ECE 5630: Information Theory for Data Transmission, Security and Machine Learning 2/2/20

Lectures 1 and 2

Lecturer: Prof. Ziv Goldfeld

Scriber: Adeel Mahmood, *Net ID*: am2384

Assistant Editor: Kia Khezeli

Introduction

What is information theory?

- A mathematical framework for quantifying and rigorously reasoning about **uncertainty** and **information** (information is the resolution of uncertainty).
- It leverages this framework to study fundamental properties of **operations** one can perform on **information sources**.

Examples of information sources are images, text files, audio files, etc.

Examples of operations one can perform on information sources are:

- **compression:** Exploiting redundancy and structure to reduce the number of bits needed to store or transmit data.
- **transmission:** Converting source information to codewords for transmission over a communication channel (source coding and channel coding).
- **encryption:** Encoding information to prevent unauthorized access.

List of Topics

(1) Background on probability theory

(2) f-divergence

Divergence is a function capable of measuring the distance/proximity between probability distributions.

Let $\mathcal{X}$ be a space. Let $\mathcal{P}(\mathcal{X})$ be the set of all probability measures on $\mathcal{X}$. Then, f-divergence is a mapping $D_f : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \mapsto \mathbb{R}_+$. In other words, f-divergence, which is induced by some convex $f : (0, \infty) \mapsto \mathbb{R}$, is a function that takes two probability measures as input and outputs a nonnegative real number. Depending on the choice of the function f, f-divergence may or may not be a distance metric. However, any divergence measure D does satisfy $D(P\|Q) = 0 \iff P = Q$ for all $P, Q \in \mathcal{P}(\mathcal{X})$.

Other divergence measures include: KL divergence, Total variation (TV) distance, χ^2 distance, etc.

(3) Information Measures

Examples of information measures:

- Shannon Entropy
- Differential Entropy
- Mutual Information

(4) Letter Typical Sequences

Let $\mathcal{X} = \{0, 1\}$ and define $\mathcal{X}^n = \underbrace{\mathcal{X} \times \mathcal{X} \times \cdots \times \mathcal{X}}_{\text{n times}}$. Alternatively, $\mathcal{X}^n$ is the set of all binary sequences of length n. Size of $\mathcal{X}^n$ is $|\mathcal{X}^n| = 2^n$.

Let $\mathcal{P}(\{0, 1\})$ be the set of all probability measures on $\mathcal{X}$. Let's fix $\underline{p} \in \mathcal{P}(\{0, 1\})$ as $\underline{p} =$ Bern(p) where $p \in \left(0, \frac{1}{2}\right)$. p is fixed.

Let $Y^n \sim \underline{p}^{\otimes n}$ be a random vector of length n. The notation means that the components $Y_1^n, Y_2^n, ..., Y_n^n$ of the random vector Y^n are an i.i.d sequence of random variables with distribution $\underline{p}$.

We denote by $\tau_n(\underline{p}) \subseteq \mathcal{X}^n$, the *letter typical set* associated with distribution $\underline{p}^{\otimes n}$. Qualitatively, the letter typical set is the set of equiprobable outcomes such that the set is small in size and occurs with high probability. More precisely, it satisfies the following properties.

- $\tau_n(\underline{p})$ is very small in size compared to the sample set (the set of all possible outcomes), i.e.,

$$\frac{|\tau_n(\underline{p})|}{|\mathcal{X}^n|} \to 0 \text{ as } n \to \infty$$

- $\mathbb{P}\left(Y^n \in \tau_n(\underline{p})\right) \to 1$ as $n \to \infty$ where $Y^n \sim \underline{p}^{\otimes n}$ as before.
- $\forall\ y^n \in \tau_n(\underline{p})$, we have:

$$\mathbb{P}\left(Y^n = y^n\right) \approx \frac{1}{|\tau_n(\underline{p})|}$$

Such a set $\tau_n(\underline{p})$ is called a typical set.

(5) Reliable Communication over noisy channels

Discuss operational setup, Shannon's channel coding theorem (1948).

Proof (direct: using typical sequences and converse using properties of information measures).

