Information Theory and Statistical Learning

Adithya Ganesh

August 25, 2019

Acknowledgments

Thank you to Yuval Widgerson from the Stanford Mathematics Department for useful discussions.

Contents

1 The Source Coding Theorem

Definition. *An* **ensemble** X *is a triple* $(x, \mathcal{A}_X, \mathcal{P}_X)$ *where the outcome* x *is the value of a random variable, which takes on one of a set of possible values,* $\mathcal{A}_X = \{a_1, a_2, \ldots, a_i, \ldots, a_I\}$ *having probabilities* $\mathcal{P}_X = \{p_1, p_2, \ldots, p_I\}$*, with* $P(x = a_i) = p_i, p_i \geq 0$ *and* $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$.

Definition. *We define the* **Shannon information content** *of the outcome* $x = a_i$ *to be*

$$h(x = a_i) \equiv \log_2 \frac{1}{p_i}.$$

Definition. *We define the* **entropy** *of the ensemble to be*

$$H(X) = \sum_i p_i \log_2 \frac{1}{p_i}.$$

Intuition. The outcome of a random experiment is guaranteed to be most informative if the probability distribution over outcomes is uniform.

1.1 A basic example

What's the smallest number of yes/no questions needed to identify an integer x between 0 and 63?

Intuitively, the best questions successively divide the 64 possibilities into equal sized sets. One strategy is to ask the following questions.

- Is $x \geq 32$?
- Is $x \mod 32 \geq 16$?
- Is $x \mod 16 \geq 8$?
- Is $x \mod 8 \geq 4$?
- Is $x \mod 4 \geq 2$?
- Is $x \mod 2 = 1$.

The answers to these questions if encoded in binary, give the expansion of x, for example $35 \implies 100011$. If all values of x are equally likely, then the answers to the questions are independent, and each has Shannon information content $\log_2(1/0.5) = 1$ bit.

The Shannon information content in this setting measures the length of a binary file that encodes x.

Similarly, refer to the submarine game example (pg. 71, MacKay).

Definition. *The* **raw bit content** *of X is*

$$H_0(X) = \log_2 |\mathcal{A}_X|,$$

which is a lower bound for the number of binary questions that are guaranteed to identify an outcome from the ensemble X.

Definition. *The* **smallest δ-sufficient subset** *S_δ is the smaller subset of $\mathcal{A}_x$ satisfying*

$$P(x \in S_\delta) \geq 1 - \delta$$

Definition. *The essential bit content of X is*

$$H_\delta(X) = \log_2 |S_\delta|.$$

Theorem (Shannon's source coding theorem)**.** *Let X be an ensemble with entropy $H(X) = H$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for $N > N_0$,*

$$\left| \frac{1}{N} H_\delta(X^N) - H \right| < \epsilon.$$

Theorem (Chebyshev's inequality 1)**.** *Let t be a non-negative real random variable, and let α be a positive rela number. Then*

$$P(t \geq a) \leq \frac{\bar{t}}{\alpha}.$$

Theorem (Chebyshev's inequality 2)**.** *Let x be a random variable, and let α be a positive real number. Then*

$$P((x-\overline{x})^2 \geq \alpha) \leq \sigma_x^2/\alpha.$$

Theorem (Weak law of large numbers)**.** *Take x to be the average of N independent random variables $h_1, \ldots, h_N$, having common mean $\overline{h}$ and common variance σ_h^2: $x = \frac{1}{N}\sum_{n=1}^{N} h_n$. Theno*

$$P((x-\overline{h}^2) \geq \alpha) \leq \sigma_h^2/\alpha N.$$

Theorem (Asymptotic equipartition principle.)**.** *For an ensemble of N independent identically distributed (i.i.d.) random variable $X^N \equiv (X_1, X_2, \ldots, X_N)$, with N sufficiently large, the outcome $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ is almost certain to belong to a subset of $\mathcal{A}_X^N$ having only $2^{NH(X)}$ members, each having probability "close" to $2^{-NH(X)}$. (The term equipartition is chosen to describe the idea that the members of the typical set have roughly equal probability.)*

Proof of source coding theorem.

Verbally, the source coding theorem states that N i.i.d. random variables each with entropy $H(X)$ can be compressed into more than $NH(X)$ with negligible risk of information loss as $N \to \infty$. Conversely, if they are compressed into fewer than $NH(X)$ bits, it is virtually certain that information will be lost.

A long string of N symbols will usually contain about p_iN occurences of the i-th symbol, so that the probability of this "typical" string is roughly

$$P(\mathbf{x})_{typ} \approx p_1^{p_1N} p_2^{p_2N} \cdots p_l^{p_lN},$$

so that the information content of a typical string is

$$\log_2 \frac{1}{P(\mathbf{x})} \approx N \sum_i p_i \log_2 \frac{1}{p_i} = NH.$$

First, apply the weak law of large numbers to the random variable $\frac{1}{N}\log_2\frac{1}{P(x)}$. Define the *typical set* with parameters N and β as follows:

$$T_{N\beta} = \left\{ x \in \mathcal{A}_X : \left[\frac{1}{N}\log_2\frac{1}{P(x)} - H\right]^2 < \beta^2 \right\}.$$

For all $x \in T_{N\beta}$, the probability of x satisfies

$$2^{-N(H+\beta)} < P(x) < 2^{-N(H-\beta)}.$$

By the law of large numbers, $P(x \in T_{N\beta}) \geq 1 - \sigma^2/(\beta^2 N)$.

This means that as N increases, the probability that $\mathbf{x}$ falls in $T_{N\beta}$ approaches 1, for any β.

Now, we will relate $T_{N\beta}$ to $H_\delta(X^N)$. Our strategy is to show that for any given δ, there is a sufficiently large N such that $H_\delta(X^N) \equiv NH$.

Part 1. $\frac{1}{N}H_\delta(X^N) < H + \epsilon$.

Since the total probability contained by $T_{N\beta}$ can't be larger than 1, we hve that

$$|T_{N\beta}|2^{-N(H+\beta)} < 1,$$

that is

$$|T_{N\beta}| < 2^{N(H+\beta)}.$$

Setting $\beta = \epsilon$, and choosing N_0 such that $\frac{\sigma^2}{\epsilon^2 N_0} \leq \delta$, then $P(T_{N\beta}) \geq 1 - \delta$, and analyzing the set $T_{N\beta}$ implies

$$H_\delta(X^N) \leq \log_2 |T_{N\beta}| < N(H + \epsilon).$$

Part 2. $\frac{1}{N}H_\delta(X^N) > H - \epsilon$.

We set $\beta = \epsilon/2$, so it suffices to show that that a subset S' having $|S'| \leq 2^{N(H-beta)}$ and achieving $P(\mathbf{x} \in S') \geq 1 - \delta$ cannot exist.

The probability of the subset S' is

$$P(\mathbf{x} \in S') = P(\mathbf{x} \in S' \cap T_{N\beta}) + P(\mathbf{x} \in S' \cap \overline{TN_{N\beta}}),$$

where $\overline{T_{N\beta}}$ denotes the complement of the typical set.

IThe maximum value of the first term is found if $S' \cap T_{N\beta}$ contains $2^{N(H-2\beta)}$ outcomes all with the maximum probability $2^{-N(H-\beta)}$. The maximum value the second term can have is $P(\mathbf{x} \notin T_{N\beta})$.

Thus:

$$P(\mathbf{x} \in S') \leq 2^{N(H-2\beta)}2^{-N(H-\beta)} + \frac{\sigma^2}{\beta^2 N} = 2^{-N\beta} + \frac{\sigma^2}{\beta^2 N}.$$

We can now set $\beta = \frac{\epsilon}{2}$ and N_0 such that $P(\mathbf{x} \in S') < 1 - \delta$, which shows that S' does not satisfy the desired conditions.

