ECE 598 - Fall 2025

Lecture 1: Course Overview

August 26, 2025

Instructor: Shubhanshu Shekhar

In this opening lecture, we begin with short informal introductions to the fields of information theory and statistical inference. We then discuss the tentative structure of the course, and then conclude with a discussion of some "guessing games" that motivate the definitions of the three key information measures that we will study formally in the next lecture.

1 Short Introduction to Information Theory

The field of information theory was initiated by Shannon (1948) as a probabilistic framework to model and analyze the fundamental limits of communication engineering systems. At its core, it involves two main problems:

- Source Coding(or data compression): how to represent data using as few bits as possible.
- Channel Coding (or reliable transmission): how to add structured redundancy so that data can be tranmitted reliably over a noisy medium.

Interestingly, these two problems are duals of each other: source coding aims to remove all redundancy from data, while channel coding adds redundancy to protect against the noise injected by the channel.

Source Coding or Data Compression. In compression, we want to encode the outputs of a source (e.g., an English document) into bitstrings of minimal lengths, so that the originial symbols can be recovered from these bitstrings (decoder) with zero or vanishing probability of error. For example, a zero-error variable-rate source coding system for an $\mathcal{X}^n$-valued random variable $X^n \overset{i.i.d.}{\sim} P_X$, consists of an encoder-decoder pair (e, f), such that with $\{0,1\}^* := \cup_{k=0}^{\infty}\{0,1\}^k$:

$$\underbrace{e : \mathcal{X}^n \to \{0,1\}^*}_{\text{encoder}}, \quad \underbrace{f : \{0,1\}^* \to \mathcal{X}^n}_{\text{decoder}}, \quad \text{and} \quad \underbrace{\mathbb{P}_{X^n \sim P_X^n}\left(f \circ e(X^n) \neq X^n\right) = 0}_{\text{zero-error property}}.$$

Here $\mathcal{X}$ could be the English alphabet, and n could represent the number of letters in the document. The key attribute of such a coding system is its *rate*, defined as

$$R(n, e, f, P_X) = \frac{\mathbb{E}_{X^n \sim P_X^n}\left[|e(X^n)|\right]}{n} \equiv \text{ average \# of bits per symbol},$$

and a fundamental result of Shannon was to prove that the optimal achievable rate in lossless compression (with i.i.d. sources) is characterized by a quantity called the entropy: $R^* = H(X) =$

$\sum_{x\in\mathcal{X}} -p_X(x)\log p_X(x)$. This optimal rate is achieved by codes that *assign longer codewords to rare symbols*, and in fact the optimal codeword lengths are

$$|e^*(x)| \approx \lceil \log 1/p_X(x) \rceil, \quad \text{for all } x \in \mathcal{X}.$$

There are several generalizations to this simple formulation: lossy compression, non-iid sources, universal compression etc.

Channel Coding or Data Transmission. The second important problem is that of transmitting messages over a noisy medium called the channel, which can be represented as follows:

$$\underbrace{W_n \in [M_n]}_{\text{Message}} \xrightarrow{\text{encoder}} \underbrace{X^n}_{\text{channel input}} \xrightarrow{\text{noisy channel}} \underbrace{Y^n}_{\text{channel output}} \xrightarrow{\text{decoder}} \underbrace{\hat{W}_n}_{\text{decoded message}}$$

- The messages are simply referenced by their index from 1 to M_n (where M_n is the total number of messages). The meanings associated with these messages are irrelevant, and they could refer to documents, images, videos, or any collection of objects agreed upon by the sender and receiver.
- The encoder assigns to each message a channel input of length n: so$\{x^n[1], \ldots, x^n[M]\} \subset \{0,1\}^n$ forms the codebook which is again agreed upon by both the sender and receiver.
- The encoded message is then transmitted over a noisy channel which introduces some noise to the transmitted codeword. Formally, a channel can be thought of as a stochastic kernel or transformation, and in the simplest case of a so-called discrete memoryless channel (DMC), we can represent it by a transition probability matrix $P_{Y|X}$, so that $\mathbb{P}(Y^n = y^n | X^n = x^n) = \prod_{i=1}^n P_{Y|X}(y_i|x_i)$.
- The decoder then observes Y^n, and makes its best guess about the transmitted message: $\hat{W}_n = f(Y^n)$.

