E_{i_m}\right) + \cdots$$

Union of Events:

- $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2)$. (Venn Diagram)
- More generally, we have the inclusion-exclusion principle:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathbb{P}(E_i) - \sum_{i<j} \mathbb{P}(E_i \cap E_j) + \sum_{i<j<k} \mathbb{P}(E_i \cap E_j \cap E_k)$$

$$\cdots + (-1)^{\ell+1} \sum_{i_1<i_2<\cdots<i_\ell} \mathbb{P}\left(\bigcap_{m=1}^{\ell} E_{i_m}\right) + \cdots$$

- Very difficult to calculate, often rely on the union bound:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(E_i) \ .$$

Independence:

- Two events E_1 and E_2 are independent if

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2) \ .$$

Independence:

- Two events E_1 and E_2 are independent if

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2) \ .$$

- The events $E_1, \ldots, E_n$ are mutually independent (or just independent) if, for all subsets $\mathcal{I} \subseteq \{1, \ldots, n\}$,

$$\mathbb{P}\left(\bigcap_{i \in \mathcal{I}} E_i \right) = \prod_{i \in \mathcal{I}} \mathbb{P}(E_i) \ .$$

Conditional Probability:

- The conditional probability that event E_1 occurs given that event E_2 occurs is

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)} \ .$$

  Note that this is well-defined only if $\mathbb{P}(E_2) > 0$.

Conditional Probability:

- The conditional probability that event E_1 occurs given that event E_2 occurs is

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)} \ .$$

  Note that this is well-defined only if $\mathbb{P}(E_2) > 0$.

- Notice that if E_1 and E_2 are independent and $\mathbb{P}(E_2) > 0$,

$$\mathbb{P}(E_1|E_2) = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)} = \frac{\mathbb{P}(E_1)\mathbb{P}(E_2)}{\mathbb{P}(E_2)} = \mathbb{P}(E_1) \ .$$

Law of Total Probability:

- If $E_1, E_2, \ldots$ are disjoint events such that $\Omega = \bigcup_{i=1}^{\infty} E_i$, then for any event A

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap E_i) = \sum_{i=1}^{\infty} \mathbb{P}(A|E_i)\mathbb{P}(E_i) \ .$$

Bayes' Law:

- If $E_1, E_2, \ldots$ are disjoint events such that $\Omega = \bigcup_{i=1}^{\infty} E_i$, then for any event A

$$\mathbb{P}(E_j|A) = \frac{\mathbb{P}(A|E_j)\mathbb{P}(E_j)}{\sum_{i=1}^{\infty} \mathbb{P}(A|E_i)\mathbb{P}(E_i)} \ .$$

Probability Review: Random Variables

- A random variable X on a sample space Ω is a real-valued function, $X : \Omega \to \mathbb{R}$.
- Cumulative Distribution Function (cdf): $F_X(x) = \mathbb{P}(X \leq x)$.

Discrete Random Variables:

Probability Review: Random Variables

- A random variable X on a sample space Ω is a real-valued function, $X : \Omega \to \mathbb{R}$.
- Cumulative Distribution Function (cdf): $F_X(x) = \mathbb{P}(X \leq x)$.

Discrete Random Variables:

- X is discrete if it only takes values on a countable subset $\mathcal{X}$ of $\mathbb{R}$.
- Probability Mass Function (pmf): For discrete random variables, we can define the pmf $p_X(x) = \mathbb{P}(X = x)$.
- Example 1: Bernoulli with parameter q.

$$p_X(x) = \begin{cases} 1-q & x = 0 \\ q & x = 1 \end{cases}.$$

- Example 2: Binomial with parameters n and q.

$$p_X(k) = \binom{n}{k} q^k (1-q)^{n-k} \qquad k = 0, 1, \ldots, n$$

Continuous Random Variables:

- A random variable X is called continuous if there exists a nonnegative function $f_X(x)$ such that

$$\mathbb{P}(a < X \leq b) = \int_a^b f_X(x)dx \qquad \text{for all } -\infty < a < b < \infty \ .$$

  This function $f_X(x)$ is called the probability density function (pdf) of X.

- Example 1: Uniform with parameters a and b.

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

- Example 2: Gaussian with parameters μ and σ^2.

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Example 3: Exponential with parameter λ.

