EE 229A Lecture Notes

Ahmed Shakil

These notes were compiled while I took EE 229A, Information Theory and Coding. The class was taught by Professor Kannan Ramchandran in the fall of 2025.

Contents

Part I:

Introduction to Information Theory

08/28/25

Lecture 1

Intro to Information Theory, Historical background

Information theory answers two fundamental questions: What are the fundamental limits of data compression (answer: the entropy H) and what are the fundamental limits of reliable communication (answer: the channel capacity C). Although it is usually considered a subset of communication theory, information theory has made fundamental contributions to many other fields.

The field of information theory was founded in the 1940s by Claude Shannon. He proved that it is possible to send information at a positive rate with negligible probability of error (near zero) for all communication rates below channel capacity. He also argued that random processes have an irreducible complexity below which the signal cannot be compressed, he called this the *entropy*. Shannon argued that if the entropy of the source is less than the capacity of the channel, then asymptotically error-free communication can be achieved. Although we now know the ultimate limits of communication thanks to Shannon's work, achieving these limits have not been easy or computationally practical in most cases.

09/02/25

Lecture 2

Intro to Entropy and Mutual Information

Definition 2.0.1: Entropy

Entropy is a measure of the uncertainty of a random variable. Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p_X(x) = \Pr\{X = x\}$, $x \in \mathcal{X}$. The *entropy* $H(X)$ of a discrete random variable X is defined by

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{1}{p(x)} = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) = \mathbb{E}_p \left[\log_2 \frac{1}{p(X)} \right].$$

The units are bits/symbol. Entropy is label invariant, it is a functional of the distribution of X. Hence, sometimes we write it as $H(p(x))$ or $H(p)$. A couple properties of entropy are listed below.

1. $H(X) \geqslant 0$
2. $H_b(X) = (\log_b a) H_a(X)$
3. It is a concave function of the distribution.

Definition 2.0.2: Joint Entropy

The *joint entropy* $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint probability mass function $p_{X,Y}(x, y)$ is defined as

$$H(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)} = \mathbb{E} \left[\log \frac{1}{p(X, Y)} \right].$$

If X and Y are independent, $p(x, y) = p(x) \cdot p(y)$. Then we have

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

Definition 2.0.3: Conditional Entropy

Consider two discrete random variables $(X, Y) \sim p(x, y)$, the *conditional entropy* $H(Y \mid X)$ is defined as

$$\begin{aligned} H(Y \mid X) &= \sum_{x \in \mathcal{X}} p(x) H(Y \mid X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y \mid x) \log p(y \mid x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y \mid x) \\ &= -\mathbb{E} \left[\log p(Y \mid X) \right]. \end{aligned}$$

Theorem 2.0.4: Chain rule

$$H(X, Y) = H(X) + H(Y \mid X).$$

Proof.

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)\\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x)p(y\mid x)\\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y\mid x)\\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y\mid x)\\
&= H(X) + H(Y\mid X).
\end{aligned}$$

□

Corollary 2.0.5

$$H(X,Y\mid Z) = H(X\mid Z) + H(Y\mid X,Z).$$

2.1 Mutual Information

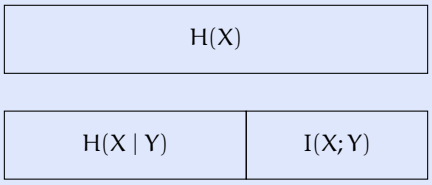
Below we define mutual information. It is a measure of the amount of information that one random variable contains about another random variable. It is the reduction in the uncertainty of one random variable due to the knowledge of the other.

Definition 2.1.1: Mutual Information

Let $(X,Y)\sim p(x,y)$. The *mutual information* is defined by the following:

$$I(X;Y) \coloneqq H(X) - H(X\mid Y) = I(Y;X).$$

Intuitively, the mutual information $I(X;Y)$ is the reduction in the uncertainty of X due to the knowledge of Y. The image below should give a slight idea on how the mutual information and entropy relate.



2.2 Convexity and Jensen's Inequality

In the following we give some definitions and theorems related to convex functions. For the most part, we will relegate the proofs to the textbook.

Definition 2.2.1: Convexity

A real-valued function $f(x)$ is said to be *convex* over an interval (a,b) if for every $x_1, x_2 \in (a,b)$ and $0 \leqslant \lambda \leqslant 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leqslant \lambda f(x_1) + (1-\lambda) f(x_2)$$

A function f is said to be *strictly* convex if equality only holds if $\lambda = 0$ and $\lambda = 1$. A function f is *concave* if $-f$ is convex.