There are two important properties associated with the above problem:

$$\text{Rate} \equiv R_n = \frac{\log M_n}{n}, \quad \text{and} \quad p_n(e,f) = \mathbb{P}\left(\hat{W}_n \neq W_n\right).$$

For a fixed blocklength n, increasing the rate M_n will lead to an increase in the probability of error p_n. A rate $R > 0$ is achievable, if there exist a sequence of channel codes $(e_n, f_n)_{n\geq 1}$, such that

$$\liminf_{n\to\infty} \frac{\log M_n}{n} \geq R, \quad \text{and} \quad \lim_{n\to\infty} p_n(e_n, f_n) = 0.$$

Shannon's channel coding theorem established that the largest achievable rate with vanishing error probability is a quantity called channel capacity, which for the case of DMC is

$$C \equiv C(P_{Y|X}) = \sup_{P_X} H(Y) - H(Y|X) = \sup_{P_X} I(X;Y).$$

Remark 1.1. Almost all important information theoretic results consist of two parts: the *achievability part* that involves showing the existence of good schemes (for example, "for every $R < C$, there exists a coding scheme with rate R and vanishing p_n"), and a *converse part* that requires us to show the impossibility of certain performance criteria ("any coding scheme with vanishing p_n cannot have rate larger than C").

2 Short Introduction to Statistical Inference

The general goal in statistical inference is to look at some data, and use it to learn or infer some property of the source that generated the data.

$$\underbrace{\{P_\theta : \theta \in \Theta\}}_{\text{family of dists.}} \longrightarrow P_\theta \longrightarrow \underbrace{(X_1, \ldots, X_n) \overset{i.i.d.}{\sim} P_\theta}_{\text{data}} \longrightarrow \text{Inference about } \theta.$$

Depending on the assumptions made on the parameter set Θ, the problem can either be parametric (finite dimensional Θ) or nonparametric (usually infinite dimensional Θ). Broadly speaking, statistical inference involves three fundamental tasks, and we will informally illustrate these via examples below.

Point estimation. The goal in point estimation is to construct our best guess of the true parameter θ, based on the observation X^n. Thus, it can be represented by a mapping $\widehat{\theta}_n : \mathcal{X}^n \to \Theta$ (more generally, we can allow for randomized rules in which case the estimator will be a stochastic kernel).

For example, if $\Theta = \mathbb{R}$, $P_\theta = N(\theta, 1)$, and suppose $X^n = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} P_\theta$, then, a natural estimate of θ given X^n is the sample mean

$$\widehat{\theta}(X^n) = \frac{1}{n} \sum_{i=1}^{n} X_i =: \bar{X}_n.$$

The quality of the estimator $\widehat{\theta}$ is often measured by its expected squared error (or mean squared error): $\mathbb{E}_{X^n \overset{i.i.d.}{\sim} P_\theta}[|\widehat{\theta}(X^n) - \theta|^2] = 1/n$. Later on in the course, we will study the optimality of such procedures: is $1/n$ the best rate achievable or can their exist methods with better rates? Such questions will naturally lead to our discussion of information-theoretic tools for proving impossibility results.

Confidence Sets. For a parameter $\alpha \in (0, 1)$, a level-$(1-\alpha)$ confidence set for a parameter θ is a set $C_n \equiv C_n(X^n, \alpha)$ constructed using only the data X^n and the parameter α, such that

$$\mathbb{P}_{X^n \overset{i.i.d.}{\sim} P_\theta}(\theta \in C_n(X^n, \alpha)) \geq 1 - \alpha, \quad \text{for all} \quad \theta \in \Theta.$$

In words, the above statement says that the random set C_n contains the true (fixed but unknown) parameter θ with probability (due to the randomness associated with X^n) at least $(1-\alpha)$.

Continuing with our Gaussian example, since $\widehat{\theta}_n(X^n) \sim N(\theta, 1/n)$, we can construct a level-$(1-\alpha)$ CI for θ as follows:

$$C_n(X^n, \alpha) = \left[\bar{X}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}},\ \bar{X}_n + \frac{z_{1-\alpha/2}}{\sqrt{n}}\right],$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$-quantile of the $N(0, 1)$ distribution.

