EE 376A: Information Theory
Lecture Notes

Prof. Tsachy Weissman
TA: Idoia Ochoa, Kedar Tatwawadi

January 6, 2016

Contents

Chapter 1

Introduction

Information theory is the science of operations on data such as compression, storage, and communication. It is among the few disciplines fortunate to have a precise date of birth: 1948, with the publication of Claude E. Shannon's paper entitled *A Mathematical Theory of Communication*. Shannon's Information theory had a profound impact on our understanding of the concepts in communication.

In this introductory chapter, we will look at a few representative examples which try to give a flavour of the problems which can be addressed using information theory. However note that, communication theory, is just one of the numerous fields which had a dramatic shift in the understanding due to information theory.

1.1 Lossless Compression

Consider a source that emits a sequence of symbols $U_1, U_2, \ldots$ with $U_i \in \{a, b, c\}$. The U_i are i.i.d (independently and identically distributed) according to the probability mass function

$$
\begin{aligned}
P(U = a) &= 0.7 \\
P(U = b) &= P(U = c) = 0.15
\end{aligned}
$$

Our task is to encode the source sequence into binary bits (1s and 0s). How should we do so?

The naive way is to use two bits to represent each symbol, since there are three possible symbols. For example, we can use 00 to represent a, 01 to represent b and 10 to represent c. This scheme has an expected codeword length of 2 bits per source symbol. Can we do better? One natural improvement is to try to use fewer bits to represent symbols that appear more often. For example, we can use the single bit 0 to represent a since a is the most common symbol, and 10 to represent b and 11 to represent c since they are less common. Note that this code satisfies the prefix condition, meaning no codeword is the prefix of another codeword, which allows us to decode a message consisting of stream of bits without any ambiguity. Thus, if we see the encoded sequence, 001101001101011, we can quickly decode it as follows:

$$
\underbrace{0}_{\text{a}}\underbrace{0}_{\text{a}}\underbrace{11}_{\text{c}}\underbrace{0}_{\text{a}}\underbrace{10}_{\text{b}}\underbrace{0}_{\text{a}}\underbrace{11}_{\text{c}}\underbrace{0}_{\text{a}}\underbrace{10}_{\text{b}}\underbrace{11}_{\text{c}}
$$

If we use this encoding scheme, then $\overline{L}$, which denotes the expected number of bits we use per source symbol, is

$$\overline{L} = 1 \times P(U = a) + 2 \times (P(U = b) + P(U = c)) = 1 \times 0.7 + 2 \times (0.15 + 0.15) = 1.3.$$

This is a significant improvement over our first encoding scheme. But can we do even better? A possible improvement is to encode two values at a time instead of encoding each value individually. For example, the following table shows all the possibilities we can get if we look at 2 values, and their respective probabilities (listed in order of most to least likely pairs). A possible prefix coding scheme is also given.

source symbols	probability	encoding
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

Note that this scheme satisfies the two important properties: 1) the prefix condition and 2) more common source symbol pairs have shorter codewords. If we use the above encoding scheme, then the expected number of bits used per source symbol is

$$\overline{L} = 0.5 \times (0.49 \times 1 + 0.105 \times 4 + 0.105 \times 3 \times 3 + 0.0225 \times 6 \times 4) = 1.1975.$$

It can be proven that if we are to encode 2 values at a time, the above encoding scheme achieves the lowest average number of bits per value (*wink* Huffman encoding *wink*).

Generalizing the above idea, we can consider a family of encoding schemes indexed by an integer k. Given an integer k, we can encode k values at a time with a scheme that satisfies the prefix condition and assigns shorter codewords to more common symbols. Under some optimal encoding scheme, it seems reasonable that the expected number of bits per value will decrease as k increases.

We may ask, what is the best we can do? Is there a lower bound on $\overline{L}$? Shannon proved that given any such source, the best we can do is $H(U)$, which is called the **Entropy** of the source . By definition, the source entropy is

$$H(U) \triangleq \sum_{u \in U} p(u) \log_2 \frac{1}{p(u)} \tag{1.1}$$

Thus, Shannon proved the following statement[1],

Theorem 1. *$\forall$ families of encoding schemes, the average codeword length, $\overline{L} \geq H(U)$.*

[1]Note that the statements of the theorems here will be informal; they will be made rigorous in later lectures.

For our example, the lower bound is thus

$$0.7 \times \log_2 \frac{1}{0.7} + 2 \times 0.15 \times \log_2 \frac{1}{0.15} \approx 1.181$$

We will also show an upper bound, namely,

Theorem 2. *$\forall \varepsilon > 0$, $\exists$ family of schemes, such that the average codeword length, $\overline{L} \leq H(U) + \varepsilon$.*

1.2 Channel Coding

Suppose we have a source that emits a stream of bits $U_1, U_2, \ldots$ The $U_i \in \{0, 1\}$ are i.i.d. Bernoulli random variables with parameter 0.5, or fair coin flips.

We want to transmit the bits U_i through a channel. Suppose the bits that are transmitted are $X_1, X_2, \ldots$ The channel is noisy and flips each bit with probability $q < 1/2$. Therefore, if $Y_1, Y_2, \ldots$ is the sequence of bits that we receive, we have

$$Y_i = X_i \oplus W_i, W_i \sim \mathbf{Ber}(q)$$

where $\oplus$ is the XOR operator.

We want to know how accurate we can be when transmitting the bits. The simplest approach is to let $X_i = U_i$, and to decode the received bit by assuming $Y_i = X_i$. Let p_e be the probability of error per source bit. Then in this case, $p_e = q < 1/2$.

Can we decrease p_e? One approach may be to use repetition encoding, i.e., send each bit k times for some k, and then decode the received bit as the value that appeared most among the k received symbols. For example, if $k = 3$, then p_e is simply the probability that the channel flipped 2 or more of the bits, which is

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

However, we need to send 3 times as many bits. To quantify this, we introduce the notion of **bit rate**, denoted by R, which is the ratio of the number of bits sent to the units of channel space used. For this scheme, our bit rate is $\frac{1}{3}$, whereas our bit rate in the previous example was 1.

Generalizing the above example, we see that as we increase k, our error rate p_e will tend to 0, but our bit rate R (which is $1/k$) tends to 0 as well. Is there some scheme that has a significant positive bit rate and yet allows us to get reliable communication (error rate tends to 0)? Again, Shannon provides the answer.

Theorem 3. *$\exists C > 0$ and $\exists$ family of schemes with $R < C$ satisfying $p_e \to 0$.*

In fact, the largest such C is known as the **channel capacity** of a channel, which represents the largest bit rate (the largest C) that still allows for reliable communication. This was a very significant and a startling revelation for the world of communication, as it was thought that zero error probability is not achievable with a non-zero bit rate.

As examples, we will consider the channel capacity of the binary symmetric channel and the additive white gaussian noise channel.

Binary Symmetric Channel

The channel capacity of a binary symmetric channel with bit-flipping probability q is

$$1 - H(X), X \sim \mathbf{Ber}(q). \tag{1.2}$$

Moreover, if we let $X \sim \mathbf{Ber}(q)$ and $Y \sim \mathbf{Ber}(p_e)$, we will see that a bit rate R such that

$$R < \frac{1 - H(X)}{1 - H(Y)}, \tag{1.3}$$

is **achievable**, whereas

$$R > \frac{1 - H(X)}{1 - H(Y)}, \tag{1.4}$$

is **unachievable**.

Additive White Gaussian Noise (AWGN) Channel

Suppose we have a source that emits a sequence of bits $U_1, U_2, \ldots, U_N$, where each U_i is i.i.d. according to $U \sim \mathbf{Ber}(\frac{1}{2})$.

However, we can only transmit real numbers $X_1, X_2, \ldots, X_n$. Also, the channel contains some noise. Specifically, if $Y_1, Y_2, \ldots, Y_n$ is the sequence of values we receive, we have

$$Y_i = X_i + N_i, N_i \sim \mathcal{N}(0, \sigma^2)$$

The **rate of transmission** is the ratio $\dfrac{N}{n}$ (which is the ratio of the number of source bits to the number of uses of the channel). We want to develop a scheme so that we can reliably reconstruct U_i from the given Y_i. One way, if we have no usage power constraint, is to make X_i a large positive value if $U_i = 1$ and X_i a large negative value if $U_i = 0$. In this manner, the noise from N_i will be trivial relative to the signal magnitude, and will not impact reconstruction too much. However, suppose there is an additional constraint on the average power of the transmitted signal, such that we require

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

for a given value p. In fact, we will see that

Theorem 4. *If the rate of transmission is* $< \frac{1}{2}\log_2\left(1 + \frac{p}{\sigma^2}\right)$*, then* $\exists$ *family of schemes that communicate reliably. And if the rate of transmission is* $> \frac{1}{2}\log_2\left(1 + \frac{p}{\sigma^2}\right)$*, then there is no family of schemes which communicates reliably.*

The ratio $\dfrac{p}{\sigma^2}$ is referred to as the signal-to-noise ratio (SNR).

1.3 Lossy Compression

Suppose we have a source that emits a sequence of values $U_1, U_2, \ldots$, where the U_i is i.i.d random variables. according to $U \sim \mathcal{N}(0, \sigma^2)$. Suppose we want to encode the source using one bit per value. Since we are representing continuous variables with discrete bits, we are employing lossy compression. Can we come up with a scheme that reconstructs the original signal as accurately as possible, based on the bits sent?

Let $B_1, B_2, \ldots$ be the bits sent. One natural scheme is to set

$$B_i = \begin{cases} 1 & \text{if } U_i \geq 0 \\ 0 & \text{if } U_i < 0. \end{cases}$$

After receiving the bits, let $V_1, V_2, \ldots$ be the reconstructed values. The **distortion** of the scheme is defined as

$$D \triangleq \mathbb{E}[(U_i - V_i)^2] \tag{1.5}$$

The optimal estimation rule for minimum mean squared error is the conditional expectation. Therefore, to minimize distortion, we should reconstruct via $V_i = \mathbb{E}[U_i \mid B_i]$. This results in

$$\begin{aligned} D &= \mathbb{E}[(U_i - V_i)^2] \\ &= Var(U_i \mid B_i) \\ &= 0.5 \times Var(U_i \mid B_i = 1) + 0.5 \times Var(U_i \mid B_i = 0) \text{ (because } U \text{ is symmetric)} \\ &= Var(U_i \mid B_i = 1) \\ &= \mathbb{E}[U_i^2 \mid B_i = 1] - (\mathbb{E}[U_i \mid B_i = 1])^2 \\ &= \sigma^2 \left(1 - \frac{2}{\pi}\right) \\ &\approx 0.363\sigma^2. \end{aligned}$$

We will see in fact that $0.363\sigma^2$ can be improved considerably, as such:

Theorem 5. *Consider a Gaussian memoryless source with mean μ and variance σ^2. $\forall \varepsilon > 0$, $\exists$ family of schemes such that $D \leq \sigma^2/4 + \varepsilon$. Moreover, $\forall$ families of schemes, $D \geq \sigma^2/4$.*

As we saw, the few examples signify the usefulness of information theory to the field of communications. In the next few chapters, we will try to build the mathematical foundations for the theory of information theory, which will make it much more convenient for us to use them later on.

Chapter 2

Entropy, Relative Entropy, and Mutual Information

In this chapter ,we will introduce certain key measures of information, that play crucial roles in theoretical and operational characterizations throughout the course. These include the entropy, the mutual information, and the relative entropy. We will also exhibit some key properties exhibited by these information measures.

Notation

A quick summary of the notation

1. **Random Variables (objects):** used more “loosely", i.e. X, Y, U, V
2. **Alphabets:** $\mathcal{X}, \mathcal{Y}, \mathcal{U}, \mathcal{V}$
3. **Specific Values:** x, y, u, v

For discrete random variable (object), U has p.m.f: $P_U(u) \triangleq P(U = u)$. Often, we'll just write $p(u)$. Similarly: $p(x, y)$ for $P_{X,Y}(x, y)$ and $p(y|x)$ for $P_{Y|X}(y|x)$, etc.

2.1 Entropy

Before we understand entropy, let us take a look at the *"surprise function"*, which will give us more intuition into the definition of entropy.

Definition 6. *"Surprise Function":*

$$s(u) \triangleq \log \frac{1}{p(u)}$$

The surprise function represents the amount of *surprise* or the amount of information a particular symbol u of a distribution holds. Intuitively the definition can be understood as follows: we would be surprised if a rare symbol ($p(u)$ is small) is observed. Thus, lower the $p(u)$, higher the *surprise*, which is what achieved by the above definition.

Definition 7. ***Entropy***: *Let U a discrete R.V. taking values in $\mathcal{U}$. The* ***entropy*** *of U is defined by:*

$$H(U) \triangleq \sum_{u \in U} p(u) \log \frac{1}{p(u)} \triangleq E[s(u)] \tag{2.1}$$

The Entropy represents the expected value of *surprise* a distribution holds. Intuitively, the more the *expected surprise* or the entropy of the distribution, the harder it is to represent.

Note: The entropy $H(U)$ is not a random variable. In fact it is not a function of the object U, but rather a functional (or property) of the underlying distribution $P_U^{(u)}, u \in \mathcal{U}$. An analogy is $E[U]$, which is also a number (the mean) corresponding to the distribution.

Properties of Entropy

Although almost everyone would have encountered the **Jensen's Inequality** in their calculus class, we take a brief look at it in a form most useful for information theory. **Jensen's Inequality:** Let Q denote a *convex* function, and X denote any random variable. Jensen's inequality states that

$$E[Q(X)] \geq Q(E[X]). \tag{2.2}$$

Further, if Q is strictly convex, equality holds iff X is deterministic. *Example*: $Q(x) = e^x$ is a convex function. Therefore, for a random variable X, we have by Jensen's inquality:

$$E[e^X] \geq e^{E[X]}$$

Conversely, if Q is a *concave* function, then

$$E[Q(X) \leq Q(E[X]). \tag{2.3}$$

Example: $Q(x) = \log x$ is a concave function. Therefore, for a random variable $X \geq 0$,

$$E[\log X] \leq \log E[X] \tag{2.4}$$

W.L.O.G suppose $\mathcal{U}$ = {1,2,...,m}

1. $H(U) \leq \log m$, with equality iff $P(u) = \frac{1}{m} \forall u$ (i.e. uniform).

   Proof:

$$\begin{aligned} H(U) &= E[\log \frac{1}{p(u)}] && (2.5)\\ &\leq \log E[\frac{1}{p(u)}] \text{ (Jensen's inequality, since log is concave)} && (2.6)\\ &= \log \sum_u p(u) \cdot \frac{1}{p(u)} && (2.7)\\ &= \log m. && (2.8) \end{aligned}$$

   Equality in Jensen, iff $\frac{1}{p(u)}$ is deterministic, iff $p(u) = \frac{1}{m}$

2. $H(U) \geq 0$, with equality iff U is deterministic.

   Proof:

$$H(U) = E[\log \frac{1}{p(u)}] \geq 0 \text{ since } \log \frac{1}{p(u)} \geq 0 \tag{2.9}$$

   The equality occurs iff $\log \frac{1}{P(U)} = 0$ with probability 1, iff P(U) = 1 w.p. 1 iff U is deterministic.

3. For a PMF q, defined on the same alphabet as p, define

$$H_q(U) \triangleq \sum_{u \in \mathcal{U}} p(u) \log \frac{1}{q(u)}. \tag{2.10}$$

   Note that this is the expected surprise function, but instead of the surprise associated with p, it is the surprise associated U, which is distributed according to PMF p, but incorrectly assumed to be having the PMF of q. The following result stipulates, that we will (on average) be more surprised if we had the wrong distribution in mind. This makes intuitive sense! Mathematically,

$$H(U) \leq H_q(U), \tag{2.11}$$

   with equality iff $q = p$.

   Proof:

$$H(U) - H_q(U) = E\left[\log \frac{1}{p(u)}\right] - E\left[\log \frac{1}{q(u)}\right] \tag{2.12}$$

$$H(U) - H_q(U) = E\left[\log \frac{q(u)}{p(u)}\right] \tag{2.13}$$

   By Jensen's, we know that $E\left[\log \frac{q(u)}{p(u)}\right] \leq \log E\left[\frac{q(u)}{p(u)}\right]$, so

$$H(U) - H_q(U) \leq \log E\left[\frac{q(u)}{p(u)}\right] \tag{2.14}$$

$$= \log \sum_{u \in \mathcal{U}} p(u) \frac{q(u)}{p(u)} \tag{2.15}$$

$$= \log \sum_{u \in \mathcal{U}} q(u) \tag{2.16}$$

$$= \log 1 \tag{2.17}$$

$$= 0 \tag{2.18}$$

   Therefore, we see that

$$H(U) - H_q(U) \leq 0.$$

   Equality only holds when Jensen's yields equality. That only happens when $\frac{q(u)}{p(u)}$ is deterministic, which only occurs when $q = p$, i.e. the distributions are identical.

Definition 8. ***Relative Entropy.*** *An important measure of distance between probability measures is relative entropy, or the Kullback–Leibler divergence:*

$$D(p||q) \triangleq \sum_{u \in \mathcal{U}} p(u) \log \frac{p(u)}{q(u)} = E\left[\log \frac{p(u)}{q(u)}\right] \tag{2.19}$$

Note that property 3 is equivalent to saying that the relative entropy is always greater than or equal to 0, with equality iff $q = p$ (convince yourself).

4. If $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) \tag{2.20}$$

Proof:

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= E\left[\log \frac{1}{p(x_1, x_2, \ldots, x_n)}\right] && (2.21) \\ &= E\left[-\log p(x_1, x_2, \ldots, x_n)\right] && (2.22) \\ &= E\left[-\log p(x_1)p(x_2)\ldots p(x_n)\right] && (2.23) \\ &= E\left[-\sum_{i=1}^{n} \log p(x_i)\right] && (2.24) \\ &= \sum_{i=1}^{n} E\left[-\log p(x_i)\right] && (2.25) \\ &= \sum_{i=1}^{n} H(X_i). && (2.26) \end{aligned}$$

Therefore, the entropy of independent random variables is the sum of the individual entropies. This is also intuitive, since the uncertainty (or surprise) associated with each random variable is independent.

2.2 Conditional and Joint Entropy

We defined the entropy of a random variable U. We also saw that when U is a joint random variable of independent variables, then $H(U)$ is the sum of the individual entropies. Can we say anything more in general for a joint random variable?

Definition 9. ***Conditional Entropy of*** X ***given*** Y

$$\begin{aligned} H(X|Y) &\triangleq \mathrm{E}\,[\log \frac{1}{P(X|Y)}] && (2.27) \\ &= \sum_{x,y} Pr[x, y] \frac{1}{\log P(x|y)} && (2.28) \\ &= \sum_{y} P(y) [\sum_{x} P(x|y) \frac{1}{\log P(x|y)}] && (2.29) \\ &= \sum_{y} P(y) H(X|y). && (2.30) \end{aligned}$$

Note: The conditional entropy is a functional of the joint distribution of (X, Y). Note that this is also a number, and denotes the "average" surprise in X when we observe Y. Here, by definition, we also average over the realizations of Y. Note that the conditional entropy is NOT a function of the random variable Y. In this sense, it is very different from a familar object in probability, the conditional expectation $E[X|Y]$ which is a random variable (and a function of Y).