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

Expectation:

- Discrete rvs: $\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) p_X(x)$
- Continuous rvs: $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$

Special Cases

Mean:

- Discrete rvs: $\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p_X(x)$
- Continuous rvs: $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$

Expectation:

- Discrete rvs: $\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) p_X(x)$
- Continuous rvs: $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$

Special Cases

Mean:

- Discrete rvs: $\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p_X(x)$
- Continuous rvs: $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$
-

Bernoulli	Binomial	Uniform	Gaussian	Exponential
p	np	$\frac{a+b}{2}$	μ	$\frac{1}{\lambda}$

Special Cases Continued

m^{th} Moment:

- Discrete rvs: $\mathbb{E}[X^m] = \sum_{x \in \mathcal{X}} x^m p_X(x)$
- Continuous rvs: $\mathbb{E}[X^m] = \int_{-\infty}^{\infty} x^m f_X(x) dx$

Variance:

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

Special Cases Continued

m^{th} Moment:

- Discrete rvs: $\mathbb{E}[X^m] = \sum_{x \in \mathcal{X}} x^m p_X(x)$
- Continuous rvs: $\mathbb{E}[X^m] = \int_{-\infty}^{\infty} x^m f_X(x) dx$

Variance:

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

Bernoulli	Binomial	Uniform	Gaussian	Exponential
$p(1-p)$	$np(1-p)$	$\frac{(b-a)^2}{12}$	σ^2	$\frac{1}{\lambda^2}$

Pairs of Random Variables (X, Y):

- Joint cdf: $F_{XY}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$
- Joint pmf: $p_{XY}(x, y) = \mathbb{P}(X = x, Y = y)$ (for discrete rvs)
- Joint pdf: If f_{XY} satisfies

$$\mathbb{P}(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f_{XY}(x, y) dy dx$$

  for all $-\infty < a < b < \infty$ and $-\infty < c < d < \infty$ then f_{XY} is called the joint pdf of (X, Y). (for continuous rvs)

- Marginalization:

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

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx$$

n-tuples of Random Variables $(X_1, \ldots, X_n)$:

- Joint cdf: $F_{X_1 \cdots X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$
- Joint pmf: $p_{X_1 \cdots X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$
- Joint pdf: If $f_{X_1 \cdots X_n}$ satisfies

$$\begin{aligned} &\mathbb{P}(a_1 < X_1 \leq b_1, \ldots, a_n < X_n \leq b_n) \\ &= \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f_{X_1 \cdots X_n}(x_1, \ldots, x_n) dx_n \cdots dx_1 \end{aligned}$$

for all $-\infty < a_i < b_i < \infty$ then $f_{X_1 \cdots X_n}$ is called the joint pdf of $(X_1, \ldots, X_n)$.

Independence of Random Variables:

- $X_1, \ldots, X_n$ are independent if
  $F_{X_1 \cdots X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$
  $\forall x_1, x_2, \ldots, x_n$
- Equivalently, we can just check if

$$p_{X_1 \cdots X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n) \quad \text{(discrete rvs)}$$

$$f_{X_1 \cdots X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) \quad \text{(continuous rvs)}$$

Conditional Probability Densities:

- Given discrete rvs X and Y with joint pmf $p_{XY}(x, y)$, the conditional pmf of X given $Y = y$ is defined to be

$$p_{X|Y}(x|y) = \begin{cases} \frac{p_{XY}(x,y)}{p_Y(y)} & p_Y(y) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

- Given continous rvs X and Y with joint pdf $f_{XY}(x, y)$, the conditional pdf of X given $Y = y$ is defined to be

$$f_{X|Y}(x|y) = \begin{cases} \frac{f_{XY}(x,y)}{f_Y(y)} & f_Y(y) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

- Note that if X and Y are independent, then $p_{X|Y}(x|y) = p_X(x)$ or $f_{X|Y}(x|y) = f_X(x)$.

Linearity of Expectation:

- $\mathbb{E}[a_1 X_1 + \cdots + a_n X_n] = a_1 \mathbb{E}[X_1] + \cdots + a_n \mathbb{E}[X_n]$ even if the X_i are dependent.

Expectation of Products:

- If $X_1, \ldots, X_n$ are independent, then $\mathbb{E}[g_1(X_1) \cdots g_n(X_n)] = \mathbb{E}[g_1(X_1)] \cdots \mathbb{E}[g_n(X_n)]$ for any deterministic functions g_i.

Conditional Expectation:

- Discrete rvs: $\mathbb{E}[g(X)|Y=y] = \sum_{x \in \mathcal{X}} g(x) p_{X|Y}(x|y)$
- Continuous rvs: $\mathbb{E}[g(X)|Y=y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) dx$

Conditional Expectation:

- Discrete rvs: $\mathbb{E}[g(X)|Y=y] = \sum_{x \in \mathcal{X}} g(x) p_{X|Y}(x|y)$
- Continuous rvs: $\mathbb{E}[g(X)|Y=y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) dx$
- $\mathbb{E}[Y|X=x]$ is a number. This number can be interpreted as a function of x.
- $\mathbb{E}[Y|X]$ is a random variable. It is in fact a function of the random variable X. (Note: A function of a random variable is a random variable.)
- Lemma: $\mathbb{E}_X\big[\mathbb{E}[Y|X]\big] = \mathbb{E}[Y]$.

