Information Theory	**Prof. I-Hsiang Wang**
Lecture Notes: Source Coding	
October 20, 2015	**Scribe: Chi-Ning Chou**

This chapter is focusing on the fundamental limit of source coding. And we are going to find the fundamental limit for different setting. On the other hand, some useful techniques will also be introduced such as typical sequence, Fano's inequality. The outline is as follow:

1. Overview and intuitions.
2. Typical method & AEP.
3. Memoryless source coding.
4. Source with memory.

Details proof will be left in the appendix.

1 Overview and intuitions about source coding

1.1 Abstract model

When it comes to design a communication system, there are two important elements: **mechanism**, and **criteria**. Mechanism describes the basic components, and functions in the system while the criteria provides a way for us to evaluate the performance of the system.

In the context of source coding, the basic mechanism is: encoder and decoder.

- **Encoder**: A function maps (encodes) each source string to binary codeword.

$$Enc : \mathcal{X}^n \to \{0,1\}^K$$

- **Decoder**: A function maps (decodes) each binary codeword to source string.

$$Dec : \{0,1\}^n \to \mathcal{X}^n$$

Figure 1 shows the framework of source coding.

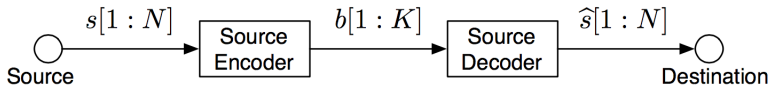

Figure 1: Framework of source coding.

Next, we are going to examine what's the criteria of source coding. Actually, there are two of them, one for encoding and one for decoding: efficiency and correctness.

- **Efficiency**: How efficient can we represent source string with binary codeword?

$$R := \frac{K}{N}$$

- **Correctness**: How well the scheme can produce correct output string?

$$Pr[S \neq \hat{S}]$$

1.2 Mechanism

To construct encoder and decoder, we have to first build up the environment. Here we consider a so called **block-to-variable** source coding scheme, and it has two important intuitions:

- **Block**: We view input source as a block of message instead of s sequence of character string.
- **Variable**: We allow the binary codeword to have different length, i.e. variable length.

1.3 Criteria

As we mentioned in Section 1.1, there are two important criteria: efficiency and correctness.

Efficiency The efficiency can be evaluated by the code rate $R = \frac{K}{N}$. That is, if R is small, it means that we can use smaller amount of binary bits to represent the information from the source terminal under certain correctness criteria. Intuitively, it's the compression rate of the source code.

As a result, here we might be interested in the following questions:

Q: What's the fundamental limit of R when given a certain correctness criteria?

Q: How can we achieve such efficiency?

Q: Is our implementation really efficiency? (In other aspects)

Correctness The correctness of the source coding scheme can have various definition, but it all directly relates to the error probability. Note that the error here refers to the decoding error. That is $Pr[S \neq Enc[Dec(S)]]$. And there are three common correctness criteria:

- **Exact**: $\forall s^N$, $Pr[s^N \neq \hat{s}^N] = 0$, $\forall N$.
- **Lossy**: Control $Pr[s^N \neq \hat{s}^N]$ in a range.
- **Lossless**: Control $Pr[s^N \neq \hat{s}^N]$ in a range, but asking $\lim_{N\to} Pr[s^N \neq \hat{s}^N] = 0$.

1.4 Conclusion

2 Typical method & AEP

Before we formally introduce the typical method and AEP, let's first conduct some thought experience about encoding and decoding scheme.

	Scheme	Criteria
Encoder	Block	Efficiency: $R = \frac{K}{N}$
Decoder	Variable	Correctness: exact, lossy, lossless

Table 1: Overview of source coding.

2.1 Encoding and decoding

The job for encoding and decoding is to build a mapping between alphabet sequence set and bit sequence set. That is: $\mathcal{X}^N \to \{0,1\}^K$. By simple counting, we can find out that there are $|\mathcal{X}|^N$ possible alphabet sequences to be encoded and 2^K bit sequences as a codebook. Clearly that as $|\mathcal{X}|^N \geq 2^K$, we can construct an one-to-one mapping and yield a exact correct mechanism. However, this is extremely a waste of resource due to the redundancy of source sequence. Most of the time, there will be some sequences that have very small probability to show up. If we allocate a position in the codebook to them, these codes will seldom be used and result in a waste.

On the other hand, we want to make the encoding sequences looked like an **uniform** source. To fulfill this goal, we have to let the probability of the sequence in the codebook roughly be the same. And such criteria requires the uniformity of encoded sequence. That is, we want the sequences to have similar probability.

Intuition (Encoding and decoding goal)

To be more efficient, we will not encode the whole sequence space $|\mathcal{X}|^N$. Instead, we will only encode a subset (category) of sequences. Moreover, we want this category to have two properties:

- The total probability of this category is large.
- The probability of each sequence to show up is roughly the same.