Definition 10. ***Joint Entropy of*** X ***and*** Y

$$H(X, Y) \triangleq \mathrm{E}\left[\log \frac{1}{P(X, Y)}\right] \tag{2.31}$$

$$= \mathrm{E}\left[\log \frac{1}{P(X)P(Y|X)}\right] \tag{2.32}$$

Properties of conditional and Joint entropy

1. $H(X|Y) \leq H(X)$, equal iff $X \perp Y$

   Proof:

$$H(X) - H(X|Y) = \mathrm{E}\left[\log \frac{1}{P(X)}\right] - \mathrm{E}\left[\log \frac{1}{P(X|Y)}\right] \tag{2.33}$$

$$= \mathrm{E}\left[\log \frac{P(X|Y)}{P(X)} \frac{P(Y)}{P(Y)}\right] = \mathrm{E}\left[\log \frac{P(X, Y)}{P(X)P(Y)}\right] \tag{2.34}$$

$$= \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \tag{2.35}$$

$$= D(P_{x,y}||P_x \times P_y) \tag{2.36}$$

$$\geq 0 \qquad \text{equal iff } X \perp Y. \tag{2.37}$$

   The last step follows from the non-negativity of relative entropy. Equality holds iff $P_{x,y} \equiv P_x \times P_y$, i.e. X and Y are independent.

2. Chain rule for entropy:

$$H(X, Y) = H(X) + H(Y|X) \tag{2.38}$$

$$= H(Y) + H(X|Y) \tag{2.39}$$

3. Sub-additivity of entropy

$$H(X, Y) \leq H(X) + H(Y), \tag{2.40}$$

   with equality iff $X \perp Y$ (follows from the property that conditioning does not increase entropy)

2.3 Mutual Information

Definition 11. ***Mutual information between X and Y***

We now define the mutual information between random variables X and Y distributed according to the joint PMF $P(x, y)$:

$$I(X,Y) \triangleq H(X) + H(Y) - H(X,Y) \tag{2.41}$$
$$= H(Y) - H(Y|X) \tag{2.42}$$
$$= H(X) - H(X|Y) \tag{2.43}$$
$$= D(P_{x,y}||P_x \times P_y) \tag{2.44}$$

The mutual information is a canonical measure of the information conveyed by one random variable about another. The definition tells us that it is the reduction in average surprise, upon observing a correlated random variable. The mutual information is again a functional of the joint distribution of the pair (X, Y). It can also be viewed as the relative entropy between the joint distribution, and the product of the marginals

1. $I(X;Y) \geq 0$, coming from the fact that $H(Y) \geq H(Y|X)$.
2. $I(X;Y) \leq \min\{H(X), H(Y)\}$, since the conditional entropies are non-negative. The equality occurs iff there exists a deterministic function f s.t. $Y = f(X)$ or $X = f(Y)$ (so that either $H(Y|X)$ or $H(X|Y)$, respectively, is zero).
3. Properties for Markov Chains:
   We introduce the notation $X - Y - Z$ to reflect that

$$\begin{aligned}
&X \text{ and } Z \text{ are conditionally independent given } Y\\
\Leftrightarrow&(X, Y, Z) \text{ is a Markov triplet}\\
\Leftrightarrow&p(x, z|y) = p(x|y)p(z|y)\\
\Leftrightarrow&p(x|y, z) = p(x|y)\\
\Leftrightarrow&p(z|y, x) = p(z|y)
\end{aligned}$$

   For example, let X, W_1, W_2 be three independent Bernoulli random variables, with $Y = X \oplus W_1$ and $Z = Y \oplus W_2$. Then, X and Z are conditionally independent given Y, i.e., $X - Y - Z$. Intuitively, Y is a noisy measurement of X, and Z is a noisy measurement of Y. Since the noise variables W_1 and W_2 are independent, we only need Y to infer X.
   We can also show that if $X - Y - Z$, then

   (a) $H(X|Y) = H(X|Y, Z)$
   (b) $H(Z|Y) = H(Z|X, Y)$
   (c) $H(X|Y) \leq H(X|Z)$
   (d) $I(X;Y) \geq I(X;Z)$, and $I(Y;Z) \geq I(X;Z)$

   Intuitively, $X - Y - Z$ indicates that X and Y are more closely related than X and Z. Therefore $I(X;Y)$ (i.e., the dependency between X and Y) is no smaller than $I(X;Z)$, and $H(X|Y)$ (the uncertainty in X given knowledge Y) is no greater than

Chapter 3

Asymptotic Equipartition Properties

In this chapter, we will try to understand how the distribution of n-length sequences generated by memoryless sources behave as we increase n. We observe that a set of small fraction of all the possible n-length sequences occurs with probability almost equal to 1. Thus, this makes the compression of n-length sequences easier as we can then concentrate on this set.

We begin by introducting some important notation:

- For a set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality (number of elements contained on the set). For example, let $\mathcal{U} = \{1, 2, \ldots, M\}$, then $|\mathcal{U}| = M$.
- $u^n = (u_1, \ldots, u_n)$ is an n-tuple of u.
- $\mathcal{U}^n = \{u^n | \; u_i \in \mathcal{U}; \; i = 1, \ldots, n\}$. It is easy to see that $|\mathcal{U}^n| = |\mathcal{U}|^n$.
- U_i generated by a memoryless source U" implies $U_1, U_2, \ldots$ i.i.d. according to U (or P_U). That is,

$$p(u^n) = \prod_{i=1}^{n} p(u_i)$$

3.1 Asymptotic Equipartition Property (AEP)

Definition 12. *The sequence u^n is ϵ-typical for a memoryless source U for $\epsilon > 0$, if*

$$\left| -\frac{1}{n} \log p(u^n) - H(U) \right| \le \epsilon$$

or equivalently,

$$2^{-n(H(U)+\epsilon)} \le p(u^n) \le 2^{-n(H(U)-\epsilon)}$$

Let $A_\epsilon^{(n)}$ denote the set of all ϵ-typical sequences, called the typical set.

So a length-n typical sequence would assume a probability approximately equal to $2^{-nH(U)}$. Note that this applies to memoryless sources, which will be the focus on this course[1].

Theorem 13 (AEP). $\forall \epsilon > 0$, *$P\left(U^n \in A_\epsilon^{(n)}\right) \to 1$ as $n \to \infty$.*

[1] For a different definition of typicality, see e.g. [1]. For treatment of non-memoryless sources, see e.g. [2], [3].

Proof This is a direct application of the Law of Large Numbers (LLN).

$$\begin{aligned}
P\left(U^n \in A_\epsilon^{(n)}\right) &= P\left(\left|-\frac{1}{n}\log p(U^n) - H(U)\right| \le \epsilon\right) \\
&= P\left(\left|-\frac{1}{n}\log \prod_{i=1}^{n} p(U_i) - H(U)\right| \le \epsilon\right) \\
&= P\left(\left|\frac{1}{n}\left[\sum_{i=1}^{n} -\log p(U_i)\right] - H(U)\right| \le \epsilon\right) \\
&\to 1 \text{ as } n \to \infty
\end{aligned}$$

where the last step is due to the Law of Large Numbers (LLN), in which $-\log p(U_i)$'s are i.i.d. and hence their arithmetic average converges to their expectation $H(U)$.

□

This theorem tells us that with very high probability, we will generate a typical sequence. But how large is the typical set $A_\epsilon^{(n)}$?

Theorem 14. $\forall \epsilon > 0$ *and sufficiently large* n*,*

$$(1-\epsilon)2^{n(H(U)-\epsilon)} \le \left|A_\epsilon^{(n)}\right| \le 2^{n(H(U)+\epsilon)}$$

Proof The upper bound:

$$1 \ge P\left(U^n \in A_\epsilon^{(n)}\right) = \sum_{u^n \in A_\epsilon^{(n)}} p(u^n) \ge \sum_{u^n \in A_\epsilon^{(n)}} 2^{-n(H(U)+\epsilon)} = \left|A_\epsilon^{(n)}\right| 2^{-n(H(U)+\epsilon)},$$

which gives the upper bound. For the lower bound, by the AEP theorem, for any $\epsilon > 0$, there exists sufficiently large n such that

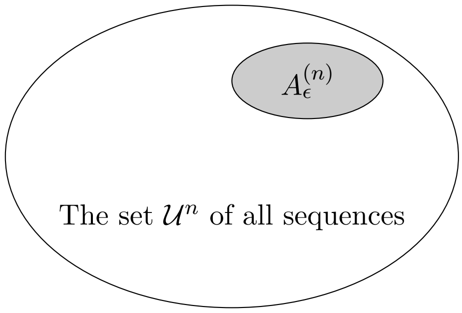
$$1-\epsilon \le P\left(U^n \in A_\epsilon^{(n)}\right) = \sum_{u^n \in A_\epsilon^{(n)}} p(u^n) \le \sum_{u^n \in A_\epsilon^{(n)}} 2^{-n(H(U)-\epsilon)} = \left|A_\epsilon^{(n)}\right| 2^{-n(H(U)-\epsilon)}.$$

□

The intuition is that since all typical sequences assume a probability about $2^{-nH(U)}$ and their total probability is almost 1, the size of the typical set has to be approximately $2^{nH(U)}$. Although $\left|A_\epsilon^{(n)}\right|$ grows exponentially with n, notice that it is a relatively small set compared to $\mathcal{U}^n$. For some $\epsilon > 0$, we have

$$\frac{\left|A_\epsilon^{(n)}\right|}{|\mathcal{U}^n|} \le \frac{2^{n(H(U)+\epsilon)}}{2^{n\log|\mathcal{U}|}} = 2^{-n(\log|\mathcal{U}|-H(U)-\epsilon)} \to 0 \text{ as } n \to \infty$$

given that $H(U) < \log|\mathcal{U}|$ (with strict inequality!), i.e., the fraction that the typical set takes up in the set of all sequences vanishes exponentially. Note that $H(U) = \log|\mathcal{U}|$ only if the source is uniformly distributed, in which case all the possible sequences are typical.



$$P\left(U^n \in A_\epsilon^{(n)}\right) \approx 1$$

$$u^n \in A_\epsilon^{(n)} \Leftrightarrow p(u^n) \approx 2^{-nH(U)}$$

$$\left|A_\epsilon^{(n)}\right| \approx 2^{nH(U)}$$

Figure 3.1: Summary of AEP

In the context of lossless compression of the source U, the AEP tells us that we may only focus on the typical set, and we would need about $nH(U)$ bits, or $H(U)$ bits per symbol, for a good representation of the typical sequences.

We know that the set, $A_\epsilon^{(n)}$ has probability 1 as n increases. However is it the smallest such set? The next theorem gives a definitive answer to the question.

Theorem 15. *For all $\delta > 0$ and all sequences of sets $B^{(n)} \subseteq \mathcal{U}^n$ such that $\left|B^{(n)}\right| \leq 2^{n[H(U)-\delta]}$,*

$$\lim_{n\to\infty} P\left(U^n \in B^{(n)}\right) = 0 \tag{3.1}$$

A visualization of Theorem 15 is shown in Figure 3.1.

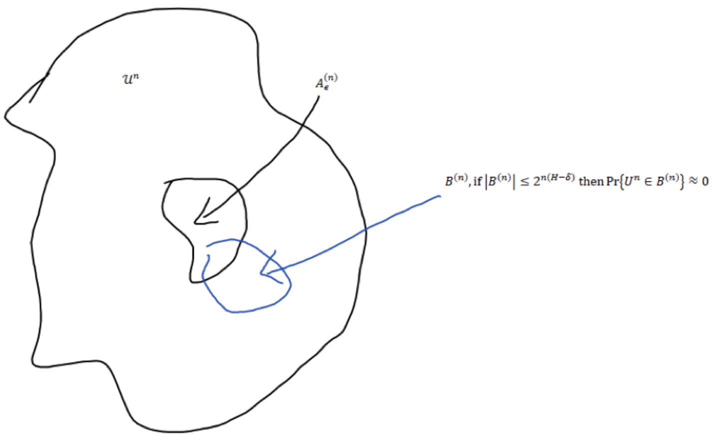


Figure 3.2: Visualization of all source sequences and ε-typical sequences.

We can justify the theorem in the following way: As n increases $|B^{(n)} \cap A_\epsilon^{(n)}| \approx 2^{-n\delta}|A_\epsilon^{(n)}|$. As every typical sequence has probability of $\approx 2^{-nH(U)}$, and is the same for every sequence, $P\left(U^n \in B^{(n)}\right) = 0$

We will next look at a simple application of the AEP for the compression of symbols generated by a discrete memoryless source.

3.2 Fixed Length (Near) Lossless Compression

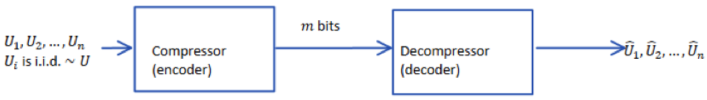


Figure 3.3: Block Diagram for Losless Compression

Suppose we have a source $U_1, \dots, U_n$ i.i.d. with distribution U. We wish to devise a compression scheme, as shown in Figure 3.2. The compressor takes a block of n source symbols and converts them into m binary bits. The decompressor does the inverse process. The rate of such a scheme (compression and decompression) is defined to be $\frac{m}{n}$ bits/source symbol.

We relax our requirements slightly: rather than insisting on strictly lossless compression, we will simply require the probability of error to be small. That is,

$$P_e = P\left(\hat{U}^n \neq U^n\right) \ll 1 \tag{3.2}$$

Definition 16 (Achievable rate)**.** *R is an achievable rate if for all $\varepsilon > 0$, there exists a scheme (n, m, compressor, decompressor) whose rate $\frac{m}{n} \leq R$ and whose probability of error $P_e < \varepsilon$.*

We are interested in the question: What is the lowest achievable rate? Theorems 17 and 18 tell us the answer.

Theorem 17 (Direct theorem)**.** *For all $R > H(U)$, R is achievable.*

Proof Fix $R > H(U)$ and $\varepsilon > 0$. Set $\delta = R - H(U) > 0$ and note that for all n sufficiently large, by Theorem 13,

$$P\left(U^n \notin A_\delta^{(n)}\right) < \varepsilon, \tag{3.3}$$

and by Theorem 14,

$$\left|A_\delta^{(n)}\right| \leq 2^{n[H(U)+\delta]} = 2^{nR}. \tag{3.4}$$

Consider a scheme that enumerates sequences in $A_\delta^{(n)}$. That is, the compressor outputs a binary representation of the index of U^n if $U^n \in A_\delta^{(n)}$; otherwise, it outputs $(0, 0, \dots, 0)$. The decompressor maps this binary representation back to the corresponding sequence in $A_\delta^{(n)}$. For this scheme, the probability of error is bounded by

$$P_e \leq P\left(U^n \notin A_\delta^{(n)}\right) < \varepsilon \tag{3.5}$$

and the rate is equal to

$$\frac{\log\left|A_\delta^{(n)}\right|}{n} \leq \frac{nR}{n} = R \tag{3.6}$$

Hence, R is an achievable rate. □

Theorem 18 (Converse theorem)**.** *If $R < H(U)$, R is not achievable.*

Proof For a given scheme of rate $r \leq R$ (and block length n), let $B^{(n)}$ denote the set of possible reconstruction sequences $\hat{U}_n$. Note that $\left|B^{(n)}\right| \leq 2^m = 2^{nr} \leq 2^{nR}$. So if $R < H(U)$, by Theorem 15,

$$P_e \geq P\left(U^n \notin B^{(n)}\right) \to 1, \text{ as } n \to \infty \tag{3.7}$$

Hence, increasing n cannot make the probability of error arbitrarily small. Furthermore, there is clearly a nonzero probability of error for any finite n, so R is not achievable. Conceptually, if the rate is too small, it can't represent a large enough set. $\square$

Bibliography

[1] A. E. Gamal and Y.-H. Kim, *Network Information Theory*, Cambridge Univ. Press, UK, 2012.

[2] R. M. Gray, *Entropy and Information Theory*, Springer-Verlag, New York, 1990.

[3] T. S. Han, *Information-Spectrum Methods in Information Theory*, Springer, 2003.

Chapter 4

Lossless Compression

4.1 Uniquely decodable codes and prefix codes

Last lecture, we talked about how using the AEP, entropy emerges when you want to describe source symbols in fixed length at nearly lossless compression. In fixed-length compression, you map source sequences to representations 1:1. We also said that if you use variable length coding, there is a way to achieve H bits/source symbol with perfect lossless compression, where H is the entropy. How can we achieve such a code? The next few lectures will be devoted to that question.

Let us start with a simple code. Let $l(u)$ represent the length of a binary codeword representing u, $u \in \mathcal{U}$. We can then write $\bar{l} = El(u) = \sum_{u \in \mathcal{U}} p(u)l(u)$ where $\bar{l}$ is the expected length of a codeword.

Example 19. Let $\mathcal{U} = \{a, b, c, d\}$ and let us try to come up with a simple code for this alphabet.

u	p(u)	Codeword	l(u)
a	1/2	0	1
b	1/4	10	2
c	1/8	110	3
d	1/8	111	3

Figure 4.1: Code I

Note: here $l(u) = -\log p(u)$
$\Rightarrow \bar{l} = \mathrm{E}[l(u)] = \mathrm{E}[-\log p(u)] = H(u)$

This code satisfies the prefix condition since no codeword is the prefix for another codeword. It also looks like the expected code length is equal to the entropy. Is the entropy the limit for variable length coding? Can we do better? Let us try a better code.

Here is a "better" code, where $\bar{l} < H$
However, the code in Figure 4.2 is not uniquely decodable. For instance, both 'abd' and 'cbb' can be represented by the code 0111. These codes are not useful. This motivates the notion of uniquely decodable schemes.

u	Codeword	l(u)
a	0	1
b	1	1
c	1	2
d	11	2

Figure 4.2: Better code with regard to $l(u)$

Definition 20. *A code is uniquely decodable(UD) if every sequence of source symbols is mapped to a distinct binary codeword.*

Definition 21.

Prefix Condition: When no codeword is the prefix of any other.

Prefix Code: A code satisfying the prefix condition.

Codes that satisfy the prefix condition are decodable on the fly. Codes that do not satisfy the prefix condition can also be uniquely decodable, but they are less useful.

Exercise 22. Consider Code II in Figure 4.3

u	Codeword
a	10
b	00
c	11
d	110

Figure 4.3: Code II

Prove that this code is UD.
Let us construct binary trees to represent codes. Here, the terminal nodes represent source symbols, and the path from the root to each terminal node represents the codeword for that source symbol. We can construct binary trees for all UD codes, as we will see later.

Here are Binary trees for Code I and Code II:

From here on, let us restrict our attention to prefix codes. In fact, we will see that for any non-prefix code with a given expected code length, we can always find a prefix code with at least as small of a code length.

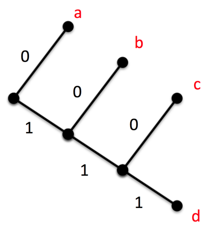


Figure 4.4: Binary tree for Code I

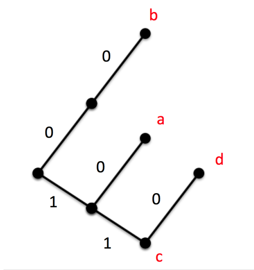


Figure 4.5: Binary tree for Code II

4.2 Prefix code for dyadic distributions

We would like to systematically construct uniquely decodable prefix codes for any alphabet with arbitrary probability mass functions. We will start with dyadic distributions.