Therefore, for large enough N, the function $\frac{1}{N}H_\delta(X^N)$ is essentially a constant function of δ for $0 < \delta < 1$. In particular, this shows us that regardless of our specific tolerance for error, the number of bits per symbol needed to specify $\mathbf{x}$ is H bits.

Figure:

(fill in)

2 Maximum Entropy Principle

Due to E.T. Jaynes in 1957, where he explored the correspondence between statistical mechanics and information theory. Take precisely stated prior data or testable information about a probability distribution function. The distribution with maximal entropy is the best choice to encode the prior data.

- The exponential distribution for which the density function is

  $$p(x|\lambda) = \begin{cases} \lambda e^{-\lambda x}; & x \geq 0 \\ 0; & x < 0, \end{cases}$$

  is the maximum entropy distribution among all continuous distributions supported in $[0, \infty)$ that have a specified mean of $\frac{1}{\lambda}$.
- The normal distribution $\mathcal{N}(\mu, \sigma^2)$ for which the density function is

  $$p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

  has maximum entropy among all real-valued distributions supported on $(-\infty, \infty)$ with specified variance σ^2. Therefore: the assumption of normality imposes the minimal prior structural constraint.

To do: watch David Tse talk + talk on Information Theory on deep learning (Stanford).

3 Core ideas in information theory

1. Overview
   (a) Compression (lossless vs. lossy)
   (b) Communication (reliable vs. communication with loss [also joint source channel coding)
2. Course goals
   (a) Measures of information (entropy, relative entropy, mutual information, chain rules)
   (b) Compression, storage, communication
   (c) Fundamental limits
   (d) Concrete schemes for compression and communication
   (e) Existence proofs via random constructions (random coding)
   (f) Typical sequences interplay between info theory, probability, and stats

Example 1. Lossless compression.

Consider the source: $U_1, U_2, \ldots$ iid $\sim U \in \{A, B, C\}$.

Further, suppose

$$P(U = A) = 0.7, P(U = B) = P(U = C) = 0.15.$$

Approach 1. Consider $A \to 00$, $B \to 01, C \to 11$. But too wasteful, since A occurs more frequently.

Approach 2. Better is $A \to 0, B \to 01, C \to 11$. Note that this is 'prefix code': no code forms the prefix of another; this makes code easy to decode.

Expected number of bits per source symbol:

$$\overline{L} = 0.7 \cdot 1 + 0.15 \cdot 2 + 0.15 \cdot 2 = 1.3 \text{ bits / symbol}$$

Approach 3. In fact, we can do better. Consider pairs of source symbols. Namely, let us examine

Pair	Probability	Code Word
AA	0.49	0
AB	0.105	100
AC	0.105	111
BA	0.105	101
CA	0.105	1100
BB	0.0225	110100
BC	0.0225	110101
CB	0.0225	110110
CC	0.0225	110111

Satisfies prefix code. Encoding and decoding done in linear time. Later: we will see that this is optimal for these source symbols.

Again, let's compute expected bits per symbol:

$$\overline{L} = \frac{1}{2}(0.49 \cdot 1 + 0.105 \cdot 3 \cdot 3 + 0.105 \cdot 4 + 0.0225 \cdot 6 \cdot 4) = 1.1975 \text{ bits / symbol}$$

Entropy. For any scheme, the value $\overline{L} \geq H(U)$, where the source entropy

$$H(U) = \sum_{u \in \mathcal{U}} P(u) \log_2 \frac{1}{P(u)}.$$

In the above case:

$$H(U) \approx 1.18129.$$

On the other hand for all $\epsilon > 0$, there exists a scheme such that

$$\overline{L} \leq H(U) + \epsilon.$$

Example 2. Consider a source

$$U_1, U_2, \dots, \quad \text{iid} ; \quad P(U_i = 0) = P(U_i = 1) = \frac{1}{2}.$$

Suppose further that a channel flips each bit w.p. $q < \frac{1}{2}$.

Output of channel:

$$Y_i = X_i \bigoplus_2 W_i,$$

where $W_i \sim \text{Ber}(q)$. Note that the source symbol U_i is different from the encoding X_i.

Approach 1. We can let

$$X_i = U_i,$$

we will get probability of error per source bit, $P_e = q$.

Approach 2. Alternatively, can repeat 3 times:

if $U = 0110$, then we can let $X = 000111111000$.

In this case:

$$\text{rate} = \frac{1}{3} \text{ bits / channel use}$$

The upside, is that the probability of error becomes

$$P_e = 3q^2(1-q) + q^3 < q.$$

So probability of error has dropped, at the cost of requiring more space.

4 Dyadic U and symbol counting

Lemma. Suppose U is dyadic with $|U| \geq 2$, and let $n_{max} = \max_{u \in \mathcal{U}} n_u$. The number of symbols with $n_u = n_{max}$ is even.

Proof. Observe that

$$\begin{aligned} 1 = \sum_u p(u) &= \sum_u 2^{-n_u} \\ &= \sum_{n=1}^{n_{max}} (\text{# of letters } u \text{ with } n_u = n) \cdot 2^{-n} \end{aligned}$$

Therefore,

$$\begin{aligned} 2^{n_{max}} &= \sum_{n=1}^{n_{max}} (\text{ of letters } u \text{ with } n_u = n) \cdot 2^{n_{max}-n}. \\ &= \sum_{n=1}^{n_{max}-1} (\text{# of letters } u \text{ with } n_u = n) \cdot 2^{n_{max}-n} + (\text{# of letters } u \text{ with } n_u = n_{max}). \end{aligned}$$

By parity, it follows that # of letters u with $n_u = n_{max}$ must be even.

5 Optimality of Huffman Codes

Construction of Huffman Codes. Exactly the same as that for dyadic sources. Recall that the procedure identifies the symbols with the smallest probabilities and merges them in a binary tree structure.

Example. (Senary Source) Consider the alphabet

Add D1

u	$p(u)$
a	0.25
b	0.25
c	0.2
d	0.15
e	0.1
f	0.05

Theorem. Huffman code is an optimal prefix code.

(Note that we say an optimal and not "the optimal" because there may be more than one construction. Even within the construction of Huffman, the way we break ties is arbitrary. We can also choose to split the binary tree in one direction via a 1 vs. 0. So we can have many different schemes, though they are all essentially equivalent, in terms of the length function.)

When we use the term "optimality" here, we mean in terms of minimizing the expected length $\bar{l}$.

Proof. Assume without loss of generality that $U \sim P$ over an alphabet $\mathcal{U} = \{1, 2, \ldots, r\}$. Further, suppose that $p(1) \geq p(2) \cdots \geq p(r)$ (i.e. they are arranged in descending probabilities).

Let V denote the random variable with $\mathcal{V} = \{1, 2, \ldots, r-1\}$ obtained from U by merging $r-1$ and r.

Let $\{c(i)\}_{i=1}^{r-1}$ be a prefix code for V. Then we can obtain $\{\tilde{c}\}_{i=1}^{r}$ which is a prefix code that *splitting* the last codeword $c(r-1)$.

Observation. The Huffman code for U is obtained from the Huffman code from V by splitting.

Lemma. Suppose that $\{c(i)\}_{i=1}^{r-1}$ is an optimal prefix code for V. If $\{\tilde{c}(i)\}_{i=1}^{r}$ is obtained from $\{c(i)\}_{i=1}^{r-1}$ by splitting, then $\{\tilde{c}(i)\}_{i=1}^{r}$ is an optimal prefix code for U.