Conditional Independence:

- X and Y are conditionally independent given Z if

$$p_{XY|Z}(x, y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z) \quad \text{(discrete rvs)}$$

$$f_{XY|Z}(x, y|z) = f_{X|Z}(x|z)f_{Y|Z}(y|z) \quad \text{(continuous rvs)}$$

Markov Chains:

- Random variables X, Y, and Z are said to form a Markov chain $X \to Y \to Z$ if the conditional distribution of Z depends only on Y and is conditionally independent of X.

Markov Chains:

- Random variables X, Y, and Z are said to form a Markov chain $X \to Y \to Z$ if the conditional distribution of Z depends only on Y and is conditionally independent of X.
- Specifically, the joint pmf (or pdf) can be factored as

$$p_{XYZ}(x,y,z) = p_X(x)p_{Y|X}(y|x)p_{Z|Y}(z|y) \quad \text{(discrete rvs)}$$
$$f_{XYZ}(x,y,z) = f_X(x)f_{Y|X}(y|x)f_{Z|Y}(z|y) \quad \text{(continuous rvs)} \ .$$

Markov Chains:

- Random variables X, Y, and Z are said to form a Markov chain $X \to Y \to Z$ if the conditional distribution of Z depends only on Y and is conditionally independent of X.
- Specifically, the joint pmf (or pdf) can be factored as

$$p_{XYZ}(x,y,z) = p_X(x)p_{Y|X}(y|x)p_{Z|Y}(z|y) \quad \text{(discrete rvs)}$$
$$f_{XYZ}(x,y,z) = f_X(x)f_{Y|X}(y|x)f_{Z|Y}(z|y) \quad \text{(continuous rvs)} \ .$$

- $X \to Y \to Z$ if and only if X and Z are conditionally independent given Y.
- $X \to Y \to Z$ implies $Z \to Y \to X$ (and vice versa).
- If Z is a deterministic function of Y, i.e. $Z = g(Y)$, then $X \to Y \to Z$ automatically.

Convexity:

- A set $\mathcal{X} \subseteq \mathbb{R}^n$ is convex if, for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$, we have that $\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2 \in \mathcal{X}$.

Convexity:

- A set $\mathcal{X} \subseteq \mathbb{R}^n$ is convex if, for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$, we have that $\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2 \in \mathcal{X}$.
- A function g on a convex set $\mathcal{X}$ is convex if, for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$, we have that

$$g(\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2) \leq \lambda g(\mathbf{x}_1) + (1 - \lambda) g(\mathbf{x}_2) \ .$$

Convexity:

- A set $\mathcal{X} \subseteq \mathbb{R}^n$ is convex if, for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$, we have that $\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2 \in \mathcal{X}$.
- A function g on a convex set $\mathcal{X}$ is convex if, for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$, we have that

$$g(\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2) \leq \lambda g(\mathbf{x}_1) + (1 - \lambda) g(\mathbf{x}_2) \ .$$

- A function g is concave if $-g$ is convex.

Convexity:

- A set $\mathcal{X} \subseteq \mathbb{R}^n$ is convex if, for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$, we have that $\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2 \in \mathcal{X}$.
- A function g on a convex set $\mathcal{X}$ is convex if, for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$, we have that

$$g(\lambda \mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2) \leq \lambda g(\mathbf{x}_1) + (1 - \lambda) g(\mathbf{x}_2) \ .$$

- A function g is concave if $-g$ is convex.

Jensen's Inequality:

- If g is a convex function and X is a random variable, then

$$g\big(\mathbb{E}[X]\big) \leq \mathbb{E}\big[g(X)\big]$$

Markov's Inequality:

- Let X be a non-negative random variable. For any $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t} \ .$$

Markov's Inequality:

- Let X be a non-negative random variable. For any $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t} \ .$$

Chebyshev's Inequality:

- Let X be a random variable. For any $\epsilon > 0$,

$$\mathbb{P}\Big(\big|X - \mathbb{E}[X]\big| > \epsilon\Big) \leq \frac{\mathsf{Var}(X)}{\epsilon^2} \ .$$

Weak Law of Large Numbers (WLLN):

- Let X_i be a sequence of independent and identically distributed (i.i.d.) random variables with finite mean, $\mu = \mathbb{E}[X_i] < \infty$.
- Define the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- For any $\epsilon > 0$, the WLLN implies that

$$\lim_{n\to\infty} \mathbb{P}\Big(\big|\bar{X}_n - \mu\big| > \epsilon\Big) = 0 \ .$$

- That is, the sample mean converges (in probability) to the true mean.

Strong Law of Large Numbers (SLLN):

- Let X_i be a sequence of independent and identically distributed (i.i.d.) random variables with finite mean, $\mu = \mathbb{E}[X_i] < \infty$.
- Define the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- The SLLN implies that

$$\mathbb{P}\left(\left\{ \lim_{n\to\infty} \bar{X}_n = \mu \right\}\right) = 1 \ .$$

- That is, the sample mean converges (almost surely) to the true mean.