Definition 23. *A dyadic distribution has* $p(u) = 2^{-n_u}$, $\forall u \in \mathcal{U}$, *where* n_u *are integers.* $(*)$

Note: If we find a UD code with $l(u) = n_u = -\log p(u)$, then $\bar{l} = H$.
We claim that we can always find a UD code for a dyadic distribution.

Lemma 24. *Assume* $(*)$ *and* $n_{max} = \max_{u\in\mathcal{U}} n_u$, *considering that we have a nontrivial distribution (where* $p(u)$ *is not 1 at one value and 0 everywhere else). The number of symbols* u *with* $n_u = n_{max}$ *is even.*

Proof

$$
\begin{aligned}
1 &= \sum_{u\in\mathcal{U}} p(u) \\
&= \sum_{u\in\mathcal{U}} 2^{-n_u} \\
&= \sum_{n=1}^{n_{max}} (\text{\# of symbols } u \text{ with } n_u = n) \cdot 2^{-n} \\
\Rightarrow 2^{n_{max}} &= \sum_{n=1}^{n_{max}} (\text{\# of symbols } u \text{ with } n_u = n) \cdot 2^{n_{max}-n} \\
&= \sum_{n=1}^{n_{max}-1} (\text{\# of symbols } u \text{ with } n_u = n) \cdot 2^{n_{max}-n} + (\text{\# of symbols } u \text{ with } n_u = n_{max})
\end{aligned}
\tag{4.1}
$$

Since all terms except (# of symbols u with $n_u = n_{max}$) are even, the number of elements in the alphabet with the smallest probability has to be even. □

Now with this lemma in hand, we can prove our claim that we can find a UD code for a dyadic distribution. Consider the following procedure:

- Choose 2 symbols with $n_u = n_{max}$ and merge them into one symbol with (twice the) probability $2^{-n_{max}+1}$
- The new source distribution is also dyadic.
- Repeat the procedure until left with one symbol.

Note: This procedure induces a binary tree.
E.g.:

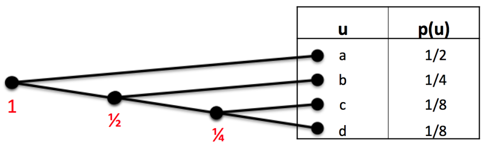


Figure 4.6: Induced binary tree using the procedure

Also note that the symbol with $p(u) = 2^{-n_u}$ has distance n_u from root. This means that the induced prefix code satisfies $l(u) = n_u = -\log p(u)$

4.3 Shannon codes

How can we get a good code for a non-dyadic distribution? We can attempt to use the above principles. Let us look for a code with

$$l(u) = \lceil -\log p(u) \rceil = n_u^* \quad \forall u \in \mathcal{U}$$

Here, we take the ceiling of $-\log p(u)$ as the length of the codeword for the source symbol u. Because the ceiling of a number is always within 1 of the actual number, the expected code length $\bar{l} = \mathrm{E}[-\log p(u)]$ is within 1 of $H(u)$.

Let us consider the "PMF" $p^*(u) = 2^{-n_u^*}$. This "PMF" is a dyadic distribution because all probabilities are a power of 2. We put PMF in quotes because for a non-dyadic source, $\sum_{u\in\mathcal{U}}^{p^*(u)}$ is less than 1, and so the PMF is not a true PMF. See the following:

$$\begin{aligned}\sum_{u\in\mathcal{U}} p^*(u) &= \sum_{u\in\mathcal{U}} 2^{-n_u^*} = \sum_{u\in\mathcal{U}} 2^{-\lceil -\log p(u)\rceil} \\ &< \sum_{u\in\mathcal{U}} 2^{-(-\log p(u))} = 1\end{aligned} \tag{4.2}$$

To make it a true PMF, let us add fictitious source symbols so that $\mathcal{U}^* \supseteq \mathcal{U}$ and $\sum_{u \in \mathcal{U}^*} p^*(u) = 1$. We can thus construct a prefix code for $\mathcal{U}^*$. Because this is a dyadic distribution, we can use the binary tree principle above to construct a code for all $u \in \mathcal{U}^*$, and we can consider only the source symbols u in the original alphabet U. The lengths of each codeword will satisfy

$$l(u) = -\log p^*(u) = n_u^* = \lceil -\log p(u) \rceil \quad \forall u \in \mathcal{U}$$

The expected code length for a Shannon code can be expressed as the following:

$$\bar{l} = \sum_u p(u) l(u) = \sum_u p(u) \lceil -\log p(u) \rceil \leq \sum_u p(u)(-\log p(u) + 1) = H(U) + 1$$

Therefore, the expected code length is always less or equal to the entropy plus 1. This result could be good or bad depending on how large $H(U)$ is to start with. If the extra "1" is too much, alternatively, we can construct a Shannon code for the multi-symbol $u^n = (u_1, u_2, ...u_n)$, where u_i is memoryless. Then,

$$\bar{l}_n \leq H(U^n) + 1 \text{ or } \frac{1}{n}\bar{l}_n \leq \frac{1}{n}H(U^n) + \frac{1}{n} = H(U) + \frac{1}{n}$$

Now we can make it arbitrarily close to the entropy. In the end, there is a trade-off between ideal code length and memory since the code map is essentially a lookup table. If n gets too large, the exponential increase in lookup table size could be a problem.

4.4 Average codelength bound for uniquely decodable codes

We looked at a way of obtaining a prefix code for any given distribution. We will now try to understand the bounds on the average codelength for any generic uniquely decodable code for a distribution.

Theorem 25 (Kraft-McMillan Inequality)**.** *For all uniquely decodable (UD) codes,*

$$\sum_{u \in \mathcal{U}} 2^{-\ell(u)} \leq 1 \tag{4.3}$$

Conversely, any integer-valued function satisfying (4.3) is the length function of some UD code.

To see the "conversely" statement, note that we know how to generate a UD code (in fact, a prefix code) with length function satisfying (4.3), using Huffman Codes. Here, we prove the first claim of the Kraft-McMillan Inequality.

Proof Take any UD code and let $\ell_{\max} = \max_u \ell(u)$. Fix any integer k and note the following.

$$\begin{aligned}
\left(\sum_{u\in\mathcal{U}} 2^{-\ell(u)}\right)^k &= (\sum_{u_1} 2^{-\ell(u_1)}) \cdot (\sum_{u_2} 2^{-\ell(u_2)}) \cdot \ldots \cdot (\sum_{u_k} 2^{-\ell(u_k)}) \\
&= \sum_{u_1}\sum_{u_2}\cdots\sum_{u_k}\prod_{i=1}^{k} 2^{-\ell(u_i)} \\
&= \sum_{(u_1,\ldots,u_k)} 2^{-\sum_{i=1}^k \ell(u_i)} \\
&= \sum_{u^k} 2^{-\ell(u^k)} \\
&= \sum_{i=1}^{k\cdot\ell_{\max}} \left|\left\{u^k \mid \ell(u^k) = i\right\}\right| \cdot 2^{-i} \\
&\leq \sum_{i=1}^{k\cdot\ell_{\max}} 2^i \cdot 2^{-i} \\
&= k \cdot \ell_{\max}
\end{aligned}$$

Note that the inequality in the second to last line arises because we know that our code is one-to-one so there can be at most 2^i symbols whose codewords have length i. Finally, we can see the theorem through the following inequality.

$$\sum_u 2^{-\ell(u)} \leq \lim_{k\to+\infty} (k\ell_{\max})^{1/k} = 1$$

□

Now, we can prove the important theorem relating UD codes to the binary entropy.

Theorem 26. *For all UD codes, $\bar{\ell} \geq H(U)$*

Proof

$$\begin{aligned}
Consider, H(U) - \bar{\ell} &= H(U) - E\ell(U) \\
&= E\left[\log\frac{1}{p(U)} - \ell(U)\right] \\
&= E\left[\log\frac{2^{-\ell(U)}}{p(U)}\right] \\
&\leq \log E\left[\frac{2^{-\ell(U)}}{p(U)}\right] \\
&= \log\sum_u 2^{-\ell(u)} \leq \log 1 = 0
\end{aligned}$$

□

Note that we used the earlier proved Kraft-McMillan Inequality for UD codes in the proof.

As an aside, suppose $\ell(u) = -\log q(u)$ for some pmf q. Then,

$$\begin{aligned}\bar{\ell} - H &= E[\ell(U) - \log \frac{1}{p(u)}] \\ &= \sum_u p(u) \log \frac{p(u)}{q(u)} \\ &= D(p\|q)\end{aligned}$$

Thus, $D(p\|q)$ can be thought of as the "cost of mismatch", in designing a code for a distribution q, when the actual distribution is p.

4.5 Huffman Coding

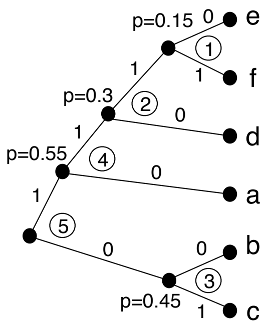
We earlier looked at Shannon code, which is a pretty good construction of a prefix code for a given distribution. However, the best prefix code for a general source code distribution is the Huffman Code.

The construction of the huffman code follows is very similar to that of the dyadic code. To find the code c(u), we follow these steps:

1. Find 2 symbols with the smallest probability and then merge them to create a new "node" and treat it as a new symbol.
2. Then merge the next 2 symbols with the smallest probability to create a new "node"
3. Repeat steps 1 and 2 until there is only 1 symbol left. At this point, we created a binary tree. The paths traversed from the root to the leaves are the prefix codes.

We consider an example of huffman code construction:

Example 27. Prefix code for a senary source (six letters) is given below:



u	p(u)	c(u)	l(u)
e	0.1	1110	4
f	0.05	1111	4
d	0.15	110	3
a	0.25	10	2
b	0.2	00	2
c	0.2	01	2

We will next try to understand why Huffman codes are the optimal prefix codes.

4.6 Optimality of Huffman codes

Suppose we have a random object $U \sim P$ with alphabet $\mathcal{U} = \{1, 2, \ldots, r\}$. We let $c(u), \ell(u)$ be the codeword and length associated with some $u \in \mathcal{U}$, respectively. Recall, the average length is denoted as $\bar{\ell} = E\ell(U)$. Assume, WLOG, that the symbols in $\mathcal{U}$ are ordered in decreasing order of probability according to p, i.e. $p(1) \geq p(2) \geq \cdots \geq p(r)$. Let U_{r-1} denote a random variable over $\mathcal{U}_{r-1} = \{1, 2, \ldots, r-1\}$ and

$$p(U_{r-1} = i) = \begin{cases} p(i) & 1 \leq i \leq r-2 \\ p(r-1) + p(r) & i = r-1 \end{cases}$$

and $c_{r-1}(u_{r-1}), \ell_{r-1}(u_{r-1})$ are the codeword and length of $u_{r-1} \in \mathcal{U}_{r-1}$, respectively. Again, $\bar{\ell}_{r-1} = E\ell_{r-1}(U_{r-1})$.

"Splitting a prefix code c_{r-1}": creating a prefix code for U by

$$\begin{cases} c(i) = c_{r-1}(i) & 1 \leq i \leq r-2 \\ c(r-1) = c_{r-1}(r-1)0 \\ c(r) = c_{r-1}(r-1)1 \end{cases}$$

We will use the following lemma to justify the optimality of Huffman Codes. Intuitively, we will show that if we start with an optimal code on $r-1$ symbols, splitting gives us an optimal code over r symbols. We can use an inductive argument, starting with a binary object to prove that Huffman Codes are optimal for alphabets with any number of symbols.

Lemma 28. *Let $c_{opt,r-1}$ be an optimal prefix code for U_{r-1}. Let c be the code obtained from $c_{opt,r-1}$ by splitting. Then c is an optimal prefix code for U.*

Proof Note there is an optimal prefix code for U satisfying:

1. $\ell(1) \leq \ell(2) \leq \ldots \leq \ell(r)$ Otherwise, we could rearrange the codes to satisfy this property, and the result would be at least as good due to the ordergin we have assumed on the probabilities.

2. $\ell(r-1) = \ell(r) = \ell_{\max}$
   Suppose in an optimal code the two longest codewords were not of the same length. Since the prefix property holds, no codeword is a prefix of the longest codeword. The longest codeword can be truncated, preserving the prefix property but achieving lower expected codeword length. Since the code was optimal, this leads to a contradiction, so the two longest codewords must be of the same length.

3. $c(r-1)$ and $c(r)$ differ only in the last bit
   Given any optimal code, rearranging the code can result in a code with this property. Suppose there is a codeword of maximal length such that it does not have a 'sibling'. The last bit of this codeword can be deleted, preserving the prefix property but achieving lower expeted codeword length. This leads to a contradiction, so every codeword of maximal length must have a 'sibling' in an optimal code. By rearranging the assignments of the maximal length codewords, we can ensure that the two least likely symbols are assigned a pair of sibling codewords. This rearrangement maintains the expected codeword length, and achieves property.

Note, by the second and third properties, there is an optimal prefix code for U that is the result of splitting a prefix code for U_{r-1} (where the code for U_{r-1} can be seen as our original code with the the codeword $c(r-1)$ is truncated before the final bit). Let ℓ_{r-1} be the length function of a prefix code for U_{r-1} and ℓ be the length function of a prefix code for U obtained by splitting U_{r-1}. Then

$$\begin{aligned}
\bar{\ell} &= \sum_{i=1}^{r} p(i)l(i) \\
&= \sum_{i=1}^{r-2} p(i)l(i) + p(r-1)l(r-1) + p(r)l(r) \\
&= \sum_{i=1}^{r-2} p(U_{r-1} = i)l_{r-1}(i) + (p(r-1) + p(r))l(r-1) \\
&= \sum_{i=1}^{r-2} p(U_{r-1} = i)l_{r-1}(i) + (p(r-1) + p(r))(l_{r-1}(r-1) + 1) \\
&= \sum_{i=1}^{r-2} p(U_{r-1} = i)l_{r-1}(i) + p(U_{r-1} = r-1)l_{r-1}(r-1) + (p(r-1) + p(r)) \\
&= \sum_{i=1}^{r-1} p(U_{r-1} = i)l_{r-1}(i) + (p(r-1) + p(r)) \\
&= \bar{\ell}_{r-1} + p(r-1) + p(r)
\end{aligned}$$

$$\bar{\ell} = \bar{\ell}_{r-1} + p(r-1) + p(r)$$

because the expectation sums differ only in the final two terms, where an additional bit is added for symbols $r-1$ and r. We can see, by this simple relationship, that if we want to optimize $\bar{\ell}$ for some fixed probability distribution, it suffices to optimize $\bar{\ell}_{r-1}$. So if $\bar{\ell}_{r-1}$ is optimal, then so is $\bar{\ell}$. Thus, we have that an optimal prefix code for U is obtained by splitting an optimal prefix code for U_{r-1}.

$\square$

Bibliography

[1] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, UK, 2003.

[2] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

Chapter 5

Communication and Channel Capacity

5.1 The Communication Problem

We model communication as the transmission of an input consisting of n symbols (denoted X^n) through a *noisy channel*. This channel will emit an output of n symbols (denoted Y^n). Think of the channel as corrupting or adding noise to the input.

The behavior of the channel is characterized by the conditional probability distribution $P_{Y^n|X^n}$, which tells us the distribution over the outputs that it will emit for any given input.

Conceptually, we have the following picture:

$$X^n \longrightarrow \boxed{\begin{array}{c}\text{noisy channel}\\ P_{Y^n|X^n}\end{array}} \longrightarrow Y^n$$

Our goal, then, is to find an encoder e and a decoder d such that we can take an input of m bits (called B^m), encode it using e to get an encoding of length n denoted X^n, pass the encoding through the channel to get Y^n, and then decode Y^n using d to recover an estimate of the original string of bits, denoted $\hat{B}^m$. Pictorially, that is:

$$(B_1, B_2, \ldots, B_m) = B^m \longrightarrow \boxed{\text{encoder } (e)} \xrightarrow{X^n} \boxed{\begin{array}{c}\text{noisy channel}\\ P_{Y^n|X^n}\end{array}} \xrightarrow{Y^n} \boxed{\text{decoder } (d)} \longrightarrow \hat{B}^m = (\hat{B}_1, \hat{B}_2, \ldots, \hat{B}_m)$$

Ideally, n will be small relative to m (we will not have to send a lot of symbols through the noisy channel), and the probability that the received message $\hat{B}^m$ does not match the original message B^m will be low. We will now make rigorous these intuitively good properties.

Definition 29. *We define a* scheme *to be a pair of encoder and decoder, denoted* (e, d).

Note that the definition of a scheme does not include the noisy channel itself. We must take the channel as it is given to us: we cannot modify it, but we can choose what symbols X^n to transmit through it.

Definition 30. *We define the* rate *of a scheme, denoted R, to be the number of bits communicated per use of the channel. This is equal to m/n in the notation of the diagram above.*

Definition 31. *We define the* probability of error *for a scheme, denoted P_e, to be the probability that the output of the decoder does not exactly match the input of the encoder. That is,*

$$P_e = P(B^m \neq \hat{B}^m).$$

Definition 32. *For a given channel, we say that a rate R is* achievable *if there exists a sequence of schemes $(e_1, d_1), (e_2, d_2), \ldots$, such that:*

1. *For all $n = 1, 2, \ldots$, scheme (e_n, d_n) has rate at least R, and*
2. $$\lim_{n\to\infty} P_e^{(n)} = 0,$$
   where $P_e^{(n)}$ denotes the probability of error of the n^{th} scheme.

5.2 Channel Capacity

We want to know what the best possible performance is under a particular noisy channel. This is essentially what channel capacity tells us.

Definition 33. *For a given channel, the* channel capacity *(denoted C) is the theoretical limit on the number of bits that can be reliably communicated (i.e., communicated with arbitrarily low probability of error) in one channel use.*

That is,

$$C := \sup\{R : R \text{ is achievable}\}$$

We assume a "memoryless channel": $P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i)$ exhibiting a "single letter channel" characteristic of output symbol given input symbol. Restated, the i-th channel output only cares about the ith channel input.

$$X \sim P_x \longrightarrow \boxed{P_{Y|X}} \longrightarrow \text{random } Y$$

With this single letter channel, we now examine $I(X;Y)$. What distribution of X will maximize $I(X;Y)$ over all possible channel inputs?

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

Theorem 34. ***Channel Coding Theorem:***

$$C = C^{(I)} = \max_{X} I(X;Y) \quad (\textit{sometimes written as } \max_{P_X})$$

Proof:
Direct Theorem: If $R < C^{(I)}$, then the rate R is achievable.
Converse Theorem: If $R > C^{(I)}$, then R is not achievable.
The direct part and the converse part of the proof are given at the end of this chapter.