This observation coupled with the lemma directly implies the theorem. We can iterate this argument to merely need establish optimality of Huffman code for binary alphabet ($r = 2$), which is trivially true.

Proof of Lemma. Note there is an optimal prefix code for U that satisfies:

1. $\bar{l}(1) \leq l(2) \leq \cdots \leq l(r-1) \leq l(r) \triangleq l_{max}$ (lengths are in increasing order).

2. $l(r-1) = l(r)$.

   (Otherwise, we would be able to "chop off" the final part of the last code word to achieve $l(r-1) = l(r)$ and improve the code.)

3. The last two code words differ only in the last bit.

   (Otherwise, we can swap out the last code word. This follows since the first $r-1$ codewords comprise a prefix code.) This ensures that the prefix code for U is obtained by splitting on the code for V.

Recall the following:

$$\mathbb{E}l_{split}(U) = \mathbb{E}l(V) + p(r-1) + p(r).$$

Therefore: an optimal prefix code for U is obtained by splitting an optimal prefix code for V. ■

Further reading on lossless compression:

- Shannon-Fano-Elias coding (5.9 of Cover and Thomas)
- Arithmetic coding (13.3)
- Lempel-Ziv coding (13.4)

Note that optimally applying Huffman codes requires working in blocks of symbols. And the table of symbols is exponential in the block length n. The Shannon-Fano-Elias and Arithmetic coding permit constructions that scale gracefully in the block length n. Lempel-Ziv coding is elegant algorithmically, and is guaranteed to be optimal even without the source being memoryless and even without knowing the probability distribution! Indeed, `gzip` at its heart is implemented in terms of the Lempel-Ziv coding scheme.

6 Channel Capacity

Given a channel with inputs X and outputs Y:

$$X \to [P(Y|X)] \to Y$$

Define: Channel capacity C is the maximal rate of reliable communication (over memoryless channel characterized by $P(Y|X)$).

Further, recall the following definition:

$$C^{(I)} = \max_{P_X} I(X;Y).$$

Theorem. Channel capacity is limited by maximum mutual information.

$$C = C^{(I)}.$$

Proof: We will see this proof soon.

- This theorem is important because C is challenging to optimize over, whereas $C^{(I)}$ is a tractable optimization problem.

6.0.1 Examples

Example I. Channel capacity of a Binary Symmetric Channel (BSC).

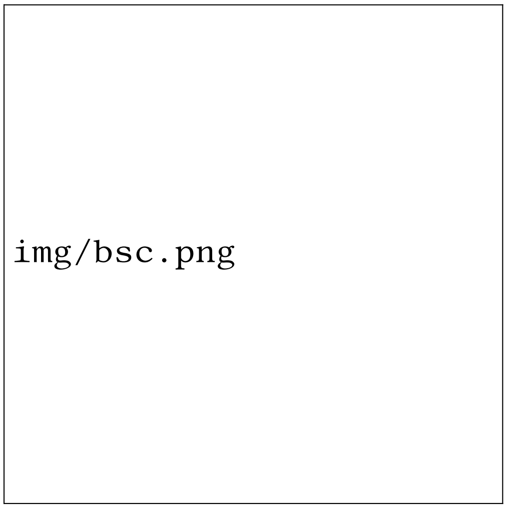
Define alphabets $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. A BSC is defined by the PMF:

$$P_{Y|X}^{(y|x)} = \begin{cases} p & y \neq x \\ 1 - p & y = x. \end{cases}$$

This is equivalent to a channel matrix

$$\begin{pmatrix} 1 - p & p \\ p & 1 - p \end{pmatrix}$$

And the graph representation



This can also be expressed in the form of additive noise.

$$Y = X \bigoplus_2 Z, \text{ where } Z \sim \text{Ber}(p).$$

To determine the channel capacity of a BSC, by the theorem we must maximize the mutual information.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - H(X \bigoplus_2 Z|X) \end{aligned}$$

Because only the random noise can't be modeled by conditioning on X, we can simplify the second term:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Z) \\ &= H(Y) - h_2(p) \leq 1 - h_2(p). \end{aligned}$$

Taking $X \sim \text{Ber}(\frac{1}{2})$ achieves equality: $I(X;Y) = 1 - h_2(p)$.

Example II. Channel capacity of a Binary Erasure Channel (BEC).

Define alphabets $\mathcal{X} = \mathcal{Y} = \{0,1\}$. Any input symbol X_i has a probability of $1-\alpha$ of being retained in the output sequence and a probability of α of being erased. Schematically, we have:

img/bec.png

Examining the mutual information, we have that

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - [H(X|Y=e)P(Y=e) + H(X|Y=0)P(Y=0) + H(X|Y=1)P(Y=1)] \\ &= H(X) - [H(X)\cdot\alpha + 0\cdot P(Y=0) + 0\cdot P(Y=1)] \\ &= (1-\alpha)H(X) \end{aligned}$$

Because the entropy of a binary variable can be no larger than 1:

$$(1-\alpha)H(X) \leq 1-\alpha$$

Equality is achieved when $H(X) = 1$, that is $X \sim \text{Ber}(\frac{1}{2})$.

6.1 Information of Continuous Random Variables

Definition: The relative entropy between two probability density functions f and g is given by

$$D(f||g) = \int f(x) \log \frac{f(x)}{g(x)}\, dx.$$

Exercise: Show that $D(f||g) \geq 0$ with equality if and only if $f = g$.

Proof. Observe that that

$$\begin{aligned} D(f||g) &= \int f(x) \log \frac{f(x)}{g(x)} dx \\ &= -\int f(x) \log \frac{g(x)}{f(x)} dx \\ &= -\mathbb{E}\left[\log \frac{g(x)}{f(x)}\right] \\ &\geq -\log \mathbb{E}\left[\frac{g(x)}{f(x)}\right] \\ &= -\log \int f(x) \frac{g(x)}{f(x)}\, dx \\ &= 0. \end{aligned}$$

Equality occurs in the manner of Jensen's when $f = g$.

Definition: The mutual information between X and Y that have a joint probability density function $f_{X,Y}$ is

$$I(X;Y) = D(f_{X,Y}||f_X f_Y).$$

Definition: The differential entropy of a continuous random variable X with probability density function f_X is

$$h(X) = -\int f_X(x) \log f_X(x)\, dx = \mathbb{E}\left[-\log f_X(X)\right]$$

If X, Y have joint density $f_{X,Y}$, the conditional differential entropy is

$$h(X|Y) = -\int f_{X,Y}(x,y) \log f_{X|Y}(x|y)\, dx\, dy = \mathbb{E}[-\log f_{X|Y}(X|Y)],$$

and the joint differential entropy is

$$h(X,Y) = \int f_{X,Y}(x,y) \log f_{X,Y}(x,y)\, dx\, dy = \mathbb{E}[-\log f_{X,Y}(X,Y)].$$

6.2 Exercises

Exercise 1. Show that

$$h(X|Y) \leq h(X)$$

with equality iff X and Y are independent.

Proof. This follows since $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \leq f_X(x)$ and log is monotonic. The equality condition is true since we have equality in $f_{X|Y}(x|y) = f_X(x)$ iff X and Y are independent.