Hypothesis Testing. Here, the goal is to ask if the unknown parameter θ satisfies certain properties or not, instead of learning the entire parameter. Thus, if $\Theta_0, \Theta_1 \subset \Theta$ denote two disjoint subsets of the parameter space Θ, given $X^n \overset{i.i.d.}{\sim} P_\theta$, we may be interested in testing

$$H_0 : \theta \in \Theta_0, \quad \text{versus} \quad H_1 : \theta \in \Theta_1.$$

So, here the decision rule is a mapping $\Psi : \mathcal{X}^n \to \{0, 1\}$ or more generally $\Psi : \mathcal{X}^n \to [0, 1]$ (to allow for randomization). The performance of this test is measured by its errors (e_I, e_{II}), where $e_I = \mathbb{P}_{H_0}(\Psi(X^n) = 1)$ is the type-I error, and $e_{II} = \mathbb{P}_{H_1}(\Psi(X^n) = 0)$ is the type-II error.

A simple example in the Gaussian case is if $\Theta_0 = \{-1\}$ and $\Theta_1 = \{1\}$. In this case, a natural idea might be to reject H_0 if the average $\bar{X}_n > 0$: that is, $\Psi(X^n) = \mathbf{1}_{\bar{X}_n > 0}$.

Remark 2.1. In our discussion in this section, we have assumed throughout that the parameter of interest is a **fixed but unknown** quantity. An alternative Bayesian approach is to treat all quantities as random variables, and use Bayes rule to update our beliefs associated with the parameters. We will discuss this approach briefly in Lecture 5.

3 An Overview of the Course

Information Measures and their properties. We will begin the course by studying the three key information measures: entropy, relative entropy (KL divergence), and mutual information. We will introduce these definitions in the technically simple setting of discrete distributions (often with finite support), and prove some important properties: such as Gibbs inequality, convexity/concavity, chain rules, data processing inequalities. These results will form of the core of our toolbox for the rest of the course and will be used repeatedly. We will illustrate the power of these results by studying the problem of generating fair random coin tosses from a sequence of biased coin tosses with unknown bias.

We will also see how the definitions of relative entropy and mutual information (but not entropy) for discrete distributions are essentially without loss of generality, by introducing some variational representations for general alphabets. To conclude this module, we will introduce the notion of f-divergences that generalizes relative entropy, and discuss some of their properties.

(4 Lectures)

Statistical Decision Theory. For the next couple of lectures, we will introduce the necessary formalism to study statistical inference problems in a unified manner. In particular, we will introduce the decision theoretic framework of Wald to study inference tasks such as hypothesis testing and estimation. To compare the quality of different statistical procedures, we will introduce the minimax (or worst case) and Bayesian (or average case) frameworks. These two frameworks can be related to each other via standard convex duality techniques. Finally, we will also look at a sequential decision-making formalism, that can be used to analyze problems such as bandits, reinforcement learning, active learning etc.

We will discuss several instantiations of the general framework including density estimation, nonparametric regression, K-armed bandits, reinforcement learning, and active learning problems and design and analyze certain procedures for some of them.

(2 Lectures)

Information Theoretic Techniques for Proving Optimality of Statistical Procedures. Having constructed some statistical procedures and derived bounds on their minimax and Bayesian risks, a natural question to ask is if we can do better? Information theoretic techniques play an important role here, by establishing impossibility or converse results. In this module, we will take a detailed look at several classes of techniques for proving such lower bounds or converse results.

We will first begin by presenting a general technique that relates the minimax risk of estimation (or more general decision problems) to the probability of error in an M-ary hypothesis testing problem. Then, we consider the simplest case of $M = 2$, which is often referred to as LeCam's method, and look at some applications. The main drawback of this method is that it often does not capture the right dimensional dependence, and this motivates us to look at the general $M > 2$ version of this problem. There are two popular techniques for $M > 2$: constructing problems indexed by the Hamming cube (or Assouad's method) and more general packing set construction (local Fano's method). We will explore the applications and drawbacks of both these techniques, and then go to a technique of Yang-Barron that relies on the metric entropy of the parameter set.