5.2.1 Channel Capacity of various discrete channels

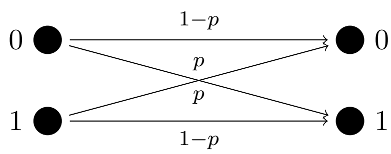
Example 35. Binary Symmetric Channel (BSC)

$$\text{Let: } \mathcal{X} = \mathcal{Y} = \{0, 1\}$$

and the crossover probability be

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

$$\iff \text{channel matrix: } \begin{array}{c|cc} & 0 & 1 \\ \hline 0 & 1-p & p \\ 1 & p & 1-p \end{array}$$



$\iff$ bipartite graph:

$$\iff Y = X \oplus_2 Z \leftarrow \text{ber(p)}$$

To compute the capacity of the BSC, we first examine mutual information

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) && \text{definition of mutual information} \\ &= H(Y) - H(X \oplus_2 Z|X) && \text{substitute } Y \text{ with } X \oplus_2 Z \\ &= H(Y) - H(Z) && \text{given } X, X \oplus_2 Z \text{ is simply } Z \\ &= H(Y) - h_2(p) \leq 1 - h_2(p) && h_2(p) \text{ is binary entropy} \end{aligned}$$

To achieve equality, $H(Y) = 1$, i.e. Y is bernoulli $\frac{1}{2}$. Taking $X \sim Ber(\frac{1}{2})$ produces this desired Y and therefore gives $I(X;Y) = 1 - h_2(p)$

$$\implies C = 1 - h_2(p)$$

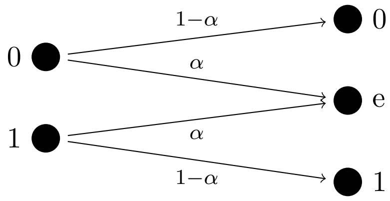
$(1 - h_2(p))n$ bits of information can be communicated reliably.

Example 36. Binary Symmetric Channel (BEC)

$$\text{Let: } \mathcal{X} = \{0, 1\}, \mathcal{Y} = 0, 1, e$$

and the crossover probability be

$$P_{Y|X}(y|x) = \begin{cases} \alpha & y = e \\ 1-\alpha & y = x \end{cases}$$



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

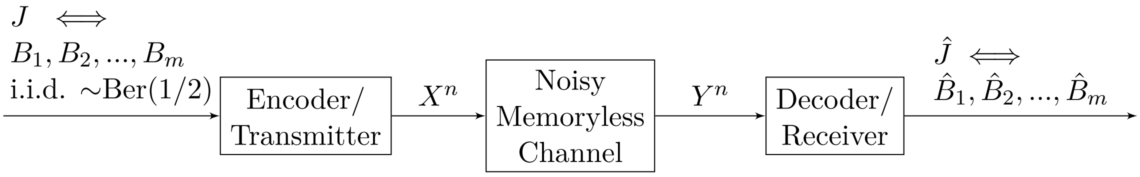
To achieve equality, $H(X) = 1$, i.e. X is bernoulli $\frac{1}{2}$.

$$I(X;Y) = 1 - \alpha$$

$$\implies C = 1 - \alpha$$

$(1-\alpha)n$ bits of information can be communicated reliably.

5.2.2 Recap



- Rate $= \frac{m}{n} \; \frac{bits}{channel\ use}$
- $P_e = \mathbb{P}(\hat{J} \neq J)$
- R is achievable if $\forall\ \epsilon > 0,\ \ \exists$ a scheme $(m,\ n,\ encoder,\ decoder)$ with $\frac{m}{n} \geq R$ and $P_e < \epsilon$.
- Capacity: $C = \sup\{$R: R is achievable$\}$.
- **Channel Coding Theorem:** $C = \max_X I(X;Y)$.

Note: The Channel Coding Theorem is equally valid for analog signals, e.g., the AWGN channel. However, we must extend our definition of the various information measures such as entropy, mutual information, etc.

Next we extend the information measures to continue random variables, and analyze the AWGN channel.

5.3 Information Measures for Continuous Random Variables

In this chapter we extend the information measures to continues random variables.

Definition 37. ***Relative Entropy*** *for two PDFs f and g, is defined as*

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

Can similarly define for $x \in \mathbb{R}^n$.
Note: When integral on the right hand side is not well defined, $D(f||g) = \infty$

Definition 38. ***Mutual Information*** *between two continuous r.v. X, Y with joint pdf $f_{X,Y}$ (i.e., $X,\ Y \sim f_{X,Y}$) is*

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

where $f_X \times f_Y$ is the product of the marginal distributions.

Definition 39. ***Differential entropy*** *of*

1. *$X \sim f_X$:*
$$h(x) \triangleq E[-\log f_X(x)]$$

2. *$X,\ Y \sim f_{X,Y}$:*
$$h(X,Y) \triangleq E[-\log f_{X,Y}(X,Y)] \qquad \text{(“Joint Differential Entropy”)}$$
$$h(X|Y) \triangleq E[-\log f_{X|Y}(X|Y)] \qquad \text{(“Conditional Differential Entropy”)}$$

Each of the above definitions is totally analogous to the discrete case.

In the homework we will show the following result:

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

This is the main/only interest in differential entropy.

Note: Unlike discrete entropy H(X), differential entropy can be positive or negative. This is not the only way in which they differ.

$$\begin{aligned} h(X+c) &= h(X), && \text{for constant c} \\ h(X \cdot c) &= h(X) + \log |c|, && c \neq 0 \end{aligned}$$

5.3.1 Examples

Example 40.

$$X \sim U[a,b]$$
$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}$$
$$h(x) = \log(b-a)$$

Example 41.

$$X \sim N(0,\sigma^2), f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}}$$

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

Note: differential entropies can be either positive or negative. The more correlated the random variable the more negative

Example 42. Significance of Differential entropy
Many Information theorists would argue none whatsoever. However, some others offer a different perspective.
If you discretize $X \sim f$ into X_Δ with time period Δ,

$$P(x_\Delta = i) = \int_{i\Delta-\Delta/2}^{i\Delta+\Delta/2} f(x)dx \approx f(i\Delta)\cdot\Delta$$

$$\begin{aligned} H(X_\Delta) &= \sum_i P_i \log\frac{1}{P_1} \\ &\approx \sum_i f(i\Delta)\cdot\Delta\cdot\log\frac{1}{\Delta f(i\Delta)} \\ &= \log\frac{1}{\Delta} + \sum_i\left(f(i\Delta)\log\frac{1}{f(i\Delta)}\right)\Delta \end{aligned}$$

$$H(X_\Delta) - \log\frac{1}{\Delta} = \sum_i f(i\Delta)\log\frac{1}{f(i\Delta)}\Delta \xrightarrow{\Delta\to 0} \int f(x)\log\frac{1}{f(x)}dx = h(x)$$

$$\Longrightarrow H(X_\Delta) \approx \log \frac{1}{\Delta} + h(x)$$

Majority of the entropy for discretized system is accounted for with $\log \frac{1}{\Delta}$. The rest of it is $h(x)$ the differential entropy

5.3.2 Gaussian Distribution

Claim: The Gaussian distribution has maximal differential entropy, i.e.,:
If $X \sim f_X$ with $E[X^2] \leq \sigma^2$ (second moment), and $G \sim N(0, \sigma^2)$,
Then $h(X) \leq h(G)$, with equality iff $X \sim N(0, \sigma^2)$.

Note: If $E[X^2] \leq \sigma^2$ and $Var(X) = \sigma^2$, then necessarily $E[X] = 0$.

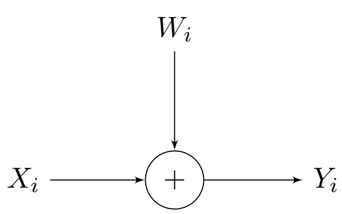
Proof of Claim:

$$f_G(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}X^2}, \quad -\log f_G(X) = \log \sqrt{2\pi\sigma^2} + \frac{\frac{1}{2\sigma^2}X^2}{\ln 2}$$

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

$$\therefore h(X) \leq h(G), \text{ with equality iff } D(f_X || f_G) = 0, \ i.e., \ X \sim G$$

5.4 Channel Capacity of the AWGN Channel (Additive White Gaussian Noise)



Note: The AWGN channel is memoryless.

- Transmission is restricted to power P:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \leq P, \quad \forall n \in \mathbb{N}$$

- R is achievable with power P if: $\forall\ \epsilon > 0, \quad \exists$ scheme restricted to power P and with rate $\frac{m}{n} \geq R$ and probability of error $P_e < \epsilon$.

- Channel Capacity: $C(P) = \sup\{$R: R is achievable with power P$\}$

5.4.1 Channel Coding Theorem for this Setting

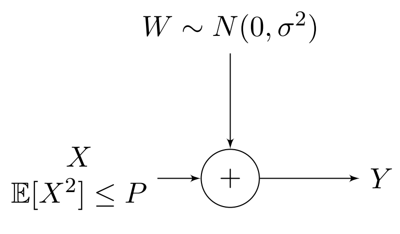
$$C(P) = \max_{\mathbb{E}[X^2]\leq P} I(X;Y)$$

Note: We could instead have considered the restriction that $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i^2] \leq P$. This constitutes a relaxed constraint. However, it turns out that even with the relaxation, you cannot perform any better in terms of the fundamental limit.

5.4.2 An Aside: Cost Constraint

More generally, we can consider an arbitrary cost function constraint on X, rather than the above power constraint. We can denote this cost function by $\phi(X_i)$. The cost constraint is then $\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) \leq \alpha$. This means that the average cost cannot exceed the core parameter α, so we consider $C(\alpha)$. In this case, the coding theorem becomes $C(\alpha) = \max_{\mathbb{E}[\phi(X)]\leq\alpha} I(X;Y)$.

5.4.3 The Example



$$
\begin{aligned}
I(X;Y) &= h(Y) - h(Y|X) \\
&= h(Y) - h(Y-X|X) && \text{(given X, X is a constant, so we can use invariance of differential entropy to constant shifts)} \\
&= h(Y) - h(W|X) \\
&= h(Y) - h(W) && \text{(since W and X are independent)} \\
&\leq h(N(0, P+\sigma^2)) - h(N(0,\sigma^2)) && (Var(Y) = Var(X+W) = Var(X) + Var(W) \leq P + \sigma^2) \\
&= \frac{1}{2}\log 2\pi e(P+\sigma^2) - \frac{1}{2} 2\pi e \sigma^2 \\
&= \frac{1}{2}\log \frac{P+\sigma^2}{\sigma^2} \\
&= \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right)
\end{aligned}
$$

So in conclusion,

$$
I(X;Y) \leq \frac{1}{2} log \left(1+\frac{P}{\sigma^2}\right)
$$

with equality

$$
\begin{aligned}
&\iff \quad Y \sim N(0, P+\sigma^2) \\
&\iff \quad X \sim N(0, P)
\end{aligned}
$$

Therefore, equality is achievable. So,

$$
C(P) = \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right)
$$

(i.e., the capacity of the AWGN channel.)

5.4.4 Rough Geometric Interpretation (Picture)

- Transmission Power Constraint: $\sqrt{\sum_{i=1}^n X_i^2} \leq \sqrt{np}$
- Noise: $\sqrt{\sum_{i=1}^n W_i^2} \approx \sqrt{n\sigma^2}$
- Channel Output Signal:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^n Y_i^2\right] &= \mathbb{E}\left[\sum_{i=1}^n (X_i + W_i)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n X_i^2 + \sum_{i=1}^n W_i^2\right] \\
&\text{(independence} \Rightarrow \text{cross-terms have zero expectation)} \\
&\leq nP + n\sigma^2 \\
&= n(P+\sigma^2)
\end{aligned}
$$

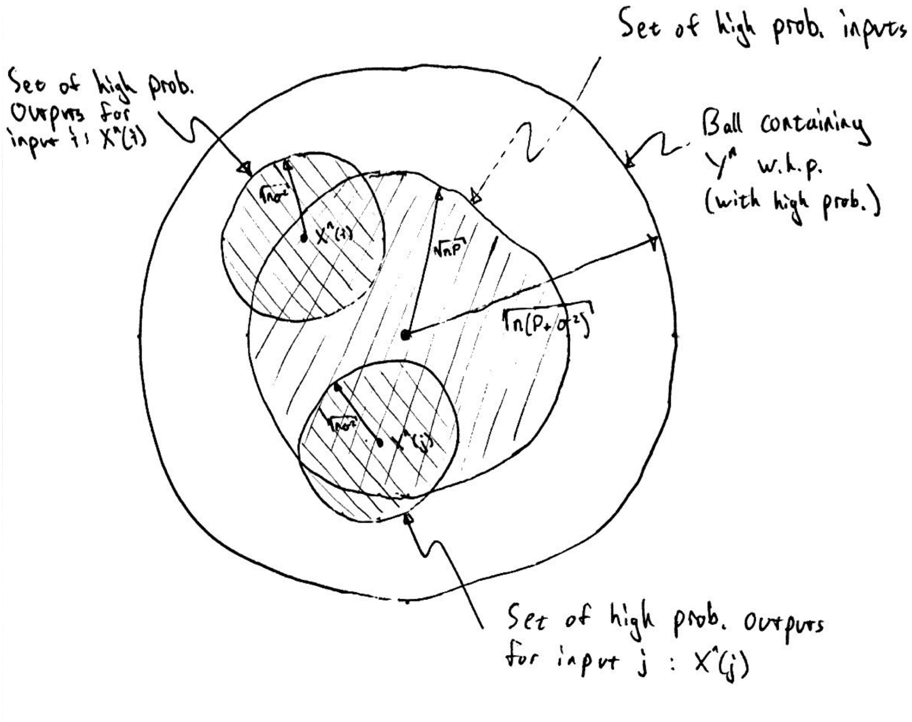


Figure 5.1: Geometrically, consider the input and output sequences as points in $\mathbb{R}^n$.

See **Figure 7.2** for the geometric interpretation of this problem. We want the high probability output balls to not intersect. This way, we can uniquely distinguish the input sequences associated with any given output sequence.

$$\#\text{ messages} \leq \frac{\text{Vol(n-dim ball of radius } \sqrt{n(P+\sigma^2)})}{\text{Vol(n-dim ball of radius } \sqrt{n\sigma^2})}$$

This inequality is due to inefficiencies in the packing ratio. Equality corresponds to perfect packing, i.e. no dead-zones. So,

$$\#\text{ of bits} = \frac{K_n(\sqrt{n(P+\sigma^2)})^n}{K_n(\sqrt{n\sigma^2})^n} = \left(1 + \frac{P}{\sigma^2}\right)^{n/2}$$

$$\Rightarrow \text{rate} = \frac{\log \#\text{ of messages}}{n} \leq \frac{1}{2}\log\left(1 + \frac{P}{Q}\right)$$

The achievability of the equality indicates that in high dimension, can pack the balls very effectively.

5.5 Joint Asymptotic Equipartition Property (AEP)

Let X, Y be jointly random variables with alphabets $\mathcal{X}$, $\mathcal{Y}$, respectively. Let the source be memoryless so that (X_i, Y_i) are i.i.d. $\sim P_{X,Y}$. That is,

$$P(x^n) = \prod_{i=1}^{n} P_X(x_i), \quad P(y^n) = \prod_{i=1}^{n} P_Y(y_i), \quad P(X^n, Y^n) = \prod_{i=1}^{n} P_{X,Y}(x_i, y_i) \tag{5.1}$$

5.5.1 Set of Jointly Typical Sequences

Let $A_\epsilon^{(n)}(X,Y)$ denote the set of jointly typical sequences. That is,

$$A_\epsilon^{(n)}(X,Y) = \{(X^n, Y^n) : |-\frac{1}{n}\log P(x^n) - H(X)| < \epsilon, \tag{5.2}$$

$$|-\frac{1}{n}\log P(y^n) - H(Y)| < \epsilon, \tag{5.3}$$

$$|-\frac{1}{n}\log P(x^n, y^n) - H(X,Y)| < \epsilon\} \tag{5.4}$$

Theorem 43. *If (X^n, Y^n) are formed by i.i.d. $(X_i, Y_i) \sim P_{X,Y}$, then*

1.

$$\lim_{n\to\infty} P((X^n, Y^n) \in A_\epsilon^{(n)}(X,Y)) = 1 \tag{5.5}$$

By the AEP, we have that X^n is typical, Y^n is typical, and (X^n, Y^n) is typical too.

2. $\forall \epsilon > 0, \exists n_0 \in \mathbb{N}$ *such that* $\forall n > n_0$

$$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |A_\epsilon^{(n)}(X,Y)| \leq 2^{n(H(X,Y)+\epsilon)} \tag{5.6}$$

Theorem 44. *If $(\tilde{X}^n, \tilde{Y}^n)$ are formed by i.i.d. $(\tilde{X}_i, \tilde{Y}_i) \sim (\tilde{X}, \tilde{Y})$, where $P_{\tilde{X},\tilde{Y}} = P_X \cdot P_Y$, then $\forall \epsilon > 0, \exists n_0 \in \mathbb{N}$ such that $\forall n > n_0$*

$$(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \leq P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}(X,Y)) \leq 2^{-n(I(X;Y)-3\epsilon)} \tag{5.7}$$

Intuition:

$$|A_\epsilon^{(n)}(\tilde{X},\tilde{Y})| \approx 2^{nH(\tilde{X},\tilde{Y})} \tag{5.8}$$

$$= 2^{n(H(X)+H(Y))} \tag{5.9}$$

$$= 2^{nH(X)} \cdot 2^{nH(Y)} \tag{5.10}$$

$$\approx |A_\epsilon^{(n)}(X)| \cdot |A_\epsilon^{(n)}(Y)| \tag{5.11}$$

Note that $(\tilde{X}, \tilde{Y})$ are distributed uniformly within a set of size $|A_\epsilon^{(n)}(X)| \cdot |A_\epsilon^{(n)}(Y)|$

$$\Rightarrow P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}(X,Y)) = \frac{|A_\epsilon^{(n)}(X,Y)|}{|A_\epsilon^{(n)}(X)| \cdot |A_\epsilon^{(n)}(Y)|} \tag{5.12}$$

$$\approx \frac{2^{nH(X,Y)}}{2^{nH(X)} \cdot 2^{nH(Y)}} \tag{5.13}$$

$$= 2^{-nI(X;Y)} \tag{5.14}$$

Proof:

$$P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}(X,Y)) = \sum_{(\tilde{x}^n, \tilde{y}^n) \in A_\epsilon^{(n)}(X,Y)} P(\tilde{x}_n) \cdot P(\tilde{y}_n) \tag{5.15}$$

$$\leq \sum_{(\tilde{x}^n, \tilde{y}^n) \in A_\epsilon^{(n)}(X,Y)} 2^{-n(H(X)-\epsilon)} \cdot 2^{-n(H(Y)-\epsilon)} \tag{5.16}$$

$$= |A_\epsilon^{(n)}(X,Y)| \cdot 2^{-n(H(X)+H(Y)-2\epsilon)} \tag{5.17}$$

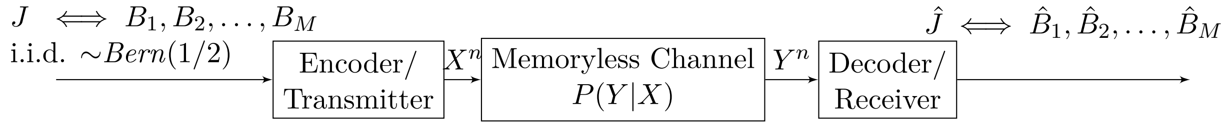
$$\leq 2^{n(H(X,Y)-\epsilon)} \cdot 2^{-n(H(X)+H(Y)-2\epsilon)} \tag{5.18}$$

$$\leq 2^{n(H(X,Y)-H(X)-H(Y)-3\epsilon)} \tag{5.19}$$

$$= 2^{-nI(X;Y)-3\epsilon} \tag{5.20}$$

5.6 Direct Theorem

Recall the setting under consideration:



J is uniformly distributed on $\{1, 2, \ldots, M\}$. We define our scheme as follows:

Encoder (also known as "codebook"): $\{1, 2, \ldots, M\} \to X^n$.
That is, codebook $c_n = \{X^n(1), X^n(2), \ldots, X^n(M)\}$

Decoder: $\hat{J}(\cdot) : Y^n \to \{1, 2, \ldots, M\}$

Rate: Bits per channel use $= \log(M)/n = \log(|c_n|)/n$

Theorem 45. *(Direct Theorem) If $R < \max_{P_X} I(X;Y)$, then R is achievable. Equivalently, if $\exists\ P_X$ s.t. $R < I(X;Y)$, then R is achievable.*

Proof

Fix P_X and a rate $R < I(X;Y)$. Choose $\epsilon = (I(X;Y) - R)/4$. This means that $R < I(X;Y) - 3\epsilon$. Generate codebook C_n of size $M = \lceil 2^{nR} \rceil$.
$X^n(k)$ are i.i.d. with distribution P_X, $\forall\ k = 1, 2, \ldots, M$. Then

$$\hat{J}(Y^n) = \begin{cases} j & \text{if} \quad (X^n(j), Y^n) \in A_\epsilon^n(X,Y) \quad \text{and} \quad (X^n(k), Y^n) \notin A_\epsilon^n(X,Y), \forall j \neq k \\ error & \text{otherwise} \end{cases} \tag{5.21}$$

Denote probability of error using a codebook c^n as $P_e(c^n)$. Thus, $P_e(c^n) = P(\hat{J} \neq J | C_n = c_n)$

$$
\begin{aligned}
E[P_e(C_n)] &= P(\hat{J} \neq J) && (5.22)\\
&= \sum_{i=1}^{M} P(\hat{J} \neq J | J = i) P(J = i) && (5.23)\\
&= P(\hat{J} = J | J = 1) && (5.24)
\end{aligned}
$$

This follows because, by symmetry, $P(\hat{J} \neq J | J = i) = P(\hat{J} \neq J | J = j), \forall i, j$, and $P(J = i) = 1/M, \forall i$

By union bound, it follows that
$P(\hat{J} \neq J | J = 1)$
$\leq P((X^n(1), Y^n) \notin A_\epsilon^{(n)}(X,Y)) + \sum_{k=2}^{M} P((X^n(k), Y^n) \in A_\epsilon^{(n)}(X,Y))$
The first term on the right tends to zero as n tends to infinity. Therefore,

$$
\begin{aligned}
P(\hat{J} \neq J | J = 1) &\leq \sum_{k=2}^{M} P((X^n(k), Y^n) \in A_\epsilon^{(n)}(X,Y)) && (5.25)\\
&\leq \sum_{k=2}^{M} P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}(X,Y)) && (5.26)\\
&\leq (M-1) \cdot 2^{-n(I(X;Y)-3\epsilon)} && (5.27)\\
&\leq 2^{nR} \cdot 2^{-n(I(X;Y)-3\epsilon)} && (5.28)\\
&\leq 2^{-n(I(X;Y)-3\epsilon-R)} && (5.29)
\end{aligned}
$$

Since $R < I(X;Y) - 3\epsilon$, the expression tends to zero as n tends to infinity.