Exercise 2. Show that

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \\ &= h(X) + h(Y) - h(X,Y). \end{aligned}$$

Proof.

$$\begin{aligned} I(X;Y) &= \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} dxdy \\ &= \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x)} dxdy - \int f_{X,Y}(x,y) \log f_Y(y) dxdy \\ &= \int f_X(x) \left[\int f_{Y|X}(y|x) \log f_{Y|X}(y|x) dy \right] dx - \int f_Y(y) \log f_Y(y) dy \\ &= H(Y) - H(Y|X). \end{aligned}$$

Symmetrically the same can be shown for $I(X;Y) = H(X) - H(X|Y)$. Also

$$\begin{aligned} I(X;Y) &= \int f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} dxdy \\ &= \int f_{X,Y}(x,y) \log f_{X,Y}(x,y) dxdy - \int f_{X,Y}(x,y) \log f_X(x) dxdy - \int f_{X,Y}(x,y) \log f_Y(y) dxdy \\ &= H(X,Y) - H(X) - H(Y). \end{aligned}$$

Exercise 3. Show that

$$h(X + c) = h(X).$$

and

$$h(c \cdot X) = h(X) + \log|c|, c \neq 0.$$

Proof.

Note that

$$h(X + c) = \mathbb{E}[-\log f_X(X + c)] = \mathbb{E}[-\log f_X(X)] = h(X);,$$

since we are integrating over the same probabilities, we are integrating over the same probabilities, the expectation of the log-density is invariant to constant shifts.

Further, note that

$$h(c \cdot X) = \mathbb{E}[-\log f_X]$$

.

To compute $h(c \cdot X)$, we start by considering the density function $p(c \cdot X)$. Set $y = c \cdot X$, yielding $dy = cdx$. We must have

$$\int p(y)\, dy = 1 = \int p(cx) \cdot c\, dx.$$

To satisfy this equality, it follows that $p(y) = \frac{p(x)}{c}$.

Therefore,

$$\begin{aligned} h(Y) &= -\int p(y)\log p(y)\,dy \\ &= -c\int p(cx)\log p(|cx|)\,dx \\ &= -c\int \frac{p(x)}{c}\log\frac{p(x)}{|c|}\,dx \\ &= -\int p(x)[\log p(x) - \log(|c|)]\,dx \\ &= h(X) + \log|c|. \end{aligned}$$

We have introduced the absolute value on c to satisfy the domain of the logarithm function.

6.3 Examples

***Example I:* Differential entropy of a uniform random variable $U \sim$ Uni(a, b).**

- Remember that the distribution of a uniform random variable is

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

The differential entropy is simply:

$$h(X) = \mathbb{E}[-\log f_X(x)] = \log(b-a)$$

- Notice that the differential entropy can be negative or positive depending on whether $b-a$ is less than or greater than 1. In practice, because of this property, differential entropy is usually used as means to determine mutual information rather than by itself.

***Example II:* Differential entropy of a Gaussian random variable $X \sim \mathcal{N}(0, \sigma^2)$.**

- Remember that the distribution of a Gaussian random variable is $f(X) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}x^2}$.

The differential entropy is:

$$h(X) = \mathbb{E}[-\log f(X)]$$

For simplicity, convert the base to e:

$$\begin{aligned}
h(X) &= \frac{1}{\ln 2}\mathbb{E}[-\ln f(X)] \\
&= \frac{1}{\ln 2}\mathbb{E}\left[\frac{1}{2}\ln 2\pi\sigma^2 + \frac{1}{2\sigma^2}X^2\right] \\
&= \frac{1}{\ln 2}\left[\frac{1}{2}\ln 2\pi\sigma^2 + \mathbb{E}\left[\frac{1}{2\sigma^2}X^2\right]\right] \\
&= \frac{1}{\ln 2}\left[\frac{1}{2}\ln 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sigma^2\right] \\
&= \frac{1}{\ln 2}\left[\frac{1}{2}\ln 2\pi e\sigma^2\right] = \frac{1}{2}\log 2\pi e\sigma^2
\end{aligned}$$

- Per Exercise 3, differential entropies are invariant to constant shifts. Therefore this expression represents the differential entropy of all Gaussian random variables regardless of mean.
- *Claim:* The Gaussian distribution has maximal differential entropy, i.e. for all random variables $X \sim f_X$ with second moment $E[X^2] \leq \sigma^2$ and Gaussian random variable $G \sim \mathcal{N}(0, \sigma^2)$ then $h(X) \leq h(G)$. Equality holds if and only if $X \sim \mathcal{N}(0, \sigma^2)$.

  Proof:

  $$\begin{aligned}
  0 \leq D(f_X \| G) &= \mathbb{E}\left[\log \frac{f_X(X)}{f_G(X)}\right] \\
  &= -h(X) + \mathbb{E}\left[\log \frac{1}{f_G(X)}\right] \\
  D(f_X \| G) &= -h(X) + \mathbb{E}\left[\log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{\frac{X^2}{2\sigma^2}}{\ln 2}\right]
  \end{aligned}$$

  Because the second moment of X is upper bounded by the second moment of G:

  $$\begin{aligned}
  0 \leq D(f_X \| G) &\leq -h(X) + \mathbb{E}\left[\log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{\frac{G^2}{2\sigma^2}}{\ln 2}\right] \\
  &\leq -h(X) + \mathbb{E}\left[\log \frac{1}{f_G(G)}\right] = -h(X) + h(G)
  \end{aligned}$$

  Rearranging:

  $$h(X) \leq h(G)$$

  □

Example III: Channel capacity of an Additive White Gaussian Noise channel (AWGN) that is restricted by power p

- Power constraint upper bounds the second moment of X_i, i.e. $p \geq E\left[X_i^2\right]$.
- Remember that the AWGN channel is a channel in which inputs X_i are corrupted by a sequence of iid additive Gaussian noise terms $W_i \sim \mathcal{N}(0, \sigma^2)$ to produce outputs Y_i.
- The Channel Coding Theorem in this setting states that:

  $$C(p) = \max_{E[X^2] \leq p} I(X;Y)$$

  Where $C(p)$ represents the 'capacity'; the maximal rate of reliable communication when constrained to power p.

7 Constraints and communication theory

Note that the encoder is equivalent to a "codebook" framed as follows:

$$c_n = \{X^n(1), X^n(2), \ldots, X^n(M)\}.$$

Here, the decoder is equivalent to the mapping $\hat{J}(.)$.

In this context, a scheme is defined as an "encoder-decoder" pair. Equivalently, this can be framed in terms of a "codebook-mapping" pair.

Definition. The *rate* is defined as

$$\text{rate} = \frac{\log M}{n} = \frac{\log |C_n|}{n} \frac{\text{bits}}{\text{channel use}}.$$

where M is the number of messages, and n is the number of channel uses. Note that M is equivalent to the size of the codebook $|C_n|$.

The probability of error can be computed as

$$P_e = P(\hat{J} \neq J).$$

Sometimes we also have a transmission constraint:

$$\frac{1}{n}\sum_{i=1}^{n} \Lambda(X_i) \leq P,$$

where Λ defines a cost function.

Example. The most common physically meaningful cost constraint pertains to the power of an electromagnetic signal. In particular, in wireless communication, we have:

$$\Lambda(x) = x^2.$$

Another example is magnetic storage media, which might have a different cost of encoding.

Recall the notion of capacity, where

$$C = \text{maximal rate of reliable communication.}$$

Further, we had the informational capacity, defined as follows:

$$C^{(I)} = \begin{cases} \max\limits_{P_X} I(X;Y); & \text{without a transmission constraint.} \\ \max\limits_{P_X : \mathbb{E}\Lambda(X) \leq P} I(X;Y); & \text{with a constraint.} \end{cases}$$

Theorem. Recall the channel coding theorem, which states the remarkable fact that

$$C = C^{(I)}.$$

Recall the following results.

1. If $G \sim \mathbb{N}(0, \sigma^2)$, then $h(G) = \frac{1}{2} \log 2\pi e \sigma^2$.
2. If X is any random variable such that $\mathbb{E}[X^2] \leq \sigma^2$ (i.e. the second moment is constrained), then $h(X) \leq h(G)$.