Theorem 2.2.2

If the function f has a second derivative that is nonnegative (positive) over an interval, the function is convex (strictly convex) over that interval.

Theorem 2.2.3: Jensen's Inequality

If X is a real-valued random variable and $f(x)$ is a convex function, then

$$\mathbb{E}\left[f(X)\right] \geqslant f\left(\mathbb{E}[X]\right).$$

It is important to note that the inequality flips if the function f is concave instead.

Theorem 2.2.4: Properties of Mutual Information

Recall the definition of mutual information:

$$I(X;Y) \coloneqq H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) = H(X) + H(Y) - H(X,Y) = I(Y;X).$$

Below we list some properties of $I(X;Y)$.

1. $I(X;Y)$ is symmetric, this can be worked out by manipulating the definitions.
2. $I(X;Y) \geqslant 0$
3. $I(X;Y) = 0 \iff X$ and Y are independent

The figure below tries to illustrate the relationships between entropy and mutual information.

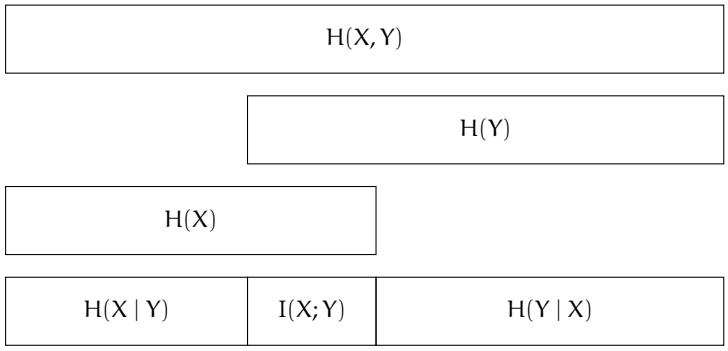


Figure 2.1: The big picture.

Theorem 2.2.5

A useful tool to have is the following inequality:

$$H(X) \leqslant \log |X|.$$

Proof. The following uses Jensen's inequality and the fact that the log function is concave,

$$H(X) = \mathbb{E}\left[\log \frac{1}{p(X)}\right] \leqslant \log \mathbb{E}\left[\frac{1}{p(X)}\right] = \log |X|\,.$$

□

09/04/25

Lecture 3

Cross Entropy and Relative Entropy

3.0.1: Recap of the previous lecture

We defined entropy as

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}.$$

We defined conditional entropy as

$$H(X \mid Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x \mid y)}.$$

We also have the following chain rule for entropy

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1} H(X_i \mid X^{i-1}).$$

For any convex function $f(x)$, Jensen's inequality states

$$\mathbb{E}[f(x)] \geqslant f(\mathbb{E}[X]).$$

Lastly, we defined mutual information as

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) = H(X) + H(Y) - H(X,Y) = I(Y;X) \geqslant 0.$$

Some properties of entropy are listed below:

1. $H(X) \geqslant 0$
2. $H(X) \leqslant \log_2 |\mathcal{X}|$, the upper bound is only attained when the distribution is uniform

As a general rule, the uniform distribution maximizes entropy for finite alphabet sources.

Theorem 3.0.2: Chain Rule for Mutual Information

$$I(X; Y_1, Y_2) = I(X; Y_1) + I(X; Y_2 \mid Y_1).$$

In general, we have the following:

$$I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X^{i-1}).$$

Definition 3.0.3: Cross Entropy

Cross entropy is the amount of information required on average to describe a random variable $X \sim p(x)$ if a description of another random variable with pmf $q(x)$ is used instead. Let $p(x)$ and $q(x)$ be two probability mass functions over the same alphabet $\mathcal{X}$. The *cross entropy* between $p(x)$ and $q(x)$ is defined as

$$H(p(x), q(x)) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)} = \mathbb{E}_p \left[\log \frac{1}{q(x)} \right].$$

Intuitively, we should think of this quantity as quantifying the mismatch of $p(x)$ and $q(x)$. In a sense we are

pretending $q(x)$ is $p(x)$, when in reality the truth is $p(x)$. Some properties of cross-entropy:

1. $H(p, q) \geqslant 0$
2. $H(p) = H(p, p)$
3. $H(p, q) \geqslant H(p)$

3.1 Relative Entropy

Definition 3.1.1: Relative Entropy

Let $p(x)$ and $q(x)$ be two probability mass functions over the same alphabet $\mathcal{X}$. The *relative entropy* or *Kullback-Leibler divergence* (KL divergence), between $p(x)$ and $q(x)$ is defined as

$$D(p(x) \,\|\, q(x)) \coloneqq \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log \frac{p(x)}{q(x)} \right] = H(p(x), q(x)) - H(p(x)).$$

For brevity, sometimes we will use the notation $D(p \,\|\, q)$ for relative entropy. Although we intuitively take this as a distance between distributions, this is not a metric as it does not satisfy the triangle inequality and is not symmetric. Two important properties to note about this quantity is the following

1. $D(p \,\|\, p) = 0$
2. $D(p \,\|\, q) \geqslant 0$ for all p, q with equality if and only if $p = q$.

Theorem 3.1.2: Gibbs Inequality

Let $p(x)$ and $q(x)$, where $x \in \mathcal{X}$ and $\mathcal{X}$ is a finite alphabet, be two probability mass functions. Then

$$D(p \,\|\, q) \geqslant 0$$

with equality if and only if $p(x) = q(x)$ for every $x \in \mathcal{X}$.

Proof. Let $A = \{x : p(x) > 0\}$ be the support set of $p(x)$. Then

$$\begin{aligned}
-D(p \,\|\, q) &= -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} && (3.1)\\
&= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} && (3.2)\\
&\leqslant \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} && (3.3)\\
&= \log \sum_{x \in A} q(x) && (3.4)\\
&\leqslant \log \sum_{x \in \mathcal{X}} q(x) && (3.5)\\
&= \log 1 && (3.6)\\
&= 0, && (3.7)
\end{aligned}$$

where (3.3) follows from Jensen's inequality. Since the log function is strictly concave, we have equality in (3.3) if and only if $q(x)/p(x)$ is constant everywhere. Thus, $\sum_{x \in A} q(x) = c \sum_{x \in A} p(x) = c$. We have equality

in (3.5) if and only if $\sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x) = 1$, which implies that $c = 1$. Hence, we have $D(p \parallel q) = 0$ if and only if $p(x) = q(x)$ for every x. □

Corollary 3.1.3: Minimum Cross Entropy

$$H(p(x), q(x)) \geqslant H(p(x))$$

with equality if and only if $p(x) = q(x)$.

Proof. Follows from $D(p \parallel q) = H(p, q) - H(p)$ and $D(p \parallel q) \geqslant 0$. □

Corollary 3.1.4

$$I(X; Y) \geqslant 0$$

Proof.

$$I(X; Y) = \mathbb{E}\left[\log \frac{p(x, y)}{p(x)p(y)}\right] = D(p(x, y) \parallel p(x) \cdot p(y)) \geqslant 0$$

□

Corollary 3.1.5

The following implies conditioning reduces entropy on average.

$$H(X \mid Y) \leqslant H(X)$$

Proof.

$$I(X; Y) = H(X) - H(X \mid Y) \geqslant 0$$

□

09/09/25

Lecture 4

Properties of Relative Entropy & Mutual Information, Data Processing

It is important to remember that the entropy of X is maximized by a uniform probability mass function on a finite alphabet $\mathcal{X}$.

Theorem 4.0.1

$$H(X) \leqslant \log |\mathcal{X}|$$

with equality if and only if $p_X(x) = \frac{1}{|\mathcal{X}|}$ for every $x \in \mathcal{X}$.

Proof. Consider a random variable X defined on $\mathcal{X}$ with $|\mathcal{X}| = n$. Let U be the uniform distribution on $\mathcal{X}$.

$$\begin{aligned} H(U) - H(X) &= \log_2 |\mathcal{X}| - \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &= \sum_{x \in \mathcal{X}} p(x) \log |\mathcal{X}| - \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{\frac{1}{|\mathcal{X}|}} \\ &= D(p(x) \,\|\, u(x)) \geqslant 0 \end{aligned}$$

Hence, $H(U) \geqslant H(X)$ with equality if and only if X is uniformly distributed. □

Theorem 4.0.2: Convexity of Relative Entropy

$D(p(x) \,\|\, q(x))$ is convex in the pair $(p(x), q(x))$; that is, if $(p_1(x), q_1(x))$ and $(p_2(x), q_2(x))$ are two pairs of probability mass functions on $\mathcal{X}$, then

$$D(\lambda p_1 + (1-\lambda)p_2 \,\|\, \lambda q_1 + (1-\lambda)q_2) \leqslant \lambda D(p_1 \,\|\, q_1) + (1-\lambda) D(p_2 \,\|\, q_2)$$

for all $0 \leqslant \lambda \leqslant 1$.