This means that,

$\exists c_n$ s.t. $|c_n| \geq 2^{nR}$ and $P_e(c_n) \leq E[P_e(C_n)]$
$\Rightarrow \exists c_n$ s.t. $|c_n| \geq 2^{nR}$ and $\lim_{n\to\infty} P_e(c_n) = 0$
$\Rightarrow$ R is achievable.

$\square$

5.7 Fano's Inequality

Theorem 46 (Fano's Inequality)**.** *Let X be a discrete random variable and $\hat{X} = \hat{X}(Y)$ be a guess of X based on Y. Let $P_e := P(\hat{X} \neq X)$. Then $H(X|Y) \leq h(P_e) + P_e \log(|\mathcal{X}| - 1)$.*

Proof Let $V = \mathbf{1}_{\{\hat{X} \neq X\}}$.

$$\begin{aligned} H(X|Y) &\leq H(X,V|Y) \quad \text{(Data Processing Inequality)} && (5.30) \\ &= H(V|Y) + H(X|V,Y) \quad \text{(Chain Rule)} && (5.31) \\ &\leq H(V) + H(X|V=0,Y=y)P(V=0,Y=y) \\ &\quad + H(X|V=1,Y=y)P(V=1,Y=y) && (5.32) \end{aligned}$$

We can simplify terms in (5.32). First, $H(V) = h(P_e)$, the entropy of a binary random variable with success probability P_e. Furthermore, X is deterministic given $V = 0$ and $Y = y$, so $H(X|V = 0, Y = y) = 0$. Finally, if $V = 1$ and $Y = y$, then $\hat{X}$ is known and X can take up to $|\mathcal{X}| - 1$ values. Thus $H(X|V = 1, Y = y) \leq \log(|\mathcal{X}| - 1)$. Putting these facts together, we arrive at:

$$\begin{aligned} H(X|Y) &\leq h(P_e) + \log(|\mathcal{X}| - 1)P(V = 1) && (5.33) \\ &= h(P_e) + P_e \log(|\mathcal{X}| - 1) && (5.34) \end{aligned}$$

□

A weaker version of Fano's Inequality uses the facts that $h(P_e) \leq 1$ and $\log(|\mathcal{X}| - 1) \leq \log(|\mathcal{X}|)$:

$$H(X|Y) \leq 1 + P_e \log(|\mathcal{X}|) \tag{5.35}$$

or equivalently,

$$P_e \geq \frac{H(X|Y) - 1}{\log(\mathcal{X})} \tag{5.36}$$

5.8 Converse Theorem

Theorem 47 (Converse Theorem). *If $R > C^{(I)}$, then rate R is not achievable.*

Proof

$$\begin{aligned} \log M - H(J|Y^n) &= H(J) - H(J|Y^n) && (5.37) \\ &= I(J; Y^n) && (5.38) \\ &= H(Y^n) - H(Y^n|J) && (5.39) \\ &= \sum_i H(Y_i|Y^{i-1}) - \sum_i H(Y_i|Y^{i-1}, J) && (5.40) \\ &\leq \sum_i H(Y_i) - \sum_i H(Y_i|Y^{i-1}, J, X^n) \quad \text{(conditioning reduces entropy)} && (5.41) \\ &= \sum_i H(Y_i) - \sum_i H(Y_i|X_i) \quad \text{(memorylessness)} && (5.42) \\ &= \sum_i I(X_i; Y_i) && (5.43) \\ &\leq nC^{(I)} && (5.44) \end{aligned}$$

Thus, for schemes with rate ($= \frac{\log M}{n}$) $\geq R$, we have

$$P_e \geq \frac{H(J|Y^n) - 1}{\log M} \geq \frac{\log M - nC^{(I)} - 1}{\log M} \geq 1 - \frac{C^{(I)}}{R} - \frac{1}{nR} \xrightarrow{n \to \infty} 1 - \frac{C^{(I)}}{R} \tag{5.45}$$

If $R > C^{(I)}$, then P_e is bounded from below by a positive constant, so it does not approach 0. Therefore, $R > C^{(I)}$ is not achievable. □

5.9 Some Notes on the Direct and Converse Theorems

5.9.1 Communication with Feedback: $X_i(J, Y^{i-1})$

Even if the encoder gets feedback of what has been received on the other side of the channel, one can verify that the proof of converse carries over verbatim; $C = C^{(I)}$ with or without feedback! But, feedback can help improve simplicity and reliability of schemes to achieve the best rate. Here is an example:

Example 48. Communicating Through Erasure Channel

Recall that the capacity of the erasure channel (Fig. 7.2) is $C = 1 - \alpha$ bits/channel use. If feedback exists, the transmitter can repeat each information bit until it goes through unerased. On average, one needs $1/(1-\alpha)$ channel uses per information bit. This means that the rate achieved by this scheme is $1 - \alpha$ bits/channel use. This simple scheme is completely reliable since the probability of error is equal to zero (every bit will eventually be error-free).

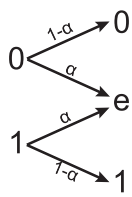


Figure 5.2: An Erasure Channel

5.9.2 Practical Schemes

In the proof of the direct part, we showed mere existence of C_n (a codebook achieving the rate equivalent to the channel capacity) with a size $|C_n| \geq 2^{nR}$, and small P_e. Even if such C_n is given, encoding and decoding using this codebook for large n is not practical. For practical schemes, see:

1. LDPC Codes: "Low Density Parity Check Codes", Gallager 1963 Thesis [3].
2. Polar Codes: "Channel Polarization", Arikan 2009 [4].
3. Or, take EE388 – Modern Coding Theory.

5.9.3 P_e vs. $P_{\max}$

In our discussion so far, our notion of reliability has been the (average) probability of error, which is defined as:

$$P_e = P(\hat{J} \neq J) = \frac{1}{M}\sum_{j=1}^{M} P(\hat{J} \neq J | J = j) \tag{5.46}$$

A more stringent notion of reliability is the maximal probability of error P_{max}, which is defined as:

$$P_{\max} = \max_{1 \leq j \leq M} P(\hat{J} \neq j | J = j) \tag{5.47}$$

It turns out that our results, i.e., direct and converse theorems, are still valid for this more stringent notion of reliability. The converse theorem is clear. If arbitrarily small P_e cannot be achieved, arbitrarily small $P_{\max}$ cannot be achieved either, therefore the converse theorem holds for $P_{\max}$. We now show that the result of the direct proof holds for vanishing $P_{\max}$. Note that with application of the Markov inequality, we have:

$$|\{1 \leq j \leq M : P(\hat{J} \neq j | J = j) \leq 2P_e\}| \geq \frac{M}{2} \tag{5.48}$$

Given C_n with $|C_n| = M$ and P_e, there exists C_n' with $|C_n'| = M/2$ and $P_{\max} \leq 2P_e$. By extracting a better half of C_n, one can construct C_n'. The rate of C_n' is:

$$\text{Rate of } C_n' \geq \frac{\log(M/2)}{n} = \frac{\log M}{n} - \frac{1}{n} \tag{5.49}$$

This implies that if there exists schemes of rate $\geq R$ with $P_e \to 0$, then for any $\epsilon > 0$, there exists schemes of rate $\geq R - \epsilon$ with $P_{\max} \to 0$

Bibliography

[1] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, UK, 2003.

[2] S. Geman and D. Geman, “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[3] R. G. Gallager, *Low-Density Parity-Check Codes*, Cambridge, MA: MIT Press, 1963.

[4] E. Arikan, “Channel polarization: A method for constructing capacity achieving codes for symmetric binary-input memoryless channels,” *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3051–3073, July 2009.

Chapter 6

Method of Types

For additional material on the Method of Types, we refer the reader to [1] (Section 11.1).

6.1 Method of Types

Denote $x^n = (x_1, \ldots, x_n)$, with $x_i \in \mathcal{X} = \{1, 2, \ldots, r\}$.

Definition 49. *The empirical distribution or type of x^n is the vector $(P_{x^n}(1), P_{x^n}(2), \ldots P_{x^n}(r))$ of relative frequencies $P_{x^n}(a) = \dfrac{N(a|x^n)}{n}$, where $N(a|x^n) = \sum_{i=1}^n \mathbb{1}_{\{x_i = a\}}$.*

Definition 50. *$\mathscr{P}_n$ denotes the collection of all empirical distributions of sequences of length n.*

Example 51. $\mathcal{X} = \{0, 1\}$

$$\mathscr{P}_n = \left\{(0,1), \left(\frac{1}{n}, \frac{n-1}{n}\right), \left(\frac{2}{n}, \frac{n-2}{n}\right), \ldots, (1,0)\right\}.$$

Definition 52. *If $P \in \mathscr{P}_n$ (probabilities are integer multiples of $1/n$), the type class or type of P is $T(P) = \{x^n : P_{x^n} = P\}$. The type class of x^n is $T_{x^n} = T(P_{x^n}) = \{\tilde{x} : P_{\tilde{x}^n} = P_{x^n}\}$.*

Example 53. $\mathcal{X} = \{a, b, c\}$, $n = 5$, $x^n = (aacba)$

Then $P_{x^n} = (3/5, 1/5, 1/5)$

$T_{x^n} = \{aaabc, aaacb, \ldots, cbaaa\}$

$|T_{x^n}| = \binom{5}{3\ 1\ 1} = \dfrac{5!}{3!1!1!} = 20.$

Theorem 54. $|\mathscr{P}_n| \le (n+1)^{r-1}$

Proof Type of x^n is determined by $(N(1|x^n), N(2|x^n), \ldots, N(r|x^n))$. Each component can assume no more than $n+1$ values ($0 \le N(i|x^n) \le n$) (and the last component is dictated by the others). $\square$

Example 55. For $\mathcal{X} = \{0,1\}$, $\quad |\mathscr{P}_n| = n+1 = (n+1)^{r-1}$.

Notation:

- $Q = \{Q(x)\}_{x\in\mathcal{X}}$ is a PMF, write $H(Q)$ for $H(X)$ when $X \sim Q$.
- $Q^n(x^n) = \prod_{i=1}^n Q(x_i)$, $S \subseteq \mathcal{X}^n$, $Q^n(S) = \sum_{x^n\in S} Q^n(x^n)$.

Theorem 56. $\forall x^n : Q^n(x^n) = 2^{-n[H(P_{x^n})+D(P_{x^n}||Q)]}$.

Proof

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

$\square$

Theorem 57. $\forall P \in \mathscr{P}_n$,

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

Note: $|T(P)| = \binom{n}{nP(1)\ nP(2)\ \ldots\ nP(r)} = \dfrac{n!}{\prod_{a\in\mathcal{X}}(nP(a))!}$.

Proof UPPER BOUND:

$$
\begin{aligned}
1 \geq P^n\left(T(P)\right) &= \sum_{x^n \in T(P)} P^n(x^n) \\
&= |T(P)| 2^{-n[H(P)+D(P||P)]} \\
&= |T(P)| 2^{-nH(P)}.
\end{aligned}
$$

$\square$

For the lower bound we will use a Lemma.

Lemma: $\forall P, Q \in \mathscr{P}_n : P^n\left(T(P)\right) \geq P^n\left(T(Q)\right)$.

Proof

$$
\begin{aligned}
\frac{P^n\left(T(P)\right)}{P^n\left(T(Q)\right)} = \frac{|T(P)| \prod_{a\in\mathcal{X}} P(a)^{[nP(a)]}}{|T(Q)| \prod_{a\in\mathcal{X}} P(a)^{[nQ(a)]}} &= \frac{\binom{n}{nP(1)\ \ldots\ nP(r)}}{\binom{n}{nQ(1)\ \ldots\ nQ(r)}} \prod_{a\in\mathcal{X}} P(a)^{n[P(a)-Q(a)]} \\
&= \prod_{a\in\mathcal{X}} \frac{(nQ(a))!}{(nP(a))!} P(a)^{n[P(a)-Q(a)]}
\end{aligned}
$$

Note: $\dfrac{m!}{n!} \geq n^{m-n}$

If $m > n$, then $\dfrac{m!}{n!} = m(m-1)\ldots(n+1) \geq n^{m-n}$.

If $n > m$, then $\dfrac{m!}{n!} = \dfrac{1}{n(n-1)\ldots(m+1)} \geq \left(\dfrac{1}{n}\right)^{n-m} = n^{m-n}$.

Therefore,

$$
\begin{aligned}
\prod_{a\in\mathcal{X}} \frac{(nQ(a))!}{(nP(a))!} P(a)^{n[P(a)-Q(a)]} &\geq \prod_{a\in\mathcal{X}} \left(nP(a)\right)^{n[Q(a)-P(a)]} P(a)^{n[P(a)-Q(a)]} \\
&= \prod_{a\in\mathcal{X}} n^{n[P(a)-Q(a)]} \\
&= n^{n\sum_{a\in\mathcal{X}}[P(a)-Q(a)]} \\
&= 1.
\end{aligned}
$$

$\square$

Proof PROOF OF LOWER BOUND:

$$
\begin{aligned}
1 = \sum_{Q\in\mathscr{P}_n} P^n\left(T(Q)\right) &\leq |\mathscr{P}_n| \max_Q P^n\left(T(Q)\right) \\
&= |\mathscr{P}_n| P^n\left(T(P)\right) \\
&= |\mathscr{P}_n| |T(P)| 2^{-n[H(P)+D(P||P)]} \\
&\leq (n+1)^{r-1} |T(P)| 2^{-nH(P)}.
\end{aligned}
$$

$\square$

Theorem 58. $\forall P \in \mathscr{P}_n, Q$,

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

Proof $\quad Q^n\left(T(P)\right) = |T(P)| 2^{-n[H(P)+D(P||Q)]}$.

Now bound $|T(P)|$ as in previous theorem. $\square$

Note: We will write $\alpha_n \doteq \beta_n$: "equality to first order in the exponent"

$\iff \frac{1}{n} \log \frac{\alpha_n}{\beta_n} \xrightarrow{n\to\infty} 0$

$\iff |\frac{1}{n} \log \alpha_n - \frac{1}{n} \log \beta_n| \xrightarrow{n\to\infty} 0$

E.g. : $\alpha_n \doteq 2^{nJ} \iff \alpha_n = 2^{n(J+\epsilon_n)}$ where $\epsilon_n \xrightarrow{n\to\infty} 0$.

6.1.1 Recap on Types

Consider the sequence $x^n \in \mathcal{X}^n$, where $\mathcal{X}$ is a finite alphabet. Let P_{x^n} be the empirical distribution and $\mathscr{P}_n$ the set of all empirical distributions over sequences of length n. Then we define the type to be:

$$T(P) = \{x^n : P_{x^n} = P\},$$

for $P \in \mathscr{P}_n$. We have shown that:

1. $|\mathscr{P}_n| \leq (n+1)^{|x|}$
2. $Q^n(x^n) = 2^{-n[H(P_{x^n})+D(P_{x^n}||Q)]}$
3. $\frac{1}{(n+1)^{|x|}} 2^{nH(p)} \leq T(P) \leq 2^{nH(p)}$, or equivalently, $|T(P)| \doteq 2^{nH(P)}, \quad P \in \mathscr{P}_n$.
4. $\frac{1}{(n+1)^{|x|}} 2^{-nD(P_{x^n}||Q)} \leq Q(T(P)) \leq 2^{-nD(P_{x^n}||Q)}$, or equivalently, $Q^n\left(T(P)\right) \doteq 2^{-nD(P||Q)}$.

6.2 A Version of Sanov's Theorem

Next we prove a version of Sanov's Theorem, which bounds the probability that a function's empirical mean exceeds some value α.

Theorem 59. *Sanov's Theorem. For sufficiently large n, we have*

$$\frac{1}{(n+1)^{|x|}} 2^{-n\min D(P||Q)} \leq \left(\frac{1}{n}\sum_{i=1}^{n} f\left(X_i\right)) \geq \alpha\right) \leq (n+1)^{|x|} 2^{-n\min D(P||Q)},$$

where the min *is over the set* $\{P : P \in \mathcal{P}_n, \langle P_{x^n}, f\rangle \geq \alpha\}$.