We now go back to example 3 from the previous section

Example III. Consider the additive white Gaussian noise (AWGN) channel, defined as D2

(In particular, when we draw diagrams with perpendicular inputs as we have done here, we mean that X and W are independent.)

And further, suppose that transmission is restricted to a power p. Namely, suppose

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 \leq p.$$

Let $C(P)$ denote the maximal rate of reliable communication constrained to power P. The channel coding theorem states that

$$C(P) = \max_{P_X : \mathbb{E}[X^2] \leq p} I(X;Y)$$

If $\mathbb{E}[X^2] \leq p$, then

$$I(X;Y) = h(Y) - h(Y|X)$$

Since differential entropy is invariant to constant shifts, we can write:

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y - X|X) \\ &= h(Y) - h(W|X) \\ &= h(Y) - h(W). \end{aligned}$$

Since $\text{Var}(Y) = \text{Var}(X) + \text{Var}(W) \leq P + \sigma^2$,

$$\begin{aligned} &\leq h(\mathbb{N}(0, p + \sigma^2)) - h(\mathbb{N}(0, \sigma^2)) \\ &= \frac{1}{2} \log 2\pi e(p + \sigma^2) - \frac{1}{2} \log 2\pi e \sigma^2 \\ &= \frac{1}{2} \log \left(1 + \frac{p}{\sigma^2}\right). \end{aligned}$$

We now try to find a distribution where this bound is achieved. To achieve equality, we require $\text{Var}(Y) = \text{Var}(X) + \text{Var}(W) = p + \sigma^2$.

In particular, let $X \sim \mathbb{N}(0, p)$, which satisfies the equality, i.e.

$$X \sim \mathbb{N}(0, p) \implies C(p) = \frac{1}{2} \log \left(1 + \frac{p}{\sigma^2}\right)$$

Note that $\frac{p}{\sigma^2}$ is known as the *signal-to-noise ratio.*

Rough geometric intuition:

Note that the power constraint can be expressed as

$$\sqrt{\sum_{i=1}^{n} X_i^2} \leq \sqrt{np}.$$

Think of $X^n(i)$ as points in n-dimensional Euclidean space. Then they lie on a sphere of radius $\sqrt{np}$.

Then, observe that

$$\frac{1}{n} \sum_{i=1}^{n} W_i^2 \approx \sigma^2 \Leftrightarrow \sqrt{\sum_{i=1}^{n} W_i^2} \approx \sqrt{n\sigma^2}.$$

The channel output can be expressed as

$$\mathbb{E}\left[\sum_{i=1}^{n} Y_i^2\right] = \sum_{i=1}^{n} \mathbb{E}[X_i^2] + \mathbb{E}[W_i^2] + \underbrace{\mathbb{E}[X_i W_i]}_{=0} \leq np + n\sigma^2.$$

Geometrically, we would like the "noise balls" to be disjoint; i.e. they should not intersect, so we can reliably discern which message point is sent.

We now want to consider bounds on the number of messages we can send. In particular, consider

fix

$$\begin{aligned}
\text{\# of messages} &\leq \frac{\text{Vol(Ball of radius } \sqrt{n(p+\sigma^2)})}{\text{Vol(Vall of radius } \sqrt{n\sigma^2})}. \\
&= \frac{k_n(\sqrt{n(p+\sigma^2)})^2}{k_n(\sqrt{n\sigma^2})^n} = \left(\frac{p+\sigma^2}{\sigma^2}\right)^{n/2} = \left(1+\frac{p}{\sigma^2}\right)^{n/2}
\end{aligned}$$

Therefore, the rate can be bounded by

$$\text{rate} = \frac{1}{2}\log\frac{\text{\# of messages}}{n} \leq \frac{1}{2}\log\left(1+\frac{p}{\sigma^2}\right).$$

7.1 Joint Asymptotic Equipartition Principle

Consider X, Y which have finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, where

$$(X,Y) \sim P_{X,Y}; \quad X \sim P_X; \quad Y \sim P_Y.$$

Here, the pairs

$$(X_i, Y_i);\ \text{iid}\ \sim (X,Y),$$

where

$$p(x^n) = \prod_{i=1}^{n} P_X(x_i),$$

$$p(y^n) = \prod_{i=1}^{n} P_Y(y_i).$$

$$p(x^n, y^n) = \prod_{i=1}^{n} P_{X,Y}(x_i, y_i).$$

Definition. The set of jointly typical sequences is defined as

$$A_\epsilon^n(X,Y) = \{(x^n,y^n) : \left|-\frac{1}{n}\log P(x^n) - H(X)\right| \leq \epsilon; \quad \left|-\frac{1}{n}\log P(y^n) - H(Y)\right| \leq \epsilon; \quad \left|-\frac{1}{n}\log P(x^n,y^n) - $$

Part A. If (X^n, Y^n) are formed by iid $(X_i, Y_i) \sim (X,Y)$, then

1. $P((X^n, Y^n) \in A_\epsilon^{(n)}(X,Y)) \to 1$, as $n \to \infty$ (basically follows directly from the original AEP on each subpart of the definition).
2. $2^{n(H(X,Y)-\epsilon)} \leq |A_\epsilon^{(n)}(X,Y)| \leq 2^{n(H(X,Y)+\epsilon)}$ (proof left to scribers, basically follows from original AEP).

Part B. If $(\tilde{X}^n, \tilde{Y}^n)$ are formed by iid $(\tilde{X}_i, \tilde{Y}_i) \sim (\tilde{X}, \tilde{Y})$ where $P_{\tilde{X},\tilde{Y}} = P_X P_Y$.

Then:

$$(1-\epsilon)2^{-nI(X,Y)+3\epsilon} \leq P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}(X,Y)) \leq 2^{-nI(X,Y)-3\epsilon};$$

for all $\epsilon > 0$ and (analytical details to be covered next time). Requires large n.

Intuition. Suppose $\tilde{X}, \tilde{Y}$ are generated independently, how likely is it to look like it came from a joint distribution? Answer: Exponentially unlikely.

8 Channel Capacity Theorem

Recall: the communication problem setting.

img/com.png

Rate of communication: number of bits per channel use, i.e.

$$\text{rate} = \frac{\log M}{n} \frac{\text{bits}}{\text{channel use}}$$

Define probability of error as

$$P_e = P(\hat{J} \neq J).$$

Main result:

$$C = \max_{P_X} I(X; Y).$$

Here, we will not concern ourselves with power / cost constraint.

We will break down this result into two sub-results. Equivalent to:

- Direct part: If $R < \max_{P_X} I(X;Y)$, then R is achievable. This means, that there exist schemes with rate $\geq R$, and $P_e \to 0$.
- Converse part: If $R > \max_{P_X} I(X;Y)$, then R is not achievable.

In this section, we will prove the direct part of the theorem.