4.1 Data Processing Inequality (DPI)

Definition 4.1.1

Say we have random variables X, Y, Z which form a Markov Chain ($X \to Y \to Z$). Recall that

$$p(x, y, z) = p(x)p(y \mid x)p(z \mid y, x) = p(x)p(y \mid x)p(z \mid y).$$

The last equality follows from the Markov property. Some simple consequences are as follows:

1. $X \to Y \to Z$ if and only if X and Z are conditionally independent given Y. Markovity implies this conditional independence since

$$p(x, z \mid y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z \mid y)}{p(y)} = p(x \mid y)p(z \mid y).$$

2. $X \to Y \to Z$ implies that $Z \to Y \to X$. Thus, we can simply write $X - Y - Z$.
3. $I(X; Z \mid Y) = 0$
4. If $Z = f(Y)$, then $X \to Y \to Z$ is a Markov chain.

We now prove an important and useful theorem demonstrating that no processing of Y, deterministic or random, can increase the information that Y contains about X.

Theorem 4.1.2: Data-processing Inequality

If $X \to Y \to Z$, then $I(X; Y) \geqslant I(X; Z)$ with equality if and only if $I(X; Y \mid Z) = 0$. Intuitively, no amount of processing on Y, deterministic or randomized, can make you learn more about X.

Proof. By the chain rule, we can expand mutual information in two different ways:

$$\begin{aligned} I(X; Y, Z) &= I(X; Y) + I(X; Z \mid Y) \\ &= I(X; Z) + I(X; Y \mid Z). \end{aligned}$$

Since X and Z are conditionally independent given Y, we have $I(X; Z \mid Y) = 0$. Since $I(X; Y \mid Z) \geqslant 0$, we have

$$I(X; Y) \geqslant I(X; Z).$$

We have equality if and only if $I(X; Y|Z) = 0$. Similarly, one can prove that $I(Y; Z) \geqslant I(X; Z)$. □

Corollary 4.1.3

If $X - Y - Z$ forms a Markov chain, then the dependency between X and Y is decreased by observation of a "downstream" random variable Z. In other words,

$$I(X; Y \mid Z) \leqslant I(X; Y).$$

STOPPED EDITING HERE.

4.2 Asymptotic Equipartition Property

Entropy is directly related to the fundamental limits of data compression.

1. For a sequence of n i.i.d random variables $\{X_i\}_{i=1}^n \sim B(\frac{1}{2})$, we need $nH(X_1) = n$ bits to describe the sequence.
2. What if $X_i \sim B(0.11)$ instead of being a fair coin flip? Now, we should need about $n/2 = nH(0.11) = nH(X_1)$ bits.

Ratio of the number of typical sequences to the total number of sequences is

$$\frac{s^{nH(X)}}{2^n} \overset{n\to\infty}{\to} 0$$

Part II:

Compression (Source Coding)

09/11/25

Lecture 5

AEP, Data compression

Readings: C & T Ch 3, 4.1-2

Example 5.0.1: Recap from last lecture

Take $n = 1000$ and $p = 0.11$. This means that $H(p) = 0.5$. We call a typical sequence to have np ones and $n\overline{p}$ zeroes. The number of typical sequences is $2^{nH(X)} = 2^{500}$ while the probability of observing a typical sequence is 2^{-500}.