(6 Lectures)

Data Compression, Gambling, and Portfolio Optimization. In the next module, we will look at the problem of data compression, and see how its ideas naturally extend to the case of gambling and portfolio optimization. With some elementary examples, we will see how all these three problems can reduce to the "probability assignment" or "prediction with log loss" problem.

With that background, we proceed to the universal versions of these problems. In particular, we look at two different formulations: stochastic and individual sequence prediction. In each case, we will develop some general algorithmic (or universal compression) schemes and establish their optimality.

(4 Lectures)

From Universal Information Theory to Sequential Inference The universal prediction/compression schemes serve as a natural bridge to the field of sequential anytime-valid inference. We will begin by introducing the philosophical idea of "testing by betting", that will serve as the general guiding principle for designing powerful sequential testing, estimation, and change detection schemes. In short, this principle allows us to translate sequential inference problems into a repeated betting/portfolio optimization games, wherein we can leverage the tools from the previous module naturally.

We will begin by focusing on the simplest case of testing the means of bounded random variables, and discuss a general methodology based on universal portfolio algorithm of Cover. We will extend the same idea to other testing problems, such as two-sample testing, independence testing, conditional independence testing etc., and then to other problems like constructing confidence sequences, and change detection schemes.

(6 Lectures)

Information Projections and Large Deviations Theory. In all the problems in the previous module, we will observe that the optimal procedures will be characterized by certain reverse information projection terms. We will focus on this aspect in this next module, and in particular look at the definitions and properties of forward and reverse information projections. The forward information projection terms arise naturally in the theory of large deviations, and we will look at some elementary results in that area, while the reverse information projection arises in bandit theory and sequential inference.

(4 Lectures)

Machine Learning and Stochastic Optimization Applications. We will conclude the course by looking at some applications of information theoretic techniques in machine learning and stochastic optimization. In particular, we will study how the variational definitions of relative entropy result in a natural approach for proving concentration inequalities and generalization bounds in statistical learning theory. Then, we will conclude with a discussion of the information theoretic analysis techniques in bandits and RL theory.

($\leq$ 3 Lectures)

4 Guessing Games

We consider a collection of simple guessing games (or a 20 questions games) that will motivate the definitions of the three main information measures: entropy, relative entropy, and mutual information.

Question 4.1 (Guessing Game I)**.** *In the simplest setting, suppose there are n identical bins, and a ball is placed uniformly at random in one of the bins. Denote the position of the ball with the random variable $X \sim$ Uniform $(\mathcal{X})$, where $\mathcal{X} = \{0, 1, \ldots, n-1\}$. Suppose, we can make binary queries of the form:* "Is X in the set A?" *for $A \subset \mathcal{X}$. Our objective is to design a strategy of asking a series of such questions, such that the average number of questions required to identify the bin containing the ball (i.e., the value of the random variable X) is small.*

A simple strategy could be to ask the series of questions: is $X = 0$, is $X = 1$, and so on. With this strategy, we can check that the average number of questions needed are equal to

$$L = \sum_{i=0}^{n-1} \mathbb{P}(X = i)(i+1) = \frac{1}{n} \sum_{i=0}^{n-1} (i+1) = \frac{1}{n} (1 + 2 + \ldots + n) = \frac{n}{2}.$$

Thus, the number of yes/no questions required by this strategy (called the linear search) is $\Omega(n)$. This strategy is quite inefficient, since with every query we reduce the search space by one. As we see next, we can do significantly better.

An optimal strategy for the above problem is the *binary search*, in which the queries are specifically designed to reduce the size of search space by half with each query. In particular, consider the case of $n = 4$. Then, the binary search decision tree is shown in Figure 1. Assigning the values $0 \leftarrow Y$ and $1 \leftarrow N$, we see that the series of questions to ascertain a value of $X = i$ is equivalent to the binary encoding or representation of i.

$$0 \equiv (YY) \equiv (00), \quad 1 \equiv (YN) \equiv (01), \quad 2 \equiv (NY) \equiv (10), \quad 3 \equiv (NN) \equiv (11).$$