8.1 Joint AEP

Recall the setting. Consider a pair of random variables $(X, Y) \sim P_{X,Y}$ with finite alphabets $\mathcal{X}$, $\mathcal{Y}$. This implies that the pair

$$(X, Y) \text{ has alphabet } \mathcal{X} \times \mathcal{Y},$$

here $\times$ represents the Cartesian product over sets. The jointly typical set

$$\begin{aligned} A_\epsilon^{(n)}(X,Y) = \{(X^n, Y^n) :& \left| -\frac{1}{n} \log p(X^n) - H(X) \right| \leq \epsilon, \\ & \left| -\frac{1}{n} \log p(Y^n) - H(Y) \right| \leq \epsilon, \\ & \left| -\frac{1}{n} \log p(X^n, Y^n) - H(X,Y) \right| \leq \epsilon\} \end{aligned}$$

Note that Part A of the joint AEP states that:

- If $(X_i, Y_i) \sim (X, Y)$, then for any $\epsilon > 0$,

$$P((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1.$$

- $(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |A_\epsilon^{(n)}(X,Y)| \leq 2^{nH(X,Y)+\epsilon}$ essentially for all large n.
- Suppose now $\tilde{X}^n \overset{d}{=} X^n$ and $\tilde{Y}^n \overset{d}{=} Y^n$ and $\tilde{X}$ and $\tilde{Y}$ are independent. Then

$$\tilde{X}^n \approx \text{uniformly distributed on } A_\epsilon^n(X).$$

$$\tilde{Y}^n \approx \text{uniformly distributed on } A_\epsilon^n(Y).$$

  and, since $\tilde{X}^n$ and $\tilde{Y}^n$ are independent, the joint distribution

$$(\tilde{X}^n, \tilde{Y}^n) \approx \text{uniformly distributed on } A_\epsilon^n(X,Y).$$

  It follows that

$$P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^n(X,Y)) \approx \frac{|A_\epsilon^n(X,Y)|}{|A_\epsilon^n(X) \times A_\epsilon^n(Y)|} \approx \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{-nI(X;Y)}.$$

- Formally stated, we find that for all $\epsilon > 0$, for sufficient large n,

$$(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^n(X,Y)) \leq 2^{-n(I(X;Y)-3\epsilon)}.$$

Interpretation of mutual information: quantifies "how unlikely two sequences that are independent appear that they are jointly typical?"

8.2 Relation of AEP to Communication Problem

Idea. (Proof of direct part of Communication Theorem)

- Randomly select codewords of the codebook from the typical set $A_\epsilon^{(n)}(X)$.

  img/random-cw.pdf

- Suppose we encode a codeword as $X^n(J)$. Then

$$P(Y^n \text{ is jointly typical with } X^n(J)) \approx 1.$$

  Further,

$$P(Y^n \text{ is jointly typical with some } X^n(i) \text{ for a particular } i \text{ not set}) \approx 2^{-nI(X;Y)}.$$

  Applying the previous result with a union bound:

$$P(Y^n \text{ is jointly typical with any of the codewords not sent}) \approx \text{very small, provided that:}$$

$$R < I(X;Y).$$

- Implies that: Joint typicaly decoding will be reliable, for $R < I(X;Y)$ (i.e. get you very small probability of error).

Proof of direct part. Fix P_X and $R < I(X;Y)$. We need to show that R is an achievable rate for reliable communication. Take $\epsilon > 0$ sufficiently small such that $R < I(X;Y) - 3\epsilon$. Generate a codebook C_n of size $M = \lceil 2^{nR} \rceil$ randomly:

$$\text{take } X^n(1), X^n(2), \dots, X^n(m) \text{ iid, each iid} \sim P_X.$$

Then, the jointly typical decoding rule states that

$$\hat{J} = (\hat{Y^n}) = \begin{cases} j; & \text{if } (X^n(j), Y^n) \in A_\epsilon^{(n)}(X,Y) \text{ and } (X^n(k), Y^n) \notin A_\epsilon^{(n)}(X,Y) \quad \forall k \neq j \\ e \quad \text{(error);} & \text{otherwise.} \end{cases}$$

Our rough discussion states that with very high probability, we will find the true code word that was sent. Consider one possible codebook c_n and a decoding rule. Let the probability of error be

$$P_e(c_n) = P(\hat{J} \neq J | C_n = c_n) :$$

Then

$$\mathbb{E}[P_e(c_n)] = P(\hat{J} \neq J) = \sum_{j=1}^{M} P(\hat{J} \neq J | J = j) P(J = j) = P(\hat{J} \neq J | J = 1).$$

$$\leq P((X^n(1), Y^n) \notin A_\epsilon^{(n)}(X,Y) | J = 1) + \sum_{j=2}^{M} P((X^n(j), Y^n) \in A_\epsilon^{(n)}(X,Y) | J = 1)$$

In the last inequality, we have used a union bound: either

- the Y sequence is not jointly typical with the message sent,
- or it is jointly typical with one of the other codewords sent.

This last quantity is equal to

$$\begin{aligned} & P((X^n(1), Y^n) \notin A_\epsilon^{(n)}(X,Y) | J = 1) + \sum_{j=2}^{M} P((X^n(j), Y^n) \in A_\epsilon^{(n)}(X,Y) | J = 1) \\ & = P((X^n, Y^n) \notin A_\epsilon^{(n)}(X,Y)) + (M-1) P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}(X,Y)) \\ & \leq 2^{-n(I(X;Y) - 3\epsilon - R)}. \end{aligned}$$

Note that in particular there exists a codebook c_n such that $|c_n| \leq 2^{nR}$ and $P_e(c_n) \leq \mathbb{E}[P_e(c_n)]$.

This implies that there exists a sequence of codebooks, $\{c_n\}_{n \geq 1}$ with $|c_n| \geq 2^{nR}$ and vanishing $P_e(c_n) \to 0$. And in particular, this means that R is an achievable rate for reliable communication. ■

There are a couple of problematic aspects of this proof:

- We have shown the existence of the codebooks, but not constructed one explicitly.
- Even if you were to find the codebook, they don't necessarily have good structure (codebook might be exponentially large, and have other undesirable properties.)

Note, our notation of reliability is

$$P_e = P(\hat{J} \neq J) = \sum_{j=1}^{M} P(\hat{J} \neq J | J = j) P(J = j).$$

One can consider a more stringent criterion:

$$P_{max} = \max_{1 \leq j \leq m} P(\hat{J} \neq J | J = j).$$

Exercise. Given c_n with $P_e(c_n)$, there exists a codebook c_n' such that $|c_n'| \geq \frac{1}{2}|c_n|$ and $P_{max}(c_n') \leq 2P_e(c_n)$. In this case, if $|c_n| = 2^{nR}$, then $|c_n'| \geq \frac{1}{2}2^{nR} \implies$ rate $\geq \frac{\log \frac{1}{2}2^{nR}}{n} = R - \frac{1}{n}$.

Next week: we will discuss practical constructions of these codebooks. We still need to prove the converse part as well.

9 Channel Coding Theorem; Converse Part

In this section, we will discuss the proof of our main theorem in the communication setting.

Recall the communication setting:

$$J \sim \text{Unif}\{1, 2, \ldots, m\} \rightarrow \text{encoder } (X_n) \rightarrow \text{memoryless channel } P_{Y|X}; Y^n \rightarrow \text{decoder } \hat{J}$$

Main result:

$$C = C^{(I)} = \max_{P_X} I(X; Y).$$

Last week, we showed that R is achievable if $R < C^{(I)}$. In this section, we will show the converse, i.e. if $R > C^{(I)}$, then R is not achievable.

Theorem. *(Fano's inequality) Let X be a discrete random variable, and $\hat{X}(Y)$ be a guess of X based on Y. Let $P_e = P(X \neq \hat{X})$. Then:*

$$H(X|Y) \leq h_2(P_e) + P_e \log(|\mathcal{X}| - 1).$$

Proof. Intuition: Fano's inequality relates the notion of conditional entropy and the probability of error.