Proof First observe that we can write

$$\begin{aligned}
\frac{1}{n}\sum_{i=0}^{n} f(x) &= \frac{1}{n}\sum_{a\in\mathcal{X}} N(a|x^n)f(a) \\
&= \sum_{a\in\mathcal{X}} P_{x^n}(a)f(a) \\
&= \langle P_{x^n}, f\rangle
\end{aligned}$$

where we have used the Euclidean inner product, defined as $\langle a, b\rangle := \sum_{i=1}^{n} a_i b_i$ for $a, b \in \mathbf{R}$. Then by the Law of Large numbers,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} f(X_i) &\approx \mathbf{E}_{X\sim Q} f(X) \\
&= \sum_{a\in\mathcal{X}} Q(a)f(a) \\
&= \langle Q, f\rangle
\end{aligned}$$

We can proceed to find the desired upper bound.

$$\begin{aligned}
P\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \alpha\right) &= P(\langle P_{x^n}, f\rangle) \\
&= Q^n\left(\bigcup_{P\in\mathcal{P}_n, \langle P_{x^n}, f\rangle\geq\alpha} T(P)\right) \\
&= \sum_{P\in\mathcal{P}_n, \langle P_{x^n}, f\rangle\geq\alpha} Q^n(T(P)) \\
&\leq |\mathcal{P}_n| \max Q^n(T(P)) \\
&\leq (n+1)^{|x|} \max 2^{-nD(P||Q)} \\
&= (n+1)^{|x|} 2^{-n\min D(P||Q)}
\end{aligned}$$

Now we solve for the lower bound:

$$\begin{aligned}
P\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \alpha\right) &\geq \max Q^n(T(P)) \\
&\geq \max \frac{1}{(n+1)^{|x|}} 2^{-nD(P||Q)} \\
&= \frac{1}{(n+1)^{|x|}} 2^{-n\min D(P||Q)}
\end{aligned}$$

Note that we are taking the min, max over the set $\{P : P \in \mathcal{P}_n, \langle P_{x^n}, f \rangle \geq \alpha\}$. Therefore, we have, up to a polynomial in n, that $P\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \alpha\right) \doteq 2^{-nD^*(\alpha)}$ where $D^*(\alpha) = \min D(P||Q)$. This is exactly what we were looking for. $\square$

Example 60. Take $X_i \sim \text{Ber}\left(\frac{1}{2}\right)$. Then:

$$P\left(\text{fraction of ones in } X_1 X_2 \cdots X_n \geq \alpha\right) = P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right) \doteq 2^{-nD^*(\alpha)}$$

where $D^*(\alpha) = \min D(\text{Ber}(\alpha)||\text{Ber}(1/2))$. This gives:

$$D^*(\alpha) = \begin{cases} 0 & 0 \leq \alpha \leq \frac{1}{2} \\ D(\text{Ber}(\alpha)||\text{Ber}(1/2)) & \frac{1}{2} < \alpha \leq 1 \\ \infty & \alpha > 1 \end{cases}$$

and since

$$\begin{aligned} D(\text{Ber}(p)||\text{Ber}(1/2)) &= \alpha \log \frac{\alpha}{1/2} + (1-\alpha)\log\frac{1-\alpha}{1/2} \\ &= 1 - h(\alpha) \end{aligned}$$

where $h(\cdot)$ is the binary entropy function. Thus we can write

$$D^*(\alpha) = \begin{cases} 0 & 0 \leq \alpha \leq \frac{1}{2} \\ 1 - h(\alpha) & \frac{1}{2} < \alpha \leq 1 \\ \infty & \alpha > 1 \end{cases}$$

Interestingly, this function explodes at $\alpha = 1$, which makes sense because the probability that the mean of random variables which take values up to 1 is greater than 1 is impossible. Furthermore, we have a region where the cost of mismatch is zero, since we are guaranteed that one of the probabilities is always going to be $\geq 1/2$, so we would expect our mean to be so as well.

Bibliography

[1] Cover, Thomas M., and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Chapter 7

Conditional and Joint Typicality

Notation:
We always write a sequence of symbols by small letter. For example x^n is an individual sequence without any probability distribution assigned to it. We use capital letter for random variables, e.g. X^n i.i.d. according to some distribution. Throughout, $\mathcal{X}$ will denote the set of possible values a symbol can take.

Before going forward, we define empirical distribution.

Definition 61. *For any sequence x^n, empirical distribution is the probability distribution derived for letters of the alphabet based on frequency of appearance of that specific letter in the sequence. More precisely:*

$$p_{x^n}(a) = \frac{1}{n}\sum 1(x_i = a), \quad \forall a \in \mathcal{X} \tag{7.1}$$

7.1 Typical Set (again)

A rough intuition for typical sets is that if one picks a sequence from an i.i.d distribution, $p \sim X^n$, then the typical set $\mathcal{T}(X)$ is a set of length n sequences with the following properties:
1. A sequence chosen at random will be in the typical set with probability almost one.
2. All the elements of the typical set have (almost) equal probabilities.
More precisely, if $\mathcal{T}(X)$ is the typical set, $|\mathcal{T}(X)| \approx 2^{nH(X)}$ and probability of each sequence inside the typical set is $\sim 2^{-nH(X)}$. So a random sequence chosen from the set looks like one chosen uniformly from the typical set.

7.2 δ–strongly typical set

Definition 62. *A sequence x^n is said to be strongly δ typical with respect to the pmf P if,*

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

Where $\mathcal{X}$ is the support of the distribution.

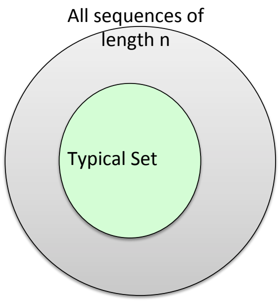


Figure 7.1: Space of all sequences and typical set inside.

Definition 63. *The strongly δ-typical set, $T_\delta(P)$, is the set of all strongly δ typical sequences. That is*

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

For reference, recall that the weakly ϵ-typical set is:

$$A_\epsilon(P) = \left\{x^n : \left|-\frac{1}{n}\log P(x^n) - H(P)\right| \le \epsilon\right\}$$

Example 64. Consider the following extreme example where $P(a) = \frac{1}{|\mathcal{X}|}$ is uniform with $a \in \mathcal{X}$. So for all x^n,

$$\begin{aligned} P(x^n) &= \frac{1}{|\mathcal{X}|^n} \\ &= 2^{-n\log|\mathcal{X}|} \\ &= 2^{-nH(p)} \end{aligned}$$

Therefore, for all $\epsilon > 0$, $A_\epsilon^{(n)}(P) = \mathcal{X}^n$! This makes sense because this is a uniform distribution so you would always expect the typical sequence to include all possibilities regardless of n.

Note that this "strongly typical" set is different from the other "(weakly) typical" set defined in previous chapters. This new notion is stronger, but as we will see it retains all the desirable properties of typicality, and more. The following example illustrates difference of this strong notion and weak notion defined earlier:

Example 65. Suppose that alphabet is $\mathcal{X} = \{a, b, c\}$ with the probabilities $p(a) = 0.1$, $p(b) = 0.8$ and $p(c) = 0.1$. Now consider two strings of length 1000:

$$x_{strong}^{1000} = (100a, 800b, 100c)$$

$$x_{weak}^{1000} = (200a, 800b)$$

In this example, these two sequences have same probability, so they are both identical in the weak notion of typical set $A_{1000}^{(\epsilon)}$ (for some ϵ). But it is not hard to see that x_{strong}^{1000} is a δ- strong typical set while the other is not (for sufficiently small δ). The strong notion is sensitive to frequency of different letters and not only the total probability of sequence.

We will show in the homework that that we preserve important properties of typical sets in the new strong notion. Specifically, we will show the following results:

1. $\forall\, \delta > 0$, there exists $\epsilon = \delta H(p)$ such that $T_\delta(P) \subseteq A_\epsilon(P)$ (i.e., strong typical sets are inside weak typical sets).
2. Empirical probability p_{X^n} is almost equal to the probability distribution p. Therefore $p(x^n) \approx 2^{-nH(X)}$ for x^n in the strong typical set.
3. There exists $\epsilon(\delta)$ such that for all n sufficiently large:

$$2^{n[H(P)-\epsilon(\delta)]} \leq |T_\epsilon(P)| \leq 2^{n[H(P)+\epsilon(\delta)]},$$

   where $\epsilon(\delta) \to 0$ as $\delta \to 0$. Thus $|\mathcal{T}_\delta(X)| \approx 2^{nH}$.
4. $P(x^n \in \mathcal{T}_\delta(X)) \to 1$ as $n \to \infty$

7.3 δ–jointly typical set

In this section we extend the notion of δ - typical sets to pair of sequences $x^n = (x_1, ..., x_n)$ and $y^n = (y_1, ..., y_n)$ from alphabets $\mathcal{X}$ and $\mathcal{Y}$.

Definition 66. *For $x^n = (x_1, x_2, \dots x_n) \in \mathcal{X}^n$ and $y^n = (y_1, y_2, \dots y_n) \in \mathcal{Y}^n$, the joint empirical distribution is defined as*

$$\begin{aligned} P_{x^n,y^n}(x,y) &= \frac{1}{n} \left|\{i \in \{1, \dots, n\} : x_i = x, y_i = y\}\right| \\ &= \frac{1}{n} N\left(x, y \mid x^n, y^n\right) \end{aligned}$$

where we have defined $N\left(x, y \mid x^n, y^n\right) := \left|\{i \in \{1, \dots, \} : x_i = x, y_i = y\}\right|$

Definition 67. *A pair of sequences (x^n, y^n) is said to be δ - jointly typical with respect to a pmf P_{XY} on $\mathcal{X} \times \mathcal{Y}$ if:*

$$|P_{x^n,y^n}(x,y) - P(x,y)| \leq \delta P(x,y),$$

where $P_{x^n,y^n}(x,y)$ is the empirical distribution.

Then, we can make the following definition.

Definition 68. *For $(X, Y) \sim P$, the jointly δ typical set is given by*

$$T_\delta(P) = \{(X^n, Y^n) : |P_{x^n,y^n}(x,y) - P(x,y)| < \delta P(x,y),\ \forall (x,y) \in \mathcal{X} \times \mathcal{Y}\}$$

If we look carefully, nothing is really very new. We just require that empirical distribution of pair of sequences be δ - close to the pmf P_{XY}.

Often as is the case with $H(X)$ vs $H(P)$, we will write $T_\delta(X)$ for $T_\delta(P)$ when $P \sim X$ and $T_\delta(X,Y)$ for $T_\delta(P)$ when $P \sim (X,Y)$.

The notion of strong vs weak typicality is important. For example, consider the random variable $G_n = \frac{1}{n}\sum_{i=1}^n g(x_i)$. If X^n is strongly typical, then G_n is close to $\mathbf{E}[g(x)]$ for large n. On the other hand, this would not necessarily have been the case if X^n were only weakly typical.

It is easy to see that size of typical set is:

$$|\mathcal{T}_\delta(X,Y)| \approx 2^{nH(X,Y)}$$

and

$$P((x^n, y^n) \in \mathcal{T}_\delta(X,Y)) \approx 2^{-nH(X,Y)}$$

In Figure 7.2 we have depicted the sequences of length n from alphabet $\mathcal{X}$ on the x axis and sequences from alphabet $\mathcal{Y}$ on the y axis.

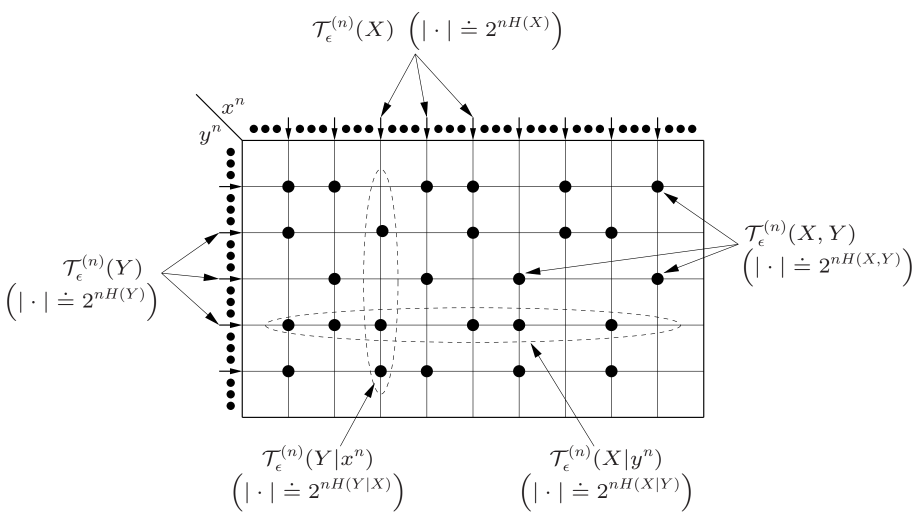


Figure 7.2: A useful diagram depicting typical sets, from El Gamal and Kim (Chapter 2)

Then we can look inside the table and any point corresponds to a pair of sequences. We have marked the jointly typical sets with dots. It is easy to see that if a set is jointly typical then both of the sequences in the set are typical as well. Also we will see in the next section that number of dots in the column corresponding to a typical x^n sequence is approximately $2^{nH(Y|X)}$ (a similar statement is also correct for rows). We can quickly check that this is consistent by a counting argument: we know that number of typical x^n sequences is $2^{nH(X)}$ and each column there are $2^{nH(Y|X)}$ jointly typical pairs. So in total there are $2^{nH(X)} \cdot 2^{nH(Y|X)}$ number of typical pairs. But $2^{nH(X)} \cdot 2^{nH(Y|X)} = 2^{n(H(X)+H(Y|X))} = 2^{nH(X,Y)}$ which is consistent to the fact that total number of jointly typical pairs is equal to $2^{nH(X,Y)}$.

7.4 δ–conditional typicality

In many applications of typicality, we know one sequence (say, the input to a channel) and want to know typical values for some other sequence. In this case, a useful concept is conditional typicality.

Definition 69. *Fix $\delta > 0, x^n \in \mathcal{X}^n$. Then, the conditional typical set $\mathcal{T}_\delta(Y|x^n)$ is defined by*

$$\mathcal{T}_\delta(Y|x^n) \triangleq \{y^n \in \mathcal{Y}^n : (x^n, y^n) \in \mathcal{T}_\delta(X, Y)\}.$$

As usual, it is useful to have bounds on the size of this set. For all $\delta' < \delta$ (and n large enough, as usual), $x^n \in \mathcal{T}_{\delta'}(X) \Rightarrow |\mathcal{T}_\delta(Y|x^n)| \doteq 2^{nH(Y|X)}$ (below, we will be a bit more careful, and we will see that this exponent should depend on a function $\epsilon(\delta)$ which vanishes as $\delta \to 0$). Note that this asymptotic behavior is does not depend on the specific sequence x^n, as long as x^n is typical! This is all in accordance with the intuition we have developed: all typical sequences behave roughly similarly. If $x^n \notin \mathcal{T}_\delta(X)$, then $|\mathcal{T}_\delta(Y|x^n)| = 0$, as (x^n, y^n) cannot be jointly typical if x^n is not typical.

To illustrate the methods used to describe the asymptotic behavior of $|T_\delta(Y|x^n)|$, we find an upper bound on this value. If $x^n \in \mathcal{T}_{\delta'}(X) \subseteq \mathcal{T}_\delta(X)$ and $y^n \in \mathcal{T}_\delta(Y|x^n)$, then by definition of the conditional typical set, $(x^n, y^n) \in \mathcal{T}_\delta(X, Y)$. Since strong typicality implies weak typicality,

$$(1-\delta)H(X,Y) \leq -\frac{1}{n}\log p(x^n, y^n) \leq (1+\delta)H(X,Y), \tag{7.2}$$

$$(1-\delta)H(X) \leq -\frac{1}{n}\log p(x^n) \leq (1+\delta)H(X). \tag{7.3}$$

So, fix some $x^n \in \mathcal{T}_{\delta'}(X)$. Then,

$$\begin{aligned}
1 &\geq \sum_{y^n \in \mathcal{T}_\delta(Y|x^n)} p(y^n|x^n) = \sum_{y^n \in \mathcal{T}_\delta(Y|x^n)} \frac{p(x^n, y^n)}{p(x^n)} \\
&\geq |\mathcal{T}_\delta(Y|x^n)| \cdot 2^{-n(H(X,Y)-H(X)+\epsilon(\delta))} = |\mathcal{T}_\delta(Y|x^n)| \cdot 2^{-n(H(Y|X)+\epsilon(\delta))} \\
&\Rightarrow |T_\delta(Y|x^n)| \leq 2^{n(H(Y|X)+\epsilon(\delta))}.
\end{aligned}$$

7.5 Encoding – Decoding Schemes for Sending Messages

We now explain the idea of how these concepts relate to sending messages over channels. Consider a discrete memoryless channel (DMC), described by conditional probability distribution $p(y|x)$, where x is the input to the channel and y is the output. Then,

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

Say we want to send one of M messages over a channel. We encode each $m \in \{1, \dots, M\}$ into a codeword $X^n(m)$. We then send the codeword over the channel, obtaining Y^n. Finally, we use a decoding rule $\hat{M}(Y^n)$ which yields $\hat{M} = m$ with high probability.

The conditional typicality lemma (proved in the homework) characterizes the behavior of this channel for large n: if we choose a typical input, then the output is essentially chosen uniformly at random from $\mathcal{T}_\delta(Y|x^n)$. More precisely, for all $\delta' < \delta$, $x^n \in \mathcal{T}_{\delta'}(X)$ implies that

$$P(y^n \in T_\delta(Y|x^n)) = P((x^n, Y^n) \in \mathcal{T}_\delta(X, Y)) \to 1,$$

as $n \to \infty$; furthermore, the probability of obtaining each element of $\mathcal{T}_\delta(Y|x^n)$ is essentially $\frac{1}{|\mathcal{T}_\delta(Y|x^n)|} \approx 2^{-nH(Y|X)}$.

This lemma limits the values of M for which there is an encoding – decoding scheme that can succeed with high probability. As illustrated in Figure 7.3, what we are doing is choosing a typical codeword $x^n \in \mathcal{T}_\delta(X)$, and receiving some element $Y^n \in T_\delta(Y|x^n)$. We can think of this latter set as a "noise ball": it is the set of outputs y^n that we could typically expect to receive, given that our input is x^n. If these noise balls corresponding to different inputs overlap significantly, then we have no hope for being able to obtain m from Y^n with high probability, as multiple inputs give indistinguishable outputs. Since, for any input, the output will (with high probability) be typical – that is, $Y^n \in T_\delta(Y)$, the number of messages we can send is limited by the number of noise balls we can fit inside of $T_\delta(Y)$. Since the number of elements of $T_\delta(Y)$ is (approximately) $2^{nH(Y)}$ and the number of elements of $T_\delta(Y|x^n)$ is (approximately) $2^{nH(Y|X)}$, it follows that the number of messages we can send over this channel is at most

$$\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)} \leq 2^{nC},$$

where C is the channel capacity. Note that this argument does not give a construction that lets us attain this upper bound on the communication rate. The magic of the direct part of Shannon's channel coding theorem is that random coding lets us attain this upper bound.