(6) Distribution Simulation

- **Exact Simulation:** Given some $\underline{p} \in \mathcal{P}(\mathcal{X})$ (some probability distribution on space $\mathcal{X}$).

  Suppose we have i.i.d samples $y_1, y_2, ...$ from Bernoulli distribution (with $p = \frac{1}{2}$).

  Design an algorithm $\mathcal{A}$ which takes $y_1, y_2, ...$ as inputs and generates new samples $z_1, z_2, ...$ such that $z_i \sim \underline{p}$. This is called distribution simulation.

- **Approximate Simulation:** Suppose the target distribution is again $\underline{p}$ and the algorithm $\mathcal{A}$ generates samples from the distribution $Q_{\mathcal{A}}$. Let D be some divergence measure. The essence of approximate simulation is that $D\left(Q_{\mathcal{A}}, \underline{p}\right) \to 0$ as the number of input samples to the algorithm $\mathcal{A}$ goes to infinity.

(7) Information-theoretic security

- Shannon's cipher system
- Wiretap Channel (Wyner 1975)
- Active Wiretap Channel.

(8) Information Theory and Machine Learning

- **IT → ML:** design ML algorithms for learning useful representations via information bottleneck methods (e.g. deep variational information bottleneck framework).
- **ML → IT:** use neural networks and stochastic gradient descent to estimate mutual information (also called MINE: Mutual Information Neural Estimation).

1 Background on Probability Theory

We start by defining probability spaces.

Definition 1.1 (Probability space) *A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where:*

1. *Ω is an arbitrary nonempty set, called the* sample set*.*
2. *$\mathcal{F}$ is a collection of subsets of Ω called the σ-algebra, which satisfies:*
   - *(i)* $\Omega \in \mathcal{F}$
   - *(ii)* $A \in \mathcal{F} \implies A^c := \Omega \setminus A \in \mathcal{F}$ *(closed under complements).*
   - *(iii)* $A_1, A_2, ... \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ *(closed under countable unions).*
3. *$\mathbb{P} : \mathcal{F} \to [0,1]$ is the probability measure, satisfying:*
   - *(i)* $\mathbb{P}(\Omega) = 1$
   - *(ii) σ-additivity: let $A_1, A_2, ... \in \mathcal{F}$ be disjoint (i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$), then*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

Remark 1.1 (Interpretation) *Ω is understood as the set of all possible outcomes of a random experiment. The elements of the set Ω can be arbitrary, e.g., numbers, functions, symbols, etc. The σ-algebra $\mathcal{F}$ can be thought of as the collection of questions one can ask about the experiment's outcomes. Questions are of the form 'what is the probability that a certain event happens?'. The probability measure $\mathbb{P}$ is interpreted as the 'answers' to those questions.*

Example 1.1 (Dice) *Consider the random experiment of drawing a dice, i.e., $\Omega = \{1,2,3,4,5,6\}$. Many σ-algebras can be defined on Ω. For instance $\mathcal{F} = \{\Omega, \emptyset, \{1,3\}, \{2,4,5,6\}\}$ is a valid choice. However, when Ω is countable (or finite, as in this example), the most common σ-algebra is $\mathcal{F} = 2^{\Omega}$, where 2^{Ω}. is the power set of Ω. To endow $(\Omega, \mathcal{F})$ with a probability measure, we may set $\mathbb{P}(A) := \frac{|A|}{|\Omega|}$, for all $A \in \mathcal{F}$. This $\mathbb{P}$ is called the uniform measure on Ω and corresponds to a fair dice.*

There are two instances of probability spaces that we will mostly focus on, discrete and continuous probability spaces.

1.1 Discrete Probability Space

- Sample set: The sample set Ω is at most countable (finite or countably infinite).
- σ-algebra: The common σ-algebra in this case is $\mathcal{F} = 2^{\Omega}$. This is also the largest possible σ-algebra.