Theorem 5.0.2: AEP

If we have the following independently and identically distributed random variables $X_1, \ldots, X_n \sim p(x)$ then

$$\frac{1}{n} \log_2 \frac{1}{p(x_1, \ldots, x_n)} \to H(X) \quad \textit{in probability.}$$

Proof. Take $Y_i = \frac{1}{n} \log_2 \frac{1}{p(x_i)}$. Then we have

$$\frac{1}{n} \log \frac{1}{p(x_1, \ldots, x_n)} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{1}{p(x_i)} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

By WLLN,

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow{p} \mathbb{E}[Y].$$

□

Definition 5.0.3: Convergence in probability

Given random variables $X_1, X_2, \ldots$ we say convergence in probability to the random variable X happens if

$$\forall \epsilon > 0 \lim_{n \to \infty} \Pr\{|X_n - X| > \epsilon\} = 0.$$

Example 5.0.4: ϵ—typical set

$$A_\epsilon^{(n)} = \left\{ x^n : 2^{-n(H(X)+\epsilon)} \leqslant p(x^n) \leqslant 2^{-n(H(X)-\epsilon)} \right\}$$

5.0.5: Properties of AEP

1. $\Pr(x^n \in A_\epsilon^{(n)}) \geqslant 1 - \epsilon$
2. $\left| A_\epsilon^{(n)} \right| \leqslant 2^{n(H(X)+\epsilon)}$
3. $A_\epsilon^{(n)} \geqslant (1 - \epsilon) 2^{n(H(X)-\epsilon)}$

5.1 Data Compression

There exists a code that maps input sequences $\{x^n\}$ to binary strings such that the mapping is one-to-one (and therefore reversible). Let $l(x^n)$ denote the length of the binary codeword for x^n. Divide your sequences into two sets, typical and atypical. This division is dictated by the sequences which are in the $A_\epsilon^{(n)}$.

$$\#\text{ of bits needed to represent a typical sequence } \leqslant n(H(X)+\epsilon)+1$$

$$\#\text{ of bits needed to represent an atypical sequence } \leqslant n\log_2|\mathcal{X}|+1$$

Essentially, if the given sequence falls into the typical set, simply send the codeword for the sequence. Otherwise, send the uncompressed sequence. When sending sequences, an extra header bit needs to be added to the sequence in order to indicate if the sequence is compressed or not.

$$\begin{aligned}\mathbb{E}[l(x^n)] &= \mathbb{E}[l(x^n)|x^n \in A_\epsilon^{(n)}]\Pr(A_\epsilon^{(n)}) + \mathbb{E}[l(x^n)|x^n \notin A_\epsilon^{(n)}]\Pr\left[(A_\epsilon^{(n)})^C\right] \\ &= [n(H+\epsilon)+2]\Pr(A_\epsilon^{(n)}) + \left[n\log_2|\mathcal{X}|+2\right]\Pr\left[(A_\epsilon^{(n)})^C\right] \\ &\leqslant n(H+\epsilon) + \epsilon n\log_2|\mathcal{X}| + 2 \\ &= n(H(X)+\epsilon') \text{ where } \epsilon' = \epsilon + \epsilon\log_2|\mathcal{X}| + \frac{2}{n}\end{aligned}$$

Theorem 5.1.1

For this scheme specifically we have the following:

$$\forall\ \epsilon > 0\ \ \exists\ n \text{ such that } \mathbb{E}[l(x^n)] \leqslant n\left[H(X)+\epsilon\right].$$

Definition 5.1.2

Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} p(x)$. For each $n = 1, 2, ..$, define $B_n \subset \mathcal{X}^n$ to be the smallest set with $\Pr\{x^n \in B_n\} \geqslant 1-\epsilon$. We can show that $\left\{B_n \cap A_\epsilon^{(n)}\right\}$ is significant, and that they have essentially the same number of elements.

Thrm 3.3.1

Converse of achievability:

The optimal source code would assign shorter descriptions to more probable outcomes. We will construct a minimum expected length code that is unique (so-called "non-singular") $n-$code as follows.

1. Order X^n according to their probabilities
2. Assign codeword 0 to the highest probability sequence
3. Assign codeword 1 to the second highest probability sequence
4. Assign codeword 00 to the third highest probability sequence
5. And so on and so forth

This procedure maps the $(2^{l+1}-2)$ most likely sequences into unique codewords of length $\leqslant l$.

Now set $l = n(H(X)-2\epsilon)$. B_n is the set of $2^{l+1}-2$ most likely sequences of length n. This means $|B_n| = 2^{n(H(X)-2\epsilon)+1}-2$.

Fill in details:

$$\begin{aligned}\Pr\{X^n \in B_n\} &= \Pr(B_n \cap (A_\epsilon^{(n)})) + \Pr(B_n \cap A_{\epsilon^{(n)}}) \\ &\leqslant \epsilon + |B_n|\,2^{-n(H-\epsilon)} \\ &\leqslant a\epsilon \text{ where } a > 1\end{aligned}$$