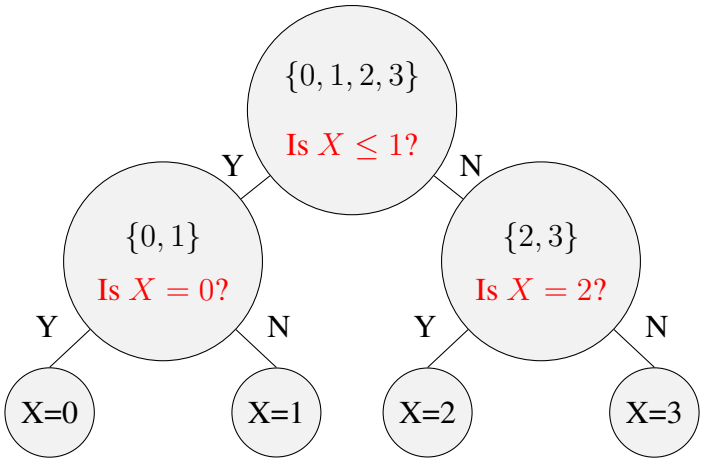

Figure 1: The figure shows the decision tree for the optimal strategy (i.e., binary search) for the guessing game with X drawn uniformly at random from $\mathcal{X} = \{0, 1, 2, 3\}$. Each query is chosen to ensures that the outcomes (i.e., Y or N) are equiprobable. In other words, this strategy proceeds by greedily selecting a query whose outcome is the most uncertain.

Denote the number of questions needed to verify $X = i$ by ℓ_i. Then, the binary search scheme asks $\ell_i = 2$ questions for all $i \in \mathcal{X}$ (in other words, it represents each $i \in \mathcal{X}$ with a *codeword* of length $\ell_i = 2$). Interestingly, 2 is also equal to the negative of the logarithm of the probability assigned to each value $i \in \mathcal{X}$ to the base 2; that is, $2 = \log(1/p_i) = \log(4)$. The average number of questions required by this (optimal) strategy is then equal to

$$L = 2 = \sum_{i=0}^{3} p_i \ell_i = \sum_{i=0}^{3} p_i \log(1/p_i).$$

Thus, the above discussion suggests that the number of binary questions needed to completely remove the uncertainty about the value of $X \sim \text{Uniform}(\mathcal{X})$ is $\log(|\mathcal{X}|)$. What is the analog of this quantity for non-uniform distribution over $\mathcal{X}$? We consider this question in the next version of the guessing game.

Question 4.2 (Guessing game II)**.** *Consider the same setting as Question 4.1 with $\mathcal{X} = \{0, 1, 2, 3\}$, but assume that X is drawn from the following distribution (instead of uniformly):*

$$P_X = (p_0, p_1, p_2, p_3) = \left(\frac{1}{8}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}\right).$$

What is the optimal sequence of binary questions to identify the true value of X?

We will develop a strategy for this problem, motivated by the 'halving' property of the binary search scheme for the uniform case. In particular, we will take the strategy which can be summarized as follows:

make queries whose outcomes are equally likely (or in other words, are most uncertain).

Note that when X is uniformly distributed, the above strategy reduces exactly to the binary search. The decision tree of this strategy for the distribution of Question 4.2 is shown in Figure **??**. Unlike the previous game, the decision tree is not balanced — it asks more questions of the less likely values of i. The expected number of questions in this case is equal to

$$L = \sum_{i=0}^{3} p_i \ell_i = \frac{1}{8} \times 3 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 + \frac{1}{2} \times 1 = \frac{15}{8}$$
$$= -\sum_{i=1}^{n} p_i \log p_i := H(P_X).$$

Again the average number of yes/no questions needed to learn X is characterized by the quantity, $-\sum_{i \in \mathcal{X}} p_i \log p_i$. This functional of the probability distribution is called its *entropy*, also called its self-information. As we will see later, it is a fundamental limit on the average number of yes/no questions needed to learn the value of X (or equivalently, the average length of a binary lossless representation of all realizations of X).