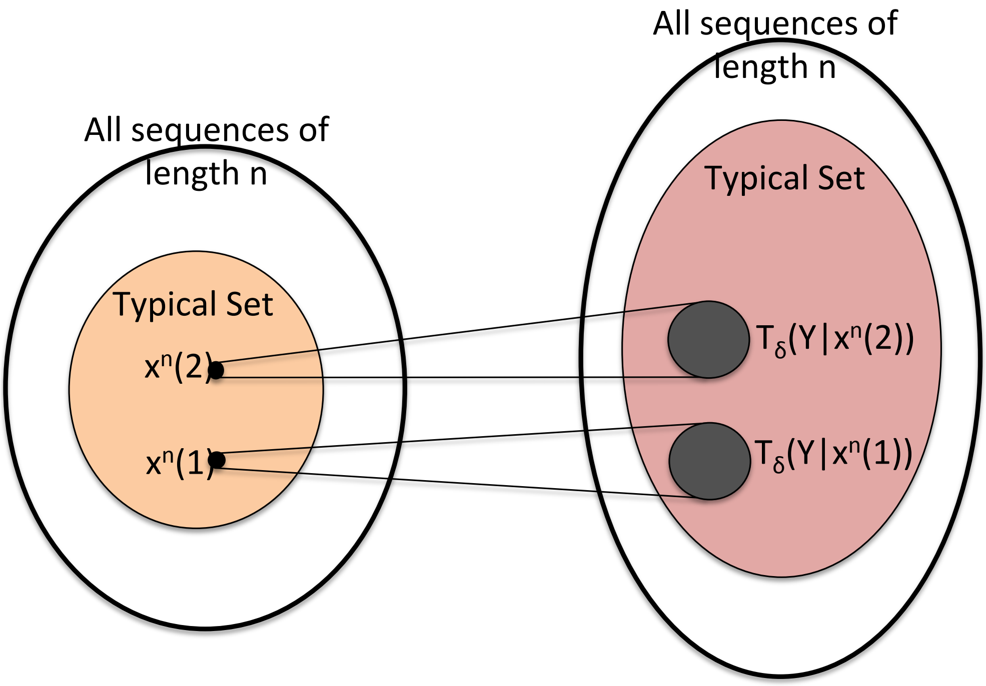


Figure 7.3: The typical sets $T_\delta(X), T_\delta(Y)$, and $T_\delta(Y|x^n)$.

7.6 Joint Typicality Lemma

In this final section, we discuss the joint typicality lemma, which tells us how well we can guess the output without knowing the input. Intuitively, if X and Y are strongly correlated, then we might expect that not knowing the input could strongly impair our ability to guess the output, and if X and Y are independent then not knowing the input should not at all impair our ability to guess the output. So, say that the actual input is x^n. We will look for a bound on the probability that an output will be in the conditional typical set $\mathcal{T}_\delta(Y|x^n)$ – that is, the probability that we'll guess that x^n was the input – in terms of the mutual information $I(X;Y)$.

Fix any $\delta > 0$ and $\delta' < \delta$, and fix $x^n \in T_{\delta'}(X)$. Choose $\tilde{Y}^n \in \mathcal{Y}^n$ by choosing each $\tilde{Y}_i$ i.i.d. according to the marginal distribution $p(y)$ (so, intuitively we've forgotten what we sent as input to the channel, and are simulating the output). Then, noting that

$$y^n \in \mathcal{T}_\delta(Y|x^n) \Rightarrow y^n \in \mathcal{T}_\delta(Y) \Rightarrow p(y^n) \leq 2^{-n(H(Y)-\epsilon(\delta))},$$

where $\epsilon(\delta)$ is a function that approaches 0 as $\delta \to 0$, we have

$$\begin{aligned}
P(\tilde{Y}^n \in \mathcal{T}_\delta(Y|x^n)) &= P((x^n, \tilde{Y}^n) \in \mathcal{T}_\delta(X,Y)) \\
&= \sum_{y^n \in \mathcal{T}_\delta(Y|x^n)} p(y^n) \\
&\leq |T_\delta(Y|x^n)| \cdot 2^{-n(H(Y)-\epsilon(\delta))} \\
&\leq 2^{nH(Y|X)+\epsilon(\delta)} \cdot 2^{-n(H(Y)-\epsilon(\delta))} \\
&= 2^{-n(H(Y)-H(Y|X)-\tilde{\epsilon}(\delta))} \\
&= 2^{-n(I(X;Y)-\tilde{\epsilon}(\delta))},
\end{aligned}$$

where $\tilde{\epsilon}(\delta) \to 0$ as $\delta \to 0$.

Intuitive argument for joint typicality lemma The joint typicality lemma asserts that the probability of observing two random x^n and y^n sequences is roughly $2^{-nI(X;Y)}$. Observe that there are roughly $2^{nH(X)}$ typical x^n sequences, and $2^{nH(Y)}$ typical y^n sequences. The total number of jointly typical sequences is $2^{nH(X,Y)}$. Thus, what is the probability that two randomly chosen sequences are jointly typical?

$$\approx \frac{2^{nH(X,Y)}}{2^{nH(X)} \times 2^{nH(Y)}} = 2^{-nI(X;Y)} \tag{7.4}$$

Chapter 8

Lossy Compression & Rate Distortion Theory

8.1 Definitions and main result

$$\xrightarrow{\mathcal{U}^N = U_1, \ldots, U_N} \boxed{\text{Encoder}} \xrightarrow{\quad J \in \{1, 2, \ldots, M\} \quad} \boxed{\text{Decoder}} \xrightarrow{\mathcal{V}^N = V_1, \ldots, V_N}$$

where $U_i \sim U$, i.i.d.

A scheme is characterized by:

- N, M
- An encoder, i.e., a mapping from $\mathcal{U}^N$ to $J \in \{1, 2, ..., M\}$ ($\log M$ bits used to encode a symbol sequence, where a symbol sequence is U^N and a symbol is U_i)
- A decoder, i.e., a mapping from $J \in \{1, 2, ..., M\}$ to $\mathcal{V}^N$

In working with lossy compression, we examine two things:

1. Rate $R = \frac{\log(M)}{N} \ \frac{\text{bits}}{\text{source symbol}}$
2. Expected distortion (figure of merit) $= d(U^N, V^N) = E[\frac{1}{N}\sum_{i=1}^{N} d(U_i, V_i)]$ (we always specify distortion on a per-symbol basis, and then average the distortions to arrive at $d(U^N, V^N)$)

There's a trade-off between rate and $\frac{\text{distortion}}{\text{symbol}}$. Distortion theory deals with this trade-off.

Definition 70. *(R,D) is* <u>*achievable*</u> *if* $\forall \epsilon > 0 \ \exists$ *scheme (N,M,encoder,decoder) such that* $\frac{\log M}{N} \leq R + \epsilon$ *and* $\mathbb{E}[d(U^N, V^N)] \leq D + \epsilon$

Definition 71. $R(D) \triangleq \inf\{R' : (R', D)$ *is achievable*$\}$

Definition 72. $R(D)^{(I)} \triangleq \min_{\mathbb{E}[d(U,V)] \leq D} I(U; V)$

Theorem 73. $R(D) = R^{(I)}(D)$.

Proof

$$\Leftrightarrow \begin{cases} \text{Direct Part} : R(D) \leq R^{(I)}D \\ \text{Converse Part} : R(D) \geq R^{(I)}D \end{cases}$$

The proof of the direct part and the converse part are given below.

□

Note that $R(D)$ is something we can't solve for (solution space is too large!), but $R^{(I)}(D)$ is something we can solve for (solution space is reasonable).

Theorem 74. *$R(D)$ is convex, i.e.,*
$\forall 0 < \alpha < 1 \ , \ D_0 \ , \ D_1 \ : \ R(\alpha D_0 + (1-\alpha)D_1) \leq \alpha R(D_0) + (1-\alpha)R(D_1)$

Sketch of proof: We consider a "time-sharing" scheme for encoding N bits. We encode the first αN bits using a "good" scheme for distortion $D = D_0$ and encode the last $(1-\alpha)N$ bits using a "good" scheme for $D = D_1$. Overall, the number of bits in the compressed message is $N\alpha R(D_0) + N(1-\alpha)R(D_1)$, so that the rate is $\alpha R(D_0) + (1-\alpha)R(D_1)$. Further, the expected distortion is the average, weighted by α between the distortions between the two different schemes, i.e. $\alpha D_0 + (1-\alpha)D_1$. We therefore have constructed a scheme which achieves distortion $\alpha D_0 + (1-\alpha)D_1$ with rate $\alpha R(D_0) + (1-\alpha)R(D_1)$, and the optimal scheme can only do better. That is

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

as desired.

8.2 Examples

Example 75.
Consider $U \sim \text{Ber}(p)$, $p \leq \frac{1}{2}$ and Hamming distortion. That is

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

Claim:

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

Proof: We will not be overly pedantic by worrying about small ϵ factors in the proof.

Note we can achieve distortion p without sending any information by setting $V = 0$. Therefore, for $D > p$, $R(D) = 0$, as claimed. For the remainder of the proof, therefore, we assume $D \leq p \leq \frac{1}{2}$.

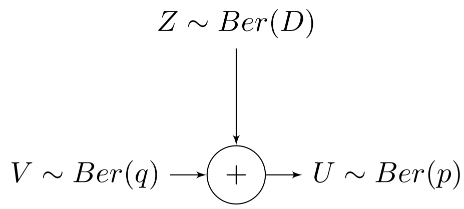
Consider U, V such that $U \sim \text{Ber}(p)$ and $\mathbb{E}d(U,V) = P(U \neq V) \leq D$. We can show that $R(D)$ is lower-bounded by $h_2(p) - h_2(D)$ by noting

$$\begin{aligned} I(U;V) &= H(U) - H(U|V) \\ &= H(U) - H(U \ominus_2 V|V) \\ &\geq H(U) - H(U \ominus_2 V) \\ &= h_2(p) - h_2(P(U \neq V)) \\ &\geq h_2(p) - h_2(D) \end{aligned}$$

In the second line we have used the fact that $H(U|V) = H(U \ominus_2 V|V)$ because there is a one to one mapping $(U,V) \leftrightarrow (U \ominus_2 V, V)$. In the third line, we have used that conditioning reduces entropy, so $H(U \ominus_2 V|V) \leq H(U \ominus_2 V)$. Finally, in the last line we have used that h_2 is increasing on $[0, \frac{1}{2}]$ and that $P(U \neq V) \leq D \leq p \leq \frac{1}{2}$. This establishes that $R(D) \geq h_2(p) - h_2(D)$.

Now we must show equality can be achieved. The first and second inequalities above demonstrate that we get equality if and only if

1. $U \ominus_2 V$ is independent of V.
2. $U \ominus_2 V \sim \text{Ber}(D)$.



Denoting $U \ominus_2 V \triangleq Z$, this is equivalent to finding q such that if $V \sim \text{Ber}(q)$ and $Z \sim \text{Ber}(D)$ is independent of V, $U = V \oplus_2 Z \sim \text{Ber}(p)$. Because $V \oplus_2 Z$ is binary, it is Bernoulli, with

$$\begin{aligned} p &= P(U = 1) \\ &= P(V = 1)P(Z = 0) + P(V = 0)P(Z = 1) \\ &= q(1 - D) + (1 - q)D \end{aligned}$$

Solving for q gives

$$q = \frac{p - D}{1 - 2D}$$

Because $D \leq p \leq \frac{1}{2}$, both the numerator and denominator are positive. Further, because $p \leq \frac{1}{2}$, we have $q \leq \frac{1/2 - D}{1 - 2D} = \frac{1}{2}$, which shows that q is a valid probability. This completes the proof.

Example 76. Consider $U \sim \text{N}(0, \sigma^2)$ and distortion given by: $d(u,v) = (U - V)^2$

Claim:

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

Proof: First note we may achieve distortion σ^2 without transmitting any information by setting $V = 0$ with certainty. Therefore, $R(D) = 0$ for $D > \sigma^2$. For the remainder of the proof, therefore, we assume that $D \leq \sigma^2$.

For any U,V such that $U \sim \mathrm{N}(0, \sigma^2)$ and $\mathbb{E}(U - V)^2 \leq D$, we assume $D \leq \sigma^2$.

We can find the lower-bound by noting

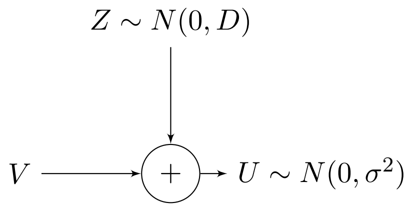
$$\begin{aligned}
I(U;V) &= h(U) - h(U|V) \\
&= h(U) - h(U - V|V) \\
&\geq h(U) - h(U - V) \\
&\geq h(U) - h(N(0, D)) \\
&= \frac{1}{2}\log 2\pi E\sigma^2 - \frac{1}{2}\log 2\pi e D \\
&= \frac{1}{2}\log \frac{\sigma^2}{D}
\end{aligned}$$

For the first inequality we have used that conditioning reduces even the differential entropy, and in the second inequality we have used the result, proved earlier in the course that the maximum differential entropy of a distribution constrained by $\mathrm{Var}(U - V) \leq D$ is achieved when $U - V \sim \mathrm{N}(0, D)$. This establishes that $R(D) \geq \frac{1}{2}\log\frac{\sigma^2}{D}$.

Now we must show that equality can be achieved. The first and second inequalities above demonstrate that we get equality if and only if

1. $U - V$ is independent of V.

2. $U - V \sim \mathrm{N}(0, D)$.



Denoting $U - V \triangleq Z$, we want to find a distribution for V such that Z independent of Z and distributed $\mathrm{N}(0, D)$ makes $V + Z \sim \mathrm{N}(0, \sigma^2)$. We see that this is possible for $V \sim \mathrm{N}(0, \sigma^2 - D)$, which is a valid distribution because $D \leq \sigma^2$. This completes the proof, and $R(D) = \frac{1}{2}\log\frac{\sigma^2}{D}$.

8.3 Proof of Direct Part $R(D) \leq R^{(I)}(D)$

8.3.1 An Equivalent Statement

First, we are going to show the equivalence of the following statements

$$\begin{aligned}
R(D) \leq R^{(I)}(D) &\iff R(D) \leq \min\{I(U;V) : U, V \text{ s.t. } \mathrm{E}\, d(U, V) \leq D\} \\
&\iff \text{If } U, V \text{ s.t. } \mathrm{E}\, d(U, V) \leq D, \text{ then } R(D) \leq I(U;V) \\
&\iff \text{If } U, V \text{ s.t. } \mathrm{E}\, d(U, V) \leq D, \text{ then } (R, D) \text{ is achievable for any } R > I(U;V).
\end{aligned}$$

Proof The first and second lines follow the definition of $R^{(I)}(D)$. For the last line, it only suffices to show

$$R(D) \leq I(U;V) \Longleftrightarrow (R,D) \text{ is achievable for any } R > I(U;V).$$

- For the $\Rightarrow$ part, consider any $R > I(U;V)$,

$$R > I(U;V) \geq R(D) = \inf\{R' : (R',D) \text{ is achievable}\}$$

  thus (R,D) is achievable.

- For the $\Leftarrow$ part, consider some $R'' = I(U;V) + \epsilon$. By the assumption (R,D) is achievable for any $R > I(U;V)$, implying that (R'',D) is achievable, and thereafter

$$R(D) = \inf\{R' : (R',D) \text{ is achievable}\} \leq R'' = I(U;V) + \epsilon.$$

  Since ϵ can be arbitrarily small, we must have $R(D) \leq I(U;V)$.

$\square$

Hence we can prove the equivalent statement instead of $R(D) \leq R^{(I)}(D)$. That's to show

$$(R,D) \text{ is the achievable for fixed } U,V \text{ s.t. } \mathrm{E}[d(U,V)] \leq D \text{ and fixed } R > I(U;V).$$

8.3.2 Two Useful Lemmas

The proof of the equivalent statement uses two lemmas appearing in the homeworks.Let's recall them in advance.

Lemma 77. ***(Joint Typicality Lemma)*** *Suppose* $u^n \in \mathcal{T}_{\delta'}(U)$, $0 < \delta' < \delta$ *and* V_i*'s* $\overset{i.i.d.}{\sim} V$,

$$2^{-n(I(U;V)+\epsilon(\delta))} \leq \mathbb{P}\left((u^n, V^n) \in \mathcal{T}_\delta(U,V)\right)$$

for sufficiently large n *and some* $\epsilon(\delta) > 0$ *where* $\lim_{\delta \to 0} \epsilon(\delta) = 0$.

Lemma 78. ***(Typical Average Lemma)***

$$(u^n, v^n) \in \mathcal{T}_\delta(U,V) \Longrightarrow d(u^n, v^n) \triangleq \frac{1}{n}\sum_{i=1}^{n} d(u_i, v_i) \leq (1+\delta)\, \mathrm{E}\, d(U,V)$$

8.3.3 Proof of the Equivalent Statement

For fixed U,V s.t. $\mathrm{E}\, d(U,V) \leq D$ and $R > I(U;V)$, we are going to show (R,D) is achievable.

Proof Take $M = \lfloor 2^{nR} \rfloor$. Denote by $C_n = \{V^n(1), V^n(2), ..., V^n(M)\}$ the random codebook which is generated by V_i's $\overset{i.i.d.}{\sim} V$ and independent of U. Let $d(u^n, C_n) = \min_{V^n \in C_n} d(u^n, V^n)$.