2.2 Typical set & AEP

What we would like to focus on is the source sequences that are in a more likely to show up category. In other words, the total probability of this category to show up is very high. And the sequences in the category are almost the same or with high symmetry. For example, consider a Bernoulli random variable with head probability 0.7. The probability for sequence 11111 to show up is greater than 11100. But if we categorize the sequence with the number of 1, we can find out that the probability of category with three 1s will be greater than that of the category with five 1s.

Combine the above example with the intuition in the previous section, now we can formally define the property we want. In information theory, such property is called AEP (Asymptotic Equipartition Property)

Definition 1 (AEP) *A sequence of sets $\mathcal{T}^n$ is said to have AEP if the following properties holds:*

- $\lim_{n\to\infty} Pr[X^n \in \mathcal{T}^n] = 1$
- $\forall x^n \in \mathcal{X}^n,\ \lim_{n\to\infty} |Pr[x^n] - \frac{1}{|\mathcal{T}^n|}| = 0$

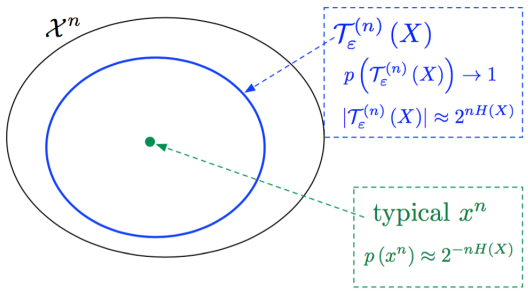

Figure 2: AEP

Intuition (Typical set & AEP)

For now, the definition is still rough, but the central ideas are clear. That is,

- The asymptotic probability of typical set is 1
- The probability of each element is asymptotically the same.

2.3 Typicality and Weakly typicality

With the intuition in the previous section, now we can give two real examples of typical set: typicality and weakly typicality.

Definition 2 (typicality) *A ϵ-typical set of random process $\{X_t\}$ is*

$$\mathcal{T}_\epsilon^n := \{x^n \in \mathcal{X}^n : |\pi(a|x^n) - p_X(a)| \leq \epsilon p_X(a),\ \forall a \in \mathcal{X}\}$$

, where $\pi(\cdot|x^n)$ is the empirical distribution of x^n over alphabet set $\mathcal{X}$.

Definition 3 (weakly typicality) *A weakly ϵ-typical of random process $\{X_t\}$ is*

$$\mathcal{A}_\epsilon^n := \{x^n : 2^{-n(H(X)+\delta(\epsilon)} \leq Pr[x^n] \leq 2^{-n(H(X)-\delta(\epsilon)}\}$$

Remark Typicality is often used in network and weakly typicality is often used in source with memory.

We can easily check that both typicality and weakly typicality implies AEP. See Appendix A. However, these two concepts have some intrinsically different intuition:

	Instance's factor	Objective factor	Notation
Typicality	Empirical distribution: $\pi(\cdot, x^n)$	$p_X(\cdot)$	$\mathcal{T}_\epsilon^n$
Weakly typicality	Probability mass: $-\frac{1}{n}\log P(x^n)$	$H(X)$	$\mathcal{A}_\epsilon^n$

Table 2: Typicality v.s. weakly typicality.

Note that $\mathcal{T}_\epsilon^n \subseteq \mathcal{A}_\epsilon^n$ but the converse doesn't hold.

Before we present and prove the source coding theorem, let's give a formal introduction to AEP:

Proposition 4 (AEP) *Both $\mathcal{T}_\epsilon^n$ and $\mathcal{A}_\epsilon^n$ satisfy the following asymptotic equipartition properties:*

1. $\forall x^n \in \mathcal{T}_\epsilon^n,\ 2^{-n(H(X)+\delta(\epsilon)} \leq Pr[x^n] \leq 2^{-n(H(X)-\delta(\epsilon)}$. *Here,* $\delta(\epsilon) = \epsilon H(X)$
2. $\lim_{n\to\infty} Pr[X^n \in \mathcal{T}_\epsilon^n] = 1$
3. $|\mathcal{T}_\epsilon^n| \leq 2^{n(H(X)+\delta(\epsilon)}$
4. $|\mathcal{T}_\epsilon^n| \geq (1-\epsilon)2^{n(H(X)-\delta(\epsilon)}$

3 Memoryless Source Coding

We start with the source coding theorem:

Theorem 5 (source coding) *Let $\{X_t\}$ be a memoryless i.i.d. source with finite support $\mathcal{X}$ and follows the p.m.f. p_X. Then, the optimal compression rate for* **lossless** *source coding is $R^* = H(X)$.*

Intuitively, Theorem 5 tells us that we can achieve lossless source coding with compression rate $H(X)$. Moreover, once we want to have lossless scheme, the compression rate cannot be smaller than $H(X)$. Somehow, this gives us an operational meaning for entropy.