Let $V = 1\left\{X \neq \hat{X}\right\}$. By the data processing inequality, we have that

$$\begin{aligned}
H(X|Y) &\leq H(X, V|Y) \\
&= H(V|Y) + H(X|V, Y) && \text{(chain rule)} \\
&\leq H(V) + \sum_{v,y} H(X|V = v, Y = y)P(V = v, Y = y) \\
& && \text{(conditioning reduces entropy)} \\
&= H(V) + \sum_{y} \underbrace{H(X|V = 0, Y = y)P(V = 0, Y = y)}_{0} + \sum_{y} \underbrace{H(X|V = 1, Y = y)}_{\leq \log(|\mathcal{X}|-1)} P(V = 1, Y = y) \\
&\leq H(V) + P(V = 1)\log(|\mathcal{X}| - 1) \\
&= h_2(P_e) + P_e \log(|\mathcal{X}| - 1)
\end{aligned}$$

□

Remark. *Often, the weakened version of Fano's inequality is often used:*

$$H(X|Y) \leq 1 + P_e \log |\mathcal{X}|,$$

or equivalently

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}.$$

Proof. (Proof of converse part of channel coding theorem.)

For any scheme, consider

$$\begin{aligned}
\log M - H(J|Y^n) &= H(J) - H(J|Y^n) \\
&= I(J; Y^n) \\
&= H(Y^n) - H(Y^n|J) \\
&= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, J) && \text{(by the chain rule)} \\
&\leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|Y^{i-1}, X_i, J) && \text{(conditioning reduces entropy)} \\
&= \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) \\
& \quad \text{(memorylessness of the channel, implying that } Y_i - X_i - (Y^{i-1}, J)\text{)} \\
&= \sum_{i=1}^{n} I(X_i; Y_i) \\
&\leq nC^{(I)} && \text{(since } C^{(I)} \text{ is the maximal mutual information)}
\end{aligned}$$

Now, consider any scheme with a rate $\frac{\log M}{n} \geq R$. By the weakened version of Fano, we have

$$\begin{aligned} P_e &\geq \frac{H(J|Y^n) - 1}{\log M} \\ &\geq \frac{\log M - nC^{(I)} - 1}{\log M} \\ &\geq 1 - \frac{C^{(I)}}{R} - \frac{1}{nR} \to 1 - \frac{C^{(I)}}{R}. && \text{(as } n \to \infty\text{)} \end{aligned}$$

But notice that if $R > C^{(I)}$, then the P_e must be lower bounded by a positive value. So this sequence of schemes cannot have a nonvanishing probability of error.

$$\boxed{\text{If } R > C^{(I)} \text{ then } R \text{ is not achievable.}}$$

This concludes the proof. □

Remark. *(Some notes on this proof).*

1. *Communication with feedback: $X_i(J, Y^{i-1})$. This is a perhaps more powerful encoder - since the encoder can adapt to what it has seen so far. However, one can verify that the proof from before holds verbatim. Therefore,*

   $$C = C^{(I)} \qquad \text{(with or without feedback)}$$

   However - P_e will vanish much more quickly, and the resulting schemes will be much more simple.

   Consider the example of communicating to the erasure channel with feedback. Earlier, we found that the capacity is given by

   $$C = 1 - \alpha \frac{\textit{bits}}{\textit{channel use}}$$

   With feedback, just repeat each information bit until it gets through. Then, one average, we will need $\frac{1}{1-\alpha}$ channel uses per information bit that we want to send. Hence, the rate achieved will be

   $$1 - \alpha \frac{\textit{bits}}{\textit{channel use}}$$

   This protocol has 0 probability of error, since we can just wait until the bit gets through.

2. *In the proof of the direct part, we showed mere existence of schemes; i.e. existence of codebooks c_n with size $|c_n| \geq 2^{nR}$ and small P_e. For practical schemes, note that LDPC codes and polar codes are concrete ways to construct these codebooks (see EE388).*

3. *Note that the proof of the direct part assumed finite alphabets. This carries over to a general case by approximation / quantization.*

4. *How do communication limits change if we want the maximal probability of error P_{max} to be small, instead of the average probability of error P_e? Recall the definitions:*

$$P_e = P(\hat{J} \neq J) = \frac{1}{m}\sum_{j=1}^{m} P(\hat{J} \neq j | J = j).$$

$$P_{max} = \max_{1 \leq j \leq m} P(\hat{J} \neq j | J = j).$$

But: let's look at the "better half" of the codebook. Consider the set of messages

$$|\left\{1 \leq j \leq M : P(\hat{J} \neq j | J = j) \geq 2P_e\right\}| \geq \frac{M}{2} \qquad \text{(by Markov's inequality)}$$

Given c_n with $|c_n| = M$ and P_e, there exists c_n' with $|c_n'| \geq \frac{M}{2}$ and $P_{max} \leq 2P_e$ - just take the messages in this better set. Then:

$$\textit{rate of } c_n' \geq \frac{\log \frac{M}{2}}{n} = \frac{\log M}{n} - \frac{1}{n}.$$

If there exist schemes of rate $\geq R$ with $P_e \to 0$, then there exist schemes of rate $\geq R - \epsilon$ with $P_{max} \to 0$.

In conclusion,

$$C = C^{(I)} \qquad \text{(under either } P_e \text{ or } P_{max}\text{)}$$

10 Lossy Compression & Rate Distortion Theory

10.1 Lossy compression problem setting

- Let U_i iid $\sim U$.
- Let the compressor compress the source to n bits.

$$U_1, U_2, \ldots, U_n \to \text{compressor / encoder} \to \text{decoder} \to V_1, V_2, \ldots, V_n$$

- Compression rate is defined as

$$\frac{n}{N}\frac{\text{bits}}{\text{source symbol}}$$

- Specify a distortion criterion d, and we will look at the expected per-symbol distortion; referred to as the "distortion" achieved.

$$D = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} d(U_i, V_i)\right]$$

- In lossy compression, we may allow D to be positive, but we want to constrain D.
- In general, there will be a tension between the distortion and the rate. We would like to identify the tradeoff. Of course, if we force distortion $D = 0$, then the best rate is the entropy. More generally, if we agree to incur a positive distortion, we can get away with smaller rate (less than the entropy).
- Concretely, when we parametrize a scheme, we need to specify:

$$\text{scheme} = (N, n, \text{encoder}, \text{decoder}).$$

Definition. *A pair (R, D) is said to be achievable if for all $\epsilon > 0$ there exists a scheme such thatits rate*

$$\frac{n}{N} \leq R + \epsilon \quad \textit{and} \quad \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} d(U_i, V_i)\right] \leq D + \epsilon.$$

Definition. *The rate distortion function $R(D)$ is defined to be*

$$R(D) = \inf\left\{R' : (R', D) \textit{ is achievable }\right\}.$$

Note that the rate distortion function is the minimal rate, optimized across all the possible schemes in the world.

Definition. *The informational rate distortion function is given by*

$$R^{(I)}(D) = \min_{\mathbb{E}d(U,V)\leq D} I(U;V)$$

Note that in the setting when you have continuous data, the notion of rate distortion is perhaps more important, because it does not make sense to talk about lossless compression.

Theorem. *(Main result)*

$$R(D) = R^{(I)}(D)$$

10.2 Qualitative analysis of $R(D)$

What does $R(D)$ look like (qualitatively?) Assume a discrete source, which we can compress losslessly with a rate equal to the entropy.

img/rdf.png

Note that $R(D)$ is monotone decreasing and convex. This is intuitive – we can always compress at at least the same rate if allowed higher distortion. $R(D)$ takes its maximal value at $D = 0$. On the other end, we see that $R(D)$ reaches its minimal value of 0 at $D_{max} = \min_v \mathbb{E}[d(U, v)]$. If we are willing to accept distortion of D_{max} we can simply encode 0 bits and always decode as v.