- Probability measure: Over discrete probability space, $\mathbb{P}$ can be built from a simpler function, called the Probability Mass Function (PMF). Let $p : \Omega \mapsto [0,1]$ be a PMF on Ω, i.e., a function satisfying $\sum_{\omega\in\Omega} p(\omega) = 1$. The probability measure $\mathbb{P}_p$ induced by p is defined as

$$\mathbb{P}_p(A) = \sum_{\omega\in A} p(\omega), \quad \forall A \in \mathcal{F}.$$

Exercise 1.1 *Verify that $\mathbb{P}_p$ is a valid probability measure.*

In summary, given a countable Ω and a PMF p, we will always consider the discrete probability space $\big(\Omega, 2^\Omega, \mathbb{P}_p\big)$. Note that this triple is completely specified by Ω and p.

1.2 Continuous Probability Space

- Sample set: The sample set Ω is uncountably infinite. We will exclusively consider the case $\Omega = \mathbb{R}^d$.

- σ-algebra: Usually, we want the largest possible σ-algebra. So ideally, $\mathcal{F} = 2^{\mathbb{R}^d}$. However, Vitali's theorem (a fundamental result in measure theory), states that there does not exist a non-trivial translation invariant σ-additive measure from $2^{\mathbb{R}^d}$ to the extended reals. More specifically, there exist non-measurable subsets of $\mathbb{R}^d$ that preclude the existence of a σ-additive measure μ satisfying $\mu\big([0,1]^d\big) = 1$ and $\mu(A) = \mu(x + A)$, for all $x \in \mathbb{R}^d$ and $A \subseteq \mathbb{R}^d$, where $x + A = \{x + y : \ y \in A\}$.

  Therefore, we need a smaller σ-algebra which excludes such pathological sets but still includes all interesting subsets of $\mathbb{R}^d$. To this end we will adopt the **Borel σ-algebra**, denoted by $\mathcal{B}(\mathbb{R}^d)$. We next give a high-level description of its construction. The idea is to identify a well-chosen collection of subsets of $\mathbb{R}^d$ (called the 'generating set') and use them to generate the σ-algebra.

Theorem 1 (Generated σ-algebra) *For an arbitrary collection $\mathcal{C}$ of subsets of Ω, there exists a unique, smallest σ-algebra, denoted by $\sigma(\mathcal{C})$ and called the σ-algebra* generated by *$\mathcal{C}$, that contains every element of $\mathcal{C}$. That is, $\sigma(\mathcal{C})$ is the unique σ-algebra such that if $\mathcal{H}$ is another σ-algebra that satisfies $\mathcal{C} \subseteq \mathcal{H}$, then $\sigma(\mathcal{C}) \subseteq \mathcal{H}$.*

Proof: Let $\{\mathcal{F}_i\}_{i\in I}$ be the collection of all σ-algebras containing $\mathcal{C}$. Note that the set is not empty since the 2^Ω is always in it. Define

$$\sigma(\mathcal{C}) := \bigcap_{i\in I} \mathcal{F}_i.$$

We need to shown that $\sigma(\mathcal{C})$ is the unique σ-algebra that contains $\mathcal{C}$, and that it is the smallest one. For the former, we rely on Exercise 1.2. Next, note that $\sigma(\mathcal{C})$ trivially contains $\mathcal{C}$ since every $\mathcal{F}_i$ contains $\mathcal{C}$. For smallness, if $\mathcal{H}$ is any other σ-algebra containing $\mathcal{C}$, then $\mathcal{H} = \mathcal{F}_i$, for some $i \in I$. Therefore, $\sigma(\mathcal{C}) \subseteq \mathcal{H}$. Finally, uniqueness follows by smallness, since if both $\mathcal{H}$ and $\mathcal{G}$ are smallest, the $\mathcal{H} \subseteq \mathcal{G}$ and $\mathcal{G} \subseteq \mathcal{H}$, which implies set equality. ■

Exercise 1.2 *Show that an arbitrary intersections of σ-algebras is a σ-algebra.*

Example 1.2 (σ-algebra generated by a single set) *The σ-algebra $\sigma(\mathcal{A})$ is also given by the collection of all countable union/intersections/complements of the elements of $\mathcal{A}$. For instance, for a single set $A \subseteq \Omega$, $\sigma(\{A\}) = \{\emptyset, \Omega, A, A^c\}$.*