To prove Theorem 5, we have to verify two directions:

- **Achievability**: Provide a scheme with compression rate $H(X)$ such that it is lossless.
- **Converse**: Show that $R^* = H(X)$ is optimal.
  - Weak converse: $\forall R \geq R^*,\ P_e^{(N)} \to 0$
  - Strong converse: $\forall R < R^*,\ P_e^{(N)} \to 1$

3.1 Achievability

Here, we use typical set $\mathcal{T}_\epsilon^n$ to achieve compression rate $R^* = H(X)$. The construction has four steps:

1. Create codebook.
2. Encoding scheme.
3. Decoding scheme.
4. Error analysis.

Now, let's start showing the achievability!

Consider compression rate $R^* = H(X)$ and block size N, we take code length $K = \lceil NR \rceil$.

Create codebook We simply take $\{0,1\}^K$ as our codebook.

Encoding scheme Consider the ϵ-typical set $\mathcal{T}_\epsilon^N$, by property 3 of AEP, we have $|\mathcal{T}_\epsilon^N| \leq 2^{N(H(X)+\epsilon)}$. As ϵ small enough, $|\mathcal{T}_\epsilon^N| \leq 2^K$. That is, we can map each sequence in $\mathcal{T}_\epsilon^N$ to $\{0,1\}^K$ without overlapping. As to the sequence not in $\mathcal{T}_\epsilon^n$, we just simply map them to 0^K.

Decoding sceme Here we simply take the direct decoding function from the encoding scheme and ignore the 0^K case.

Error analysis We can clearly see that the error occurs iff the input source sequence is not in the typical set $\mathcal{T}_\epsilon^n$. Thus, we can upper bound the error probability by the probability for a sequence not being in $\mathcal{T}_\epsilon^n$. That is,

$$P_e^N \leq Pr[X^n \notin \mathcal{T}_\epsilon^n] \to 0 \; as \; N \to \infty$$

Which coincides the lossless criteria.

Intuition (Acievability)

Since typical set has two important AEP:

- $Pr[X^n \in \mathcal{T}_\epsilon^n] \to 1$
- The probability of each sequence to show up is approximately the same.

It suffices to consider only in $\mathcal{T}_\epsilon^N$.

3.2 Converse

4 Source with memory

5 Useful tools

5.1 Fano's inequality

The traditional Fano's inequality is in the following form:

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

Proof: Let $U = \mathbf{1}_{\{X \neq Y\}}$, then we have $P_e = Pr[X \neq Y] = Pr[U = 1]$

$$\begin{aligned} H(X|Y) &= H(X,U|Y) = H(X|U,Y) + H(U) \\ &= Pr[U=1] \times H(X|U=1,Y) + Pr[U=0] \times H(X|U=0,Y) + H_b(P_e) \\ &= P_e \times H(X|U=1,Y) + (1-P_e) \times 0 + H_b(P_e) \\ &\leq P_e \log |\mathcal{X}| + H_b(P_e) \end{aligned}$$

■

Intuition (Fano's inequality)

There are two ways to view Fano's inequality:

1. Lower bound for error probability.
2. Upper bound for conditional entropy.

Note that the bound is tight when the two random variables are close to **independent** since the lost happen in $H(X|U=1,Y) \leq \log|\mathcal{X}|$.

There is variation of Fano's inequality in the form of mutual information as one of X, Y is uniformly distributed:

$$I(X;Y) \geq P_s \log|\mathcal{X}| - H_b(P_s)$$

Proof: Let $V = \mathbf{1}_{\{X=Y\}}$ and without loss of generality assume that X is uniform. Then

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) = \log|\mathcal{X}| - H(X|Y) \\ \text{(by Fano's)} &\geq \log|\mathcal{X}| - P_e \log|\mathcal{X}| - H_b(P_e) \\ &= P_s \log|\mathcal{X}| - H_b(P_s) \end{aligned}$$

■

Intuition (Fano's inequality for mutual information)

There are also two ways to view this variation:

- Upper bound for similarity probability.
- Lower bound for mutual information.

Note that the bound is tight when X and Y are close to independent.

A Proof of AEP's properties

Since we say that x^n is a typical sequence, we have

$$(1-\epsilon)p_X(a) \leq \pi(a|x^n) \leq (1+\epsilon)p_X(a),\ \forall a \in \mathcal{X}$$

$$\begin{aligned} p(x^n) &= \prod_{a\in\mathcal{X}} p_X(a)^{n\pi(a|x^n)} \in [\prod_{a\in\mathcal{X}} p_X(a)^{n(1+\epsilon)p_X(a)}, \prod_{a\in\mathcal{X}} p_X(a)^{n(1-\epsilon)p_X(a)}] \\ &= [2^{\sum_{a\in\mathcal{X}} n(1+\epsilon)p_X(a)\log p_X(a)}, 2^{\sum_{a\in\mathcal{X}} n(1-\epsilon)p_X(a)\log p_X(a)}] \\ &= [2^{-n(1+\epsilon)H(X)}, 2^{-n(1-\epsilon)H(X)}] \end{aligned}$$