Claim. *$R(D)$ is convex, i.e. for all $0 \leq \alpha \leq 1$, D_0, D_1, we have that*

$$R(\alpha D_0 + (1-\alpha)D_1) \leq \alpha R(D_0) + (1-\alpha)R(D_1).$$

Proof outline. Consider the "time sharing" scheme for encoding the source symbols $(U_1, \ldots, U_N)$. Take the first αN source symbols and encode them with optimal distortion D_0, and the last $(1 - \alpha)N$ source symbols and encode them with optimal distortion D_1. The total expected distortion is $\alpha D_0 + (1-\alpha)D_1$. Then the minimal rate across all schemes is at most the rate for this particular scheme:

$$R(\alpha D_0 + (1-\alpha)D_1) \leq \alpha R(D_0) + (1-\alpha)R(D_1).$$

10.3 Examples

1. Let $U \sim \text{Ber}(p)$ with $p \leq \frac{1}{2}$ and define the Hamming distortion as

$$d(u,v) = \begin{cases} 0; & u = v \\ 1; & u \neq v \end{cases}$$

In this setting U and V take values in $\mathcal{U}$ and $\mathcal{V}$, where $\mathcal{U} = \mathcal{V} = \{0,1\}$. We claim that

$$R(D) = \begin{cases} h_2(p) - h_2(D); & 0 \leq D \leq p \\ 0; \quad D > p. \end{cases}$$

This function is convex, since in the region $0 \leq D \leq p$, the function takes the value of a constant minus the binary entropy function (which is concave). When $D > p$, we can take the reconstruction to be all zeros.

Conditioning reduces entropy, so we obtain

Proof. Consider the case when $0 \leq D \leq p$. For any U, V such that $U \sim \text{Ber}(p)$ and $\mathbb{E}[d(U,v)] = P(U \neq V) \leq D \leq p \leq 1/2$, consider

$$I(U;V) = H(U) - H(U|V) = H(U) - H(U \oplus_2 V|V)$$

Conditioning reduces entropy, so we obtain

$$\begin{aligned} H(U) - H(U \oplus_2 V|V) &\geq H(U) - H(U \oplus_2 V) \\ &= h_2(p) - h_2(P(U \neq V)) \end{aligned}$$

Equality in the above inequality is achieved when $U \oplus_2 V$ and V are independent.

Since the binary entropy function h_2 is monotonic increasing on the interval $[0, \frac{1}{2}]$, we know that

$$h_2(p) - h_2(P(U \neq V)) \geq h_2(p) - h_2(D).$$

Thus, $I(U;V) \geq h_2(p) - h_2(D)$, implying that

$$\begin{aligned} R(D) &= R^{(I)}(D) \\ &= \min_{\mathbb{E}[d(U,V) \leq D]} I(U;V) \\ &\geq h_2(p) - h_2(D). \end{aligned}$$

To show that equality is achievable, we can demonstrate that the two equality conditions above are satisfied. This is straightforward - essentially we have to find U, V such that

- $U \oplus_2 V$ is independent of V and
- $U \oplus_2 V \sim \text{Ber}(D)$.

□

2. Now, consider $U \sim \mathbb{N}(0, \sigma^2)$. We claim that

$$R(D) = \begin{cases} \frac{1}{2}\log(\sigma^2/D); & 0 < D \leq \sigma^2; \\ 0; & D > \sigma^2. \end{cases}$$

   Note that this function is convex, and for allowed distortion D greater than the variance σ^2, we don't need any bits to describe the reconstruction, since it can be taken to be always zero.

   Since this is an analog source, the entropy is infinite, so we can't expect to describe it and get zero distortion for a fixed number of bits per source symbol.

   We will ccomplete the proof of this result in the next section.

11 Method of Types

Notation: Denote $x^n = \{x_1, \ldots, x_n\}$ with $x_i \in \mathcal{X} = \{1, \ldots, r\}$ and

$$N(a|x^n) = \sum_{i=1}^{n} \mathbf{I}_{\{x_i = a\}}$$
$$\mathrm{P}_{x^n}(a) = \frac{N(a|x^n)}{n}.$$

Definition. *The empirical distribution of x^n is the probability vector ($P_{x^n}(1), \ldots, P_{x^n}(r)$).*

Definition. $\mathbf{P}_n$ *denotes the collection of all empirial distributions of sequences of length n.*

Definition. *For P $\in \mathbb{P}_n$, the type class or type of P is $T(P) = \{x^n : P_{x^n} = P\}$.*

Theorem. *The number of type classes for sequences of length n, $|\mathbb{P}_n|$, satisfies*

$$|\mathbb{P}_n| \leq (n+1)^{r-1}$$

Proof. Every empirical distribution P_{x^n} is determined by a vector $N(1|x^n), N(2|x^n) \ldots, N(r-1|x^n)$. This is a vector of length $r-1$, and each element can take up to $n+1$ values. Therefore, there are at most $(n+1)^{r-1}$ possibilities.

Note that for $r \geq 3$ the bound is not tight since we did not include the constraint $\sum_{a=1}^{r-1} N(a|x^n) \geq n$. □

More notation:

- For a probability mass function $Q = \{Q(x)\}_{x \in \mathcal{X}}$, we will write $H(Q)$ to denote $H(X)$ where $X \sim Q$.
- Let $Q(x^n) = \prod_{i=1}^{n} Q(x_i)$. For $S \subset \mathcal{X}^n$, we write $Q(S) = \sum_{x^n \in S} Q(x^n)$.

Theorem. *For all x^n, we have $2^{-n[H(P_{x^n})+D(P_{x^n}||Q)}$, where $H(P_{x^n})$ is referred to as the empirical entropy of x^n.*

Proof. This is a few straightforward manipulations of definitions.

$$\begin{aligned} Q(x^n) &= \prod_{i=1}^{n} Q(x_i) \\ &= 2^{\sum_{i=1}^{n} \log Q(x_i)} \\ &= 2^{\sum_{a\in\mathcal{X}} N(a|x^n)\log Q(a)} \\ &= 2^{-n[\sum_{a\in\mathcal{X}} \frac{N(a|x^n)}{n} \log \frac{1}{Q(a)}]} \\ &= 2^{-n\left[\sum_{a\in\mathcal{X}} P_{x^n}(a)\log\left(\frac{1}{Q(a)}\frac{P_{x^n}(a)}{P_{x^n}(a)}\right)\right]} \\ &= 2^{-n[H(P_{x^n})+D(P_{x^n}||Q)]} \end{aligned}$$

□

Theorem. *For all $P \in \mathbb{P}_n$, we have that*

$$\frac{1}{(n+1)^{r-1}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}$$

Proof. Proof is straightforward. □

Fill in later

Theorem. *For any probability mass function Q and any empirical distribution $P \in \mathbb{P}_n$,*

$$\frac{1}{(n+1)^{r-1}} 2^{-nD(P||Q)} \leq Q(T(P)) \leq 2^{-nD(P||Q)}.$$

That is, on an exponential scale - the probability that the sequence looks like it came from source P if the data is generated iid from distribution Q is very unlikely.

Note that in the expression above, $D(P||Q)$ is between P, the "wrong" source and Q the "true" source. This is different from the cost of mismatch in lossless compression; $D(p||q)$ is such that p is the true source and q is the wrong source.

12 Strong, Conditional, and Joint Typicality

Definition. *A sequence $x^n \in \mathcal{X}^n$ is strongly δ-typical with respect to a probability mass function $\mathcal{P} \in \mathcal{M}(\mathcal{X})$ if*

$$|P_{x^n}(a) - P(a)| \leq \delta P(a); \qquad \forall a \in \mathcal{X}.$$

Definition. *The strongly δ-typical set of p, $T_\delta(P)$ is defined as the set of all sequences that are strongly δ-typical with respect to P, that is*

$$T_\delta(P) = \{x^n : |P_{x^n}(A)\}$$