Definition 1.2 (Borel σ-algebra) *Let $\mathcal{C}_0 := \big\{(-\infty, a_1] \times \cdots \times (-\infty, a_d] : \ a_1, \ldots, a_d \in \mathbb{R}\big\}$. The Borel σ-algebra, denoted by $\mathcal{B}(\mathbb{R})$, is the σ-algebra generated by $\mathcal{C}_0$, i.e., $\mathcal{B}(\mathbb{R}^d) := \sigma(\mathcal{C}_0)$.*

- Probability measure: Over continuous probability space, $\mathbb{P}$ can be built from an Lebesgue integrable function, called the probability density function (PDF). Let $f : \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$ be such that $\int_{\mathbb{R}^d} f(\mathbf{x})d\mathbf{x} = 1$. The probability measure $\mathbb{P}_f$ induced by f is

$$\mathbb{P}_f(B) = \int_B f(\mathbf{x})d\mathbf{x}, \quad \forall B \in \mathcal{B}(\mathbb{R}^d).$$

In summary, when $\Omega = \mathbb{R}^d$, we always choose the σ-algebra as $\mathcal{F} = \mathcal{B}(\mathbb{R}^d)$. Given a PDF f, we then consider the probability space $\left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P}_f\right)$

Remark 1.2 (Notation) *We use $\mathcal{P}(\Omega)$ to denote the set of all probability measures over Ω.*

- *if Ω is countable, then $\mathcal{F} = 2^\Omega$.*
- *if $\Omega = \mathbb{R}^d$, then $\mathcal{F} = \mathcal{B}(\mathbb{R}^d)$.*

1.3 Properties of Probability Measure

The following are several important properties of probability measures.

Proposition 1.1 (Properties of probability measure) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.*

1. $\mathbb{P}(\emptyset) = 1 - \mathbb{P}(\Omega) = 0$
2. *Law of complement probability:* $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ *for all $A \in \mathcal{F}$.*
3. *Monotonocity: For all $A, B \in \mathcal{F}$, $A \subseteq B$, we have $\mathbb{P}(A) \leq \mathbb{P}(B)$*
4. *Union bound:*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

   for all $A_1, A_2, ... \in \mathcal{F}$.

5. *Continuity of probability measure: Let $A_1 \subseteq A_2 \subseteq \cdots$ be a sequence of increasing events converging to $A \in \mathcal{F}$ in the sense that $A = \cup_{n=1}^{\infty} A_n$, then*

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$$

   Same for decreasing sequence.

1.4 Conditional Probability Spaces

Given a probability space, one may generate multiple conditional probability spaces by different conditional probability measures.

Definition 1.3 (Conditional probability measure) *Given $(\Omega, \mathcal{F}, \mathbb{P})$ and some $A \in \mathcal{F}$ such that $\mathbb{P}(A) > 0$, define $\mathbb{P}(\cdot|A) : \mathcal{F} \to [0, 1]$ by*

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \quad B \in \mathcal{F}.$$

Exercise 1.3 *Show that $(\Omega, \mathcal{F}, \mathbb{P}(.|A))$ is also a probability space.*

The above definition allows generating multiple (conditional) probability space for a given space $(\Omega, \mathcal{F}, \mathbb{P})$. What is the relation between $\mathbb{P}(.|A)$ and $\mathbb{P}(.|B)$, for some $A, B \in \mathcal{F}$ such that $\mathbb{P}(A), \mathbb{P}(B) > 0$? The relation is encoded in the Bayes theorem.

Theorem 2 (Bayes Theorem) *Let $A, B \in \mathcal{F}$ such that $\mathbb{P}(A), \mathbb{P}(B) > 0$. Then*

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

Definition 1.4 *Let Ω be a set, and $\{B_n\}$ a countable collection of subsets of Ω. We say the sets B_n form a partition of Ω if and only if the following conditions hold:*

1. *(Pairwise disjoint) $B_i \cap B_j = \varnothing$ for all i, j such that $i \neq j$.*
2. *(Cover) $\bigcup_n B_n = \Omega$*

Proposition 1.2 (Law of Total Probability) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $B_1, B_2, B_3, \ldots, \in \mathcal{F}$ form a partition of Ω. Then, for all $A \in \mathcal{F}$,*

$$\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(B_n)\mathbb{P}(A|B_n)$$

Exercise 1.4 *Prove this.*