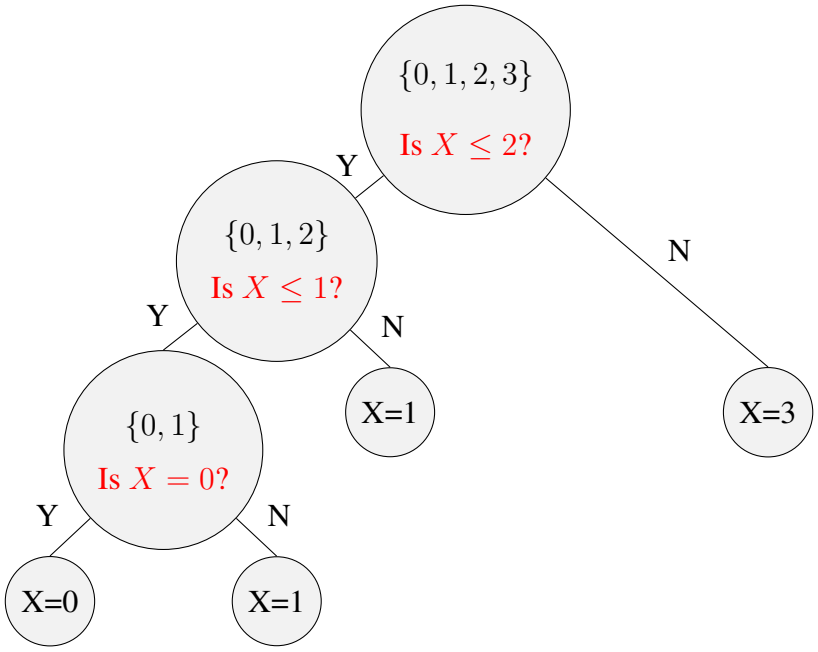

Figure 2: The figure shows the decision tree for the optimal strategy for a non-uniform probability distribution over $\boldsymbol{X}$. The key observation is that this strategy assigns fewer questions (or a shorter codeword) to the higher probability symbol (3), and more questions to the less probable symbols, such as 0 and 1.

In both the previous games, we assumed that the player knows the true distribution of X exactly. We now consider a situation where there is a mismatch between the true and the assumed distribution of X.

Question 4.3. *Consider the same setting as Question 4.2, but now assume that the true distribution (P_X) of X is not known to the player, and instead he believes that $X \sim Q_X$, where $Q_X = (1/2, 1/8, 1/8, 1/4)$. What is the effect of this distribution mismatch?*

	$X=0$	$X=1$	$X=2$	$X=3$
$Y=a$	0	1/8	1/8	0
$Y=b$	1/8	0	1/8	1/2

Table 1: Joint distribution P_{XY} of (X, Y).

Under the assumption that the true distribution is Q_X, the optimal codeword assignment is

$$0 \equiv (Y) \equiv (0), \quad 1 \equiv (NNY) \equiv (110), \quad 2 \equiv (NNN) \equiv (111), \quad 3 \equiv (NY) \equiv (10).$$

Again, as before, in this example, the number of questions needed to ascertain that $X = i$ is equal to $\log(1/q_i)$. The average number of questions needed to learn the value of X is

$$\begin{aligned} L &= \sum_{i=0}^{3} p_i \log(1/q_i) = -\sum_{i=1}^{3} p_i \log(p_i) + \sum_{i=0}^{3} p_i \log(p_i/q_i) \\ &= H(P_X) + D_{\text{KL}}(P_X \parallel Q_X). \end{aligned}$$

The second term in the display is called the relative entropy or KL divergence between P_X and Q_X, and it denotes the price paid by the player for using the wrong model for asking the yes/no questions (or the extra average codeword length incurred due to the ignorance of the true distribution). As we will see later, this quantity is always non-negative.

Question 4.4. *Finally, we now consider a case of guessing another random variable Y on the set $\{a, b\}$. The joint distribution of X and Y is stated in Table1.*

Suppose we want to ask a series of yes/no questions to find out the true value of both X and Y. Consider two strategies:

- *Player 1 develops two independent strategies for learning X and Y*
- *Player 2 develops a joint strategy for learning X and Y together.*

Who does better in terms of the average number of questions needed to learning both X and Y? How much is the improvement?