For sufficient small $0 < \delta' < \delta$ which appear in Lemma 77, the assumption $R > I(U,V)$ implies $R > I(U;V) + \epsilon(\delta)$. For any $u^n \in \mathcal{T}_{\delta'}(U)$ and sufficiently large n,

$$
\begin{aligned}
\mathbb{P}\left(d(u^n, C_n) > D(1+\delta)\right) &= \mathbb{P}\left(d(u^n, V_n(i)) > D(1+\delta) \text{ for } i = 1, 2, ..., M\right) \\
&\hspace{20em} \text{(Definition of } d(u^n, C_n)) \\
&= \mathbb{P}\left(d(u^n, V_n(1)) > D(1+\delta)\right)^M && (V_i \overset{i.i.d.}{\sim} V) \\
&\leq \mathbb{P}\left(d(u^n, V_n(1)) > \mathrm{E}\, d(U,V)(1+\delta)\right)^M && \text{(Assumption of } \mathrm{E}\, d(U,V) \leq D) \\
&\leq \mathbb{P}\left((u^n, V_n(1)) \notin \mathcal{T}_\delta(U,V)\right)^M && \text{(Inverse-negative of Lemma 78)} \\
&= \left[1 - \mathbb{P}\left((u^n, V_n(1)) \in \mathcal{T}_\delta(U,V)\right)\right]^M \\
&\leq \left[1 - 2^{-n(I(U;V)+\epsilon(\delta))}\right]^M && \text{(Lemma 77 with } u^n \in \mathcal{T}_{\delta'}(U) \text{ and large } n) \\
&\leq \exp\left(-M \cdot 2^{-n(I(U;V)+\epsilon(\delta))}\right) && (1 - x \leq e^{-x})
\end{aligned}
$$

So far, we have an upper bound of $\mathbb{P}\left(d(u^n, C_n) > D(1+\delta)\right)$ for any $u^n \in \mathcal{T}_{\delta'}(U)$ and sufficiently large n.

$$
\mathbb{P}\left(d(u^n, C_n) > D(1+\delta)\right) \leq \exp\left(-M \cdot 2^{-n(I(U;V)+\epsilon(\delta))}\right) \tag{8.1}
$$

Then for $U_i \overset{i.i.d.}{\sim} U$,

$$
\begin{aligned}
\mathbb{P}\left(d(U^n, C_n) > D(1+\delta)\right) &= \sum_{u^n \in \mathcal{T}_{\delta'}(U)} \mathbb{P}\left(d(u^n, C_n) > D(1+\delta), U^n = u^n\right) \\
&\quad + \sum_{u^n \notin \mathcal{T}_{\delta'}(U)} \mathbb{P}\left(d(u^n, C_n) > D(1+\delta), U^n = u^n\right) \\
&\leq \sum_{u^n \in \mathcal{T}_{\delta'}(U)} \mathbb{P}\left(d(u^n, C_n) > D(1+\delta)\right) \mathbb{P}\left(U^n = u^n\right) \\
&\hspace{20em} (U^n \text{ independent of } C_n) \\
&\quad + \mathbb{P}\left(U^n \notin \mathcal{T}_{\delta'}(U)\right) \\
&\leq \exp\left(-M \cdot 2^{-n(I(U;V)+\epsilon(\delta))}\right) + \mathbb{P}\left(U^n \notin \mathcal{T}_{\delta'}(U)\right) \\
&\hspace{20em} \text{(Upper bound in Eq. 8.1)}
\end{aligned}
$$

where the first term goes to 0 as $n \to \infty$ because

$$
M = \lfloor 2^{nR} \rfloor, \quad R > I(U;V) + \epsilon(\delta),
$$

and the second term goes to 0 as $n \to \infty$ because of AEP. Thus for $U_i \overset{i.i.d.}{\sim} U$,

$$
\mathbb{P}\left(d(U^n, C_n) > D(1+\delta)\right) \to 0 \text{ as } n \to \infty \tag{8.2}
$$

Further, let $d(C_n) = \mathrm{E}\left(d(U^n, C_n) | C_n\right)$ be the average distortion by random codebook C_n, and thus $d(c_n) = \mathrm{E}\left(d(U^n, c_n) | C_n = c_n\right) = \mathrm{E}\left(d(U^n, c_n)\right)$ (C_n is independent of U^n) is the average distortion by a realization c_n of C_n.

$$\begin{aligned}
\mathrm{E}\, d(C_n) &= \mathrm{E}\left[\mathrm{E}\left(d(U^n, C_n)|C_n\right)\right] \\
&= \mathrm{E}\left(d(U^n, C_n)\right) && \text{(tower property)} \\
&\leq \mathbb{P}\left(d(U^n, C^n) > D(1+\delta)\right) D_{max} && (D_{max} \triangleq \max_{u\in\mathcal{U}, v\in\mathcal{V}} d(u,v)) \\
&\quad + \mathbb{P}\left(d(U^n, C^n) \leq D(1+\delta)\right) D(1+\delta) \\
&\to D(1+\delta) \text{ as } n \to \infty && \text{(Limiting result in Eq. 8.2)}
\end{aligned}$$

It implies that

$$\mathrm{E}\, d(C_n) < D + 2\delta D_{max} \text{ for sufficiently large } n,$$

which further implies existence of c_n, a realization of C_n, satisfying

$$d(c_n) \leq \mathrm{E}\, d(C_n) < D + 2\delta D_{max} \text{ for sufficiently large } n.$$

Taking arbitrarily small δ and sufficiently large n, we can get the average distortion $d(c_n)$ arbitrarily close to D. And the size of codeword lists

$$|c_n| = M = \lfloor 2^{nR} \rfloor \leq 2^{nR}.$$

(R, D) is achieved by the codebook c_n. □

8.4 Proof of the converse

Proof

Fix a scheme satisfying $E[d(U^N, V^N)] \leq D$, then $H(V^N) \leq \log M$ for V^N taking M different values.

$$\begin{aligned}
\log M &\geq H(V^N) \\
&\geq H(V^N) - H(V^N \mid U^N) \\
&= I(U^N; V^N) \\
&= H(U^N) - H(U^N \mid V^N) \\
&= \sum_{i=1}^{N} H(U_i) - H(U_i \mid U^{i-1}, V^N) \text{ (by chain rule)} \\
&\geq \sum_{i=1}^{N} H(U_i) - H(U_i \mid U^{i-1}, V_i) \text{ (conditioning reduces entropy)} \\
&= \sum_{i=1}^{N} I(U_i; V_i) \\
&\geq \sum_{i=1}^{N} R^{(I)}(E[d(U_i, V_i)]) \text{ (by definition of} R^{(I)}(D)) \\
&= N \sum_{i=1}^{N} \frac{1}{N} R^{(I)}(E[d(U_i, V_i)]) \text{ (average of} R^{(I)}(D) \text{over all } i) \\
&\geq N R^{(I)}(\frac{1}{N} \sum_{i=1}^{N} E[d(U_i, V_i)]) \text{ (By the convexity of } R^{(I)}(D)) \\
&\geq N R^{(I)}(D) \ (R^{(I)}(D) \text{ is nonincreasing}) \\
\text{rate} &= \frac{\log M}{N} \geq R^{(I)}(D)
\end{aligned}$$

□

8.5 Geometric Interpretation

$I(U;V)$ is the expected distortion if both U, V are in jointly typical set, as we just proved. The following figures will give a geometric interpretation to the results.

How large does a codebook has to be so that every source sequence in the typical set has a reconstruction, which it is jointly typical? Let $T(U \mid V^N(i))$ be the set of source sequences jointly typical with the reconstruction sequence $V^N(i)$. Therefore, to cover every source sequence, we need a codebook of at least the size of typical set of the input divided by the number of source sequences one reconstruction can cover.
The size of the codebook $= \dfrac{|T(U)|}{|T(U \mid V^N(i))|} \approx \dfrac{2^{NH(U)}}{2^{NH(U|V)}} = 2^{NI(U;V)}$. This is showed in Fig. 8.1 on the distortion function.

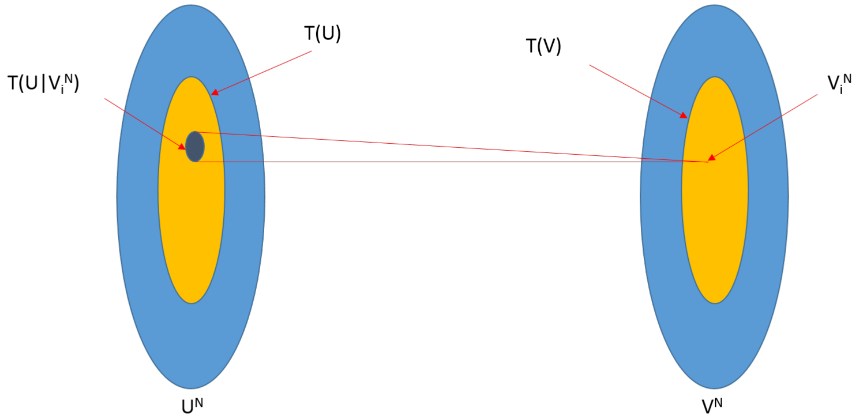


Figure 8.1: Distortion Function

Achievability: generate $V^N(i)$ i.i.d. $\sim V$, $P((U^N, V^N(i)) \in T(U,V)) \approx 2^{-nI(U;V)}$. Therefore, in order for all typical sequences to be described, $P((U^N, V^N(i)) \in T(U,V), i = 1,\ 2,\ \cdots,\ M) \approx 1$ if the codebook is sufficiently large, i.e., if $R > I(U;V)$, as the codebook is of size $\lfloor 2^{NR} \rfloor$

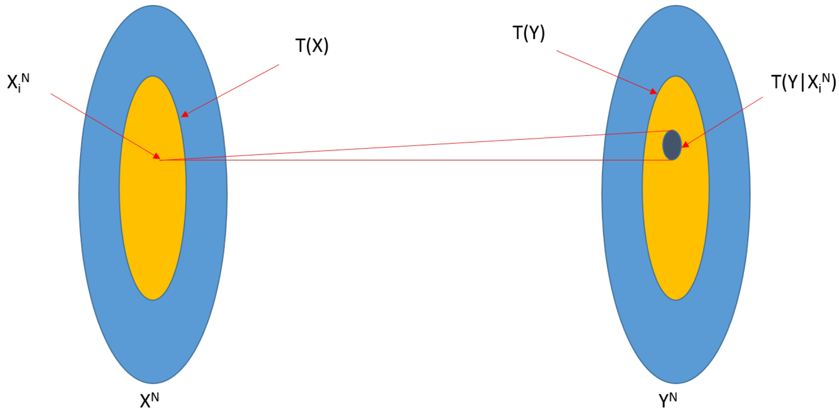


Figure 8.2: Communication Channel

The communication problem has a similar setup. In order to achieve reliable communication, the number of messages $\leq \dfrac{|T(Y)|}{|T(Y \mid X^n(i))|} = \dfrac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{-nI(X;Y)}$. This communication channel is shown in Fig. 8.2, as the distortion occurs on $\mathcal{V}^N$.

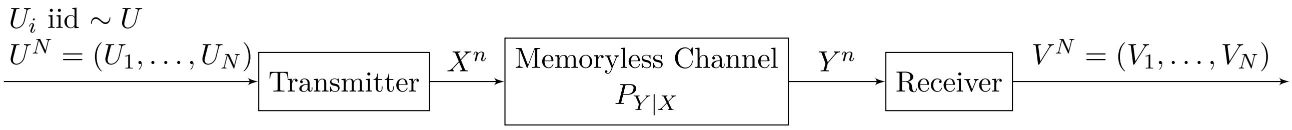
Achievability: $\forall i$ s.t. $P(Y^n \in T(Y^n \mid X^n(i)|i \neq j) \approx 2^{-nI(X;Y)}$. Therefore, because the number of messages is $\lfloor 2^{NR} \rfloor$, in order to guarantee that $P(Y^n \in T(Y^n \mid X^n(i))$for any $i \neq j) \approx 0$, $R < I(X;Y)$.

Chapter 9

Joint Source Channel Coding

9.1 Joint Source Channel Coding

Now that we understand lossy compression as well as the communication problem, we can combine them into a joint source-channel coding theorem. A schematic of this setup is shown below:



With this channel description, the goal is to communicate the $U^N = (U_1, U_2, \ldots, U_N)$ through the memoryless channel given by $P_{Y|X}$ with small expected distortion, measured by $\mathbb{E}\left[d(U^N, V^N)\right]$. In other words, the goal is to find the best possible distortion given some rate and some noise during transmission. Note that the U_i are not necessarily bits.

The rate of communication is then

$$\text{rate} = \frac{N}{n} \frac{\text{source symbols}}{\text{channel use}}$$

We also allow an expected distortion $\mathbb{E}\left[d(U^N, V^N)\right]$, with

$$d(U^N, V^N) = \frac{1}{N} \sum_{i=1}^{N} d\left(U_i, V_i\right).$$

Definition 79. *A rate-distortion pair (ρ, D) is achievable if* $\forall \epsilon > 0, \exists$ *a scheme with* $\dfrac{N}{n} \geq \rho - \epsilon$ *and* $E[d(U^N, V^N)] \leq D + \epsilon$.

Note: under any scheme, $E[d(U^N, V^N)] \leq D$, and $U^N \to X^n \to Y^n \to V^N$ forms a Markov chain. Therefore,

$$
\begin{aligned}
&nC \geq I(X^n; Y^n) \text{ (proven in channel coding converse theorem)}\\
&I(X^n; Y^n) \geq I(U^N; V^N)) \text{ (Data processing inequality)}\\
&I(U^N; V^N) \geq NR(D) \text{ (proven in converse of rate distortion theorem)}\\
&\frac{NR(D)}{n} = \text{Rate} \cdot R(D) \leq C.
\end{aligned}
$$

Thus, if (ρ, D) is achievable $\Rightarrow \rho R(D) \leq C$.

Consider the following “Separation Scheme"

$\xrightarrow{U^N}$ Good Distortion Compressor $\xrightarrow{N \cdot R(D) \text{ bits}}$ Reliable Channel Encoder $\xrightarrow{X^n}$ Memoryless Channel

$\xrightarrow{Y^N}$ Reliable Channel Decoder $\xrightarrow{N \cdot R(D) \text{ bits}}$ Good Distortion Compressor $\xrightarrow{V^N}$

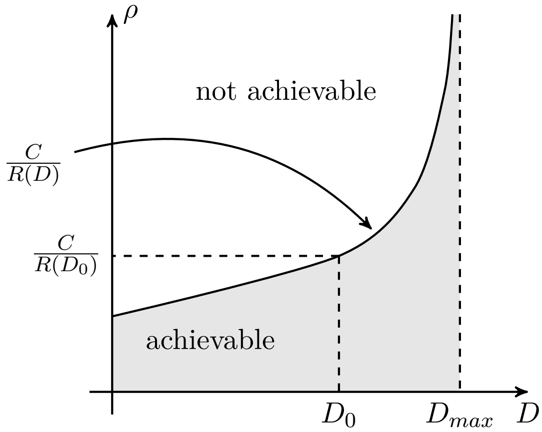
All these pieces work correctly to ensure that distortion and channel noise are handled properly.

It is guaranteed that $E[d(U^N, V^N)] \approx D$ provided that $n \cdot C \geq NR(D) \cdot C \geq \frac{N}{n} \cdot R(D) =$ rate $\cdot R(D)$. Thus, if $R(D) \leq C$, then (ρ, D) is achievable.

9.2 Source – Channel Separation Theorem

Theorem 80. *(ρ, D) is achievable if and only if $\rho \cdot R(D) \leq C$.*

Essentially, we have separated the problem of compression from the problem of transmission and have proven that a separated solution is optimal. There is no need nor advantage to address both problems simultaneously.



We can achieve points on the curve above by first thinking about representing the data as bits in an efficient manner (compression) with rate $R(D)$ and then transmitting these bits losslessly across the channel with rate C. Note that the distortion without sending anything over the channel is D_{max}.

9.3 Examples

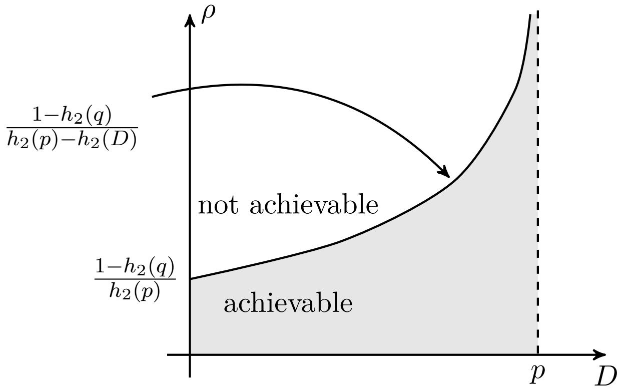
Example 81. Binary Source and Binary Channel

Source: $U \sim \mathrm{Ber}(p), \quad 0 \le p \le 1/2$
Channel: $\mathrm{BSC}(q), \quad 0 \le q \le 1/2$
Distortion: Hamming

Recall that for $U \sim \mathrm{Ber}(p)$, the rate distortion function is $R(D) = h_2(p) - h_2(D)$ and that a binary symmetric channel with crossover probability q has capacity $C = 1 - h_2(q)$

So, we see that if we want distortion $\le D$, then (for $D \le p$) the maximum achievable rate is:

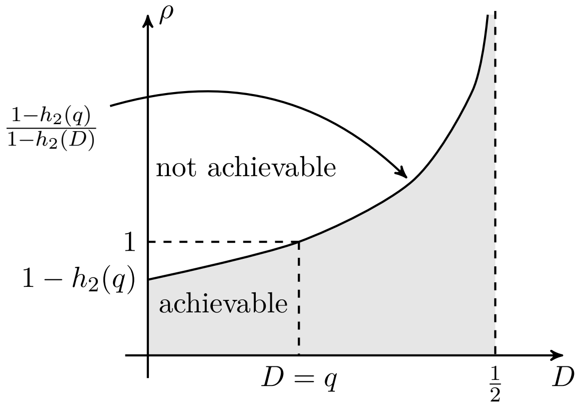
$$\rho = \frac{1 - h_2(q)}{h_2(p) - h_2(D)}$$



Note that the communication problem corresponds to $D = 0$.

In particular, if $p = 1/2$, then if we want distortion $\le D$, the maximum rate we can transmit at is:

$$\rho = \frac{1 - h_2(q)}{1 - h_2(D)}$$



Consider the following scheme for rate=1:

Channel input: $X_i = U_i$
reconstruction: $V_i = Y_i$

The expected distortion is then $P(U_i \neq V_i) = P(X_i \neq Y_i) = q$

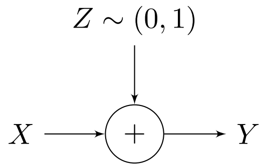
$\longrightarrow$ this scheme is optimal, since $\rho = (1 - h_2(q))/(1 - h_2(D = q)) = 1$.

In this particular case, it is possible to achieve the optimal rate using a scheme that individually encodes and transmits each symbol.

Example 82. Gaussian Source and Gaussian Channel

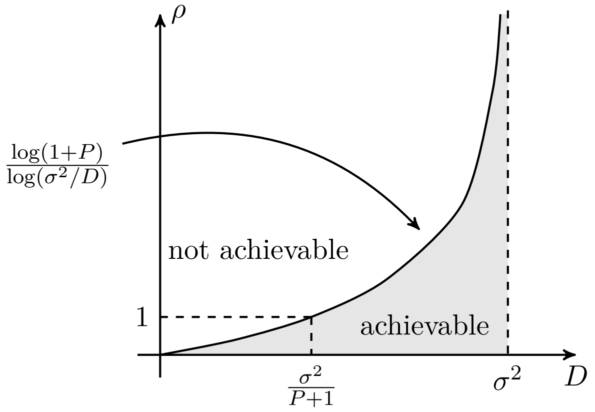
Source: $U \sim \mathcal{N}(1, \sigma^2)$

Channel: AWGN (Additive White Gaussian Noise Channel) with power constraint P



distortion: squared error

Recall that for $U \sim \mathcal{N}(1, \sigma^2)$, the rate distortion function is $R(D) = \frac{1}{2}\log(\frac{\sigma^2}{D})$ (for $0 \leq D \leq \sigma^2$) and that the AWGN channel with power constraint P has capacity $C = \frac{1}{2}\log(1 + P)$



Then, for a given distortion $D \leq \sigma^2$, the maximum achievable rate is

$$\rho = \frac{\log(1 + P)}{\log(\sigma^2/D)}$$

Consider the following scheme at rate=1:

transmit: $X_i = \sqrt{\frac{P}{\sigma^2}} U_i$

receive: $Y_i = X_i + Z_i = \sqrt{\frac{P}{\sigma^2}} U_i + Z_i$
reconstruction: $V_i = \mathbb{E}[U_i|V_i]$

The distortion is squared error, so we know that reconstruction using the expected value is optimal. Thus, we take $V_i = \mathbb{E}[U_i|V_i]$.

The expected distortion is then:

$$
\begin{aligned}
\mathbb{E}[U_i|V_i] &= \mathrm{Var}(U_i|V_i) \\
&= \mathrm{Var}\left(U_i \middle| \sqrt{\frac{\sigma^2}{P}} Y_i\right) \\
&= \mathrm{Var}\left(U_i \middle| U_i + \sqrt{\frac{\sigma^2}{P}} Z_i\right) \\
&\overset{(a)}{=} \frac{\sigma^2(\sigma^2/P)}{\sigma^2 + \sigma^2/P} \\
&= \frac{\sigma^2}{P+1}
\end{aligned}
$$

where (a) follows from the fact that for $X \sim \mathcal{N}(0, \sigma_1^2)$ independent from $Y \sim \mathcal{N}(0, \sigma_2^2)$:

$$
\mathrm{Var}(X|X+Y) = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}
$$

Now, at rate $= 1$:
The optimal D satisfies

$$
\begin{aligned}
\frac{\log(1+P)}{\log(\sigma^2/D)} &= 1 \\
\rightarrow 1 + P &= \frac{\sigma^2}{D}
\end{aligned}
$$

So, in the specific case of rate $= 1$ we see that the simple scheme above is optimal, just as the simple scheme for the Binary Source and Channel was also optimal when rate $= 1$.