From the table, we can check that Y has a distribution $P_Y = (1/4, 3/4)$ over the set $\{a, b\}$. The marginal distribution of X is the same as in the previous two questions. Hence, we expect that the average number of yes/no questions need by player 1 (denoted by L_1) is

$$L_1 = H(P_X) + H(P_Y) \approx 1.75 + 0.811 \text{ bits} = 2.561 \text{ bits}$$

For the second player, who develops a joint strategy for querying about (X, Y), we expect the average number of yes/no questions (denoted by L_2) to be

$$L_2 = H(P_X, P_Y) = 4 \times \frac{1}{8} \times 3 + \frac{1}{2} \times 1 = 2 \text{ bits}.$$

Thus, the player who jointly considers the two random variables requires fewer questions. This is because, knowing the value of Y also provides information about the X value. For instance, if we know that $Y = a$, then we know that X cannot be 0 or 3. Exploiting this leads to the reduced number of questions needed by player 2. The amount of improvement, $L_1 - L_2$, is equal to

$$L_1 - L_2 = H(X) + H(Y) - H(X, Y) := I(X; Y).$$

The term $I(X; Y)$ is called the *mutual information* between the random variables X and Y, and it precisely quantifies the amount of information that X contains about Y (or equivalently, Y contains about X; since $I(X; Y)$ is symmetric).

Summary. The above discussion can be summarized as follows:

- Entropy $H(X) = -\sum_i p_i \log(p_i)$ quantifies the information content of a random variable X. It is also equal to the minimum average number of yes/no questions needed to learn the value of X.
- The relative entropy is a measure of discrepancy between two distributions P_X and Q_X. It quantifies the additional (on an average) number of yes/no questions needed to learn about X, under wrong model assumptions.
- The mutual information $I(X; Y)$ is measure of dependence between X and Y. It quantifies how much information about X is contained in the random variable Y.

Bibliographic Notes

As mentioned earlier, the field of information theory of initiated in a rather complete form by Shannon (1948). The textbook by Cover and Thomas (2006) is considered a standard reference, and it presents the basics of the subject in a clear style, while also highlighting its connections to various other fields. Csiszár and Körner (2011) develop a combinatorial approach to proving the main results in information theory, while the more recent book by Polyanskiy and Wu (2025) presents a modern and mathematically rigorous treatment of the subject, with a significant emphasis on the statistical applications. For a historical perspective, Gallager (2002) offers an engaging retrospective on Shannon's life and his technical contributions, while Verdu (1998) surveys the progress in information theory following Shannon's seminal paper.

The foundations of modern mathematical statistics were developed in the late 19th and early 20th century, with key contributions from Pearson, Gosset, Galton, Pearson (E.), Neyman, and most notably, Fisher. Rao (1992) contains an interesting account of these early years with a particular focus on Fisher. Wald (1945) initiated the formal study of sequential statistics, and important subsequent advances in sequential methods were made by Chernoff, Robbins, Siegmund, and Lai. A popular introductory textbook on statistical inference is by Casella and Berger (2024). More specialized treatments include (Lehmann and Casella, 1998) on point estimation, (Lehmann and Romano, 2005) on hypothesis testing, Van der Vaart (2000) on asymptotic theory, Tsybakov (2009) on nonparametric estimation, and Berger (2013) on Bayesian decision theory.

References

J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

G. Casella and R. Berger. *Statistical inference*. Chapman and Hall/CRC, 2024.

T. M. Cover and J. Thomas. *Elements of information theory*. John Wiley & Sons, 2006.

I. Csiszár and J. Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

R. G. Gallager. Claude E. Shannon: A retrospective on his life, work, and impact. *IEEE Transactions on Information Theory*, 47(7):2681–2695, 2002.

E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer, 1998.

E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer, 2005.

Y. Polyanskiy and Y. Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

C. R. Rao. RA Fisher: The founder of modern statistics. *Statistical science*, pages 34–48, 1992.

C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27 (3):379–423, 1948.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

S. Verdu. Fifty years of Shannon theory. *IEEE Transactions on information theory*, 44(6):2057–2078, 1998.

A. Wald. Sequential method of sampling for deciding between two courses of action. *Journal of the American Statistical Association*, 40(231):277–306, 1945.